|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



Nguyễn trung tuyến

Bùi trung quốc

xÂY dựng hệ thống đề xuất phim

ĐỒ ÁN Trí tuệ nhân tạo

HÀ NỘI, NĂM 2024

**NGUYỄN TRUNG TUYẾN - BÙI TRUNG QUỐC ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HÀ NỘI, NĂM 2024**

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI

NGUYỄN TRUNG TUYẾN

BÙI TRUNG QUỐC

**XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM**

|  |  |
| --- | --- |
| Ngành: | Trí tuệ nhân tạo và khoa học dữ liệu |
| Mã số: | ---------- |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | TS. Tạ Quang Chiểu |

HÀ NỘI, NĂM 2024

|  |  |
| --- | --- |
| logo | TRƯỜNG ĐẠI HỌC THUỶ LỢI  **KHOA CÔNG NGHỆ THÔNG TIN**  BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TTNT |

**TÊN ĐỀ TÀI:** **XÂY DỰNG HỆ THỐNG ĐỀ XUẤT PHIM**

|  |  |
| --- | --- |
| *Sinh viên thực hiện:* | Nguyễn Trung Tuyến  Bùi Trung Quốc |
| *Lớp:* | 63TTNT |
| *Giáo viên hướng dẫn:* | TS. Tạ Quang Chiểu |

**TÓM TẮT ĐỀ TÀI**

Hệ thống đề xuất phim dựa trên nội dung đóng vai trò quan trọng trong việc cung cấp trải nghiệm tốt hơn cho người dùng bằng cách gợi ý những bộ phim phù hợp với sở thích cá nhân. Với sự gia tăng nhanh chóng của các bộ phim và nội dung số, người dùng thường gặp khó khăn trong việc chọn lựa phim để xem. Các phương pháp truyền thống như dựa trên đánh giá của người dùng hoặc đề xuất ngẫu nhiên không đáp ứng được nhu cầu cá nhân hóa cao của người dùng.

Ngày nay, với sự phát triển của công nghệ thông tin, học máy và xử lý ngôn ngữ tự nhiên, nhiều phương pháp mới đã được áp dụng để giải quyết vấn đề này. Một trong những phương pháp hiệu quả là sử dụng TF-IDF (Term Frequency-Inverse Document Frequency) để xây dựng bộ hồ sơ cho từng phim, giúp phân tích và gợi ý phim dựa trên nội dung.

Trong đồ án môn học này, chúng em sẽ nghiên cứu và triển khai hệ thống đề xuất phim dựa trên nội dung.

**CÁC MỤC TIÊU CHÍNH**

Mục tiêu 1: Hoàn thành báo cáo tổng kết của đề tài

Mục tiêu 2: Thu thập được các dữ liệu về phim cần sử dụng cho đề tài

Mục tiêu 3: Xây dựng được hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác dựa trên nội dung sử dụng TF-IDF.

**KẾT QUẢ DỰ KIẾN**

Xây dựng được hệ thống đề xuất phim hoàn chỉnh và chạy được trên giao diện website.

MỤC LỤC

[MỞ ĐẦU 5](#_Toc123524127)

[CHƯƠNG 1 GIỚI THIỆU TỔNG QUAN 7](#_Toc123524128)

[1.1 Lý do chọn đề tài 7](#_Toc123524129)

[1.2 Mục tiêu đồ án 7](#_Toc123524130)

[1.2.1 Mục tiêu tổng quát 7](#_Toc123524131)

[1.2.2 Mục tiêu cụ thể 8](#_Toc123524132)

[1.2.3 Nhiệm vụ nghiên cứu 8](#_Toc123524133)

[1.3 Đối tượng và phạm vi nghiên cứu 8](#_Toc123524134)

[1.4 Phương pháp nghiên cứu 9](#_Toc123524135)

[CHƯƠNG 2 TÌM HIỂU MỘT SỐ MÔ HÌNH HỌC MÁY 10](#_Toc123524136)

[2.1 Tổng quan về Machine Learning 10](#_Toc123524137)

[2.1.1 Tổng quan 10](#_Toc123524138)

[2.2 Phương pháp hồi quy tuyến tính 11](#_Toc123524139)

[2.2.1 Mô hình hồi quy tuyến tính 11](#_Toc123524140)

[2.2.2 Xây dựng và tối ưu hàm mất mát 12](#_Toc123524141)

[2.3 Tổng quan về Deep Learning 14](#_Toc123524142)

[2.3.1 Deep Learning là gì? 14](#_Toc123524143)

[2.3.2 Cách thức hoạt động của Deep Learning 15](#_Toc123524144)

[2.4 Mạng ANN 16](#_Toc123524145)

[2.5 Mạng RNN 18](#_Toc123524146)

[2.6 Phương pháp Long Short-Term Memory (LSTM) 19](#_Toc123524147)

[2.6.1 Các ô nhớ có cổng 20](#_Toc123524148)

[2.6.2 Ô nhớ tiềm năng 22](#_Toc123524149)

[2.6.3 Ô nhớ 22](#_Toc123524150)

[2.6.4 Các trạng thái ẩn 23](#_Toc123524151)

[CHƯƠNG 3 XÂY DỰNG MÔ HÌNH DỰ BÁO MỰC NƯỚC TRÊN SÔNG NHẬT LỆ 24](#_Toc123524152)

[3.1 Các công cụ, thư viện dùng để xây dựng mô hình học máy 24](#_Toc123524153)

[3.1.1 Ngôn ngữ lập trình Python 24](#_Toc123524154)

[3.1.2 Các thư viện sử dụng 26](#_Toc123524155)

[CHƯƠNG 4 ĐÁNH GIÁ KẾT QUẢ MÔ HÌNH 27](#_Toc123524156)

[4.1 Chuẩn bị bộ dữ liệu 27](#_Toc123524157)

[4.2 Phương pháp đánh giá 29](#_Toc123524158)

[4.3 Bài toán dự đoán mực nước ở 3 trạm Lệ Thủy, Đồng Hới, Kiến Giang 32](#_Toc123524159)

[4.3.1 Mô hình hồi quy tuyến tính 33](#_Toc123524160)

[4.3.2 Mô hình LSTM 43](#_Toc123524161)

[4.3.3 Đánh giá kết quả thực nghiệm 52](#_Toc123524162)

[CHƯƠNG 5 KẾT LUẬN 54](#_Toc123524163)

[5.1 Kết quả đạt được 54](#_Toc123524164)

[5.2 Hướng phát triển 54](#_Toc123524165)

[Tài liệu tham khảo 55](#_Toc123524166)

DANH MỤC HÌNH ẢNH

[Hình 1.1 Địa hình sông suối tỉnh Quảng Bình 10](file:///D:\\Doan\\BaocaoDATN_VAT.docx" \l "_Toc123524167)

[Hình 2.1 Mô hình mạng ANN 17](#_Toc123524168)

[Hình 2.2 Mô hình mạng RNN 19](file:///D:\\Doan\\BaocaoDATN_VAT.docx" \l "_Toc123524169)

[Hình 2.3 Cấu trúc của một module LSTM 20](file:///D:\\Doan\\BaocaoDATN_VAT.docx" \l "_Toc123524170)

[Hình 2.4 Các phép tính tại cổng đầu vào, cổng quên và cổng đầu ra trong một đơn vị LSTM. 22](#_Toc123524171)

[Hình 2.5 Các phép tính toán trong ô nhớ tiềm năng của LSTM. 23](#_Toc123524172)

[Hình 2.6 : Các phép tính toán trong ô nhớ của LSTM. Ở đây, ta sử dụng phép nhân theo từng phần 24](#_Toc123524173)

[Hình 2.7 Các phép tính của trạng thái ẩn. Phép tính nhân được thực hiện trên từng phần tử. 24](#_Toc123524174)

[Hình 4.1 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 1 - LR 35](#_Toc123524175)

[Hình 4.2 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR 35](#_Toc123524176)

[Hình 4.3 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 2 - LR 36](#_Toc123524177)

[Hình 4.4 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 7 (k=7) - LR 36](#_Toc123524178)

[Hình 4.5 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 1 – Mô hình hồi quy tuyền tính 37](#_Toc123524179)

[Hình 4.6 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 1 - LR 38](#_Toc123524180)

[Hình 4.7 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR 39](#_Toc123524181)

[Hình 4.8 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 2 - LR 40](#_Toc123524182)

[Hình 4.9 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 1 (k=1) - LR 40](#_Toc123524183)

[Hình 4.10 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình hồi quy tuyền tính 41](#_Toc123524184)

[Hình 4.11 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Đồng Hới sử dụng bộ dữ liệu trường hợp 1 - LR 42](#_Toc123524185)

[Hình 4.12 Kết quả tốt nhất ở trạm Đồng Hới so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR 42](#_Toc123524186)

[Hình 4.13 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Đồng Hới sử dụng bộ dữ liệu trường hợp 2 - LR 43](#_Toc123524187)

[Hình 4.14 Kết quả tốt nhất ở trạm Đồng Hới so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 9 (k=9) - LR 43](#_Toc123524188)

