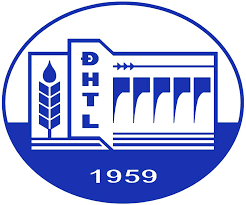
**BỘ GIÁO DỤC VÀ ĐÀO TẠO   
BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG THÔN**

**TRƯỜNG ĐẠI HỌC THỦY LỢI**



CAO THỊ MAI PHƯƠNG

**Dự đoán chất lượng nước**

HỌC PHẦN: ĐỒ ÁN TRÍ TUỆ NHÂN TẠO

Hà Nội, ngày tháng 6 năm 2024

|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG THÔN**  **TRƯỜNG ĐẠI HỌC THỦY LỢI**  CAO THỊ MAI PHƯƠNG  **Dự đoán chất lượng nước**  CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO  GIẢNG VIÊN HƯỚNG DẪN: TS. TẠ QUANG CHIỂU  Hà Nội, ngày tháng 6 năm 2024 |

**LỜI CAM ĐOAN**

Em xin cam đoan đây là bài nghiên cứu của bản thân sinh viên. Các kết quả nghiên cứu và các kết luận trong báo cáo là trung thực, 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 theo đúng quy định.

**Sinh viên**

**Báo cáo bài tập lớn**

**Cao Thị Mai Phương**

**LỜI CẢM ƠN**

Em xin gửi lòng biết ơn sâu sắc đến những tâm huyết và tri thức mà thầy cô trong Công nghệ tjoong tin– Trường Đại học Thủy Lợi đã chỉ bảo cho em. Đặc biệt em xin gửi lời cảm ơn đến thầy TS. Tạ Quang Chiểu– Người đã tận tình hướng dẫn, hỗ trợ về chuyên môn và theo dõi sát sao quá trình thực tập của em.

Trong quá trình học tập em đã học hỏi được nhiều kinh nghiệm hơn trong công việc nhưng vẫn không thể tránh khỏi các sai sót. Em rất mong sẽ nhận được các ý kiến đóng góp quý báu của thầy/cô cũng như ban lãnh đạo của công ty để em có thể hoàn thiện hơn trong tương lai.

Sau cùng, em xin kính chúc thầy cô trong Khoa Công nghệ thông tin dồi dào sức khỏe và công tác tốt.

MỤC LỤC

[CHƯƠNG I: GIỚI THIỆU TỔNG QUAN BÀI TOÁN 6](#_Toc169664950)

[CHƯƠNG 2: CÁC KĨ THUẬT SỬ DỤNG TRONG BÀI TOÁN 7](#_Toc169664951)

[2.1. Tổng quan về học máy 7](#_Toc169664952)

[***2.1.1.*** ***Khái niệm học máy*** 7](#_Toc169664953)

[***2.1.2.*** ***Phân loại nhóm thuật toán trong học máy*** 8](#_Toc169664954)

[***2.1.3.*** ***Các mô hình sử dụng trong bài toán*** 13](#_Toc169664955)

[2.2. Các phương pháp đánh giá mô hình 18](#_Toc169664956)

[CHƯƠNG 3: THỰC NGHIỆM 20](#_Toc169664957)

[3.1. Mô tả bài toán 20](#_Toc169664958)

[3.2. Tiền xử lý dữ liệu 21](#_Toc169664959)

[***3.2.1.*** ***Xử lý dữ liệu thiếu*** 21](#_Toc169664960)

[***3.2.2.*** ***Xử lý ngoại lệ*** 21](#_Toc169664961)

[***3.2.3.*** ***Xử lý mất cân bằng dữ liệu*** 22](#_Toc169664962)

[3.3. Huấn luyện mô hình 23](#_Toc169664963)

[***3.3.1.*** ***Mô hình LSTM*** 23](#_Toc169664964)

[***3.3.2.*** ***Mô hình LSTM\_GWO*** 23](#_Toc169664965)

[***3.3.3.*** ***Mô hình LSTM\_PSO*** 24](#_Toc169664966)

[3.4. Đánh giá 25](#_Toc169664967)

[3.5. Giao diện 25](#_Toc169664968)

[KẾT LUẬN 28](#_Toc169664969)

[TÀI LIỆU THAM KHẢO 29](#_Toc169664970)

# **CHƯƠNG I: GIỚI THIỆU TỔNG QUAN BÀI TOÁN**

Bài toán "Dự đoán chất lượng nước" là một trong những thách thức quan trọng trong lĩnh vực quản lý tài nguyên nước và bảo vệ môi trường. Bài toán này đặt ra mục tiêu xây dựng các mô hình dự đoán dựa trên dữ liệu liên quan đến các chỉ số và thông số về chất lượng nước, từ đó giúp nhận biết và dự báo các vấn đề tiềm ẩn hoặc thay đổi trong chất lượng nước một cách nhanh chóng và chính xác hơn.

**Ý nghĩa của bài toán:**

**Tầm quan trọng của chất lượng nước:** Nước là tài nguyên sống còn và quan trọng đối với mọi sinh hoạt, sản xuất và môi trường tự nhiên. Chất lượng nước không chỉ ảnh hưởng trực tiếp đến sức khỏe con người mà còn đến sự phát triển bền vững của nền kinh tế và xã hội.

**Khả năng ứng dụng của học máy:** Các kỹ thuật học máy và khoa học dữ liệu ngày càng phát triển, mang lại khả năng xử lý và phân tích dữ liệu lớn. Áp dụng học máy trong bài toán dự đoán chất lượng nước giúp tạo ra các mô hình mạnh mẽ, từ đó cải thiện khả năng dự báo và giám sát môi trường một cách hiệu quả hơn.

Việc có thể dự đoán chất lượng nước sớm và chính xác giúp các cơ quan quản lý và chính phủ có thể đưa ra các biện pháp can thiệp kịp thời để bảo vệ tài nguyên nước và ngăn ngừa các vấn đề môi trường nghiêm trọng. Bài toán này không chỉ hướng đến các nhà nghiên cứu và chuyên gia môi trường mà còn đưa ra các công cụ và thông tin hữu ích cho cộng đồng để họ có thể tham gia vào việc bảo vệ và quản lý tài nguyên nước một cách tự chủ và chủ động hơn. Việc nghiên cứu và áp dụng học máy trong dự đoán chất lượng nước không chỉ mang lại lợi ích ngay lập tức mà còn mở ra cánh cửa cho sự phát triển bền vững và sáng tạo trong các lĩnh vực môi trường và khoa học dữ liệu.

