

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**



NGUYỄN THÀNH LỘC

**DỰ ĐOÁN MỨC ĐỘ ẢNH HƯỞNG CỦA BIẾN ĐỔI
KHÍ HẬU ĐẾN CÁC VÙNG KINH TẾ TẠI VIỆT
NAM ỨNG DỤNG MACHINE LEARNING**

**KHÓA LUẬN TỐT NGHIỆP
NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

TP. HỒ CHÍ MINH, 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH



NGUYỄN THÀNH LỘC

DỰ ĐOÁN MỨC ĐỘ ẢNH HƯỚNG CỦA BIẾN ĐỔI
KHÍ HẬU ĐẾN CÁC VÙNG KINH TẾ TẠI VIỆT
NAM ỨNG DỤNG MACHINE LEARNING

Mã số sinh viên: 2154050160

KHÓA LUẬN TỐT NGHIỆP
NGÀNH HỆ THÔNG THÔNG TIN QUẢN LÝ

Giảng viên hướng dẫn: ThS. HỒ HƯỚNG THIÊN

TP. HỒ CHÍ MINH, 2025

TRƯỜNG ĐẠI HỌC MỞ CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
THÀNH PHỐ HỒ CHÍ MINH **Độc lập – Tự do – Hạnh phúc**
KHOA CÔNG NGHỆ THÔNG TIN

—
GIẤY XÁC NHẬN

Tôi tên là: Nguyễn Thành Lộc

Ngày sinh: 24/12/2003

Nơi sinh: Bến Tre

Chuyên ngành: Hệ thống thông tin quản lý

Mã sinh viên: 2154050160

Tôi đồng ý cung cấp toàn văn thông tin đồ án/khoa luận tốt nghiệp hợp lệ về bản quyền cho Thư viện Trường Đại học Mở Thành phố Hồ Chí Minh. Thư viện Trường Đại học Mở Thành phố Hồ Chí Minh sẽ kết nối toàn văn thông tin đồ án/khoa luận tốt nghiệp vào hệ thống thông tin khoa học của Sở Khoa học và Công nghệ Thành phố Hồ Chí Minh.

Ký tên
(Ghi rõ họ và tên)

**Ý KIẾN CHO PHÉP BẢO VỆ ĐỒ ÁN/KHÓA LUẬN TỐT NGHIỆP CỦA
GIẢNG VIÊN HƯỚNG DẪN**

Giảng viên hướng dẫn: Thạc sĩ Hồ Hướng Thiên

Sinh viên thực hiện: Nguyễn Thành Lộc Lớp: DH21IM01

Ngày sinh: 24/12/2003 Nơi sinh: Bến Tre

TÊN ĐỀ TÀI: DỰ ĐOÁN MỨC ĐỘ ẢNH HƯỞNG CỦA BIẾN ĐỔI KHÍ HẬU ĐẾN CÁC VÙNG KINH TẾ TẠI VIỆT NAM ÚNG DỤNG MACHINE LEARNING

Ý kiến của giảng viên hướng dẫn về việc cho phép sinh viên được bảo vệ đồ án/khoa luận trước Hội đồng:

.....
.....
.....
.....
.....
.....
.....
.....
.....

Thành phố Hồ Chí Minh, ngày ... tháng ... năm

Người nhận xét

Thạc sĩ Hồ Hướng Thiên

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn sâu sắc đến Thạc sĩ Hồ Hướng Thiên, người đã tận tình hướng dẫn, định hướng và đóng góp nhiều ý kiến quý báu trong suốt quá trình thực hiện khóa luận. Nhờ sự chỉ dẫn tận tâm của thầy, em đã có thể hoàn thành nghiên cứu này một cách trọn vẹn và hiệu quả nhất.

Em cũng xin chân thành cảm ơn Ban Giám hiệu cùng Quý thầy cô Trường Đại học Mở Thành phố Hồ Chí Minh, đặc biệt là Khoa Công nghệ Thông tin, những người đã tạo điều kiện thuận lợi, truyền đạt kiến thức và kỹ năng, giúp em xây dựng nền tảng vững chắc trong suốt quá trình học tập.

Bên cạnh đó, em xin gửi lời tri ân đến gia đình, bạn bè và đồng nghiệp, những người đã luôn ủng hộ, động viên và tiếp thêm động lực để em hoàn thành khóa luận một cách tốt nhất.

Mặc dù đã nỗ lực hết mình, nhưng chắc chắn không thể tránh khỏi những thiếu sót trong quá trình nghiên cứu. Em rất mong nhận được những góp ý quý báu từ Quý thầy cô để có thể tiếp tục hoàn thiện và tích lũy thêm kinh nghiệm trong tương lai.

Xin trân trọng cảm ơn!

Nguyễn Thành Lộc

Thành phố Hồ Chí Minh, tháng 5 năm 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

TÓM TẮT NỘI DUNG NGHIÊN CỨU BẰNG TIẾNG VIỆT

Biến đổi khí hậu đang ngày càng tác động rõ rệt đến các lĩnh vực kinh tế, xã hội trên toàn cầu, trong đó Việt Nam là một trong những quốc gia chịu ảnh hưởng nặng nề nhất. Trong bối cảnh đó, việc đánh giá và dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế trọng điểm là một yêu cầu cấp thiết nhằm hỗ trợ đưa ra các ý tưởng giúp phát triển thêm các chính sách phát triển bền vững. Khóa luận tập trung vào việc dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến sáu vùng kinh tế trọng điểm tại Việt Nam thông qua việc phân tích dữ liệu khí hậu lịch sử từ năm 2011 đến 2023.

Mục tiêu chính của nghiên cứu là xác định mức độ ảnh hưởng của các yếu tố khí hậu, các yếu tố công nghiệp và tác động của yếu tố con người đến từng vùng kinh tế nhằm nhận diện các khu vực đang chịu rủi ro cao.

Khóa luận sử dụng phương pháp tiếp cận định lượng, kết hợp các kỹ thuật thống kê mô tả và các mô hình học máy như Random Forest, XGBoost, Logistic Regression và Naive Bayes để xây dựng mô hình dự đoán và phân loại mức độ ảnh hưởng.

Kết quả nghiên cứu cho thấy có sự khác biệt rõ rệt giữa mức độ ảnh hưởng đến các vùng kinh tế tại Việt Nam. Một số vùng kinh tế như Đồng bằng sông Cửu Long và Đông Nam Bộ cho thấy xu hướng gia tăng nhiệt độ và mức độ phát thải khí nhà kính ảnh hưởng đến đời sống người dân nước ta. Mô hình học máy được lựa chọn để áp dụng đã đạt được độ chính xác cao trong việc phân loại mức độ khí hậu của các vùng.

Bài nghiên cứu mang lại ý nghĩa thực tiễn quan trọng trong việc hỗ trợ, đề xuất ý tưởng đến các cơ quan quản lý nhà nước, nhà hoạch định chính sách nhằm xây dựng các chiến lược thích ứng phù hợp, giảm thiểu tác động tiêu cực của biến đổi khí hậu và hướng tới phát triển bền vững cho từng vùng kinh tế.

TÓM TẮT NỘI DUNG NGHIÊN CỨU BẰNG TIẾNG ANH

Climate change is increasingly having a clear impact on economic and social sectors worldwide, with Vietnam being one of the countries most affected. In this context, assessing and predicting the extent of climate change impact on key economic regions is a pressing need to help propose ideas for sustainable development policies. This thesis focuses on predicting the extent of climate change effects on six key economic regions in Vietnam through the analysis of historical climate data from 2011 to 2023.

The primary objective of this study is to identify the impact of climate factors, industrial elements, and human influences on each region in order to pinpoint areas at high risk.

The research applies a quantitative approach, combining descriptive statistical techniques and machine learning models such as Random Forest, XGBoost, Logistic Regression, and Naive Bayes to develop prediction and classification models for assessing the impact levels.

The research findings highlight significant differences in the impact levels across the regions of Vietnam. Economic regions such as the Mekong Delta and Southeast region show a trend of increasing temperature and greenhouse gas emissions, affecting the livelihoods of the population. The selected machine learning model achieved high accuracy in classifying the climate impact levels of the regions.

This research offers valuable practical insights for supporting state agencies and policymakers in formulating adaptive strategies that mitigate the negative impacts of climate change, promoting sustainable development for each economic region.

MỤC LỤC

GIẤY XÁC NHẬN	
Ý KIẾN CHO PHÉP BẢO VỆ ĐỒ ÁN/KHÓA LUẬN TỐT NGHIỆP CỦA GIẢNG VIÊN HƯỚNG DẪN	
LỜI CẢM ƠN	
NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN	
TÓM TẮT NỘI DUNG NGHIÊN CỨU BẰNG TIẾNG VIỆT	
TÓM TẮT NỘI DUNG NGHIÊN CỨU BẰNG TIẾNG ANH	
MỤC LỤC	
DANH MỤC CHỮ VIẾT TẮT	
DANH SÁCH CÁC BẢNG, SƠ ĐỒ VÀ HÌNH VẼ	
CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI	1
1.1 Sơ lược về đề tài	1
1.2 Nội dung thực hiện	2
1.3 Phạm vi đề tài	3
1.4 Phương pháp nghiên cứu	4
1.5 Bố cục bài báo cáo	4
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	6
2.1 Học máy	6
2.2 Chuẩn bị dữ liệu	7
2.2.1 Tiền xử lý dữ liệu	7
2.2.2 Kỹ thuật tạo đặc trưng	8
2.2.3 Chuẩn hóa đặc trưng	10
2.2.4 Lựa chọn đặc trưng	12
2.3 Mô hình cây quyết định	13
2.3.1 Khái niệm	13
2.3.2 Cây phân loại	14
2.3.3 Cây hồi quy	15
2.4 Học tổ hợp	15
2.4.1 Khái niệm	15
2.4.2 Bagging	16

2.4.3 Stacking.....	17
2.4.4 Boosting	18
2.5 Các mô hình thuật toán dự đoán.....	20
2.5.1 Random Forest	20
2.5.2 XGBoost.....	23
2.5.3 Logistic Regression.....	26
2.5.4 Naive Bayes	27
2.6 Các chỉ số đánh giá chất lượng mô hình dự đoán	28
2.6.1 Ma trận nhầm lẫn	29
2.6.2 Các chỉ số đánh giá cơ bản.....	31
2.6.3 Đường cong ROC và chỉ số AUC trong phân loại đa lớp	33
CHƯƠNG 3: THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ.....	35
3.1 Bộ dữ liệu và phương pháp đề xuất.....	35
3.1.1 Giới thiệu tập dữ liệu	35
3.1.2 Bộ dữ liệu sử dụng	35
3.1.3 Phương pháp đề xuất.....	39
3.1.4 Quy trình xây dựng mô hình	43
3.2 Thực nghiệm và phân tích kết quả	44
3.2.1 Bộ dữ liệu nghiên cứu	45
3.2.2 Tiền xử lý dữ liệu	46
3.2.3 Phân tích và trực quan hóa	55
3.2.4 Xây dựng mô hình học máy	67
3.2.5 So sánh và đánh giá mô hình	72
3.2.6 Dự đoán và phân tích kết quả.....	81
CHƯƠNG 4: KẾT LUẬN VÀ KIẾN NGHỊ.....	88
4.1 Kết quả đạt được	88
4.2 Mặt hạn chế	88
4.3 Hướng phát triển cho tương lai	89
TÀI LIỆU THAM KHẢO.....	91
PHỤ LỤC	94

DANH MỤC CHỮ VIẾT TẮT

STT	Chữ viết tắt	Diễn giải
1	NASA	National Aeronautics and Space Administration
2	ROC	Receiver Operating Characteristic
3	XGBoost	Extreme Gradient Boosting
4	LightGBM	Light Gradient Boosting Machine
5	AUC	Area Under The Curve
6	API	Application Programming Interface
7	OvR	One-vs-Rest
8	AdaBoost	Adaptive Boosting
9	CatBoost	Categorical boosting
10	IIP	Index Of Industrial Production
11	kPa	Kilopascal
12	MJ	Megajoule

DANH SÁCH CÁC BẢNG, SƠ ĐỒ VÀ HÌNH VẼ

Bảng 2.1: Ví dụ về Label Encoding	9
Bảng 2.2: Ví dụ về Label Encoding (tiếp theo)	9
Bảng 2.3: Ví dụ về One-hot Encoding	10
Bảng 3.1: Danh sách các đặc trưng trong tập dữ liệu	37
Bảng 3.2: Danh sách các tỉnh thuộc các vùng kinh tế.....	38
Bảng 3.3: Danh sách thư viện sử dụng trong quá trình phân tích dữ liệu.....	43
Bảng 3.4: Kiểu dữ liệu của tập dữ liệu địa lý - khí hậu	47
Bảng 3.5: Kiểm tra giá trị thiếu của tập dữ liệu địa lý - khí hậu.....	48
Bảng 3.6: Mã hóa vùng kinh tế theo phương pháp Label Encoding.....	50
Bảng 3.7: Kiểu dữ liệu của tập dữ liệu môi trường, công nghiệp và nhân khẩu học	52
Bảng 3.8: Kiểm tra giá trị thiếu của tập dữ liệu môi trường, công nghiệp và nhân khẩu học	52
Bảng 3.9: Các chỉ số đánh giá của các thuật toán sử dụng	73
Hình 2.1: Sơ đồ cây quyết định.....	14
Hình 2.4: Sơ đồ phương pháp Bagging.....	16
Hình 2.5: Sơ đồ phương pháp Staking	18
Hình 2.6: Sơ đồ phương pháp Boosting.....	19
Hình 2.7: Sơ đồ thuật toán Random Forest.....	21
Hình 2.8: Sơ đồ thuật toán XGBoost	24
Hình 2.9: Cấu trúc ma trận nhầm lẫn phân loại nhị phân	30
Hình 2.10: Ví dụ minh họa ma trận nhầm lẫn phân loại đa lớp	31
Hình 3.1: Sơ đồ mô tả tổng quát quá trình thực nghiệm đề tài	40
Hình 3.2: Giới thiệu tập dữ liệu địa lý - khí hậu ban đầu	46
Hình 3.3: Tập dữ liệu địa lý - khí hậu sau khi tiền xử lý	50
Hình 3.4: Giới thiệu tập dữ liệu môi trường, công nghiệp và nhân khẩu học ban đầu	51
Hình 3.5: Tập dữ liệu môi trường, công nghiệp và nhân khẩu học sau khi tiền xử lý	54

Hình 3.6: Biểu đồ trung bình nhiệt độ trung bình các vùng kinh tế năm 2011-2023	55
Hình 3.7: Biểu đồ trung bình nhiệt độ lớn nhất các vùng kinh tế năm 2011-2023...56	
Hình 3.8: Biểu đồ trung bình nhiệt độ nhỏ nhất các vùng kinh tế năm 2011-2023 ..56	
Hình 3.9: Biểu đồ trung bình tổng lượng mưa các vùng kinh tế năm 2011-202357	
Hình 3.10: Biểu đồ trung bình độ ẩm các vùng kinh tế năm 2011-202358	
Hình 3.11: Biểu đồ trung bình độ ẩm đất các vùng kinh tế năm 2011-202358	
Hình 3.12: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Bắc Trung bộ và Duyên hải miền Trung và vùng Trung du và miền núi phía Bắc59	
Hình 3.13: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Tây Nguyên và vùng Đông Nam Bộ60	
Hình 3.14: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Đồng bằng sông Cửu Long và vùng Đồng bằng sông Hồng60	
Hình 3.15: Biểu đồ trung bình áp suất khí quyển các vùng kinh tế năm 2011-202361	
Hình 3.16: Biểu đồ trung bình bức xạ mặt trời các vùng kinh tế năm 2011-2023 ...61	
Hình 3.17: Biểu đồ trung bình tổng lượng mây rùng cả nước năm 2011-202362	
Hình 3.18: Biểu đồ trung bình tổng lượng khí thải nhà kính cả nước năm 2011-202363	
Hình 3.19: Biểu đồ trung bình IIP cả nước năm 2011-2023.....64	
Hình 3.20: Bản đồ so sánh mật độ dân số của Việt Nam năm 2011 và năm 2023 ...65	
Hình 3.21: Bản đồ so sánh dân số trung bình của Việt Nam năm 2011 và năm 202366	
Hình 3.22: Khoảng giá trị của các yếu tố đầu vào chưa chuẩn hóa67	
Hình 3.23: Biểu đồ thể hiện mức độ phân bố chỉ số ảnh hưởng của biến đổi khí hậu69	
Hình 3.24: Kết quả phân lớp mức độ ảnh hưởng cho tập dữ liệu ban đầu70	
Hình 3.25: So sánh chỉ số Precision theo từng lớp giữa các mô hình.....74	
Hình 3.26: So sánh chỉ số Recall theo từng lớp giữa các mô hình74	
Hình 3.27: So sánh chỉ số F1-Score theo từng lớp giữa các mô hình.....75	
Hình 3.28: Ma trận nhầm lẫn của mô hình Random Forest.....76	
Hình 3.29: Ma trận nhầm lẫn của mô hình XGBoost76	
Hình 3.30: Ma trận nhầm lẫn của mô hình Logistic Regression77	
Hình 3.31: Ma trận nhầm lẫn của mô hình Naive Bayes77	

Hình 3.32: Biểu đồ đường cong ROC theo phân loại đa lớp (OvR).....	79
Hình 3.33: Minh họa DataFrame compare_xgboost_df.....	81
Hình 3.34: Mức độ ảnh hưởng của biến đổi khí hậu theo từng vùng kinh tế	82
Hình 3.35: Bản đồ mức độ ảnh hưởng khí hậu các vùng kinh tế Việt Nam	84
Hình 3.36: Yếu tố ảnh hưởng mạnh đến biến đổi khí hậu theo XGBoost	86

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1 Sơ lược về đề tài

Biến đổi khí hậu đang trở thành mối đe dọa nghiêm trọng đối với môi trường sống và sự phát triển bền vững của nền kinh tế toàn cầu. Đặc biệt, Việt Nam đang là một trong những quốc gia chịu ảnh hưởng nặng nề nhất bởi các hiện tượng khí hậu cực đoan, bao gồm sự gia tăng nhiệt độ, mực nước biển dâng và các đợt thời tiết khắc nghiệt, dẫn đến nguy cơ ngập lụt, xâm nhập mặn và suy thoái tài nguyên thiên nhiên. Theo báo cáo của Cục Biến đổi khí hậu, trong giai đoạn 1958–2018, nhiệt độ trung bình năm của cả nước đã tăng khoảng $0,89^{\circ}\text{C}$, lượng mưa trung bình tăng 2,1% và mực nước biển trung bình tại các trạm hải văn ghi nhận mức tăng khoảng 2,74 mm mỗi năm. [1] Những biến động này không chỉ tác động trực tiếp đến hệ sinh thái biển và đất liền, mà còn ảnh hưởng sâu rộng đến các lĩnh vực kinh tế - xã hội, đặc biệt là nông nghiệp, công nghiệp và đời sống người dân nước ta.

Đứng trước thực trạng đó, việc nghiên cứu và phân tích xu hướng biến đổi khí hậu là hết sức cần thiết nhằm đánh giá hiện trạng, đưa ra dự báo và đề xuất các giải pháp ứng phó phù hợp trong giai đoạn hiện nay. Đặc biệt, việc ứng dụng các phương pháp khoa học hiện đại như mô hình học máy trong việc phân tích và dự đoán khí hậu đang mở ra nhiều triển vọng trong việc nâng cao độ chính xác và hiệu quả của các mô hình phân tích.

Đề tài dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế trọng điểm tại Việt Nam sẽ tập trung vào việc ứng dụng các mô hình học máy để đưa ra dự đoán về mức độ phân loại ảnh hưởng của từng vùng kinh tế. Phương pháp tiếp cận của nghiên cứu bao gồm việc thu thập dữ liệu về những đặc trưng có liên quan đến các yếu tố khí hậu, môi trường và những tác động của con người trong giai đoạn từ tháng 01 năm 2011 đến tháng 12 năm 2023. Ban đầu, bộ dữ liệu sẽ được phân tích theo phạm vi 63 tỉnh thành tại Việt Nam, sau đó phân nhóm thành 6 vùng kinh tế trọng điểm nhằm tối ưu hóa quá trình xử lý dữ liệu và xác định khu vực chịu ảnh hưởng nghiêm trọng nhất bởi biến đổi khí hậu. Các thuật toán học máy như Random

Forest, XGBoost, Logistic Regression và Naive Bayes sẽ được triển khai để huấn luyện, dự đoán và đánh giá hiệu suất của mô hình.

Những mô hình này được kỳ vọng sẽ cung cấp các thông tin khoa học có giá trị, hỗ trợ các cơ quan quản lý hay các nhà hoạch định chính sách trong việc đưa ra các biện pháp ứng phó kịp thời và hiệu quả, từ đó góp phần bảo vệ hệ sinh thái và nâng cao sự phát triển bền vững về kinh tế, xã hội, môi trường.

1.2 Nội dung thực hiện

Trước tiên, nghiên cứu sẽ tiến hành thu thập dữ liệu thứ cấp từ các nguồn chính thống như Tổng cục Thống kê Quốc gia, Bộ Tài nguyên và Môi trường, cùng với các tổ chức quốc tế như NASA (National Aeronautics and Space Administration), Global Forest Watch. Dữ liệu được thu thập bao gồm các chỉ số thuộc yếu tố khí hậu như nhiệt độ, lượng mưa, độ ẩm, mật độ dân số, chỉ số khí thải nhà kính, chỉ số sản xuất công nghiệp,... Những yếu tố trên vô cùng quan trọng để tạo nên bài phân tích, đặc biệt khi con người cũng đang là một nhân tố đặc biệt quan trọng góp phần vào sự biến đổi khí hậu hiện nay.

Sau khi thu thập, các tập dữ liệu sẽ được xử lý và làm sạch nhằm loại bỏ các giá trị nhiễu hoặc thiếu, đồng thời chuẩn hóa về mặt định dạng và đơn vị đo. Tiếp đến, dữ liệu sẽ được phân tích theo từng tỉnh thành tổng hợp lại thành 6 vùng kinh tế trọng điểm để thuận tiện hơn cho quá trình so sánh và dự đoán.

Trong giai đoạn xây dựng mô hình, các thuật toán học máy sẽ được huấn luyện và kiểm tra trên tập dữ liệu lịch sử, với mục tiêu là có thể dự đoán được mức độ ảnh hưởng theo từng vùng kinh tế tại Việt Nam.

Đồng thời, các chỉ số đánh giá mô hình được sử dụng để đánh giá và so sánh hiệu quả của các thuật toán được sử dụng trong bài toán. Từ đó có thể chọn ra được thuật toán có hiệu suất tốt nhất để tiến hành thực hiện bước dự đoán của đề tài.

Thông qua việc ứng dụng các công cụ phân tích hiện đại và tiếp cận dữ liệu theo hướng thời gian – không gian, bài nghiên cứu được kỳ vọng sẽ mở ra phương hướng

nghiên cứu mới, góp phần trong việc đưa ra những kết luận có giá trị thực tiễn cao, đóng góp vào công tác quản lý môi trường và hoạch định chính sách phát triển bền vững trong bối cảnh biến đổi khí hậu ngày càng diễn biến phức tạp như hiện nay.

1.3 Phạm vi đề tài

Khóa luận tập trung chủ yếu vào việc ứng dụng các kỹ thuật học máy để phân tích và dự đoán xu hướng biến đổi khí hậu tại Việt Nam trong giai đoạn từ tháng 1 năm 2011 đến tháng 12 năm 2023. Nghiên cứu sử dụng dữ liệu theo hướng thời gian – không gian, bao gồm 63 tỉnh thành và sau đó được phân nhóm thành 6 vùng kinh tế trọng điểm nhằm tối ưu hóa hiệu quả phân tích.

Phạm vi dữ liệu bao gồm các đặc trưng và nhóm đặc trưng chính như:

Nhóm đặc trưng nhiệt độ gồm Nhiệt độ trung bình (Average Temperature), Nhiệt độ cao nhất (Max Temperature), Nhiệt độ nhỏ nhất (Min Temperature).

Nhóm đặc trưng lượng mưa và độ ẩm gồm Tổng lượng mưa trung bình (Total Recipitation), Độ ẩm trung bình (Relative Humidity), Độ ẩm đất (Soil Moisture)

Nhóm đặc trưng gió và áp suất gồm Tốc độ gió (Wind Speed), hướng gió (Wind Direction), Áp suất bề mặt (Surface Pressure).

Đặc trưng lượng bức xạ mặt trời (Solar Radiation).

Nhóm đặc trưng môi trường gồm Diện tích mất rừng (Tree Cover Loss), Lượng khí thải nhà kính (Green House Gas).

Đặc trưng Chỉ số sản xuất công nghiệp (Index Of Industrial Production).

Nhóm đặc trưng nhân khẩu học và địa lý gồm Diện tích (Area), Dân số trung bình (Average Population), Mật độ dân số (Population Density).

Các đặc trưng trên sẽ thu thập từ các nguồn uy tín nhằm giúp khóa luận đảm bảo được mức độ tin cậy về mặt dữ liệu. Sau khi thu thập, các tập dữ liệu sẽ được xử lý thông qua các giai đoạn như tiền xử lý, phân tích và trực quan hóa. Sau đó sử dụng các mô hình học máy để tiến hành xây dựng mô hình dự đoán mức độ ảnh hưởng của biến đổi khí hậu.

Phạm vi đánh giá và so sánh giữa các thuật toán với nhau bao gồm việc sử dụng các chỉ số hiệu suất như Accuracy, Precision, Recall, F1-Score cùng với các biểu đồ như Ma trận nhầm lẫn (Confusion Matrix) và Biểu đồ đường cong ROC (Receiver Operating Characteristic Curve). Sau đó tiến hành đánh giá mô hình có hiệu suất tốt nhất, phục vụ cho việc dự đoán của đề tài.

1.4 Phương pháp nghiên cứu

Để đạt được mục tiêu nghiên cứu là phân tích sự biến đổi khí hậu và dự đoán mức độ ảnh hưởng của nó đến các khu vực kinh tế tại Việt Nam, đề tài áp dụng phương pháp nghiên cứu định lượng kết hợp với kỹ thuật phân tích mô tả và phân tích dự đoán, đặc biệt là áp dụng các mô hình học máy như Random Forest, XGBoost, Logistic Regression và Naive Bayes. Việc dự đoán ảnh hưởng của biến đổi khí hậu sẽ được thực hiện bằng ngôn ngữ Python, trên phần mềm Google Colab.

1.5 Bộ cục bài báo cáo

Bài báo cáo được chia thành 4 phần như sau:

Chương 1: Mục tiêu giới thiệu tổng quan, nội dung thực hiện đề tài, nêu rõ giới hạn và các phương pháp nghiên cứu được áp dụng khi thực hiện khóa luận.

Chương 2: Trình bày các cơ sở lý thuyết có liên quan đến đề tài, các thuật toán được sử dụng trong việc huấn luyện và dự đoán kết quả.

Chương 3: Giới thiệu tổng quan bộ dữ liệu sử dụng, các phương pháp đề xuất và quy trình xây dựng các mô hình. Đồng thời mô tả chi tiết các giai đoạn thực hiện phân tích và dự đoán, đánh giá và so sánh hiệu quả các mô hình học máy.

Chương 4: Đưa ra kết quả đạt được, chỉ ra những mặt hạn chế khi thực hiện đề tài và đề xuất hướng phát triển trong tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Học máy

Học máy (Machine Learning) [2] là một phương pháp nằm trong trí tuệ nhân tạo (Artificial Intelligence), tập trung vào việc phát triển các thuật toán và mô hình có khả năng học hỏi từ dữ liệu mà không cần phải thực hiện theo từng bước cụ thể. Thay vì dựa hoàn toàn vào các quy tắc do con người thiết lập, hệ thống học máy có thể tự động nhận diện mẫu hình, rút trích thông tin và cải thiện hiệu suất theo thời gian thông qua quá trình huấn luyện với dữ liệu đầu vào.

Mục tiêu chính của học máy là giúp hệ thống đưa ra quyết định chính xác hơn khi tiếp xúc với dữ liệu mới, bằng cách xây dựng các mô hình dựa trên các tập dữ liệu lịch sử. Một đặc điểm nổi bật là khả năng thích nghi liên tục, cho phép các mô hình điều chỉnh và tối ưu hóa dự đoán khi điều kiện đầu vào thay đổi. Điều này đặc biệt hữu ích trong bối cảnh cần phải phân tích phức tạp hoặc có các tập dữ liệu biến động liên tục theo thời gian như khí hậu, tài chính, hay các hành vi của người dùng.

Học máy được chia thành bốn phương pháp chính, bao gồm:

Học có giám sát (Supervised Learning): Mô hình được huấn luyện trên tập dữ liệu đã gán nhãn, cho phép dự đoán đầu ra từ đầu vào với độ chính xác cao.

Học không giám sát (Unsupervised Learning): Dữ liệu huấn luyện sẽ là dữ liệu không được gán nhãn, thuật toán phải tự tìm ra cấu trúc và các nhóm ẩn trong dữ liệu để đưa ra kết quả dự đoán.

Học bán giám sát (Semi-supervised Learning): Các mô hình học máy sẽ kết hợp giữa các tập dữ liệu có nhãn và không có nhãn lại với nhau.

Học củng cố (Reinforcement Learning): Mô hình học thông qua quá trình thử nghiệm và nhận phản hồi từ môi trường, qua nhiều lần thử và sai, mô hình sẽ học được cách hoạt động sao cho đạt được kết quả tốt nhất.

Với tiềm năng to lớn trong việc tự động hóa, phân tích xu hướng và tối ưu hóa quyết định, học máy đã và đang được áp dụng rộng rãi trong nhiều lĩnh vực như y tế, công nghiệp, giáo dục hay môi trường. Trong bài nghiên cứu này, học máy sẽ được sử dụng như công cụ nền tảng để xây dựng mô hình dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế trọng điểm tại Việt Nam, từ đó hỗ trợ lập kế hoạch và ra quyết định hiệu quả hơn trong tương lai.

2.2 Chuẩn bị dữ liệu

2.2.1 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu (Data Preprocessing) là một phương pháp quan trọng trong quá trình khám phá và phân tích dữ liệu. Đây là tập hợp các kỹ thuật được thực hiện trước khi áp dụng các phương pháp phân tích dữ liệu, nhằm làm sạch, chuẩn hóa và điều chỉnh thông tin sao cho phù hợp với yêu cầu của từng loại thuật toán khác nhau.

