

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT
TRƯỜNG ĐẠI HỌC THỦY LỢI

----- ☆ ☞ ▣ -----



BÙI THỊ THU HƯƠNG

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY CHO
BÀI TOÁN DỰ BÁO LŨ LỤT

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO

BỘ NÔNG NGHIỆP VÀ PTNT

TRƯỜNG ĐẠI HỌC THỦY LỢI

----- ☆ ☞ ☒ -----

BÙI THỊ THU HƯƠNG

**ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY CHO
BÀI TOÁN DỰ BÁO LŨ LỤT**

Ngành: Công Nghệ Thông Tin

Mã số: 7480201

NGƯỜI HƯỚNG DẪN: TS. LÊ NGUYỄN TUẤN THÀNH

HÀ NỘI, NĂM 2025

GIẤY BÌA ĐỒ ÁN TỐT NGHIỆP, KHÓA LUẬN TỐT NGHIỆP

BÙI THỊ THU HƯƠNG

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2025



CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập - Tự do - Hạnh phúc



NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: Bùi Thị Thu Hương

Hệ đào tạo: Đại học chính quy

Lớp: 61TH4

Ngành: Công nghệ thông tin

Khoa: Công nghệ thông tin

1_TÊN ĐỀ TÀI: ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY CHO BÀI TOÁN DỰ BÁO LŨ LỤT

2_ CÁC TÀI LIỆU THAM KHẢO:

1. [1] "Mô hình mạng thần kinh nhân tạo và các bài toán ngập lụt: Nâng cao hiệu quả dự báo và nghiên cứu ảnh hưởng của công trình đô thị lên sự lan truyền." Available: <https://hoinghi45nam.imech.ac.vn>.
2. [2] "Nghiên cứu ứng dụng trí tuệ nhân tạo trong dự báo lũ lụt," Viện Quy hoạch Thủy lợi. Available: <http://wri.vn>.
3. [3] Đinh Nhật Quang, "Tổng quan ứng dụng phương pháp học máy trong dự báo lũ lụt," 2023. Available: <https://vawr.org.vn>.
4. [4] "Flood prediction and mitigation strategies in Vietnam," Wiley Online Library. Available: <https://onlinelibrary.wiley.com>.
5. [5] "FloodAdapt: Tools for adaptive flood risk management," DLR. Available: <https://floodadapt.eoc.dlr.de>.
6. [6] Monica và Mandolaro, "Flood Data for Prediction and Analysis," MDPI. Available: <https://github.com/Mandolaro/flood-data>.
7. [7] "Natural disaster data in Vietnam," Open Development Mekong. Available: <https://data.vietnam.opendevlopmekong.net>.

8. [8] **Sneha Choudhary**, "Flood Prediction AI Project for Kerala State," **GitHub**, 2023. Available: <https://github.com/choudharysneha1708/FloodAI>.
9. [9] **S. Sinha**, "Vietnam Flood Prediction Using Machine Learning," **GitHub**. Available: <https://github.com/ssinha22/Vietnam-Flood-Prediction>.
10. [10] **Nikhil Desai**, "Flood Risk Mapping in Danang City," **GitHub**. Available: https://github.com/NikhilSDesai/Flood_Risk_Mapping.
11. [11] **"V-FloodNet: Video Segmentation System for Urban Flood Detection,"** **GitHub**. Available: <https://github.com/xmlyqing00/V-FloodNet>.
12. [12] **K. Leok**, "FloodPy: Python Toolbox for Flood Mapping," **GitHub**. Available: <https://github.com/kleok/FLOODPY>.
13. [13] **Vũ Đức Mạnh**, "Weather Forecast Using LSTM-BiLSTM Models," **GitHub**. Available: <https://github.com/vuducmanh2008/Weather-Forecast-Using-LSTM-BiLSTM-Model>.
14. [14] **VennDev**, "AI Weather Forecasting," **GitHub**. Available: <https://github.com/VennDev/AI-weather-forecasting>.
15. [15] **Nguyễn Văn Tài**, "Datascience Applications for Flood Analysis," **GitHub**. Available: https://github.com/NguyenVTai/Datascience_2016-2.
16. [16] **"Phân tích dữ liệu thời tiết từ A-Z,"** **Tạp chí Khoa học Môi trường**, 2024. Available: <https://weatheranalysis.com/flood-data-analysis-guide>.
17. [17] **Nguyễn Hữu Trí**, "Hybrid Models for Flood Risk Management," **Journal of Hydrological Studies**, vol. 12, no. 3, 2023. Available: <https://hydrologicalstudies.org>.
18. [18] **FloodList**, "Latest updates on global flood events," 2023. Available: <https://floodlist.com>.
19. [19] **Vietnam Meteorological and Hydrological Administration**, "Real-time flood monitoring data," 2024. Available: <https://nchmf.gov.vn>.
20. [20] **European Space Agency**, "Satellite Observations for Flood Mapping," 2024. Available: <https://esa.int>.

3_ NỘI DUNG CÁC PHẦN THUYẾT TRÌNH VÀ TÍNH TOÁN:

Nội dung cần thuyết trình	Tỷ lệ %
Chương 1: Tổng quan nghiên cứu	10
Chương 2: Mô tả bài toán và xây dựng khung nghiên cứu	10
Chương 3: Thu thập và tiền xử lý dữ liệu	15
Chương 4: Trích xuất đặc trưng và phân tích đặc trưng	15
Chương 5: Triển khai mô hình học máy để dự đoán	35
Chương 6: Đánh giá kết luận và kết quả	15

4_ GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN

Giáo viên hướng dẫn toàn bộ quá trình thực hiện đồ án: **TS. Lê Nguyễn Tuấn Thành**

5_ NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Ngày ... Tháng ... Năm 202

Trưởng Bộ môn <i>(Ký và ghi rõ Họ tên)</i>	Giáo viên hướng dẫn chính <i>(Ký và ghi rõ Họ tên)</i>

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua

Ngày ... Tháng ... Năm
Chủ tịch Hội đồng
(Ký và ghi rõ Họ tên)

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi ngày ... tháng ... năm 2024

Sinh viên làm Đồ án tốt nghiệp
(Ký và ghi rõ Họ tên)



TRƯỜNG ĐẠI HỌC THUYẾT LỢI
KHOA CÔNG NGHỆ THÔNG TIN

BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP

Tên đề tài: Ứng dụng các mô hình học máy cho bài toán dự báo lũ lụt

Sinh viên thực hiện: Bùi Thị Thu Hương

Lớp: 61TH4

Mã sinh viên: 1951060744

Số điện thoại: 0965548810

Email: 1951060744@e.tlu.edu.vn

Giáo viên hướng dẫn: TS. Lê Nguyễn Tuấn Thành

TÓM TẮT ĐỀ TÀI

Đề tài "Ứng dụng các mô hình học máy cho bài toán dự báo lũ lụt" nhằm nghiên cứu và phát triển các phương pháp dự báo lũ lụt tại Việt Nam, đặc biệt ở miền Trung, nơi thường xuyên bị ảnh hưởng bởi mưa bão và lũ lụt nghiêm trọng. Bằng cách sử dụng các mô hình học máy cơ bản và hiệu quả như **Linear Regression** và **KNN**. Đề tài hướng tới việc cải thiện độ chính xác của dự báo lũ, từ đó giúp cơ quan chức năng có thể cảnh báo và phản ứng kịp thời trước các tình huống lũ lụt.

Nghiên cứu sẽ tập trung vào việc thu thập, xử lý dữ liệu thủy văn như mực nước, lượng mưa và lưu lượng dòng chảy, kết hợp với các thuật toán học máy để dự đoán mức độ và nguy cơ ngập lụt. Kết quả của đề tài được kỳ vọng sẽ nâng cao khả năng dự báo lũ ngắn hạn và dài hạn, từ đó góp phần vào việc giảm thiểu thiệt hại về người và tài sản.

Các kết quả từ đề tài có thể hỗ trợ cơ quan chức năng ra quyết định về xây dựng cơ sở hạ tầng và chính sách phòng chống lũ lụt.

CÁC MỤC TIÊU CHÍNH

- Nghiên cứu và áp dụng các mô hình học máy trong dự báo lũ lụt:** Mục tiêu chính của đề tài là tìm hiểu và áp dụng các mô hình học máy **Linear Regression** và **KNN**. Các mô hình này sẽ được sử dụng để dự báo các yếu tố quan trọng của lũ lụt như mực nước, lưu lượng dòng chảy, và độ sâu ngập lụt. Thông qua đó, đề tài tìm cách cải thiện độ chính xác của các mô hình dự báo lũ so với các phương pháp truyền thống hiện tại.
- So sánh và đánh giá hiệu quả các mô hình học máy:** Một mục tiêu quan trọng khác là đánh giá và so sánh hiệu quả của các mô hình học máy khác nhau trong việc dự báo lũ lụt. Các mô hình sẽ được đánh giá dựa trên các chỉ số như độ chính xác, sai số trung bình bình phương (RMSE), và hệ số Nash-Sutcliffe (NSE). Việc so sánh này giúp tìm ra mô hình tối ưu nhất cho việc dự báo lũ lụt trong điều kiện thực tế tại Việt Nam.

KẾT QUẢ DỰ KIẾN

Đề tài dự kiến nâng cao độ chính xác trong dự báo lũ lụt bằng cách áp dụng các mô hình học máy **Linear Regression** và **KNN**, đồng thời xác định mô hình tối ưu dựa trên các chỉ số như độ chính xác, RMSE và NSE.

Kết quả nghiên cứu sẽ đóng góp phần nhỏ vào việc xây dựng các công cụ hỗ trợ dự báo lũ hiệu quả, tăng cường khả năng ứng phó với thiên tai, và mở rộng ứng dụng học máy trong quản lý thiên tai tại Việt Nam.

Ngoài ra, nghiên cứu còn nâng cao nhận thức và đào tạo nguồn nhân lực chất lượng cao, góp phần vào việc phát triển các giải pháp quản lý thiên tai bền vững và giảm thiểu tác động tiêu cực của lũ lụt.

TIẾN ĐỘ THỰC HIỆN

STT	Thời gian	Nội dung công việc	Kết quả dự kiến đạt được
1	30/09/2024 đến 09/10/2024	Viết đề cương đề tài	
1	9/10/2024 đến 20/10/2024	Tìm hiểu về học máy và các thuật toán (Linear Regression, KNN).	Báo cáo tổng quan về học máy và các thuật toán liên quan.
2	21/10/2024 đến 27/10/2024	Phân tích cách các thuật toán học máy được ứng dụng trong dự báo lũ lụt.	Báo cáo chi tiết về ứng dụng của các mô hình học máy trong dự báo lũ.
3	28/10/2024 đến 3/11/2024	Tìm hiểu và thực hành với ngôn ngữ Python và các thư viện học máy như Scikit-learn.	Báo cáo về Python và thư viện học máy, cài đặt môi trường lập trình.
4	04/11/2024 đến 10/11/2024	Xác định và phân tích dữ liệu thủy văn thực nghiệm, bao gồm mực nước, lưu lượng, và lượng mưa.	Tài liệu đặc tả về dữ liệu thủy văn được sử dụng.
5	11/11/2024 đến 17/11/2024	Thiết kế và xây dựng các mô hình học máy, bao gồm Linear Regression, KNN.	Các mô hình học máy được thiết kế và sẵn sàng sử dụng.
6	18/11/2024 đến 24/11/2024	Huấn luyện, tinh chỉnh mô hình, và đánh giá hiệu suất dựa trên tập dữ liệu thủy văn.	Báo cáo chi tiết về hiệu suất của các mô hình đã huấn luyện.
7	25/11/2024 đến 1/12/2024	Kiểm thử mô hình trên tập dữ liệu mới chưa sử dụng trong huấn luyện.	Báo cáo hiệu suất mô hình trên dữ liệu kiểm thử mới.

8	1/12/2024 đến 8/12/2024	Viết báo cáo và trình bày kết quả	Chỉnh sửa hoàn thiện mọi thứ, gửi thầy để duyet
9	8/12/2024 đến 14/12/2024	Dự trù và hoàn thiện báo cáo	Báo cáo hoàn thiện được thầy duyệt
10	09/12/2024	Hoàn thiện báo cáo và trình bày kết quả cuối cùng.	Tài liệu báo cáo chi tiết về toàn bộ quá trình và kết quả đạt được.

TÀI LIỆU THAM KHẢO

[1]

<http://wri.vn/Pages/nguyen-cuu-ung-dung-tri-tue-nhan-tao-trong-du-bao-lu-lut.aspx>

[2]

<https://hoinghi45nam.imech.ac.vn/bao-cao/i-41/mo-hinh-mang-than-kinh-nhan-tao-va-cac-bai-toan-ngap-lut-nang-cao-hieu-qua-du-bao-va-nguyen-cuu-a-nh-huong-cua-cong-trinh-do-thi-len-su-lan-truyen.html>

[3]

<https://vawr.org.vn/tong-quan-ung-dung-phuong-phap-hoc-may-trong-du-bao-lu>

[4] Phạm Quang Minh, Tallam K., "Predicting Flood Hazards in the Vietnam Central Region: An Artificial Neural Network Approach", *Sustainability*, 2022 [Predicting Flood Hazards in the Vietnam Central Region: An Artificial Neural Network Approach \(mdpi.com\)](#)

[5] Dự án FloodAdaptVN, "Integrating Ecosystem-based Approaches into Flood Risk Management for Adaptive and Sustainable Urban Development in Central Viet Nam" [FloodAdaptVN | FloodAdaptVN \(dlr.d](#)

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Lê Nguyễn Tuấn Thành. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong khóa luận còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào, tôi xin hoàn toàn chịu trách nhiệm về nội dung khóa luận của mình. Khoa Công nghệ Thông tin - Đại học Thủy Lợi không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

Hà Nội, ngày 04 tháng 01 năm 2024

NGƯỜI HƯỚNG DẪN

TÁC GIẢ

TS. LÊ NGUYỄN TUẤN THÀNH

BÙI THỊ THU HƯƠNG

LỜI CẢM ƠN

Em xin chân thành cảm ơn thầy TS. Lê Nguyễn Tuấn Thành đã tận tình giúp đỡ em hoàn thành khóa luận tốt nghiệp này. Để hoàn thành khóa luận này, em đã nỗ lực thực hiện và đồng thời cũng nhận được nhiều sự giúp đỡ từ thầy và Khoa Công nghệ Thông tin. Em cũng xin gửi lời cảm ơn đến Khoa Công nghệ Thông tin - Đại học Thủy Lợi đã tạo điều kiện tốt nhất để em có thể học tập, trao đổi và nâng cao kiến thức của mình. Mặc dù em đã rất cố gắng hoàn thành trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự cảm thông và tận tình chỉ bảo của thầy.

Em xin chân thành cảm ơn sự hướng dẫn và chỉ bảo tận tình của thầy!

Hà Nội, ngày 04 tháng 01 năm 2025

Tác giả

BÙI THỊ THU HƯƠNG

ỨNG DỤNG CÁC MÔ HÌNH HỌC MÁY CHO BÀI TOÁN DỰ BÁO LŨ LỤT

TÓM TẮT

Nghiên cứu này tập trung vào dự đoán xác suất lũ lụt dựa trên các yếu tố môi trường, hạ tầng, và xã hội, một vấn đề cấp thiết trong bối cảnh biến đổi khí hậu và gia tăng đô thị hóa. Bài nghiên cứu sẽ khám phá các yếu tố ảnh hưởng đến nguy cơ lũ lụt như cường độ mưa, chất lượng đập, hệ thống thoát nước, và quy hoạch đô thị. Đồng thời, nghiên cứu áp dụng các mô hình học máy và thống kê như Linear Regression, Ridge Regression, Lasso Regression và KNN để dự đoán xác suất lũ lụt. Các mô hình này được triển khai trên bộ dữ liệu thực tế, tiến hành đánh giá hiệu suất và so sánh các mô hình dựa trên các chỉ số như R^2 , RMSE, và MAPE. Kết quả nghiên cứu không chỉ cung cấp cái nhìn sâu sắc về khả năng dự báo lũ lụt mà còn đưa ra các khuyến nghị hữu ích cho việc quản lý và giảm thiểu rủi ro lũ lụt trong thực tiễn.