**Đối tượng và phương pháp nghiên cứu:**

Bài toán này có thể áp dụng cho một loạt các đối tượng như các hồ, sông, giếng khoan, và các nguồn nước khác. Các phương pháp nghiên cứu bao gồm thu thập dữ liệu đa dạng về các chỉ số nước như pH, độ đục, hàm lượng oxy hòa tan, các kim loại nặng, chất hữu cơ, độ dẫn điện,vi khuẩn có hại,…. Dữ liệu này sau đó sẽ được sử dụng để huấn luyện các mô hình học máy như học sâu (deep learning), cây quyết định (decision trees), hoặc mạng nơ-ron nhân tạo (artificial neural networks) để dự đoán chất lượng nước và phân loại các trạng thái nước.

Bài toán "Dự đoán chất lượng nước" không chỉ là một lĩnh vực nghiên cứu thú vị mà còn có ý nghĩa sâu sắc trong việc giải quyết các vấn đề môi trường và bảo vệ tài nguyên nước. Việc áp dụng các kỹ thuật học máy có thể thúc đẩy sự phát triển bền vững và mang lại lợi ích rõ rệt cho cả cộng đồng và môi trường tự nhiên.

# **CHƯƠNG 2: CÁC KĨ THUẬT SỬ DỤNG TRONG BÀI TOÁN**

* 1. **Tổng quan về học máy**
     1. ***Khái niệm học máy***

Ở thời đại 4.0 hiện nay, trí tuệ nhân tạo (AI) đánh dấu một bước tiến lớn trong sự phát triển công nghệ. Theo các báo cáo và nghiên cứu gần đây của WIPO [1] từ năm 2013 đến 2018, mức đầu tư toàn cầu cho AI đã tăng từ 2 tỷ USD lên hơn 24 tỷ USD. Con số này cho thấy tiềm năng phát triển vượt bậc của AI và thể hiện tiềm năng và nhu cầu lớn trong nhiều lĩnh vực khác nhau.

Về lý thuyết, AI mô phỏng được khả năng tư duy và học tập của con người bằng cách sử dụng máy tính và thuật toán. Hiện nay AI không chỉ là một lĩnh vực nghiên cứu mà còn là một động lực mạnh mẽ đằng sau nhiều tiến bộ trong cuộc sống hàng ngày ảnh hưởng to lớn đến đời sống của chúng ta, từ các ứng dụng cụ thể đến sự biến đổi toàn diện của cộng đồng toàn cầu trên mọi lĩnh vực.

Khi nhắc tới AI thì chúng ta không thể nhắc tới Machine Learning (*ML*), ML là một phần thuộc AI. ML là các thuật toán của máy tính có khả năng học từ dữ liệu và kinh nghiệm mà không cần được lập trình cụ thể, có thể giúp máy tính phân tích, dự đoán và tối ưu hóa các kết quả một cách nhanh chóng và chính xác. Và theo giáo sư [Tom](https://vi.wikipedia.org/w/index.php?title=Tom_Mitchell&action=edit&redlink=1) [Mitchell](https://vi.wikipedia.org/w/index.php?title=Tom_Mitchell&action=edit&redlink=1), trường [Đại học Carnegie Mellon](https://vi.wikipedia.org/wiki/%C4%90%E1%BA%A1i_h%E1%BB%8Dc_Carnegie_Mellon) đã định nghĩa về ML như sau [3]: "Một chương trình máy tính CT được coi là học cách thực hiện một lớp nhiệm vụ NV thông qua trải nghiệm KN, đối với thang đo năng lực NL nếu như dùng NL ta đo thấy năng lực thực thi của chương trình có tiến bộ sau khi trải qua KN". Nói cách khác, ML được học và xây dựng một mô hình (models) để máy tính có thể tự động học và đưa ra các dự đoán hoặc quyết định một cách chính xác. Quá trình này dựa trên một khối lượng lớn dữ liệu đầu vào, còn gọi là dữ liệu huấn luyện. Từ dữ liệu huấn luyện, các thuật toán ML sẽ tự phân tích các mẫu mã và mối quan hệ trong dữ liệu để xây dựng nên các mô hình dự đoán. Chính điều này tạo nên sự thông minh và khả năng tự động hóa của ML. Nhờ ML, con người có thể đưa ra câu trả lời chính xác cho nhiều câu hỏi khó hoặc thậm chí chưa từng đặt ra.

Về cơ bản ML đã thay đổi cách chúng ta xử lý thông tin. Các thuật toán ML có khả năng dự đoán xu hướng, phân loại dữ liệu, và thậm chí tạo ra mô hình dự đoán cho tương lai. Các ví dụ điển hình của việc ứng dụng ML ở thời điểm hiện tại phải kể đến như [nhận](https://www.wikihow.vn/Vi%E1%BA%BFt-m%E1%BB%99t-%C4%91o%E1%BA%A1n-v%C4%83n) [diện khuôn mặt, giọng nói, văn bản, trợ lý ảo, chẩn đoán bệnh, dự báo thời tiết](https://www.wikihow.vn/Vi%E1%BA%BFt-m%E1%BB%99t-%C4%91o%E1%BA%A1n-v%C4%83n) v.v. [ML](https://giasutatdat.edu.vn/tin-tuc/lam-the-nao-de-viet-van-hay-6-tip-can-luu-y-khi-viet-van) [cũng có khả năng dự đoán xu hướng giá BĐS trong tương lai dựa trên các yếu tố như vị](https://giasutatdat.edu.vn/tin-tuc/lam-the-nao-de-viet-van-hay-6-tip-can-luu-y-khi-viet-van) [trí, diện tích, tiện ích, v.v](https://giasutatdat.edu.vn/tin-tuc/lam-the-nao-de-viet-van-hay-6-tip-can-luu-y-khi-viet-van).

* + 1. ***Phân loại nhóm thuật toán trong học máy***

Machine Learning bao gồm 3 nhóm thuật toán chính là: Học có giám sát (*Supervised Learning*), Học không giám sát (*Unsupervised Learning*), Học bán giám sát (*Semi- Supervised Learning*) và ngoài ra con có Học tăng cường (*Reinforcement Learning*). Trong đó 2 nhóm Học có giám sát và Học không giám sát là 2 nhóm thuật toán thường gặp nhất trong ML [4] [5].

* + - 1. *Học có giám sát (Supervised Learning)*

Học có giám sát (Supervised Learning) là một thuật toán học máy dự đoán giá trị đầu ra của những bộ dữ liệu mới, trong đó một mô hình được huấn luyện dựa trên một tập dữ liệu mẫu.