Bộ dữ liệu thô ban đầu thường gặp nhiều vấn đề như thiếu giá trị, chứa các thông tin thừa hoặc không nhất quán, khiến việc sử dụng trực tiếp các thuật toán khai thác trở nên không hiệu quả hoặc không khả thi. Việc thực hiện tiền xử lý dữ liệu trở nên cần thiết hơn bao giờ hết. Quá trình này giúp nâng cao chất lượng dữ liệu đầu vào, góp phần gia tăng hiệu quả và độ chính xác của các mô hình thuật toán được áp dụng phía sau. [3]

Việc tiền xử lý dữ liệu được chia thành một số kỹ thuật chính như:

Xử lý dữ liệu không hoàn hảo (Imperfect Data): Trong phân tích dữ liệu, dữ liệu hoàn hảo là tập dữ liệu có chứa đầy đủ thông tin và không bị nhiễu. Tuy nhiên, dữ liệu thực tế thường không đáp ứng được các điều kiện đó. Chính vì vậy, trong quá trình tiền xử lý dữ liệu, việc áp dụng các kỹ thuật để loại bỏ dữ liệu nhiễu hoặc điền vào các giá trị bị thiếu là rất cần thiết.

Giảm chiều dữ liệu (Dimensionality Reduction): Các tập dữ liệu lớn về số lượng các thuộc tính dự đoán thường tiêu tốn nhiều thời gian huấn luyện mô hình và dung lượng lưu trữ. Do vậy, đối với những bộ dữ liệu có kích thước lớn, cần phải tiến hành giảm chiều dữ liệu nhằm rút gọn những thuộc tính không cần thiết.

Giảm số lượng mẫu (Instance Reduction): Việc giảm số lượng mẫu giúp loại bỏ đi những mẫu không cần thiết hoặc dư thừa trong tập dữ liệu, từ đó giảm kích thước dữ liệu mà vẫn giữ được các thông tin quan trọng.

2.2.2 Kỹ thuật tạo đặc trưng

Kỹ thuật tạo đặc trưng (Feature Engineering) là quá trình lựa chọn, trích xuất và biến đổi các đặc trưng có sẵn từ dữ liệu thô sang loại dữ liệu có thể sử dụng hiệu quả trong các mô hình học máy. Việc áp dụng Feature Engineering giúp cải thiện độ chính xác và kết quả của các mô hình dự đoán.

Các loại kỹ thuật phổ biến nhất có thể kể đến như Label Encoding và One-hot Encoding [4], giúp mã hóa các biến phân loại thành dạng số để thuận tiện hơn trong quá trình huấn luyện mô hình dự đoán.

2.2.2.1 Label Encoding

Label Encoding là quá trình gán nhãn cho một đặc trưng trong tập dữ liệu theo dạng số nguyên không trùng lặp và theo thứ tự bảng chữ cái.

Label Encoding được sử dụng cho các đặc trưng có dữ liệu phân loại theo tính Logic Ví dụ: Một tập dữ liệu ban đầu có các thuộc tính và giá trị như sau:

Color.ID	Color
FF0000	Red
0000FF	Blue
008000	Green

008000	Green
FFFFFF	White

Bảng 2.1: Ví dụ về Label Encoding

Sau khi áp dụng kỹ thuật gán nhãn Label Encoding, các giá trị tại thuộc tính Color sẽ được chuyển hóa từ dữ liệu dạng chữ sang dạng số nguyên như sau:

Color.ID	Color
FF0000	2
0000FF	0
008000	1
008000	1
FFFFFF	3

Bảng 2.2: Ví dụ về Label Encoding (tiếp theo)

Việc sử dụng Label Encoding giúp tiết kiệm được phần lớn bộ nhớ so với One-hot Encoding khi chỉ tạo ra một cột giá trị mới dựa trên đặc trưng cũ.

2.2.2.2 One-hot Encoding

Với kỹ thuật One-hot Encoding, thuộc tính ban đầu sẽ được phân tách thành các cột giá trị và được mã hóa theo dạng nhị phân (0 và 1).

Tiếp tục với ví dụ và bảng dữ liệu ban đầu, bộ dữ liệu khi áp dụng kỹ thuật One-hot Encoding sẽ được biểu diễn như sau:

Color.ID	Red	Blue	Green	White
FF0000	1	0	0	0

0000FF	0	1	0	0
008000	0	0	1	0
008000	0	0	1	0
FFFFFF	0	0	0	1

Bảng 2.3: Ví dụ về One-hot Encoding

Các giá trị nhị phân được tạo ra tại các thuộc tính mới được hiểu như 1 (Thỏa điều kiện) và 0 (Không thỏa điều kiện).

Sử dụng One-hot Encoding sẽ tránh được tình trạng thứ tự giả do mỗi giá trị khi mã hóa sẽ tạo thành các thuộc tính riêng biệt. Nhưng cũng chính vì vậy mà đã vô tình làm tăng kích thước của tập dữ liệu. Càng nhiều đặc trưng cần mã hóa thì càng sinh ra nhiều thuộc tính khác nhau.

2.2.3 Chuẩn hóa đặc trưng

2.2.3.1 Khái niệm

Chuẩn hóa đặc trưng (Feature Scaling) là một phương pháp thường được sử dụng để chuẩn hóa phạm vi của các biến độc lập hoặc các đặc trưng trong tập dữ liệu.

Thông thường, một bộ dữ liệu sẽ chứa nhiều loại thuộc tính với các đơn vị đo lường khác nhau. Chẳng hạn, ở thuộc tính A có các giá trị nằm trong khoảng 10 đến 100, thuộc tính B có giá trị trong khoảng 1000 đến 1000. Việc áp dụng Feature Scaling nhằm đưa các giá trị về một khoảng xác định có cùng phạm vi, tránh việc một số thuộc tính có giá trị quá lớn chi phối quá trình dự đoán.

Hai thuật toán phổ biến khi sử dụng phương pháp Feature Scaling bao gồm Normalization và Standardization: [5]

2.2.3.2 Normalization (Min-Max Scaling)

Normalization, còn được gọi là Min-Max Scaling, là phương pháp đơn giản nhất trong việc chuẩn hóa các đặc trưng để đưa giá trị về khoảng 0 đến 1. Công thức Normalization (Min-Max Scaling) được biểu diễn như sau:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Trong đó:

- x' thể hiện giá trị sau khi chuẩn hóa
- x thể hiện giá trị gốc của đặc trưng
- $\min(x)$ thể hiện giá trị lớn nhất của đặc trưng
- $\max(x)$ thể hiện giá trị nhỏ nhất của đặc trưng

Min-Max Scaling phù hợp với các tập dữ liệu chưa xác định phân phối hoặc không tuân theo phân phối chuẩn (Phân phối Gaussian), do khi áp dụng Min-Max Scaling sẽ đưa các giá trị về một phạm vi nhất định mà không làm thay đổi tỷ lệ giữa các giá trị với nhau.

2.2.3.3 Standardization (Z-Scores)

Phương pháp Standardization, còn được gọi là Z-Scores, là phương pháp chuẩn hóa sao cho các giá trị trong đặc trưng sau khi chuẩn hóa có trung bình xấp xỉ bằng 0 và độ lệch chuẩn xấp xỉ bằng 1. Công thức Standardization (Z-Scores) được biểu diễn như sau:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Trong đó:

- x' thể hiện giá trị sau khi chuẩn hóa
- x thể hiện giá trị gốc của đặc trưng
- \bar{x} thể hiện giá trị trung bình của đặc trưng
- σ thể hiện độ lệch chuẩn của đặc trưng

Với Standardization sẽ phù hợp hơn với các tập dữ liệu tuân theo phân phối chuẩn (phân phối Gaussian) do nó giúp duy trì các đặc tính của phân phối (trung bình, độ lệch chuẩn) giúp mô hình học máy hoạt động hiệu quả hơn.

2.2.4 Lựa chọn đặc trưng

Lựa chọn đặc trưng (Feature Selection) là quá trình trích chọn những đặc trưng quan trọng nhất và loại bỏ những đặc trưng không cần thiết trong việc phân tích và dự đoán. Dựa vào mức độ liên quan của các đặc trưng trong tập dữ liệu, các đặc trưng được phân loại thành các thành 4 mức độ sau [6]:

1. Đặc trưng bị nhiễu và không liên quan
2. Đặc trưng dư thừa và có mức độ liên quan thấp
3. Đặc trưng có độ liên quan thấp nhưng không dư thừa
4. Đặc trưng có mức độ liên quan mạnh mẽ

Mục tiêu quan trọng nhất của việc chọn lọc đặc trưng đó là giữ lại những đặc trưng quan trọng, có liên quan và loại bỏ đi những đặc trưng không cần thiết. Từ đó giúp cải thiện độ chính xác của mô hình.

Dựa vào những giá trị sẵn có của thông tin nhãn (label), việc lựa chọn những đặc trưng phù hợp để ứng dụng vào các mô hình máy học được chia thành các nhóm như: Phương pháp lựa chọn đặc trưng có giám sát (Supervised) và lựa chọn đặc trưng không giám sát (Unsupervised). [6]

Phương pháp lựa chọn đặc trưng có giám sát (Supervised) sẽ tận dụng những thông tin có dán nhãn và tiến hành lựa chọn những đặc trưng có khả năng phân biệt cao và có liên quan trực tiếp đến biến mục tiêu. Phương pháp này thường được sử dụng trong các bài toán phân loại hoặc hồi quy, với mục tiêu lựa chọn tập đặc trưng tối ưu để cải thiện hiệu suất dự đoán của các mô hình học máy.

Phương pháp lựa chọn đặc trưng không giám sát (Unsupervised) không sử dụng những thông tin về biến mục tiêu. Thay vào đó, chúng thường

dựa vào các chỉ số thống kê như hệ số tương quan để loại bỏ các đặc trưng dữ thừa hoặc không cần thiết.

Việc lựa chọn phương pháp phù hợp dựa trên thông tin nhãn là yếu tố quan trọng trong quá trình chọn lọc đặc trưng cần thiết, nhằm tối ưu hiệu quả các mô hình.

2.3 Mô hình cây quyết định

2.3.1 Khái niệm

Cây quyết định (Decision Tree) là một mô hình trực quan hóa quá trình ra quyết định, thể hiện các lựa chọn khác nhau cùng với những kết quả có thể xảy ra tương ứng. Nhờ cấu trúc phân nhánh rõ ràng, cây quyết định giúp người dùng có thể dễ dàng phân tích và lựa chọn được phương án tối ưu trong quá trình giải quyết vấn đề.

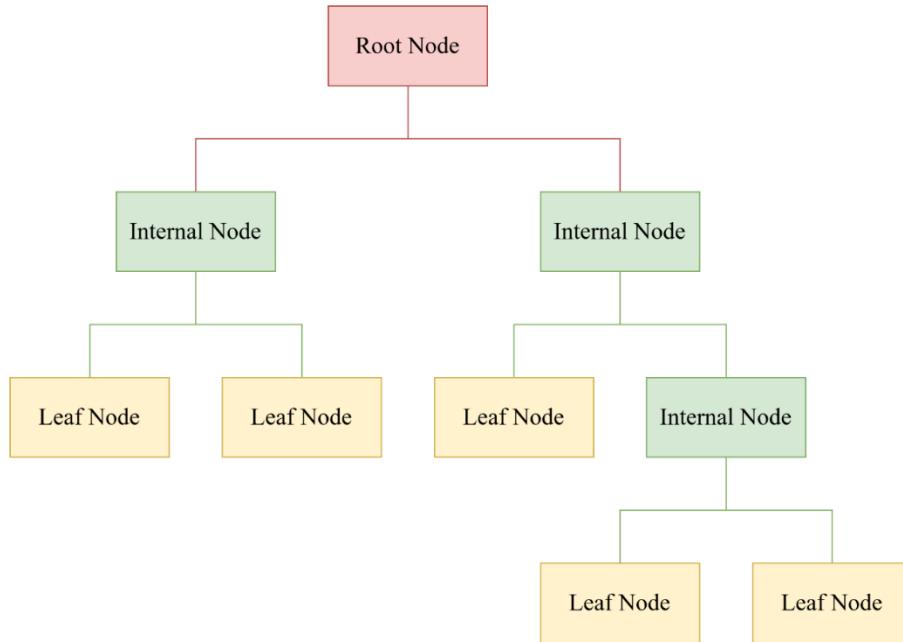
Các thành phần của cây quyết định bao gồm:

Root Node (Nút gốc): Là nơi bắt đầu quá trình phân cấp.

Branches (Các nhánh): Là các đường nối giữa các nút, thể hiện quá trình di chuyển từ một quyết định này đến một kết quả khác.

Internal Nodes (Các nút trung gian): Là các điểm nút nơi điều kiện phân chia được áp dụng để chia nhỏ dữ liệu thành các nhóm cụ thể hơn dựa trên các thuộc tính đầu vào.

Leaf Nodes (Các nút lá): Là các điểm nút cuối cùng trong cây quyết định, đại diện cho các kết quả hay các nhãn phân loại cuối cùng của mô hình.



Hình 2.1: Sơ đồ cây quyết định

Mô hình cây quyết định được sử dụng với mục đích chính hiểu rõ mối quan hệ giữa các đặc trưng đầu vào và kết quả đầu ra, thông qua cấu trúc phân nhánh trực quan thể hiện rõ cách từng đặc trưng ảnh hưởng đến quyết định cuối cùng. Bên cạnh đó, mô hình này còn đóng vai trò như một công cụ hỗ trợ ra quyết định hiệu quả, cho phép người dùng phân tích các tình huống một cách có hệ thống, từ đó lựa chọn được phương án tối ưu dựa trên các điều kiện cụ thể.

Dựa vào bản chất cuối cùng của biến mục tiêu, mô hình cây quyết định được phân thành 2 loại, bao gồm: Cây phân loại (Classification Trees) và cây hồi quy (Regression Trees). [7]

2.3.2 Cây phân loại

Cây phân loại (Classification Trees) là một dạng cây quyết định được sử dụng với các bài toán có biến mục tiêu là các giá trị phân loại. Mô hình này giúp dự đoán nhãn của một đối tượng dựa trên các đặc trưng đầu vào của nó.

Quá trình hoạt động của cây phân loại bắt đầu từ nút gốc, nơi một câu hỏi đầu tiên được đưa ra dựa trên đặc trưng của dữ liệu. Các câu hỏi tiếp theo sẽ tiếp tục phân chia dữ liệu thành các nhóm con nhỏ hơn dựa trên các đặc trưng còn lại. Quá trình này tiếp tục cho đến khi cây đạt đến nút lá, sẽ là nơi kết quả dự đoán cuối cùng được đưa ra.

Với mục tiêu của là dự đoán mức độ ảnh hưởng của biến đổi khí hậu, mô hình cây phân loại sẽ được sử dụng để đưa ra kết quả cho bài toán.

2.3.3 Cây hồi quy

Cây hồi quy (Regression Trees) là một dạng cây quyết định được sử dụng khi biến mục tiêu là một giá trị liên tục thay vì phân loại rời rạc. Mục tiêu của cây hồi quy là dự đoán giá trị số của một đối tượng đầu vào dựa trên các đặc trưng của nó.

Quá trình xây dựng cây hồi quy bắt đầu từ nút gốc, nơi toàn bộ tập dữ liệu được xem xét và một điều kiện phân chia được lựa chọn sao cho sai số dự đoán được giảm xuống thấp nhất. Việc phân chia này tiếp tục diễn ra cho đến khi tới nút lá. Tại mỗi nút lá, cây hồi quy sẽ trả về một giá trị dự đoán cụ thể.

2.4 Học tổ hợp

2.4.1 Khái niệm

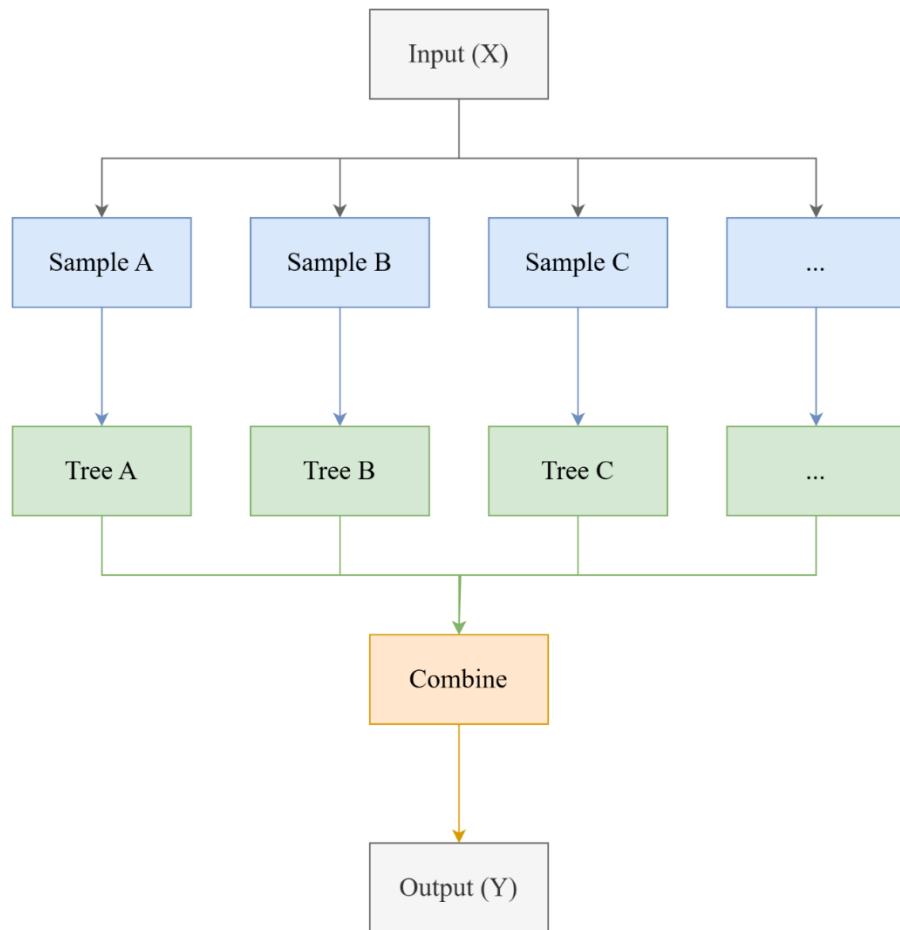
Học tổ hợp (Ensemble Learning) là một phương pháp trong học máy nhằm nâng cao độ chính xác và độ ổn định của mô hình dự đoán thông qua việc kết hợp nhiều mô hình học máy khác nhau thay vì phụ thuộc vào một mô hình đơn lẻ. Bằng cách khai thác các điểm mạnh của từng mô hình, Ensemble Learning giúp giảm thiểu sai số, cải thiện hiệu suất tổng thể và tăng cường tính mạnh mẽ của mô hình trong các bài toán phân tích dữ liệu.

Các nhóm phương pháp chính của Ensemble Learning gồm Bagging (Bootstrap Aggregating), Stacking (Stacked Generalization) và Boosting. [8]

2.4.2 Bagging

Bagging, còn gọi là Bootstrap Aggregating, là một kỹ thuật nhằm cải thiện hiệu suất dự đoán của mô hình bằng cách giảm phương sai và ngăn hiện tượng quá khóp dữ liệu.

Cơ chế hoạt động của Bagging bao gồm việc tạo ra nhiều tập dữ liệu huấn luyện con thông qua lấy mẫu bootstrap, sau đó huấn luyện từng mô hình trên từng tập dữ liệu này. Các mô hình sẽ đưa ra dự đoán độc lập và kết quả cuối cùng được tổng hợp bằng phương pháp bỏ phiếu đa số (majority voting) trong bài toán phân loại hoặc tính trung bình (averaging) trong bài toán hồi quy.



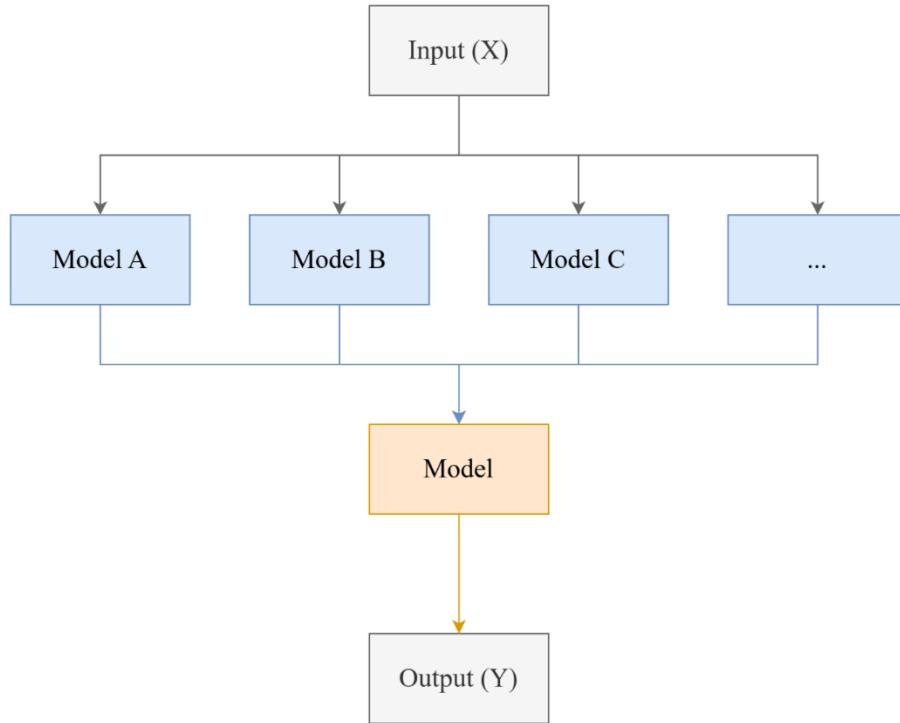
Hình 2.4: Sơ đồ phương pháp Bagging

Sự đa dạng trong các mô hình thành phần chính là yếu tố quan trọng giúp Bagging hoạt động hiệu quả. Nhờ đó, phương pháp này không chỉ giúp giảm độ lệch chuẩn trong dự đoán mà còn nâng cao khả năng tổng quát hóa của mô hình. Bagging là nền tảng cho nhiều thuật toán học máy tổ hợp phổ biến như Bagged Decision Trees, Random Forest, Extra Trees.

2.4.3 Stacking

Stacking, còn gọi là Stacked Generalization, là một kỹ thuật nhằm cải thiện hiệu suất dự đoán bằng cách kết hợp nhiều mô hình học máy khác nhau thông qua một mô hình tổng hợp. Thay vì sử dụng một thuật toán duy nhất như trong Bagging, Stacking tận dụng sự đa dạng về kiến trúc mô hình bằng cách huấn luyện nhiều mô hình khác nhau trên cùng một tập dữ liệu đầu vào.

Quy trình của Stacking thường bao gồm hai tầng mô hình. Tầng đầu tiên bao gồm nhiều mô hình khác nhau nhằm tạo ra những dự đoán ban đầu từ tập dữ liệu gốc. Sau đó, các dự đoán này được sử dụng làm đặc trưng đầu vào cho tầng thứ hai, thường là một mô hình đơn giản để học cách kết hợp tối ưu các đầu ra từ tầng đầu tiên.



Hình 2.5: Sơ đồ phương pháp Stacking

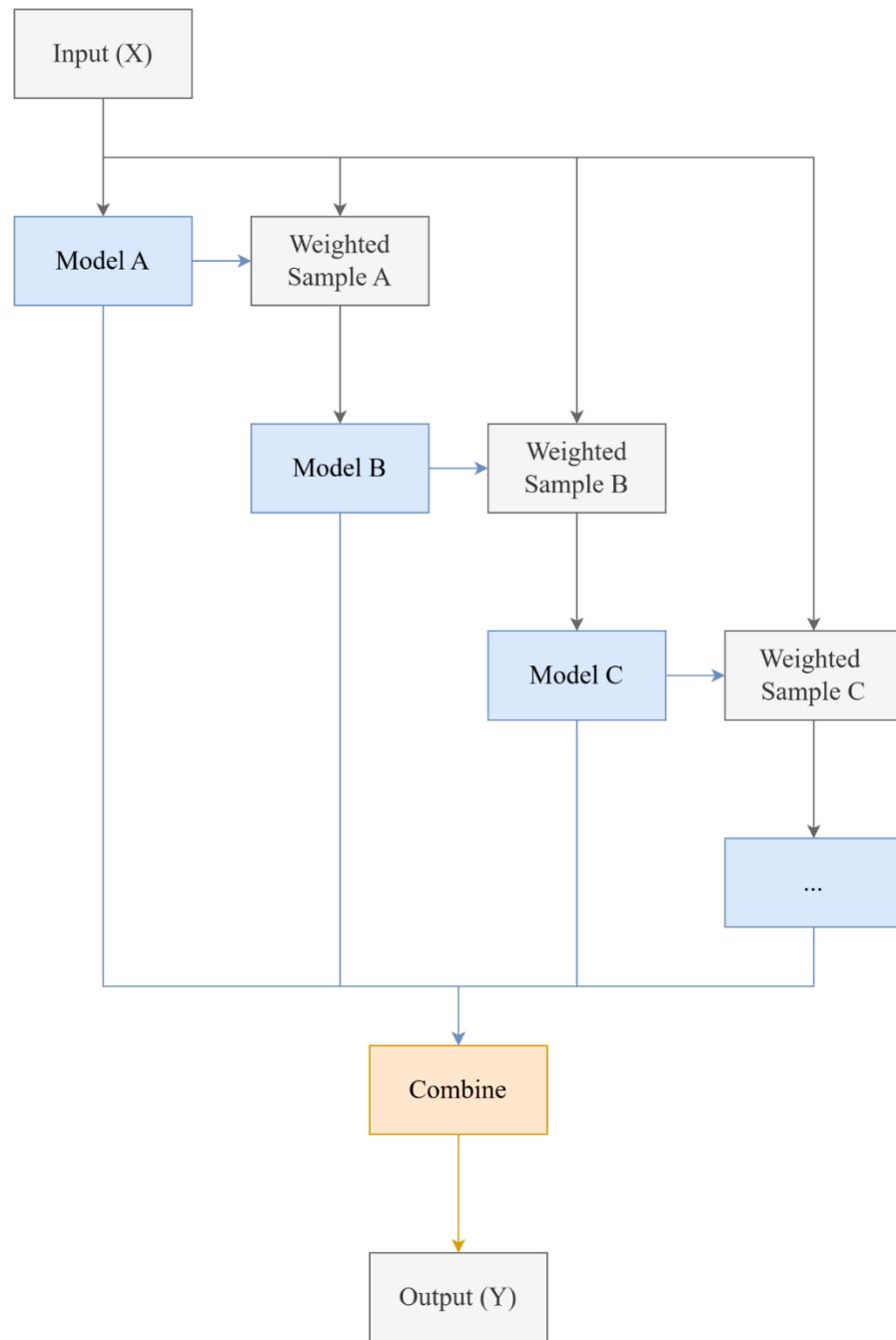
Với tính linh hoạt và hiệu quả cao, Stacking được xem là một trong những phương pháp phổ biến và mạnh mẽ nhất.

2.4.4 Boosting

Boosting là một kỹ thuật học máy tổng hợp nhằm nâng cao hiệu suất của mô hình dự đoán thông qua việc kết hợp tuần tự nhiều mô hình học yếu với nhau. Mục đích cốt lõi của Boosting là xây dựng một mô hình học máy mạnh mẽ bằng cách lần lượt huấn luyện các mô hình yếu hơn, trong đó mỗi mô hình mới được thêm vào sẽ cố gắng sửa các lỗi dự đoán mà các mô hình trước đã mắc phải.

Trong quá trình thực hiện Boosting, tập dữ liệu huấn luyện ban đầu không thay đổi, nhưng thuật toán sẽ điều chỉnh trọng số của các mẫu dữ liệu sau mỗi vòng huấn luyện. Điều này giúp mô hình tổng hợp cải thiện khả năng dự đoán đối với những trường hợp khó. Dù từng mô hình riêng lẻ có độ chính xác không cao, nhưng khi

được kết hợp lại bằng các kỹ thuật như trung bình có trọng số hoặc bỏ phiếu có trọng số, chúng có thể tạo ra một mô hình có hiệu suất vượt trội hơn trước.



Hình 2.6: Sơ đồ phương pháp Boosting

Một trong những thuật toán Boosting tiêu biểu là AdaBoost (Adaptive Boosting), được xem là bước tiến quan trọng giúp Boosting trở thành một phương pháp tổng hợp

hiệu quả. Sau đó, nhiều biến thể mới hơn đã được hình thành và phát triển, bao gồm Gradient Boosting Machines, XGBoost, LightGBM, và CatBoost.

2.5 Các mô hình thuật toán dự đoán

2.5.1 Random Forest

2.5.1.1 Khái niệm

Random Forest [9] là một thuật toán học máy mạnh mẽ, được phát triển bởi Leo Breiman và Adele Cutler và thuộc nhóm mô hình học có giám sát (Supervised Learning). Thuật toán này hoạt động bằng cách xây dựng một khu rừng gồm nhiều cây quyết định, sau đó kết hợp đầu ra của các cây để đưa ra một dự đoán chính xác và ổn định hơn. [10] Đây là một mô hình thuộc phương pháp ensemble learning, cụ thể là sử dụng kỹ thuật bagging, nghĩa là kết hợp nhiều mô hình học độc lập để nâng cao hiệu suất tổng thể.

Sự phô biến rộng rãi của Random Forest đến từ tính linh hoạt, dễ sử dụng và hiệu quả cao. Điều này khiến nó trở thành một trong những thuật toán được sử dụng nhiều nhất trong lĩnh vực học máy.

Một điểm nổi bật của Random Forest là khả năng xử lý được cả hai loại dữ liệu, bao gồm: Biến liên tục trong bài toán hồi quy (regression), Biến phân loại trong bài toán phân loại (classification).

Nhờ khả năng xử lý tốt các tập dữ liệu lớn, có nhiều đặc trưng, và giảm thiểu hiện tượng quá khớp dữ liệu, Random Forest được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau.