[Hình 4.15 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình hồi quy tuyền tính 44](#_Toc123524189)

[Hình 4.16 Biểu đồ so sánh các thông số trong 7 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 1 - LSTM 45](#_Toc123524190)

[Hình 4.17 Biểu đồ so sánh các thông số trong 6 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 2 - LSTM 46](#_Toc123524191)

[Hình 4.18 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình LSTM 47](#_Toc123524192)

[Hình 4.19 Biểu đồ so sánh các thông số trong 7 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 1 - LSTM 48](#_Toc123524193)

[Hình 4.20 Biểu đồ so sánh các thông số trong 6 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 2 - LSTM 49](#_Toc123524194)

[Hình 4.21 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình LSTM 52](#_Toc123524195)

[Hình 4.22 Biểu đồ so sánh kết quả thực nghiệm tốt nhất giữa 2 trường hợp trong mô hình hồi quy tuyến tính 53](#_Toc123524196)

[Hình 4.23 Biểu đồ so sánh kết quả thực nghiệm nhất giữa 2 trường hợp trong mô hình hồi LSTM 54](#_Toc123524197)

[Hình 4.24 Biểu đồ so sánh trường hợp có kết quả thực nghiệm tốt nhất của hai mô hình hồi quy tuyến tính và mô hình LSTM 54](#_Toc123524198)

**DANH MỤC BẢNG BIỂU**

[Bảng 4.1 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1 37](#_Toc123528286)

[Bảng 4.2 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1 38](#_Toc123528287)

[Bảng 4.3 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Lệ Thủy sử dụng mô hình Hồi quy tuyến tính 39](#_Toc123528288)

[Bảng 4.4 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1 40](#_Toc123528289)

[Bảng 4.5 : Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1 42](#_Toc123528290)

[Bảng 4.6 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Kiến Giang sử dụng mô hình Hồi quy tuyến tính 43](#_Toc123528291)

[Bảng 4.7 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. 44](#_Toc123528292)

[Bảng 4.8 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. 45](#_Toc123528293)

[Bảng 4.9 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Đồng Hới sử dụng mô hình Hồi quy tuyến tính 46](#_Toc123528294)

[Bảng 4.10 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1. Với Epochs = 350 47](#_Toc123528295)

[Bảng 4.11 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1. Với Epochs = 450 48](#_Toc123528296)

[Bảng 4.12 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Lệ Thủy sử dụng mô hình LSTM 49](#_Toc123528297)

[Bảng 4.13 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1. Với Epochs = 350 50](#_Toc123528298)

[Bảng 4.14 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1. Với Epochs = 450 51](#_Toc123528299)

[Bảng 4.15 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. Với Epochs = 350 52](#_Toc123528300)

[Bảng 4.16 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. Với Epochs = 450 53](#_Toc123528301)

[Bảng 4.17 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Đồng Hới sử dụng mô hình LSTM 54](#_Toc123528302)

[Bảng 4.18 So sánh kết quả thực nghiệm tốt nhất ở các trạm trên tất cả các trường hợp 55](#_Toc123528303)

**DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
|  | Từ viết tắt | Ý nghĩa |
| 1 | ĐATN | Đồ án tốt nghiệm |
| 2 | LR | Linear Regression (Mô hình hồi quy tuyến tính) |
| 3 | LSTM | Long Short-Term Memory (Mô hình bộ nhớ dài ngắn hạn) |
| 4 | ML | Machine Learning |
| 5 | AI |  |
| 6 | ANN |  |
| 7 | RNN |  |

# MỞ ĐẦU

Trí tuệ nhân tạo (Artificial Intelligence - AI) ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau của đời sống. AI có thể được áp dụng trong ngành Thủy lợi để dự báo mực nước hoặc lưu lượng dòng chảy trong sông nhằm cảnh báo lũ, góp phần giảm nhẹ và phòng ngừa tác hại của lũ lụt [3]. Việc dự báo sớm mực nước sông giúp dự đoán cấp độ lũ và hỗ trợ con người người trong việc xây dựng phương án phòng chống lũ, hạn chế các tác hại do lũ gây ra. Đặc biệt trong trận lũ từ ngày 16 đến 22 tháng 10 năm 2020, đã gây ngập lụt trên diện rộng và gây thiệt hại nghiêm trọng, nhất là địa bàn huyện Lệ Thủy, Quảng Ninh và các địa phương lưu vực sông Gianh. Để giảm thiểu tối đa những thiệt hại do thiên tai gây ra con người chúng ta đã đẩy mạnh đầu tư phát triển mạng lưới quan trắc hiện đại, đầu tư, cải tiến các phương pháp dự báo, mô hình dự báo... để kịp thời ứng phó với biến đổi thời tiết thất thường. Tuy nhiên, để dự đoán đúng, trúng thiên tai xảy ra theo thời gian thực là bài toán khó, nhất là trong bối cảnh công nghệ mới và biến đổi khí hậu.

Từ trước đến nay, các mô hình thủy lực được áp dụng rộng rãi ở Việt Nam cũng như trên thế giới để tính toán dòng chảy trong sông vùng chịu ảnh hưởng của thủy triều, phục vụ cho công tác quy hoạch phòng chống lũ lụt hay hạn hán, xâm nhập mặn. Tuy nhiên, các mô hình này thường yêu cầu một lượng lớn các dữ liệu đầu vào như địa hình, lượng mưa, các lưu lượng chảy qua cống lấy nước hay thoát nước... Các thông số của mô hình phải được hiệu chỉnh và kiểm định dựa trên nhiều tài liệu thực đo, do đó không dễ dàng lựa chọn các thông số phù hợp. Trong trường hợp không có đủ tài liệu địa hình, địa mạo và tùy theo nhu cầu thực tế, mô hình mạng nơ-ron nhân tạo (ANN) có thể được áp dụng để dự báo các yếu tố thủy văn, thủy lực của dòng chảy. Các mô hình ANN dựa trên mối quan hệ giữa các dữ liệu đầu vào và đầu ra để học các quy luật trong quá khứ từ đó đưa ra các dự báo trong tương lai. Trên thế giới, các mô hình ANN đã được sử dụng để dự báo lũ từ năm 1990. Sau đó, các nhà khoa học đã áp dụng và cải tiến các thuật toán trong mô hình ANN nhằm tăng độ chính xác trong dự báo (Lê Xuân Hiền, Hồ Việt Hùng, 2018) [3]. Asaad Y. Shamseldin (2010) đã khai thác sử dụng mạng nơ-ron nhân tạo để dự báo lưu lượng dòng chảy sông Nile xanh ở Sudan. Trong nghiên cứu này, tác giả đã sử dụng giá trị lưu lượng dòng chảy trung bình ngày của sông Nile xanh đo tại Eldeim gần biên giới Sudan - Ethiopia và dữ liệu lượng mưa trung bình ngày trong khu vực cho giai đoạn bốn năm 1992-1995 để dự báo. Chen, J.F. và nnk (2014) đã sử dụng thuật toán Cuckoo Search để dự báo lưu lượng dòng chảy đến hồ Hòa Bình, Việt Nam. Ngoài ra, có thể kể đến các nghiên cứu ở trong nước như: Trương Xuân Nam, Nguyễn Thanh Tùng (2016) đã sử dụng phương pháp Học sâu để dự báo lưu lượng nước đến hồ Hòa Bình. Trong nghiên cứu này, các tác giả đã sử dụng dữ liệu đầu vào là lưu lượng mùa kiệt từ năm 1964 đến 2002 với bước thời gian quan trắc 10 ngày và đưa ra dự báo lưu lượng dòng chảy cho 10 ngày sau đó. Lê Xuân Hiền, Hồ Việt Hùng (2018) đã ứng dụng mô hình bộ nhớ gần xa, Long Short-Term Memory (LSTM) để dự báo mực nước tại các trạm thủy văn ở Hải Phòng, khu vực bị ảnh hưởng của thủy triều. Các tác giả đã sử dụng dữ liệu là mực nước giờ để dự báo mực nước trong tương lai từ 1 giờ đến 5 giờ, tại trạm Quang Phục và Cửa Cấm [3]. Hồ Việt Hùng và nnk (2018) đã xây dựng một mô hình toán dựa trên mạng nơ-ron hồi quy (Recurrent Neural Network - RNN), một dạng ứng dụng chuyên sâu của mô hình ANN, để dự báo lưu lượng lũ sông Đà tại Lai Châu trước 1 ngày và dự báo lưu lượng dòng chảy sông Hồng tại trạm Sơn Tây trước 1 ngày, 2 ngày và 3 ngày. Mô hình mà các tác giả đề xuất không đòi hỏi các dữ liệu về địa hình và thảm phủ bề mặt lưu vực sông, chỉ cần các dữ liệu về lưu lượng dòng chảy. Lê Xuân Hiền và nnk (2018) đã sử dụng mô hình LSTM để dự báo lưu lượng nước về hồ Hòa Bình trước 6 giờ, 12 giờ và 18 giờ. Những nghiên cứu kể trên cho thấy rằng, các mô hình dự báo dòng chảy và mực nước dựa trên ANN ngày càng phổ biến và có thể ứng dụng rộng rãi trong ngành Thủy lợi. Dựa trên thuật toán Hồi quy tuyến tính và Bộ nhớ dài ngắn hạn để đề xuất trong đồ án này trình bày xây dựng chương trình ứng dụng minh họa. [3]

Bố cục của đồ án được trình bày trong 5 chương như sau:

**Chương 1:** Giới thiệu

**Chương 2:** Tìm hiểu một số mô hình học máy

**Chương 3:** Xây dựng mô hình dự báo mực nước trên sông Nhật Lệ

**Chương 4:** Đánh giá kết quả mô hình

**Chương 5:** Kết luận

# GIỚI THIỆU TỔNG QUAN

## Lý do chọn đề tài

Hệ thống đề xuất phim dựa trên nội dung đóng vai trò quan trọng trong việc cung cấp trải nghiệm tốt hơn cho người dùng bằng cách gợi ý những bộ phim phù hợp với sở thích cá nhân. Với sự gia tăng nhanh chóng của các bộ phim và nội dung số, người dùng thường gặp khó khăn trong việc chọn lựa phim để xem. Các phương pháp truyền thống như dựa trên đánh giá của người dùng hoặc đề xuất ngẫu nhiên không đáp ứng được nhu cầu cá nhân hóa cao của người dùng.