Trong quá trình huấn luyện, thuật toán được cung cấp các cặp dữ liệu gồm đầu vào và nhãn *(labels)* tương ứng cho biết đầu ra. Mục tiêu là tìm ra một hàm ánh xạ giữa đầu vào và nhãn sao cho mô hình có thể dự đoán chính xác nhãn cho những đầu vào mới chưa được gán nhãn. Nếu nói theo phương diện toán học, thì Học có giám sát là khi chúng ta có một tập hợp biến đầu vào 𝑋 = {𝑥1, 𝑥2, … , 𝑥𝑁} và một tập hợp nhãn tương ứng 𝑌 = {𝑦1, 𝑦2, … , 𝑦𝑁}, trong đó 𝑥𝑖, 𝑦𝑖 là các vector. Các cặp dữ liệu biết trước (𝑥𝑖, 𝑦𝑖) ∈ 𝑋 × 𝑌 được gọi là tập training data (dữ liệu huấn luyện). Từ tập training data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập 𝑋 sang một phần tử (xấp xỉ) tương ứng của tập 𝑌:

𝑦𝑖 ≈ 𝑓(𝑥𝑖), ∀𝑖 = 1,2, … , 𝑁 *(2-1)*

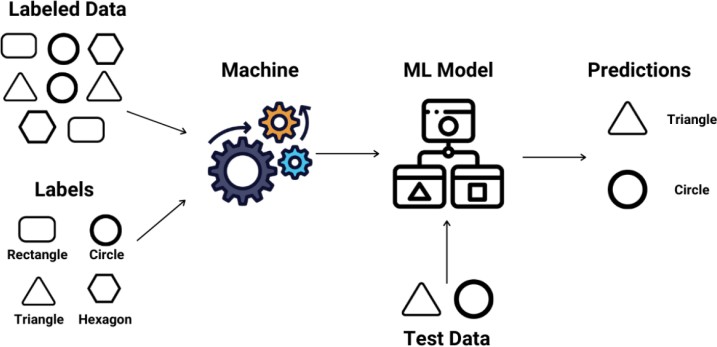
Mục đích là xấp xỉ hàm số 𝑓 thật tốt để khi có một dữ liệu 𝑥 mới, chúng ta có thể tính được nhãn tương ứng của nó 𝑦 = 𝑓(𝑥).

Ưu điểm của Học có giám sát là độ chính xác cao do được huấn luyện trên dữ liệu đã được gán nhãn. Phương pháp này đang rất phổ biến và hiệu quả trong nhiều ứng dụng dự báo, phân loại dữ liệu thực tế.

Trong Học có giám sát, chúng ta có thể phân loại được 2 bài toán thường gặp nhất là: Classification (Phân loại) và Regression (Hồi quy).

**Classification (Phân loại)**

Một bài toán thuộc nhóm classification nếu các *label* của *input* data được chia làm một số hữu hạn nhóm. Lấy ví dụ như hình 2.3, Giả sử ta có một bài toán là phân loại và nhận dạng các hình dạng khác nhau, tập dữ liệu huấn luyện gồm có hai phần: đầu vào và đầu ra. Đầu vào *(input)* là một hình ảnh của một hình dạng nào đó, ví dụ như hình chữ nhật, hình tròn, hình tam giác hay hình lục giác. Đầu ra hay *label* là tên của hình dạng đó, ví dụ như “hình chữ nhật”, “hình tròn”, “hình tam giác” hay “hình lục giác”. Trước tiên, ta cần "chú thích" cho máy tính bằng cách nói cho nó biết mỗi hình dạng là gì. Điều này giống như việc “ghi nhãn” cho từng hình ảnh. Ta nói với máy tính rằng hình này là hình chữ nhật, hình kia là hình tròn, và tiếp tục như vậy. Quá trình này được gọi là “chú thích”. Sau khi máy tính được chú thích đầy đủ, nó sẽ tự học cách dự đoán tên của các hình dạng khi nó nhìn thấy những hình ảnh mà nó chưa từng thấy trước đó. Ví dụ, nếu ta đưa cho máy tính một hình ảnh mới, nó có thể dự đoán đó là hình chữ nhật, hình tròn, hình tam giác hoặc hình lục giác dựa trên những gì nó đã học từ việc chú thích trước đó.



*Hình 2.1 Mô tả quá trình hoạt động của bài toán Classification*

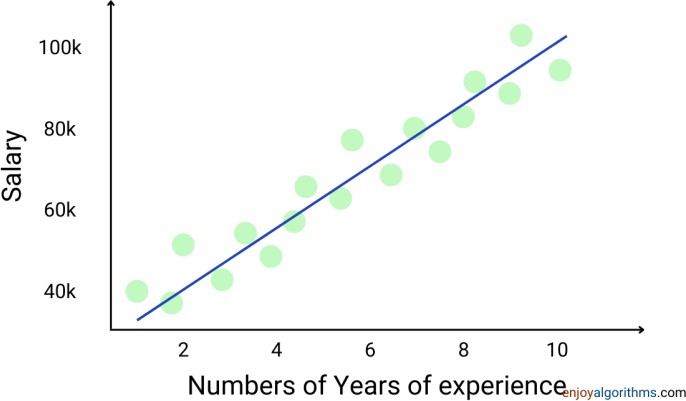
Ngoài ra còn các ví dụ khác như: Trong Gmail muốn kiểm tra xem email có phải là spam hay không; các ngân hàng muốn xác định xem một khách hàng sẽ có khả năng trả nợ

các khoản vay cho ngân hàng hay không; đưa các bức ảnh vào và xác định tương ứng với chữ số nào.

**Regression (Hồi quy)**

Ngược lại với bài toán Classification với số lượng *label* là hữu hạn, thì ở bài toán Regression số lượng *label* là vô hạn và có thể là bất kì giá trị nào và không có khoảng xác định. Ví dụ như hình 2.2, giả sử chúng ta có tập dữ liệu thể hiện mối liên hệ giữa số năm kinh nghiệm làm việc (trục X) và mức lương hàng tháng (trục Y) của một số người. Mỗi điểm dữ liệu trên biểu đồ đại diện cho kinh nghiệm và lương của một người. Bài toán nhằm tìm ra một hàm ánh xạ liên tục từ dữ liệu *input* và *label* và để kết nối những điểm này, như đường màu xanh trên hình ảnh. Đây chính là bài toán hồi quy (regression) trong machine learning.

Nếu giả sử mối quan hệ giữa hai biến là tuyến tính (tức đường thẳng), thuật toán sẽ tìm ra phương trình đường thẳng màu xanh sao cho phù hợp nhất với dữ liệu, thể hiện xu hướng chung. Hay còn được gọi là mô hình hồi quy tuyến tính.



*Hình 2.2 Mô tả quá trình hoạt động của bài toán Regression*

Ngoài ra còn các ví dụ khác như: Dự báo mực nước lũ tại trạm khí tượng thủy văn [6]; dự đoán tuổi thọ và nguy cơ tử vong của người dân ở Đan Mạch, v.v. Đặc biệt, Regression là lựa chọn phù hợp khi chúng ta muốn dự đoán giá nhà dựa trên các biến đầu vào khác nhau, do đó ta sẽ tập trung vào những thuật toán của bài toán này.

