TRẦN HẢI VINH

HỆ THỐNG THÔNG TIN

2022

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM
KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Trần Hải Vinh

NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU TRONG DỰ BÁO MỘT SỐ THÔNG SỐ KHÍ QUYỂN

LUẬN VĂN THẠC SĨ CHUYÊN NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội - 2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VN

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Trần Hải Vinh

NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU TRONG DỰ BÁO MỘT SỐ THÔNG SỐ KHÍ QUYỂN

Chuyên ngành: Hệ thống thông tin
Mã số: 8480104

LUẬN VĂN THẠC SĨ NGÀNH MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC :
TS. NGUYỄN XUÂN ANH

Hà Nội - 2022

LỜI CAM ĐOAN

Tôi là Trần Hải Vinh, học viên khóa 2020B, ngành Máy tính, chuyên ngành Hệ Thống Thông Tin. Tôi xin cam đoan đề tài nghiên cứu trong luận văn này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận văn là trung thực nếu sai tôi hoàn chịu trách nhiệm.

Tác giả luận văn

Trần Hải Vinh

LỜI CẢM ƠN

Đầu tiên, tôi xin chân thành cảm ơn TS. Nguyễn Xuân Anh, người thầy, lãnh đạo cơ quan của tôi, đã hướng dẫn, dìu dắt tôi trong suốt quá trình làm luận văn. Nhờ sự chỉ báo tận tình của thầy giúp cho tôi có kiến thức nghiên cứu những vấn đề được để cập trong luận văn và giải quyết bài toán đưa ra một cách khoa học.

Tiếp theo, tôi xin trân trọng cảm ơn các thầy cô ở Học viện khoa học và công nghệ Việt Nam cũng như các thầy cô tại Viện công nghệ thông tin, Viện Hàn lâm khoa học và công nghệ đã giảng dạy tận tình, trang bị cho tôi những kiến thức quý báu. Các thầy cô đã tạo ra một môi trường học tập, nghiên cứu khoa học cực kỳ nghiêm chỉnh nhưng cũng rất năng động giúp cho tôi có những kiến thức chuyên môn nền tảng làm cơ sở để hoàn thành khóa luận này.

Ngoài ra, tôi xin trân trọng cảm ơn Ban Lãnh đạo, phòng Đào tạo, các phòng chức năng của Học viện khoa học công nghệ Việt Nam đã tạo các điều kiện cho tôi được học tập và hoàn thành khóa luận một cách thuận lợi.

Tôi cũng xin gửi lời cảm ơn tới người thân, bàn bè và đồng nghiệp đã luôn ủng hộ, động viên, tạo mọi điều kiện giúp tôi hoàn thành khóa luận này.

Trong quá trình học tập và hoàn thành khóa luận, tuy đã thực hiện và học tập với một tính thần hết sức nghiêm túc nhưng chắc chắn sẽ không thẻ tránh khỏi những sai sót. Rất mong nhận được sự thông cảm và chỉ bảo tận tình đến từ thầy cô và các bạn

Hà Nội, ngày 08 tháng 08 năm 2022

Tác giả

Trần Hải Vinh

MỤC LỤC

DANH MỤC VIẾT TẮT

ANN	Artificial Neural Network
BPN	Back Propagation Neural Network
K-NN	K-Nearest Neighbors
NWP	Numerical Weather Prediction
SVM	Support Vector Machine

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

MỞ ĐẦU

Khí tượng học là môn khoa học nghiên cứu về khí quyển với mục tiêu chủ yếu là theo dõi và dự báo thời tiết. Dự báo thời tiết được thực hiện bằng cách thu thập dữ liệu liên quan đến trạng thái thời tiết hiện tại như nhiệt độ, độ ẩm, áp suất, gió,mưa, sương mù,… sau đó dựa vào các mô hình khí quyển chạy trên các siêu máy tính để đưa ra dự báo các điều kiện khí quyển trong tương lai cho một, vài giờ, hoặc nhiều ngày tiếp theo tại khu vực cụ thể. Ngày nay, do sự phát phát triển mạnh mẽ các kỹ thuật quan sát nên có rất nhiều các loại số liệu khí tượng như dữ liệu vệ tinh, các trạm mặt đất ngày càng nhiều. Việc khai thác các nguồn số liệu (kỹ thuật khai phá dữ liệu) trong lĩnh vực này vẫn còn mới và đang dần được phát triển và hoàn thiện.

Lĩnh vực dự báo thời tiết trên thế giới và Việt Nam hiện nay chủ yếu sử dụng mô hình dự báo thời tiết hiện đại như WRF (The Weather Research and Forecasting Model) với đầu vào là các bộ số liệu thời tiết như GFS, CFS,…Phương pháp dự báo số dựa trên công cụ siêu máy tính ngày một hoàn thiện. Ngày nay, các loại số liệu đầu vào cho các mô hình này rất đa dạng và nhiều loại. Các vệ tinh đời mới cho phép quan trắc các trường khí tượng ở nhiều giải phổ và với độ phân giải cao hơn. Các trạm mặt đất ngày càng dày đặc, nhiều thiết bị quan trắc mớiđược đưa vào sử dụng tạo ra nhiều loại số liệu cả đa dạng về chủng loại, chất lượng và số lượng. Những năm gần đây, các nhà khoa học đã bắt đầu quan tâm nghiên cứu việc sử dụng các kỹ thuật khai phá dữ liệu trong lĩnh vực dự báo thời và đã đạt được các kết quả nhất định [1]. Tuy nhiên các nghiên cứu cần tiếp tục hoàn thiện cả về phương pháp lẫn ứng dụng cụ thể.

Dự báo thời tiết chính xác mang một ý nghĩa rất quan trọng trong tất cả các lĩnh vực đời sống, vì vậy việc sử dụng các kỹ thuật khai phá dữ liệu trong lĩnh vực này có tính cấp thiết cao đặc biệt Việt Nam là một trong những nước bị ảnh hưởng mạnh mẽ của thiên tai và biến đổi khí hậu.

Trong quá trình nghiên cứu, luận văn sử dụng một số phương pháp:

1. Sử dụng các tài liệu được thầy hướng dẫn cung cấp, tìm kiếm tài liệu trên mạng,

2. Đọc, chọn lọc, phân tích và tổng hợp tài liệu

3. So sánh, đối chiếu, đưa ra kết luận

4. Áp dụng từng bước lý thuyết vào ứng dụng kỹ thuật khai phá dữ liệu vào một bài toán dự báo thời tiết cụ thể.

Với mục tiêu nghiên cứu, tìm hiểu các kỹ thuật khai phá dữ liệu được ứng

dụng trong dự báo thời tiết, học viên lựa chọn đề tài: "**Nghiên cứu, ứng dụng kỹ thuật khai phá dữ liệu trong dự báo một số thông số khí quyển**" với nhưng nội dung sau:

1. Tìm hiểu, nghiên cứu các kỹ thuật khai phá dữ liệu trong dự báo thời tiết

2. Phân tích bài dự báo dự báo nhiệt độ cảm nhận từ nhiệt độ và độ ẩm

3. Thực hiện xử lý số liệu sau đó ứng dụng các kỹ thuật khai phá dữ liệu vào giải quyết bài toán.

4. Đánh giá kết quả đạt được.

Luận văn sẽ đưa ra các kỹ thuật khai phá dữ liệu ứng dụng chúng vào trong lĩnh vực dự báo thời tiết giúp hoàn thiện các bộ số liệu thời tiết để có thế tăng cao chất lượng, độ chính xác của kết quả dự báo thời tiết.

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Khai phá dữ liệu

Khai phá dữ liệu (data mining) [2] là quá trình để tìm ra các mẫu hữu ích từ một lượng lớn dữ liệu. Khai phá dữ liệu cũng có thể được định nghĩa là quy trình truy xuất thông tin và kiến thức tiềm ẩn, chưa biết trước đây và hữu ích từ một lượng lớn dữ liệu nhiễu, không rõ ràng, ngẫu nhiên, không đầy đủ để ứng dụng vào trong thực tế. Khai phá dữ liệu sử dụng máy học, kỹ thuật thống kê và trực quan để khám phá, dự đoán kiến thức ở dạng dễ hiểu đối với người dùng.

Khai phá dữ liệu là một bước trong quá trình khai phá tri thức (Knowledge Discovery Process) [2], bao gồm:

- Làm sạch và tiền xử lý dữ liệu (data cleaning & preprocessing): Loại bỏ các dữ liệu nhiễu và các dữ liệu không nhất quán.

- Tích hợp dữ liệu: (Data integration): quá trình hợp nhất nhiều nguồn dữ liệu thành những kho dữ.

- Trích chọn dữ liệu (Data selection): truy xuất nhưng dữ liệu liên quan đến nhiệm vụ phân tích từ cơ sở dữ liệu

