TRƯỜNG ĐẠI HỌC HUTECH
KHOA CÔNG NGHỆ THÔNG TIN

Lâm Trần Tấn Phát, Trần Công Hậu

ỨNG DỤNG MÔ HÌNH CVAE - CONDITIONAL VARIATIONAL AUTOENCODER ĐỂ SINH DỮ LIỆU KHẢO SÁT NGƯỜI ĐI LÀM

ĐỒ ÁN MÔN HỌC
THỐNG KÊ MÁY TÍNH & ỨNG DỤNG

GIÁO VIÊN HƯỚNG DẪN:

HOÀNG VĂN QUÝ

Hồ Chí Minh, ngày 20 tháng 12 năm 2025

LỜI MỞ ĐẦU

Trong bối cảnh các nghiên cứu xã hội và giáo dục ngày càng ứng dụng mạnh mẽ các phương pháp học máy, chất lượng và cấu trúc của dữ liệu khảo sát đóng vai trò then chốt đối với độ tin cậy của kết quả nghiên cứu. Tuy nhiên, dữ liệu khảo sát thực tế thường gặp nhiều hạn chế như mất cân bằng nghiêm trọng giữa các nhóm đối tượng, khó thu thập, tốn kém chi phí và tiềm ẩn rủi ro về quyền riêng tư. Xuất phát từ thực tiễn đó, đề tài này tập trung nghiên cứu và ứng dụng mô hình **Conditional Variational Autoencoder (CVAE)** nhằm sinh dữ liệu khảo sát có điều kiện, phục vụ mục tiêu cải thiện chất lượng dữ liệu huấn luyện và nâng cao hiệu quả dự đoán nghề nghiệp. Nghiên cứu được thực hiện với định hướng vừa đảm bảo tính khoa học, vừa hướng đến khả năng ứng dụng thực tế trong các bài toán hướng nghiệp và phân tích dữ liệu khảo sát.

LỜI CAM KẾT

Tôi xin cam kết rằng toàn bộ nội dung trình bày trong báo cáo này là kết quả của quá trình nghiên cứu nghiêm túc và độc lập, dựa trên dữ liệu, mô hình và các thí nghiệm do chính nhóm chúng tôi thực hiện. Các số liệu, kết quả và nhận định đều được trình bày trung thực, có đối chiếu và đánh giá rõ ràng; mọi tài liệu tham khảo đều được trích dẫn minh bạch. Tôi hoàn toàn chịu trách nhiệm về tính chính xác và trung thực của báo cáo này.

Người cam kết:

Lâm Trần Tấn Phát

Trần Công Hậu

Ngày 24 tháng 12 năm 2025

Mục lục

Danh mục hình ảnh (Hậu)

1. CHƯƠNG 1: TỔNG QUAN

1.1. Đặt vấn đề

Dữ liệu khảo sát xã hội học có vai trò quan trọng nhưng thường gặp khó khăn về chi phí thu thập, bảo mật và đặc biệt là sự mất cân bằng giữa các nhóm nghề, dẫn đến mô hình dự đoán bị thiên lệch. Các phương pháp oversampling truyền thống (như SMOTE) hạn chế do không phù hợp với đặc thù dữ liệu rời rạc (Likert, nhị phân) của khảo sát. Do đó, nghiên cứu này đề xuất sử dụng **Conditional Variational Autoencoder (CVAE)** để sinh dữ liệu giả lập theo từng nhóm nghề. Phương pháp này giúp khắc phục tình trạng mất cân bằng lớp đồng thời bảo toàn cấu trúc thống kê của dữ liệu gốc, nhằm nâng cao hiệu quả cho các mô hình tư vấn hướng nghiệp.

1.2. Mục tiêu đề tài

Tập dữ liệu khảo sát người đi làm được sử dụng trong nghiên cứu gồm tổng cộng *759* mẫu, được phân chia thành 6 nhóm nghề nghiệp khác nhau. Phân phối số lượng mẫu giữa các nhóm nghề cho thấy sự mất cân bằng rõ rệt và mang tính cấu trúc.

Cụ thể, nhóm nghề chiếm tỷ lệ lớn nhất là nhóm 0 với 398 mẫu, tương đương khoảng *52,4%* tổng số dữ liệu. Các nhóm nghề tiếp theo gồm nhóm 1 với 141 mẫu (*18,6%*), nhóm 3 với 111 mẫu (*14,6%*) và nhóm 2 với 92 mẫu (*12,1%*). Trong khi đó, hai nhóm nghề còn lại là nhóm 4 và nhóm 5 chỉ có lần lượt 14 mẫu *(1,8%*) và 3 mẫu (*0,4%*), chiếm tỷ lệ rất nhỏ trong toàn bộ tập dữ liệu.

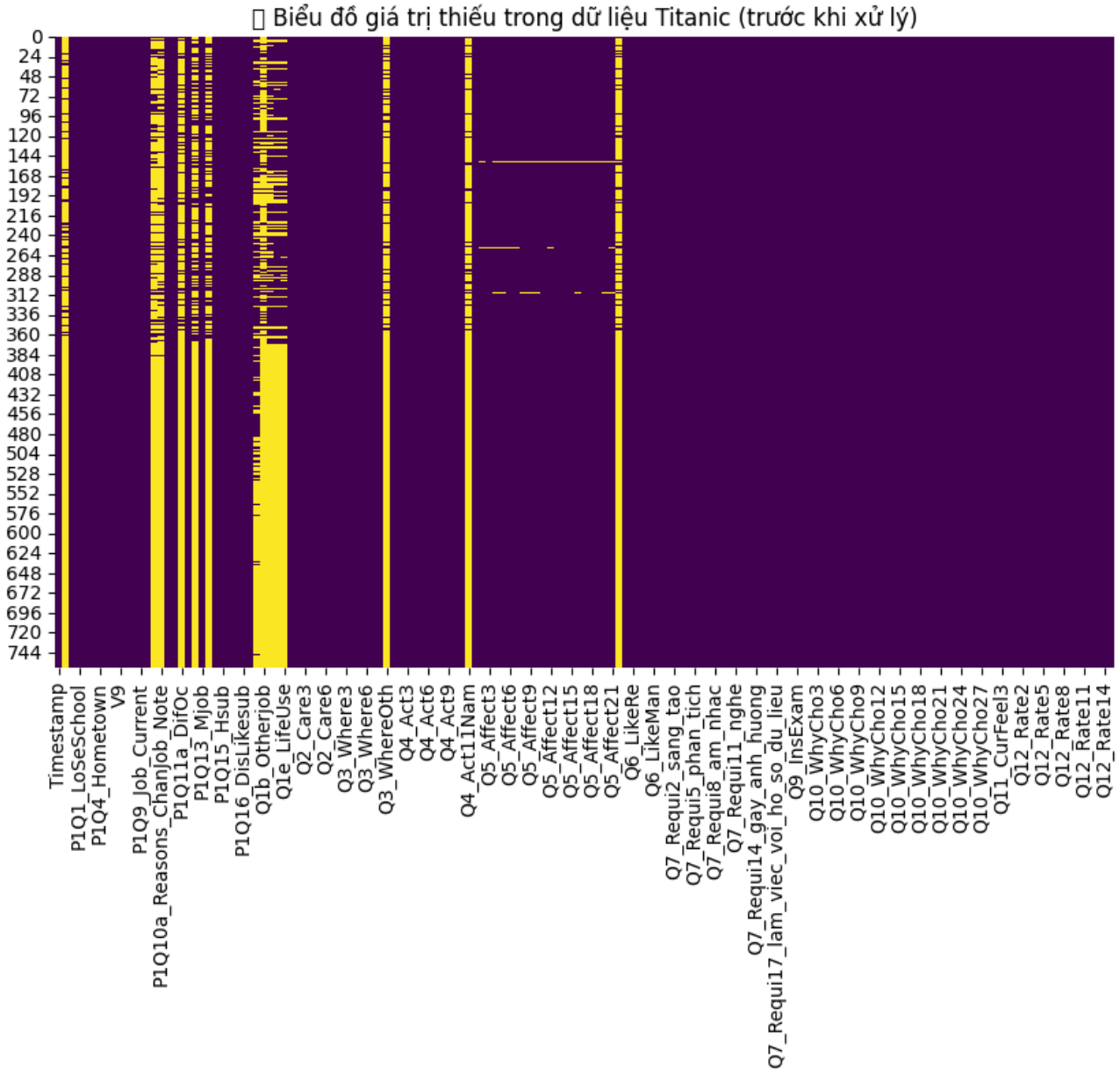
Hệ quả trực tiếp là khả năng dự đoán nghề nghiệp của mô hình trở nên thiên lệch, đặc biệt đối với nhóm nghề 4 và nhóm nghề 5, vốn chỉ chiếm chưa đến 3% tổng số dữ liệu. Trong bối cảnh nghiên cứu hướng đến ứng dụng hướng nghiệp, việc mô hình không dự đoán tốt các nhóm nghề hiếm sẽ làm giảm đáng kể giá trị ứng dụng thực tế.

Do đó mục tiêu đề tài là tìm kiếm một phương pháp xử lý mất cân bằng phù hợp với đặc thù dữ liệu khảo sát. Bằng cách sinh dữ liệu từ dữ liệu có thật.

2. CHƯƠNG 2 : TỔNG QUAN DỮ LIỆU KHẢO SÁT

2.1. Xử lý dữ liệu

2.1.1. Lọc nhiễu và loại bỏ dữ liệu



Hình 2.1.1 : Biểu đồ giá trị thiếu

Để đảm bảo chất lượng dữ liệu đầu vào cho mô hình, nghiên cứu áp dụng hai chiến lược xử lý:

- Lọc tự động: Các cột có tỷ lệ giá trị thiếu vượt quá 80% được tự động loại bỏ. Ngưỡng này được lựa chọn nhằm loại bỏ những biến không đủ thông tin thống kê, tránh gây nhiễu và sai lệch trong quá trình huấn luyện mô hình.
- Loại bỏ thủ công: Một số biến định danh và dữ liệu văn bản không mang ý nghĩa dự đoán, *Timestamp*, *Tên trường*, *Quê quán*, được chủ động loại bỏ. Việc này giúp giảm nhiễu, đồng thời đảm bảo dữ liệu tập trung vào các biến khảo sát có giá trị phân tích và mô hình hóa.

Sau bước xử lý này, tập dữ liệu thu được có cấu trúc gọn hơn, ít nhiễu.

