**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**



**BÀI BÁO CÁO ĐỒ ÁN**

**Môn: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**Đề tài:**

**DỰ ĐOÁN KHẢ NĂNG TỐT NGHIỆP CỦA SINH VIÊN TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**GVHD: ThS. Nguyễn Thị Anh Thư**

**Nhóm thực hiện: Nhóm 1**

**Sinh viên thực hiện:**

**Trần Tuyết Nhi – 21520383**

**Nguyễn Trọng Ân - 21520548**

**Hoàng Kim Ngọc Anh - 21520560**

**Trần Khánh Duy - 21520787**

Tp. Hồ Chí Minh, tháng 5 năm 2023

**MỤC LỤC**

[**Chương 1: MỞ ĐẦU 3**](#_Toc137219695)

[**1.1. Đặt vấn đề 3**](#_Toc137219696)

[**1.2. Mục tiêu 4**](#_Toc137219697)

[**1.3. Đối tượng và phạm vi nghiên cứu 5**](#_Toc137219698)

[**1.4. Nội dung và phương pháp thực hiện 5**](#_Toc137219699)

[**1.4.1. Xác định câu hỏi nghiên cứu 5**](#_Toc137219700)

[**1.4.2. Tạo bảng dữ liệu từ bộ dữ liệu thô 5**](#_Toc137219701)

[**1.4.3. Làm sạch dữ liệu 6**](#_Toc137219702)

[**1.4.4. Phân tích dữ liệu 6**](#_Toc137219703)

[**1.4.5. Chuẩn hóa dữ liệu và xây dựng các mô hình phân loại để dự đoán kết quả tốt nghiệp của sinh viên. 8**](#_Toc137219704)

[**1.4.6. Giải thích kết quả của các mô hình phân loại 8**](#_Toc137219705)

[**1.4.7. Kết quả dự kiến 8**](#_Toc137219706)

[**1.4.8. Kế hoạch thực hiện: 10 tuần 9**](#_Toc137219707)

[**Chương 2: TỔNG QUAN 10**](#_Toc137219708)

[**2.1. Tình hình nghiên cứu trên thế giới 10**](#_Toc137219709)

[**2.2. Tình hình nghiên cứu trong nước 11**](#_Toc137219710)

[**Chương 3: DỰ ĐOÁN KHẢ NĂNG TỐT NGHIỆP CỦA SINH VIÊN TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN 11**](#_Toc137219711)

[**3.1 . Tìm hiểu dữ liệu 11**](#_Toc137219712)

[**3.2. Phương pháp giải quyết bài toán 21**](#_Toc137219713)

[**3.3. Phân loại nhóm tốt nghiệp sinh viên 22**](#_Toc137219714)

[**3.4. Xử lý dữ liệu thô 24**](#_Toc137219715)

[**3.5. Tiền xử lý dữ liệu 27**](#_Toc137219716)

[**3.5.1. Mô tả chi tiết các bước xử lý trong tiền xử lý dữ liệu 30**](#_Toc137219717)

[**3.6. Xây dựng mô hình 30**](#_Toc137219718)

[**3.6.1 . Chia dữ liệu huấn luyện, dữ liệu kiểm tra 30**](#_Toc137219719)

[**3.6.2. Mã hóa dữ liệu 31**](#_Toc137219720)

[**3.6.3. Chuẩn hóa dữ liệu 31**](#_Toc137219721)

[**3.6.4. Huấn luyện các mô hình dự đoán dựa trên các thuật toán máy học 32**](#_Toc137219722)

[**3.6.4.1. Naïve Bayes 32**](#_Toc137219723)

[**3.6.4.2. Support Vector Machine (SVM) 33**](#_Toc137219724)

[**3.6.4.3. Random Forest 34**](#_Toc137219725)

[**3.6.4.4. Decision Tree 35**](#_Toc137219726)

[**3.6.2. Kiểm định mô hình dự đoán 36**](#_Toc137219727)

[**3.6.3. Các bước còn lại 36**](#_Toc137219728)

[**3.7. Dự đoán thực tế 36**](#_Toc137219729)

[**3.8. Kết quả thực nghiệm 37**](#_Toc137219730)

# **Chương 1: MỞ ĐẦU**

## **1.1. Đặt vấn đề**

Hiện nay, công nghệ thông tin đang phát triển thần tốc và được áp dụng rộng rãi trong hầu hết các lĩnh vực, điều đó góp phần tạo nên một cuộc cách mạng công nghiệp lần thứ tư sôi động. Trong lĩnh vực giáo dục, rất nhiều trường học và trung tâm đã áp dụng công nghệ thông tin để số hóa quá trình giảng dạy của mình nhằm mục đích đưa tri thức đến mọi người một cách nhanh và tiện lợi hơn. Sau khi đợi dịch bệnh lớn bùng phát vào cuối năm 2019, quá trình số hóa này lại càng được đẩy mạnh thông qua việc học trực tuyến. Điều này đồng nghĩa với việc theo thời gian, lượng dữ liệu về giáo dục được lưu trữ tại các trường học tăng lên rất nhiều. Nếu lượng dữ liệu này được xử lý, phân tích, khai phá một cách khoa học thì sẽ góp một phần rất lớn vào việc hỗ trợ học tập cho các bạn học viên cũng như điều chỉnh phương pháp giảng dạy của các thầy cô giáo. Chính vì vậy mà lĩnh vực phân tích và khai phá dữ liệu giáo dục đã và đang thu hút được nhiều sự quan tâm. Khá nhiều vấn đề nhức nhối của giáo dục trước đây đã được giải quyết bằng máy vi tính nhờ các công nghệ khai phá dữ liệu nói riêng và khoa học dữ liệu nói chung.

Hiện nay, tại một số cơ sở đào tạo sau trung học phổ thông như các trường đại học, cao đẳng đang áp dụng cơ chế học tín chỉ. Cơ chế này giúp cho các bạn sinh viên chủ động hơn trong việc lựa chọn thời khóa biểu học tập. Trong cơ chế học tín chỉ này, sinh viên có thể tùy ý lựa chọn đăng ký học nhiều môn học nếu cảm thấy mình đủ khả năng, thời gian để tiếp thu hết các môn đã đăng ký và đăng ký ít môn học nếu cảm thấy mình không có đủ khả năng tiếp thu nhiều môn hoặc không thể dành nhiều thời gian để học nhiều môn. Tuy nhiên trên thực tế thì các bạn sinh viên thường bị lúng túng khi lựa chọn môn học do có nhiều môn được giảng dạy trong môn học kỳ. Khi đó, bên cạnh khả năng tự tìm hiểu thì sinh viên cần đến sự trợ giúp của cố vấn học tập cùng với đội ngũ thầy cô giáo nhà trường để tư vấn giúp giải quyết triệt để tình trạng. Do đó các thầy cô cần được tra cứu thêm kết quả học tập của từng sinh viên, tìm hiểu để tư vấn, đưa ra các lời cảnh báo sớm cho các sinh viên. Với một trường đại học có hàng ngàn sinh viên đang theo học thì việc làm này không hề dễ triển khai. Chính vì vậy đồ án này được thực hiện hướng đến việc khai phá dữ liệu giáo dục, sau đó từ kết quả khai phá này xây dựng nên phần mềm hỗ trợ cho công tác giáo dục.

## **1.2. Mục tiêu**

* Mục tiêu đề tài: Dựa vào dữ liệu tốt nghiệp các khóa trước, dự đoán được khả năng tốt nghiệp của sinh viên (chưa tốt nghiệp / tốt nghiệp) tại trường Đại học Công nghệ Thông Tin – Đại học Quốc gia Thành phố Hồ Chí Minh
* Mục đích nghiên cứu: Tìm ra những yếu tố ảnh hưởng đến khả năng tốt nghiệp của sinh viên Đại học Công nghệ Thông Tin – Đại học Quốc gia Thành phố Hồ Chí Minh để đưa ra lưu ý và phương pháp học tập hiệu quả cho các sinh viên khóa sau.

## **1.3. Đối tượng và phạm vi nghiên cứu**

* Dữ liệu đề tài nghiên cứu giới hạn trong trường đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh gồm:
  + Tập dữ liệu tuyển sinh: dữ liệu của sinh viên trước khi vào trường: Kết quả đầu vào(điểm thi THPTQG, điểm đánh giá năng lực, điểm xét tuyển học bạ, sinh viên được tuyển thẳng/ ưu tiên xét tuyển), các thông tin cá nhân của sinh viên, các chứng chỉ sinh viên đã đạt được(nếu có).
  + Tập dữ liệu kết quả trong tiến trình học: Dữ liệu điểm của các môn học trong từng học kỳ theo từng năm của sinh viên, điểm rèn luyện của sinh viên từng học kỳ, các chứng chỉ, những xử lý học vụ, bảo lưu của sinh viên.

## **1.4. Nội dung và phương pháp thực hiện**

### 1.4.1. Xác định câu hỏi nghiên cứu

Câu hỏi nghiên cứu là phân tích và so sánh khả năng dự đoán tốt nghiệp của các mô hình máy học khác nhau dựa trên dữ liệu của sinh viên. Để xác định câu hỏi nghiên cứu cần dựa vào kết quả quá trình tìm hiểu toàn bộ dữ liệu thô (Education Dataset), từ những thông tin có sẵn để tìm ra câu hỏi nghiên cứu.

### 1.4.2. Tạo bảng dữ liệu từ bộ dữ liệu thô

Chọn ra các thuộc tính cần thiết cho đề tài từ các bảng dữ liệu trong bộ dữ liệu Education Dataset và kết hợp chúng thành một bảng dữ liệu dùng cho đề tài.

Phương pháp thực hiện:

* Pandas dataframe.groupby() được sử dụng để nhóm dữ liệu theo danh mục và áp dụng chức năng cho danh mục nhằm tổng hợp dữ liệu một cách hiệu quả.
* Pandas dataframe.merge() được sử dụng để hợp nhất hai đối tượng DataFrame bằng thao tác nối kiểu cơ sở dữ liệu. Việc nối được thực hiện trên các cột hoặc chỉ mục.

