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

PHÂN TÍCH VÀ DỰ ĐOÁN NĂNG SUẤT LAO ĐỘNG CỦA CÔNG NHÂN TRONG NGÀNH CÔNG NGHIỆP MAY MẶC

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Trường Thịnh	20520783
2	Nguyễn Minh Tâm	20520748

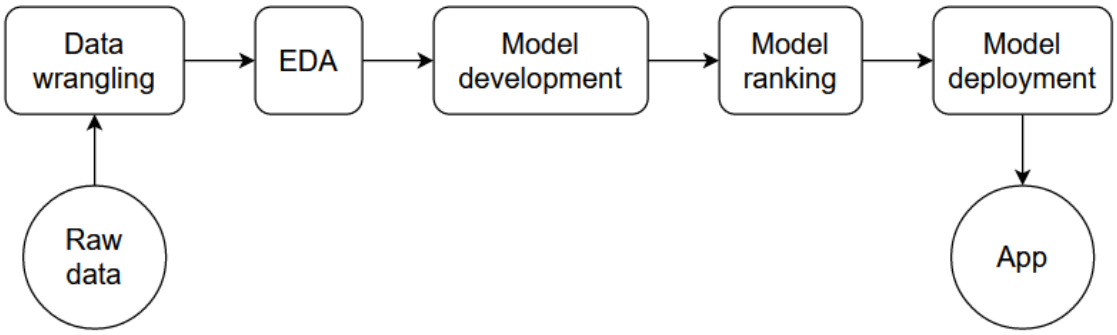
TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

Bài toán phân tích năng suất lao động của công nhân có ý nghĩa thực tiễn giúp các nhà quản lý nhận định các yếu tố cốt lõi có tác động đến hiệu quả làm việc của công nhân trong công ty, từ đó đưa ra được những giải pháp cải tiến phù hợp. Trong đồ án này, nhóm nghiên cứu sử dụng bộ dữ liệu Garment Employees Dataset được cung cấp bởi tác giả Abdullah Al Imran năm 2019 [1]. Bộ dữ liệu sơ bộ mô tả các yếu tố ảnh hưởng đến năng suất lao động thực tế của công nhân (actual productivity) trong ngành công nghiệp may mặc, được thu thập trong giai đoạn 01/01/2015 đến 11/03/2015 tại một công ty lớn ở Bangladesh. Dựa trên bộ dữ liệu, bằng phương pháp phân tích thăm dò (exploratory data analysis) và phân tích phương sai một yếu tố (oneway-ANOVA), nhóm đã đưa ra kết luận các biến độc lập nào có tương quan thực sự với biến đầu ra. Ở giai đoạn tiếp theo, nhóm tiến hành xây dựng một số mô hình dự đoán năng suất lao động của công nhân dựa trên các biến đầu vào, chủ yếu theo hướng tiếp cận của máy học (machine learning). Các mô hình xây dựng được đánh giá trên 3 thang đo là root mean squared error (RMSE), mean absolute error (MAE) và mean absolute percentage error (MAPE) nhằm lựa chọn ra mô hình tốt nhất. Cuối cùng, nhóm đã xây dựng một pipeline hoàn chỉnh để dự đoán và triển khai trên web app với sự hỗ trợ của thư viện streamlit. Ngôn ngữ lập trình được sử dụng trong đồ án là python.

2. NỘI DUNG

Các nội dung chính của đồ án được trình bày như trong Hình 1.



Hình 1. Quy trình phân tích dữ liệu và xây dựng mô hình dự đoán.

2.1. Bộ dữ liệu

Bộ dữ liệu *Productivity of Garment Employees* được thu thập trong giai đoạn từ 01/01/2015 đến 11/03/2015 tại một công ty may mặc ở Bangladesh. Tác giả của bộ dữ liệu gốc, ông Abdullah Al Imran, đã công bố chính thức trên UCI Machine Learning Repository vào năm 2019. Đường dẫn truy cập bộ dữ liệu như sau: https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees [1].

Bộ dữ liệu thô có 1197 mẫu với 15 thuộc tính, gồm cả thuộc tính dạng số và dạng phân loại. Trong đó, có duy nhất một thuộc tính bị khuyết giá trị là '*wip*', với 506 giá trị khuyết, chiếm 42,2 % số lượng mẫu. Mô tả chi tiết ý nghĩa các thuộc tính được trình bày trong Bảng 1 dưới đây.

Bảng 1. Mô tả chi tiết các thuộc tính trong bộ dữ liệu.