Ngày nay, với sự phát triển của công nghệ thông tin, học máy và xử lý ngôn ngữ tự nhiên, nhiều phương pháp mới đã được áp dụng để giải quyết vấn đề này. Một trong những phương pháp hiệu quả là sử dụng TF-IDF (Term Frequency-Inverse Document Frequency) để xây dựng bộ hồ sơ cho từng phim, giúp phân tích và gợi ý phim dựa trên nội dung.

Trong đồ án môn học này, chúng em sẽ nghiên cứu và triển khai hệ thống đề xuất phim dựa trên nội dung.

## Mục tiêu đồ án

### Mục tiêu tổng quát

Nghiên cứu và ứng dụng các kỹ thuật học máy trong việc xây dựng hệ thống đề xuất phim dựa trên nội dung nhằm cung cấp các gợi ý phim phù hợp với sở thích người dùng.

### Mục tiêu cụ thể

Mục tiêu 1: Hoàn thành báo cáo tổng kết của đề tài

Mục tiêu 2: Thu thập được các dữ liệu về phim cần sử dụng cho đề tài

Mục tiêu 3: Xây dựng được hệ thống đề xuất phim sử dụng phương pháp lọc cộng tác dựa trên nội dung sử dụng TF-IDF.

### Nhiệm vụ nghiên cứu

- Nghiên cứu các khái niệm và phương pháp liên quan đến TF-IDF và hệ thống đề xuất dựa trên nội dung.

- Thu thập và tiền xử lý dữ liệu để có thể sử dụng cho hệ thống đề xuất phim.

- Phát triển hệ thống đề xuất phim dựa trên nội dung sử dụng TF-IDF.

## Đối tượng và phạm vi nghiên cứu

## Nghiên cứu tập trung vào các bộ phim và dữ liệu mô tả nội dung của phim bao gồm các thông tin như mô tả cốt truyện, thể loại, tác giả,… Bộ dữ liệu được thu thập từ trang web đánh giá phim nổi tiếng là IMDB.

## Phương pháp nghiên cứu

Phương pháp lý thuyết: Tìm hiểu và nghiên cứu các mô hình TF-IDF và các thuật toán học máy liên quan đến hệ thống đề xuất dựa trên nội dung.

Phương pháp thực nghiệm: Xây dựng và thử nghiệm hệ thống đề xuất phim dựa trên nội dung sử dụng TF-IDF.

# TÌM HIỂU MỘT SỐ MÔ HÌNH HỌC MÁY

## Tổng quan về Machine Learning

# XÂY DỰNG MÔ HÌNH DỰ BÁO MỰC NƯỚC TRÊN SÔNG NHẬT LỆ

## Các công cụ, thư viện dùng để xây dựng mô hình học máy

### Ngôn ngữ lập trình Python

**Python** là một [ngôn ngữ lập trình](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh) bậc cao cho các mục đích lập trình đa năng, do [Guido](https://vi.wikipedia.org/w/index.php?title=Guido_van_Rossum&action=edit&redlink=1) [van Rossum](https://vi.wikipedia.org/w/index.php?title=Guido_van_Rossum&action=edit&redlink=1) tạo ra và lần đầu ra mắt vào năm 1991 [11]. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình và là ngôn ngữ lập trình dễ học; được dùng rộng rãi trong phát triển [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o). Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu. Vào tháng 7 năm 2018, van Rossum đã từ chức lãnh đạo trong cộng đồng ngôn ngữ Python sau 30 năm làm việc.

Python hoàn toàn [tạo kiểu động](https://vi.wikipedia.org/w/index.php?title=T%E1%BA%A1o_ki%E1%BB%83u_%C4%91%E1%BB%99ng&action=edit&redlink=1) và dùng cơ chế [cấp phát bộ nhớ tự động](https://vi.wikipedia.org/wiki/Qu%E1%BA%A3n_l%C3%BD_b%E1%BB%99_nh%E1%BB%9B); do vậy nó tương tự như [Perl](https://vi.wikipedia.org/wiki/Perl), [Ruby,](https://vi.wikipedia.org/wiki/Ruby_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)) [Scheme](https://vi.wikipedia.org/wiki/Scheme), [Smalltalk](https://vi.wikipedia.org/wiki/Smalltalk), và [Tcl](https://vi.wikipedia.org/wiki/Tcl). Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý.

Python là ngôn ngữ lập trình hướng đối tượng đơn giản, dễ học, mạnh mẽ, cấp cao.  
Python có cấu trúc cú pháp ít hơn các ngôn ngữ khác.

* **Python được thông dịch**: Python được trình thông dịch xử lý trong thời gian chạy. Bạn không cần phải biên dịch chương trình của mình trước khi thực hiện nó. Nó tương tự với PERL và PHP.
* **Python là tương tác (Interactive)**: Tại một dấu nhắc Python (command line) bạn có thể tương tác trực tiếp với trình thông dịch để viết chương trình Python.
* **Python là hướng đối tượng**: Python hỗ trợ kỹ thuật lập trình hướng đối tượng hoặc kỹ thuật lập trình đóng gói mã trong các đối tượng.
* **Python là ngôn ngữ của người mới bắt đầu**: Python là ngôn ngữ tuyệt vời cho các lập trình viên mới bắt đầu và hỗ trợ phát triển một loạt các ứng dụng từ xử lý văn bản đơn giản, lập trình web, cho đến lập trình game.

Đặc điểm:

* Là ngôn ngữ mã nguồn mở.
* Vừa hướng thủ tục, vừa hướng đối tượng
* Hỗ trợ module và hỗ trợ gói (package)
* Xử lý lỗi bằng ngoại lệ (exception)
* Kiểu dữ liệu động ở mức cao
* Có khả năng tương tác với các module viết bằng ngôn ngữ lập trình khác
* Có thể nhúng vào ứng dụng như một giao tiếp kịch bản (scripting interface) Python có 2 chế độ thực thi
* Chế độ thực thi: chỉ ra chương trình cần thực hiện
  + Trình dịch python sẽ nạp, dịch và chạy chương trình đó.
* Chế độ dòng lệnh: chạy từng lệnh một
  + Chế độ thực thi: “python abc.py” chạy file abc.py

Biên dịch mã python:

* Trường hợp cần thiết, mã python có thể được biên dịch, kết quả dịch là chương trình dạng bytecode cho máy ảo python
* Tương tự như trường hợp của ngôn ngữ java
* Mã lệnh dịch được lưu vào file với đuôi. pyc
* Việc biên dịch có nhiều lợi điểm, chẳng hạn như khi sử dụng câu lệnh import một thư viện nào đó, thì có thể sử dụng luôn mã pyc có sẵn thay vì phải dịch lại từ đầu

Trong Python:

* Biến không cần khai báo trước, không cần chỉ kiểu
* Dữ liệu chuỗi nằm trong cặp nháy đơn ('), nháy kép ("), hoặc ba dấu nháy (""") – nếu viết nhiều dòng
* Sử dụng chuỗi thoát (escape sequence) để khai báo các ký tự đặc biệt
* Sử dụng chuỗi “trần”: r"nội dung"
* Dùng dấu thăng (#) để viết dòng chú thích
* Dùng hàm print để in dữ liệu
* Dùng hàm input để nhập dữ liệu
* Có thể kết hợp với hàm chuyển đổi kiểu

Ngoài ra Python hỗ trợ kiểu số rất mạnh và nhiều loại phép tính phong phú

* Sử dụng if cho tất cả các nhu cầu rẽ nhánh
* Phép toán if cho phép viết lệnh một cách tự nhiên
* Vòng lặp while tương tự như các ngôn ngữ khác
* Ngoại trừ việc có thể có thêm khối else
* Vòng lặp for cho phép lần lượt thực hiện lặp với các giá trị nhận được từ một danh sách
* Sử dụng từ khóa def để định nghĩa một hàm, hàm có thể có các tham số mặc định

### Các thư viện sử dụng

Ngôn ngữ sử dụng là Python. Phiên bản 3.9.6

Tải python tại: <https://www.python.org/downloads/>chọn phiên bản cần cài đặt.