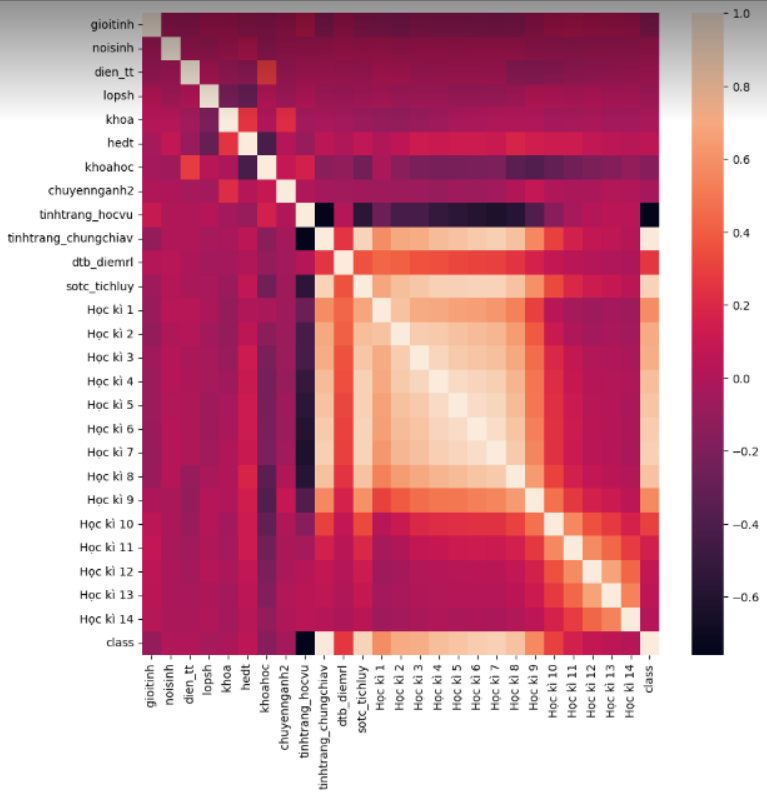
### 1.4.3. Làm sạch dữ liệu

Làm sạch dữ liệu để chuẩn bị cho phân tích. Nội dung này thường liên quan đến việc:

* Xóa dữ liệu trùng lặp và bất thường.
  + Sử dụng pandas dataframe.duplicated().sum() để kiểm tra và thống kê các giá trị trùng lặp trong bảng dữ liệu.
  + Sử dụng pandas dataframe.drop\_duplicates() để xóa các trùng lặp
  + Sử dụng pandas dataframe.isna().sum() để kiểm tra và thống kê các giá trị thiếu trong bảng dữ liệu.
  + Sử dụng pandas dataframe.dropna(how="all") để xóa các hàng dữ liệu thiếu tất cả các cột.
* Điều chỉnh sự không nhất quán:
  + Sử dụng pandas dataframe.replace() để thay thế giá trị được chỉ định bằng một giá trị được chỉ định khác. Phương thức này sẽ tìm kiếm trên toàn bộ Data Frame và thay thế mọi trường hợp của giá trị đã chỉ định.
* Chuẩn hóa cấu trúc và định dạng dữ liệu
  + Sử dụng Label Encoding để chuyển đổi các cột dữ liệu kiểu object thành các cột số để phục vụ cho việc trực quan hoá dữ liệu và các mô hình học máy chỉ lấy dữ liệu số.
* Xử lý khoảng trắng và các lỗi cú pháp khác.

### 1.4.4. Phân tích dữ liệu

Sau khi trực quan hóa dữ liệu ta có ma trận sau:

****

Từ ma trận trên cho thấy rằng:

* Các giá trị trong bảng đều có tỉ lệ với nhau, từ màu nhạt dần biểu thị tỉ lệ thuận đến màu đen biểu thị tỉ lệ nghịch
* Đối với class có hai giá trị là “tốt nghiệp” và “không tốt nghiệp” tương ứng là 1 và 0 thì có những thuộc tính tương quan mạnh yếu xuất hiện chẳng hạn như:
* Thuộc tính gioitinh, noisinh,dien\_tt,lopsh,khoa,hedt,khoahoc có ít ảnh hưởng đến hai nhãn của bài toán phân loại.
* Các cột “ Học kì 11 ” đến “ Học kì 14 ” do ít sinh viên nào học đến những học kỳ đó nên dữ liệu ít phản ánh đến các lớp.
* Thuộc tính tinhtrang\_hocvu tỉ lệ nghịch với khả năng tốt nghiệp của sinh viên ( ví dụ: tinhtrang\_hocvu=0 là không vi phạm, tinhtrang\_hocvu=5 là bị buộc thôi học)
* Thuộc tính tinhtrang\_chungchiav tỉ lệ thuận với khả năng tốt nghiệp của sinh viên do nếu tinhtrang\_chungchiav =1 tức là sinh viên đã đạt chuẩn anh văn đầu ra và ngược lại, từ đó tỉ lệ tốt nghiệp cao.
* Từ những thuộc tính sotc\_tichluy cho đến thuộc tính “ Học kì 10 ” cho thấy rằng các số tín chỉ tích lũy hay điểm của sinh viên trong những học kỳ đó nếu cao thì tăng tỉ lệ tốt nghiệp.

### 1.4.5. Chuẩn hóa dữ liệu và xây dựng các mô hình phân loại để dự đoán kết quả tốt nghiệp của sinh viên.

* Sử dụng sklearn sklearn preprocessing.StandardScaler() để chuẩn hóa dữ liệu.
* Sử dụng các mô hình phân loại như:
* Decision Tree
* Naïve Bayes
* SVM(Support Vector Machine)
* Random Forest

### 1.4.6. Giải thích kết quả của các mô hình phân loại

Sử dụng các độ đo đánh giá để đánh giá kết quả của các mô hình dự đoán:

* Accuracy: Tỷ lệ các trường hợp được dự báo đúng trên tổng số các trường hợp là bao nhiêu
* Confusion Matrix: Tổng kết giữa nhãn thực tế của dữ liệu với nhãn do mô hình dự đoán
* F1 score: Là số dung hòa Recall và Precision giúp ta có căn cứ để lựa chọn model

### 1.4.7. Kết quả dự kiến

* Nắm được các phương pháp khoa học dữ liệu khác nhau để giải quyết bài toán trong giáo dục.
* Áp dụng được các phương pháp tiền xử lý đặc trưng cho dữ liệu giáo dục.
* Xây dựng được mô hình máy học để dự đoán sinh viên có tốt nghiệp hay không, tiến hành so sánh các mô hình và tối ưu các mô hình(nếu có).
* Biết cách sử dụng nhuần nhuyễn các mô hình để đưa ra những nhận định hay cải tiến(nếu có).
* Dự báo từ sớm các vấn đề từ đó đưa ra các cảnh báo kịp thời dành cho sinh viên.

### 1.4.8. Kế hoạch thực hiện: 10 tuần

|  |  |  |
| --- | --- | --- |
| Thời gian | Nội dung | Công việc |
| 21/03/2023 | Nhận thông tin về báo cáo đồ án môn học Khai thác dữ liệu và ứng dụng |  |
| 23/03/2023 | Họp nhóm bàn về bài tập 1 của báo cáo đồ án | Chọn đề tài, tìm hiểu dữ liệu và phân tích vấn đề |
| 30/04/2023 | Cá nhân nộp báo cáo từng phần cho bài tập 1 của báo cáo đồ án |  |
| 03/05/2023 | Báo cáo bài tập 1 và nhận sự hướng dẫn từ giảng viên |  |
| 04/05/2023 | Nhận dữ liệu thực hành mới |  |
| 05/05/2023 | Họp nhóm thực hiện lại bài tập 1 | Tìm hiểu dữ liệu và phân tích vấn đề |
| 06/05/2023 | Họp nhóm thực hiện bài tập 2 của báo cáo đồ án | Tìm hiểu về thuật toán sử dụng, trường dữ liệu và sử dụng dữ liệu mẫu |
| 13/05/2023 | Họp lần 2 thực hiện bài tập 2 | Thay đổi mục tiêu đề tài, tối ưu hóa và rút gọn dữ liệu, lên kế hoạch cuối cho báo cáo đồ án |
| 16/05/2023 | Hoàn thành mô tả đồ án đăng kí | Chuẩn bị nội dung file word về báo cáo đồ án |

# **Chương 2: TỔNG QUAN**

## **2.1. Tình hình nghiên cứu trên thế giới**

Trong những năm gần đây, xu hướng áp dụng khai phá dữ liệu – khoa học dữ liệu vào lĩnh vực giáo dục đang phát triển mạnh mẽ. Cụ thể là từ năm 2008, lĩnh vực khai phá dữ liệu giáo dục đã đón chào sự ra đời của hội thảo khoa học chuyên ngành Khai Phá Dữ Liệu Giáo dục (International Conference on Data Mining) và sau đó là sự ra đời của tạp chí chuyên ngành Khai phá Dữ Liệu Giáo dục (International Journal on Data Mining) [1 ]. Cùng với đó là các công trình khảo sát tổng quan về lĩnh vực này cũng được nghiên cứu và công bố rộng rãi [ 2], [ 3], [4 ], [5 ]. Cụ thể, việc áp dụng data mining để dự đoán khả năng tốt nghiệp của sinh viên đang trở thành một lĩnh vực nghiên cứu hứa hẹn, với mục tiêu tăng cường hiệu suất học tập và cải thiện quản lý giáo dục. Trong năm 2020, Tác giả Francesca Del Bonifro cùng các cộng sự của mình đã thực hiện dự đoán sinh viên không tốt nghiệp trong nghiên cứu “Student Dropout prediction” và đã cho ra kết quả độ đo Accuracy cao nhất là 87% với thuật toán Random Forest [ ]. Trước đó, tác giả Lovenoor Aulck cùng các cộng sự đã thực hiện đoán việc thôi học của sinh viên trong công trình “Predicting Student Dropout in Higher Education” tuy không đạt được hiệu quả cao về việc dự đoán (accuracy cao nhất là 0.6659) những công trình đã phát hiện được điểm môn Tâm lý học ảnh hưởng đến hơn 53% khả năng bỏ học của sinh viên [ ]. Ngoài ra cũng có một đề tài nghiên cứu có liên quan đến hiệu suất học tập của sinh viên đó là “Adaptive recommendation system using machine learning algorithms for predicting student’s best academic program” [ ]. trong công trình này, tác giả Mohamed Ezz cùng các cộng sự đã dùng hệ khuyến nghị để tìm ra chương trình học tập giúp sinh viên đạt được hiệu suất học tập tốt nhất và tránh phải thôi học.