Thuộc tính	**Miền giá trị**	**Ý nghĩa**
date	1/1/2015 → 11/3/2015	Ngày thu thập dữ liệu, theo định dạng mm/dd/yyyy.
quarter	Quarter1, Quarter2, Quarter3, Quarter4, Quarter5	Một phần của tháng, cứ 7 ngày liên tiếp tạo thành một quarter. Quarter 1 tính từ ngày 1 của tháng.
department	sewing, finishing	Phòng ban làm việc của tổ công nhân.
day	Sunday, Monday, Tuesday, Wednesday, Thursday, Saturday	Thứ trong tuần.
team	1 → 12	Số thứ tự của tổ công nhân.
targeted_productivity	0.07 → 0.80	Năng suất lao động mục tiêu được đặt ra bởi người quản lý đối với mỗi tổ công nhân vào mỗi ngày.
smv	2.90 → 54.56	(standard minute value) Thời gian được phân bổ cho một tác vụ, tính theo phút.
wip	7 → 23122	(work in progress) Số lượng đồ chưa hoàn thành đối với một sản phẩm.
over_time	0 → 25920	Thời gian làm việc quá giờ của một tổ công nhân (tính theo phút).
incentive	0 → 3600	Tiền thưởng khuyến khích công nhân cho một công đoạn (tính theo BDT).
idle_time	0 → 300	Thời gian rảnh rỗi mà dây chuyền bị gián đoạn do một vài lý do nào đó (tính theo phút).
idle_men	0 → 45	Số lượng công nhân rảnh rỗi do dây chuyền bị gián đoạn.
no_of_style_change	0 → 2	Số lượng thay đổi quy cách đối với một sản phẩm nào đó.
no_of_workers	2 → 89	Số lượng công nhân trong mỗi tổ.
actual_productivity	0.23 → 1.12	Năng suất lao động thực tế ước tính.

2.2. Tiền xử lý dữ liệu

Ở nội dung này, nhóm tiến hành xử lý một vài dữ liệu bị nhầm lẫn hoặc sai kiểu dữ liệu và điền dữ liệu bị khuyết.

Đối với thuộc tính '*department*' có kiểu phân loại, sau khi kiểm tra các giá trị unique thì nhận thấy có 3 giá trị là 'sweing', 'finishing', 'finishing '. Giá trị 'sweing' sai lỗi chính tả nên được sửa lại thành 'sewing', giá trị 'finishing ' vốn dĩ trùng với 'finishing' nhưng do thừa khoảng trắng ở cuối nên cần điều chỉnh cho khớp.

Đối với thuộc tính '*team*' khi load dữ liệu bằng thư viện pandas thì tự động chuyển sang kiểu int64, do miền giá trị của biến này là các số từ 1 đến 12. Để tránh nhầm lẫn khi phân tích, nhóm đặt lại các giá trị này theo kiểu biến phân loại từ 'Team1' đến "Team12'.

Đối với thuộc tính '*wip*', các giá trị bị khuyết sẽ được điền theo phương pháp k-Nearest Neighbor dựa trên tất cả các dữ liệu kiểu số. Nhóm sử dụng class KNNImputer() mặc định được hỗ trợ bởi thư viện scikit-learn. Sau khi xử lý, các biến kiểu số đều được trả về kiểu float64.

Tất cả các tác vụ tiền xử lý này đều được đóng gói trong một hàm data_loader().

2.3. Phân tích thăm dò dữ liệu

A) Thống kê mô tả trên bộ dữ liệu

Đối với các biến kiểu số, tiến hành tính toán một số giá trị thống kê đặc trưng như trung bình, trung vị, min, max, các tứ phân vị, độ lệch chuẩn. Đối với các biến kiểu phân loại, tiến hành xác định số loại giá trị, mode, tần số của mode. Kết quả được trình bày ở Bảng 2 và Bảng 3.

Bảng 2. Thống kê mô tả các biến kiểu số.

