BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

#   



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

**XÂY DỰNG WEBSITE GIỚI THIỆU KHOA**

**TÍCH HỢP HỆ THỐNG GỢI Ý CHỌN NGÀNH HỌC**

**Sinh viên thực hiện : Nguyễn Đăng Khoa**

**Mã số : B1611129**

**Khóa : 42**

Cần Thơ, ??/????

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

#   



**LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC**

**NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

XÂY DỰNG WEBSITE GIỚI THIỆU KHOA

TÍCH HỢP HỆ THỐNG GỢI Ý CHỌN NGÀNH HỌC

**Sinh viên thực hiện:**

**Nguyễn Đăng Khoa**

**Mã số : B1611129**

**Khóa: 42**

**Giảng viên hướng dẫn:**

**Ths. Phạm Xuân Hiền**

Cần Thơ, ??/????

NHẬN XÉT CỦA GIẢNG VIÊN

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

-----------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

-----------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

-----------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

-----------------------------------------------------------------------------

------------------------------------------------------------------------------

------------------------------------------------------------------------------

-----------------------------------------------------------------------------

------------------------------------------------------------------------------

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn Khoa Công Nghệ Thông Tin và Truyền Thông đã tạo điều kiện thuận lợi cho chúng em học tập và thực hiện đề tài niên luận này. Chúng em xin bày tỏ lòng biết ơn sâu sắc đến cô Phạm Xuân Hiền đã tận tình hướng dẫn chỉ bảo chúng em trong quá trình thực hiện đề tài.

Em xin giử lời cảm ơn chân thành đến các Thầy Cô Giảng viên Đại học Cần Thơ, đặc biệt là các Thầy Cô ở Khoa CNTT & TT, những người đã truyển đạt những kiến thức quý báu trong thời gian qua.

Trong phạm vi giới hạn về thời gian cũng như kiến thức, chắc chắn bài làm của em còn nhiều thiếu sót. Em rất mong được sự góp ý và chỉ dẫn của các Thầy Cô để bài làm được hoàn chỉnh hơn.

Em xin chân thành cảm ơn.

Cần Thơ, ngày ... tháng … năm … Người viết

Nguyễn Đăng Khoa

MỤC LỤC

[**PHẦN GIỚI THIỆU 6**](#_bookmark0)

1. [Đặt vấn đề 6](#_bookmark1)
2. [Lịch sử giải quyết vấn đề 6](#_bookmark2)
3. [Mục tiêu đề tài 6](#_bookmark3)
4. [Đối tượng và phạm vi nghiên cứu 6](#_bookmark4)
5. [Phương pháp nghiên cứu 6](#_bookmark5)
6. [Kết quả đạt được 6](#_bookmark6)
7. [Bố cục luận văn 6](#_bookmark7)

[**PHẦN NỘI DUNG 7**](#_bookmark8)

[**CHƯƠNG 1 7**](#_bookmark9)

[**MÔ TẢ BÀI TOÁN 7**](#_bookmark10)

1. [Mô tả chi tiết bài toán 7](#_bookmark11)
2. [Vấn đề và giải pháp liên quan đến bài toán 7](#_bookmark12)
   1. [Vấn đề 1 7](#_bookmark13)
      1. [Ý 1 của vấn đề 1 7](#_bookmark14)
      2. [Ý 2 của vấn đề 1 7](#_bookmark15)
   2. [Vấn đề 2 7](#_bookmark16)

[1.2.1 Ý 1 của vấn đề 2 7](#_bookmark17)

1. [Mô tả giải pháp cho bài toán 7](#_bookmark18)

[**THIẾT KẾ VÀ CÀI ĐẶT 8**](#_bookmark19)

1. [Thiết kế hệ thống 8](#_bookmark20)
2. [Thiết kế và cài đặt giải thuật 8](#_bookmark21)
3. [Giao diện hệ thống 8](#_bookmark22)

[**CHƯƠNG 3 9**](#_bookmark23)

[**KIỂM THỬ VÀ ĐÁNH GIÁ 9**](#_bookmark24)

1. [Mục tiêu 9](#_bookmark25)
2. [Nghi thức kiểm tra 9](#_bookmark26)
3. [Kết quả kiểm tra 9](#_bookmark27)

[**PHẦN KẾT LUẬN 10**](#_bookmark28)

1. [Kết quả đạt được 10](#_bookmark29)
2. [Hướng phát triển 10](#_bookmark30)

[**TÀI LIỆU THAM KHẢO 11**](#_bookmark31)

DANH MỤC HÌNH

Hình 1 : 10

Hình 2 : 12

Hình 3 17

DANH MỤC BẢNG

Bảng 1 15

Bảng 2 : 23

Bảng 3 : 23

ABSTRACT

TÓM TẮT

**PHẦN GIỚI THIỆU**

1. **Đặt vấn đề**

Theo số liệu thống kê của Bộ Giáo dục và Đào tạo [số] [số] [số], tính đến năm 2019 cả nước có 236 trường đại học, học viện (không tính các trường thuộc khối ngành quốc phòng – an ninh) trong đó bao gồm 171 trường công lập, 60 trường tư thục và dân lập. Với số lượng đăng ký vào năm 2019 là 887.173 thí sinh trong khi tổng chi tiêu cả nước chỉ 489.637 thí sinh. Tình hình này dẫn đến áp lực nặng nề cho các học sinh trong kỳ thi THPT Quốc gia. Hầu hết các em có mong muốn vào trường đại học (kể cả những em có học lực yếu) trong khi xã hội đang cần rất nhiều công nhân kỹ thuật lành nghề tham gia trực tiếp trong một số lĩnh vực nghề nghiệp. Ước mơ của các em đôi khi còn rất xa với thực tế lao động, hoạt động nghề nghiệp, chưa thấy được giá trị đích thực của các nghề. Các em có kỳ vọng quá cao vào một số nghề nhưng khi tiếp xúc với nghề nghiệp trong thực tế thường làm các em thất vọng. Chọn nghề mà hiểu biết quá ít, thậm chí không hiểu nghề định chọn thì sớm muộn sẽ gặp trở ngại lớn trong hoạt động nghề nghiệp của cá nhân, tạo sự hẫng hụt, bi quan chán nản, miễn cưỡng trong lao động.

Ở Việt Nam hiện nay do cơ chế quản lý kinh tế đang có nhiều thay đổi, ảnh hưởng cả tích cực và tiêu cực tới thị trường lao động nên vấn đề việc làm rất được quan tâm. Để phát triển và xây dựng kinh tế của đất nước, chất lượng nguồn nhân lực phải được xem là yếu tố then chốt để phát huy mọi tiềm lực của đất nước. Vẫn là bài toán cũ nhưng vẫn chưa có cách giải quyết, tình trang thất nghiệp sau đại học vẫn không có xu hướng giảm xuống. Theo bản tin cập nhật thị trường lao động Việt Nam số 21, quý một năm 2019 của Bộ lao động – Thương binh và Xã hội, số người thất nghiệp có trình độ “cao đẳng” là 65,1 nghìn người, nhóm trình độ “trung cấp” là 52.7 nghìn người và nhóm có trình độ “đại học” là 124,5 nghìn người và ngoài ra còn một số ngành nghề khác[số]. Điều đó cũng có nghĩa hàng năm, hơn 100.000 sinh viên tốt nghiệp mà không được làm việc theo đúng ngành nghề đã chọn, hệ lũy đầu tiên của việc này có thể kể đến là lãng phí thời gian, cùng số lượng lớn tiền bạc bỏ ra nhưng không thể áp dụng kiến thức đã học vào thực tế và thực trạng thất nghiệp đã trở thành một đặc trưng của nền kinh tế thị trường, điều này ảnh hưởng đến sự phát triển, làm suy giảm sự tăng trưởng kinh tế của một đất nước. Hằng năm, nước ta có một lực lượng lớn trình độ đại học, cao đẳng phải chấp nhận làm việc trái ngành nghề, không phù hợp với trình độ đã được đào tạo và nguyện vọng của bản thân. Thậm chí, có một nghịch lý hết sức phổ biến hiện nay là lao động phổ thông dể tìm việc hoặc tìm được những công việc có thu nhập cao. Trong khi đó, lao động có trình độ đại học, cao đẳng lại khó xin việc do tâm lý kén chọn, quá kỳ vọng vào tấm bằng đại học mà mình đang có. Nguyên nhân khách quan dẫn đến tình trạng cử nhân thất nghiệp ngày càng nhiều là do trong nhiều năm qua công tác tuyên truyền tư vấn nghề nghiệp, định hướng nghề nghiệp chưa đạt yêu cầu hoặc còn nhiều thiếu sót, cộng thêm nếp suy nghĩ truyền thống của xã hội khiến các sinh viên hiện nay vẫn quyết tâm có được bằng đại học bằng mọi giá… Hệ quả là sau khi tốt nghiệp sinh viên rất khó thích nghi với môi trường làm việc mới trong khi yêu cầu của thị trường lao động ngày càng biến động. Về nguyên nhân chủ quan, đa phần học sinh, sinh viên chưa đánh giá đúng kỹ năng, sở trường, sở đoản của bản thân nên dẫn đến túng trong quá trình lựa chọn hệ học, ngành học, việc chọn nghề theo mong muốn của cha mẹ, theo ý kiến cá nhân của người khác mà không căn cứ vào khả năng của bản thân và nhu cầu của xã hội, dẫn đến đưa ra quyết định sai lâm trong bước đầu chọn lựa nghề nghiệp. Chính vì vậy, định hướng nghề nghiệp cho sinh viên ngay khi còn ngồi trên ghế nhà trường là vấn đề vô cùng quan trọng và cấp thiết để tạo ra nguồn nhân lực có chẩt lượng, phù hợp với mong muốn cá nhân và nhu cầu xã hội đặt ra.

Việc làm sau khi tốt nghiệp của sinh viên, đặt biệt là sinh viên đại học là một vấn đề quan trọng ảnh hưởng đến chất lượng nguồn nhân lực của Việt Nam. Vì thế tìm hiểu được nhu cầu, nguyện vọng việc làm sau khi tốt nghiệp của sinh viên là góp phần giải quyết được bài toán “nóng” hiện nay của học sinh – sinh viên.

Xuất phát từ những vấn đề đã đặt ra, đề tài: **“ xây dựng website giới thiệu khoa tích hợp hệ thống gợi ý chọn ngành học”** với mục tiêu tìm hiểu các yếu tố ảnh hưởng đến lựa chọn nghề nghiệp của học sinh – sinh viên hiện nay, qua đó đưa ra gợi ý về ngành học góp phần hỗ trợ học sinh THPT lựa chọn ngành nghề phù hợp với sở thích cũng như khả năng của bản thân.

1. **Lịch sử giải quyết vấn đề**
2. **Các nghiên cứu trong nước**

Đã có rất nhiều đề tài nghiên cứu về yếu tố ảnh hưởng đến quyết định chọn ngành của học sinh – sinh viên nhưng với nhiều góc độ khác nhau, dưới đây là một số đề tài nghiên cứu vấn đề này:

Bài báo ***“Đánh giá các yếu tố ảnh hưởng đến định hướng nghề nghiệp của học sinh Trung học phổ thông tại Nghệ An”*** ( Nghiên cứu của Trường Đại học Vinh, tác giả Nguyễn Thị Kim Nhung – Lương Thị Thành Vinh )[số]. Mục tiêu đề tài là xác định được các yếu tố ảnh hướng đến quyết định chọn ngành nghề của học sinh THPT cư trú tại Thành Phố Nghệ An. Kết quả thu được, nghiên cứu đã xây dựng và kiểm chứng mô hình các yếu tố ảnh hưởng đến định hướng nghề nghiệp (ĐHNN) của học sinh THPT tỉnh Nghệ An với 7 nhân tố đại diện. Theo kết quả khảo sát, đánh giá của người học trong việc xác định mức độ quan trọng của từng yếu tố đối với ĐHNN[vt] của mình, yếu tố cá nhân người học có vị trí dẫn đầu (7,26/10 điểm); thứ 2 là yếu tố gia đình(5,52/10 điểm); thứ 3 là yếu tố xu hướng phát triển Kinh tế - Xã hội(5,46/10 điểm); thứ tư là yếu tố nhà trường(4,57/10 điểm); thứ 5 là yếu tố hoạt động hướng nghiệp của các trường ĐH – CĐ; thứ 6 là hoạt động truyền thống và cuối cùng là yếu tố bạn bè. Nhưng khi kiểm định thang đo thì chỉ còn 2 yếu tố có ý nghĩa quan trọng nhất là yếu tố *cá nhân người học* và *nhân tố gia đình*, trong đó nhân tố cá nhân người học giữ vai trò quyết định. Mô hình nghiên cứu chính đã giải thích được 88,9% cho tổng thể về mối liên hệ của 2 nhân tố *cá nhân người học* và *gia đình*, đồng thời khẳng định mối quan hệ đồng biến giữa 2 nhân tố này với biến phụ thuộc.

Đề tài nghiên cứu ***“Phân tích các nhân tố ảnh hưởng đến quyết định chọn ngành, trường đại học của học sinh THPT vùng ĐBSCL”*** (Khoa Kinh tế, trường Đại học Cần Thơ năm 2017)[số]. Với mục tiêu phân tích các nhân tố ảnh hưởng đến quyết định chọn ngành và trường đại học của học sinh THPT vùng ĐBSCL, từ đó đề xuất giải pháp cho công tác hướng nghiệp cho học sinh THPT. Sơ lược về đề tài, phạm vi nghiên cứu của đề tài bao gồm 4 tỉnh, thành phố: Bạc Liêu, Tiền Giang, An Giang, Cần Thơ. Kết quả nghiên cứu cho thấy, trong các nhân tố thì nhân tố tâm lý ảnh hưởng đáng kể nhất đến quyết định chọn ngành và trường đại học của học sinh. Những yếu tố tâm lý bên trong con người như: nhu cầu, nhận thức, thái độ,…đặc biệt là “động cơ” có ảnh hưởng quan trọng đến quyết định chọn ngành của học sinh, các động cơ phải kể đến là “Khả năng có việc làm cao”, “phù hợp với năng lực học tập”. Theo đối tượng nghiên cứu thì nhân tố này được đánh giá là quan trọng nhất vì các em kỳ vọng nhiều vào ngành học mà mình đã chọn, các em tin rằng việc chọn ngành giúp họ chắc chắn sẽ tìm được việc làm sau khi tốt nghiệp ra trường để có nguồn thu thập ổn định đảm bảo cho những nhu cầu cơ bản của con người. Một yếu tố thuộc động cơ chọn ngành cũng không kém phần quan trọng là “sở thích ngành nghề”. Nếu việc lựa chọn đối với các thí sinh có học lực trung bình, là dựa trên “năng lực học tập” thì đối với các thí sinh có học lực khá - giỏi, là yếu tố quan trọng nhất trong việc theo đuổi con đường học vấn, cũng như thành công nghề nghiệp trong tương lai. Khi chọn trường đại học dự thi, động cơ thúc đẩy học sinh ra quyết định là nhân tố “Trường bạn thích”, “trường có nhiều ngành nghề”. Theo đối tượng nghiên cứu thì hai nhân tố này được đánh giá là quan trọng nhất. Qua đó, ta có thể kết luận về tính cách của các bạn học sinh THPT hiện nay là thích đề cao danh tiếng vì vậy họ chú trọng nhiều đến nhu cầu được quý trọng. Nhân tố có sự ảnh hưởng thấp nhất đến quyết định chọn trường là “thi theo bạn bè”. (trích dẫn nghiên cứu)

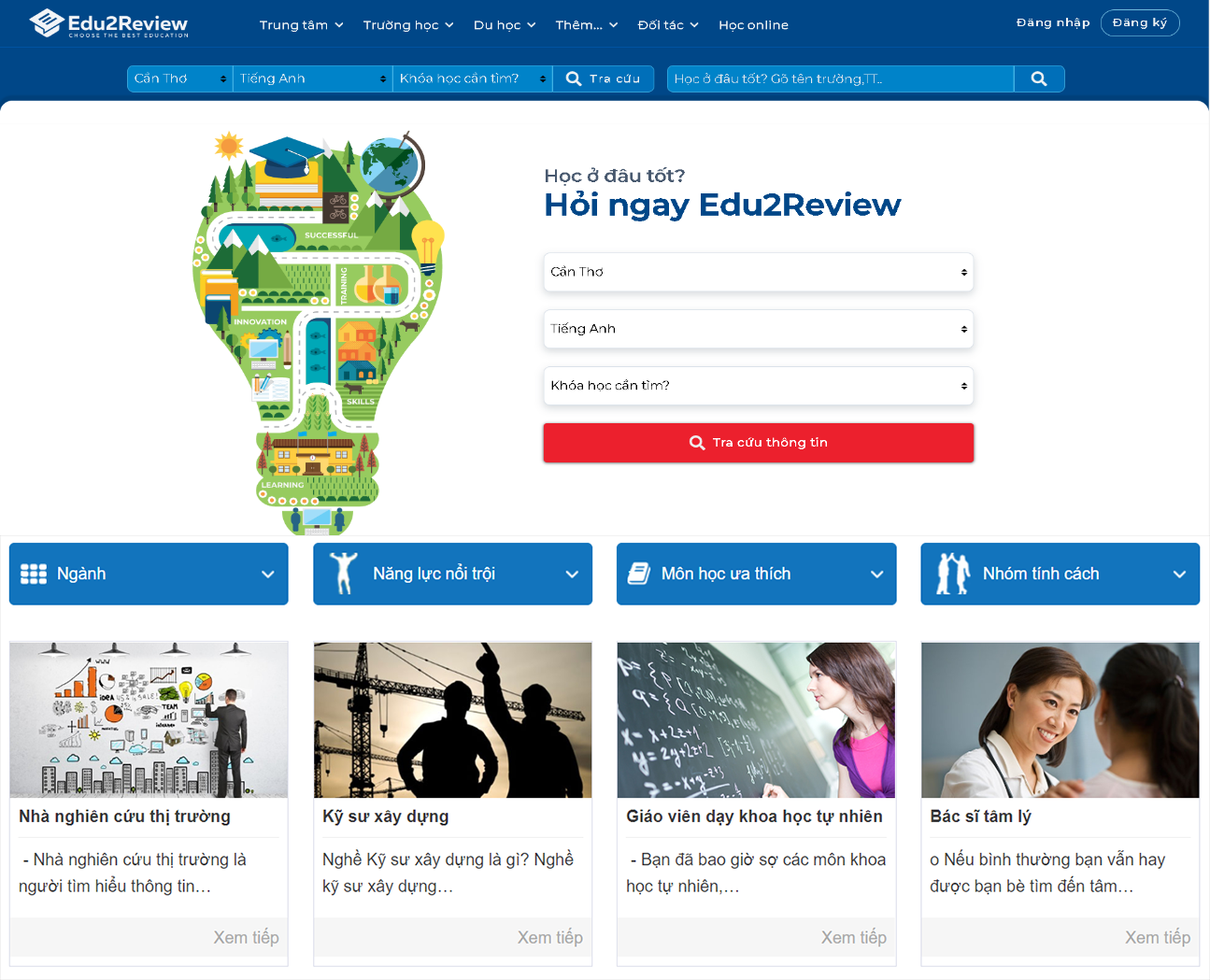
Theo công trình nghiên cứu “***Các yếu tố ảnh hưởng tới quyết định chọn trường đại học của học sinh THPT tại Việt Nam: Bằng chứng khảo sát năm 2020***”. Với số lượng mẫu khảo sát là 200 người, nghiên cứu đã xác định được 5 nhân tố ảnh hưởng đến quyết điịnh chọn trường đại học bao gồm: Yếu tố bản thân; thông tin và quảng cáo; ý kiến tham khảo; thương hiệu và việc làm; học phí và cơ sở vật chất. Kết quả mô hình cũng cho thấy những người tham gia khảo sát khá hài lòng với quyết định lựa chọn ngành học, trong đó các nhân tố công tác truyền thông quãng bá thông tin, chất lượng đào tạo, thương hiệu và sự tương thích phù hợp giữa học sinh với chương trình là yếu tố quan trọng trong việc lựa chọn ngành nghề phù hợp đối với học sinh THPT

