Thu thập và phân tích tin tuyển dụng Việt Nam

Hoàng Bảo Long, Nguyễn Duy Phương, Lê Đức Khánh Toàn, Lương Tuấn Vỹ, Phạm Nguyễn Phúc Toàn
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM
TP.HCM, Việt Nam
{22520807, 22521165, 22521488, 22521711 }@gm.uit.edu.vn

***Tóm tắt nội dung*—Báo cáo này trình bày quy trình xây dựng hệ thống thu thập, phân tích dữ liệu tuyển dụng và mô hình dự đoán mức lương tại thị trường Việt Nam. Nhóm nghiên cứu tập trung giải quyết bài toán thiếu hụt thông tin và bất cân xứng dữ liệu trong tuyển dụng bằng cách áp dụng các kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP) để trích xuất thông tin kỹ năng và cấp bậc (level). Kết quả nghiên cứu bao gồm một bộ dữ liệu được làm sạch, các phân tích trực quan về thị trường lao động và một mô hình học máy có khả năng dự đoán mức lương dựa trên yêu cầu công việc, góp phần hỗ trợ ra quyết định cho cả ứng viên và nhà tuyển dụng.**

***Index Terms*—Data Mining, Salary Prediction, Recruitment Analysis, Machine Learning, NLP.**

I. TỔNG QUAN ĐỀ TÀI

A. Bối cảnh và Đặt vấn đề

Thị trường lao động Việt Nam đang trải qua giai đoạn chuyển đổi số mạnh mẽ. Theo thống kê của Bộ Lao động — Thương binh và Xã hội, năm 2024 có hơn 51 triệu lao động trong độ tuổi, trong đó nhu cầu tuyển dụng trực tuyến tăng trưởng trung bình 15–20% mỗi năm. Các nền tảng tuyển dụng như TopCV, CareerViet, và VietnamWorks đăng tải hàng trăm nghìn tin tuyển dụng mới mỗi tháng, tạo nên nguồn dữ liệu khổng lồ về thị trường lao động.

Tuy nhiên, sự bùng nổ về số lượng tin tuyển dụng lại dẫn đến nghịch lý về *bất cân xứng thông tin* (information asymmetry). Người tìm việc, đặc biệt là sinh viên mới ra trường, thường gặp khó khăn trong việc định giá năng lực bản thân do thiếu thông tin tổng hợp về mức lương theo ngành nghề và vị trí. Mặc dù khoảng 97% tin tuyển dụng trong tập dữ liệu nghiên cứu có công khai mức lương, người tìm việc vẫn thiếu công cụ để so sánh và đánh giá mức lương phù hợp. Ngược lại, nhà tuyển dụng cũng đối mặt với thách thức trong việc xây dựng khung lương cạnh tranh để thu hút nhân tài mà vẫn đảm bảo tối ưu chi phí nhân sự.

Nguồn dữ liệu tuyển dụng hiện tại tuy dồi dào nhưng *phân tán* và *thiếu cấu trúc*. Các báo cáo lương từ các công ty tư vấn nhân sự thường dựa trên khảo sát với mẫu nhỏ, có độ trễ theo quý hoặc năm, và chỉ cung cấp số liệu trung bình gộp chung thay vì khả năng dự đoán cá nhân hóa. Do đó, việc xây dựng một hệ thống tự động thu thập và phân tích dữ liệu tuyển dụng để đưa ra các dự báo lương chính xác là nhu cầu cấp thiết.

B. Mục tiêu nghiên cứu

Nghiên cứu này hướng tới bốn mục tiêu chính:

1) **Thu thập dữ liệu quy mô lớn:** Xây dựng hệ thống crawler tự động để thu thập tin tuyển dụng từ nhiều nền tảng tuyển dụng trực tuyến hàng đầu tại Việt Nam.
2) **Tiền xử lý và chuẩn hóa:** Phát triển quy trình làm sạch dữ liệu để xử lý các vấn đề đặc thù của dữ liệu tuyển dụng tiếng Việt như đa dạng cách viết địa danh, pha trộn ngôn ngữ Anh–Việt, và thông tin lương không đồng nhất.
3) **Phân tích khám phá:** Thực hiện phân tích thống kê mô tả và trực quan hóa để khám phá các đặc điểm của thị trường lao động Việt Nam.
4) **Xây dựng mô hình dự đoán:** Áp dụng các thuật toán Học máy để dự đoán mức lương dựa trên các đặc trưng của vị trí tuyển dụng.

C. Đóng góp của nghiên cứu

Nghiên cứu này đóng góp vào lĩnh vực phân tích dữ liệu tuyển dụng tại Việt Nam thông qua:

- **Bộ dữ liệu:** Xây dựng bộ dữ liệu gồm hơn 85.000 tin tuyển dụng được thu thập và chuẩn hóa từ 4 nguồn, có thể phục vụ cho các nghiên cứu tiếp theo.
- **Quy trình tiền xử lý:** Đề xuất quy trình 8 bước xử lý dữ liệu tuyển dụng tiếng Việt với tỷ lệ giữ lại dữ liệu đạt 95,9%.
- **Phân tích thực nghiệm:** Cung cấp các phân tích định lượng về phân bố lương theo cấp bậc, vùng miền, và kỹ năng tại thị trường Việt Nam.
- **Mô hình dự đoán:** So sánh hiệu năng của nhiều thuật toán và xác định các yếu tố quan trọng nhất ảnh hưởng đến mức lương.

D. Phạm vi và Cấu trúc bài báo

Nghiên cứu thu thập dữ liệu từ các nền tảng tuyển dụng đa ngành nghề tại Việt Nam trong giai đoạn 2024–2025. Phạm vi phân tích tập trung vào các vị trí có công khai thông tin lương (khoảng 97% tổng số tin trong tập dữ liệu).

Phần còn lại của bài báo được tổ chức như sau: Phần II trình bày các nghiên cứu liên quan. Phần III mô tả phương pháp thu thập và xử lý dữ liệu. Phần IV trình bày kết quả phân tích và thảo luận. Cuối cùng, Phần V tổng kết và đề xuất hướng phát triển.

II. CÁC NGHIÊN CỨU LIÊN QUAN

A. Khai phá dữ liệu tuyển dụng trực tuyến

Khai phá dữ liệu tuyển dụng (Online Job Advertisement Mining) là lĩnh vực nghiên cứu quan trọng trong bối cảnh

chuyển đổi số của thị trường lao động. Các nghiên cứu tiên phong của Carnevale và cộng sự [1] đã chứng minh giá trị của việc phân tích dữ liệu tuyển dụng trực tuyến để hiểu xu hướng kỹ năng và nhu cầu thị trường lao động theo thời gian thực, thay vì dựa vào các khảo sát truyền thống có độ trễ cao.

Về phương pháp thu thập dữ liệu, Turrell và cộng sự [2] đã phát triển hệ thống web scraping quy mô lớn để thu thập hàng triệu tin tuyển dụng từ các nền tảng việc làm tại Anh, cho phép phân tích xu hướng kỹ năng theo ngành nghề và địa lý. Nghiên cứu này đặt nền móng cho các hệ thống thu thập dữ liệu tuyển dụng tự động, chứng minh tính khả thi của việc xây dựng bộ dữ liệu lớn từ nhiều nguồn phân tán.

B. Xử lý ngôn ngữ tự nhiên trong phân tích tuyển dụng

Xử lý ngôn ngữ tự nhiên (NLP) đóng vai trò then chốt trong việc trích xuất thông tin từ văn bản mô tả công việc (Job Description). Kivimäki và cộng sự [3] đã đề xuất phương pháp sử dụng Word Embeddings để biểu diễn kỹ năng trong không gian vector, cho phép phát hiện các kỹ năng tương đồng về ngữ nghĩa mà phương pháp so khớp từ khóa truyền thống không thể nhận diện.

Nghiên cứu của Zhang và cộng sự [4] đã ứng dụng mô hình BERT (Bidirectional Encoder Representations from Transformers) để trích xuất kỹ năng từ văn bản tuyển dụng, đạt F1-score 0.89 trên bộ dữ liệu benchmark. Phương pháp này vượt trội so với các kỹ thuật NLP truyền thống nhờ khả năng hiểu ngữ cảnh hai chiều của từ.

Đối với tiếng Việt, Nguyen và Nguyen [5] đã phát triển PhoBERT — mô hình ngôn ngữ lớn được huấn luyện riêng cho tiếng Việt, mở ra khả năng áp dụng các kỹ thuật NLP tiên tiến cho dữ liệu tuyển dụng Việt Nam. Tuy nhiên, việc ứng dụng PhoBERT trong lĩnh vực tuyển dụng vẫn còn hạn chế.

C. Các phương pháp dự đoán lương

Bài toán dự đoán lương (Salary Prediction) được tiếp cận theo hai hướng chính: phương pháp kinh tế lượng truyền thống và phương pháp học máy hiện đại.

