**BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ NÔNG NGHIỆP VÀ PTNT**



**PHẠM VĂN ÁNH ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HÀ NỘI, NĂM 2025**

TRƯỜNG ĐẠI HỌC THỦY LỢI

PHẠM VĂN ÁNH

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ QUẢN LÝ LƯU LƯỢNG NƯỚC HỒ THỦY ĐIỆN TUA SRAH**

ĐỒ ÁN TỐT NGHIỆP

|  |  |  |  |
| --- | --- | --- | --- |
| HÀ NỘI, NĂM 2025   |  |  | | --- | --- | | **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** | |  |

TRƯỜNG ĐẠI HỌC THỦY LỢI

PHẠM VĂN ÁNH

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ QUẢN LÝ LƯU LƯỢNG NƯỚC HỒ THỦY ĐIỆN TUA SRAH**

|  |  |
| --- | --- |
| Ngành : | Hệ thống thông tin |
| Mã sinh viên: | 2051063673 |

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | TS. TRẦN MẠNH TUẤN  PGS.TS. TRẦN KIM CHÂU |

HÀ NỘI, NĂM 2025

|  |  |
| --- | --- |
|  | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập - Tự do - Hạnh phúc**  \*  **NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP** |

**Họ tên sinh viên:** PHẠM VĂN ÁNH **Hệ đào tạo:** Đại học chính quy

**Lớp:** 62HT **Ngành:** Hệ thống thông tin

**Khoa:** Công nghệ thông tin

**1- TÊN ĐỀ TÀI:**

Ứng dụng khai phá dự liệu để quản lý lượng nước hồ thủy điện Tua Srah

**2- CÁC TÀI LIỆU CƠ BẢN:**

[1]Quy định của pháp luật trong quản lý tài nguyên nước: [Văn bản hợp nhất 34/VBHN-VPQH 2020 Luật Tài nguyên nước (thuvienphapluat.vn)](https://thuvienphapluat.vn/van-ban/Tai-nguyen-Moi-truong/Van-ban-hop-nhat-34-VBHN-VPQH-2020-Luat-Tai-nguyen-nuoc-469844.aspx)

[2]Bài báo về việc áp dụng công nghệ cao, trí thông minh nhân tạo trong quản lý tài nguyên nước cũng như phòng chống thiên tai: [Nghiên cứu, ứng dụng công nghệ cảnh báo dự báo thiên tai: Tạo bước đột phá đổi mới (baotainguyenmoitruong.vn)](https://baotainguyenmoitruong.vn/nghien-cuu-ung-dung-cong-nghe-canh-bao-du-bao-thien-tai-tao-buoc-dot-pha-doi-moi-351872.html)

[3] Bài báo về sử dụng trí tuệ nhân tạo trong sự báo lưu lượng nước: [221 (vap.ac.vn)](https://proceeding.vap.ac.vn/index.php/proceedingvap/article/viewFile/2016.00015/221)

[4] Bài báo về dự báo lưu lượng nước sử dụng Random Forest: [Áp dụng mô hình trí tuệ nhân tạo vào dự báo lưu lượng đến hồ lưu vực sông Ba - TaiLieu.VN](https://tailieu.vn/doc/ap-dung-mo-hinh-tri-tue-nhan-tao-vao-du-bao-luu-luong-den-ho-luu-vuc-song-ba-2213989.html)

[5] LSTM: [sundermeyer12\_interspeech.pdf (isca-archive.org)](https://www.isca-archive.org/interspeech_2012/sundermeyer12_interspeech.pdf)

[6] Python: [Welcome to Python.org](https://www.python.org/)

**3 - NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN:**Tỷ lệ %

|  |  |
| --- | --- |
| **Nội dung các phần** |  |
| Chương 1: Tổng quan cơ sở lý thuyết | 20 |
| Chương 2: Xây dựng phương pháp dự báo | 40 |
| Chương 3: Xây dựng hệ thống mô phỏng và kết quả thực nghiệm Xây dựng hệ thống  Mô tả hệ thống  Các thuật toán, thư viện sử dụng Kết quả thực nghiệm | 40 |

**4. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN**

|  |  |
| --- | --- |
| **Phần** | **Họ và tên giáo viên hướng** |
|  | **dẫn** |
| Chương 1: Tổng quan cơ sở lý thuyết | TS. Trần Mạnh Tuấn |
| Chương 2: Xây dựng phương pháp | TS. Trần Mạnh Tuấn  PGS.TS Trần Kim Châu |
| Chương 3: Xây dựng hệ thống mô phỏng và kết quả thực nghiệm | TS. Trần Mạnh Tuấn |

1. **NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP**

Ngày ............ tháng năm 20

|  |  |
| --- | --- |
| **Trưởng Bộ môn**  *(Ký và ghi rõ Họ tên)* | **Giáo viên hướng dẫn chính**  *(Ký và ghi rõ Họ tên)* |

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

Ngày. . . . .tháng. . . . .năm 20....

**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 20...

**Sinh viên làm Đồ án tốt nghiệp**

*(Ký và ghi rõ Họ tên)*

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

TÊN ĐỀ TÀI: Ứng dụng khai phá dữ liệu để quản lý lưu lượng nước hồ thủy điện Tua Srah

*Sinh viên thực hiện*: Phạm Văn Ánh

*Lớp*: 62HT

*Mã sinh viên: 2051063673*

*Số điện thoại:0944053537*

*Email:* phamvananh010102@gmail.com

*Giáo viên hướng dẫn*: TS. Trần Mạnh Tuấn, PGS.TS Trần Kim Châu

**TÓM TẮT ĐỀ TÀI**

**Bài toán: Ứng dụng khai phá dữ liệu để quản lý lưu lượng nước hồ thủy điện Tua Srah**

Việc quản lý hiệu quả tài nguyên nước, đặc biệt là tại các hồ thủy điện, đóng vai trò quan trọng trong phát triển kinh tế - xã hội và phòng chống thiên tai. Trong bối cảnh biến đổi khí hậu, nhu cầu về các giải pháp dự báo lưu lượng nước chính xác và hiệu quả ngày càng cấp thiết[1].

Vì vậy ngoài những phương pháp tính toán thực tế thì ngày nay đã áp dụng những phương pháp học máy hay là sử dụng mô hình AI vào việc dự báo lưu lượng nước trong hồ thủy điện[2]. Đã có những bài toán sử dụng mô hình mạng nơ-ron nhân tạo ANN, DNN[3] hay sử dụng các giải thuật học máy như Random forest[4].

Đối với bài toán sử dụng mô hình học máy trong quản lý lưu lượng nước hồ thủy điện Tua Srah thì dữ liệu bao gồm Q( lượng nước thực tế vào hộ), Z( mực nước hồ thực tế), q( lượng nước xả thực tế). Do vậy có thể sử dụng mô hình LSTM( Long Short Term Memory), để huấn luyện mô hình đưa ra kết quả đó là dự đoán q xả. Mô hình LSTM[5] là một mô hình phù hợp với dữ liệu chuỗi thời gian với mối quan hệ dài hạn.

Kết quả thu được là một mô hình dự đoán được q xả khi biết tham số đầu vào là Q và Z. Sau khi đã có mô hình sẽ tiếp tục sử dụng Q ở tương lai để cho việc tinh toán Z từ đó có 2 tham số đầu vào cần thiết để dự đoán q xả.

Việc này mang ý nghĩa:

**Tối ưu hóa vận hành**: Học máy giúp tối ưu hóa quy trình vận hành các tua bin và xả nước, đảm bảo cung cấp điện liên tục, đồng thời giảm thiểu tác động xấu đến môi trường.

**Phòng chống thiên tai**: Các hệ thống dự báo lũ lụt và hạn hán dựa trên học máy giúp cảnh báo sớm, từ đó đưa ra biện pháp phòng tránh và giảm thiểu thiệt hại.

**Tăng cường hiệu quả kinh tế**: Nhờ dự báo chính xác và quản lý tốt, các nhà máy thủy điện có thể tăng sản lượng điện, giảm lãng phí nước và tiết kiệm chi phí vận hành.

**CÁC MỤC TIÊU CHÍNH**

* Tìm hiểu về tầm quan trọng của dự báo lưu lượng nước trong thực tế.
* Tìm hiều nội dung và cách hoạt động của các mô hình học máy.
* Phát triển mô hình học máy phù hợp để dự báo lưu lượng nước trong hồ.
* Đánh giá hiệu quả của mô hình.

**NỘI DUNG CHÍNH**

* Thu thập dữ liệu thực tế từ hồ: các thông số Q( lượng nước vào hồ thực tế), Z( độ cao mực nước hồ) và q( lượng nước xả thực tế).
* Tiền xử lý dữ liệu: Dữ liệu được kiểm tra và điều chỉnh kiểu dữ liệu tương ứng đúng với từng thuộc tính, đồng thời loại bỏ những mẫu dữ liệu không hợp lệ hoặc dữ liệu thiếu.
* Lựa chọn các mô hình phù hợp cho bài toán và dữ liệu.
* Tối ưu hóa mô hình: Tinh chỉnh và tối ưu hóa các tham số của mô hình để cải thiện độ chính xác.
* Đánh giá mô hình.
* Xây dựng hệ thống tích hợp hệ thống dự báo lượng nước cần xả trong tương lai.

**PHƯƠNG PHÁP DỰ ĐỊNH THỰC HIỆN**

Công nghệ dự kiến sử dụng Python[3].

Thu thập và xây dựng tập dữ liệu liên quan đến lưu lượng nước:

Sử dụng các thư viện Python như Pandas, NumPy để thu thập và xử lý dữ liệu về mực nước, lưu lượng nước từ các nguồn dữ liệu như cảm biến hoặc cơ sở dữ liệu công cộng.

Sử dụng công cụ trực quan hóa dữ liệu như Matplotlib và Seaborn để kiểm tra và làm sạch dữ liệu.

**Huấn luyện mô hình học máy dự báo lưu lượng nước**

Lựa chọn mô hình học máy phù hợp với đề tài.

Sử dụng TensorFlow hoặc PyTorch để xây dựng và huấn luyện mô hình, đồng thời tối ưu hóa các siêu tham số như số lượng lớp, số lượng nơ-ron, tốc độ học (learning rate) để đạt độ chính xác cao.

**Phát triển thuật toán dự báo lưu lượng nước dựa trên mô hình học máy**

Thu thập dữ liệu đầu vào( Q tương lai) sau đó đưa ra kết quả là lượng nước mà hồ cần xả.

Sử dụng các kỹ thuật học máy khác (như hồi quy tuyến tính, cây quyết định) để kết hợp và so sánh kết quả dự báo.

**Đánh giá hiệu quả của mô hình trên tập dữ liệu thử nghiệm:**

Đánh giá hiệu suất mô hình bằng các chỉ số như **MSE (Mean Squared Error)**, **RMSE (Root Mean Squared Error)**, **MAE (Mean Absolute Error)** để đo độ lệch giữa giá trị dự báo và giá trị thực tế.

**KẾT QUẢ DỰ KIẾN**

* Demo mô hình dự báo lưu lượng nước.
* Báo cáo tổng kết của đồ án tốt nghiệp

# LỜI CAM ĐOAN

Tác giả xin cam đoan đây là Đồ án tốt nghiệp của bản thân tác giả. Các kết quả trong Đồ án tốt nghiệp này là trung thực, và không sao chép từ bất kỳ một nguồn nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu (nếu có) đã được thực hiện trích dẫn và ghi nguồn tài liệu tham khảo đúng quy định.

**Tác giả ĐATN**

*Chữ ký*

**Phạm Văn Ánh**

# LỜI CẢM ƠN

Trong suốt bốn năm rưỡi học tập và nghiên cứu tại Khoa Công nghệ thông tin, Trường Đại học Thủy Lợi, em đã nhận được sự quan tâm, dìu dắt và hướng dẫn tận tình từ quý thầy cô. Thầy cô và bạn bè đã luôn đồng hành, động viên và giúp đỡ em trên hành trình chinh phục tri thức. Không chỉ được trau dồi kiến thức chuyên môn, em còn được lắng nghe những câu chuyện, những chia sẻ quý báu, đầy tâm huyết từ thầy cô, anh chị và bạn bè.

Em đã có cơ hội được thử sức, trải nghiệm và phát triển bản thân mình trên nhiều sân chơi bổ ích như: nghiên cứu khoa học, các cuộc thi ý tưởng khởi nghiệp, Olympic Toán, Olympic Tin học,… Qua mỗi cuộc thi, mỗi bài tập lớn, mỗi môn học trên lớp, em ngày càng trưởng thành và tự tin hơn.

Em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến quý thầy cô trong Khoa Công nghệ thông tin nói chung, và Tiến sĩ Trần Mạnh Tuấn nói riêng. Thầy đã tận tình hướng dẫn, chỉ bảo, định hướng và giúp đỡ em hoàn thành tốt đồ án tốt nghiệp này. Em vô cùng trân trọng và biết ơn những kiến thức, kinh nghiệm cùng sự quan tâm mà thầy đã dành cho em.

Em xin bày tỏ lòng biết ơn sâu sắc đến tất cả quý thầy cô trong Khoa Công nghệ thông tin đã dìu dắt, truyền đạt cho em những kiến thức chuyên môn quý báu và rèn luyện cho em những kỹ năng mềm cần thiết. Nhờ sự tận tâm của thầy cô, em đã có định hướng rõ ràng hơn cho con đường tương lai và hiểu rõ bản thân cần phải làm gì để đạt được mục tiêu của mình. Những kiến thức và kỹ năng ấy chính là nền tảng vững chắc để em tự tin bước vào đời, bước ra ngoài xã hội rộng lớn.

Đồ án tốt nghiệp là kết quả của quá trình nỗ lực, cố gắng và là tâm huyết của em trong học phần cuối cùng này. Mặc dù đã cẩn thận thực hiện và cố gắng hết sức, nhưng đồ án chắc chắn không tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu từ quý thầy cô và các bạn để đồ án được hoàn thiện hơn nữa.

Em xin chân thành cảm ơn!

**MỤC LỤC**

[LỜI CAM ĐOAN i](#_Toc187072787)

[LỜI CẢM ƠN ii](#_Toc187072788)

[MỞ ĐẦU 1](#_Toc187072789)

[CHƯƠNG 1 TỔNG QUAN VỀ BÀI TOÁN DỰ BÁO LƯU LƯỢNG NƯỚC 3](#_Toc187072790)

[1.1 Mô tả bài toán 3](#_Toc187072791)

[1.2 Tình hình nghiên cứu quốc tế và trong nước hiện nay 4](#_Toc187072792)

[1.3 Khai phá dữ liệu 5](#_Toc187072793)

[1.3.1Tổng quan 5](#_Toc187072794)

[1.3.2 Quy trình khai phá dữ liệu 7](#_Toc187072795)

[1.3.3 Các phương pháp khai phá dữ liệu 9](#_Toc187072796)

[1.3 Ngôn ngữ sử dụng 14](#_Toc187072797)

[1.4.1. Tổng quan về ngôn ngữ lập trình Python 14](#_Toc187072798)

[1.4.2. Ưu, nhược điểm của ngôn ngữ lập trình Python 15](#_Toc187072799)

[CHƯƠNG 2 MÔ HÌNH KHAI PHÁ DỮ LIỆU ĐỂ DỰ ĐOÁN LƯU LƯỢNG NƯỚC ĐẾN HỒ 16](#_Toc187072800)

[2.1 Mô hình tổng quát 16](#_Toc187072801)

[2.1.1. Bộ dữ liệu ban đầu 16](#_Toc187072802)

[2.1.2. Tiền xử lý dữ liệu 17](#_Toc187072803)

[2.1.3. Khám phá dữ liệu 17](#_Toc187072804)

[2.2 Chi tiết các bước thực hiện 19](#_Toc187072805)

[2.3 Mạng Nơ-ron( Neural Network) 21](#_Toc187072806)

[2.3.1 Giới thiệu về mạng nơ-ron (Neural Network) 21](#_Toc187072807)

[2.3.3 Ứng dụng của Neural Network 22](#_Toc187072808)

[2.4 Mô hình kết hợp LSTM- RF 23](#_Toc187072809)

[2.4.1 Mạng LSTM (Long Short-Term Memory): 23](#_Toc187072810)

[2.4.2 Giải thuật Random Forest ( RF): 25](#_Toc187072811)

[2.4.2 Mô hình LSTM-RF 27](#_Toc187072812)

[2.5 Kết hợp LSTM và RF 29](#_Toc187072813)

[2.5.1 Quá trình dự báo 29](#_Toc187072814)

[2.5.2 Ưu điểm của mô hình LSTM-RF 30](#_Toc187072815)

[2.6 Nhận xét và đánh giá 30](#_Toc187072816)

[CHƯƠNG 3 CÀI ĐẶT THỰC NGHIỆM 32](#_Toc187072817)

[3.1 Mô tả dữ liệu 32](#_Toc187072818)

[3.2 Tiền xử lý dữ liệu 33](#_Toc187072819)

[3.2.1 Tiền xử lý dữ liệu 33](#_Toc187072820)

[3.2.2 Phân tích tương quan giữa các thuộc tính 33](#_Toc187072821)

[3.2.3 Phân tích thành phần chính (PCA) 33](#_Toc187072822)

[3.2.4 Đánh giá hiệu suất dự báo 34](#_Toc187072823)

[3.2.5 Kết luận và đề xuất 34](#_Toc187072824)

[3.3 Độ đo đánh giá 35](#_Toc187072825)

[3.4 Môi trường thực nghiệm 38](#_Toc187072826)

[3.5 Kết quả thực nghiệm 40](#_Toc187072827)

[3.5.1 Cài đặt môi trường thực nghiệm 40](#_Toc187072828)

[3.5.2 Đánh giá 42](#_Toc187072829)

[3.6 Ứng dụng 46](#_Toc187072830)

[3.6.1 Tổng quan hệ thống 46](#_Toc187072831)

[3.6.3 Triển khai chương trình thử nghiệm 50](#_Toc187072832)

[3.7 Nhận xét và đánh giá 51](#_Toc187072833)

[KẾT LUẬN 54](#_Toc187072834)

[TÀI LIỆU THAM KHẢO 57](#_Toc187072835)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Hình ảnh thủy điện Buôn Tua Srah 3](#_Toc186615628)

[Hình 2. Biểu đồ quy trình khai phá tri thức 8](#_Toc186615629)

[Hình 3. Sơ đồ mô hình tổng quát 17](#_Toc186615630)

[Hình 4. Mô hình mạng nơ-ron 20](#_Toc186615631)

[Hình 5. Mô hình LSTM 22](#_Toc186615632)

[Hình 6. Sơ đồ LSTM kết hợp RF 28](#_Toc186615633)

[Hình 7. Dữ liệu demo 35](#_Toc186615634)

[Hình 8. Biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế của LSTM 44](#_Toc186615635)

[Hình 9. Biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế của LSTM-RF 45](#_Toc186615636)

[Hình 10. Thang điểm của LSTM 46](#_Toc186615637)

[Hình 11. Thang điểm của LSTM- RF 46](#_Toc186615638)

[Hình 12. Giao diện ứng dụng 49](#_Toc186615639)

[Hình 13. Kết quả trả về và hiển thị thông báo dự đoán 50](#_Toc186615640)

[Hình 14. Kết quả trả về vào file excel 51](#_Toc186615641)

**DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ**

|  |  |
| --- | --- |
| Long Short Term Memory |  |
| DATN | đồ án tốt nghiệp |
| Random Forest |  |
| Mean absolute error | Sai số trung bình tuyệt đối |
| Mean squared error | Lỗi bình phương trung bình |
| Root mean squared error | Căn bình phương sai số |
| Radial Basis Function | Hàm cơ sở xuyên tâm |

# MỞ ĐẦU

Hồ chứa nước đóng vai trò quan trọng trong việc trữ nước và điều hòa sự phân bố nguồn nước tự nhiên, góp phần giảm thiểu những biến động bất lợi của môi trường. Không chỉ dự trữ nguồn nước mưa để cung cấp cho sản xuất thủy điện và tưới tiêu nông nghiệp, các hồ chứa còn có khả năng giảm thiểu dòng chảy mạnh gây lũ lụt và cung cấp nước trong điều kiện hạn hán. Tuy nhiên, dự báo dòng chảy đến các hồ chứa là một nhiệm vụ phức tạp bởi tính chất phi tuyến, cường độ thay đổi cao, và hành vi không cố định của dòng chảy. Quá trình vận hành hồ chứa đòi hỏi phải đưa ra các quyết định kịp thời và chính xác để tối ưu hóa lưu trữ, xả nước và đảm bảo an toàn vận hành.