2.1.2. Chuẩn hóa dữ liệu

Sau bước loại bỏ dữ liệu thiếu và lọc nhiễu, dữ liệu tiếp tục được chuẩn hóa nhằm đảm bảo tính nhất quán trong quá trình xử lý và phân tích. Cụ thể, toàn bộ tên cột được chuyển về dạng chữ thường thông qua hàm df.columns.str.lower(). Đồng thời, các biến dữ liệu dạng chuỗi được chuẩn hóa bằng cách chuyển về chữ thường và loại bỏ khoảng trắng dư thừa ở đầu và cuối chuỗi.

2.1.3. Xử lý Logic nghiệp vụ

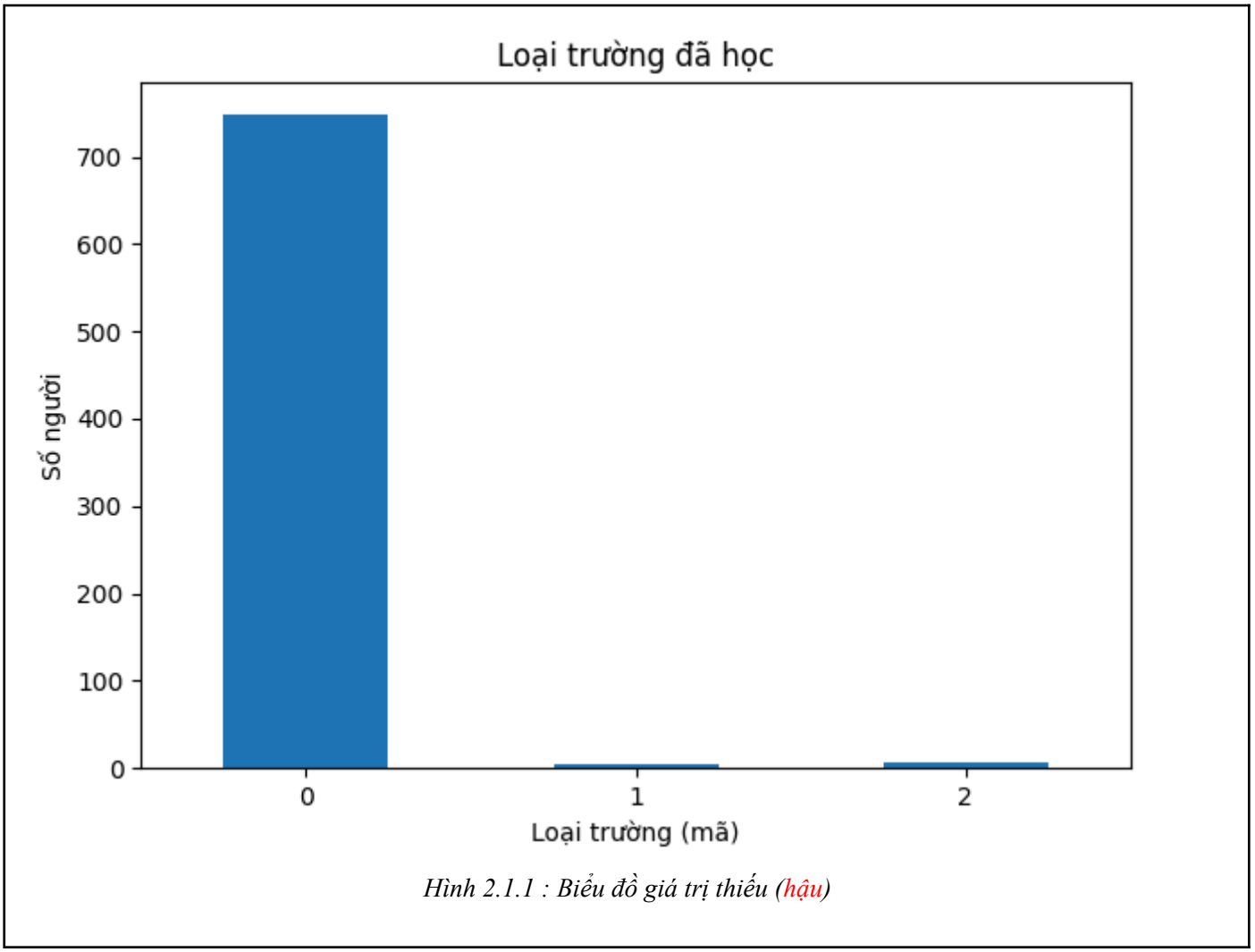
Đối với câu hỏi định hướng nghề nghiệp, nếu người trả lời chọn "Không" ở câu hỏi nền tảng q1_trainjob = 2, thì toàn bộ các câu hỏi chi tiết liên quan phía sau được gán giá trị -2, biểu thị trạng thái "Không áp dụng".
Đối với các giá trị còn thiếu nhưng không thuộc trường hợp trên, dữ liệu được gán giá trị -1, tương ứng với "Không trả lời/Không nhớ". Cách tiếp cận này cho phép phân biệt rõ ràng giữa các trạng thái dữ liệu khác nhau, thay vì loại bỏ hoặc gộp chung các giá trị thiếu. Nhờ đó, dữ liệu đầu vào vẫn giữ được cấu trúc logic và ý nghĩa nghiệp vụ, đồng thời phù hợp hơn cho các bước mã hóa và mô hình hóa sau này.

2.1.4. Mã hóa nhị phân

Đối với các câu hỏi cho phép chọn nhiều đáp án, đặc biệt là nhóm câu hỏi về yêu cầu kỹ năng nghề nghiệp (Câu 7), nghiên cứu áp dụng phương pháp mã hóa nhị phân nhằm chuyển đổi dữ liệu dạng chuỗi phức hợp sang dạng số có thể sử dụng cho mô hình học máy.
Cụ thể, các câu trả lời chứa nhiều lựa chọn được xử lý bằng cách quét theo từng nhóm nghề (Kỹ thuật, Nghiên cứu, Nghệ thuật, Xã hội, Quản lý, Nghiệp vụ). Thông qua việc sử dụng hàm kiểm tra chuỗi (str.contains), mỗi lựa chọn được tách thành một cột nhị phân riêng biệt, với giá trị 1 biểu thị lựa chọn được chọn và 0 biểu thị không được chọn.
Cách mã hóa này giúp bảo toàn đầy đủ thông tin của các câu trả lời đa lựa chọn, đồng thời chuyển đổi dữ liệu về dạng có cấu trúc rõ ràng, thuận lợi cho cả phân tích thống kê và huấn luyện các mô hình học máy ở các bước tiếp theo.