1. **Các nghiên cứu ngoài nước**

Đề tài nghiên cứu ***“Factors affecting career development of senior secondary school students in Ethiope east local government area, Delta state, Nigeria*** (Các yếu tố ảnh hưởng đến sự phát triển nghề nghiệp của học sinh trung học phổ thông ở khu vực chính quyền địa phương phía đông Ethiope, bang Delta, Nigeria – Phòng hướng dẫn và tư vấn, Đại học ban Delta, University, năm 2010)**”** vớimục tiêu nghiên cứu là tìm hiểu yếu tố tác động đến việc lựa chọn ngành nghề của học sinh tại ban Delta. Và theo nghiên cứu trên, có 4 yếu tố tập chung ảnh hưởng đến quyết định lựa chọn nghề nghiệp bao gồm: yếu tố tâm lý (Sở thích, tính cách và lòng tự trọng), yếu tố xã hội học(bạn bè, người bảo hộ, chính sách nhà nước), yếu tố kinh tế(nghèo đói, đòi hỏi và mong muốn), yếu tố di truyền(ngoại hình và năng khiếu) và yếu tố giáo dục(vị trí trường học trong và ngoài nước và các kỹ năng thu được)

1. **Ứng dụng của máy học vào định hướng nghề nghiệp**

Thực tế hiện nay, việc ứng dụng khoa học, công nghệ thông tin vào giáo dục đang được phát triển rất phổ biến. Không khó để ta có thể tìm kiếm thông tin về trường học cũng như ngành học nào đó. Và đặc biệt hơn việc sử dụng công nghệ để giúp định hướng nghề nghiệp hiện đang là một xu hướng mà xã hội đang tiến đến. Có thể kể đến như: webisite ***Edu2Review***[số]… Ngoài giúp người đọc có cái nhìn mới về việc tìm hiểu thông tin giáo vụ tại trường bằng phương pháp sàng lọc thì ***Edu2Review*** còn có thể đưa ra gợi ý về ngành nghề mà học sinh có thể đăng ký dự thi. Nhưng website kể trên chỉ xoay quanh về việc gợi ý dựa theo chỉ số *sở thích, nguyện vọng gia đình* mà bỏ qua các yếu tố ảnh hưởng về khả năng, cụ thể là học lực từng cá thể học sinh, trong khi đó trình độ thực tế của học sinh là một trong những yếu tố quan trọng ảnh hưởng đến kết xuất với độ chính xác cao. Cho nên mặc dù đã đáp ứng đúng nhu cầu của xã hội nhưng tổng quan vẫn chưa hỗ trợ chính xác nhu cầu thực tế.



Hình 1: Trang web edu2review.com

1. **Mục tiêu đề tài**

*Đề tài nghiên cứu được thực hiện với mục tiêu:*

* Xác định và đánh giá yếu tố ảnh hưởng đến quyết định chọn trường của học sinh THPT trong kỳ thi tuyển sinh Đại học – Cao đẳng.
* Đưa ra gợi ý về ngành nghề phù hợp với các yếu tố của từng cá nhân góp phần nâng cao hiệu quả công tác hướng nghiệp cho học sinh lớp THPT trong việc quyết định lựa chọn ngành dự thi THPT quốc gia.

*Để đạt được mục tiêu đặt ra, nghiên cứu cần tập trung giải quyết các vấn đề sau:*

* Tiến hành khảo sát mẫu học sinh - sinh viên, thu thập thông tin các *yếu tố ảnh hưởng đến tác động đến quyết định chọn ngành của học sinh – sinh viên.*
* Lựa chọn giải thuật phù hợp đưa ra mô hình gợi ý.
* Phân tích sự khác biệt của các tác động nêu trên giữa các nhóm học sinh về đặc điểm của từng cá nhân.

1. **Đối tượng và phạm vi nghiên cứu**

***Đối tượng nghiên cứu:*** Nghiên cứu phương pháp để xây dựng hệ thống gợi ý, chủ yếu tập trung vào kỹ thuật lọc theo nội dung (Content Based). Bên cạnh đó, phương pháp lọc theo nội dung cũng sẽ sử dụng để xây dựng hệ thống gợi ý. Nghiên cứu công nghệ xây dụng website có khả năng tích hợp hệ thống gợi ý. Ngoài ra nghiên cứu đối tượng học sinh – sinh viên với mục tiêu tìm hiểu các yếu tố ảnh hưởng đến quyết định chọn ngành.

***Đối tượng được gợi ý***: học sinh THPT, là những đối tượng có nhu cầu tìm hiểu ngành nghề, cần được định hướng nghề nghiệp.

***Thời gian nghiên cứu:*** Khảo sát được tiến hành trong năm học 2020 – 2021.

1. **Phương pháp nghiên cứu**
2. **Phương pháp thu thập thông tin bằng phiếu câu hỏi:**

Tiến hành thu thập tài liệu từ nhiều nguồn khác nhau: các bài báo, các đề tài nghiên cứu và các tài liệu khác có liên quan. Thông qua phân tích tổng hợp lý thuyết, phân loại hệ thống hoá và khái quát hoá lý thuyết từ đó rút ra các kết luận khoa học là cơ sở lý luận cho đề tài.

Phiếu câu hỏi được thiết kế dựa theo mô hình nghiên cứu của đề tài nhằm thu thập thông tin, số lượng câu hỏi và các câu hỏi được chọn lọc dựa theo các ***yếu tố ảnh hưởng đến quyết định chọn ngành, nghề***. Các thông tin đưa vào phân tích nhầm cho ra kết quả với độ chính xác thực tế cao nhất.

1. **Phương pháp tiền xử lý dữ liệu và xử lý mất cân bằng dữ liệu:**

Phần lớn dữ liệu thu thập được sẽ theo một hướng đối tượng nghề nghiệp nhất định, trong khi đó mục tiêu đề tài hướng đến là đại đa số các học sinh – sinh viên đều sử dụng được trên tất cả các lĩnh vực ngành, nghề, nên tập dữ liệu sau khi thu thập sẽ mắc phải tình trạng mất cân bằng dữ liệu. Áp dụng những phương pháp tiền xử lý dữ liệu và xử lý mất cân bằng dữ liệu để có được tập dữ liệu hoàn chỉnh cho mô hình, gớp phần tăng độ tin cậy ở kết quả.

1. **Kiến thức**

* Tìm hiểu cách xây dựng hệ thống Website bằng Angular Framework với các ngôn ngữ lập trình HTML, CSS, Javascript… và các thư viện hỗ trợ liên quan.
* Tìm hiểu về cơ sở dữ liệu quan hệ PostgreSQL và hệ quản trị cơ sở dữ liệu Parse Server.
* Tìm hiểu giải thuật Demographic Filter và giải thuật Content Based đưa ra gợi ý ngành học.
* Ngoài ra, tìm kiếm các phương pháp xử lý mất cân bằng dữ liệu hỗ trợ bước tiền xử lý dữ liệu.

1. **Kết quả đạt được**

* Xây dựng thành công hệ thống Webservice quảng bá khoa CNTT & TT.
* Xây dựng mô hình gợi ý ngành học hỗ trợ học sinh trong việc lựa chọn ngành nghề phù hợp với sở thích và khả năng của từng cá nhân.

1. **Bố cục luận văn**

**Phần giới thiệu**

Giới thiệu tổng quát về đề tài gồm: đặt vấn đề, lịch sử giải quyết vấn đề, mục tiêu đề tài, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu, kết quả đạt được

**Phần nội dung**

**Chương 1** : Mô tả bài .

**Chương 2** : Thiết kế, cài đặt giải thuật, biễu diễn cơ sở dữ liệu, trình bày các bước xây dựng hệ thống bằng phương pháp lọc cộng tác.

**Chương 3** : Kiểm thử hệ thống và đánh giá độ chính xác.

**Phần kết luận**

Trình bày kết quả đạt được và hướng phát triển hệ thống

**PHẦN NỘI DUNG**

CHƯƠNG 1 CƠ SỞ LÝ THUYẾT

1. **Mô tả chi tiết bài toán**

Hiện nay ứng dụng Công nghệ thông tin vào giáo dục ngày càng phổ biến, Việc áp dụng máy học vào giáo dục đang được phát triển rộng rãi và có chiều sâu cụ thể là sử dụng hệ thống gợi ý vào đề suất các khóa học, sách và tài liệu học tập phù hợp với từng cá nhân học sinh – sinh viên. Mục đích chung của việc ứng dụng máy học vào giáo dục mang hướng cá nhân hóa, từ dữ liệu và giải thuật đưa ra gợi ý hữu ích đến với học sinh – sinh viên.

1. **Vấn đề và giải pháp liên quan đến bài toán** 
   1. **Giải pháp liên quan đến bài toán**

Trong nhiều năm qua đã có rất nhiều công trình nghiên cứu trên thế giới và ở Việt Nam đề cập đến vấn đề quyết định lựa chọn của học sinh vào việc chọn ngành/nghề để theo học. Sau đây là những công trình nghiên cứu nổi bật và giải pháp cho bài toán.

Ở công trình nghiên cứu ***“Phân tích các nhân tố ảnh hưởng đến quyết định chọn ngành, trường đại học của học sinh THPT vùng ĐBSCL”*** (Khoa Kinh tế, trường Đại học Cần Thơ năm 2017) được giới thiệu trong phần lịch sử giải quyết vấn đề. Với phương pháp thực hiện phân tích thống kê mô tả để đánh giá xu hướng hội tụ, những chỉ tiêu thống kê sử dụng phổ biến để miêu tả khuynh hướng hội tụ của một phân phối là số trung bình, trung vị và mode (trích dẫn báo cáo), trung bình số học của một mẫu được xác định theo công thức:

Trong đó:

**μ** là trung bình số học

là giá trị quan sát thứ i

**N** là số lượng quan sát tổng thể trên toàn mẫu

Ở nghiên cứu ***“Các yếu tố ảnh hưởng tới quyết định chọn trường đại học của học sinh THPT tại Việt Nam: Bằng chứng khảo sát năm 2020”*** nghiên cứu nhằm xác định và đánh giá mức độ ảnh hưởng của các yếu tố trong việc đưa ra quyết định chọn trường đại học của học sinh THPT. Thu thập dữ liệu bằng cách phỏng vấn trực tiếp kết hợp với bảng hỏi, trong bảng hỏi điều tra có 27 câu hỏi được phân trong 6 nhóm đề mục gồm: (1) nhà trường/ từ nhà trường, (2) đặc điểm bản thân, (3) tiếp cận – quảng bá thông tin, (4) tư vấn từ người xung quanh, (5) mức độ hài lòng, (6) thông tin cá nhân. Nhóm từ 1 đến 5 sử dụng thang đo với dãy giá trị từ 1 – 5 để đo lường mức độ cảm nhận của đối tượng khảo sát theo mức độ: hoàn toàn không đồng ý, không đồng ý, bình thường, đồng ý, hoàn toàn đồng ý. Với kích thước mẫu là 200 sinh viên, đối tượgn là sinh viên năm nhất và tiến hành thu thập trong khoảng thời gian tháng 2 và tháng 3 năm 2020. Sử dụng mô hình hồi quy tuyến tính ước lượng bình phương nhỏ nhất để nhận diện các yếu tố quan trọng ảnh hưởng đến mức độ hài lòng và chắn chắn về sự lựa chọn ngành nghề của sinh viên. *(trích dẫn tài liệu khảm khảo)*

1. **Giải pháp cho bài toán**

Theo các nghiên cứu trên, có rất nhiều yếu tố ảnh hưởng đến quyết định chọn ngành/nghề của học sinh THPT, Những mô hình nghiên cứu lý thuyết đã được kiểm nghiệm sẽ là cơ sở để hình thành mô hình thực nghiệm trong nghiên cứu này. Dựa vào cơ sở lý thuyết đã được các tác giả trên thế giới và Việt Nam phát triển, hoàn thiện và kết hợp với các yếu tố đặc trung của học sinh THPT tại Việt Nam, nghiên cứu này đề xuất 4 nhóm yếu tố ảnh hưởng đến quyết định chọn ngành/nghề. (1) nhóm yếu tố cá nhân, (2) nhóm chịu tác động từ các yếu tố bên ngoài, (3) nhóm năng lực học tập và cuối cùng (4) nhóm yếu tố quản bá.

* **Nhóm (1)** bao gồm các nhân tố cá nhân: Giới tính, sở thích cá nhân, ngành nghề có đúng với nguyện vọng của học sinh hay không, mức thu nhập cá nhân
* **Nhóm (2)** bao gồm các yếu tố chịu tác động từ các yếu tố bên ngoài: Mức độ nổi tiếng của trường học có ảnh hưởng đến lựa chọn, yếu tố này liên quan đến các yếu tố con (trang thiết bị trường học, trình độ giảng dạy, học phí …), tác động từ gia đình-thầy cô-bạn bè, tác động từ các yếu tố xã hội, mức độ cần thiết của nghề nghiệp trong tương lai.
* **Nhóm (3)** bao gồm các yếu tố liên quan đến năng lực học tập thực tế của mỗi học sinh: điểm số các môn học và học lực cụ thể dựa trên điểm số.
* **Nhóm (4)** bao gồm các yếu tố tìm hiểu thông tin: Mức độ tìm hiểu ngành nghề, các thông tin ngành nghề tìm hiểu được từ đâu.
  1. **Thu thập dữ liệu**

Dựa vào các giả thuyết trên, tiến hành thu thập số liệu trong thời gian trong năm 2020. Phương pháp thu thập sử dụng biểu mẫu câu hỏi trực tuyến với đối tượng nghiên cứu là sinh viên thuộc các tỉnh Đồng bằng sông Cửu Long, trong biểu mẫu có tổng 19 câu hỏi thuộc 4 nhóm yếu tố trên, hình dưới là nội dung câu hỏi trong biểu mẫu.

(thu thập dữ liệu)

* 1. **Kết quả thu thập dữ liệu**

Kết quả thu thập được 1066 lượt trả lời câu hỏi từ các sinh viên đang học tập tại các trường Đại học- Cao đẳng- Nghề, tập dữ liệu có tên **dataset.**

1. **Giới thiệu tập dữ liệu**
   1. **Giới thiệu tập dữ liệu**

Tập dữ liệu **dataset** được thu thập từ sinh viên ở các trường Đại học, Cao đẳng, Trung cấp thuộc các tỉnh Đồng bằng sông Cửu Long. Nội dung dung thu thập bao gồm các thông tin về sở thích, nguyện vọng, các yếu tố liên quan đến việc quyết định chọn ngành. Dữ liệu có được bằng cách lập biểu mẫu các câu hỏi và giá trị dữ liệu là các câu trả lời, ý kiến phản hồi từ đối tượng, thời gian thu thập dữ liệu từ năm 2019 đến năm 2020.

* 1. **Mô tả tập dữ liệu**

***Kích thước:*** Tập dữ liệu gồm 1062 dòng và 13 cột, số dòng tương ứng với lượt trả lời câu hỏi từ người dùng, số cột tương ứng với số lượng câu hỏi đặt ra trong biểu mẫu cũng là số thuộc tính có trong tập dữ liệu.

***Nhãn***: 1 cột ( Major ).

***Thuộc tính:*** 12 cột (Gender – Interest – RFCM – Time – Math\_score – Physics\_score – Chemistry\_score – English\_score – Literature\_score – History\_score – Geography\_score – Influence ).

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Mô tả** | **Ví dụ** |
| **Gender** | Giới tính | Nam hoặc Nữ |
| **Interest** | Sở thích | Du lịch, Phim ảnh, Âm nhạc, …. |
| **RFCM** | Lý do chọn Ngành/nghề | Năng lực bản thân hoặc sở thích hoặc nguyện vọng gia đình … |
| **Time** | Thời gian tìm hiểu ngành/nghề | Ít hoặc vừa phải hoặc nhiều … |
| **Math\_score** | Điểm môn Toán | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Physics\_score** | Điểm môn Vật lý | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Chemistry\_score** | Điểm môn Hóa học | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **English\_score** | Điểm môn Ngoại ngữ | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Literature\_score** | Điểm môn Ngữ văn | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **History\_score** | Điểm môn Lịch sử | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Geography\_score** | Điểm môn Địa lý | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Influence** | Mức độ nổi tiếng của trường đại học có ảnh hưởng đến quyết định chọn ngành/nghề | ảnh hưởng ít hoặc ảnh hưởng nhiều … |

Bảng 2: Các thuộc tính của tập dữ liệu dataset

1. **Tiền xử lý dữ liệu**
   1. **Giới thiệu tiền xử lý dữ liệu**

Trong thực tế, công việc xử lý dữ liệu trước khi đưa vào mô hình là rất cần thiết, bước này làm cho dữ liệu có được ban đầu qua thu thập dữ liệu (gọi là dữ liệu gốc ) có thể áp dụng được với các mô hình học máy. Các kỹ thuật tiền xử lý dữ liệu phổ biến hiện nay bao gồm: Xử lý dữ liệu bị khuyết (Missing data), mã hóa các biến nhóm (ending categorical varibles), chuẩn hóa dữ liệu (standardization data), co giãn dữ liệu (scaling data)…

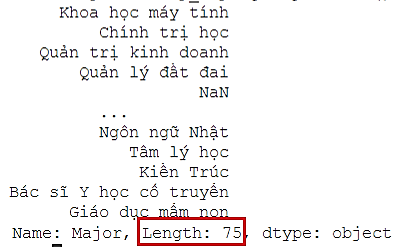
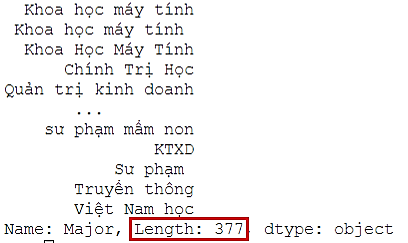
***Chuẩn hóa dữ liệu (Standardization data)***  là một trong những bước tiền xử lý dữ liệu mục tiêu ở bước này là đảm bảo rằng dữ liệu nhất quán nội bộ, có nghĩa là mỗi kiểu dữ liệu có cùng nội dung định dạng, Các giá trị được chuẩn hóa theo mục đích của yêu cầu bài toán có nhiều cách để chuẩn hóa hóa dữ liệu, ví dụ như: chuyển đổi dữ liệu nam – nữ sang giá trị 0 – 1, Chuyển đổi kiểu dữ liệu liên tục sang kiểu dữ liệu rời rạc,….[vimentor]

***Co giãn dữ liệu (Scaling data)*** là một phương pháp chuẩn hóa phạm vi của các đặc trưng dữ liệu và được thực hiện trong suốt quá trình tiền xử lý dữ liệu. Phạm vi của dữ liệu thô rất rộng, trong khi đối với một số thuật toán học máy, các hàm mục tiêu sẽ không hoạt động đúng như yêu cầu bài toán đặt ra, vậy nên để dữ liệu có thể hoạt động đúng mục đích thì cần phải áp dụng các giải thuật toán học để giải tuyết. Tiêu biểu là những công thức tính khoảng cách giữa các vector, các điểm dữ liệu góp phần vào quá trình tính toán độ dài, từ đó đưa ra phương pháp co giãn độ dài phù hợp. Các giải thuật phổ biến thường được sử dụng trong việc co giãn dữ liệu là ***KNN (K Nearest Neighbors), K-Means (K- Means Cluster), SVM (Support Vector Machine)…***[vimentor][7]

**Áp dụng chuẩn hóa dữ liệu vào bài toán:**

***Loại bỏ thuộc tính Null, dư thừa, gây nhiễu:*** Tập dữ liệu gốc bao gồm những giá trị dư thừa trong quá trình thu thập, loại bỏ các giá trị mà bài toán không yêu cầu.