Trong bối cảnh biến đổi khí hậu ngày càng nghiêm trọng, việc quản lý các hồ chứa nước đang đối mặt với nhiều thách thức lớn, bao gồm lũ lụt, hạn hán và những biến động khó lường về thời tiết. Các hiện tượng cực đoan này không chỉ ảnh hưởng đến an toàn hồ đập mà còn gây nguy cơ cho sản xuất nông nghiệp, đời sống của người dân và sự ổn định của nguồn năng lượng thủy điện. Do đó, việc phát triển một hệ thống dự báo lưu lượng nước đến hồ chứa là một yêu cầu cấp thiết để hỗ trợ công tác quản lý và vận hành hồ chứa trong điều kiện phức tạp hiện nay.

Hồ thủy điện Tua Srah, nằm trên sông Đồng Nai tại tỉnh Đắk Nông, là một công trình thủy điện quan trọng. Ngoài chức năng phát điện, hồ còn điều tiết nước cho tưới tiêu nông nghiệp, sinh hoạt và kiểm soát lũ lụt tại khu vực. Tuy nhiên, do địa hình rừng núi hiểm trở, hồ chứa này chịu tác động lớn từ các điều kiện thời tiết khắc nghiệt, bao gồm lũ quét trong mùa mưa và khô hạn trong mùa khô. Điều này đặt ra nhu cầu cấp bách về một hệ thống dự báo chính xác lưu lượng nước đến hồ chứa, giúp người quản lý đưa ra quyết định tối ưu trong việc vận hành hồ, lưu trữ và xả nước.

Nhiều nghiên cứu trước đây đã tập trung vào việc xây dựng các mô hình dự báo lưu lượng nước, từ các phương pháp truyền thống dựa trên mô hình vật lý đến các phương pháp hiện đại dựa trên học máy. Trong bối cảnh đó, đồ án này đề xuất sử dụng mô hình LSTM (Long Short-Term Memory), một phương pháp học sâu hiệu quả trong việc xử lý dữ liệu chuỗi thời gian[1], để dự báo lưu lượng nước đến hồ Tua Srah. Mô hình LSTM có khả năng ghi nhớ thông tin dài hạn và ngắn hạn, giúp xử lý tốt các dữ liệu biến động phức tạp và phi tuyến.

Mục tiêu của đồ án là xây dựng một mô hình dự báo lưu lượng nước chính xác, ứng dụng trên bộ dữ liệu lịch sử của hồ Tua Srah, bao gồm các thông số về mực nước hồ, lưu lượng nước đến, lưu lượng xả, và dữ liệu khí tượng từ các trạm mưa lân cận. Mô hình sẽ được kiểm chứng qua các thước đo hiệu quả như MAE, RMSE và R². Kết quả thu được sẽ là cơ sở để phát triển một hệ thống hỗ trợ quyết định cho việc quản lý hồ chứa Tua Srah, góp phần giảm thiểu rủi ro và tối ưu hóa nguồn tài nguyên nước trong bối cảnh biến đổi khí hậu hiện nay.

# CHƯƠNG 1 TỔNG QUAN VỀ BÀI TOÁN DỰ BÁO LƯU LƯỢNG NƯỚC

## 1.1 Mô tả bài toán

Nằm trong vùng địa hình rừng núi hiểm trở của tỉnh Đắk Nông, thủy điện Tua Srah chịu ảnh hưởng lớn từ các điều kiện thời tiết khắc nghiệt[2]. Vào mùa khô, dòng chảy của nhiều sông suối trong khu vực giảm mạnh do nhiệt độ cao và lượng mưa thấp, khiến nguy cơ thiếu nước trở nên rõ rệt. Ngược lại, mùa mưa kéo dài thường xuyên mang đến các đợt lũ lụt, bao gồm cả lũ quét, gây ảnh hưởng nghiêm trọng đến hoạt động sản xuất và đời sống của người dân ở vùng hạ du.



Hình 1. Hình ảnh thủy điện Buôn Tua Srah

Hồ thủy điện Tua Srah, nằm trên sông Đồng Nai, là một công trình thủy điện quan trọng của tỉnh Đắk Nông, Việt Nam. Vai trò của hồ chứa này không chỉ dừng lại ở việc phát điện, mà còn điều tiết nước cho nông nghiệp, sinh hoạt và kiểm soát lũ. Tuy nhiên, để vận hành hiệu quả, việc dự báo chính xác lưu lượng nước đến hồ là cần thiết. Điều này giúp đảm bảo an toàn hồ đập, tối ưu hóa việc lưu trữ và xả nước, đồng thời giảm thiểu thiệt hại do biến đổi khí hậu và thời tiết thất thường gây ra.

Bài toán đặt ra là xây dựng một mô hình dự báo lưu lượng nước đến hồ chứa Tua Srah, sử dụng các dữ liệu lịch sử thu thập từ quá trình vận hành hồ. Các thuộc tính dữ liệu bao gồm:

* Mực nước hồ (m).
* Lưu lượng nước đến hồ (m³/s).
* Tổng lưu lượng nước xả ra khỏi hồ (m³/s).

Đồ án đề xuất sử dụng một mô hình học máy mạng học sâu, cụ thể là mô hình **LSTM ( Long Short Term Memory)** nhằm giải quyết bài toán.

## 1.2 Tình hình nghiên cứu quốc tế và trong nước hiện nay

Dự báo lưu lượng nước đến hồ chứa là một lĩnh vực nghiên cứu quan trọng trong quản lý tài nguyên nước và vận hành các công trình thủy lợi, thủy điện. Trước những diễn biến khó lường của biến đổi khí hậu, việc phát triển các mô hình dự báo lưu lượng nước với độ chính xác cao đóng vai trò quan trọng hơn bao giờ hết.

**Nghiên cứu quốc tế**

Trong những năm gần đây, các nhà nghiên cứu quốc tế đã đạt được những tiến bộ đáng kể trong việc áp dụng các mô hình học máy và trí tuệ nhân tạo để dự báo lưu lượng nước. Chẳng hạn, các mô hình như mạng nơ-ron nhân tạo (ANN), máy hỗ trợ vectơ hồi quy (SVR), cùng với các phương pháp lai khác đã cho thấy hiệu quả trong việc nâng cao độ chính xác của dự báo. Những mô hình này có khả năng xử lý lượng dữ liệu lớn, phát hiện các mối quan hệ phi tuyến và dự đoán chính xác trong các tình huống phức tạp.

Nghiên cứu tại một số nơi đã sử dụng mạng nơ-ron nhân tạo để mô hình hóa quá trình mưa-dòng chảy, cho thấy khả năng tái hiện chính xác dòng chảy sông và hỗ trợ quản lý nguồn nước hiệu quả [3]. Tương tự, các kỹ thuật học máy như MLP, SVR và Random Forest đã được áp dụng để dự báo dòng chảy sông, đạt độ chính xác cao và giúp quản lý nguồn nước trong điều kiện khắc nghiệt [4]. Ngoài ra, Extreme Learning Machine (ELM), một mô hình học máy tiên tiến, đã chứng minh khả năng vượt trội trong việc dự báo dòng chảy sông tại các khu vực nhiệt đới [5].

Những nghiên cứu này khẳng định rằng các phương pháp học máy tiên tiến không chỉ nâng cao hiệu quả dự báo mà còn cung cấp công cụ mạnh mẽ để quản lý tài nguyên nước trong điều kiện biến đổi khí hậu.

**Nghiên cứu tại Việt Nam**

Tại Việt Nam, việc dự báo lưu lượng nước vẫn chủ yếu dựa trên các mô hình vật lý truyền thống. Các nghiên cứu thường tập trung vào các mô hình thủy văn như MIKE SHE, MIKE NAM và HEC-HMS. Ví dụ, mô hình MIKE SHE được áp dụng tại lưu vực sông Trà Khúc-Sông Vệ, sử dụng dữ liệu mưa dự báo từ hệ thống IFS để dự báo lưu lượng nước [6]. Tương tự, các nghiên cứu tại hồ Sông Hinh và Ban Chat đã sử dụng các mô hình HEC-HMS để dự báo lũ và hỗ trợ quản lý vận hành hồ chứa [7].

Mặc dù các nghiên cứu này đã mang lại những kết quả đáng khích lệ, việc ứng dụng các mô hình học máy tiên tiến vẫn còn hạn chế. Điều này chỉ ra một khoảng cách lớn giữa nghiên cứu trong nước và quốc tế.

**Hướng nghiên cứu mới**

Để thu hẹp khoảng cách này, việc ứng dụng các mô hình học máy hiện đại vào dự báo lưu lượng nước là một hướng đi mới và tiềm năng. Trong đề tài này, mô hình LSTM (Long Short-Term Memory) được đề xuất do khả năng xử lý tốt các dữ liệu phi tuyến và biến động cao. LSTM có ưu điểm vượt trội trong việc ghi nhớ cả thông tin dài hạn và ngắn hạn, giúp giải quyết bài toán dự báo lưu lượng nước đến hồ chứa với độ chính xác cao hơn so với các phương pháp truyền thống.

Nghiên cứu tập trung vào hồ chứa thủy điện Tua Srah tại tỉnh Đắk Nông, sử dụng bộ dữ liệu lịch sử bao gồm mực nước hồ, lưu lượng nước đến, lưu lượng xả, và các thông số khí tượng thủy văn từ các trạm đo lân cận. Mô hình LSTM được đào tạo và kiểm nghiệm qua các thước đo hiệu quả như MAE, RMSE, và R² để đánh giá độ chính xác và hiệu quả trong dự báo lưu lượng nước.

Kết quả của nghiên cứu không chỉ hỗ trợ công tác quản lý hồ chứa và vận hành thủy điện mà còn góp phần giảm thiểu rủi ro do lũ lụt và hạn hán, đồng thời đóng góp vào việc bảo vệ môi trường và phát triển bền vững nguồn tài nguyên nước tại Việt Nam.

## Khai phá dữ liệu

### 1.3.1Tổng quan

Khai phá dữ liệu là quá trình tìm hiểu, phân tích và hiểu rõ dữ liệu từ các nguồn dữ liệu không có cấu trúc hoặc có cấu trúc nhưng chưa được khám phá [8].

Mục tiêu chính của khai phá dữ liệu là tìm ra thông tin, mối quan hệ và xu hướng tiềm ẩn trong dữ liệu một cách tự nhiên, không giả định trước. Quá trình này giúp khám phá đặc điểm dữ liệu, đồng thời xác định các mối quan hệ, xu hướng và sự biến động trong đó. Kết quả của khai phá dữ liệu có thể tạo ra các đặc trưng mới, loại bỏ nhiễu và cung cấp một cái nhìn sâu hơn về bản chất dữ liệu [9].

Gần đây, khai phá dữ liệu đã chứng kiến sự phát triển vượt bậc, mở ra nhiều lợi ích và cơ hội lớn hơn so với các phương pháp phân tích truyền thống[10]. Các lĩnh vực như cơ sở dữ liệu, học máy, trí tuệ nhân tạo, lý thuyết thông tin, thống kê xác suất và tính toán hiệu năng cao đóng vai trò cốt lõi trong việc phát triển các kỹ thuật khai phá dữ liệu.

Dưới đây là một số kỹ thuật khai phá dữ liệu phổ biến:

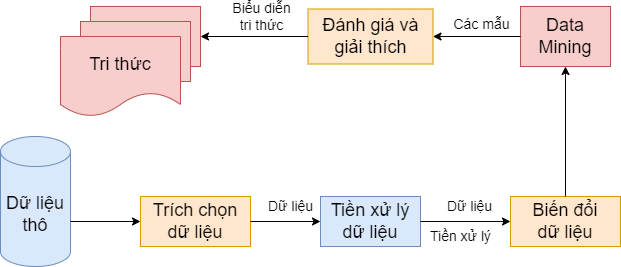
* **Học máy (Machine Learning):** Sử dụng các thuật toán như học sâu (Deep Learning), máy vector hỗ trợ (Support Vector Machines - SVM), cây quyết định (Decision Trees) và học tăng cường (Reinforcement Learning) để khai thác các mẫu ẩn và xây dựng mô hình dự đoán từ dữ liệu. Học máy được chứng minh là một công cụ quan trọng trong việc phân tích các hệ thống phức tạp, giúp nâng cao độ chính xác trong dự đoán. [11].
* **Trí tuệ nhân tạo (Artificial Intelligence):** Các kỹ thuật như mạng nơ-ron nhân tạo (Artificial Neural Networks), xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), và các hệ thống thông minh được áp dụng rộng rãi để nhận diện mẫu, phân loại và dự đoán từ dữ liệu. LeCun và cộng sự (2015) đã chứng minh rằng mạng nơ-ron sâu đặc biệt hiệu quả trong việc xử lý dữ liệu phi tuyến tính và không có cấu trúc rõ ràng.[12].
* **Xác suất thống kê (Probability and Statistics):** Các công cụ thống kê, bao gồm hồi quy, kiểm định giả thuyết và phân phối xác suất, thường được áp dụng để phân tích dữ liệu và phát hiện mối liên hệ giữa các biến. Montgomery & Runger (2014) nhấn mạnh vai trò của phân tích thống kê trong việc hiểu và dự báo các xu hướng dữ liệu [13].
* **Lý thuyết thông tin (Information Theory):** Các khái niệm như entropy, mutual information và các độ đo tương tự được sử dụng để phân tích thông tin và đo lường sự không chắc chắn trong dữ liệu. Shannon (1948) đã đặt nền tảng cho lý thuyết thông tin, giúp khai thác hiệu quả dữ liệu trong các hệ thống phức tạp [14].
* **Cơ sở dữ liệu (Database):** Hệ quản trị cơ sở dữ liệu (DBMS) và ngôn ngữ truy vấn như SQL đóng vai trò quan trọng trong việc quản lý, lưu trữ và truy xuất dữ liệu một cách hiệu quả, tạo nền tảng cho các hoạt động khai phá dữ liệu. Silberschatz và cộng sự (2010) đã có những đóng góp quan trọng trong việc phát triển các phương pháp xử lý và quản lý dữ liệu. [15].
* **Tính toán hiệu năng cao (High-Performance Computing):** Các công nghệ tính toán hiệu năng cao, như hệ thống phân tán, xử lý song song và sử dụng GPU, đã được áp dụng để xử lý khối lượng dữ liệu lớn và phức tạp một cách nhanh chóng. Dean & Ghemawat (2008) giới thiệu MapReduce, một trong những công cụ quan trọng trong tính toán phân tán, giúp tăng tốc đáng kể trong xử lý dữ liệu lớn [16].

Những kỹ thuật này phối hợp với nhau, đóng vai trò quan trọng trong việc khai phá dữ liệu và tạo ra những thông tin giá trị, hỗ trợ các quyết định quản lý và chiến lược trong nhiều lĩnh vực.

### 1.3.2 Quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu là một chuỗi các giai đoạn liên tục và có hệ thống nhằm tìm kiếm, xử lý, và phân tích dữ liệu để rút ra những thông tin có ý nghĩa và giá trị thực tiễn. Quy trình này bao gồm nhiều bước liên kết chặt chẽ với nhau, từ việc chuẩn bị dữ liệu thô ban đầu đến việc áp dụng các phương pháp phân tích và mô hình hóa để trích xuất các kết quả hữu ích. Những bước này thường được thực hiện theo một trình tự hợp lý, mặc dù đôi khi có thể phải quay lại các bước trước đó để điều chỉnh và cải thiện quy trình.

Thông qua quy trình khai phá dữ liệu, người thực hiện không chỉ làm sạch và chuyển đổi dữ liệu thành dạng có thể sử dụng mà còn khám phá các mẫu, mối quan hệ, hay xu hướng tiềm ẩn trong dữ liệu, từ đó cung cấp cơ sở cho việc ra quyết định dựa trên dữ liệu.

Hình 2. Biểu đồ quy trình khai phá tri thức

#### 1.3.2.1. Dữ liệu thô

Dữ liệu thô là điểm xuất phát trong quy trình khai phá dữ liệu, được thu thập từ các nguồn như cơ sở dữ liệu, tệp log, cảm biến, bản ghi giao dịch, hoặc các nguồn trực tuyến. Dữ liệu này có thể gồm cả định dạng có cấu trúc như dữ liệu bảng và không có cấu trúc là văn bản, hình ảnh hoặc video. Thông thường, dữ liệu thô chứa nhiều thông tin không cần thiết hoặc nhiễu. Mục tiêu ở giai đoạn này là thu thập một tập dữ liệu đầy đủ, đại diện, đảm bảo độ chính xác cho kết quả khai phá.

#### 1.3.2.2. Trích chọn dữ liệu

Trích chọn dữ liệu là quá trình sàng lọc và lựa chọn những dữ liệu phù hợp từ tập dữ liệu thô để chuẩn bị cho quá trình khai phá dữ liệu. Mục đích là đảm bảo dữ liệu được chọn có liên quan, đầy đủ và hỗ trợ mục tiêu dự án.

**Chi tiết**:

* **Lọc dữ liệu**: Loại bỏ dữ liệu nhiễu hoặc không liên quan.
* **Chọn lọc thuộc tính**: Xác định các thuộc tính quan trọng cần thiết cho phân tích.