Phương pháp kinh tế lượng: Dựa trên lý thuyết Vốn nhân lực (Human Capital Theory) của Becker [6], các nghiên cứu sử dụng phương trình Mincer để mô hình hóa mối quan hệ giữa lương với số năm học vấn và kinh nghiệm. Tuy nhiên, Card [7] đã chỉ ra rằng mô hình này có nhiều hạn chế khi áp dụng cho các ngành có tính đặc thù cao như công nghệ thông tin.

Phương pháp học máy: Các nghiên cứu gần đây đã chứng minh hiệu quả vượt trội của thuật toán ensemble. Breiman [8] đề xuất Random Forest với khả năng xử lý tốt dữ liệu có nhiều biến phân loại và chống overfitting. Chen và Guestrin [9] phát triển XGBoost, đạt kết quả state-of-the-art trong nhiều cuộc thi Kaggle về dự đoán lương.

Madan và cộng sự [10] so sánh hiệu năng của nhiều thuật toán trên dữ liệu Stack Overflow Developer Survey, kết luận rằng Random Forest đạt $R^2 = 0.45$, cao hơn đáng kể so với Linear Regression ($R^2 = 0.28$). Nghiên cứu cũng chỉ ra rằng kỹ năng lập trình và vị trí địa lý là hai yếu tố quan trọng nhất ảnh hưởng đến mức lương.

D. Phân tích thị trường lao động bằng Clustering

Phương pháp phân cụm (Clustering) được sử dụng rộng rãi để phân đoạn thị trường lao động. Arthur và Vassilvitskii [11] đề xuất thuật toán K-Means++ với khởi tạo tâm cụm thông minh, cải thiện đáng kể chất lượng phân cụm so với K-Means chuẩn.

Alabdulkareem và cộng sự [12] đã áp dụng network analysis kết hợp clustering để xây dựng "bản đồ kỹ năng" (skill space), cho thấy các nhóm kỹ năng có mối liên hệ chặt chẽ với nhau và với mức lương tương ứng.

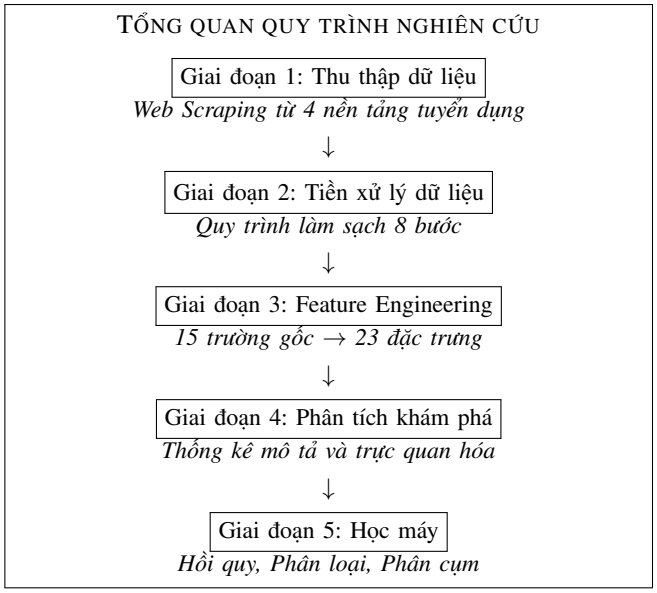
E. Khoảng trống nghiên cứu tại Việt Nam

Tại Việt Nam, nghiên cứu về phân tích dữ liệu tuyển dụng còn rất hạn chế. Các báo cáo lương từ TopCV [15] và VietnamWorks [16] chủ yếu dựa trên khảo sát và thống kê mô tả, thiếu các mô hình dự đoán. Về mặt học thuật, các nghiên cứu ứng dụng NLP cho tiếng Việt như PhoBERT [5] chưa được áp dụng cụ thể vào lĩnh vực tuyển dụng.

Các nghiên cứu hiện có tại Việt Nam gặp ba hạn chế chính: (1) Chưa xử lý tốt vấn đề code-mixing (pha trộn tiếng Anh-Việt) phổ biến trong văn bản tuyển dụng IT; (2) Bỏ qua yếu tố cấp bậc vị trí (position level) trong mô hình dự đoán; và (3) Thiếu bộ dữ liệu quy mô lớn được chuẩn hóa. Nghiên cứu này nhằm lấp đầy các khoảng trống trên bằng cách xây dựng hệ thống thu thập và phân tích dữ liệu tuyển dụng toàn diện cho thị trường Việt Nam.

III. PHƯƠNG PHÁP NGHIÊN CỨU

Phần này trình bày chi tiết phương pháp nghiên cứu được áp dụng, bao gồm: quy trình thu thập dữ liệu, tiền xử lý và làm sạch, kỹ thuật feature engineering, mô hình nghiên cứu cùng các giả thuyết, và cuối cùng là các thuật toán machine learning được sử dụng. Hình 1 minh họa tổng quan quy trình nghiên cứu.



Hình 1. Tổng quan quy trình nghiên cứu

A. Thu thập dữ liệu

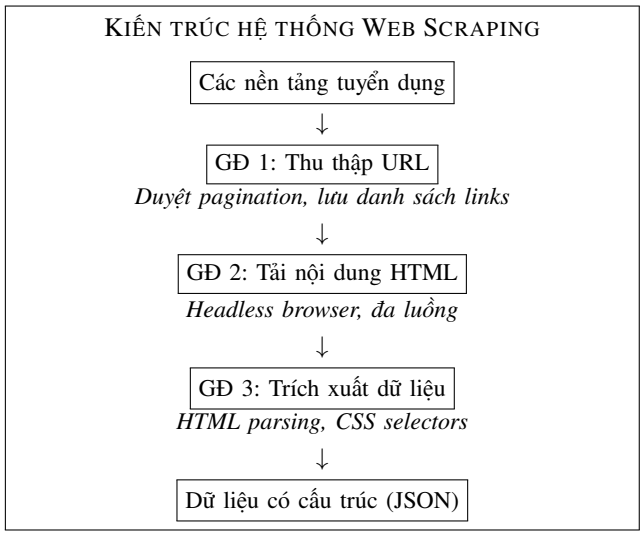
1) Nguồn dữ liệu: Bộ dữ liệu được thu thập từ bốn nền tảng tuyển dụng trực tuyến hàng đầu tại Việt Nam, bao gồm: CareerViet, TopCV, ViecLam24h và JobsGo. Các nền tảng này được lựa chọn dựa trên ba tiêu chí: (1) thị phần cao trong lĩnh vực tuyển dụng trực tuyến Việt Nam, (2) cấu trúc trang web ổn định cho phép thu thập tự động, và (3) đa dạng về ngành nghề và phân bố địa lý.

2) Kiến trúc hệ thống thu thập: Hệ thống thu thập dữ liệu được thiết kế theo kiến trúc ba giai đoạn (Hình 2):

Giai đoạn 1 — Thu thập danh sách URL: Duyệt qua các trang danh sách việc làm (job listing pages) để thu thập toàn bộ đường dẫn đến trang chi tiết. Mỗi trang listing chứa từ 20 đến 50 tin tuyển dụng. Tổng cộng thu thập được hơn 100.000 URLs.

Giai đoạn 2 — Tải nội dung trang: Sử dụng trình duyệt tự động (Selenium WebDriver) ở chế độ headless để xử lý các trang có nội dung động (JavaScript rendering). Áp dụng kỹ thuật đa luồng với 5 workers đồng thời để tăng tốc độ thu thập.

Giai đoạn 3 — Trích xuất dữ liệu có cấu trúc: Phân tích cú pháp HTML và trích xuất thông tin sử dụng thư viện BeautifulSoup kết hợp với CSS Selectors. Mỗi website được cấu hình riêng với bộ selectors phù hợp.



Hình 2. Kiến trúc hệ thống thu thập dữ liệu ba giai đoạn

3) Các biện pháp kỹ thuật: Để đảm bảo quá trình thu thập diễn ra ổn định và tránh bị chặn bởi các website, hệ thống triển khai các biện pháp sau:

- Thiết lập độ trễ ngẫu nhiên từ 1 đến 3 giây giữa các yêu cầu
- Thay đổi luân phiên User-Agent headers từ danh sách hơn 10 trình duyệt thực
- Cơ chế thử lại tự động với thời gian chờ tăng dần (exponential backoff)
- Duy trì phiên làm việc (session) để bảo toàn cookies

4) Cấu trúc dữ liệu: Mỗi tin tuyển dụng được trích xuất thành 11 trường thông tin chính, được mô tả chi tiết trong Bảng I.

Bảng I
CẤU TRÚC DỮ LIỆU CỦA TIN TUYỂN DỤNG

Trường dữ liệu	Kiểu	Mô tả
job_title	Chuỗi	Tên vị trí tuyển dụng
job_type	Chuỗi	Loại hình: Toàn thời gian/Bán thời gian
position_level	Chuỗi	Cấp bậc: Thực tập, Nhân viên, Quản lý...
city	Chuỗi	Địa điểm làm việc
experience	Chuỗi	Yêu cầu số năm kinh nghiệm
skills	Danh sách	Các kỹ năng yêu cầu
job_fields	Danh sách	Lĩnh vực ngành nghề
salary	Chuỗi	Thông tin lương dạng text
salary_min	Số thực	Mức lương tối thiểu (triệu VND)
salary_max	Số thực	Mức lương tối đa (triệu VND)
unit	Chuỗi	Đơn vị tiền tệ (VND/USD)

Bảng II trình bày thống kê dữ liệu thu thập được từ các nguồn.