2.5.1.2 Cách thức hoạt động

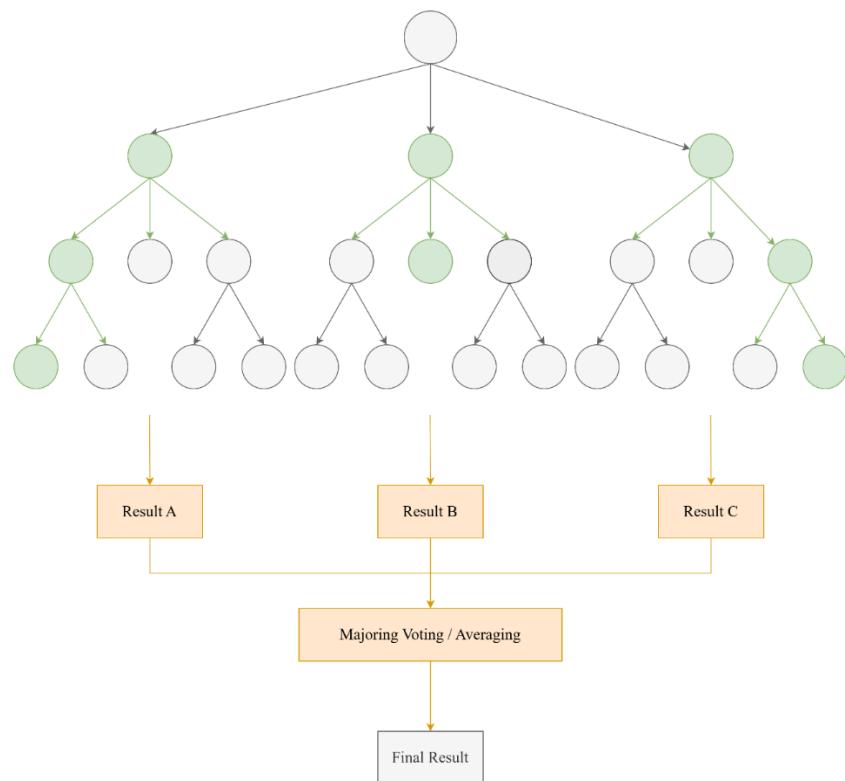
Thuật toán Random Forest xây dựng một tập hợp các cây quyết định độc lập dựa trên việc lựa chọn ngẫu nhiên tập dữ liệu và đặc trưng, sau đó tổng hợp kết quả từ các cây này để đưa ra dự đoán cuối cùng. Cụ thể, quy trình hoạt động của thuật toán gồm 4 bước chính như sau: [9]

Bước 1: Thuật toán bắt đầu bằng cách chọn ngẫu nhiên một phần nhỏ của dữ liệu dựa trên tập dữ liệu gốc. Đồng thời, nó cũng chọn ngẫu nhiên một vài đặc trưng để xây dựng từng cây quyết định. Mỗi cây trong rừng được xây dựng từ một phần dữ liệu và một số đặc trưng được chọn ngẫu nhiên, nên không có cây nào là giống nhau.

Bước 2: Với mỗi tập dữ liệu đã chọn, một cây quyết định sẽ được xây dựng riêng biệt. Các cây này hoạt động độc lập với nhau và có cấu trúc khác nhau do dữ liệu và đặc trưng đầu vào là khác nhau.

Bước 3: Khi đưa vào một dữ liệu mới, mỗi cây trong rừng sẽ dự đoán kết quả riêng của nó. Những kết quả này có thể giống hoặc khác nhau giữa các cây.

Bước 4: Cuối cùng, Random Forest sẽ tổng hợp tất cả các dự đoán. Nếu là bài toán phân loại, thuật toán sẽ chọn kết quả được nhiều cây dự đoán nhất. Đối với bài toán hồi quy, thuật toán sẽ lấy trung bình kết quả dự đoán từ tất cả các cây để ra kết quả cuối cùng.



Hình 2.7: Sơ đồ thuật toán Random Forest

2.5.1.3 *Ưu điểm*

Tính linh hoạt cao: Thuật toán Random Forest có khả năng áp dụng hiệu quả cho cả hai bài toán phân loại (classification) và hồi quy (regression). Ngoài ra, thuật toán còn hỗ trợ tốt trong việc đánh giá mức độ quan trọng của các đặc trưng đầu vào, giúp người dùng hiểu rõ hơn yếu tố nào đang ảnh hưởng nhiều nhất đến mô hình dự đoán.

Giảm thiểu tình trạng quá khớp dữ liệu (Overfitting): Nhờ vào việc kết hợp nhiều cây quyết định, thuật toán Random Forest sẽ giảm thiểu được việc quá khớp dữ liệu so với việc sử dụng một cây riêng lẻ. Khi số lượng cây đủ lớn, mô hình sẽ có xu hướng tổng quát hóa tốt hơn và ít bị phụ thuộc vào dữ liệu huấn luyện cụ thể.

Khả năng xử lý lượng dữ liệu lớn: Random Forest có khả năng mở rộng tốt và xử lý hiệu quả các tập dữ liệu lớn nhờ vào cơ chế xây dựng nhiều cây quyết định riêng lẻ, độc lập nhau. Các cây quyết định có thể thực thi song song, giúp rút ngắn thời gian khi tập dữ liệu có khoảng hàng triệu dòng.

2.5.1.4 *Nhược điểm*

Tốc độ dự đoán chậm: Để đạt được dự đoán chính xác, Random Forest thường cần xây dựng nhiều cây quyết định. Điều này làm tăng mức sử dụng bộ nhớ và có thể khiến mô hình trở nên chậm hơn, đặc biệt khi triển khai trong các hệ thống yêu cầu dự đoán theo thời gian thực. Mặc dù quá trình huấn luyện mô hình diễn ra tương đối nhanh, nhưng việc đưa ra dự đoán sau khi huấn luyện lại mất nhiều thời gian hơn so với một số mô hình khác.

Không giải thích được mối quan hệ giữa các biến: Random Forest là một công cụ dùng để dự đoán chứ không phải là mô hình mô tả. Mô hình hoạt động bằng cách học các mẫu từ dữ liệu để đưa ra dự đoán, nhưng không cung cấp thông tin rõ ràng về mối quan hệ giữa các biến đầu vào, điều này có thể gây khó khăn nếu mục tiêu là phân tích hoặc giải thích kết quả.

2.5.2 XGBoost

2.5.2.1 Khái niệm

Extreme Gradient Boosting [11], còn được viết tắt là XGBoost là một mô hình học máy mạnh mẽ nằm trong phương pháp Ensemble Learning. Đây là một phiên bản tối ưu hóa của phương pháp Gradient Boosting, được thiết kế để giải quyết các bài toán học có giám sát như phân loại (classification) và hồi quy (regression).

XGBoost xây dựng mô hình dự đoán bằng cách kết hợp nhiều mô hình con, thường là các cây quyết định, theo phương thức tuần tự. Trong quá trình huấn luyện, các mô hình yếu được thêm dần vào hệ thống, mỗi mô hình mới có nhiệm vụ khắc phục sai số của các mô hình trước nhằm đưa ra kết quả dự đoán chính xác hơn.

2.5.2.2 Cách thực hiện

XGBoost là một thuật toán Boosting nâng cao, hoạt động bằng cách xây dựng các mô hình học yếu, thường là cây quyết định, theo trình tự và mỗi mô hình mới sẽ tập trung vào việc sửa lỗi của mô hình trước đó. Quá trình huấn luyện của XGBoost có thể được chia thành các bước như sau: [12]

Bước 1: Thuật toán bắt đầu bằng việc tạo ra một mô hình đầu tiên đơn giản, thường là một mô hình cây quyết định. Trong bài toán hồi quy, mô hình ban đầu có thể chỉ đơn giản là dự đoán giá trị trung bình của biến mục tiêu đối với toàn bộ tập dữ liệu.

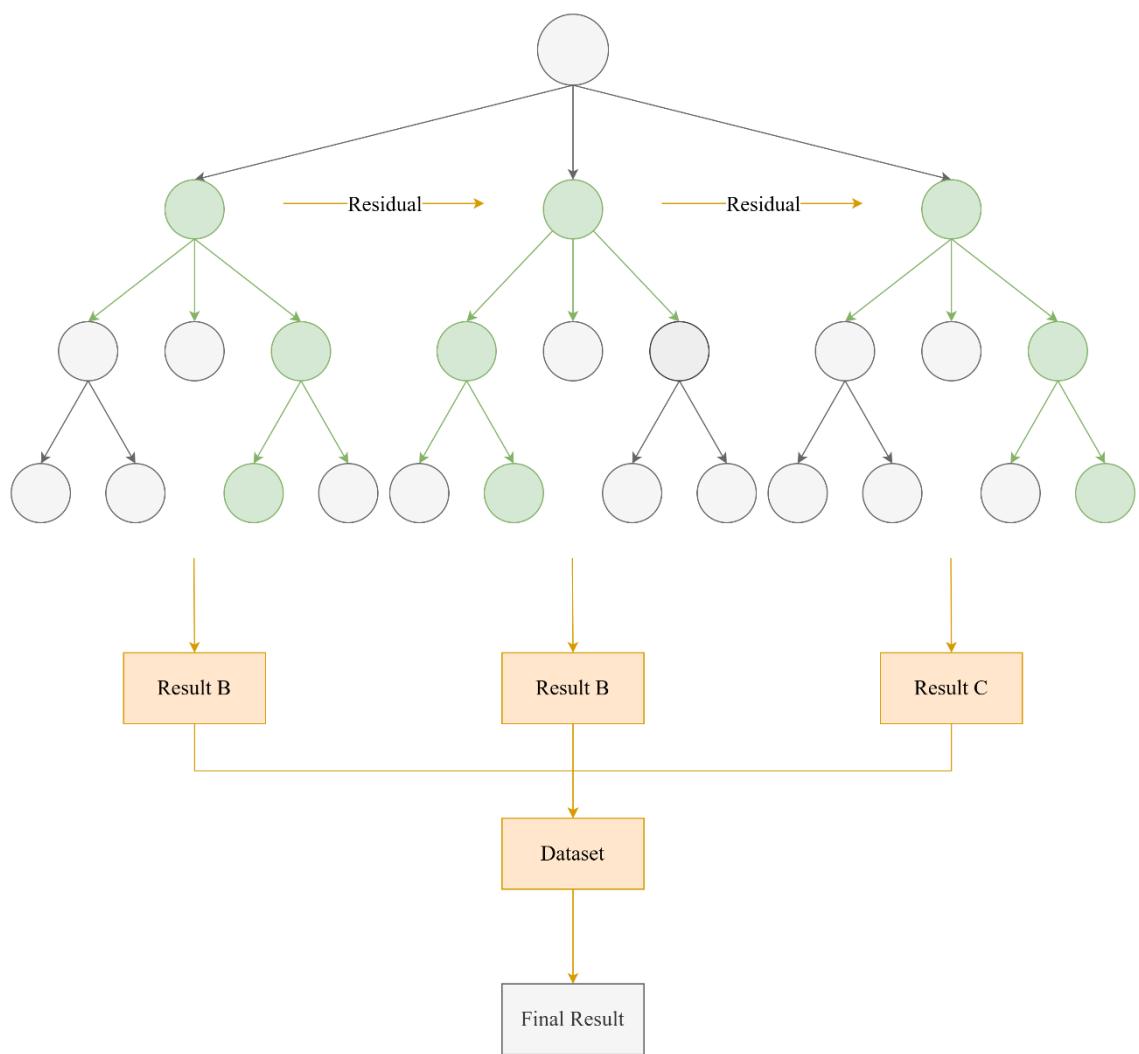
Bước 2: Sau khi mô hình đầu tiên đưa ra kết quả dự đoán, thuật toán sẽ tiến hành tính toán sai số giữa giá trị dự đoán và giá trị thực tế. Sai số này được xem như một biến mục tiêu mới để huấn luyện cho cây quyết định tiếp theo.

Bước 3: Cây quyết định tiếp theo được huấn luyện dựa trên sai số từ cây trước đó. Thay vì dự đoán đầu ra ban đầu, cây mới tập trung vào việc học và sửa các lỗi mà mô hình hiện tại chưa xử lý tốt, giúp cải thiện tính chính xác tổng thể của mô hình.

Bước 4: Quá trình huấn luyện tiếp tục được lặp lại nhiều lần, với mỗi cây mới được thêm vào đều cố gắng giảm bớt sai số còn lại từ mô hình trước đó. Việc lặp lại

này diễn ra cho đến khi mô hình đạt đến một tiêu chí dừng nhất định, chẳng hạn như số lượng cây tối đa hoặc khi sai số không còn đáng kể.

Bước 5: Sau khi toàn bộ số lượng cây được xây dựng, mô hình XGBoost sẽ kết hợp các dự đoán từ tất cả các cây lại với nhau để đưa ra dự đoán cuối cùng. Trong bài toán hồi quy, các dự đoán của các cây thường được cộng lại (hoặc tính trung bình) để tạo thành đầu ra cuối cùng. Đối với bài toán phân loại, XGBoost có thể tính xác suất của mỗi lớp và chọn lớp có xác suất cao nhất.



Hình 2.8: Sơ đồ thuật toán XGBoost

2.5.2.3 *Ưu điểm*

Hiệu suất cao: XGBoost nổi bật với khả năng tạo ra các mô hình có độ chính xác cao, nhờ vào cơ chế boosting, trong đó các cây quyết định được xây dựng tuần tự, mỗi cây mới học từ lỗi của cây trước. Quá trình này giúp mô hình cải thiện hiệu suất dự đoán một cách liên tục và hiệu quả.

Khả năng xử lý bộ dữ liệu lớn: XGBoost hỗ trợ xử lý song song và tối ưu bộ nhớ, cho phép huấn luyện mô hình nhanh chóng ngay cả với những tập dữ liệu có kích thước lớn. Điều này giúp tiết kiệm thời gian và tài nguyên, đồng thời duy trì hiệu suất cao khi làm việc với dữ liệu thực tế phức tạp và đa dạng.

Linh hoạt cao: XGBoost là một thuật toán có tính linh hoạt cao nhờ khả năng áp dụng cho nhiều loại bài toán khác nhau như phân loại (classification), hồi quy (regression), và xếp hạng (ranking). Ngoài ra, XGBoost còn hỗ trợ tùy chỉnh các siêu tham số quan trọng như độ sâu cây, tốc độ học, số lượng cây,..., giúp người dùng dễ dàng điều chỉnh mô hình để phù hợp với dữ liệu và mục tiêu cụ thể.

2.5.2.4 *Nhược điểm*

Yêu cầu tài nguyên tính toán cao: XGBoost là một thuật toán mạnh mẽ nhưng đồng thời cũng đòi hỏi tài nguyên phần cứng đáng kể, đặc biệt khi áp dụng trên các tập dữ liệu lớn hoặc có nhiều đặc trưng. Quá trình huấn luyện mô hình có thể tiêu tốn nhiều bộ nhớ và thời gian, do thuật toán thực hiện nhiều bước như tạo cây, tính toán hàm mất mát và cập nhật gradient liên tục. Điều này có thể gây trở ngại khi làm việc trong môi trường có giới hạn về phần cứng hoặc khi cần triển khai mô hình theo thời gian thực.

Khó khăn trong việc tinh chỉnh các siêu tham số: XGBoost có nhiều siêu tham số như n_estimators, max_depth, learning_rate,... Viết tinh chỉnh các tham số này để đạt được hiệu quả tối ưu đòi hỏi phải thử nghiệm nhiều lần. Quá trình này không chỉ phức tạp mà còn rất tốn thời gian, đặc biệt với những người mới bắt đầu hoặc khi mô hình cần cập nhật thường xuyên theo dữ liệu mới.

2.5.3 Logistic Regression

2.5.3.1 Khái niệm

Hồi quy Logistic (Logistic Regression) [13] là một thuật toán học máy thuộc nhóm học có giám sát, được sử dụng phổ biến trong các bài toán phân loại nhị phân, với biến mục tiêu sẽ trả về một trong hai giá trị, chẳng hạn như ‘True’ hoặc ‘False’. Khác với hồi quy tuyến tính (Linear Regression), vốn dự đoán giá trị liên tục, mô hình hồi quy Logistic nhằm mục đích dự đoán xác suất của một biến phụ thuộc rơi vào một trong hai lớp. Thuật toán sử dụng hàm Sigmoid để chuyển đổi đầu ra tuyến tính thành một xác suất nằm trong khoảng từ 0 đến 1, từ đó giúp thực hiện phân loại.

Với tính đơn giản và dễ triển khai, mô hình hồi quy Logistic thường được sử dụng như một mô hình cơ bản trong nhiều bài toán nghiên cứu và phân loại thực tế.

2.5.3.2 Ưu điểm

Đơn giản và dễ hiểu: Logistic Regression là một trong những mô hình học máy cơ bản nhất với cấu trúc rõ ràng, nên rất thuận tiện trong việc triển khai và giải thích kết quả.

Hiệu quả với dữ liệu tuyến tính: Khi mối quan hệ giữa các đặc trưng và nhãn phân loại có dạng tuyến tính, Logistic Regression thường cho kết quả dự đoán tốt. Trong những trường hợp như vậy, mô hình có thể đạt được độ chính xác cao mà không cần đến các thuật toán phức tạp hơn.

Ít yêu cầu về tài nguyên tính toán: Logistic Regression không đòi hỏi thời gian huấn luyện dài hạn như các mô hình XGBoost hay Neural Network, nhờ đó rất phù hợp với các hệ thống có giới hạn tài nguyên hoặc yêu cầu phản hồi nhanh theo thời gian thực.

Có khả năng đưa ra xác suất dự đoán: Thay vì chỉ trả về nhãn phân loại, Logistic Regression còn cung cấp xác suất tương ứng, giúp người dùng đánh giá mức độ tin cậy của dự đoán. Đây là một lợi thế quan trọng trong các bài toán như phân tích rủi ro, ra quyết định theo ngưỡng xác suất hoặc phân tầng đối tượng. [14]

2.5.3.3 Nhược điểm

Dễ bị ảnh hưởng bởi dữ liệu ngoại lai: Khi các giá trị trong tập dữ liệu có khoảng cách chênh lệch quá lớn (outlier), có thể làm chênh lệch trọng số của mô hình, dẫn đến sai lệch trong quá trình học và giảm hiệu quả dự đoán.

Hiệu suất kém với dữ liệu phi tuyến: Trong các bài toán phức tạp mà xu hướng phân chia lớp là phi tuyến, Logistic Regression không thể học hiệu quả như các mô hình phức tạp hơn như Random Forest hoặc XGBoost. Các mô hình này có khả năng biểu diễn mối quan hệ phi tuyến tốt hơn nhờ cấu trúc linh hoạt hoặc khả năng học đặc trưng phức tạp.

Cần xử lý kỹ lưỡng các đặc trưng đầu vào: Để mô hình Logistic Regression hoạt động tốt, cần chuẩn hóa dữ liệu và lựa chọn đặc trưng phù hợp. Đặc biệt khi các đặc trưng có phân phối không đồng đều hoặc đơn vị đo lường khác nhau, việc không xử lý đúng có thể làm giảm thiểu đáng kể hiệu quả mô hình.

2.5.4 Naive Bayes

2.5.4.1 Khái niệm

Naive Bayes [15] là một tập hợp các thuật toán phân loại học có giám sát dựa trên định lý Bayes. Điểm đặc biệt của nó là giả định rằng các đặc trưng trong dữ liệu đều độc lập với nhau – nghĩa là đặc trưng này không ảnh hưởng gì đến đặc trưng kia. Tên “Naive” xuất phát từ giả định mạnh mẽ rằng tất cả các đặc trưng đều độc lập với nhau, một điều hiếm khi đúng trong thực tế nhưng lại cho kết quả rất tốt trong nhiều trường hợp, đặc biệt hiệu quả với các bài toán phân loại.

2.5.4.2 Ưu điểm

Hiệu suất xử lý cao, phù hợp với dữ liệu lớn: Thuật toán Naive Bayes có thời gian huấn luyện nhanh và không yêu cầu nhiều tài nguyên tính toán. Nhờ vậy, nó đặc biệt thích hợp cho các bài toán có quy mô dữ liệu lớn, giúp tiết kiệm đáng kể thời gian và chi phí triển khai.

Cấu trúc đơn giản, dễ triển khai: Với nguyên lý hoạt động rõ ràng và mô hình hóa đơn giản, Naive Bayes dễ dàng được cài đặt và tích hợp vào các hệ thống hiện có mà không đòi hỏi quá nhiều công đoạn phức tạp.

Phù hợp với dữ liệu rời rạc: Thuật toán thể hiện hiệu quả vượt trội trong việc xử lý các đặc trưng rời rạc như từ ngữ hay nhãn phân loại. Điều này lý giải tại sao Naive Bayes thường được ứng dụng phổ biến trong các bài toán xử lý ngôn ngữ tự nhiên, điển hình là phân loại văn bản và lọc email.

Hoạt động tốt với lượng dữ liệu hạn chế: Ngay cả trong trường hợp dữ liệu huấn luyện không nhiều, Naive Bayes vẫn có thể cho kết quả phân loại khá chính xác, là lựa chọn phù hợp cho các tình huống dữ liệu thu thập còn hạn chế.

2.5.4.3 Nhược điểm

Giả định độc lập giữa các đặc trưng: Thuật toán giả định rằng các đặc trưng đầu vào là độc lập với nhau, điều này hiếm khi đúng trong thực tế. Khi tồn tại sự phụ thuộc giữa các đặc trưng, hiệu quả của mô hình có thể bị suy giảm đáng kể.

Độ tin cậy của xác suất đầu ra thấp: Mặc dù thường đưa ra kết quả phân loại đúng, nhưng xác suất dự đoán mà mô hình cung cấp không phản ánh chính xác mức độ tin cậy. Nguyên nhân chính là do giả định về tính độc lập giữa các đặc trưng dẫn đến sai lệch trong ước lượng xác suất.

Hạn chế trong xử lý dữ liệu liên tục: Naive Bayes không xử lý tốt dữ liệu dạng liên tục nếu không có bước tiền xử lý phù hợp như phân khoang hoặc giả định phân phối xác suất. Do đó, việc chuẩn bị dữ liệu đóng vai trò quan trọng khi áp dụng thuật toán này cho các tập dữ liệu không rời rạc.

2.6 Các chỉ số đánh giá chất lượng mô hình dự đoán

Đối với mục tiêu của bài toán là phân loại đa lớp, việc đánh giá chất lượng của các mô hình học máy là một bước quan trọng nhằm xác định khả năng dự đoán chính xác của mô hình trên dữ liệu thực tế. Việc lựa chọn các chỉ số đánh giá phù hợp không chỉ giúp so sánh hiệu quả của các mô hình một cách khách quan, mà còn hỗ trợ trong

việc lựa chọn thuật toán tối ưu phục vụ cho giai đoạn dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế tại Việt Nam. Khóa luận sẽ tiến hành xem xét các chỉ số đánh giá như Accuracy, Precision, Recall, F1-Score; Ma trận nhầm lẫn; Biểu đồ đường cong ROC và chỉ số AUC trong phân loại đa lớp.

Trong đó, các chỉ số như Precision, Recall và F1-score sẽ được sử dụng để đánh giá độ chính xác của mô hình đối với từng lớp trong tập dữ liệu. Để đảm bảo tính công bằng, đặc biệt trong trường hợp dữ liệu có sự phân phối không đồng đều giữa các lớp với nhau, khóa luận sẽ sử dụng phương pháp macro average, tức là tính trung bình không trọng số của các chỉ số này trên tất cả các lớp.

2.6.1 Ma trận nhầm lẫn

Ma trận nhầm lẫn (Confusion Matrix) là một ma trận dạng bảng tổng hợp, cho phép đánh giá hiệu suất của các thuật toán phân loại bằng cách so sánh, đối chiếu các giá trị trong nhãn dự đoán của mô hình so với các giá trị trong nhãn thực tế của tập dữ liệu. Ma trận sẽ thể hiện số lượng mẫu dự đoán đúng hoặc sai đối với từng lớp, từ đó có thể cung cấp cái nhìn chi tiết về mức độ chính xác, cũng như xu hướng nhầm lẫn giữa các lớp với nhau trong bài toán phân loại.

Các trường hợp phân loại có trong ma trận nhầm lẫn như True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN):

True Positive (TP) đại diện thông tin số lượng mẫu mà mô hình dự đoán chính xác so với giá trị thực tế dương.

False Positive (FP) đại diện thông tin số lượng mẫu mà mô hình dự đoán là dương, nhưng thực tế giá trị chính xác lại là âm. False Positive còn được biết đến là lỗi loại I (Type I Error).

True Negative (TN) đại diện thông tin số lượng mẫu dự đoán đúng một cách gián tiếp. Có nghĩa là số lượng mẫu mà mô hình dự đoán chính xác là âm, mô hình đã nhận dạng đúng các trường hợp không thuộc lớp dương.

False Negative (FN) đại diện thông tin số lượng mẫu dự đoán sai một cách gián tiếp. Có nghĩa là giá trị được mô hình dự đoán là âm, nhưng trong thực tế trường hợp đó là dương. *False Negative* còn được biết đến là lỗi loại II (Type II Error).

Tùy thuộc vào tính chất của bài toán mà ma trận nhầm lẫn được chia thành 2 loại, gồm Ma trận nhầm lẫn cho bài toán phân loại nhị phân (Binary Classification) và Ma trận nhầm lẫn cho bài toán phân loại đa lớp (Multi-Class Classification).

2.6.1.1 Bài toán phân loại nhị phân (Binary Classification)

Ma trận nhầm lẫn trong bài toán phân loại nhị phân có dạng một bảng 2x2, thể hiện số lượng dự đoán đúng và sai của mô hình đối với 2 lớp: lớp dương (Positive) và lớp âm (negative). Cấu trúc của ma trận nhầm lẫn với bài toán phân loại nhị phân được hiển thị như sau:

		Predicted Classes	
		Positive	Negative
Actual Classes	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)
		Positive	Negative

Hình 2.9: Cấu trúc ma trận nhầm lẫn phân loại nhị phân

2.6.1.2 Bài toán phân loại đa lớp (Multi-Class Classification)

Đối với bài toán phân loại đa lớp, khi có nhiều hơn hai nhãn lớp, ma trận nhầm lẫn sẽ được mở rộng thành một bảng dạng NxN, trong đó N là số lượng các lớp. Các

hàng (i) trong ma trận biểu diễn các nhãn thực tế, còn các cột (j) sẽ biểu diễn các nhãn được mô hình dự đoán. Mỗi ô (i, j) trong ma trận biểu thị số lượng mẫu có nhãn thực tế là lớp i nhưng được mô hình dự đoán là lớp j.

Giả sử mô hình phân loại được áp dụng để dự đoán các loại quả, với 3 lớp gồm: A (Táo), B (Cam) và C (Chanh). Sau khi tiến hành dự đoán, kết quả ma trận có thể được hiển thị như sau:

	A	45	3	2
	B	4	40	6
	C	1	5	44
Predicted Classes	A	B	C	
				Actual Classes

Hình 2.10: Ví dụ minh họa ma trận nhầm lẫn phân loại đa lớp

Tại ô (A, A) có giá trị là 45, nghĩa là có 45 mẫu thực tế là quả táo và được mô hình dự đoán chính xác là quả táo.

Tại ô (B, A) có giá trị là 4, nghĩa là thực tế có 4 mẫu là quả cam, nhưng mô hình đã dự đoán nhầm thành hoa táo.

Tại ô (C, B) có giá trị là 5, nghĩa là thực tế có 5 mẫu là quả chanh, nhưng mô hình đã dự đoán nhầm thành quả cam.

2.6.2 Các chỉ số đánh giá cơ bản

Từ các thông số được thể hiện tại ma trận nhầm lẫn, có thể tiến hành tính toán các chỉ số cơ bản nhằm đánh giá và so sánh hiệu suất hoạt động của các mô hình.

2.6.2.1 Accuracy

Chỉ số Accuracy cho biết độ chính xác tổng thể của mô hình dự đoán dựa theo tỷ lệ giữa tổng số mẫu được mô hình dự đoán chính xác, so với tổng số mẫu nằm trong tập kiểm tra. Tuy nhiên chỉ số Accuracy có thể sẽ không phản ánh đúng hiệu quả khi giá trị của các lớp phân bố không đồng đều.

Công thức tính chỉ số Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

2.6.2.2 Precision

Precision được tính riêng cho từng lớp nhằm đánh giá mức độ chính xác của mô hình khi dự đoán các mẫu thuộc lớp đó. Cụ thể, chỉ số Precision đo lường tỷ lệ các mẫu được mô hình dự đoán là thuộc một lớp nhất định và thực tế cũng thực sự thuộc về lớp đó.

Công thức tính chỉ số Macro Precision:

$$\text{Macro Precision} = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

2.6.2.3 Recall

Chỉ số Recall trong phân loại đa lớp phản ánh khả năng của mô hình trong việc phát hiện đầy đủ các mẫu thực sự thuộc về một lớp cụ thể. Chỉ số này được tính toán dựa trên tỷ lệ giữa số mẫu được dự đoán đúng với tổng số mẫu thực tế của lớp đó.

Công thức tính chỉ số Macro Recall:

$$\text{Macro Recall} = \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

2.6.2.4 F1-Score

F1-Score là trung bình điều hòa giữa Precision và Recall đối với từng lớp. Cả 2 chỉ số trên có mối liên hệ chặt chẽ với chỉ số F1-Score. Nếu Precision hoặc Recall thấp thì chỉ số F1-Score cũng sẽ thấp theo.

Công thức tính giá trị trung bình F1-Score của tất cả các lớp như sau:

$$\text{Macro F1 - Score} = \frac{1}{N} \sum_{i=1}^N \frac{2 * \text{Precision}_i * \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

2.6.3 Đường cong ROC và chỉ số AUC trong phân loại đa lớp

2.6.3.1 Đường cong ROC và phương pháp One-vs-Rest (OvR)

Đường cong ROC (Receiver Operating Characteristic Curve) là một công cụ quan trọng trong việc đánh giá và so sánh hiệu suất của các mô hình phân loại, đặc biệt với bài toán phân loại nhị phân. Đường cong sẽ biểu hiện mối quan hệ giữa tỷ lệ dương đúng (True Positive Rate - TPR) với tỷ lệ âm sai (False Positive Rate - FPR) qua các ngưỡng phân loại khác nhau. Trong bài toán phân loại nhị phân, đường cong ROC giúp đánh giá được khả năng phân biệt giữa các lớp dương và âm của mô hình.

Khi so sánh nhiều mô hình với nhau, ROC cho phép so sánh và đánh giá mô hình nào phân biệt tốt hơn giữa các lớp. Một mô hình hoàn hảo sẽ có ROC đi qua điểm (0,1), tức là TPR = 1 và FPR = 0, có nghĩa tất cả các trường hợp dương đều được dự đoán đúng và không có trường hợp âm nào bị dự đoán sai. Ngược lại, một mô hình ngẫu nhiên sẽ có ROC gần như nằm trên đường chéo của biểu đồ, với TPR = FPR. Điều này cho biết mô hình không phân biệt giữa các lớp dương và lớp âm tốt. Bằng cách so sánh đường cong ROC giữa các mô hình với nhau, khóa luận có thể xác định mô hình nào có hiệu suất phân biệt rõ ràng và tốt hơn giữa các lớp.

Trong bài toán phân loại đa lớp sử dụng phương pháp One-vs-Rest (OvR), mô hình sẽ được huấn luyện cho mỗi lớp, trong đó lớp hiện tại sẽ được coi là lớp dương và các lớp còn lại sẽ là lớp âm. Đường cong ROC sẽ được xây dựng cho từng lớp và sau đó tổng hợp kết quả để đánh giá hiệu suất phân loại tổng thể.

Công thức tính True Positive Rate và False Positive Rate như sau:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

2.6.3.1 Chỉ số AUC trong phân loại đa lớp

Diện tích dưới đường cong AUC (Area Under the Curve) là một chỉ số quan trọng dùng để đánh giá khả năng phân biệt giữa các lớp trong mô hình phân loại. Các chỉ số AUC được giải thích như sau:

Giá trị AUC bằng 1 chứng tỏ mô hình càng có khả năng phân biệt hoàn hảo giữa lớp dương và lớp âm.