Ngôn ngữ python có hệ thống các gói rất phong phú, hỗ trợ nhiều lĩnh vực khác nhau, từ xây dựng ứng dụng, xử lý web, xử lý ảnh, xử lý text…

Sử dụng pip để tải các gói mới về từ internet.

Một số thư viện sử dụng trong xây dựng mô hình:

* **NumPy** (Numerical Python): là gói chuyên về xử lý dữ liệu số (nhiều chiều), gói cũng chứa các hàm số tuyến tính cơ bản, biến đổi fourier, sinh số ngẫu nhiên nâng cao, …
* **SciPy** (Scientific Python): dựa trên NumPy, cung cấp các công cụ mạnh cho khoa học và kỹ nghệ, chẳng hạn như biến đổi fourier rời rạc, đại số tuyến tính, tối ưu hóa và ma trận thưa.
* **Matplotlib:** chuyên sử dụng để vẽ biểu đồ, hỗ trợ rất nhiều loại biểu đồ khác nhau
* **Pandas:** chuyên sử dụng cho quản lý và tương tác với dữ liệu có cấu trúc, được sử dụng rộng rãi trong việc thu thập và tiền xử lý dữ liệu.
* **Scikit Learn:** chuyên về học máy, dựa trên NumPy, SciPy và matplotlib, thư viện này có sẵn nhiều công cụ hiệu quả cho việc học máy và thiết lập mô hình thống kê chẳng hạn như các thuật toán phân lớp, hồi quy, phân cụm, giảm chiều dữ liệu.
* **Seaborn:** dự trên matplotlib, cung cấp các công cụ diễn thị (visualization) dữ liệu thống kê đẹp và hiệu quả, mục tiêu của gói là sử dụng việc diễn thị như là trọng tâm của khám phá và hiểu dữ liệu
* **TensorFlow:** gói chuyên dùng cho học máy của Google, đặc biệt là các mạng thần kinh nhân tạo
* **Keras:** thư viện cấp cao chuyên về học máy, sử dụng Theano, TensorFlow hoặc CNTK làm phụ trợ.

# ĐÁNH GIÁ KẾT QUẢ MÔ HÌNH

## Chuẩn bị bộ dữ liệu

Lưu vực sông Nhật Lệ bao gồm 03 nhánh sông chính là Kiến Giang, Long Đại và Nhật Lệ với tổng diện tích của toàn lưu vực là 2.612 km2. Trên lưu vực nghiên cứu hiện chỉ có 03 trạm đo mực nước (Kiến Giang, Lệ Thủy và Đồng Hới) và 3 trạm đo mưa (Lệ Thủy, Kiến Giang và Đồng Hới) có dữ liệu đo đạc liên tục đến nay; trong đó số liệu theo giờ chỉ có trong một thời đoạn ngắn trong từng đợt lũ. Trong nghiên cứu này bộ số liệu quan trắc mưa và mực nước ngày tại 3 trạm trong giai đoạn 1977-2020 đã được thu thập. Do nghiên cứu chỉ tập trung vào bài toán dự báo mực nước ngày trên sông trong mùa lũ nên 5,368 bộ số liệu mưa, mực nước ngày và độ chênh lệch giữa mực nước và lượng mưa tại 3 trạm trong mùa lũ (tháng 9 - tháng 12) trong 44 năm đã được sử dụng. Để thiết lập và đánh giá các mô hình học máy, bộ dữ liệu trên được chia thành bộ dữ liệu huấn luyện và kiểm định, trong đó 80% dành cho huấn luyện và 20% còn lại dùng để kiểm định.

Bộ dữ liệu bao gồm các thuộc tính:

* **WL\_LeThuy:** Mực nước được ghi nhận trong ngày tại trạm Lệ Thủy tính bằng (m).
* **WL\_KienGiang:** Mực nước được ghi nhận trong ngày tại trạm Kiến Giang tính bằng (m).
* **WL\_DongHoi:** Mực nước được ghi nhận trong ngày tại trạm Đồng Hới tính bằng (m).
* **RF\_LeThuy:** Lượng mưa được ghi nhận trong ngày tại trạm Lệ Thủy tính bằng (m).
* **RF\_KienGiang:** Lượng mưa được ghi nhận trong ngày tại trạm Kiến Giang tính bằng (m).
* **RF\_DongHoi:** Lượng mưa được ghi nhận trong ngày tại trạm Đồng Hới tính bằng (m).
* **DWL\_LeThuy:** Mực nước chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Kiến Giang tính bằng (m).
* **DWL\_KienGiang:** Mực nước chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Kiến Giang tính bằng (m).
* **DWL\_DongHoi:** Mực nước chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Đồng Hới tính bằng (m).
* **DRF\_LeThuy:** Lượng mưa chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Lệ Thủy tính bằng (m).
* **DRF\_KienGiang:** Lượng mưa chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Kiến Giang tính bằng (m).
* **DRF\_DongHoi:** Lượng mưa chênh lệch được ghi nhận giữa ngày thứ t và ngày thứ t-1 tại trạm Đồng Hới tính bằng (m).

A picture containing table

Description automatically generated

Hình 4.1 Biểu đồ tương quan giữa các thuộc tính trong bộ dữ liệu

## Phương pháp đánh giá

**Mean Squared Error** [1]

Mean Squared Error (MSE) có lẽ là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Về cơ bản, nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính - nó luôn không âm và các giá trị càng gần 0 càng tốt.

(4-1)

trong đó n là số điểm dữ liệu, yᵢ là giá trị quan sát và ŷ ᵢ là giá trị dự đoán.

Trong phân tích hồi quy, vẽ biểu đồ là một cách tự nhiên hơn để xem xu hướng chung của toàn bộ dữ liệu. Đơn giản MSE cho bạn biết mức độ gần của đường hồi quy với một tập hợp các điểm. Nó thực hiện điều này bằng cách lấy khoảng cách từ các điểm đến đường hồi quy (những khoảng cách này là “sai số”) và bình phương chúng. Bình phương là rất quan trọng để giảm độ phức tạp với các dấu hiệu tiêu cực. Nó cũng tạo ra nhiều trọng lượng hơn cho sự khác biệt lớn hơn.

Để giảm thiểu MSE, mô hình có thể chính xác hơn, có nghĩa là mô hình gần với dữ liệu thực tế hơn. Một ví dụ về hồi quy tuyến tính sử dụng phương pháp này là - phương pháp bình phương nhỏ nhất đánh giá sự phù hợp của mô hình hồi quy tuyến tính với tập dữ liệu hai biến, nhưng giới hạn của nó liên quan đến phân phối dữ liệu đã biết.

MSE càng thấp thì dự báo càng tốt.

**Mean Absolute Error**

Mean Absolute Error (MAE) đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau.

(4-2)

trong đó n là số điểm dữ liệu, xᵢ là giá trị thực và yᵢ là giá trị dự đoán.

Có thể diễn đạt MAE là tổng hòa của hai thành phần: Bất đồng về số lượng và Bất đồng về phân bổ.

MAE được biết đến là mạnh mẽ hơn đối với các yếu tố ngoại lai so với MSE. Lý do chính là trong MSE bằng cách bình phương các sai số, các giá trị ngoại lai (thường có sai số cao hơn các mẫu khác) được chú ý nhiều hơn và chiếm ưu thế trong sai số cuối cùng và tác động đến các tham số của mô hình.

**Root Mean Square Error**

Root Mean Square Error (RMSE) [6] hoặc Root Mean Square Deviation (RMSD) là căn bậc hai của mức trung bình của các sai số bình phương. RMSE là độ lệch chuẩn của các phần dư (sai số dự đoán).

Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy; RMSE là thước đo mức độ dàn trải của những phần dư này, nói cách khác, nó cho bạn biết mức độ tập trung của dữ liệu xung quanh đường phù hợp nhất.

(4-3)

Ảnh hưởng của mỗi lỗi đối với RMSE tỷ lệ với kích thước của lỗi bình phương; do đó các sai số lớn hơn có ảnh hưởng lớn đến RMSE một cách không cân xứng. Do đó, RMSE nhạy cảm với các yếu tố ngoại lai. Sai số bình phương trung bình gốc thường được sử dụng trong khí hậu học, dự báo và phân tích hồi quy để xác minh kết quả thực nghiệm.

Khi các quan sát và dự báo chuẩn hóa được sử dụng làm đầu vào RMSE, có mối quan hệ trực tiếp với hệ số tương quan . Ví dụ, nếu hệ số tương quan là 1, RMSE sẽ bằng 0, bởi vì tất cả các điểm nằm trên đường hồi quy (và do đó không có sai số).

RMSE luôn không âm và giá trị 0 (hầu như không bao giờ đạt được trong thực tế) sẽ chỉ ra sự phù hợp hoàn hảo với dữ liệu. Nói chung, RMSE thấp hơn sẽ tốt hơn RMSE cao hơn.