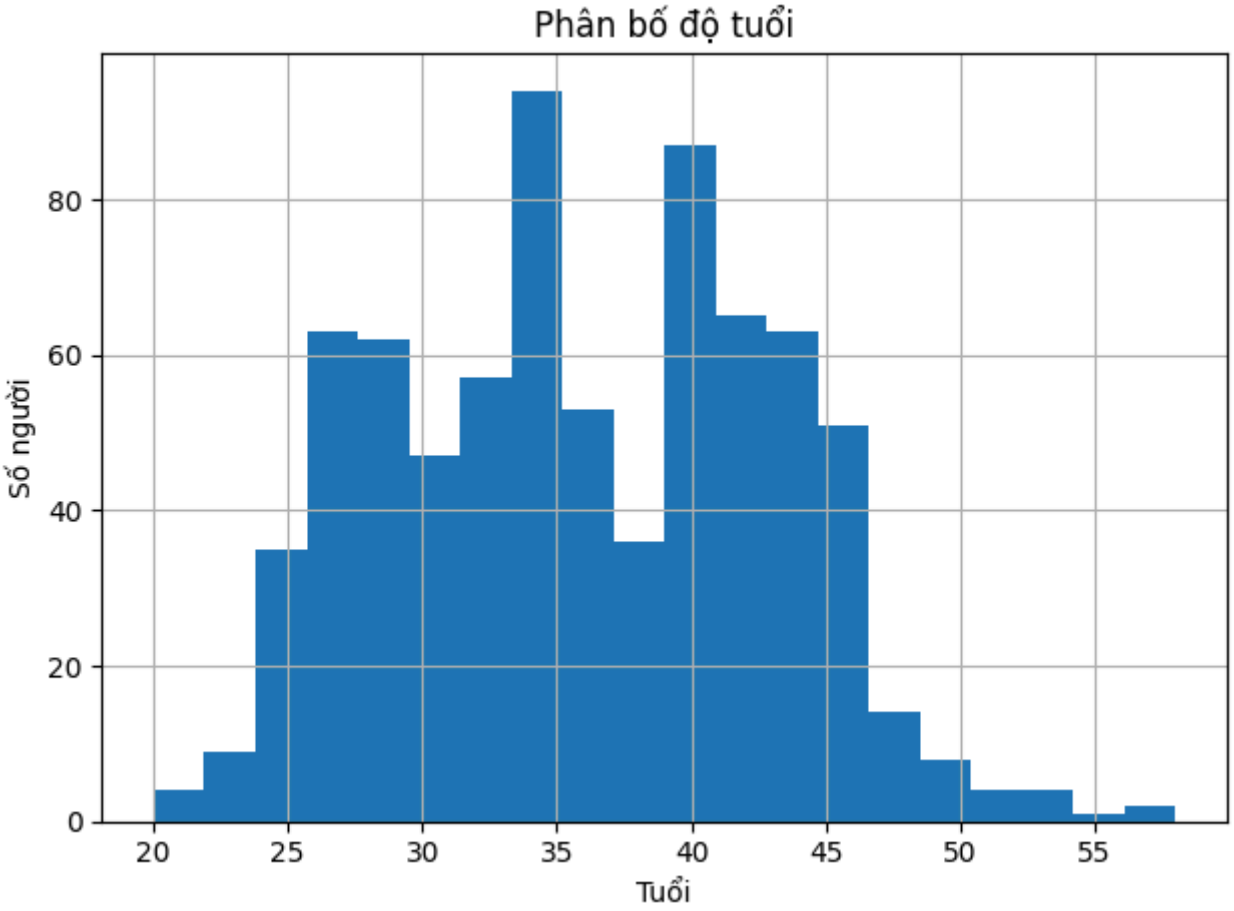
2.2. Trực quan dữ liệu

2.2.1. Tổng quan đặc điểm người tham gia khảo sát



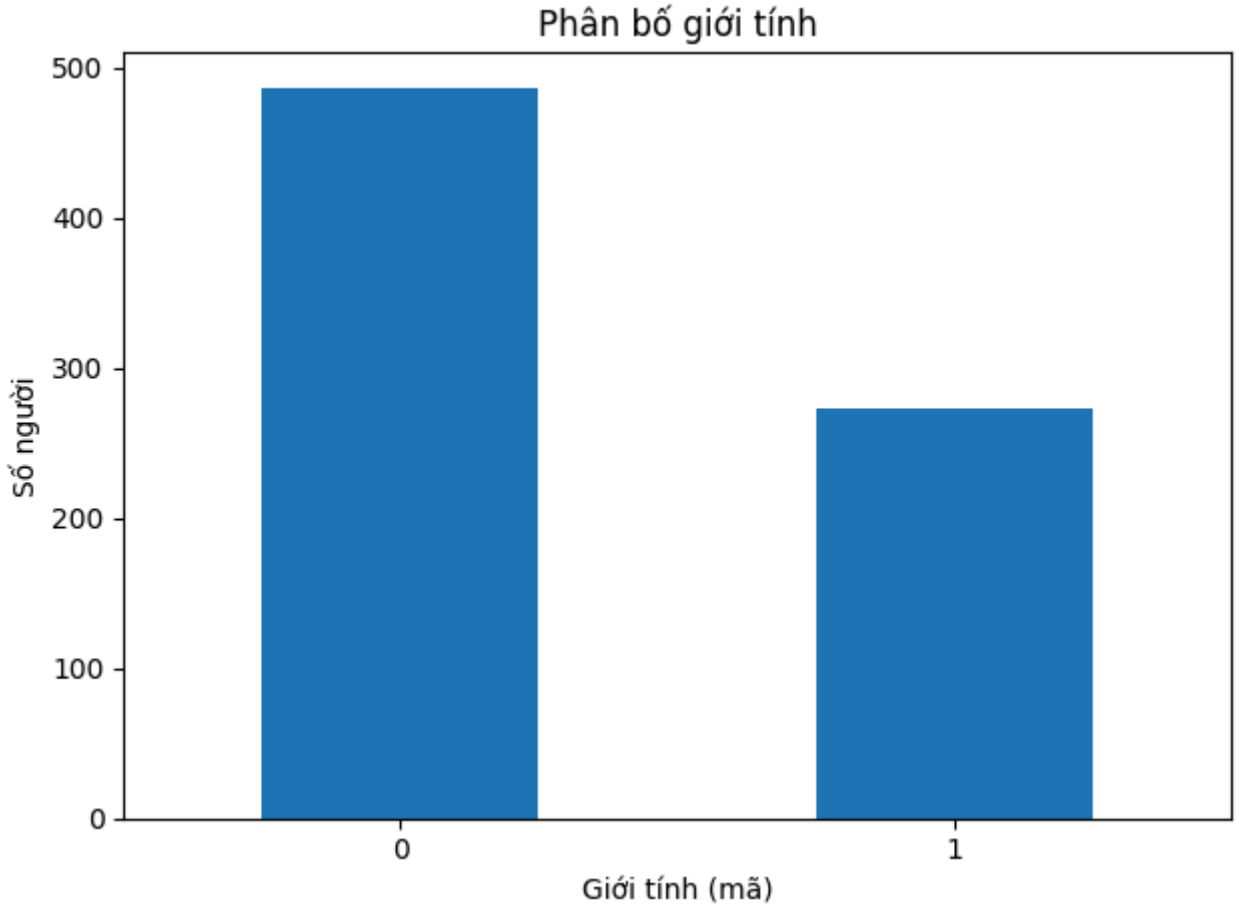
Hình 2.1.1 : Biểu đồ giá trị thiếu (hậu)

Trước hết, xét về loại trường đã từng theo học, phần lớn người tham gia xuất thân từ cùng một nhóm loại hình trường học, trong khi các loại hình khác chỉ chiếm tỷ lệ rất nhỏ. Điều này cho thấy mẫu khảo sát có sự tập trung cao vào một nhóm nền tảng giáo dục chủ đạo, phản ánh đặc điểm thực tế của đối tượng được tiếp cận trong quá trình thu thập dữ liệu.



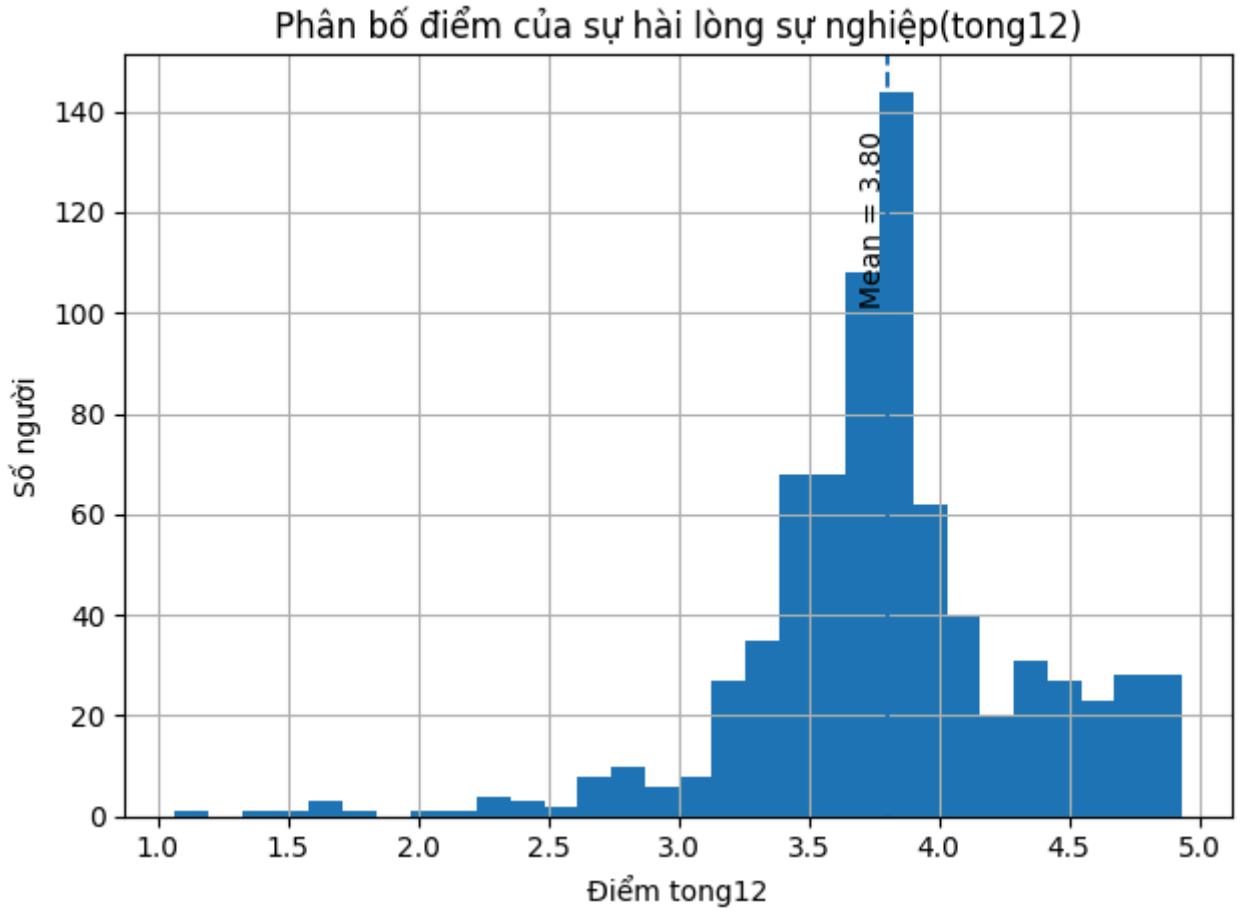
Hình 2.1.1 : Biểu đồ giá trị thiếu

Về độ tuổi, phân bố cho thấy người tham gia chủ yếu nằm trong độ tuổi lao động trung bình, tập trung nhiều ở nhóm từ khoảng cuối 20 đến đầu 40 tuổi. Phân bố tuổi không đồng đều hoàn toàn mà có xu hướng tập trung quanh một số mốc tuổi nhất định, cho thấy mẫu khảo sát nghiêng về nhóm người đã có thời gian làm việc và trải nghiệm nghề nghiệp tương đối ổn định.



Hình 2.1.1 : Biểu đồ giá trị thiếu

Xét theo **giới tính**, dữ liệu cho thấy sự chênh lệch nhất định giữa hai nhóm, với một nhóm chiếm tỷ lệ cao hơn. Tuy nhiên, cả hai giới đều được đại diện trong tập dữ liệu, đảm bảo dữ liệu phản ánh được sự đa dạng giới tính ở mức cơ bản.



Hình 2.1.1 : Biểu đồ giá trị thiếu

Đối với mức độ hài lòng về sự nghiệp, phân bố điểm cho thấy đa số người tham gia đánh giá mức độ hài lòng ở mức trung bình đến khá. Giá trị trung bình của thang đo tập trung quanh mức cao hơn trung vị, cho thấy phần lớn người được khảo sát có cảm nhận tích cực về con đường sự nghiệp hiện tại, trong khi vẫn tồn tại một số ít trường hợp đánh giá ở mức thấp hoặc rất cao.

Tổng hợp các đặc điểm trên cho thấy nhóm người tham gia khảo sát chủ yếu là người đi làm trong độ tuổi lao động chính, có nền tảng giáo dục tương đối đồng nhất và nhìn chung có mức độ hài lòng nghề nghiệp khá ổn định.

2.2.2. Nghề sau học xong và nghề hiện tại

Hình 2.1.1 : Biểu đồ giá trị thiếu

Biểu đồ so sánh cho thấy sự khác biệt rõ rệt giữa lựa chọn nghề nghiệp của người tham gia ngay sau khi hoàn thành bậc THCS và nghề nghiệp hiện tại của họ tại thời điểm khảo sát.