#### 1.3.2.3. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước chuẩn bị và biến đổi dữ liệu ban đầu trước khi áp dụng các phương pháp phân tích hoặc xây dựng mô hình. Mục tiêu chính là đảm bảo tính nhất quán, chất lượng và đồng nhất của dữ liệu.

Quá trình tiền xử lý bao gồm:

* **Làm sạch dữ liệu**: Xử lý giá trị thiếu, ngoại lệ, và nhiễu.
* **Rút gọn dữ liệu**: Giảm kích thước hoặc số chiều của dữ liệu bằng cách chọn thuộc tính hoặc lấy mẫu.
* **Chuẩn hóa dữ liệu**: Đưa dữ liệu về cùng một phạm vi (ví dụ: [0, 1] hoặc [-1, 1]).
* **Rời rạc hóa dữ liệu**: Biến đổi các biến liên tục thành các khoảng rời rạc.

#### 1.3.2.4. Biến đổi dữ liệu

Biến đổi dữ liệu nhằm xác định các đặc trưng quan trọng, tạo biểu diễn mới và có ý nghĩa hơn cho dữ liệu. Quá trình này bao gồm:

* **Chuyển đổi thuộc tính**: Tạo thuộc tính mới từ các thuộc tính hiện có.
* **Chuyển đổi dữ liệu**: Thay đổi cấu trúc dữ liệu để phù hợp với thuật toán khai phá.
* **Biểu đồ hóa dữ liệu**: Minh họa dữ liệu qua đồ thị để nhận biết xu hướng hoặc mô hình.

#### 1.3.2.5. Khai phá dữ liệu (Data Mining)

Đây là bước chính trong quy trình, sử dụng các thuật toán để phát hiện tri thức từ dữ liệu:

* **Phân loại và dự đoán**: Dự đoán lớp hoặc giá trị của dữ liệu mới.
* **Phân cụm**: Nhóm dữ liệu có đặc điểm tương đồng.
* **Tìm luật kết hợp**: Phát hiện mối quan hệ giữa các biến.
* **Phân tích độ liên kết**: Tìm mô hình liên kết giữa các đối tượng.

#### 1.3.2.6. Đánh giá và giải thích

Kết quả khai phá cần được đánh giá để kiểm tra độ chính xác, tính hiệu quả và ý nghĩa thực tế.

**Chi tiết**:

* **Đánh giá mô hình**: Kiểm tra hiệu suất qua bộ dữ liệu kiểm thử.
* **Giải thích mô hình**: Làm rõ cách mô hình hoạt động và lý do cho các kết quả.
* **Trực quan hóa**: Sử dụng biểu đồ để minh họa tri thức.

#### 1.3.2.7. Biểu diễn tri thức

Tri thức cuối cùng cần được trình bày một cách rõ ràng và dễ hiểu để người dùng cuối có thể ứng dụng.

**Phương pháp biểu diễn tri thức**:

* **Đồ thị và biểu đồ**: Minh họa mối quan hệ giữa các đối tượng.
* **Cây quyết định**: Hữu ích trong phân loại và quy tắc quyết định.
* **Bảng biểu**: Tổng hợp và trình bày dữ liệu.
* **Luật kết hợp**: Biểu diễn các mẫu liên hệ dưới dạng văn bản.

**Đánh giá tri thức**:

* **Độ chính xác**: Đảm bảo phản ánh đúng mô hình dữ liệu.
* **Độ hữu ích**: Xác định giá trị thực tế của tri thức.
* **Tính khả thi**: Đánh giá khả năng ứng dụng tri thức vào thực tiễn.

### Các phương pháp khai phá dữ liệu

#### 1.3.3.1. Phân tích hồi quy (Regression Analysis)

Phân tích hồi quy là một phương pháp thống kê hữu ích để khám phá mối liên hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mục tiêu chính của phương pháp này là xây dựng một mô hình dự đoán giá trị của biến phụ thuộc dựa trên các yếu tố tác động từ các biến độc lập.

##### 1.3.3.1.1. Các bước thực hiện phân tích hồi quy

* **Thu thập dữ liệu:** Tập hợp các thông tin liên quan đến biến phụ thuộc và các biến độc lập nhằm phục vụ cho quá trình phân tích và xây dựng mô hình.
* **Xây dựng mô hình:**
  + Lựa chọn loại hồi quy (hồi quy tuyến tính, đa biến, phi tuyến).
  + Xác định rõ biến độc lập và biến phụ thuộc.
* **Tiền xử lý dữ liệu**: Tiến hành làm sạch dữ liệu, xử lý các giá trị bị thiếu, ngoại lệ, hoặc các vấn đề không phù hợp khác để đảm bảo chất lượng dữ liệu đầu vào.
* **Phân tích mô hình**: Đánh giá hiệu quả của mô hình thông qua các phương pháp như phân tích đường hồi quy, kiểm định giả thuyết, và các chỉ số đánh giá như R² hoặc lỗi trung bình bình phương (MSE).
* **Dự đoán và đánh giá:**
  + Sử dụng mô hình để dự đoán giá trị biến phụ thuộc.
  + Đánh giá hiệu quả bằng các chỉ số như RMSE hoặc R-squared.
* **Diễn giải kết quả:** Phân tích ý nghĩa các hệ số hồi quy và so sánh với kiến thức thực tiễn.

##### 1.3.3.1.2. Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính giả định rằng mối quan hệ giữa biến phụ thuộc và các biến độc lập là tuyến tính, tức là giá trị của biến phụ thuộc có thể được biểu diễn dưới dạng một tổ hợp tuyến tính của các biến độc lập với các hệ số tương ứng. Công thức tổng quát:

y = β0 + β1x + ε (1-1)

Trong đó:

* y là giá trị dự đoán của biến phụ thuộc.
* x là giá trị của biến độc lập.
* β0 là hệ số góc (intercept), đại diện cho giá trị y khi x = 0.
* β1 là hệ số hồi quy (slope), đại diện cho tốc độ thay đổi của y theo x.
* ε là sai số ngẫu nhiên (error term), biểu thị sự không chính xác của dự đoán.

##### 1.3.3.1.3. Hồi quy đa biến (Multiple Regression)

Hồi quy đa biến mở rộng hồi quy tuyến tính với nhiều biến độc lập. Công thức tổng quát:

y = β₀ + β₁x₁ + β₂x₂ + ... + βₚxₚ + ɛ (1-2)

Trong đó:

* y là giá trị dự đoán của biến phụ thuộc.
* x₁, x₂, ..., xₚ là giá trị của các biến độc lập.
* β₀ là hệ số góc (intercept), biểu thị giá trị của biến phụ thuộc (y) khi tất cả các biến độc lập bằng 0.
* β₁, β₂, ..., βₚ là hệ số hồi quy (slopes), thể hiện mức độ thay đổi của y ứng với mỗi đơn vị thay đổi của từng biến độc lập, giữ các biến khác không đổi.
* ɛ là sai số ngẫu nhiên (error term), phản ánh sự khác biệt giữa giá trị thực tế và giá trị dự đoán, bao gồm các yếu tố không được mô hình hóa.  
  Phương pháp hồi quy tuyến tính cho phép phân tích và đánh giá tác động đồng thời của nhiều biến độc lập đến biến phụ thuộc, giúp hiểu rõ hơn về mối quan hệ giữa các biến và xây dựng mô hình dự đoán chính xác.

##### 1.3.3.1.4. Hồi quy phi tuyến (Nonlinear Regression)

Hồi quy phi tuyến là một phương pháp thống kê được sử dụng để mô hình hóa mối quan hệ không tuyến tính giữa biến phụ thuộc và một hoặc nhiều biến độc lập. Phương pháp này phù hợp khi dữ liệu không thể được biểu diễn chính xác bằng một đường thẳng. Ví dụ:

y = β0 + β1xa + ɛ (1-3)

Trong đó:

* y là giá trị dự đoán của biến phụ thuộc.
* x là giá trị của biến độc lập.
* β0 là hệ số góc (intercept), đại diện cho giá trị y khi x = 0.
* Β1 là hệ số hồi quy (slope), đại diện cho tốc độ thay đổi của y theo x.
* a là một tham số mũ, cho biết mức độ phi tuyến của mô hình.
* ɛ là sai số ngẫu nhiên (error term), biểu thị sự không chính xác của dự đoán.  
  Phương pháp này hữu ích khi dữ liệu không đáp ứng giả định tuyến tính.

##### 1.3.3.1.5. Ưu điểm và hạn chế

* **Ưu điểm:**
  + Dễ hiểu, dễ áp dụng.
  + Hữu ích trong việc dự đoán và phân tích tác động của các biến độc lập.
* **Hạn chế:**
  + Phụ thuộc vào các giả định thống kê.
  + Dễ bị ảnh hưởng bởi dữ liệu nhiễu hoặc hiện tượng overfitting.

#### 1.3.3.2. Phân tích chuỗi thời gian (Time Series Analysis)

##### 1.3.3.2.1. Tổng quan

Phân tích chuỗi thời gian tập trung vào dữ liệu theo thứ tự thời gian để hiểu rõ xu hướng, mối quan hệ, và sự biến động. Các kỹ thuật chính:

* **Mô hình hóa chuỗi:** Dùng ARIMA, GARCH hoặc Exponential Smoothing.
* **Dự đoán:** Sử dụng dữ liệu lịch sử để ước lượng giá trị tương lai.
* **Phân tích thành phần:** Tách dữ liệu thành xu hướng, chu kỳ, và nhiễu.
* **Phát hiện ngoại lệ:** Xác định các điểm bất thường để cải thiện mô hình dự đoán.

##### 1.3.3.2.2. Ưu điểm và hạn chế

* **Ưu điểm:**
  + Khai thác tốt xu hướng và chu kỳ trong dữ liệu.
  + Hỗ trợ hiệu quả trong dự báo và lập kế hoạch.
* **Hạn chế:**
  + Dữ liệu bị nhiễu hoặc thiếu có thể gây khó khăn.
  + Yêu cầu chuyên môn cao trong chọn và xây dựng mô hình.

#### 1.3.3.3. Phân tích thành phần chính (PCA - Principal Component Analysis)

Phân tích thành phần chính (PCA) là một kỹ thuật giảm chiều dữ liệu được sử dụng rộng rãi trong học máy và thống kê. Phương pháp này chuyển đổi các biến gốc có liên quan thành một tập hợp các thành phần chính không tương quan, giúp giữ lại phần lớn thông tin quan trọng của dữ liệu.

**Quy trình thực hiện PCA:**

* **Chuẩn bị dữ liệu:** Loại bỏ giá trị bị thiếu và chuẩn hóa dữ liệu.
* **Tính ma trận hiệp phương sai:** Đánh giá mối tương quan giữa các biến.
* **Tính toán các thành phần chính:** Dựa trên trị riêng và vector riêng.
* **Lựa chọn thành phần chính:** Giữ lại các thành phần giải thích phần lớn biến động.
* **Chiếu dữ liệu:** Chuyển đổi dữ liệu ban đầu sang không gian mới.

**Ứng dụng của PCA:**

* **Giảm chiều dữ liệu:** Tăng tốc xử lý và giảm độ phức tạp.
* **Phát hiện tương quan:** Làm rõ các mối quan hệ ẩn giữa các biến.
* **Trực quan hóa:** Giúp hiểu rõ cấu trúc dữ liệu qua biểu diễn 2D hoặc 3D.

PCA là một công cụ mạnh mẽ, đặc biệt hữu ích trong việc giảm chiều dữ liệu, phân tích, mô hình hóa và trực quan hóa. Phương pháp này không chỉ giúp tối ưu hóa hiệu quả xử lý mà còn làm sáng tỏ các mẫu ẩn trong dữ liệu phức tạp, hỗ trợ đưa ra quyết định tốt hơn.

## Ngôn ngữ sử dụng

### 1.4.1. Tổng quan về ngôn ngữ lập trình Python

Python là một ngôn ngữ lập trình thông dịch, hướng đối tượng và bậc cao, được phát triển bởi Guido van Rossum và ra mắt vào năm 1991. Với thiết kế đơn giản, dễ sử dụng và thân thiện, Python đã thu hút cả người mới học lập trình lẫn các nhà phát triển chuyên nghiệp. Tên "Python" được lấy cảm hứng từ nhóm hài kịch Monty Python của Anh, thể hiện mục tiêu tạo ra một ngôn ngữ lập trình thú vị và dễ tiếp cận.

Python hiện là một trong những ngôn ngữ lập trình phổ biến nhất trên thế giới, đặc biệt được ưa chuộng trong các lĩnh vực như:

* **Phát triển web:** Xây dựng các ứng dụng web phía máy chủ.
* **Phát triển phần mềm:** Tạo các ứng dụng phức tạp.
* **Phân tích dữ liệu và học máy:** Nhờ vào các thư viện mạnh mẽ và công cụ dễ sử dụng.
* **Tự động hóa:** Viết script để tự động hóa các tác vụ thường ngày.

Python nổi bật nhờ:

* **Cấu trúc dữ liệu tích hợp cao cấp,** kiểu dữ liệu động, và liên kết động.
* **Cộng đồng mã nguồn mở:** Cho phép hàng nghìn lập trình viên liên tục xây dựng thư viện và chức năng mở rộng.

Python được sử dụng rộng rãi trong các lĩnh vực học máy (Machine Learning) và trí tuệ nhân tạo (AI) nhờ vào hệ sinh thái phong phú các thư viện và công cụ mạnh mẽ như TensorFlow, PyTorch, Scikit-learn, Keras, và Numpy. Ngôn ngữ này cung cấp khả năng phát triển các mô hình học máy, từ các thuật toán cơ bản đến mạng nơ-ron sâu, đồng thời hỗ trợ tối ưu hóa và triển khai trong môi trường thực tế. Cú pháp của Python đơn giản, dễ đọc và dễ triển khai, giúp giảm độ phức tạp trong lập trình. Ngôn ngữ này được trang bị các thư viện như:

* **NumPy:** Xử lý mảng và các phép toán số học.
* **Pandas:** Phân tích và thao tác dữ liệu.
* **Scikit-learn:** Triển khai các thuật toán học máy.
* **TensorFlow và PyTorch:** Xây dựng và huấn luyện mạng nơ-ron trong học sâu.

Python còn sở hữu một cộng đồng sôi nổi với hàng loạt tài liệu, diễn đàn và giải pháp giúp các nhà phát triển giải quyết vấn đề một cách nhanh chóng. Sự kết hợp giữa tính dễ học, linh hoạt và mạnh mẽ đã biến Python trở thành ngôn ngữ hàng đầu cho lập trình viên và nhà nghiên cứu, đặc biệt trong lĩnh vực AI và học máy.

### 1.4.2. Ưu, nhược điểm của ngôn ngữ lập trình Python

Python là một ngôn ngữ lập trình mạnh mẽ với nhiều ưu điểm vượt trội, tuy nhiên cũng có một số nhược điểm cần lưu ý.

**Ưu điểm của Python:**

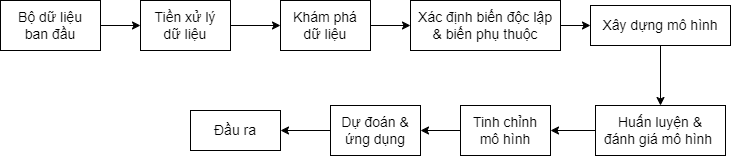
* **Dễ học:**
  + Cú pháp đơn giản, rõ ràng, gần gũi với ngôn ngữ tự nhiên.
  + Thân thiện với người mới bắt đầu.
* **Đa năng:**
  + Phù hợp với nhiều ứng dụng khác nhau, từ phát triển web, phân tích dữ liệu, đến các lĩnh vực như học máy, trí tuệ nhân tạo và tự động hóa.
* **Hệ sinh thái thư viện phong phú:**
  + Sự hỗ trợ mạnh mẽ từ các thư viện như NumPy, Pandas, Matplotlib, TensorFlow và PyTorch, giúp người dùng thực hiện các tác vụ học máy, học sâu và phân tích dữ liệu một cách hiệu quả.
* **Khả năng tích hợp cao:**
  + Python dễ dàng tích hợp với các ngôn ngữ và công cụ khác như C++, Java, và R.
  + Điều này cho phép kết hợp các thành phần từ nhiều ngôn ngữ trong một dự án.
* **Cộng đồng lớn và hỗ trợ tốt:**
  + Cộng đồng lập trình viên Python luôn sẵn sàng chia sẻ kinh nghiệm và tài liệu.
  + Các diễn đàn, hướng dẫn và khóa học trực tuyến phong phú.

**Nhược điểm của Python:**

* **Tốc độ thực thi:**
  + Python có thể chậm hơn so với các ngôn ngữ như C++ hay Java, đặc biệt là khi thực hiện các tác vụ yêu cầu hiệu suất cao.
  + Ngôn ngữ này không lý tưởng cho các ứng dụng cần xử lý thời gian thực với tốc độ cực kỳ nhanh.
* **Quản lý bộ nhớ:**
  + Python sử dụng cơ chế quản lý bộ nhớ tự động, điều này có thể ảnh hưởng đến hiệu suất khi làm việc với các tập dữ liệu lớn.
* **Giới hạn trên thiết bị di động:**
  + Python không phải lựa chọn tối ưu cho phát triển ứng dụng di động do hiệu suất kém và hạn chế tài nguyên.

# CHƯƠNG 2 MÔ HÌNH KHAI PHÁ DỮ LIỆU ĐỂ DỰ ĐOÁN LƯU LƯỢNG NƯỚC ĐẾN HỒ

## 2.1 Mô hình tổng quát



Hình 3. Sơ đồ mô hình tổng quát

### 2.1.1. Bộ dữ liệu ban đầu

* **Dữ liệu cần thu thập**:
  + **Lưu lượng nước đến**: Giá trị đã được tính toán từ các số liệu quan trắc (ví dụ: lưu lượng mưa, dòng chảy từ các nhánh sông).
  + **Mực nước hồ Tua Srah**: Số liệu quan trắc thực tế tại hồ trong các khoảng thời gian cụ thể.
  + **Lưu lượng xả**: Giá trị đầu ra cần dự đoán, được lấy từ các quan sát thực tế trong hồ.
* **Nguồn dữ liệu**:
  + Hệ thống đo lường tự động tại hồ.
  + Báo cáo thủy văn từ các cơ quan khí tượng thủy văn.

### 2.1.2. Tiền xử lý dữ liệu

* **Xử lý dữ liệu thô**:
  + Kiểm tra và loại bỏ giá trị bị thiếu hoặc không hợp lệ trong dữ liệu (ví dụ: các thời điểm mất số liệu quan trắc).
  + Loại bỏ các giá trị ngoại lệ (outliers), như lưu lượng bất thường không phù hợp với thực tế.
  + Chuẩn hóa dữ liệu để đảm bảo các biến đầu vào có cùng quy mô và giảm độ chênh lệch giữa các đặc tính (normalization hoặc standardization).
* **Biến đổi dữ liệu thời gian**:
  + Tạo các đặc trưng mới dựa trên dữ liệu thời gian, như giá trị trung bình hoặc mức thay đổi của lưu lượng nước đến và mực nước hồ trong các khoảng thời gian khác nhau (ví dụ: 1 giờ, 6 giờ, 24 giờ).

