ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH BỘ DỮ LIỆU TUYỂN DỤNG VIỆC LÀM VÀ XÂY DỰNG MÔ HÌNH HỒI QUY DỰ ĐOÁN MỨC LƯƠNG

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Lê Tuấn Hưng	21520250
2	Nguyễn Trọng Mạnh	21520343
3	Tô Trường Long	21521101

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Hiện nay, việc lựa chọn mức lương chi trả cho những công việc cụ thể đang là một bài toán khó đối với các nhà tuyển dụng do ảnh hưởng trực tiếp đến ngân sách và nhân sự của một công ty, tổ chức. Nhận thấy được điều đó, nhóm xây dựng một mô hình dự đoán mức lương cho các công việc dựa trên các yêu tố như: vị trí, dạng trả lương, loại công việc, học vấn.... Để xây dựng mô hình dự đoán mức lương nhóm bắt đầu với việc thu thập dữ liệu trên trang web vieclamtot.com [1] (một trang web con của chotot.com [2]) nơi các nhà tuyển dụng có thể đăng bài để tìm kiếm nguồn nhân lực phù hợp cho nhu cầu của mình.

Bộ dữ liệu được thu thập bằng 2 thư viện BeautifulSoup [3] và Selenium [4] chạy trên ngôn ngữ lập trình Python. Quá trình tiền xử lý và làm sạch dữ liệu, nhóm sử dụng ngôn ngữ Python với môi trường xử lý dữ liệu là Google Colab. Ngoài ra, nhóm áp dụng các kỹ thuật trực quan hóa trên một số thuộc tính có trong bộ dữ liệu và sử dụng Power BI [5] tiến hành trực quan theo kiểu dashboard nhằm đưa ra được một góc nhìn sinh động nhất về bộ dữ liệu.

Nhóm đã tiến hành phân tích tìm ra các mối tương quan, phân phối giữa các thuộc tính tới biến mục tiêu. Nhóm cũng thực nghiệm trên nhiều mô hình, sau đó chọn ra những mô hình tốt nhất kết hợp lại thành mô hình Voting [6] dự đoán mức lương với kết quả R2 score tốt nhất là 0.8395.

Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế, thu thập tại vieclamtot.com [1], không dựa trên đề tài nào khác và được dùng làm đồ án môn học môn DS105 - Phân tích và trực quan hóa dữ liệu trong học kỳ I năm học 2023/2024

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu được thu thập trên trang web vieclamtot.com bằng 2 thư viện là Selenium [4] và BeautifulSoup [3]. Bộ dữ liệu thu thập được có 19 thuộc tính và 2344 dòng dữ liệu với 3 biến dạng số và 16 biến phân loại.

Tên thuộc tính	Mô tả thuộc tính	Kiểu dữ liệu của thuộc tính
link	Link của bài tuyển dụng	Object
title	Tiêu đề bài tuyển dụng	Object
priceNormal	Mức lương	Object
companyName	Tên công ty/nhà tuyển dụng	Object
location	Tỉnh thành tuyển dụng công việc	Object
companyType	Loại công ty	Object

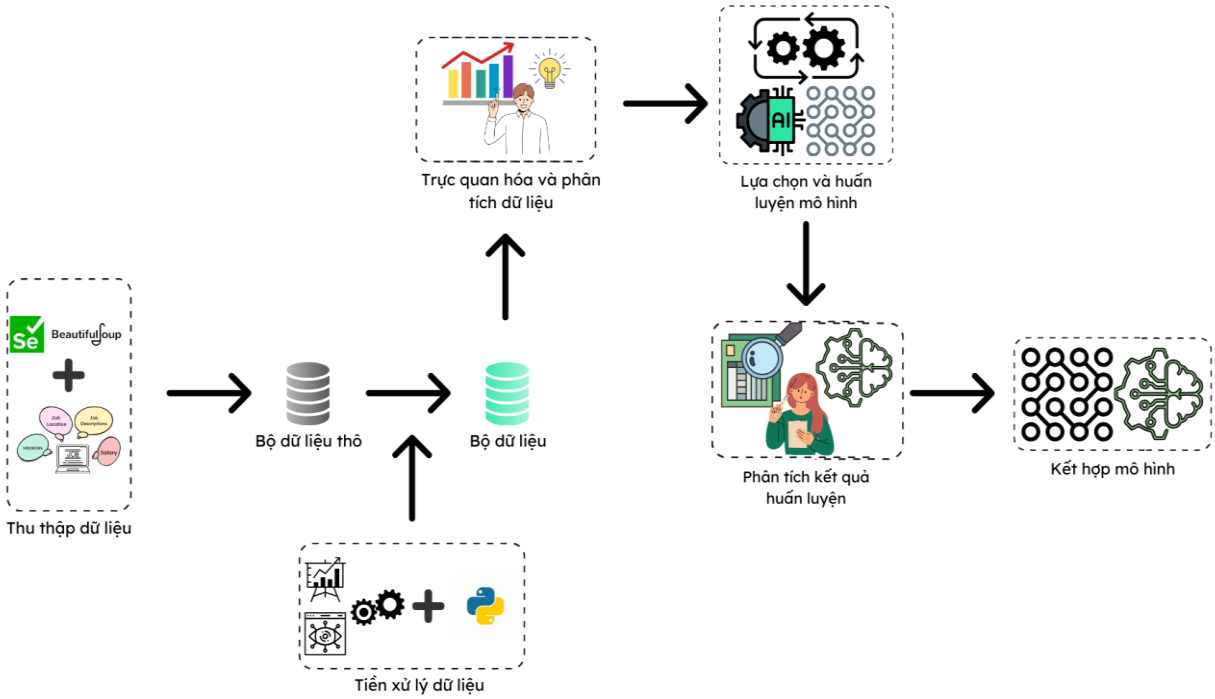
salaryType	Phương thức trả lương	Object
contractType	Loại hợp đồng làm việc	Object
jobType	Công việc tuyển dụng	Object
experience	Kinh nghiệm yêu cầu	Object
gender	Giới tính ứng viên	Object
education	Trình độ học vấn	Object
Skill	Kỹ năng chuyên môn	Object
Partner	Có là đối tác của vieclamtot.com hay không	Object
Description	Mô tả công việc	Object