* + - 1. *Học không giám sát (Unsupervised Learning)*

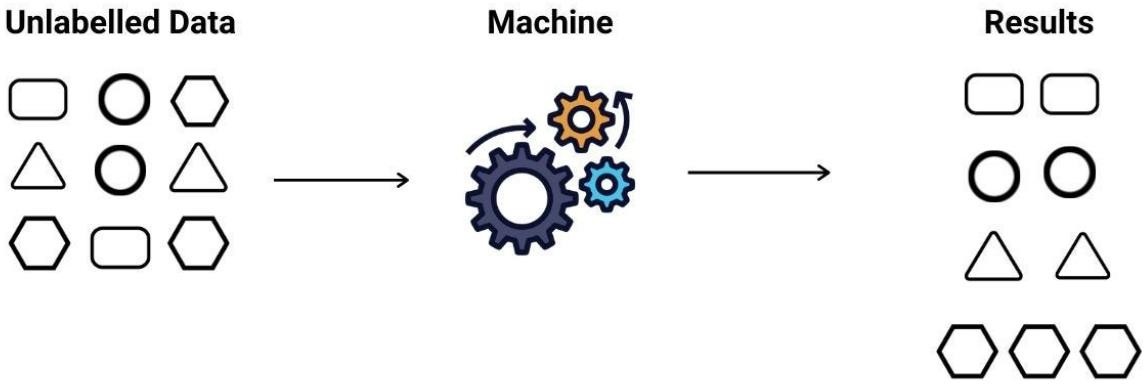
Khác với Học có giám sát là chúng ta có thể biết được giá trị đầu ra hay nhãn *(label)* của dữ liệu huấn luyện, thì đối với Học không giám sát chúng ta không biết được giá trị đầu ra *(outcome)* hay *label* mà chỉ có dữ liệu đầu vào. Thuật toán Học không giám sát sẽ phải tự học từ dữ liệu mà không có sự hướng dẫn cụ thể nào về kết quả mong muốn. Thay vì dựa vào dữ liệu có nhãn để học, Học không giám sát giúp máy tính tự phát triển cấu trúc và hiểu biết về dữ liệu thông qua quy luật ẩn. Theo phương diện toán học, Học không giám sát là khi chúng ta chỉ có dữ liệu vào 𝑋 mà không biết *nhãn* 𝑌 tương ứng.

Ưu điểm của phương pháp này là có thể xử lý được các tập dữ liệu rất lớn mà không đòi hỏi gán nhãn, tiết kiệm nhân lực. Tuy nhiên chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là số 1 hay 2

Trong Học không giám sát, chung ta có thể phân loại được 2 bài toán thường gặp nhất là: Clustering (phân nhóm) và Association

**Clustering (phân nhóm)**

Ở bài toán này, Ta có dữ liệu huấn luyện đầu vào *(input)* X và mục tiêu của bài toán là phân thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ nhìn vào hình 2.3 bên dưới, ta có dữ liệu *input* bao gồm các hình ảnh có hình dạng khác nhau như “hình tròn”, “hình vuông”, v.v. Và thuật toán sẽ cố gắng tìm ra sự giống nhau giữa các hình ảnh khác dựa trên các giá trị pixel màu, kích thước và hình dạng và tạo thành các nhóm làm đầu ra trong đó các trường hợp đầu vào tương tự nằm ở đó. Và dễ nhận thấy ở phần kết quả, các hình vuông được nhóm lại với nhau, và tương tự, hình tròn và hình lục giác



*Hình 2.3 Mô tả quá trình hoạt động của bài toán Clustering*

Các ví dụ điển hình ở bài toán này như: Phân nhóm khách hàng dựa trên những đặc điểm chung về thói quen mua sắm, sở thích để có chiến lược tiếp thị và bán hàng hiệu quả cho từng nhóm; Nhóm các bức ảnh có nội dung tương đồng (cùng một đối tượng/người) dựa trên các đặc trưng trích xuất từ ảnh; Phân loại văn bản thành các chủ đề chính trong nghiên cứu văn hóa v.v.

**Association**

Association hay dịch ra tiếng việt là “liên kết”, là một bài toán mà mục tiêu cần tìm ra những mối liên hệ và quy tắc kết hợp giữa các item trong tập dữ liệu. Association sử dụng các thuật toán thống kê và dựa vào phân tích tần suất xuất hiện của các item để phát hiện ra xu hướng mua hàng cùng nhau giữa các sản phẩm hay dịch vụ. Ban đầu, chúng ta sẽ dễ lầm tưởng bài toán trông tượng bài toán Phân nhóm *(Clustering)***.** Nhưng tuy nhiên, bài toán Phân nhóm là tìm mối quan hệ giữa các điểm dữ liệu, còn Association là tìm mối quan hệ giữa các thuộc tính của điểm dữ liệu đó. Ví dụ như trong giỏ hàng mua sắm online, association có thể phát hiện ra những sản phẩm thường được mua cùng nhau. Ví dụ khi mua tã bỉm thì cũng hay mua kèm sữa. Hay khi xem sản phẩm nào thì thường có xu hướng xem cả sản phẩm kia.

* + - 1. *Học bán giám sát (Semi-Supervised Learning)*

Học bán giám sát là một nhóm thuật toán nằm ở giữa Học có giám sát và Học không giám sát, đại diện cho sự kết hợp thông minh giữa hai loại học máy này. Trong Học bán giám sát, chúng ta sử dụng một lượng lớn dữ liệu đầu vào X, nhưng chỉ một phần nhỏ của dữ liệu này được gán nhãn. Điều này có nghĩa là mô hình sẽ học từ cả dữ liệu có nhãn và không nhãn, tận dụng sự đa dạng của cả hai loại dữ liệu để huấn luyện một mô hình mạnh mẽ và linh hoạt.

Hãy theo dõi ví dụ ở hình 2.4 bên dưới để hiểu hơn về cách hoạt động của Học bán giám sát. Giả sử chúng ta có tập dữ liệu hình ảnh gồm 3 loại trái cây là táo, cam và chuối. Trong số đó, chỉ có hình ảnh của cam và chuối được gán nhãn, còn hình ảnh của táo thì không có nhãn. Ban đầu mô hình được huấn luyện trên dữ liệu đã được gán nhãn gồm chỉ hình ảnh cam và chuối. Sau đó mô hình này được dùng để dự đoán nhãn cho các hình ảnh của táo chưa có nhãn.