### 2.1.3. Khám phá dữ liệu

* **Thống kê và trực quan hóa**:
  + Biểu đồ phân bố của lưu lượng nước đến, mực nước hồ và lưu lượng xả.
  + Mối tương quan giữa các biến (ví dụ: lưu lượng nước đến có mối quan hệ như thế nào với lưu lượng xả).
  + Phát hiện xu hướng (trend) hoặc tính mùa vụ (seasonality) trong dữ liệu, như sự thay đổi lưu lượng xả theo mùa mưa/lũ.

**2.1.4. Xác định biến độc lập và biến phụ thuộc**

* **Biến độc lập** (đầu vào):
  + Lưu lượng nước đến (đã xử lý).
  + Mực nước hồ.
* **Biến phụ thuộc** (đầu ra):
  + Lưu lượng xả của hồ.

**2.1.5. Xây dựng mô hình**

* Lựa chọn mô hình học sâu để xây dựng:
  + **Mô hình LSTM (Long Short-Term Memory)**: Phù hợp với dữ liệu thời gian liên tục, có khả năng ghi nhớ và khai thác các mối quan hệ dài hạn giữa các biến.
  + **Mô hình LSTM-RF (LSTM kết hợp với Random Forest)**: LSTM dùng để trích xuất các đặc trưng từ chuỗi thời gian, Random Forest dùng để dự đoán giá trị lưu lượng xả.
* **Đầu vào mô hình**:
  + Chuỗi thời gian của các biến độc lập (lưu lượng nước đến, mực nước hồ) trong một khoảng thời gian nhất định.
* **Đầu ra mô hình**:
  + Giá trị lưu lượng nước xả tương ứng.

**2.1.6. Huấn luyện và đánh giá mô hình**

* **Dữ liệu huấn luyện và kiểm tra**:
  + Chia dữ liệu thành tập huấn luyện (training set) và kiểm tra (test set), ví dụ: 80% dữ liệu để huấn luyện và 20% để kiểm tra.
* **Chỉ số đánh giá mô hình**:
  + MAE, MSE, RMSE: Đo lường sai số giữa giá trị dự đoán và thực tế.
  + R²: Đánh giá mức độ phù hợp của mô hình với dữ liệu.
* **Quy trình đánh giá**:
  + Sử dụng **cross-validation** (kiểm định chéo) để kiểm tra khả năng tổng quát hóa của mô hình.

**2.1.7. Tinh chỉnh mô hình**

* Tinh chỉnh các siêu tham số (hyperparameters) của mô hình, ví dụ:
  + Số lượng tầng LSTM và số lượng nơ-ron mỗi tầng.
  + Kích thước batch và tốc độ học (learning rate).
  + Số lượng cây trong mô hình Random Forest (n\_estimators).
* Kết hợp các phương pháp để giảm overfitting:
  + Dropout.
  + Regularization (L1, L2).

**2.1.8. Dự đoán và ứng dụng**

* Áp dụng mô hình đã huấn luyện để dự đoán **lưu lượng nước xả** trong tương lai dựa trên dữ liệu **lưu lượng nước đến** và **mực nước hồ** hiện tại.
* Triển khai mô hình vào hệ thống dự báo vận hành hồ, tích hợp với các hệ thống giám sát tự động.

**2.1.9. Đầu ra**

* Kết quả dự đoán:
  + **Lưu lượng xả** trong các khoảng thời gian tới.
* Trực quan hóa:
  + Biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế.
  + Báo cáo phân tích kết quả cho các cơ quan quản lý hồ.

Dự báo lưu lượng nước xả của hồ Tua Srah có thể được xây dựng bằng cách sử dụng các phương pháp học sâu. Mô hình này sẽ dự đoán lưu lượng nước xả từ hồ Tua Srah dựa trên các biến đầu vào như mực nước hồ và lưu lượng nước đến. Trong đó, lưu lượng nước đến đã được tính toán trước từ các số liệu quan trắc và xử lý dữ liệu ban đầu.

## 2.2 Chi tiết các bước thực hiện

Các bước chính trong mô hình tổng quát bao gồm thu thập dữ liệu, tiền xử lý dữ liệu, khám phá dữ liệu, xác định biến độc lập và biến phụ thuộc, xây dựng mô hình, huấn luyện và đánh giá mô hình, tinh chỉnh mô hình, và cuối cùng là dự đoán và ứng dụng.

* **Bộ dữ liệu ban đầu:**  
  Trong quá trình thu thập dữ liệu, cần thu thập thông tin về lưu lượng nước xả từ hồ Tua Srah, mực nước hồ, và lưu lượng nước đến hồ trong khoảng thời gian quan tâm. Dữ liệu này được lưu trữ dưới dạng bảng dữ liệu, trong đó mỗi hàng đại diện cho một mẫu dữ liệu với các cột tương ứng với các biến độc lập và biến phụ thuộc.
* **Tiền xử lý dữ liệu:**  
  Trong bước này, cần xử lý các vấn đề như:
  + **Dữ liệu thiếu:** Có thể xử lý dữ liệu thiếu bằng cách điền vào các giá trị thay thế, ví dụ như giá trị trung bình, giá trị trung vị, hoặc sử dụng các mô hình dự đoán để ước tính giá trị thiếu.
  + **Giá trị ngoại lệ:** Có thể xác định và xử lý bằng các phương pháp như loại bỏ, thay thế, hoặc biến đổi.
  + **Chuẩn hóa dữ liệu:** Đảm bảo các biến có cùng phạm vi và đơn vị đo bằng cách chuẩn hóa dữ liệu (ví dụ: Min-Max Scaling hoặc Standard Scaling).
* **Khám phá dữ liệu:**  
  Bước này giúp hiểu rõ hơn về các mẫu và quan hệ trong dữ liệu. Sử dụng các phương pháp như:
  + **Phân tích đơn biến:** Thống kê mô tả, biểu đồ phân bố.
  + **Phân tích đa biến:** Ma trận tương quan, biểu đồ tương quan giữa mực nước hồ, lưu lượng nước đến, và lưu lượng nước xả để tìm ra mối quan hệ tiềm năng.
* **Xác định biến độc lập và biến phụ thuộc:**
  + **Biến độc lập (input):** Mực nước hồ, lưu lượng nước đến, và các yếu tố thời gian (như ngày, giờ).
  + **Biến phụ thuộc (output):** Lưu lượng nước xả từ hồ Tua Srah.
* **Xây dựng mô hình:**  
  Sử dụng mạng LSTM để xây dựng mô hình dự báo. LSTM là mạng nơ-ron hồi tiếp (RNN) có khả năng ghi nhớ và xử lý dữ liệu chuỗi thời gian, rất phù hợp với bài toán này. Kiến trúc mô hình có thể bao gồm:
  + Một hoặc nhiều lớp LSTM.
  + Các lớp Dropout để tránh overfitting.
  + Các lớp Dense ở đầu ra để dự đoán lưu lượng nước xả.
* **Huấn luyện và đánh giá mô hình:**
  + **Huấn luyện:** Dữ liệu huấn luyện được chia thành các tập chuỗi thời gian, từ đó điều chỉnh các tham số của mô hình sao cho tối ưu.
  + **Đánh giá:** Hiệu suất của mô hình được đánh giá trên dữ liệu kiểm tra thông qua các chỉ số như MAE, RMSE hoặc R², nhằm đảm bảo tính khả thi và độ tin cậy của mô hình.
* **Tinh chỉnh mô hình:**
  + Tinh chỉnh các siêu tham số như số lượng lớp LSTM, số nút trong mỗi lớp, learning rate, batch size.
  + Thử nghiệm các phương pháp tối ưu hóa như tăng cường dữ liệu hoặc kỹ thuật ensemble để cải thiện hiệu suất mô hình.
* **Dự đoán và ứng dụng:**
  + **Dự đoán:** Sử dụng mô hình LSTM để dự báo lưu lượng nước xả từ hồ Tua Srah dựa trên dữ liệu mực nước và lưu lượng nước đến.
  + **Ứng dụng thực tế:** Kết quả dự báo được áp dụng trong việc quản lý và điều tiết nước của hồ, đảm bảo an toàn và ổn định hệ thống thủy lợi và cấp nước.

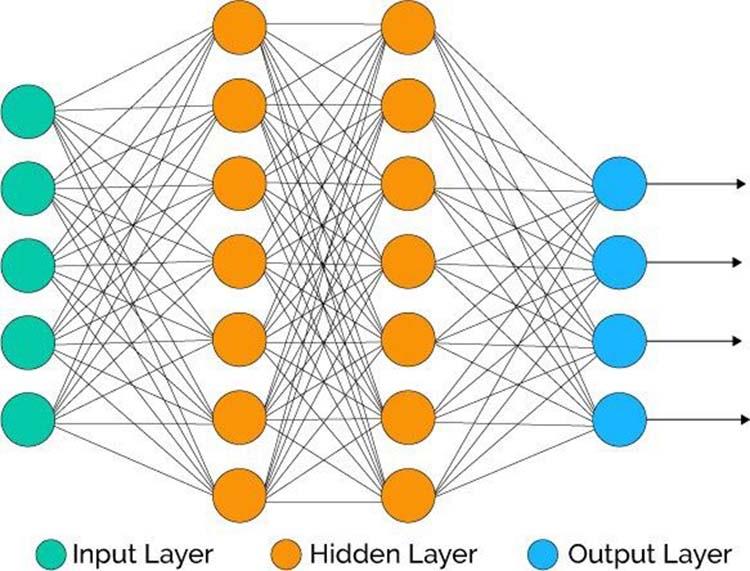
**Đầu ra:**  
Đầu ra của mô hình là lưu lượng nước xả từ hồ Tua Srah trong các thời điểm dự báo.

## 2.3 Mạng Nơ-ron( Neural Network)

### 2.3.1 Giới thiệu về mạng nơ-ron (Neural Network)

[Neural Network](https://bizflycloud.vn/tin-tuc/neural-network-la-gi-20220512174329187.htm) hay mạng nơ-ron thần kinh nhân tạo là một mô hình toán học phức tạp được đưa ra để xử lý thông tin, giải quyết các vấn đề phổ biến trong lĩnh vực học máy, trí tuệ nhân tạo AI, deep learning và tìm kiếm tất cả các mối quan hệ cơ bản trong một tập hợp các dữ liệu. Tất cả điều này được Neural Network thực hiện thông qua việc bắt chước các hoạt động của của tế bào thần kinh và khớp thần kinh từ bộ não của con người.

Hiểu đơn giản, Neural Network tương tự như bộ não của con người, nó sẽ kết nối các nút đơn giản và các nút này sẽ tạo thành một mạng lưới. Do đó, nó được xtác giả là hệ thống của các tế bào thần kinh nhân tạo. Trong Mạng Nơ-ron (Neural Network), một loạt các thuật toán được sử dụng để xác định và nhận diện các mối quan hệ trong các tập dữ liệu. Mạng Nơ-ron có khả năng thích ứng linh hoạt với mọi thay đổi ngay từ đầu vào, vì vậy không cần phải thiết kế lại các tiêu chí đầu ra mà vẫn có thể đưa ra kết quả tối ưu.2.3.2 Kiến trúc mạng nơron



Hình 4. Mô hình mạng nơ-ron

Một mạng nơ-ron cơ bản bao gồm ba lớp nơ-ron nhân tạo được liên kết với nhau:

* **Lớp đầu vào (Input Layer)**: Thông tin từ môi trường bên ngoài được đưa vào mạng nơ-ron qua lớp đầu vào. Các nút đầu vào sẽ xử lý dữ liệu, phân tích hoặc phân loại và sau đó chuyển dữ liệu sang lớp tiếp theo.
* **Lớp ẩn (Hidden Layer)**: Dữ liệu từ lớp đầu vào hoặc các lớp ẩn trước đó sẽ được chuyển vào lớp ẩn. Mạng nơ-ron có thể có một hoặc nhiều lớp ẩn. Mỗi lớp ẩn sẽ phân tích dữ liệu đầu ra từ lớp trước, xử lý thêm và truyền dữ liệu sang lớp tiếp theo.
* **Lớp đầu ra (Output Layer)**: Lớp này cung cấp kết quả cuối cùng sau khi tất cả dữ liệu đã được xử lý bởi mạng nơ-ron. Lớp đầu ra có thể có một hoặc nhiều nút. Ví dụ, trong bài toán phân loại nhị phân (có/không), lớp đầu ra sẽ chỉ có một nút, với kết quả là 1 hoặc 0. Trong trường hợp phân loại đa lớp, lớp đầu ra sẽ có nhiều nút, mỗi nút tương ứng với một lớp phân loại.

### 2.3.3 Ứng dụng của Neural Network

Với số lượng các ứng dụng được triển khai ngày càng lớn, Neural Network, máy học và trí tuệ nhân tạo được ứng dụng trong nhiều công việc khác nhau như:

**Nhận dạng chữ viết tay**: Mạng nơ-ron nhân tạo (Neural Network) được sử dụng để chuyển đổi các ký tự viết tay thành các ký tự kỹ thuật số, giúp máy tính dễ dàng nhận dạng và xử lý các ký tự này.

**Dự đoán giao dịch chứng khoán**: Việc theo dõi thị trường chứng khoán thường gặp khó khăn do có nhiều yếu tố tác động đến giá trị cổ phiếu. Neural Network có thể phân tích các yếu tố này và dự đoán giá cổ phiếu hàng ngày, mang lại lợi ích lớn cho các nhà môi giới chứng khoán.

**Vấn đề đi lại của nhân viên bán hàng**: Neural Network giúp tìm ra các con đường tối ưu cho các nhân viên bán hàng di chuyển giữa các thành phố trong cùng một khu vực, từ đó tăng doanh thu và giảm chi phí.

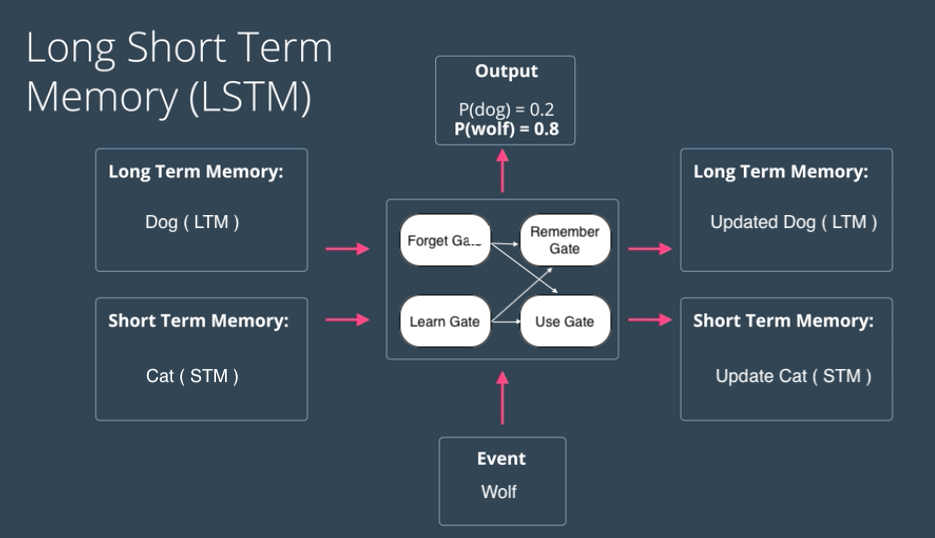
**Nén hình ảnh**: Neural Network có thể sử dụng để nén hình ảnh bằng cách lưu trữ, mã hoá và tái tạo các hình ảnh một cách hiệu quả. Việc nén hình ảnh này giúp tối ưu hóa kích thước dữ liệu, tiết kiệm dung lượng bộ nhớ.

Ngoài các ứng dụng trên, Neural Network còn được sử dụng rộng rãi trong nhiều lĩnh vực và công nghệ khác như thị giác máy tính, trò chơi điện tử, dịch tự động, lọc mạng xã hội, nhận dạng giọng nói và chẩn đoán y tế. Đặc biệt, Neural Network cũng được ứng dụng trong các hoạt động sáng tạo và nghệ thuật như hội hoạ.

## 2.4 Mô hình kết hợp LSTM- RF

### 2.4.1 Mạng LSTM (Long Short-Term Memory):

#### 2.4.1.1 Ưu điểm khi sử dụng LSTM trong bài toán



Hình 5. Mô hình LSTM

* **Xử lý dữ liệu chuỗi thời gian:**  
  Mạng LSTM (Long Short-Term Memory) được thiết kế đặc biệt để xử lý và học từ dữ liệu chuỗi thời gian, chẳng hạn như lưu lượng đến và mực nước hồ, thông qua khả năng ghi nhớ thông tin từ các bước thời gian trước đó. Điều này giúp mô hình phát hiện được các mẫu (patterns) dài hạn và ngắn hạn trong dữ liệu.
* **Dự đoán lưu lượng xả:**  
  LSTM được sử dụng để dự đoán lưu lượng xả dựa trên dữ liệu chuỗi mực nước hồ và lưu lượng đến trong quá khứ, đảm bảo rằng dự đoán phản ánh các biến động thời gian thực tế.

#### 2.4.1.2 Quy trình hoạt động

**Khởi tạo các tham số:**  
Tất cả trọng số (W) và bias (b) của LSTM được khởi tạo ngẫu nhiên. Adam cũng khởi tạo hai giá trị:

* **mt​:** Giá trị trung bình luỹ thừa bậc nhất của gradient.
* **vt​:** Giá trị trung bình luỹ thừa bậc hai của gradient.

**Tính toán hàm mất mát (Loss Function):**

LSTM dự đoán đầu ra ​ và so sánh với giá trị thực tế .

Hàm mất mát, chẳng hạn MSE (Mean Squared Error), được tính:

Loss = (2-1)

**Tính gradient của các tham số:**  
Sử dụng **Backpropagation Through Time (BPTT)** để tính gradient của hàm mất mát theo từng tham số trong mạng LSTM.

**Cập nhật các tham số với Adam:**

**Cập nhật giá trị trung bình:**

= \*

= \*

Trong đó:

* là gradient tại thời điểm ttt.
* là các hệ số giảm dần (thường là ,

**Chỉnh sửa các giá trị bias:**

là các giá trị được chỉnh sửa để tránh độ chệch ban đầu.

**Cập nhật trọng số:**

Trong đó:

* là trọng số tại thời điểm ttt.
* là tốc độ học (learning rate).
* là giá trị nhỏ để tránh chia cho 0 (thường là =).

**Lặp lại:**  
Quy trình được lặp lại qua nhiều epoch để tối ưu hóa các tham số của LSTM, cho đến khi hàm mất mát hội tụ hoặc đạt đến số lần huấn luyện tối đa.