***Chuẩn hóa các giá trị nhãn:*** dữ liệu được thu thập qua biểu mẫu online, và dữ liệu về ngành nghề dã chọn (nhãn) là loại được phép nhập trên biểu mẫu, Nên các giá trị này không thống nhất, nghĩa là cùng một đáp án nhưng lại có nhiều cách nhập khác nhau, ví dụ như: “Khoa học máy tính”, “Khoa Học Máy Tính”,, “ Khoa học máy tính ”,… Để dữ liệu có thể sử dụng được ta phải chuẩn hóa về dạng thống nhất. giá trị nhãn ban đầu chưa qua chuẩn hóa là 377 loại, sau khi chuẩn hóa kết quả còn 75 loại và đó cũng là số lượng nhãn của tập dữ liệu. Hình bên dưới thể hiện quá trình chuẩn hóa giá trị nhãn (ngành học)



***Chuẩn hóa dữ liệu Gender (Giới tính):*** chuyển đổi giá trị dữ liệu Nam – Nữ về 2 giá trị 0 – 1. Sau khi chuẩn hóa ta được tập dữ liệu với cột thuộc tính Gender có giá trị là 0 – 1 tương ứng với 0 là giá trị Nam, 1 là giá trị nữ.

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính**  **tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Nam** | 0 | 410 | 1877 |
| **Nữ** | 1 | 584 | 2798 |

***Chuẩn hóa dữ liệu RFCM (Nguyên nhân tác động lớn nhất đến việc chọn ngành học):*** chuyển đổi giá trị thuộc tính ở cột RFCM (Sở thích, Năng lực bản thân, Nguyện vọng gia đình, nhu cầu cao của thị trường, Yêu thích và tư vấn từ gia đình, Mức thu thập của gia đình) thành các giá trị số (lần lượt là: 0, 1, 2 ,3 ,4 ,5).

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính**  **tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Mức thu nhập của gia đình** | 0 | 38 | 179 |
| **Nguyện vọng gia đình** | 1 | 129 | 599 |
| **Nhu cầu cao của thị trường** | 2 | 225 | 1052 |
| **Năng lực bản thân** | 3 | 212 | 1000 |
| **Sở thích** | 4 | 345 | 1680 |
| **Tư vấn từ thầy cô bạn bè** | 5 | 45 | 165 |

***Chuẩn hóa dữ liệu Influence (Mức độ nổi tiếng của trường đại học có ảnh htưởng đến quyết định chọn ngành ):*** chuyển đổi giá trị thuộc tính ở cột RFCM (Không ảnh hưởng, Ảnh hưởng ít, Khá ảnh hưởng, Ảnh hưởng lớn, Ảnh hưởng rất lớn ) sang các giá trị số (lần lượt là: 1, 2, 3, 4, 5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Không ảnh hưởng** | 1 | 71 | 317 |
| **Ảnh hưởng ít** | 2 | 99 | 311 |
| **Khá ảnh hưởng** | 3 | 241 | 950 |
| **Ảnh hưởng lớn** | 4 | 377 | 2078 |
| **Ảnh hưởng rất lớn** | 5 | 206 | 1019 |

***Chuẩn hóa dữ liệu Time (Thời gian tìm hiểu ngành học):*** chuyển đổi giá trị thuộc tính ở cột Time (Rất ít, Ít, Vừa phải, Nhiều, Rất nhiều) sang các giá trị số ( lần lượt là: 1, 2, 3, 4, 5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Rất ít** | 1 | 56 | 218 |
| **ít** | 2 | 85 | 349 |
| **Vừa phải** | 3 | 307 | 1323 |
| **Nhiều** | 4 | 383 | 2070 |
| **Rất nhiều** | 5 | 163 | 715 |

***Chuẩn hóa dữ liệu Maths\_score ( điểm toán)***: chuyển đổi giá trị thuộc tính ở cột Maths\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 16 | 35 |
| **5 – 6.5** | 5.75 | 41 | 160 |
| **6.5 – 8** | 7.25 | 304 | 1466 |
| **8 – 9** | 8.5 | 456 | 2141 |
| **9 – 10** | 9.5 | 165 | 741 |
| **Điểm số do người dùng nhập vào** | 8.0 | 1 | 3 |
|  | 6.8 | 2 | 14 |
|  | 8.3 | 2 | 7 |
|  | 9.7 | 2 | 7 |
|  | 6.9 | 1 | 3 |
|  | 9.2 | 1 | 2 |
|  | 7.1 | 1 | 1 |
|  | 8.4 | 1 | 1 |
|  | 10.0 | 1 | 2 |
|  | 7.6 | 1 | 9 |
|  | 9.6 | 1 | 17 |
|  | 8.8 | 1 | 3 |
|  | 8.9 | 1 | 10 |
|  | 9.4 | 1 | 0 |
|  | 5.0 | 0 | 53 |

***Chuẩn hóa dữ liệu Physics\_score ( điểm vật lý):*** chuyển đổi giá trị thuộc tính ở cột Physics\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 19 | 52 |
| **5 - 6.5** | 5.75 | 90 | 291 |
| **6.5 - 8** | 7.25 | 374 | 1826 |
| **8 - 9** | 8.5 | 397 | 1889 |
| **9 - 10** | 9.5 | 135 | 469 |
| **Điểm số do người dùng nhập vào** | 5.0 | 0 | 91 |
|  | 8.8 | 1 | 17 |
|  | 6.5 | 1 | 13 |
|  | 8.4 | 1 | 10 |
|  | 8.7 | 1 | 6 |
|  | 7.9 | 1 | 3 |
|  | 6.6 | 1 | 3 |
|  | 9.0 | 1 | 2 |
|  | 6.3 | 1 | 1 |
|  | 7.4 | 1 | 1 |
|  | 7.1 | 1 | 1 |

***Chuẩn hóa dữ liệu Chemistry\_score ( điểm Hóa học )***: chuyển đổi giá trị thuộc tính ở cột Chemistry\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 27 | 55 |
| **5 - 6.5** | 5.75 | 98 | 368 |
| **6.5 - 8** | 7.25 | 371 | 1685 |
| **8 - 9** | 8.5 | 383 | 1746 |
| **9 - 10** | 9.5 | 140 | 700 |
| **Điểm số do người dùng nhập vào** | 5.0 | 0 | 67 |
|  | 9.4 | 1 | 17 |
|  | 5.4 | 1 | 13 |
|  | 7.0 | 1 | 6 |
|  | 9.8 | 1 | 4 |
|  | 7.5 | 1 | 3 |
|  | 8.8 | 1 | 3 |
|  | 9.6 | 1 | 3 |
|  | 9.1 | 1 | 2 |
|  | 5.7 | 1 | 1 |
|  | 6.5 | 1 | 1 |
|  | 9.2 | 1 | 1 |

***Chuẩn hóa dữ liệu English\_score ( điểm Ngoại ngữ )***: chuyển đổi giá trị thuộc tính ở cột English\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 24 | 90 |
| **5 - 6.5** | 5.75 | 149 | 586 |
| **6.5 - 8** | 7.25 | 428 | 2054 |
| **8 - 9** | 8.5 | 310 | 1437 |
| **9 - 10** | 9.5 | 102 | 386 |
| **Điểm số do người dùng nhập vào** | 6.2 | 1 | 13 |
|  | 7.1 | 1 | 6 |
|  | 7.6 | 1 | 1 |
|  | 9.4 | 1 | 3 |
|  | 8.6 | 1 | 17 |
|  | 8.8 | 1 | 2 |
|  | 7.4 | 1 | 10 |
|  | 5.0 | 1 | 0 |
|  | 9.2 | 1 | 0 |
|  | 6.7 | 1 | 3 |
|  | 5.0 | 1 | 67 |

***Chuẩn hóa dữ liệu Literature\_score ( điểm Ngữ văn )***: chuyển đổi giá trị thuộc tính ở cột Literature\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 21 | 83 |
| **5 - 6.5** | 5.75 | 112 | 478 |
| **6.5 - 8** | 7.25 | 505 | 2467 |
| **8 - 9** | 8.5 | 280 | 1358 |
| **9 - 10** | 9.5 | 37 | 167 |
| **Điểm số do người dùng nhập vào** | 5.0 | 28 | 62 |
|  | 7.8 | 1 | 18 |
|  | 7.4 | 1 | 13 |
|  | 7.2 | 1 | 10 |
|  | 6.5 | 2 | 9 |
|  | 8.7 | 1 | 3 |
|  | 8.1 | 1 | 3 |
|  | 7.6 | 1 | 2 |
|  | 7.9 | 1 | 1 |
|  | 5.6 | 1 | 1 |

***Chuẩn hóa dữ liệu History\_score ( điểm Lịch Sử ):*** chuyển đổi giá trị thuộc tính ở cột History\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 14 | 26 |
| **5 - 6.5** | 5.75 | 59 | 174 |
| **6.5 - 8** | 7.25 | 377 | 1896 |
| **8 - 9** | 8.5 | 335 | 1596 |
| **9 - 10** | 9.5 | 166 | 795 |
| **Điểm số do người dùng nhập vào** | 5.0 | 33 | 129 |
|  | 8.9 | 2 | 8 |
|  | 7.5 | 2 | 4 |
|  | 9.7 | 1 | 3 |
|  | 8.4 | 1 | 13 |
|  | 8.7 | 1 | 10 |
|  | 8.8 | 1 | 3 |
|  | 7.7 | 1 | 1 |
|  | 8.6 | 1 | 17 |

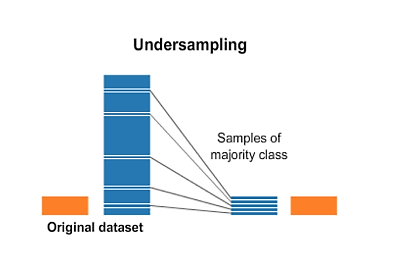
***Chuẩn hóa dữ liệu Geography\_score ( điểm Địa lý ):*** chuyển đổi giá trị thuộc tính ở cột Literature\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

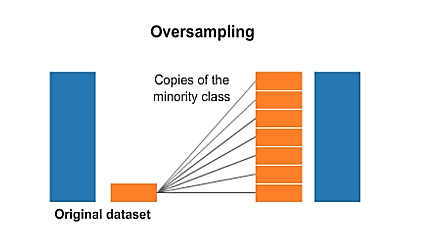
|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 16 | 29 |
| **5 - 6.5** | 5.75 | 40 | 114 |
| **6.5 - 8** | 7.25 | 336 | 1658 |
| **8 - 9** | 8.5 | 389 | 1906 |
| **9 - 10** | 9.5 | 177 | 832 |
| **Điểm số do người dùng nhập vào** | 5.0 | 26 | 80 |
|  | 9.2 | 3 | 19 |
|  | 8.6 | 1 | 17 |
|  | 7.6 | 2 | 14 |
|  | 7.8 | 2 | 4 |
|  | 9.1 | 1 | 2 |
|  | 8.5 | 1 | 0 |

* 1. **Xử lý mất cân bằng dữ liệu**

Mất cân bằng dữ liệu là một trong những hiện tượng phổ biến trong quá trình thu thập dữ liệu, Mất cân bằng dữ liệu là khi dữ liệu bị thiếu hụt ở một số nhóm nhất định, ví dụ: Khi thu thập dữ liệu ngành học, tỷ lệ dữ liệu thu thập được giữa sinh viên Công nghệ thông tin và sinh viên Kinh tế có sự chênh lệch đáng kể là 80:20. Dữ liệu sẽ được xem là cân bằng nếu tỷ lệ đạt được 50:50, hầu hết các bộ dữ liệu đều khó đạt được trạng thái cân bằng mà luôn có sự khác biệt về tỷ lệ giữa 2 hoặc nhiều nhóm trong bộ dữ liệu. Đối với những trường hợp có tỷ lệ chênh lệch dữ liệu là 60:40, mặc dù vần được xem là mất cân bằng nhưng không ảnh hưởng đáng kể đến khả năng dự báo mô hình. Tuy nhiên nếu xảy ra hiện tượng mất cân bằng, chẳng hạn như tỷ lệ dữ liệu 90:10 thì chất lượng kết quả mô hình so với độ chính xác thực tế sẽ sai lệch nghiêm trọng. Giải pháp khắc phục hiện tượng mất cân bằng dữ liệu hiện đang được nhiều nhà nghiên cứu hướng đến là **Over Sampling** và **Under Sampling**. [8]

*Under Sampling* là việc gia giảm số lượng quan sat của nhóm có tỷ lệ dữ liệu trội hơn, nhiều hơn hoặc những nhóm có lượng dữ liệu tăng một cách đột biến … để dữ liệu số lượng nhiều chở nên cân bằng so với nhóm dữ liệu thiểu số. Ưu điểm của giải thuật này là làm cân bằng dữ liệu một cách nhanh chóng, dể dàng thực hiện. Nhưng nhược điểm ở giải thuật này là kích thước tập dữ liệu sẽ bị giảm đáng kể. Giả sử tỷ lệ giữa sinh viên Công nghệ thông tin và sinh viên Kinh tế là 1000:500 (sinh viên), nếu sử dụng Under Sampling, sau khi cân bằng dữ liệu sẽ bị giảm 500 (sinh viên). Dẫn đến tình trạng tập dữ liệu huấn luyện sẽ bị thu hẹp. Cho nên giải thuật Under Sampling sẽ không phù hợp nếu áp dụng với bộ dữ liệu có kích thước nhỏ.

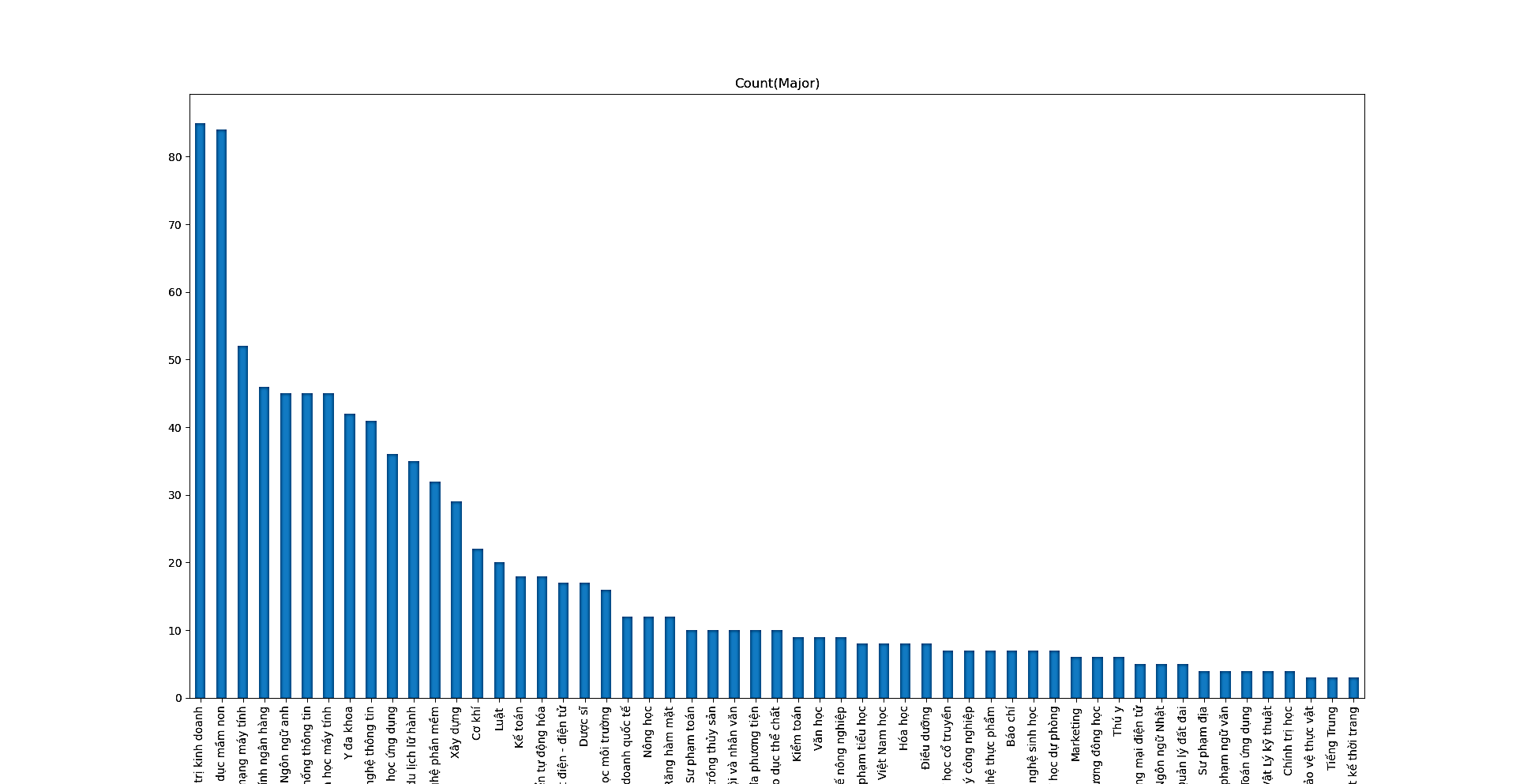


*Over Sampling* là phương pháp giúp giải quyết hiện tượng mất cân bằng mẫu bằng cách gia tăng kích thước mẫu thuộc nhóm dữ liệu thiểu số sao cho dữ liệu thiểu số cân bằng so với dữ liệu có số lượng đa số, cách làm này khắc phục nhược điểm so với giải thuật Under Sampling là có thể áp dụng cho tập dữ liệu có kích thước nhỏ, ngoài ra còn mở rộng quy mô dữ liệu lớn hơn để mô hình đạt được độ chính xác cao trên nhóm dữ liệu thiểu số, hình ảnh dưới đây minh họa cho phương pháp xử lý mất cân bằng dữ liệu Over Sampling.