Tiếp theo, người quan sát sẽ xem xét các dự đoán cho hình ảnh táo và gán nhãn chính xác là "táo" cho các hình có dự đoán đúng. Sau cùng, mô hình được huấn luyện lại trên cả dữ liệu được gán nhãn ban đầu và thêm phần hình ảnh táo đã được gán nhãn vừa rồi.

A diagram of machine learning model

Description automatically generated

*Hình 2.4 Mô tả quá trình hoạt động của Học bán giám sát*

Ưu điểm của Học bán giám sát là kết hợp thêm dữ liệu không nhãn giúp mô hình học được nhiều hơn từ trong dữ liệu cho nên tăng được hiệu suất huấn luyện so với Học có giám sát thuần túy, đồng thời giảm chi phí và nguồn lực lao động gán nhãn so với Học có giám sát.

* + 1. ***Các mô hình sử dụng trong bài toán***
       1. *LSTM*

LSTM, viết tắt của Long Short-Term Memory, là một loại mạng nơ-ron hồi quy (Recurrent Neural Network - RNN) được thiết kế để xử lý và dự đoán chuỗi dữ liệu. LSTM được giới thiệu bởi Hochreiter và Schmidhuber vào năm 1997 và đã được cải tiến bởi nhiều người khác trong các nghiên cứu sau này. Điểm mạnh của LSTM là khả năng học được các phụ thuộc dài hạn trong dữ liệu chuỗi, điều mà các mô hình RNN truyền thống thường gặp khó khăn.

Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng 𝑡𝑎𝑛ℎ*tanh*.

A diagram of a diagram

Description automatically generated

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

A diagram of a plant

Description automatically generated

Ở sơ đồ trên, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác. Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau.

Cấu trúc của LSTM

LSTM bao gồm một tập hợp các đơn vị LSTM (LSTM cells), mỗi đơn vị LSTM có cấu trúc phức tạp hơn so với một nơ-ron đơn giản trong RNN truyền thống. Một đơn vị LSTM chứa ba thành phần chính:

1. **Cổng quên (Forget Gate)**: Quyết định thông tin nào từ trạng thái trước đó cần được giữ lại hoặc loại bỏ.
2. **Cổng đầu vào (Input Gate)**: Quyết định thông tin mới nào sẽ được thêm vào trạng thái.
3. **Cổng đầu ra (Output Gate)**: Quyết định giá trị nào từ trạng thái hiện tại sẽ được xuất ra ngoài.

Mỗi đơn vị LSTM duy trì hai trạng thái: trạng thái tế bào (cell state) và trạng thái ẩn (hidden state). Các cổng này hoạt động thông qua các hàm kích hoạt như sigmoid và tanh để điều chỉnh luồng thông tin.

Ứng dụng của LSTM

LSTM được ứng dụng rộng rãi trong nhiều lĩnh vực xử lý chuỗi dữ liệu, bao gồm:

* **Xử lý ngôn ngữ tự nhiên (NLP)**: Dịch máy, phân loại văn bản, tạo văn bản.
* **Dự đoán chuỗi thời gian**: Dự báo giá cổ phiếu, dự báo nhu cầu sản phẩm.
* **Nhận dạng giọng nói**: Chuyển đổi giọng nói thành văn bản.
* **Xử lý video**: Phân tích hành động trong video, nhận diện khuôn mặt.

Ưu điểm của LSTM

* **Giữ được thông tin dài hạn**: Khả năng duy trì thông tin qua nhiều bước thời gian giúp LSTM xử lý tốt các vấn đề liên quan đến phụ thuộc dài hạn.
* **Giảm thiểu vấn đề biến mất gradient**: Nhờ cấu trúc cổng phức tạp, LSTM giảm thiểu tình trạng gradient biến mất, một vấn đề phổ biến trong RNN truyền thống khi đào tạo với các chuỗi dài.

Nhược điểm của LSTM

* **Tính toán phức tạp**: Cấu trúc của LSTM phức tạp hơn RNN truyền thống, yêu cầu nhiều tài nguyên tính toán hơn.
* **Cần nhiều dữ liệu để đào tạo**: Để phát huy hết khả năng của LSTM, thường cần một lượng lớn dữ liệu để đào tạo.

LSTM là một công cụ mạnh mẽ trong việc xử lý và dự đoán dữ liệu chuỗi, vượt trội hơn RNN truyền thống nhờ khả năng giữ được thông tin dài hạn và giải quyết các vấn đề gradient biến mất.

* + - 1. *Giải thuật GWO*

GWO, viết tắt của **Grey Wolf Optimizer** (Giải thuật tối ưu hoá bầy sói xám), là một thuật toán tối ưu hoá dựa trên hành vi săn mồi và cấu trúc xã hội của sói xám trong tự nhiên. Giải thuật này được đề xuất bởi Mirjalili và cộng sự vào năm 2014. Mục đích của GWO là tìm kiếm lời giải tốt nhất cho các vấn đề tối ưu hoá thông qua việc mô phỏng hành vi săn mồi của bầy sói xám.

Cấu trúc của bầy sói trong GWO

Bầy sói xám trong GWO được phân thành bốn cấp độ:

1. **Alpha (α)**: Sói đầu đàn, có nhiệm vụ dẫn dắt cả bầy và ra quyết định.
2. **Beta (β)**: Sói cấp thứ hai, hỗ trợ alpha và có thể thay thế vị trí alpha nếu cần.
3. **Delta (δ)**: Sói cấp thứ ba, báo cáo cho alpha và beta, điều hành sói cấp dưới.
4. **Omega (ω)**: Sói cấp dưới cùng, làm theo lệnh của tất cả các cấp trên.

Quá trình tối ưu hoá trong GWO

Giải thuật GWO mô phỏng quá trình săn mồi của bầy sói qua các bước sau:

1. **Khởi tạo**: Ban đầu, một tập hợp các cá thể (sói) được khởi tạo ngẫu nhiên trong không gian tìm kiếm.
2. **Cập nhật vị trí của alpha, beta và delta**: Trong quá trình tối ưu hoá, ba con sói mạnh nhất (alpha, beta và delta) dẫn dắt quá trình săn mồi. Các sói khác (omega) sẽ cập nhật vị trí của mình dựa trên vị trí của ba con sói dẫn đầu này.
3. **Mô phỏng hành vi săn mồi**:
   * **Bao vây con mồi**: Sói xám bao vây con mồi dựa trên vị trí hiện tại của alpha, beta và delta. Vị trí mới của sói được tính toán dựa trên khoảng cách đến con mồi.
   * **Tấn công con mồi**: Sau khi bao vây con mồi, bầy sói tiến hành tấn công. Vị trí của các sói được cập nhật để tiếp cận con mồi gần hơn.
4. **Cập nhật vị trí**: Vị trí của mỗi con sói được cập nhật dựa trên công thức toán học mô phỏng hành vi săn mồi và vị trí của alpha, beta, delta.
5. **Kiểm tra điều kiện dừng**: Quá trình trên lặp lại cho đến khi đạt được số lần lặp tối đa hoặc đáp ứng điều kiện dừng trước khi thuật toán dừng lại và trả về kết quả tối ưu.