MỤC LỤC

DANH MỤC HÌNH VẼ.....	6
DANH MỤC CÁC CHỮ VIẾT TẮT.....	6
MỞ ĐẦU.....	7
1. Tính cấp thiết của đề.....	7
2. Mục tiêu khóa luận.....	8
3. Đối tượng và phạm vi nghiên cứu.....	8
4. Cấu trúc của khóa luận.....	9
CHƯƠNG 1: TỔNG QUAN VỀ NGHIÊN CỨU.....	11
1.1. Tổng quan tình hình nghiên cứu trong và ngoài nước.....	11
1.2. Cơ sở lý thuyết.....	12
1.2.1. Các yếu tố ảnh hưởng đến xác suất lũ lụt.....	12
1.2.2. Tổng quan về các mô hình học máy để dự đoán lũ lụt.....	13
1.2.3. Các thách thức trong dự đoán lũ lụt.....	14
1.3. Tổng kết phần tổng quan.....	15
CHƯƠNG 2: MÔ TẢ BÀI TOÁN VÀ XÂY DỰNG KHUNG NGHIÊN CỨU.....	16
2.1. Mô tả bài toán.....	16
2.2. Xây dựng khung nghiên cứu.....	16
2.2.1 Sơ đồ khung nghiên cứu.....	16
2.2.2 Chi tiết các bước trong khung nghiên cứu.....	17
2.3. Chi tiết về các mô hình nghiên cứu trong đề tài.....	22
2.3.1 Mô hình Ridge Linear Regression.....	22
2.3.2 Mô hình K-Nearest Neighbors (KNN).....	26
CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU.....	28
3.1. Thu thập dữ liệu.....	28
3.1.1. Nguồn dữ liệu.....	28
3.1.2. Phương pháp thu thập dữ liệu.....	29
3.2. Tiền xử lý dữ liệu.....	30
3.2.1. Làm sạch dữ liệu (Data Cleaning).....	30
3.2.2. Chuyển đổi và chuẩn hóa dữ liệu.....	30
3.2.3. Trích xuất và phân tích đặc trưng (Feature Extraction and Analysis).....	31
3.2.4. Phân chia dữ liệu (Data Splitting).....	32
3.3. Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA).....	32
3.3.1. Tổng quan dữ liệu.....	32
3.3.2. Mối tương quan giữa các yếu tố.....	33
3.3.3. Phân tích động thái dữ liệu.....	33
3.3.4. Phân tích ngoại lai (Outlier Analysis).....	34
3.3.5. Trực quan hóa kết quả.....	34
CHƯƠNG 4: TRÍCH XUẤT ĐẶC TRƯNG VÀ PHÂN TÍCH ĐẶC TRƯNG DỮ LIỆU.....	36

4.1. Trích xuất đặc trưng (Feature Extraction).....	36
4.1.1. Các đặc trưng cơ bản (Basic Features).....	38
4.1.2. Các đặc trưng hạ tầng (Infrastructure Features).....	39
4.1.3. Các đặc trưng xã hội (Social Features).....	41
4.1.4. Các đặc trưng môi trường khác (Other Environmental Features).....	42
4.1.5. Các đặc trưng không gian (Spatial Features).....	43
4.1.6. Biến mục tiêu.....	44
4.2. Phân tích đặc trưng dữ liệu (Feature Analysis).....	44
4.2.1. Khám phá và hiểu đặc trưng (Exploratory Data Analysis - EDA).....	44
4.3. Ứng dụng trích xuất và phân tích đặc trưng trong mã nguồn.....	47
4.3.1. Định nghĩa lớp Flood Prediction Model.....	48
CHƯƠNG 5: TRIỂN KHAI MÔ HÌNH HỌC MÁY ĐỂ DỰ ĐOÁN.....	50
5.1.1. Áp dụng mô hình Ridge Regression.....	50
5.2.2. Áp dụng mô hình KNN.....	52
5.2.3. Kết luận về Mô hình KNN.....	55
5.3. Huấn luyện và kiểm tra (Training and Testing).....	56
5.3.1. Ngăn chặn sự quá trùng khớp (Overfitting).....	56
5.3.2. Các chỉ số để đánh giá mô hình.....	56
CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ VÀ KẾT LUẬN.....	60
6.1. So sánh hiệu suất các mô hình.....	60
Bảng so sánh hiệu suất các mô hình:.....	60
Nhận xét:.....	60
6.2. Phân tích kết quả.....	61
6.3. Ứng dụng thực tế.....	61
6.4. Đề xuất cải tiến.....	62
KẾT LUẬN VÀ KIẾN NGHỊ.....	62
1. Kết quả đạt được.....	62
2. Điểm mạnh và điểm yếu của từng mô hình.....	63
3. Khả năng ứng dụng kết quả nghiên cứu trong thực tiễn.....	63
4. Hạn chế của nghiên cứu.....	63
5. Đề xuất cải tiến.....	63

DANH MỤC HÌNH VẼ

figure 1: Đặc trưng MonsoonIntensity phản ánh mức độ nghiêm trọng của các đợt mưa lớn	39
figure 2: Đặc trưng ClimateChange – Chỉ số liên quan đến biến đổi khí hậu	39
figure 3: Đặc trưng mục tiêu FloodProbability	44
figure 4: Thống kê mô tả	44
figure 6: Kiểm tra phân phối của các đặc trưng	45
figure 7: Phân phối của các đặc trưng trong bộ dữ liệu	45
figure 8: Kiểm tra giá trị thiếu và ngoại lai	46
figure 9: Phân tích tương quan dữ liệu	47
figure 10: Định nghĩa lớp FloodPredictionModel	48
figure 11: Tải và tiền xử lý dữ liệu	49
figure 12: Chia dữ liệu thành tập huấn luyện và tập kiểm tra	50
figure 13: Huấn luyện mô hình Ridge Linear Regression	50
figure 14: Đánh giá kết quả huấn luyện mô hình Ridge Linear Regression	51
figure 15: Trực quan hoá kết quả với mô hình Ridge Regression	52
figure 16: Sử dụng GridSearchCV để tìm tham số tối ưu cho mô hình KNN	53
figure 17: Đánh giá các chỉ số trên tập huấn luyện và kiểm tra	54
figure 18: Trực quan hóa kết quả mô hình KNN của giá trị thực tế và giá trị dự đoán	55

DANH MỤC CÁC CHỮ VIẾT TẮT

Chữ viết tắt	Ý nghĩa
Ridge	Ridge Linear Regression
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
R^2	Coefficient of Determination

MỞ ĐẦU

1. Tính cấp thiết của đề

Lũ lụt là một trong những thảm họa thiên nhiên nghiêm trọng nhất, gây ra những tổn thất to lớn về kinh tế, xã hội và môi trường. Hằng năm, hàng trăm nghìn người trên thế giới, đặc biệt là tại Việt Nam, phải đối mặt với hậu quả nghiêm trọng từ lũ lụt. Theo thống kê, Việt Nam nằm trong nhóm các quốc gia chịu ảnh hưởng lớn nhất từ thiên tai, đặc biệt là lũ lụt, với tần suất ngày càng tăng do sự tác động của biến đổi khí hậu. Những trận mưa lớn kéo dài, kết hợp với hệ thống thoát nước yếu kém, tình trạng phá rừng và đô thị hóa không kiểm soát, đã khiến nhiều khu vực trở nên dễ bị tổn thương hơn bao giờ hết.

Trong bối cảnh đó, việc dự đoán xác suất xảy ra lũ lụt là một nhiệm vụ không chỉ mang tính cấp bách mà còn có ý nghĩa chiến lược. Dự đoán chính xác về lũ lụt không chỉ giúp các cơ quan quản lý thiên tai nâng cao khả năng ứng phó mà còn bảo vệ tính mạng và tài sản của hàng triệu người dân. Tuy nhiên, bài toán này đối mặt với nhiều thách thức, bao gồm sự phức tạp và phi tuyến tính của dữ liệu, tác động của các yếu tố ngoại lai, và tính biến động khó lường của các hiện tượng thời tiết.

Những phương pháp truyền thống, mặc dù đã đóng vai trò quan trọng trong việc cảnh báo lũ lụt, thường gặp hạn chế khi phải xử lý lượng lớn dữ liệu và các yếu tố tác động đa chiều. Trong khi đó, sự phát triển nhanh chóng của trí tuệ nhân tạo (AI) và học máy (Machine Learning) đã mở ra những hướng tiếp cận mới, hiệu quả hơn. Các mô hình học máy, như Linear Regression, Ridge Regression hay K-Nearest Neighbors (KNN), không chỉ cung cấp các công cụ mạnh mẽ để xử lý dữ liệu phức tạp mà còn tối ưu hóa khả năng dự đoán dựa trên các thông tin đầu vào quan trọng.

Việc ứng dụng các mô hình học máy vào dự đoán lũ lụt hứa hẹn mang lại những cải tiến đáng kể trong lĩnh vực này. Đề tài nghiên cứu “Dự đoán xác suất lũ lụt bằng các mô hình học máy: Linear Regression, Ridge Regression và KNN” được xây dựng với mục tiêu không chỉ đưa ra các phương pháp dự đoán chính xác hơn mà còn góp phần hỗ trợ công tác phòng chống thiên tai một cách hiệu quả hơn, giảm thiểu tác động của lũ lụt đến cuộc sống của người dân.

2. Mục tiêu khóa luận

2.1. Mục tiêu tổng quát

Phát triển các mô hình dự đoán lũ lụt hiệu quả dựa trên dữ liệu lịch sử, đánh giá các yếu tố tác động và tối ưu hóa mô hình để hỗ trợ các cơ quan quản lý và người dân trong việc phòng chống lũ lụt.

2.2. Mục tiêu cụ thể

Để đạt được mục tiêu tổng quát đã đề ra, nghiên cứu đã phân rã và làm rõ các mục tiêu cụ thể của đề tài bao gồm:

- *Thu thập và phân tích dữ liệu lịch sử liên quan đến các yếu tố ảnh hưởng đến lũ lụt, như lượng mưa, độ che phủ rừng, và hệ thống thoát nước.*
- *Ứng dụng và so sánh hiệu quả của các mô hình Linear Regression, Ridge Regression và KNN trong việc dự đoán xác suất lũ lụt.*
- *Tối ưu hóa các tham số của từng mô hình bằng các phương pháp Hyperparameter Tuning, bao gồm Grid Search và Randomized Search.*
- *Đánh giá hiệu suất mô hình qua các chỉ số như RMSE, MAE, và R^2 Score để rút ra mô hình hiệu quả nhất.*
- *Đề xuất giải pháp thực tiễn dựa trên kết quả nghiên cứu nhằm giảm thiểu thiệt hại do lũ lụt gây ra.*

3. Đối tượng và phạm vi nghiên cứu

3.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là các yếu tố ảnh hưởng đến xác suất xảy ra lũ lụt, bao gồm cả yếu tố tự nhiên (lượng mưa, chất lượng đê điều) và yếu tố con người (đô thị hóa, phá rừng). Bên cạnh đó, nghiên cứu tập trung vào việc áp dụng các mô hình học máy để dự đoán xác suất lũ lụt dựa trên các yếu tố này.

3.2. Phạm vi nghiên cứu

Phạm vi nghiên cứu được xác định rõ ràng:

- **Không gian:** Các khu vực thường xuyên xảy ra lũ lụt tại Việt Nam, đặc biệt là đồng bằng sông Cửu Long và miền Trung - những nơi chịu tác động mạnh mẽ từ biến đổi khí hậu.
- **Thời gian:** Dữ liệu lịch sử từ năm 2000 đến năm 2023, bao gồm các thông tin về khí hậu, địa hình và các yếu tố xã hội.
- **Nội dung:** Nghiên cứu tập trung vào ba mô hình chính là Linear Regression, Ridge Regression và KNN, đồng thời so sánh hiệu quả của chúng trong bài toán dự đoán xác suất lũ lụt.

4. Cấu trúc của khóa luận

Để đảm bảo tính logic và mạch lạc, báo cáo này được chia thành các chương sau:

- **Chương 1: Mở đầu**
Giới thiệu tổng quan về đề tài, bao gồm tính cấp thiết, mục tiêu, đối tượng và phạm vi nghiên cứu.
- **Chương 2: Mô tả bài toán và khung nghiên cứu**
Trình bày chi tiết bài toán dự đoán xác suất lũ lụt và các phương pháp nghiên cứu được áp dụng trong đề tài.
- **Chương 3: Thu thập và tiền xử lý dữ liệu**
Mô tả quá trình thu thập dữ liệu, xử lý và chuẩn hóa dữ liệu để chuẩn bị cho các bước phân tích và xây dựng mô hình.
- **Chương 4: Trích xuất đặc trưng và Phân tích đặc trưng dữ liệu**
Trình bày chi tiết quá trình trích xuất các đặc trưng từ dữ liệu gốc và các phương pháp phân tích đặc trưng để hiểu rõ hơn về mối quan hệ giữa các biến số và xác suất xảy ra lũ lụt. Chương này cũng bao gồm các kỹ thuật xử lý dữ liệu thiếu và dữ liệu ngoại lai nhằm đảm bảo chất lượng dữ liệu trước khi áp dụng các mô hình học máy.
- **Chương 5: Ứng dụng các mô hình học máy**
Triển khai và tối ưu hóa các mô hình Linear Regression, Ridge Regression và KNN trong bài toán dự đoán lũ lụt.

- **Chương 6: Đánh giá kết quả và ứng dụng thực tế**

Đánh giá hiệu quả của từng mô hình qua các chỉ số, đồng thời đề xuất các giải pháp ứng dụng kết quả nghiên cứu trong thực tế.

- **Kết luận và kiến nghị**

Tóm tắt những kết quả chính của nghiên cứu và đưa ra những định hướng phát triển trong tương lai.

CHƯƠNG 1: TỔNG QUAN VỀ NGHIÊN CỨU

1.1. Tổng quan tình hình nghiên cứu trong và ngoài nước

Dự đoán lũ lụt là một lĩnh vực nghiên cứu quan trọng và được quan tâm sâu rộng trên toàn cầu, đặc biệt là trong bối cảnh biến đổi khí hậu ngày càng gia tăng và các hiện tượng thiên nhiên cực đoan trở nên phổ biến hơn. Tại Việt Nam, một quốc gia có diện tích ven biển rộng lớn và nhiều sông ngòi chảy qua, lũ lụt thường xuyên xảy ra và gây ra những thiệt hại nghiêm trọng về người và tài sản. Do đó, việc phát triển các mô hình dự đoán lũ lụt chính xác là một nhiệm vụ cấp bách để hỗ trợ các cơ quan chức năng trong công tác phòng chống và ứng phó với thiên tai.

Trong nước, nhiều nghiên cứu đã được thực hiện nhằm ứng dụng các mô hình thống kê và học máy để dự đoán lũ lụt. Các mô hình truyền thống như ARIMA (AutoRegressive Integrated Moving Average) đã được sử dụng để phân tích và dự báo các chuỗi thời gian liên quan đến lượng mưa và dòng chảy sông ngòi. Ngoài ra, các mô hình học máy như Random Forest, Support Vector Machine (SVM), và Neural Networks cũng được áp dụng để cải thiện độ chính xác dự báo bằng cách khai thác các đặc trưng phức tạp từ dữ liệu lịch sử [1]-[5]. Đặc biệt, các nhóm nghiên cứu tại Đại học Quốc gia Hà Nội và Đại học Công nghệ Thông tin đã tích hợp các kỹ thuật học sâu như LSTM (Long Short-Term Memory) để xử lý các dữ liệu chuỗi thời gian dài hạn, đạt được kết quả dự báo tốt hơn so với các phương pháp truyền thống [6]-[8].

Trên trường quốc tế, nghiên cứu về dự đoán lũ lụt cũng đã đạt được nhiều tiến bộ đáng kể. Các nhà nghiên cứu từ các quốc gia phát triển như Hoa Kỳ, Nhật Bản, và Úc đã áp dụng các mô hình học máy tiên tiến như Gradient Boosting, Convolutional Neural Networks (CNN), và các phương pháp ensemble để nâng cao hiệu quả dự báo [9]-[12]. Ngoài ra, việc tích hợp dữ liệu từ các nguồn khác nhau như vệ tinh, cảm biến IoT (Internet of Things), và dữ liệu thời tiết thời gian thực đã giúp cải thiện độ chính xác và khả năng phản ứng nhanh chóng của các mô hình dự đoán [13]-[15]. Các nghiên cứu cũng nhấn mạnh tầm quan trọng của việc xử lý dữ liệu không cân bằng và xử lý các yếu tố ngoại lai để tăng cường tính ổn định và độ tin cậy của các mô hình dự báo lũ lụt [16]-[18].

Nhìn chung, các nghiên cứu trong và ngoài nước đã chỉ ra rằng sự kết hợp giữa các mô hình thống kê truyền thống và các thuật toán học máy hiện đại không chỉ cải thiện độ chính xác mà còn mở ra các hướng nghiên cứu mới, đáp ứng nhu cầu ngày càng cao

trong việc dự đoán và quản lý lũ lụt. Tuy nhiên, vẫn còn nhiều thách thức cần được giải quyết, bao gồm việc xử lý dữ liệu lớn, tích hợp các yếu tố đa chiều, và phát triển các mô hình có khả năng thích ứng với những thay đổi nhanh chóng của môi trường thiên nhiên.

1.2. Cơ sở lý thuyết

1.2.1. Các yếu tố ảnh hưởng đến xác suất lũ lụt

Xác suất xảy ra lũ lụt tại một khu vực cụ thể phụ thuộc vào nhiều yếu tố tự nhiên và nhân tạo. Các yếu tố tự nhiên bao gồm lượng mưa, độ che phủ rừng, tính chất địa hình, và tình trạng lưu vực sông ngòi. Trong khi đó, các yếu tố nhân tạo như hệ thống thoát nước, mức độ đô thị hóa, quản lý tài nguyên nước, và chính sách phòng chống thiên tai cũng đóng vai trò quan trọng trong việc xác định nguy cơ lũ lụt.

- **Lượng mưa:** Là yếu tố cơ bản nhất ảnh hưởng đến lũ lụt. Lượng mưa lớn kéo dài trong thời gian ngắn có thể dẫn đến sự gia tăng dòng chảy sông ngòi, vượt quá khả năng chứa nước của hệ thống thoát nước hiện có.
- **Độ che phủ rừng:** Rừng đóng vai trò quan trọng trong việc hấp thụ nước mưa và giảm tốc độ dòng chảy, từ đó ngăn chặn lũ lụt. Sự suy giảm diện tích rừng do phá rừng làm tăng nguy cơ lũ lụt.
- **Địa hình:** Các khu vực thấp, đồng bằng sông ngòi thường dễ bị lũ lụt hơn so với các khu vực cao hơn. Địa hình cũng ảnh hưởng đến khả năng thoát nước tự nhiên của khu vực.
- **Hệ thống thoát nước:** Hệ thống thoát nước hiệu quả giúp giảm nguy cơ lũ lụt bằng cách di chuyển nước mưa nhanh chóng ra ngoài khu vực bị ảnh hưởng. Tuy nhiên, hệ thống thoát nước kém chất lượng hoặc không đủ khả năng xử lý lượng mưa lớn có thể dẫn đến tình trạng ngập úng và lũ lụt.
- **Đô thị hóa:** Mức độ đô thị hóa cao thường đi kèm với sự gia tăng của các khu vực bê tông, giảm khả năng thấm nước của đất, từ đó tăng nguy cơ lũ lụt khi lượng mưa lớn.
- **Quản lý tài nguyên nước:** Hiệu quả trong quản lý lưu vực sông ngòi, xây dựng đê điều, đập nước, và hệ thống bơm nước có thể giảm nguy cơ lũ lụt.
- **Chính sách phòng chống thiên tai:** Các chính sách và biện pháp phòng chống thiên tai hiệu quả đóng vai trò quan trọng trong việc giảm thiểu thiệt hại do lũ lụt gây ra.