**Nash-Sutcliffe-Efficiency (NSE)**

Nash-Sutcliffe-Efficiency (NSE) là một độ đo chuẩn hóa xác định độ lớn tương đối của phương sai dư ("nhiễu") so với phương sai dữ liệu đo được ("thông tin") (Nash và Sutcliffe, 1970).

NSE cho biết đồ thị của dữ liệu được quan sát so với dữ liệu được mô phỏng phù hợp với đường 1:1 như thế nào.

Hiệu suất Nash-Sutcliffe nằm trong khoảng từ -Inf đến 1. Về cơ bản, càng gần 1, mô hình càng chính xác.

-) NSE = 1, tương ứng với sự phù hợp hoàn hảo của mô hình hóa với dữ liệu được quan sát.

-) NSE = 0, cho biết các dự đoán của mô hình chính xác như giá trị trung bình của dữ liệu được quan sát,

-) -Inf < NSE < 0, chỉ ra rằng giá trị trung bình được quan sát là yếu tố dự báo tốt hơn mô hình.

(4-4)

**Coefficient of Determination Definition (R2)**

**R-bình phương (R2)** [6] là phần trăm mức độ dao động của biến đầu ra (response variable) có thể được giải thích bởi mô hình tuyến tính (linear model). **R2** luôn nằm trong khoảng 0 tới 100%. **R2** là một giá trị đo lường thống kê để đo mức độ gần giữa dữ liệu với đường hồi quy của mô hình tuyến tính. Nó còn được gọi là **hệ số xác định** (coefficient of determination) hay **hệ số xác định bội** (coefficient of multiple determination) trong mô hình hồi quy đa biến.

(4-5)

Một cách tổng quan, nếu giá trị **R2** càng lớn thì mô hình càng phù hợp với dữ liệu. Tuy nhiên, có một số điều kiện quan trọng chúng ta cần kiểm tra trước khi đưa ra kết luận mức độ tốt của mô hình dựa trên giá trị **R2**. Một trong những yếu tố cần kiểm tra là biểu đồ sai số (residual plot) để kiểm tra liệu có bất kì pattern nào hay không. Nếu biểu đồ này có pattern, thì mô hình của chúng ta đang có vấn đề và cần phải xem xét lại. [6]

## Bài toán dự đoán mực nước ở 3 trạm Lệ Thủy, Đồng Hới, Kiến Giang

Dự đoán mực nước ở 3 trạm Lệ Thủy, Kiến Giang và Đồng Hới trong mùa lũ (từ tháng 9 đến tháng 12) dựa vào lượng mưa và mực nước ở 3 trạm Lệ Thủy, Kiến Giang và Đồng Hới.

Bài toán 1: Biết thông tin về lượng mưa, mực nước và độ chênh lệch (lượng mưa, mực nước) của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1. So sánh trường hợp dùng độ chênh lệch với không dùng độ chênh lệch.

Bài toán 2: Biết thông tin về lượng mưa, mực nước và độ chênh lệch (lượng mưa, mực nước) của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1. So sánh trường hợp dùng độ chênh lệch với không dùng độ chênh lệch.

Bài toán 3: Biết thông tin về lượng mưa, mực nước và độ chênh lệch (lượng mưa, mực nước) của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. So sánh trường hợp dùng độ chênh lệch với không dùng độ chênh lệch.

**Đầu vào:** Dữ liệu 44 năm quan sát lượng mưa và mực nước trên 3 trạm: Kiến Giang, Lệ Thủy và Đồng Hới trên sông Nhật Lệ.

**Đầu ra:** Dự đoán mực nước ở ngày hôm sau.

### Mô hình hồi quy tuyến tính

**Đầu vào:** Cho tập dữ liệu huấn luyện gồm N mẫu. Mỗi mẫu là một cặp (xi,yi):

* xi: vector đặc trưng
* yi: giá trị của vector đặc trưng xi

**Đầu ra:** Hàm tuyến tính có dạng f(xi) = wxi

**Phương pháp:** Cần tìm hệ số w của hàm f(xi) sao cho trung bình sai số giữa yi và f(xi) là nhỏ nhất. Nghĩa là, tìm w để hàm số sau đạt giá trị nhỏ nhất.

Diagram, text, schematic

Description automatically generated

#### Kết quả thực nghiệm bài toán 1

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Lệ Thủy ngày thứ k+1

Kết quả thực nghiệm trường hợp 1 với bộ dữ liệu mực nước và lượng mưa ở các trạm

Bảng 4.1 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | Nash-Sutcliffe efficiency (NSE) | Coefficient of determination (R2) | Mean absolute error (MAE) | Root mean square error (RMSE) | MAX error | Time (sec) |
| 1 | 0.903223 | 0.904585 | 0.085972 | 0.183744 | 1.908615 | 0.046873 |
| 2 | 0.909891 | 0.91094 | 0.085458 | 0.177346 | 1.810657 | 0.031252 |
| 3 | 0.910878 | 0.911817 | 0.084917 | 0.176402 | 1.803174 | 0.031279 |
| 4 | 0.910879 | 0.911537 | 0.08545 | 0.176442 | 1.800205 | 0.046789 |
| 5 | 0.911015 | 0.911656 | 0.086098 | 0.176363 | 1.794749 | 0.046885 |
| 6 | 0.911102 | 0.911678 | 0.085921 | 0.176343 | 1.794366 | 0.062498 |
| 7 | 0.911387 | 0.911971 | 0.08617 | 0.176135 | 1.789947 | 0.078089 |
| 8 | 0.911361 | 0.91192 | 0.086467 | 0.176241 | 1.789196 | 0.078227 |
| 9 | 0.911187 | 0.911746 | 0.08687 | 0.176495 | 1.786895 | 0.093744 |
| 10 | 0.911606 | 0.912156 | 0.086429 | 0.176118 | 1.779317 | 0.093763 |

Hình 4.2 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 1 - LR

**Kết luận:** Như kết quả ở bảng 4.1 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 10 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Chart, line chart

Description automatically generated

Hình 4.3 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước và độ chệnh lệch lượng mưa, mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Lệ Thủy ngày thứ k+1.

Kết quả thực nghiệm trường hợp 2 với bộ dữ liệu mực nước và lượng mưa và bộ dự liệu chênh lệch mực nước, lượng mưa ở các trạm.

Bảng 4.2 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | (NSE) | (R2) | (MAE) | (RMSE) | MAX error | Time (sec) |
| 1 | 0.909828 | 0.910869 | 0.085515 | 0.177363 | 1.81097 | 0.032002 |
| 2 | 0.911064 | 0.912212 | 0.085324 | 0.176189 | 1.8063 | 0.063002 |
| 3 | 0.909287 | 0.909378 | 0.083961 | 0.177969 | 1.810812 | 0.076924 |
| 4 | 0.911284 | 0.91168 | 0.085222 | 0.176041 | 1.806946 | 0.083978 |
| 5 | 0.910338 | 0.911192 | 0.087628 | 0.177033 | 1.789969 | 0.100695 |
| 6 | 0.911197 | 0.911578 | 0.085896 | 0.176249 | 1.792987 | 0.109289 |
| 7 | 0.911629 | 0.912013 | 0.086139 | 0.175894 | 1.788821 | 0.125005 |
| 8 | 0.910704 | 0.911053 | 0.086811 | 0.176892 | 1.770324 | 0.140544 |
| 9 | 0.910872 | 0.911528 | 0.087249 | 0.176807 | 1.775057 | 0.179335 |
| 10 | 0.909785 | 0.910447 | 0.088495 | 0.177922 | 1.78694 | 0.171875 |

Hình 4.4 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 2 - LR

**Kết luận:** Như kết quả ở bảng 4.2 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 7 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Chart, line chart

Description automatically generated

Hình 4.5 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 7 (k=7) - LR

Bảng 4.3 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Lệ Thủy sử dụng mô hình Hồi quy tuyến tính

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** | **Time (sec)** |
| **TH1** | 0.911606 | 0.912156 | 0.086429 | 0.176118 | 1.779317 | 0.093763 |
| **TH2** | 0.911629 | 0.912013 | 0.086139 | 0.175894 | 1.788821 | 0.125005 |

Hình 4.6 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 1 – Mô hình hồi quy tuyền tính

#### Kết quả thực nghiệm bài toán 2

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Kiến Giang ngày thứ k+1

Kết quả thực nghiệm bài toán 2 với bộ dữ liệu mực nước và lượng mưa ở các trạm.

