

TRƯỜNG ĐẠI HỌC CNTT
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

Đề tài: Phân tích và dự đoán năng suất lao động của công nhân trong ngành công nghiệp may mặc

Nhóm 3: Nguyễn Trường Thịnh - 20520783

Nguyễn Minh Tâm - 20520748

OUTLINE

01

Giới thiệu

Giới thiệu nguồn và nội dung bộ dữ liệu

02

Bộ dữ liệu

Thông tin cơ bản về bộ dữ liệu và các thuộc tính

03

Tiền xử lý

Các bước tiền xử lý dữ liệu

04

EDA

Thống kê mô tả và phân tích thăm dò các biến

05

Mô hình

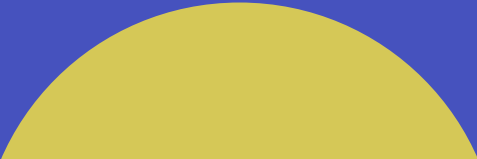
Các bước xây dựng mô hình và kết quả

06

Tổng kết

Nhận xét kết quả, hạn chế và hướng phát triển

1. Giới thiệu

- Bộ dữ liệu Garment Employees Dataset từ UCI
 - Bộ dữ liệu mô tả các yếu tố ảnh hưởng đến năng suất lao động thực tế của công nhân 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
- 

15 thuộc tính

1 thuộc tính bị khuyết giá trị là 'wip'

1197 mẫu



2. Bộ dữ liệu

2. Bộ dữ liệu

Bảng 1. Mô tả 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
quarter	Quarter1, Quarter2, Quarter3, Quarter4, Quarter5	Một phần của tháng, cứ 7 ngày liên tiếp tạo thành một quarter.
department	sewing, finishing	Phòng ban làm việc
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 mỗi ngày
smv	2.90 → 54.56	Thời gian phân bổ cho 1 tác vụ (phút)

Thuộc tính	Miền giá trị	Ý nghĩa
wip	7 → 23122	Số lượng phụ liệu chưa hoàn thành
over_time	0 → 25920	Thời gian làm việc quá giờ (phút)
incentive	0 → 3600	Tiền thưởng khuyến khích (BDT)
idle_time	0 → 300	Thời gian dây chuyền gián đoạn (phút)
idle_men	0 → 45	Số công nhân rảnh rỗi do dây chuyền gián đoạn
no_of_style_change	0 → 2	Số lần thay đổi thiết kế sản phẩm
no_of_workers	2 → 89	Số lượng công nhân mỗi tổ
actual_productivity	0.23 → 1.12	Năng suất lao động thực tế

3. Tiền xử lý

Preprocess



Sửa lỗi

Thuộc tính 'department' có 3 giá trị là 'sweing', 'finishing', 'finishing'. Sửa 'sweing' → 'sewing', 'finishing' → 'finishing'



Sửa kiểu dữ liệu

Sửa giá trị thuộc tính 'team' từ số nguyên 1 đến 12 thành biến phân loại từ 'Team1' đến 'Team12' (kiểu dữ liệu int64 → object)



Điền khuyết

Điền khuyết giá trị cho thuộc tính 'wip' bằng KNNImputer từ thư viện scikit-learn