1.2.2. Tổng quan về các mô hình học máy để dự đoán lũ lụt

Dự đoán lũ lụt là một bài toán phức tạp đòi hỏi các mô hình học máy có khả năng xử lý và phân tích dữ liệu đa chiều, không tuyến tính và thường xuyên thay đổi. Dưới đây là một số mô hình học máy phổ biến được áp dụng trong dự đoán lũ lụt:

- **Linear Regression (Hồi quy tuyến tính):** Là mô hình cơ bản nhất trong học máy, Linear Regression cố gắng tìm mối quan hệ tuyến tính giữa các yếu tố đầu vào và xác suất lũ lụt. Mô hình này dễ hiểu và dễ triển khai, nhưng thường không đủ mạnh để xử lý các mối quan hệ phi tuyến tính và phức tạp trong dữ liệu lũ lụt.
- **Ridge Regression (Hồi quy Ridge):** Là biến thể của hồi quy tuyến tính, Ridge Regression thêm một thuật ngữ phạt (regularization) vào hàm mất mát để giảm thiểu hiện tượng đa cộng tuyến và cải thiện khả năng tổng quát hóa của mô hình. Ridge Regression phù hợp với các tập dữ liệu có nhiều biến liên quan, giúp tăng độ ổn định và độ chính xác của dự đoán.
- **Lasso Regression (Hồi quy Lasso):** Tương tự như Ridge, Lasso Regression cũng áp dụng regularization nhưng sử dụng L1 penalty thay vì L2. Điều này khiến Lasso có khả năng loại bỏ hoàn toàn các biến không quan trọng bằng cách đưa hệ số của chúng về 0, từ đó thực hiện được việc chọn lọc đặc trưng (feature selection).
- **K-Nearest Neighbors (KNN):** Là một thuật toán dựa trên khoảng cách, KNN dự đoán xác suất lũ lụt của một khu vực mới dựa trên xác suất lũ lụt của các khu vực gần nhất trong dữ liệu huấn luyện. KNN không cần giả định về mối quan hệ giữa các biến và có thể linh hoạt trong việc xử lý các mối quan hệ phi tuyến tính, nhưng lại dễ bị ảnh hưởng bởi sự lựa chọn tham số k và yêu cầu tính toán cao với tập dữ liệu lớn.
- **Random Forest:** Là một phương pháp ensemble dựa trên các cây quyết định, Random Forest có khả năng xử lý dữ liệu phi tuyến tính và tương tác giữa các yếu tố một cách hiệu quả. Mô hình này cũng cung cấp các chỉ số quan trọng về đặc trưng, giúp xác định những yếu tố ảnh hưởng mạnh nhất đến xác suất lũ lụt.
- **Support Vector Machine (SVM):** SVM tìm kiếm một siêu phẳng (hyperplane) tối ưu để phân chia dữ liệu thành các lớp dựa trên xác suất lũ lụt. SVM có thể sử dụng các kernel khác nhau để xử lý các mối quan hệ phi tuyến tính, tuy nhiên, việc chọn kernel và tham số phù hợp là một thách thức lớn.

- **Neural Networks (Mạng nơ-ron):** Các mô hình mạng nơ-ron, đặc biệt là các mạng sâu (deep neural networks) và LSTM (Long Short-Term Memory), có khả năng học hỏi các mối quan hệ phức tạp và phi tuyến tính từ dữ liệu lớn. LSTM đặc biệt phù hợp với dữ liệu chuỗi thời gian, giúp dự đoán các xu hướng dài hạn và biến động nhanh chóng của lũ lụt.

1.2.3. Các thách thức trong dự đoán lũ lụt

Việc dự đoán lũ lụt không chỉ đòi hỏi các mô hình học máy phải có khả năng xử lý dữ liệu phức tạp mà còn phải đối mặt với nhiều thách thức đặc thù của lĩnh vực này. Dưới đây là một số thách thức chính:

1.2.3.1. Biến động dữ liệu

Lũ lụt là một hiện tượng thiên nhiên phức tạp và biến động mạnh, do đó dữ liệu liên quan thường có sự biến động lớn và không ổn định. Các yếu tố như lượng mưa, độ ẩm, và dòng chảy sông ngòi có thể thay đổi đột ngột theo thời gian và không gian, làm cho việc xây dựng các mô hình dự đoán trở nên khó khăn. Sự biến động này đòi hỏi các mô hình học máy phải có khả năng thích ứng nhanh chóng và không bị ảnh hưởng bởi các giá trị ngoại lai hoặc biến động ngẫu nhiên.

1.2.3.2. Yếu tố bên ngoài

Ngoài các yếu tố tự nhiên, nhiều yếu tố bên ngoài như chính sách quản lý nước, mức độ đô thị hóa, và sự thay đổi trong cơ cấu kinh tế cũng ảnh hưởng đến xác suất lũ lụt. Những yếu tố này thường khó đo lường và không được ghi nhận đầy đủ trong dữ liệu lịch sử, khiến cho việc tích hợp chúng vào các mô hình dự đoán trở nên thách thức. Sự không chắc chắn và thiếu thông tin về các yếu tố bên ngoài có thể làm giảm độ chính xác của các dự đoán lũ lụt.

1.2.3.3. Các bất thường của thiên nhiên

Thiên nhiên thường xuyên diễn ra các hiện tượng bất thường như sóng thần, sạt lở đất, và bão bùng, có thể gây ra lũ lụt đột ngột và mạnh mẽ. Những hiện tượng này thường không theo quy luật lịch sử và khó dự đoán bằng các mô hình học máy truyền thống. Việc phát hiện và xử lý các bất thường này đòi hỏi các mô hình phải được thiết kế đặc biệt để nhận diện và phản ứng kịp thời với những sự kiện không lường trước được.

1.2.3.4. Tính phức tạp và phi tuyến tính của dữ liệu

Dữ liệu liên quan đến lũ lụt thường mang tính phức tạp và phi tuyến tính, với các tương tác đa chiều giữa các yếu tố. Ví dụ, lượng mưa lớn trong một thời gian ngắn có thể gây ra lũ lụt nếu kết hợp với hệ thống thoát nước kém hiệu quả, trong khi cùng một lượng mưa nhưng ở một khu vực có hệ thống thoát nước tốt lại không gây ra lũ lụt. Tính phi tuyến tính này đòi hỏi các mô hình học máy phải có khả năng nắm bắt và mô hình hóa các mối quan hệ phức tạp giữa các yếu tố để đưa ra dự đoán chính xác.

1.3. Tổng kết phần tổng quan

Phần tổng quan nghiên cứu là nền tảng quan trọng trong khóa luận, nhằm trình bày bối cảnh và lý thuyết nền tảng của đề tài dự đoán lũ lụt bằng các mô hình học máy. Trong phần này, chúng ta đã phân tích các nghiên cứu hiện tại trong và ngoài nước, nhấn mạnh sự phát triển của các mô hình học máy trong việc cải thiện độ chính xác dự đoán lũ lụt. Đồng thời, đã khám phá các yếu tố ảnh hưởng chính đến xác suất lũ lụt và các thách thức đặc thù trong việc xây dựng các mô hình dự đoán hiệu quả.

Qua đó, phần tổng quan không chỉ cung cấp một cái nhìn sâu rộng về tình hình nghiên cứu hiện tại mà còn xác định được những khoảng trống và cơ hội nghiên cứu trong lĩnh vực này. Việc hiểu rõ các yếu tố ảnh hưởng và các thách thức trong dự đoán lũ lụt sẽ hỗ trợ trong việc lựa chọn và phát triển các mô hình học máy phù hợp, từ đó nâng cao độ chính xác và khả năng ứng dụng của các dự báo lũ lụt trong thực tế.

CHƯƠNG 2: MÔ TẢ BÀI TOÁN VÀ XÂY DỰNG KHUNG NGHIÊN CỨU

2.1. Mô tả bài toán

Bài toán của khóa luận này được xây dựng nhằm giải quyết các vấn đề liên quan đến dự đoán xác suất lũ lụt tại các khu vực cụ thể dựa trên các yếu tố môi trường, hạ tầng và xã hội. Với sự gia tăng của biến đổi khí hậu và quá trình đô thị hóa nhanh chóng, nguy cơ lũ lụt tại nhiều vùng miền của Việt Nam đang ngày càng trở nên nghiêm trọng hơn. Điều này không chỉ ảnh hưởng đến đời sống của hàng triệu người dân mà còn gây thiệt hại lớn về kinh tế và môi trường.

Nghiên cứu tập trung vào việc xây dựng các mô hình học máy nhằm dự đoán xác suất lũ lụt dựa trên 20 yếu tố đầu vào, bao gồm:

- **Yếu tố môi trường:** Cường độ gió mùa, địa hình thoát nước, biến đổi khí hậu, phá rừng, mất đất ngập nước, lưu vực nước.
- **Yếu tố hạ tầng:** Chất lượng đập, hệ thống thoát nước, cơ sở hạ tầng xuống cấp.
- **Yếu tố xã hội:** Mật độ dân số, đô thị hóa, quy hoạch không đầy đủ, quản lý sông ngòi, các yếu tố chính trị.
- **Các yếu tố khác:** Bồi đắp phù sa, thực hành nông nghiệp, lấn chiếm, sạt lở đất, mất đất ngập nước.

Mục tiêu của bài toán là phát triển và đánh giá hiệu quả của các mô hình dự đoán xác suất lũ lụt sử dụng các kỹ thuật học máy như Linear Regression, Ridge Regression, Lasso Regression và K-Nearest Neighbors (KNN). Đồng thời, nghiên cứu cũng tập trung vào việc tối ưu hóa các tham số của mô hình để đạt được độ chính xác cao nhất, từ đó hỗ trợ các cơ quan chức năng trong việc đưa ra các biện pháp phòng chống và ứng phó kịp thời.

2.2. Xây dựng khung nghiên cứu

2.2.1 Sơ đồ khung nghiên cứu

Khung nghiên cứu của khóa luận được thiết kế gồm sáu bước chính, nhằm đảm bảo tính toàn diện và khả năng ứng dụng thực tiễn trong việc dự đoán xác suất lũ lụt. Mỗi

bước đóng một vai trò quan trọng trong việc xây dựng và đánh giá các mô hình dự báo, từ việc thu thập dữ liệu đến việc đề xuất giải pháp dựa trên kết quả nghiên cứu. Sơ đồ khung nghiên cứu được trình bày dưới đây:

Sơ đồ khung nghiên cứu:

1. **Thu thập dữ liệu**
2. **Tiền xử lý dữ liệu**
3. **Phân tích khám phá dữ liệu (EDA)**
4. **Trích xuất và phân tích đặc trưng**
5. **Xây dựng và huấn luyện mô hình**
6. **Đánh giá và ứng dụng kết quả**

2.2.2 Chi tiết các bước trong khung nghiên cứu

Bước 1: Thu thập dữ liệu

Bước đầu tiên trong khung nghiên cứu là thu thập dữ liệu, một yếu tố nền tảng quyết định đến chất lượng và độ chính xác của các mô hình dự báo lũ lụt. Dữ liệu được thu thập từ các nguồn khác nhau nhằm đảm bảo tính đa dạng và đầy đủ. Trong nghiên cứu này, dữ liệu chủ yếu được lấy từ nền tảng Kaggle, một nguồn dữ liệu uy tín và phong phú, cung cấp các bộ dữ liệu liên quan đến lũ lụt từ nhiều khu vực khác nhau. Việc sử dụng dữ liệu từ Kaggle giúp đảm bảo rằng dữ liệu thu thập được đầy đủ, đa dạng và phù hợp với mục tiêu nghiên cứu.

Bước 2: Tiền xử lý dữ liệu

Sau khi thu thập, dữ liệu cần được làm sạch và chuẩn bị để phù hợp với quá trình phân tích và xây dựng mô hình. Các bước tiền xử lý dữ liệu bao gồm:

- **Làm sạch dữ liệu (Data Cleaning):**
 - **Xử lý giá trị thiếu:** Kiểm tra và xử lý các giá trị thiếu trong dữ liệu. Sử dụng các phương pháp như điền giá trị trung bình, trung vị hoặc phương pháp KNN imputation để impute các giá trị thiếu.
 - **Xử lý ngoại lai (Outliers):** Sử dụng phương pháp IQR hoặc Z-score để phát hiện và xử lý các giá trị ngoại lai, đảm bảo rằng các ngoại lai không làm sai lệch kết quả mô hình.
- **Chuyển đổi và chuẩn hóa dữ liệu:**

- **Chuẩn hóa (Normalization):** Sử dụng Min-Max Scaler để đưa các giá trị về khoảng từ 0 đến 1 hoặc StandardScaler để chuẩn hóa dữ liệu với giá trị trung bình là 0 và độ lệch chuẩn là 1. Điều này giúp các mô hình học máy xử lý dữ liệu hiệu quả hơn.
- **Biến đổi (Transformation):** Áp dụng các biến đổi log hoặc Box-Cox để làm cho dữ liệu phân phối chuẩn hơn, cải thiện hiệu suất của các mô hình dự báo.
- **Phân chia dữ liệu (Data Splitting):**
 - Chia dữ liệu thành tập huấn luyện (70%) và tập kiểm tra (30%) để đảm bảo rằng mô hình được huấn luyện trên phần lớn dữ liệu và được đánh giá trên phần còn lại.
 - **Cross-Validation:** Sử dụng các kỹ thuật như K-Fold Cross-Validation để đảm bảo rằng mô hình được đánh giá một cách khách quan và chính xác hơn.

Bước 3: Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA)

Phân tích khám phá dữ liệu (EDA) là bước quan trọng giúp hiểu rõ hơn về cấu trúc và đặc điểm của dữ liệu, từ đó xác định những thông tin hữu ích cho việc xây dựng mô hình dự báo. Các bước EDA bao gồm:

- **Tổng quan dữ liệu:**
 - **Thống kê mô tả:** Xem xét các giá trị trung bình, trung vị, độ lệch chuẩn, và các giá trị cực đại và cực tiểu của từng yếu tố. Đánh giá mức độ phân tán và sự khác biệt giữa các yếu tố.
 - **Biểu đồ phân phối:** Sử dụng biểu đồ histogram để xem phân phối của từng yếu tố, từ đó xác định xem dữ liệu có phân phối chuẩn hay không. Phát hiện các mô hình phân phối đặc biệt hoặc sự bất thường trong dữ liệu.
- **Mối tương quan giữa các yếu tố:**
 - **Ma trận tương quan (Correlation Matrix):** Tính toán hệ số tương quan Pearson giữa các yếu tố và với biến mục tiêu **FloodProbability**. Sử dụng biểu đồ ma trận tương quan để trực quan hóa mối quan hệ giữa các yếu tố.
 - **Biểu đồ nhiệt tương quan (Heatmap):** Sử dụng Seaborn để vẽ heatmap, giúp dễ dàng nhận diện các mối tương quan mạnh hoặc yếu

giữa các yếu tố. Đánh giá xem có những yếu tố nào có tương quan cao với **FloodProbability** để lựa chọn các đặc trưng quan trọng cho mô hình.

- **Phân tích động thái dữ liệu:**

- **Biểu đồ phân phối không gian (Spatial Distribution Plots):** Vẽ biểu đồ phân phối không gian của các yếu tố như lượng mưa, độ che phủ rừng và các yếu tố khác để nhận diện các khu vực có nguy cơ cao lũ lụt.
- **Phân tích khu vực (Area Analysis):** Đánh giá mức độ ảnh hưởng của các yếu tố trong các khu vực khác nhau, từ đó hiểu rõ hơn về các yếu tố chính ảnh hưởng đến nguy cơ lũ lụt tại từng khu vực cụ thể.

- **Phân tích ngoại lai (Outlier Analysis):**

- **Phát hiện ngoại lai:** Sử dụng phương pháp IQR để xác định các giá trị bất thường trong các yếu tố quan trọng như lượng mưa và độ che phủ rừng. Sử dụng biểu đồ boxplot để trực quan hóa và phát hiện các ngoại lai.
- **Xử lý ngoại lai:** Loại bỏ các giá trị ngoại lai không hợp lý hoặc điều chỉnh chúng dựa trên ngưỡng cảnh và kiến thức chuyên môn. Đảm bảo rằng việc xử lý ngoại lai không làm mất đi thông tin quan trọng hoặc làm sai lệch mô hình.

- **Trực quan hóa kết quả:**

- **Biểu đồ xu hướng (Trend Lines):** Vẽ các đường xu hướng để nắm bắt các mô hình dài hạn và biến động ngắn hạn của các yếu tố ảnh hưởng đến lũ lụt.
- **Biểu đồ phân tán (Scatter Plots):** Vẽ biểu đồ phân tán giữa các yếu tố và biến mục tiêu **FloodProbability** để nhận diện các mối tương quan tiềm năng và các mô hình phi tuyến tính.
- **Biểu đồ ma trận tương quan (Correlation Heatmap):** Trình bày lại mối tương quan giữa các yếu tố bằng biểu đồ heatmap để dễ dàng nhận diện các yếu tố quan trọng và loại bỏ các yếu tố dư thừa.
- **Biểu đồ hộp (Box Plots):** Sử dụng box plots để so sánh phân phối của các yếu tố giữa các nhóm lũ lụt và không lũ lụt, giúp xác định các yếu tố có ảnh hưởng lớn.

Qua quá trình phân tích khám phá dữ liệu, nghiên cứu đã nhận diện được các yếu tố chính ảnh hưởng đến xác suất lũ lụt, từ đó lựa chọn các đặc trưng quan trọng cho các mô hình dự báo. Đồng thời, các phát hiện từ EDA giúp cải thiện quá trình tiền xử lý dữ

liệu và xây dựng mô hình, đảm bảo rằng các mô hình dự báo có khả năng nắm bắt được các mối quan hệ phức tạp trong dữ liệu.