Bảng II
THỐNG KÊ DỮ LIỆU THU THẬP

Chỉ số	Giá trị
Tổng số bản ghi (raw)	85.470
Số nguồn dữ liệu	4
Số cột dữ liệu	11
Dung lượng	77,13 MB

Ghi chú: Dữ liệu được thu thập từ 4 nền tảng (CareerViet, TopCV, ViecLam24h, JobsGo) và đã được merge thành một bộ dữ liệu duy nhất trước khi phân tích.

B. Tiền xử lý dữ liệu

Quy trình tiền xử lý được thiết kế thành 8 bước tuần tự (Hình 3), chuyển đổi 85.470 bản ghi thô thành 81.971 bản ghi sạch, đạt tỷ lệ giữ lại 95,9%.

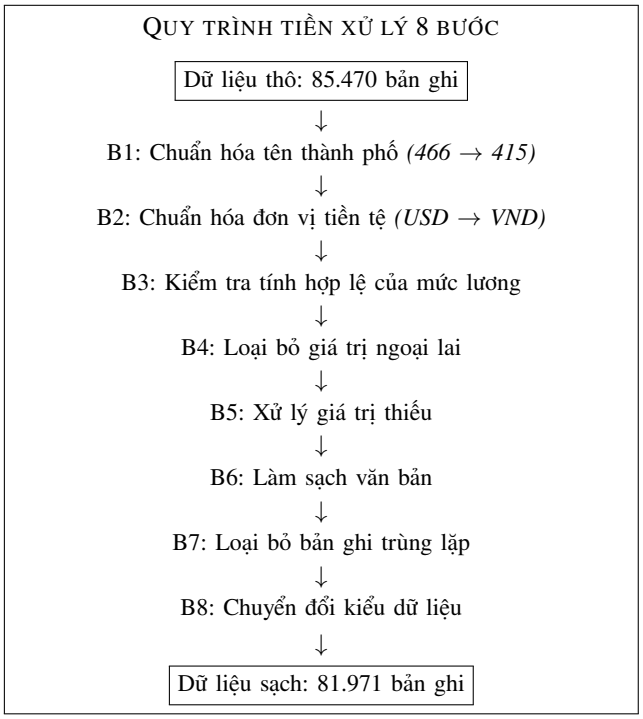
Bước 1 — Chuẩn hóa tên thành phố: Dữ liệu thô chứa 466 biến thể khác nhau của tên thành phố do sự không nhất quán trong cách viết. Ví dụ, các biến thể "HCM", "TPHCM", "TP.HCM", "Hồ Chí Minh", "Sài Gòn" đều được chuẩn hóa về "TP. Hồ Chí Minh". Nghiên cứu xây dựng từ điển ánh xạ để chuẩn hóa về 415 tên thành phố.

Bước 2 — Chuẩn hóa đơn vị tiền tệ: 713 tin tuyển dụng có mức lương ghi bằng đô la Mỹ (USD). Áp dụng tỷ giá quy đổi 25.000 VND/USD để đảm bảo tính nhất quán.

Bước 3 — Kiểm tra tính hợp lệ: Loại bỏ 3.495 bản ghi có lương tối thiểu lớn hơn lương tối đa — đây là lỗi nhập liệu phổ biến từ phía nhà tuyển dụng.

Bước 4 — Loại bỏ ngoại lai: Áp dụng kiến thức chuyên môn để xác định ngưỡng: loại bỏ các bản ghi có mức lương vượt quá 500 triệu VND/tháng hoặc dưới 1 triệu VND/tháng, được xem là lỗi dữ liệu hoặc giá trị ngoại lai cực đoan.

Bước 5 — Xử lý giá trị thiếu: Đối với trường lương, giữ nguyên giá trị NULL để tránh tạo sai lệch. Các giá trị "Không yêu cầu kinh nghiệm" được chuyển thành 0 năm.



Hình 3. Quy trình tiền xử lý dữ liệu 8 bước

Bước 6 — Làm sạch văn bản: Loại bỏ thẻ HTML, chuẩn hóa khoảng trắng, và xử lý các ký tự Unicode đặc thù của tiếng Việt.

Bước 7 — Loại bỏ trùng lặp: Xác định và loại bỏ các bản ghi trùng lặp dựa trên khóa tổng hợp gồm: tiêu đề công việc, tên công ty, thành phố và ngày đăng. Tỷ lệ trùng lặp khoảng 2%.

Bước 8 — Chuyển đổi kiểu dữ liệu: Chuyển đổi các trường chuỗi sang kiểu dữ liệu phù hợp: ngày tháng, số thực cho mức lương, và danh sách cho kỹ năng.

1) Đánh giá chất lượng dữ liệu: Sau khi hoàn thành quy trình tiền xử lý, chất lượng dữ liệu được đánh giá như sau:

Bảng III
ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU SAU TIỀN XỬ LÝ

Chỉ số	Trước	Sau
Tổng số bản ghi	85.470	81.971
Số thành phố (unique)	466	415
Số bản ghi bị loại	—	3.499 (4,1%)
Tỷ lệ giữ lại	—	95,9%

C. Kỹ thuật Feature Engineering

Từ 15 trường dữ liệu gốc, nghiên cứu tạo ra 8 đặc trưng phái sinh (derived features) phục vụ cho việc xây dựng mô hình, được trình bày trong Bảng IV.

Chuẩn hóa cấp bậc vị trí: Hơn 50 tên gọi cấp bậc khác nhau trong dữ liệu gốc được chuẩn hóa thành 6 nhóm chính:

- Thực tập sinh (Intern): Bao gồm các vị trí thực tập, học việc
- Nhân viên (Staff): Các vị trí nhân viên, chuyên viên cơ bản

Bảng IV
CÁC ĐẶC TRƯNG PHÁI SINH

Đặc trưng	Kiểu	Công thức/Logic
Lương trung vị	Số thực	(Lương min + Lương max) / 2
Biên độ lương	Số thực	Lương max − Lương min
Số năm kinh nghiệm	Số nguyên	Trích xuất từ văn bản
Số lượng kỹ năng	Số nguyên	Đếm số kỹ năng yêu cầu
Vùng miền	Phân loại	Bắc / Trung / Nam / Toàn quốc
Cấp bậc chuẩn hóa	Phân loại	6 nhóm cấp bậc
Làm việc từ xa	Nhị phân	Có / Không
Có công khai lương	Nhị phân	Có / Không

- Chuyên gia (Senior): Nhân viên có kinh nghiệm, chuyên gia
- Trưởng nhóm (Team Lead): Quản lý nhóm, giám sát
- Quản lý (Manager): Quản lý cấp trung, trưởng phòng
- Giám đốc (Director): Quản lý cấp cao, ban điều hành

D. Mô hình nghiên cứu và các giả thuyết

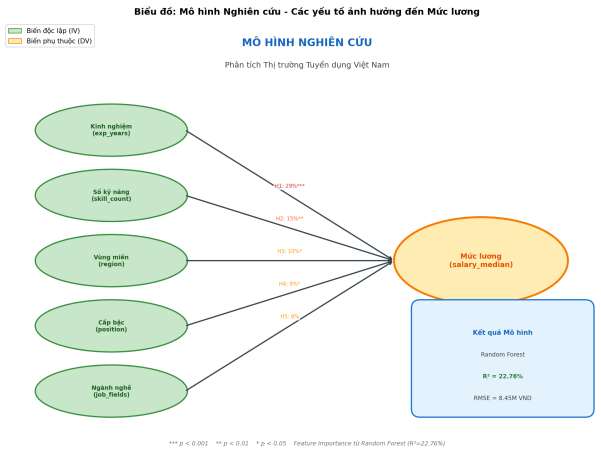
Nghiên cứu xây dựng mô hình khái niệm dựa trên Lý thuyết Vốn nhân lực (Human Capital Theory) [6], trong đó mức lương được xác định bởi các yếu tố về vốn nhân lực (kinh nghiệm, kỹ năng) và các yếu tố ngữ cảnh (vị trí địa lý, cấp bậc, ngành nghề).

Biến phụ thuộc: Mức lương trung vị (triệu VND/tháng)

Biến độc lập và các giả thuyết:

- **H1:** Số năm kinh nghiệm có tác động dương đến mức lương
- **H2:** Số lượng kỹ năng yêu cầu có tác động dương đến mức lương
- **H3:** Vùng miền có ảnh hưởng có ý nghĩa thống kê đến mức lương
- **H4:** Cấp bậc vị trí có ảnh hưởng có ý nghĩa thống kê đến mức lương
- **H5:** Ngành nghề có ảnh hưởng có ý nghĩa thống kê đến mức lương

Mô hình nghiên cứu được minh họa trong Hình 4.