Biến	mean	std	min	Q1	med	Q3	max
targeted_productivity	0.73	0.10	0.07	0.70	0.75	0.80	0.80
smv	15.06	10.94	2.90	3.94	15.26	24.26	54.56
wip	1069.04	1414.80	7	773	983	1119	23122
over_time	4567.46	3348.82	0	1440	3960	6960	25920
incentive	38.21	160.18	0	0	0	50	3600
idle_time	0.73	12.71	0	0	0	0	300
idle_men	0.37	3.27	0	0	0	0	45
no_of_style_change	0.15	0.43	0	0	0	0	2
no_of_workers	34.61	22.20	2	9	34	57	89
actual_productivity	0.74	0.17	0.23	0.65	0.77	0.85	1.12

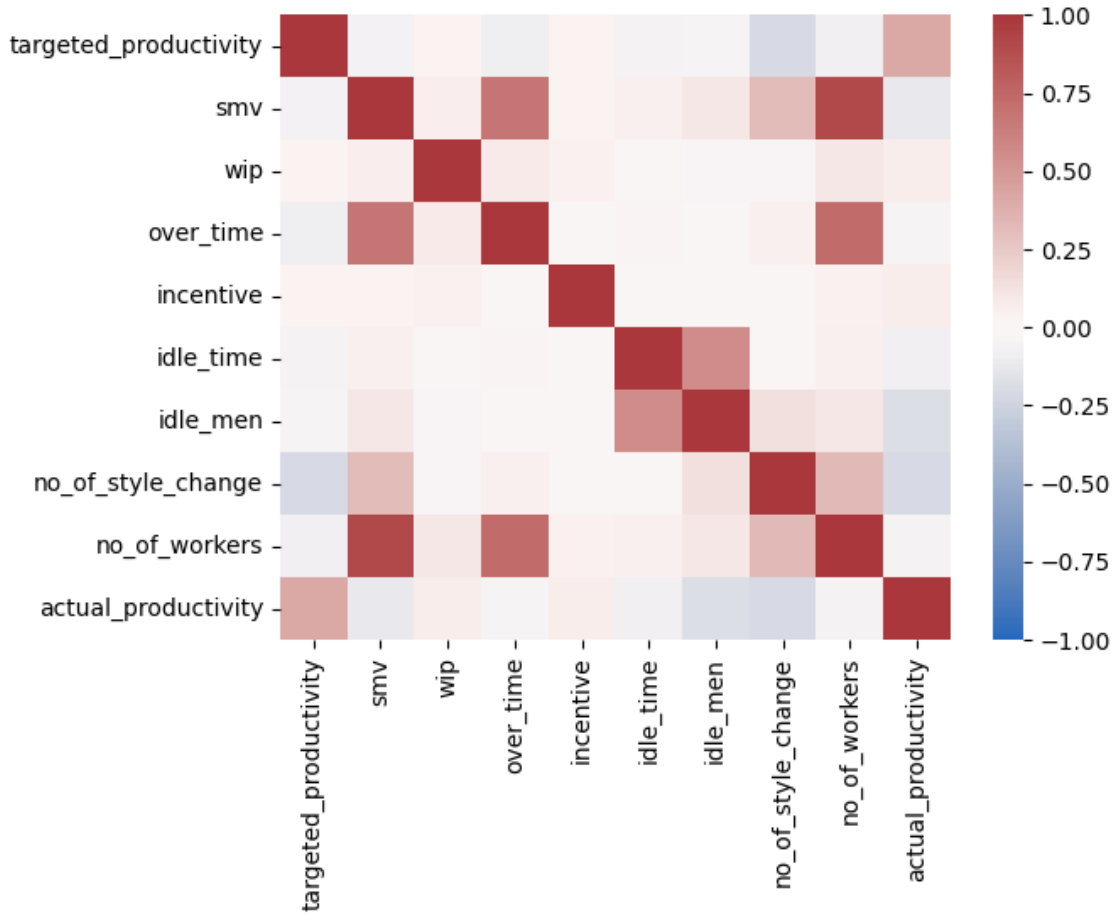
Bảng 3. Thống kê mô tả các biến kiểu phân loại.

Biến	unique	top	freq
quarter	5	Quarter1	360
department	2	sewing	691
day	6	Wednesday	208
team	12	Team8	109

Thống kê các đặc trưng phân phối của biến phụ thuộc '*actual_productivity*' cho thấy skewness = –0.81 < 0 nên có phân phối lệch sang trái, đồng thời kurtosis = 0.33 < 3 nên có đỉnh thấp hơn phân phối chuẩn, ít bị ảnh hưởng của giá trị ngoại lệ.