Dữ liệu dạng số:

Tên thuộc tính	**Mô tả thuộc tính**	**Kiểu dữ liệu của thuộc tính**
vacacies	Số lượng cần tuyển	Int
minAge	Độ tuổi tối thiểu	Int
maxAge	Độ tuổi tối đa	Int

Sau khi loại bỏ các cột không cần thiết (link, companyName, title, Description) và chuẩn hóa một số cột, dữ liệu còn lại 11 biến phân loại và 4 biến số. Các giá trị bị khuyết lần lượt nằm ở cột experience (633 giá trị), education (628 giá trị), maxAge (202 giá trị). Các giá trị trống sau đó được nhóm sử dụng thuật toán KNN [6] để tiến hành điền khuyết theo hai hướng bao gồm điền khuyết dựa trên toàn bộ bộ dữ liệu và điền khuyết dựa theo mỗi loại nghề nghiệp.

3. TỔNG QUAN VỀ PHƯƠNG PHÁP PHÂN TÍCH



Hình 1: Quy trình xây dựng mô hình và phân tích trên bộ dữ liệu

Mô tả: Nhóm thu thập dữ liệu từ trang web, sau đó sẽ tiến hành làm sạch, chuẩn hóa, loại bỏ các cột không có giá trị và tiến hành điền khuyết. Sau đó nhóm bước vào quá trình thăm dò, phân tích dữ liệu thông qua các biểu đồ trực quan về sự phân bố của biến mục tiêu, sự phân bố của các thuộc tính dựa trên biến mục tiêu… Khi đã có cái nhìn tổng quan về các biến, nhóm sẽ dùng các phương pháp để tìm ra những thuộc tính ảnh hưởng nhất đến biến mục tiêu rồi tiến hành đưa vào mô hình huấn luyện và đánh giá. Cuối cùng lựa chọn ra tập hợp mô hình tốt nhất kết hợp thành mô hình Voting Regressor.

3.1. Xác định bài toán, thu thập và làm sạch dữ liệu:

Với các ứng viên thì ưu tiên cao nhất khi tìm việc luôn là mức lương, còn với các nhà tuyển dụng, họ sẽ cần đưa ra mức lương cho công việc bao nhiêu sẽ là phù hợp nhất. Nắm bắt tâm lý đó, nhóm quyết định xây dựng mô hình hồi quy dự đoán mức lương dựa trên các thuộc tính như kinh nghiệm, học vấn, địa điểm… Dữ liệu trên trang web sẽ có hai loại đặc biệt là nhà tuyển dụng đó có phải là đối tác của web vieclamtot.com hay không, việc một nhà tuyển dụng/công ty là đối tác sẽ khiến cho dữ liệu đáng tin cậy hơn. Nhóm sẽ thu thập dữ liệu của đối tác và không phải sau đó thêm một cột phân loại.