AUC nằm trong khoảng 0,7 đến 0,9 thể hiện mô hình có khả năng phân biệt tốt giữa lớp dương và lớp âm.

Trong khi giá trị AUC dưới 0,7 phản ánh mô hình có hiệu suất phân loại kém.

Macro Average AUC là cách tính trung bình cộng giá trị AUC của tất cả các lớp mà không xét đến số lượng mẫu trong từng lớp. Phương pháp này đảm bảo mỗi lớp đều có trọng số đóng góp ngang nhau, từ đó mang lại cái nhìn công bằng hơn về hiệu suất tổng thể của mô hình. Macro Average đặc biệt tốt trong các bộ dữ liệu không cân bằng, khi số lượng mẫu giữa các lớp có sự chênh lệch với nhau.

CHƯƠNG 3: THỰC NGHIỆM VÀ PHÂN TÍCH KẾT QUẢ

3.1 Bộ dữ liệu và phương pháp đề xuất

3.1.1 Giới thiệu tập dữ liệu

Biến đổi khí hậu đang ngày càng trở thành một trong những thách thức lớn nhất đối với sự phát triển bền vững, đặc biệt tại các quốc gia có địa hình và khí hậu đa dạng như Việt Nam. Việc theo dõi và phân tích các chỉ số khí hậu theo không gian và thời gian là nền tảng quan trọng giúp đánh giá mức độ ảnh hưởng của biến đổi khí hậu đối với từng khu vực, từ đó hỗ trợ xây dựng các chính sách ứng phó hiệu quả.

Bộ dữ liệu được sử dụng trong bài nghiên cứu thể hiện các chỉ số khí hậu đã thu thập được của 63 tỉnh thành ở Việt Nam trong khoảng thời gian từ tháng 1 năm 2011 đến tháng 12 năm 2023. Thông tin được thể hiện theo dạng bảng, với tổng cộng 25 thuộc tính khác nhau thể hiện thông tin các tỉnh thành, các chỉ số khí hậu và tác động của con người. Thông tin các tháng trong năm của mỗi tỉnh sẽ tương ứng với từng giá trị trong tập dữ liệu.

Bộ dữ liệu được thu thập từ các nguồn khí tượng chính thống, giúp đảm bảo được được tính liên tục và độ chính xác cao. Với độ bao phủ rộng, tập dữ liệu cho phép thực hiện các phân tích chuỗi thời gian, so sánh theo vùng, và đánh giá xu hướng biến đổi khí hậu trên phạm vi toàn quốc.

Qua việc khai thác bộ dữ liệu, nghiên cứu nhằm mục tiêu xác định mức độ và xu hướng ảnh hưởng của biến đổi khí hậu tại từng vùng kinh tế, từ đó đưa ra dự đoán những khu vực nào chịu tác động nặng nề nhất. Kết quả phân tích sẽ là cơ sở để đề xuất các giải pháp thích ứng phù hợp với từng vùng, hướng đến sự phát triển bền vững trong bối cảnh khí hậu đang ngày càng biến đổi nhanh chóng và khó lường.

3.1.2 Bộ dữ liệu sử dụng

3.1.2.1 Cấu trúc và các thuộc tính dữ liệu

Bộ dữ liệu bao gồm 25 thuộc tính riêng biệt, thể hiện qua bảng sau:

STT	Tên thuộc tính	Ý nghĩa
1	Province	Tên tỉnh thành
2	Month	Tháng
3	Year	Năm
4	Region	Vùng kinh tế
5	Region Encode	Vùng kinh tế được mã hóa
6	Latitude	Kinh độ
7	Longitude	Vĩ độ
8	Average Temperature	Nhiệt độ trung bình
9	Max Temperature	Nhiệt độ cao nhất
10	Min Temperature	Nhiệt độ thấp nhất
11	Total Precipitation	Tổng lượng mưa trung bình
12	Relative Humidity	Độ ẩm trung bình
13	Wind Speed	Tốc độ gió
14	Wind Direction	Hướng gió
15	Surface Pressure	Áp suất bề mặt
16	Solar Radiation	Lượng bức xạ mặt trời
17	Soil Moisture	Độ ẩm đất
18	Tree Cover Loss	Diện tích mất rừng
19	Green House Gas	Lượng khí thải nhà kính
20	Index Of Industrial Production	Chỉ số sản xuất công nghiệp

21	Area	Diện tích
22	Average Population	Dân số trung bình
23	Population Density	Mật độ dân số
24	Climate Change Impact Score	Chỉ số ảnh hưởng biến đổi khí hậu
25	Impact Level	Mức độ ảnh hưởng biến đổi khí hậu

Bảng 3.1: Danh sách các đặc trưng trong tập dữ liệu

Danh mục các tỉnh thành được gom nhóm vào các vùng kinh tế trọng điểm ở Việt Nam:

STT	Vùng kinh tế	Tên tỉnh thành
1	TRUNG DU VÀ MIỀN NÚI PHÍA BẮC	Bắc Giang, Bắc Kạn, Cao Bằng, Điện Biên, Hà Giang, Hòa Bình, Lai Châu, Lạng Sơn, Lào Cai, Phú Thọ, Sơn La, Thái Nguyên, Tuyên Quang, Yên Bai
2	ĐỒNG BẰNG SÔNG HỒNG	Bắc Ninh, Hà Nam, Hà Nội, Hải Dương, Hải Phòng, Hưng Yên, Nam Định, Ninh Bình, Quảng Ninh, Thái Bình, Vĩnh Phúc
3	BẮC TRUNG BỘ VÀ DUYÊN HẢI MIỀN TRUNG	Bình Định, Bình Thuận, Đà Nẵng, Khánh Hòa, Nghệ An, Ninh Thuận, Phú Yên, Quảng Bình, Quảng Nam, Quảng Ngãi, Quảng Trị, Thanh Hóa, Thừa Thiên Huế, Hà Tĩnh

4	TÂY NGUYÊN	Đăk Lăk, Đăk Nông, Kon Tum, Gia Lai, Lâm Đồng
5	ĐÔNG NAM BỘ	Bà Rịa Vũng Tàu, Bình Dương, Bình Phước, Đồng Nai, Tây Ninh, Thành phố Hồ Chí Minh
6	ĐỒNG BẰNG SÔNG CỬU LONG	An Giang, Bạc Liêu, Bến Tre, Cà Mau, Cần Thơ, Đồng Tháp, Hậu Giang, Kiên Giang, Long An, Sóc Trăng, Tiền Giang, Trà Vinh, Vĩnh Long

Bảng 3.2: Danh sách các tỉnh thuộc các vùng kinh tế

3.1.2.2 Mục tiêu sử dụng dữ liệu

Bộ dữ liệu được sử dụng nhằm mục tiêu trước tiên là phân tích mức độ ảnh hưởng của biến đổi khí hậu tại các vùng kinh tế của Việt Nam. Cụ thể, dữ liệu sẽ được sử dụng để đánh giá mức độ biến động của các đặc trưng, từ đó có thể tính toán và phân loại mức độ ảnh hưởng theo từng tỉnh thành.

Sau khi xác định và phân loại mức độ ảnh hưởng giữa các tỉnh thành, tập dữ liệu sẽ tiếp tục được sử dụng để xây dựng và huấn luyện các mô hình học máy nhằm đưa ra được kết quả dự đoán mức độ ảnh hưởng tại các vùng kinh tế ở nước ta. Việc đưa ra dự đoán sẽ giúp xác định, đề xuất các giải pháp phát triển bền vững và thích ứng với điều kiện môi trường ngày càng thay đổi trong tương lai.

3.1.3 Phương pháp đề xuất

3.1.3.1 Quy trình nghiên cứu tổng thể

Quy trình nghiên cứu được thực hiện theo các bước sau:

Bước 1: Thu thập dữ liệu

Dữ liệu thô được thu thập và tập hợp thành 2 tập chính, một tập chứa thông tin các tỉnh thành và các chỉ số về biến đổi khí hậu; một tập chứa các thông tin về rừng, khí thải, phát triển công nghiệp và con người. Dữ liệu sẽ được tổng hợp trong giai đoạn 2011 – 2023.

Bước 2: Tiền xử lý dữ liệu

Các tập dữ liệu sẽ được kiểm tra, chuẩn hóa, xử lý các giá trị thiếu, định dạng lại theo cấu trúc bảng dữ liệu tổng hợp và gộp theo khóa tỉnh thành và năm. Một số biến định lượng được phân bố theo tháng (từ giá trị trung bình năm), trong khi các biến ổn định theo thời gian sẽ được giữ nguyên.

Bước 3: Phân tích và trực quan hóa

Tiến hành phân tích mô tả để hiểu rõ xu hướng dữ liệu, xác định những vùng có dấu hiệu bị ảnh hưởng nặng bởi biến đổi khí hậu thông qua biểu đồ, bản đồ nhiệt và thống kê mô tả.

Bước 4: Xây dựng mô hình học máy

Sau khi xác định được mức độ ảnh hưởng của các tỉnh thành, bốn mô hình học máy bao gồm Random Forest, XGBoost, Logistic Regression và Naive Bayes sẽ được huấn luyện và xây dựng mô hình dự đoán.

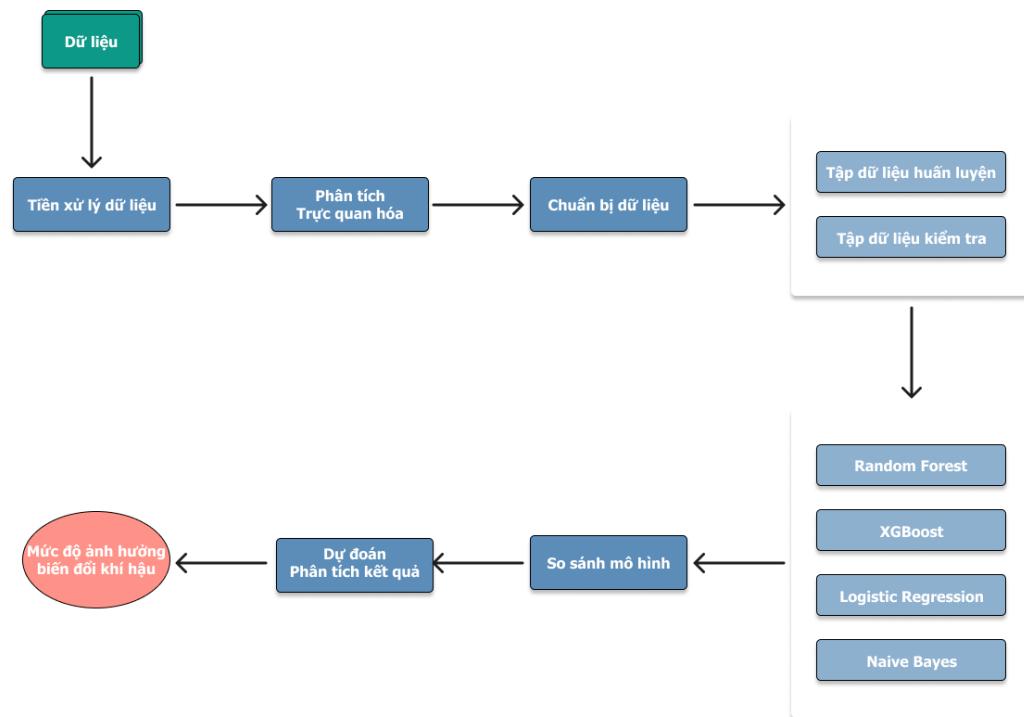
Bước 5: So sánh và đánh giá mô hình

Các mô hình được huấn luyện và đánh giá bằng các chỉ số như Accuracy, Precision, Recall và F1-Score. Ngoài ra còn kết hợp với việc nhận xét Ma trận nhầm lẫn, đường cong ROC và chỉ số AUC để xác định và lựa chọn mô hình có hiệu suất dự đoán cao.

Bước 6: Dự đoán và phân tích kết quả

Sau khi chọn được mô hình tối ưu, thực hiện phân tích và dự đoán mức độ ảnh hưởng của biến đổi khí hậu, sau đó tiến hành đưa ra kết quả dự đoán dựa trên mô hình tốt nhất.

Chi tiết quá trình thực nghiệm sẽ được mô tả thông qua sơ đồ sau:



Hình 3.1: Sơ đồ mô tả tổng quát quá trình thực nghiệm để tài

3.1.3.2 Mô hình học máy sử dụng

Trong bài nghiên cứu, 4 mô hình học máy được lựa chọn sử dụng để xây dựng và huấn luyện gồm Random Forest, XGBoost, Logistic Regression và Naive Bayes. Một số lý do để lựa chọn và sử dụng các thuật toán trên như:

Mô hình Random Forest: Mô hình Random Forest được lựa chọn nhờ vào khả năng xử lý dữ liệu hiệu quả với độ chính xác cao trong các bài toán phân loại đa lớp. Với cơ chế kết hợp nhiều cây quyết định với nhau giúp hạn chế hiện tượng quá khớp, đồng thời tăng cường khả năng tổng quát hóa mô hình.

Mô hình cũng cho phép đánh giá mức độ quan trọng của từng đặc trưng đầu vào, từ đó hỗ trợ việc phân tích các yếu tố có ảnh hưởng lớn đến hiện tượng biến đổi khí hậu, từ đó có thể đề xuất giải pháp ứng phó phù hợp.

Mô hình XGBoost: XGBoost là mô hình học máy theo phương pháp Boosting, thường được sử dụng với các tập dữ liệu phức tạp và đa chiều. Trong bài nghiên cứu, XGBoost được lựa chọn nhờ khả năng học nhanh, có hiệu suất cao và khả năng kiểm soát hiện tượng quá khớp dữ liệu tốt. Mô hình này cũng xử lý tốt dữ liệu thiếu và không yêu cầu chuẩn hóa phức tạp, giúp tối ưu quy trình huấn luyện.

Mô hình Logistic Regression: Dù là một thuật toán tuyến tính đơn giản nhưng được sử dụng trong nghiên cứu để làm nền tảng so sánh với các mô hình có độ phức tạp lớn hơn. Ưu điểm chính của mô hình là dễ diễn giải và thời gian triển khai ngắn, giúp bài toán nghiên cứu hiểu rõ hơn về mối quan hệ giữa các biến độc lập với biến mục tiêu.

Mô hình Naive Bayes: Naive Bayes được sử dụng vì khả năng hoạt động hiệu quả ngay cả khi dữ liệu huấn luyện không quá lớn. Nhờ cơ chế tính toán xác suất rõ ràng, thuật toán cho phép ước lượng khả năng các vùng kinh tế bị ảnh hưởng ở các mức độ khác nhau với tốc độ xử lý nhanh và độ tin cậy nhất định. Dù có giới hạn do giả định độc lập, mô hình Naive Bayes vẫn sẽ mang lại góc nhìn hữu ích khi so sánh với các mô hình học nâng cao hơn.

Các mô hình trên sẽ được huấn luyện trên cùng một tập dữ liệu, bao gồm các yếu tố về khí hậu, các yếu tố môi trường công nghiệp và con người. Việc sử dụng nhiều mô hình khác nhau cho phép so sánh và lựa chọn phương pháp dự đoán tối ưu nhất, phục vụ cho việc phân tích và dự đoán ảnh hưởng của biến đổi khí hậu đến lãnh thổ Việt Nam.

3.1.3.3 Phương pháp đánh giá mô hình

Để đánh giá hiệu quả dự đoán của các mô hình học máy được sử dụng trong bài nghiên cứu, các chỉ số phổ biến trong bài toán phân loại sẽ được sử dụng như Accuracy, Precision, Recall, F1-Score kết hợp với việc so sánh các ma trận nhầm lẫn, đường cong ROC cùng với chỉ số AUC của các thuật toán với nhau.

Ngoài ra, bộ dữ liệu sẽ được chia thành hai tập gồm tập huấn luyện (Training set) 70% và tập kiểm tra (Testing set) 30%. Các mô hình sẽ được huấn luyện dựa trên tập huấn luyện và đánh giá dựa trên tập kiểm tra nhằm đảm bảo tính khách quan và tránh trường hợp quá khớp dữ liệu.

3.1.3.4 Công cụ và ngôn ngữ sử dụng

Trong quá trình nghiên cứu, ngôn ngữ lập trình Python được lựa chọn là ngôn ngữ chính trong việc phân tích dữ liệu và xây dựng các mô hình học máy. Đây là một ngôn ngữ phổ biến trong việc phân tích dữ liệu nhờ vào cú pháp đơn giản, dễ đọc và hệ sinh thái thư viện hỗ trợ phong phú.

Một số thư viện quan trọng được sử dụng trong bài phân tích, bao gồm:

STT	Tên thư viện	Diễn giải
1	pandas	Xử lý dữ liệu dạng bảng
2	numpy	Hỗ trợ tính toán số học
3	matplotlib.pyplot	Trực quan hóa cơ bản
4	seaborn	Trực quan hóa nâng cao

5	sklearn.model_selection.train_test_split	Chia tập dữ liệu huấn luyện và kiểm tra
6	sklearn.ensemble.RandomForestClassifier	Mô hình phân loại Random Forest
7	xgboost.XGBClassifier	Mô hình phân loại XGBoost
8	sklearn.linear_model.LogisticRegression	Mô hình hồi quy Logistic Regression
9	sklearn.naive_bayes.GaussianNB	Mô hình phân loại Naive Bayes
10	sklearn.metrics.classification_report	Báo cáo hiệu suất mô hình
11	shapely.geometry.shape	Phân tích hình học không gian
12	geopandas.GeoDataFrame	Xử lý và hiển thị dữ liệu bản đồ

Bảng 3.3: Danh sách thư viện sử dụng trong quá trình phân tích dữ liệu

Toàn bộ quá trình phân tích và dự đoán trong khóa luận được thực hiện trên nền tảng Google Colab, là một môi trường lập trình trực tuyến do Google phát triển, hỗ trợ ngôn ngữ Python. Google Colab cho phép người dùng viết, thực thi mã nguồn, lưu trữ và chia sẻ trực tiếp trên Google Drive, đồng thời còn tích hợp sẵn nhiều thư viện phục vụ cho việc phân tích dữ liệu.

3.1.4 Quy trình xây dựng mô hình

Quy trình xây dựng mô hình học máy dự đoán được thực hiện qua 5 bước sau:

Bước 1: Chuẩn bị dữ liệu đầu vào

Thực hiện các thao tác đưa tập dữ liệu về khoảng giá trị nhất định, sau đó tính toán chỉ số trung bình và tiến hành việc phân lớp mức độ ảnh hưởng khí hậu của từng tinh thành theo từng mốc thời gian.

Bước 2: Chia tập dữ liệu

Bộ dữ liệu sẽ được chia thành hai phần, bao gồm: Tập huấn luyện (Training Set) gồm 70% dữ liệu và tập kiểm tra (Testing Set) gồm 30% dữ liệu. Tập kiểm tra là phần dữ liệu không được sử dụng trong quá trình huấn luyện, nhằm đánh giá khách quan hiệu quả mô hình dựa trên phần dữ liệu mới.

Bước 3: Xây dựng mô hình học máy

Các mô hình học máy được lựa chọn sẽ được huấn luyện dựa trên tập dữ liệu huấn luyện (Training Set). Trong quá trình huấn luyện, có thể thực hiện tinh chỉnh thông tin các tham số nhằm giúp nâng cao độ chính xác và khả năng tổng quát của các mô hình.

Bước 4: So sánh và đánh giá mô hình

Sau khi thực hiện huấn luyện mô hình, các mô hình sẽ được đánh giá dựa trên tập kiểm tra (Testing Set) thông qua các chỉ số đánh giá như Accuracy, Precision, Recall và F1-Score, kết hợp với việc nhận xét các ma trận nhầm lẫn, đường cong ROC và chỉ số AUC.

Bước 5: Dự đoán và phân tích kết quả

Kết quả dự đoán của bài toán sẽ được thực hiện dựa trên mô hình học máy tốt nhất được lựa chọn. Mục tiêu của dự đoán là giúp so sánh được mức độ ảnh hưởng của biến đổi khí hậu của từng khu vực với nhau, từ đó có thể có cái nhìn tổng quan hơn và đề xuất các giải pháp phù hợp để hạn chế sự gia tăng biến đổi khí hậu ở các vùng.

3.2 Thực nghiệm và phân tích kết quả

3.2.1 Bộ dữ liệu nghiên cứu

Bộ dữ liệu được sử dụng trong bài nghiên cứu là sự kết hợp giữa hai bộ dữ liệu được thu thập từ nhiều nguồn khác nhau, bao gồm các tổ chức có chuyên môn cao về khí hậu, môi trường và hoạt động phát triển kinh tế. Nguồn gốc các đặc trưng trong bộ dữ liệu như sau:

Nhóm đặc trưng địa lý ('Province', 'Latitude', 'Longitude'): Được thu thập từ Nominatim, một dịch vụ mã hóa cung cấp các thông tin địa lý dựa trên OpenStreetMap (OSM). [16]

Nhóm đặc trưng khí hậu ('Average Temperature', 'Max Temperature', 'Min Temperature', 'Total Precipitation', 'Relative Humidity', 'Wind Speed', 'Wind Direction', 'Surface Pressure', 'Solar Radiation', 'Soil Moisture'): Được thu thập từ NASA POWER (Prediction Of Worldwide Energy Resources) [17], một trung tâm nghiên cứu trực thuộc NASA, cung cấp các nguồn dữ liệu khí hậu, năng lượng và môi trường trên phạm vi toàn cầu.

Nhóm đặc trưng môi trường ('Tree Cover Loss', 'Green House Gas'): Được thu thập từ Global Forest Watch [18], nền tảng giám sát rừng toàn cầu do Viện Tài nguyên Thế giới (World Resources Institute) phát triển, cung cấp các thông tin về mất rừng và khí thải nhà kính.

Nhóm đặc trưng công nghiệp, nhân khẩu học ('Index of Industrial Production', 'Area', 'Average Population', 'Population Density'): Được thu thập từ Tổng cục Thống kê Việt Nam (General Statistics Office of Vietnam) [19], một cơ quan nhà nước thuộc Bộ Kế hoạch và Đầu tư, cung cấp các thông tin thống kê quốc gia về các lĩnh vực kinh tế, xã hội và môi trường.

Về cách thức thu thập dữ liệu, phần lớn dữ liệu được thu thập tự động bằng ngôn ngữ lập trình Python, thông qua việc sử dụng các thư viện hỗ trợ để gọi API từ các trang web chính thức. Đối với một số thuộc tính không có API công khai, dữ liệu

được thu thập thủ công dưới dạng các tệp .csv hoặc .xlsx, sau đó được xử lý và chuẩn hóa để đảm bảo tính nhất quán và phù hợp với cấu trúc dữ liệu tổng thể.

3.2.2 Tiềm xử lý dữ liệu

3.2.2.1 Tiềm xử lý dữ liệu địa lý - khí hậu

a. Giới thiệu tập dữ liệu ban đầu

Tập dữ liệu địa lý - khí hậu ban đầu bao gồm 9828 dòng và chứa 15 đặc trưng tương ứng, được thu thập bằng cách gửi yêu cầu đến API của trang web OpenStreetMap và NASA POWER.

Các đặc trưng trong tập dữ liệu thô ban đầu gồm: ‘Province’, ‘Region’, ‘Month’, ‘Latitude’, ‘Longitude’, ‘Average Temperature’, ‘Max Temperature’, ‘Min Temperature’, ‘Total Precipitation’, ‘Relative Humidity’, ‘Wind Speed’, ‘Wind Direction’, ‘Surface Pressure’, ‘Solar Radiation’, ‘Soil Moisture’.

Thông số tập dữ liệu trước khi thực hiện tiềm xử lý như sau:

	Average Temperature	Max Temperature	Min Temperature	Total Precipitation	Relative Humidity	Wind Speed	Wind Direction	Surface Pressure	Solar Radiation	Soil Moisture
count	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000
mean	24.594001	32.401581	18.331553	4.820337	81.515832	2.981459	129.972477	98.459176	15.985534	0.748202
std	4.240931	3.750281	6.197989	4.881987	7.349054	1.170668	70.804518	2.986345	4.035015	0.134384
min	7.120000	16.590000	-1.660000	0.000000	46.350000	0.740000	0.500000	88.240000	4.130000	0.410000
25%	22.450000	30.270000	14.767500	1.150000	78.867500	2.070000	70.300000	96.780000	13.320000	0.640000
50%	25.960000	32.320000	20.165000	3.580000	83.640000	2.950000	115.900000	99.610000	16.680000	0.750000
75%	27.540000	34.682500	23.180000	7.070000	86.330000	3.690000	187.525000	100.770000	18.820000	0.860000
max	33.240000	43.230000	28.330000	61.040000	93.620000	8.860000	357.600000	102.070000	24.970000	0.990000

Hình 3.2: Giới thiệu tập dữ liệu địa lý - khí hậu ban đầu

b. Kiểm tra dữ liệu

Kiểm tra kiểu dữ liệu:

Các đặc trưng như tên tỉnh (‘Province’) và vùng kinh tế (‘Region’) sẽ có kiểu dữ liệu dạng văn bản (object), thuộc tính tháng (‘Month’) sẽ có giá trị kiểu số nguyên (int), các chỉ số còn lại đều theo kiểu số thực (float). Chi tiết các kiểu dữ liệu theo từng đặc trưng được tổng hợp chi tiết trong bảng sau:

STT	Đặc trưng	Kiểu dữ liệu
1	Province, Region	object
2	Month	int64
3	Latitude, Longitude, Average Temperature, Max Temperature, Min Temperature, Total Precipitation, Relative Humidity, Wind Speed, Wind Direction, Surface Pressure, Solar Radiation, Soil Moisture	float64

Bảng 3.4: Kiểu dữ liệu của tập dữ liệu địa lý - khí hậu

Nhìn chung, kết quả kiểm tra cho thấy tất cả các cột đều có kiểu dữ liệu phù hợp cho việc phân tích và dự đoán.

Kiểm tra dữ liệu thiếu:

Kiểm tra giá trị thiếu là một bước quan trọng trong việc đảm bảo rằng dữ liệu không bị gián đoạn hoặc thiếu sót, có thể ảnh hưởng đến chất lượng phân tích. Kết quả kiểm tra như sau:

STT	Đặc trưng	Giá trị thiếu
1	Province	0
2	Region	0
3	Month	0
4	Latitude	0
5	Longitude	0
6	Average Temperature	0
7	Max Temperature	0

8	Min Temperature	0
9	Total Precipitation	0
10	Relative Humidity	0
11	Wind Speed	0
12	Wind Direction	0
13	Surface Pressure	0
14	Solar Radiation	0
15	Soil Moisture	0

Bảng 3.5: Kiểm tra giá trị thiếu của tập dữ liệu địa lý - khí hậu

Kết quả kiểm tra cho thấy rằng tỷ lệ thiếu dữ liệu trong các cột là không có dữ liệu thiếu (0%). Có thể do đây là bộ dữ liệu được thu thập trực tiếp từ việc truy vấn API từ các trang web như OpenStreetMap và NASA POWER.

Dữ liệu trùng lặp và giá trị bất thường:

Sau khi khóa luận tiến hành kiểm tra, bộ dữ liệu cho ra kết quả không có giá trị nào trùng lặp thông tin.

Trong quá trình phân tích dữ liệu, việc kiểm tra và xử lý các giá trị ngoại lệ (outliers) thường được thực hiện nhằm loại bỏ các điểm dữ liệu bất hợp lý, tránh ảnh hưởng xấu đến kết quả phân tích và mô hình dự đoán. Tuy nhiên, trong phạm vi đề tài này, việc kiểm tra và loại bỏ outlier không được áp dụng vì tập dữ liệu phản ánh tình hình thực tế một cách khách quan. Các giá trị trong cả hai tập dữ liệu đều được thu thập từ các nguồn chính thống và đáng tin cậy. Do đó, những giá trị cao hay thấp bất thường vẫn có thể là phản ánh chính xác thực trạng của từng tỉnh thành trong từng thời điểm cụ thể.

c. Chuẩn hóa dữ liệu

Chuẩn hóa thuộc tính Month thành 2 thuộc tính Year – Month riêng biệt:

Ban đầu, giá trị dữ liệu thời gian được lưu trữ tại thuộc tính có tên ‘Month’ với định dạng ‘yyyyMM’. Tuy nhiên, để thể hiện rõ ý nghĩa của đặc trưng này, cũng như hạn chế sự nhầm lẫn, tên đặc trưng sẽ được đổi thành ‘Time’.

Để phục vụ cho việc phân tích dữ liệu theo từng mốc thời gian cụ thể (theo tháng hoặc theo năm), dựa vào cột ‘Time’ sẽ được tách thành hai thuộc tính riêng biệt bao gồm: tháng (‘Month’) và năm (‘Year’).

Sau khi tách dữ liệu, đặc trưng ‘Time’ sẽ được xóa khỏi tập dữ liệu, đồng thời 2 đặc trưng ‘Month’ và ‘Year’ sẽ được di chuyển lên phía trên bảng dữ liệu để thuận tiện cho việc theo dõi và quan sát.

Gom nhóm tỉnh theo vùng kinh tế:

Danh sách 63 tỉnh thành ở Việt Nam sẽ được gom nhóm vào các vùng kinh tế theo thứ tự bảng chữ cái quốc tế, lần lượt gồm: Bắc Trung Bộ và Duyên hải miền Trung; Trung du và miền núi phía Bắc; Tây Nguyên; Đông Nam Bộ; Đồng bằng sông Cửu Long; Đồng bằng sông Hồng.

Sau khi phân nhóm các tỉnh thành, một đặc trưng mới sẽ được thêm vào nhằm giúp lưu trữ thông tin các khu vực trọng điểm dưới dạng mã hóa. Thuộc tính này sẽ được đặt tên là ‘Region Encode’, phương pháp mã hóa Label Encoding sẽ được áp dụng để tiến hành mã hóa các vùng kinh tế như sau:

STT	Vùng kinh tế	Label Encoding
1	TRUNG DU VÀ MIỀN NÚI PHÍA BẮC	1
2	ĐỒNG BẰNG SÔNG HỒNG	5

3	BẮC TRUNG BỘ VÀ DUYÊN HẢI MIỀN TRUNG	0
4	TÂY NGUYÊN	2
5	ĐÔNG NAM BỘ	3
6	ĐÔNG BẮNG SÔNG CỦU LONG	4

Bảng 3.6: Mã hóa vùng kinh tế theo phương pháp Label Encoding

d. Dữ liệu sau khi xử lý

Tập dữ liệu địa lý - khí hậu sau khi thực hiện các bước tiền xử lý và chuẩn hóa dữ liệu sẽ có 9828 dòng và 17 thuộc tính riêng biệt, bao gồm: ‘Province’, ‘Month’, ‘Year’, ‘Region’, ‘Region Encode’, ‘Latitude’, ‘Longitude’, ‘Average Temperature’, ‘Max Temperature’, ‘Min Temperature’, ‘Total Precipitation’, ‘Relative Humidity’, ‘Wind Speed’, ‘Wind Direction’, ‘Surface Pressure’, ‘Solar Radiation’, ‘Soil Moisture’.