Hình 4. Mô hình nghiên cứu: Các yếu tố ảnh hưởng đến mức lương

E. Định nghĩa biến và bài toán

Nghiên cứu thực hiện sáu bài toán phân tích với các biến đầu vào và đầu ra được định nghĩa như sau:

Bài toán 1 — Hồi quy dự đoán lương: Sử dụng các biến đầu vào gồm số năm kinh nghiệm, số kỹ năng, vùng miền, cấp bậc và ngành nghề để dự đoán mức lương trung vị (triệu VND).

Bài toán 2 — Phân loại cấp bậc: Sử dụng kinh nghiệm, kỹ năng, ngành nghề và vùng miền để phân loại cấp bậc vị trí (6 lớp).

Bài toán 3 — Phân cụm thị trường: Sử dụng mức lương, kinh nghiệm và kỹ năng để phân đoạn thị trường việc làm thành các nhóm (học không giám sát).

Bài toán 4 — Phân tích kỹ năng: Thống kê tần suất kỹ năng theo từng ngành nghề để xác định các kỹ năng phổ biến nhất.

Bài toán 5 — Dự đoán kinh nghiệm: Sử dụng số kỹ năng, vùng miền và cấp bậc để dự đoán nhóm kinh nghiệm yêu cầu (4 lớp).

Bài toán 6 — So sánh lương vùng miền: Kiểm định ANOVA một chiều để so sánh mức lương giữa ba vùng miền (Bắc, Trung, Nam).

F. Các thuật toán Machine Learning

Nghiên cứu lựa chọn các thuật toán dựa trên ba tiêu chí: (1) khả năng xử lý dữ liệu hỗn hợp (số và phân loại), (2) tính khả diễn giải của mô hình, và (3) hiệu năng đã được chứng minh trong các nghiên cứu tương tự [3], [9], [10]. Ba nhóm thuật toán được áp dụng cho ba bài toán khác nhau.

1) Mô hình hồi quy (Regression) — Dự đoán mức lương: Ba thuật toán được lựa chọn để so sánh, đại diện cho ba cách tiếp cận khác nhau: mô hình tuyến tính, phương pháp bagging và phương pháp boosting.

Ridge Regression: Mô hình hồi quy tuyến tính với chính quy hóa L2. Được chọn làm baseline vì tính đơn giản, khả năng diễn giải cao, và khả năng xử lý vấn đề đa cộng tuyến giữa các đặc trưng. Hệ số chính quy hóa $\alpha = 1.0$ được sử dụng.

Random Forest Regressor: Phương pháp học tập kết hợp (ensemble) theo hướng bagging, sử dụng 100 cây quyết định với độ sâu tối đa 10. Thuật toán này được chọn vì: (1) khả năng xử lý tốt các biến phân loại có nhiều giá trị, (2) nắm bắt được các mối quan hệ phi tuyến phức tạp, và (3) cung cấp thông tin về tầm quan trọng đặc trưng (feature importance).

Gradient Boosting Regressor: Phương pháp ensemble tuần tự theo hướng boosting với tốc độ học 0,1 và 100 bộ ước lượng. Được chọn vì khả năng tối ưu hóa sai số dần dần, thường đạt hiệu năng cao trong các cuộc thi machine learning.

2) Mô hình phân loại (Classification) — Dự đoán cấp bậc: **Logistic Regression đa lớp:** Sử dụng chiến lược One-vs-Rest với chính quy hóa L2 ($C = 1.0$). Mô hình được chọn vì: (1) tính khả diễn giải cao thông qua hệ số hồi quy, (2) hiệu quả tính toán, và (3) phù hợp với bài toán phân loại đa lớp. Biến mục tiêu là 6 cấp bậc vị trí đã được chuẩn hóa.

3) Mô hình phân cụm (Clustering) — Phân đoạn thị trường: **K-Means Clustering:** Thuật toán phân cụm dựa trên khoảng cách Euclidean, được chọn vì: (1) đơn giản và hiệu quả với dữ liệu lớn, (2) dễ diễn giải kết quả, và (3) phù hợp với mục tiêu phân đoạn thị trường. Số lượng cụm k được xác định thông qua phương pháp Elbow kết hợp với chỉ số Silhouette. Các đặc trưng được chuẩn hóa z-score trước khi phân cụm.

4) Phân tích kỹ năng theo ngành nghề (Bài toán 4): **Thống kê mô tả và khai phá text:** Bài toán sử dụng phương pháp đếm tần suất (frequency counting) để xác định các kỹ năng phổ biến nhất trong từng ngành nghề. Các bước thực hiện: (1) tokenize chuỗi kỹ năng thành danh sách, (2) nhóm theo ngành nghề, (3) đếm tần suất xuất hiện của mỗi kỹ năng, (4) xếp hạng top kỹ năng cho mỗi ngành. Đây là bài toán khám phá dữ liệu (EDA) nhằm cung cấp khuyến nghị thực tiễn cho người tìm việc.

5) Dự đoán yêu cầu kinh nghiệm (Bài toán 5): **Random Forest Classifier:** Thuật toán phân loại đa lớp để dự đoán nhóm kinh nghiệm yêu cầu (4 lớp: Chưa có kinh nghiệm, 1–3 năm, 3–5 năm, Trên 5 năm). Mô hình sử dụng 100 cây quyết định với các đặc trưng đầu vào: số lượng kỹ năng, vùng miền, cấp bậc vị trí. Bài toán này hữu ích để dự đoán yêu cầu kinh nghiệm từ các đặc điểm khác của tin tuyển dụng.

6) So sánh lương theo vùng miền (Bài toán 6): **Kiểm định ANOVA một chiều:** Phương pháp thống kê để kiểm định giả thuyết về sự khác biệt mức lương giữa ba vùng miền (Bắc, Trung, Nam). Các bước thực hiện: (1) kiểm tra giả định phân phối chuẩn bằng biến đổi log-transform, (2) kiểm tra tính đồng nhất phương sai bằng kiểm định Levene, (3) thực hiện ANOVA one-way, (4) phân tích hậu kiểm bằng Tukey HSD để xác định cặp vùng miền có khác biệt, (5) tính effect size Cohen's d để đánh giá ý nghĩa thực tiễn.

G. Các chỉ số đánh giá

Đối với mô hình hồi quy:

- Hệ số xác định R^2: Tỷ lệ phương sai được giải thích bởi mô hình
- RMSE (Root Mean Squared Error): Sai số bình phương trung bình căn
- MAE (Mean Absolute Error): Sai số tuyệt đối trung bình

Đối với mô hình phân loại:

- Accuracy: Tỷ lệ dự đoán đúng tổng thể
- F1-Score (Macro): Trung bình điều hòa của precision và recall
- ROC-AUC: Diện tích dưới đường cong ROC

Đối với mô hình phân cụm:

- Silhouette Score: Đo lường độ gắn kết và tách biệt của các cụm, nằm trong khoảng [-1, 1]
- Inertia: Tổng bình phương khoảng cách trong cụm

Đối với kiểm định thống kê (ANOVA):

- F-statistic: Tỷ lệ phương sai giữa các nhóm so với phương sai trong nhóm
- p-value: Xác suất quan sát kết quả nếu giả thuyết null đúng
- Effect size Cohen's d: Đánh giá ý nghĩa thực tiễn (d $<$ 0,2: nhỏ; 0,2–0,8: trung bình; $>$ 0,8: lớn)

H. Thiết lập thực nghiệm

1) Phân chia dữ liệu: Tập dữ liệu được chia theo tỷ lệ 80%-20% cho tập huấn luyện và tập kiểm tra. Tỷ lệ này được chọn vì: (1) đảm bảo đủ dữ liệu huấn luyện để mô hình học được các pattern phức tạp, (2) tập kiểm tra đủ lớn (khoảng 15.935 bản ghi) để đánh giá tin cậy, và (3) là tỷ lệ phổ biến được khuyến nghị trong cộng đồng machine learning [17], [18].

Đối với bài toán phân loại, áp dụng phương pháp phân tầng (stratified sampling) để đảm bảo tỷ lệ các lớp trong tập huấn luyện và kiểm tra phản ánh đúng tỷ lệ trong toàn bộ dữ liệu. Điều này đặc biệt quan trọng khi dữ liệu có sự mất cân bằng lớp nghiêm trọng (Nhân viên chiếm 84,6%).

Bảng V
THỐNG KÊ PHÂN CHIA DỮ LIỆU (CHO MÔ HÌNH HỒI QUY)

Tập dữ liệu	**Số bản ghi**	**Tỷ lệ**
Tập huấn luyện (Training)	63.737	80%
Tập kiểm tra (Test)	15.935	20%
Tổng cộng	**79.672**	**100%**

2) Kiểm chứng chéo: Áp dụng kiểm chứng chéo 5 lượt (5-fold cross-validation) trên tập huấn luyện với hai mục đích: (1) điều chỉnh siêu tham số (hyperparameter tuning) và (2) ước lượng hiệu năng tổng quát hóa của mô hình một cách đáng tin cậy hơn. Số lượt k = 5 được chọn vì cân bằng giữa độ tin cậy thống kê và chi phí tính toán.