Bộ dữ liệu sau khi thu thập sẽ tiến hành làm sạch bằng cách loại bỏ đi những cột không có giá trị, loại bỏ nhiễu như chuẩn hóa những mức lương không hợp lý (VD: 6 triệu/giờ, 1000 triệu/tháng). Cột ‘Salary’ ban đầu là dạng phân loại với format như 6 – 12 triệu/ tháng, 800.000/giờ, nhóm sẽ tiến hành tách số và biến cột ‘Salary’ sang dạng số. Trong bộ dữ liệu có tồn tại khoảng lương như “Từ 6 triệu - 12 triệu/tháng”, nhóm sẽ

tiến hành tách số và lấy trung bình của khoảng này làm giá trị. Một số cột không có tính khai thác như ‘companyName’, ‘index’, ‘link’, ‘title’ sẽ bị loại bỏ khỏi bộ dữ liệu.

Cột ‘skill’ sẽ bao gồm những yêu cầu về kỹ năng chuyên môn nhưng cũng có những ô trống do nhà tuyển dụng không yêu cầu. Nhóm sẽ chuẩn hóa cột này về dạng nhị phân nếu có bất cứ yêu cầu từ nhà tuyển dụng thì giá trị sẽ là ‘Có’ còn nếu bỏ trống thì sẽ được điền là ‘Không’.

Dữ liệu sau khi loại bỏ các cột không liên quan và chuẩn hóa sẽ được điền khuyết, nhóm sẽ dùng thuật toán KNN [6] tiến hành điền khuyết. Các cột bị khuyết bao gồm ‘education’, ‘experience’, ‘maxAge’, vì mỗi ngành nghề sẽ có mức độ về kinh nghiệm hay độ tuổi tối đa khác nhau nên nhóm sẽ tiến hành điền khuyết theo từng nhóm nghề bằng cách chia tập dữ liệu thành nhiều tập nhỏ theo loại ngành nghề rồi áp dụng KNN [6] để tiến hành điền khuyết. Bên cạnh đó, nhóm cũng sẽ điền khuyết bằng KNN [6] cho toàn bộ tập dữ liệu mà không chia theo từng loại ngành nghề để so sánh với tập dữ liệu được điền khuyết theo từng loại ngành nghề.

3.2. Phân tích thăm dò dữ liệu:

Phân tích biến mục tiêu (Salary): Nhóm sẽ phân tích phân phối của Salary thông qua các biểu đồ như histogram để hiểu rõ hơn về phạm vi và xu hướng của dữ liệu lương.

Phân tích các thuộc tính khác: Khám phá mối quan hệ giữa Salary và các biến khác như ‘experience’, ‘education’, và ‘location’ thông qua biểu đồ boxplot, violin plot và các loại biểu đồ phân tích khác.

3.3. Lựa chọn mô hình, phân tích và đánh giá kết quả

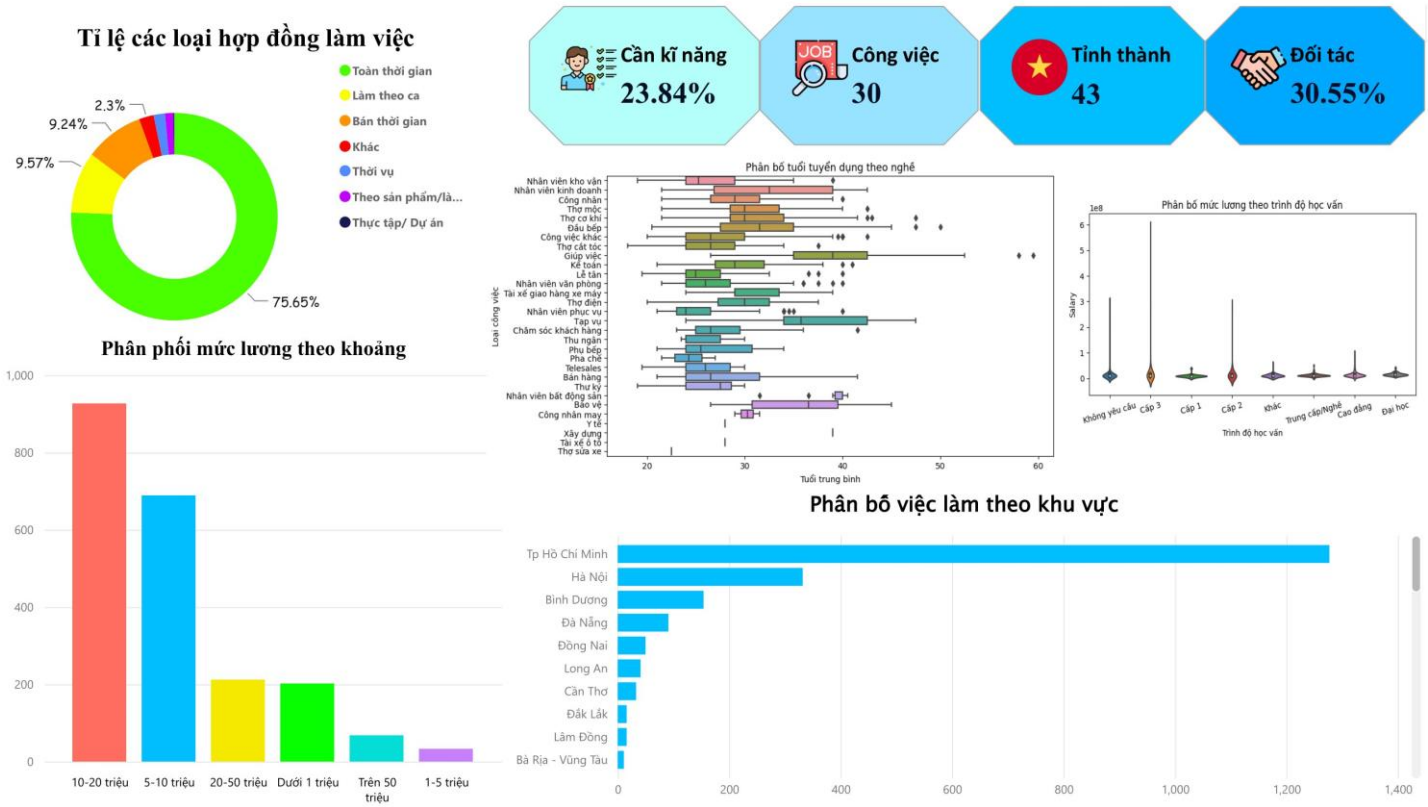
Phân tích tính tương quan: Để có thể đạt được kết quả tốt nhất cho mô hình, nhóm sẽ sử dụng các phương pháp phân tích để tìm ra thuộc tính nào có mối quan hệ tương quan tới biến mục tiêu (Salary). Đối với biến dạng số, nhóm sẽ dùng phương pháp pearsonr [7] tìm ra 2 giá trị p và r với p-value < 0.05 và | r | > 0.3. Với biến phân loại, nhóm sẽ dùng phương pháp ANOVA với p-value < 0.05.