B) Phân tích thăm dò các biến kiểu số

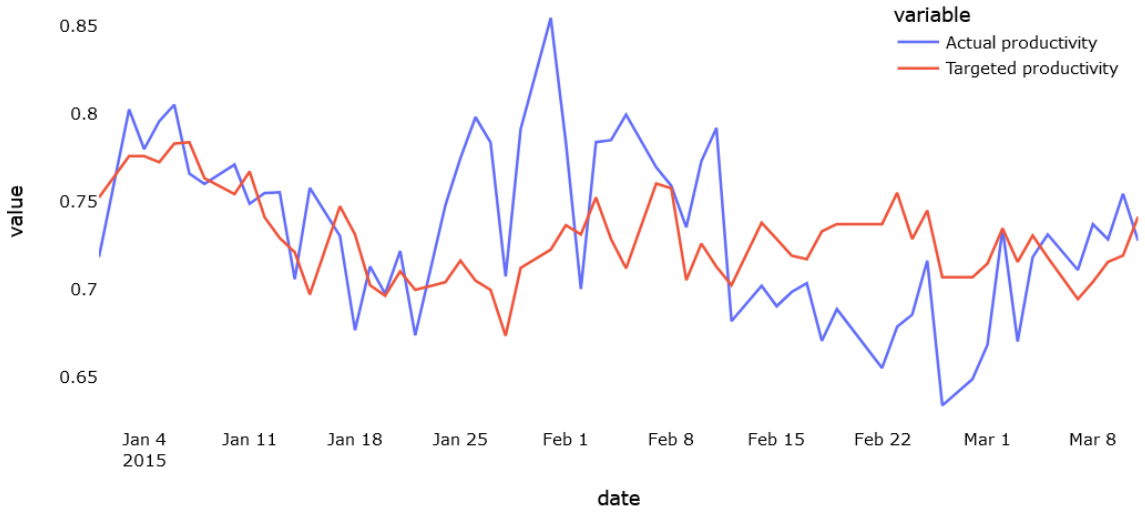
Phương pháp được áp dụng là tính toán mức độ tương quan giữa các biến bằng thống kê Pearson. Kết quả phân tích tương quan được trình bày ở Hình 2.



Hình 2. Mức độ tương quan giữa các biến kiểu số.

Biểu đồ trên cho thấy các biến kiểu số hầu như không có tương quan nào rõ rệt với biến mục tiêu '*actual_productivity*', ngoại trừ biến '*targeted_productivity*' cho tương quan thuận mức độ yếu (r = 0,42) và độ tin cậy khá chắc chắn (p-value ≈ 0). Điều này cũng đúng trong thực tế, vì khi một target đã được chỉ định bởi người quản lý thì tất cả

các tổ công nhân phải cố gắng để đạt mục tiêu đặt ra. Hình 3 trực quan mối liên hệ giữa *'actual_productivity'* với *'targeted_productivity'* theo thời gian.



Hình 3. So sánh 'actual_productivity' với 'targeted_productivity' theo thời gian.