Bảng 4.4 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | Nash-Sutcliffe efficiency (NSE) | Coefficient of determination (R2) | Mean absolute error (MAE) | Root mean square error (RMSE) | MAX error | Time (sec) |
| 1 | 0.60101 | 0.601288 | 0.226712 | 0.451623 | 4.129331 | 0.062528 |
| 2 | 0.619489 | 0.620343 | 0.222292 | 0.440756 | 4.105189 | 0.046826 |
| 3 | 0.61924 | 0.619987 | 0.222467 | 0.440923 | 4.106446 | 0.03124 |
| 4 | 0.616668 | 0.61727 | 0.223417 | 0.442591 | 4.105933 | 0.046898 |
| 5 | 0.617517 | 0.618041 | 0.22537 | 0.442302 | 4.115047 | 0.062532 |
| 6 | 0.617393 | 0.61779 | 0.226213 | 0.442572 | 4.133277 | 0.078151 |
| 7 | 0.619386 | 0.619799 | 0.226869 | 0.44161 | 4.123722 | 0.093778 |
| 8 | 0.618643 | 0.618992 | 0.227199 | 0.442225 | 4.120523 | 0.093791 |
| 9 | 0.61806 | 0.618373 | 0.227775 | 0.442739 | 4.117787 | 0.109558 |
| 10 | 0.620111 | 0.620411 | 0.22633 | 0.441091 | 4.100738 | 0.109605 |

Hình 4.7 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 1 - LR

**Kết luận:** Như kết quả ở bảng 4.4 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 10 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Chart, line chart, histogram

Description automatically generated

Hình 4.8 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước và độ chệnh lệch lượng mưa, mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Kiến Giang ngày thứ k+1.

Kết quả thực nghiệm bài toán 2 với bộ dữ liệu mực nước và lượng mưa và bộ dự liệu chênh lệch mực nước, lượng mưa ở các trạm.

Bảng 4.5 : Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | Nash-Sutcliffe efficiency (NSE) | Coefficient of determination (R2) | Mean absolute error (MAE) | Root mean square error (RMSE) | MAX error | Time (sec) |
| 1 | 0.619751 | 0.62065 | 0.222593 | 0.440889 | 4.105465 | 0.031282 |
| 2 | 0.622446 | 0.623479 | 0.221982 | 0.43904 | 4.122192 | 0.04698 |
| 3 | 0.614862 | 0.61519 | 0.222798 | 0.443451 | 4.110269 | 0.078171 |
| 4 | 0.616836 | 0.617173 | 0.224392 | 0.442493 | 4.126152 | 0.078162 |
| 5 | 0.615511 | 0.615798 | 0.228761 | 0.44346 | 4.115722 | 0.093717 |
| 6 | 0.616798 | 0.617036 | 0.228214 | 0.442916 | 4.160373 | 0.109429 |
| 7 | 0.618882 | 0.619325 | 0.228757 | 0.441902 | 4.12707 | 0.125058 |
| 8 | 0.616116 | 0.616751 | 0.230052 | 0.443688 | 4.099053 | 0.171871 |
| 9 | 0.619906 | 0.62037 | 0.227199 | 0.441668 | 4.101483 | 0.171977 |
| 10 | 0.61938 | 0.619528 | 0.226574 | 0.441515 | 4.080248 | 0.187591 |

Hình 4.9 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 2 - LR

**Kết luận:** Như kết quả ở bảng 4.5 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 1 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Chart, line chart, histogram

Description automatically generated

Hình 4.10 Kết quả tốt nhất ở trạm Lệ Thủy so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 1 (k=1) - LR

Bảng 4.6 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Kiến Giang sử dụng mô hình Hồi quy tuyến tính

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** | **Time (sec)** |
| **TH1** | 0.620111 | 0.620411 | 0.22633 | 0.441091 | 4.100738 | 0.109605 |
| **TH2** | 0.619751 | 0.62065 | 0.222593 | 0.440889 | 4.105465 | 0.031282 |

Hình 4.11 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình hồi quy tuyền tính

#### Kết quả thực nghiệm bài toán 3

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Đồng Hới ngày thứ k+1

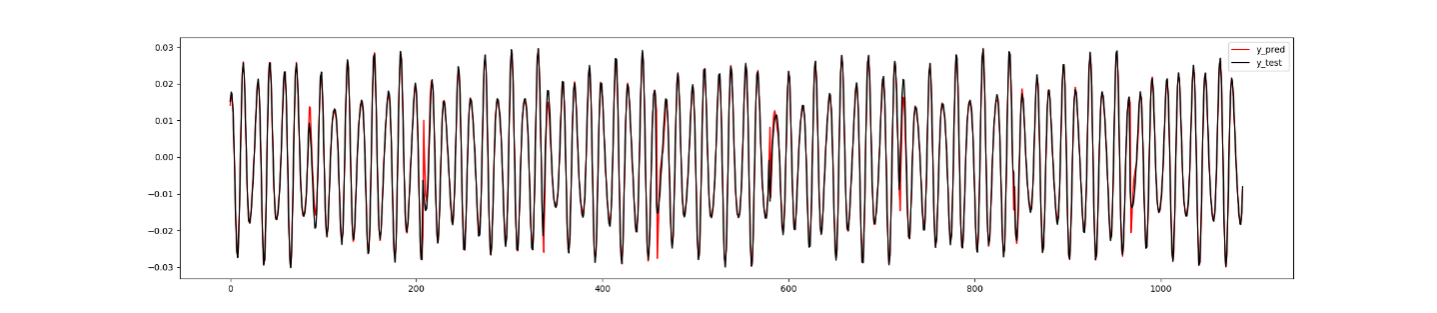
Kết quả thực nghiệm bài toán 3 với bộ dữ liệu mực nước và lượng mưa ở các trạm.

Bảng 4.7 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | Nash-Sutcliffe efficiency (NSE) | Coefficient of determination (R2) | Mean absolute error (MAE) | Root mean square error (RMSE) | MAX error | Time (sec) |
| 1 | 0.795227 | 0.795232 | 0.006157 | 0.007194 | 0.027465 | 0.015632 |
| 2 | 0.980771 | 0.980808 | 0.001126 | 0.002204 | 0.024918 | 0.031269 |
| 3 | 0.98105 | 0.981083 | 0.001276 | 0.002188 | 0.024512 | 0.06251 |
| 4 | 0.98107 | 0.981102 | 0.001271 | 0.002187 | 0.024521 | 0.046893 |
| 5 | 0.981075 | 0.981105 | 0.001269 | 0.002187 | 0.024568 | 0.046886 |
| 6 | 0.981691 | 0.981717 | 0.0012 | 0.002152 | 0.024928 | 0.062516 |
| 7 | 0.983109 | 0.983133 | 0.001046 | 0.002068 | 0.025436 | 0.078147 |
| 8 | 0.984393 | 0.984413 | 0.000889 | 0.001989 | 0.025815 | 0.068008 |
| 9 | 0.985028 | 0.985044 | 0.000806 | 0.001948 | 0.025922 | 0.093767 |
| 10 | 0.985088 | 0.985101 | 0.000794 | 0.001945 | 0.025921 | 0.1094 |

Hình 4.12 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Đồng Hới sử dụng bộ dữ liệu trường hợp 1 - LR

**Kết luận:** Như kết quả ở bảng 4.7 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 10 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.



Hình 4.13 Kết quả tốt nhất ở trạm Đồng Hới so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 10 (k=10) - LR

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước và độ chệnh lệch lượng mưa, mực nước của các trạm trong k ngày liên tiếp, dự đoán mực nước ở Lệ Thủy ngày thứ k+1.

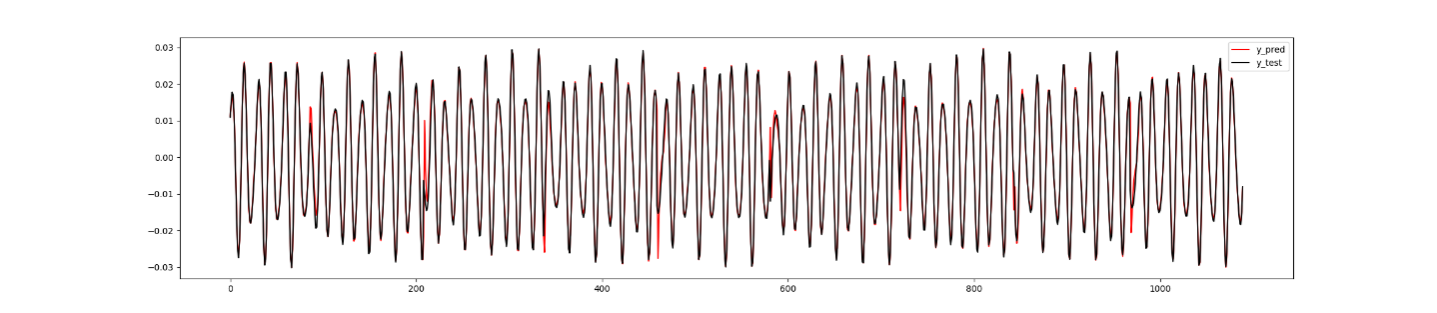
Kết quả thực nghiệm bài toán 3 với bộ dữ liệu mực nước và lượng mưa và bộ dự liệu chênh lệch mực nước, lượng mưa ở các trạm.