- Các mô hình nhóm dự tính sử dụng sẽ từ những mô hình hồi quy cơ bản như Linear Regression, Ridge Regression, Lasso Regression cho tới những mô hình phức tạp hơn như Random Forest, Extra Trees… [6]. Bên cạnh đó, nhóm sử dụng GridsearchCV[6] để tìm ra siêu tham số phù hợp cho mỗi mô hình. Sau khi thực nghiệm tìm ra những mô hình tốt nhất với tham số tốt nhất của mỗi mô hình để đưa vào mô hình Voting Regressor [6] để tạo ra mô hình mạnh mẽ hơn.
- Vì bộ dữ liệu khá nhỏ (2138 dòng) nhóm sẽ sử dụng phương pháp đánh giá chéo k-fold validation [6] để có cái nhìn tin cậy nhất về hiệu suất và giảm độ bias của mô hình.
- Sau khi xây dựng các mô hình, nhóm tiến hành đánh giá chúng dựa trên các chỉ số như R2 Score và Mean Squared Error (MSE), cùng với việc phân tích kỹ lưỡng sự phù hợp của mỗi mô hình với dữ liệu. R2 Score để đánh giá mức độ mô hình giải thích được sự biến thiên của dữ liệu. Một R2 Score cao sẽ

cho thấy mô hình phù hợp tốt với dữ liệu. MSE dùng đo lường sai số trung bình bình phương, giúp nhận diện mô hình có xu hướng dự đoán chính xác hơn.

- Nhóm lựa chọn các mô hình có kết quả tốt nhất rồi tiến hành kết hợp theo dạng soft voting, nghĩa là tính trung bình hoặc trọng số kết quả dự đoán của các mô hình để đưa ra dự đoán cuối cùng. Các mô hình có độ tin cậy cao được đánh giá cao hơn trong quá trình này.

4. PHÂN TÍCH THĂM DÒ/SƠ BỘ



Hình 2: Dashboard tổng hợp các biểu đồ khi thăm dò dữ liệu, bao gồm 5 biểu đồ và những thông tin thống kê ở bộ dữ liệu

Biểu đồ 1: Tỉ lệ các loại hợp đồng làm việc

- Biểu đồ này cho thấy tỉ lệ các loại hợp đồng làm việc được đăng tuyển trên vieclamtot.com trong năm 2023. Theo đó, hợp đồng toàn thời gian chiếm tỉ lệ cao nhất, với 75.65%. Tiếp theo là hợp đồng làm theo ca với 9.57%. Hợp đồng thực tập/dự án và theo sản phẩm/làm tại nhà chiếm tỉ lệ thấp nhất, lần lượt là 0.14% và 1.31%.