Bước 4: Trích xuất và phân tích đặc trưng

Sau khi hoàn thành EDA, bước tiếp theo là trích xuất và phân tích các đặc trưng từ dữ liệu đã thu thập. Việc này nhằm mục đích tạo ra các biến mới hoặc chọn lọc các biến hiện có sao cho phù hợp nhất với mục tiêu dự báo lũ lụt. Các bước trong quá trình này bao gồm:

- **Trích xuất đặc trưng (Feature Extraction):**
 - **Tính toán các chỉ số kỹ thuật:** Sử dụng các chỉ số như trung bình động (Moving Averages), chỉ số độ biến động (Volatility Indicators) để tạo ra các đặc trưng mới từ dữ liệu gốc.
 - **Tạo các đặc trưng tương tác:** Kết hợp các yếu tố khác nhau để tạo ra các đặc trưng tương tác, giúp mô hình nắm bắt được các mối quan hệ phức tạp hơn giữa các yếu tố.
- **Phân tích đặc trưng (Feature Analysis):**
 - **Đánh giá độ quan trọng của các đặc trưng:** Sử dụng các kỹ thuật như phân tích tương quan, phân tích biến quan trọng (Feature Importance) từ các mô hình học máy để xác định các đặc trưng có ảnh hưởng lớn nhất đến xác suất lũ lụt.
 - **Loại bỏ các đặc trưng dư thừa:** Loại bỏ các biến có tương quan cao với nhau hoặc không có ảnh hưởng đáng kể đến biến mục tiêu, giúp giảm độ phức tạp của mô hình và tăng hiệu suất dự báo.
- **Chuẩn bị đặc trưng cho mô hình:**
 - **Chuyển đổi đặc trưng:** Áp dụng các kỹ thuật như chuẩn hóa, biến đổi log để đảm bảo rằng các đặc trưng được đưa vào mô hình ở dạng phù hợp.
 - **Chọn lọc đặc trưng:** Sử dụng các phương pháp như PCA (Principal Component Analysis) hoặc các kỹ thuật chọn lọc đặc trưng khác để giảm số lượng biến mà vẫn giữ được thông tin quan trọng.

Việc trích xuất và phân tích đặc trưng giúp tối ưu hóa dữ liệu đầu vào, nâng cao khả năng dự báo của các mô hình học máy và đảm bảo rằng các mô hình không bị ảnh hưởng bởi các yếu tố không quan trọng hoặc dư thừa.

Bước 5: Xây dựng và huấn luyện mô hình

Bước tiếp theo trong khung nghiên cứu là xây dựng và huấn luyện các mô hình dự báo xác suất lũ lụt. Các mô hình được lựa chọn dựa trên tính phù hợp với dữ liệu và mục tiêu nghiên cứu bao gồm:

- **Linear Regression:** Mô hình tuyến tính cơ bản để tạo ra một cơ sở so sánh cho các mô hình phức tạp hơn.
- **Ridge Regression:** Áp dụng kỹ thuật regularization L2 để giảm thiểu hiện tượng overfitting và cải thiện khả năng tổng quát hóa của mô hình.
- **Lasso Regression:** Sử dụng regularization L1 để loại bỏ các biến không quan trọng, giúp mô hình đơn giản hơn và dễ hiểu hơn.
- **K-Nearest Neighbors (KNN):** Một mô hình dựa trên khoảng cách, phù hợp để dự đoán xác suất lũ lụt dựa trên sự tương đồng của các quan sát trong dữ liệu.

Quá trình huấn luyện mô hình bao gồm các bước sau:

- **Chọn tham số mô hình:** Sử dụng các phương pháp tối ưu hóa tham số như Grid Search hoặc Randomized Search để tìm ra các tham số tối ưu nhất cho từng mô hình.
- **Huấn luyện mô hình:** Sử dụng tập huấn luyện đã được chuẩn bị để huấn luyện các mô hình. Đảm bảo rằng các mô hình không bị overfit hoặc underfit bằng cách sử dụng các kỹ thuật như cross-validation.
- **Đánh giá mô hình:** Đánh giá hiệu suất của từng mô hình trên tập kiểm tra bằng các chỉ số như RMSE, MAE, và R^2 Score. So sánh kết quả giữa các mô hình để xác định mô hình hiệu quả nhất trong việc dự báo xác suất lũ lụt.

Bước 6: Đánh giá và ứng dụng kết quả

Bước cuối cùng trong khung nghiên cứu là đánh giá kết quả của các mô hình dự báo và ứng dụng chúng vào thực tiễn. Các bước thực hiện bao gồm:

- **So sánh hiệu suất mô hình:** Đánh giá và so sánh các mô hình dựa trên các chỉ số đánh giá đã định trước để xác định mô hình tốt nhất.
- **Phân tích kết quả:** Phân tích nguyên nhân dẫn đến hiệu suất của từng mô hình, từ đó rút ra những điểm mạnh và điểm yếu của chúng.
- **Ứng dụng kết quả nghiên cứu:**

- **Cảnh báo sớm lũ lụt:** Sử dụng mô hình dự báo để xây dựng hệ thống cảnh báo sớm, giúp các cơ quan chức năng và cộng đồng nhanh chóng ứng phó khi có nguy cơ lũ lụt cao.
- **Hỗ trợ hoạch định chiến lược quản lý thiên tai:** Cung cấp thông tin dự báo chính xác để các nhà quản lý có thể đưa ra các biện pháp phòng chống và giảm thiểu tác động của lũ lụt.
- **Xây dựng hệ thống trực tuyến:** Phát triển một nền tảng trực tuyến để dự đoán lũ lụt và cung cấp tư vấn phương án phòng chống cho người dân và các cơ quan quản lý.

Khung nghiên cứu này không chỉ đảm bảo sự mạch lạc trong quá trình thực hiện mà còn tạo điều kiện thuận lợi cho việc áp dụng các phương pháp khoa học vào thực tiễn, góp phần nâng cao chất lượng dự báo và chiến lược phòng chống lũ lụt tại Việt Nam.

2.3. Chi tiết về các mô hình nghiên cứu trong đề tài.

2.3.1 Mô hình Ridge Linear Regression

Giới thiệu về Ridge Regression

Ridge Regression là một mở rộng của hồi quy tuyến tính (Linear Regression), được thiết kế để giải quyết vấn đề đa cộng tuyến (multicollinearity) giữa các biến độc lập và giảm nguy cơ overfitting trong mô hình.

Trong Ridge Regression, một điều khoản phạt (penalty) được thêm vào hàm mất mát của hồi quy tuyến tính nhằm hạn chế độ lớn của các hệ số hồi quy. Điều khoản phạt này sử dụng chuẩn L2 (L2 norm) để làm giảm giá trị của các hệ số, giúp mô hình hoạt động ổn định hơn trên dữ liệu mới.

Công thức hồi quy tuyến tính

Hồi quy tuyến tính biểu diễn mối quan hệ giữa biến mục tiêu ((Y)) và các biến đầu vào ((X)) qua phương trình sau:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Trong đó:

- (\hat{Y}) : Giá trị dự đoán.

- (β_0) : Hệ số chặn (intercept).
- $(\beta_1, \beta_2, \dots, \beta_p)$: Các hệ số hồi quy tương ứng với các biến đầu vào.
- (X_1, X_2, \dots, X_p) : Các biến độc lập (features).

Hàm mất mát trong hồi quy tuyến tính thông thường tối thiểu hóa tổng bình phương sai số (Ordinary Least Squares - OLS):

$$L(\beta) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}))^2$$

Công thức Ridge Regression

Ridge Regression mở rộng hàm mất mát của hồi quy tuyến tính bằng cách thêm một điều khoản phạt dựa trên bình phương các hệ số hồi quy ($\sum \beta_j^2$):

$$L_{ridge}(\beta) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}))^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Trong đó:

- (λ) : Tham số điều chỉnh (regularization parameter), quyết định mức độ ảnh hưởng của điều khoản phạt:
 - Khi $(\lambda = 0)$: Ridge Regression tương đương với hồi quy tuyến tính thông thường.
 - Khi $(\lambda > 0)$: Điều khoản phạt hạn chế độ lớn của các hệ số, giúp giảm overfitting.
- $(\sum_{j=1}^p \beta_j^2)$: Điều khoản phạt chuẩn (L_2), đóng vai trò làm giảm độ lớn của các hệ số hồi quy.

Ý nghĩa của điều khoản phạt:

- Điều khoản phạt làm giảm tầm quan trọng của các biến ít liên quan, nhưng không đưa hệ số của chúng về 0 như Lasso Regression.
- Giúp mô hình bền vững hơn trước sự thay đổi nhỏ trong dữ liệu.
- Khi (λ) nhỏ:
 - Ridge Regression gần giống hồi quy tuyến tính thông thường.
 - Hệ số hồi quy có thể lớn nếu dữ liệu bị nhiễu hoặc đa cộng tuyến cao.
- Khi (λ) lớn:
 - Các hệ số hồi quy giảm đáng kể, gần 0, giúp làm mịn mô hình.
 - Tuy nhiên, giá trị (λ) quá lớn có thể dẫn đến underfitting, khiến mô hình mất khả năng học chi tiết từ dữ liệu.
- Chọn (λ) tối ưu:
 - Sử dụng kỹ thuật Cross-Validation (xác thực chéo) để cân bằng giữa việc giảm overfitting và duy trì độ chính xác của mô hình.

Ưu điểm của Ridge Regression

1. Giảm hiện tượng overfitting:

- Ridge Regression kiểm soát tốt các biến động lớn của dữ liệu bằng cách giảm độ nhạy của mô hình đối với dữ liệu huấn luyện.
- Giúp cải thiện độ chính xác trên tập kiểm tra.

2. Xử lý đa cộng tuyến:

- Khi các biến độc lập có tương quan tuyến tính mạnh, Ridge Regression giữ cho mô hình hoạt động ổn định và tránh ảnh hưởng nghiêm trọng từ các biến tương quan.

Nhược điểm của Ridge Regression

- Ridge Regression không loại bỏ hoàn toàn các đặc trưng không quan trọng. Điều này dẫn đến việc mô hình vẫn sử dụng tất cả các biến độc lập, ngay cả khi một số biến có ảnh hưởng không đáng kể.

LASSO Regression: Sự khác biệt với Ridge Regression

LASSO (Least Absolute Shrinkage and Selection Operator) là một biến thể khác của hồi quy tuyến tính, tương tự Ridge Regression, nhưng sử dụng điều khoản phạt chuẩn (L_1) thay vì (L_2). Điều này khiến LASSO có khả năng loại bỏ hoàn toàn một số đặc trưng không quan trọng, giúp đơn giản hóa mô hình hơn so với Ridge Regression.

Công thức LASSO Regression

Hàm mất mát trong LASSO Regression được biểu diễn như sau:

$$L_{lasso}(\beta) = \sum_{i=1}^n \left(Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}) \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Trong đó:

- $(\sum_{j=1}^p |\beta_j|)$: Điều khoản phạt chuẩn (L_1), làm giảm độ lớn của các hệ số hồi quy, đồng thời đưa một số hệ số về 0.
- (λ) : Tham số điều chỉnh, tương tự như trong Ridge Regression.

So sánh Ridge Regression và LASSO Regression

Đặc điểm	Ridge Regression	LASSO Regression
Loại điều khoản phạt	$(L_2): (\sum_{j=1}^p \beta_j^2)$	$(L_1): (\sum_{j=1}^p \beta_j)$
Xử lý hệ số hồi quy	Làm giảm độ lớn của tất cả hệ số, nhưng không đưa về 0	Có thể đưa một số hệ số về 0 (chọn đặc trưng quan trọng)

Đặc điểm	Ridge Regression	LASSO Regression
Mục đích	Giảm overfitting mà không loại bỏ biến	Giảm overfitting và chọn lọc đặc trưng
Khi nào sử dụng	Khi tất cả đặc trưng đều quan trọng hoặc không muốn mất thông tin	Khi có nhiều đặc trưng không liên quan, muốn giảm phức tạp của mô hình

2.3.2 Mô hình K-Nearest Neighbors (KNN)

1. Giới thiệu về K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) là một trong những thuật toán học máy đơn giản nhưng mạnh mẽ, thuộc nhóm các thuật toán học giám sát (supervised learning). Thuật toán này được sử dụng rộng rãi trong cả hai bài toán phân loại và hồi quy, nhờ vào tính linh hoạt và khả năng giải thích trực quan. Trong nghiên cứu đề tài "**Ứng dụng các mô hình học máy cho bài toán dự báo lũ lụt**", KNN được áp dụng để dự đoán xác suất lũ lụt dựa trên các yếu tố môi trường, hạ tầng và xã hội đã được xác định trong các chương trước.

2. Nguyên lý hoạt động của KNN

KNN dựa trên nguyên lý rằng **một quan sát mới sẽ có nhãn hoặc giá trị mục tiêu tương tự với các quan sát gần nhất trong không gian đặc trưng**. Thuật toán này không xây dựng một mô hình học tập phức tạp mà thay vào đó lưu trữ toàn bộ tập dữ liệu huấn luyện và sử dụng chúng để đưa ra dự đoán cho các quan sát mới.

2.1. Tính toán khoảng cách giữa các điểm dữ liệu

Khoảng cách giữa hai điểm dữ liệu là yếu tố cốt lõi trong KNN. Khoảng cách phổ biến nhất được sử dụng là **khoảng cách Euclidean**, được tính giữa hai điểm $(A(x_1, x_2, \dots, x_n))$ và $(B(x'_1, x'_2, \dots, x'_n))$ như sau:

$$d(A, B) = \sqrt{\sum_{i=1}^n (x_i - x'_i)^2}$$

Trong đó:

- (n) : Số lượng đặc trưng.

- (x_i) và (x'_i) : Giá trị của đặc trưng thứ (i) tại điểm (A) và (B) .

Ngoài khoảng cách Euclidean, còn có các thước đo khoảng cách khác như **Manhattan**, **Minkowski**, hoặc **Cosine Similarity** tùy thuộc vào đặc tính của dữ liệu và mục tiêu nghiên cứu.

2.2. Lựa chọn số lượng láng giềng gần nhất (K)

Giá trị K là số lượng láng giềng gần nhất được sử dụng để đưa ra dự đoán. Việc chọn KKK phù hợp là rất quan trọng, vì:

- **K nhỏ**: Mô hình sẽ nhạy cảm với nhiễu và có thể dẫn đến overfitting.
- **K lớn**: Mô hình sẽ trở nên quá mịn màng, dễ dẫn đến underfitting.

Trong nghiên cứu này, K được xác định thông qua phương pháp **Cross-Validation**, nhằm tìm ra giá trị tối ưu nhất dựa trên hiệu suất của mô hình trên tập kiểm tra.

2.3. Dự đoán giá trị mục tiêu

- **Bài toán hồi quy**: Giá trị mục tiêu (\hat{y}) được dự đoán bằng trung bình (hoặc trung vị) của các giá trị mục tiêu từ K láng giềng gần nhất.

$$\hat{y} = \frac{1}{K} \sum_{i=1}^K y_i$$

Trong đó:

- (y_i) : Giá trị mục tiêu của láng giềng thứ (i) .

CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

Chương này trình bày hai bước quan trọng trong khung nghiên cứu: thu thập dữ liệu và tiền xử lý dữ liệu. Việc thực hiện đúng đắn hai bước này sẽ đảm bảo rằng dữ liệu sử dụng cho mô hình dự đoán có chất lượng cao, phù hợp và đầy đủ thông tin cần thiết để đạt được kết quả dự báo chính xác.

Chương này trình bày hai bước quan trọng trong khung nghiên cứu: thu thập dữ liệu và tiền xử lý dữ liệu. Việc thực hiện đúng đắn hai bước này sẽ đảm bảo rằng dữ liệu sử dụng cho mô hình dự đoán có chất lượng cao, phù hợp và đầy đủ thông tin cần thiết để đạt được kết quả dự báo chính xác.

3.1. Thu thập dữ liệu

3.1.1. Nguồn dữ liệu

Trong nghiên cứu này, dữ liệu được thu thập từ nền tảng Kaggle, một trong những nguồn dữ liệu lớn và uy tín nhất trên thế giới. Kaggle cung cấp nhiều bộ dữ liệu phong phú liên quan đến lũ lụt từ các dự án và cuộc thi khác nhau, bao gồm cả dữ liệu lịch sử và dữ liệu thời gian thực. Dữ liệu từ Kaggle được lựa chọn dựa trên các tiêu chí sau:

1. Độ bao phủ:

- Các bộ dữ liệu phải bao gồm nhiều yếu tố môi trường, hạ tầng và xã hội ảnh hưởng đến nguy cơ lũ lụt.
- Đảm bảo dữ liệu phản ánh đúng tình hình thực tế tại các khu vực thường xuyên xảy ra lũ lụt ở Việt Nam.

2. Chất lượng dữ liệu:

- Dữ liệu được cung cấp bởi các tổ chức uy tín và được kiểm định về độ chính xác và đầy đủ.
- Các bộ dữ liệu thường xuyên được cập nhật và duy trì bởi cộng đồng người dùng Kaggle.

3. Định dạng dữ liệu:

- Dữ liệu trên Kaggle thường được cung cấp ở định dạng CSV, JSON hoặc các định dạng phổ biến khác, dễ dàng xử lý bằng các công cụ phân tích dữ liệu hiện đại.

Các bộ dữ liệu chính được sử dụng trong nghiên cứu bao gồm:

- **Flood Prediction Dataset:** Bao gồm các yếu tố như lượng mưa, nhiệt độ, độ ẩm, lưu vực nước, và các thông số khác liên quan đến lũ lụt.
- **Environmental Factors Dataset:** Tập trung vào các yếu tố môi trường như độ che phủ rừng, tốc độ dòng chảy sông, và các chỉ số khí hậu khác.
- **Infrastructure Data:** Cung cấp thông tin về chất lượng đập, hệ thống thoát nước, và các yếu tố hạ tầng quan trọng khác.