## **2.2. Tình hình nghiên cứu trong nước**

Ở trong nước, tác giả Nguyễn Hữu Phùng [ ] cùng các cộng sự đã thực hiện phân lớp tình trạng học tập dựa trên điểm số của từng học kỳ, tại đây nhóm tác giả đã sử dụng 6 thuật toán gồm Naive Bayes, Neural Network, Support Vector Machine (SVM), K-Nearest Neighbors (kNN), Decision Tree (C45), Random Forest và cho ra kết quả khả quan với việc áp dụng phương pháp cân bằng dữ liệu SMOTE [ ] với thuật toán Random Forest cho ra độ chính xác là gần 81%. Các nghiên cứu khác về cùng vấn đề chủ yếu tiếp cận bằng phương pháp thống kê kinh điển hoặc ở mức ứng dụng ban đầu của machine learning để khai phá dữ liệu.

# **Chương 3: DỰ ĐOÁN KHẢ NĂNG TỐT NGHIỆP CỦA SINH VIÊN TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

## **3.1 . Tìm hiểu dữ liệu**

Dữ liệu sử dụng cho bài toán dự đoán

Nhóm chúng em xin được cảm ơn thầy cô cho phép chúng em sử dụng bộ dữ liệu này.

Dữ liệu dùng cho bài toán này bao gồm gồm 8 bảng dữ liệu từ bộ dữ liệu Education Dataset. Các dữ liệu này bao gồm từ khoá 8 (2013) đến khoá 14 (2019) với tất cả các dữ liệu của các sinh viên từ thời điểm năm 2013-2014 cho đến hết học kỳ 2 năm học 2019-2020 cùng với đó là một vài dữ liệu đầu học kỳ 1 năm học 2020-2021.

Chi tiết là tại:

**Bảng thứ nhất (SINHVIEN)** bao gồm các thông tin chung về lý lịch sinh viên. Trong đó có 13 cột tương ứng với 13 loại thông tin của sinh viên và 8295 dòng ứng với 8295 sinh viên trong dữ liệu đang xét.

Chi tiết 13 cột đó là:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| id | Mã để phân biệt từng dòng trong bảng | object |  |
| masv\_tham so dau | 4 số đầu của mã số sinh viên |  |  |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| namsinh | Năm sinh của sinh viên | object |  |
| gioitinh | Giới tính của sinh viên | object | Có 2 giá trị:  - 1: Nam  - 0: Nữ |
| noisinh | Nơi sinh của sinh viên | object |  |
| lopsh | Lớp sinh hoạt của mỗi sinh viên | object |  |
| khoa | Khoa của sinh viên theo học | object |  |
| hedt | Hệ đào tạo sinh viên theo học | object |  |
| khoahoc | Khoá học thứ | object |  |
| chuyennganh2 | Mã chuyên ngành của sinh viên | object |  |
| tinhtrang | Tình trạng của từng sinh viên | object |  |
| diachi\_tinhtp | Địa chỉ của sinh viên | object | Địa chỉ thường trú = diachi\_thuongtru + diachi\_phuongxa + diachi\_quanhuyen + diachi\_tinhtp |

**Bảng thứ hai (DIEM\_THU)** bao gồm các thông tin liên quan đến các điểm đạt được của sinh viên ứng với từng môn học. Trong đó có 14 cột ứng với 14 loại thông tin về môn học, mã lớp của môn học, điểm số tương ứng của mỗi bạn sinh viên. Trong bảng này có 674274 dòng dữ liệu.

Chi tiết 14 cột đó là:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| mamh | Mã định danh cho từng môn học trong trường | object |  |
| malop | Mã định danh cho từng lớp học trong trường | object |  |
| sotc | Số tín chỉ phải học của từng môn học | int64 |  |
| hocky | Học kỳ mà môn học được mở | int64 |  |
| namhoc | Năm học mà môn học được mở | int64 |  |
| diem\_qt | Điểm được lấy trong quá trình học của sinh viên | float64 |  |
| diem\_th | Điểm thực hành môn học của sinh viên | float64 |  |
| diem\_gk | Điểm thi giữa kỳ môn học của sinh viên | float64 |  |
| diem\_ck | Điểm thi cuối kỳ môn học của sinh viên | float64 |  |
| diem\_hp | Điểm tổng kết môn học của sinh viên | float64 |  |
| trangthai | Biểu thị trạng thái qua môn của sinh viên | int64 | 0: hủy  1: bình thường  2: trả nợ  3: cải thiện  4: Miễn  5: Hoãn |
| tinhtrang |  | int64 | Có 5 giá trị: 0, 2, 3, 1, -1 |
| mamh\_tt | Mã định danh trước đó của môn học | object |  |

**Bảng thứ 3 (SINHVIEN\_CHUNGCHI)** bao gồm các thông tin liên quan tình trạng nộp chứng chỉ anh văn của sinh viên. Trong đó có 14 cột tương ứng với 14 loại thông tin về chứng chỉ anh văn của sinh viên. Trong bảng này có 3401 dòng ứng với 3401 sinh viên đã nộp chứng chỉ anh văn.

|  |  |  |
| --- | --- | --- |
| Data Field | Description | Data Type |
| id | Số thứ tự trong bảng dữ liệu, tăng dần | object |
| masv | Mã sinh viên, mỗi sinh viên chỉ có một mã sinh viên duy nhất và không trùng nhau | Object |
| ngaythi | Ngày thi chứng chỉ được nộp | Object |
| url | Địa chỉ dẫn tới hình ảnh của bằng được nộp trên trang hệ thống | Object |
| loaixn | Loại chứng chỉ của thí sinh nộp lên trên hệ thống | Object |
| url | Địa chỉ dẫn tới hình ảnh của bằng được nộp trên trang hệ thống | Float64 |
| loaixn | Loại chứng chỉ của thí sinh nộp lên trên hệ thống | Float64 |
| Listening | Điểm bài thi nghe dựa trên chứng chỉ tương ứng | Float64 |
| Speaking | Điểm bài thi nói dựa trên chứng chỉ tương ứng | Float64 |
| Reading | Điểm bài thi đọc dựa trên chứng chỉ tương ứng | Object |
| Writing | Điểm bài thi viết dựa trên chứng chỉ tương ứng | Object |
| Tongdiem | Tổng điểm các bài thi dựa trên chứng chỉ tương ứng | Object |
| Lydo | Lý do chứng chỉ không được chấp nhận | object |
| Trangthai | Thời gian cuối cùng hệ thống ghi nhận sinh viên nộp chứng chỉ | Object |
| ngayxl | Ngày xếp loại chứng chỉ của sinh viên | Float64 |

**Bảng thứ tư (THISINH)** bao gồm các thông tin liên quan đến trạng thái trúng tuyển vào trường của sinh viên. Bảng có 6 cột ứng với 6 loại thông tin trúng tuyển của sinh viên, gồm 8235 dòng dữ liệu ứng với 8235 sinh viên.

Chi tiết 6 cột trong bảng dữ liệu là:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| dien\_tt | Diện trúng tuyển là phương thức sinh viên chọn để trúng tuyển vào trường | object |  |
| diem\_tt | Điểm trúng tuyển xét theo diện trúng tuyển vào trường của thí sinh | float64 |  |
| lop12\_matinh | Tỉnh thành sinh viên học lớp 12 | object |  |
| lop12\_matruong | Mã trường THPT nơi sinh viên theo học lớp 12 | float64 |  |
| TEN\_TRUONG | Tên trường THPT sinh viên đã theo học | object |  |

**Bảng thứ năm (XLHV)** bao gồm các thông tin liên quan đến trạng thái xử lý học vụ của sinh viên. Trong đó có 8 cột là 8 loại thông tin xử lý học vụ của sinh viên, có 3453 dòng ứng với 3453 việc xử lý học vụ đã xảy ra.

Chi tiết bộ dữ liệu như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| Id | Mã số định danh cho mỗi quyết định | object |  |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| tinhtrang | Tình trạng của từng sinh viên | float64 | Có 5 giá trị:  - 2: Cảnh cáo  - 5: Thôi học  - 7: Gia hạn  - 8: Tự do |
| lydo | Lý do bị xử lý học vụ | object |  |
| hocky | Học kỳ bị xử lý học vụ | object |  |
| namhoc | Năm học bị xử lý học vụ | object |  |
| soqd | Số quyết định | object |  |
| ngayqd | Ngày đưa ra quyết định | object |  |

**Bảng thứ sáu (DIEMRL)** bao gồm các thông tin liên quan đến điểm rèn luyện trong quá trình học tại trường của sinh viên. Trong đó có 7 cột là 7 loại thông tin điểm rèn luyện của sinh viên, có 111979 dòng.

Chi tiết bộ dữ liệu như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| id | Mã để phân biệt từng dòng trong bảng | object |  |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| lopsh | Lớp sinh hoạt của mỗi sinh viên | object |  |
| hocky | Học kỳ tính điểm rèn luyện | float64 |  |
| namhoc | Năm học tính điểm rèn luyện | float64 |  |
| drl | Giá trị điểm rèn luyện của mỗi sinh viên | object |  |
| ghichu | Ghi chú thêm những gì cần chú ý | object |  |

**Bảng thứ bảy (SINHVIEN\_DTB\_TOANKHOA)** bao gồm các thông tin liên quan đến điểm trung bình toàn khóa học của sinh viên. Trong đó có 4 cột là 4 loại thông tin điểm trung bình toàn khóa của sinh viên, có 13971 dòng ứng với 13971 sinh viên.

Chi tiết bộ dữ liệu như sau:

|  |  |  |
| --- | --- | --- |
| Data Field | Description | Data Type |
| masv | Mã sinh viên, mỗi sinh viên chỉ có một mã sinh viên duy nhất và không trùng nhau | Object |
| dtb\_toankhoa | Là điểm trung bình của sinh viên toàn khóa học trên số tín chỉ kể cả các môn không đạt | Float64 |
| dtb\_tichluy | Là điểm trung bình của sinh viên toàn khóa học trên số tín chỉ nhưng chỉ tính các môn đạt | Float64 |
| sotc\_tichluy | Là số tín chỉ của sinh viên tích lũy toàn khóa học chỉ tính những môn đạt | Float64 |

**Bảng thứ tám (TOTNGHIEP)** bao gồm các thông tin liên quan đến trạng thái tốt nghiệp của sinh viên. Trong đó có 5 cột là 5 loại thông tin tốt nghiệp của sinh viên, có 1848 dòng ứng với 1848 sinh viên.