Biểu đồ 2: Phân bố tuổi tuyển dụng theo nghề

- Biểu đồ này thể hiện phân bố tuổi tuyển dụng theo nghề. Theo đó, dữ liệu cho thấy một xu hướng chung là độ tuổi tuyển dụng ở các nghề đều có xu hướng trẻ hóa, ta có thể thấy điều này qua min của biểu đồ boxplot, tất cả các nghề đều có mức tuyển dụng ở độ tuổi 18-20.
- Xét về độ tuổi trung bình nghề giúp việc có độ tuổi tuyển dụng trung bình cao nhất, với 39.14 tuổi. Nghề giúp việc không yêu cầu kinh nghiệm làm việc chuyên môn cao. Do đó, những người mới bắt đầu tìm việc làm giúp việc thường ở độ tuổi cao hơn, chủ yếu là những người lớn tuổi di chuyển lên thành phố kiếm việc làm.
- Nghề pha chế có độ tuổi tuyển dụng trung bình thấp nhất, với 24.25 tuổi. Điều này có thể được giải thích bởi: nghề pha chế không yêu cầu trình độ học vấn cao, nghề pha chế đòi hỏi kỹ năng thực hành, khả năng sáng tạo và giao tiếp tốt. Những phẩm chất này thường được tìm thấy ở những người trẻ tuổi.

Biểu đồ 3: Phân bố việc làm ở các tỉnh thành

- Bản đồ cho chúng ta thấy phân bố việc làm của bộ dữ liệu trong cả nước, dữ liệu công việc trải dài cả 3 miền Bắc – Trung – Nam Việt Nam, tuy nhiên tập trung nhiều nhất vẫn là ở miền Nam, đặc biệt là thành phố Hồ Chí Minh và Bình Dương. Hà Nội chiếm tỉ lệ việc làm cao nhất ở miền Bắc nhưng mức độ vẫn kém hơn nhiều so với thành phố Hồ Chí Minh. Ở miền Trung, thành phố Đà Nẵng có lượng công việc nhiều nhất. Có thể thấy đa số nhu cầu tuyển dụng đều nằm ở các thành phố lớn trong cả nước.

Biểu đồ 4: Phân phối mức lương theo khoảng

- Biểu đồ này cho thấy phân phối mức lương theo khoảng khác nhau. Theo đó, mức lương từ 10 – 20 triệu có nhiều giá trị nhất trong bộ dữ liệu, theo sau là mức lương từ 1 đến 5 triệu. Mức lương từ 1 đến 5 triệu có ít giá trị nhất trong bộ dữ liệu vì đây thường là các công việc bán thời gian hoặc thời vụ.

Biểu đồ 5: Phân bố mức lương theo trình độ học vấn

- Những người có trình độ " cấp 3" có phạm vi lương rộng nhất, với một số outliers có mức lương rất cao. Trình độ "không yêu cầu" và "cấp 1", “cấp 2” có sự chênh lệch lớn giữa các mức lương, nhưng không có outliers nào quá cao như ở nhóm "cấp 3". Nhóm "khác" và "nghề" có mức lương tương đối thấp và đồng đều. Nhóm "cao đẳng" và "đại học" có mức lương thấp nhất trong tất cả các nhóm, nhưng lại có phạm vi phân phối lương khá hẹp, điều này cho thấy sự đồng đều trong thu nhập của nhóm này.
- Mức lương không hoàn toàn tăng theo trình độ học vấn. Điều này có thể do nhiều yếu tố khác như kinh nghiệm làm việc, ngành nghề, hoặc thậm chí là tính chất công việc đặc thù không cần bằng cấp cao. Mức lương cao nhất

không thuộc về nhóm có trình độ học vấn cao nhất, điều này chỉ ra rằng có những cơ hội kiếm tiền tốt không cần bằng cấp đại học. Sự chênh lệch lớn về mức lương trong một số nhóm có thể chỉ ra rằng có sự phân hóa về thu nhập trong cùng một trình độ học vấn.

Về tổng thể, dashboard này cho thấy thị trường việc làm trong năm 2023 khá sôi động, với nhiều cơ hội việc làm ở các lĩnh vực khác nhau (30 ngành nghề riêng biệt). Mức lương đa phần ở mức trung bình hoặc trung bình thấp cho thấy mức độ đòi hỏi của nhà tuyển dụng trên web vieclamtot.com với ứng viên về kinh nghiệm hay học vấn không quá cao với chỉ 23.84% cần kĩ năng chuyên môn. Mặt khác nhu cầu tuyển dụng việc làm thường chỉ tập trung ở các thành phố lớn, đông dân cư đi kèm với đó là mức độ trẻ hóa về độ tuổi lao động.