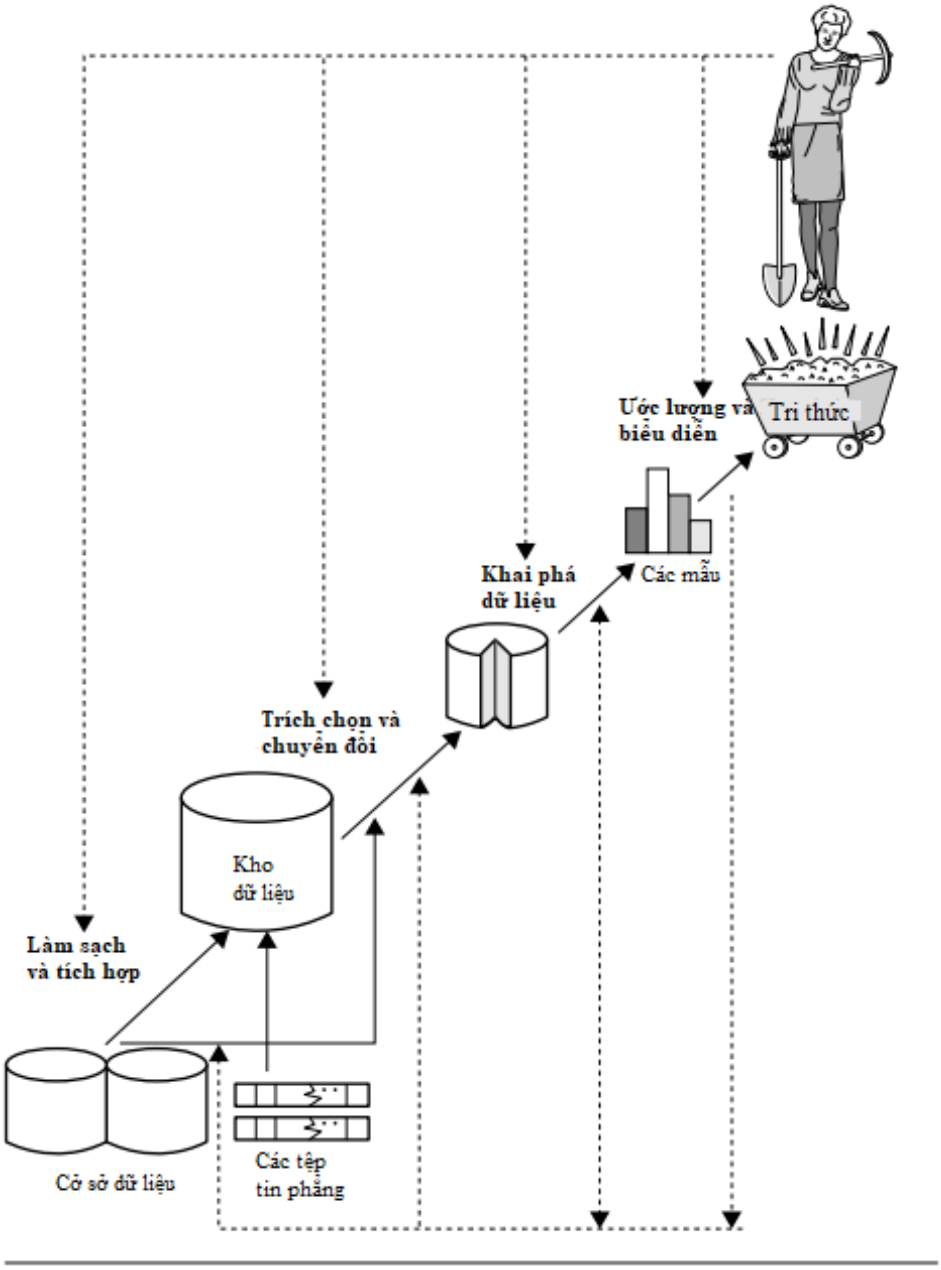
- Chuyển đổi dữ liệu (Data transformation): Dữ liệu được chuyển đổi và hợp nhất thành các dạng thích hợp để khai thác bằng các thực hiện các hoạt động tóm tắt hoặc tổng hợp

- Khai phá dữ liệu(Data mining): Là một quy trình thiết yếu, áp dụng các phương pháp thông mình để trích xuất các mẫu dữ liệu.

- Ước lượng mẫu (Patern evaluation): Là quá trình đánh giá các kết quả đặt được qua một độ đo nhất định.

- Biểu diễn tri thức (Knowledge presentation): Quá trình trình bày kiến thức đã khai phá cho người dùng bằng các kỹ thuật trực quan hóa và trình bày tri thức.

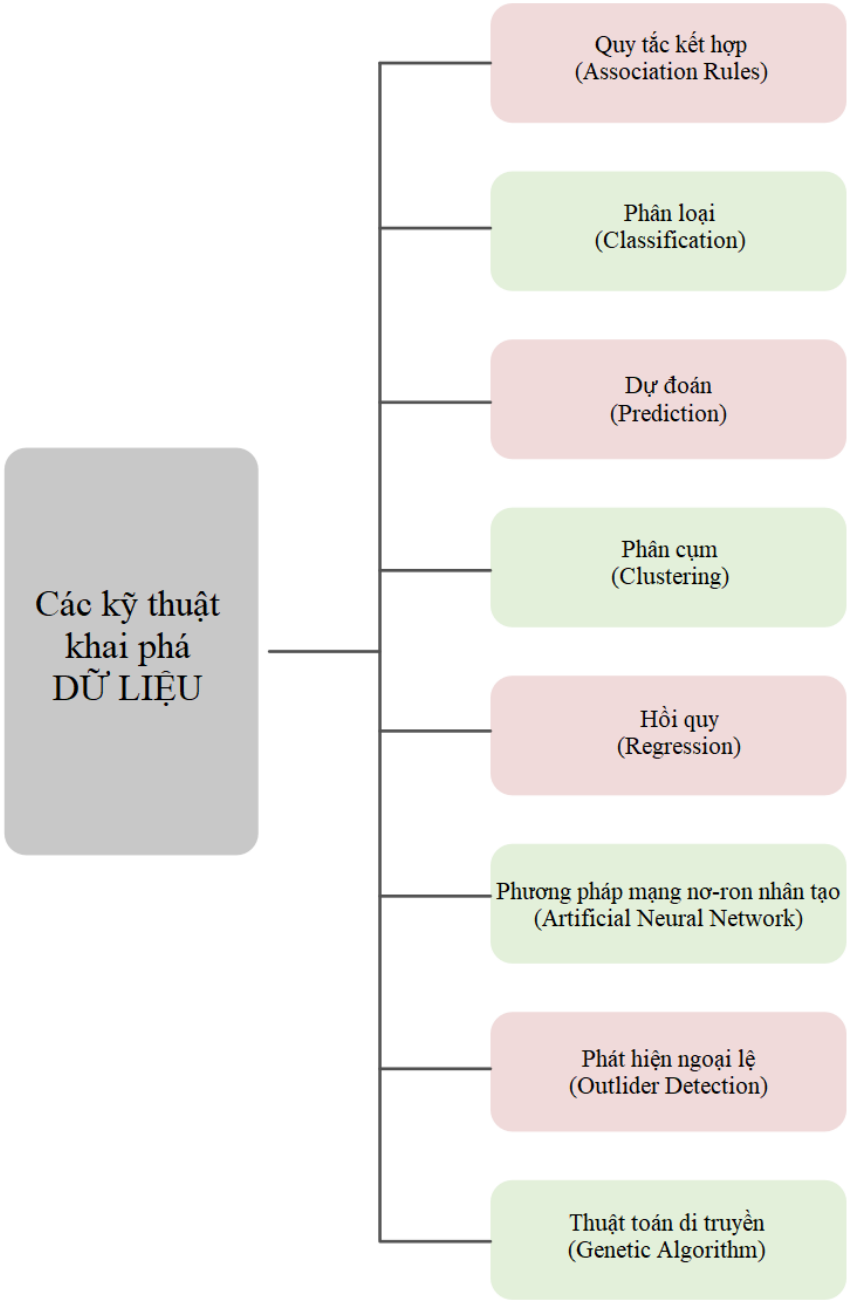
Các bước trong quá trình khai phá trí thức sẽ diễn ra tuần tự từ đầu tới cuối và lặp lại liên tục.



Hình 1.1. Quá trình khai phá tri thức [2]

1.2. Các kỹ thuật khai phá dữ liệu

Các kỹ thuật khai phá dữ liệu [2] được thể hiện dưới ảnh 1.2.



Hình 1.2. Các kỹ thuật khai phá dữ liệu

1.2.1. Quy tắc kết hợp (Association Rules)

Phân tích kết hợp là việc tìm kiếm các quy tắc kết hợp cho thấy các điều kiện thuộc tính-giá trị thường xuyên xảy ra cùng nhau trong một tập dữ liệu nhất định. Phân tích kết hợp được sử dụng rộng rãi cho thị trường hoặc phân tích dữ liệu giao dịch. Khai thác quy tắc kết hợp là một lĩnh vực nghiên cứu khai phá dữ liệu quan trọng và đặc biệt năng động. Một phương pháp phân loại dựa trên kết hợp, được gọi

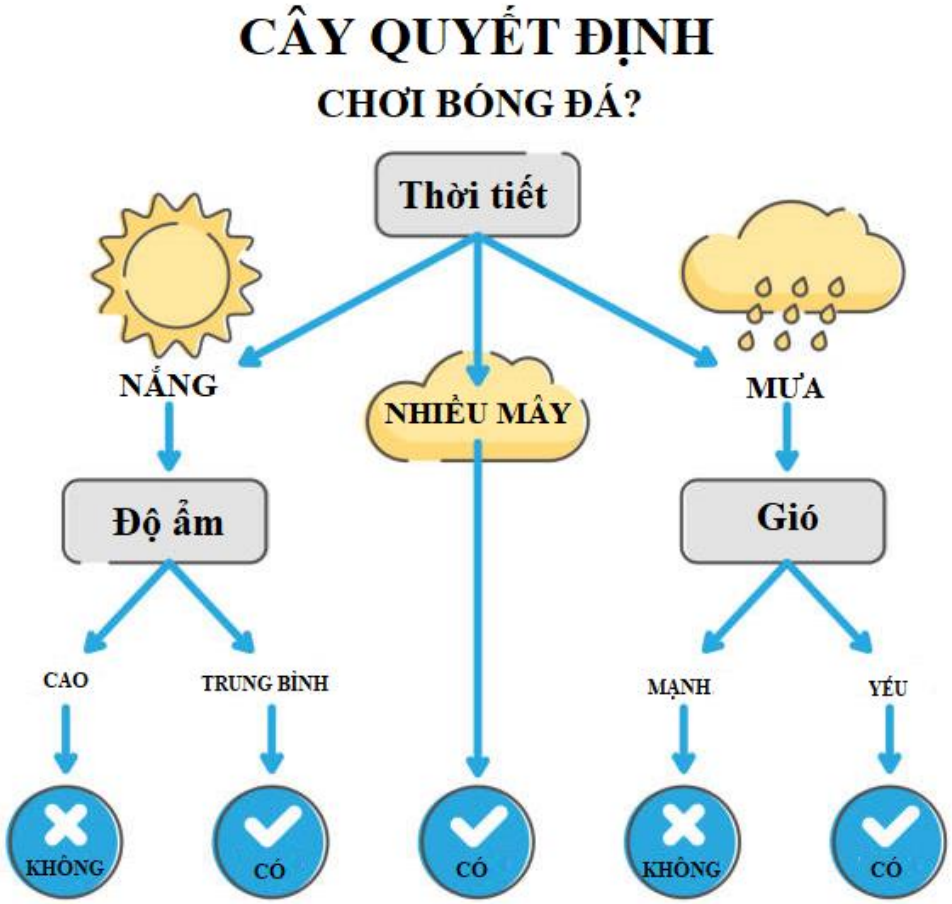
là phân loại kết hợp, bao gồm hai bước. Trong bước chính, các lệnh kết hợp được tạo bằng cách sử dụng phiên bản sửa đổi của thuật toán khai thác quy tắc kết hợp tiêu chuẩn được gọi là Apriori. Bước thứ hai xây dựng một bộ phân loại dựa trên các quy tắc kết hợp được phát hiện.