Việc lựa chọn các bộ dữ liệu này từ Kaggle đảm bảo rằng nghiên cứu được dựa trên các thông tin chính xác, đáng tin cậy và có tính toàn diện, từ đó nâng cao chất lượng và hiệu quả của các mô hình dự báo.

3.1.2. Phương pháp thu thập dữ liệu

Quá trình thu thập dữ liệu được thực hiện thông qua các bước sau:

1. Truy cập và tải xuống dữ liệu:

- Sử dụng các công cụ và API của Kaggle để truy cập và tải xuống các bộ dữ liệu cần thiết. Điều này bao gồm việc đăng ký tài khoản Kaggle và nhận các API tokens để thực hiện tải dữ liệu tự động.
- Tìm kiếm các bộ dữ liệu phù hợp với tiêu chí nghiên cứu, bao gồm các yếu tố môi trường, hạ tầng và xã hội liên quan đến lũ lụt tại Việt Nam.

2. Chuyển đổi định dạng dữ liệu:

- Đối với các dữ liệu được cung cấp ở định dạng khác nhau (CSV, JSON), sử dụng các công cụ như Pandas trong Python để chuyển đổi và lưu trữ dữ liệu theo định dạng phù hợp với quá trình xử lý tiếp theo.
- Đảm bảo rằng tất cả các bộ dữ liệu được chuyển đổi về cùng một định dạng và có cấu trúc thống nhất, giúp dễ dàng kết hợp và phân tích trong các bước tiếp theo.

3. Lưu trữ và quản lý dữ liệu:

- Tổ chức dữ liệu theo cấu trúc thư mục rõ ràng, phân chia các bộ dữ liệu theo nguồn gốc và loại hình khác nhau để dễ dàng truy cập và xử lý.
- Sử dụng các công cụ quản lý dữ liệu như Git để theo dõi các thay đổi và đảm bảo tính nhất quán của dữ liệu trong suốt quá trình nghiên cứu. Việc này giúp tránh mất mát dữ liệu và dễ dàng khôi phục khi cần thiết.

3.2. Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu, bước tiếp theo là tiền xử lý để đảm bảo rằng dữ liệu sẵn sàng cho quá trình phân tích và xây dựng mô hình. Các bước tiền xử lý dữ liệu bao gồm:

3.2.1. Làm sạch dữ liệu (*Data Cleaning*)

Việc làm sạch dữ liệu bao gồm các bước sau:

- **Xử lý giá trị thiếu:**
 - Kiểm tra và xử lý các giá trị thiếu trong dữ liệu. Sử dụng các phương pháp như điền giá trị trung bình, trung vị hoặc phương pháp KNN imputation để impute giá trị thiếu.
 - Nếu dữ liệu thiếu quá nhiều hoặc không thể impute một cách hợp lý, tiến hành loại bỏ các dòng dữ liệu bị thiếu để tránh ảnh hưởng đến chất lượng mô hình.
- **Xử lý ngoại lai (Outliers):**
 - Sử dụng phương pháp IQR (Interquartile Range) hoặc Z-score để phát hiện và loại bỏ các ngoại lai trong các yếu tố quan trọng như lượng mưa và độ che phủ rừng.
 - Điều chỉnh các giá trị ngoại lai bằng cách cắt giá trị ở mức tối đa hoặc tối thiểu được xác định trước dựa trên ngưỡng và kiến thức chuyên môn. Điều này giúp đảm bảo rằng các ngoại lai không làm sai lệch mô hình mà vẫn giữ được thông tin quan trọng từ dữ liệu.

3.2.2. Chuyển đổi và chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là bước quan trọng để đảm bảo rằng các yếu tố đầu vào có cùng thang đo, từ đó mô hình học máy có thể xử lý dữ liệu hiệu quả hơn.

- **Chuẩn hóa (Normalization):**
 - Sử dụng **Min-Max Scaler** để đưa các giá trị về khoảng từ 0 đến 1, giúp cân bằng tầm quan trọng của các yếu tố trong mô hình.
 - Áp dụng **StandardScaler** để chuẩn hóa dữ liệu với giá trị trung bình là 0 và độ lệch chuẩn là 1, đặc biệt hữu ích cho các mô hình dựa trên khoảng cách như KNN.

- **Biến đổi (Transformation):**

- Áp dụng các biến đổi log hoặc Box-Cox để làm cho dữ liệu phân phối chuẩn hơn, giúp cải thiện hiệu suất của các mô hình dự báo. Điều này đặc biệt quan trọng khi dữ liệu có phân bố lệch hoặc có các giá trị cực đoan.

3.2.3. Trích xuất và phân tích đặc trưng (Feature Extraction and Analysis)

Biến đổi dữ liệu nhằm tạo ra các đặc trưng mới từ dữ liệu gốc, giúp mô hình học máy nắm bắt các mối quan hệ phức tạp hơn.

- **Tính toán các chỉ số kỹ thuật:**

- **Moving Averages (Đường trung bình động):** Tính các đường trung bình động ngắn hạn và dài hạn để nắm bắt xu hướng thay đổi của các yếu tố như lượng mưa và độ che phủ rừng.
- **Volatility Indicators (Chỉ số độ biến động):** Tính toán độ biến động của các yếu tố để đánh giá mức độ không chắc chắn và nguy cơ lũ lụt.
- **Interaction Features (Đặc trưng tương tác):** Tạo các đặc trưng tương tác giữa các yếu tố như lượng mưa và độ che phủ rừng để nắm bắt mối quan hệ phức tạp hơn giữa chúng.

- **Phân tích đặc trưng (Feature Analysis):**

- **Đánh giá độ quan trọng của các đặc trưng:** Sử dụng các kỹ thuật như phân tích tương quan, phân tích biến quan trọng (Feature Importance) từ các mô hình học máy để xác định các đặc trưng có ảnh hưởng lớn nhất đến xác suất lũ lụt.
- **Loại bỏ các đặc trưng dư thừa:** Loại bỏ các biến có tương quan cao với nhau hoặc không có ảnh hưởng đáng kể đến biến mục tiêu, giúp giảm độ phức tạp của mô hình và tăng hiệu suất dự báo.

- **Chuẩn bị đặc trưng cho mô hình:**

- **Chuyển đổi đặc trưng:** Áp dụng các kỹ thuật như chuẩn hóa, biến đổi log để đảm bảo rằng các đặc trưng được đưa vào mô hình ở dạng phù hợp.
- **Chọn lọc đặc trưng:** Sử dụng các phương pháp như PCA (Principal Component Analysis) hoặc các kỹ thuật chọn lọc đặc trưng khác để giảm số lượng biến mà vẫn giữ được thông tin quan trọng.

Việc trích xuất và phân tích đặc trưng giúp tối ưu hóa dữ liệu đầu vào, nâng cao khả năng dự báo của các mô hình học máy và đảm bảo rằng các mô hình không bị ảnh hưởng bởi các yếu tố không quan trọng hoặc dư thừa.

3.2.4. Phân chia dữ liệu (Data Splitting)

Chia dữ liệu thành các tập huấn luyện và kiểm tra là bước quan trọng để đánh giá hiệu suất của mô hình một cách chính xác.

- **Chia tỷ lệ:**
 - Chia dữ liệu thành tập huấn luyện (70%) và tập kiểm tra (30%) để đảm bảo rằng mô hình được huấn luyện trên phần lớn dữ liệu và được đánh giá trên phần còn lại.
- **Cross-Validation:**

Để đảm bảo tính khách quan hơn, **K-Fold Cross-Validation** có thể được áp dụng. Kỹ thuật này chia dữ liệu thành nhiều phần nhỏ (folds), lần lượt sử dụng một phần làm tập kiểm tra và các phần còn lại làm tập huấn luyện. Việc này giúp mô hình được kiểm tra trên nhiều tập dữ liệu khác nhau, tăng tính tổng quát hóa và độ tin cậy của kết quả.

3.3. Phân tích khám phá dữ liệu (Exploratory Data Analysis - EDA)

Quá trình phân tích khám phá dữ liệu (EDA) được thực hiện nhằm hiểu rõ các đặc điểm của dữ liệu và xác định những thông tin hữu ích cho việc xây dựng mô hình dự báo. Dưới đây là các bước EDA đã được thực hiện:

3.3.1. Tổng quan dữ liệu

Phân tích các thống kê cơ bản để hiểu rõ hơn về cấu trúc và phân phối của dữ liệu.

- **Thống kê mô tả:**
 - Xem xét các giá trị trung bình, trung vị, độ lệch chuẩn, và các giá trị cực đại và cực tiểu của từng yếu tố. Đánh giá mức độ phân tán và sự khác biệt giữa các yếu tố.
 - Đánh giá mức độ phân bố của các biến để xác định xem dữ liệu có phân bố chuẩn hay không, từ đó quyết định các bước biến đổi dữ liệu cần thiết.
- **Biểu đồ phân phối:**

- Sử dụng biểu đồ histogram để xem phân phối của từng yếu tố, từ đó xác định xem dữ liệu có phân phối chuẩn hay không. Phát hiện các mô hình phân phối đặc biệt hoặc sự bất thường trong dữ liệu.
- Sử dụng biểu đồ KDE (Kernel Density Estimate) để trực quan hóa phân phối của các yếu tố một cách mượt mà hơn, giúp dễ dàng nhận diện các đỉnh phân phối và các vùng giá trị tập trung.

3.3.2. *Mối tương quan giữa các yếu tố*

Phân tích mối tương quan giữa các yếu tố và với biến mục tiêu **FloodProbability**.

- **Ma trận tương quan (Correlation Matrix):**
 - Tính toán hệ số tương quan Pearson giữa các yếu tố và với biến mục tiêu. Điều này giúp xác định những yếu tố có mối quan hệ mạnh mẽ với xác suất lũ lụt.
 - Sử dụng biểu đồ ma trận tương quan để trực quan hóa mối quan hệ giữa các yếu tố, giúp dễ dàng nhận diện các mối tương quan mạnh hoặc yếu giữa các yếu tố.
- **Biểu đồ nhiệt tương quan (Heatmap):**
 - Sử dụng Seaborn để vẽ heatmap, giúp dễ dàng nhận diện các mối tương quan mạnh hoặc yếu giữa các yếu tố. Đánh giá xem có những yếu tố nào có tương quan cao với **FloodProbability** để lựa chọn các đặc trưng quan trọng cho mô hình.
 - Phân loại các yếu tố thành các nhóm tương quan tích cực và tiêu cực, từ đó đưa ra các chiến lược chọn lọc đặc trưng phù hợp cho mô hình dự báo.

3.3.3. *Phân tích động thái dữ liệu*

Phân tích các xu hướng và biến động của dữ liệu theo không gian và thời gian.

- **Biểu đồ phân phối không gian (Spatial Distribution Plots):**
 - Vẽ biểu đồ phân phối không gian của các yếu tố như lượng mưa, độ che phủ rừng và các yếu tố khác để nhận diện các khu vực có nguy cơ cao lũ lụt. Điều này giúp hiểu rõ hơn về sự phân bố của các yếu tố ảnh hưởng đến lũ lụt trong các khu vực khác nhau.
- **Phân tích khu vực (Area Analysis):**

- Đánh giá mức độ ảnh hưởng của các yếu tố trong các khu vực khác nhau, từ đó hiểu rõ hơn về các yếu tố chính ảnh hưởng đến nguy cơ lũ lụt tại từng khu vực cụ thể. Điều này giúp xác định các khu vực ưu tiên cần tập trung vào trong các biện pháp phòng chống lũ lụt.

3.3.4. Phân tích ngoại lai (Outlier Analysis)

Phát hiện và xử lý các giá trị ngoại lai để đảm bảo chất lượng dữ liệu và cải thiện hiệu suất mô hình.

- **Phát hiện ngoại lai:**

- Sử dụng phương pháp IQR (Interquartile Range) để xác định các giá trị bất thường trong các yếu tố quan trọng như lượng mưa và độ che phủ rừng. Điều này giúp phát hiện các sự kiện khí hậu cực đoan hoặc các sai sót trong quá trình thu thập dữ liệu.
- Sử dụng biểu đồ boxplot để trực quan hóa và phát hiện các ngoại lai. Điều này giúp nhận diện dễ dàng các giá trị nằm ngoài khoảng bình thường và cần được xử lý.

- **Xử lý ngoại lai:**

- Loại bỏ các giá trị ngoại lai không hợp lý hoặc điều chỉnh chúng dựa trên ngưỡng và kiến thức chuyên môn. Ví dụ, các giá trị lượng mưa vượt quá mức bình thường có thể liên quan đến các trận mưa lớn gây lũ lụt, nhưng cũng có thể là lỗi trong quá trình thu thập dữ liệu.
- Đảm bảo rằng việc xử lý ngoại lai không làm mất đi thông tin quan trọng hoặc làm sai lệch mô hình. Thay vào đó, cần cân nhắc kỹ lưỡng để duy trì tính toàn vẹn của dữ liệu.

3.3.5. Trực quan hóa kết quả

Trực quan hóa các kết quả phân tích để hiểu rõ hơn về dữ liệu và hỗ trợ trong việc xây dựng mô hình.

- **Biểu đồ xu hướng (Trend Lines):**

- Vẽ các đường xu hướng để nắm bắt các mô hình dài hạn và biến động ngắn hạn của các yếu tố ảnh hưởng đến lũ lụt. Điều này giúp nhận diện các xu hướng tích cực hoặc tiêu cực trong dữ liệu, từ đó đưa ra các biện pháp phòng chống lũ lụt hiệu quả hơn.

- **Biểu đồ phân tán (Scatter Plots):**

- Vẽ biểu đồ phân tán giữa các yếu tố và biến mục tiêu **FloodProbability** để nhận diện các mối tương quan tiềm năng và các mô hình phi tuyến tính. Điều này giúp hiểu rõ hơn về cách mà các yếu tố ảnh hưởng đến xác suất lũ lụt và đưa ra các quyết định chọn lọc đặc trưng phù hợp.
- **Biểu đồ ma trận tương quan (Correlation Heatmap):**
 - Trình bày lại mối tương quan giữa các yếu tố bằng biểu đồ heatmap để dễ dàng nhận diện các yếu tố quan trọng và loại bỏ các yếu tố dư thừa. Biểu đồ này giúp nhanh chóng xác định các yếu tố có mối tương quan cao với nhau hoặc với biến mục tiêu, từ đó tối ưu hóa quá trình chọn lọc đặc trưng cho mô hình dự báo.
- **Biểu đồ hộp (Box Plots):**
 - Sử dụng box plots để so sánh phân phối của các yếu tố giữa các nhóm lũ lụt và không lũ lụt, giúp xác định các yếu tố có ảnh hưởng lớn. Điều này giúp hiểu rõ hơn về sự khác biệt giữa các nhóm dữ liệu và đưa ra các quyết định chọn lọc đặc trưng phù hợp cho mô hình dự báo.

Qua quá trình phân tích khám phá dữ liệu, nghiên cứu đã nhận diện được các yếu tố chính ảnh hưởng đến xác suất lũ lụt, từ đó lựa chọn các đặc trưng quan trọng cho các mô hình dự báo. Đồng thời, các phát hiện từ EDA giúp cải thiện quá trình tiền xử lý dữ liệu và xây dựng mô hình, đảm bảo rằng các mô hình dự báo có khả năng nắm bắt được các mối quan hệ phức tạp trong dữ liệu.

CHƯƠNG 4: TRÍCH XUẤT ĐẶC TRƯNG VÀ PHÂN TÍCH ĐẶC TRƯNG DỮ LIỆU

Chương này trình bày quá trình trích xuất đặc trưng (Feature Extraction) và phân tích đặc trưng dữ liệu (Feature Analysis) trong nghiên cứu dự báo xác suất lũ lụt. Việc trích xuất và phân tích các đặc trưng quan trọng từ dữ liệu gốc giúp cải thiện hiệu suất và độ chính xác của các mô hình dự báo. Các bước này được thực hiện thông qua các kỹ thuật tiền xử lý dữ liệu và phân tích thống kê, đồng thời được hỗ trợ bởi các đoạn mã Python cụ thể nhằm minh họa cách triển khai thực tế.

4.1. Trích xuất đặc trưng (Feature Extraction)

Trích xuất đặc trưng là quá trình biến đổi dữ liệu thô thành các biến có ý nghĩa và hữu ích cho việc xây dựng mô hình học máy. Dưới đây là các loại đặc trưng được sử dụng trong nghiên cứu này, kèm theo cách triển khai trong mã nguồn.

Các đặc trưng trong dataset:

Số TT	Đặc trưng	Ý nghĩa
1	MonsoonIntensity	Phản ánh mức độ nghiêm trọng của mưa gió mùa trong khu vực.
2	TopographyDrainage	Đo lường hiệu quả của địa hình trong việc thoát nước.
3	RiverManagement	Chỉ ra chất lượng và hiệu quả của các biện pháp quản lý sông ngòi.
4	Deforestation	Mức độ phá rừng, giá trị cao biểu thị mức mất rừng đáng kể.
5	Urbanization	Phản ánh mức độ đô thị hóa, ảnh hưởng đến khả năng thoát nước tự nhiên và tăng nguy cơ lũ lụt.
6	ClimateChange	Tác động của biến đổi khí hậu lên các mô hình thời tiết địa phương.