Ưu điểm của GWO

* **Dễ triển khai**: Cấu trúc và thuật toán của GWO đơn giản và dễ hiểu, dễ dàng triển khai vào các bài toán khác nhau.
* **Hiệu quả cao**: GWO có thể tìm kiếm giải pháp tối ưu một cách hiệu quả trong không gian tìm kiếm rộng lớn.
* **Không cần gradient**: GWO không cần tính toán gradient của hàm mục tiêu, do đó có thể áp dụng cho các bài toán không liên tục và không khả vi.

Ứng dụng của GWO

GWO được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, bao gồm:

* **Tối ưu hoá các tham số của mô hình học máy**: Điều chỉnh các tham số của mạng nơ-ron, SVM, v.v.
* **Tối ưu hoá hệ thống năng lượng**: Tối ưu hoá lịch trình sản xuất và phân phối năng lượng.
* **Xử lý ảnh**: Phân đoạn ảnh, nâng cao chất lượng ảnh.
* **Quản lý chuỗi cung ứng**: Tối ưu hoá các tuyến đường vận chuyển, quản lý tồn kho.

Giải thuật GWO là một công cụ mạnh mẽ và linh hoạt cho các bài toán tối ưu hoá phức tạp. Nó mô phỏng hành vi tự nhiên của bầy sói xám, từ đó cung cấp các giải pháp tối ưu thông qua quá trình tương tác và săn mồi trong không gian tìm kiếm.

* + - 1. *Giải thuật PSO*

Giải thuật PSO, viết tắt của Particle Swarm Optimization (Giải thuật tối ưu hóa bầy đàn), là một thuật toán tối ưu hóa lấy cảm hứng từ hành vi di cư và tìm kiếm thức ăn của các đàn chim và đàn cá. Thuật toán này được phát triển bởi Eberhart và Kennedy vào năm 1995. Mục đích của PSO là tìm kiếm lời giải tốt nhất cho các vấn đề tối ưu hóa thông qua sự tương tác và phối hợp giữa các cá thể trong bầy đàn.

Cấu trúc của PSO

Trong PSO, mỗi cá thể trong bầy đàn được gọi là một hạt (particle). Mỗi hạt đại diện cho một lời giải tiềm năng trong không gian tìm kiếm và có hai thuộc tính chính:

1. **Vị trí (position)**: Biểu thị lời giải hiện tại của hạt.
2. **Vận tốc (velocity)**: Biểu thị tốc độ và hướng di chuyển của hạt trong không gian tìm kiếm.

Mỗi hạt di chuyển trong không gian tìm kiếm dựa trên kinh nghiệm của chính nó và kinh nghiệm của các hạt khác trong bầy đàn.

Quá trình tối ưu hóa trong PSO

Thuật toán PSO mô phỏng hành vi tìm kiếm thức ăn của đàn chim/cá qua các bước sau:

1. **Khởi tạo**: Ban đầu, một tập hợp các hạt được khởi tạo ngẫu nhiên trong không gian tìm kiếm với vị trí và vận tốc ban đầu.
2. **Cập nhật vận tốc**: Vận tốc của mỗi hạt được cập nhật dựa trên vị trí tốt nhất mà hạt đã tìm được (pBest) và vị trí tốt nhất của toàn bộ bầy đàn (gBest). Công thức cập nhật vận tốc bao gồm thành phần ngẫu nhiên để đảm bảo sự khám phá không gian tìm kiếm.
3. **Cập nhật vị trí**: Vị trí của mỗi hạt được cập nhật dựa trên vận tốc mới.
4. **Đánh giá và cập nhật pBest và gBest**: Đánh giá chất lượng lời giải hiện tại của mỗi hạt và cập nhật pBest và gBest nếu tìm được lời giải tốt hơn.
5. **Kiểm tra điều kiện dừng**: Quá trình trên lặp lại cho đến khi đạt được số lần lặp tối đa hoặc đáp ứng điều kiện dừng.

Ưu điểm của PSO

* **Dễ triển khai**: Cấu trúc và thuật toán của PSO đơn giản và dễ hiểu, dễ dàng triển khai vào các bài toán khác nhau.
* **Hiệu quả cao**: PSO có thể tìm kiếm giải pháp tối ưu một cách hiệu quả trong không gian tìm kiếm rộng lớn.
* **Không cần gradient**: PSO không cần tính toán gradient của hàm mục tiêu, do đó có thể áp dụng cho các bài toán không liên tục và không khả vi.

Ứng dụng của PSO

PSO được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, bao gồm:

* **Tối ưu hóa các tham số của mô hình học máy**: Điều chỉnh các tham số của mạng nơ-ron, SVM, v.v.
* **Tối ưu hóa hệ thống năng lượng**: Tối ưu hóa lịch trình sản xuất và phân phối năng lượng.
* **Xử lý ảnh**: Phân đoạn ảnh, nâng cao chất lượng ảnh.
* **Quản lý chuỗi cung ứng**: Tối ưu hóa các tuyến đường vận chuyển, quản lý tồn kho.

Giải thuật PSO là một công cụ mạnh mẽ và linh hoạt cho các bài toán tối ưu hóa phức tạp. Nó mô phỏng hành vi tự nhiên của đàn chim/cá, từ đó cung cấp các giải pháp tối ưu thông qua quá trình tương tác và phối hợp giữa các cá thể trong bầy đàn. Cấu trúc đơn giản nhưng hiệu quả của PSO giúp nó trở thành lựa chọn ưu tiên trong nhiều ứng dụng thực tiễn.

* 1. **Các phương pháp đánh giá mô hình**

Để đánh giá mô hình có dự đoán có chính xác hay không chúng ta sử dụng 3 độ đo và một phương pháp biểu diễn đồ thị để đánh giá [10], bao gồm:

* Hệ số xác định *(*𝑅2*)*: Là thước đo tỉ lệ phần trăm biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập trong mô hình. Giá trị càng gần 1 thì mô hình càng tốt.

(2-5)

* Sai số tuyệt đối trung bình (Mean Absolute Error - *MAE*): Là trung bình cộng sai số tuyệt đối giữa các giá trị dự đoán và thực tế. Giá trị càng nhỏ thì mô hình càng tốt.

(2-6)

* Lỗi trung bình phương gốc (Root Mean Square Error - *RMSE*): Tương tự MAE nhưng lấy căn bậc 2 của sai số trung bình. RMSE nhạy cảm hơn với các sai số lớn. Giá trị càng nhỏ thì mô hình càng tốt.