### 2.4.2 Giải thuật Random Forest ( RF):

#### 2.4.2.1 Random Forest là gì?

* **Random Forest** là một mô hình học tập tập hợp (**ensemble learning**) bao gồm nhiều cây quyết định độc lập.
* Kết quả cuối cùng của Random Forest được đưa ra bằng cách:
* **Bài toán phân loại (classification):** Bỏ phiếu đa số từ các cây.
* **Bài toán hồi quy (regression):** Lấy trung bình các đầu ra từ các cây.

#### 2.4.2.2 Ưu điểm của giải thuật random forest

* **Tăng cường độ chính xác:**  
  Random Forest là một mô hình học máy mạnh mẽ, sử dụng sự kết hợp của nhiều cây quyết định (Decision Trees) để cải thiện độ chính xác của các dự đoán. Mỗi cây quyết định học một phần dữ liệu khác nhau, và kết quả cuối cùng được tổng hợp để đưa ra dự đoán chính xác hơn.
* **Trích xuất các đặc trưng phi tuyến:**  
  Random Forest có khả năng xử lý các mối quan hệ phi tuyến trong dữ liệu, giúp mô hình hiểu được các tương tác phức tạp giữa lưu lượng đến, mực nước hồ, và lưu lượng xả. Điều này đặc biệt hữu ích trong việc phân tích dữ liệu thực tế có nhiều yếu tố ảnh hưởng.

#### 2.4.2.3 Cách thức hoạt động

**Bước 1: Xây dựng từng cây quyết định**

* **Tạo các tập dữ liệu con (Bootstrap Sampling):**
  + Từ tập dữ liệu huấn luyện ban đầu, Random Forest áp dụng phương pháp **Bootstrap Sampling** để tạo ra các tập dữ liệu con, bằng cách lấy mẫu ngẫu nhiên với sự thay thế từ dữ liệu gốc.
  + Từ tập dữ liệu huấn luyện ban đầu, Random Forest sử dụng phương pháp **bagging** (Bootstrap Aggregating), trong đó nhiều cây quyết định được xây dựng trên các mẫu dữ liệu con khác nhau. Sau đó, kết quả của các cây quyết định này được kết hợp lại để đưa ra dự đoán cuối cùng, giúp tăng cường độ chính xác và giảm thiểu sự dao động.
  + Mỗi cây quyết định trong Random Forest được xây dựng trên một tập dữ liệu con, với các đặc trưng được chọn lọc ngẫu nhiên, giúp tăng tính đa dạng của các cây và cải thiện khả năng tổng quát của mô hình.
* **Chọn ngẫu nhiên các đặc trưng (Random Feature Selection):**
  + Tại mỗi nút của cây, chỉ một tập con nhỏ các đặc trưng được xem xét để tìm ra điều kiện chia tách (split).
  + Điều này giúp tăng tính đa dạng giữa các cây và giảm khả năng overfitting.
* **Xây dựng cây quyết định:**
  + Mỗi cây được xây dựng trên tập dữ liệu con với một tập hợp các đặc trưng chọn lọc.
  + Quá trình dừng lại khi đạt đến số lá tối thiểu hoặc khi không thể cải thiện chia tách.

**Bước 2: Tổng hợp kết quả từ các cây**

**Bài toán hồi quy:** Lấy trung bình các dự đoán từ tất cả các cây.

**Trong đó:**

* + ​: Kết quả dự đoán cuối cùng.
  + ​: Kết quả dự đoán từ cây thứ iii.
  + N: Tổng số cây trong Random Forest.

### 2.4.2 Mô hình LSTM-RF

Mô hình LSTM-RF kết hợp sức mạnh của LSTM trong việc xử lý dữ liệu chuỗi thời gian và trích xuất đặc trưng với Random Forest để đưa ra dự đoán cuối cùng. LSTM học các đặc trưng tiềm ẩn từ dữ liệu chuỗi thời gian, và Random Forest sử dụng các đặc trưng này để thực hiện phân loại hoặc hồi quy. Quá trình huấn luyện điều chỉnh trọng số của LSTM để tối ưu hóa khả năng trích xuất đặc trưng, và mô hình Random Forest học cách kết hợp các đặc trưng đó để đưa ra kết quả tốt nhất.

Quy trình thực hiện được mô tả như sau:

**Bắt đầu:** Quá trình bắt đầu.

**Khởi tạo các tham số LSTM:**

* Các tham số của mô hình LSTM được khởi tạo, bao gồm:
  + Số lượng hidden units (đơn vị ẩn).
  + Số lớp (layers).
  + Tốc độ học (learning rate).
  + Tỷ lệ dropout.
  + Và các tham số khác (nếu có).

**Tập Huấn luyện:**

* Tập dữ liệu huấn luyện được đưa vào mô hình.

**Trích xuất đặc trưng từ LSTM:**

* Mô hình LSTM xử lý dữ liệu chuỗi thời gian từ tập huấn luyện.
* **Sử dụng hidden states cuối cùng:** Các hidden states (trạng thái ẩn) cuối cùng của LSTM được sử dụng làm đặc trưng đại diện cho dữ liệu đầu vào. Đây là thông tin đã được mô hình học và mã hóa từ chuỗi thời gian.

**Random Forest:**

* Các đặc trưng được trích xuất từ LSTM (hidden states) được đưa vào mô hình Random Forest.
* Mô hình Random Forest sử dụng các đặc trưng này để huấn luyện.

**Huấn luyện mô hình LSTM:**

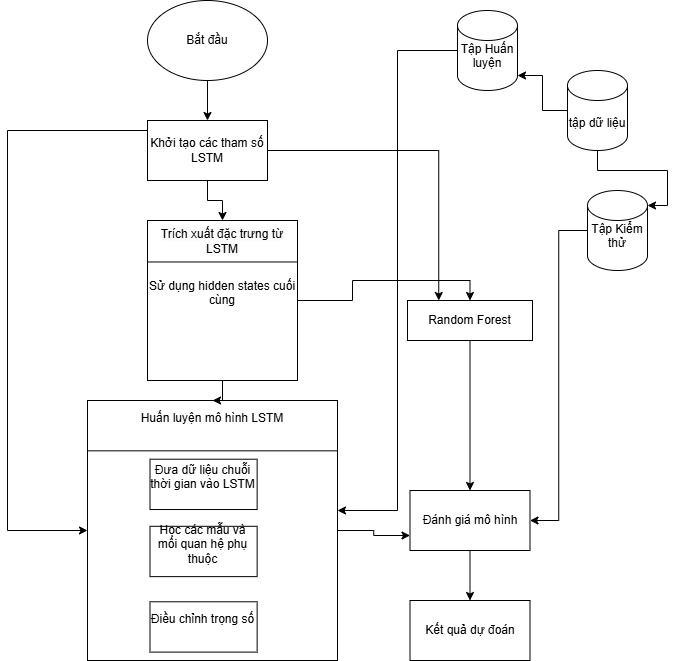
* **Đưa dữ liệu chuỗi thời gian vào LSTM:** Dữ liệu chuỗi thời gian từ tập huấn luyện được đưa vào mô hình LSTM.
* **Học các mẫu và mối quan hệ phụ thuộc:** Mô hình LSTM học các mẫu (patterns) và mối quan hệ phụ thuộc theo thời gian trong dữ liệu, giúp nó nắm bắt được sự thay đổi và xu hướng trong các chuỗi thời gian.
* **Điều chỉnh trọng số:** Các trọng số (weights) của mô hình LSTM được điều chỉnh thông qua quá trình lan truyền ngược (backpropagation) dựa trên các đặc trưng đã học và hàm mất mát (loss function) để tối thiểu hóa sai số dự đoán.

**Đánh giá mô hình:**

* Mô hình (cụ thể là Random Forest sau khi nhận đặc trưng từ LSTM) được đánh giá trên tập kiểm thử (Test set).
* Quá trình đánh giá sử dụng các độ đo (metrics) phù hợp để xác định hiệu suất của mô hình (ví dụ: độ chính xác, RMSE, MAE,...).

**Kết quả dự đoán:**

* Mô hình LSTM-RF đưa ra kết quả dự đoán dựa trên dữ liệu đầu vào (có thể là dữ liệu mới hoặc dữ liệu từ tập kiểm thử).



Hình 6. Sơ đồ LSTM kết hợp RF

## 2.5 Kết hợp LSTM và RF

Mô hình LSTM-RF là sự kết hợp giữa hai mô hình Long Short-Term Memory (LSTM) và Random Forest (RF). Mô hình này tận dụng khả năng xử lý dữ liệu chuỗi thời gian của LSTM để trích xuất các đặc trưng quan trọng, sau đó sử dụng các đặc trưng này làm đầu vào cho mô hình Random Forest để đưa ra dự đoán cuối cùng.

### 2.5.1 Quá trình dự báo

* **Trích xuất đặc trưng từ LSTM**
* **Huấn luyện LSTM:** Mô hình LSTM được huấn luyện trên dữ liệu chuỗi thời gian. LSTM học các mẫu (patterns) và mối quan hệ phụ thuộc theo thời gian trong dữ liệu, từ đó điều chỉnh các trọng số (weights) nội tại của nó.
* **Đặc trưng từ Hidden States:** Sau khi huấn luyện, các hidden states (trạng thái ẩn) cuối cùng của LSTM (hoặc một cách kết hợp các hidden states) được sử dụng làm đặc trưng đại diện cho dữ liệu đầu vào. Các hidden states này chứa đựng thông tin quan trọng đã được LSTM học và mã hóa từ chuỗi thời gian.
* **Dự báo với Random Forest:**
* **Đầu vào cho Random Forest:** Các đặc trưng được trích xuất từ mô hình LSTM sẽ được sử dụng làm đầu vào cho mô hình Random Forest để thực hiện dự đoán hoặc phân loại.
* **Huấn luyện Random Forest:** Mô hình Random Forest được huấn luyện trên tập dữ liệu mới với đầu vào là các đặc trưng từ LSTM và đầu ra là giá trị cần dự đoán.
* **Dự đoán cuối cùng:** Random Forest sử dụng các đặc trưng được trích xuất từ LSTM để đưa ra dự đoán cuối cùng. Mô hình này kết hợp các dự đoán từ nhiều cây quyết định, giúp nâng cao độ chính xác và giảm thiểu hiện tượng overfitting.

### 2.5.2 Ưu điểm của mô hình LSTM-RF

* **Kết hợp sức mạnh của LSTM và RF:** LSTM hiệu quả trong việc xử lý dữ liệu chuỗi thời gian và trích xuất đặc trưng, trong khi Random Forest mạnh mẽ trong việc phân loại và hồi quy dựa trên các đặc trưng đầu vào.
* **Cải thiện hiệu suất:** Việc sử dụng các đặc trưng từ LSTM làm đầu vào cho Random Forest giúp cải thiện độ chính xác của dự đoán so với việc sử dụng trực tiếp dữ liệu thô.
* **Giảm chiều dữ liệu:** LSTM trích xuất và nén thông tin quan trọng từ chuỗi thời gian thành các hidden states, giúp giảm chiều dữ liệu đầu vào cho Random Forest, từ đó có thể tăng tốc độ huấn luyện và cải thiện hiệu suất.

## 2.6 Nhận xét và đánh giá

**Nhận xét và đánh giá mô hình LSTM-RF**

**Ưu điểm:**

* **Xử lý hiệu quả dữ liệu chuỗi thời gian:** LSTM là một loại mạng nơ-ron hồi tiếp (RNN) được thiết kế đặc biệt để xử lý và học từ dữ liệu chuỗi thời gian, nhờ khả năng ghi nhớ thông tin dài hạn và ngắn hạn. Nó có khả năng nắm bắt các mối quan hệ phụ thuộc dài hạn trong chuỗi, điều mà các mô hình truyền thống như Linear Regression và SVR gặp khó khăn.
* **Trích xuất đặc trưng mạnh mẽ:** LSTM tự động học và trích xuất các đặc trưng quan trọng từ dữ liệu chuỗi thời gian thông qua các hidden states. Việc sử dụng hidden states cuối cùng làm đặc trưng giúp nắm bắt được thông tin cô đọng và hữu ích nhất từ chuỗi.
* **Kết hợp sức mạnh của LSTM và Random Forest:**
  + **LSTM:** Cung cấp khả năng trích xuất đặc trưng hiệu quả từ dữ liệu chuỗi thời gian.
  + **Random Forest:** Cung cấp khả năng phân loại/hồi quy mạnh mẽ dựa trên các đặc trưng được trích xuất, đồng thời giảm thiểu overfitting nhờ cơ chế kết hợp nhiều cây quyết định.
* **Giảm chiều dữ liệu:** LSTM giúp giảm chiều dữ liệu đầu vào cho Random Forest bằng cách nén thông tin chuỗi thời gian thành các vector đặc trưng có kích thước nhỏ hơn.
* **Tăng độ chính xác:** Việc kết hợp LSTM và RF thường mang lại độ chính xác cao hơn so với việc sử dụng riêng lẻ từng mô hình, đặc biệt là trên các tập dữ liệu chuỗi thời gian phức tạp.

**Hạn chế:**

* **Độ phức tạp:** Mô hình LSTM-RF phức tạp hơn so với các mô hình đơn giản như Linear Regression. Việc huấn luyện và tối ưu hóa mô hình yêu cầu nhiều thời gian và tài nguyên tính toán hơn.
* **Khó diễn giải:** Giống như hầu hết các mô hình học sâu, LSTM khó diễn giải hơn so với các mô hình tuyến tính. Việc hiểu rõ cách mô hình đưa ra quyết định không phải lúc nào cũng đơn giản.
* **Lựa chọn siêu tham số:** Cả LSTM và Random Forest đều có nhiều siêu tham số cần điều chỉnh. Việc tìm ra bộ siêu tham số tối ưu đòi hỏi thử nghiệm và có thể tốn kém.
* **Phụ thuộc vào chất lượng dữ liệu:** Mô hình LSTM-RF, đặc biệt là phần LSTM, rất nhạy cảm với chất lượng dữ liệu đầu vào. Dữ liệu bị nhiễu hoặc thiếu sót có thể làm giảm hiệu suất của mô hình.

**Kết luận:**

Mô hình LSTM-RF là một lựa chọn mạnh mẽ cho các bài toán dự đoán trên dữ liệu chuỗi thời gian, đặc biệt là khi dữ liệu có các mối quan hệ phi tuyến tính và phụ thuộc dài hạn. Tuy nhiên, cần lưu ý đến độ phức tạp, khó khăn trong việc diễn giải và yêu cầu cao về tài nguyên tính toán của mô hình. Việc chọn lựa mô hình phù hợp nhất vẫn phụ thuộc vào đặc điểm của từng bài toán và tập dữ liệu cụ thể.

# CHƯƠNG 3 CÀI ĐẶT THỰC NGHIỆM

## 3.1 Mô tả dữ liệu

* **Mực nước hồ (m):**
* Giá trị trung bình (478.38 m) cho biết mực nước hồ bình quân trong thời gian theo dõi, cho thấy hồ thường ở mức nước cao so với mực nước tối thiểu.
* Độ lệch chuẩn (7.61 m) chỉ ra mức độ dao động của mực nước. Độ lệch thấp cho thấy mực nước hồ tương đối ổn định.
* Giá trị nhỏ nhất (464.10 m) cho thấy mức nước thấp nhất từng ghi nhận, có thể là trong mùa khô hoặc khi mức xả lớn.
* Giá trị lớn nhất (487.49 m) là mức nước cao nhất, có thể xảy ra vào mùa mưa hoặc khi có lượng nước đổ về lớn.
* **Lưu lượng đến hồ (m³/s):**
* Giá trị trung bình (96.33 m³/s) cho biết mức lưu lượng nước trung bình chảy vào hồ, phản ánh sức mạnh của nguồn nước cung cấp cho hồ. Giá trị này có thể thay đổi theo mùa (mùa khô và mùa mưa).
* Độ lệch chuẩn cao (78.53 m³/s) cho thấy sự biến động lớn trong lượng nước đến hồ, có thể do các yếu tố thời tiết như lũ lụt, mưa lớn, hoặc hạn hán.
* Giá trị lớn nhất (1,012.89 m³/s) chỉ ra mức nước đổ về hồ cao nhất, có thể là trong những trận lũ lớn, điều này rất quan trọng trong việc điều hành xả lũ và đảm bảo an toàn cho hồ.
* **Tổng lưu lượng xả (m³/s) [Thực tế]:**
* Giá trị trung bình (95.94 m³/s) cho thấy mức nước trung bình xả ra từ hồ, tương đối cân bằng với lưu lượng đến hồ (96.33 m³/s). Điều này cho thấy hồ vận hành ổn định và xả lượng nước phù hợp với lượng nước đến.
* Giá trị lớn nhất (503.00 m³/s) có thể liên quan đến các giai đoạn xả lũ hoặc điều chỉnh mực nước hồ để tránh ngập lụt.
* Độ lệch chuẩn (51.63 m³/s) thể hiện sự biến đổi trong lưu lượng xả, quan trọng khi đánh giá cách hồ ứng phó với sự thay đổi về lưu lượng đến

## 3.2 Tiền xử lý dữ liệu

### 3.2.1 Tiền xử lý dữ liệu

**Bước 1: Chuẩn hóa dữ liệu thời gian**

* Dữ liệu ban đầu có cột Ngày ở định dạng chuỗi. Để dễ dàng xử lý và khai thác thông tin thời gian, cột này được chuyển đổi sang định dạng datetime. Điều này cho phép thực hiện các thao tác xử lý thời gian như lọc dữ liệu theo ngày, tháng hoặc tính toán các giá trị trung bình theo khoảng thời gian.

**Bước 2: Kiểm tra và xử lý dữ liệu thiếu**

* Trong dữ liệu, một số giá trị có thể bị thiếu do các nguyên nhân khác nhau. Những giá trị bị thiếu này được kiểm tra và điền vào bằng giá trị trung bình của các mẫu trong ngày tương ứng. Phương pháp này đảm bảo tính liên tục của dữ liệu mà không làm mất đi xu hướng hoặc đặc điểm quan trọng.

**Bước 3: Loại bỏ dữ liệu không hợp lệ**

* Các mẫu dữ liệu có giá trị bất hợp lý (ví dụ: giá trị âm hoặc vượt mức giới hạn vật lý) được loại bỏ. Việc này giúp giảm nhiễu và đảm bảo rằng dữ liệu đầu vào chính xác và đáng tin cậy.

**Bước 4: Chuẩn hóa giá trị**

* Các thuộc tính như mực nước hồ, lưu lượng đến, và lưu lượng xả được chuẩn hóa để đảm bảo tất cả giá trị nằm trong cùng một phạm vi. Điều này giúp cải thiện hiệu quả của các thuật toán học máy, đồng thời giảm thiểu tác động của sự chênh lệch giá trị giữa các thuộc tính.