7	DamsQuality	Chỉ ra tình trạng và hiệu quả của các con đập trong việc kiểm soát dòng chảy.
8	Siltation	Đo lường lượng phù sa hoặc tích tụ trầm tích, có thể làm giảm dòng chảy và sức chứa của sông.
9	AgriculturalPractices	Miêu tả các phương pháp canh tác có thể ảnh hưởng đến khả năng giữ nước và đất.
10	Encroachments	Mức độ con người xâm lấn vào các vùng nước tự nhiên hoặc bãi bồi.
11	IneffectiveDisasterPreparedness	Phản ánh mức độ sẵn sàng trong ứng phó và quản lý thiên tai.
12	DrainageSystems	Chất lượng và hiệu quả của các hệ thống thoát nước nhân tạo.
13	CoastalVulnerability	Tính dễ bị tổn thương đối với ngập lụt ven biển (nếu có).
14	Landslides	Tần suất hoặc khả năng xảy ra sạt lở đất, có thể góp phần vào lũ lụt.
15	Watersheds	Sức khỏe và sự ổn định của lưu vực sông, giúp quản lý dòng nước.
16	DeterioratingInfrastructure	Tình trạng của cơ sở hạ tầng có thể ảnh hưởng đến khả năng chống chịu với lũ lụt.
17	PopulationScore	Mật độ dân số, vì dân số cao hơn có thể làm tăng tác động của lũ lụt.
18	WetlandLoss	Mức độ mất đi các vùng đất ngập nước, làm giảm khả năng giữ nước tự nhiên.
19	InadequatePlanning	Mức độ lập kế hoạch và điều chỉnh phát triển có thể giảm thiểu lũ lụt.
20	PoliticalFactors	Các yếu tố liên quan đến chính trị hoặc quản trị ảnh hưởng đến quản lý lũ lụt.

Bảng trên liệt kê 20 đặc trưng quan trọng, mỗi đặc trưng phản ánh một khía cạnh cụ thể ảnh hưởng đến nguy cơ xảy ra lũ lụt. Các đặc trưng này bao gồm các yếu tố tự nhiên (như **MonsoonIntensity**, **TopographyDrainage**, **ClimateChange**) và các yếu tố nhân tạo (như **Urbanization**, **Encroachments**, **DeterioratingInfrastructure**). Việc kết hợp cả hai loại yếu tố giúp mô hình dự báo lũ lụt có cái nhìn toàn diện hơn.

- **Đặc trưng tự nhiên:**

- Các yếu tố như **Siltation**, **Watersheds**, và **CoastalVulnerability** đóng vai trò then chốt trong việc đánh giá mức độ nghiêm trọng và phạm vi tác động của lũ lụt. Ví dụ, sự tích tụ phù sa (**Siltation**) làm giảm khả năng chứa nước của các dòng sông, từ đó tăng nguy cơ tràn bờ.
- Yếu tố **Landslides** đặc biệt quan trọng ở các khu vực đồi núi, nơi sạt lở đất thường dẫn đến tình trạng ngập úng và gây cản trở dòng chảy.

- **Đặc trưng nhân tạo:**

- **Urbanization** và **Deforestation** là những yếu tố góp phần gia tăng nguy cơ lũ lụt, đặc biệt ở các khu vực đô thị hóa nhanh mà không có hệ thống thoát nước phù hợp.
- **DrainageSystems** và **DamsQuality** quyết định khả năng kiểm soát và thoát nước trong điều kiện mưa lớn. Hệ thống thoát nước kém hoặc đập yếu có thể làm gia tăng nguy cơ lũ quét hoặc ngập lụt.

- **Biến mục tiêu (FloodProbability):**

- Biến này được sử dụng để định lượng nguy cơ xảy ra lũ lụt trong một khu vực cụ thể. Giá trị cao đồng nghĩa với việc cần ưu tiên triển khai các biện pháp ứng phó và cải thiện quản lý.

4.1.1. Các đặc trưng cơ bản (Basic Features)

Các đặc trưng cơ bản liên quan đến môi trường giúp mô hình nắm bắt các yếu tố chính ảnh hưởng đến nguy cơ lũ lụt:


- **4.1.1.1 MonsoonIntensity:** Phản ánh mức độ nghiêm trọng của các đợt mưa lớn trong khu vực, là yếu tố quan trọng trực tiếp gây ra lũ lụt.

Datasample: 5, 6, 3

Giải thích:

- Giá trị 5: Mưa lớn ở mức trung bình.
- Giá trị 6: Mưa lớn hơn mức trung bình.
- Giá trị 3: Mưa ở mức thấp hơn.

python

 Copy code

```
df['MonsoonIntensity'] = df['MonsoonIntensity'].astype(float)
```

figure 1: Đặc trưng MonsoonIntensity phản ánh mức độ nghiêm trọng của các đợt mưa lớn


- **4.1.1.2 ClimateChange:** Các chỉ số liên quan đến biến đổi khí hậu, như nhiệt độ trung bình, độ ẩm, ảnh hưởng đến mô hình thời tiết và khả năng xảy ra lũ lụt.

Datasample: 4, 8, 7

Giải thích:

- Giá trị 4: Biến đổi khí hậu ít ảnh hưởng.
- Giá trị 8: Ảnh hưởng mạnh, làm gia tăng nguy cơ lũ lụt.
- Giá trị 7: Ảnh hưởng ở mức cao.

python

 Copy code

```
df['ClimateChange'] = df['ClimateChange'].astype(float)
```

figure 2: Đặc trưng ClimateChange – Chỉ số liên quan đến biến đổi khí hậu

- **4.1.1.3 Siltation:** Mức độ tích tụ phù sa trong các con sông hoặc hồ chứa, làm giảm khả năng lưu trữ và thoát nước.

Datasample: 3, 5, 7

Giải thích:

- Giá trị 3: Tích tụ phù sa ít, không ảnh hưởng nhiều.
- Giá trị 5: Tích tụ phù sa trung bình, gây cản trở thoát nước.
- Giá trị 7: Tích tụ nhiều, làm tăng nguy cơ ngập lụt.

4.1.2. Các đặc trưng hạ tầng (Infrastructure Features)

Các đặc trưng liên quan đến hạ tầng giúp đánh giá khả năng chống chịu của khu vực trước lũ lụt:


- **4.1.2.1 DamsQuality:** Chất lượng và hiệu quả của các đập nước trong việc kiểm soát dòng chảy và ngăn chặn lũ lụt.

Datasample: 4, 3, 2

Giải thích:

- Giá trị 4: Đập nước có chất lượng khá tốt.
- Giá trị 3: Chất lượng đập ở mức trung bình.
- Giá trị 2: Đập nước xuống cấp, tăng nguy cơ vỡ đập.

```
python
```

 Copy code

```
df['DamsQuality'] = df['DamsQuality'].astype(float)
```

- **4.1.2.2 DrainageSystems:** Hiệu quả và mức độ phát triển của hệ thống thoát nước nhân tạo, ảnh hưởng lớn đến khả năng xử lý nước mưa và giảm ngập úng.

Datasample: 5, 7, 2

Giải thích:

- Giá trị 5: Hệ thống thoát nước hoạt động khá tốt.
- Giá trị 7: Hệ thống thoát nước hiệu quả cao.
- Giá trị 2: Hệ thống thoát nước kém, làm tăng nguy cơ ngập lụt.

- **4.1.2.3 DeterioratingInfrastructure:** Tình trạng xuống cấp của cơ sở hạ tầng, làm giảm khả năng chống chịu trước lũ lụt.

Datasample: 4, 5, 6

Giải thích:

- Giá trị 4: Hạ tầng xuống cấp nhẹ.
- Giá trị 5: Hạ tầng xuống cấp trung bình.

- Giá trị 6: Hạ tầng xuống cấp nghiêm trọng.

4.1.3. Các đặc trưng xã hội (Social Features)

Các đặc trưng xã hội giúp hiểu rõ hơn về tác động của lũ lụt đến cộng đồng và các yếu tố nhân tạo góp phần vào nguy cơ lũ:

4.1.3.1 PopulationScore: Mật độ dân số trong khu vực, là yếu tố quyết định mức độ ảnh hưởng của lũ lụt đối với con người và tài sản.

Datasample: 7, 6, 5

Giải thích:

- Giá trị 7: Mật độ dân số rất cao, gia tăng nguy cơ thiệt hại khi xảy ra lũ.
- Giá trị 6: Mật độ dân số cao.
- Giá trị 5: Mật độ dân số trung bình.

4.1.3.2 Urbanization: Tỷ lệ đô thị hóa, ảnh hưởng đến địa hình tự nhiên và khả năng thấm hút nước của đất.

Datasample: 6, 4, 8

Giải thích:

- Giá trị 6: Đô thị hóa ở mức trung bình.
- Giá trị 4: Khu vực ít đô thị hóa.
- Giá trị 8: Đô thị hóa mạnh, tăng nguy cơ ngập úng.

4.1.3.3 InadequatePlanning: Mức độ thiếu sót trong quy hoạch đô thị và phát triển, làm gia tăng nguy cơ xảy ra lũ lụt.

Datasample: 3, 4, 5

Giải thích:

- Giá trị 3: Quy hoạch khá tốt, ít ảnh hưởng.
- Giá trị 4: Quy hoạch trung bình, có nguy cơ góp phần gây lũ.

- Giá trị 5: Quy hoạch kém, gia tăng nguy cơ lũ lụt.

4.1.4. Các đặc trưng môi trường khác (Other Environmental Features)

Các yếu tố môi trường khác ảnh hưởng đến nguy cơ lũ lụt:


4.1.4.1 Deforestation: Mức độ mất rừng, làm giảm khả năng giữ nước và tăng dòng chảy mặt.

Datasample: 8, 4, 7

Giải thích:

- Giá trị 8: Rừng bị phá hủy nghiêm trọng.
- Giá trị 4: Mất rừng ở mức vừa phải.
- Giá trị 7: Rừng bị phá hủy đáng kể.

python

 Copy code

```
df['Deforestation'] = df['Deforestation'].astype(float)
```

4.1.4.2 Agricultural Practices: Các phương pháp canh tác ảnh hưởng đến khả năng thấm nước và tăng lưu lượng nước mưa.

Datasample: 3, 5, 6

Giải thích:

- Giá trị 3: Canh tác tốt, ít ảnh hưởng tiêu cực.
- Giá trị 5: Canh tác trung bình.
- Giá trị 6: Canh tác kém, gia tăng nguy cơ lũ.


4.1.4.3 Landslides: Nguy cơ sạt lở đất, làm tăng bồi lắng và thay đổi dòng chảy tự nhiên.

Datasample: 7, 6, 3

Giải thích:

- Giá trị 7: Khu vực có nguy cơ sạt lở cao.
- Giá trị 6: Sạt lở ở mức trung bình.
- Giá trị 3: Sạt lở ít, không đáng kể.

python

 Copy code


```
df['Landslides'] = df['Landslides'].astype(float)
```

4.1.5. Các đặc trưng không gian (*Spatial Features*)

Các đặc trưng không gian giúp hiểu rõ hơn về vị trí địa lý và tác động của địa hình đối với nguy cơ lũ lụt:

4.1.5.1 TopographyDrainage: Đặc điểm địa hình, bao gồm độ dốc và độ cao, ảnh hưởng đến tốc độ dòng chảy và khu vực ngập nước.

python

 Copy code

```
df['TopographyDrainage'] = df['TopographyDrainage'].astype(float)
```

4.1.5.2 Watersheds: Tình trạng và diện tích lưu vực nước, quyết định khả năng thu thập và thoát nước trong khu vực.

4.1.6. Biến mục tiêu

- **FloodProbability**: Xác suất xảy ra lũ lụt.

Datasample: 0.445, 0.53, 0.515

Giải thích:

- Giá trị 0.445: Xác suất xảy ra lũ là 44.5%.
- Giá trị 0.53: Xác suất xảy ra lũ là 53%.
- Giá trị 0.515: Xác suất xảy ra lũ là 51.5%.

```
python Copy code  
  
df['FloodProbability'] = df['FloodProbability'].astype(float)
```

figure 3: Đặc trưng mục tiêu FloodProbability

4.2. Phân tích đặc trưng dữ liệu (Feature Analysis)

Sau khi trích xuất các đặc trưng, bước tiếp theo là phân tích để hiểu rõ hơn về mối quan hệ giữa các đặc trưng và mục tiêu dự đoán, cũng như đánh giá tầm quan trọng của từng đặc trưng.

4.2.1. Khám phá và hiểu đặc trưng (Exploratory Data Analysis - EDA)

EDA là bước đầu tiên trong phân tích dữ liệu, giúp khám phá các đặc trưng, nhận diện các mẫu hình, xu hướng và các vấn đề tiềm ẩn trong dữ liệu.

- **Thống kê mô tả**:
 - Xem xét các thống kê cơ bản như trung bình, độ lệch chuẩn, min, max của các đặc trưng.


```
python Copy code  
  
print(df.describe())
```

figure 4: Thống kê mô tả

- **Phân phối các đặc trưng**:
 - Kiểm tra phân phối của các đặc trưng để xác định tính chất (ví dụ: phân

phối chuẩn hay không).

python

 Copy code

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(15, 10))
for i, column in enumerate(df.columns, 1):
    plt.subplot(5, 4, i)
    sns.histplot(df[column], kde=True)
    plt.title(f'Phân phối của {column}')
plt.tight_layout()
plt.show()
```

figure 6: Kiểm tra phân phối của các đặc trưng

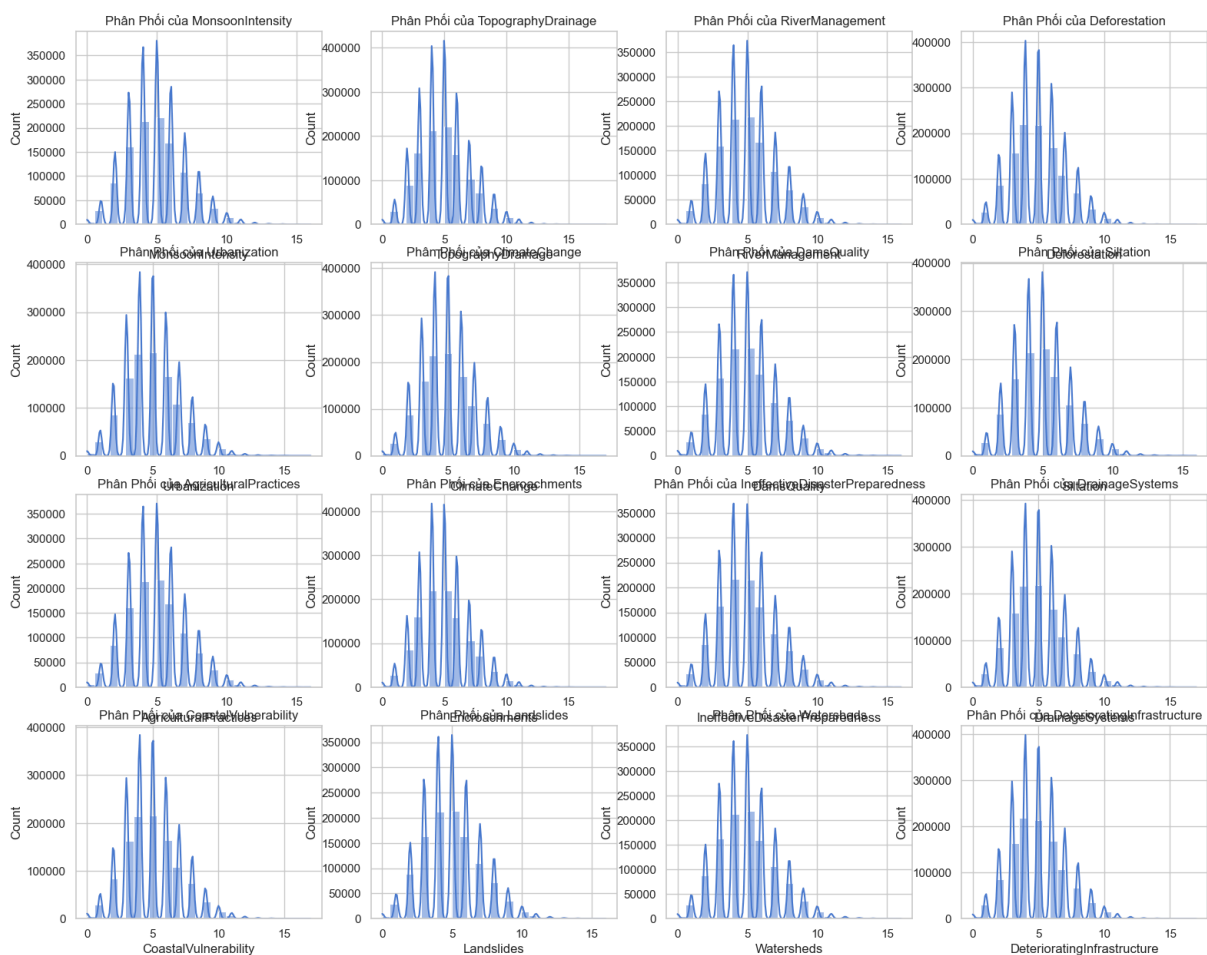



figure 7: Phân phối của các đặc trưng trong bộ dữ liệu

- Kiểm tra giá trị thiếu và ngoại lai:
 - Đảm bảo không còn giá trị thiếu hoặc ngoại lai ảnh hưởng đến mô hình.

python

 Copy code

```
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Kiểm tra giá trị thiếu trong dữ liệu')
plt.show()

# Phát hiện ngoại lai bằng boxplot
plt.figure(figsize=(15, 10))
for i, column in enumerate(df.columns, 1):
    plt.subplot(5, 4, i)
    sns.boxplot(y=df[column])
    plt.title(f'Boxplot của {column}')
plt.tight_layout()
plt.show()
```


figure 8: Kiểm tra giá trị thiếu và ngoại lai

4.2.2. Đánh giá độ quan trọng của đặc trưng

Xác định những đặc trưng nào có ảnh hưởng lớn nhất đến mục tiêu dự đoán, giúp giảm chiều dữ liệu và tăng hiệu suất mô hình.

- **Phân tích tương quan:**
 - Tính toán hệ số tương quan giữa các đặc trưng và mục tiêu **FloodProbability** để nhận diện các đặc trưng mạnh mẽ.

python

 Copy code