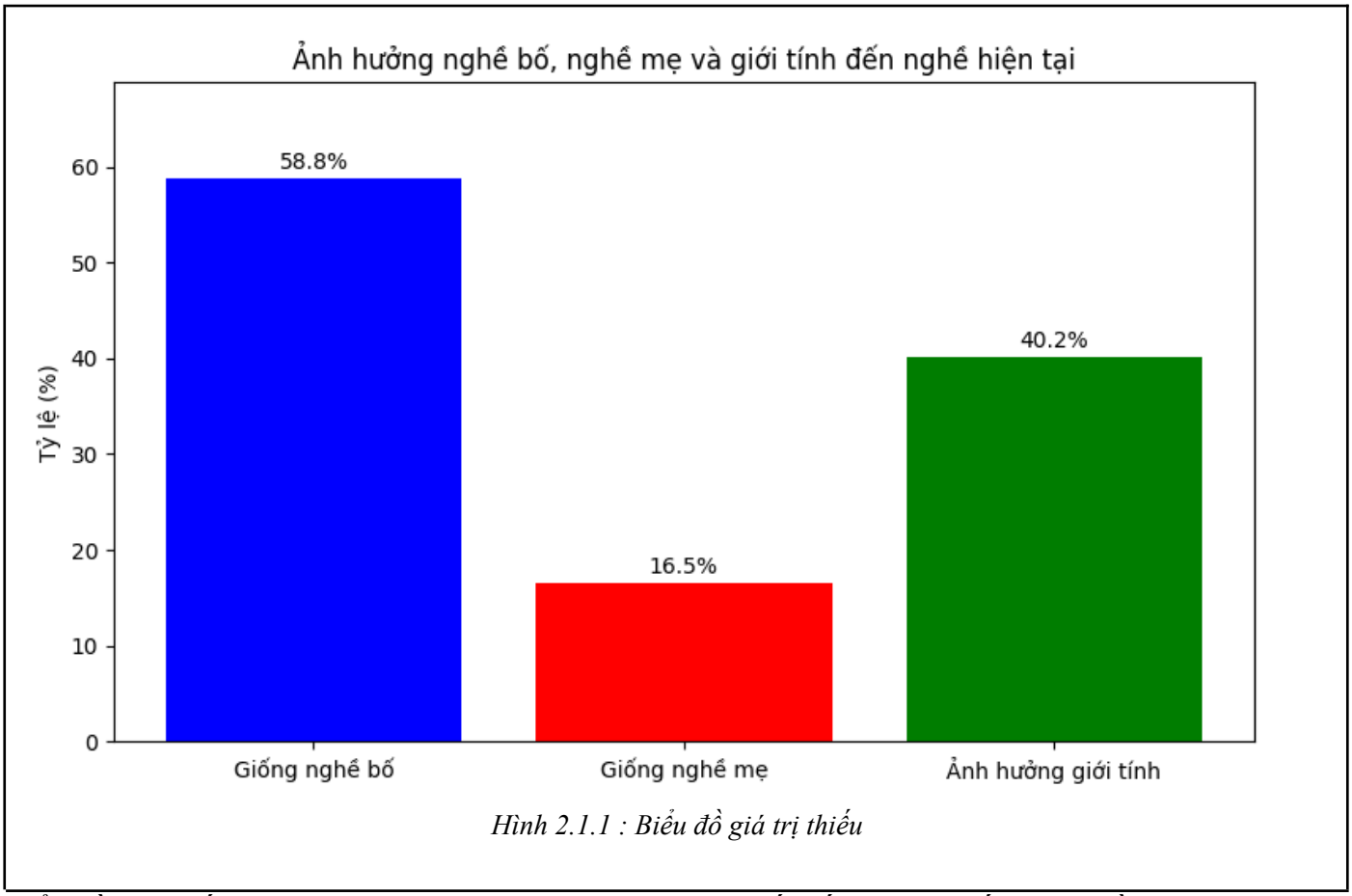
Trước hết, có thể quan sát rằng ở thời điểm sau khi học xong THCS, lựa chọn nghề của người tham gia tập trung rất mạnh vào một nhóm nghề cụ thể, trong khi các nhóm nghề còn lại hầu như không được lựa chọn hoặc chỉ xuất hiện với số lượng rất nhỏ.

Ngược lại, phân bố nghề nghiệp hiện tại cho thấy mức độ đa dạng cao hơn rõ rệt. Số lượng người làm việc ở các nhóm nghề khác nhau tăng lên, đặc biệt ở những nhóm nghề trước đó gần như không được lựa chọn ở giai đoạn sau THCS..

Sự chênh lệch giữa hai phân bố cho thấy rằng lựa chọn nghề nghiệp ban đầu không hoàn toàn quyết định nghề nghiệp sau này, mà có thể chỉ đóng vai trò như một định hướng sớm.

Biểu đồ này cung cấp một góc nhìn trực quan quan trọng về khoảng cách giữa định hướng nghề nghiệp sớm và thực tế nghề nghiệp sau này, đồng thời nhấn mạnh tính động và biến đổi của quá trình hình thành nghề nghiệp trong suốt vòng đời học tập và làm việc.

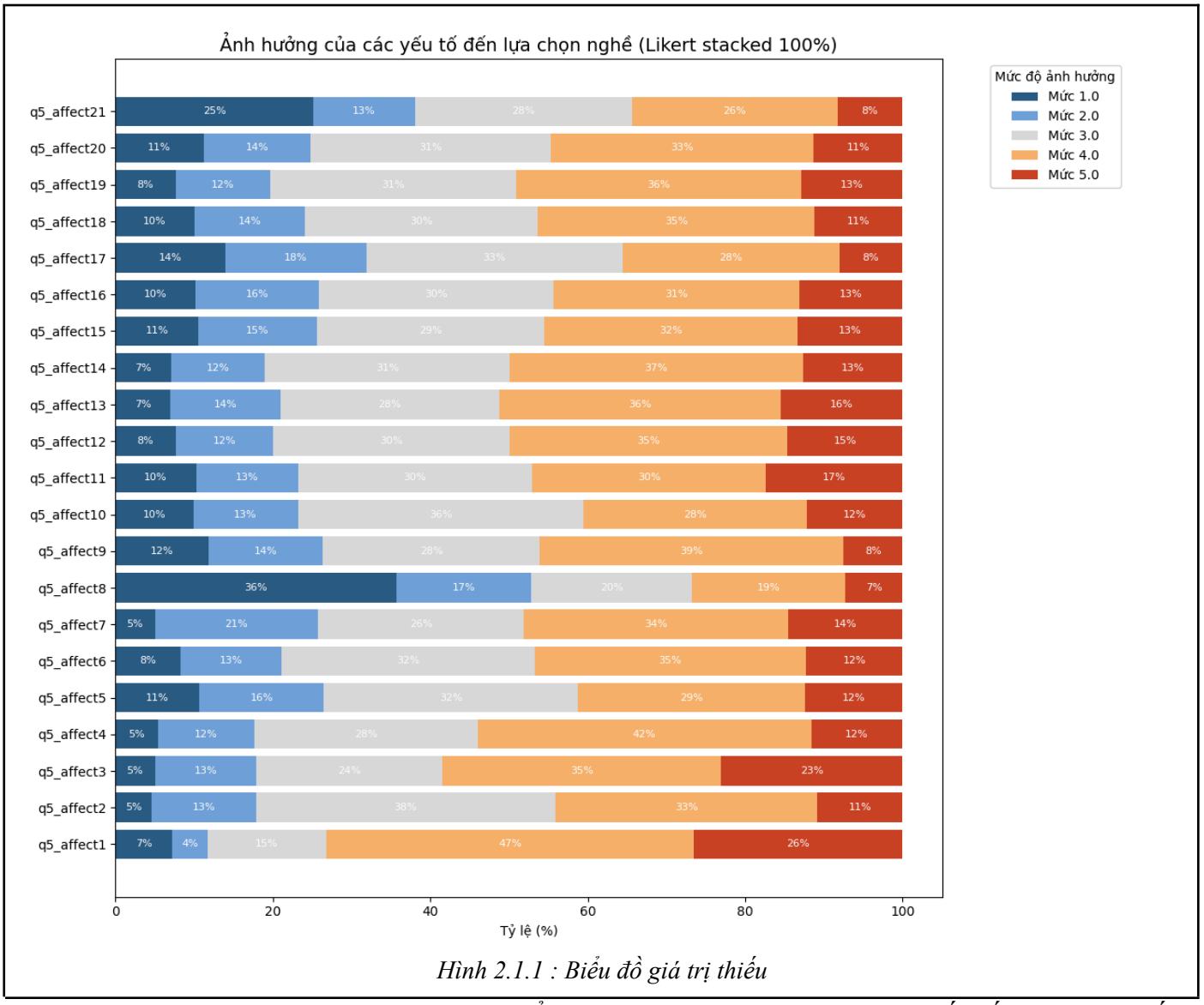
2.2.3. Yếu tố ảnh hưởng đến nghề hiện tại



Hình 2.1.1 : Biểu đồ giá trị thiếu

Biểu đồ cho thấy mức độ ảnh hưởng khác nhau của các yếu tố gia đình đối với nghề nghiệp hiện tại của người tham gia khảo sát. Trong đó, nghề nghiệp của bố có tỷ lệ giống cao nhất, với khoảng 58,8%.

Ngược lại, giống nghề nghiệp của mẹ thể hiện ở mức thấp hơn đáng kể, chỉ chiếm khoảng 16,5%. Bên cạnh đó, khoảng 40,2% giới tính sẽ quyết định nghề nghiệp hiện tại sẽ giống bố hay mẹ.



Hình 2.1.1 : Biểu đồ giá trị thiếu

Hình trực quan dạng Likert stacked 100% thể hiện mức độ ảnh hưởng của 21 yếu tố khác nhau đến quyết định lựa chọn nghề nghiệp của người được khảo sát trong giai đoạn học THCS. Mỗi yếu tố được đánh giá theo 5 mức độ, từ *Không ảnh hưởng* đến *Ảnh hưởng mạnh*, cho phép quan sát đồng thời cả xu hướng chung và sự phân hóa mức độ tác động.

Kết quả cho thấy, đa số các yếu tố có tỷ lệ cao ở các mức "Tương đối ảnh hưởng" và "Ảnh hưởng mạnh", phản ánh việc lựa chọn nghề nghiệp là kết quả của nhiều tác động kết hợp, thay vì chỉ phụ thuộc vào một yếu tố đơn lẻ.