1.2.2. Phân loại (Classification)

Phân loại là quá trình xử lý tìm kiếm một tập hợp các mô hình (hoặc chức năng) mô tả và phân biệt các lớp hoặc khái niệm dữ liệu, nhằm mục đích có thể sử dụng mô hình để dự đoán lớp của các đối tượng mà nhãn lớp chưa biết. Mô hình được xác định phụ thuộc vào việc điều tra tập hợp thông tin dữ liệu huấn luyện (tức là các đối tượng dữ liệu có nhãn lớp được biết đến). Mô hình dẫn xuất có thể được biểu diễn dưới nhiều dạng khác nhau, chẳng hạn như quy tắc phân loại (nếu - thì), cây quyết định và mạng nơ-ron. Khai phá dữ liệu có một số loại phân loại khác nhau:

- Cây quyết định (Decision Tree).
- Máy Vector hỗ trợ (Support Vector Machine)
- Mô hình tuyến tính tổng quát (Generalized Linear Models)
- Phân loại Naive Bayes (Bayesian classification)
- Phân loại theo lan truyền ngược (Classification by Backpropagation)
- K láng giềng gần nhất (K- Nearest Neighbor Classifier)
- Phân loại dựa trên quy tắc (Rule-Based Classification)
- Phân loại dưa trên mẫu thường xuyên (Frequent-Pattern Based Classification)
- Hệ thống logic mờ (Fuzzy Logic)

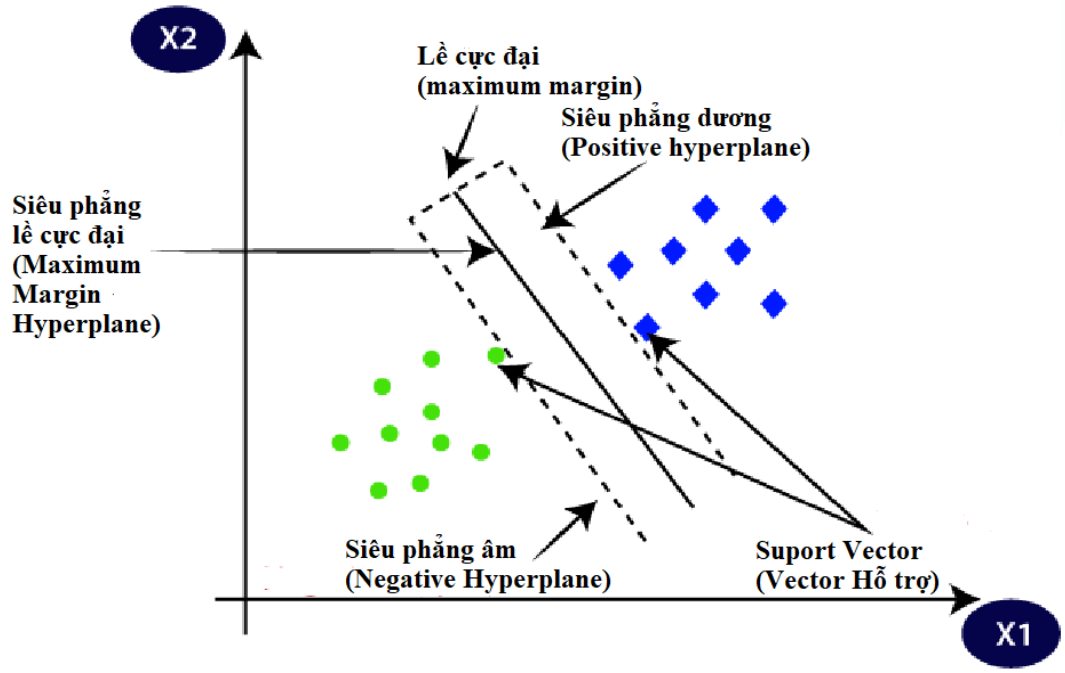
Cây quyết định: Cây quyết định là một cấu trúc cây giống như biểu đồ luồng, trong đó mỗi nút biểu thị một phép thử trên một giá trị thuộc tính, mỗi nhánh biểu thị kết quả của một phép thử và các lá cây biểu thị các lớp hoặc phân phối lớp. Cây quyết định có thể dễ dàng chuyển thành các quy tắc phân loại. Cây quyết định là một phương pháp luận phi tham số để xây dựng các mô hình phân loại. Nói cách khác, nó không yêu cầu bất kỳ giả định trước nào về loại phân phối xác suất được thỏa mãn bởi lớp và các thuộc tính khác. Cây quyết định, đặc biệt là cây có kích thước nhỏ hơn, tương đối dễ hiểu. Độ chính xác của các cây cũng có thể so sánh với hai kỹ thuật phân loại khác cho một tập dữ liệu đơn giản hơn nhiều. Chúng cung cấp một biểu diễn rõ ràng để học các hàm có giá trị rời rạc.



Hình 1.3. Ví dụ mô hình phân loại cây quyết định

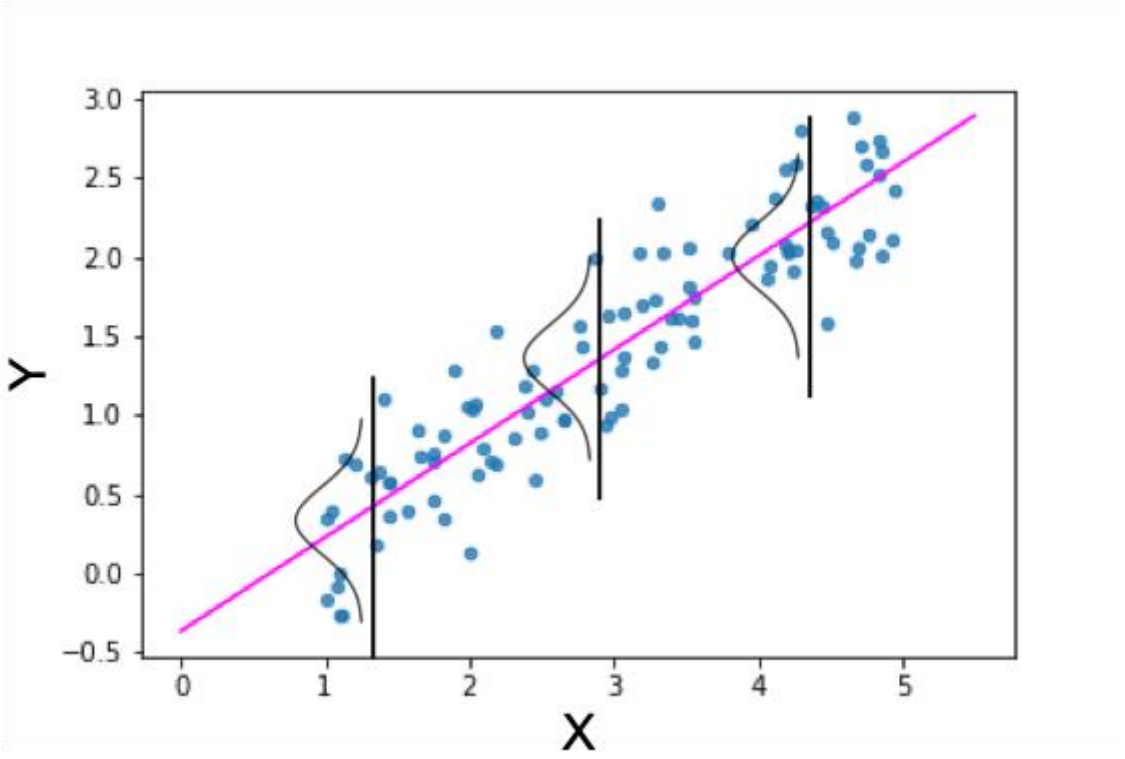
Máy Vector hỗ trợ (SVM) Phương pháp phân loại: Máy Vector hỗ trợ là một chiến lược học tập có giám sát được sử dụng để phân loại và bổ sung được sử dụng để hồi quy. Khi đầu ra của máy vectơ hỗ trợ là một giá trị liên tục, phương pháp học được yêu cầu thực hiện hồi quy; và phương pháp học sẽ dự đoán một nhãn danh mục của đối tượng đầu vào, nó được gọi là phân loại. Các biến độc lập có thể có hoặc không thể là định lượng. Phương trình lõi là các hàm biến đổi thông tin tuyến tính không phân tách được trong một miền này thành một miền khác tại bất cứ nơi nào các cá thể trở nên có thể phân chia tuyến tính. Các phương trình lõi cũng là tuyến tính, bậc hai, hoặc bất cứ thứ gì đạt được mục đích cụ thể này. Một kỹ thuật phân loại tuyến tính có thể là một bộ phân loại sử dụng một hàm tuyến tính của các đầu vào để làm cơ sở đưa ra quyết định. Việc áp dụng các phương trình lõi sẽ sắp xếp các mẫu thông tin theo cách sao cho trong các khoảng thời gian ở không gian đa chiều, có một siêu mặt phẳng ngăn cách các thể hiện tri thức của một loại này với các thể hiện khác. Ưu điểm của Máy vectơ hỗ trợ là chúng sẽ sử dụng một số hạt nhân nhất định để biến đổi vấn đề, như vậy chúng ta có thể áp dụng kỹ thuật phân loại tuyến tính cho kiến thức phi tuyến. Một khi quản lý để phân chia thông tin

thành hai lớp khác nhau, mục đích là bao gồm siêu mặt phẳng hiệu quả nhất để phân tách hai loại cá thể.



Hình 1.4. Mô hình thuật toán máy Vector hỗ trợ

Mô hình tuyến tính tổng quát là một kỹ thuật thống kê, dành cho mô hình tuyến tính. Mô hình tuyến tính tổng quát cung cấp thống kê hệ số và thống kê mô hình mở rộng, cũng như chẩn đoán hàng. Nó cũng hỗ trợ giới hạn độ tin cậy.