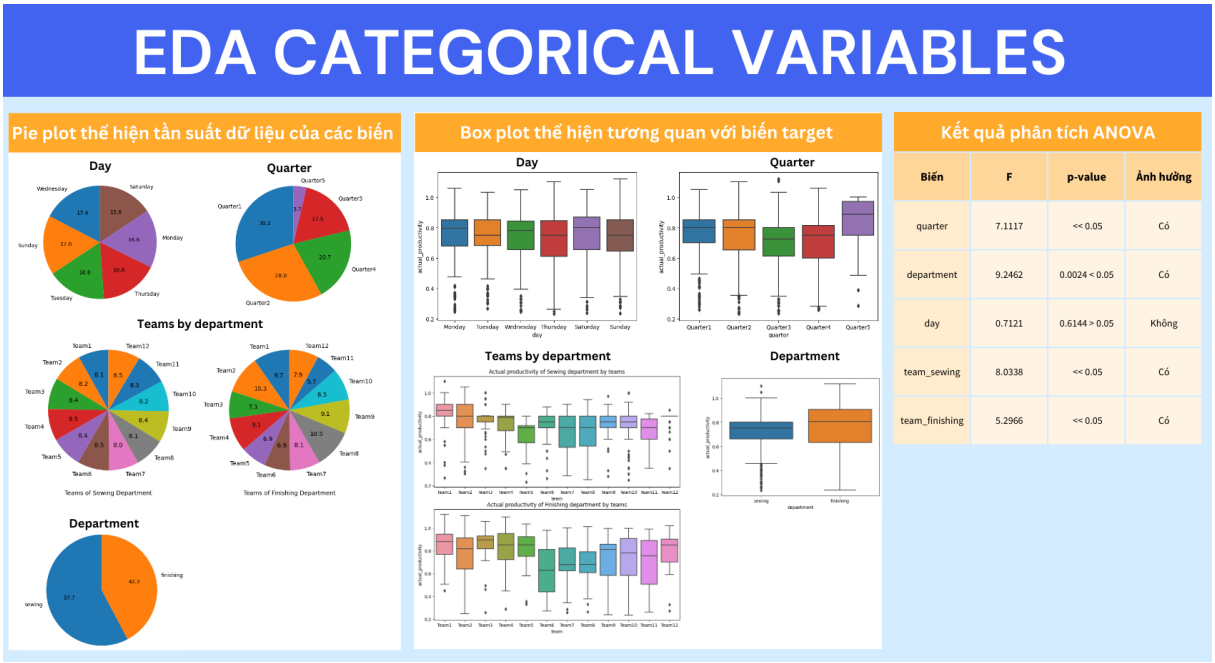
Quan sát các biến khác, nhận thấy có sự tương quan thuận mức độ từ trung bình đến mạnh trong nhóm biến '*smv*', '*over_time*', '*no_of_workers*' theo từng cặp. Điều này có thể giải thích là do, khi một tác vụ có lượng thời gian phân bổ nhiều, chứng tỏ rằng tác vụ đó có độ phức tạp cao và cần nhiều công nhân xử lý, và có thể phải làm thêm giờ để hoàn thành tác vụ đó. Ngoài ra còn có tương quan thuận yếu giữa '*no_of_style_change*' với '*smv*' và '*no_of_worker*', có thể giải thích rằng khi xuất hiện sự thay đổi trong quy cách sản xuất, tác vụ đó có xu hướng sẽ được phân bổ thời gian nhiều hơn để công nhân làm quen với quy cách mới, và có nhiều công nhân thực hiện hơn để được đào tạo theo quy cách mới. Tương quan trung bình giữa '*idle_time*' với '*idle_men*' là hiển nhiên, vì thời gian rảnh và công nhân rảnh xuất hiện cùng một lúc một khi dây chuyền sản xuất gặp sự cố và bị gián đoạn.

C) Phân tích thăm dò các biến kiểu phân loại

Có 4 biến kiểu phân loại trong bộ dữ liệu là '*quarter*', '*department*', '*day*' và '*team*'. Kết quả phân tích thăm dò các biến này được minh họa ở Hình 4.

Tần suất các phân loại trong biến '*day*' và '*team*' tương đối đều nhau. Ở biến '*quarter*' có loại 'Quarter5' xuất hiện rất ít. Điều này là do 'Quarter5' rơi vào các ngày 29-30-31/1/2015, còn các ngày từ 1-28 đã được chia đều vào 4 quarter trước, tháng 2 chỉ có 28 ngày nên không có 'Quarter5', còn tháng 3 chỉ thu thập đến ngày 11. Ở biến '*department*', dữ liệu của phòng 'sewing' nhiều hơn phòng 'finishing', có thể vì công đoạn 'sewing' tốn nhiều nhân công hơn.

Tiếp theo, để đánh giá một cách khách quan sự ảnh hưởng của các biến phân loại đến biến mục tiêu '*actual_productivity*', nhóm triển khai phương pháp phân tích phương sai một yếu tố (oneway-ANOVA), kiểm định xác suất giả thiết không với mức ý nghĩa $\alpha = 5\%$. Kết quả cho thấy có 3 biến thực sự có ảnh hưởng đến biến mục tiêu là '*quarter*', '*department*' và '*team*'.

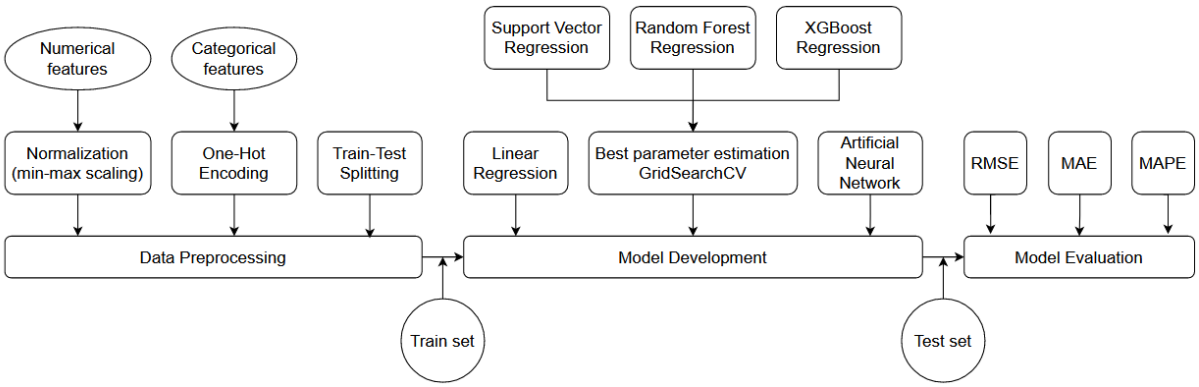


Hình 4. Tóm tắt kết quả phân tích thăm dò các biến kiểu phân loại.