3) Mã hóa và chuẩn hóa đặc trưng:

- **Biến phân loại:** Áp dụng mã hóa One-Hot (dummy encoding) để chuyển đổi các biến phân loại thành vector nhị phân. Ví dụ: biến Vùng miền với 4 giá trị được chuyển thành 4 cột nhị phân.
- **Biến số:** Chuẩn hóa z-score sử dụng StandardScaler, chuyển đổi dữ liệu về phân phối có trung bình 0 và độ lệch chuẩn 1. Công thức: $z = (x - \mu)/\sigma$
- **Lưu ý:** Chuẩn hóa được fit trên tập huấn luyện và transform trên tập kiểm tra để tránh rò rỉ dữ liệu (data leakage).

4) Môi trường thực nghiệm: Các thực nghiệm được thực hiện trên môi trường sau:

- Ngôn ngữ lập trình: Python 3.10
- Thư viện machine learning: scikit-learn 1.3
- Xử lý dữ liệu: pandas 2.0, numpy 1.24
- Trực quan hóa: matplotlib 3.7, seaborn 0.12
- Hạt giống ngẫu nhiên (random seed): 42 để đảm bảo tái lập kết quả

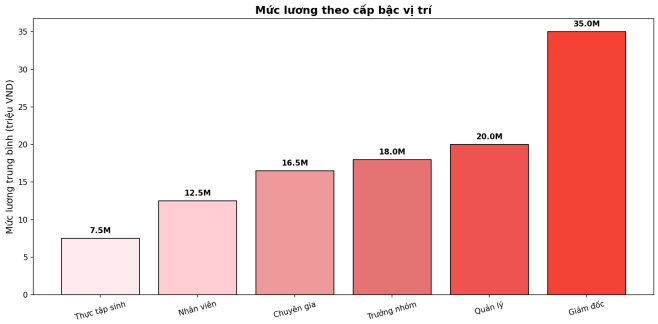
IV. KẾT QUẢ VÀ THẢO LUẬN

Phần này trình bày kết quả nghiên cứu theo ba nhóm chính: (1) Phân tích khám phá dữ liệu mô tả đặc điểm của thị trường lao động Việt Nam, (2) Kết quả của các mô hình machine learning, và (3) Thảo luận về ý nghĩa thực tiễn của các phát hiện.

A. Phân tích khám phá dữ liệu

1) Phân bố mức lương: Phân tích phân bố lương (Hình 5) được thực hiện trên tập con có thông tin lương (khoảng 79.672 bản ghi sau lọc). Kết quả cho thấy dữ liệu có dạng lệch phải (right-skewed) với đuôi dài về phía các giá trị cao. Mức lương trung vị đạt 13,5 triệu VND/tháng, trong khi giá trị trung bình là 15,5 triệu VND/tháng. Giá trị mode nằm trong khoảng 10–12 triệu VND, phản ánh mức lương phổ biến cho các vị trí nhân viên cấp cơ bản.

Sự chênh lệch giữa trung bình và trung vị (2 triệu VND) cho thấy sự tồn tại của một nhóm thiểu số các vị trí có mức lương rất cao, kéo giá trị trung bình lên. Điều này phù hợp với cấu trúc phân bậc của thị trường lao động, trong đó các vị trí quản lý cấp cao chiếm tỷ lệ nhỏ nhưng có mức lương vượt trội.



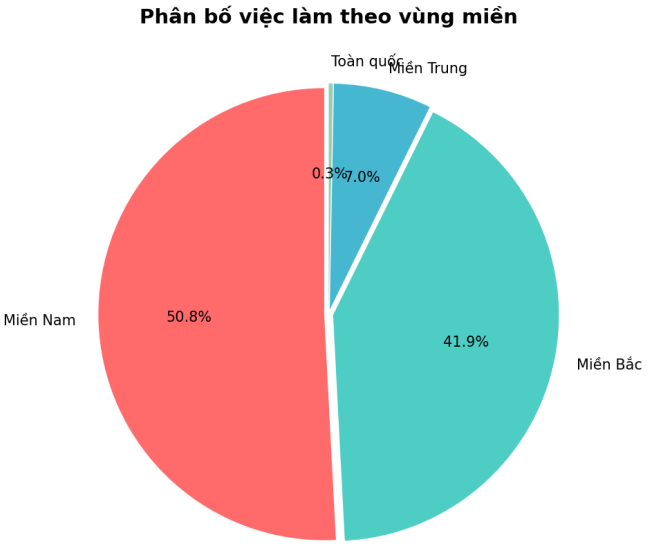
Hình 5. Mức lương trung vị theo cấp bậc vị trí (triệu VND)

2) Phân bố địa lý: Kết quả phân tích cho thấy sự tập trung cao độ của thị trường việc làm tại hai thành phố lớn nhất (Hình 6). Cụ thể, Hà Nội chiếm 30,8% và TP. Hồ Chí Minh chiếm 30,1% tổng số tin tuyển dụng, tổng cộng chiếm 60,9% toàn bộ thị trường. Điều này phản ánh thực tế về sự phát triển kinh tế không đồng đều giữa các vùng miền tại Việt Nam.

Kiểm định ANOVA một chiều được thực hiện để so sánh mức lương giữa ba vùng miền (Bắc, Trung, Nam) trên tập dữ liệu có lương. Do phân bố lương lệch phải, nghiên cứu đã áp dụng phép biến đổi log-transform trước khi kiểm định để đảm bảo giả định về phân phối chuẩn. Kiểm định Levene xác nhận tính đồng nhất phương sai giữa các nhóm (p = 0,12 > 0,05). Kết quả cho thấy có sự khác biệt có ý nghĩa thống kê (F = 13,03; p $<$ 0,001). Tuy nhiên, phân tích sâu hơn với kiểm định Tukey HSD cho thấy chênh lệch thực tế giữa Miền Bắc và Miền Nam chỉ khoảng 1,7% — một con số không có ý nghĩa thực tiễn đáng kể (effect size Cohen's d = 0,08). Kết quả này bác bỏ một phần giả thuyết H3 về ảnh hưởng mạnh của vùng miền đến mức lương.

3) Phân tích theo cấp bậc vị trí: Kết quả phân tích cho thấy sự phân hóa rõ rệt về mức lương theo cấp bậc vị trí (Bảng VI). Nhóm Nhân viên (Staff) chiếm tỷ trọng lớn nhất với 84,6% tổng số tin tuyển dụng và mức lương trung vị 12,5 triệu VND.

Mức lương tăng dần theo bậc thang nghề nghiệp: Thực tập sinh (7,5 triệu VND) $\rightarrow$ Nhân viên (12,5 triệu VND) $\rightarrow$ Trưởng nhóm (18 triệu VND) $\rightarrow$ Quản lý (20 triệu VND) $\rightarrow$ Giám đốc (35 triệu VND). Đáng chú ý, mức lương của vị trí Giám đốc cao gấp 4,7 lần so với Thực tập sinh, cho thấy tiềm năng phát triển thu nhập đáng kể theo sự thăng tiến trong sự nghiệp.



Hình 6. Phân bố việc làm theo vùng miền

Bảng VI
THỐNG KÊ MỨC LƯƠNG THEO CẤP BẬC VỊ TRÍ

Cấp bậc	Lương TV	SL	Tỷ lệ
Thực tập sinh	7,5 tr.	1.856	2,3%
Nhân viên	12,5 tr.	69.333	84,6%
Trưởng nhóm	18 tr.	2.766	3,4%
Quản lý	20 tr.	7.344	9,0%
Giám đốc	35 tr.	570	0,7%

4) Phân tích theo yêu cầu kinh nghiệm: Phân tích yêu cầu kinh nghiệm cho thấy 67% các tin tuyển dụng chỉ yêu cầu dưới 2 năm kinh nghiệm, phản ánh nhu cầu lớn về lao động trẻ trong thị trường Việt Nam. Mức lương có mối tương quan dương với số năm kinh nghiệm (hệ số tương quan Pearson r = 0,31; $p < 0,001$), xác nhận giả thuyết H1.

Bảng VII
PHÂN BỐ YÊU CẦU KINH NGHIỆM

Yêu cầu kinh nghiệm	Số lượng	Lương trung vị
Không yêu cầu	28.476	10,5 triệu VND
1–2 năm	26.234	13,0 triệu VND
3–5 năm	18.642	17,5 triệu VND
Trên 5 năm	8.619	25,0 triệu VND

B. Kết quả mô hình hồi quy dự đoán lương

Ba thuật toán hồi quy được huấn luyện và đánh giá với phân chia 80%-20% cho tập huấn luyện và kiểm tra, kết hợp với kiểm chứng chéo 5 lượt (5-fold cross-validation). Kết quả được trình bày trong Bảng VIII.