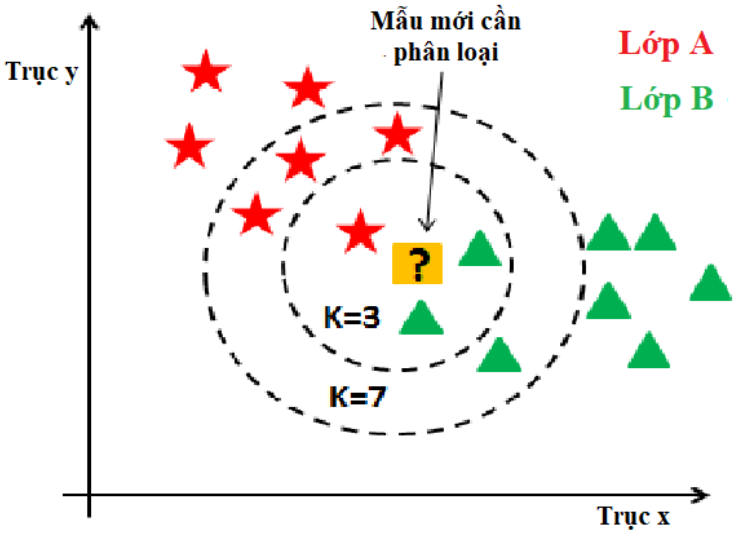


Hình 1.5. Mô hình tuyến tính tổng quát

Phân loại Bayes: Phân loại Bayes là một bộ phân loại thống kê, có thể dự đoán xác suất thành viên của lớp, chẳng hạn, xác suất mà một mẫu nhất định thuộc về một lớp cụ thể. Phân loại Bayes được tạo ra dựa trên định lý Bayes. Các nghiên cứu so sánh các thuật toán phân loại đã tìm thấy một bộ phân loại được gọi là bộ phân loại Bayes đơn giản để có thể so sánh về hiệu suất với các bộ phân loại cây quyết định và mạng nơ-ron. Bộ phân loại Bayes cũng đã hiển thị độ chính xác và tốc độ cao khi áp dụng cho cơ sở dữ liệu lớn. Bộ phân loại Naive Bayesian chấp nhận rằng giá trị thuộc tính chính xác trên một lớp nhất định là độc lập với giá trị của các thuộc tính khác. Giả định này được gọi là độc lập có điều kiện của lớp. Nó được tạo ra để đơn giản hóa các phép tính liên quan và được coi là "ngây thơ". Các mạng niềm tin Bayes là các bản sao đồ họa, không giống như các bộ phân loại Naive Bayes cho phép mô tả sự phụ thuộc giữa các tập con của các thuộc tính.

Phân loại theo lan truyền ngược: Một lan truyền ngược học bằng cách xử lý lặp đi lặp lại một tập hợp các mẫu đào tạo, so sánh ước tính của mạng cho từng mẫu với nhãn lớp thực tế đã biết. Đối với mỗi mẫu đào tạo, trọng số được sửa đổi để giảm thiểu sai số bình phương trung bình giữa dự đoán của mạng và lớp thực tế. Những thay đổi này được thực hiện theo hướng “lùi lại”, tức là từ lớp đầu ra, qua từng lớp ẩn xuống đến lớp ẩn đầu tiên (do đó có tên là backpropagation). Mặc dù nó không được đảm bảo, nhưng nói chung, các trọng số cuối cùng sẽ hội tụ, và quá trình kiến thức dừng lại.

Phương pháp phân loại K-Nearest Neighbor (K-NN): Trình phân loại k-láng giềng gần nhất được coi là một trình phân loại dựa trên ví dụ, có nghĩa là các tài liệu đào tạo được sử dụng để so sánh thay vì minh họa lớp chính xác , giống như các cấu hình lớp được sử dụng bởi các bộ phân loại khác. Do đó, không có phần đào tạo thực sự. khi một tài liệu mới phải được phân loại, k tài liệu tương tự nhất (hàng xóm) được tìm thấy và nếu một tỷ lệ đủ lớn trong số chúng được phân bổ cho một lớp chính xác, thì tài liệu mới cũng được chỉ định vào lớp hiện tại, ngược lại thì không. Ngoài ra, việc tìm kiếm những người hàng xóm gần nhất được thực hiện nhanh chóng bằng cách sử dụng các chiến lược phân loại truyền thống.



Hình 1.6. Ví dụ minh họa phân loại K-Nearest Neighbor

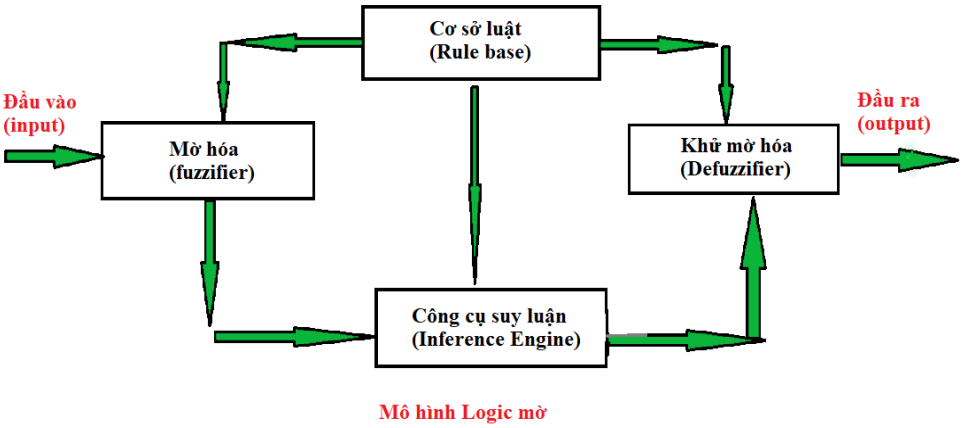
Phân loại dựa trên quy tắc: Phân loại dựa trên quy tắc biểu diễn kiến thức dưới dạng các quy tắc Nếu-Thì. Đánh giá quy tắc được đánh giá theo độ chính xác và mức độ phù hợp của trình phân loại. Nếu nhiều quy tắc được kích hoạt thì chúng ta cần giải quyết xung đột trong phân loại dựa trên quy tắc. Giải quyết xung đột có thể được thực hiện trên ba tham số khác nhau: Thứ tự kích thước, Thứ tự dựa trên lớp và thứ tự dựa trên quy tắc.

Phân loại dựa trên mẫu thường xuyên: Khám phá mẫu thường xuyên (phát hiện mẫu thường xuyên hoặc khai thác tập hợp mục thường xuyên) là một phần của khai phá dữ liệu. Nó mô tả nhiệm vụ tìm kiếm các mẫu liên quan và thường xuyên nhất trong các tập dữ liệu lớn. Ý tưởng lần đầu tiên được trình bày cho cơ sở dữ liệu giao dịch khai thác. Các mẫu thường xuyên được định nghĩa là các tập hợp con (tập hợp mục, chuỗi con hoặc cấu trúc con) xuất hiện trong tập dữ liệu với tần suất không thấp hơn ngưỡng do người dùng chỉ định hoặc tự động xác định.

Hệ thống logic mờ: Các hệ thống dựa trên quy tắc để phân loại có nhược điểm là chúng liên quan đến việc cắt giảm mạnh đối với các thuộc tính liên tục. Logic mờ rất có giá trị cho các khuôn khổ khai phá dữ liệu thực hiện phân nhóm / phân loại. Nó cung cấp lợi ích của việc làm việc ở mức độ trừu tượng cao. Nói chung, việc sử dụng logic mờ trong các hệ thống dựa trên quy tắc liên quan đến những điều sau:

Giá trị thuộc tính được thay đổi thành giá trị mờ.

Đối với một tập dữ liệu / ví dụ mới nhất định, có thể áp dụng nhiều hơn một quy tắc mờ. Mọi quy tắc hiện hành đều đóng góp một phiếu bầu cho tư cách thành viên trong các hạng mục. Thông thường, các giá trị chân lý cho mỗi danh mục dự kiến được tính tổng.



Hình 1.7. Mô hình hệ thống logic mờ

1.2.3. Dự đoán (Prediction)

Dự đoán dữ liệu là một quá trình gồm hai bước, tương tự như quá trình phân loại dữ liệu. Dù vậy, đối với dự đoán, ta không sử dụng cụm từ của "Thuộc tính nhãn lớp" vì thuộc tính mà các giá trị đang được dự đoán luôn có giá trị nhất định (có thứ tự) thay vì phân loại (có giá trị rời rạc và không có thứ tự). Thuộc tính có thể được gọi đơn giản là thuộc tính dự đoán. Dự đoán có thể được xem như việc xây dựng và sử dụng một mô hình để đánh giá lớp của một đối tượng không được gắn nhãn hoặc để đánh giá giá trị hoặc phạm vi giá trị của một thuộc tính mà một đối tượng nhất định có khả năng có.

1.2.4. Phân cụm (Clustering)

Không giống như phân loại và dự đoán, phân tích các đối tượng hoặc thuộc tính dữ liệu được gắn nhãn lớp, phân cụm phân tích các đối tượng dữ liệu mà không cần tham khảo nhãn lớp đã xác định. Nói chung, các nhãn lớp không tồn tại trong dữ liệu huấn luyện đơn giản vì chúng không được biết đến từ đầu. Phân cụm có thể được sử dụng để tạo các nhãn này. Các đối tượng được phân nhóm dựa trên nguyên tắc tối đa hóa sự tương đồng giữa các lớp và giảm thiểu sự tương tự giữa các lớp. Nghĩa là, các cụm đối tượng được tạo ra để các đối tượng bên trong một cụm có độ tương phản cao, tương phản với nhau, nhưng lại là các đối tượng khác nhau trong các cụm khác. Mỗi cụm được tạo ra có thể được xem như một lớp đối tượng, từ đó có thể suy ra các quy tắc. Phân cụm cũng có thể tạo điều kiện thuận lợi cho việc hình thành phân loại, nghĩa là, việc tổ chức các quan sát thành một hệ thống phân cấp của các lớp nhóm các sự kiện tương tự lại với nhau.