Viết hàm

Tất cả các tác vụ tiền xử lý được đóng gói thành hàm `data_loader()`

4.1. Thống kê mô tả

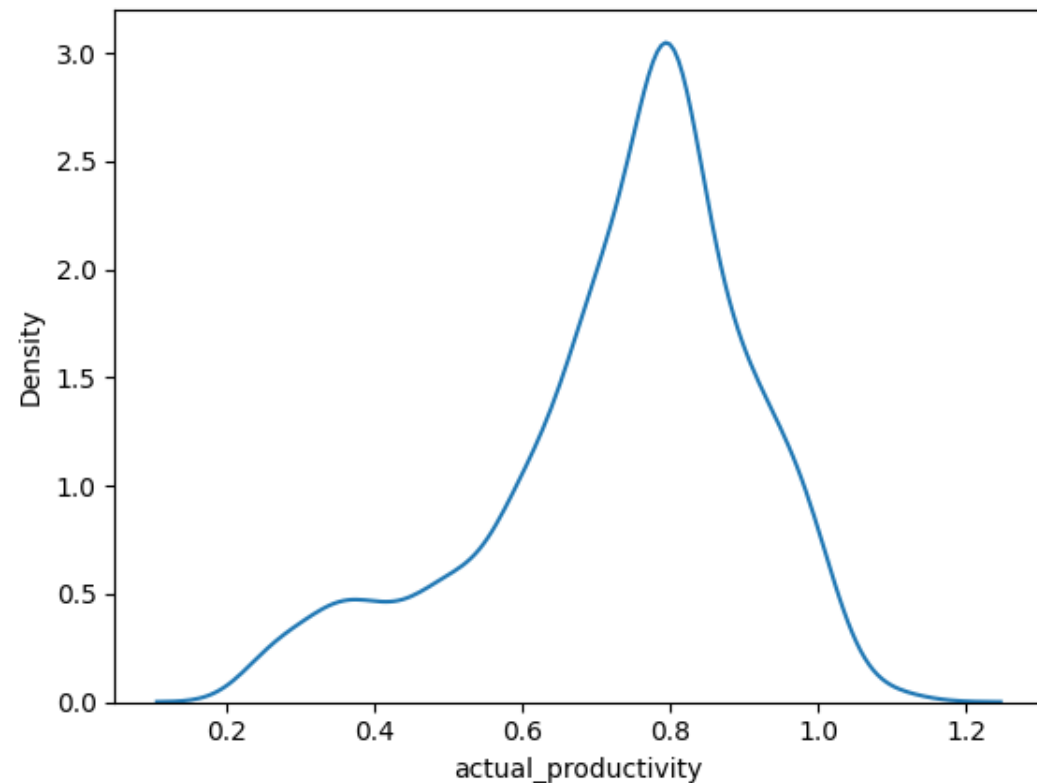
Biến	mean	std	min	Q1	med	Q3	max
targeted_productivity	0.73	0.10	0.07	0.70	0.75	0.80	0.80
smv	15.06	10.94	2.90	3.94	15.26	24.26	54.56
wip	1069.04	1414.80	7	773	983	1119	23122
over_time	4567.46	3348.82	0	1440	3960	6960	25920
incentive	38.21	160.18	0	0	0	50	3600
idle_time	0.73	12.71	0	0	0	0	300
idle_men	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

Biến kiểu số

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

Biến phân loại

4.2. Phân tích thăm dò biến mục tiêu

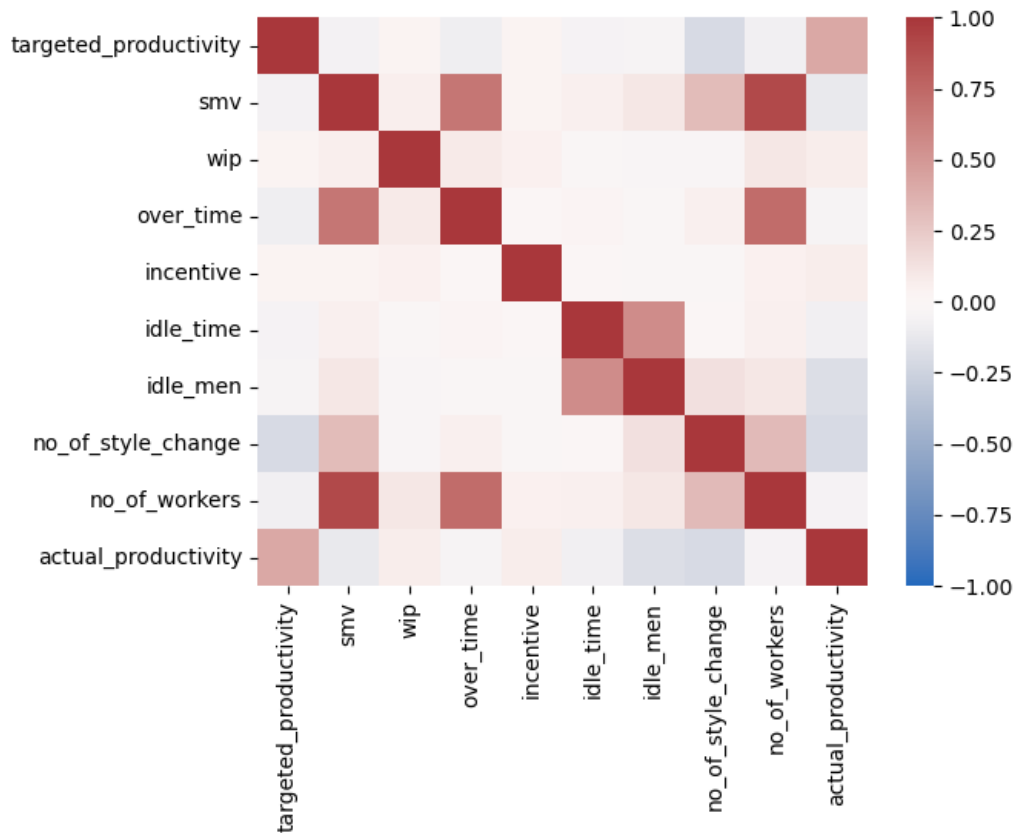


KDE plot của biến mục tiêu 'actual_productivity'