Các thông số trong tập dữ liệu sau khi thực hiện tiền xử lý:

	Average Temperature	Max Temperature	Min Temperature	Total Precipitation	Relative Humidity	Wind Speed	Wind Direction	Surface Pressure	Solar Radiation	Soil Moisture
count	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000	9828.000000
mean	24.594001	32.401581	18.331553	4.820337	81.515832	2.981459	129.972477	98.459176	15.985534	0.748202
std	4.240931	3.750281	6.197989	4.881987	7.349054	1.170668	70.804518	2.986345	4.035015	0.134384
min	7.120000	16.590000	-1.660000	0.000000	46.350000	0.740000	0.500000	88.240000	4.130000	0.410000
25%	22.450000	30.270000	14.767500	1.150000	78.867500	2.070000	70.300000	96.780000	13.320000	0.640000
50%	25.960000	32.320000	20.165000	3.580000	83.640000	2.950000	115.900000	99.610000	16.680000	0.750000
75%	27.540000	34.682500	23.180000	7.070000	86.330000	3.690000	187.525000	100.770000	18.820000	0.860000
max	33.240000	43.230000	28.330000	61.040000	93.620000	8.860000	357.600000	102.070000	24.970000	0.990000

Hình 3.3: Tập dữ liệu địa lý - khí hậu sau khi tiền xử lý

3.2.2.2 Tiền xử lý dữ liệu môi trường, công nghiệp và nhân khẩu học

a. Giới thiệu tập dữ liệu ban đầu

Đối với tập dữ liệu thứ 2 gồm 819 dòng và chứa 8 đặc trưng được thu thập thủ công bằng cách tải về từ các nguồn như Global Forest Watch và Tổng cục Thống kê Việt Nam. Các đặc trưng trong tập dữ liệu thô bao gồm: ‘Province’, ‘Year’, ‘Tree Cover Loss’, ‘Green House Gas’, ‘Index Of Industrial Production’, ‘Area’, ‘Average Population’, ‘Population Density’.

Thông số của tập dữ liệu trước khi tiền xử lý như sau:

	Tree Cover Loss	Green House Gas	Index Of Industrial Production	Area	Average Population	Population Density
count	786.000000	7.950000e+02	756.000000	819.000000	819.000000	819.000000
mean	3388.256312	2.530549e+06	109.644974	5255.102808	1497.253114	502.218559
std	4355.800695	3.410550e+06	14.565106	3650.805072	1366.606488	636.692647
min	0.072174	8.745795e+00	43.300000	822.700000	298.700000	43.300000
25%	98.569727	7.263959e+04	104.800000	2358.900000	865.950000	136.100000
50%	1690.865478	1.087292e+06	108.500000	4701.200000	1205.700000	269.800000
75%	5164.276242	3.791548e+06	112.925000	6871.500000	1628.400000	666.600000
max	29413.874700	2.074131e+07	322.800000	16493.700000	9456.700000	4513.100000

Hình 3.4: Giới thiệu tập dữ liệu môi trường, công nghiệp và nhân khẩu học ban đầu

b. Kiểm tra dữ liệu

Các thao tác kiểm tra dữ liệu sẽ được thực hiện tương tự như với tập dữ liệu ở trên.

Kiểm tra kiểu dữ liệu:

Các đặc trưng như tên tỉnh (‘Province’) sẽ có kiểu dữ liệu dạng văn bản (object), năm (‘Year’) có dữ liệu dạng số nguyên, các đặc trưng còn lại sẽ có giá trị là các số thực (float).

STT	Đặc trưng	Kiểu dữ liệu
1	Province	object
2	Year	int64

3	Tree Cover Loss, Green House Gas, Index Of Industrial Production, Area, Average Population, Population Density	float64
---	--	---------

Bảng 3.7: Kiểu dữ liệu của tập dữ liệu môi trường, công nghiệp và nhân khẩu học

Kiểm tra dữ liệu thiếu:

Số lượng giá trị thiếu trong mỗi đặc trưng như sau:

STT	Đặc trưng	Giá trị thiếu
1	Province	0
2	Year	0
3	Tree Cover Loss	33
4	Green House Gas	24
5	Index Of Industrial Production	63
6	Area	0
7	Average Population	0
8	Population Density	0

Bảng 3.8: Kiểm tra giá trị thiếu của tập dữ liệu môi trường, công nghiệp và nhân khẩu học

Sau khi kiểm tra, kết quả cho thấy trong tập dữ liệu, đặc trưng ‘Tree Cover Loss’ thiếu 33 giá trị (chiếm 4,03%), ‘Green House Gas’ thiếu 24 giá trị (chiếm 2,93%) và đặc trưng ‘Index Of Industrial Production (IIP)’ thiếu 63 giá trị (chiếm 7,69%).

Dữ liệu trùng lặp và giá trị bất thường:

Sau khi kiểm tra, tập dữ liệu trên không chứa các giá trị trùng lặp.

c. Chuẩn hóa dữ liệu

Xử lý giá trị thiếu:

Hiện tại, bộ dữ liệu đang chứa ba đặc trưng thiếu dữ liệu, bao gồm ‘Tree Cover Loss’, ‘Green House Gas’ và ‘Index Of Industrial Production’. Các giá trị thiếu sẽ được điền vào bằng giá trị trung vị (median) của từng nhóm dữ liệu được lấy theo đặc trưng tỉnh (‘Province’).

Cụ thể, bộ dữ liệu sẽ được nhóm lại theo cột ‘Province’, nghĩa là phân chia dữ liệu thành các nhóm tương ứng với từng tỉnh. Sau đó, đối với mỗi nhóm, giá trị thiếu trong các thuộc tính cần xử lý được thay thế bằng giá trị trung vị của nhóm đó. Phương pháp sử dụng trung vị để lấp đầy những giá trị thiếu được lựa chọn vì trung vị ít bị ảnh hưởng bởi các outlier, đảm bảo tính ổn định và chính xác của dữ liệu.

Chia tập dữ liệu từ định dạng tháng sang định dạng năm:

Trong quá trình chuẩn hóa dữ liệu phục vụ cho mô hình dự báo, việc điều chỉnh định dạng thời gian từ dữ liệu theo năm sang dữ liệu theo tháng là bước cần thiết để có thể phân tích xu hướng biến động chi tiết hơn trong từng khoảng thời gian ngắn. Đặc biệt, với các chỉ số như ‘Tree Cover Loss’, ‘Green House Gas’, và ‘Index Of Industrial Production’, việc phân rã dữ liệu năm thành 12 tháng cho phép mô hình học máy nhận diện được các xu hướng theo từng giai đoạn cụ thể trong năm.

Để thực hiện điều này, mỗi giá trị năm được chia thành 12 phần tương ứng với các tháng trong năm. Tuy nhiên, để phản ánh thực tế rằng sự biến động giữa các tháng không hoàn toàn đồng đều, mỗi tháng được gán một hệ số ngẫu nhiên nằm trong khoảng từ 80% đến 120% của giá trị trung bình tháng. Sau đó, các hệ số này được tính toán sao cho tổng giá trị của 12 tháng bằng đúng giá trị của năm ban đầu, nhằm đảm bảo tính chính xác tổng thể của dữ liệu.

Đối với các chỉ số ít có sự biến động theo tháng như ‘Area’, ‘Average Population’ hay ‘Population Density’, các giá trị sẽ được giữ nguyên cố định cho tất cả các tháng trong năm. Điều này phản ánh thực tế rằng các chỉ số này thường chỉ thay đổi theo

năm hoặc theo các mốc thời gian lớn, nên việc lặp lại giá trị theo tháng không làm sai lệch dữ liệu mà còn giúp giữ tính ổn định cho mô hình phân tích.

d. Dữ liệu sau khi xử lý

Tập dữ liệu môi trường, công nghiệp và nhân khẩu học sau khi thực hiện các bước tiền xử lý và chuẩn hóa dữ liệu bao gồm 9828 dòng và 9 thuộc tính riêng biệt như ‘Province’, ‘Year’, ‘Month’, ‘Tree Cover Loss’, ‘Green House Gas’, ‘Index Of Industrial Production’, ‘Area’, ‘Average Population’, ‘Population Density’.

Các thông số trong tập dữ liệu sau khi thực hiện tiền xử lý:

	Tree Cover Loss	Green House Gas	Index Of Industrial Production	Area	Average Population	Population Density
count	9828.000000	9.828000e+03	9828.000000	9828.000000	9828.000000	9828.000000
mean	270.563176	2.045236e+05	9.127884	5255.102808	1497.253114	502.218559
std	362.933040	2.858966e+05	1.567684	3648.761208	1365.841408	636.336202
min	0.004730	5.852862e-01	3.025992	822.700000	298.700000	43.300000
25%	5.655664	3.962039e+03	8.148828	2358.300000	865.700000	136.000000
50%	122.486266	8.192767e+04	9.025471	4701.200000	1205.700000	269.800000
75%	411.622103	3.033588e+05	9.942971	6871.500000	1630.600000	669.000000
max	2876.623106	2.007705e+06	34.291136	16493.700000	9456.700000	4513.100000

Hình 3.5: Tập dữ liệu môi trường, công nghiệp và nhân khẩu học sau khi tiền xử lý

3.2.2.3 Kết hợp tập dữ liệu

Sau khi hoàn tất quá trình kiểm tra và xử lý dữ liệu cho từng bộ dữ liệu riêng lẻ, khóa luận tiến hành bước tiếp theo là gộp các tập dữ liệu này nhằm tạo ra một bảng dữ liệu tổng thể phục vụ cho phân tích các mối quan hệ và xây dựng mô hình dự đoán. Cụ thể, hai tập dữ liệu chính bao gồm tập dữ liệu địa lý - khí hậu và tập dữ liệu liên quan đến môi trường, công nghiệp và nhân khẩu học đã được tổng hợp lại với nhau.

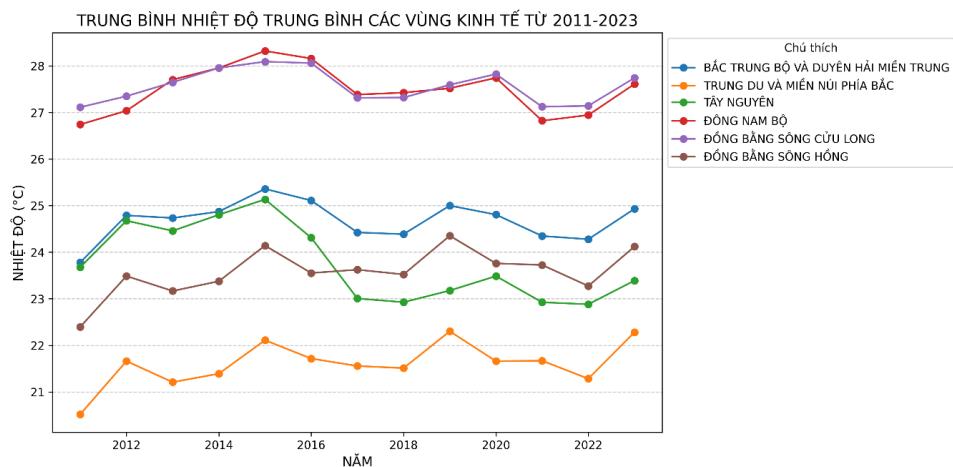
Phương pháp nối dữ liệu theo chiều ngang được áp dụng để gộp dữ liệu dựa trên các cột khóa chung. Các khóa chính được sử dụng bao gồm tên tỉnh (‘Province’), tháng (‘Month’) và năm (‘Year’), nhằm đảm bảo rằng mỗi bản ghi trong tập dữ liệu cuối cùng đều phản ánh chính xác thông tin tương ứng theo từng tỉnh và từng năm. Sau khi gộp thông tin, tập dữ liệu cuối cùng bao gồm 9828 bản ghi và 23 thuộc tính riêng lẻ.

3.2.3 Phân tích và trực quan hóa

Sau khi hoàn tất các bước tiền xử lý, tập dữ liệu sẽ được đưa vào giai đoạn phân tích kết hợp với trực quan hóa. Mục tiêu của giai đoạn này là giúp mô tả rõ ràng các đặc trưng trong tập dữ liệu, hỗ trợ nhận diện các xu hướng tiềm ẩn thông qua việc sử dụng các biểu đồ hay đồ thị trực quan. Giai đoạn này không chỉ giúp tăng khả năng hiểu và diễn giải dữ liệu, mà còn hỗ trợ định hướng cho quá trình xây dựng mô hình dự đoán.

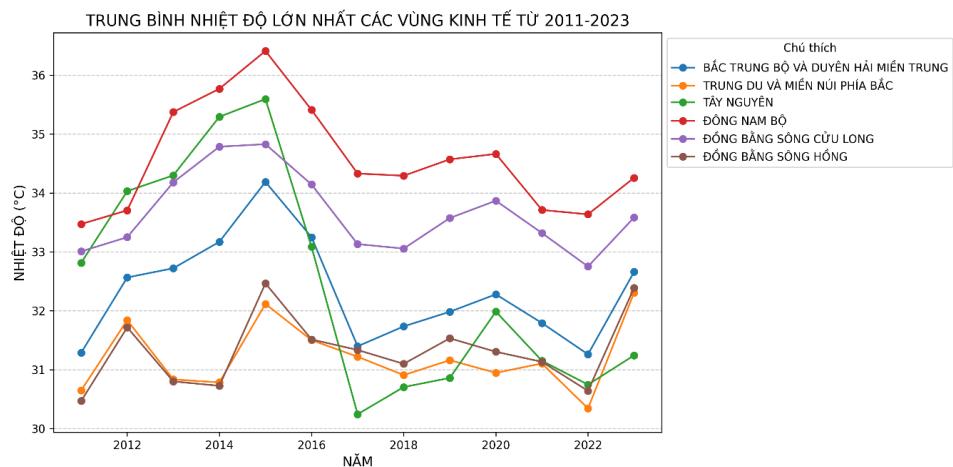
Bên cạnh đó, việc phân tích và trực quan hóa còn đóng vai trò như một bước kiểm tra chất lượng dữ liệu sau tiền xử lý. Qua đó, có thể phát hiện kịp thời các sai lệch, giá trị bất thường hoặc các vấn đề tiềm ẩn trong dữ liệu, từ đó có những điều chỉnh phù hợp nhằm nâng cao độ chính xác và hiệu quả của mô hình học máy trong các bước tiếp theo.

3.2.3.1 Nhóm đặc trưng nhiệt độ



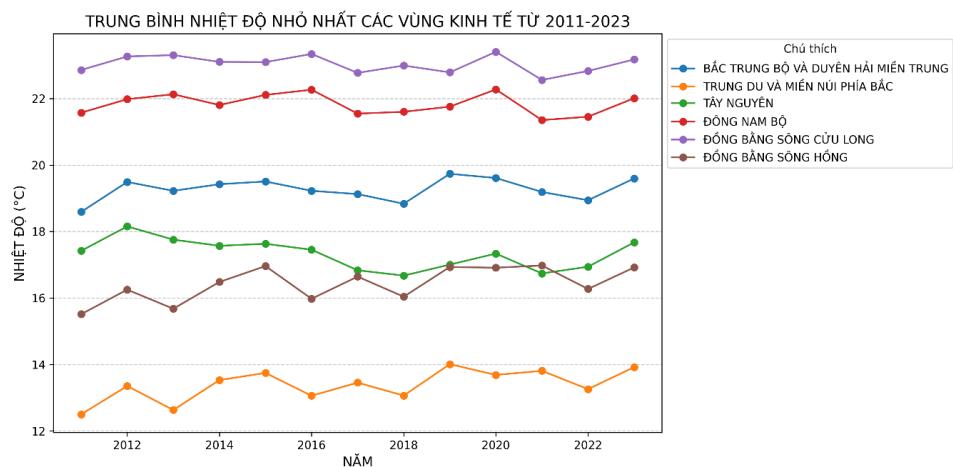
Hình 3.6: Biểu đồ trung bình nhiệt độ trung bình các vùng kinh tế năm 2011-2023

Biểu đồ trên cho thấy xu hướng biến động nhiệt độ rõ rệt ở hầu hết các vùng trong khoảng các năm gần đây. Trong khi Đông Nam Bộ và Đồng bằng sông Cửu Long luôn duy trì mức nhiệt cao trên khoảng 27°C thì các khu vực vốn có khí hậu mát mẻ như Trung du và miền núi phía Bắc, Đồng bằng sông Hồng cũng ghi nhận sự gia tăng đáng kể. Riêng khu vực Tây Nguyên ghi nhận nhiều biến động giữa các năm, nhưng nhìn chung vẫn giữ được mức nhiệt trung bình so với năm 2011 là 23,7°C.



Hình 3.7: Biểu đồ trung bình nhiệt độ lớn nhất các vùng kinh tế năm 2011-2023

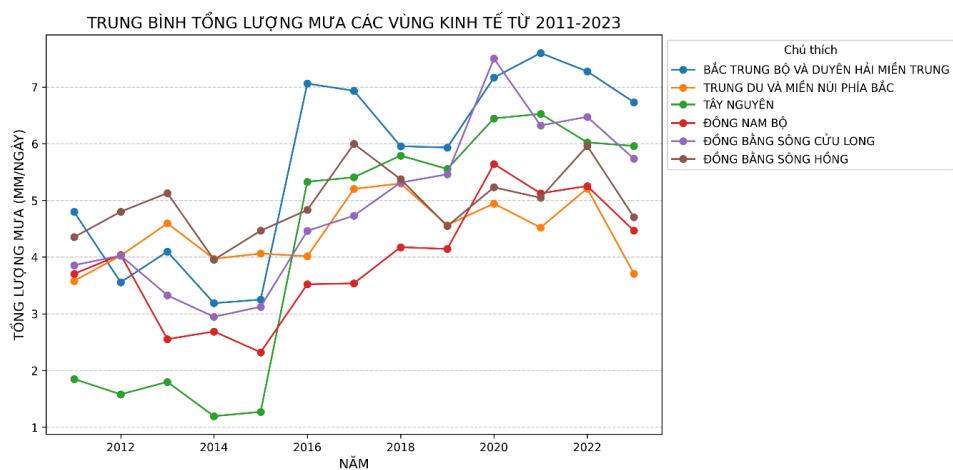
Biểu đồ trên thể hiện xu hướng nhiệt độ tối đa trung bình từ năm 2011 đến 2023 tại các vùng kinh tế, cho thấy sự biến động nhiệt độ rõ rệt ở hầu hết các khu vực, đặc biệt là giai đoạn 2011-2015. Trong đó, Đông Nam Bộ là vùng ghi nhận mức nhiệt tối đa cao nhất, có thời điểm đạt ngưỡng 36,4°C vào năm 2015. Ngược lại, Trung du và miền núi phía Bắc cùng với Đồng bằng sông Hồng là các khu vực có nhiệt độ tối đa thấp hơn, thường dao động quanh khoảng 30-32°C. Đặc biệt, Tây Nguyên là khu vực có khoảng cách nhiệt độ cao nhất khi đạt 35,6°C vào năm 2015, sau đó giảm chỉ còn 30,2°C vào năm 2017.



Hình 3.8: Biểu đồ trung bình nhiệt độ nhỏ nhất các vùng kinh tế năm 2011-2023

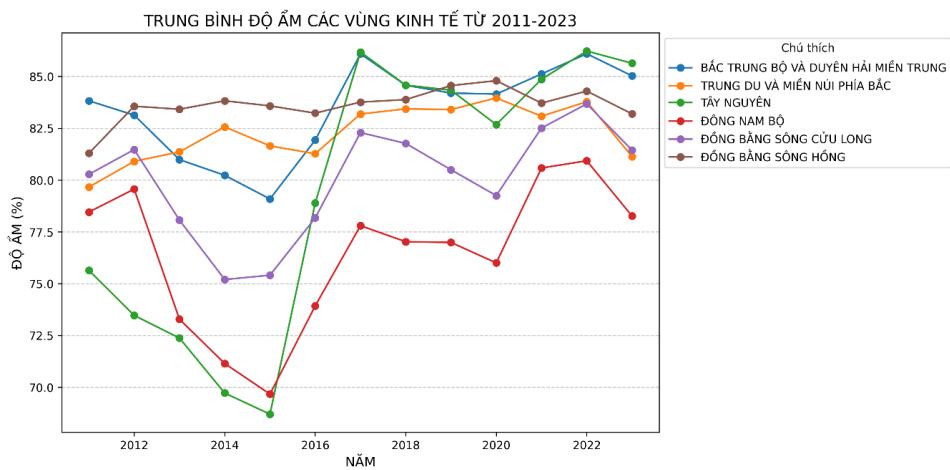
Biểu đồ đường thể hiện sự thay đổi nhiệt độ nhỏ nhất trung bình từ năm 2011 đến 2023, cho thấy xu hướng tăng nhẹ nhưng dao động không quá lớn qua các năm. Trong đó, vùng Đồng bằng sông Cửu Long và Đông Nam Bộ duy trì mức nhiệt tối thiểu cao nhất, luôn dao động trên mức 21°C , phản ánh tính chất khí hậu nhiệt đới ổn định, ít biến động về đêm. Ngược lại, tại khu vực Trung du và miền núi phía Bắc có mức nhiệt trung bình tối thiểu dưới 14°C , cho thấy khí hậu có tính chất cận nhiệt đới, lạnh nhiều về đêm và vào mùa đông. Dù có dao động nhẹ theo từng năm nhưng biểu đồ đã cho thấy sự ổn định tương đối về nền nhiệt tối thiểu tại từng vùng.

3.2.3.2 Nhóm đặc trưng lượng mưa và độ ẩm



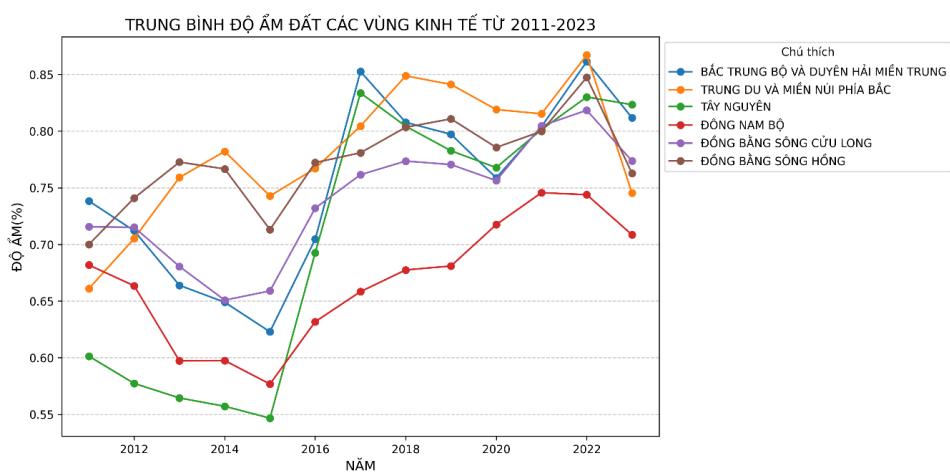
Hình 3.9: Biểu đồ trung bình tổng lượng mưa các vùng kinh tế năm 2011-2023

Biểu đồ trung bình tổng lượng mưa từ năm 2011 đến 2023 cho thấy sự gia tăng rõ rệt về lượng mưa ở hầu hết các vùng kinh tế, đặc biệt từ sau năm 2015 trở đi. Bắc Trung Bộ và Duyên hải miền Trung là khu vực có mức độ tăng trưởng mạnh nhất, từ khoảng 3,2 mm/ngày năm 2014 lên đến 7,6 mm/ngày năm 2021, cho thấy tác động rõ nét của biến đổi khí hậu đến tần suất và cường độ mưa tại khu vực ven biển miền Trung. Riêng Đông Nam Bộ ghi nhận sự tăng trưởng đều và tương đối ổn định, cho thấy khu vực này đang chịu ảnh hưởng bởi sự gia tăng của các trận mưa lớn.



Hình 3.10: Biểu đồ trung bình độ ẩm các vùng kinh tế năm 2011-2023

Từ năm 2011 đến 2023, độ ẩm trung bình có xu hướng biến động khác nhau giữa các khu vực. Vùng Đồng bằng sông Hồng và Trung du và miền núi phía Bắc duy trì độ ẩm ổn định ở mức cao, dao động trong khoảng 80-85%, cho thấy khí hậu ít biến động và tương đối ẩm ướt quanh năm. Ngược lại, các khu vực phía Nam có xu hướng biến động mạnh hơn. Tây Nguyên là vùng có sự thay đổi rõ rệt nhất khi độ ẩm trung bình giảm xuống mức thấp nhất 68,7% vào năm 2015, sau đó tăng vọt lên đến 86,2% vào năm 2022. Tương tự, Đồng Nam Bộ cũng ghi nhận sự gia tăng đáng kể về độ ẩm từ mức dưới 75% trong giai đoạn 2013-2015, sau đó tăng lên đến hơn 80% sau năm 2020.

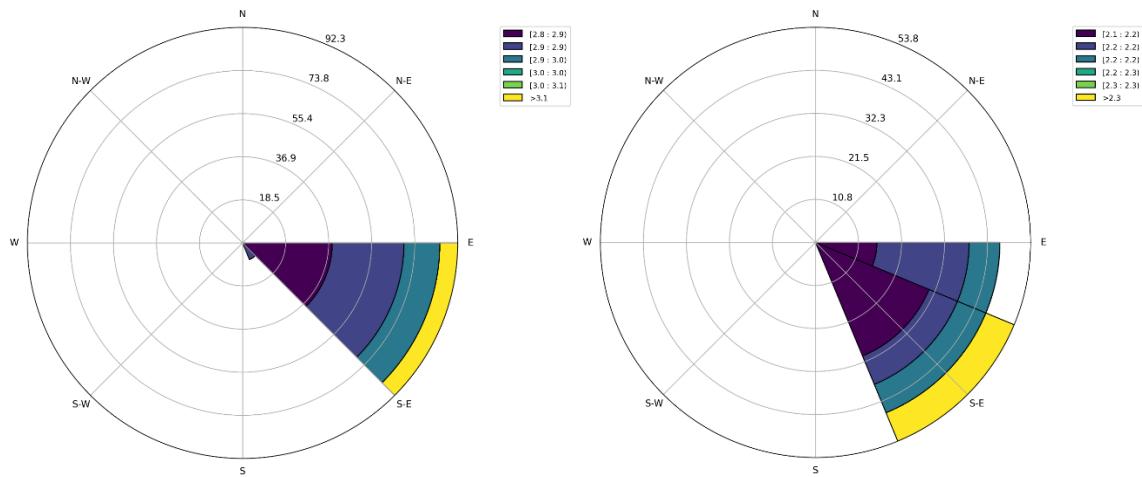


Hình 3.11: Biểu đồ trung bình độ ẩm đất các vùng kinh tế năm 2011-2023

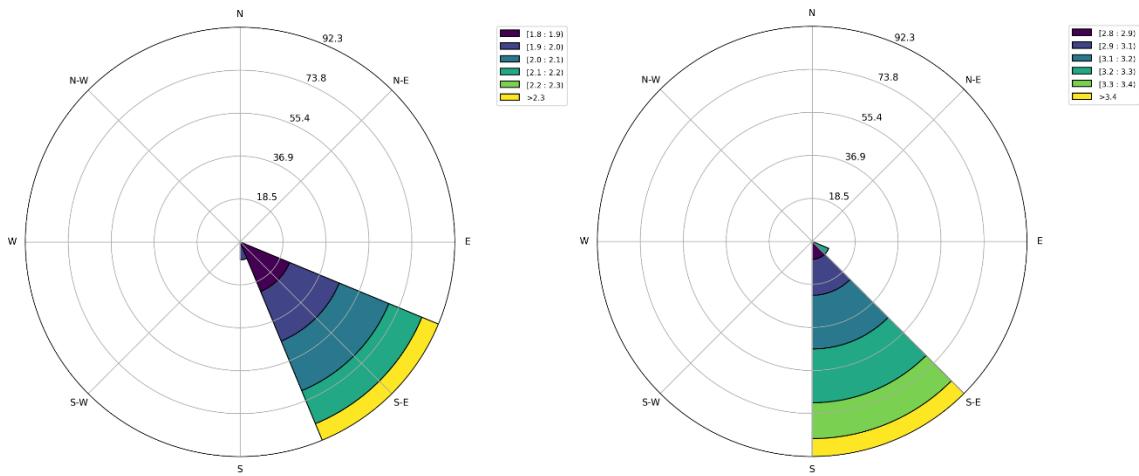
Dựa vào biểu đồ trung bình độ ẩm đất, có thể thấy ở hầu hết các khu vực đều có xu hướng tăng cao, đặc biệt là giai đoạn sau năm 2015. Một số khu vực như Tây Nguyên và Đông Nam Bộ có độ ẩm đất thấp hơn so với các vùng khác trong những năm đầu từ 2011-2015, nhưng sau đó tăng mạnh và ổn định hơn từ năm 2016 trở đi. Tại các khu vực như Bắc Trung Bộ và Duyên hải miền Trung cùng Đồng bằng sông Hồng có xu hướng giữ độ ẩm đất ở mức cao và ổn định qua các năm. Đáng chú ý, vào các năm 2017 và 2022, nhiều vùng ghi nhận độ ẩm cao nhất trong chuỗi thời gian.

3.2.3.3 Nhóm đặc trưng gió và áp suất

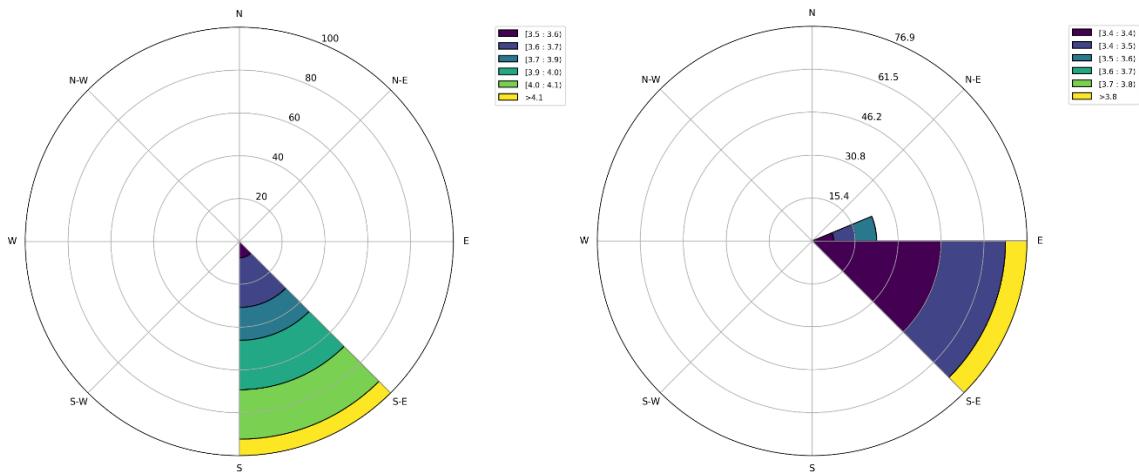
Việc sử dụng biểu đồ hoa gió nhằm mục đích trực quan hóa thông tin về hướng và tốc độ gió tại một khu vực trong một khoảng thời gian xác định. Biểu đồ này hỗ trợ nhận diện các hướng gió chủ đạo, tần suất xuất hiện của từng hướng gió, đồng thời thể hiện phân bố tốc độ gió tương ứng theo từng hướng.