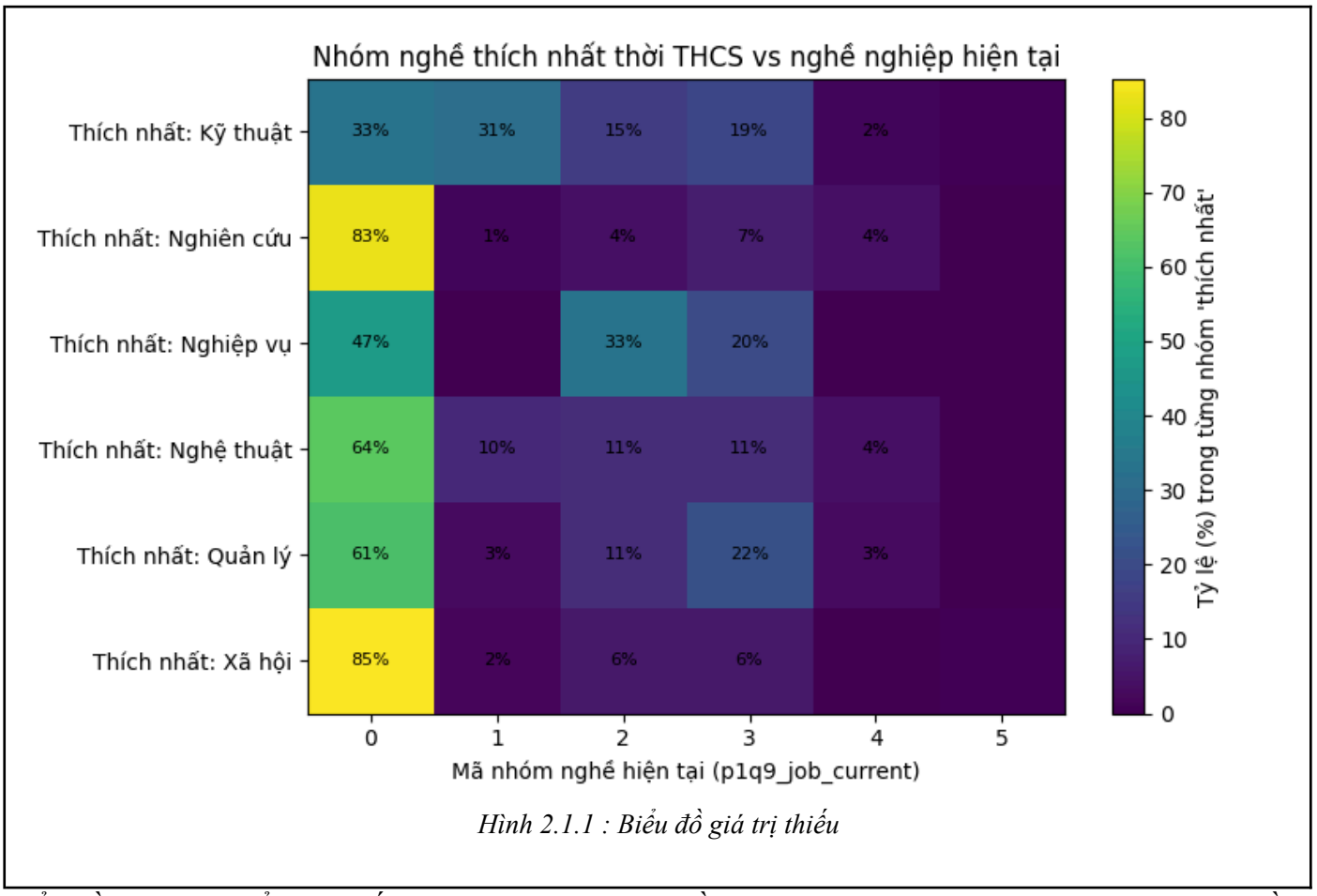
Các nhóm yếu tố liên quan đến bản thân học sinh như:

- *Am hiểu về kiến thức, kỹ năng nghề nghiệp* (effect1),
- *Phù hợp với hứng thú, năng khiếu, sở trường* (effect2)
- *Chủ động, độc lập khi lựa chọn nghề* (effect4),

đều có tỷ lệ đánh giá cao ở hai mức ảnh hưởng mạnh nhất, cho thấy vai trò quan trọng của nhận thức cá nhân và sự tự chủ trong định hướng nghề nghiệp từ sớm.

Bên cạnh đó, yếu tố kinh tế – xã hội như *thu nhập nghề nghiệp* (effect12), *nhu cầu nhân lực xã hội* (effect11), và *cơ hội thăng tiến, phát triển nghề* (effect18) cũng được đánh giá có mức ảnh hưởng đáng kể.

Ngược lại, một số yếu tố mang tính ngoại cảnh hoặc cảm tính như *a dua theo bạn bè* (effect8) hay *truyền thông, quảng cáo* (effect17) có tỷ lệ cao hơn ở các mức ảnh hưởng thấp và trung bình, cho thấy vai trò của các yếu tố này tuy tồn tại nhưng không mang tính chi phối.



Hình 2.1.1 : Biểu đồ giá trị thiếu

Biểu đồ heatmap thể hiện mối liên hệ giữa nhóm nghề yêu thích khi còn học THCS và nhóm nghề mà người khảo sát đang làm hiện tại. Giá trị trong mỗi ô biểu diễn tỷ lệ (%) phân bố nghề hiện tại trong từng nhóm yêu thích ban đầu.

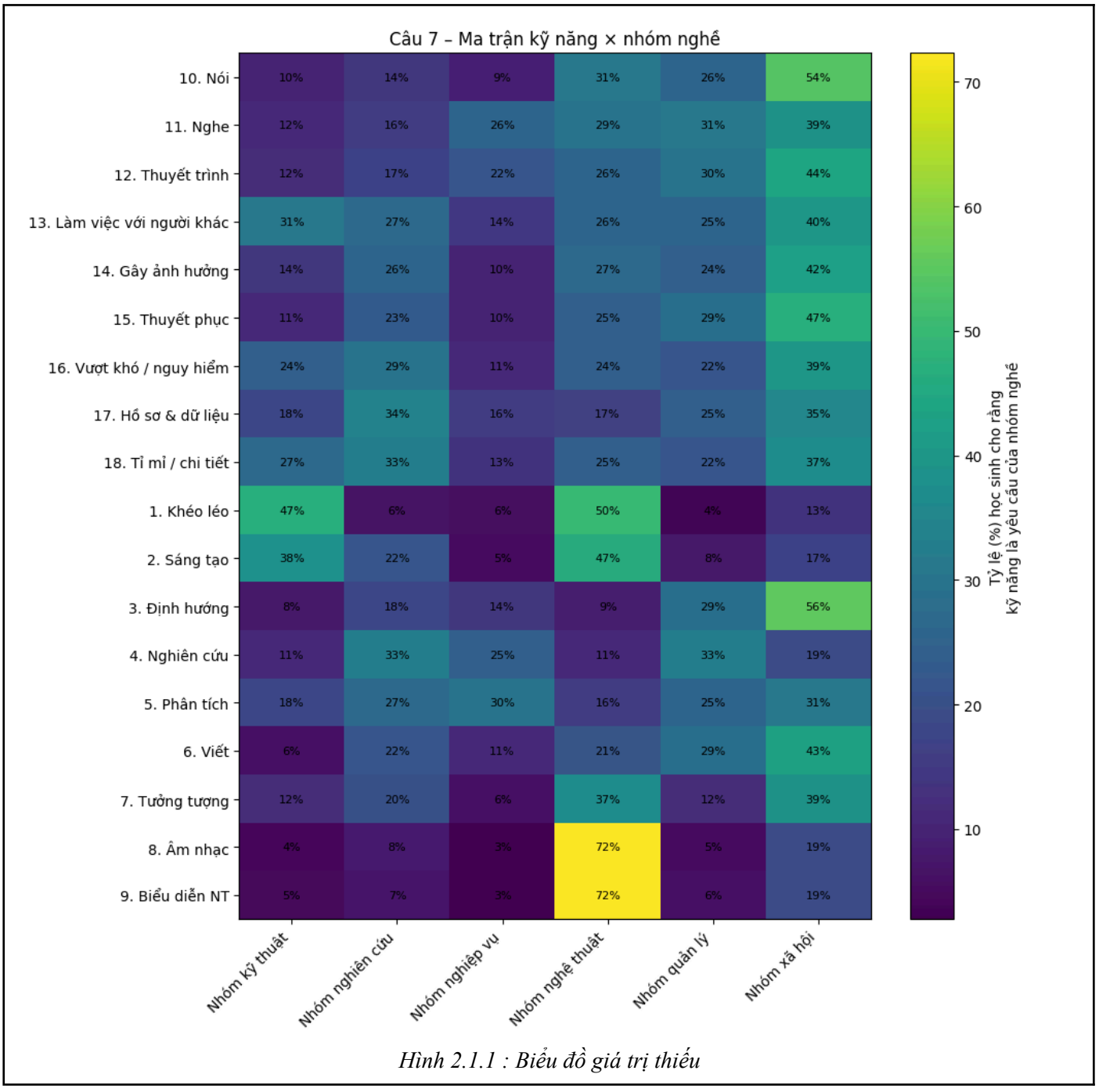
Biểu đồ heatmap thể hiện mối liên hệ giữa nhóm nghề yêu thích khi còn học THCS và nhóm nghề mà người khảo sát đang làm hiện tại. Giá trị trong mỗi ô biểu diễn tỷ lệ (%) phân bố nghề hiện tại trong từng nhóm yêu thích ban đầu.

Một số nhóm thể hiện mức độ "khớp" tương đối cao hơn, chẳng hạn:

- Nhóm Nghiệp vụ có tỷ lệ đáng kể chuyển sang nghề nghiệp cùng nhóm,
- Nhóm Quản lý có phân bố đa dạng hơn, trải sang nhiều nhóm nghề hiện tại.

Ngược lại, các nhóm như Nghiên cứu và Xã hội cho thấy sự chênh lệch rất lớn, với trên 80% người khảo sát cuối cùng làm nghề thuộc nhóm khác so với sở thích ban đầu.