Chi tiết bộ dữ liệu như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Data Field | Description | Data Type | Ghi chú |
| id | Mã để phân biệt từng dòng trong bảng | object |  |
| masv | Mã số định danh được cấp cho từng sinh viên trong trường | object |  |
| xeploai | Xếp loại tốt nghiệp của sinh viên | object |  |
| soquyetdinh | Số của quyết định cho sinh viên tốt nghiệp | object |  |
| ngaycapvb | Ngày cấp văn bằng | object |  |

## **3.2. Phương pháp giải quyết bài toán**

Nhóm thực hiện giải quyết bài toán “Dự đoán khả năng tốt nghiệp của sinh viên Trường Đại học Công nghệ Thông tin – ĐHQG TP.Hồ Chí Minh” theo hướng data mining, cụ thể ở đây là áp dụng kỹ thuật phân lớp trong máy học để dự đoán sinh viên “tốt nghiệp” hoặc “không tốt nghiệp”. Và các cột thuộc tính trong dữ liệu sử dụng cho bài toán phân lớp chính là các cột thông tin về lý lịch, điểm học tập … của các bạn sinh viên từ khoá 8 đến khoá 14. Chúng em sẽ tiến hành lập trình các giải thuật xử lý dữ liệu trong lĩnh vực khai phá dữ liệu để biến các bảng dữ liệu thô nêu trên thành một bảng dữ liệu (data) theo ý muốn. Trong bảng (data) này sẽ có một cột dữ liệu là cột nhãn (‘class’) để dự đoán, hay có thể gọi là cột nhãn để phân lớp trong bài toán phân lớp. Các cột còn lại sẽ được chọn từ các bảng dữ liệu thô để làm các thuộc tính huấn luyện cho mô hình dự đoán. Sẽ lần lượt có các bảng dữ liệu theo thời gian từ bảng có đầy đủ các thuộc tính và bảng có các thuộc tính theo từng học kỳ (từ vừa vào trường đến học kỳ 2 năm 2) theo 2 cách xử lý thông tin khác nhau. Lý do cho việc lấy dữ liệu theo thời gian như vậy là bởi vì trong thực tế chúng ta cần dự đoán chính xác từ sớm và cập nhật theo thời gian để nhà trường và thầy cô có thể chú ý và đưa ra những biện pháp phù hợp để đảm bảo khả năng tốt nghiệp của sinh viên.

## **3.3. Phân loại nhóm tốt nghiệp sinh viên**

Khả năng tốt nghiệp của sinh viên: tốt nghiệp hoặc không tốt nghiệp là các nhãn trong bài toán phân lớp này. Như vậy, trong mục này nhóm sẽ tập trung trình bày cách tạo ra các nhãn trong bài toán dự đoán kết quả tốt nghiệp của sinh viên trường Đại học Công nghệ Thông tin - ĐHQG.

Để có được những nhãn cho bài toán phân lớp này, nhóm thực hiện như sau:

Tính toán số lượng sinh viên có dữ liệu tốt nghiệp trên tổng số sinh viên của toàn trường (SINHVIEN) là 22.24%. Chi tiết số lượng loại tốt nghiệp của sinh viên như sau:

|  |  |
| --- | --- |
| **Xếp loại** | **Số lượng sinh viên** |
| Khá | 1125 |
| Giỏi | 447 |
| Xuất sắc | 7 |
| Trung bình Khá | 266 |
| NaN (Không có dữ liệu tốt nghiệp) | 6450 |
| Tổng số | 8295 |

Giá trị ‘NaN’ biểu thị hai ý nghĩa: một là, sinh viên chưa tốt nghiệp do vẫn đang hoàn thành chương trình đào tạo. Hai là, sinh viên vì một số lý do đã bỏ học hoặc bỏ bằng tốt nghiệp. Do vậy cần gán nhãn lại cho các dữ liệu ‘NaN’. Để tăng thêm kích thước dữ liệu, nhóm sẽ thực hiện gán nhãn thêm cho các sinh viên mà xếp loại tốt nghiệp là ‘NaN’ nhưng đã đủ điều kiện tốt nghiệp (những sinh viên này chưa đăng ký tốt nghiệp).

Để đơn giản hoá đề tài và tăng độ chính xác cho các mô hình phân loại nhóm thực hiện bài toán dự đoán khả năng tốt nghiệp với hai lớp: tốt nghiệp và không tốt nghiệp. Trong đó, nhãn tốt nghiệp là những sinh viên đã có kết quả tốt nghiệp hoặc những sinh viên đủ điều kiện tốt nghiệp nhưng chưa đăng ký xét tốt nghiệp. Nhãn không tốt nghiệp được gán cho những trường hợp sau: sinh viên bị nhà trường buộc thôi học, sinh viên tự nghỉ học, sinh viên bỏ bằng (thường là những học kỳ đầu có điểm học tập rất cao nhưng những học kỳ cuối điểm thấp do đã đi làm, hoặc không đủ điều kiện tốt nghiệp do không nộp bằng anh văn và điểm rèn luyện thấp).

Khi thực hiện phân loại dữ liệu theo hai nhóm lớn: tốt nghiệp và không tốt nghiệp, nhóm có bảng thống kê 2 nhóm như sau:

|  |  |  |
| --- | --- | --- |
| **Tên nhóm** | **Số lượng sinh viên** | **Tỷ lệ** |
| Tốt nghiệp | 2190 | 65% |
| Không tốt nghiệp | 1179 | 35% |
| Tổng số | 3369 | 100% |

Khi thực hiện gán nhãn và tạo ra bộ dữ liệu để huấn luyện theo hai nhãn trên, kết quả dự đoán tốt và bộ dữ liệu có kích thước khá ổn với tỷ lệ 2 class tương đối cân bằng. Vì vậy nhóm sẽ chọn hai nhãn này làm nhãn cho bài toán phân loại khả năng tốt nghiệp của sinh viên.

|  |  |  |
| --- | --- | --- |
| **Tên nhóm** | **Mô tả** | **Nhãn trong bài toán phân lớp** |
| Tốt nghiệp | 1. Những sinh viên có kết quả tốt nghiệp  2. Những sinh viên đủ điều kiện tốt nghiệp bao gồm các điều kiện:   * Đủ 138 tín chỉ tích luỹ * Đã có chứng chỉ anh văn * Điểm rèn luyện >= 50 | 1 |
| Không tốt nghiệp | Những sinh viên có một trong các dấu hiệu sau:   * Bị buộc thôi học (tình trạng học vụ là 5) * Điểm trung bình điểm rèn luyện là ‘NaN’ (biểu thị học sinh tự nghỉ học) hoặc bé hơn 50 (không đủ điều kiện tốt nghiệp) * Số tín chỉ tích lũy bằng 0 | 0 |

## **3.4. Xử lý dữ liệu thô**

Trong phần này, nhóm sẽ trình bày chi tiết phần xử lý dữ liệu thô. Mục tiêu của xử lý dữ liệu thô là tạo ra bảng dữ liệu chính cho đề tài bao gồm các cột thuộc tính và một cột nhãn.

Để áp dụng phân lớp máy học vào bảng dữ liệu này thì mỗi dòng trong bảng chính là thông tin của 1 sinh viên và mỗi cột chính là các thuộc tính liên quan đến bạn sinh viên đó như: giới tính, nơi sinh,... Trong đó cột cuối cùng của bảng chính là cột nhãn (class). Danh sách các bảng dữ liệu dùng cho đề tài đã được nhóm giới thiệu ở những phần trước. Bây giờ việc cần làm là xác định các cột thuộc tính nào sẽ có mặt trong bảng dữ liệu.

Để phù hợp với mục tiêu dự đoán khả năng tốt nghiệp của sinh viên, đặc biệt là ở các giai đoạn đầu khi mới vào trường thì nhóm xác định các cột thuộc tính cần có của bảng dữ liệu chính như sau:

|  |  |
| --- | --- |
| **Loại thông tin** | **Các cột thuộc tính cần có trong bảng dữ liệu** |
| Thông tin cơ bản về lý lịch sinh viên | Các cột thuộc tính này được lấy từ bảng SINHVIEN và sẽ cung cấp cho các mô hình máy học những thông tin bổ ích về hoàn cảnh của sinh viên. Các cột ấy là:   1. Cột mssv 2. Cột gioitinh 3. Cột noisinh 4. Cột lopsh 5. Cột khoa 6. Cột hedt 7. Cột khoahoc 8. Cột chuyennganh2 9. Cột xeploai (class) |
| Thông tin cơ bản về tình trạng bị xử lý học vụ của sinh viên | Cột thuộc tính ‘tinhtrang’ trong bảng XLHV sẽ cung cấp thông tin sinh viên đó có bị xử lý học vụ lần nào chưa và mức độ xử lý học vụ. Thông tin này ảnh hưởng mạnh đến kết quả bài toán phân lớp. |
| Thông tin cơ bản của sinh viên khi trúng tuyển vào trường | Các cột thuộc tính này được lấy từ bảng THISINH và sẽ cung cấp cho các mô hình máy học những thông tin bổ ích về nền tảng học tập cơ bản khi sinh viên mới trúng tuyển vào trường. Các cột ấy là:   1. Cột dien\_tt 2. Cột diem\_tt |
| Thông tin cơ bản về điều kiện anh văn đầu ra | Cột thuộc tính ‘url’ trong bảng SINHVIEN\_CHUNGCHI sẽ cung cấp thông tin sinh viên đó có đạt chuẩn đầu ra anh văn hay không thông qua link minh chứng bằng anh văn nộp về cho trường. |
| Thông tin cơ bản về điểm rèn luyện trung bình của các học kỳ của sinh viên | Cột thuộc tính ‘drl’ trong bảng DIEMRL thông qua các bước xử lý dữ liệu sẽ cung cấp thông tin điểm rèn luyện trung bình các kỳ học của sinh viên đó từ đó phản ánh thái độ học tập và tham gia các hoạt động trong trường của sinh viên. |
| Thông tin cơ bản về số tín chỉ tích lũy của sinh viên | Cột ‘sotc\_tichluy’ trong bảng DTB\_TOANKHOA sẽ cung cấp thông tin sinh viên đó có đạt chuẩn số tín chỉ cần thiết để tốt nghiệp hay không. |
| Thông tin cơ bản về điểm học tập trung bình của từng học kỳ của sinh viên | Từ bảng DIEMTHU nhóm thực hiện tính toán điểm trung bình từng học kỳ cho mỗi sinh viên. Do có những sinh viên học nhiều học kỳ mới đủ điều kiện tốt nghiệp nên nhóm sẽ tính điểm cho 14 học kỳ để đảm bảo tính chính xác cao hơn. Những sinh viên tốt nghiệp trước 14 học kỳ sẽ được điền dữ liệu thiếu bằng các cách được trình bày ở các phần sau.   1. Học kì 1 2. Học kì 2 3. Học kì 3 4. Học kì 4 5. Học kì 5 6. Học kì 6 7. Học kì 7 8. Học kì 8 9. Học kì 9 10. Học kì 10 11. Học kì 11 12. Học kì 12 13. Học kì 13 14. Học kì 14 |