Bảng 4.8 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Numdays | Nash-Sutcliffe efficiency (NSE) | Coefficient of determination (R2) | Mean absolute error (MAE) | Root mean square error (RMSE) | MAX error | Time (sec) |
| 1 | 0.980578 | 0.980619 | 0.001132 | 0.002215 | 0.024926 | 0.031153 |
| 2 | 0.981056 | 0.98109 | 0.001277 | 0.002188 | 0.02451 | 0.062485 |
| 3 | 0.981025 | 0.981058 | 0.001274 | 0.00219 | 0.024518 | 0.062608 |
| 4 | 0.981029 | 0.981064 | 0.001272 | 0.002189 | 0.024572 | 0.093873 |
| 5 | 0.98165 | 0.981672 | 0.001204 | 0.002154 | 0.024894 | 0.093746 |
| 6 | 0.983016 | 0.98304 | 0.001051 | 0.002073 | 0.0254 | 0.125001 |
| 7 | 0.984351 | 0.98437 | 0.000892 | 0.00199 | 0.025796 | 0.12491 |
| 8 | 0.984993 | 0.985009 | 0.000813 | 0.00195 | 0.025916 | 0.171853 |
| 9 | 0.985096 | 0.985109 | 0.000796 | 0.001944 | 0.025899 | 0.281243 |
| 10 | 0.985092 | 0.985107 | 0.000799 | 0.001945 | 0.0259 | 0.187438 |

Hình 4.14 Biểu đồ so sánh các thông số trong 10 ngày tại trạm Đồng Hới sử dụng bộ dữ liệu trường hợp 2 - LR

**Kết luận:** Như kết quả ở bảng 4.8 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 9 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.



Hình 4.15 Kết quả tốt nhất ở trạm Đồng Hới so sánh giá trị thực tế (y\_test) và giá trị dự đoán (y\_pred) với số ngày dùng để dự đoán là 9 (k=9) - LR

Bảng 4.9 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Đồng Hới sử dụng mô hình Hồi quy tuyến tính

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** | **Time (sec)** |
| **TH1** | 0.620111 | 0.620411 | 0.22633 | 0.441091 | 4.100738 | 0.109605 |
| **TH2** | 0.619751 | 0.62065 | 0.222593 | 0.440889 | 4.105465 | 0.031282 |

Hình 4.16 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình hồi quy tuyền tính

### Mô hình LSTM

**Đầu vào:** Đầu vào của LSTM là dữ liệu ba chiều, có nghĩa là mỗi mẫu có hình dạng

dấu thời gian × thuộc tính.

Đầu ra:

* Mặc định: Last Hidden State (Trạng thái ẩn của bước thời gian cuối cùng)
* return\_sequences=True : Tất cả các trạng thái ẩn (Trạng thái ẩn của TẤT CẢ các bước thời gian)
* return\_state=True : Trạng thái ẩn cuối cùng+ Trạng thái ẩn cuối cùng (again [2] [3]!) + Trạng thái ô cuối cùng (Trạng thái ô của bước thời gian cuối cùng)
* return\_sequences=True + return\_state=True: Tất cả trạng thái ẩn (Trạng thái ẩn của TẤT CẢ các bước thời gian) + Trạng thái ẩn cuối cùng + Trạng thái ô cuối cùng (Trạng thái ô của bước thời gian cuối cùng)

**Phương pháp:** Mô hình LSTM được thiết kế gồm 50 cells, epochs lần lượt là [TH1: 350; TH2: 450], batch size là 256 và sử dụng thuật toán tối ưu Adam.

#### Kết quả thực nghiệm bài toán 1

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 350 để dự đoán mực nước ở Lệ Thủy ngày thứ k+1

Bảng 4.10 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1. Với Epochs = 350

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 350 | 1 | 0.899215 | 0.902997 | 0.078322 | 0.187883 | 1.870338 |
| 2 | 0.908662 | 0.91282 | 0.081848 | 0.178861 | 1.727957 |
| 3 | 0.913453 | 0.916254 | 0.072632 | 0.174107 | 1.719936 |
| 4 | 0.913039 | 0.917531 | 0.07246 | 0.174523 | 1.71418 |
| 5 | 0.913762 | 0.917314 | 0.071938 | 0.173792 | 1.706022 |
| 6 | 0.913732 | 0.917796 | 0.072044 | 0.173823 | 1.718797 |
| 7 | 0.914556 | 0.916208 | 0.075002 | 0.17299 | 1.74566 |

Hình 4.17 Biểu đồ so sánh các thông số trong 7 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 1 - LSTM

**Kết luận:** Như kết quả ở bảng 4.10 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 7 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 450 để dự đoán mực nước ở Lệ Thủy ngày thứ k+1

Bảng 4.11 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Lệ Thủy vào ngày thứ k+1. Với Epochs = 450

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 450 | 1 | 0.9081 | 0.912269 | 0.073523 | 0.17941 | 1.73766 |
| 2 | 0.910531 | 0.916139 | 0.072931 | 0.177022 | 1.727802 |
| 3 | 0.913317 | 0.915489 | 0.075583 | 0.174244 | 1.735433 |
| 4 | 0.912574 | 0.914943 | 0.07624 | 0.174989 | 1.721642 |
| 5 | 0.913174 | 0.915619 | 0.072614 | 0.174384 | 1.728761 |
| 6 | 0.914503 | 0.916595 | 0.075229 | 0.173044 | 1.698606 |

Hình 4.18 Biểu đồ so sánh các thông số trong 6 ngày tại trạm Lệ Thủy sử dụng bộ dữ liệu trường hợp 2 - LSTM

**Kết luận:** Như kết quả ở bảng 4.11 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 5 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Bảng 4.12 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Lệ Thủy sử dụng mô hình LSTM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** |
| **TH1** | 0.914556 | 0.916208 | 0.075002 | 0.17299 | 1.74566 |
| **TH2** | 0.913174 | 0.915619 | 0.072614 | 0.174384 | 1.728761 |

Hình 4.19 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình LSTM

#### Kết quả thực nghiệm bài toán 2

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 350 để dự đoán mực nước ở Kiến Giang ngày thứ k+1

Bảng 4.13 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1. Với Epochs = 350

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 350 | 1 | 0.597866 | 0.623024 | 0.191307 | 0.457685 | 4.073642 |
| 2 | 0.618499 | 0.651034 | 0.183784 | 0.445789 | 4.019323 |
| 3 | 0.615027 | 0.648687 | 0.184467 | 0.447813 | 4.042028 |
| 4 | 0.619268 | 0.651645 | 0.184481 | 0.445339 | 4.065321 |
| 5 | 0.622207 | 0.651129 | 0.183312 | 0.443619 | 4.098009 |
| 6 | 0.618567 | 0.646775 | 0.185216 | 0.44575 | 4.235851 |
| 7 | 0.620299 | 0.647218 | 0.186369 | 0.444738 | 4.201457 |

[1]Hình 4.20 Biểu đồ so sánh các thông số trong 7 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 1 - LSTM

**Kết luận:** Như kết quả ở bảng 4.13 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 7 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 450 để dự đoán mực nước ở Kiến Giang ngày thứ k+1

Bảng 4.14 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Kiến Giang vào ngày thứ k+1. Với Epochs = 450

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 450 | 1 | 0.619549 | 0.652371 | 0.186099 | 0.445175 | 3.954488 |
| 2 | 0.626801 | 0.655635 | 0.18259 | 0.440912 | 3.977892 |
| 3 | 0.617831 | 0.652382 | 0.186023 | 0.446179 | 4.032021 |
| 4 | 0.617934 | 0.651362 | 0.18788 | 0.446119 | 4.077421 |
| 5 | 0.642726 | 0.65422 | 0.193749 | 0.431403 | 4.035924 |
| 6 | 0.627605 | 0.650162 | 0.186161 | 0.440438 | 4.232785 |

Hình 4.21 Biểu đồ so sánh các thông số trong 6 ngày tại trạm Kiến Giang sử dụng bộ dữ liệu trường hợp 2 - LSTM

**Kết luận:** Như kết quả ở bảng 4.14 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 5 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

*Bảng 4.15:* So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Lệ Thủy sử dụng mô hình LSTM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** |
| **TH1** | 0.620299 | 0.647218 | 0.186369 | 0.444738 | 4.201457 |
| **TH2** | 0.642726 | 0.65422 | 0.193749 | 0.431403 | 4.035924 |

`

#### Kết quả thực nghiệm bài toán 3

**Trường hợp 1:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 350 để dự đoán mực nước ở Đồng Hới ngày thứ k+1

Bảng 4.15 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. Với Epochs = 350

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 350 | 1 | 0.724183 | 0.747176 | 0.074032 | 0.127374 | 0.864288 |
| 2 | 0.746803 | 0.767239 | 0.071528 | 0.122039 | 0.850521 |
| 3 | 0.743861 | 0.767824 | 0.071816 | 0.122746 | 0.84492 |
| 4 | 0.741341 | 0.76775 | 0.07247 | 0.123348 | 0.839713 |
| 5 | 0.741883 | 0.773586 | 0.07332 | 0.123166 | 0.854501 |
| 6 | 0.743209 | 0.775107 | 0.073685 | 0.12285 | 0.8553 |
| 7 | 0.735081 | 0.771456 | 0.075214 | 0.124779 | 0.855668 |

**Kết luận:** Như kết quả ở bảng 4.13 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 7 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

**Trường hợp 2:** Biết thông tin về lượng mưa và mực nước của các trạm trong k ngày liên tiếp và epochs là 450 để dự đoán mực nước ở Đồng Hới ngày thứ k+1

Bảng 4.16 Biết thông tin về bộ dữ liệu của các trạm trong k ngày liên tiếp, dự đoán mực nước ở trạm Đồng Hới vào ngày thứ k+1. Với Epochs = 450

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Epochs** | **Used data (Days)** | **NSE** | **R2** | **MAE** | **RMSE** | **MaxError** |
| 450 | 1 | 0.743466 | 0.769034 | 0.07296 | 0.122841 | 0.845113 |
| 2 | 0.737245 | 0.77405 | 0.075088 | 0.124321 | 0.863687 |
| 3 | 0.760216 | 0.76967 | 0.068147 | 0.118762 | 0.81966 |
| 4 | 0.756651 | 0.765515 | 0.068661 | 0.119642 | 0.837021 |
| 5 | 0.762222 | 0.77638 | 0.069277 | 0.118214 | 0.839503 |
| 6 | 0.767883 | 0.773198 | 0.067967 | 0.116799 | 0.836905 |

**Kết luận:** Như kết quả ở bảng 4.14 cho chúng ta thấy kết quả khi sử dụng dữ liệu của 5 ngày để dự đoán cho kết quả các thông số dự đoán tốt nhất.