## 3.2.2 Phân tích tương quan giữa các thuộc tính

* Sau khi tiền xử lý, dữ liệu được phân tích để đánh giá mối quan hệ giữa các thuộc tính. Ma trận tương quan được sử dụng để xác định mức độ tương quan tuyến tính giữa các thuộc tính như mực nước hồ, lưu lượng đến, và lưu lượng xả.
* Kết quả từ ma trận tương quan cho thấy rằng hai thuộc tính lưu lượng đến và lưu lượng xả có mối quan hệ tương quan tuyến tính đáng kể. Trong khi đó, mối quan hệ giữa các thuộc tính khác như mực nước hồ với các thuộc tính còn lại có mức độ tương quan yếu hơn.

### 3.2.3 Phân tích thành phần chính (PCA)

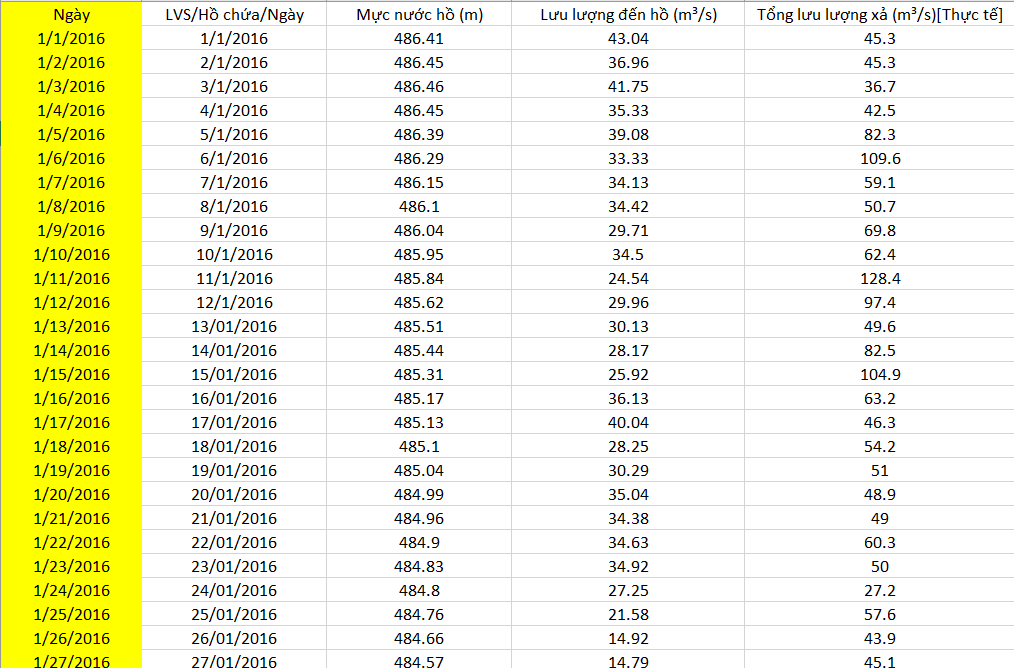
* Phân tích thành phần chính (PCA) được áp dụng để giảm chiều dữ liệu và xem xét sự biến thiên của các thuộc tính, giúp xác định các yếu tố chính đóng góp vào sự biến đổi của dữ liệu, từ đó rút gọn số lượng thuộc tính mà vẫn giữ được phần lớn thông tin quan trọng. PCA giúp xác định xem liệu tất cả các thuộc tính có thực sự cần thiết hay không, dựa trên mức độ đóng góp vào tổng phương sai của dữ liệu.
* Kết quả từ PCA chỉ ra rằng giữ lại cả ba thuộc tính sẽ giữ được nhiều thông tin hơn so với giảm chiều dữ liệu xuống còn hai thành phần chính. Do đó, việc sử dụng đầy đủ cả ba thuộc tính được khuyến nghị.

### 3.2.4 Đánh giá hiệu suất dự báo

* Để đánh giá hiệu suất dự báo, dữ liệu được đưa vào mô hình dự báo với hai trường hợp:
  1. Sử dụng đầy đủ ba thuộc tính (mực nước hồ, lưu lượng đến, lưu lượng xả).
  2. Sử dụng hai thuộc tính chính từ kết quả PCA.
* Hiệu suất của mô hình được đánh giá bằng chỉ số Mean Squared Error (MSE). Kết quả cho thấy rằng sử dụng đầy đủ ba thuộc tính mang lại độ chính xác dự báo tốt hơn so với giảm chiều dữ liệu xuống còn hai thuộc tính.

### 3.2.5 Kết luận và đề xuất

* Phân tích tương quan và PCA cho thấy rằng ba thuộc tính mực nước hồ, lưu lượng đến, và lưu lượng xả đều đóng vai trò quan trọng trong dự báo. Việc giảm chiều dữ liệu không mang lại lợi ích đáng kể và làm mất một phần thông tin quan trọng.
* Do đó, việc sử dụng đầy đủ ba thuộc tính là lựa chọn phù hợp nhất để đảm bảo hiệu suất dự báo tối ưu.

****

Hình 7. Dữ liệu demo

## 3.3 Độ đo đánh giá

Để đánh giá tính hiệu quả của mô hình LSTM-RF trong việc dự báo lưu lượng nước về hồ thủy điện Tua Srah, các độ đo sau đây sẽ được sử dụng:

**3.3.1 Sai số trung bình tuyệt đối (Mean Absolute Error - MAE):**

* **Định nghĩa:** **MAE (Mean Absolute Error)** là giá trị trung bình của các sai số tuyệt đối giữa giá trị dự đoán và giá trị thực tế.
* **Công thức:**
* MAE = (i=1,....,N) (2-3)

Trong đó:

* yi: Giá trị thực tế của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i.
* ŷi: Giá trị dự đoán của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i bởi mô hình LSTM-RF.
* N: Tổng số mẫu dự đoán.

**Ý nghĩa:** MAE cho biết sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế của lưu lượng xả, giúp đánh giá độ chính xác của mô hình dự báo. MAE càng nhỏ, mô hình dự đoán càng chính xác.

**Ưu điểm:**

* Dễ hiểu và diễn giải, có cùng đơn vị với biến mục tiêu (lưu lượng xả).
* Chống chịu tốt với ngoại lệ (outliers) do sử dụng giá trị tuyệt đối.

**Nhược điểm:**

* Không có đạo hàm tại mọi điểm, gây khó khăn cho một số thuật toán tối ưu.
* Xem xét mọi sai số cá nhân như nhau, không phân biệt sai số lớn hay nhỏ.

**3.3.2 Sai số bình phương trung bình (Mean Squared Error - MSE):**

* **Định nghĩa:** **MSE (Mean Squared Error)** là giá trị trung bình của bình phương sai số giữa giá trị dự đoán và giá trị thực tế.
* **Công thức:**

MSE = (2-4)

Trong đó:

* yi: Giá trị thực tế của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i.
* ŷi: Giá trị dự đoán của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i bởi mô hình LSTM-RF.
* N: Tổng số mẫu dự đoán.

**Ý nghĩa:** **MSE** đo lường mức độ chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế của lưu lượng xả, đồng thời đặt trọng số cao hơn cho các sai số lớn, giúp nhấn mạnh các sai số lớn hơn trong quá trình đánh giá. MSE càng nhỏ, mô hình càng chính xác.

**Ưu điểm:**

* Có đạo hàm tại mọi điểm, thuận lợi cho các thuật toán tối ưu.

**Nhược điểm:**

* Đơn vị là bình phương của đơn vị biến mục tiêu (lưu lượng xả), khó diễn giải hơn.
* Nhạy cảm với ngoại lệ (outliers) do sử dụng bình phương sai số.

**3.3.3 Căn bậc hai của sai số bình phương trung bình (Root Mean Squared Error - RMSE):**

* **Định nghĩa:** RMSE là căn bậc hai của MSE.
* **Công thức:**
* RMSE = (2-5)

**Ý nghĩa:** RMSE có cùng ý nghĩa với MSE nhưng có cùng đơn vị với biến mục tiêu (lưu lượng xả), giúp dễ dàng diễn giải hơn. RMSE càng nhỏ, mô hình càng chính xác.

**Ưu điểm:**

Dễ hiểu và diễn giải, có cùng đơn vị với biến mục tiêu (lưu lượng xả).

**Nhược điểm:**

* Vẫn nhạy cảm với ngoại lệ (outliers) dù đã giảm bớt so với MSE.
* Phải thêm bước tính căn bậc hai.

**3.3.4 Hệ số xác định (Coefficient of Determination - R²):**

* **Định nghĩa:** **R²** đo lường tỷ lệ phần trăm phương sai của biến mục tiêu (lưu lượng xả hồ thủy điện Tua Srah) được mô hình giải thích và dự đoán, giúp đánh giá mức độ phù hợp của mô hình trong việc dự báo lưu lượng xả dựa trên các yếu tố đầu vào như mực nước hồ và lưu lượng đến.
* **Công thức:**

R² = (2-6)

Trong đó:

* yi: Giá trị thực tế của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i.
* ŷi: Giá trị dự đoán của **lưu lượng xả** hồ thủy điện Tua Srah tại thời điểm i bởi mô hình LSTM-RF.
* ȳ: Giá trị trung bình của **lưu lượng xả** hồ thủy điện Tua Srah.
* N: Tổng số mẫu dự đoán.

**Ý nghĩa:** R² có giá trị từ 0 đến 1. **R²** càng gần 1, mô hình càng hiệu quả, vì nó giải thích được phần lớn sự biến động của lưu lượng xả. R² gần 0 nghĩa là mô hình không giải thích được nhiều. R² âm nghĩa là mô hình dự đoán kém hơn cả việc sử dụng giá trị trung bình.

**Ưu điểm:**

* Dễ hiểu và diễn giải, cho biết mức độ phù hợp của mô hình với dữ liệu.

**Nhược điểm:**

* Không cung cấp thông tin về sai số dự đoán (như MAE, MSE, RMSE).
* Có thể cao giả tạo khi mô hình bị overfitting.

## 3.4 Môi trường thực nghiệm

**Môi trường và công cụ phát triển**

Đề tài sử dụng môi trường phát triển **Visual Studio Code (VS Code)** được cài đặt **Python** cùng các thư viện hỗ trợ Machine Learning, Deep Learning và phân tích dữ liệu. Mô hình **LSTM-RF** sẽ được triển khai và đánh giá trong môi trường này.

* **Visual Studio Code (VS Code):** Là một trình soạn thảo mã nguồn mở mạnh mẽ và phổ biến, hỗ trợ nhiều ngôn ngữ lập trình, bao gồm cả Python. VS Code cung cấp nhiều tính năng hữu ích như:
  + Gợi ý mã thông minh (IntelliSense).
  + Gỡ lỗi (debugging).
  + Tích hợp Git.
  + Hỗ trợ nhiều phần mở rộng (extensions) hữu ích, bao gồm các extension cho Python và Jupyter Notebook.
* **Python:** Là ngôn ngữ lập trình chính được sử dụng trong đề tài. Python nổi bật với cú pháp đơn giản, dễ học, và đặc biệt là hệ sinh thái thư viện phong phú dành cho Machine Learning, Deep Learning và khoa học dữ liệu.

**Các thư viện chính:**

* **TensorFlow/Keras:** Là thư viện hàng đầu cho Deep Learning, cung cấp các công cụ để xây dựng, huấn luyện và triển khai các mô hình mạng nơ-ron, bao gồm cả LSTM. Keras là một API cấp cao, giúp đơn giản hóa việc sử dụng TensorFlow.
* **Scikit-learn:** Thư viện phổ biến cho Machine Learning trong Python. Scikit-learn cung cấp các thuật toán Machine Learning truyền thống, như Random Forest, cùng với các công cụ hữu ích cho việc tiền xử lý dữ liệu, đánh giá mô hình và lựa chọn tham số.
* **Pandas:** là một thư viện mạnh mẽ để xử lý và phân tích dữ liệu dạng bảng, cung cấp cấu trúc dữ liệu DataFrame linh hoạt, cho phép thao tác, biến đổi và phân tích dữ liệu một cách hiệu quả.
* **NumPy:** Thư viện nền tảng cho tính toán khoa học trong Python. NumPy cung cấp các đối tượng mảng đa chiều (ndarray) hiệu suất cao và các hàm toán học để thao tác với các mảng này.
* **Matplotlib:** Thư viện hỗ trợ trực quan hóa dữ liệu, cho phép tạo ra các biểu đồ, đồ thị đa dạng và tùy chỉnh cao, giúp dễ dàng theo dõi, phân tích và trình bày kết quả.

**Cụ thể:**

Đề tài sử dụng ngôn ngữ lập trình Python, với sự hỗ trợ của các thư viện TensorFlow/Keras để xây dựng và huấn luyện mô hình LSTM, và Scikit-learn cho mô hình Random Forest. Pandas và NumPy được sử dụng trong các bước tiền xử lý, xử lý và phân tích dữ liệu. Kết quả huấn luyện và dự báo sẽ được trực quan hóa bằng các đồ thị, biểu đồ thông qua thư viện Matplotlib. Toàn bộ quá trình nghiên cứu, phát triển và kiểm thử mô hình sẽ được thực hiện, quản lý trên môi trường VS Code với sự hỗ trợ của Jupyter Notebook.

|  |  |
| --- | --- |
| **Loại thiết bị** | **Chủng loại** |
| CHIP | AMD Ryzen 5 4000 series |
| RAM | 16GB DDR4 3200 MHz |
| SSD | 512GB |
| GPU | Nvidia Geforce 1650 |
| OS | Windows 11 x64 |

## 3.5 Kết quả thực nghiệm

### 3.5.1 Cài đặt môi trường thực nghiệm

**Thực nghiệm mô hình LSTM-RF**

Mô hình LSTM-RF được triển khai và đánh giá trên tập dữ liệu lưu lượng nước của hồ thủy điện Tua Srah. Quá trình thực nghiệm bao gồm các bước: tiền xử lý dữ liệu, chuẩn bị dữ liệu cho LSTM, xây dựng và huấn luyện mô hình LSTM, trích xuất đặc trưng, huấn luyện mô hình Random Forest, và đánh giá hiệu suất của mô hình.

#### 3.5.1.1 Tiền xử lý dữ liệu:

* **Load dữ liệu:** Dữ liệu được load từ file CSV.
* **Chọn thuộc tính:** Dữ liệu đầu vào gồm 2 cột mực\_nước\_hồ và lưu\_lượng\_đến, cột đầu ra (mục tiêu) là lưu\_lượng\_xả.
* **Chuẩn hóa dữ liệu:** Dữ liệu được chuẩn hóa theo thang 0-1 bằng cách sử dụng MinMaxScaler giúp tăng tốc độ hội tụ và cải thiện hiệu suất của mô hình.
* **Chia tập dữ liệu:** Dữ liệu được phân chia thành hai tập, với 80% dành cho tập huấn luyện và 20% dành cho tập kiểm tra.

#### 3.5.1.2. Chuẩn bị dữ liệu cho LSTM:

* **Tạo chuỗi thời gian:** Dữ liệu được biến đổi thành các chuỗi thời gian có độ dài look\_back (số bước thời gian trong quá khứ được sử dụng để dự đoán) thông qua hàm create\_dataset.
* **Reshape dữ liệu:** Dữ liệu đầu vào được điều chỉnh thành định dạng 3D [samples, time steps, features] phù hợp với đầu vào của lớp LSTM.
* **Tăng cường dữ liệu (Data Augmentation):** Áp dụng kỹ thuật tăng cường dữ liệu cho tập huấn luyện bằng cách thêm nhiễu ngẫu nhiên vào các mẫu, giúp mô hình học tốt hơn và giảm thiểu hiện tượng overfitting.

#### 3.5.1.3. Xây dựng và huấn luyện mô hình LSTM:

* **Kiến trúc:** Mô hình LSTM bao gồm 2 lớp LSTM, mỗi lớp có 50 đơn vị (units), theo sau là các lớp Dropout với tỷ lệ 0.2 để tránh overfitting, và cuối cùng là một lớp Dense đầu ra có 1 unit cho dự đoán (regression).
* **Trình tối ưu hóa:** Sử dụng trình tối ưu hóa Adam.
* **Hàm mất mát:** Sử dụng hàm mất mát Mean Squared Error (MSE).
* **Tham số huấn luyện:**
  + Learning Rate: 0.0001.
  + Batch Size: 32.
  + Epochs: 100.
* **Đánh giá:** Mô hình được đánh giá trong quá trình huấn luyện sử dụng tập validation.
* **Kết quả:** Mô hình đã huấn luyện được lưu để sử dụng sau này.

#### 3.5.1.4. Trích xuất đặc trưng từ LSTM:

* **Sử dụng mô hình LSTM đã huấn luyện:** Dùng mô hình LSTM đã huấn luyện để dự đoán trên tập huấn luyện và tập kiểm tra.
* **Đầu ra của lớp LSTM cuối cùng:** Kết quả dự đoán này (đầu ra của lớp LSTM cuối cùng) được sử dụng làm đặc trưng cho mô hình Random Forest.

#### 3.5.1.5. Huấn luyện Random Forest:

* **Mô hình:** Sử dụng mô hình RandomForestRegressor.
* **Tham số:**
  + n\_estimators: 100 (số lượng cây quyết định).
  + random\_state: 42 (đảm bảo kết quả tái lập được).
* **Đầu vào:** Các đặc trưng được trích xuất từ LSTM ở bước 4.
* **Đầu ra:** Giá trị dự đoán lưu\_lượng\_xả.
* **Kết quả:** Mô hình Random Forest đã huấn luyện được lưu lại để sử dụng sau này.

#### 3.5.1.6. Dự đoán và đánh giá mô hình:

Dự đoán: Sử dụng mô hình Random Forest đã huấn luyện để dự đoán *lưu\_lượng\_xả* trên tập kiểm tra.

Độ đo đánh giá: Sử dụng các độ đo sau để đánh giá hiệu suất mô hình:

* R² (R-squared): Hệ số xác định, đánh giá mức độ phù hợp của mô hình với dữ liệu.
* MSE (Mean Squared Error): Sai số bình phương trung bình, đo lường trung bình bình phương của sai số giữa giá trị dự đoán và giá trị thực tế.
* RMSE (Root Mean Squared Error): Căn bậc hai của sai số bình phương trung bình, có cùng đơn vị đo với biến mục tiêu.
* MAE (Mean Absolute Error): Sai số tuyệt đối trung bình, đo lường trung bình của giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực tế.

#### 3.5.1.7. Trực quan hóa kết quả:

* Vẽ biểu đồ so sánh giá trị lưu\_lượng\_xả thực tế và giá trị dự đoán từ mô hình Random Forest trên tập kiểm tra để trực quan hóa kết quả sau đó đánh giá hiệu suất mô hình.