## **3.5. Tiền xử lý dữ liệu**

|  |  |
| --- | --- |
| Phương thức xử lý | Mô tả |
| Chọn những thuộc tính cần thiết từ 2 bảng sinhvien.csv và totnghiep.csv | Để tạo ra các cột thuộc tính của bảng (data) như đã nên trên thì nhóm cần chọn ra những dữ liệu:  Ánh xạ 2 bảng sinhvien.csv và totnghiep.csv theo mssv, nhóm chọn ra những cột: ‘mssv’, ‘gioitinh’, ‘noisinh’, ‘lopsh’, ‘khoa’, ‘hedt’, ‘khoahoc’, ‘chuyennganh2’, ‘xeploai’.  Do dữ liệu của 2 bảng không tương đương nên có 42 học sinh bị thiếu dữ liệu, sau khi drop dữ liệu bị thiếu giữ lại được 8295 dòng, trong đó có 6450 sinh viên chưa được xếp loại, đây là những sinh viên không có trong bảng totnghiep.csv |
| Đồng nhất dữ liệu ‘noisinh’ | Có hiện tượng không nhất quán trong cột dữ liệu ‘noisinh’ và có đến 231 nơi sinh khác nhau. Sau khi đồng nhất dữ liệu thì cột ‘noisinh’ còn 68 giá trị unique ( Mô tả chi tiết cách xử lý tại mục 3.5.1 ) |
| Chọn thuộc tính cần thiết từ bảng XLHV.csv | Kết dữ liệu theo ‘mssv’  Lấy thông tin từ cột ‘tinhtrang’ trong bảng XLHV.csv và đổi tên thành ‘tinhtrang\_hocvu’ thể hiện tình trạng xử lý học vụ của sinh viên. Đổi tên cột ‘xeploai’ thành ‘class’ |
| Đồng nhất dữ liệu ‘tinhtrang\_hocvu’ | Các sinh viên không bị xử lý học vụ sẽ không có trong bảng XLHV.csv, nhóm chọn cách xử lý điền cho các sinh viên đó giá trị 0 ( Mô tả chi tiết cách xử lý tại mục 3.5.1 ) |
| Chọn thuộc tính cần thiết từ bảng thisinh.csv | Kết dữ liệu theo ‘mssv’  Trong bảng thisinh.csv, chọn ra 2 thuộc tính là ‘dientt’ và ‘diemtt’. |
| Đồng nhất dữ liệu ‘dientt’ | Những giá trị bị thiếu trong cột ‘dientt’ sẽ được điền bằng “Không xác định”, sau đó được thay bằng giá trị có số lượng nhiều nhất là “THPT” ( Mô tả chi tiết cách xử lý tại mục 3.5.1) |
| Xử lý ‘diemtt’ | - Chuẩn hoá điểm trúng tuyển về thang 30 cho những sinh viên có điểm ĐGNL  - Điền tất cả điểm thi ĐGNL bị thiếu bằng điểm trung bình ĐGNL  - Điền tất cả những điểm bị thiếu còn lại của các diện ưu tiên, tuyển thẳng… bằng -1 thể hiện không có thông tin |
| Chọn thuộc tính cần thiết từ bảng sinhvien\_chungchi.csv | Kết dữ liệu theo ‘mssv’  Chọn cột ‘url’  Đổi tên cột ‘url’ thành ‘tinhtrang\_chungchiav’, |
| Xử lý dữ liệu ‘tinhtrang\_chungchiav’ | Xử lý những thông tin sinh viên bị trùng lặp do nộp nhiều chứng chỉ bằng cách xoá dữ liệu trùng lặp và giữ lại dữ liệu trễ nhất  Những giá trị bị thiếu điền thành giá trị 0 thể hiện sinh viên chưa có chứng chỉ, còn lại điền 1 cho những sinh viên đã có chứng chỉ anh văn  Điều chỉnh dữ liệu không nhất quán, có những sinh viên đã tốt nghiệp nhưng ‘tinhtrang\_chungchiav’ có giá trị 0, điền bằng giá trị 1 cho những sinh viên này |
| Chọn thuộc tính cần thiết từ bảng diemrl.csv | Kết dữ liệu theo ‘mssv’  Chọn cột ‘drl’ đổi tên thành ‘dtb\_diemrl’ |
| Tính trung bình điểm rèn luyện của sinh viên | Nhóm dữ liệu theo mssv và tính trung bình điểm rèn luyện cho từng sinh viên, có 69 sinh viên bị thiếu dữ liệu drl, nhóm chọn giữ lại để phân lớp sinh viên ( Mô tả chi tiết cách xử lý tại mục 3.5.1 ) |
| Chọn thuộc tính cần thiết từ bảng sinhvien\_dtb\_toankhoa.csv | Kết dữ liệu theo ‘mssv’  Lấy cột ‘sotc\_tichluy’ |
| Xử lý dữ liệu ‘sotc\_tichluy’ | Điền giá trị -100 cho những sinh viên bị thiếu dữ liệu  Đồng nhất dữ liệu: Những sinh viên có kết quả tốt nghiệp nhưng ‘sotc\_tichluy’ = -100, nhóm thay thế bằng giá trị 148 (giá trị trung bình số tín chỉ tốt nghiệp của các khoa tính từ khoá 7) ( Mô tả chi tiết cách xử lý tại mục 3.5.1 ) |
| Chọn thuộc tính cần thiết trong bảng diem\_Thu.csv | Bỏ các thuộc tính không cần thiết: ‘mamh’, ‘malop’, ‘diem\_qt’, ‘diem\_th’, ‘diem\_gk’, ‘diem\_ck’, ‘tinhtrang’, ‘mamh\_tt’  Điền lần lượt giá các giá trị trung bình theo từng trạng thái của môn học cho các sinh viên có ‘diem\_hp’ trống dựa trên ‘trangthai’  Bỏ đi các môn có ‘trangthai’ là 0, 4, 5  Nhóm các môn lại thành các cột biểu diễn điểm trung bình và số tín chỉ sinh viên đó đạt được trong các học kỳ  Lấy điểm trung bình và số tín chỉ của sinh viên theo từng học kỳ từ 1 đến 14 |
| Gán nhãn cho sinh viên | Gán nhãn không tốt nghiệp cho những sinh viên có 1 trong 3 điều kiện sau:  - ‘tinhtrang\_hocvu’ = 5  - ‘dtb\_diemrl’ = NaN hoặc < 50  - ‘sotc\_tichluy’ <=0  Gán nhãn tốt nghiệp cho những sinh viên có đủ 3 điều kiện sau:  - ‘sotc\_tichluy’ >= 138  - ‘tinhtrang\_chungchiav’ = 1  - ‘dtb\_diemrl’ >= 50 |
| Hoàn thành bảng dữ liệu | Xoá những sinh viên chưa được gán nhãn  Đồng bộ hoá dữ liệu ‘class’, gán nhãn ‘tốt nghiệp’ cho những sinh viên có kết quả tốt nghiệp ‘xuất sắc’ , ‘giỏi’ , ‘khá’, ‘trung bình khá’ |

Link tiền xử lý dữ liệu: [link](https://colab.research.google.com/drive/1Y-Bll12UoXUlF9QbGriWnqktOL8BkKNd?usp=sharing)