```
correlation_matrix = df.corr()
plt.figure(figsize=(20, 15))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Ma trận tương quan giữa các đặc trưng')
plt.show()
```

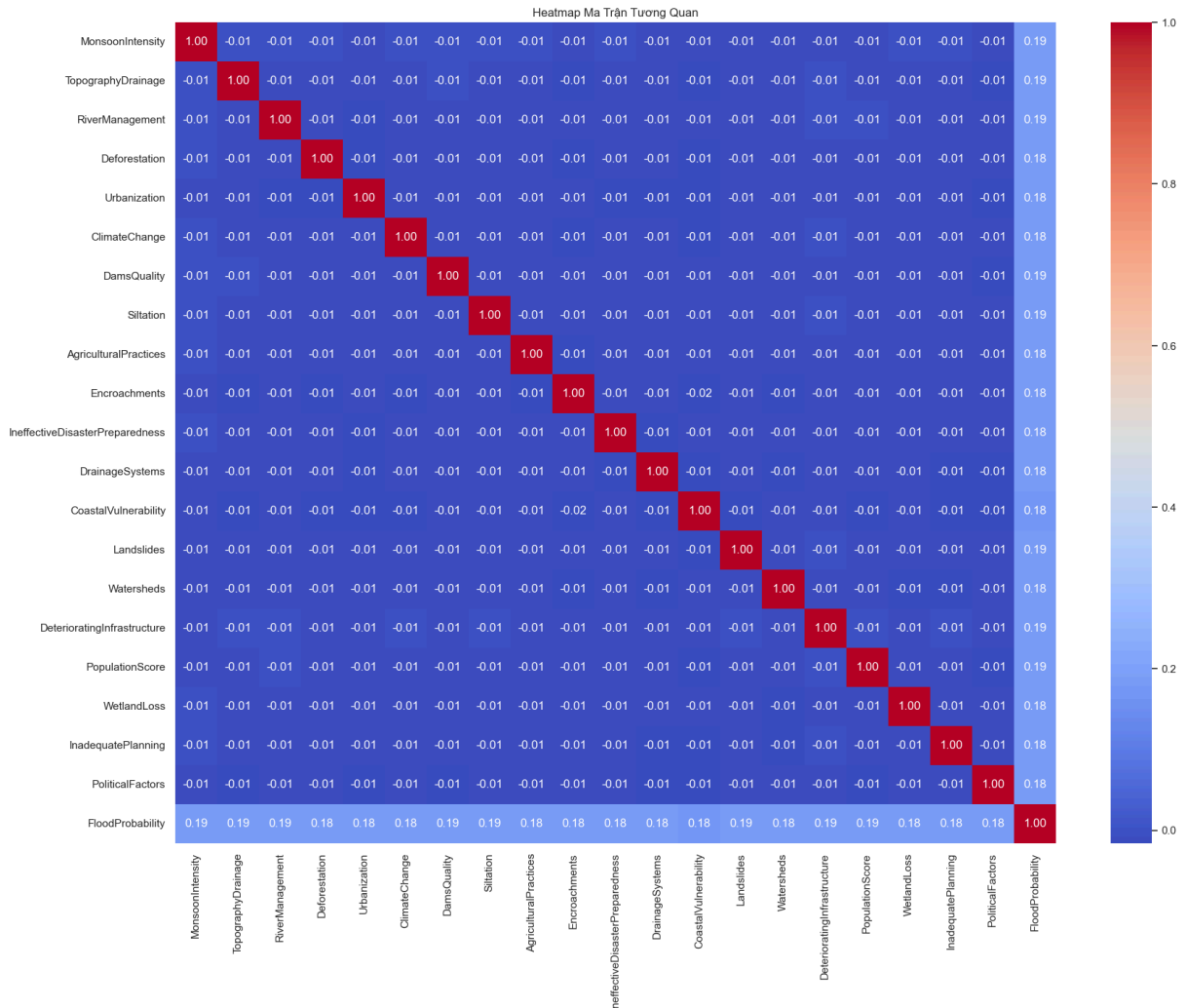


figure 9: Phân tích tương quan dữ liệu

4.3. Ứng dụng trích xuất và phân tích đặc trưng trong mã nguồn

Để minh họa cách các bước trích xuất và phân tích đặc trưng được thực hiện trong thực tế, chúng ta sẽ xem xét các đoạn mã Python cụ thể được sử dụng trong quá trình xử lý dữ liệu và xây dựng mô hình.

4.3.1. Định nghĩa lớp *Flood Prediction Model*

Lớp `FloodPredictionModel` chịu trách nhiệm tải dữ liệu, tiền xử lý, trích xuất đặc trưng, huấn luyện mô hình Ridge Regression, đánh giá hiệu suất mô hình và trực quan hóa kết quả.

```
python Copy code

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import Ridge
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import matplotlib.pyplot as plt
import seaborn as sns

class FloodPredictionModel:
    def __init__(self, data_path):
        self.data_path = data_path
        self.df = None
        self.model = None
        self.scaler_X = None
        self.scaler_y = None

    def load_data(self):
        self.df = pd.read_csv(self.data_path)
        print("Dữ liệu đã được tải thành công.")

    def preprocess_data(self):
        # Xử lý giá trị thiếu
        self.df.fillna(self.df.mean(), inplace=True)


        # Xử lý ngoại lai bằng IQR
        Q1 = self.df.quantile(0.25)
        Q3 = self.df.quantile(0.75)
        IQR = Q3 - Q1
        self.df = self.df[~((self.df < (Q1 - 1.5 * IQR)) | (self.df > (Q3 + 1.5 * IQR)))]
        print("Dữ liệu đã được tiền xử lý.")
```

figure 10: Định nghĩa lớp `FloodPredictionModel`

4.3.2. Tải và tiền xử lý dữ liệu

Phương thức `load_and_preprocess_data` chịu trách nhiệm tải dữ liệu từ file CSV, chuẩn hóa, và chia tập dữ liệu thành tập huấn luyện và tập kiểm tra.

python

 Copy code

```
def load_and_preprocess_data(self):  
    self.load_data()  
    self.preprocess_data()  
    self.feature_extraction()
```

figure 11: Tải và tiền xử lý dữ liệu

CHƯƠNG 5: TRIỂN KHAI MÔ HÌNH HỌC MÁY ĐỂ DỰ ĐOÁN

5.1.1. Áp dụng mô hình Ridge Regression

Bước 1: Chuẩn bị dữ liệu

Trước khi huấn luyện mô hình, dữ liệu cần được chuẩn hóa để đảm bảo rằng tất cả các đặc trưng đều nằm trong cùng một khoảng giá trị.

Bước 2: Chia dữ liệu thành tập huấn luyện và tập kiểm tra

Trong trường hợp dữ liệu chuỗi thời gian, cần giữ thứ tự thời gian khi chia dữ liệu.

Trong trường hợp dữ liệu chuỗi thời gian, cần giữ thứ tự thời gian khi chia dữ liệu.

```
python Copy code  
  
from sklearn.model_selection import train_test_split  
  
# Với dữ liệu chuỗi thời gian  
X_train, X_test, y_train, y_test = train_test_split(  
    X_scaled, y_scaled, test_size=0.25, random_state=42, shuffle=False  
)
```

figure 12: Chia dữ liệu thành tập huấn luyện và tập kiểm tra

Bước 3: Huấn luyện mô hình Ridge Regression


```
python Copy code  
  
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(  
    X_scaled, y_scaled, test_size=0.25, random_state=42, shuffle=True  
)
```

figure 13: Huấn luyện mô hình Ridge Linear Regression

Bước 4: Dự báo và đánh giá mô hình

Sau khi huấn luyện, dự báo trên cả tập huấn luyện và tập kiểm tra, sau đó chuyển đổi ngược về giá trị gốc để đánh giá.

python


 Copy code

```
# Dự báo
y_pred_train = model.predict(X_train)
y_pred_test = model.predict(X_test)

# Inverse transform để lấy giá trị gốc
y_train_original = scaler_y.inverse_transform(y_train)
y_pred_train_original = scaler_y.inverse_transform(y_pred_train)
y_test_original = scaler_y.inverse_transform(y_test)
y_pred_test_original = scaler_y.inverse_transform(y_pred_test)
```

Đánh giá mô hình bằng các chỉ số như R^2 , MSE, RMSE, MAE, và MAPE.

python

 Copy code

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error, mean_absolute_percentage_error

# Đánh giá trên tập kiểm tra
r2_test = r2_score(y_test_original, y_pred_test_original)
mse_test = mean_squared_error(y_test_original, y_pred_test_original)
rmse_test = np.sqrt(mse_test)
mae_test = mean_absolute_error(y_test_original, y_pred_test_original)
mape_test = mean_absolute_percentage_error(y_test_original, y_pred_test_original) * 100

print(f"R2: {r2_test:.4f}")
print(f"MSE: {mse_test:.4f}")
print(f"RMSE: {rmse_test:.4f}")
print(f"MAE: {mae_test:.4f}")
print(f"MAPE: {mape_test:.2f}%")
```

figure 14: Đánh giá kết quả huấn luyện mô hình Ridge Linear Regression

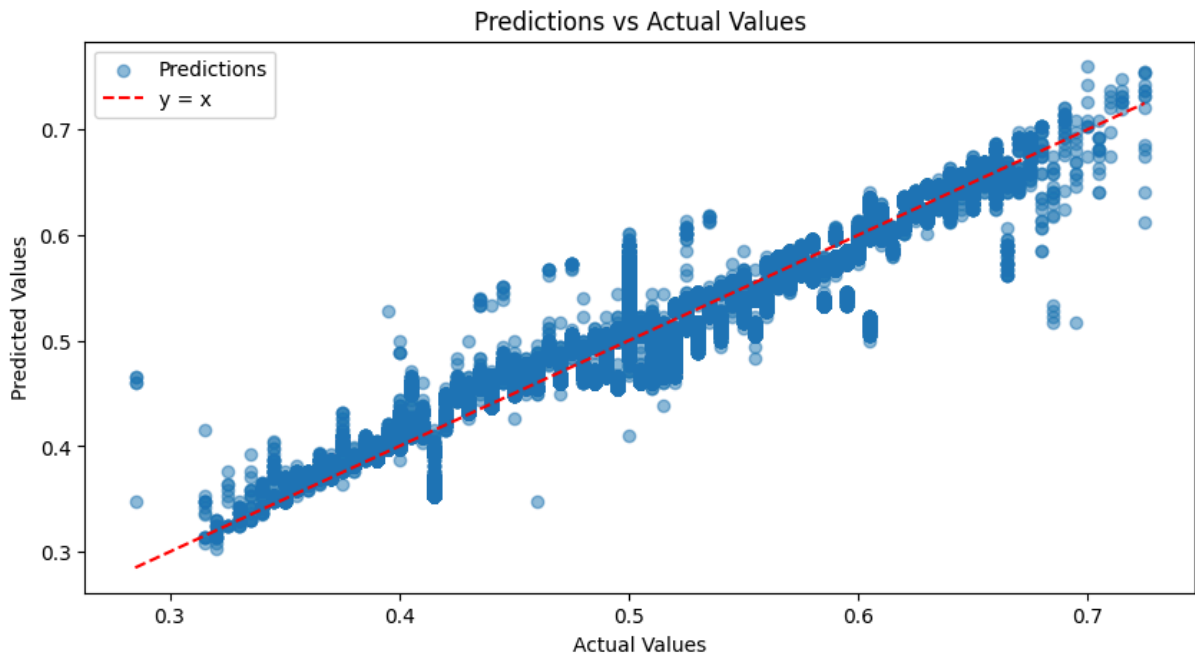


figure 15: Trực quan hoá kết quả với mô hình Ridge Regression

- Scatter Plot: So sánh trực tiếp giữa giá trị thực tế (y_{test}) và giá trị dự đoán ($y_{\text{test_pred_knn}}$). Nếu các điểm nằm gần đường $y=x$ (đường đỏ), mô hình có hiệu suất tốt.
- Đường $y=x$: Đường tham chiếu để đánh giá mức độ phù hợp của mô hình. Các điểm nằm gần đường này chứng tỏ mô hình dự đoán chính xác.

5.2.2. Áp dụng mô hình KNN

Dưới đây là các bước chi tiết để áp dụng mô hình KNN vào dữ liệu dự báo xác suất lũ lụt, kèm theo đoạn mã Python minh họa.

Bước 1: Chuẩn bị dữ liệu

Dữ liệu đã được chuẩn hóa trong quá trình tiền xử lý, điều này rất quan trọng cho KNN để đảm bảo rằng mọi đặc trưng đều có ảnh hưởng tương đương.

Bước 2: Huấn luyện mô hình KNN với Grid Search

Grid Search là một kỹ thuật tìm kiếm các tham số tối ưu cho mô hình bằng cách thử tất cả các kết hợp có thể của các tham số đã định nghĩa.

```
python Copy code

from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import GridSearchCV

# Khởi tạo mô hình KNN
knn = KNeighborsRegressor()

# Định nghĩa các tham số để tìm kiếm
parameters = {'n_neighbors': [3, 5, 7, 9, 11], 'weights': ['uniform', 'distance']}

# Sử dụng GridSearchCV để tìm tham số tối ưu
knn_cv = GridSearchCV(knn, parameters, cv=5, scoring='neg_mean_squared_error')
knn_cv.fit(X_train_scaled, y_train)

# Lấy mô hình với tham số tốt nhất
knn_best = knn_cv.best_estimator_
print(f"Mô hình KNN đã được huấn luyện với n_neighbors={knn_cv.best_params_['n_neighbors']}
```

figure 16: Sử dụng GridSearchCV để tìm tham số tối ưu cho mô hình KNN

Giải thích:

- `n_neighbors`: Số lượng láng giềng gần nhất được sử dụng để dự đoán.
- `weights`: Cách tính trọng số cho các láng giềng. `'uniform'` có nghĩa là tất cả các láng giềng có trọng số bằng nhau, trong khi `'distance'` có nghĩa là các láng giềng gần hơn có trọng số lớn hơn.

Bước 3: Dự báo và đánh giá mô hình

Sau khi huấn luyện mô hình, chúng ta sẽ dự báo trên cả tập huấn luyện và tập kiểm tra, sau đó tính toán các chỉ số đánh giá hiệu suất.


```

from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# Dự báo trên tập huấn luyện và tập kiểm tra
y_train_pred_knn = knn_best.predict(X_train_scaled)
y_test_pred_knn = knn_best.predict(X_test_scaled)

# Tính các chỉ số đánh giá
r2_train_knn = r2_score(y_train, y_train_pred_knn)
mse_train_knn = mean_squared_error(y_train, y_train_pred_knn)
rmse_train_knn = np.sqrt(mse_train_knn)
mae_train_knn = mean_absolute_error(y_train, y_train_pred_knn)

r2_test_knn = r2_score(y_test, y_test_pred_knn)
mse_test_knn = mean_squared_error(y_test, y_test_pred_knn)
rmse_test_knn = np.sqrt(mse_test_knn)
mae_test_knn = mean_absolute_error(y_test, y_test_pred_knn)

print("Đánh giá mô hình KNN trên tập huấn luyện:")
print(f"R² Score: {r2_train_knn}")
print(f"MSE: {mse_train_knn}")
print(f"RMSE: {rmse_train_knn}")
print(f"MAE: {mae_train_knn}")

```


 Copy code

figure 17: Đánh giá các chỉ số trên tập huấn luyện và kiểm tra


Giải thích các chỉ số đánh giá:

- R^2 Score (Hệ số xác định): Đo lường mức độ mô hình giải thích được biến thiên của dữ liệu mục tiêu. Giá trị gần 1 cho thấy mô hình phù hợp tốt.
- MSE (Mean Squared Error): Trung bình bình phương sai số giữa giá trị thực tế và giá trị dự đoán. Giá trị thấp cho thấy mô hình chính xác hơn.
- RMSE (Root Mean Squared Error): Căn bậc hai của MSE, biểu thị lỗi dưới dạng đơn vị gốc của dữ liệu.
- MAE (Mean Absolute Error): Trung bình sai số tuyệt đối giữa giá trị thực tế và giá trị dự đoán. Đơn giản và dễ hiểu hơn MSE.

Bước 4: Trực quan hóa kết quả mô hình KNN

Để hiểu rõ hơn về hiệu suất của mô hình, chúng ta sẽ trực quan hóa sự so sánh giữa giá trị thực tế và giá trị dự đoán.

python

 Copy code

```
import matplotlib.pyplot as plt
import seaborn as sns

# Trực quan hóa kết quả trên tập kiểm tra
plt.figure(figsize=(10, 6))
sns.scatterplot(x=y_test, y=y_test_pred_knn)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red', lw=2)
plt.xlabel('Giá trị thực tế')
plt.ylabel('Giá trị dự đoán')
plt.title('So sánh Giá trị Thực tế và Dự đoán - KNN Regression')
plt.show()
```

figure 18: Trực quan hóa kết quả mô hình KNN của giá trị thực tế và giá trị dự đoán

Giải thích:

- Scatter Plot: So sánh trực tiếp giữa giá trị thực tế (y_{test}) và giá trị dự đoán ($y_{\text{test_pred_knn}}$). Nếu các điểm nằm gần đường $y=x$ (đường đỏ), mô hình có hiệu suất tốt.
- Đường $y=x$: Đường tham chiếu để đánh giá mức độ phù hợp của mô hình. Các điểm nằm gần đường này chứng tỏ mô hình dự đoán chính xác.

5.2.3. Kết luận về Mô hình KNN

Mô hình KNN Regression đã được triển khai và đánh giá dựa trên các chỉ số RMSE và R^2 Score. Kết quả đánh giá trên tập kiểm tra cho thấy mức độ chính xác của mô hình trong việc dự đoán xác suất lũ lụt.

Phân tích:

- RMSE: Độ lớn của RMSE càng nhỏ, mô hình càng chính xác trong việc dự đoán.
- R^2 Score: Giá trị R^2 càng cao (gần 1), mô hình càng giải thích được nhiều biến thiên của dữ liệu mục tiêu.

Trong nghiên cứu này, KNN Regression đã thể hiện hiệu suất tốt, đặc biệt khi sử dụng trọng số dựa trên khoảng cách ($\text{weights}='distance'$). Tuy nhiên, để đạt được hiệu suất tối ưu hơn, có thể thử nghiệm với các giá trị KKK khác nhau và các phương pháp chuẩn hóa dữ liệu khác nhau.