Để có những insight chi tiết hơn, nhóm tiến hành tiếp phân tích hậu nghiệm (post-hoc analysis) bằng phương pháp TukeyHSD. Kết quả cho thấy ở biến '*quarter*', công nhân có xu hướng làm việc năng suất hơn vào thời điểm đầu tháng (Quarter1). Trong 2 phòng ban, công nhân ở phòng ban '*finishing*' nhìn chung có năng suất lao động cao hơn công nhân ở phòng ban '*department*'. Ở cả hai phòng ban, team số 1 và số 3 nhìn chung có năng suất làm việc tốt nhất, trong khi team số 7 và số 8 có năng suất thấp nhất.

2.4. Xây dựng mô hình dự đoán

Từ kết quả phân tích thăm dò, nhóm nhận thấy chỉ có 1 biến kiểu số có tương quan yếu và 3 biến kiểu phân loại có ảnh hưởng đến biến đầu ra. Do đó, nếu chỉ sử dụng các biến này để xây dựng mô hình thì dễ dẫn đến hiện tượng chưa khớp (underfitting) do quá ít thuộc tính. Vì vậy, nhóm vẫn giữ tất cả các biến để xây dựng mô hình dự đoán, ngoại trừ biến '*date*' là ngày/tháng/năm sẽ không được đưa vào mô hình. Các bước xây dựng mô hình được minh họa ở Hình 5.



Hình 5. Các bước xây dựng mô hình.

A) Tiền xử lý dữ liệu

Các biến kiểu số được normalization bằng class MinMaxScaler() của sklearn để đưa về miền giá trị [0, 1]. Đối với các biến kiểu phân loại, nhóm áp dụng class OneHotEncoder() để chuyển kiểu phân loại về dạng số có thể đưa được vào mô hình. Sau khi thực hiện, dữ liệu mới gồm có 34 thuộc tính. Tiến hành phân chia 2 tập huấn luyện và tập kiểm thử theo tỉ lệ 8 : 2, trong đó tập train chứa 957 mẫu dữ liệu dùng để phát triển mô hình, tập test chứa 240 mẫu dữ liệu dùng cho việc đánh giá mô hình.

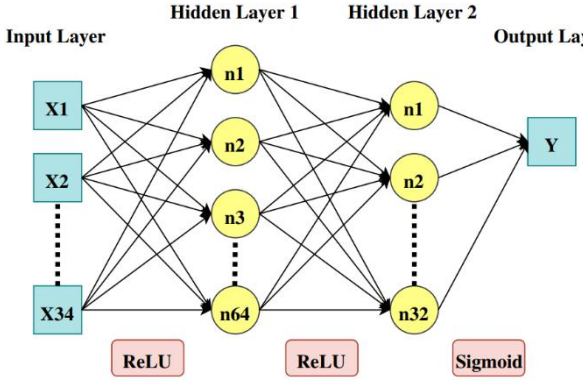
B) Phát triển mô hình

Mô hình đầu tiên được xây dựng là Linear Regression đa biến. Sau khi huấn luyện trên tập train, kết quả cho thấy R^2 tương đối thấp (0.3051), đây là dấu hiệu của underfitting. Nhóm tiếp tục lựa chọn những mô hình nâng cao hơn như Support Vector Regression, các mô hình ensemble như Random Forest và XGBoost. Để chọn bộ siêu tham số (hyperparameters) tốt nhất, nhóm thực hiện chiến lược GridSearchCV trên tập train với số fold k = 5 và lựa chọn mô hình dựa trên scoring = ‘r2’. Các siêu tham số tìm thấy cho từng mô hình được miêu tả trong Bảng 4.

Bảng 4. Các siêu tham số được đề xuất cho các mô hình và kết quả tìm kiếm.