### 3.5.1. Mô tả chi tiết các bước xử lý trong tiền xử lý dữ liệu

* Đồng nhất dữ liệu ‘noisinh’:
  + Từ bảng sinh viên ta có thể kiểm tra, truy xuất được nơi sinh của sinh viên nhưng xảy ra tình huống các tên thành phố hay tỉnh thành được nhập vào không đồng nhất, có giá trị null ở trong bảng sinh viên hay giá trị rỗng. Ta rà soát rồi trực quan hóa và đồng bộ dữ liệu để các nơi sinh không đồng nhất mặt cấu trúc thành cùng một nơi. Sau khi đồng bộ hóa tất cả nơi sinh thì chỉ còn 68 nơi sinh ( bao gồm cả các tỉnh,thành phố cũ và mới ).
  + Đối với các dữ liệu có giá trị null và rỗng dựa trên phân tích dữ liệu cho thấy hầu hết sinh viên có nơi sinh là ở TP. Hồ Chí Minh nên chúng ta thay thế chúng thành TP. Hồ Chí Minh.
  + Tổng hợp số lượng sinh viên ở nơi sinh theo thứ tự giảm dần như sau:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nơi sinh | Số lượng | Nơi sinh | Số lượng | Nơi sinh | Số lượng | Nơi sinh | Số lượng |
| TP. Hồ Chí Minh | 1665 | Tỉnh Bình Phước | 160 | TP. Đà Nẵng | 58 | Tỉnh Hà Nam Ninh | 8 |
| Tỉnh Đồng Nai | 523 | Tỉnh Hà Tĩnh | 151 | Tỉnh Cà Mau | 50 | Tỉnh Quảng Ninh | 5 |
| Tỉnh Đăk Lăk | 505 | Tỉnh Quảng Trị | 151 | Tỉnh Bạc Liêu | 47 | Tỉnh Lạng Sơn | 4 |
| Tỉnh Bình Định | 422 | Tỉnh Nghệ An | 144 | Tỉnh Thái Bình | 44 | Tỉnh Phú Thọ | 3 |
| Tỉnh Lâm Đồng | 355 | Tỉnh Thừa Thiên Huế | 122 | Tỉnh Sông Bé | 31 | Liên Bang Nga | 3 |
| Tỉnh Quảng Ngãi | 312 | Tỉnh Tây Ninh | 121 | Tỉnh Ninh Bình | 26 | Tỉnh Thái Nguyên | 3 |
| Tỉnh Quảng Nam | 283 | Tỉnh Bình Dương | 116 | TP. Hà Nội | 25 | Tỉnh Yên Bái | 2 |
| Tỉnh Tiền Giang | 242 | Tỉnh Vĩnh Long | 113 | Tỉnh Hải Dương | 24 | Tỉnh Hòa Bình | 2 |
| Tỉnh Bà Rịa - Vũng Tàu | 229 | Tỉnh Kiên Giang | 107 | Tỉnh Bắc Giang | 17 | Tỉnh Lai Châu | 2 |
| Tỉnh Khánh Hòa | 226 | Tỉnh Thanh Hóa | 100 | Tỉnh Hưng Yên | 16 | Tỉnh Tuyên Quang | 1 |
| Tỉnh An Giang | 208 | Tỉnh Quảng Bình | 94 | Tỉnh Hà Nam | 16 | Tỉnh Hà Sơn Bình | 1 |
| Tỉnh Phú Yên | 199 | Tỉnh Ninh Thuận | 75 | Tỉnh Đăk Nông | 15 | Tỉnh Hải Hưng | 1 |
| Tỉnh Gia Lai | 188 | Tỉnh Nam Định | 72 | TP. Hải Phòng | 14 | Campuchia | 1 |
| Tỉnh Bình Thuận | 180 | Tỉnh Sóc Trăng | 68 | Tỉnh Vĩnh Phúc | 13 | Tỉnh Cao Bằng | 1 |
| Tỉnh Bến Tre | 172 | Tỉnh Trà Vinh | 68 | Tỉnh Bắc Ninh | 12 | Tỉnh Bắc Kạn | 1 |
| Tỉnh Long An | 169 | TP. Cần Thơ | 65 | Australia | 10 | Tỉnh Minh Hải | 1 |
| Tỉnh Đồng Tháp | 163 | Tỉnh Kon Tum | 59 | Tỉnh Hậu Giang | 8 | Cộng Hòa Séc | 1 |

* Đồng nhất dữ liệu ‘tinhtrang\_hocvu’:
  + Cột ‘tinhtrang\_hocvu’ được lấy từ bảng XLHV sau khi map 3 bảng sinhvien + totnghiep + XLHV

A screenshot of a computer

Description automatically generated with medium confidence

* + Ta thấy có khá nhiều dữ liệu bị thiếu và bị trùng
  + Những giá trị thiếu là những MSSV không có trong bảng XLHV, ta thay bằng giá trị 0 thể hiện sinh viên không bị xử lý học vụ
  + Những giá trị trùng lặp là các MSSV bị xử lý học vụ nhiều lần, ta chỉ cần giữ lại giá trị tình trạng xử lý học vụ cao nhất, ta lược bỏ đi một số dòng có MSSV trùng nhau và giữ lại dòng cuối cùng

-> Bảng sau khi đã xử lý dữ liệu thiếu và dữ liệu trùng:

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| tinhtrang\_hocvu | Loại xử lý | Số lượng |
| 0 | Không bị xử lý | 6432 |
| 2 | Cảnh cáo | 800 |
| 5 | Thôi học | 877 |
| 7 | Gia hạn | 25 |
| 8 | Sinh viên tự do | 151 |

* Đồng nhất dữ liệu ‘dientt’**:** 
  + Cột dien\_tt mô tả diện trúng tuyển của thí sinh bao gồm 8 giá trị:

|  |  |  |
| --- | --- | --- |
| Diện xét tuyển | Ý nghĩa | Số lượng |
| THPT | Sinh viên trúng tuyển theo diện điểm thi Trung học Phổ thông Quốc gia | 6668 |
| ĐGNL | Sinh viên trúng tuyển theo điểm của kì thi đánh giá năng lực | 370 |
| ƯT-ĐHQG | Sinh viên trúng tuyển học sinh giỏi theo quy định của Đại học Quốc gia – Thành phố Hồ Chí Minh | 287 |
| TT-Bộ | Sinh viên tuyển thẳng theo quy chế của Bộ Giáo dục và Đào tạo | 42 |
| ƯT-Bộ | Sinh viên trúng tuyển học sinh giỏi theo quy định của Bộ Giáo dục và Đào tạo | 21 |
| CUTUYEN | Sinh viên trúng tuyển diện cử nhân | 20 |
| CCQT | Sinh viên trúng tuyển theo phương thức Chứng chỉ quốc tế đánh giá kiến thức khoa học tự nhiên, xã hội | 1 |
| 30A | Sinh viên trúng tuyển diện tuyển thẳng huyện nghèo | 1 |
| Không xác định | Giá trị không xác định hoặc bị trống | 72 |

* Giá trị không xác định sẽ được fill bằng với giá trị có số lượng nhiều nhất là THPT, ta có biểu đồ biểu diễn sau khi fill lA screenshot of a graph

  Description automatically generated with medium confidence

## **3.6. Xây dựng mô hình**

### 3.6.1 . Chia dữ liệu huấn luyện, dữ liệu kiểm tra

Trong bảng dữ liệu (Data) ta có 3369 dòng dữ liệu tương ứng 3369 sinh viên. Nhóm sẽ dự đoán khả năng tốt nghiệp dựa trên thời gian sinh viên học tại trường đó là: Sinh viên mới vào trường, sinh viên học tới học kì 1, sinh viên học tới học kì 2, sinh viên học tới học kì 3, sinh viên học tới học kì 4.

Dữ liệu kiểm tra sẽ là dữ liệu của của sinh viên theo từng học kì và dữ liệu test cũng là dữ liệu của từng học kì cho một khóa sinh viên đã tốt nghiệp cụ thể để huấn luyện cho bộ test của nhóm đánh giá khả năng tốt nghiệp của sinh viên. Trước khi đưa dữ liệu vào bộ train và bộ test, nhóm cũng đã điều chỉnh cho phù hợp các nhãn bị trống thuộc bảng dữ liệu (Data) và chuẩn hóa dữ liệu về tỉ lệ tương đương để đảm bảo độ chính xác cao khi huấn luyện và kiểm định mô hình dự đoán.

### 3.6.2. Mã hóa dữ liệu

Đối với dữ liệu huấn luyện thì:

* Các cột thuộc tính chứa giá trị rời rạc được mã hóa theo phương pháp Label Encoder
* Các cột thuộc tính chứa giá trị số thì được giữ nguyên
* Các nhãn sẽ được đặt tên dễ hiểu, sắp xếp đúng thứ tự và đồng nhất.

Đối với dữ liệu kiểm tra thì:

* Các cột thuộc tính chứa giá trị rời rạc được mã hóa theo phương pháp Label Encoder
* Các cột thuộc tính chứa giá trị số thì được giữ nguyên
* Các nhãn sẽ được đặt tên dễ hiểu, sắp xếp đúng thứ tự và đồng nhất.

### 3.6.3. Chuẩn hóa dữ liệu

Với cả dữ liệu huấn luyện và dữ liệu kiểm tra thì:

* Các cột thuộc tính về nơi sinh
* Các cột thuộc tính về diện trúng tuyển
* Các cột thuộc tính về lớp sinh hoạt
* Các cột thuộc tính về khóa
* Các cột thuộc tính về hệ đào tạo
* Các cột thuộc tính về chuyên ngành 2
* Các cột thuộc tính về xếp loại tốt nghiệp được nhóm chuẩn hóa theo phương pháp Standard Scaler:
  + Phương thức chuẩn hóa bằng cách loại bỏ giá trị trung bình và cân bằng tỉ lệ theo tỉ lệ phương sai.
  + Điểm chuẩn hóa của một mẫu x được tính bằng công thức:

Z = (x-u)/s

Trong đó:

* u là trung bình của các mẫu huấn luyện hoặc bằng 0 nếu with\_mean = False
* s là độ lệch chuẩn của các mẫu huấn luyện hoặc bằng 1 nếu with\_std=False

Cân bằng và tỉ lệ xảy ra riêng biệt với mỗi phương thức bằng cách tính toán số liệu thống kê có liên quan trong các mẫu của bảng huấn luyện. Giá trị trung bình và độ lệch chuẩn được lưu lại và tái sử dụng khi hàm transform được gọi lên.

Chuẩn hóa dữ liệu cho một bảng dữ liệu là điều cần thiết cho nhiều mô hình máy học: Những mô hình này vận hành không được tốt nếu những phương thức riêng lẻ này không được chú trọng đến để điều chỉnh những dữ liệu quan trọng.

Ví dụ với những yếu tố sử dụng trong hàm chủ động trong thuật toán (ví dụ như là RBF kernel của Support Vector Machine hay là bộ điều chỉnh L1 và L2 của mô hình tiệm cận), những thuật toán này yêu cầu tất cả các tính năng cân bằng xung quanh giá trị 0 và sắp xếp các phương sai trong cùng một trình tự. Nếu một tính năng có một phương sao mà thứ tự của nó có kích cỡ lớn hơn các phương sai khác, điều này dẫn tới sự áp đảo trong hàm chủ động và làm bộ tính toán không thể học từ các tính năng khác một cách chính xác.

Bộ điều chỉnh tỉ lệ này có thể áp dụng để giảm bớt ma trận CSR hoặc CSC thông qua gọi lên with\_mean = False để tránh làm hỏng cấu trúc thưa thớt trong dữ liệu.

### 3.6.4. Huấn luyện các mô hình dự đoán dựa trên các thuật toán máy học

#### **3.6.4.1. Naïve Bayes**

Thuật toán Naive Bayes là một thuật toán phân loại trong lĩnh vực học máy. Nó dựa trên nguyên tắc của định lý Bayes để tính xác suất xảy ra của một sự kiện dựa trên các đặc trưng có sẵn.

Thuật toán Naive Bayes giả định rằng các đặc trưng đầu vào là độc lập nhau, tức là không có sự tương quan hoặc ảnh hưởng lẫn nhau. Mặc dù giả định này thường không đúng trong thực tế, thuật toán vẫn cho kết quả tốt trong nhiều bài toán và được sử dụng rộng rãi trong phân loại văn bản, phân loại email và các bài toán khác.