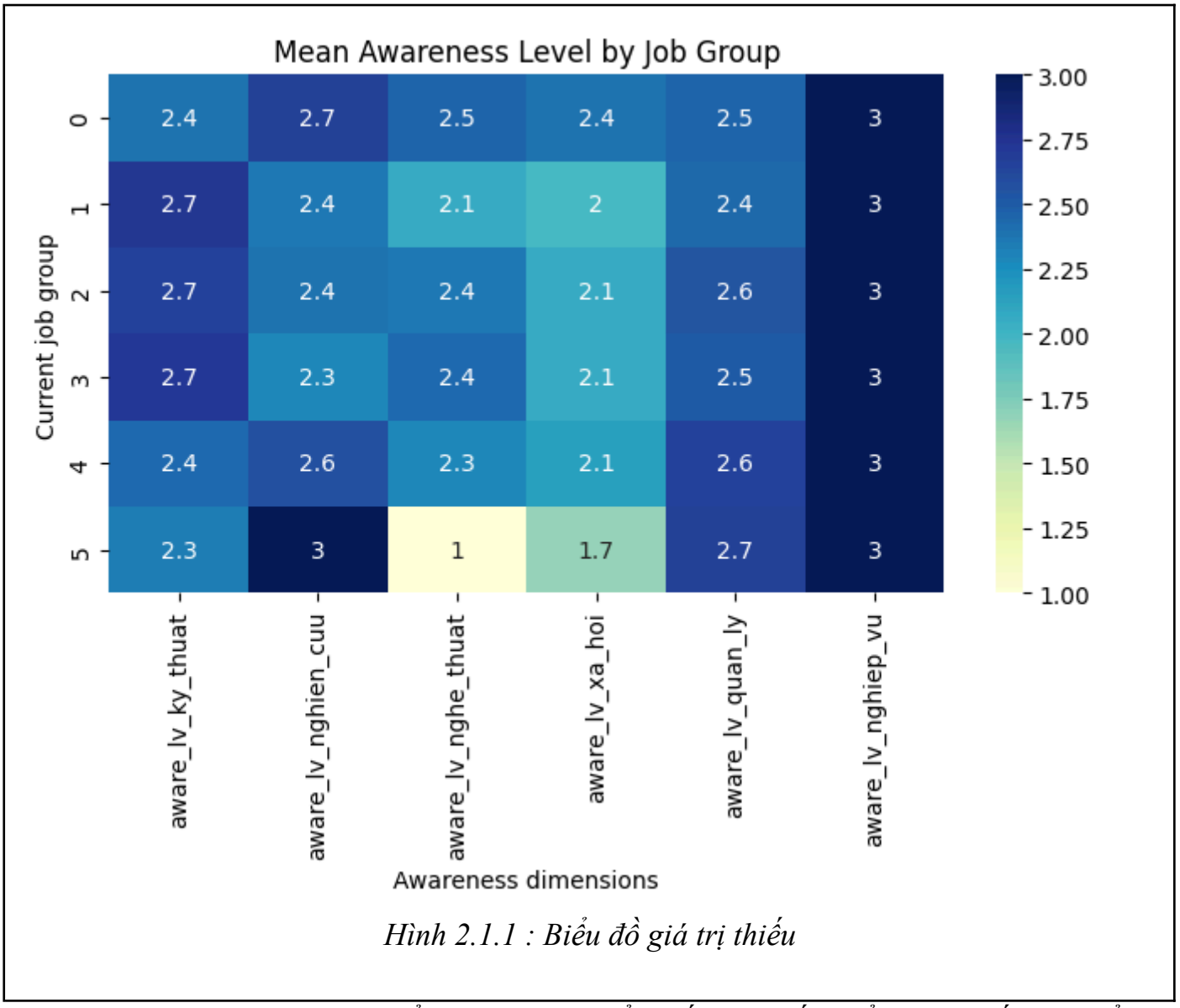
2.2.4. Nhận thức về "yêu cầu của các nhóm nghề" khi học THCS



Hình 2.1.1 : Biểu đồ giá trị thiếu

Nhận xét chính:

- Nhóm Nghệ thuật được nhận diện khá rõ bằng các kỹ năng đặc thù: các dòng "Âm nhạc" và "Biểu diễn nghệ thuật" có tỷ lệ rất cao ở cột Nghệ thuật (trên biểu đồ thể hiện mức ~72%).
- Nhóm Xã hội nổi bật ở các kỹ năng liên quan giao tiếp – con người như "Nói", "Thuyết trình", "Làm việc với người khác", "Gây ảnh hưởng", "Thuyết phục". Cột Xã hội có xu hướng cao nhất ở các kỹ năng mềm, phản ánh nhận thức tương đối đúng về bản chất nhóm nghề này.
- Nhóm Nghiên cứu và Nhóm Kỹ thuật có xu hướng "chồng lấn" nhận thức ở các kỹ năng nền tảng như "Nghiên cứu", "Phân tích", "Viết", "Định hướng". Đây là điểm quan trọng: với các nghề thiên về tư duy và quy trình, người trả lời thường hiểu theo hướng "cần nhiều kỹ năng học thuật" nhưng khó tách rạch ròi giữa nhóm nghề.
- Nhóm Nghiệp vụ (hành chính – hồ sơ – dữ liệu – tỉ mỉ/chi tiết) có các dòng liên quan "hồ sơ & dữ liệu", "tỉ mỉ/chi tiết" tương đối cao, nhưng không "đột biến" như Nghệ thuật/Xã hội. Điều này thường xảy ra vì các kỹ năng nghiệp vụ có tính "đời thường" và dễ bị xem như kỹ năng chung.



Hình 2.1.1 : Biểu đồ giá trị thiếu

Nhìn chung, nhóm Nghiệp vụ thể hiện mức độ hiểu biết cao nhất và ổn định nhất, với điểm trung bình đạt 3.0 ở tất cả các nhóm nghề hiện tại, cho thấy nhận thức rõ ràng và nhất quán về yêu cầu công việc.

Nhóm Kỹ thuật và Quản lý duy trì mức độ hiểu biết ở mức Khá, với điểm trung bình dao động khoảng 2.4–2.7, phản ánh nhận thức tương đối tốt nhưng chưa đồng đều tuyệt đối giữa các nhóm nghề.

Nhóm Xã hội có mức hiểu biết trung bình ở mức Khá (≈2.5–2.7), tuy nhiên nhóm nghề hiện tại 5 ghi nhận mức thấp hơn so với các nhóm còn lại, cho thấy sự hạn chế về nhận thức trong một bộ phận đối tượng.

Đối với nhóm Nghệ thuật, mặc dù điểm trung bình chung nằm quanh mức 2.3–2.5, sự chênh lệch là đáng kể. Đặc biệt, nhóm nghề 5 chỉ đạt mức 1.0, là giá trị thấp nhất toàn bộ bảng, phản ánh mức độ hiểu biết rất hạn chế về yêu cầu nghề nghệ thuật.

Tổng hợp lại, nhóm Nghiệp vụ nổi bật về mức độ hiểu biết, trong khi nhóm Nghệ thuật và Xã hội xuất hiện sự phân hóa mạnh, đặc biệt ở các nhóm nghề có quy mô mẫu nhỏ. Kết quả này cho thấy nhận thức nghề nghiệp không đồng đều và chịu ảnh hưởng bởi quỹ đạo nghề nghiệp hiện tại của người khảo sát.

3. CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

3.1. Giới thiệu chung

Các mô hình sinh (generative models) đã trở thành một trong những trụ cột quan trọng nhất của học sâu hiện đại, mở ra khả năng cho máy móc không chỉ phân tích mà còn có thể sáng tạo ra dữ liệu mới. Từ việc tạo ra những hình ảnh chân thực đến tổng hợp giọng nói tự nhiên, các mô hình này đang định hình lại ranh giới của trí tuệ nhân tạo. Báo cáo này sẽ đi sâu khám phá quá trình tiến hóa của một họ mô hình sinh đặc biệt mạnh mẽ: kiến trúc autoencoder. Chúng ta sẽ bắt đầu từ nền tảng với Autoencoder (AE) cơ bản, sau đó tiến tới mô hình xác suất Variational Autoencoder (VAE), và cuối cùng là Conditional Variational Autoencoder (CVAE) — một kiến trúc tinh vi cho phép tạo sinh dữ liệu một cách có kiểm soát và theo điều kiện.
Để có thể nắm bắt đầy đủ sức mạnh và sự phức tạp của các mô hình tiên tiến như VAE và CVAE, việc hiểu rõ các nguyên tắc nền tảng của Autoencoder truyền thống là bước đi đầu tiên và thiết yếu.

3.2. Autoencoder (AE)

Về mặt chiến lược, Autoencoder (AE) không chỉ là một công cụ nén dữ liệu đơn thuần mà còn là một kiến trúc nền tảng cho việc học biểu diễn (representation learning). Mục tiêu của nó là học một biểu diễn nén, hay còn gọi là mã hóa, của dữ liệu đầu vào trong một không gian có số chiều thấp hơn, sau đó sử dụng biểu diễn này để tái tạo lại dữ liệu gốc một cách chính xác nhất có thể.
Cấu trúc kỹ thuật của một AE bao gồm hai thành phần chính:

- Bộ mã hóa (Encoder): Ánh xạ dữ liệu đầu vào x thành một vector biểu diễn trong không gian ẩn (latent space), thường có số chiều thấp hơn. Vector này được gọi là vector mã h.
- Bộ giải mã (Decoder): Nhận vector mã h làm đầu vào và cố gắng tái tạo lại dữ liệu ban đầu, tạo ra $\hat{x}$ sao cho càng giống x càng tốt.

Mặc dù hiệu quả trong việc giảm chiều và khử nhiễu, AE bộc lộ những hạn chế cố hữu khi được sử dụng như một mô hình sinh. Vấn đề cốt lõi nằm ở không gian tiềm ẩn (latent space) mà nó học được. Không gian này thường rời rạc và không liên tục; các điểm mã hóa cho các lớp dữ liệu khác nhau có thể nằm ở những vùng tách biệt, không có sự chuyển tiếp mượt mà giữa chúng. Điều này khiến cho việc lấy mẫu ngẫu nhiên một điểm từ không gian tiềm ẩn và đưa qua bộ giải mã để tạo ra một mẫu dữ liệu mới, có ý nghĩa trở nên cực kỳ khó khăn và không hiệu quả. Về cơ bản, AE chỉ học cách sao chép, không phải cách sáng tạo.

Để vượt qua rào cản này, cần có một phương pháp có thể tạo ra một không gian tiềm ẩn trơn tru và có cấu trúc, nơi việc nội suy giữa các điểm dữ liệu trở nên khả thi, mở đường cho sự ra đời của VAE.