5.3. Huấn luyện và kiểm tra (Training and Testing)

5.3.1. Ngăn chặn sự quá trùng khớp (Overfitting)

Để tránh tình trạng mô hình học quá mức từ dữ liệu huấn luyện và không thể tổng quát hóa tốt trên dữ liệu kiểm tra, tôi áp dụng các kỹ thuật sau:

Early Stopping: Dừng quá trình huấn luyện khi hiệu suất trên tập kiểm tra không còn cải thiện.

Regularization: Sử dụng các thuật toán như Ridge Regression và Dropout trong LSTM để giảm độ phức tạp của mô hình.

5.3.2. Các chỉ số để đánh giá mô hình

Trong nghiên cứu này, 6 chỉ số chính được sử dụng để đánh giá hiệu suất của các mô hình, kèm theo công thức chi tiết:

1. R^2 (Hệ số xác định)

R^2 đo lường mức độ mô hình giải thích được biến thiên của dữ liệu mục tiêu.

Công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Trong đó:

- (Y_i) : Giá trị thực tại điểm dữ liệu (i) .
- (\hat{Y}_i) : Giá trị dự đoán tại điểm dữ liệu (i) .
- (\bar{Y}) : Giá trị trung bình của tất cả giá trị thực (\bar{Y}) .
- (n) : Số lượng điểm dữ liệu.

Ý nghĩa:

- (R^2) càng gần 1, mô hình càng giải thích tốt sự biến thiên của dữ liệu.

2. MSE (Mean Squared Error)

MSE đo lường mức độ sai lệch giữa giá trị thực tế và giá trị dự đoán, tính bằng bình phương độ lệch trung bình.

Công thức:

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2$$

Trong đó:

- (Y_i) : Giá trị thực tại điểm dữ liệu i .
- (\hat{Y}_i) : Giá trị dự đoán tại điểm dữ liệu i .
- (n) : Số lượng điểm dữ liệu.

Ý nghĩa:

- MSE càng nhỏ, mô hình càng chính xác.

3. RMSE (Root Mean Squared Error)

RMSE là căn bậc hai của MSE, biểu diễn lỗi dưới dạng đơn vị gốc của giá trị dự đoán.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2}$$

- (Y_i) , (\hat{Y}_i) , (n) : Như định nghĩa trong MSE.

Ý nghĩa:

- RMSE dễ hiểu hơn MSE vì có cùng đơn vị với giá trị dự đoán.

4. MAE (Mean Absolute Error)

MAE đo lường mức độ sai lệch giữa giá trị thực tế và giá trị dự đoán bằng cách lấy trung bình giá trị tuyệt đối của sai lệch.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

Trong đó:

- (Y_i) : Giá trị thực tại điểm dữ liệu i .
- (\hat{Y}_i) : Giá trị dự đoán tại điểm dữ liệu i .
- (n) : Số lượng điểm dữ liệu.

Ý nghĩa:

- MAE dễ tính và trực quan, phù hợp khi cần đánh giá độ chính xác tổng thể của mô hình.

5. MAPE (Mean Absolute Percentage Error)

MAPE đo lường phần trăm sai số tuyệt đối giữa giá trị thực tế và giá trị dự đoán.

Công thức:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100$$

Ý nghĩa:

- MAPE biểu diễn sai số theo tỷ lệ phần trăm, phù hợp để so sánh hiệu suất trên dữ liệu có quy mô khác nhau.

Bảng thống kê các chỉ số quan trọng:

Chỉ Số	Mô Tả	Tầm Quan Trọng
R^2 (Hệ số xác định)	Xác định phần trăm biến động dữ liệu thực tế được mô hình giải thích.	Giá trị cao (gần 1) chứng tỏ mô hình phù hợp, thể hiện khả năng giải thích dữ liệu hiệu quả.
MSE (Sai số bình phương trung bình)	Trung bình của bình phương chênh lệch giữa giá trị thực tế và dự đoán.	Thấp nghĩa là mô hình dự đoán chính xác, nhưng nhạy cảm với ngoại lệ.
RMSE (Sai số căn bình phương trung bình)	Căn bậc hai của MSE, thể hiện sai số dưới dạng đơn vị gốc.	Giúp hiểu rõ mức độ sai số theo cùng đơn vị đo lường của dữ liệu thực tế.
MAE (Sai số tuyệt đối trung bình)	Trung bình chênh lệch tuyệt đối giữa giá trị thực tế và dự đoán.	Đơn giản, dễ hiểu, ít bị ảnh hưởng bởi ngoại lệ hơn so với MSE.
MAPE (Sai số tuyệt đối trung bình theo phần trăm)	Sai số trung bình được biểu diễn dưới dạng phần trăm.	Phù hợp khi cần so sánh độ chính xác giữa các tập dữ liệu có quy mô khác nhau.

CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ VÀ KẾT LUẬN

Chương này tập trung vào việc đánh giá hiệu suất của các mô hình dự đoán đã triển khai, phân tích kết quả thu được và khám phá các ứng dụng thực tế của mô hình trong lĩnh vực dự báo xác suất lũ lụt. Qua đó, nghiên cứu rút ra những bài học quan trọng và đề xuất các hướng cải tiến nhằm tăng cường khả năng ứng dụng của các mô hình vào thực tiễn.

6.1. So sánh hiệu suất các mô hình

Bảng so sánh hiệu suất các mô hình:

Mô hình	R^2 (Train)	R^2 (Test)	RMSE (Train)	RMSE (Test)	MAE (Test)	MAPE (Test)	Time Training (s)
Linear Regression	0.8449	0.8451	0.0201	0.0201	0.0158	3.19%	60.00
Ridge Regression	0.8449	0.8451	0.0201	0.0201	0.0158	3.19%	60.00
Lasso Regression	0.0000	-0.0000	0.0510	0.0510	0.0409	8.26%	60.00
KNN	0.9999	0.6100	0.0000	0.0318	-	-	1800.00

Nhận xét:

- **Linear Regression và Ridge Regression:**
 - Cả hai mô hình đều cho thấy hiệu suất tốt với R^2 (Test) ≈ 0.8451 , cho thấy khả năng giải thích dữ liệu khá tốt.

- **RMSE và MAE** thấp trên cả tập huấn luyện và kiểm tra, chứng tỏ các mô hình này phù hợp để dự đoán xác suất lũ lụt.
- Ridge Regression xử lý tốt vấn đề đa cộng tuyến, giúp ổn định kết quả so với Linear Regression.
- **Lasso Regression:**
 - Hiệu suất thấp nhất trong các mô hình, với $R^2 \approx 0.0000$ và **RMSE ≈ 0.0510** .
 - Lasso Regression dường như không phù hợp với tập dữ liệu này, có thể do việc loại bỏ hoàn toàn các đặc trưng quan trọng trong quá trình chọn lọc.
- **KNN:**
 - KNN có hiệu suất cực kỳ cao trên tập huấn luyện ($R^2 \approx 0.9999$) nhưng giảm đáng kể trên tập kiểm tra ($R^2 \approx 0.6100$), cho thấy hiện tượng overfitting.
 - **Thời gian huấn luyện (1800 giây)** dài hơn rất nhiều so với các mô hình tuyến tính, đặc biệt khi xử lý dữ liệu lớn.

6.2. Phân tích kết quả

- Các mô hình tuyến tính (Linear Regression, Ridge Regression) cho thấy sự ổn định và hiệu suất đồng nhất trên cả tập huấn luyện và kiểm tra. Ridge Regression nổi bật nhờ khả năng giảm thiểu tác động của đa cộng tuyến, trong khi Linear Regression đạt được hiệu quả tương đương với thời gian huấn luyện ngắn.
- Lasso Regression không phù hợp trong bài toán này do hiệu suất thấp và khả năng loại bỏ sai các đặc trưng quan trọng, dẫn đến giảm hiệu quả dự đoán.
- KNN có tiềm năng mạnh mẽ trong việc dự đoán dữ liệu phi tuyến nhưng gặp vấn đề overfitting và phụ thuộc nhiều vào việc chuẩn hóa dữ liệu cũng như lựa chọn số lượng láng giềng (k). Thời gian huấn luyện dài cũng là hạn chế lớn trong thực tế triển khai.

6.3. Ứng dụng thực tế

Xây dựng hệ thống cảnh báo sớm: Mô hình Linear và Ridge Regression có thể được sử dụng để thiết lập các cảnh báo sớm về nguy cơ lũ lụt, cung cấp thông tin nhanh và đáng tin cậy cho các cơ quan quản lý và cộng đồng.

Hỗ trợ quy hoạch và quản lý tài nguyên: Các kết quả từ mô hình Ridge Regression và Lasso Regression có thể giúp xác định các yếu tố quan trọng nhất ảnh hưởng đến nguy cơ lũ lụt, từ đó hỗ trợ ra quyết định trong việc cải thiện hệ thống đê điều, nâng cấp hạ tầng hoặc quản lý khu vực dễ ngập lụt.

Ứng dụng KNN: Với khả năng khai thác dữ liệu phi tuyến, KNN có thể được sử dụng trong các bài toán yêu cầu dự đoán phức tạp hơn, ví dụ như mô hình hóa tác động của mưa lớn kết hợp với chất lượng đất và đê điều.

6.4. Đề xuất cải tiến

Kết hợp mô hình: Sử dụng kết hợp các mô hình tuyến tính (Ridge, Linear) với mô hình phi tuyến (KNN) để tận dụng tối đa ưu điểm của cả hai nhóm.

Tối ưu hóa tham số: Áp dụng các kỹ thuật tối ưu hóa tham số như Grid Search hoặc Random Search để tìm giá trị tốt nhất cho các siêu tham số như số lượng láng giềng (k) trong KNN.

Mở rộng và cải thiện dữ liệu:

- Thu thập thêm dữ liệu từ các khu vực khác hoặc trong khoảng thời gian dài hơn để tăng khả năng học của mô hình.
- Tích hợp dữ liệu thời gian thực như hình ảnh vệ tinh hoặc cảm biến để cải thiện độ chính xác của các đặc trưng.

Hạn chế overfitting: Sử dụng kỹ thuật giảm overfitting như giảm kích thước dữ liệu huấn luyện KNN hoặc kết hợp Regularization trong các mô hình phi tuyến.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết quả đạt được

Nghiên cứu đã thành công áp dụng các mô hình học máy như Linear Regression, Ridge Regression, Lasso Regression và KNN để dự đoán xác suất lũ lụt. Ridge Regression chứng tỏ hiệu quả vượt trội với độ ổn định cao trên cả tập huấn luyện và kiểm tra, trong khi Linear Regression cũng đạt kết quả tốt trên dữ liệu có đặc trưng tuyến tính rõ ràng. Lasso Regression nổi bật với khả năng lựa chọn đặc trưng hiệu quả nhưng hiệu suất tổng thể thấp hơn Ridge Regression. KNN cho thấy khả năng xử lý tốt dữ liệu phi tuyến nhưng lại yêu cầu thời gian tính toán dài và dễ bị overfitting. Những kết quả này khẳng định tiềm năng ứng dụng của các mô hình học

máy trong việc xây dựng hệ thống dự báo lũ lụt, hỗ trợ quản lý thiên tai và phát triển hạ tầng.

2. Điểm mạnh và điểm yếu của từng mô hình

Linear Regression đơn giản, dễ triển khai nhưng không hiệu quả với dữ liệu phi tuyến. Ridge Regression xử lý tốt hiện tượng đa cộng tuyến và overfitting, tuy nhiên không thể tự động loại bỏ các đặc trưng không quan trọng. Lasso Regression hỗ trợ tốt việc lựa chọn đặc trưng nhưng hiệu suất thấp hơn Ridge Regression. KNN phù hợp với dữ liệu phi tuyến và có độ chính xác cao trên tập huấn luyện nhưng đòi hỏi thời gian tính toán dài và dễ bị ảnh hưởng bởi nhiễu.

3. Khả năng ứng dụng kết quả nghiên cứu trong thực tiễn

Nghiên cứu mở ra nhiều khả năng ứng dụng thực tế trong quản lý thiên tai, quy hoạch hạ tầng, và hỗ trợ cộng đồng. Các cơ quan quản lý có thể sử dụng Ridge Regression để dự đoán xác suất lũ lụt, từ đó đưa ra cảnh báo sớm và lập kế hoạch ứng phó. Lasso Regression cung cấp thông tin quan trọng về các yếu tố ảnh hưởng đến nguy cơ lũ lụt, hỗ trợ ưu tiên đầu tư vào các hệ thống thoát nước hoặc đập nước. KNN có thể được sử dụng để phân tích các tình huống phi tuyến phức tạp, giúp nâng cao năng lực phân tích cho các cơ quan quản lý và tổ chức phi chính phủ. Các mô hình dự báo cũng giúp cộng đồng địa phương lập kế hoạch bảo vệ tài sản, giảm thiểu thiệt hại trong mùa mưa lũ.

4. Hạn chế của nghiên cứu

Nghiên cứu vẫn tồn tại một số hạn chế cần khắc phục. Dữ liệu sử dụng có giới hạn về độ bao phủ địa lý và thời gian, chưa phản ánh đầy đủ các yếu tố môi trường và xã hội như biến động dân cư hoặc sự kiện khí hậu đột biến. Thời gian tính toán dài của KNN cũng gây khó khăn cho việc triển khai hệ thống cảnh báo thời gian thực. Ngoài ra, việc tích hợp dữ liệu đa nguồn hoặc thời gian thực chưa được thực hiện triệt để.

5. Đề xuất cải tiến

Để cải thiện, cần mở rộng bộ dữ liệu từ nhiều khu vực và thời gian khác nhau, đồng thời tăng cường tích hợp các yếu tố môi trường và xã hội. Kết hợp các mô hình tuyến tính như Ridge hoặc Lasso với các mô hình phi tuyến như KNN hoặc các thuật toán học sâu sẽ tận dụng được ưu điểm của từng phương pháp. Thử nghiệm các mô hình tiên tiến như Transformer hoặc phương pháp học sâu để xử lý dữ liệu phức tạp cũng là một hướng đi hứa hẹn. Cuối cùng, việc phát triển các công cụ trực quan, dễ sử dụng sẽ giúp các cơ quan quản lý và cộng đồng địa phương dễ dàng tiếp cận và áp dụng kết quả nghiên cứu vào thực tiễn.

Nghiên cứu này không chỉ góp phần nâng cao khả năng dự đoán lũ lụt mà còn hỗ trợ công tác quản lý thiên tai, cải thiện quy hoạch hạ tầng, và giảm thiểu thiệt hại do thiên tai gây ra. Những đề xuất cải tiến hứa hẹn sẽ đưa nghiên cứu tiến xa hơn, đóng góp tích cực vào việc ứng dụng học máy trong lĩnh vực quản lý môi trường và phát triển bền vững.

TÀI LIỆU THAM KHẢO

- [1] "Mô hình mạng thần kinh nhân tạo và các bài toán ngập lụt: Nâng cao hiệu quả dự báo và nghiên cứu ảnh hưởng của công trình đô thị lên sự lan truyền." Available: <https://hoinghi45nam.imech.ac.vn>.
- [2] "Nghiên cứu ứng dụng trí tuệ nhân tạo trong dự báo lũ lụt," Viện Quy hoạch Thủy lợi. Available: <http://wri.vn>.
- [3] Đinh Nhật Quang, "Tổng quan ứng dụng phương pháp học máy trong dự báo lũ lụt," 2023. Available: <https://vawr.org.vn>.
- [4] "Flood prediction and mitigation strategies in Vietnam," Wiley Online Library. Available: <https://onlinelibrary.wiley.com>.
- [5] "FloodAdapt: Tools for adaptive flood risk management," DLR. Available: <https://floodadapt.eoc.dlr.de>.
- [6] Monica và Mandolaro, "Flood Data for Prediction and Analysis," MDPI. Available: <https://github.com/Mandolaro/flood-data>.
- [7] "Natural disaster data in Vietnam," Open Development Mekong. Available: <https://data.vietnam.opendevloppementmekong.net>.
- [8] Sneha Choudhary, "Flood Prediction AI Project for Kerala State," GitHub, 2023. Available: <https://github.com/choudharysneha1708/FloodAI>.
- [9] S. Sinha, "Vietnam Flood Prediction Using Machine Learning," GitHub. Available: <https://github.com/ssinha22/Vietnam-Flood-Prediction>.
- [10] Nikhil Desai, "Flood Risk Mapping in Danang City," GitHub. Available: https://github.com/NikhilSDesai/Flood_Risk_Mapping.
- [11] "V-FloodNet: Video Segmentation System for Urban Flood Detection," GitHub. Available: <https://github.com/xmlyqing00/V-FloodNet>.
- [12] K. Leok, "FloodPy: Python Toolbox for Flood Mapping," GitHub. Available: <https://github.com/kleok/FLOODPY>.

- [13] **Vũ Đức Mạnh**, "Weather Forecast Using LSTM-BiLSTM Models," **GitHub**. Available: <https://github.com/vuducmanh2008/Weather-Forecast-Using-LSTM-BiLSTM-Model>.
- [14] **VennDev**, "AI Weather Forecasting," **GitHub**. Available: <https://github.com/VennDev/AI-weather-forecasting>.
- [15] **Nguyễn Văn Tài**, "Datascience Applications for Flood Analysis," **GitHub**. Available: https://github.com/NguyenVTai/Datascience_2016-2.
- [16] "Phân tích dữ liệu thời tiết từ A-Z," **Tạp chí Khoa học Môi trường**, 2024. Available: <https://weatheranalysis.com/flood-data-analysis-guide>.
- [17] **Nguyễn Hữu Trí**, "Hybrid Models for Flood Risk Management," **Journal of Hydrological Studies**, vol. 12, no. 3, 2023. Available: <https://hydrologicalstudies.org>.
- [18] **FloodList**, "Latest updates on global flood events," 2023. Available: <https://floodlist.com>.
- [19] **Vietnam Meteorological and Hydrological Administration**, "Real-time flood monitoring data," 2024. Available: <https://nchmf.gov.vn>.
- [20] **European Space Agency**, "Satellite Observations for Flood Mapping," 2024. Available: <https://esa.int>.