- Skewness = $-0.8 < 0$: biến lệch trái (negatively-skewed).
- Excess Kurtosis = $0.33 > 0$: biến có phần đuôi rộng (leptokurtic), có nhiều giá trị ngoại lệ.

4.2 Phân tích thăm dò thuộc tính kiểu số

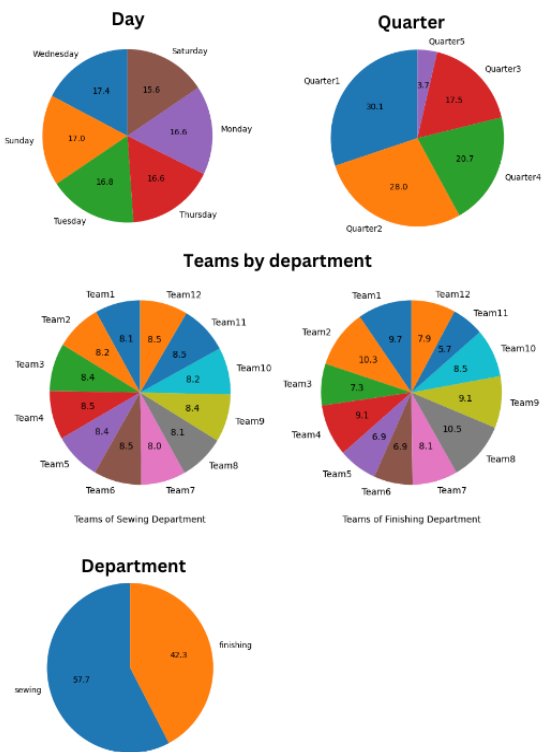
Hình 1. Mức độ tương quan giữa các biến kiểu số



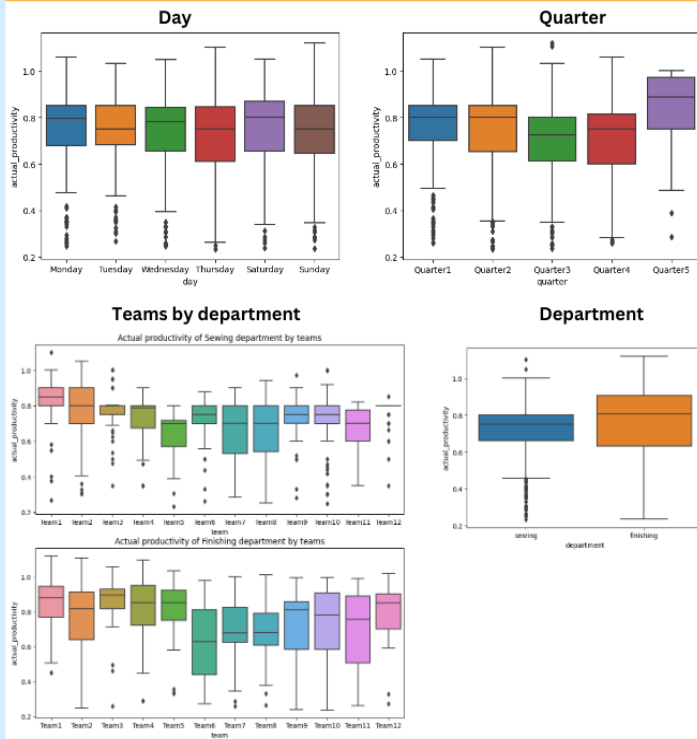
Nhận xét: Chỉ có biến 'targeted_productivity' có tương quan yếu với biến target 'actual_productivity'