3.3. Variational Autoencoder (VAE)

Variational Autoencoder (VAE) là một mô hình sinh dữ liệu xác suất, được đề xuất nhằm khắc phục những hạn chế của Autoencoder (AE) truyền thống trong việc tạo ra dữ liệu mới. Khác với AE, nơi mỗi mẫu dữ liệu được mã hóa thành một điểm xác định trong không gian tiềm ẩn, VAE mô hình hóa không gian tiềm ẩn dưới dạng phân phối xác suất liên tục, cho phép sinh dữ liệu đa dạng và có cấu trúc.

3.3.1. Kiến trúc và cơ chế hoạt động

Trong VAE, bộ mã hóa (Encoder) không tạo ra một vector tiềm ẩn duy nhất mà sinh ra hai vector:

- Vector trung bình μ
- Vector log–phương sai $log(\sigma^2)$

Hai vector này cùng nhau xác định phân phối hậu nghiệm xấp xỉ của biến tiềm ẩn:

$$q\phi(z|x) = N(z; \mu\phi(x), diag(\sigma\phi^{2}))$$

Từ phân phối này, một vector tiềm ẩn z được lấy mẫu và đưa vào bộ giải mã (Decoder) để tái tạo dữ liệu đầu vào. Mô hình sinh của VAE được định nghĩa như sau:

$$p\theta(x) = \int p\theta(x|z)p(z)dz$$

Trong đó, phân phối tiên nghiệm p(z)p(z)p(z) thường được chọn là phân phối Gaussian chuẩn:

$$p(z) = N(0, I)$$

Cách tiếp cận này buộc không gian tiềm ẩn phải liên tục và có cấu trúc, cho phép nội suy mượt mà giữa các điểm dữ liệu và sinh ra các biến thể mới mang ý nghĩa thống kê.

3.3.2. Kỹ thuật tái tham số hóa (Reparameterization Trick)

Một thách thức quan trọng của VAE là thao tác lấy mẫu từ phân phối $q(z|x)$, vốn là một quá trình ngẫu nhiên và không khả vi, gây cản trở lan truyền ngược gradient. Để giải quyết vấn đề này, kỹ thuật Reparameterization Trick được sử dụng.

Thay vì lấy mẫu trực tiếp, biến tiềm ẩn được biểu diễn như sau:

$$z = \mu\phi(x) + \sigma\phi(x) \odot \varepsilon, \varepsilon \sim N(0, I)$$

Bằng cách tách thành phần ngẫu nhiên ε khỏi các tham số học được (μ và σ), gradient có thể lan truyền ngược qua encoder, cho phép huấn luyện mô hình theo cách end-to-end.

3.3.3. Hàm mất mát của VAE

Hàm mất mát của VAE được xây dựng dựa trên cận dưới ELBO (Evidence Lower Bound), bao gồm hai thành phần chính:

$$L_{\text{VAE}}(x) = -E_{q\phi(z|x)}\left[log\ p\theta\ (x|z)\right] + KL(q\phi(z|x) \parallel p(z))$$

Trong đó:

- Reconstruction Loss đo lường mức độ sai khác giữa dữ liệu gốc x và dữ liệu tái tạo $\widehat{x}$.
- KL Divergence đóng vai trò điều chuẩn, buộc phân phối tiềm ẩn học được gần với phân phối chuẩn, giúp không gian tiềm ẩn trơn tru và tránh hiện tượng ghi nhớ dữ liệu.

3.4. Conditional Variational Autoencoder (CVAE)

Conditional Variational Autoencoder (CVAE) là phần mở rộng của VAE, được đề xuất nhằm giải quyết bài toán sinh dữ liệu có điều kiện. Bằng cách đưa thêm biến điều kiện y vào mô hình, CVAE cho phép định hướng quá trình sinh dữ liệu theo các thuộc tính hoặc nhãn mong muốn.

3.4.1. Mô hình xác suất có điều kiện

Khác với VAE mô hình hóa $p(x)$, CVAE mô hình hóa phân phối có điều kiện $p(x|y)$:

$$p\theta(x|y) = \int p\theta(x|z,y)p(z)dz$$

Trong đó, phân phối hậu nghiệm xấp xỉ có điều kiện được định nghĩa là:

$$q\phi(z|x,y)$$

và thường được giả định có dạng Gaussian:

$$q\phi(z|x,y) = N(z; \mu\phi(x,\ y), diag(\sigma\phi^{\,2}\ (x,\ y)))\ )$$

Biến điều kiện y được đưa vào cả encoder và decoder, cho phép mô hình học được mối quan hệ giữa dữ liệu và điều kiện tương ứng.

3.4.2. Hàm mất mát của CVAE

Mục tiêu huấn luyện của CVAE là tối đa hóa log-likelihood có điều kiện $log\ p\theta(x|y)$. Cận dưới ELBO có điều kiện được viết như sau:

$$log\ p\theta(x|y) \geq E_{q\phi(z|x,y)}\ [log\ p\theta(x|z,y)]\ -\ KL(q\phi(z|x,y) \parallel p(z))$$

Tương ứng, hàm mất mát cần tối thiểu hóa là:

$$L_{cVAE}\ (x,\ y) =\ \ -\ E_{q\phi(z|x,y)}\ [log\ p\theta(x|z,y)]\ -\ \beta\ \cdot KL(q\phi(z|x,y) \parallel p(z))$$

Hệ số β cho phép điều chỉnh mức độ ảnh hưởng của thành phần KL Divergence, giúp cân bằng giữa khả năng tái tạo và tính trơn tru của không gian tiềm ẩn.

4. CHƯƠNG 4: PHƯƠNG PHÁP THỰC HIỆN

4.1. Chia tập huấn luyện và kiểm tra

Trong giai đoạn huấn luyện CVAE, nghiên cứu sử dụng chiến lược hold-out split:

- 80% dữ liệu cho huấn luyện
- 20% dữ liệu cho kiểm tra

Việc lựa chọn hold-out (thay vì k-fold) được đưa ra do:

- Một số lớp nghề có số lượng mẫu rất nhỏ
- K-fold có thể dẫn đến việc một fold chỉ còn 1–2 mẫu cho lớp hiếm, khiến việc huấn luyện mô hình sinh không ổn định

4.2. Kiến trúc mô hình

4.2.1. Lý do lựa chọn CVAE

Mô hình được xây dựng là một CVAE (Conditional Variational Autoencoder) được thiết kế đặc biệt để xử lý dữ liệu dạng bảng hỗn hợp (Mixed-type Tabular Data), bao gồm các biến nhị phân (binary), thang đo Likert (likert), và biến phân loại (categorical).

Điểm đặc biệt của kiến trúc này là cơ chế Conditional (Có điều kiện): Biến mục tiêu y (nghề nghiệp hiện tại) được đưa vào cả Encoder và Decoder, giúp mô hình học cách sinh ra dữ liệu *đặc thù cho từng nhóm nghề*.

4.2.2. Chiến lược sinh dữ liệu cho các lớp hiếm

Thay vì huấn luyện một CVAE chung cho toàn bộ dữ liệu, nghiên cứu áp dụng chiến lược:

- Sinh dữ liệu chỉ cho các lớp hiếm (class 4 và class 5)

- Mục tiêu cân bằng "gần đều": đưa số mẫu của các lớp hiếm lên mức tối thiểu của các lớp phổ biến (0–3)
- Áp dụng giới hạn an toàn (cap factor = 40). Tránh hiện tượng mô hình sinh dữ liệu quá nhiều nhưng không có giá trị.

Cách tiếp cận này giúp:

- Giảm nguy cơ làm méo phân phối chung của dữ liệu
- Giảm nguy cơ overfitting đối với các lớp có số mẫu thật cực nhỏ

4.2.3. Kiến trúc mô hình CVAE rời rạc (Discrete CVAE)

Mô hình CVAE được hiện thực dưới dạng CVAE cho dữ liệu rời rạc, với các thành phần chính:

Encoder

- Nhận đầu vào gồm:

  Biến nhị phân

  Biến Likert (one-hot 5 mức)

  Biến phân loại (one-hot theo số danh mục)

  Vector điều kiện (nhãn nghề)

- Xuất ra hai vector:

  μ (trung bình)

  $log(\sigma^2)$ (log-phương sai)

Latent space

- Kích thước latent: 32
- Phân phối tiên nghiệm: Gaussian chuẩn $N(0, I)$

Decoder đa đầu ra (Multi-head Decoder)

- Binary head: Bernoulli (sigmoid + BCE loss)
- Likert head: Softmax 5 lớp + Cross-Entropy
- Categorical head: Softmax K lớp + Cross-Entropy

Thiết kế này cho phép mỗi loại biến được sinh theo đúng bản chất thống kê của nó.

4.2.4. Hàm mất mát và chiến lược huấn luyện

Hàm mất mát của CVAE bao gồm:

- Reconstruction loss cho từng loại biến
- KL Divergence để điều chuẩn latent space

Các điểm tinh chỉnh quan trọng:

- KL warm-up: hệ số β tăng dần theo epoch, giúp tránh hiện tượng posterior collapse
- Early stopping dựa trên validation loss
- Huấn luyện riêng cho từng lớp hiếm để tránh gradient bị chi phối bởi lớp lớn

4.2.5. Kiểm soát đầu ra của mô hình

Sau khi sinh dữ liệu, nghiên cứu không sử dụng trực tiếp mà tiến hành kiểm tra thông qua:

- Dữ liệu sinh ra được kiểm soát chặt chẽ về miền giá trị
- So sánh phân phối giữa dữ liệu thật và dữ liệu sinh (real_vs_synthetic_for_bar.csv)
- Phân tích trực quan các biến rời rạc

Cách tiếp cận này đảm bảo dữ liệu sinh ra có giá trị sử dụng, thay vì chỉ "giống về mặt hình thức".

5. CHƯƠNG 5: THỰC NGHIỆM VÀ KẾT QUẢ

5.1. Kết quả sinh dữ liệu

5.1.1. Mục tiêu sinh dữ liệu

Mục tiêu của việc sinh dữ liệu trong nghiên cứu này không phải là thay thế dữ liệu thật, mà là bổ sung có kiểm soát dữ liệu cho các nhóm nghề hiếm, nhằm giảm mức độ mất cân bằng lớp và cải thiện khả năng học của mô hình phân loại nghề nghiệp.

Việc sinh dữ liệu được thực hiện bằng mô hình Conditional Variational Autoencoder (CVAE), với điều kiện là nhãn nghề nghiệp, và chỉ áp dụng cho các lớp hiếm (class 4 và class 5).