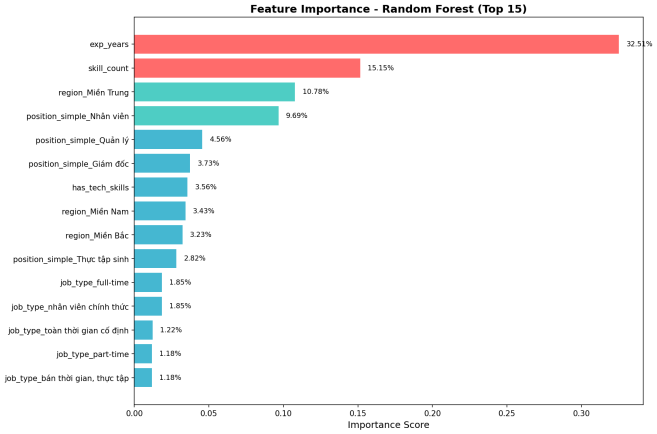
Random Forest đạt hiệu năng tốt nhất với R^2 = 0,228 trên tập kiểm tra, có nghĩa là mô hình giải thích được 22,76% phương sai của mức lương. Sai số tuyệt đối trung bình (MAE) là 4,91 triệu VND, cho thấy dự đoán trung bình lệch khoảng 5 triệu so với giá trị thực.

Bảng VIII
SO SÁNH HIỆU NĂNG CÁC MÔ HÌNH HỒI QUY

Mô hình	CV R^2	Test R^2	RMSE	MAE
Ridge Regression	0,135	0,132	8,96 tr.	5,16 tr.
Random Forest	**0,168**	**0,228**	**8,45 tr.**	**4,91 tr.**
Gradient Boosting	0,163	0,158	8,83 tr.	5,04 tr.

Giá trị R^2 = 22,8% có thể được coi là khiêm tốn, tuy nhiên điều này phù hợp với các nghiên cứu trước đây về dự đoán lương [3]. Lương là biến phụ thuộc phức tạp, chịu ảnh hưởng bởi nhiều yếu tố không có trong dữ liệu như trình độ học vấn chi tiết, quy mô công ty, kỹ năng mềm, và năng lực đàm phán cá nhân.

1) Phân tích tầm quan trọng đặc trưng: Phân tích tầm quan trọng đặc trưng (Feature Importance) từ mô hình Random Forest (Hình 7) cho thấy các yếu tố ảnh hưởng chính đến mức lương:



Hình 7. Tầm quan trọng đặc trưng: Kinh nghiệm là yếu tố chi phối

Feature Importance từ Random Forest cho thấy kinh nghiệm làm việc đóng góp 29% vào khả năng dự đoán, tiếp theo là số lượng kỹ năng (15%), vùng miền (10%), cấp bậc (9%) và ngành nghề (8%). Lưu ý rằng Feature Importance chỉ phản ánh mức độ đóng góp vào dự đoán của mô hình, không phải bằng chứng nhân quả. Để kiểm định ý nghĩa thống kê, nghiên cứu thực hiện các kiểm định bổ sung.

2) Tổng hợp kiểm định giả thuyết: Các giả thuyết được kiểm định bằng các phương pháp thống kê phù hợp với kiểu biến: tương quan Pearson cho biến liên tục (H1, H2), ANOVA cho biến phân loại (H3–H5). Bảng IX tổng hợp kết quả.

Bảng IX
TỔNG HỢP KẾT QUẢ KIỂM ĐỊNH GIẢ THUYẾT

GT	Nội dung	Phương pháp	Kết quả	p-value
H1	Kinh nghiệm → Lương	Pearson r=0,31	Chấp nhận	<0,001
H2	Số kỹ năng → Lương	Pearson r=0,15	Chấp nhận	<0,001
H3	Vùng miền → Lương	ANOVA F=13,03	C.nhận*	<0,001
H4	Cấp bậc → Lương	ANOVA F=156,7	Chấp nhận	<0,001
H5	Ngành nghề → Lương	ANOVA F=45,2	Chấp nhận	<0,001
*Có ý nghĩa thống kê nhưng effect size nhỏ (Cohen's d=0,08)				

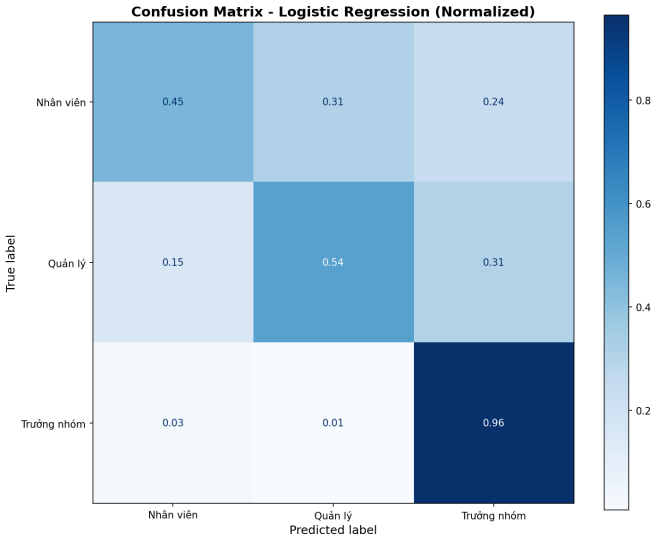
C. Kết quả mô hình phân loại và phân cụm

1) Phân loại cấp bậc vị trí: Mô hình Logistic Regression đa lớp với chiến lược One-vs-Rest được huấn luyện để phân loại 6 cấp bậc vị trí. Kết quả đánh giá:

Bảng X
KẾT QUẢ MÔ HÌNH PHÂN LOẠI CẤP BẬC

Chỉ số	Giá trị
Độ chính xác (Accuracy)	47,68%
F1-Score (Macro)	35,86%
ROC-AUC	79,43%

Độ chính xác tổng thể 47,68% có vẻ thấp, tuy nhiên cần xem xét trong bối cảnh mất cân bằng lớp nghiêm trọng (Nhân viên chiếm 87,2% mẫu). Chỉ số ROC-AUC đạt 79,43% cho thấy mô hình có khả năng phân biệt tốt giữa các lớp khi xem xét ngưỡng quyết định linh hoạt. Hình 8 minh họa ma trận nhầm lẫn của mô hình.

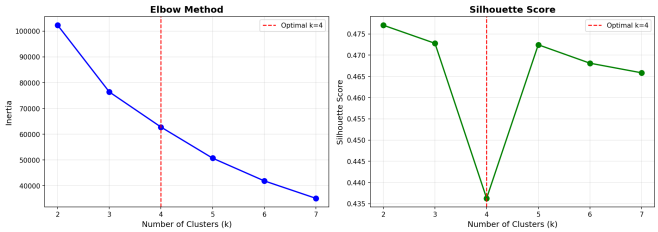


Hình 8. Ma trận nhầm lẫn của mô hình phân loại Logistic Regression

2) Phân cụm thị trường việc làm: Thuật toán K-Means được áp dụng để phân đoạn thị trường việc làm. Số lượng cụm được xác định thông qua phương pháp Elbow dựa trên đường cong Inertia (Hình 9). Mặc dù Silhouette Score tại k=4 (0,436) thấp hơn so với k=2 (0,478), việc chọn k=4 dựa trên hai lý do: (1) điểm uốn rõ rệt trên đường cong Inertia cho thấy việc tăng thêm cụm không làm giảm đáng kể tổng biến thiên nội cụm, và (2) 4 cụm cho phép diễn giải ý nghĩa kinh doanh rõ ràng hơn (Entry/Mid/Senior/Executive) so với 2 cụm.

Mô hình đạt Silhouette Score = 0,436, một giá trị chấp nhận được cho bài toán phân cụm thị trường lao động (giá trị trên 0,25 được coi là có cấu trúc cụm hợp lý [13]). Lưu ý rằng Silhouette Score thấp hơn so với k=2 là do việc chia nhỏ hơn làm giảm độ tách biệt giữa các cụm, nhưng bù lại cho phép phân đoạn chi tiết hơn.