### 3.5.2 Đánh giá

**Tham số mô hình**

Mô hình LSTM-RF bao gồm hai mô hình chính: LSTM và Random Forest. Các tham số quan trọng của từng mô hình được thiết lập như sau:

#### 3.5.2.1. LSTM:

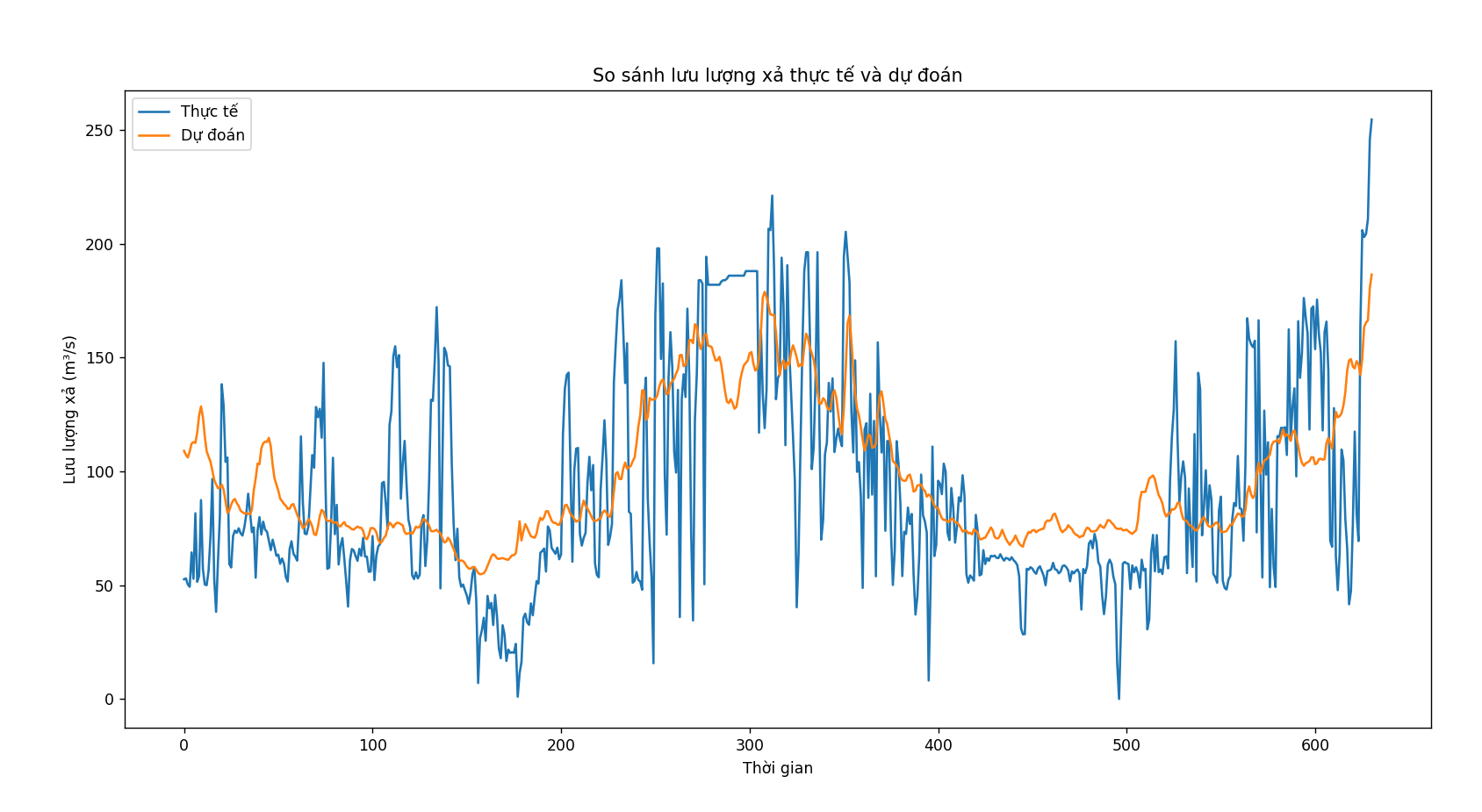
* **Kiến trúc mạng:**
  + **Số lớp LSTM:** 2
  + **Số đơn vị (units) mỗi lớp:** 50
  + **Dropout rate:** 0.2 (cho mỗi lớp Dropout)
  + **Hàm kích hoạt (activation function):** tanh (mặc định cho LSTM)
  + **Lớp đầu ra:** Dense layer với 1 unit, sử dụng hàm kích hoạt tuyến tính (linear) cho bài toán hồi quy.
* **Tham số huấn luyện:**
  + **Learning Rate:** 0.0001 (cho trình tối ưu hóa Adam)
  + **Batch Size:** 32
  + **Epochs:** 100
  + **Trình tối ưu hóa (Optimizer):** Adam
  + **Hàm mất mát (Loss Function):** Mean Squared Error (MSE)
* **Tham số khác:**
  + **Look back:** 3 (số bước thời gian trong quá khứ được sử dụng để dự đoán)

#### 3.5.2.2. Random Forest:

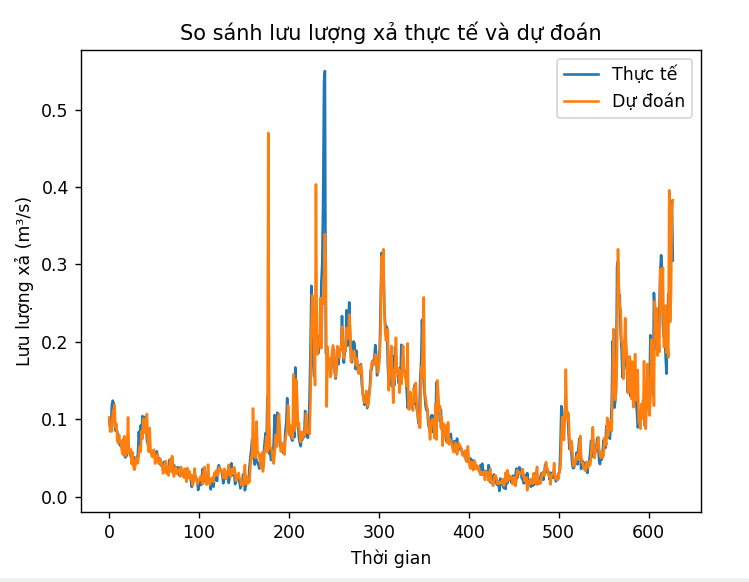
* **Số lượng cây (n\_estimators):** 100
* **Random State:** 42 (để đảm bảo kết quả tái lập được)
* **Các tham số khác:** Sử dụng giá trị mặc định của thư viện scikit-learn

#### 3.5.2.3 Đánh giá hiệu suất của LSTM và LSTM- RF:

Do dự báo lưu lượng nước xả phụ thuộc chủ yếu vào lưu lượng đến của hồ Tua Srah. Các mô hình LSTM và LSTM- RF được cài đặt và kết quả được biểu diễn trong các hình:

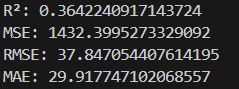


Hình 8. Biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế của LSTM

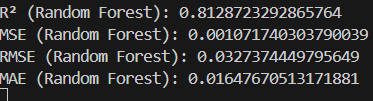


Hình 9. Biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế của LSTM-RF

Sau khi chạy thực nghiệm, nhóm tác giả cũng đã tiến hành so sánh kết quả của các phương pháp về cả các độ đo đánh giá, cũng như là thời gian chạy. Giá trị độ đo được trình bày cụ thể như sau:



Hình 10. Thang điểm của LSTM



Hình 11. Thang điểm của LSTM- RF

**Kết quả và Nhận xét**

Kết quả thực nghiệm trên mô hình LSTM và LSTM-RF được thể hiện trong Bảng X và Bảng Y (bạn cần thay X và Y bằng số thứ tự bảng tương ứng). Các bảng này cung cấp các giá trị MAE, MSE, RMSE và R² của từng mô hình trên tập kiểm tra, với các kịch bản đầu vào khác nhau (bạn cần mô tả rõ các kịch bản đầu vào).

**Phân tích kết quả:**

* **So sánh LSTM và LSTM-RF:** Kết quả cho thấy mô hình **LSTM-RF** cho ra kết quả dự báo khả quan hơn so với mô hình **LSTM thuần túy** trên tất cả các độ đo đánh giá (MAE, MSE, RMSE và R²) và ở tất cả các kịch bản đầu vào (cần bổ sung số liệu cụ thể từ bảng kết quả để củng cố nhận định này). Điều này chỉ ra rằng việc **kết hợp giữa LSTM và Random Forest đã cải thiện đáng kể hiệu suất dự báo** so với việc chỉ sử dụng riêng lẻ mô hình LSTM.
* **Hiệu quả của trích xuất đặc trưng:** Mô hình LSTM-RF tận dụng khả năng **trích xuất đặc trưng hiệu quả của LSTM** từ dữ liệu chuỗi thời gian. Các **hidden states cuối cùng** của LSTM đã mã hóa thông tin quan trọng từ chuỗi dữ liệu đầu vào, giúp cung cấp các đặc trưng có ý nghĩa cho mô hình Random Forest.
* **Sức mạnh của Random Forest:** Mô hình **Random Forest** với khả năng **kết hợp dự đoán từ nhiều cây quyết định** đã cho thấy hiệu quả trong việc **nâng cao độ chính xác và giảm thiểu overfitting** khi dự báo lưu lượng xả dựa trên các đặc trưng từ LSTM.
* **Hạn chế của LSTM thuần túy:** Mô hình LSTM thuần túy, trong trường hợp này, có thể **không đủ mạnh để tự mình đưa ra dự đoán chính xác** cho bài toán dự báo lưu lượng xả, có thể do **khả năng khái quát hóa kém hơn** so với LSTM-RF, hoặc do **chưa tối ưu được tham số** cho mô hình LSTM.

**Kết luận và Hướng phát triển:**

Mô hình LSTM-RF đã cho thấy **tiềm năng và hiệu quả** trong việc dự báo lưu lượng xả hồ thủy điện Tua Srah, **vượt trội hơn so với mô hình LSTM thuần túy**. Việc kết hợp giữa khả năng trích xuất đặc trưng từ dữ liệu chuỗi thời gian của LSTM và khả năng dự báo mạnh mẽ của Random Forest đã mang lại kết quả khả quan.

Tuy nhiên, mô hình vẫn cần được cải thiện để **nâng cao độ chính xác, đặc biệt là ở các điểm có lưu lượng xả đột biến**. Một số hướng phát triển trong tương lai bao gồm:

* **Tối ưu hóa tham số:** Tiếp tục **tinh chỉnh các tham số** của cả LSTM và Random Forest (ví dụ: learning rate, số units, số cây, độ sâu cây) để tìm ra bộ tham số tối ưu.
* **Mở rộng tập dữ liệu:** Bổ sung thêm dữ liệu, đặc biệt là các dữ liệu về các trường hợp lưu lượng xả đột biến, để mô hình có thể học được đầy đủ các quy luật biến động.
* **Thử nghiệm các kiến trúc LSTM khác:** Khám phá các kiến trúc LSTM phức tạp hơn (ví dụ: Bidirectional LSTM, Stacked LSTM) để nâng cao khả năng học của mô hình.
* **Thử nghiệm với các mô hình khác:** Nghiên cứu áp dụng các mô hình Deep Learning tiên tiến hơn cho bài toán chuỗi thời gian như Transformer, Attention Mechanism để so sánh và có thể cải thiện hơn nữa độ chính xác của mô hình.
* **Tăng cường phần Random Forest**: Thử nghiệm các thuật toán khác ngoài Random Forest như Gradient Boosting Machine (GBM) để xem xét liệu chúng có thể cải thiện kết quả hay không.
* **Xử lý dữ liệu bất thường**: Bao gồm các phương pháp phát hiện và xử lý điểm bất thường (outlier detection and handling) trong dữ liệu chuỗi thời gian, có thể giúp mô hình học tốt hơn và dự báo chính xác hơn trong các tình huống cực đoan.

## 3.6 Ứng dụng

Hệ thống này tập trung vào việc khai thác và dự báo lưu lượng nước đối với hồ Tua Srah. Quá trình thực hiện bao gồm việc thu thập và tiền xử lý dữ liệu, sau đó áp dụng mô hình dự báo kết hợp giữa LSTM và Random Forest. Mục tiêu của nghiên cứu là cung cấp các dự đoán chính xác về lưu lượng nước, từ đó hỗ trợ việc quản lý tài nguyên nước một cách hiệu quả và bền vững.

### 3.6.1 Tổng quan hệ thống

Chức năng “Dự đoán trên số liệu nhập tay”: Ở chức năng này, người dùng sẽ tự nhập số liệu giả định vào các ô thông tin đầu vào và sẽ được trả về kết quả dự báo lưu lượng nước xả của hồ Tua Srah tương ứng với số liệu đã nhập trước đó. Và sẽ đưa ra trạng thái của lưu lượng nước xả của hồ Tua Srah thời điểm dự báo phía sau3.6.2 Giao diện hệ thống

**Công nghệ thiết kế ứng dụng dự đoán lưu lượng xả và mực nước**

Ứng dụng "Dự đoán Lưu lượng Xả và Mực nước" là một công cụ được phát triển để hỗ trợ dự báo các thông số quan trọng trong quản lý hồ chứa, cụ thể là lưu lượng xả và mực nước hồ. Giao diện người dùng của ứng dụng được thiết kế đơn giản, trực quan, cho phép người dùng dễ dàng nhập dữ liệu đầu vào và nhận kết quả dự đoán.

**Ngôn ngữ lập trình và thư viện:**

Ứng dụng được xây dựng bằng Python, một ngôn ngữ lập trình mạnh mẽ, phổ biến và dễ sử dụng trong các lĩnh vực khoa học dữ liệu và học máy. Python cung cấp nhiều thư viện hỗ trợ mạnh mẽ cho việc xử lý dữ liệu và phát triển các mô hình học máy, giúp tối ưu hóa quá trình triển khai và phân tích. Python cung cấp nhiều thư viện hỗ trợ cho việc phát triển các ứng dụng khoa học và kỹ thuật, bao gồm:

* **Thư viện giao diện người dùng (GUI):** Dựa vào hình ảnh, có thể dự đoán ứng dụng sử dụng một trong các thư viện GUI phổ biến của Python là **Tkinter:** Thư viện GUI tiêu chuẩn đi kèm với Python, đơn giản, dễ sử dụng và có sẵn trên hầu hết các nền tảng.
* **Thư viện tính toán và khoa học dữ liệu:**
  + **NumPy:** thư viện cơ bản cho tính toán khoa học trong Python, cung cấp các công cụ mạnh mẽ để làm việc với mảng đa chiều (ndarray) và các hàm toán học hiệu suất cao. Nó hỗ trợ các phép toán số học, đại số tuyến tính, thống kê, và xử lý ma trận, giúp tối ưu hóa hiệu suất và khả năng tính toán trong các ứng dụng khoa học dữ liệu và học máy.
  + **Pandas:** Thư viện mạnh mẽ cho việc xử lý và phân tích dữ liệu dạng bảng, cho phép thao tác, biến đổi và trích xuất thông tin từ dữ liệu dễ dàng.
  + **Scikit-learn:** thư viện mạnh mẽ trong Python dành cho học máy, cung cấp các thuật toán phân loại, hồi quy, phân cụm, và nhiều phương pháp tiền xử lý dữ liệu. Thư viện này cũng hỗ trợ lựa chọn mô hình, đánh giá hiệu suất, và tối ưu hóa các tham số. Với giao diện dễ sử dụng và khả năng tích hợp cao, Scikit-learn là công cụ lý tưởng cho việc xây dựng và triển khai các mô hình học máy trong các bài toán khoa học dữ liệu và trí tuệ nhân tạo.



Hình 12. Giao diện ứng dụng

Chức năng thông báo kết quả được dự đoán với mực nước đầu vào được xcs định theo công thức:

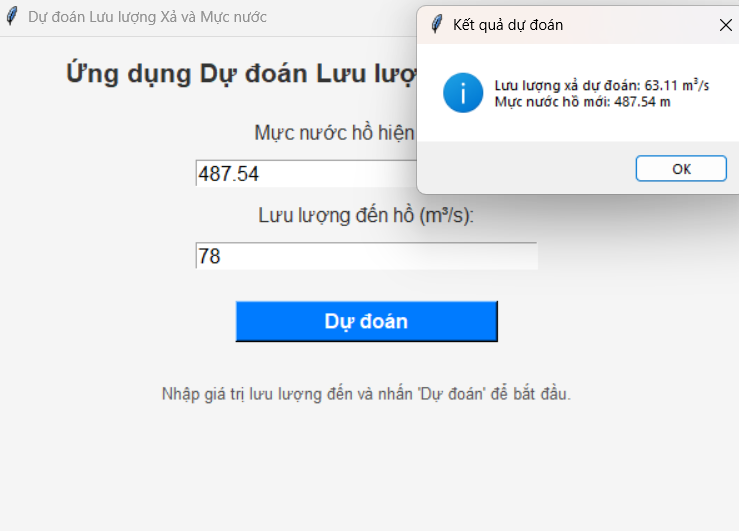
**Công thức tính mực nước hồ:**

V2 = V1 + (\*t (3-1)

**Giải thích các thành phần trong công thức:**

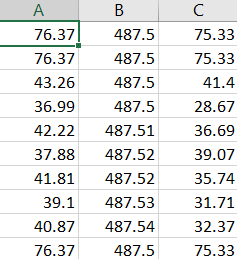
* **Q**: Lưu lượng đến *(m³/s)* – Đây là lượng nước chảy vào hồ từ thượng nguồn hoặc các nhánh sông, suối.
* **q**: Lưu lượng xả *(m³/s)* – Đây là lượng nước được xả ra từ hồ chứa và là giá trị cần được dự đoán trong quá trình vận hành.
* **t**: Thời gian *(s)* – Khoảng thời gian được sử dụng để tính toán mực nước hồ thay đổi.
* **V₁**: Thể tích hồ tại thời điểm ban đầu *(triệu m³)* – Thể tích nước trong hồ trước khi tiến hành xả.
* **V₂**: Thể tích hồ tại thời điểm sau khi xả *(triệu m³)* – Đây là giá trị thể tích nước trong hồ cần tính toán sau quá trình xả.

Từ đó có thể xác định được mực nước đầu vào, và xác định được lưu lượng xả.



Hình 13. Kết quả trả về và hiển thị thông báo dự đoán

Chức năng “Dự đoán trên dữ liệu file excel”: Ở chức năng này, người dùng sẽ gửi lên 1 file excel chứa số liệu của các dữ liệu đầu vào giống mẫu. Ở đây người dùng có thể gửi lên 1 bản ghi hoặc nhiều bản ghi trong excel.



Hình 14. Kết quả trả về vào file excel

### 3.6.3 Triển khai chương trình thử nghiệm

Chương trình thử nghiệm được thiết kế để đánh giá hiệu quả của mô hình LSTM-RF trong việc dự báo lưu lượng xả hồ thủy điện Tua Srah. Hệ thống thử nghiệm bao gồm hai phần chính: **Hệ thống khai thác dữ liệu** và **Hệ thống dự báo**.