5. KẾT QUẢ PHÂN TÍCH

5.1/ Kết quả của tổng thể dữ liệu:

Kết quả phân tích sẽ bao gồm 2 bảng trước và sau khi bỏ các thuộc tính thông qua phân tích ảnh hưởng bằng ANOVA và pearsonr [7]. Các biến sau khi chạy pearsonr và anova được cho không ảnh hưởng tới biến mục tiêu gồm: 'minAge', 'maxAge', 'vacacies', 'skill'. Các biến được giữ lại bao gồm: ‘location’, ‘companyType’, ‘salaryType’, ‘contractType’, ‘jobType’, ‘experience’, ‘gender’, ‘education’, ‘Partner’. Chúng tôi còn muốn phân tích điền khuyết ảnh hưởng như thế nào đến kết quả mô hình nên sẽ có hai phiên bản dữ liệu (điền khuyết bằng cách groupby theo jobType rồi dùng KNN và điền khuyết bằng cách cho tất cả các thuộc tính rồi dùng KNN)

Kết quả của tập dữ liệu điền khuyết theo gom nhóm jobType rồi qua KNN

Mô hình	Trước khi bỏ các thuộc tính		Sau khi bỏ các thuộc tính	
	MSE	R2 Score	MSE	R2 Score
Linear Regression	9.81E+14	0.2333	1.01E+15	0.2105
Ridge Regression	9.81E+14	0.2331	1.01E+15	0.2099
Lasso Regression	9.81E+14	0.2333	1.01E+15	0.2105
Decision Tree	4.04E+14	0.6845	8.35E+14	0.3475
AdaBoost	3.67E+14	0.7129	6.92E+14	0.4588
Gradient Boosting	4.99E+14	0.6101	7.20E+14	0.4368
XG Boost	2.90E+14	0.7734	6.90E+14	0.4605
LightGBM	4.60E+14	0.6403	**5.91E+14**	**0.5380**
Random Forest	2.85E+14	0.7772	6.10E+14	0.5232
Extra Trees	**2.71E+14**	**0.7881**	6.27E+14	0.5097
K-Nearest Neighbors	4.91E+14	0.6157	6.74E+14	0.4730

Dữ liệu điền khuyết bằng tất cả thuộc tính cho qua KNN:

Mô hình	Trước khi bỏ các thuộc tính		Sau khi bỏ các thuộc tính	
	MSE	R2 score	MSE	R2 score
Linear Regression	8.19E+14	0.2496	8.46E+14	0.2245

Ridge Regression	8.19E+14	0.2493	8.46E+14	0.2243
Lasso Regression	8.19E+14	0.2496	8.46E+14	0.2245
Decision Tree	3.90E+14	0.6428	6.03E+14	0.4476
AdaBoost	2.92E+14	0.7323	6.00E+14	0.4496
Gradient Boosting	2.85E+14	0.7386	5.73E+14	0.4748
XG Boost	2.57E+14	0.7648	5.90E+14	0.4594
LightGBM	3.04E+14	0.7215	**5.17E+14**	**0.5261**
Random Forest	2.81E+14	0.7426	5.28E+14	0.5163
Extra Trees	2.15E+14	0.8033	5.28E+14	0.5163
K-Nearest Neighbors	**1.96E+14**	**0.8203**	6.14E+14	0.4369

Kết quả của các mô hình cho thấy sự khác biệt rõ rệt giữa việc bỏ hay không các thuộc tính được phân tích là có ảnh hưởng tới biến mục tiêu. Ngạc nhiên là khi bỏ các thuộc tính trên, mô hình lại cho kết quả rất thấp, giảm từ 0.02 cho đến 0.3 với R2 score. Việc R2 score giảm cho thấy rằng, sau khi loại bỏ các thuộc tính, mô hình mới có khả năng giải thích ít hơn về sự biến đổi của mức lương. Các thuộc tính bị loại bỏ có thể không có mối quan hệ mạnh với mức lương theo phân tích ANOVA, nhưng chúng vẫn có thể cung cấp thông tin hữu ích đóng góp vào mô hình tổng thể, đặc biệt khi tương tác với các biến khác.

Trong một số trường hợp, các biến độc lập tương tác với nhau, và loại bỏ một biến có thể ảnh hưởng đến mức độ tương tác đó, làm giảm chất lượng của mô hình. Các mô hình như Random Forest và Extra Trees có thể phát hiện và tận dụng tốt các tương tác này, và do đó việc loại bỏ các biến có thể làm giảm hiệu quả của chúng.

Việc điền khuyết cũng ảnh hưởng đến kết quả khi có đến 7 trong số 10 mô hình trong quá trình điền khuyết bằng KNN khi cho tất cả thuộc tính cho kết quả tốt hơn kiểu điền khuyết còn lại. Tuy nhiên khi tính ANOVA để xem có sự khác biệt giữa hai phương pháp điền khuyết hay không, kết quả cho thấy p_value = 0.7 nên không có đủ bằng chứng kết luận có sự khác nhau giữa hai kiểu điền khuyết.