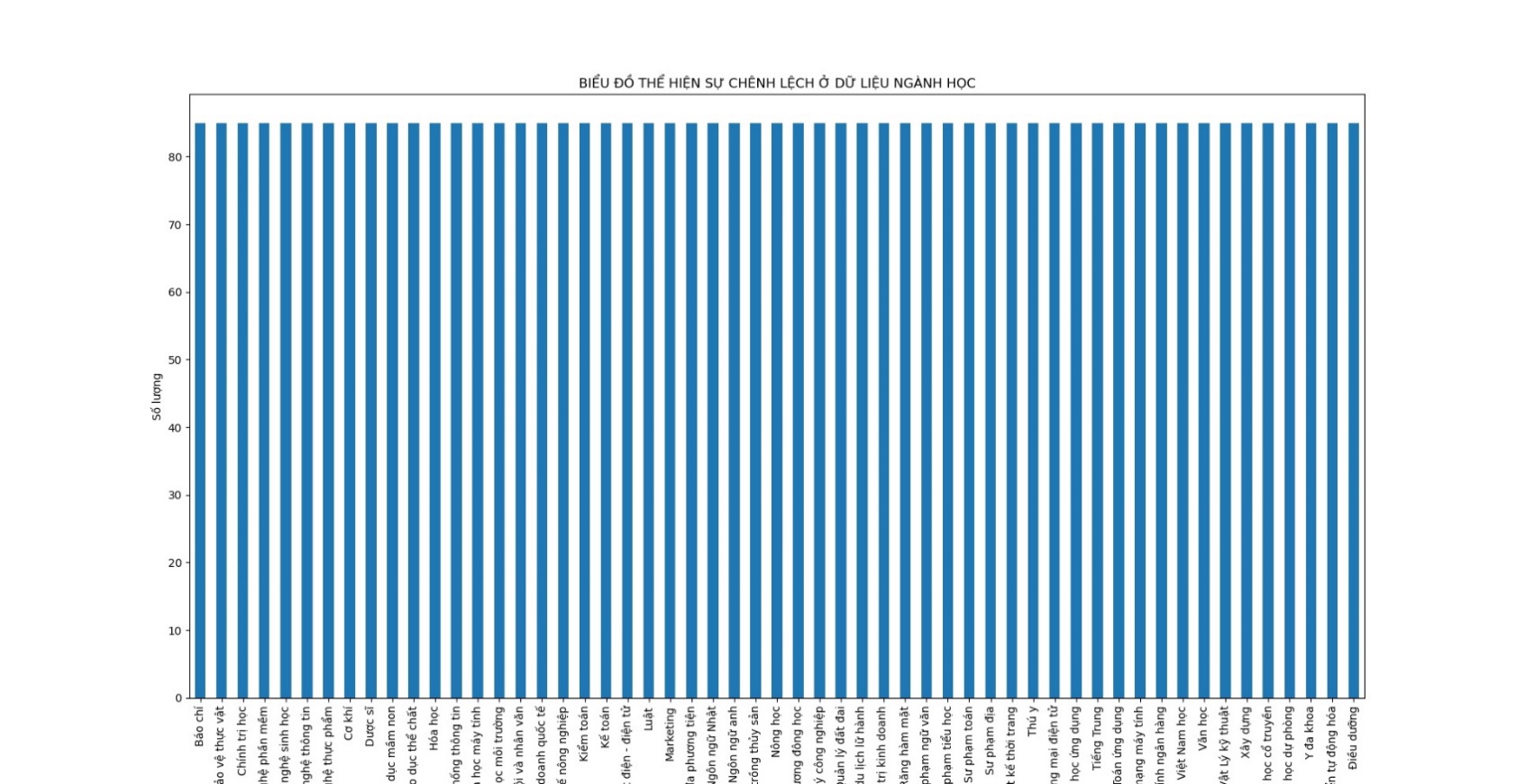
**Áp dụng Oversampling vào bài toán cụ thể:**

OverSampling – Random Forest: ý tưởng ở phương pháp này là tạo ra giá trị thuộc tính mới dựa vào giá trị thuộc tính gốc bằng cách áp dụng giải thuật Random Forest. Đặc điểm ở phương pháp này là giá trị thuộc tính mới sẽ được tăng ngẫu nhiên.

***Trước cân bằng:***

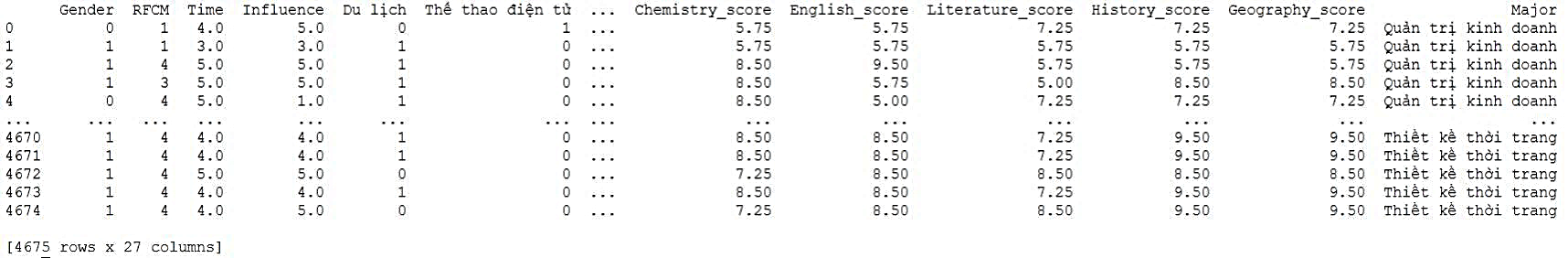


***Sau cân bằng:***

**

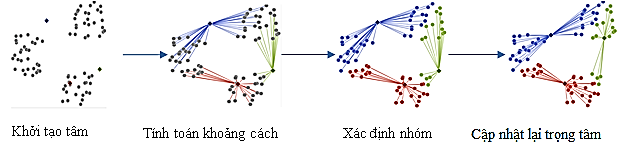
**Kết quả cân bằng dữ liệu:** sau quá trình cân bằng dữ liệu, bộ dữ liệu đã có những thay đổi sau, các giá trị được thể hiện ở hình bên dưới:

* Kích thước bộ dữ liệu từ 996 dòng mở rộng ra 4675 dòng và 27 cột.
* Giá trị thuộc tính cột nhãn được cân bằng, số lượng mỗi giá trị nhãn đồng giá trị là 85.
* Các giá trị thuộc tinh sau cân bằng được thể hiện ở bảng…..



1. **Lọc nhân khẩu học (Demographic Filter)**

Lọc nhân khẩu học (DF): nhằm mục đích phân loại người dùng dựa trên các thuộc tính cá nhân hổ trợ cho hệ thống gợi ý đưa ra các đề xuất dựa trên các nhóm người dùng đã phân loại. Mục tiêu ở giải thuật này là tìm ra các nhóm người dùng cùng thích một sản phẩm nào nhất. Ví dụ: nhóm người dùng nhóm C thích một sản phẩm s, người dùng C1 thuộc nhóm người dùng C chưa biết đến sản phẩm s, thì sản phẩm s này có thể gợi ý cho người dùng C1. Có nhiều phương pháp để xác định và phân nhóm dữ liệu như: SVM (Support vector machine), KNN (K – Nearest neighbor ), K-Means Clustering … Tùy vào tập dữ liệu mà mỗi phương pháp đều có ưu điểm riêng, được sử dụng phổ biến là giái thuật K-Means do giải thuật này không cần biết giá trị nhãn, ý tưởng chính là xác đinh k trung tâm, với k là số lượng nhóm cố định cho trước, sau đó tính toán khoảng cách giữa các giá trị thuộc tính so với giá trị k trung tâm, lập lại nhiều lần, sau đó sinh ra nhóm dữ liệu, các bước của giải thuật được minh họa ở hình xxx bên dưới.

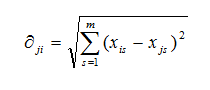


Các bước thực hiện của giải thuật K – means Clustering như sau: giả sử có tập dữ liệu với n phần tử, mỗi phần tử có m thuộc tính. Số chiều của tọa độ được xác định dựa trên số thuộc tính, xem mỗi thuộc tính của đối tượng như một tọa độ của không gian m chiều và biểu diễn đối tượng như một điểm của không gian m chiều.

Bước 1:Xác định số cụm k cho tập dữ liệu ban đầu gồm n phần tử

Bước 2: khởi tạo số điểm trung tâm bằng với số lượng k

Bước 3: Đối với mỗi đối tượng trong tập dữ liệu, tính toán khảng cách đến trọng tâm của các cụm bằng công thức sau:



Với các giá trị biểu thức:

: là đối tượng thứ i cần được tính khoảng cách và phân loại

: là đối tượng trung tâm cụm j

: là thuộc tính thứ s của nhóm

: là thuộc tính thứ s của đối tượng trung tâm của cụm

Bước 4: Cập nhật lại tâm của đối tượng trung tâm k (k nhóm) bằng phương pháp tính trung bình cộng vector của các đối tượng trong mỗi cụm. giá trị trọng tâm được xác định bằng công thức sau:

Với các giá trị biểu thức:

: là tọa độ thứ j của đối tượng trung tâm nhóm I ( cần cập nhật lại)

: là thuộc tính thứ j của phần tử s

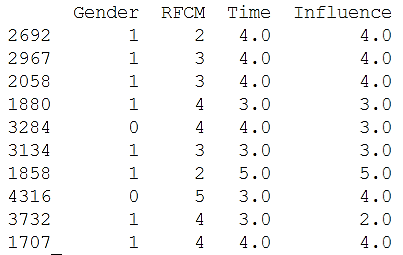
***t*:** là số phần tử hiện có ở nhóm i

Bước 5: Lặp lại các bước 3 và bước 4 cho đến khi các trọng tâm của cụm không có sự thay đổi.

***Áp dụng giải thuật K-Means để xác định nhóm người dùng trên tập dữ liệu dataset.***

Thuật toán K-Means clustering là một trong những thuật toán học không giám sát cơ đơn giản nhất để giải quyết vấn đề phân cụm. với mục đích chính là phân tập dữ liệu định sẵn thành k cụm với k là số cho trước. Kết quả của quá trình phân cụm là k nhóm dữ liệu, áp dụng thuật toán K-Means vào bài toán được mô tả trong ví dụ bên dưới.

***Ví dụ:*** Tập dữ liệu có 10 dòng và 4 cột, 10 dòng tương ứng với 10 đối tượng, 4 cột tương ứng với số thuộc tính và sẽ có 4 chiều dữ liệu. Phân lớp tập dữ liệu.



Lấy trọng tâm là 2 điểm: 2692[1; 2; 4; 5] và 3134[1; 3; 3; 3] (phân thành 2 lớp dữ liệu)

Lần lập 1:

Áp dụng công thức [CT]: tính khoảng cách giữa các điểm so với trọng tâm 2692:

KC[2692:2967] = = 1

Tương tự ta có khoảng cách:

KC[2692:2085] = 1

KC[2692:1880] = 2.449

KC[2692:3284] = 2.449

KC[2692:1858] = 1.414

KC[2692:4316] = 3.316

KC[2692:3732] = 3

KC[2692:1707] = 2

***Áp dụng công thức [ơss]: tính khoảng cách giữa các điểm so với trọng tâm 3134:***

KC[3134:2967] = = 1.414

Tương tự ta có khoảng cách:

KC[3134:2058] = 1.414

KC[3134:1880] = 1

KC[3134:3284] = 2.449

KC[3134:1858] = 3

KC[3134:4316] = 2.449

KC[3134:3732] = 1.414

KC[3134:1707] = 1.732

Hình dưới thể thống kê các giá trị khoảng cách giữa các điểm so với 2 trọng tâm, so sánh độ lớn khoảng cách các điểm so với trọng tâm, nếu độ lớn nhỏ tương ứng với khoảng cách giữa điểm đó so với trọng tâm nhỏ và ngược lại, các bước thực hiện trong lần lập 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Đối tượng** | **Khoảng cách Euclidean** | | **Lần lập 1** | |
| **Trọng tâm**  **2692** | **Trọng tâm**  **3134** | **Cụm 1 gần 2692** | **Cụm 2 gần 3134** |
| **2967** | 1 | 1.414 | **x** |  |
| **2058** | 1 | 1.414 | **x** |  |
| **1880** | 2.449 | 1 |  | **y** |
| **3284** | 2.449 | 2.449 | x | y |
| **1858** | 1.414 | 3 | x |  |
| **4316** | 3.316 | 2.449 |  | y |
| **3732** | 3 | 1.414 |  | y |
| **1707** | 2 | 1.732 |  | y |

***Sau khi tính khoảng cách các điểm và xác định các điểm gần trọng tâm, cập nhật lại vị trí trọng tâm bằng công thức [CT]***

*Cập nhật trọng tâm cụm 1 gần điểm 2692:*

*Điểm mới là: c\_1[0.6, 4.2, 3.4, 3.2]*

*Cập nhật trọng tâm cụm 1 gần điểm 3134:*

*Điểm mới là: c\_2[0.75, 3, 3.5, 4]*

Sau khi có được điểm trọng tâm mới, tiếp tục tính khoản cách giữa tất cả các điểm so với điểm c1 và c2 và cập nhật lại trọng tâm c1n và c2n cho đến khi có được kêt quả phân lớp phù hợp với bài toán.

**Áp dụng trong bài toán nghiên cứu:**

Trong nghiên cứu này, sử dụng phương pháp K-Means Clustering với số cụm là 12 thuộc tính dùng để tính toán và huấn luyện bao gồm 4 cột: Gender – RFCM – Time – Influence, với số mục dữ liệu là 3740 dòng. Kết quả huấn luyện mô hình có được 12 nhóm dữ liệu người dùng dựa vào 4 thuộc tính là Gender, RFCM, Time, Infuence.

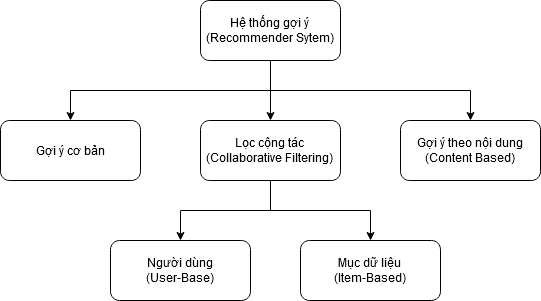
1. **Khái quát về hệ thống gợi ý (Recommender System – RS )**

***Lọc thông tin (Information Filtering)*** là lĩnh vực nghiên cứu quá trình phân tích thông tin thích hợp, ngắn ngừa và loại bỏ những thông tin không phù hợp với người dùng.

***Khái quát:*** Hệ thống gợi ý (Recommender System - RS) là một dạng của hệ thống lọc thông tin, với mục tiêu chính là lọc thông tin quan trọng ra khỏi một khối lượng lớn thông tin, nó được sử dụng để dự đoán sở thích (preference), cung cấp giải pháp phù hợp từng từng đối tượng, nhằm gợi ý các mục thông tin phù hợp với người dùng [5]. Hệ thống gợi ý được định nghĩa như một trợ lý giúp người dùng đưa ra quyết định trong môt trường thông tin phức tạp, không khó có thể bắt gặp những tính năng gợi ý trên các trang mạng xã hội, thương mại điện tử,… Chúng có lợi cho cả nhà cung cấp và người tiêu dùng, giúp giảm chi phí tìm kiếm và lựa chọn các mặt hàng trong môi trường mua sắm trực tuyến. Ngoài ra, hệ thống gợi ý giúp cho người dùng dể dàng tìm kiếm thông tin bằng cách đưa ra những từ khóa có liên quan đến nội dung tìm kiếm, hổ trợ quá trình tìm kiếm thuận tiện hơn.[số]

***Hệ thống gợi ý trong giáo dục:*** Những năm gần đây việc ứng dụng HTGY vào giáo dục đang dần chở nên phổ biến và ngày càng phát triển thêm nhiều tính năng hấp dẫn hơn, việc ứng dụng HTGY vào giáo dục giúp công tác giảng dạy trở nên thuận tiện hơn, Những ứng dụng nổi bật trong giáo dục như: Hỗ trợ sinh viên tìm kiếm tài liệu bằng cách đưa ra những gợi ý có liên quan đến nội dung tài liệu mà sinh viên tìm kiếm, Đưa ra gợi ý những khóa học phù hợp với chuyên ngành của sinh viên dựa theo lượt đánh giá cao từ người dùng trước đó.

**Phân loại**

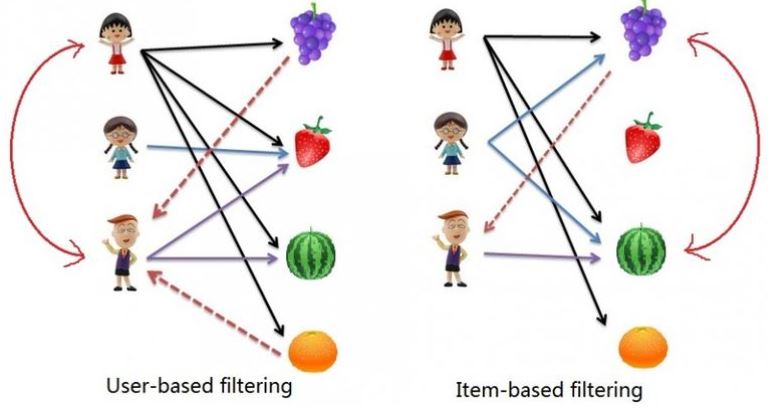


Hình 2: Sơ đồ phân loại hệ thống gợi ý

1. **Gợi ý cơ bản**

Gợi ý cơ bản là những gợi ý ban đầu được thực hiện mà chưa đề cập đến tương tác giữa người dùng với các mục dữ liệu (items) trong hệ thống. Gợi ý được biểu diễn bằng một giá trị xếp hạng trung bình tổng thể thì mô hình dự báo ban đầu kết hợp giữa người dùng và mục dữ liệu.

1. **Gợi ý theo lọc cộng tác (Collaborative Filtering Recommendation)**

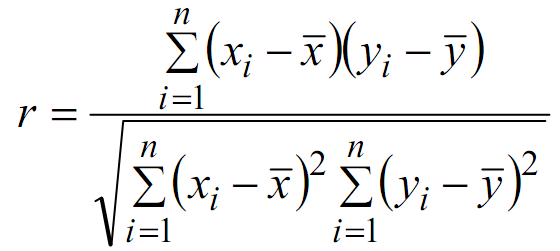
Gợi ý theo lọc cộng tác (Collaborative Filtering Recommendation) là phương pháp so sánh, tính toán đọ tương tự nhau giữa những users hay items, từ đó users sẽ được tư vấn những items được ưa chuộng nhất bởi những users có nét tương động (có cùng điểm chung). Được thể hiện như hình 3 dưới đây [số].

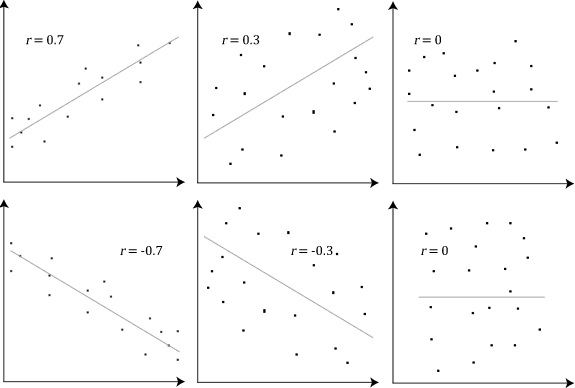
Hình 3: Giải thuật lọc cộng tác

1. **Độ tương tự (Similarity)**

Để đưa ra được danh sách gợi ý chính xác, cần xác định mối tương quan hoặc sự tương tự giữa người dùng và các mục dữ liệu bằng chỉ số tương tự.

**Chỉ số tương quan Pearson [4w]:** được biết đên như một phương pháp đo lường mối liên hệ giữa các biến bởi nó dựa trên phương pháp phát hiện phương sai, phương pháp này xác định mối tương quan, cũng như hướng của mối quan hệ, ý tưởng được mô tả bởi minh họa dưới đây.





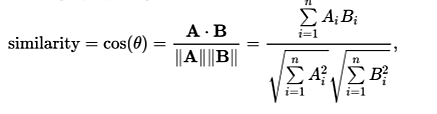
Hình 5: Minh họa ý tưởng chỉ số Pearson

***Giải thích:***

Với r là hệ số tương quan Pearson:

* **r = 0:** Hai biến không có tương quan tuyến tính.
* **r =1;** **r = -1**: Hai biến có mối tương quan tuyến tính tuyệt đối.
* **r < 0:** Hệ số tương quan âm. Giá trị biến x tăng thì giá trị biến y giảm và ngược lại, giá trị biến y tăng thì giá trị biến x giảm.
* **r > 0:** Hệ số tương quan dương. Nghĩa là giá trị biến x tăng thì giá trị biến y tăng và ngược lại, giá trị biến y tăng thì giá trị biến x cũng tăng.