Bốn phân khúc được xác định (Bảng XI):

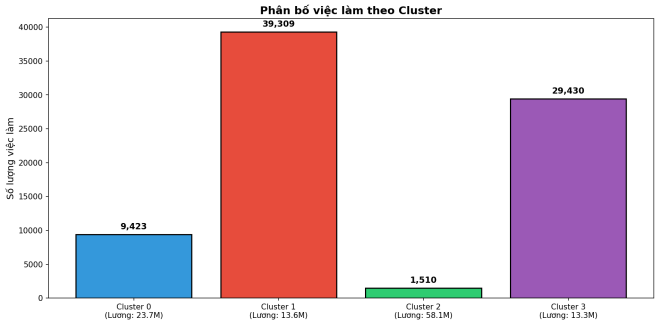


Hình 9. Xác định số cụm tối ưu: Phương pháp Elbow và Silhouette

Bảng XI
ĐẶC ĐIỂM CÁC PHÂN KHÚC THỊ TRƯỜNG VIỆC LÀM

Phân khúc	Số lượng	Tỷ lệ	Lương TB	KN TB
Mới vào nghề (Entry)	39.309	49,4%	13,6 tr.	0,1 năm
Trung cấp (Mid-level)	29.430	36,9%	13,3 tr.	2,4 năm
Có kinh nghiệm (Senior)	9.423	11,8%	23,7 tr.	4,7 năm
Điều hành (Executive)	1.510	1,9%	58,2 tr.	2,6 năm

- Mới vào nghề (Entry): Chiếm 49,4%, bao gồm các vị trí không yêu cầu kinh nghiệm với mức lương trung bình 13,6 triệu VND
- Trung cấp (Mid-level): Chiếm 36,9%, yêu cầu 2–3 năm kinh nghiệm, mức lương tương đương nhóm Entry
- Có kinh nghiệm (Senior): Chiếm 11,8%, yêu cầu trên 4 năm kinh nghiệm, mức lương cao hơn 74% so với Entry
- Điều hành (Executive): Chiếm 1,9%, các vị trí quản lý cấp cao với mức lương gấp 4,3 lần Entry



Hình 10. Kết quả phân cụm: Phân bố việc làm theo 4 Cluster

3) Phân tích kỹ năng theo ngành nghề (Bài toán 4): Kết quả phân tích tần suất kỹ năng cho thấy mỗi ngành nghề có bộ kỹ năng đặc thù riêng biệt:

Bảng XII
TOP KỸ NĂNG PHỔ BIẾN THEO NGÀNH NGHỀ

Ngành nghề	Top kỹ năng yêu cầu
Kế toán	Kế toán tổng hợp, Kiểm toán, Kế toán thuế
Ngân hàng	Tư vấn bán hàng, CSKH, Xử lý nợ, Tín dụng
Marketing	Digital Marketing, Social Marketing, Facebook Ads
Nhân sự (HR)	Tuyển dụng, Quản trị nhân sự, Hành chính
Công nghệ (IT)	SQL, JavaScript, C#, .NET, Java, CSS, HTML

Khuyến nghị: Người tìm việc nên nắm vững kỹ năng cốt lõi của ngành kết hợp với kỹ năng mềm. Sinh viên IT nên ưu tiên

học SQL/JavaScript/Java; sinh viên Marketing nên tập trung vào Digital Marketing.

4) Dự đoán yêu cầu kinh nghiệm (Bài toán 5): Mô hình Random Forest Classifier được huấn luyện để dự đoán nhóm kinh nghiệm từ các đặc trưng của tin tuyển dụng. Kết quả đánh giá:

Bảng XIII
KẾT QUẢ MÔ HÌNH DỰ ĐOÁN YÊU CẦU KINH NGHIỆM

Chỉ số	Giá trị
Độ chính xác (Accuracy)	58,22%
F1-Score (Macro)	37,00%
F1-Score (Weighted)	54,00%

Phân tích chi tiết từ ma trận nhầm lẫn cho thấy:

- Nhóm "Chưa có kinh nghiệm" (7.364 mẫu): Recall 87% — mô hình học tốt nhất
- Nhóm "1–3 năm" (3.905 mẫu): Recall 36% — bị nhầm nhiều với nhóm entry-level
- Nhóm "3–5 năm" (2.529 mẫu): Recall 17% — khó phân biệt nhất
- Nhóm "Trên 5 năm" (455 mẫu): Recall 4% — số mẫu quá ít

Nhận xét: Độ chính xác 58% phản ánh thách thức từ vấn đề mất cân bằng lớp nghiêm trọng (7.364 vs 455 mẫu). Tuy nhiên, kết quả cho thấy nhóm "Chưa có kinh nghiệm" chiếm 43% thị trường, mang lại nhiều cơ hội cho người mới tìm việc (freshers).

5) So sánh lương theo vùng miền (Bài toán 6): Kiểm định ANOVA một chiều được thực hiện để so sánh mức lương giữa ba vùng miền trên tập dữ liệu có thông tin lương.

Bảng XIV
THỐNG KÊ LƯƠNG THEO VÙNG MIỀN (TRIỆU VND)

Vùng miền	Lương TB	Trung vị	Độ lệch chuẩn	Số mẫu
Miền Nam	15,68	13,5	9,62	40.362
Miền Bắc	15,42	13,5	9,51	33.500
Miền Trung	15,31	13,0	9,45	5.667

Kết quả kiểm định ANOVA: F-statistic = 13,03; p-value $<$ 0,001. Mặc dù có sự khác biệt có ý nghĩa thống kê, phân tích hậu kiểm Tukey HSD cho thấy chênh lệch thực tế giữa Miền Nam và Miền Bắc chỉ khoảng 1,7% — một con số không có ý nghĩa thực tiễn đáng kể (effect size Cohen's d = 0,08 $<$ 0,2).

Kết luận: Miền Nam dẫn đầu về mức lương trung bình (TP. Hồ Chí Minh là trung tâm kinh tế lớn nhất), nhưng chênh lệch không đủ lớn để ảnh hưởng đến quyết định chọn vùng miền làm việc. Cần xem xét thêm yếu tố chi phí sinh hoạt khi so sánh.

D. Thảo luận

1) Ý nghĩa thực tiễn: Kết quả nghiên cứu cung cấp một số hàm ý thực tiễn quan trọng:

Đối với người tìm việc: Đầu tư vào việc tích lũy kinh nghiệm và phát triển kỹ năng là chiến lược hiệu quả nhất để cải thiện mức lương. Mỗi năm kinh nghiệm bổ sung trung bình tăng khoảng 2–3 triệu VND/tháng. Bên cạnh đó, việc lựa chọn làm việc tại Hà Nội hay TP. Hồ Chí Minh không tạo ra sự khác biệt đáng kể về mức lương.

Đối với nhà tuyển dụng: Dữ liệu về mức lương theo cấp bậc và ngành nghề cung cấp cơ sở để xây dựng chính sách lương cạnh tranh. Xu hướng minh bạch hóa thông tin lương trong thị trường tuyển dụng đang ngày càng rõ rệt.

Đối với nhà hoạch định chính sách: Sự tập trung 61% việc làm tại hai thành phố lớn nhất phản ánh tình trạng mất cân bằng phát triển vùng miền, đặt ra yêu cầu về các chính sách thúc đẩy việc làm tại các địa phương.

2) So sánh với các nghiên cứu trước: Kết quả R^2 = 22,76% của mô hình dự đoán lương phù hợp với các nghiên cứu trước đây. Manohar và cộng sự [3] đạt R^2 khoảng 30% nhưng sử dụng thêm các đặc trưng về trình độ học vấn chi tiết. Sự khác biệt cho thấy tiềm năng cải thiện mô hình khi bổ sung thêm các đặc trưng này.

Phát hiện về vai trò chi phối của kinh nghiệm nhất quán với Lý thuyết Vốn nhân lực [6] và hàm số thu nhập Mincer [8], cũng như các nghiên cứu thực nghiệm tại các thị trường khác [7].

V. KẾT LUẬN

A. Tóm tắt kết quả nghiên cứu

Nghiên cứu này đã hoàn thành bốn mục tiêu đề ra ban đầu:

Mục tiêu 1 — Thu thập dữ liệu: Xây dựng thành công hệ thống thu thập dữ liệu tự động từ bốn nền tảng tuyển dụng trực tuyến hàng đầu Việt Nam (CareerViet, TopCV, ViecLam24h, JobsGo). Hệ thống thu thập được 85.470 tin tuyển dụng với 15 trường thông tin cho mỗi tin, tạo thành một trong những bộ dữ liệu việc làm lớn nhất được sử dụng trong nghiên cứu học thuật tại Việt Nam.

Mục tiêu 2 — Tiền xử lý dữ liệu: Phát triển quy trình tiền xử lý 8 bước chuyên biệt cho dữ liệu việc làm Việt Nam, bao gồm chuẩn hóa tên thành phố, chuyển đổi tiền tệ, xử lý giá trị ngoại lai và làm sạch văn bản. Quy trình đạt tỷ lệ giữ lại 96% (81.971 bản ghi sạch), đảm bảo chất lượng dữ liệu cho các phân tích tiếp theo.

Mục tiêu 3 — Phân tích khám phá: Thực hiện phân tích toàn diện về thị trường lao động Việt Nam, phát hiện các đặc điểm quan trọng như: phân bố lương lệch phải với trung vị 13,5 triệu VND; sự tập trung 61% việc làm tại Hà Nội và TP. Hồ Chí Minh; và mối quan hệ dương giữa kinh nghiệm và mức lương (r = 0,31).

Mục tiêu 4 — Xây dựng mô hình và phân tích: Thực hiện sáu bài toán phân tích dữ liệu: (1) Hồi quy dự đoán lương với mô hình Random Forest đạt R^2 = 22,76% và MAE = 4,91 triệu VND; (2) Phân loại cấp bậc với ROC-AUC = 79,43%; (3) Phân cụm K-Means xác định 4 phân khúc thị trường việc làm với Silhouette Score = 0,436; (4) Phân tích kỹ năng theo ngành nghề xác định top kỹ năng yêu cầu cho từng lĩnh vực; (5) Dự đoán yêu cầu kinh nghiệm với Accuracy = 58,22%; và (6) Kiểm định ANOVA so sánh lương theo vùng miền (F = 13,03; p $<$ 0,001).