Hình 3.12: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Bắc Trung bộ và Duyên hải miền Trung và vùng Trung du và miền núi phía Bắc



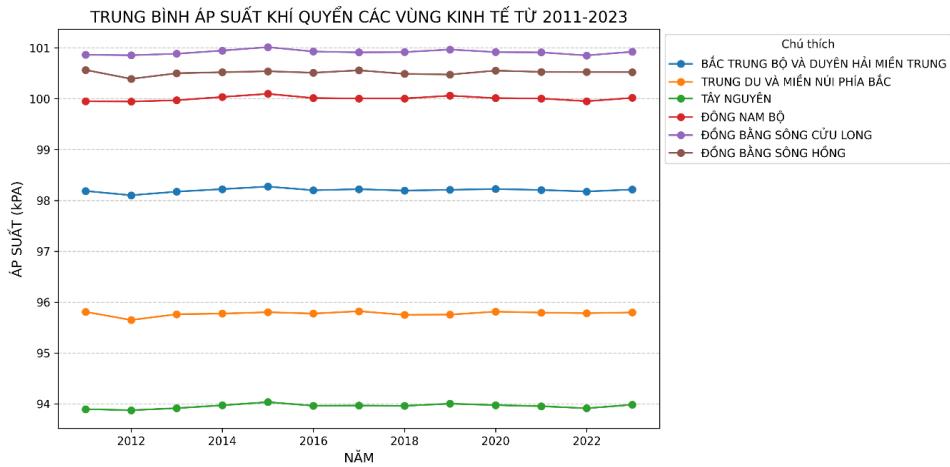
Hình 3.13: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Tây Nguyên và vùng Đông Nam Bộ



Hình 3.14: Biểu đồ thể hiện tốc độ gió và hướng gió của vùng Đồng bằng sông Cửu Long và vùng Đồng bằng sông Hồng

Các biểu đồ hình 3.12, 3.13 và 3.14 thể hiện tổng quan về tốc độ và hướng gió trung bình tại 6 khu vực trên cả nước. Kết quả cho thấy hướng gió chủ đạo tại các khu vực chủ yếu thổi theo hướng Đông và Đông Nam. Tốc độ gió có sự khác biệt giữa các khu vực, trong đó một số vùng như Đồng bằng sông Cửu Long và Đồng bằng sông Hồng ghi nhận tốc độ gió cao và ổn định hơn so với các khu vực khác. Ngược lại, một số khu vực như Tây Nguyên có tốc độ gió trung bình thấp, cho thấy hoạt động gió yếu hơn. Sự khác biệt về hướng và tốc độ gió giữa các khu vực phản ánh đặc điểm khí hậu riêng biệt của từng vùng và là cơ sở quan trọng trong việc đánh

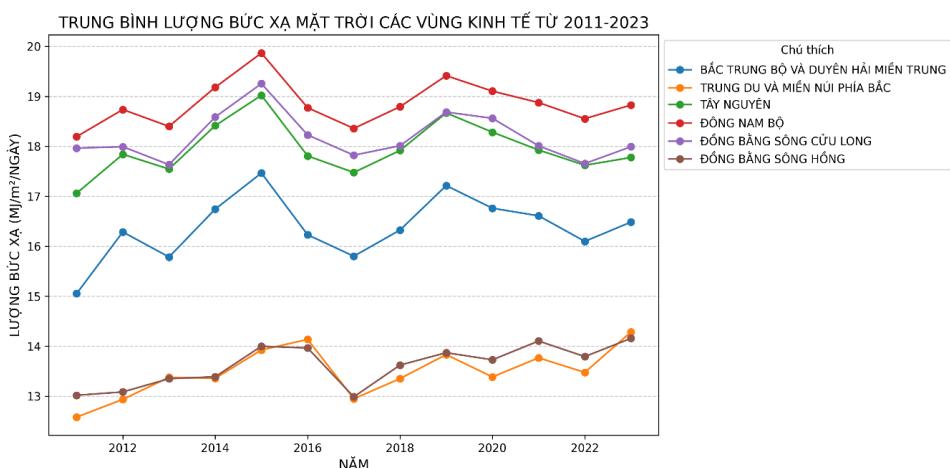
giá tiềm năng khai thác năng lượng gió, cũng như ứng phó với biến đổi khí hậu tại địa phương.



Hình 3.15: Biểu đồ trung bình áp suất khí quyển các vùng kinh tế năm 2011-2023

Biểu đồ trên thể hiện sự biến động trung bình áp suất khí quyển tại các vùng kinh tế ở nước ta. Nhìn chung, áp suất ở các vùng không có sự thay đổi đáng kể theo thời gian, cho thấy yếu tố này khá ổn định trong suốt hơn một thập kỷ qua. Các vùng có địa hình thấp như Đồng bằng sông Hồng, Đồng bằng sông Cửu Long và Đông Nam Bộ duy trì mức áp suất cao nhất tại Việt Nam, xấp xỉ hơn 100 Kilopascal (kPa), trong khi đó khu vực Tây Nguyên ghi nhận áp suất thấp hơn, khoảng 94 kPa.

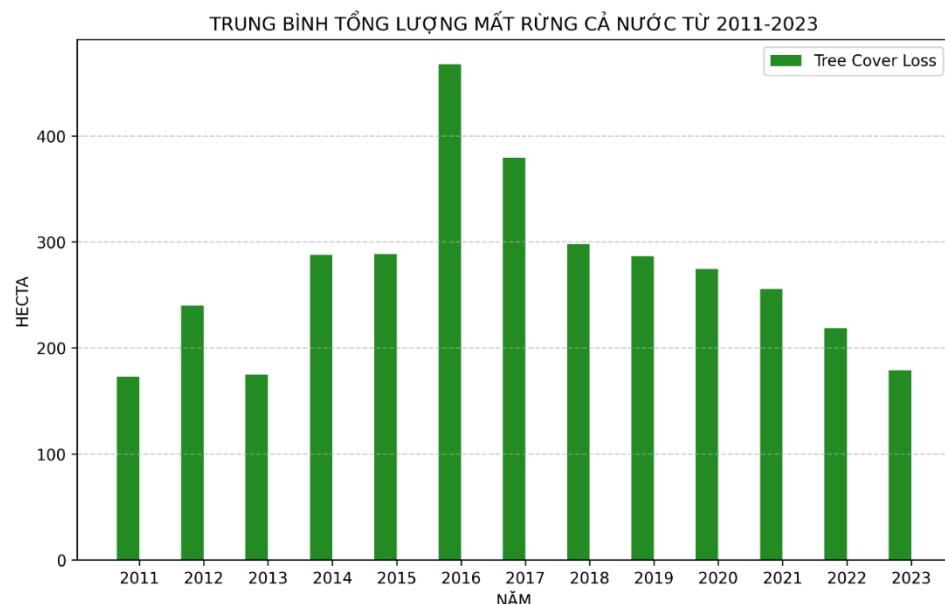
3.2.3.4 Đặc trưng bức xạ mặt trời



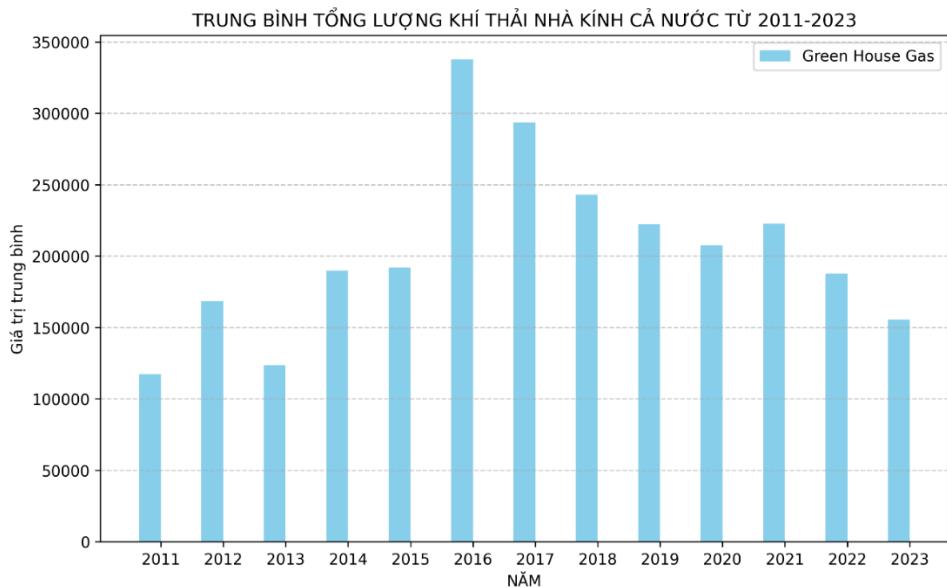
Hình 3.16: Biểu đồ trung bình bức xạ mặt trời các vùng kinh tế năm 2011-2023

Hình 3.16 thể hiện xu hướng thay đổi trung bình lượng bức xạ mặt trời tại các vùng kinh tế từ năm 2011 đến 2023. Trong suốt giai đoạn này, các khu vực như Đông Nam Bộ, Đồng bằng sông Cửu Long và Tây Nguyên ghi nhận lượng bức xạ mặt trời cao nhất cả nước, dao động trong khoảng 17-20 MJ (Megajoule)/m²/ngày, phản ánh đặc điểm khí hậu nhiệt đới khô và ít mây che phủ. Ngược lại, Trung du và miền núi phía Bắc cùng với Đồng bằng sông Hồng có lượng mức bức xạ mặt trời thấp nhất, nằm trong khoảng dưới 14 MJ/m²/ngày, do ảnh hưởng của địa hình và điều kiện thời tiết sương mù nhiều hơn. Mặc dù có một số dao động nhẹ qua các năm, nhưng xu hướng tổng thể cho thấy lượng bức xạ mặt trời khá ổn định ở từng vùng, với mức tăng nhẹ ở một vài khu vực các năm gần đây.

3.2.3.5 Nhóm đặc trưng môi trường



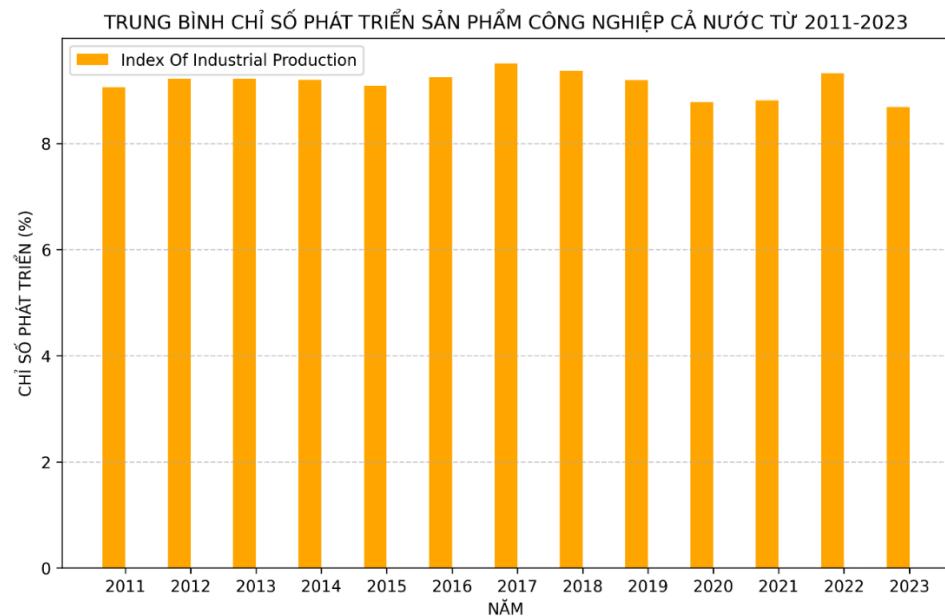
Hình 3.17: Biểu đồ trung bình tổng lượng mất rừng cả nước năm 2011-2023



Hình 3.18: Biểu đồ trung bình tổng lượng khí thải nhà kính cả nước năm 2011-2023

Biểu đồ 3.17 và 3.18 cho thấy xu hướng biến động của lượng mất rừng và khí thải nhà kính trung bình hằng năm trên phạm vi cả nước trong giai đoạn 2011-2023. Cả hai chỉ số đều có xu hướng tăng dần từ năm 2011 và đạt đỉnh vào năm 2016, trong đó lượng mất rừng đạt mức 467,5 hecta và lượng khí thải nhà kính lên tới hơn 337682,4 đơn vị. Đây có thể được xem là thời điểm môi trường chịu áp lực lớn nhất do các hoạt động phát triển kinh tế, chuyển đổi mục đích sử dụng đất và khai thác tài nguyên thiên nhiên. Tuy nhiên, kể từ năm 2017 trở đi, cả hai chỉ số này đều đang giảm dần, phản ánh phần nào hiệu quả của các chính sách bảo vệ môi trường, phát triển bền vững và giảm phát thải. Đến năm 2023, mức độ mất rừng và khí thải đã quay về mức gần tương đương với giai đoạn ban đầu, cho thấy tín hiệu tích cực trong công tác quản lý tài nguyên và ứng phó với biến đổi khí hậu của nước ta.

3.2.3.6 Đặc trưng phát triển công nghiệp

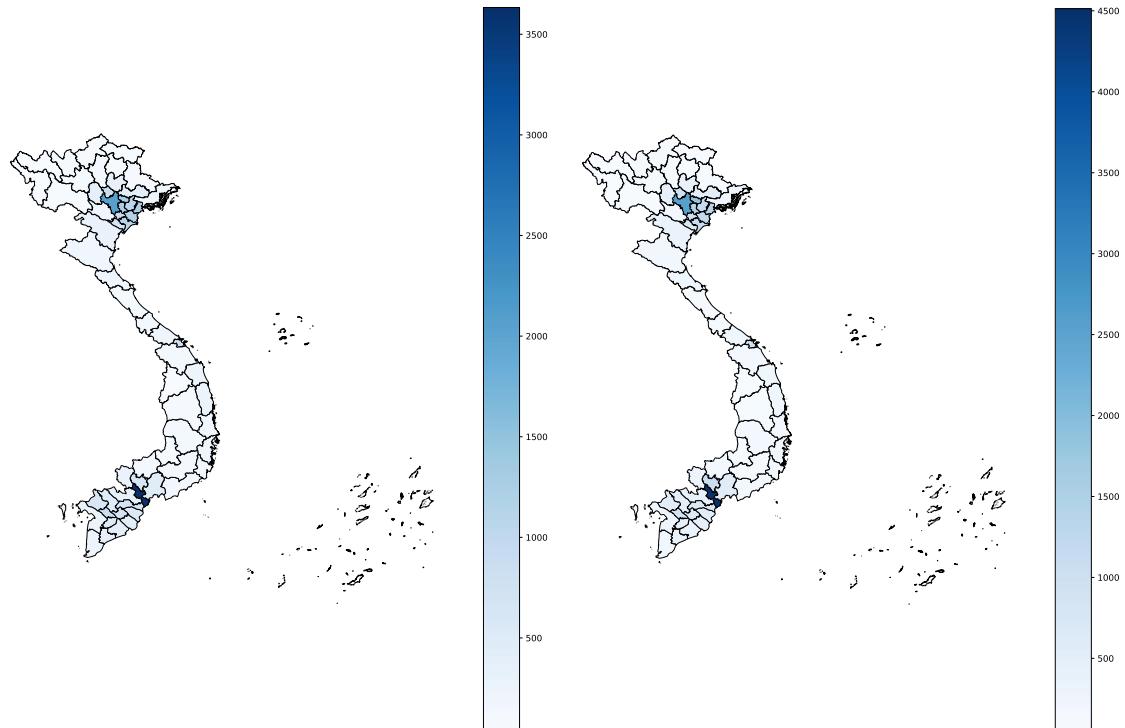


Hình 3.19: Biểu đồ trung bình IIP cả nước năm 2011-2023

Dựa trên biểu đồ thể hiện trung bình chỉ số phát triển sản phẩm công nghiệp của cả nước từ năm 2011 đến năm 2023, có thể thấy rằng mức tăng trưởng của ngành công nghiệp Việt Nam trong hơn một thập kỷ qua duy trì ở mức ổn định, dao động chủ yếu trong khoảng 9-10% mỗi năm.

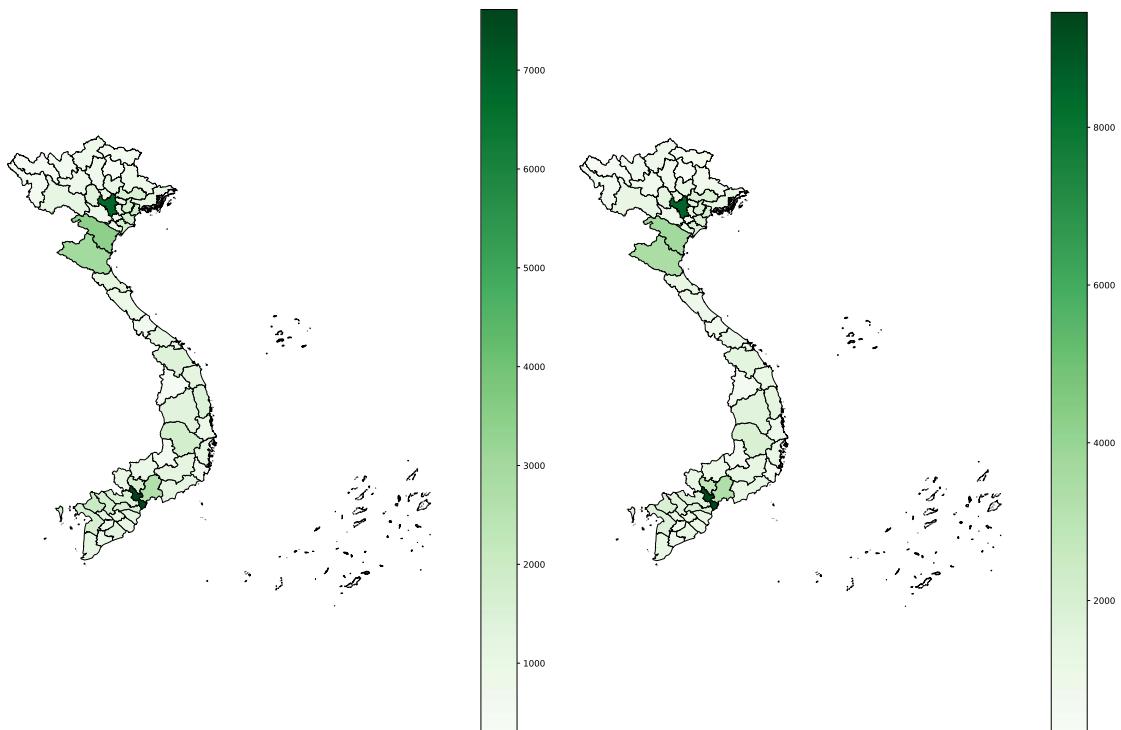
Trong giai đoạn này, không có sự thay đổi quá đột biến nào về chỉ số IIP, điều này phản ánh rằng khả năng sản xuất công nghiệp của Việt Nam nhìn chung đang ở mức bền vững, duy trì được tốc độ tăng trưởng khá đều đặn qua các năm. Tuy nhiên, bên cạnh sự ổn định, dữ liệu cũng cho thấy Việt Nam vẫn chưa đạt được bước đột phá rõ rệt nào trong tăng trưởng sản phẩm công nghiệp. Việc duy trì chỉ số ở mức tiệm cận 10% tuy tích cực, nhưng chưa phản ánh được những cú đột phá mạnh mẽ có thể tạo ra sự chuyển dịch lớn về năng suất, công nghệ, hay giá trị gia tăng trong sản xuất. Điều này gợi mở nhu cầu cần có thêm các chính sách thúc đẩy đổi mới công nghệ, đầu tư vào công nghiệp xanh, và nâng cao năng lực cạnh tranh của sản phẩm công nghiệp trong nước.

3.2.3.7 Nhóm đặc trưng nhân khẩu học và địa lý



Hình 3.20: Bản đồ so sánh mật độ dân số của Việt Nam năm 2011 và năm 2023

Hình 3.20 thể hiện bản đồ so sánh mật độ dân số của các tỉnh, thành tại Việt Nam trong hai năm 2011 và 2023. Bản đồ trên cho thấy các địa phương như Hà Nội, Thành phố Hồ Chí Minh tiếp tục duy trì mật độ dân số ở mức rất cao so với các khu vực khác trong cả hai thời điểm. Đặc biệt, Thành phố Hồ Chí Minh có sự gia tăng mật độ dân số rõ rệt, với năm 2011 là 3633,1 người/km² và năm 2023 là 4513,1 người/km², phản ánh xu hướng dân cư tiếp tục dồn về các đô thị lớn để tìm kiếm cơ hội việc làm và phát triển kinh tế. Ngược lại, các khu vực miền núi và vùng sâu, vùng xa vẫn có mật độ dân số thấp, điển hình như Lai Châu là tỉnh có mật độ dân số thấp nhất cả nước với năm 2011 là 43,3 người/km² và năm 2023 là 54 người/km², cho thấy sự chênh lệch đáng kể trong phân bố dân cư giữa các vùng.



Hình 3.21: Bản đồ so sánh dân số trung bình của Việt Nam năm 2011 và năm 2023

Hình 3.21 thể hiện bản đồ so sánh dân số trung bình của các tỉnh, thành phố tại Việt Nam giữa hai năm 2011 và 2023. Theo quan sát, có thể thấy rõ sự gia tăng dân số tại các khu vực trung tâm kinh tế như Hà Nội, Thành phố Hồ Chí Minh, và một số tỉnh lân cận như Bình Dương, Thanh Hóa, Nghệ An,... Đặc biệt, Thành phố Hồ Chí Minh tiếp tục là địa phương có dân số trung bình cao nhất với 7613,4 nghìn người vào năm 2011 và 9456,7 nghìn người ở năm 2023, cho thấy xu hướng đô thị hóa và tập trung dân cư vẫn đang tiếp diễn mạnh mẽ. Bên cạnh đó, một số tỉnh miền núi phía vẫn duy trì mức dân số trung bình thấp như Bắc Kạn năm 2011 có 298,7 nghìn người và năm 2023 có 326,5 nghìn người, phản ánh sự phân bố dân cư không đồng đều giữa các vùng.

3.2.4 Xây dựng mô hình học máy

3.2.4.1 Chuẩn bị dữ liệu đầu vào

Trong giai đoạn đầu tiên trong quá trình xây dựng mô hình học máy, việc chuẩn bị và xử lý các giá trị của bộ dữ liệu đầu vào là bước quan trọng, đóng vai trò then chốt nhằm đảm bảo chất lượng đầu ra cũng như độ chính xác của các mô hình.

Sau khi hoàn tất thao tác xử lý tập dữ liệu, thông qua việc kiểm tra các thuộc tính trong tập dữ liệu để xác định khoảng giá trị của các đặc trưng, có thể nhận thấy được các giá trị trong tập dữ liệu có mức độ phân bố không đồng đều với nhau. Cụ thể như:

	Average Temperature	Max Temperature	Min Temperature	Total Precipitation	Relative Humidity	Wind Speed	Wind Direction	Surface Pressure	Solar Radiation	Soil Moisture	Tree Cover Loss	Green House Gas	Index Of Industrial Production	Area	Average Population	Population Density
min	7.12	16.59	-1.66	0.00	46.35	0.74	0.5	88.24	4.13	0.41	0.004830	6.595076e-01	2.938035	822.7	298.7	43.3
max	33.24	43.23	28.33	61.04	93.62	8.86	357.6	102.07	24.97	0.99	2893.380909	2.070651e+06	31.740858	16493.7	9456.7	4513.1

Hình 3.22: Khoảng giá trị của các yếu tố đầu vào chưa chuẩn hóa

Việc chuẩn hóa được thực hiện nhằm đưa các giá trị về cùng một thang đo, phục vụ tốt hơn cho quá trình xây dựng và dự đoán bằng các mô hình học máy trong các giai đoạn tiếp theo. Phương pháp Min-Max Scaling sẽ được sử dụng để chuẩn hóa các thuộc tính trong tập dữ liệu. Ngoại trừ các đặc trưng dạng văn bản như ‘Province’, ‘Region’, đặc trưng mã hóa ‘Region Encode’ và các đặc trưng thời gian, tọa độ như ‘Month’, ‘Year’, ‘Latitude’, ‘Longitude’ không cần thực hiện chuẩn hóa do không ảnh hưởng đến tỷ lệ học của các mô hình.

Sau khi chuẩn hóa thành công tập dữ liệu, khóa luận tiếp tục thực hiện tính chỉ số ảnh hưởng của biến đổi khí hậu (‘Climate Change Impact Score’) bằng cách lấy giá trị trung bình cộng của các đặc trưng đầu vào, các đặc trưng được lựa chọn để tính chỉ số ảnh hưởng của biến đổi khí hậu được chia thành các nhóm như:

Nhóm đặc trưng nhiệt độ: Bao gồm các đặc trưng như ‘Average Temperature’, ‘Max Temperature’, ‘Min Temperature’, đây là nhóm đặc trưng cốt lõi, phản ánh xu hướng gia tăng nhiệt độ dài hạn, là dấu hiệu điển hình của

hiện tượng nóng lên toàn cầu. Nhiệt độ cao nhất cho thấy tần suất các đợt nắng nóng, trong khi nhiệt độ thấp nhất thể hiện tần suất cho các đợt rét lạnh kéo dài.

Nhóm đặc trưng lượng mưa và độ ẩm: Gồm các đặc trưng ‘Total Recipitation’, ‘Relative Humidity’ và ‘Soil Moisture’, các yếu tố này có vai trò trong việc đánh giá sự thay đổi chu kỳ nước và trạng thái ẩm của môi trường. Biến động bất thường của lượng mưa và độ ẩm là nguyên nhân trực tiếp dẫn đến các hiện tượng như hạn hán, lũ lụt và suy thoái đất.

Nhóm đặc trưng gió và áp suất: Gồm các đặc trưng ‘Wind Speed’, ‘Wind Direction’, ‘Surface Pressure’, nhóm này phản ánh động lực của bề mặt khí quyển, là yếu tố liên quan đến sự hình thành bão, áp thấp và thay đổi phân bố nhiệt độ, từ đó ảnh hưởng đến cấu trúc thời tiết trong khu vực.

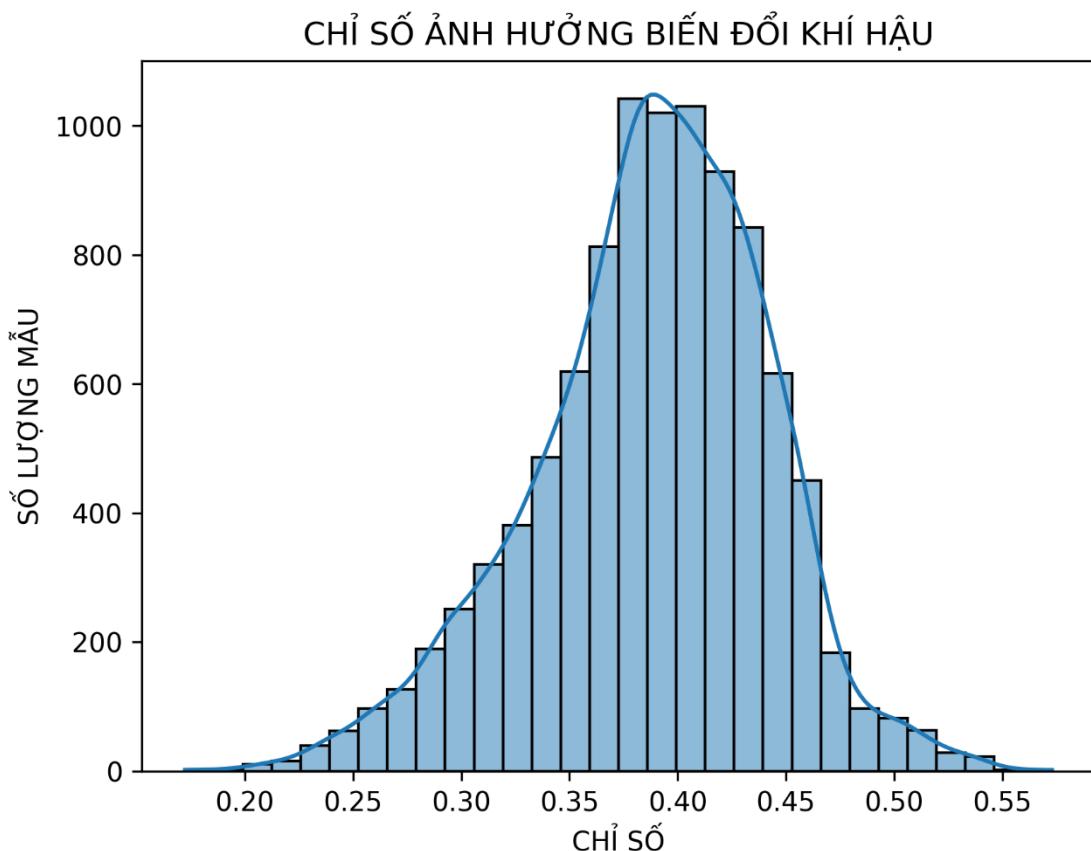
Đặc trưng bức xạ mặt trời: ‘Solar Radiation’ là nguồn năng lượng đầu vào chính của hệ thống khí hậu, chỉ số này thể hiện mức độ tích lũy nhiệt và ảnh hưởng trực tiếp đến xu thế biến đổi khí hậu toàn cầu.

Nhóm đặc trưng môi trường: Gồm các đặc trưng ‘Tree Cover Loss’ và ‘Green House Gas’, phản ánh các tác động của nhân tố con người. Mất rừng và phát thải khí nhà kính là hai nguyên nhân hàng đầu làm suy giảm khả năng điều hòa khí hậu của hệ sinh thái.

Đặc trưng phát triển công nghiệp: Chỉ số ‘Index Of Industrial Production’ dù không phải chỉ số khí hậu trực tiếp, nhưng đã phản ánh mức độ phát triển công nghiệp, là một yếu tố liên quan mật thiết đến phát thải và biến đổi khí hậu.