MSE ở mức rất cao, điều này có thể do dữ liệu Salary có sự biến thiên lớn và phạm vi rộng ở các mức lương [14 000 : 580 000 000].

Các mô hình tuyến tính (Linear, Ridge, Lasso Regression) có kết quả khá giống nhau cả về R2 và MSE, điều này có thể phản ánh rằng chúng phụ thuộc vào những tín hiệu tuyến tính tương tự từ dữ liệu. Mô hình Decision Tree và Extra Trees có R2 score tốt hơn các mô hình tuyến tính, nhưng vẫn giảm sau khi loại bỏ các thuộc tính, điều này cho thấy các thuộc tính có thể chứa thông tin không tuyến tính quan trọng. Mô hình AdaBoost, Gradient Boosting, XG Boost và LightGBM, những mô hình dựa trên boosting, cho thấy sự cải thiện so với mô hình tuyến tính, nhưng lại gặp sự giảm sút đáng kể sau khi loại bỏ các thuộc tính.

Các mô hình học máy cổ điển như Linear, Ridge và Lasso Regression cho kết quả tương tự nhau và rất thấp nên có vẻ không phù hợp với loại dữ liệu này.

Kết quả cho thấy việc loại bỏ các thuộc tính dựa trên kết quả phân tích ANOVA không phải lúc nào cũng mang lại kết quả tích cực cho mô hình học máy. Cần phải cân nhắc kỹ lưỡng trước khi thực hiện bất kỳ loại bỏ nào, đặc biệt là khi có những tương tác phức tạp không dễ dàng được phát hiện thông qua phân tích thống kê đơn giản.

Sau khi thực nghiệm và tìm ra được các mô hình và các siêu tham số tốt nhất, nhóm thực hiện kết hợp các mô hình tốt nhất theo kết quả của Bảng 2 thành mô hình Voting Regressor.

Kết quả mô hình Voting Regressor

Top	Mô hình được kết hợp	MSE	R2 score
2	M1 = (K-Nearest Neighbors, Extra Tree)	1.99E+14	0.8169
3	M2 = M1 + XGBoost	**1.75E+14**	**0.8395**
4	M3 = M2 + Random Forest	1.98E+14	0.8182
5	M4 = M3 + Gradient Boosting	2.12E+14	0.8055

Kết quả trên cho thấy kết hợp ba mô hình K-Nearest Neighbors, Extra Tree và XGBoost tốt nhất với R2 là 0.8395 cao hơn so với mô hình đơn lẻ tốt nhất K-Nearst Neighbors là 0.8203. Điều này cho thấy việc kết hợp giữa các mô hình với nhau có thể làm tăng độ chính xác việc dự đoán mức lương. Tuy nhiên khi kết hợp thêm nhiều mô hình hơn thì R2 score lại giảm xuống cho thấy mức độ mạnh yếu khác nhau của từng mô hình thành viên. Điều này cũng cho thấy việc cân nhắc kĩ khi lựa chọn mô hình kết hợp.

b/ Kết quả theo cột education:

Nhận thấy trong cột education có các giá trị outlier quá cao như ‘cấp 3’, ‘cấp 2’, ‘khác’ và việc các mức lương ở học vấn cấp 3 hay cấp 2 vượt quá đại học có thể là điều bất hợp lý. Nhóm sẽ chỉ giữ lại các dòng có giá trị education là ‘trung cấp/ nghề’, ‘cao đẳng’, ‘đại học’ sau đó loại cả những dòng có outlier trong bộ dữ liệu. Bộ dữ liệu được lọc chỉ còn 351 dòng.

Mô hình	Trước khi bỏ các thuộc tính		Sau khi bỏ các thuộc tính	
	MSE	R2 score	MSE	R2 score
Linear Regression	1.10E+13	0.2103	1.11E+13	0.2028
Ridge Regression	1.11E+13	0.2029	1.11E+13	0.1989
Lasso Regression	1.10E+13	0.2103	1.11E+13	0.2028
Decision Tree	1.58E+13	-0.1344	1.39E+13	-0.0027
AdaBoost	9.55E+12	0.3150	1.13E+13	0.1869
Gradient Boosting	**9.63E+12**	**0.3098**	1.19E+13	0.1444

XG Boost	1.08E+13	0.2257	**1.10E+13**	**0.2072**
LightGBM	1.28E+13	0.0840	1.28E+13	0.0823
Random Forest	1.02E+13	0.2664	1.14E+13	0.1826
Extra Trees	1.09E+13	0.2212	1.18E+13	0.1518
K-Nearest Neighbors	1.44E+13	-0.0292	1.31E+13	0.0540