Bảng 4.17 So sánh kết quả thực nghiệm của 2 trường hợp tại trạm Đồng Hới sử dụng mô hình LSTM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Case** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** |
| **TH1** | 0.735081 | 0.771456 | 0.075214 | 0.124779 | 0.855668 |
| **TH2** | 0.762222 | 0.77638 | 0.069277 | 0.118214 | 0.839503 |

Hình 4.22 Biểu đồ so sánh thông số đánh giá giữa hai trường hợp trong bài toán 2 – Mô hình LSTM

### Đánh giá kết quả thực nghiệm

Bảng 4.18 So sánh kết quả thực nghiệm tốt nhất ở các trạm trên tất cả các trường hợp

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case** | **Station** | **NSE** | **R2** | **MAE** | **RMSE** | **Max Error** |
| **TH1 LR** | Lệ Thủy | 0.911606 | 0.912156 | 0.086429 | 0.17611 | 1.779317 |
| Kiến Giang | 0.6201106 | 0.620411 | 0.226329603 | 0.44109141 | 4.100738402 |
| Đồng Hới | 0.9850881 | 0.9851014 | 0.000794373 | 0.00194494 | 0.025920913 |
| **TH2 LR** | Lệ Thủy | 0.9121048 | 0.9124133 | 0.086033309 | 0.17561947 | 1.764290896 |
| Kiến Giang | 0.619751 | 0.6206503 | 0.222593232 | 0.44088924 | 4.105465433 |
| Đồng Hới | 0.9851069 | 0.9851205 | 0.000794414 | 0.00194328 | 0.025908452 |
| **TH1 LSTM** | Lệ Thủy | 0.9145563 | 0.9162083 | 0.075001694 | 0.17298989 | 1.7456596 |
| Kiến Giang | 0.6202985 | 0.6472177 | 0.18636943 | 0.4447378 | 4.201457 |
| Đồng Hới | 0.7350807 | 0.7714565 | 0.07521428 | 0.12477879 | 0.8556684 |
| **TH2 LSTM** | Lệ Thủy | 0.9159794 | 0.9177832 | 0.07373159 | 0.17154323 | 1.6922672 |
| Kiến Giang | 0.6427261 | 0.65422 | 0.1937491 | 0.4314034 | 4.035924 |
| Đồng Hới | 0.7622222 | 0.7763805 | 0.069276884 | 0.11821419 | 0.8395033 |

Hình 4.23 Biểu đồ so sánh kết quả thực nghiệm tốt nhất giữa 2 trường hợp trong mô hình hồi quy tuyến tính

Như *hình 4.23* chúng ta thấy được với mô hình hồi quy tuyến tính trường hợp 2 khi chúng ta sử dụng bộ dữ liệu lượng mưa, mực nước và độ chênh lệch cho kết quả các thống số đánh giá tốt nhất.

Hình 4.24 Biểu đồ so sánh kết quả thực nghiệm nhất giữa 2 trường hợp trong mô hình hồi LSTM

Với *hình 4.24* chúng ta thấy được với mô hình LSTM trường hợp 2 khi chúng ta sử dụng bộ dữ liệu lượng mưa, mực nước và độ chênh lệch cho kết quả các thống số đánh giá tốt nhất đặc biệt với MaxError thấp hơn.

Hình 4.25 Biểu đồ so sánh trường hợp có kết quả thực nghiệm tốt nhất của hai mô hình hồi quy tuyến tính và mô hình LSTM

Với *hình 4.25* Sau khi tìm thấy được trường hợp tốt nhất trong từng mô hình để đánh giá bài toán dự báo mực nước. Ta có thể thấy trường hợp sử dụng bộ dữ liệu mực nước, lượng mưa và độ chênh lệch cho kết quả tốt hơn. Đánh giá kết quả từ hai mô hình hồi quy tuyến tính và mô hình LSTM chúng ta có thể thấy được rằng Mô hình hồi quy tuyến tính có kết quả các tham số đánh giá đều và tốt hơn.

**Kết luận:** Kết quả dự đoán mực nước tại cả 3 trạm sử dụng mô hình LR và mô hình LSTM được thể hiện trong bảng 3.16. Mực nước dự đoán bằng mô hình hồi quy LR với bộ dữ liệu gồm mực nước, lượng mưa và độ chênh lệch mực nước và lượng mưa ở các trạm cho ra kết quả tốt nhất và đồng đều ở cả 3 trạm với các thông số đánh giá tốt. Mô hình LSTM dự đoán mực nước dưới mức cũng khá tốt; tuy nhiên khi dự báo ở trạm Đồng Hới mô hình trả ra kết quả các thông số đánh giá chưa được tốt.

# KẾT LUẬN

## Kết quả đạt được

Trong luận văn đồ án, em đã xây dựng được 2 mô hình dự đoán mực nước theo ngày tại 3 trạm: Lệ Thủy, Kiến Giang và Đồng Hới trên sông Nhật Lệ dựa trên phương pháp hồi quy LR và LSTM. Bộ dữ liệu về lượng mưa và mực nước ngày tại 3 trạm trong mùa lũ của 44 năm đã được sử dụng để huấn luyện và kiểm định các mô hình. Kết quả nghiên cứu chỉ ra việc sử dụng bộ dữ liệu đầu vào gồm lượng mưa, mực nước ngày và độ chênh lệch giữ lượng mưa, mực nước giữa ngày t và (t+1) cho kết quả tốt nhất khi mô hình có thể dự báo tốt cả chân lẫn đỉnh của đường quá trình mực nước. Các chỉ số thống kê R2 , NSE, MAE và RMSE cho thấy việc ứng dụng mô hình định hướng dữ liệu là hoàn toàn khả thi và đáng tin cậy trong việc dự đoán mực nước; trong đó mô hình dự đoán mực nước bằng phương pháp hồi quy tuyến tính cho kết quả tốt hơn so với phương pháp LSTM.

## Hướng phát triển

Sau khi xây dựng mô hình và có kết quả thực nghiệm dự báo mực nước ở sông Nhật Lệ, em nhận thấy rằng kết quả các thông số đánh giá ở trạm Kiến Giang vẫn chưa được tốt vì vậy em xem xét hướng phát triển mô hình sẽ bổ sung them một số yếu tố đầu vàng khác như: dòng chảy, mực nước thủy triều, lượng mưa và mực nước ở 1 số trạm lân cận. Ngoài ra em sẽ tìm hiểu thêm các kỹ thuật học máy, thuật toán học sâu khác để áp dụng nhằm cải thiện chất lượng dự báo mực nước trong tương lai.

Vì thời gian có hạn và trình độ hiểu biết của em còn hạn chế nên đồ án này không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý của quý thầy cô để đồ án của em được hoàn thiện hơn.

Em xin chân thành cảm ơn!

# 

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | T. N. Toan, "Đánh giá model trong Machine Learning," 29 5 2022. [Online]. Available: https://viblo.asia/p/danh-gia-model-trong-machine-learing-RnB5pAq7KPG. [Accessed 29 5 2022]. |
| [2] | Z. C. L. M. l. A. J. S. Aston Zhang, D2L. |
| [3] | V. H. Tiệp, Machine Learning Cơ Bản. |
| [5] | T. Q. C. Đ. T. H. N. T. K. N. Đinh Nhật Quang, "DỰ BÁO MỰC NƯỚC TRÊN SÔNG KIẾN GIANG," *BÀI BÁO KHOA HỌC,* 2022. |
| [6] | H. V. H. Hồ Việt Tuấn, "Sử dụng mạng nơ ron nhân tạo dự báo mực nước sông chịu ảnh hưởng thủy triều," *Tạp chí Khoa học và Công nghệ thủy lợi,* pp. 52(1-9), 2019. |
| [7] | V. Tiệp, "Machine learning cơ bản," *Nhà xuất bản Khoa học và Kỹ Thuật,* 2018. |