1.2.5. Hồi quy (Regression)

Hồi quy được định nghĩa là một phương pháp mô hình thống kê, trong đó dữ

liệu thu được trước đó được sử dụng để dự đoán một đại lượng liên tục cho các quan sát mới. Có hai loại mô hình hồi quy: Mô hình hồi quy tuyến tính và mô hình hồi quy tuyến tính bội:

- Hồi quy Hồi quy tuyến tính được sử dụng chủ yếu cho mục đích mô hình hóa mối quan hệ giữa hai biến đã cho. Điều này thường được thực hiện bằng cách lắp một phương trình tuyến tính để nhận thức dữ liệu. Ngoài ra, nó cũng có thể được sử dụng để tìm mối quan hệ toán học giữa các biến. Nó là dạng đơn giản nhất của hồi quy. Công thức của hồi quy tuyến tính là:

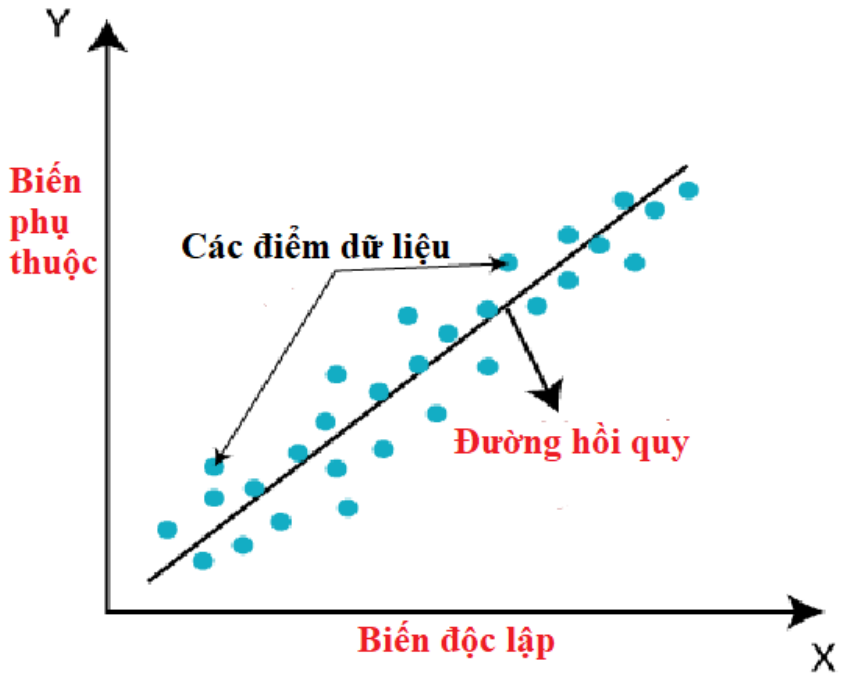
$$Y = bX + A$$

Y là mô hình của hàm tuyến tính X

b là hệ số góc của đường thẳng

A là điểm giao nhau (liên quan đến điểm X đi qua trục y).

Giá trị của Y sẽ tăng hoặc giảm theo cách mà giá trị của X sẽ thay đổi theo tuyến tính.



Hình 1.8. Đường thẳng có độ nghiêng thể hiện mối quan hệ giữa các biến trong hồi quy tuyến tính

- Hồi quy tuyến tính bội đề cập đến một kỹ thuật thống kê sử dụng hai hoặc nhiều biến độc lập để dự đoán kết quả của một biến phụ thuộc. Kỹ thuật này cho phép các nhà phân tích xác định sự thay đổi của mô hình và sự đóng góp tương đối của mỗi biến độc lập trong tổng phương sai. Hồi quy bội có thể có hai dạng là hồi quy tuyến tính và hồi quy phi tuyến tính. Công thức của hồi quy tuyến tính bội:

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta pxip + \varepsilon$$

yi là biến phụ thuộc hoặc biến dự đoán

$\beta 0$ là giao của y, tức là giá trị của y khi cả xi1 và xi2 đều bằng 0.

$\beta 1$ và $\beta 2$ là các hệ số hồi quy biểu thị sự thay đổi của y so với sự thay đổi một đơn vị trong xi1 và xi2, tương ứng.

βp là hệ số góc cho mỗi biến độc lập

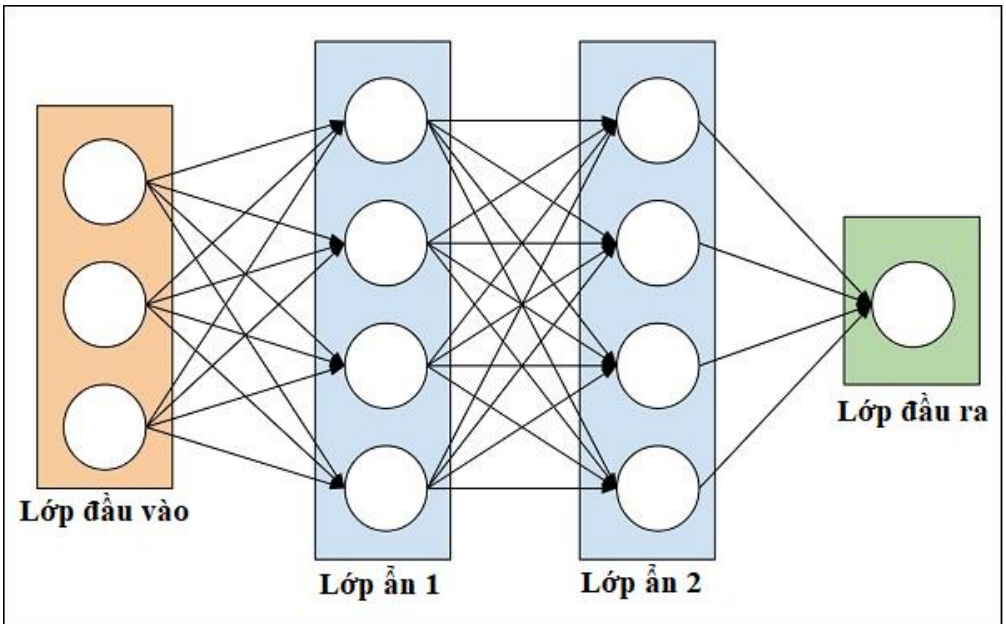
ε là số hạng sai số ngẫu nhiên (phần dư) của mô hình.

1.2.6. Phương pháp mạng nơ-ron nhân tạo (Artificial Neural Network)

Mạng nơ-ron nhân tạo (ANN) còn được gọi đơn giản là "Mạng nơ-ron" (Neural Network), là một mô hình quy trình được hỗ trợ bởi mạng nơ-ron sinh học. Nó bao gồm một tập hợp các tế bào thần kinh nhân tạo được kết nối với nhau. Mạng nơ-ron là một tập hợp các đơn vị đầu vào / đầu ra được kết nối trong đó mỗi kết nối có trọng số liên quan đến nó. Trong giai đoạn kiến thức, mạng thu nhận bằng cách điều chỉnh các trọng số để có thể dự đoán nhãn lớp chính xác của các mẫu đầu vào. Học tập mạng nơ-ron cũng được biểu thị là học tập kết nối do sự kết nối giữa các đơn vị. Mạng nơ-ron đòi hỏi thời gian đào tạo dài, do đó phương pháp này thích hợp hơn cho các ứng dụng có thể đáp ứng được điều này. Chúng yêu cầu một số tham số thường được xác định tốt nhất theo kinh nghiệm, chẳng hạn như cấu trúc liên kết mạng hoặc "cấu trúc". Mạng nơ-ron đã bị chỉ trích vì khả năng diễn giải kém do con người khó có thể hiểu được ý nghĩa biểu tượng đằng sau các trọng số đã học. Những tính năng này làm cho mạng nơ-ron ít được sử dụng hơn cho việc khai phá dữ liệu.

Tuy nhiên, ưu điểm của mạng nơ-ron là khả năng chịu đựng dữ liệu nhiễu cao cũng như khả năng phân loại các mẫu mà chúng chưa được đào tạo. Ngoài ra, một số thuật toán mới được phát triển để trích xuất các quy tắc từ các mạng nơ-ron được đào tạo. Những vấn đề này góp phần vào tính hữu ích của mạng nơ-ron để phân loại trong khai phá dữ liệu.

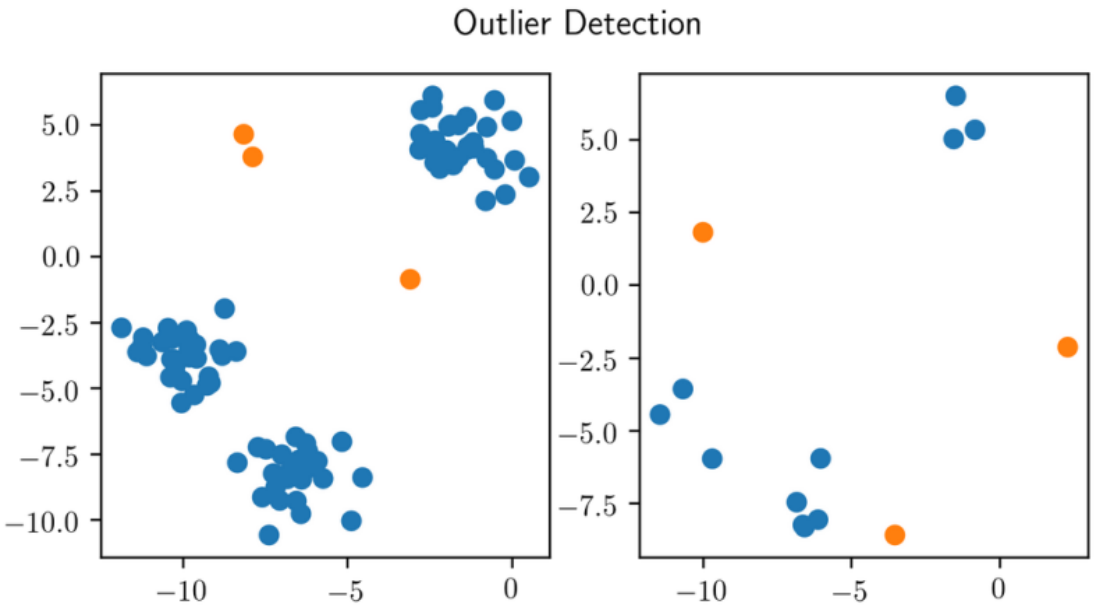
Mạng nơ-ron nhân tạo là một hệ thống phụ thuộc thay đổi thông tin được hỗ trợ cấu trúc của nó chảy qua mạng nhân tạo trong một phần học. ANN dựa trên nguyên tắc học hỏi bằng cách làm gương. Có hai loại mạng nơ-ron: Mạng nơ ron cổ điển (Perceptron) và mạng nơ ron truyền thẳng nhiều lớp (Multi-layer Perceptron).



Hình 1.9. Biểu diễn minh họa cho Mạng nơ-ron nhân tạo

1.2.7. Phát hiện ngoại lệ (Outlider Detection)

Cơ sở dữ liệu có thể chứa các đối tượng dữ liệu không tuân thủ hành vi hoặc mô hình chung của dữ liệu. Các đối tượng dữ liệu này là Ngoại lệ. Việc điều tra dữ liệu ngoại lệ được gọi là khai thác ngoại lệ (Outlider Mining). Một ngoại lệ có thể được phát hiện bằng cách sử dụng các thử nghiệm thống kê giả định mô hình phân phối hoặc xác suất cho dữ liệu hoặc sử dụng các phép đo khoảng cách trong đó các đối tượng có một phần nhỏ các hàng xóm "gần gũi" trong không gian được coi là ngoại lệ. Thay vì sử dụng các thước đo thực tế hoặc khoảng cách, các kỹ thuật dựa trên độ lệch phân biệt các ngoại lệ / ngoại lệ bằng cách kiểm tra sự khác biệt trong các thuộc tính nguyên tắc của các đơn vị trong một nhóm.