Có thể thấy dù đã loại bỏ đi các giá trị có outlier quá cao nhưng việc bỏ các cột thông qua ANOVA vẫn cho kết quả thấp hơn so với việc giữ nguyên thuộc tính. Các mô hình cho kết quả rất thấp ở bộ dữ liệu này, thậm chí có những mô hình cho giá trị R2 âm. Điều này có thể giải thích do dữ liệu quá ít cũng như số lượng dữ liệu bị loại bỏ là quá lớn, đặc biệt là nếu nó đại diện cho một phần lớn của dữ liệu, có thể ảnh hưởng đến khả năng của mô hình học hỏi và dự đoán.

6.KẾT LUẬN

Trong dự án này, nhóm đã trải qua một quá trình toàn diện từ việc thu 2thập, làm sạch, và chuẩn hóa dữ liệu nhằm mục tiêu xây dựng mô hình học máy có khả năng dự đoán mức lương dựa trên nhiều thuộc tính như kinh nghiệm, học vấn, và địa điểm làm việc. Các bước tiền xử lý dữ liệu bao gồm việc chuyển đổi các giá trị mức lương từ dạng văn bản sang số, loại bỏ nhiễu, và xử lý dữ liệu khuyết thiếu thông qua thuật toán KNN. Để hiểu rõ hơn về dữ liệu, nhóm đã thực hiện phân tích thăm dò bằng cách sử dụng nhiều loại biểu đồ trực quan.

Kết quả từ việc đánh giá các mô hình học máy cho thấy một số điểm mấu chốt sau khi loại bỏ các thuộc tính dựa trên phân tích ANOVA có p-value > 0.05:

- Các mô hình tuyến tính không cho thấy sự cải thiện đáng kể sau khi loại bỏ các thuộc tính. Các mô hình phi tuyến, bao gồm cả mô hình tăng cường và cây quyết định, cho thấy sự giảm sút trong hiệu suất, điều này có thể được giải thích thông qua việc loại bỏ thông tin quan trọng. R2 score đa số ở mức khá cho thấy các mô hình giải thích được đáng kể sự biến thiên của dữ liệu mục tiêu.
- Dựa trên những phát hiện, có thể kết luận rằng việc lựa chọn và loại bỏ thuộc tính cần được thực hiện một cách cẩn thận để không loại bỏ các thông tin quan trọng có thể cải thiện hiệu suất của mô hình. Những phân tích ANOVA có p-value cao không nhất thiết chỉ ra rằng một thuộc tính là không cần thiết, đặc biệt nếu xem xét các tương tác giữa các thuộc tính.

Để tiếp tục cải thiện mô hình và kết quả dự đoán, nhóm có thể:

- Xem xét lại quá trình tiền xử lý và đánh giá các thuộc tính để đảm bảo rằng không có thông tin quan trọng nào bị loại bỏ.
- Áp dụng các kỹ thuật chọn lọc đặc trưng tự động và phân tích độ quan trọng của đặc trưng để có cái nhìn toàn diện hơn về đóng góp của từng thuộc tính.
- Cuối cùng, nhóm cần xem xét việc thực hiện thêm các thử nghiệm thêm nhiều mô hình và siêu tham số khác nhau để tìm ra phương án tối ưu nhất cho dữ liệu cụ thể này hoặc tăng cường dữ liệu áp dụng deep learning.

TÀI LIỆU THAM KHẢO

[1] Việc Làm Tốt. Link: https://www.vieclamtot.com/ (Ngày truy cập: 23/11/2023).

[2] Chợ Tốt. Link: http: https://www.chotot.com/ (Ngày truy cập: 22/11/2023).

[3] Beautiful Soup. Link: https://beautiful-soup-4.readthedocs.io/en/latest/ (Ngày truy cập 23/11/2023)

[4] Selenium. Link: https://www.selenium.dev/ (Ngày truy cập 23/11/2023)

[5] Microsoft Power BI. Link: https://www.microsoft.com/en-us/power-platform/products/power-bi/ (Ngày truy cập 23/11/2023)

[6] Scikit-learn. Link: https://scikit-learn.org/stable/index.html (Ngày truy cập 2/12/2023)

[7] SciPy. Link: https://docs.scipy.org/doc/scipy/index.html (Ngày truy cập 23/11/2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Lê Tuấn Hưng	Xây dựng slide, viết báo cáo, đề xuất mô hình, trực quan quy trình
2	Nguyễn Trọng Mạnh	Thu thập dữ liệu, xây dựng mô hình, trực quan hóa dữ liệu
3	Tô Trường Long	Làm sạch dữ liệu, viết báo cáo, trực quan hóa dữ liệu