Quy trình xây dựng mô hình Naive Bayes bao gồm các bước sau:

1. Chuẩn bị dữ liệu: Dữ liệu huấn luyện được sắp xếp thành các mẫu với nhãn đã biết. Mỗi mẫu gồm các đặc trưng đầu vào.
2. Xác định xác suất tiên nghiệm: Xác suất tiên nghiệm của các lớp đầu ra được tính dựa trên tần suất xuất hiện của mỗi lớp trong dữ liệu huấn luyện.
3. Xác định xác suất hậu nghiệm: Dựa trên định lý Bayes, xác suất hậu nghiệm của các lớp đầu ra được tính dựa trên xác suất tiên nghiệm và xác suất có điều kiện của các đặc trưng đầu vào.
4. Dự đoán: Dựa trên xác suất hậu nghiệm đã tính, thuật toán sẽ dự đoán lớp đầu ra cho các mẫu dữ liệu mới.

Thuật toán Naive Bayes có các ưu điểm như đơn giản, dễ triển khai, hiệu suất tính toán nhanh và tốt với các tập dữ liệu lớn. Tuy nhiên, nó có giả định về tính độc lập của các đặc trưng, vì vậy nếu giả định này không chính xác, kết quả có thể bị ảnh hưởng.

#### **3.6.4.2. Support Vector Machine (SVM)**

Thuật toán Support Vector Machine (SVM) là một thuật toán phân loại và hồi quy trong lĩnh vực học máy. Nó được sử dụng phổ biến vì khả năng phân loại chính xác các dữ liệu phức tạp và tìm ra đường biên tốt nhất để phân chia các lớp.

SVM tìm ra một đường biên tối ưu giữa các điểm dữ liệu thuộc các lớp khác nhau. Đường biên này được gọi là siêu phẳng (hyperplane) và được chọn sao cho cách xa nhất so với các điểm dữ liệu của các lớp. Các điểm dữ liệu gần nhất với siêu phẳng này được gọi là các vector hỗ trợ (support vectors). SVM tận dụng các vector hỗ trợ này để xác định đường biên phân chia và thực hiện dự đoán trên các điểm dữ liệu mới.

Thuật toán SVM có các đặc điểm quan trọng sau:

1. Hàm Kernel: SVM sử dụng hàm kernel để ánh xạ dữ liệu từ không gian ban đầu sang không gian cao chiều hơn, giúp phân loại các dữ liệu không tuyến tính. Một số hàm kernel phổ biến là Linear Kernel, Polynomial Kernel và Gaussian RBF Kernel.
2. Margin lớn nhất: SVM tìm cách tạo ra một đường biên có khoảng cách lớn nhất với các điểm dữ liệu gần nhất của các lớp. Điều này giúp giảm thiểu sự chồng lấn giữa các lớp và tăng tính tổng quát hóa của mô hình.
3. Biến mềm (Soft Margin): Trong trường hợp không thể tìm ra một đường biên hoàn hảo, SVM cho phép sự chấp nhận một số lỗi phân loại bằng cách cho phép một margin nhất định.
4. Đa lớp: SVM cũng có khả năng xử lý các bài toán phân loại đa lớp bằng cách kết hợp các bộ phân loại nhị phân (binary classifier) thông qua các phương pháp như One-vs-One và One-vs-All.

SVM có nhiều ưu điểm như khả năng phân loại hiệu quả đối với dữ liệu phức tạp, khả năng xử lý các bộ dữ liệu lớn, và khả năng xử lý các bài toán phân loại đa lớp. Tuy nhiên, SVM có độ phức tạp tính toán cao và yêu cầu sự chuẩn bị dữ liệu cẩn thận.

#### **3.6.4.3. Random Forest**

Random Forest là một thuật toán học máy được sử dụng phổ biến trong các bài toán phân loại và hồi quy. Nó là một phương pháp kết hợp (ensemble method) dựa trên cây quyết định (decision tree) để tạo ra một mô hình mạnh mẽ và ổn định.

Ý tưởng chính của Random Forest là xây dựng nhiều cây quyết định độc lập và kết hợp kết quả từ các cây này để đưa ra dự đoán cuối cùng. Mỗi cây quyết định trong Random Forest được xây dựng trên một tập con ngẫu nhiên của dữ liệu huấn luyện, được gọi là bootstrap samples. Đồng thời, trong quá trình xây dựng mỗi nút trong cây, chỉ một số lượng ngẫu nhiên các thuộc tính được chọn để đánh giá độ tách biệt.

Các bước chính để xây dựng mô hình Random Forest bao gồm:

1. Bootstrap Sampling: Lựa chọn ngẫu nhiên các mẫu từ dữ liệu huấn luyện với sự thay thế (sampling with replacement) để tạo ra các tập con dữ liệu.
2. Xây dựng cây quyết định: Đối với mỗi tập con dữ liệu, một cây quyết định được xây dựng bằng cách lặp lại các bước chọn thuộc tính và tách nút cho đến khi cây hoàn thành.
3. Kết hợp dự đoán: Khi có một dữ liệu mới cần dự đoán, các cây quyết định trong Random Forest đưa ra dự đoán riêng lẻ và kết quả cuối cùng được tính dựa trên đa số phiếu bầu hoặc trung bình (trong trường hợp hồi quy).

Các ưu điểm của Random Forest bao gồm khả năng xử lý các dữ liệu có nhiễu và tương quan cao, khả năng xác định độ quan trọng của các thuộc tính, khả năng xử lý dữ liệu thiếu, và khả năng xử lý cả bài toán phân loại lẫn hồi quy. Ngoài ra, Random Forest cũng giảm thiểu nguy cơ overfitting và thường có hiệu suất dự đoán tốt.

Tuy nhiên, việc xây dựng một mô hình Random Forest có thể tốn kém về mặt tính toán và thời gian, đặc biệt là với dữ liệu lớn.

#### **3.6.4.4. Decision Tree**

Thuật toán cây quyết định (decision tree) là một phương pháp phân loại và dự đoán trong lĩnh vực học máy. Nó được sử dụng rộng rãi trong việc xây dựng mô hình dự đoán dựa trên các quyết định đơn giản hướng dẫn từ các thuộc tính của dữ liệu.

Cây quyết định được xây dựng dựa trên một quy trình phân tách tuần tự, trong đó từng nút của cây đại diện cho một quyết định dựa trên một thuộc tính của dữ liệu. Các nút gốc của cây biểu thị thuộc tính quan trọng nhất trong việc phân loại dữ liệu, trong khi các nút lá đại diện cho các nhãn hoặc kết quả cuối cùng.

Quy trình xây dựng cây quyết định bao gồm các bước sau:

1. Lựa chọn thuộc tính: Bước này liên quan đến việc chọn thuộc tính tốt nhất để phân tách dữ liệu. Các thuộc tính được đánh giá dựa trên các tiêu chí như độ tách biệt giữa các lớp, độ thuần khiết và độ tin cậy.
2. Xây dựng cây: Mỗi nút trong cây quyết định biểu diễn một quyết định dựa trên thuộc tính đã chọn. Dữ liệu được phân tách thành các nhánh con tương ứng với các giá trị của thuộc tính.
3. Tiếp tục phân tách: Quá trình xây dựng cây tiếp tục lặp lại các bước trên cho từng nhánh con cho đến khi đạt được điều kiện dừng, chẳng hạn như khi mọi nhánh đều chứa cùng một lớp hoặc không còn thuộc tính để phân tách.
4. Cắt tỉa (pruning): Để tránh việc overfitting (quá khớp) và làm cho cây tổng quát hóa tốt hơn, một quá trình cắt tỉa có thể được áp dụng để loại bỏ các nhánh không cần thiết hoặc không quan trọng.

Cây quyết định có nhiều ưu điểm, bao gồm khả năng hiểu được và giải thích tốt, khả năng xử lý dữ liệu dạng hỗn hợp và giảm chi phí tính toán. Tuy nhiên, nó có thể dễ bị ảnh hưởng bởi nhiễu trong dữ liệu và có thể dẫn đến quá khớp nếu không được điều chỉnh phù hợp. Có nhiều biến thể và cải tiến của thuật toán cây quyết định như cây quyết định ngẫu nhiên (random forest) và cây quyết định tăng cường (gradient boosting).

### 3.6.2. Kiểm định mô hình dự đoán

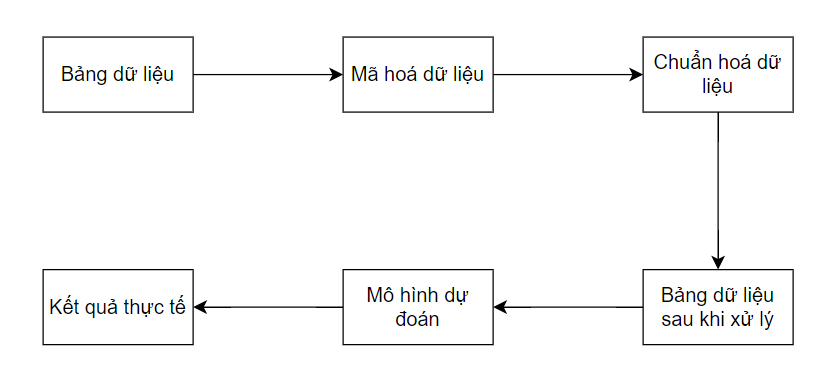
Tại đây, hai độ đo: độ chính xác (accuracy) và độ đo F1-score được nhóm sử dụng để kiểm định chất lượng của mô hình dự đoán trong các trường hợp thử nghiệm.

### 3.6.3. Các bước còn lại

Nhóm sẽ thực hiện chọn mô hình có kết quả thực nghiệm tốt nhất cho từng test và tiến hành so sánh giữa các test với nhau để chọn ra những thuộc tính quan trọng nhất cho bài toán dự đoán khả năng tốt nghiệp của sinh viên trường đại học Công nghệ Thông tin.