**Chỉ số tương tự Cosine (Cosine Similarity):** Ý tưởng của bài toán này là việc biến hai văn bản cần so sánh thành 2 vector rồi từ đó tính khoảng cách giữa 2 vector đó. Chỉ số tương tự Cosine cho phep tính toán độ tương tự giữa 2 vectơ n chiều bằng cách xác định góc giữa 2 vectơ. Ví dụ, cho 2 vectơ và , chỉ số tương tự cosine được tính theo công thức sau:

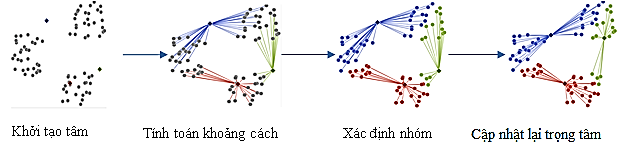


***Với các giá trị biểu thức :***

* : là giá trị thuộc tính của phần tử mới đến.
* : là giá trị thuộc tính của các phần tử trong tập dữ liệu.

1. **Lọc nhân khẩu học (Demographic Filter - DF)**

Lọc nhân khẩu học (DF): nhằm mục đích phân loại người dùng dựa trên các thuộc tính cá nhân hổ trợ cho hệ thống gợi ý đưa ra các đề xuất dựa trên các nhóm người dùng đã phân loại. Mục tiêu ở giải thuật này là tìm ra các nhóm người dùng cùng thích một sản phẩm nào nhất. Ví dụ: nhóm người dùng nhóm ***C*** thích một sản phẩm ***s***, người dùng ***C1*** thuộc nhóm người dùng ***C*** chưa biết đến sản phẩm ***s***, thì sản phẩm s này có thể gợi ý cho người dùng ***C1***. Có nhiều phương pháp để xác định và phân nhóm dữ liệu như: **SVM (Support vector machine), KNN (K – Nearest neighbor ), K-Means Clustering** … Tùy vào tập dữ liệu mà mỗi phương pháp đều có ưu điểm riêng, được sử dụng phổ biến là giái thuật **K-Means** do giải thuật này không cần biết giá trị nhãn, ý tưởng chính là xác đinh *k* trung tâm, với *k* là số lượng nhóm cố định cho trước, sau đó tính toán khoảng cách giữa các giá trị thuộc tính so với giá trị *k* trung tâm, lập lại nhiều lần, sau đó sinh ra nhóm dữ liệu, các bước của giải thuật được minh họa ở hình xxx bên dưới.

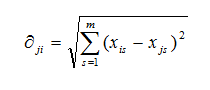


Hình 6: Giải thuật K-Means Clustering

Các bước thực hiện của giải thuật K – means Clustering như sau: giả sử có tập dữ liệu với *n* phần tử, mỗi phần tử có m thuộc tính. Số chiều của tọa độ được xác định dựa trên số thuộc tính, xem mỗi thuộc tính của đối tượng như một tọa độ của không gian *m* chiều và biểu diễn đối tượng như một điểm của không gian *m* chiều.  
*Bước 1*:Xác định số cụm *k* cho tập dữ liệu ban đầu gồm n phần tử

*Bước 2*: khởi tạo số điểm trung tâm bằng với số lượng *k*

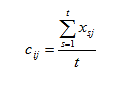
Bước 3: Đối với mỗi đối tượng trong tập dữ liệu, tính toán khảng cách đến trọng tâm của các cụm bằng công thức sau:



***Với các giá trị biểu thức:***

* : là đối tượng thứ *i* cần được tính khoảng cách và phân loại
* : là đối tượng trung tâm cụm j
* là thuộc tính thứ *s*  của nhóm
* : là thuộc tính thứ *s* của đối tượng trung tâm của cụm

*Bước 4*: Cập nhật lại tâm của đối tượng trung tâm *k* (*k* nhóm) bằng phương pháp tính trung bình cộng vector của các đối tượng trong mỗi cụm. giá trị trọng tâm được xác định bằng công thức sau:



***Với các giá trị biểu thức:***

* : là tọa độ thứ *j* của đối tượng trung tâm nhóm *I* ( cần cập nhật lại)
* : là thuộc tính thứ j của phần tử *s*
* : là số phần tử hiện có ở nhóm *i*

*Bước 5:* Lặp lại các *bước 3* và *bước 4* cho đến khi các trọng tâm của cụm không có sự thay đổi.

1. **Tiền xử lý dữ liệu (Data Preprocessing)**

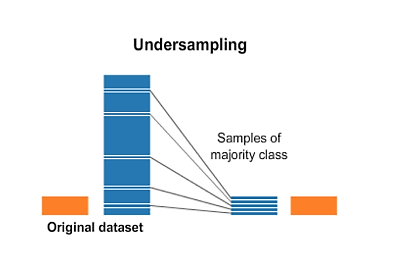
Trong thực tế, công việc xử lý dữ liệu trước khi đưa vào mô hình là rất cần thiết, bước này làm cho dữ liệu có được ban đầu qua thu thập dữ liệu (gọi là dữ liệu gốc ) có thể áp dụng được với các mô hình học máy. Các kỹ thuật tiền xử lý dữ liệu phổ biến hiện nay bao gồm: Xử lý dữ liệu bị khuyết (Missing data), mã hóa các biến nhóm (ending categorical varibles), chuẩn hóa dữ liệu (standardization data), co giãn dữ liệu (scaling data)…

***Chuẩn hóa dữ liệu (Standardization data)***  là một trong những bước tiền xử lý dữ liệu mục tiêu ở bước này là đảm bảo rằng dữ liệu nhất quán nội bộ, có nghĩa là mỗi kiểu dữ liệu có cùng nội dung định dạng, Các giá trị được chuẩn hóa theo mục đích của yêu cầu bài toán có nhiều cách để chuẩn hóa hóa dữ liệu, ví dụ như: chuyển đổi dữ liệu nam – nữ sang giá trị 0 – 1, Chuyển đổi kiểu dữ liệu liên tục sang kiểu dữ liệu rời rạc,….[vimentor]

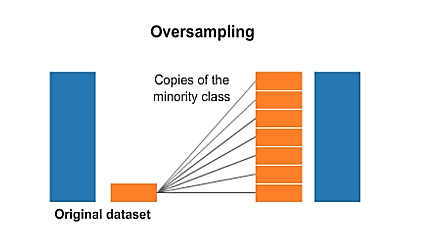
***Co giãn dữ liệu (Scaling data)*** là một phương pháp chuẩn hóa phạm vi của các đặc trưng dữ liệu và được thực hiện trong suốt quá trình tiền xử lý dữ liệu. Phạm vi của dữ liệu thô rất rộng, trong khi đối với một số thuật toán học máy, các hàm mục tiêu sẽ không hoạt động đúng như yêu cầu bài toán đặt ra, vậy nên để dữ liệu có thể hoạt động đúng mục đích thì cần phải áp dụng các giải thuật toán học để giải tuyết. Tiêu biểu là những công thức tính khoảng cách giữa các vector, các điểm dữ liệu góp phần vào quá trình tính toán độ dài, từ đó đưa ra phương pháp co giãn độ dài phù hợp. Các giải thuật phổ biến thường được sử dụng trong việc co giãn dữ liệu là ***KNN (K Nearest Neighbors), K-Means (K- Means Cluster), SVM (Support Vector Machine)…***[vimentor][7]

1. Xử lý mất cân bằng dữ liệu

Mất cân bằng dữ liệu là một trong những hiện tượng phổ biến trong quá trình thu thập dữ liệu, Mất cân bằng dữ liệu là khi dữ liệu bị thiếu hụt ở một số nhóm nhất định, ví dụ: Khi thu thập dữ liệu ngành học, tỷ lệ dữ liệu thu thập được giữa sinh viên Công nghệ thông tin và sinh viên Kinh tế có sự chênh lệch đáng kể là 80:20. Dữ liệu sẽ được xem là cân bằng nếu tỷ lệ đạt được 50:50, hầu hết các bộ dữ liệu đều khó đạt được trạng thái cân bằng mà luôn có sự khác biệt về tỷ lệ giữa 2 hoặc nhiều nhóm trong bộ dữ liệu. Đối với những trường hợp có tỷ lệ chênh lệch dữ liệu là 60:40, mặc dù vần được xem là mất cân bằng nhưng không ảnh hưởng đáng kể đến khả năng dự báo mô hình. Tuy nhiên nếu xảy ra hiện tượng mất cân bằng, chẳng hạn như tỷ lệ dữ liệu 90:10 thì chất lượng kết quả mô hình so với độ chính xác thực tế sẽ sai lệch nghiêm trọng. Giải pháp khắc phục hiện tượng mất cân bằng dữ liệu hiện đang được nhiều nhà nghiên cứu hướng đến là **Over Sampling** và **Under Sampling**. [8]

*Under Sampling* là việc gia giảm số lượng quan sat của nhóm có tỷ lệ dữ liệu trội hơn, nhiều hơn hoặc những nhóm có lượng dữ liệu tăng một cách đột biến … để dữ liệu số lượng nhiều chở nên cân bằng so với nhóm dữ liệu thiểu số. Ưu điểm của giải thuật này là làm cân bằng dữ liệu một cách nhanh chóng, dể dàng thực hiện. Nhưng nhược điểm ở giải thuật này là kích thước tập dữ liệu sẽ bị giảm đáng kể. Giả sử tỷ lệ giữa sinh viên Công nghệ thông tin và sinh viên Kinh tế là 1000:500 (sinh viên), nếu sử dụng Under Sampling, sau khi cân bằng dữ liệu sẽ bị giảm 500 (sinh viên). Dẫn đến tình trạng tập dữ liệu huấn luyện sẽ bị thu hẹp. Cho nên giải thuật Under Sampling sẽ không phù hợp nếu áp dụng với bộ dữ liệu có kích thước nhỏ.

Hình 7: Minh họa giải thuật Under Sampling

*Over Sampling* là phương pháp giúp giải quyết hiện tượng mất cân bằng mẫu bằng cách gia tăng kích thước mẫu thuộc nhóm dữ liệu thiểu số sao cho dữ liệu thiểu số cân bằng so với dữ liệu có số lượng đa số, cách làm này khắc phục nhược điểm so với giải thuật Under Sampling là có thể áp dụng cho tập dữ liệu có kích thước nhỏ, ngoài ra còn mở rộng quy mô dữ liệu lớn hơn để mô hình đạt được độ chính xác cao trên nhóm dữ liệu thiểu số, hình ảnh dưới đây minh họa cho phương pháp xử lý mất cân bằng dữ liệu Over Sampling.

Hình 8: Minh họa giải thuật Over Sampling

1. Phương pháp đánh giá

Khi xây dựng mô hình máy học, để biết được tính khả thi và độ tin cậy của mô hình cần phải có thước đo đánh giá xem mô hình sử dụng có hiệu quả không. Một hệ thống gợi ý được đánh giá bằng cách phân tích trên tập dữ liệu dùng để đánh giá hoặc lấy thông tin trực tiếp từ người sử dụng hệ thống [10][11]. Hướng sử dụng phổ biến là chia tập dữ liệu ra thành 2 phần, 1 phần để huấn luyện mô hình và phần còn lại để kiểm thử, 2 phương pháp thường được sử dụng là k-fold và hold-out. Ngoài ra Given – N là một trong những biến thể khác của phương pháp k-fold.

Nguyên tắc của phương pháp Given – N là xét tất cả các giá trị thuộc tính của tập kiểm thử bằng cách đem các giá trị đó vào mô hình dự đoán, sau khi có kết quả, đem so sánh với giá trị thực tế là nhãn của tập kiểm thử, nếu kết quả mô hình trùng hoặc có trong kết quả nhãn thực tế thì độ chính xác tăng lên 1 đơn vị và bằng 0 khi sản phẩm gợi ý không trùng với sản phẩm trong tập kiểm tra [10].

Đánh giá việc sử dụng các dự đoán:

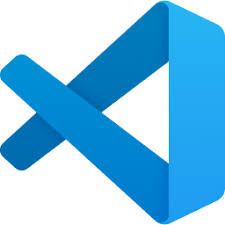
Precision: là tỷ lệ giữa số lượng gợi ý đúng so với toàn bộ số lượng nhãn trong tập kiểm thử, được tính toán bằng công thức sau:

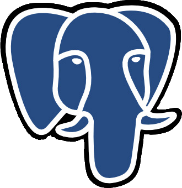
*Với:*

* TP: là số lượng gợi ý chính xác
* FN: là số lượng gợi ý không chính xác

1. Công cụ

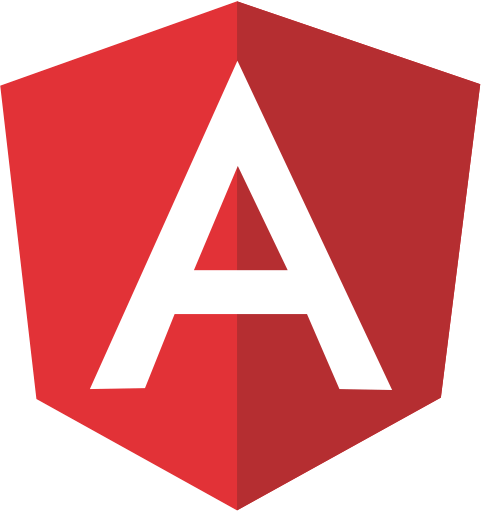
Công cụ sử dụng:

***Visual Studio Code[số]*** là một trong những công cụ soạn thảo văn bản lập trình mã nguồn mở nhẹ nhưng mạnh mẽ với thư viện hổ trợ soạn thảo vô cùng phong phú. Đây là công cụ biên tập code đang được đông đảo những lập trình viên sử dụng và yêu thích. Bên cạnh đó Visual Code hổ trợ hầu hết các ngôn ngữ lập trình đang có hiện nay và tối ưu hóa tốt giúp kết nối tốt các phần mềm khác. Hơn nữa công cụ này hổ trợ đa nền tảng hệ điều hành bao gồm: Windows, macOS và Linux.

***PostgreSQL[số]*** là một hệ thống quản trị cơ sở dữ liệu quan hệ, được phát triển bởi Khoa Điện toán, Đại học California – Hoa Kỳ. Là một phần mềm mã nguồn mở, được thiết kế để hoạt động đa nền tảng. Với lịch sử phát triển lâu dài, PostgreSQL đảm bảo đầy đủ các chức năng của một hệ quản trị cơ sở dữ liệu như: Tính ràng buộc toàn vẹn dữ liệu, tối ưu hóa kiểu dữ liệu, sử dụng đa nền tảng và hơn hết có thể kết hợp với các hệ quản trị cở sở dữ liệu khác.



***Parse Server[số]***là một ứng dụng mã nguồn mở, Parse Server cung cấp các **API**(Application Programs Interface)[số] xây dựng ứng dụng, Web Service API cho phép lưu trữ dữ liệu online, ngoài ra Parse Server còn có tính năng kết nối và quản lý cơ sở dữ liệu gốc (PostgreSQL, MongoDB, mySQL,…) với giao diện dể sử dụng. Parser Server tương thích với rất nhiều ngôn ngữ khá nhau như php, nodejs, c#,…



***Angular Framework[số]*** là một Javascript Framework dùng để viết giao diện Web, được phát triển bởi Google với phiên bản tiền nhiệm là Angular JS được phát hành vào năm 2010. Hiện tại đã được phát triển thành phiên bản Angular cũng là phiên bản được sử dụng phổ biến nhất hiện nay, hiểu đơn giản Angular là một khung là việc của Javascript được cài đặt dựa trên cấu trúc của mô hình MVVM mục tiêu hướng đến là thị trường di dộng (Mobile) và Web Services. Hoạt động dựa trên các thuộc tính mở rộgn HTML, Framework mã nguồn mở hoàn toàn miển phí nên mọi lập trình viên đều có thể tiếp cận và sử dụng.

***Bootstrap Framework [số]***

Bootstrap là một nền tảng (Framework) miển phí, mã nguồn mở,

dựa trên HTML, CSS và Javascript, với mục tiêu ban đầu là xây dựng giao diện Website tương thích với tất cả các thiết bị có kích thước màn hình khác nhau. Ban đầu bootstrap có tên là Twitter Blueprint được phát triển bởi Mark Otto và Jacob Thorton tại Twitter với phiên bản nội bộ, sau đó đã phát triển thành Bootstrap nhờ vào sự đóng góp của các thành viên trong nhóm và trở thành một Framework mã nguồn mở hoàn toàn miển phí. Bootstrap bao gồm các HTML Templaté, CSS Templates và Javascript tạo ra những cái cơ bản có sẳn như: Typography, forms, buttons, table và nhiều thứ khác, ngoài ra framework này còn hổ trợ nhiều tính năng giúp tùy chỉnh nội dung trang web tương thích với nhiều loại kích cở màn hình, hổ trợ việc thiết kế website một cách dể dàng và nhanh chóng hơn.

**Ngôn ngữ lập trình:**

***Khái quát về HTML, CSS, Javascript [5]***

HTML (Hyper Text Makup Language – Ngôn ngữ dánh dấu siêu văn bản): là ngôn ngữ dùng để xây dựng một trang web. Các thẻ HTML được thống nhất toàn cầu (W3C – World Wide Web Consortium).

Một trang web thông thường gồm có hai thành phần chính: Dữ liệu của một trang web ( văn bản, âm thanh, hình ảnh,…). Các thẻ HTML dùng để dịnhdạng mô tả cách thức dữ liệu hiển thị trên trình duyệt.

Trình duyệt web (Chrome, IE, Firefox, Safari …) là đọc tài liệu HTML và hiển thị chúng. Trình duyệt không hiển thị các thẻ HTML, nhưng sử dụng để xác định cách hiển thị tài liệu.

CSS là một dạng tài liệu chứa các thông tin về các mẫu định dạng mà tài liệu thông tin này có thể được nhiều trang web sử dụng. Các mẫu này dùng để định nghĩa phương thức hiển thị ( đường kẻ khung, khoảng cách giữa ác dòng,..) và định dạng ( màu chữ, kiểu chữ, màu nền…) phần nội dung trang web và định dạng của nó.