Mô hình	**Các siêu tham số tìm kiếm**	**Kết quả chọn**
Support Vector Regression	'kernel': ['linear', 'poly', 'rbf', 'sigmoid']	‘linear’
	'C': [0.01, 0.1, 1.0, 10]	10
	'epsilon': [0.02, 0.05, 0.1, 0.2]	0.2
Random Forest Regression	'n_estimators': [10, 20, 50, 100]	100
	'criterion': ['squared_error', 'absolute_error', 'friedman_mse', 'poisson']	‘absolute_error’
	'max_features': ['sqrt', 'log2', 1.0]	‘log2’
XGBoost Regression	'n_estimators': [10, 20, 50, 100]	20
	'max_depth': [3, 4, 5, 6, 7, 8]	3

Ngoài ra nhóm cũng xây dựng một mô hình theo hướng học sâu là Artificial Neural Network với kiến trúc như được đề xuất bởi chính tác giả [2]. Mô hình gồm 2 hidden layer với số node lần lượt là 64 và 32. Kiến trúc mô hình được minh họa ở Hình 6.



Hình 6. Kiến trúc mô hình ANN.

C) Đánh giá mô hình

Các mô hình sau khi được huấn luyện trên tập train sẽ được đánh giá trên tập test theo 3 độ đo đặc trưng của bài toán hồi quy là: RMSE (root mean squared error), MAE (mean absolute error) và MAPE (mean absolute percentage error). Kết quả đánh giá các mô hình được ghi nhận trong Bảng 5.

Bảng 5. Đánh giá các mô hình đã xây dựng trên tập kiểm thử.

Mô hình	**RMSE**	**MAE**	**MAPE**
Linear Regression	0.142	0.106	0.185
Support Vector Regression	0.142	0.104	0.187
Random Forest Regression	0.119	0.081	0.146
XGBoost Regression	0.124	0.083	0.148
Artificial Neural Network	0.145	0.100	0.181

Từ kết quả trên, nhóm nhận thấy 2 mô hình cho kết quả tương đối tốt trên cả 3 thang đo là Random Forest Regression và XGBoost Regression. Các mô hình còn lại cho kết quả tương đương nhau. Từ đó, nhóm lựa chọn mô hình Random Forest để xây dựng ứng dụng đưa ra dự đoán kết quả '*actual_productivity*' dựa trên dữ liệu nhập từ người dùng.

2.5. Xây dựng ứng dụng dự đoán

Mô hình Random Forest được kết hợp với các bộ tiền xử lý biến đổi các thuộc tính do người dùng nhập vào nhằm tạo thành 1 pipeline. Pipeline sau đó được huấn luyện và lưu lại trọng số bằng thư viện pickle. Nhóm tiến hành xây dựng ứng dụng dự đoán bằng thư viện streamlit. Để deploy ứng dụng lên web app, nhóm tạo một repository trên github để lưu trữ các dữ liệu cần thiết. Sau đó kết nối streamlit với repo và deploy ứng dụng lên web. Giao diện của ứng dụng như Hình 7.

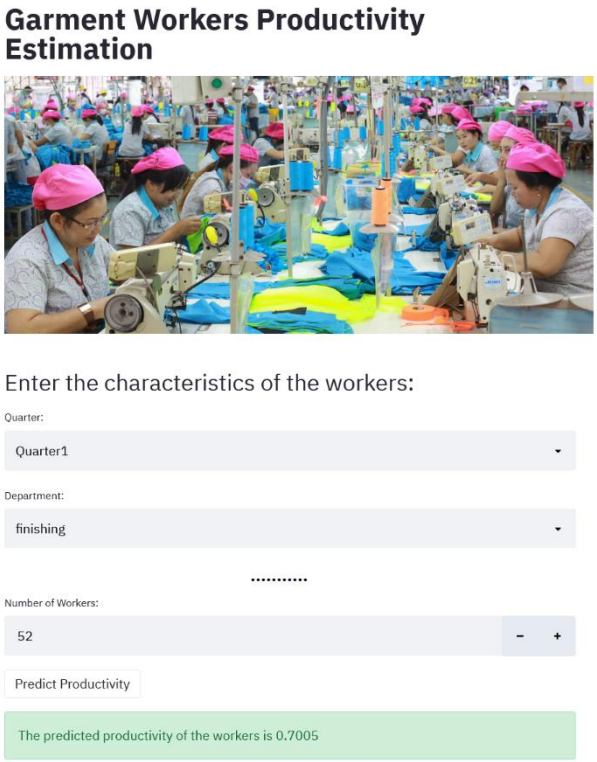
Link github: https://github.com/Tam1032/Garment-Employee-Productivity .