**1. Hệ thống khai thác dữ liệu:**

* **Nhiệm vụ:** Thu thập dữ liệu đầu vào cho mô hình dự báo, bao gồm:

**Dữ liệu vận hành hồ chứa Tua Srah:** Bao gồm mực nước hồ, lưu lượng nước đến (có thể lấy từ dữ liệu quá khứ hoặc từ hệ thống giám sát thời gian thực).

* **Phương pháp:** có thể sử dụng kỹ thuật **crawling (thu thập dữ liệu tự động)** để mô phỏng dữ liệu từ các trang web trực tuyến. Các thư viện Python như requests, Beautiful Soup có thể được sử dụng cho việc này.
* **Tần suất thu thập:** Tùy chỉnh dựa trên yêu cầu, có thể thu thập dữ liệu mỗi giờ, mỗi ngày,...
* **Lưu trữ**: Dữ liệu sau khi thu thập sẽ được lưu trữ và dùng cho quá trình huấn luyện và dự báo

**2. Hệ thống dự báo:**

* **Mô hình:** Sử dụng mô hình **LSTM-RF** đã được huấn luyện để dự báo lưu lượng xả hồ thủy điện Tua Srah.
* **Các kịch bản dự báo:**
  + Mô hình LSTM-RF dự báo lưu lượng xả hồ thủy điện Tua Srah dựa vào 2 đặc trưng đầu vào là mực nước hồ và lưu lượng nước đến.
* **Quy trình huấn luyện:**
  + Mô hình được huấn luyện định kỳ (ví dụ: hàng ngày, hàng tuần) với dữ liệu mới được cập nhật từ hệ thống khai thác dữ liệu.
  + Quá trình huấn luyện bao gồm các bước: tiền xử lý dữ liệu, chuẩn bị dữ liệu cho LSTM, huấn luyện mô hình LSTM, trích xuất đặc trưng, và huấn luyện mô hình Random Forest.
* **Quy trình dự báo:**
  + **Khởi động ứng dụng (giao diện):** Người dùng khởi động ứng dụng dự báo.
  + **Nhập dữ liệu đầu vào:** Người dùng nhập dữ liệu cho các thông tin đầu vào bao gồm: mực nước hồ Tua Srah tại thời điểm t, lưu lượng nước đến hồ Tua Srah tại thời điểm t. Dữ liệu có thể được nhập thủ công hoặc từ file.
  + **Kiểm tra dữ liệu:** Ứng dụng sẽ kiểm tra dữ liệu đầu vào có hợp lệ hay không.
  + **Tiền xử lý dữ liệu:** Dữ liệu đầu vào được tiền xử lý (ví dụ: chuẩn hóa) trước khi đưa vào mô hình.
  + **Truyền dữ liệu đến mô hình dự báo:** Dữ liệu đã được tiền xử lý được đưa vào mô hình LSTM-RF.
  + **Mô hình dự báo trả về kết quả:** Mô hình LSTM-RF thực hiện dự báo và trả về kết quả lưu lượng xả hồ Tua Srah.
  + **Hiển thị kết quả:** Ứng dụng nhận kết quả dự báo và hiển thị cho người dùng trên giao diện.

## 3.7 Nhận xét và đánh giá

Trong đề tài này, em đã giải quyết bài toán dự báo **lưu lượng xả** hồ thủy điện Tua Srah bằng cách áp dụng mô hình kết hợp **LSTM-RF**. Mô hình này đã kết hợp khả năng trích xuất đặc trưng hiệu quả từ dữ liệu chuỗi thời gian của LSTM và khả năng dự báo chính xác của Random Forest, từ đó nâng cao hiệu suất dự báo. Kết quả thực nghiệm đã chỉ ra rằng mô hình **LSTM-RF** cho kết quả **tốt hơn** so với mô hình **LSTM thuần túy** trên bộ dữ liệu của hồ thủy điện Tua Srah (cần bổ sung số liệu so sánh cụ thể: MAE, MSE, RMSE, R²).

**Ý nghĩa:**

* **Hỗ trợ vận hành hồ chứa hiệu quả:**  
  Nghiên cứu cung cấp một công cụ dự báo lưu lượng xả nước đáng tin cậy, giúp các nhà quản lý hồ thủy điện Tua Srah vận hành hiệu quả hơn, tối ưu hóa việc xả nước và phát điện, đồng thời đảm bảo an toàn cho công trình và khu vực hạ lưu.
* **Giảm thiểu rủi ro ngập lụt và thiếu nước:**  
  Với dự báo chính xác lưu lượng xả, nhà quản lý có thể đưa ra các quyết định phù hợp để hạn chế nguy cơ ngập lụt trong mùa mưa và đảm bảo cung cấp đủ nước cho sản xuất, sinh hoạt, và môi trường trong mùa khô.
* Hỗ trợ người dân vùng hạ lưu:  
  Ứng dụng giúp người dân sống gần hồ Tua Srah có thể theo dõi các dự báo lưu lượng xả nước, từ đó chủ động thực hiện các biện pháp ứng phó trước các tình huống thời tiết hoặc biến động bất thường, giảm thiểu thiệt hại về tài sản và con người.
* **Tối ưu hóa sử dụng tài nguyên nước:**  
  Dự báo lưu lượng xả chính xác giúp sử dụng tài nguyên nước một cách hợp lý hơn, tránh lãng phí và góp phần bảo vệ nguồn nước cho các mục đích dài hạn.
* Bảo vệ môi trường và hệ sinh thái:  
  Quản lý lưu lượng xả hợp lý giúp duy trì dòng chảy tự nhiên, bảo vệ hệ sinh thái sông suối ở hạ lưu hồ chứa, giảm thiểu tác động tiêu cực đến động, thực vật cũng như chất lượng nước.
* **Tăng cường năng suất và hiệu quả kinh tế:**  
  Việc vận hành hồ chứa tốt hơn không chỉ đảm bảo sản lượng điện ổn định mà còn giảm chi phí khắc phục sự cố do lũ lụt hoặc hạn hán, mang lại lợi ích kinh tế lâu dài.
* **Xây dựng hệ thống hỗ trợ quyết định thực tiễn:**  
  Hệ thống dự báo và giao diện ứng dụng cho phép người dùng nhập dữ liệu và nhận kết quả dự báo tức thời, giúp tăng cường tính khả dụng và thực tiễn trong việc hỗ trợ quản lý hồ chứa.
* **Ứng dụng trong quản lý các hồ chứa khác:**  
  Phương pháp và hệ thống này có thể được nhân rộng để áp dụng cho các hồ chứa khác trong cả nước, nâng cao năng lực quản lý tài nguyên nước ở quy mô quốc gia.

**Hạn chế:**

Mô hình LSTM-RF có độ phức tạp cao hơn so với các mô hình đơn giản, dẫn đến thời gian huấn luyện và dự báo lâu hơn. Ngoài ra, mô hình cũng cho thấy hạn chế trong việc dự báo chính xác các giá trị lưu lượng xả đột biến (cần phân tích và nêu rõ nguyên nhân).

**Hướng phát triển:**

Các nghiên cứu tiếp theo của em sẽ tập trung vào các hướng sau:

* **Cải thiện độ chính xác:**
  + **Tối ưu hóa tham số:** Tiếp tục nghiên cứu và tối ưu hóa các tham số của mô hình LSTM-RF (ví dụ: learning rate, số units, số cây, độ sâu cây,...) để nâng cao hơn nữa độ chính xác của mô hình.
  + **Thử nghiệm các kiến trúc LSTM khác:** Khám phá các kiến trúc LSTM phức tạp hơn (ví dụ: Bidirectional LSTM, Stacked LSTM) hoặc các mô hình Deep Learning tiên tiến khác cho chuỗi thời gian (ví dụ: Transformer, Attention Mechanism) để cải thiện khả năng học và dự báo.
  + **Xử lý dữ liệu:** Tập trung vào tiền xử lý dữ liệu, đặc biệt là xử lý các điểm ngoại lai (outliers) và các giá trị đột biến, cải thiện chất lượng dữ liệu đầu vào.
  + **Tăng cường dữ liệu (Data Augmentation):** Nghiên cứu các phương pháp tăng cường dữ liệu phù hợp cho bài toán chuỗi thời gian để cải thiện hiệu suất của mô hình, đặc biệt trong các tình huống dữ liệu khan hiếm hoặc có các biến động bất thường.
* **Giảm thời gian tính toán:**
  + **Tối ưu hóa code:** Tối ưu hóa mã nguồn, sử dụng các kỹ thuật lập trình hiệu quả nhằm giảm thời gian huấn luyện và tăng độ chính xác của dự báo.
  + **Nghiên cứu phần cứng:** Sử dụng phần cứng mạnh hơn (ví dụ: GPU) để tăng tốc độ tính toán.
  + **Áp dụng kỹ thuật song song:** Khai thác các kỹ thuật tính toán song song để rút ngắn thời gian huấn luyện mô hình.
  + **Giảm độ phức tạp của mô hình**: Nghiên cứu, điều chỉnh lại các tham số của LSTM và RF để cân bằng giữa độ phức tạp của mô hình với độ chính xác, cải thiện tốc độ.
* **Mở rộng hệ thống:**
  + **Tích hợp thêm nguồn dữ liệu:** Bổ sung thêm các nguồn dữ liệu liên quan khác (ví dụ: dữ liệu khí tượng thủy văn chi tiết hơn, dữ liệu vận hành của các hồ chứa lân cận) để nâng cao độ chính xác của mô hình.
  + **Phát triển hệ thống hỗ trợ ra quyết định:** Xây dựng hệ thống hỗ trợ ra quyết định dựa trên kết quả dự báo của mô hình, giúp người vận hành hồ chứa đưa ra các quyết định tối ưu trong việc xả nước và phát điện.
  + **Cải thiện mô hình Random Forest**: Thử nghiệm và thay thế bằng các mô hình khác như Gradient Boosting để tối ưu hơn.

**Về hệ thống:**

Đã xây dựng được hệ thống khai thác dữ liệu và hệ thống dự báo lưu lượng xả hồ thủy điện Tua Srah, bao gồm cả ứng dụng (client-side) để người dùng tương tác và sử dụng. Ứng dụng đã trực quan hóa thông tin đầu vào, cung cấp kết quả dự báo, có tính khả thi và ứng dụng cao. Giao diện của ứng dụng cho phép người dùng nhập các thông tin mực nước hồ, lưu lượng nước đến, và thực hiện dự báo lưu lượng xả, hỗ trợ hiệu quả cho việc quản lý hồ thủy điện Tua Srah. Ngoài ra còn giúp người dân gần khu vực hồ Tua Srah có một nguồn để tham khảo về tình trạng ở khu vực của mình, đồng thời có những biện pháp ứng phó kịp thời với những yếu tố xấu xảy ra.

# KẾT LUẬN

Sau quá trình nghiên cứu, dưới sự hướng dẫn tận tình của thầy Trần Mạnh Tuấn, em đã nắm được những kiến thức cơ bản và nâng cao về khai phá dữ liệu, học máy và học sâu, đặc biệt là các mô hình **LSTM và Random Forest**. Em đã hiểu được cách sử dụng các thuật toán và mô hình học máy để phân tích dữ liệu chuỗi thời gian, từ đó giải quyết bài toán dự báo **lưu lượng xả** hồ thủy điện, cụ thể là hồ thủy điện **Tua Srah**. Đồng thời, em đã triển khai được **ứng dụng** (client-side) để thực hiện dự báo và hiển thị kết quả.

Với đề xuất sử dụng mô hình **LSTM-RF** vào dự báo **lưu lượng xả** hồ thủy điện Tua Srah là **khả thi**, đáp ứng được yêu cầu của bài toán và có thể đưa ra kết quả dự báo với độ chính xác **cao hơn so với mô hình LSTM thuần túy** (cần bổ sung số liệu so sánh cụ thể). Mô hình đã cho thấy tiềm năng ứng dụng thực tế, giúp quản lý vận hành hồ chứa an toàn và hiệu quả hơn, hỗ trợ đưa ra các biện pháp phòng tránh và ứng phó kịp thời với biến động của lưu lượng xả. Ứng dụng đã được xây dựng bước đầu, mang đến một nguồn tham khảo hữu ích cho người quản lý, vận hành hồ chứa cũng như người dân trong khu vực. Tuy nhiên, ứng dụng cần được triển khai trên tập kiểm thử, và kiểm chứng độ chính xác với nhiều mẫu dữ liệu hơn. Mô hình cũng cần được tiếp tục nghiên cứu, cải thiện để nâng cao độ chính xác và thích ứng tốt hơn với các điều kiện vận hành khác nhau của hồ thủy điện Tua Srah.

**Kinh nghiệm thu được:**

* Thành thạo kiến thức về khai phá dữ liệu và biết cách áp dụng quy trình khai phá dữ liệu vào các bài toán thực tế một cách hiệu quả.
* Hiểu sâu hơn về học máy, học sâu, đặc biệt là các mô hình **LSTM và Random Forest**, và cách áp dụng vào bài toán dự báo chuỗi thời gian.
* Nắm bắt được các kiến thức chuyên ngành về thủy điện, lưu lượng nước và hoạt động của hồ chứa thủy điện Tua Srah.
* Có kiến thức và kinh nghiệm thực tế trong việc xây dựng mô hình học máy kết hợp (hybrid model) để cải thiện hiệu suất dự báo.
* Rèn luyện kỹ năng lập trình, đặc biệt là **Python** và các thư viện **TensorFlow/Keras, Scikit-learn, Pandas, NumPy, Matplotlib**.
* Áp dụng các kiến thức lý thuyết đã học để giải quyết vấn đề thực tiễn trong ngành thủy điện.
* Kiện thức vững vàng về các môn học như khai phá dữ liệu, học máy, trí tuệ nhân tạo, và nhập môn khoa học dữ liệu sẽ được củng cố.
* Tìm hiểu và nghiên cứu chuyên sâu về lĩnh vực thủy điện, đặc biệt là các vấn đề liên quan đến vận hành và quản lý hồ chứa.

**Hướng phát triển:**

* **Cải thiện mô hình:**
  + Tiếp tục nghiên cứu, **tối ưu hóa tham số** và **thử nghiệm các kiến trúc mạng LSTM** khác nhau để nâng cao độ chính xác của mô hình LSTM-RF trong việc dự báo lưu lượng xả hồ thủy điện Tua Srah.
  + Nghiên cứu áp dụng các **kỹ thuật tăng cường dữ liệu (data augmentation)**, **xử lý dữ liệu bất thường (outlier detection)** để nâng cao chất lượng dữ liệu đầu vào và cải thiện độ chính xác của mô hình.
  + Thử nghiệm thay thế **Random Forest** bằng các thuật toán khác như **Gradient Boosting** để tối ưu hơn
* **Nghiên cứu và phát triển mô hình:**
  + Nghiên cứu các thuật toán học máy, học sâu mới và tiên tiến hơn, đặc biệt là các mô hình chuyên biệt cho dự báo chuỗi thời gian như **Transformer, Attention Mechanism**, để tìm ra phương pháp phù hợp và hiệu quả hơn cho bài toán dự báo lưu lượng xả.
  + Tích hợp thêm các nguồn dữ liệu liên quan khác vào mô hình (ví dụ: dữ liệu khí tượng thủy văn chi tiết hơn, dữ liệu vận hành của các hồ chứa lân cận) để nâng cao độ chính xác.
* **Phát triển ứng dụng:**
  + Hoàn thiện ứng dụng, cải thiện giao diện người dùng và bổ sung thêm các tính năng hữu ích (ví dụ: biểu đồ trực quan hóa dữ liệu, cảnh báo khi lưu lượng xả vượt ngưỡng, xuất báo cáo).
  + Nghiên cứu và triển khai mô hình trên nền tảng web hoặc di động, giúp người dùng có thể dễ dàng truy cập và sử dụng ứng dụng một cách thuận tiện và hiệu quả.
* **Mở rộng ứng dụng:**
  + Nghiên cứu áp dụng mô hình LSTM-RF cho các hồ thủy điện khác, tiến tới xây dựng một hệ thống dự báo lưu lượng xả cho nhiều hồ thủy điện trong khu vực.
  + Phát triển mô hình thành một hệ thống hỗ trợ ra quyết định, cung cấp các khuyến nghị về vận hành hồ chứa dựa trên kết quả dự báo, giúp tối ưu hóa việc sử dụng tài nguyên nước và đảm bảo an toàn cho hồ đập.
* **Thử nghiệm với các bài toán khác:**
  + Nghiên cứu khả năng áp dụng mô hình LSTM-RF cho các bài toán dự báo khác trong lĩnh vực thủy điện, ví dụ như dự báo mực nước hồ, dự báo lưu lượng nước đến, dự báo sản lượng điện.

# TÀI LIỆU THAM KHẢO

[1] TNU Journal of Science and Technology. "Retrieved from" <https://jst.tnu.edu.vn/jst/article/view/3806>

[2] Môi trường & Cuộc sống. "Thủy điện Buôn Kuốp chủ động đảm bảo an toàn hồ đập trong mùa mưa bão." Retrieved from <https://moitruong.net.vn/thuy-dien-buon-kuop-chu-dong-dam-bao-an-toan-ho-dap-trong-mua-mua-bao-54466.html>

[3] S. Bazzazian et al., "Artificial Neural Networks for Rainfall-Runoff Modeling: A Case Study in Northern Italy," *Water Resources Management*, vol. 33, no. 4, pp. 1221–1236, 2019. <https://doi.org/10.1007/s11269-019-2181-y>

[4] A. Hussain et al., "Application of Machine Learning Techniques for River Flow Forecasting in Pakistan," *Journal of Hydrology*, vol. 594, p. 125668, 2021. <https://doi.org/10.1016/j.jhydrol.2020.125668>

[5] S. Shahid et al., "Extreme Learning Machine for River Flow Prediction in Tropical Climates," *Environmental Earth Sciences*, vol. 79, no. 9, p. 403, 2020. <https://doi.org/10.1007/s12665-020-09057-1>

[6] ResearchGate. "Áp dụng mô hình MIKE SHE kết hợp sử dụng sản phẩm mưa dự báo IFS dự báo lưu lượng đến hồ lưu vực sông Trà Khúc-Sông Vệ." Retrieved from [https://www.researchgate.net/publication/330382123](https://www.researchgate.net/publication/330382123" \t "_new)

[7] Tạp chí Khoa học Tự nhiên và Môi trường. "Retrieved from" <http://tapchi.hunre.edu.vn/index.php/tapchikhtnmt/article/view/27/25>

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.

[9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. <https://doi.org/10.1038/nature14539>

[13] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, Wiley, 2014.

[14] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>

[15] A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts*, McGraw-Hill, 2010.

[16] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008. <https://doi.org/10.1145/1327452.1327492>

[17] VTC Academy, "Data Mining là gì?" Retrieved from <https://plus.vtc.edu.vn/data-mining-la-gi>

[18] Nguyễn Tuấn, "Bài 3: Neural Network." Retrieved from <https://nttuan8.com/bai-3-neural-network/>