Javascript là ngôn ngữ dưới dạng script có thể gắn với các tập tin HTML được trình duyệt diễn dịch. Trình duyệt sẽ đọc Javascript dưới dạng mã nguồn. Chính vì vậy chúng ta có thể dễ dàng sử dụng chúng trên các trang web. JavaSript là ngôn ngữ dựa trên đối tượng. Tuy vậy nó không phải là ngôn ngữ hướng đối tượng do không thể hổ trợ các lớp hay tính thừa kế. Chạy trên bất cứ trình duyệt nào. Các trình duyệt web có thể hiển thị những câu lệnh Javascript được nhúng vào HTML

***Khái quát về ngôn ngữ lập trình Python***

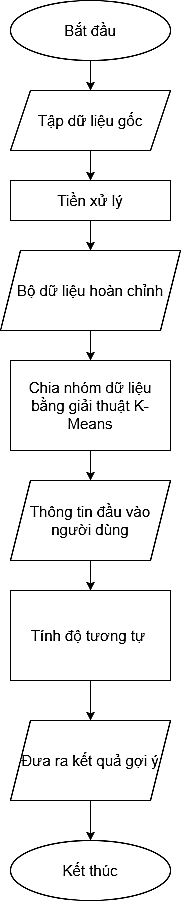
Python là một ngôn ngữ lập trình thông dịch do Guido van Rossum tạo ra năm 1990, nó được xem là ngôn ngữ có cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Nhiều năm gần đây Python được sử dụng nhiều trong máy học, trí tuệ nhân tạo, lập trình web, ứng dụng… Được phát triển trong một dự án mã nguồn mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý, Python hoàn toàn có thể sử dụng trên nhiều nền tảng hệ điều hành như: Window, Mac OS, Linux…

Với cú pháp đơn giản dể hiểu Python có rất nhiều ưu điểm để khiến các lập trình viên tin dùng, việc sử dụng Python để lập trình web (phía Server) đang được sử dụng phổ biến do tốc độ xử lý nhanh hơn những ngôn ngữ lập trình khác. Điểm quan trọng tạo nên thành công của ngôn ngữ Python là do có cộng đồng hổ trợ đông đảo, hầu hết các thắc mắc, câu hỏi liên quan đến lập trình Python điều được giải đáp trên các diễn đàn lập trình. Ngoài ra, Python còn có kho thư viện phong phú, với tiêu chuẩn cao. Python hổ trợ lập trình viên trong rất nhiều lĩnh vực khác nhau như: phân tích và xử lý dữ liệu, xử lý ảnh, đồ họa máy tính, thiết kế giao diện,… Với nhiều ưu điểm kể trên, Python là một ngôn ngữ rất đáng để cân nhắc trong quá trình quyết định ngôn ngữ lập trình cho đề tài nghiên cứu.

**CHƯƠNG 2**

**THIẾT KẾ VÀ CÀI ĐẶT**

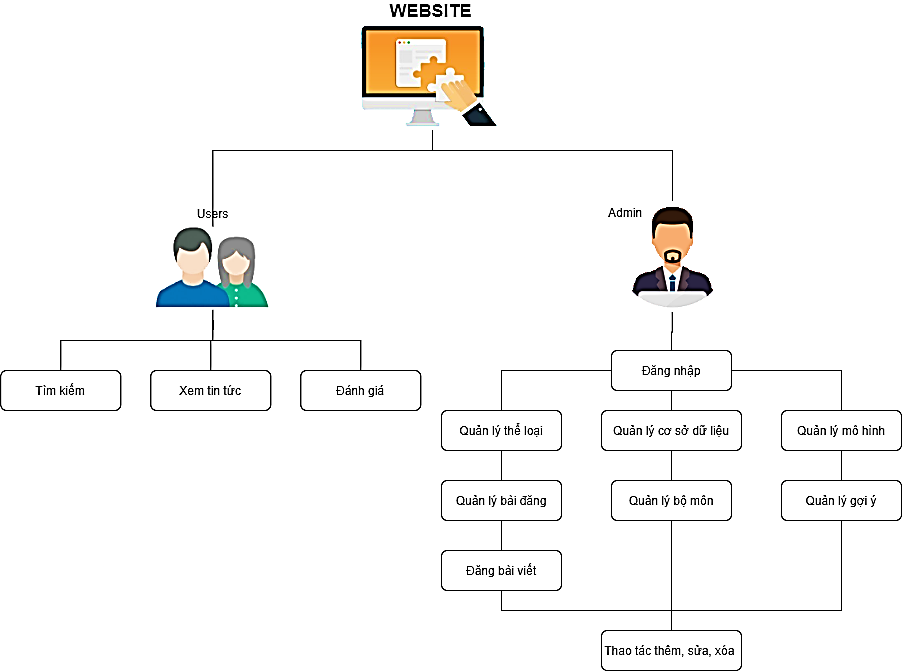
Trong chương này sẽ trình bày phân tích và cài đặt trang web tin tức kết hợp với giải thuật gợi y theo nội dung để phù hợp với nhu cầu của từng người dùng truy cập vào trang web.

1. **Sơ đồ tổng quan về chức năng**

Hình 9: Lưu đồ chức năng hệ thống

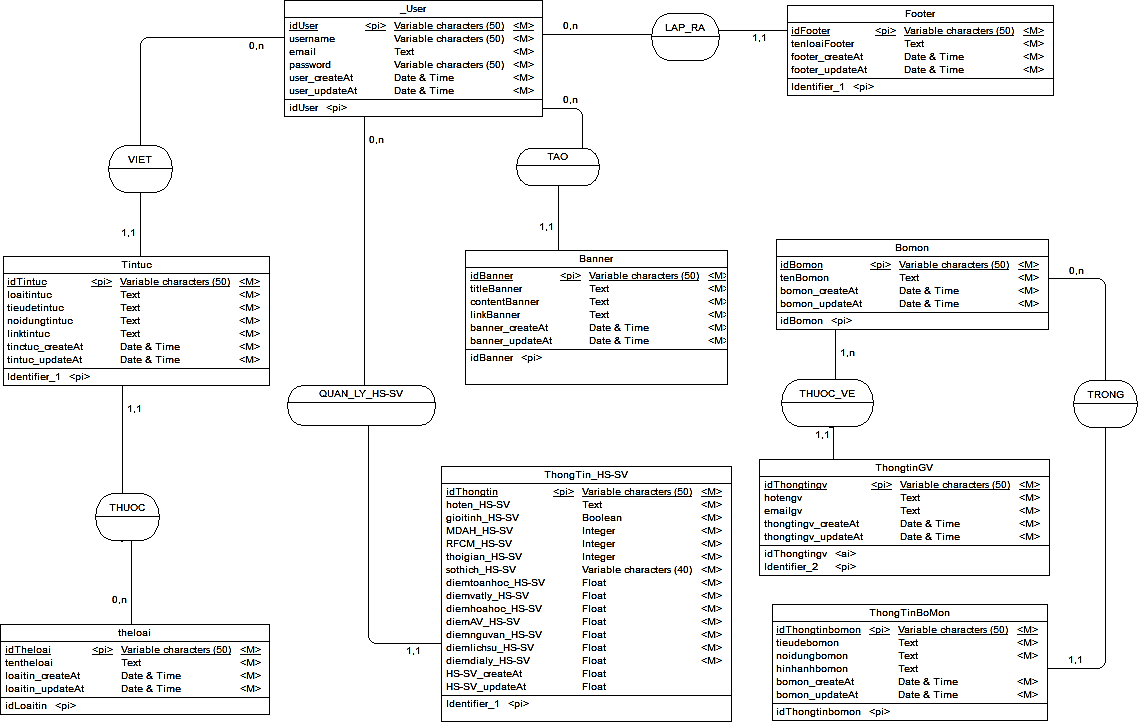
Hệ thống cung cấp chức năng cho 2 nhóm người dùng: quản trị viên (Admin) và người dùng (User) trên trang web gợi ý ngành học.

* Quản trị viên là người quản trị hệ thống, thực hiện các chức năng quản lý người dùng, quản lý tin tức, quản lý thể loại, bộ môn, quản lý bài đăng, cơ sở dữ liệu… Quản trị viên là người có quyền cao nhất trong hệ thống và được phép thêm, sửa, xóa, tìm kiếm.
* Người dùng được phép truy cập vào trang web, tìm kiếm, xem tin tức về trang web. Tổng quan về trang web sẽ được minh họa ở hình số.



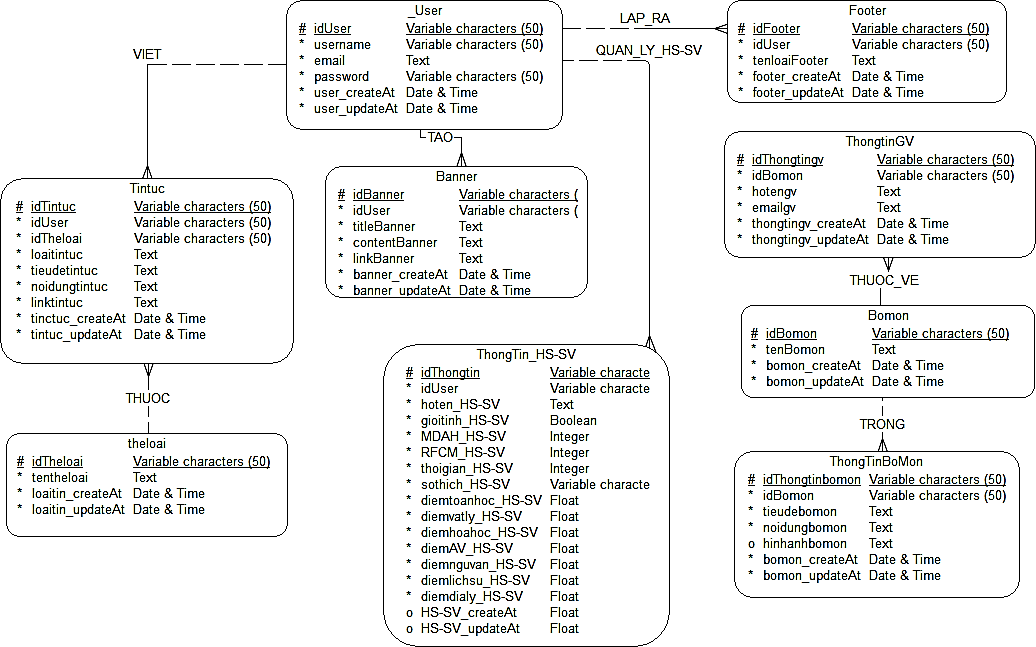
Hình 10: Sơ đồ tổng quang trang gợi ý ngành học

1. **Cơ sở dữ liệu**
   1. **Mô hình dữ liệu mức quan niệm (CDM)**



Hình 11: Mô hình dữ liệu mức quan niệm

* 1. **Mô hình vật lý (PDM)**

****

Hình 12: Mô hình vật lý (PDM)

* 1. **Giới thiệu thực thể**

**User( idUser**, username, email, password, user\_createAt, user\_updateAt**).**

**Tintuc(** idTintuc, idUser, idTheloai, loaitintuc, tieudetintuc, linktintuc, tintuc\_createAt, tintuc\_updateAt**)**

**Theloai( idTheloai**, tentheloai, loaitin\_createAt, loaitin\_updateAt**)**

**Banner(idBanner**, idUser, titleBanner, contentBanner, linkBanner, banner\_createAt, banner\_updateAt**)**

**Thongtin\_HS-SV(idThongtin**, idUser, hoten\_HS-SV, gioitinh\_HS-SV, MDAH\_HS-SV, RFCM\_HS-SV, thoigian\_HS-SV, sothich\_HS-SV, diemtoanhoc\_ HS-SV, diemvatly\_ HS-SV, diemhoahoc\_ HS-SV, diemAV\_ HS-SV, diemnguvan\_ HS-SV, diemlíchsu\_ HS-SV, diemdialy\_ HS-SV, HS-SV\_createAt, HS-SV\_updateAt**)**

**Footer( idFooter**, idUser, tenloaiFooter, footer\_createAt, footer\_updateAt**)**

**ThongtinGV(idThongtingv**, idBomon, hotengv, emailgv, thongtingv\_createAt, thongtingv\_updateAt**)**

**Bomon( idBomon**, tenBomon,bomon\_createAt, bomon\_updateAt**)**

**ThongtinBoMon(idThongtinbomon**, idBomon, tieudebomon, noidungbomon, hinhanhbomon, bomon\_createAt, bomon\_updateAt**)**

* 1. **Mô tả tập thực thể**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **stt** | **Tên thực thể** | **Thuộc tính** | **Kiểu dữ liệu** | **Kích thước** | **Not Null** | **Khóa** |
| **1** | **\_User** | **idUser** | VARCHAR | **50** |  | **X** |
| **username** | TEXT |  |  |  |
| **email** | TEXT |  |  |  |
| **Password** | TEXT | **50** |  |  |
| **User\_createAt** | DATE&TIME |  |  |  |
| **User\_updateAt** | DATE&TIME |  |  |  |
| **2** | **Thongtin\_HS-SV** | **idThongtin** | VARCHAR | **50** |  | **X** |
| **Hoten\_HS-SV** | TEXT |  |  |  |
| **Gioitinh\_HS-SV** | BOOLEAN |  |  |  |
| **MDAH\_HS-SV** | INT |  |  |  |
| **RFCM\_HS-SV** | INT |  |  |  |
| **Thoigian\_HS-SV** | INT |  |  |  |
| **Sothich\_HS-SV** | VARCHAR | **40** |  |  |
| **Diemtoanhoc** | FLOAT |  |  |  |
| **Diemvatly** | FLOAT |  |  |  |
| **DiemAV** | FLOAT |  |  |  |
| **Diemnguvan** | FLOAT |  |  |  |
| **Diemlichsu** | FLOAT |  |  |  |
| **Diemdialy** | FLOAT |  |  |  |
| **HS-SV\_createAt** | DATE&TIME |  |  |  |
| **HS-SV\_updateAt** | DATE&TIME |  |  |  |
| **3** | **Tintuc** | **idTintuc** | VARCHAR | **50** |  | **X** |
| **Loaitintuc** | TEXT |  |  |  |
| **Tieudetintuc** | TEXT |  |  |  |
| **Noidungtintuc** | TEXT |  |  |  |
| **Linktintuc** | TEXT |  |  |  |
| **Tintuc\_createAt** | DATE&TIME |  |  |  |
| **Tintuc\_updateAt** | DATE&TIME |  |  |  |
| **4** | **nhomtin** | **idNhomtin** | VARCHAR | **50** |  | **X** |
| **Tennhomtin** | TEXT |  |  |  |
| **Nhomtin\_createAt** | DATE&TIME |  |  |  |
| **Nhomtin\_updateAt** | DATE&TIME |  |  |  |
| **5** | **Banner** | **idBanner** | VARCHAR |  |  | **X** |
| **titleBanner** | TEXT |  |  |  |
| **contentBanner** | TEXT |  |  |  |
| **linkBanner** | TEXT |  |  |  |
| **Banner\_createAt** | DATE&TIME |  |  |  |
| **Banner\_updateAt** | DATE&TIME |  |  |  |
| **6** | **Footer** | **idFooter** | VARCHAR | **50** |  | **X** |
| **TenloaiFooter** | TEXT |  |  |  |
| **Footer\_createAt** | DATE&TIME |  |  |  |
| **Footer\_updateAt** | DATE&TIME |  |  |  |
| **7** | **Bomon** | **idBomon** | VARCHAR |  |  | **X** |
| **tenBomon** | Text |  |  |  |
| **Bomon\_createAt** | DATE&TIME |  |  |  |
| **Bomon\_updateAt** | DATE&TIME |  |  |  |
| **8** | **thongtinGV** | **idThongtingv** | VARCHAR | **50** |  | **X** |
| **Hotengv** | TEXT |  |  |  |
| **Emailgv** | TEXT |  |  |  |
| **Thongtin\_createAt** | DATE&TIME |  |  |  |
| **Thongtin\_updateAt** | DATE&TIME |  |  |  |
| **9** | **ThongtinBoMon** | **idThongtinbomon** | VARCHAR | **50** |  | **X** |
| **Tieudebomon** | TEXT |  |  |  |
| **Noidungbomon** | TEXT |  |  |  |
| **Hinhanhbomon** | TEXT |  | **X** |  |
| **Bomon\_createAt** | DATE&TIME |  |  |  |
| **Bomon\_updateAt** | DATE&TIME |  |  |  |

Bảng 1 Bảng mô tả tập thực thể

* 1. **Các ràng buộc toàn vẹn**
     1. **Ràng buộc toàn vẹn khóa chính**
* **Quan hệ nhomtin**

**rb\_pk1**: Mỗi nhóm tin đều được cấp cho một mã gọi là idNhomtin:

nhomtin1,nhomtin2… nhomtin, *sao cho:*

nhomtin1 nhomtin2 và nhomtin1.idNhomtin ≠ nhomtin2.idNhomtin

* **Quan hệ Tintuc**

**rb\_pk2:** Mỗi tin tức đều được cấp cho một mã gọi là idTintuc:

∀tintuc1,tintuc2… ∈ Tintuc, sao cho:

Tintuc1 ≠ tintuc2 và tintuc1.idTintuc ≠ tintuc2.idTintuc

* **Quan hệ ThongTin\_HS-SV**

**rb\_pk3**: Mỗi thông tin học sinh, sinh viên đều được cấp cho một mã gọi là idThongtin\_HS-SV:

∀Thongtin\_HS-SV1,Thongtin\_HS-SV2… ∈ Thongtin\_HS-SV, *sao cho*:

Thongtin\_HS-SV1 ≠ Thongtin\_HS-SV2 và Thongtin\_HS-SV1.idThongtin\_HS-SV ≠ Thongtin\_HS-SV2.idThongtin\_HS-SV

* **Quan hệ Banner**

**rb\_pk4**: Mỗi banner đều được cấp cho một mã gọi là idBanner:

∀banner1,banner2… ∈ Banner, *sao cho:*

banner1 ≠ banner2 và banner1.idBanner ≠ banner2.idBanner

* **Quan hệ Bomon**

**rb\_pk5**: Mỗi bộ môn đều được cấp cho một mã gọi là idBomon:

∀bomon1,bomon2… ∈ Bomon, *sao cho:*

bomon1 ≠ bomon2 và bomon1.idBomon ≠ bomon2.idBomon

* **Quan hệ ThongtinGV**

**rb\_pk6**: Mỗi thông tin giảng viên đều được cấp cho một mã gọi là idThongtingv

∀thongtinv1,thongtinv2… ∈ ThongtinGV, *sao cho:*

thongtinv1 ≠ thongtinv1 và thongtinv1.idThongtinGV ≠ thongtinv2.idThongtinGV

* **Quan hệ ThongTinBoMon**

**rb\_pk7**: Mỗi thông tin bộ môn đều được cấp cho một mã gọi là idThongtinbomon:

∀thongtinbomon1, thongtinbomon2… ∈ ThongTinBoMon, *sao cho:*

thongtinbomon1 ≠ thongtinbomon2 và thongtinbomon1.idThongtinbomon ≠ thongtinbomon2.idThongtinbomon

* **Quan hệ \_User**

**rb\_pk8**: Mỗi user đều được cấp cho một mã gọi là idUser:

∀user1, user2… ∈ \_User, *sao cho*:

user1 ≠ user2 và user1.idUser ≠ user2.idUser

* **Quan hệ Footer**

**rb\_pk9**: Mỗi footer đều được cấp cho một mã gọi là idFooter:

∀footer1,footer2… ∈ Footer, *sao cho:*

footer1 ≠ footer2 và footer1.idFooter ≠ footer2.idFooter

* + 1. **Ràng buộc toàn vẹn khóa ngoại**
* **Quan hệ Tintuc – theloai**

Một tin tức phải thuộc một thể loại tin tức.

Miêu tả: tintuc.idTintuc theloai.idTheloai

* **Quan hệ ThongtinGV – Bomon**

Một giảng viên phải thuộc duy nhất một bộ môn.