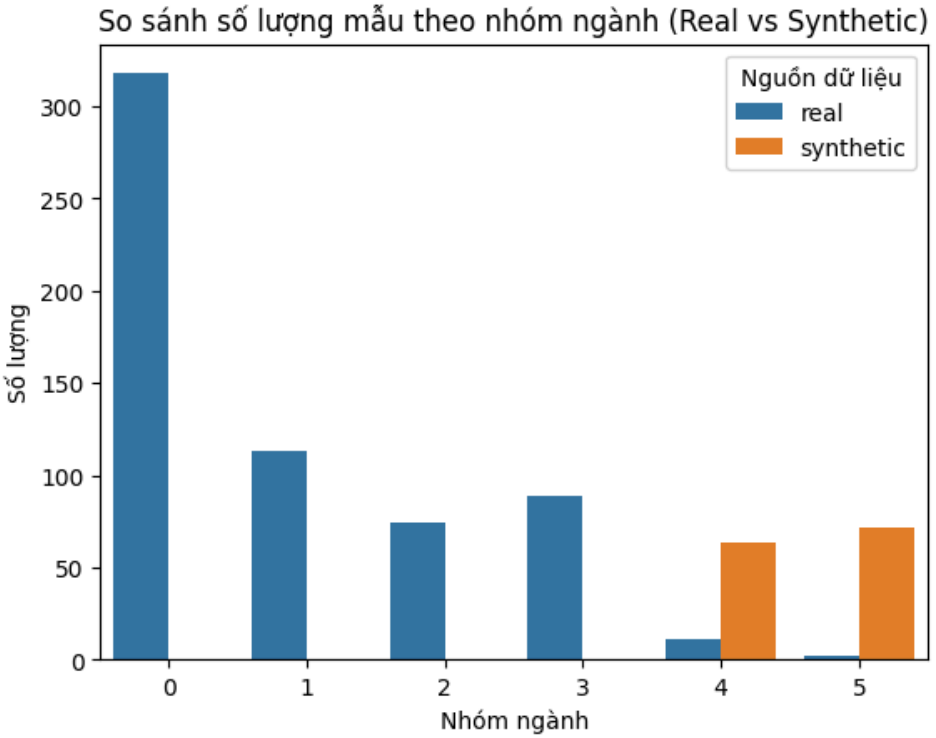
5.1.2. Số lượng dữ liệu sinh

Theo kế hoạch sinh dữ liệu được lưu trong tệp " synth_plan.csv ", số lượng dữ liệu đã sinh cho từng nhóm nghề hiếm. Cụ thể:

- Nhóm nghề 4 có số mẫu thật rất thấp so với các nhóm 0–3. Đã tăng từ 14 đến *77*, tăng thêm 63 mẫu, gấp 5.5 lần so với ban đầu.
- Nhóm nghề 5 có số mẫu thật cực kỳ ít, nghiêm trọng hơn nhóm nghề 4. Đã tăng từ 3 đến *75*, tăng thêm 72 mẫu, gấp 25 lần so với ban đầu.

- Tổng số mẫu tăng từ 759 → 894, tăng 135 mẫu (≈ *17.8%*).

Ảnh được trực quan từ file dữ liệu gốc.
(Phân bố mất cân bằng nghiêm trọng, nhóm 4,5 chiếm rất ít)



Ảnh được trực quan sau khi áp dụng mô hình sinh CVAE.
(Nhóm ngành 4, 5 đã tăng, số lượng tương đương với các nhóm ngành 1-3)

6. CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết quả đạt được

Nghiên cứu này tập trung đánh giá hiệu quả của việc sinh dữ liệu khảo sát bằng Conditional Variational Autoencoder (CVAE) trong bối cảnh dữ liệu mất cân bằng nghiêm trọng cho bài toán dự đoán nghề nghiệp của người đi làm. Hiệu quả của phương pháp đề xuất được đánh giá thông qua so sánh trực tiếp với mô hình baseline không sử dụng dữ liệu sinh.

6.1.1. So sánh định lượng giữa mô hình Baseline và CVAE

Hai thiết lập được so sánh:

- RF_BASELINE: Mô hình Random Forest huấn luyện trên dữ liệu gốc
- RF_CVAE: Mô hình Random Forest huấn luyện trên dữ liệu gốc kết hợp dữ liệu sinh bằng CVAE

Kết quả trung bình thu được như sau:

Mô hình	Balanced Accuracy	Macro F1
RF_BASELINE	0.3531	0.3404
RF_CVAE	**0.4043**	**0.3918**

6.1.2. Mức độ cải thiện

So với mô hình baseline, phương pháp sử dụng dữ liệu sinh bằng CVAE mang lại cải thiện rõ rệt và nhất quán trên cả hai thước đo đánh giá chính:

- Balanced Accuracy:
    - Tăng từ 0.3531 lên 0.4043
    - Mức tăng tuyệt đối: +0.0512

    - Mức tăng tương đối: ≈ **+14.5%**

- Macro F1-score:
    - Tăng từ 0.3404 lên 0.3918
    - Mức tăng tuyệt đối: +0.0514
    - Mức tăng tương đối: ≈ **+15.1%**

Kết quả này cho thấy việc bổ sung dữ liệu sinh bằng CVAE đã nâng cao đáng kể khả năng dự đoán cho các nhóm nghề, vốn là trọng tâm của bài toán mất cân bằng dữ liệu.

6.2. Hạn chế

Mặc dù nghiên cứu đã đạt được những kết quả tích cực và cho thấy tiềm năng rõ rệt của mô hình CVAE trong việc sinh dữ liệu khảo sát mất cân bằng, vẫn tồn tại một số hạn chế cần được nhìn nhận một cách nghiêm túc. Thứ nhất, chất lượng thống kê của dữ liệu sinh chưa được đánh giá một cách toàn diện. Việc đánh giá hiện tại chủ yếu dựa trên hiệu năng của mô hình phân loại (utility-based evaluation), trong khi các kiểm định thống kê chuyên sâu hơn như so sánh phân phối, kiểm tra mức độ bảo toàn tương quan giữa các biến hay đánh giá cấu trúc phụ thuộc đa biến giữa dữ liệu gốc và dữ liệu sinh vẫn chưa được triển khai đầy đủ. Thứ hai, không gian tiềm ẩn và kiến trúc của mô hình CVAE chưa được tối ưu hóa một cách có hệ thống. Các tham số quan trọng như kích thước latent space, trọng số của thành phần KL divergence trong hàm mất mát, hay cơ chế sampling (ví dụ temperature) chủ yếu được lựa chọn dựa trên kinh nghiệm và thử nghiệm thủ công, chưa áp dụng các chiến lược tìm kiếm tham số bài bản như grid search hoặc Bayesian optimization. Cuối cùng, nghiên cứu chưa đánh giá đầy đủ rủi ro overfitting vào dữ liệu sinh. Việc bổ sung một lượng lớn dữ liệu tổng hợp vào tập huấn luyện có thể dẫn đến nguy cơ mô hình học quá nhiều từ dữ liệu sinh nếu không có các cơ chế kiểm soát phù hợp, từ đó ảnh hưởng đến khả năng tổng quát hóa trên dữ liệu thực. Những hạn chế này đồng thời cũng chính là cơ sở quan trọng để định hướng cho các nghiên cứu tiếp theo.

6.3. Hướng phát triển

Dựa trên các kết quả đạt được và những hạn chế đã được chỉ ra, nghiên cứu này mở ra nhiều hướng phát triển tiềm năng trong tương lai. Trước hết, mô hình CVAE có thể được cải thiện hơn nữa thông qua việc tinh chỉnh hàm mất mát và chiến lược huấn luyện, chẳng hạn như điều chỉnh cơ chế chuẩn hóa loss cho dữ liệu hỗn hợp, áp dụng KL warm-up hoặc các biến thể như β-VAE nhằm nâng cao chất lượng không gian tiềm ẩn và tính đa dạng của dữ liệu sinh. Bên cạnh đó, việc đánh giá chất

lượng dữ liệu sinh cần được mở rộng ở mức độ thống kê sâu hơn, bao gồm so sánh phân phối, kiểm định thống kê và phân tích cấu trúc tương quan giữa các biến, nhằm đảm bảo dữ liệu tổng hợp không chỉ hữu ích cho mô hình học máy mà còn phản ánh đúng đặc điểm của dữ liệu gốc. Ngoài CVAE, nghiên cứu trong tương lai cũng có thể mở rộng sang việc so sánh với các mô hình sinh dữ liệu khác như SMOTE-NC, GAN cho dữ liệu rời rạc hoặc các phương pháp lai (hybrid) để đánh giá toàn diện ưu, nhược điểm của từng hướng tiếp cận. Cuối cùng, các kết quả của nghiên cứu này có thể được ứng dụng vào nhiều bài toán thực tế, đặc biệt là trong lĩnh vực hướng nghiệp, phân tích thị trường lao động và chia sẻ dữ liệu khảo sát ở dạng synthetic nhằm bảo vệ quyền riêng tư, từ đó gia tăng giá trị ứng dụng của mô hình trong bối cảnh nghiên cứu và thực tiễn.

Tài liệu

https://arxiv.org/abs/1312.6114 ,Nội dung: Giới thiệu kiến trúc VAE, Reparameterization Trick và hàm mất mát ELBO.

https://papers.nips.cc/paper/2015/hash/8d55a249e643c0506e99bc43dfcf9dbb-Abstract.html ,Nội dung: Đề xuất mô hình Conditional VAE (CVAE) để sinh dữ liệu có điều kiện (giống cách bạn đưa nhãn nghề nghiệp vào để sinh dữ liệu).

https://arxiv.org/abs/1606.05908 ,Nội dung: Tài liệu giải thích chi tiết, dễ hiểu về toán học đằng sau VAE.

https://arxiv.org/abs/1611.01144 ,Nội dung: Kỹ thuật giúp lan truyền ngược (backpropagation) qua các biến rời rạc/phân loại.

https://lup.lub.lu.se/student-papers/search/publication/9024848 ,Nội dung: Một bài viết kỹ thuật (Technical Report) thảo luận về việc thiết kế các "Head" (đầu ra) khác nhau cho VAE (Gaussian cho số thực, Bernoulli cho nhị phân).

https://arxiv.org/abs/2202.09341 , Nội dung: Tổng quan các phương pháp sinh dữ liệu cho dữ liệu hỗn hợp.

https://arxiv.org/abs/2011.06079 , Nội dung: Bài báo minh họa việc dùng CVAE để tăng cường dữ liệu có đặc tính phân loại.