(2-7)

Trong đó:

𝑆𝑆𝑟𝑒𝑠 được gọi là Tổng bình phương các sai số, 𝑆𝑆𝑡𝑜𝑡 được gọi là Tổng bình phương sai số so với trung bình

𝑦𝑖 là giá thực tế, 𝑓𝑖 là giá trị dự đoán, là giá trị trung bình của dữ liệu, 𝑛 là số lượng mẫu.

* Đường cong học tập *(Learning Curve)* [11]**:**

Đường học tập là công cụ quan trọng trong đào tạo và đánh giá mô hình máy học, biểu diễn sự biến đổi của hiệu suất trên tập huấn luyện và kiểm thử khi kích thước của tập dữ liệu tăng.

Trên tập huấn luyện, đường học tập thể hiện sự thay đổi sai số khi kích thước tập luyện tăng, có thể phản ánh việc mô hình học từ dữ liệu.

Trên tập kiểm tra, đường học tập giúp đánh giá khả năng học tổng quát của mô hình đối với dữ liệu mới. Nếu đường học tập trên tập kiểm thử không giảm hoặc tăng khi kích thước dữ liệu tăng, có thể là dấu hiệu của mô hình quá mức phức tạp hoặc overfitting.

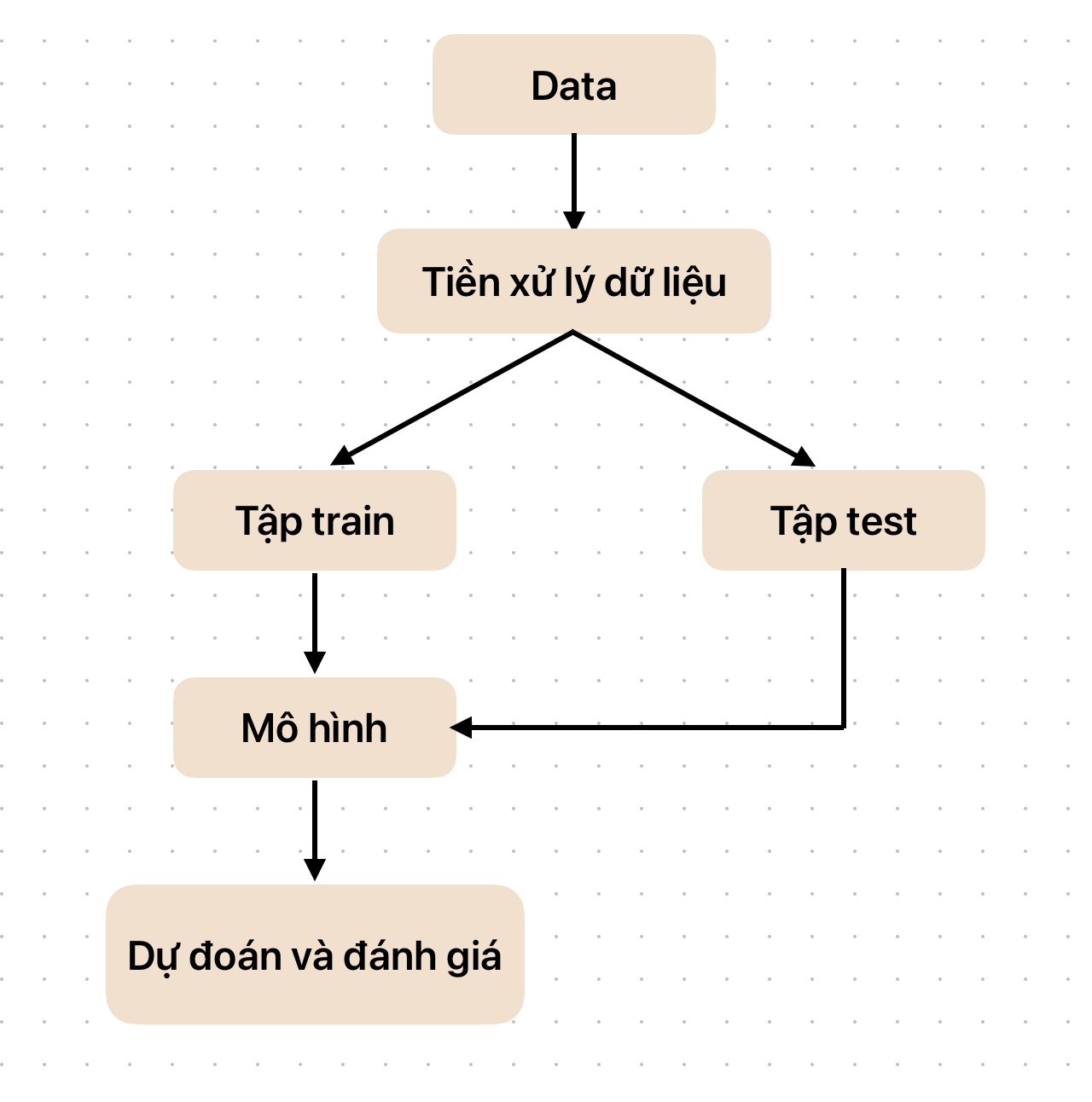
# **CHƯƠNG 3: THỰC NGHIỆM**

1. **Mô tả bài toán**

Với yêu cầu của bài, một số mục tiêu của bài toán được đặt ra là:

* Mục tiêu 1: Thu thập dữ liệu, tiền xử lý dữ liệu
* Mục tiêu 2: Phân tích dữ liệu
* Mục tiêu 3: Dùng phương pháp học sâu kết hợp với thuật toán tối ưu để dự đoán chất lượng nước
* Mục tiêu 4: Xây dựng giao diện

Quy trình thực hiện:

****

A screenshot of a graph

Description automatically generated

Dữ liệu có 9 cột thuộc tính và 1 cột nhãn

|  |  |
| --- | --- |
| **Thuộc tính** | **Ý nghĩa** |
| pH | pH của 1. nước (0 đến 14). |
| Hardness | Khả năng kết tủa xà phòng của nước tính bằng mg/L. |
| Solids | Tổng chất rắn hòa tan tính bằng ppm. |
| Chloramines | Lượng Cloramin tính bằng ppm. |
| Sulfate | Lượng Sunfat hòa tan tính bằng mg/L. |
| Conductivity | Độ dẫn điện của nước tính bằng μS/cm. |
| Organic-carbon | Lượng carbon hữu cơ tính bằng ppm. |
| Trihalomethanes | Lượng Trihalomethane tính bằng μg/L. |
| Tủbidity | Đo tính chất phát sáng của nước tính bằng NTU. |

1. **Tiền xử lý dữ liệu**
2. ***Xử lý dữ liệu thiếu***

Có 3 thuộc tính thiếu dữ liệu là: pH (491), Sulfate (781), Trihalomethanes (162)

**A screen shot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

Xử lý dữ liệu thiếu bằng phương pháp KNN

1. ***Xử lý ngoại lệ***

A graph of a bar chart

Description automatically generated with medium confidence

Tất cả các thuộc tính đề có dữ liệu thiếu. Xử lý dữ liệu thiếu bằng phương pháp IQR.