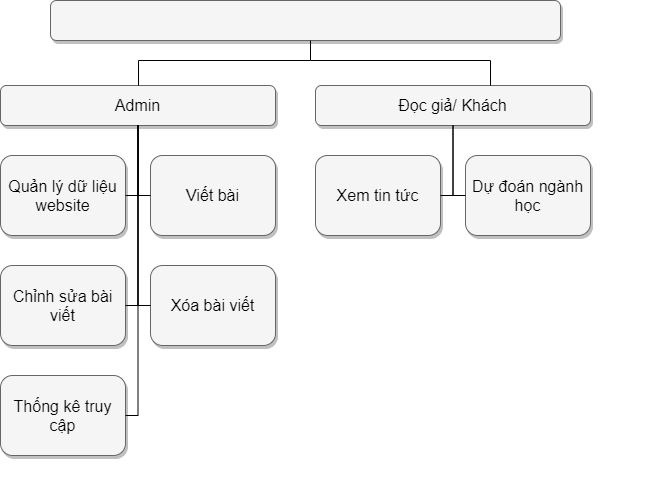
Miêu tả: thongtingv.idThongtinGV bomon.idBomon

* **Quan hệ ThongTinBoMon – Bomon**

Một thông tin bộ môn phải thuộc duy nhất một bộ môn.

Miêu tả: thongtinbomon.idThongTinBoMon bomon.idBomon

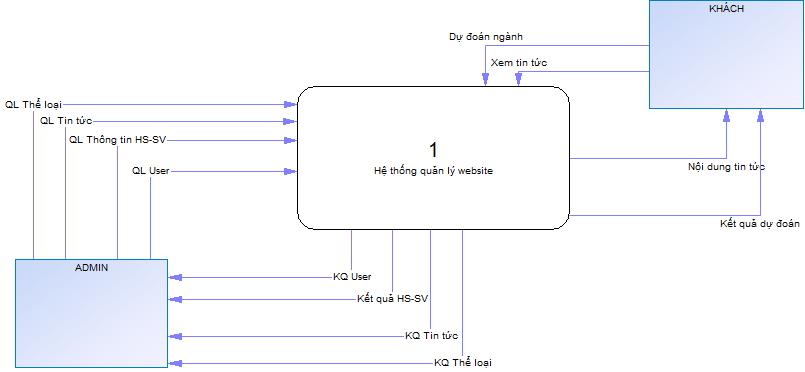
* 1. **Mô hình BFD**

****

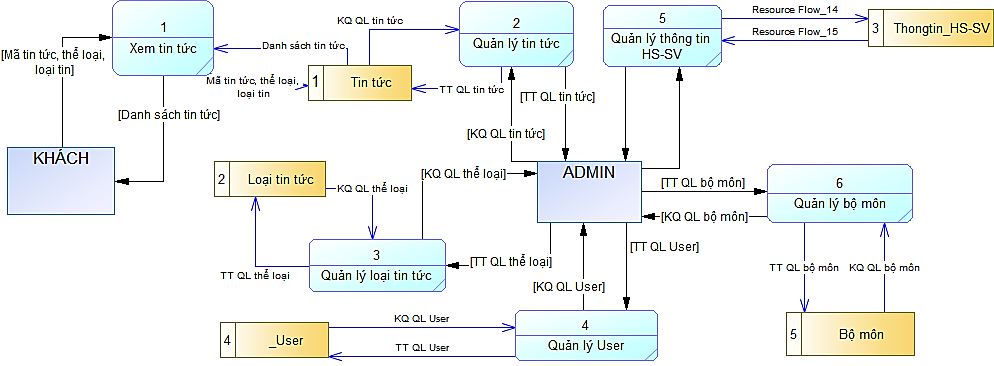
Website Gợi ý ngành học

Hình 13: Mô hình BFD

* 1. **Mô hình DFD cấp ngữ cảnh**



Hình 14: Mô hình DFD cấp ngữ cảnh

* 1. **Mô hình DFD cấp 0**

Hình 15: Mô hình DFD cấp 0

1. **Mô tả tập dữ liệu**
   1. **Giới thiệu tập dữ liệu**

Tập dữ liệu **dataset** được thu thập từ sinh viên ở các trường Đại học, Cao đẳng, Trung cấp thuộc các tỉnh Đồng bằng sông Cửu Long. Nội dung dung thu thập bao gồm các thông tin về sở thích, nguyện vọng, các yếu tố liên quan đến việc quyết định chọn ngành. Dữ liệu có được bằng cách lập biểu mẫu các câu hỏi và giá trị dữ liệu là các câu trả lời, ý kiến phản hồi từ đối tượng, thời gian thu thập dữ liệu từ năm 2019 đến năm 2020.

* 1. **Mô tả tập dữ liệu**

***Kích thước:*** Tập dữ liệu gồm 1062 dòng và 13 cột, số dòng tương ứng với lượt trả lời câu hỏi từ người dùng, số cột tương ứng với số lượng câu hỏi đặt ra trong biểu mẫu cũng là số thuộc tính có trong tập dữ liệu.

***Nhãn***: 1 cột ( Major ).

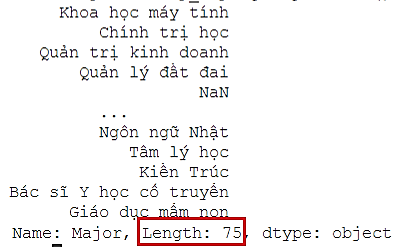
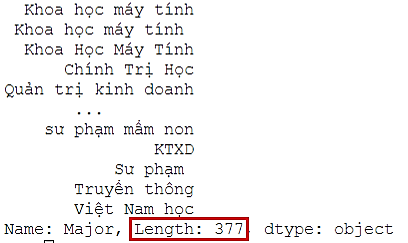
***Thuộc tính:*** 12 cột (Gender – Interest – RFCM – Time – Math\_score – Physics\_score – Chemistry\_score – English\_score – Literature\_score – History\_score – Geography\_score – Influence ).

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Mô tả** | **Ví dụ** |
| **Gender** | Giới tính | Nam hoặc Nữ |
| **Interest** | Sở thích | Du lịch, Phim ảnh, Âm nhạc, …. |
| **RFCM** | Lý do chọn Ngành/nghề | Năng lực bản thân hoặc sở thích hoặc nguyện vọng gia đình … |
| **Time** | Thời gian tìm hiểu ngành/nghề | Ít hoặc vừa phải hoặc nhiều … |
| **Math\_score** | Điểm môn Toán | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Physics\_score** | Điểm môn Vật lý | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Chemistry\_score** | Điểm môn Hóa học | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **English\_score** | Điểm môn Ngoại ngữ | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Literature\_score** | Điểm môn Ngữ văn | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **History\_score** | Điểm môn Lịch sử | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Geography\_score** | Điểm môn Địa lý | Từ 5 đến 6.5 hoặc  Từ 6.5 – 8 … |
| **Influence** | Mức độ nổi tiếng của trường đại học có ảnh hưởng đến quyết định chọn ngành/nghề | ảnh hưởng ít hoặc ảnh hưởng nhiều … |

Bảng 2: Các thuộc tính của tập dữ liệu dataset

1. **Tiền xử lý dữ liệu**
   1. **Chuẩn hóa dữ liệu**

* ***Loại bỏ thuộc tính Null, dư thừa, gây nhiễu:*** Tập dữ liệu gốc bao gồm những giá trị dư thừa trong quá trình thu thập, loại bỏ các giá trị mà bài toán không yêu cầu.
* ***Chuẩn hóa các giá trị nhãn:*** dữ liệu được thu thập qua biểu mẫu online, và dữ liệu về ngành nghề dã chọn (nhãn) là loại được phép nhập trên biểu mẫu, Nên các giá trị này không thống nhất, nghĩa là cùng một đáp án nhưng lại có nhiều cách nhập khác nhau, ví dụ như: “Khoa học máy tính”, “Khoa Học Máy Tính”,, “ Khoa học máy tính ”,… Để dữ liệu có thể sử dụng được ta phải chuẩn hóa về dạng thống nhất. giá trị nhãn ban đầu chưa qua chuẩn hóa là 377 loại, sau khi chuẩn hóa kết quả còn 75 loại và đó cũng là số lượng nhãn của tập dữ liệu.



* ***Chuẩn hóa dữ liệu Gender (Giới tính):***chuyển đổi giá trị dữ liệu Nam – Nữ về 2 giá trị 0 – 1. Sau khi chuẩn hóa ta được tập dữ liệu với cột thuộc tính **Gender** có giá trị là 0 – 1 tương ứng với 0 là giá trị Nam, 1 là giá trị nữ.

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Nam** | 0 | 410 | 1877 |
| **Nữ** | 1 | 584 | 2798 |

Bảng 3: Chuẩn hóa dữ liệu giới tính

* ***Chuẩn hóa dữ liệu RFCM (Nguyên nhân tác động lớn nhất đến việc chọn ngành học):*** chuyển đổi giá trị thuộc tính ở cột RFCM (Sở thích, Năng lực bản thân, Nguyện vọng gia đình, nhu cầu cao của thị trường, Yêu thích và tư vấn từ gia đình, Mức thu thập của gia đình) thành các giá trị số (lần lượt là: 0, 1, 2 ,3 ,4 ,5).

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính**  **tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Mức thu nhập của gia đình** | 0 | 38 | 179 |
| **Nguyện vọng gia đình** | 1 | 129 | 599 |
| **Nhu cầu cao của thị trường** | 2 | 225 | 1052 |
| **Năng lực bản thân** | 3 | 212 | 1000 |
| **Sở thích** | 4 | 345 | 1680 |
| **Tư vấn từ thầy cô bạn bè** | 5 | 45 | 165 |

Bảng 4: Chuẩn hóa dữ liệu RFCM

* ***Chuẩn hóa dữ liệu Influence (Mức độ nổi tiếng của trường đại học có ảnh htưởng đến quyết định chọn ngành ):*** chuyển đổi giá trị thuộc tính ở cột RFCM (Không ảnh hưởng, Ảnh hưởng ít, Khá ảnh hưởng, Ảnh hưởng lớn, Ảnh hưởng rất lớn ) sang các giá trị số (lần lượt là: 1, 2, 3, 4, 5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Không ảnh hưởng** | 1 | 71 | 317 |
| **Ảnh hưởng ít** | 2 | 99 | 311 |
| **Khá ảnh hưởng** | 3 | 241 | 950 |
| **Ảnh hưởng lớn** | 4 | 377 | 2078 |
| **Ảnh hưởng rất lớn** | 5 | 206 | 1019 |

Bảng 5: Chuẩn hóa dữ liệu Influence

* ***Chuẩn hóa dữ liệu Time (Thời gian tìm hiểu ngành học):*** chuyển đổi giá trị thuộc tính ở cột Time (Rất ít, Ít, Vừa phải, Nhiều, Rất nhiều) sang các giá trị số ( lần lượt là: 1, 2, 3, 4, 5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Rất ít** | 1 | 56 | 218 |
| **ít** | 2 | 85 | 349 |
| **Vừa phải** | 3 | 307 | 1323 |
| **Nhiều** | 4 | 383 | 2070 |
| **Rất nhiều** | 5 | 163 | 715 |

Bảng 6: Chuẩn hóa dữ liệu Time

* ***Chuẩn hóa dữ liệu Maths\_score ( điểm toán):*** chuyển đổi giá trị thuộc tính ở cột Maths\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 16 | 35 |
| **5 – 6.5** | 5.75 | 41 | 160 |
| **6.5 – 8** | 7.25 | 304 | 1466 |
| **8 – 9** | 8.5 | 456 | 2141 |
| **9 – 10** | 9.5 | 165 | 741 |
| **Điểm số do người dùng nhập vào** | 8.0 | 1 | 3 |
| 6.8 | 2 | 14 |
| 8.3 | 2 | 7 |
| 9.7 | 2 | 7 |
| 6.9 | 1 | 3 |
| 9.2 | 1 | 2 |
| 7.1 | 1 | 1 |
| 8.4 | 1 | 1 |
| 10.0 | 1 | 2 |
| 7.6 | 1 | 9 |
| 9.6 | 1 | 17 |
| 8.8 | 1 | 3 |
| 8.9 | 1 | 10 |
| 9.4 | 1 | 0 |
| 5.0 | 0 | 53 |

Bảng 7 Chuẩn hóa dữ liệu Toán

* **Chuẩn hóa dữ liệu *Physics\_score* ( điểm vật lý):** chuyển đổi giá trị thuộc tính ở cột Physics\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 19 | 52 |
| **5 - 6.5** | 5.75 | 90 | 291 |
| **6.5 - 8** | 7.25 | 374 | 1826 |
| **8 - 9** | 8.5 | 397 | 1889 |
| **9 - 10** | 9.5 | 135 | 469 |
| **Điểm số do người dùng nhập vào** | 5.0 | 0 | 91 |
| 8.8 | 1 | 17 |
| 6.5 | 1 | 13 |
| 8.4 | 1 | 10 |
| 8.7 | 1 | 6 |
| 7.9 | 1 | 3 |
| 6.6 | 1 | 3 |
| 9.0 | 1 | 2 |
| 6.3 | 1 | 1 |
| 7.4 | 1 | 1 |
| 7.1 | 1 | 1 |

Bảng 8: Chuẩn hóa dữ liệu điểm Vật lý

* **Chuẩn hóa dữ liệu *Chemistry\_score* ( điểm Hóa học ):** chuyển đổi giá trị thuộc tính ở cột Chemistry\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 27 | 55 |
| **5 - 6.5** | 5.75 | 98 | 368 |
| **6.5 - 8** | 7.25 | 371 | 1685 |
| **8 - 9** | 8.5 | 383 | 1746 |
| **9 - 10** | 9.5 | 140 | 700 |
| **Điểm số do người dùng nhập vào** | 5.0 | 0 | 67 |
| 9.4 | 1 | 17 |
| 5.4 | 1 | 13 |
| 7.0 | 1 | 6 |
| 9.8 | 1 | 4 |
| 7.5 | 1 | 3 |
| 8.8 | 1 | 3 |
| 9.6 | 1 | 3 |
| 9.1 | 1 | 2 |
| 5.7 | 1 | 1 |
| 6.5 | 1 | 1 |
| 9.2 | 1 | 1 |

Bảng 9: Chuẩn hóa dữ liệu điểm Hóa học

* **Chuẩn hóa dữ liệu *English\_score* ( điểm Ngoại ngữ ):** chuyển đổi giá trị thuộc tính ở cột English\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 24 | 90 |
| **5 - 6.5** | 5.75 | 149 | 586 |
| **6.5 - 8** | 7.25 | 428 | 2054 |
| **8 - 9** | 8.5 | 310 | 1437 |
| **9 - 10** | 9.5 | 102 | 386 |
| **Điểm số do người dùng nhập vào** | 6.2 | 1 | 13 |
| 7.1 | 1 | 6 |
| 7.6 | 1 | 1 |
| 9.4 | 1 | 3 |
| 8.6 | 1 | 17 |
| 8.8 | 1 | 2 |
| 7.4 | 1 | 10 |
| 5.0 | 1 | 0 |
| 9.2 | 1 | 0 |
| 6.7 | 1 | 3 |
| 5.0 | 1 | 67 |

Bảng 10: Chuẩn hóa dữ liệu điểm Ngoại ngữ

* **Chuẩn hóa dữ liệu *Literature\_score* ( điểm Ngữ văn ):** chuyển đổi giá trị thuộc tính ở cột Literature\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 21 | 83 |
| **5 - 6.5** | 5.75 | 112 | 478 |
| **6.5 - 8** | 7.25 | 505 | 2467 |
| **8 - 9** | 8.5 | 280 | 1358 |
| **9 - 10** | 9.5 | 37 | 167 |
| **Điểm số do người dùng nhập vào** | 5.0 | 28 | 62 |
| 7.8 | 1 | 18 |
| 7.4 | 1 | 13 |
| 7.2 | 1 | 10 |
| 6.5 | 2 | 9 |
| 8.7 | 1 | 3 |
| 8.1 | 1 | 3 |
| 7.6 | 1 | 2 |
| 7.9 | 1 | 1 |
| 5.6 | 1 | 1 |

Bảng 11: Chuẩn hóa dữ liệu điểm Ngữ văn

* **Chuẩn hóa dữ liệu *History\_score* ( điểm Lịch Sử ):** chuyển đổi giá trị thuộc tính ở cột History\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 14 | 26 |
| **5 - 6.5** | 5.75 | 59 | 174 |
| **6.5 - 8** | 7.25 | 377 | 1896 |
| **8 - 9** | 8.5 | 335 | 1596 |
| **9 - 10** | 9.5 | 166 | 795 |
| **Điểm số do người dùng nhập vào** | 5.0 | 33 | 129 |
| 8.9 | 2 | 8 |
| 7.5 | 2 | 4 |
| 9.7 | 1 | 3 |
| 8.4 | 1 | 13 |
| 8.7 | 1 | 10 |
| 8.8 | 1 | 3 |
| 7.7 | 1 | 1 |
| 8.6 | 1 | 17 |

Bảng 12: Chuẩn hóa dữ liệu điểm Lịch sử

* **Chuẩn hóa dữ liệu *Geography\_score* ( điểm Địa lý ):**chuyển đổi giá trị thuộc tính ở cột Literature\_score (Dưới 5; 5 - 6.5; 6.5 – 8; 8 – 9; 9 – 10) sang các giá trị số thống nhất ( lần lượt là: 2.5; 5.75; 7.25; 8.5; 9.5)

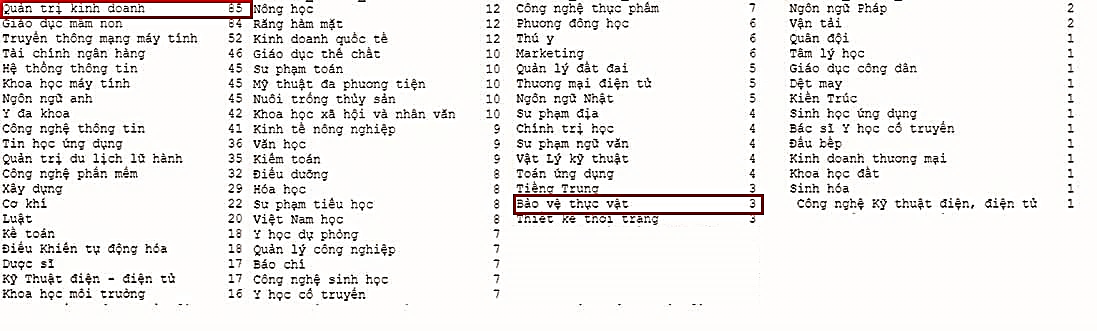
|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị thuộc tính tập dữ liệu gốc** | **Giá trị chuẩn hóa** | **Số lượng trước cân bằng** | **Số lượng sau cân bằng** |
| **Dưới 5** | 2.5 | 16 | 29 |
| **5 - 6.5** | 5.75 | 40 | 114 |
| **6.5 - 8** | 7.25 | 336 | 1658 |
| **8 - 9** | 8.5 | 389 | 1906 |
| **9 - 10** | 9.5 | 177 | 832 |
| **Điểm số do người dùng nhập vào** | 5.0 | 26 | 80 |
| 9.2 | 3 | 19 |
| 8.6 | 1 | 17 |
| 7.6 | 2 | 14 |
| 7.8 | 2 | 4 |
| 9.1 | 1 | 2 |
| 8.5 | 1 | 0 |

Bảng 13: Chuẩn hóa dữ liệu điểm địa lý

* 1. **Xử lý mất cân bằng dữ liệu**

**Phương pháp xử lý mất cân bằng dữ liệu**

Tình trạng mất cân bằng dữ liệu xảy ra khi số lượng của một phần tử nhãn thuộc tập dữ liệu có số lượng vượt trội nhiều so với các nhãn còn lại. Dữ liệu mất cân bằng ảnh hưởng đến kết quả dự báo, dẫn đến độ xác thực mô hình so với thực tế không chính xác. Hiện nay việc xử lý mất cân bằng dữ liệu có nhiều cách thực hiện, tùy theo bộ dữ liệu cần được cân bằng ở những hạn mục dữ liệu nào mà áp dụng đúng giải pháp cần thiết. Ở nghiên cứu này, bộ dữ liệu được thu thập ở quy mô nhỏ, các thuộc tính thu thập ở bộ dữ liệu không đồng đều về số lượng và có tỷ lệ chênh lệch cao, bộ dữ liệu mang kích thước nhỏ (1062 dòng). Để bộ dữ liệu có thể áp dụng vào mô hình và cho ra độ chính xác cao ta cần phải áp dụng phương pháp cân bằng làm tăng quy mô về kích thước nhưng vẫn giữ đúng mục tiêu ban đầu lúc thu thập dữ liệu. Ở tập dữ liệu dataset cũng đang mắc phải tình trạng mất cân bằng dữ liệu, theo biểu đồ số cho thấy ngành Quản trị kinh doanh có số lượng 85 nhiều vượt trội so với ngành Bảo vệ thực vật là 3.