Link ứng dụng: https://thinhnt19393-garment-employee-productivity-predictor-qvnld4.streamlit.app/ .

3. KẾT LUẬN

Trong đồ án, nhóm đã sử dụng bộ dữ liệu *Productivity of Garment Employee* để tiến hành phân tích các yếu tố ảnh hưởng đến năng suất lao động của công nhân trong ngành công nghiệp may mặc và xây dựng mô hình dự báo. Bộ dữ liệu thô đã được tiến hành tiền xử lý dữ liệu khuyết bằng phương pháp kNN, đồng thời điều chỉnh một số giá trị lỗi và chuyển kiểu dữ liệu cho phù hợp với mục tiêu phân tích.

Quá trình phân tích thăm dò (EDA) cho thấy hầu như không có biến kiểu số nào có tương quan với biến mục tiêu '*actual_productivity*', ngoại trừ duy nhất biến '*targeted_productivity*'. Các mối tương quan rõ hơn của các biến khác đều có thể giải thích được trong bối cảnh thực tế. Để xem xét mối ảnh hưởng của các biến kiểu phân loại, nhóm sử dụng phương pháp oneway ANOVA. Kết quả cho thấy 3 trên 4 biến có ảnh hưởng đến biến đầu ra là '*quarter*', '*department*' và '*team*'.



Hình 7. Giao diện của ứng dụng dự đoán.

Từ kết quả EDA, nhóm nhận thấy số lượng biến có ảnh hưởng đến biến đầu ra khá ít. Do đó, tất cả các biến ngoại trừ biến '*date*' đều được giữa lại để xây dựng mô hình. Trước khi huấn luyện, các biến kiểu số được chuẩn hóa min-max và biến kiểu phân loại được biểu diễn về dạng vector one-hot. Sau đó, huấn luyện trên tập train với 5 mô hình là Linear Regression, SVR, Random Forest, XGBoost và ANN. Kết quả đánh giá trên 3 thang đo RMSE, MAE và MAPE cho thấy có 2 mô hình vượt trội hơn hẳn là Random Forest và XGBoost. Mô hình Random Forest lưu lại và đóng gói với các bộ tiền xử lý trong một pipeline. Dựa trên pipeline này, nhóm đã xây dựng một ứng dụng đơn giản cho phép người dùng nhập dữ liệu và đưa ra dự đoán bằng streamlit.

Qua quá trình thực hiện, nhóm nhận thấy đồ án còn tồn tại một số điểm hạn chế. Đầu tiên, các mô hình dự đoán cho kết quả sai số còn cao, chênh lệch khoảng 15% so với giá trị thực tế. Mặt khác, nhóm chỉ xem xét biến đầu ra dựa trên các biến khác mà chưa khảo sát đến yếu tố chuỗi thời gian (time series) của bộ dữ liệu. Vì ngoài sự tương quan với các biến độc lập, biến đầu ra trong dữ liệu time series còn có thể có yếu tố xu hướng (trend) và yếu tố chu kỳ (seasonality), các yếu tố này để phân tích thì cần sử dụng các mô hình đặc thù. Tuy nhiên, do đặc trưng của bộ dữ liệu là ứng với một mốc thời gian ('*date*') có nhiều điểm dữ liệu khác nhau theo từng biến '*team*' và '*department*', nên để phân tích chi tiết thì cần chia ra theo từng loại, điều này gây khó khăn khi khối lượng phân tích tương đối lớn. Vì vậy, nhóm đề xuất hướng phát triển trong tương lai của đồ án là khảo sát thêm yếu tố chuỗi thời gian trong dữ liệu và xây dựng những mô hình ước lượng tốt hơn.

TÀI LIỆU THAM KHẢO

[1] UCI Machine Learning Repository, Garment Employees Data Set**,** Link: https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees. (Truy cập 21/11/2022).

[2] Abdullah Al Imran, Md Nur Aminy, Md Rifatul Islam Rifatz, Shamprikta Mehreen, Deep Neural Network Approach for Predicting the Productivity of Garment Employees, CoDIT'19, 2019.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Trường Thịnh	- Tiền xử lý dữ liệu - Phân tích thăm dò - Xây dựng mô hình - Xây dựng ứng dụng - Viết báo cáo
2	Nguyễn Minh Tâm	- Tìm và thống kê bộ dữ liệu - Tạo dashboard - Tham gia xây dựng mô hình - Viết báo cáo - Chuẩn bị slide báo cáo