A chart with blue rectangular bars

Description automatically generated with medium confidence

1. ***Xử lý mất cân bằng dữ liệu***

A screenshot of a computer

Description automatically generated

Nhãn 1 ít hơn nhãn 0. Xử lý mất cân bằng dữ liệu bằng phương pháp SMOTE. Dữ liệu sau khi xử lý mất cân bằng.

A screenshot of a computer

Description automatically generated

1. **Huấn luyện mô hình**
2. ***Mô hình LSTM***

****

Kết qủa dự đoán khi sử dụng mô hình LSTM

1. ***Mô hình LSTM\_GWO***

Các bước thực huấn luyện mô hình

**A diagram of a data system

Description automatically generated**

Sử dụng thuật toán tối ưu GWO để trích chọn các đặc trưng tốt nhất để huấn luyện trong mô hình LSTM.

***A screenshot of a cell phone

Description automatically generated***

Kết qủa dự đoán khi sử dụng mô hình LSTM\_GWO

1. ***Mô hình LSTM\_PSO***

Các bước thực huấn luyện mô hình

A diagram of a company

Description automatically generated

Sử dụng thuật toán tối ưu PSO để trích chọn các đặc trưng tốt nhất để huấn luyện trong mô hình LSTM.

***A screenshot of a phone

Description automatically generated***

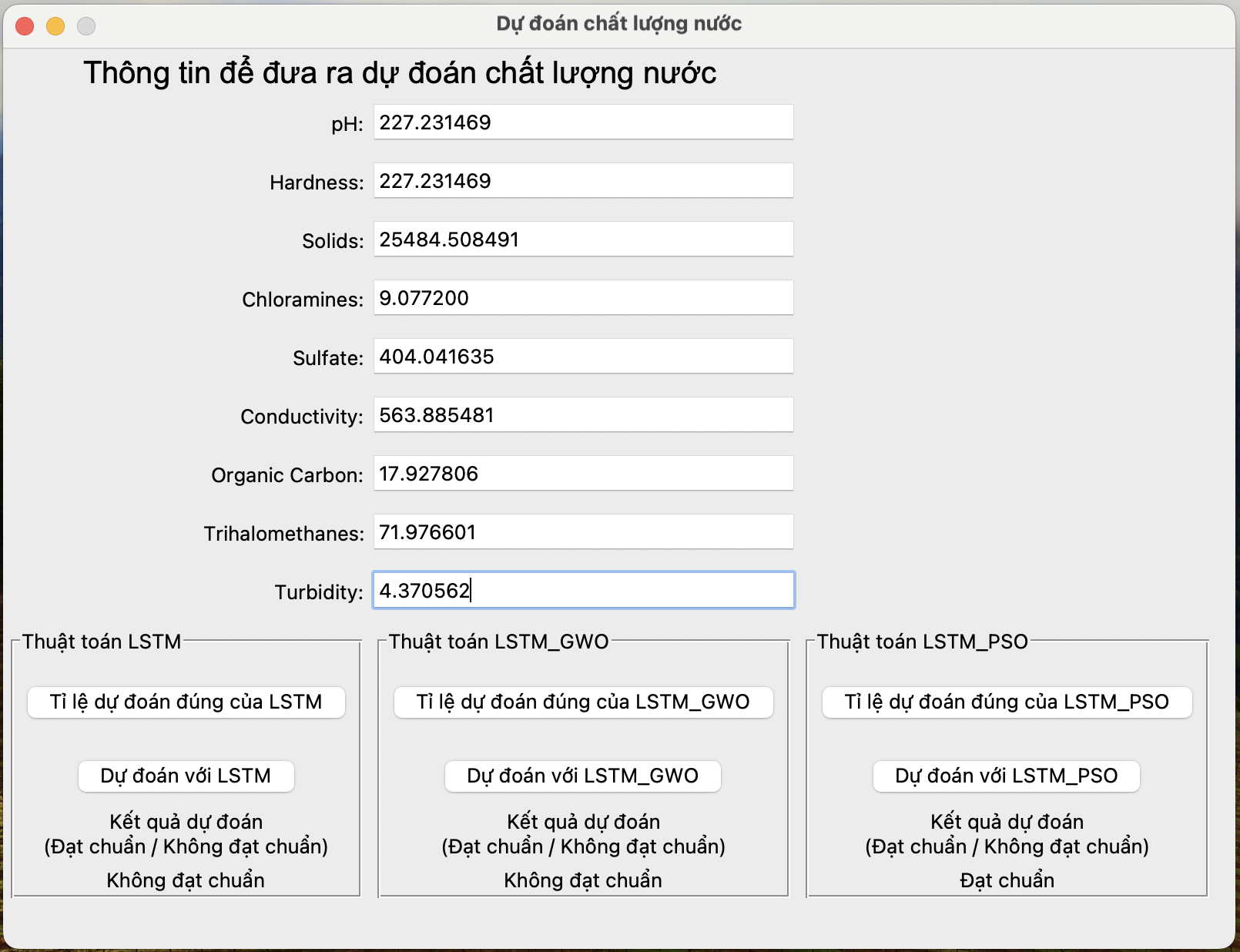
Kết qủa dự đoán khi sử dụng mô hình LSTM\_PSO

1. **Đánh giá**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Precistion** | **Recall** | **F1** | **Accuracy** |
| **LSTM** | 0.5075 | 1.0 | 1.6733 | 0.507 |
| **LSTM\_GWO** | 0.5817 | 0.7537 | 0.6567 | 0.6 |
| **LSTM\_PSO** | 0.6514 | 0.5246 | 0.5812 | 0.61635 |

1. **Giao diện**

Giao diện

****

Tỷ lệ dự đoán của mô hình LSTM

**A screenshot of a computer

Description automatically generated**

Tỷ lệ dự đoán của mô hình LSTM\_GWO

A screenshot of a computer

Description automatically generated

Tỷ lệ dự đoán của mô hình LSTM\_PSO

A screenshot of a computer

Description automatically generated

# **KẾT LUẬN**

1. Kết quả đạt được

- Sử dụng một số phương pháp học sâu để dự đoán chất lượng nước.

- Xây dựng một giao diện dự đoán chất lượng nước.

# **TÀI LIỆU THAM KHẢO**