## **3.7. Dự đoán thực tế**

Hiện tại chúng ta sẽ có bảng dữ liệu chính với các cột thuộc tính và một cột nhãn. Quy trình cần thực hiện là:



Hình 1. Quy trình để dự đoán thực tế

## **3.8. Kết quả thực nghiệm**

Trong phần này nhóm sẽ trình bày kết quả kiểm định mô hình dự đoán. Để có được kết quả kiểm định mô hình, nhóm đã lập trình bằng ngôn ngữ Python với các framework như: pandas, numpy, sklearn, …

Bộ dữ liệu chính đang có thông tin lý lịch đầy đủ và quá trình học tập của các sinh viên tại trường cho đến lúc tốt nghiệp. Bài toán dự đoán khả năng tốt nghiệp của sinh viên sẽ có ý nghĩa cao hơn khi ta dự đoán sớm được khả năng tốt nghiệp của sinh viên đó, cụ thể là thời điểm sinh viên mới nhập học vào trường và trễ nhất là sau khi học 4 học kỳ ở trường.

Vì vậy nhóm xin thử nghiệm mô hình với 5 trường hợp thử nghiệm như sau:

Thực nghiệm 0: Chỉ có thông tin sinh viên [(link)](https://colab.research.google.com/drive/1NPlSsAw5OF3V-HCfoDMBXzXcGQvHQDoh?usp=sharing)

Thực nghiệm 1: Thông tin sinh viên và các thông tin về điểm số khi sinh viên học 1 học kỳ [(link)](https://colab.research.google.com/drive/1D3cXZRgY-e6VYFrE0CPXc12FxheGhFCT?usp=sharing)

Thực nghiệm 2: Thông tin sinh viên và các thông tin về điểm số khi sinh viên học 2 học kỳ [(link)](https://colab.research.google.com/drive/1W9SH6sHs5KU8aR0ggCKmsVKZgQKnkJn_?usp=sharing)

Thực nghiệm 3: Thông tin sinh viên và các thông tin về điểm số khi sinh viên học 3 học kỳ [(link)](https://colab.research.google.com/drive/1FbvxRXZydMEyrqgy9_kV9d8LuqsT9qTj?usp=sharing)

Thực nghiệm 4: Thông tin sinh viên và các thông tin về điểm số khi sinh viên học 4 học kỳ [(link)](https://colab.research.google.com/drive/1V1RsptqTKw9yyWMxzYFUZfMcPhsvcVqx?usp=sharing)

Với mỗi trường hợp thử nghiệm, nhóm có đề xuất 2 cách xử lý dữ liệu, cụ thể là dữ liệu về điểm. Khi tổ hợp lại nhóm sẽ có 10 quy trình khai phá dữ liệu khác nhau. Nhóm thực hiện huấn luyện trên 4 mô hình dự đoán từ đó dẫn đến có 40 mô hình để so sánh với nhau.

Thực nghiệm 0 bao gồm:

Quy trình khai phá dữ liệu 1: điền các dữ liệu trống của diem\_tt bằng -1

Quy trình khai phá dữ liệu 2: điền các dữ liệu trống của diem\_tt bằng điểm trung bình của cột diem\_tt

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Quy trình 1 | | | Quy trình 2 | | |
| Acc | F1 (0) | F1 (1) | Acc | F1 (0) | F1 (1) |
| **Naive Bayes** | **72.93%** | **16%** | **84%** | 73.06% | 1% | 84% |
| SVM | 72.40% | 14% | 84% | 72.67% | 3% | 84% |
| Random Forest | 68.86% | 17% | 81% | 69.12% | 15% | 81% |
| Decision Tree | 63.07% | 27% | 75% | 63.60% | 21% | 76% |

Như vậy, kết quả của quy trình khai phá 1 cho kết quả accuracy thấp hơn một chút so với quy trình khai phá 2 nhưng F1 của lớp 0 (lớp thiểu số) lại cho kết quả tốt hơn nhiều so với quy trình khai phá 2. Vì vậy nhóm kết luận quy trình khai phá 1 tốt hơn trong thực nghiệm này. Và sẽ lấy các xử lý cột diem\_tt này áp dụng cho các thực nghiệm kế tiếp.

Và quan sát thấy mô hình máy học Naive Bayes sẽ cho kết quả cao nhất so với các thuật toán còn lại ở cả 2 quy trình khai phá dữ liệu.

Thực nghiệm 1 bao gồm:

Quy trình khai phá dữ liệu 3: điền các dữ liệu trống của ‘Học kì 1’ bằng -1

Quy trình khai phá dữ liệu 4: điền các dữ liệu trống của ‘Học kì 1’ bằng điểm trung bình của cột.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Quy trình 3 | | | Quy trình 4 | | |
| Acc | F1 (0) | F1 (1) | Acc | F1 (0) | F1 (1) |
| Naive Bayes | 79.37% | 63% | 86% | 78.71% | 63% | 85% |
| **SVM** | 81.21% | 57% | 88% | **81.47%** | **58%** | **88%** |
| Random Forest | 71.14% | 63% | 84% | 77.40% | 62% | 84% |
| Decision Tree | 67.54% | 54% | 75% | 62.94% | 47% | 72% |

Như vậy, kết quả của quy trình khai phá 3 và quy trình khai phá 4 tuy có sự chênh lệch giữa các mô hình nhưng nhìn chung là tương tự nhau. Và quan sát thấy mô hình máy học SVM ở quy trình khai phá 4 sẽ cho kết quả tốt hơn so với các mô hình còn lại.

Thực nghiệm 2 bao gồm:

Quy trình khai phá dữ liệu 5: điền các dữ liệu trống của ‘Học kì 2’ bằng -1

Quy trình khai phá dữ liệu 6: điền các dữ liệu trống của ‘Học kì 2’ bằng điểm trung bình các học kỳ trước.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Quy trình 5 | | | Quy trình 6 | | |
| Acc | F1 (0) | F1 (1) | Acc | F1 (0) | F1 (1) |
| Naive Bayes | 81.08% | 45% | 89% | 81.08% | 45% | 89% |
| **SVM** | **85.15%** | **68%** | **90%** | 84.89% | 68% | 90% |
| Random Forest | 83.44% | 71% | 88% | 83.18% | 71% | 88% |
| Decision Tree | 74.11% | 58% | 81% | 75.56% | 62% | 82% |

Như vậy, kết quả của quy trình khai phá 5 nhìn chung cho kết quả tốt hơn quy trình khai phá 6. Và quan sát thấy mô hình máy học SVM sẽ cho kết quả tốt hơn so với các thuật toán còn lại.

Thực nghiệm 3 bao gồm:

Quy trình khai phá dữ liệu 7: điền các dữ liệu trống của ‘Học kì 3’ bằng -1

Quy trình khai phá dữ liệu 8: điền các dữ liệu trống của ‘Học kì 3’ bằng điểm trung bình các học kì trước.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Quy trình 7 | | | Quy trình 8 | | |
| Acc | F1 (0) | F1 (1) | Acc | F1 (0) | F1 (1) |
| Naive Bayes | 85.28% | 73% | 90% | 85.15% | 73% | 90% |
| **SVM** | **87.12%** | **75%** | **91%** | 86.99% | 74% | 91% |
| Random Forest | 83.18% | 72% | 88% | 83.71% | 73% | 88% |
| Decision Tree | 77.66% | 65% | 84% | 78.98% | 67% | 85% |

Như vậy, kết quả của quy trình khai phá 7 và quy trình khai phá 8 tuy có sự chênh lệch giữa các mô hình nhưng nhìn chung là tương tự nhau. Và quan sát thấy mô hình máy học SVM ở quy trình khai phá 7 sẽ cho kết quả tốt hơn so với các thuật toán còn lại.

Thực nghiệm 4 bao gồm:

Quy trình khai phá dữ liệu 9: điền các dữ liệu trống của ‘Học kì 4’ bằng -1

Quy trình khai phá dữ liệu 10: điền các dữ liệu trống của ‘Học kì 4’ bằng điểm trung bình các học kì trước.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Quy trình 9 | | | Quy trình 10 | | |
| Acc | F1 (0) | F1 (1) | Acc | F1 (0) | F1 (1) |
| Naive Bayes | 88.17% | 78% | 92% | 87.52% | 77% | 91% |
| **SVM** | **89.75%** | **80%** | **93%** | 89.62% | 80% | 93% |
| Random Forest | 88.17% | 78% | 92% | 88.30% | 79% | 92% |
| Decision Tree | 88.17% | 69% | 87% | 88.30% | 64% | 84% |

Như vậy, kết quả của quy trình khai phá 9 cho kết quả accuracy thấp hơn một chút so với quy trình khai phá 10 nhưng F1 của lớp 0 (lớp thiểu số) lại cho kết quả tốt hơn so với quy trình khai phá 10. Và quan sát thấy mô hình máy học SVM ở quy trình khai phá 9 sẽ cho kết quả tốt hơn so với các mô hình còn lại.

Từ các thực nghiệm và các mô hình, nhóm rút ra các mô hình tốt nhất cho từng thực nghiệm như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Thực nghiệm** | **Quy trình** | **Thuật toán** | **Acc** | **F1(0)** | **F1(1)** |
| Thực nghiệm 0 | Quy trình 1 | Naive Bayes | 72.93% | 16% | 84% |
| Thực nghiệm 1 | Quy trình 4 | SVM | 81.47% | 58% | 88% |
| Thực nghiệm 2 | Quy trình 5 | SVM | 85.15% | 68% | 90% |
| Thực nghiệm 3 | Quy trình 7 | SVM | 87.12% | 75% | 91% |
| Thực nghiệm 4 | Quy trình 9 | SVM | 89.75% | 80% | 93% |

Nhận thấy cách điền giá trị bị thiếu cho các cột điểm học kì trống bằng một giá trị âm -1 (các quy trình khai phá đánh số lẻ) cho kết quả dự đoán tốt hơn cách xử lý còn lại.

Thuật toán SVM là thuật toán cho kết quả tốt nhất trong ⅘ thực nghiệm nhóm đã làm. Các thực nghiệm được thêm vào dần các điểm học kì biểu thị cho thời gian sinh viên mới vào trường cho đến khi học hết năm 2. Càng nhiều thông tin về quá trình học của các sinh viên thì mô hình dự đoán càng chính xác. Kết quả dự đoán cao nhất của 5 thực nghiệm có độ chính xác lên đến 89.75% khi muốn dự đoán khả năng tốt nghiệp của sinh viên khi sinh viên đó học được 4 học kì ở trường.