Hình 1.10. Minh họa kỹ thuật phát hiện ngoại lệ

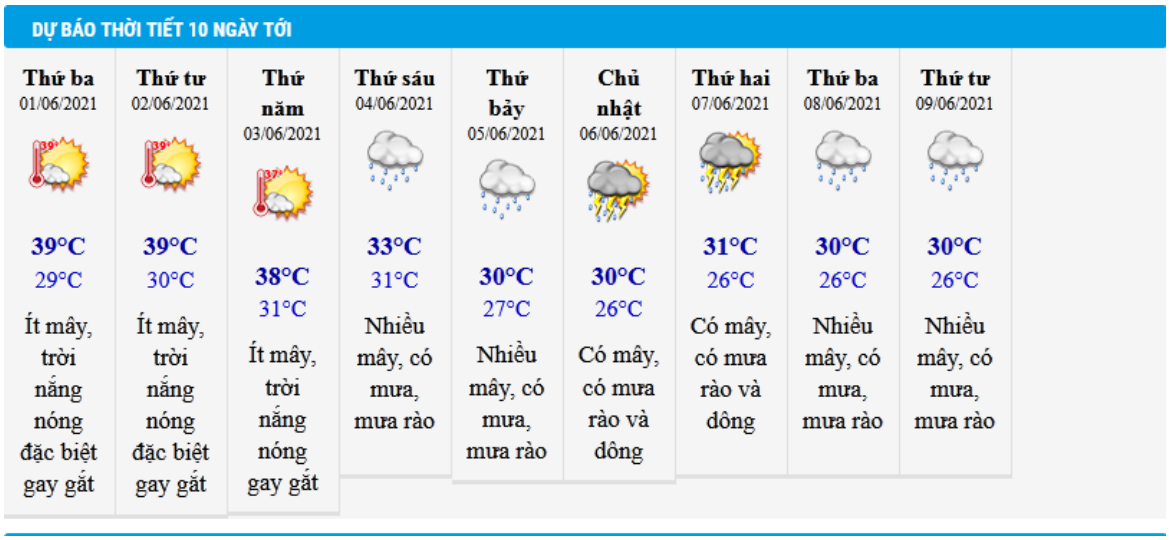
1.2.8. Thuật toán di truyền (Genetic Algorithm)

Các thuật toán di truyền là các thuật toán tìm kiếm thích ứng thuộc về phần lớn hơn của các thuật toán tiến hóa. Các thuật toán di truyền dựa trên ý tưởng của chọn lọc tự nhiên và di truyền. Đây là những khai thác thông minh của tìm kiếm ngẫu nhiên được cung cấp dữ liệu lịch sử để hướng tìm kiếm vào khu vực có hiệu suất tốt hơn trong không gian giải pháp. Chúng thường được sử dụng để tạo ra các giải pháp chất lượng cao cho các vấn đề tối ưu hóa và các vấn đề tìm kiếm. Các thuật toán di truyền mô phỏng quá trình chọn lọc tự nhiên, có nghĩa là những loài nào có thể thích nghi với những thay đổi của môi trường thì có thể tồn tại và sinh sản và chuyển sang thế hệ tiếp theo. Nói một cách đơn giản, chúng mô phỏng "sự sống sót của những người khỏe nhất" giữa các cá nhân thuộc các thế hệ liên tiếp để giải quyết một vấn đề. Mỗi thế hệ bao gồm một quần thể các cá thể và mỗi cá nhân đại diện cho một điểm trong không gian tìm kiếm và giải pháp khả thi. Mỗi cá nhân được biểu diễn dưới dạng một chuỗi ký tự / số nguyên / số thức / các bit. Chuỗi này tương tự như Nhiễm sắc thể.

1.3. Dự báo thời tiết

Dự báo thời tiết [3] là ứng dụng của khoa học và công nghệ để dự đoán các điều kiện của khí quyển cho một địa điểm và thời gian nhất định. Dự báo thời tiết được thực hiện bằng cách thu thập dữ liệu định lượng về trạng thái hiện tại của khí quyển, đất liền và đại dương và sử dụng khí tượng học để dự đoán khí quyển sẽ thay đổi như thế nào tại một địa điểm nhất định.

Trước đây, dự báo thời tiết được tính toán thủ công chủ yếu dựa trên những thay đổi về áp suất không khí, điều kiện thời tiết hiện tại và tình trạng bầu trời hoặc mây che phủ, dự báo thời tiết giờ đây dựa vào các mô hình dựa trên máy tính có tính đến nhiều yếu tố khí quyển. Con người vẫn được yêu cầu để chọn mô hình dự báo tốt nhất có thể để làm cơ sở cho dự báo, bao gồm các kỹ năng nhận dạng mẫu, kết nối từ xa, kiến thức về hiệu suất mô hình và kiến thức về sai lệch của mô hình. Sự thiếu chính xác của dự báo là do tính chất hỗn loạn của khí quyển, thiếu các thiết bị có hiệu suất cao để giải các phương trình mô tả khí quyển, đất liền và đại dương, sai số liên quan đến việc đo các điều kiện ban đầu, sự hiểu biết không đầy đủ về khí quyển và các quy trình liên quan. Do đó, các dự báo trở nên kém chính xác hơn khi chênh lệch giữa thời gian hiện tại và thời gian dự báo được thực hiện (phạm vi dự báo) tăng lên. Việc sử dụng các tập hợp và sự đồng thuận của mô hình giúp thu hẹp sai số và cung cấp mức độ tin cậy trong dự báo.



Hình 1.11. Minh họa dự báo thời tiết cho khu vực Hà Nội trong 10 ngày

Có một số phương pháp khác nhau được sử dụng để dự báo thời tiết [4]:

- Dự báo thời tiết khái quát (Synoptic weather prediction): Đây là cách tiếp cận truyền thống trong dự báo thời tiết. Sơ đồ khái quát đề cập đến việc quan sát các yếu tố thời tiết khác nhau trong thời gian quan sát cụ thể. Để theo dõi sự thay đổi của thời tiết, một trung tâm khí tượng chuẩn bị một loạt các biểu mẫu khái quát hàng ngày, đây là biểu mẫu rất cơ bản của dự báo thời tiết. Nó liên quan đến việc thu thập và phân tích khổng lồ dữ liệu quan sát thu được từ hàng nghìn trạm thời tiết.

- Dự báo thời tiết dạng số (Numerical weather prediction): Phương pháp này sử dụng sức mạnh của máy tính để dự báo thời tiết. Các chương trình máy tính phức tạp được chạy trên siêu máy tính và đưa ra các dự đoán về nhiều thông số khí quyển. Một lỗ hổng là các phương trình được sử dụng liên kết chặt chẽ với nhau. Nếu không hoàn toàn biết được giai đoạn ban đầu của thời tiết thì dự đoán sẽ không hoàn toàn chính xác.

- Dự báo thời tiết thống kê (Statistical weather prediction): Phương pháp này thường được sử dụng cùng với các phương pháp dạng số. Nó sử dụng các bản ghi dữ liệu thời tiết trong quá khứ với giả định rằng tương lai sẽ là sự lặp lại của thời tiết trong quá khứ. Mục đích chính là để tìm ra những khía cạnh thời tiết là những chỉ báo tốt về các sự kiện trong tương lai. Chỉ có thể dự báo thời tiết tổng thể theo cách này.

Dự báo thời tiết được ứng dụng vào một số lĩnh vực:

- Hàng không: Do ngành hàng không đặc biệt nhạy cảm với thời tiết nên việc dự báo thời tiết chính xác là điều cần thiết. Sương mù hoặc trần mây đặc biệt thấp có thể ngăn cản máy bay hạ cánh và cất cánh. Sự xáo trộn và đóng băng cũng là những mối nguy hiểm đáng kể trong chuyến bay. Sấm sét là một vấn đề đối với tất

cả các máy bay, đóng băng do lượng mưa lớn, cũng như mưa đá lớn, gió mạnh và sét, tất cả đều có thể gây ra thiệt hại nghiêm trọng cho một máy bay đang bay. Hàng ngày, các máy bay được định tuyến để tận dụng lợi thế của luồng gió phụ để cải thiện hiệu quả sử dụng nhiên liệu. Các phi hành đoàn được thông báo ngắn gọn trước khi cất cánh về các điều kiện dự kiến trên đường bay và tại điểm đến của họ. Ngoài ra, các sân bay thường thay đổi đường băng nào đang được sử dụng để tận dụng chiều gió. Điều này làm giảm khoảng cách cần thiết để cất cánh và loại bỏ các luồng gió chéo tiềm ẩn.

- Hàng hải: Việc sử dụng đường thủy cho mục đích thương mại và giải trí có thể bị hạn chế đáng kể bởi hướng và tốc độ gió, chu kỳ sóng và độ cao, thủy triều và lượng mưa. Các yếu tố này đều có thể ảnh hưởng đến sự an toàn của việc vận chuyển hàng hải. Do đó, nhiều loại mã đã được thiết lập để truyền một cách hiệu quả các dự báo thời tiết biển chi tiết cho các hoa tiêu tàu qua radio, ví dụ như MAFOR (dự báo hàng hải). Dự báo thời tiết điển hình có thể được nhận trên biển thông qua việc sử dụng RTTY, Navtex và Radiofax.

- Nông nghiệp: Người nông dân dựa vào dự báo thời tiết để quyết định công việc cần làm vào bất kỳ ngày cụ thể nào. Ví dụ, việc làm phơi cỏ khô chỉ khả thi khi thời tiết khô ráo. Thời gian khô hạn kéo dài có thể làm hỏng cây bông, lúa mì và ngô. Sương giá và đóng băng tàn phá mùa màng cả trong mùa xuân và mùa thu. Ví dụ, những cây đào đang nở rộ có thể bị tàn lụi bởi đợt lạnh vào mùa xuân. Vườn cam có thể bị thiệt hại đáng kể trong thời gian sương giá và đóng băng, bất kể thời điểm nào của cây.

- Lâm nghiệp: Dự báo thời tiết về gió, lượng mưa và độ ẩm là điều cần thiết để ngăn ngừa và kiểm soát cháy rừng. Các chỉ số khác nhau, như Chỉ số thời tiết cháy rừng và Chỉ số Haines, đã được phát triển để dự đoán các khu vực có nguy cơ xảy ra cháy do nguyên nhân tự nhiên hoặc do con người gây ra. Điều kiện phát triển của côn trùng có hại cũng có thể được dự đoán bằng cách dự báo diễn biến của thời tiết.

- Và rất nhiều lĩnh vực khác nữa trong cuộc sống.