Nhóm đặc trưng nhân khẩu học và địa lý: Bao gồm các đặc trưng ‘Area’, ‘Average Population’ và ‘Population Density’, nhóm đặc trưng này sẽ thể hiện sức ép từ hoạt động của con người và khả năng chống chịu tự nhiên của từng vùng. Khu vực đông dân và có mật độ cao thường dễ bị tổn thương hơn trước các cú sốc biến đổi khí hậu.

Chỉ số này được sử dụng như một đại lượng tổng hợp, phản ánh mức độ ảnh hưởng bởi biến đổi khí hậu của từng tỉnh thành theo một khoảng thời gian cụ thể. Sau khi thực hiện tính toán, nhằm đánh giá mức độ phân phối của chỉ số trên từng giá trị, một biểu đồ sẽ được tạo mới và được hiển thị như sau:

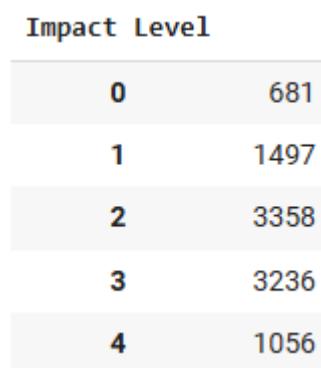


Hình 3.23: Biểu đồ thể hiện mức độ phân bố chỉ số ảnh hưởng của biến đổi khí hậu

Dựa vào kết quả chỉ số ảnh hưởng của biến đổi khí hậu phân bổ theo từng tỉnh thành theo từng mốc thời gian nhất định, để thuận tiện hơn cho giai đoạn phân lớp và dự đoán mức độ ảnh hưởng theo từng khu vực, khóa luận sẽ tiến hành phân nhánh mức độ ảnh hưởng của biến đổi khí hậu thành 5 lớp, bao gồm: Ảnh hưởng rất thấp (0) khi chỉ số ‘Climate Change Impact Score’ bé hơn 0,30; ảnh hưởng thấp (1) khi chỉ số bé hơn 0,35; ảnh hưởng trung bình (2) khi chỉ số bé hơn 0,40; ảnh hưởng cao (3) khi kết quả bé hơn 0,45 và ảnh hưởng rất cao (4) khi kết quả lớn hơn hoặc bằng 0,45.

Kết quả của việc phân nhãm mức độ ảnh hưởng sẽ được lưu vào một đặc trưng mới là Mức độ ảnh hưởng ('Impact Level').

Sau khi tiến hành gán nhãm và phân lớp, tập dữ liệu cho ra kết quả có 681 giá trị ở mức độ ảnh hưởng rất thấp, 1497 giá trị ở mức độ ảnh hưởng thấp, 3358 giá trị ở mức trung bình, 3236 ở mức cao và 1056 giá trị đang ở mức rất cao.



Hình 3.24: Kết quả phân lớp mức độ ảnh hưởng cho tập dữ liệu ban đầu

Dữ liệu sau khi chuẩn hóa và tính toán sẽ được lưu trữ trong một DataFrame mới để đảm bảo tách biệt với dữ liệu gốc, đồng thời hỗ trợ việc kiểm tra lại kết quả chuẩn hóa một cách dễ dàng và nhanh chóng nhất.

3.2.4.2 Chia tập dữ liệu

Trong quá trình xây dựng mô hình học máy, việc phân chia tập dữ liệu thành hai phần riêng biệt là một bước vô cùng cần thiết, nhằm đảm bảo được tính khách quan trong việc đánh giá hiệu suất của mô hình dự đoán. Khó khăn sẽ chia dữ liệu thành hai phần bao gồm: Tập dữ liệu huấn luyện (Training Set) chiếm tỷ lệ 70% trên tổng số dữ liệu, và tập dữ liệu kiểm tra (Testing Set) chiếm tỷ lệ 30% trên tổng số dữ liệu.

Trước khi tiến hành chia tập dữ liệu, cần xác định rõ các biến đầu vào và biến mục tiêu cần có để đạt được kết quả dự đoán như sau:

Biến đầu vào (x): Bao gồm các đặc trưng đã được lựa chọn để tính chỉ số trung bình biến đổi khí hậu như: 'Average Temperature', 'Max Temperature',

‘Min Temperature’, ‘Total Precipitation’, ‘Relative Humidity’, ‘Wind Speed’, ‘Wind Direction’, ‘Surface Pressure’, ‘Solar Radiation’, ‘Soil Moisture’, ‘Tree Cover Loss’, ‘Green House Gas’, ‘Index Of Industrial Production’, ‘Area’, ‘Average Population’, ‘Population Density’.

Biến mục tiêu (y): Là đặc trưng ‘Impact Level’ đại diện cho mức độ ảnh hưởng của biến đổi khí hậu, được phân thành các nhãn như ‘Rất thấp’, ‘Thấp’, ‘Trung bình’, ‘Cao’ và ‘Rất cao’.

Phương pháp Stratified Sampling sẽ được áp dụng nhằm giúp giữ nguyên tỷ lệ mẫu giữa các lớp của biến mục tiêu trong cả tập dữ liệu huấn luyện và tập dữ liệu kiểm tra, đảm bảo rằng mô hình sẽ không bị học lệch khi các lớp trong biến mục tiêu phân bố không đồng đều, từ đó giúp nâng cao độ tin cậy khi ứng dụng các mô hình học máy trong việc dự đoán.

Sau khi chia tập dữ liệu, số lượng mẫu và thuộc tính của tập dữ liệu huấn luyện (x_{train}, y_{train}) là 6879 dòng và 16 đặc trưng, tương tự với tập dữ liệu kiểm tra (x_{test}, y_{test}) sẽ là 2949 dòng và 16 đặc trưng.

Tập dữ liệu huấn luyện sẽ được sử dụng để huấn luyện các mô hình học máy đã đề xuất trước đó. Các mô hình này sẽ được kiểm tra dựa trên tập dữ liệu kiểm tra đã được chuẩn bị từ trước các mô hình này sẽ được đánh giá trên tập kiểm tra đã được tách ra từ trước nhằm kiểm tra khả năng dự đoán và tổng quát hóa của các mô hình. Để so sánh và đo lường hiệu suất giữa các mô hình, các chỉ số đánh giá được sử dụng bao gồm Accuracy, Precision, Recall và F1-Score, đồng thời kết hợp với việc nhận xét ma trận nhầm lẫn, đường cong ROC và chỉ số AUC.

Do bài toán phân loại trong nghiên cứu bao gồm 5 lớp mục tiêu, các chỉ số đánh giá nêu trên sẽ được tính riêng biệt cho từng lớp, sau đó tổng hợp lại bằng phương pháp Macro Average. Phương pháp Macro Average thực hiện việc tính trung bình cộng không trọng số các chỉ số trên toàn bộ các lớp, giúp đánh giá hiệu suất tổng thể của mô hình một cách công bằng, không phụ thuộc vào số lượng mẫu được phân bổ.

Việc sử dụng Macro Average đặc biệt phù hợp khi các lớp mục tiêu phân bố không đồng đều, đảm bảo rằng mỗi lớp đều được đánh giá với mức độ quan trọng như nhau, qua đó phản ánh được khả năng phân loại đồng đều của mô hình đối với tất cả các lớp với nhau.

3.2.4.3 Xây dựng mô hình học máy

Sau khi hoàn tất giai đoạn chuẩn bị bộ dữ liệu đầu vào và phân tách tập dữ liệu thành hai phần gồm tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Ở bước tiếp theo, khóa luận sẽ tiến hành xây dựng các mô hình học máy nhằm thực hiện bài toán dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế tại Việt Nam. Trong khuôn khổ đề tài, bốn thuật toán học máy được lựa chọn để xây dựng mô hình bao gồm Random Forest và XGBoost được khởi tạo với tham số random_state=42, Logistic Regression được khởi tạo với tham số max_iter=1000 và Naive Bayes được cài đặt mặc định, không tinh chỉnh tham số. Các mô hình này được xây dựng và huấn luyện dựa trên tập dữ liệu huấn luyện x_train và y_train .

3.2.5 So sánh và đánh giá mô hình

3.2.5.1 So sánh các chỉ số đánh giá cơ bản

Sau khi thực hiện thao tác xây dựng các mô hình học máy trên tập dữ liệu nghiên cứu, các chỉ số Accuracy, Macro Precision, Macro Recall và Macro F1-Score có kết quả như sau:

	Random Forest	XGBoost	Logistic Regression	Naive Bayes
Precision	88,30%	90,11%	90,50%	67,02%
Recall	85,37%	88,01%	84,36%	64,98%
F1-Score	86,73%	89,01%	86,95%	64,88%

Accuracy	87,15%	89,56%	89,11%	66,87%
-----------------	--------	--------	--------	--------

Bảng 3.9: Các chỉ số đánh giá của các thuật toán sử dụng

Bảng 3.9 thể hiện các chỉ số đánh giá hiệu suất của 4 mô hình học máy được sử dụng trong bài toán phân loại gồm chỉ số Macro Precision, Macro Recall và Macro F1-Score của Random Forest, XGBoost, Logistic Regression và Naive Bayes. Kết quả cho thấy mô hình XGBoost đạt hiệu suất cao nhất với chỉ số Precision là 90,11%, chỉ số Recall là 88,01% và chỉ số F1-Score là 89,01%. Điều này cho thấy XGBoost không chỉ cho ra kết quả nhận diện đúng các trường hợp quan trọng, mà còn duy trì được độ chính xác cao trong mô hình dự đoán.

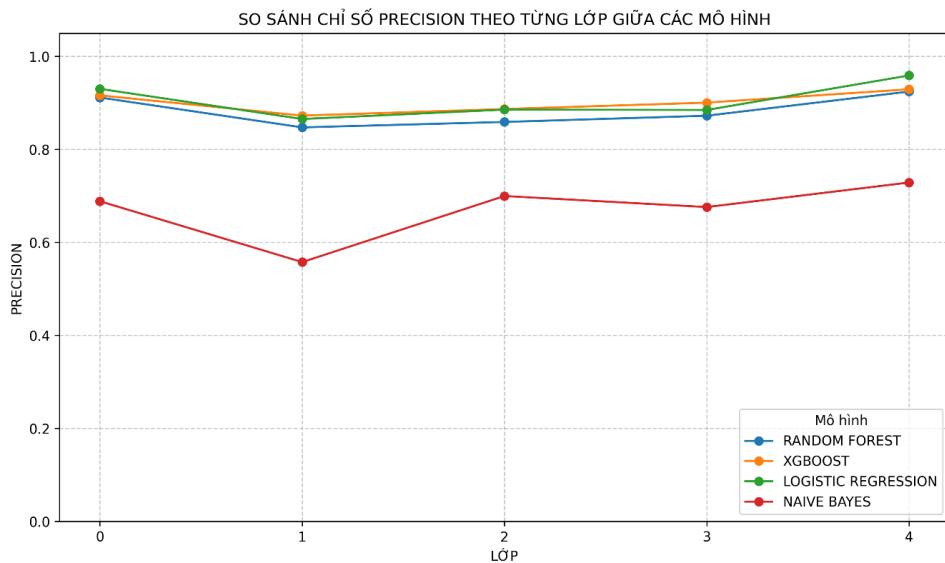
Bên cạnh đó, thuật toán Logistic Regression tuy có chỉ số Precision cao nhất với 90,50%, tuy nhiên chỉ số Recall chỉ đạt 84,36% dẫn đến kết quả F1-Score chỉ đạt 86,95%, phản ánh sự mâu thuẫn giữa việc nhận dạng các trường hợp và độ chính xác. Tương tự, thuật toán Random Forest cũng cho ra kết quả với F1-Score là 86,73%, rất sát với Logistic Regression, cho thấy đây cũng là một mô hình ổn định và hiệu quả.

Đối với mô hình Naive Bayes cho ra kết quả thấp nhất ở cả 3 chỉ số, với Precision là 67,02%, Recall là 64,98% và cuối cùng là F1-Score với 64,88%, điều đó cho thấy Naive Bayes chính là mô hình kém phù hợp nhất trong bài toán phân loại mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế ở Việt Nam.

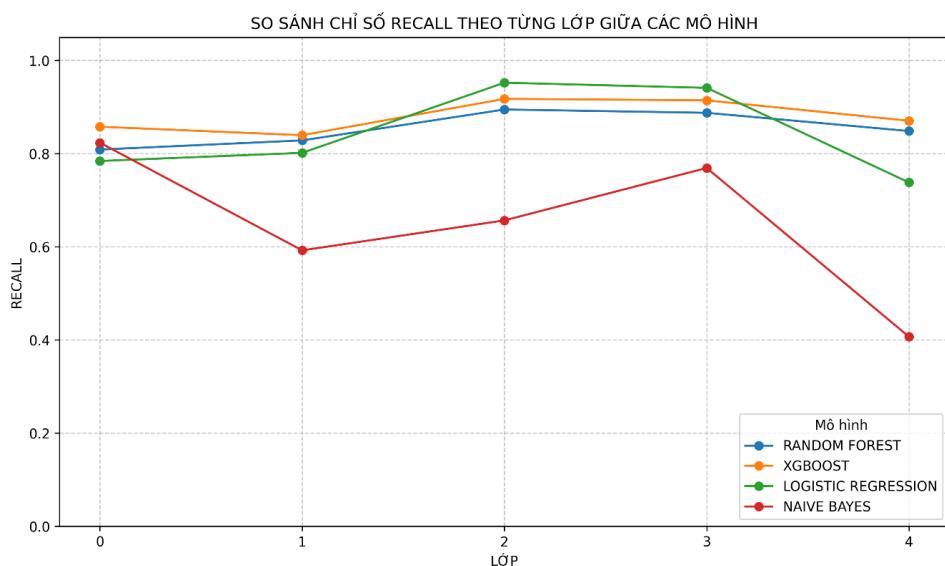
Ngoài ra, bảng trên còn cung cấp thêm thông tin chỉ số Accuracy để đánh giá tổng thể độ chính xác của các mô hình. Theo đó, XGBoost tiếp tục là mô hình hiệu quả nhất với độ chính xác đạt 89,56%, theo sau là Logistic Regression với 89,11% và Random Forest với 87,15%. Trong khi đó, Naive Bayes chỉ đạt độ chính xác 66,87%, thấp hơn đáng kể so với các mô hình còn lại. Sự nhau giữa các chỉ số Precision, Recall, F1-Score cùng với Accuracy đã cho thấy rằng XGBoost là mô hình

phù hợp và hiệu quả nhất trong việc dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế tại Việt Nam dựa trên các chỉ số đánh giá cơ bản.

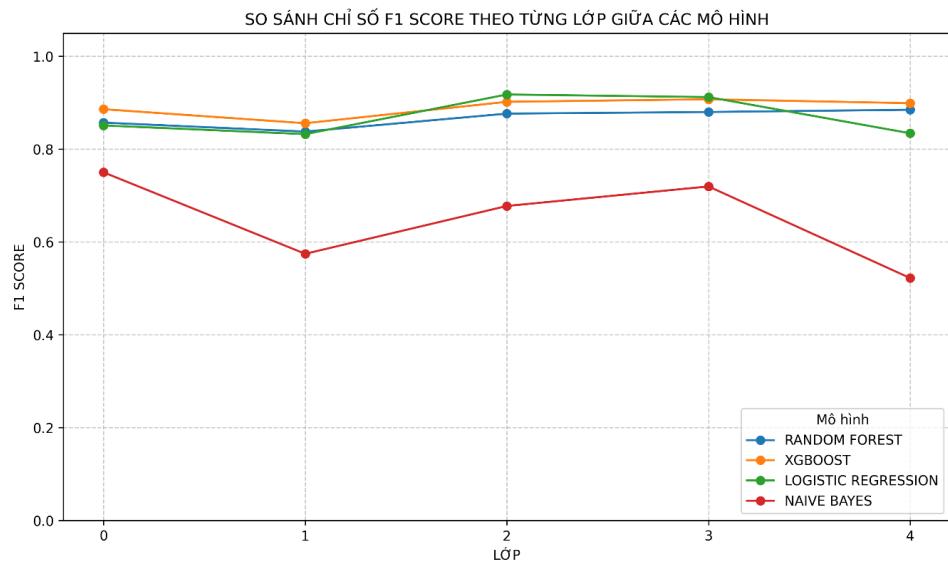
Các biểu đồ sau minh họa sự so sánh các chỉ số đánh giá hiệu suất giữa các mô hình học máy trên từng lớp riêng biệt, giúp làm nổi bật rõ sự khác biệt về các chỉ số đánh giá của các thuật toán:



Hình 3.25: So sánh chỉ số Precision theo từng lớp giữa các mô hình



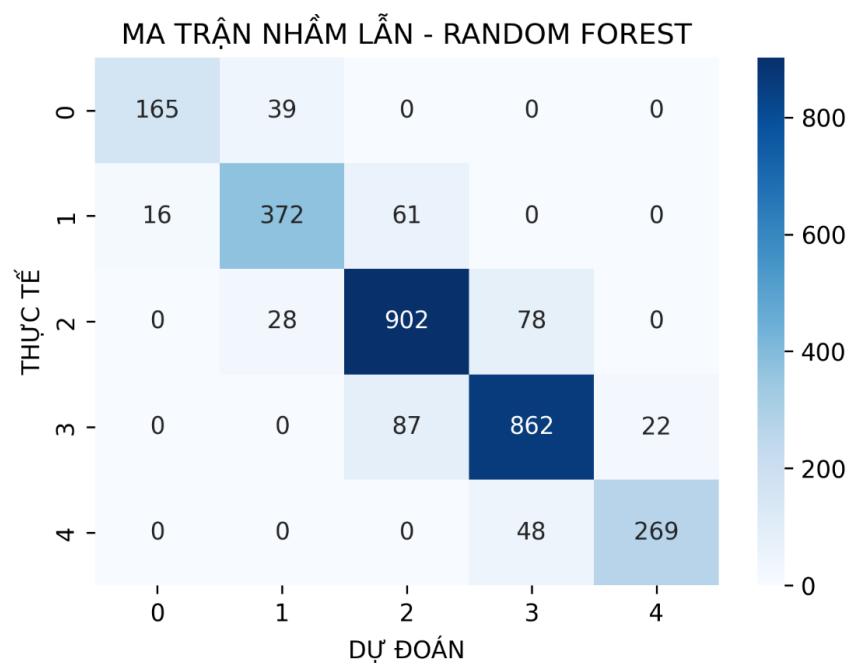
Hình 3.26: So sánh chỉ số Recall theo từng lớp giữa các mô hình



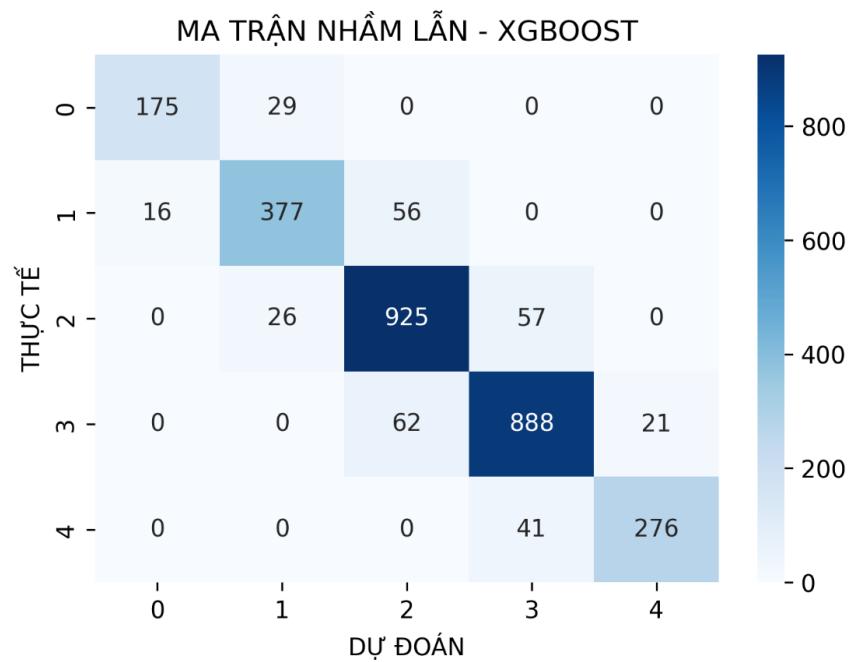
Hình 3.27: So sánh chỉ số F1-Score theo từng lớp giữa các mô hình

3.2.5.2 Đánh giá ma trận nhầm lẫn

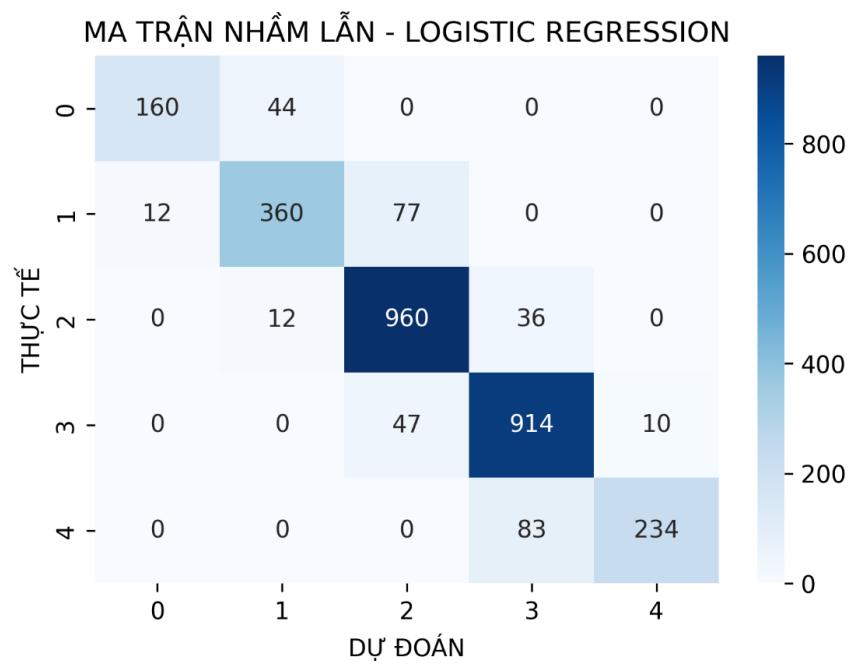
Để đánh giá hiệu suất của mô hình theo từng mức độ ảnh hưởng, các ma trận nhầm lẫn giúp cung cấp cái nhìn trực quan về khả năng dự đoán chính xác cũng như mức độ sai lệch của từng thuật toán được áp dụng trong khóa luận. Mỗi ma trận thể hiện tổng hợp kết quả dự đoán so với nhãn thực tế trên 5 lớp phân loại tương ứng với từng mô hình. Trong đó, các giá trị nằm trên đường chéo chính biểu thị số lượng mẫu mà mô hình dự đoán đúng với nhãn thực tế, trong khi các giá trị ngoài đường chéo phản ánh số lượng dự đoán sai. Việc phân tích các ma trận này giúp xác định mô hình nào hoạt động tốt hơn ở từng mức độ ảnh hưởng cụ thể.



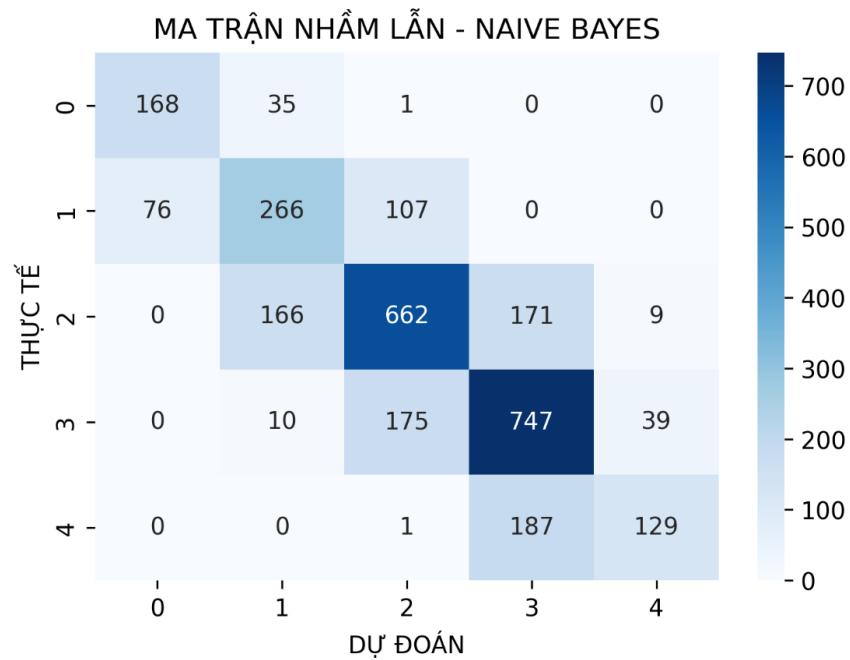
Hình 3.28: Ma trận nhầm lẩn của mô hình Random Forest



Hình 3.29: Ma trận nhầm lẩn của mô hình XGBoost



Hình 3.30: Ma trận nhầm lẫn của mô hình Logistic Regression



Hình 3.31: Ma trận nhầm lẫn của mô hình Naive Bayes

Dựa trên ma trận nhầm lẫn của các mô hình, có thể thấy rằng XGBoost và Random Forest là hai thuật toán cho ra kết quả phân loại hiệu quả nhất. Cụ thể,

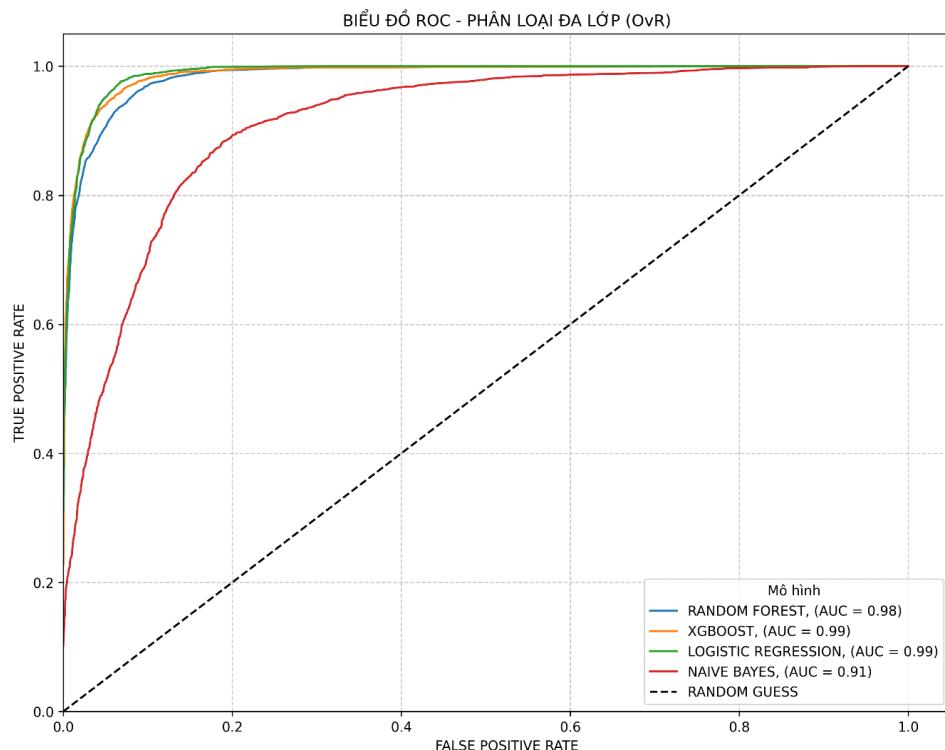
XGBoost đạt số lượng dự đoán đúng cao ở các lớp, đặc biệt là ở lớp 1 với 377 giá trị; lớp 2 với 925 giá trị và lớp 4 với 276 giá trị. Mặc dù vẫn xuất hiện hiện tượng nhầm lẫn giữa các lớp liền kề với nhau nhưng với mức độ thấp hơn so với các mô hình còn lại, chẳng hạn như lớp 0 bị nhầm 29 mẫu sang lớp 1; lớp 1 bị nhầm 16 mẫu sang lớp 0 và 56 mẫu sang lớp 2; lớp 2 nhầm 26 mẫu ở lớp 1 và 57 mẫu ở lớp 3; lớp 3 nhầm lẫn 62 mẫu ở lớp 2 và 21 mẫu ở lớp 4; cuối cùng ở lớp 4 nhầm lẫn 41 mẫu sang lớp 3. Điều này cho thấy XGBoost không chỉ là mô hình phân loại có độ chính xác cao, mà còn có thể kiểm soát tốt sự nhầm lẫn giữa các lớp liền kề nhau.

Bên cạnh đó, mô hình Random Forest cũng cho ra kết quả tương đối, với số lượng mẫu dự đoán đúng nổi bật ở lớp 2 với 902 giá trị và ở lớp 3 với 862 giá trị. Tuy nhiên mô hình này lại đưa ra mức độ nhầm lẫn giữa các lớp liền kề cao hơn so với XGBoost, cụ thể như lớp 0 bị nhầm 39 mẫu sang lớp 1; lớp 1 bị nhầm 16 mẫu sang lớp 0 và 61 mẫu sang lớp 2; lớp 2 nhầm 28 mẫu ở lớp 1 và 78 mẫu ở lớp 3; lớp 3 nhầm lẫn 87 mẫu ở lớp 2 và 22 mẫu ở lớp 4, và cuối cùng lớp 4 nhầm lẫn 48 mẫu sang lớp 3. Kết quả trên cho thấy Random Forest cũng là một mô hình cho ra hiệu quả tốt, tuy nhiên vẫn có xu hướng khó phân biệt giá trị giữa các lớp trung gian.

Logistic Regression tuy là mô hình tuyến tính nhưng lại có kết quả dự đoán đúng cao nhất trong số các mô hình học máy ở lớp 2 với 960 giá trị và lớp 3 với 914 giá trị. Tuy nhiên, mức độ nhầm lẫn lại cao hơn ở các lớp phía ngoài như lớp 0 bị nhầm 44 mẫu sang lớp 1; lớp 1 bị nhầm 12 mẫu sang lớp 0 và 77 mẫu sang lớp 2; lớp 2 nhầm 12 mẫu ở lớp 1 và 36 mẫu ở lớp 3; lớp 3 nhầm lẫn 47 mẫu ở lớp 2 và 10 mẫu ở lớp 4, và cuối cùng lớp 4 nhầm lẫn 83 mẫu sang lớp 3. Những giá trị trên cho thấy Logistic Regression rất mạnh mẽ trong việc phân loại đúng ở các lớp trung tâm nhưng lại yếu hơn khi phân loại các nhãn ở vị trí đầu hoặc cuối. Điều này cho thấy mô hình còn gặp hạn chế trong việc phân biệt các lớp có đặc điểm tương đồng và phân bố gần nhau, đặc biệt khi dữ liệu có tính phi tuyến. Vì là một mô hình tuyến tính, Logistic Regression không thể mô hình hóa được các mối quan hệ phức tạp giữa các đặc trưng, dẫn đến việc phân loại còn gặp nhiều sai sót ở các lớp liền kề.