EDA CATEGORICAL VARIABLES

Pie plot thể hiện tần suất dữ liệu của các biến



Box plot thể hiện tương quan với biến target

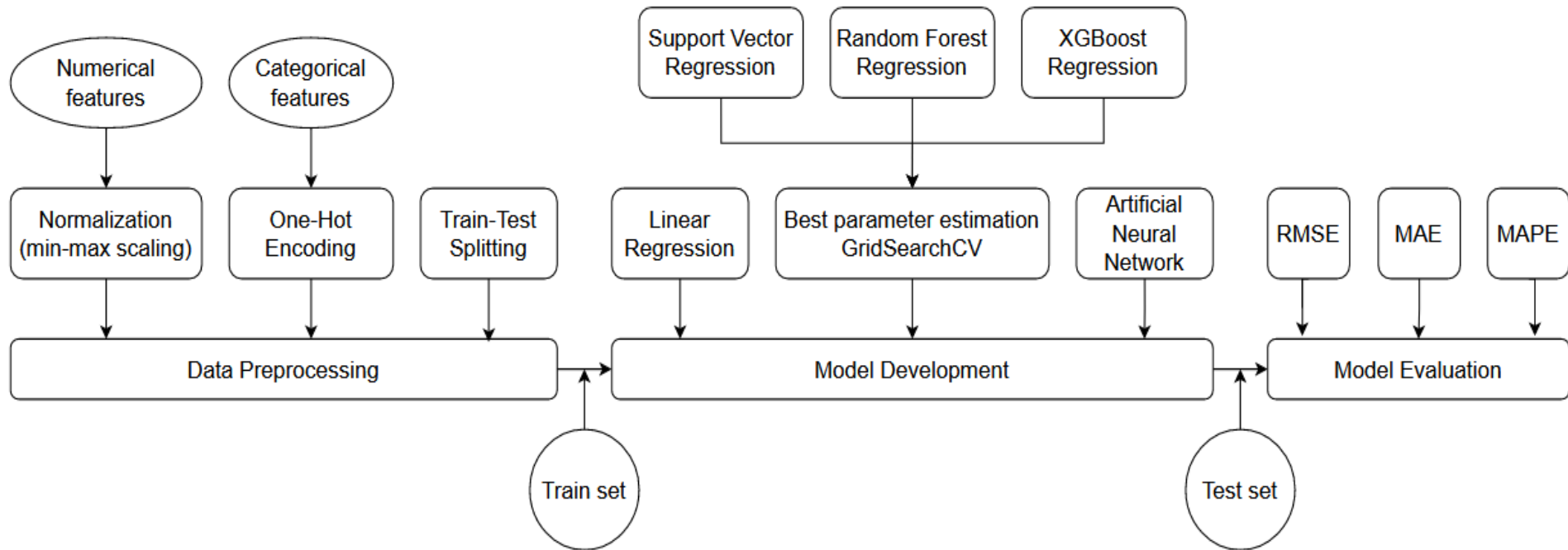


Kết quả phân tích ANOVA

Biến	F	p-value	Ảnh hưởng
quarter	7.1117	<< 0.05	Có
department	9.2462	0.0024 < 0.05	Có
day	0.7121	0.6144 > 0.05	Không
team_sewing	8.0338	<< 0.05	Có
team_finishing	5.2966	<< 0.05	Có

5. Mô hình

Hình 2. Các bước xây dựng mô hình



5. Mô hình

Đánh giá mô hình 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

5. Mô hình

Pipeline



Encode

Áp dụng `OnehotEncoder()` để chuyển kiểu phân loại về dạng số nhằm đưa vào mô hình



Scale

Áp dụng `MinMaxScaler()` để biến đổi các biến kiểu số về miền $[0, 1]$



Predict

Sử dụng mô hình `Random Forest Regression` (mô hình có kết quả tốt nhất) để đưa ra kết quả dự đoán



Pipeline

Tổng hợp tất cả các bước trên vào một pipeline. Lưu pipeline lại bằng thư viện `pickle`

6. Kết luận



EDA

Ít biến ảnh hưởng đến biến đầu ra. Do đó, nhóm giữ lại tất cả các biến trừ 'date'



Mô hình

Từ 5 mô hình, nhóm lựa chọn Random Forest Regression làm mô hình dự đoán cuối cùng.



Hạn chế

Các mô hình có sai số còn cao, chênh lệch 15% so với giá trị thực tế



Hướng phát triển

Khảo sát thêm yếu tố chuỗi thời gian trong dữ liệu và xây dựng mô hình dự đoán tốt hơn

Thanks!

Do you have any questions?

20520783@gm.uit.com.vn
+84 345 350 678



CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik

Please keep this slide for attribution