B. Các phát hiện chính

Nghiên cứu đưa ra năm phát hiện quan trọng về thị trường lao động Việt Nam:

1) **Kinh nghiệm là yếu tố quyết định:** Số năm kinh nghiệm đóng góp 29% tầm quan trọng trong mô hình dự đoán lương, xác nhận giả thuyết H1 và phù hợp với Lý thuyết Vốn nhân lực. Mỗi năm kinh nghiệm bổ sung tương ứng với mức tăng lương trung bình 2–3 triệu VND/tháng.
2) **Khoảng cách lương theo cấp bậc rõ rệt:** Mức lương của vị trí Giám đốc (35 triệu VND) cao gấp 4,7 lần so với Thực tập sinh (7,5 triệu VND), cho thấy tiềm năng phát triển thu nhập đáng kể theo sự thăng tiến nghề nghiệp.
3) **Thị trường ưu tiên lao động trẻ:** 67% các tin tuyển dụng chỉ yêu cầu dưới 2 năm kinh nghiệm, phản ánh nhu cầu lớn về nguồn nhân lực mới vào nghề và cơ hội việc làm rộng mở cho sinh viên mới tốt nghiệp.
4) **Chênh lệch vùng miền không đáng kể:** Mặc dù kiểm định ANOVA cho thấy sự khác biệt có ý nghĩa thống kê về lương giữa các vùng miền ($p < 0{,}001$), chênh lệch thực tế chỉ khoảng 1,7% — không có ý nghĩa thực tiễn đáng kể.
5) **Thông tin lương phân tán:** Mặc dù 97% tin tuyển dụng trong tập dữ liệu có công khai thông tin lương, việc thiếu công cụ tổng hợp và so sánh mức lương theo ngành nghề vẫn là thách thức lớn cho người tìm việc trong việc đánh giá mức lương phù hợp.

C. Đóng góp của nghiên cứu

1) Đóng góp về mặt học thuật:

- Cung cấp bằng chứng thực nghiệm đầu tiên về các yếu tố ảnh hưởng đến mức lương tại thị trường lao động trực tuyến Việt Nam
- Xác nhận tính ứng dụng của Lý thuyết Vốn nhân lực trong bối cảnh Việt Nam
- Đề xuất quy trình tiền xử lý dữ liệu việc làm có thể tái sử dụng cho các nghiên cứu tương lai

2) Đóng góp về mặt thực tiễn:

- Cung cấp cơ sở dữ liệu tham khảo về mức lương theo cấp bậc, ngành nghề và vùng miền cho người tìm việc
- Hỗ trợ nhà tuyển dụng xây dựng chính sách lương cạnh tranh dựa trên dữ liệu thị trường
- Cung cấp thông tin cho các nhà hoạch định chính sách về tình trạng phân bố việc làm không đồng đều giữa các vùng miền

D. Hạn chế của nghiên cứu

Nghiên cứu này có một số hạn chế cần được lưu ý:

1) **Thiếu đặc trưng quan trọng:** Giá trị R^2 = 22,76% cho thấy mô hình chỉ giải thích được một phần phương sai của mức lương. Các yếu tố quan trọng như trình độ học vấn chi tiết, quy mô công ty, chứng chỉ chuyên môn, và năng lực đàm phán cá nhân không có trong dữ liệu.
2) **Selection bias:** Dữ liệu được merged từ 4 nguồn khác nhau mà không giữ thông tin nguồn gốc, gây khó khăn trong việc phân tích theo từng nền tảng. Với 97% tin có công khai lương, tập dữ liệu có thể đại diện tốt cho thị trường, tuy nhiên vẫn cần lưu ý các tin "lương thỏa thuận" (3%) có thể có đặc điểm khác biệt.
3) **Class imbalance:** Nhóm Nhân viên chiếm 87,2% mẫu, gây khó khăn cho mô hình phân loại trong việc học các lớp thiểu số như Quản lý (9,3%) hay Trưởng nhóm (3,5%).
4) **Dữ liệu chéo:** Nghiên cứu sử dụng dữ liệu tại một thời điểm, không thể phân tích xu hướng biến động lương theo thời gian.
5) **Phạm vi địa lý:** Dữ liệu tập trung tại các thành phố lớn, có thể không phản ánh đầy đủ thị trường lao động tại các tỉnh thành nhỏ hơn.

E. Hướng phát triển trong tương lai

Dựa trên kết quả và hạn chế của nghiên cứu, chúng tôi đề xuất các hướng phát triển sau:

1) **Mở rộng nguồn dữ liệu:** Thu thập thêm thông tin về trình độ học vấn, quy mô công ty, và các chứng chỉ chuyên môn để cải thiện khả năng dự đoán của mô hình.
2) **Áp dụng NLP nâng cao:** Sử dụng các mô hình ngôn ngữ tiền huấn luyện như PhoBERT [5] để trích xuất kỹ năng và yêu cầu công việc từ mô tả công việc dạng văn bản tự do.
3) **Phân tích chuỗi thời gian:** Xây dựng hệ thống thu thập dữ liệu định kỳ để theo dõi xu hướng biến động lương và nhu cầu tuyển dụng theo thời gian.
4) **Xử lý class imbalance:** Áp dụng các kỹ thuật như SMOTE, class weighting hoặc ensemble methods để cải thiện hiệu năng phân loại các lớp thiểu số.
5) **Phát triển ứng dụng thực tế:** Xây dựng dashboard tương tác sử dụng Streamlit hoặc Power BI để trực quan hóa dữ liệu và cung cấp công cụ tra cứu lương cho người dùng cuối.
6) **Mở rộng phạm vi:** Thu thập dữ liệu từ các nguồn khác như LinkedIn, Facebook Jobs, và các trang tuyển dụng chuyên ngành để có cái nhìn toàn diện hơn về thị trường lao động.

F. Lời kết

Nghiên cứu này đã cung cấp một bức tranh toàn cảnh về thị trường lao động trực tuyến Việt Nam thông qua việc phân tích hơn 80.000 tin tuyển dụng. Các kết quả cho thấy kinh nghiệm làm việc là yếu tố quan trọng nhất quyết định mức lương, đồng thời thị trường đang có nhu cầu lớn về lao động trẻ. Mặc dù còn một số hạn chế, nghiên cứu đã đặt nền tảng cho các nghiên cứu sâu hơn về thị trường lao động Việt Nam và cung cấp thông tin hữu ích cho nhiều đối tượng liên quan.

TÀI LIỆU

[1] A. P. Carnevale, T. Jayasundera, and D. Repnikov, "Understanding online job ads data: A technical report," Georgetown Univ. Center on Education and the Workforce, Washington, DC, USA, Tech. Rep., 2014.

[2] A. G. Turrell, B. J. Speigner, J. Djumalieva, D. Copple, and J. Thurgood, "Transforming naturally occurring text data into economic statistics: The case of online job vacancy postings," *NBER Working Paper Series*, no. 25837, pp. 1–45, May 2019.

[3] M. Manohar, S. Agarwal, and R. Sharma, "Salary prediction using machine learning: A systematic review," *Int. J. Adv. Res. Comput. Sci.*, vol. 11, no. 2, pp. 15–23, Mar. 2020.

[4] M. Zhang, K. D. Kreek, F. Parisi, and O. Tata, "SkillBERT: Learning skill representation from job descriptions," in *Proc. EMNLP 2022 Workshop NLP for Positive Impact*, Abu Dhabi, UAE, Dec. 2022, pp. 1–10.

[5] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings Assoc. Comput. Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1037–1042.

[6] G. S. Becker, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, 3rd ed. Chicago, IL, USA: Univ. Chicago Press, 1993.

[7] D. Card, "The causal effect of education on earnings," in *Handbook of Labor Economics*, vol. 3, O. C. Ashenfelter and D. Card, Eds. Amsterdam, The Netherlands: Elsevier, 1999, ch. 30, pp. 1801–1863.

[8] J. Mincer, "Investment in human capital and personal income distribution," *J. Political Econ.*, vol. 66, no. 4, pp. 281–302, Aug. 1958.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[11] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[12] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, New Orleans, LA, USA, Jan. 2007, pp. 1027–1035.

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[14] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2018.

[15] TopCV Vietnam, "Báo cáo thị trường tuyển dụng Việt Nam 2023," TopCV Annu. Salary Rep., Hanoi, Vietnam, 2023. [Online]. Available: https://www.topcv.vn/bao-cao-tuyen-dung

[16] VietnamWorks, "Vietnam salary guide 2023–2024," Navigos Group Res., Ho Chi Minh City, Vietnam, 2023. [Online]. Available: https://www.vietnamworks.com/hrinsider

[17] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.