Và cuối cùng, thuật toán Naive Bayes cho ra kết quả phân loại thấp nhất so với các mô hình được sử dụng trong bài toán. Mặc dù mô hình có đường chéo dự đoán bao gồm 168 giá trị đúng nằm ở lớp 0; 266 giá trị đúng ở lớp 1; 662 giá trị ở lớp 2; 747 ở lớp 3 và 129 giá trị đúng ở lớp 4. Tuy nhiên, mức độ nhầm lẫn giữa các lớp là rất cao. Với lớp 0 bị nhầm 25 mẫu sang lớp 1 và 1 mẫu sang lớp 2; lớp 1 bị nhầm 76 mẫu sang lớp 0 và 107 mẫu sang lớp 2; lớp 2 nhầm 166 mẫu ở lớp 1, 171 mẫu ở lớp 3 và 9 mẫu ở lớp 4; lớp 3 nhầm 10 mẫu ở lớp 1, 175 mẫu ở lớp 2 và 39 mẫu ở lớp 4; cuối cùng lớp 4 nhầm 1 mẫu sang lớp 2 và 187 mẫu sang lớp 3. Việc cho ra những giá trị nhầm lẫn lớn thể hiện rằng các đặc trưng trong tập dữ liệu có mối liên hệ với nhau, trong khi thuật toán Naive Bayes lại giả định chúng là độc lập. Do đó, thuật toán này không phù hợp với dữ liệu thực tế và đưa ra kết quả kém chính xác nhất trong số các mô hình được sử dụng.

3.2.5.3 Đánh giá biểu đồ đường cong ROC theo phân loại đa lớp



Hình 3.32: Biểu đồ đường cong ROC theo phân loại đa lớp (OvR)

Kết quả biểu đồ ROC cho thấy hai mô hình gồm XGBoost, Logistic Regression đều có giá trị AUC là 0,99, mô hình Random Forest cũng có giá trị AUC bằng 0,98, điều này cho thấy các mô hình trên có khả năng phân loại vượt trội. Các đường cong ROC của ba mô hình này đều nằm rất gần trực tung, cho thấy chúng khả năng phát hiện đúng các lớp phân loại với tỷ lệ nhầm lẫn thấp.

Trong đó, thuật toán XGBoost và Random Forest có ưu thế hơn trong việc xử lý tập dữ liệu phức tạp và không tuyến tính, trong khi Logistic Regression lại có ưu điểm hơn nhờ sự đơn giản và khả năng dễ diễn giải mô hình.

Ngược lại, mô hình Naive Bayes chỉ đạt chỉ số AUC là 0.91, thấp hơn đáng kể so với ba mô hình còn lại. Mặc dù vẫn hoạt động tốt hơn mô hình đoán ngẫu nhiên, nhưng Naive Bayes lại thể hiện hiệu quả phân loại kém hơn trong bài toán trên.

3.2.5.4 Lựa chọn mô hình dự đoán

Dựa vào những chỉ số đánh giá hiệu suất mô hình như Accuracy, Precision, Recall, F1-Score, kết hợp với kết quả diễn giải của các ma trận nhầm lẫn và biểu đồ đường cong ROC, có thể thấy rằng các mô hình XGBoost, Random Forest và Logistic Regression đều đạt hiệu suất cao và tương đối đồng đều với nhau. Tuy nhiên, khi xét đến các chỉ số đánh giá chi tiết hơn, mô hình XGBoost đã thể hiện sự vượt trội hơn so với các thuật toán học máy còn lại.

Cụ thể, XGBoost đạt độ chính xác cao nhất với 89,56%, cùng với các chỉ số như Precision là 90,11%, Recall là 88,01% và F1-Score là 89,01%, cho thấy mô hình có thể duy trì hiệu suất ổn định và cân bằng giữa việc dự đoán chính xác và khả năng phát hiện đầy đủ các lớp có trong bài toán. Mô hình Random Forest và Logistic Regression cũng đạt kết quả tốt, tuy nhiên chỉ số Recall và F1-Score lại thấp hơn so với XGBoost. Trong khi đó, mô hình Naive Bayes lại yếu hơn rõ rệt, với các chỉ số Precision, Recall và F1-Score lần lượt là 67,02%, 64,98% và 64,88%, cho thấy mô hình này không phù hợp với bài toán đang thực hiện nghiên cứu.

Từ những chỉ số trên, XGBoost sẽ được lựa chọn là mô hình chính cho bài toán dự đoán trong khóa luận, nhờ khả năng học tốt trên tập dữ liệu, hiệu suất cao và sự ổn định khi phân loại các lớp trong bài toán đa lớp. Việc lựa chọn mô hình phù hợp đóng vai trò quan trọng đến bước tiếp theo của bài nghiên cứu, khi tiến hành dự đoán mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế của Việt Nam. Mô hình XGBoost sẽ hỗ trợ việc dự đoán trở nên chính xác hơn, từ đó khóa luận có thể đề xuất các giải pháp phù hợp với từng mức độ ảnh hưởng biến đổi khí hậu đến từng khu vực tại nước ta.

3.2.6 Dự đoán và phân tích kết quả

Sau khi hoàn tất giai đoạn so sánh hiệu suất và lựa chọn được mô hình hiệu quả nhất sử dụng cho việc dự đoán là XGBoost, khóa luận tiếp tục thực hiện dự đoán mức độ ảnh hưởng của biến đổi khí hậu đối với từng tỉnh và từng vùng kinh tế dựa trên mô hình.

Kết quả dự đoán của bài toán sẽ được tổng hợp vào một DataFrame, bao gồm các thông tin như: Tên tỉnh ('Province'), tháng ('Month'), năm ('Year'), mã vùng kinh tế ('Region Encode'), mức độ ảnh hưởng thực tế ('Actual Impact Level') và mức độ ảnh hưởng dự đoán ('Predicted Impact Level'). Việc tạo mới một DataFrame chứa thông tin 'Actual Impact Level' cùng với 'Predicted Impact Level' mục tiêu giúp việc so sánh giữa giá trị dự đoán và giá trị thực tế của từng mẫu dữ liệu trở nên trực quan và dễ dàng hơn.

	Province	Month	Year	Region Encode	Actual Impact Level	Predicted Impact Level
0	Bến Tre	10	2011	4	3	3
1	Phú Thọ	5	2020	1	3	3
2	Bến Tre	3	2013	4	2	2
3	Ninh Thuận	3	2018	0	1	1
4	Hà Giang	1	2018	1	0	0

Hình 3.33: Minh họa DataFrame compare_xgboost_df

Sau khi đã có kết quả dự đoán mức độ biến đổi khí hậu của từng tỉnh thành, để hướng đến mục tiêu phân tích mức độ ảnh hưởng theo các vùng kinh tế, khóa luận sẽ tiến hành tính toán giá trị trung bình của mức độ ảnh hưởng dự đoán của các tỉnh nằm trong cùng một vùng kinh tế. Quá trình sẽ được thực hiện thông qua phương pháp gom nhóm dữ liệu theo đặc trưng mã vùng kinh tế (Region Encode), nhằm cho ra kết quả tổng quan nhất theo từng vùng. Sau khi thực hiện tính toán, kết quả của các vùng kinh tế sẽ được gán nhãn thành 5 mức độ ảnh hưởng dựa vào chỉ số mức độ dự đoán ảnh hưởng trung bình ('Average Predicted Impact Level'), bao gồm: Ảnh hưởng rất thấp (0) khi kết quả bé hơn 1,2; ảnh hưởng thấp (1) khi kết quả bé hơn 1,7; ảnh hưởng trung bình (2) khi kết quả bé hơn 2,2; ảnh hưởng cao (3) khi kết quả bé hơn 2,7 và ảnh hưởng rất cao (4) khi kết quả lớn hơn hoặc bằng 2,7.

Kết quả phân loại sẽ được tổng hợp tại DataFrame với các đặc trưng như: Mã vùng kinh tế ('Region Encode'), mức độ dự đoán ảnh hưởng trung bình ('Average Predicted Impact Level') và mức độ ảnh hưởng vùng kinh tế ('Region Impact Level'). Các giá trị được thể hiện qua ảnh sau:

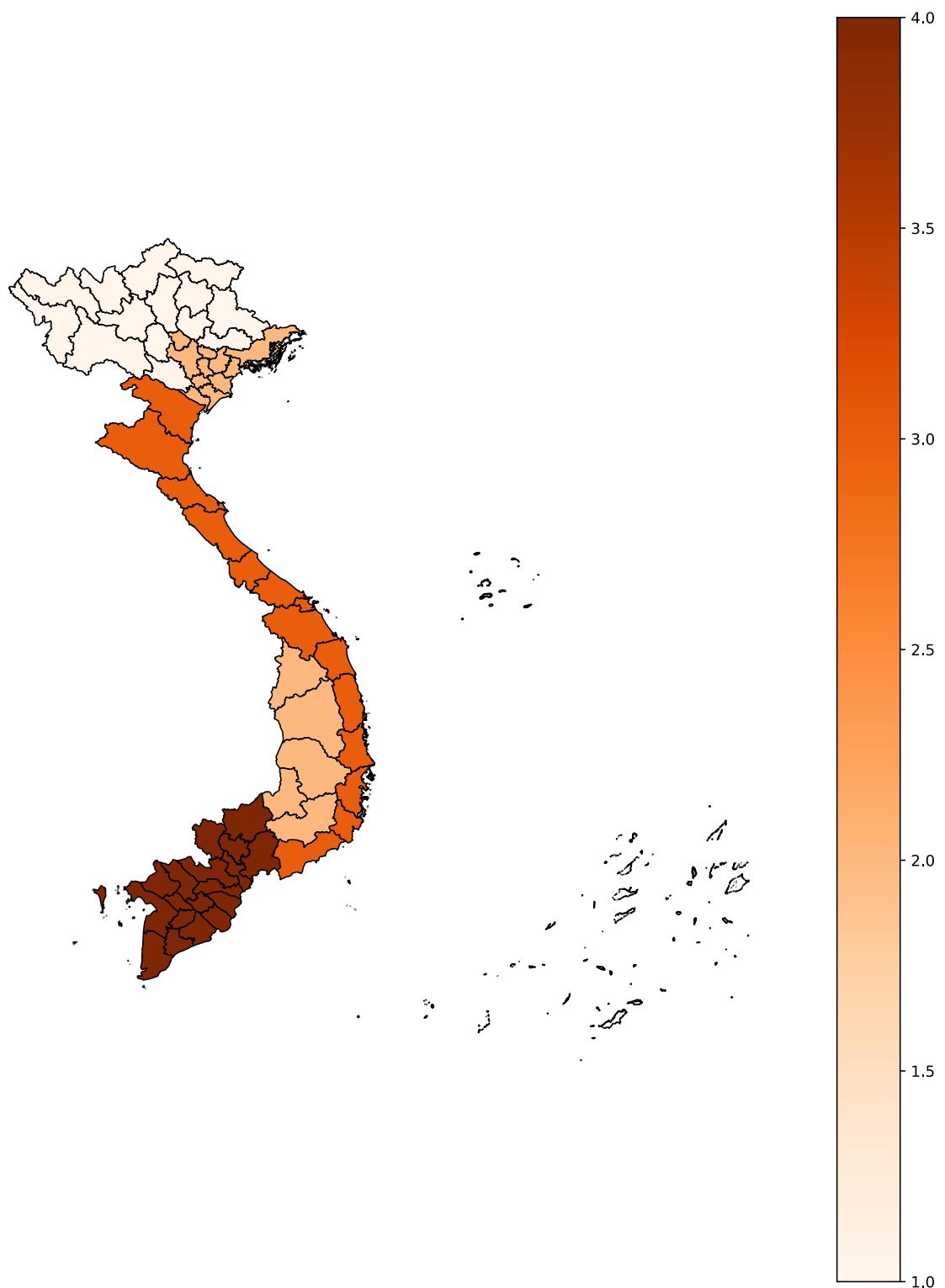
Region Encode	Average Predicted Impact Level	Region Impact Level
0	0	2.576862
1	1	1.465224
2	2	2.021368
3	3	2.832192
4	4	2.758730
5	5	2.048544

Hình 3.34: Mức độ ảnh hưởng của biến đổi khí hậu theo từng vùng kinh tế

Theo kết quả được trình bày tại hình 3.34, có hai khu vực đang chịu tác động rất cao bởi biến đổi khí hậu (ở mức 4), bao gồm vùng Đông Nam Bộ với giá trị trung bình khoảng 2,83 và vùng Đồng bằng sông Cửu Long với giá trị khoảng 2,76. Vùng Bắc Trung Bộ và Duyên hải miền Trung được xếp vào nhóm có mức ảnh hưởng cao (ở mức 3) với giá trị trung bình khoảng 2,58. Bên cạnh đó, hai khu vực có mức độ

ảnh hưởng trung bình (ở mức 2) là Tây Nguyên khoảng 2,02 và Đồng bằng sông Hồng khoảng 2,05. Trong khi đó, Trung du và miền núi phía Bắc là khu vực chịu ảnh hưởng thấp nhất (mức 1) với giá trị trung bình khoảng 1,47. Từ kết quả trên cho thấy, trong bối cảnh biến đổi khí hậu đang diễn ra ngày càng nghiêm trọng tại Việt Nam, không có khu vực kinh tế nào thuộc nhóm chịu ảnh hưởng rất thấp (ở mức 0), từ đó phản ánh tác động lan rộng và có xu hướng gia tăng tại hầu hết các vùng kinh tế trên cả nước.

Nhằm để trực quan hóa mức độ ảnh hưởng của biến đổi khí hậu đến các vùng kinh tế tại Việt Nam, khóa luận tiến hành xây dựng bản đồ thể hiện sự phân bố các mức độ tác động dựa trên kết quả dự đoán từ mô hình học máy XGBoost. Việc biểu diễn bằng bản đồ không chỉ giúp dễ dàng quan sát sự khác biệt giữa các khu vực, mà còn giúp hỗ trợ trong việc nhận diện những vùng cần ưu tiên trong công tác thích ứng và ứng phó với tình trạng biến đổi khí hậu diễn ra càng nghiêm trọng như hiện nay.



Hình 3.35: Bản đồ mức độ ảnh hưởng khí hậu các vùng kinh tế Việt Nam

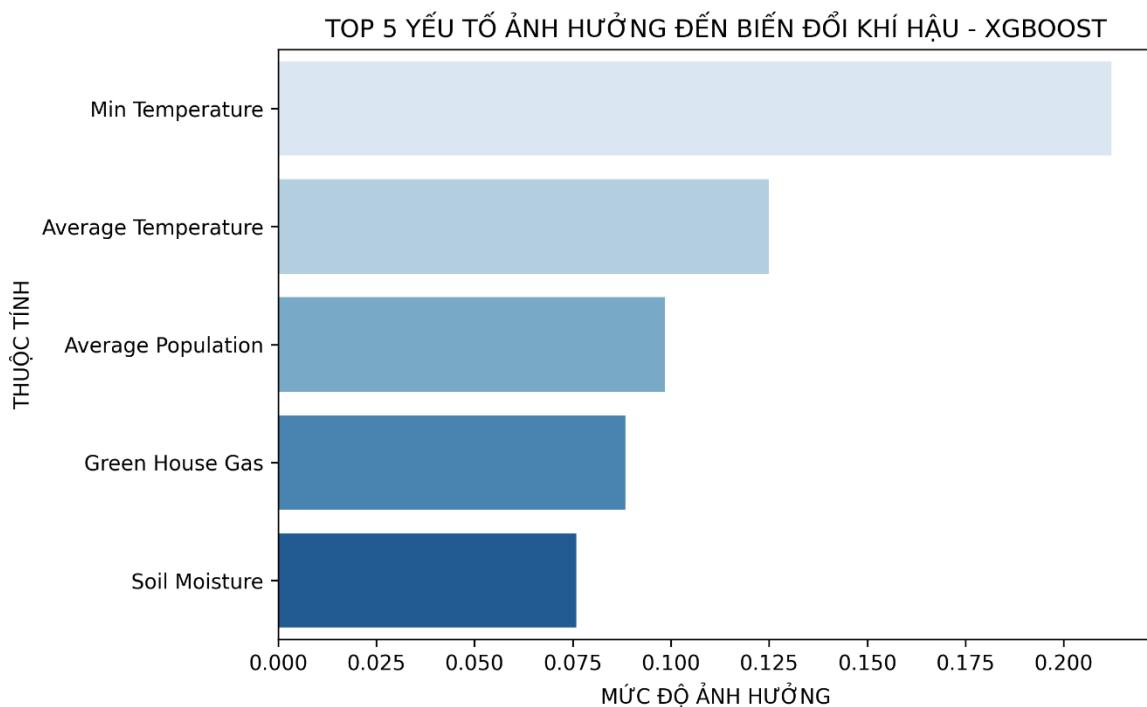
Đối với vùng Đông Nam Bộ và Đồng bằng sông Cửu Long, nơi chịu tác động rất cao của biến đổi khí hậu, các hiện tượng như xâm nhập mặn, hạn hán kéo dài,

triều cường và ngập úng đô thị đang ngày càng phổ biến. Điều này ảnh hưởng trực tiếp đến tình hình sản xuất nông nghiệp, an ninh lương thực, sinh kế của người dân, cũng như sự phát triển bền vững của hệ thống đô thị - công nghiệp. Đặc biệt tại khu vực Đồng bằng sông Cửu Long, nơi được coi là “vựa lúa” của cả nước, đang đối mặt với nguy cơ mất đất sản xuất do mực nước biển ngày càng dâng cao gây ra xói lở bờ sông, bờ biển.

Tại khu vực Bắc Trung Bộ và Duyên hải miền Trung, nằm trong nhóm chịu ảnh hưởng cao, sự tăng về tần suất và cường độ của các hiện tượng thời tiết cực đoan như bão mạnh, lũ quét và sạt lở đất đã gây thiệt hại lớn về cơ sở hạ tầng, đồng thời cả tính mạng của người dân. Với đặc điểm địa hình hẹp ngang, nhiều đồi núi và hệ thống sông ngòi ngắn và dốc khiến khu vực này dễ bị tổn thương trước các tác động của khí hậu cực đoan.

Trong khi đó, vùng Tây Nguyên và Đồng bằng sông Hồng với mức độ ảnh hưởng trung bình cũng đang ghi nhận nhiều dấu hiệu biến đổi như thay đổi chu kỳ mùa vụ, suy giảm nguồn nước và gia tăng dịch bệnh trên cây trồng, vật nuôi. Tây Nguyên còn đối mặt với nguy cơ suy thoái rừng và mất cân bằng sinh thái, trong khi khu vực Đồng bằng sông Hồng phải đối diện với vấn đề ngập úng nội đô và áp lực gia tăng dân số đô thị, ảnh hưởng đến chất lượng cuộc sống và phát triển hạ tầng khu vực.

Đối với Trung du và miền núi phía Bắc, dù được đánh giá là khu vực có mức độ ảnh hưởng thấp nhất, song vẫn phải đối mặt với những rủi ro như sạt lở đất, lũ quét và biến đổi về chế độ mưa, nhiệt độ, đặc biệt tại các khu vực thưa dân cư, người dân khó tiếp cận các nguồn lực hỗ trợ phòng chống thiên tai, bão lũ.



Hình 3.36: Yếu tố ảnh hưởng mạnh đến biến đổi khí hậu theo XGBoost

Theo mô hình XGBoost, các yếu tố có ảnh hưởng lớn nhất đến kết quả phân loại mức độ biến đổi khí hậu bao gồm: nhiệt độ tối thiểu ('Min Temperature'), nhiệt độ trung bình ('Average Temperature'), dân số trung bình ('Average Population'), lượng khí nhà kính ('Green House Gas'), và độ ẩm đất ('Soil Moisture'). Trong đó, nhiệt độ tối thiểu là yếu tố có mức độ ảnh hưởng cao nhất, cho thấy sự biến động về nhiệt độ, đặc biệt là xu hướng tăng dần vào ban đêm đang tác động rõ rệt đến môi trường và hệ sinh thái. Điều này phản ánh mức độ nóng lên toàn cầu không chỉ giới hạn trong giờ cao điểm trong ngày, gây ảnh hưởng tiêu cực đến khả năng điều hòa sinh học tự nhiên.

Yếu tố nhiệt độ trung bình cũng góp phần đáng kể vào biến đổi khí hậu. Khi nhiệt độ tăng vượt mức ngưỡng thích nghi của nhiều loài sinh vật và hệ sinh thái, các hiện tượng cực đoan như hạn hán, nắng nóng kéo dài và cháy rừng sẽ trở nên thường xuyên hơn, đe dọa đến an ninh sinh thái và phát triển bền vững.

Bên cạnh đó, dân số trung bình là một yếu tố thể hiện rõ nét tác động từ con người đến môi trường. Sự gia tăng dân số kéo theo nhu cầu tiêu dùng năng lượng, mở rộng hạ tầng và phát triển công nghiệp, từ đó làm tăng lượng khí thải và sức ép lên tài nguyên thiên nhiên. Mọi quan hệ giữa tăng trưởng dân số và biến đổi khí hậu nhấn mạnh sự cần thiết của việc quy hoạch đô thị và phát triển dân cư theo hướng bền vững.

Khí nhà kính vốn được xem là nguyên nhân trực tiếp gây hiệu ứng nhà kính, tiếp tục đóng vai trò quan trọng. Phát thải từ các ngành năng lượng, giao thông, công nghiệp và nông nghiệp cần được kiểm soát thông qua các chính sách cụ thể nhằm giảm thiểu lượng CO₂ và CH₄ trong khí quyển.

Cuối cùng, độ ẩm đất là một chỉ báo cho thấy sự biến động về điều kiện thủy văn, phản ánh khả năng giữ nước và duy trì độ phì nhiêu của đất, yếu tố then chốt trong sản xuất nông nghiệp và cân bằng sinh thái.

CHƯƠNG 4: KẾT LUẬN VÀ KIẾN NGHỊ

4.1 Kết quả đạt được

Thông qua quá trình thực hiện nghiên cứu, khóa luận đã đạt được mục tiêu đã đề ra ban đầu là dự đoán được mức độ ảnh hưởng của biến đổi khí hậu đối với các khu vực kinh tế trọng điểm tại Việt Nam.

Bắt đầu bằng việc thu thập và tiền xử lý hai bộ dữ liệu ban đầu, kết hợp với thao tác trực quan hóa các thuộc tính có liên quan bằng các biểu đồ, hình ảnh. Khóa luận đã trực quan được mức độ phân bố của từng đặc trưng có liên quan theo từng tỉnh thành dựa vào mốc thời gian cụ thể.

Ngoài ra, việc sử dụng đa dạng các mô hình học máy như Random Forest, XGBoost, Logistic Regression và Naive Bayes vào bài toán kết hợp với việc tính toán và so sánh các chỉ số như Accuracy, Precision, Recall, F1-Score; so sánh ma trận nhầm lẫn hay biểu đồ đường cong ROC. Từ đó đã lựa chọn được mô hình phù hợp nhất là XGBoost để thực hiện bước dự đoán của đề tài.

Dựa vào kết quả dự đoán theo từng tỉnh thành, khóa luận có thể tiến hành tính chỉ số trung bình theo từng khu vực kinh tế, sau đó có thể xác định được mức độ ảnh hưởng đối với từng vùng. Đặc biệt, khóa luận đã có thể kết hợp file JSON và thư viện GeoPandas để có thể trực quan hóa thông tin bằng hình ảnh dạng bản đồ Việt Nam.

4.2 Mật hạn chế

Bên cạnh những kết quả đạt được, khóa luận vẫn còn tồn tại nhiều mặt hạn chế như:

Mặc dù hai bộ dữ liệu ban đầu được thu thập từ những nguồn có độ uy tín cao, tuy nhiên số lượng mẫu của các tập dữ liệu còn nhiều hạn chế do thông tin đã thu thập chỉ giới hạn trong 63 tỉnh thành ở Việt Nam và nằm trong khoảng từ năm 2011-2023. Đặc biệt do số lượng mẫu nằm trong 2 tập dữ liệu không bằng nhau, nên phải thực

hiện thêm thao tác tiền xử lý để gộp tập dữ liệu. Do số lượng mẫu còn gấp nhiều giới hạn, nên có thể dẫn đến việc các mô hình học máy chưa thể học đủ thông tin, dẫn đến tình trạng các giá trị được dự đoán có thể còn gấp nhiều sai lầm.

Từ hạn chế trên cũng đã cho biết được việc tính toán các chỉ số của các mô hình dự đoán như Random Forest, XGBoost, Logistic Regression và Naive Bayes đưa ra kết quả chưa cao. Chẳng hạn độ chính xác tổng thể của các mô hình chỉ nằm trong khoảng từ 60% đến dưới 90%. Do đó kết quả phân loại còn gấp nhiều sai sót, đặc biệt là với các lớp liền kề với nhau.

4.3 Hướng phát triển cho tương lai

Mặc dù đã đạt được mục tiêu nghiên cứu của bài toán, những khóa luận vẫn còn nhiều mặt hạn chế đã được nêu ở phía trên. Một số hướng phát triển được đề xuất giúp bài toán ngày càng trở nên hoàn thiện hơn như:

Có thể mở rộng phạm vi của tập dữ liệu kể cả về mặt không gian, thời gian hay các đặc trưng có liên quan. Việc mở rộng phạm vi giúp các mô hình có thêm nhiều dữ liệu để huấn luyện hơn, từ đó có thể giúp nâng cao độ chính xác trong quá trình phân loại mức độ ảnh hưởng.

Khóa luận có thể phát triển theo hướng xác định mục tiêu dự đoán một cách cụ thể hơn thay vì xác định mức độ dự đoán biến đổi khí hậu cho từng khu vực một cách tổng thể. Một số chỉ số đặc trưng có thể sử dụng để tiến hành dự đoán cho tương lai như: Chỉ số phát triển sản phẩm công nghiệp, Diện tích đất rừng, nhiệt độ trung bình,... Việc dự đoán sẽ được thực hiện theo phương pháp dự đoán mô hình theo chuỗi thời gian (Times Series Forecasting).

Một hướng phát triển khác là sẽ kết hợp các đặc trưng trong tập dữ liệu với thông tin khảo sát thực tế từ các chuyên gia khí hậu, người dân hay các doanh nghiệp nhằm bổ sung mức độ thực tiễn của mô hình dựa trên bộ dữ liệu định lượng. Ngoài ra, việc mở rộng bài toán thành hệ thống hỗ trợ ra quyết định giúp người dùng có thể

nhập vào các giá trị đầu vào, từ đó hệ thống có thể cho ra kết quả dự đoán trực quan dựa trên những giá trị đã nhập.

Cùng với việc nâng cao sự kết hợp với hệ thống thông tin địa lý để trực quan các kết quả báo cáo trên bản đồ nhằm giúp các nhà quản lý, các chuyên gia về môi trường có thể nhận định nhanh được các khu vực, các tỉnh thành đang chịu ảnh hưởng mạnh, từ đó có thể đưa ra các đề xuất, giải pháp kịp thời để có thể xây dựng các chính sách phát triển bền vững hơn trong bối cảnh tình trạng biến đổi khí hậu đang biến động mạnh như hiện nay.

TÀI LIỆU THAM KHẢO

(Theo chuẩn IEEE)

- [1] “Ảnh hưởng của biến đổi khí hậu ở Việt Nam”. Truy cập: 23 Tháng Ba 2025. [Online]. Available at: <https://pilot.dcc.gov.vn/Ảnh-hưởng-của-biến-đổi-khí-hậu-ở-Việt-Nam>
- [2] “Học máy là gì?”, Intel. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://www.intel.com/content/www/vn/vi/learn/what-is-machine-learning.html>
- [3] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, Francisco Herrera, “Big data preprocessing: methods and prospects”. 2016. [Online]. Available at: <https://link.springer.com/content/pdf/10.1186/s41044-016-0014-0.pdf>
- [4] Z. Bobbitt, “Label Encoding vs. One Hot Encoding: What’s the Difference?”, Statology. Truy cập: 4 Tháng Tư 2025. [Online]. Available at: <https://www.statology.org/label-encoding-vs-one-hot-encoding/>
- [5] R. Vashisht, “Machine Learning: When to perform a Feature Scaling?”, Atoti Community. Truy cập: 4 Tháng Tư 2025. [Online]. Available at: <https://www.atoti.io/articles/when-to-perform-a-feature-scaling/>
- [6] D. Butvinik, “Feature Selection — Extended Overview”, Analytics Vidhya. Truy cập: 6 Tháng Tư 2025. [Online]. Available at: <https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c>
- [7] “Decision Tree”, GeeksforGeeks. Truy cập: 6 Tháng Tư 2025. [Online]. Available at: <https://www.geeksforgeeks.org/decision-tree/>
- [8] Brownlee J., “A Gentle Introduction to Ensemble Learning Algorithms”, MachineLearningMastery.com. Truy cập: 7 Tháng Tư 2025. [Online]. Available at: <https://www.machinelearningmastery.com/tour-of-ensemble-learning-algorithms/>

- [9] Sruthi, “Understanding Random Forest Algorithm With Examples”, Analytics Vidhya. Truy cập: 9 Tháng Tư 2025. [Online]. Available at: <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
- [10] “Random Forest: A Complete Guide for Machine Learning”, Built In. Truy cập: 9 Tháng Tư 2025. [Online]. Available at: <https://builtin.com/data-science/random-forest-algorithm>
- [11] A. Tyagi, “What is XGBoost Algorithm?”, Analytics Vidhya. Truy cập: 10 Tháng Tư 2025. [Online]. Available at: <https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/>
- [12] “XGBoost”, GeeksforGeeks. Truy cập: 10 Tháng Tư 2025. [Online]. Available at: <https://www.geeksforgeeks.org/xgboost/>
- [13] “Logistic Regression in Machine Learning”, GeeksforGeeks. Truy cập: 9 Tháng Năm 2025. [Online]. Available at: <https://www.geeksforgeeks.org/understanding-logistic-regression/>
- [14] “Hồi quy logistic là gì? - Giải thích về mô hình hồi quy logistic - AWS”, Amazon Web Services, Inc. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://aws.amazon.com/vi/what-is/logistic-regression/>
- [15] “Naive Bayes Classifiers”, GeeksforGeeks. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://www.geeksforgeeks.org/naive-bayes-classifiers/>
- [16] “OpenStreetMap”, OpenStreetMap. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://www.openstreetmap.org/>
- [17] “NASA POWER | Docs | Tutorials - NASA POWER | Docs”. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://power.larc.nasa.gov/docs/tutorials/>

- [18] “Forest Monitoring, Land Use & Deforestation Trends | Global Forest Watch”. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://www.globalforestwatch.org/>
- [19] “GSO”, General Statistics Office of Vietnam. Truy cập: 10 Tháng Năm 2025. [Online]. Available at: <https://www.nso.gov.vn/>

PHỤ LỤC