1.4. Kết chương

Chương trên đã trình bày các lý thuyết về lĩnh vực khai phá dữ liệu, các kỹ thuật khai phá dữ liệu và lĩnh vực dự báo thời tiết. Các lý thuyết này sẽ rất hữu ích trong việc "**Ứng dụng các kỹ thuật khai phá dữ liệu trong dự báo một số thông số khí quyển**". Ở chương tiếp theo, việc ứng dụng các kỹ thuật khai phá dữ liệu trong dự báo các thông số khí quyển sẽ được nói đến qua các bài báo, nghiên cứu của các tác giả trên khắp thế giới.

CHƯƠNG 2: CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU ỨNG DỤNG TRONG DỰ BÁO THỜI TIẾT

2.1. Các kỹ thuật khai phá dữ liệu được ứng dụng phổ biến trong dự báo thời tiết

Trên thế giới hiện nay đã có rất nhiều bài báo, nghiên cứu về ứng dụng của các kỹ thuật khai phá dữ liệu để dự báo thời tiết. Điển hình trong số đó có một số kỹ thuật được sử dụng phổ biến:

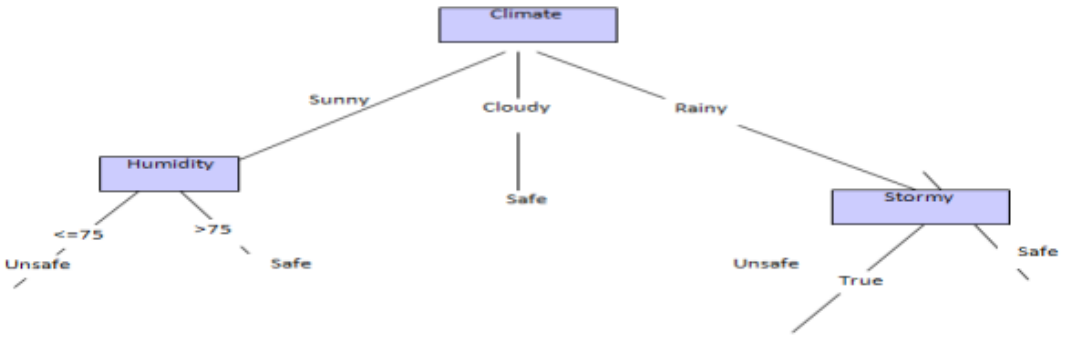
- Phân cụm (Clustering)
- Hồi quy (Regression)
- Mạng nơ-ron nhân tạo (ANN)
- Phân loại cây quyết định (Decision Tree)
- Phối hợp hai hoặc nhiều kỹ thuật khai phá dữ liệu

Một số bài báo, nghiên cứu sử dụng các kỹ thuật khai phá dữ liệu trong dự báo các thông số thời tiết

2.1.1. Cây quyết định

E.G.Petre [5] đã trình bày một ứng dụng nhỏ của thuật toán cây quyết định CART để dự báo thời tiết. Dữ liệu được ghi lại từ năm 2002 đến 2005 trên khắp Hồng Kông. Dữ liệu được sử dụng để tạo tập dữ liệu bao gồm các thông số năm, tháng, áp suất trung bình, độ ẩm tương đối, lượng mây, lượng mưa và nhiệt độ trung bình. WEKA, phần mềm khai phá dữ liệu mã nguồn mở, được sử dụng để thực hiện thuật toán cây quyết định CART. Cây quyết định, kết quả và thông tin thống kê về dữ liệu được sử dụng để tạo ra mô hình quyết định cho việc dự báo thời tiết. Việc chuyển đổi dữ liệu được yêu cầu theo thuật toán cây quyết định để WEKA sử dụng hiệu quả cho việc dự báo thời tiết.

P.Hemalatha [6] đã thực hiện các phương pháp khai phá dữ liệu để hướng dẫn đường đi của các con tàu trong quá trình ra khơi. Hệ thống Định vị Toàn cầu được sử dụng để xác định khu vực mà tàu hiện đang di chuyển. Các thuộc tính của dữ liệu thời tiết bao gồm khí hậu, độ ẩm, nhiệt độ, bão. Báo cáo thời tiết của khu vực được theo dõi được so sánh với cơ sở dữ liệu hiện có. Tập dữ liệu được phân tích được cung cấp cho thuật toán cây quyết định, C4.5 và ID3. Quyết định thu được về điều kiện thời tiết được hướng dẫn cho con tàu và con đường được chọn cho phù hợp. Phối hợp chặt chẽ giữa thống kê và tính toán cung cấp sức mạnh tổng hợp trong phân tích dữ liệu.



Hình 2.1. Cây quyết định minh họa cho bài toán của P.Hemalatha [6]

Soo-Yeon Ji và cộng sự [7] đã dự đoán thời gian lượng mưa hàng giờ ở bất kỳ khu vực địa lý nào một cách hiệu quả. Đầu tiên sẽ là xác định khả năng có mưa. Sau đó, chỉ khi có bất kỳ khả năng nào sẽ có mưa, dự đoán lượng mưa hàng giờ được thực hiện. Mặc dù có khá nhiều phương pháp luận đã được giới thiệu để dự đoán theo giờ, nhưng hầu hết chúng đều có những hạn chế về hiệu suất vì sự tồn tại của nhiều biến thể trong dữ liệu và số lượng dữ liệu hạn chế. CART và C4.5 được sử dụng để cung cấp kết quả, có thể cung cấp các mẫu ẩn và quan trọng với lý do rõ ràng. Khoảng 18 biến được sử dụng từ trạm thời tiết. Đối với mục đích xác nhận, phương pháp xác nhận chéo 10 lần được thực hiện. CART cho hiệu suất tốt hơn một chút so với C4.5.

2.1.2. Mạng nơ-ron nhân tạo

Gaurav J.Sawale và Sunil R.Gupta [8] đã đề xuất một phương pháp mạng nơ-ron nhân tạo để dự báo thời tiết tại một địa điểm nhất định. Mạng Nơ-ron lan truyền ngược (BNP) được sử dụng để tạo mô hình ban đầu. Sau đó, Mạng Hopfield được cung cấp với kết quả được xuất ra bởi mô hình BPN. Các thuộc tính bao gồm nhiệt độ, độ ẩm và tốc độ gió. Dữ liệu về thời tiết trong ba năm được thu thập bao gồm 15000 mẫu. Sai số dự đoán là rất ít và quá trình học tập diễn ra nhanh chóng. Đây có thể được coi là một giải pháp thay thế cho các phương pháp tiếp cận khí tượng truyền thống. Cả hai thuật toán được kết hợp một cách hiệu quả. Nó có thể xác định mối quan hệ phi tuyến tính tồn tại giữa các thuộc tính dữ liệu lịch sử và dự báo thời tiết trong tương lai.

A.R.W.M.M.S.C.B. Amarakoon [9] đã đề xuất một hệ thống sử dụng dữ liệu thời tiết lịch sử và áp dụng thuật toán khai phá dữ liệu KNN để phân loại các dữ liệu lịch sử này thành một khoảng thời gian cụ thể. Các khoảng thời gian gần nhất sau đó được tiếp tục lấy để dự báo thời tiết của Sri Lanka. Dữ liệu thời tiết hàng ngày được thu thập trong vòng một năm. Nó tạo ra kết quả chính xác trong một khoảng thời gian hợp lý trước hàng tháng. Kết luận rằng KNN có lợi cho dữ liệu động, dữ

liệu thay đổi hoặc cập nhật nhanh chóng và cung cấp hiệu suất tốt hơn so với các kỹ thuật khác. Tích hợp các kỹ thuật lựa chọn tính năng thậm chí có thể cho kết quả chính xác hơn.

2.1.3. Phân cụm

M.Kalyankar và S.Alaspurkar [10] đã sử dụng các kỹ thuật khai phá dữ liệu để thu thập dữ liệu thời tiết và tìm ra các mẫu ẩn bên trong tập dữ liệu lớn để chuyển thông tin thu được thành kiến thức có thể sử dụng để phân loại và dự đoán tình trạng thời tiết. Quy trình khai phá dữ liệu được áp dụng để trích xuất kiến thức từ tập dữ liệu thời tiết của thành phố Gaza. Kiến thức này có thể được sử dụng để có được những dự đoán hữu ích và hỗ trợ quá trình ra quyết định. Cần phải xây dựng các phương pháp khai phá dữ liệu động, có thể học động để phù hợp với tính chất thời tiết thay đổi nhanh chóng và đột ngột.

K.Pabreja [11] đã chứng minh nguồn gốc của các hệ thống thời tiết quy mô lưới phụ từ các sản phẩm đầu ra của mô hình Dự báo thời tiết dạng số (NWP) bằng cách sử dụng các kỹ thuật khai phá dữ liệu mà kỹ thuật MOS thông thường không thể thực hiện được. Kỹ thuật khai phá dữ liệu, phân cụm, khi được áp dụng trên sự phân kỳ và độ ẩm tương đối có thể cung cấp dấu hiệu sớm về sự hình thành của đám mây. K-mean clustering là phân cụm được sử dụng cho dữ liệu hai ngày của trường hợp mưa to trong đời thực. Một nỗ lực được thực hiện nhằm cung cấp thông tin kịp thời và có thể hành động về các sự kiện này bằng cách sử dụng các kỹ thuật khai phá dữ liệu bổ sung cho các mô hình NWP. Nhược điểm là nó không thể được sử dụng cho các dự đoán dài hạn

2.1.4. Hồi quy

P.Dutta và H.Tahbilder [12] đã dự đoán lượng mưa hàng tháng của Assam bằng cách sử dụng kỹ thuật khai phá dữ liệu. Kỹ thuật thống kê truyền thống - Hồi quy tuyến tính bội được sử dụng. Dữ liệu bao gồm 6 năm từ 2007 đến 2012 được thu thập cục bộ từ Trung tâm Khí tượng Khu vực, Guwahati, Assam, Ấn Độ. Dữ liệu được chia thành bốn tháng cho mỗi mùa. Các thông số được chọn cho mô hình là nhiệt độ tối thiểu, nhiệt độ tối đa, áp suất mực nước biển trung bình, tốc độ gió và lượng mưa. Hiệu suất của mô hình này được đo bằng hệ số xác định R bình phương hiệu chỉnh được thực hiện trong C#. Một số thông số như hướng gió không được đưa vào do những hạn chế trong việc thu thập dữ liệu làm cho dữ liệu chưa được chính xác. Mô hình dự đoán dựa trên hồi quy tuyến tính bội đưa ra kết quả độ chính xác có thể chấp nhận được.