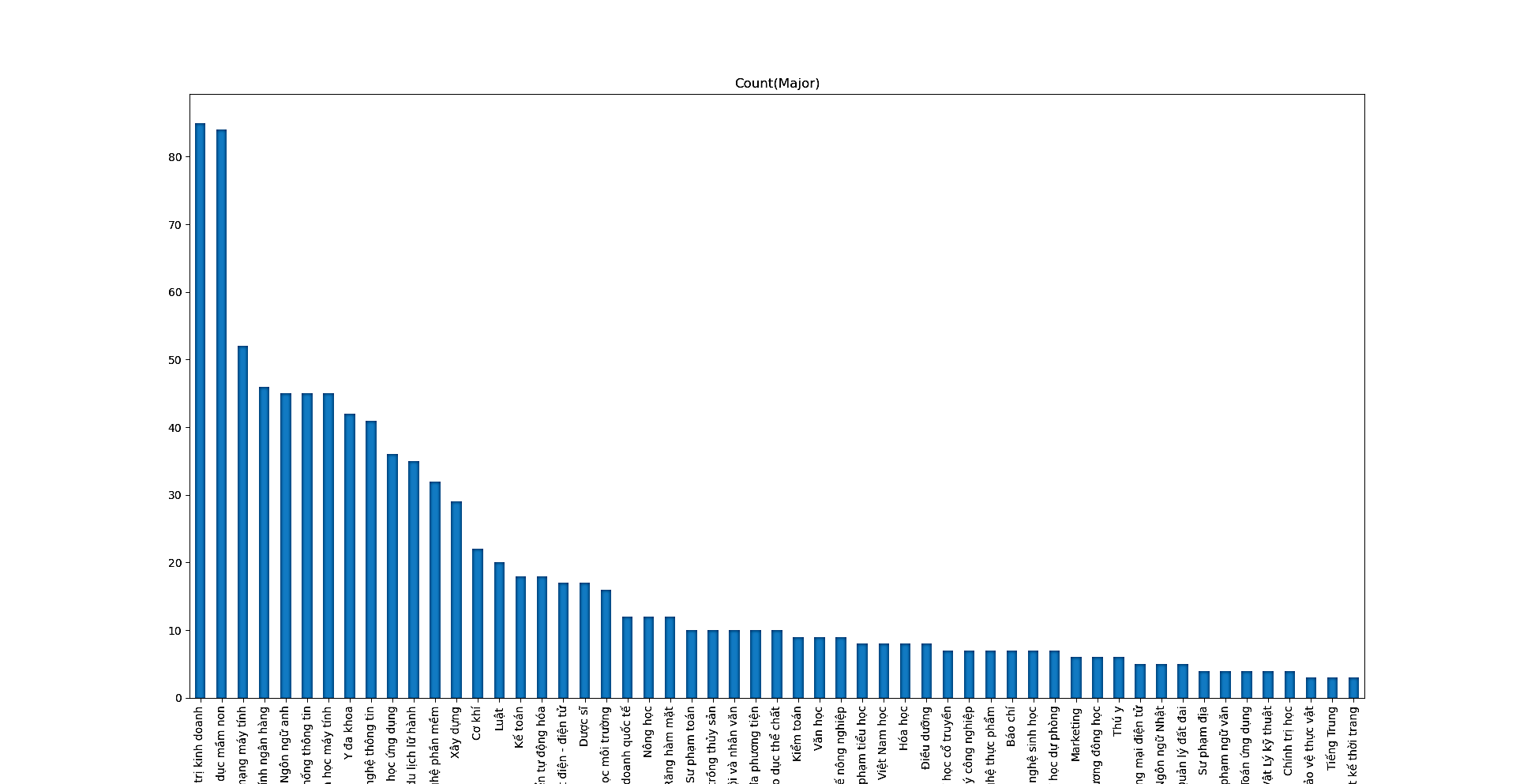
**

Hình 16: Thống kê số lượng các ngành

**Áp dụng Oversampling vào bài toán cụ thể:**

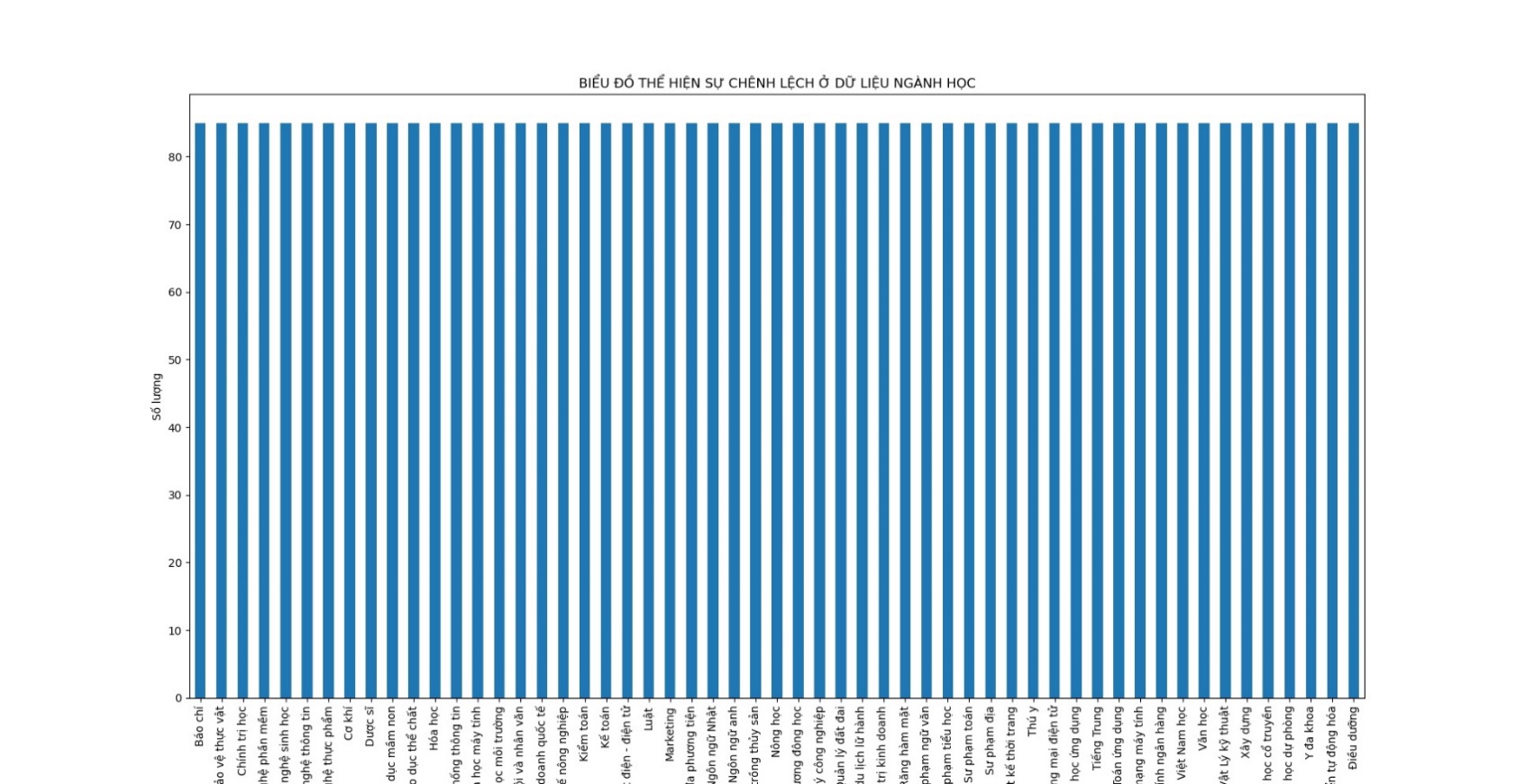
***OverSampling – Random Forest:*** ý tưởng ở phương pháp này là tạo ra giá trị thuộc tính mới dựa vào giá trị thuộc tính gốc bằng cách áp dụng giải thuật Random Forest. Đặc điểm ở phương pháp này là giá trị thuộc tính mới sẽ được tăng ngẫu nhiên.

**Trước cân bằng:**



***Trước cân bằng:*** *Biểu đồ thể hiện sự mất cân bằng ở cột thuộc tính nhãn*

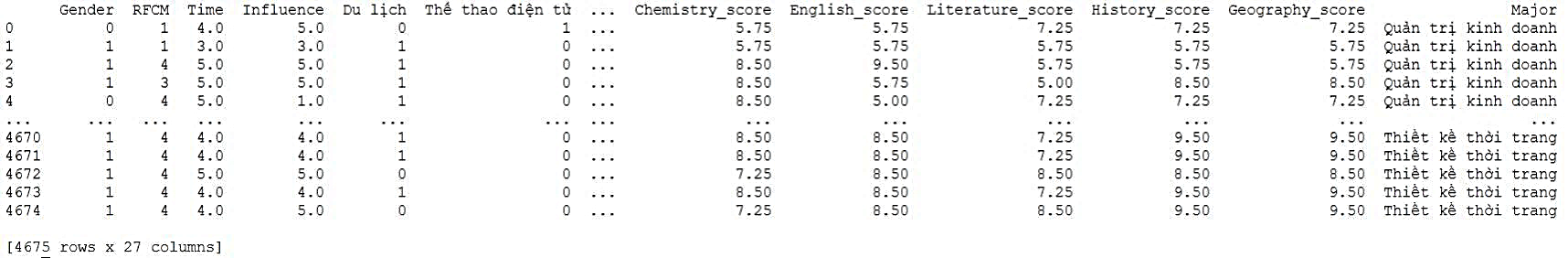
**Sau cân bằng**

**

***Sau cân bằng****: Biểu đồ thể hiện sự cân bằng ở cột thuộc tính nhãn*

**Kết quả cân bằng dữ liệu:** sau quá trình cân bằng dữ liệu, bộ dữ liệu đã có những thay đổi sau:

* Kích thước bộ dữ liệu từ 996 dòng mở rộng ra 4675 dòng và 27 cột.
* Giá trị thuộc tính cột nhãn được cân bằng, số lượng mỗi giá trị nhãn đồng giá trị là 85.



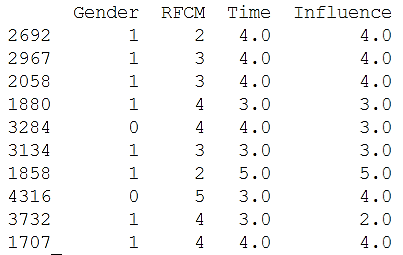
Hình 17: Thống kê số lượng các ngành sau cân bằng dữ liệu

1. **Phân loại nhóm người dùng ( Demographic Filter)**

**Áp dụng giải thuật K-Means để xác định nhóm người dùng trên tập dữ liệu dataset.**

Thuật toán K-Means clustering là một trong những thuật toán học không giám sát cơ đơn giản nhất để giải quyết vấn đề phân cụm. với mục đích chính là phân tập dữ liệu định sẵn thành *k* cụm với *k* là số cho trước. Kết quả của quá trình phân cụm là *k* nhóm dữ liệu, áp dụng thuật toán K-Means vào bài toán được mô tả trong ví dụ bên dưới.

Ví dụ: Tập dữ liệu có 10 dòng và 4 cột, 10 dòng tương ứng với 10 đối tượng, 4 cột tương ứng với số thuộc tính và sẽ có 4chiều dữ liệu. Phân lớp tập dữ liệu.



Lấy trọng tâm là 2 điểm: 2692[1; 2; 4; 5] và 3134[1; 3; 3; 3] (phân thành 2 lớp dữ liệu)

***Lần lập 1:***

***Áp dụng công thức [CT]: tính khoảng cách giữa các điểm so với trọng tâm 2692:***

KC[2692:2967] = = 1

**Tương tự ta có khoảng cách:**

KC[2692:2085] = 1

KC[2692:1880] = 2.449

KC[2692:3284] = 2.449

KC[2692:1858] = 1.414

KC[2692:4316] = 3.316

KC[2692:3732] = 3

KC[2692:1707] = 2

***Áp dụng công thức [ơss]: tính khoảng cách giữa các điểm so với trọng tâm 3134:***

KC[3134:2967] = = 1.414

**Tương tự ta có khoảng cách:**

KC[3134:2058] = 1.414

KC[3134:1880] = 1

KC[3134:3284] = 2.449

KC[3134:1858] = 3

KC[3134:4316] = 2.449

KC[3134:3732] = 1.414

KC[3134:1707] = 1.732

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Đối tượng** | **Khoảng cách Euclidean** | | **Lần lập 1** | |
| **Trọng tâm**  **2692** | **Trọng tâm**  **3134** | **Cụm 1 gần 2692** | **Cụm 2 gần 3134** |
| **2967** | 1 | 1.414 |  | y |
| **2058** | 1 | 1.414 |  | y |
| **1880** | 2.449 | 1 | x |  |
| **3284** | 2.449 | 2.449 | x | y |
| **1858** | 1.414 | 3 |  | y |
| **4316** | 3.316 | 2.449 | x |  |
| **3732** | 3 | 1.414 | x |  |
| **1707** | 2 | 1.732 | x |  |

Bảng 14: Bảng thống kê phân cụm lần 1

**Sau khi tính khoảng cách các điểm và xác định các điểm gần trọng tâm, cập nhật lại vị trí trọng tâm bằng công thức [CT]**

***Cập nhật trọng tâm cụm 1 gần điểm 2692:***

*Điểm mới là:*  ***[0.6, 4.2, 3.4, 3.2]***

***Cập nhật trọng tâm cụm 2 gần điểm 3134:***

*Điểm mới là:*  ***[0.75, 3, 3.5, 4]***

Sau khi có được điểm trọng tâm mới, tiếp tục tính khoản cách giữa tất cả các điểm so với điểm c1 và c2 và cập nhật lại trọng tâm c1n và c2n cho đến khi có được kêt quả phân lớp phù hợp với bài toán.

**Áp dụng trong bài toán nghiên cứu:**

Trong nghiên cứu này, sử dụng phương pháp K-Means Clustering với số cụm là 12 thuộc tính dùng để tính toán và huấn luyện bao gồm 4 cột: Gender – RFCM – Time – Influence, với số mục dữ liệu là 3740 dòng. Kết quả huấn luyện mô hình có được 12 nhóm dữ liệu người dùng dựa vào 4 thuộc tính là Gender, RFCM, Time, Infuence.

1. **Phương pháp gợi ý (Collabrative Filtering)**

Phương pháp lọc cộng tác dựa trên người dùng tập trung vào việc tìm kiếm người dùng có cùng sở thích, hoạt động gần giống nhau nhất [3]. Danh sách các gợi ý được xây dựng từ các mục dữ liệu được đánh giá cao bởi người dùng “tương tự”. Ví dụ sau đây áp dụng cho bài toán nghiên cứu:

Sau khi xác định nhóm người dùng mới dựa vào giải thuật K-Means, tiếp theo tính độ tương tự giữa người dùng mới và các người dùng cùng nhóm, do người dùng mới thuộc nhóm 4, nhóm 4 có người dùng số 164 nên ta có độ tương tự giữa hai người dùng là:

* Người dùng mới = [0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 8.0, 7.0, 6.5, 8.0, 5.5, 6.5, 9.5 ]
* Người dùng 164 = [1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 8.5, 8.5, 7.25, 8.5, 7.25, 7.25, 7.25 ]

Áp dụng công thức [] tính độ tương tự ta có:

**19.69**

**20.80**

***Độ tương tự giữa người dùng mới và người dùng 164:***

**Áp dụng chỉ số tương tự Cosine để xây dựng hệ thống gợi ý theo phương pháp Lọc cộng tác dựa trên người dùng.**

Sau quả trình tính độ tương tự: ta có được danh sách độ tương tự giữa người dùng mới so với các thông tin người dùng cùng nhóm trong tập dữ liệu, từ danh danh sách độ tương tự, chọn ra 3 chỉ số lớn nhất, truy xuất và đưa ra gợi ý 3 ngành học.

1. **Xây dựng hệ thống giao diện người dùng**

CHƯƠNG 3

**KIỂM THỬ VÀ ĐÁNH GIÁ**

PHẦN KẾT LUẬN

* 1. **Kết quả đạt được**
  2. **Hướng phát triển**

TÀI LIỆU THAM KHẢO

1. Bộ Lao động – Thương binh và Xã hội (2019), *Bản tin cập nhật thị trường lao động Việt Nam số 21 quý 1 năm 2019.*
2. Trường Đại học Vinh, tác giả Nguyễn Thị Kim Nhung – Lương Thị Thành Vinh (2018), *Đánh giá các yếu tố ảnh hưởng đến định hướng nghề nghiệp của học sinh trung học phổ thông tại nghệ an.*
3. Trường Đại học Cần Thơ, Khoa Kinh tế - Luận văn tốt nghiệp (2017), *Phân tích các nhân tố ảnh hưởng đến quyết định chọn ngành, Trường Đại học của học sinh THPT vùng ĐBSCL.*
4. Tác giả F.O.Isinkaye ,Y.O.Folajimi, B.A.Ojokoh (2015), *Recommendation systems: Principles, methods and eluavation.*
5. Khoa CNTT&TT ĐHCT, Nguyễn Nhựt Quỳnh (2018), *Xây dựng trang web gợi ý phim dựa vào thể loại hoặc diễn viên.*
6. Tác giả Yugantshekhar (2020), *ML – Content Based Recommendation System – geeksforgeek.org*
7. Tác giả Aniruddha Bhandari (2020), *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization*
8. Tác giả Phạm Đình Khánh (2020), *Mất cân bằng dữ liệu (Imbalanced dataset).*
9. Tác giả Iateilang, L. Chameikho, Don Bosco College Of Engineering & Technology Assam Don Bosco University (2014), *Recommender Systems: Types of Filtering Techniques (IJERT).*
10. Khoa Công nghệ thông tin & Truyền thông, Trường Đại học Cần Thơ – Tác giả: Trần Nguyễn Minh Thư, Phạm Xuân Hiền (2016) – *Các phương pháp đánh giá hệ thống gợi ý.*
11. Mortensen M.,2007. *Design and evaluation of a recommender system*.

Danh sách các website kham khảo.

1. <https://moet.gov.vn/thong-ke/Pages/thong-ko-giao-duc-dai-hoc.aspx?ItemID=6636>.
2. <http://sqladvice.com/postgresql-la-gi-so-sanh-mysql-va-postgresql/>
3. <https://vimentor.com/vi/lesson/tien-xu-ly-du-lieu-trong-linh-vuc-hoc-may-phan-1>
4. <https://phantichspss.com/he-so-tuong-quan-pearson-cach-thao-tac-phan-tich-tuong-quan-trong-spss.html\>
5. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

* Các địa chỉ web trên điều có thể truy cập cho đến ngày …..