N.Khandelwal và R.Davey [13] đã dự đoán lượng mưa của một năm bằng cách sử dụng 4 yếu tố khí hậu khác nhau là nhiệt độ, độ ẩm, áp suất và mực nước biển và từ đó sử dụng tập dữ liệu để tính toán các khả năng hạn hán ở Rajasthan. Một số yếu tố được trích xuất bằng cách sử dụng các kỹ thuật khai phá dữ liệu. Sau đó, phân tích tương quan được áp dụng trên tập dữ liệu và mối tương quan được tìm thấy trong các yếu tố. Các yếu tố có mối tương quan thuận được lựa chọn và sử dụng để phân tích hồi quy. Hồi quy tuyến tính bội được sử dụng để dự đoán lượng mưa. Sau đó, phân tích thống kê được áp dụng trên dữ liệu đó để tìm ra khả năng hạn hán. Đối với độ lệch chuẩn về khả năng hạn hán, phương sai của hệ số, chỉ số hạn hán và cảm nhận về hạn hán được sử dụng. Lượng mưa chỉ là một tham số được xem xét để phân tích tình trạng khô hạn trong khi các yếu tố khí hậu khác có thể ảnh hưởng đến tình trạng này trên một phạm vi rộng. Do đó kết quả đưa ra không được chính xác lắm.

2.1.5. Phối hợp hai hoặc nhiều kỹ thuật

S.Kannan và S.Ghosh [14] đã đóng góp vào việc phát triển phương pháp luận để dự đoán trạng thái lượng mưa ở các quy mô khác nhau cho một lưu vực sông từ dữ liệu khí hậu quy mô lớn. Một mô hình dựa trên kỹ thuật K-mean clustering kết hợp với thuật toán cây quyết định, CART, được sử dụng để tạo các trạng thái mưa từ các biến khí quyển quy mô lớn trong lưu vực sông. Trạng thái lượng mưa hàng ngày được lấy từ dữ liệu lịch sử lượng mưa hàng ngày của nhiều địa điểm bằng cách sử dụng K-mean clustering. Các biện pháp tính hợp lệ của cụm khác nhau được áp dụng cho dữ liệu lượng mưa quan sát được để có được số lượng cụm tối ưu. CART được sử dụng để đào tạo dữ liệu về trạng thái lượng mưa hàng ngày của lưu vực sông trong 33 năm. Phương pháp này được thử nghiệm cho sông Mahanadi ở Ấn Độ. Thuật toán CART tỏ ra tốt trong việc dự đoán trạng thái lượng mưa hàng ngày ở lưu vực sông bằng cách sử dụng phương pháp giảm tỷ lệ thống kê.

Olaiya Folorunsho and A.B.Adeyemo [15] đã nghiên cứu việc sử dụng các kỹ thuật khai phá dữ liệu trong việc dự đoán nhiệt độ tối đa, lượng mưa, lượng bốc hơi và tốc độ gió. Thuật toán cây quyết định C4.5 và mạng nơ-ron nhân tạo được sử dụng để dự đoán. Dữ liệu khí tượng được thu thập từ năm 2000 đến năm 2009 từ thành phố Ibadan, Nigeria. Một mô hình dữ liệu cho dữ liệu khí tượng được phát triển và được sử dụng để đào tạo các thuật toán phân loại. Hiệu suất của mỗi thuật toán được so sánh với các chỉ số hiệu suất tiêu chuẩn và thuật toán có kết quả tốt nhất được sử dụng để tạo ra các quy tắc phân loại cho các biến thời tiết. Một mô

hình mạng nơ-ron dự đoán cũng được phát triển để dự báo thời tiết và kết quả được so sánh với dữ liệu thời tiết thực tế trong khoảng thời gian dự đoán. Kết quả cho thấy rằng với đủ dữ liệu đào tạo, kỹ thuật khai phá dữ liệu có thể được sử dụng hiệu quả cho các nghiên cứu dự báo thời tiết và biến đổi khí hậu.

2.2. So sánh các kỹ thuật khai phá dữ liệu được ứng dụng trong dự báo thời tiết

Từ các kỹ thuật khai phá dữ liệu ứng dụng trong dự báo thời tiết được trình bày qua các bài báo, nghiên cứu trước đây, có thể đưa ra một số nhận xét, so sánh như sau.

- Mỗi kỹ thuật khai phá dữ liệu khác nhau được sử dụng trong các bài toán dự đoán các thông số khác nhau của thời tiết như độ ẩm, nhiệt độ, cường độ gió tùy theo yêu cầu của bài toán

- Các kỹ thuật khai phá dữ liệu khác nhau sẽ mang lại các kết quả khác nhau và mỗi kỹ thuật đều có ưu, nhược điểm của chúng. Do vây, việc chọn kỹ thuật khai phá dữ liệu phù hợp với yêu cầu bài toán là một yếu tố tiên quyết.

- Từ những bài báo, nghiên cứu ứng dụng kỹ thuật khai phá dữ liệu trong dự báo thời tiết ở trên, ta có bảng 2.1 so sánh sau:

Bảng 2.1. Bảng so sánh các kỹ thuật khai phá dữ liệu được ứng dụng trong dự báo thời tiết [16]

Tác giả	Ứng dụng	Kỹ thuật khai phá dữ liệu	Thuật toán	Các thuộc tính	Khoảng thời gian	Kích cỡ bộ dữ liệu	Độ chính xác	Ưu điểm	Nhược điểm
E. G. Petre [4]	Dự báo thời tiết	Cây quyết định	CART	Áp suất, độ ẩm lượng mây, lượng mưa, nhiệt độ	4 năm	48 mẫu	83%	Độ chính xác của dự đoán tốt	Bắt buộc chuyển đổi dữ liệu. Yêu cầu tính toán thêm
P.Hemalatha [5]	Dự báo thời tiết cho tàu thuyền ra khơi	Cây quyết định	C4.5, ID3	Khí hậu, độ ẩm, mây, mưa, nhiệt độ	4-5 khu vực	20 – 30 mẫu		Hiệu suất có thể kiểm chứng	Không xử lý trực tiếp dữ liệu phạm vi liên tục.
Soo-Yeon Ji và cộng sự [6]	Dự đoán lượng mưa hàng giờ	Cây quyết định	C4.5, CART	Nhiệt độ, hướng gió, tốc độ gió, gió giật, độ ẩm, áp suất	3 năm	26280 mẫu	99%, 93%	Độ chính xác cao	Vẫn chưa khai thác được dữ liệu nhỏ để dự đoán
Gaurav J. Sawale và Sunil R. Gupta [7]	Dự báo thười tiết chung	Mạng nơ-ron nhân tạo	BPN, Mạng Hopfiled	Nhiệt độ, độ ẩm, tốc độ gió	3 năm	15000 mẫu		Kết hợp cả 2 thuật toán để mang lại đô chính xác cao hơn	Bắt buộc phải chuẩn hóa các thuộc tính
A.R.W.M.M.S.C.B. Amarakoon [8]	Dự đoán khí hậu ở Sri Lanka	Mạng nơ-ron nhân tạo	K Nearest Neighbor	Nhiệt độ, độ ẩm, mưa, tốc độ gió	1 năm	365 mẫu		Có lợi cho dữ liệu động	Cần tích hợp các kỹ thuật lựa chọn tính năng

M. A. Kalyankar và S. J. Alaspurkar [9]	Phân tích dữ liệu khí tượng	Phân cụm	Phân cụm K-mean	Nhiệt độ, độ ẩm, mưa, tốc độ gió	4 năm	8660 mẫu		Độ chính xác tốt	Cần có các phương pháp khai phá dữ liệu động
Kavita Pabreja [10]	Dự báo mưa lớn	Phân cụm	Phân cụm K-mean	Nhiệt độ, độ ẩm	2 ngày			Bổ sung cùng với các mô hình dự báo thời tiết bằng số	Không tốt cho các dự báo dài hạn
Pinky Saikia Dutta và Hitesh Tahbilder [11]	Dự báo lượng mưa	Hồi quy	Hồi quy tuyến tính bội	Nhiệt độ cao nhất, thấp nhất, hướng gió, độ ẩm, lượng mưa	6 năm	72 mẫu	63%	Độ chính xác chấp nhận được	Cần loại bỏ một số thông số để có độ chính xác cao hơn
Neha Khandelwal và Ruchi Davey [12]	Dự báo hạn hán	Hồi quy	Hồi quy tuyến tính bội	Lượng mưa, mực nước biển, độ ẩm, nhiệt độ	1 năm	365 mẫu		Phân tích tương quan và thống kê được áp dụng	Việc xác minh chưa được thực hiện xong
S. Kannan và S. Ghosh [13]	Dự báo lượng mưa tại lưu vực sông	Cây quyết định, phân cụm	CART, phân cụm k-mean	Nhiệt độ, ấp suất, gió, lượng mưa	50 năm	432000 mẫu		Phân nhóm dữ liệu lượng mưa nhiều địa điểm trong các cụm	Vẫn chưa khai thác được dữ liệu nhỏ để dự đoán. Việc xác minh chưa được thực hiên xong
F. Oliya và A. B. Adeyemo [14]	Dự báo thời tiết và nghiên cứu biến đổi khí hậu	Cây quyết định, mạng nơ-ron nhân tạo	C4.5, CART, TLFN	Nhiệt độ, lượng mưa, lượng bay hơi, tốc độ gió	10 năm	36000 mẫu	82%	Mạng tốt nhất được chọn để dự đoán	Độ chính xác thay đổi khác nhau tùy theo kích thước của tập dữ liệu đào tạo

2.3. Kết chương

Từ những nội dung so sánh được nêu trong bảng 2.1 ở trên, có thể thấy các kỹ thuật khai phá dữ liệu được ứng dụng phổ biến trong dự báo thời tiết đều có những ưu, nhược điểm nhất định và được sử dụng tùy theo yêu cầu của bài toán. Từ những ưu, nhược điểm này sẽ giúp đưa ra quyết định sử dụng kỹ thuật khai phá dữ liệu nào khi "**Ứng dụng các kỹ thuật khai phá dữ liệu trong dự báo một số thông số khí quyển**". Ở chương tiếp theo, một số kỹ thuật khai phá dữ liệu như Phát hiện ngoại lệ, các mô hình hồi quy sẽ được ứng dụng vào một bài toán cụ thể.

CHƯƠNG 3: ỨNG DỤNG CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU VÀO MỘT BÀI TOÁN DỰ BÁO NHIỆT ĐỘ CẢM NHẬN TỪ ĐỘ ẨM VÀ NHIỆT ĐỘ TRONG NGÀY

3.1. Phân tích bài toán

Như đã trình bày ở trên, dự báo thời tiết là một công việc rất cần thiết trong mọi lĩnh vực của thế giới hiện đại ngày này, tuy nhiên việc dự đoán chính xác là rất khó và đòi hỏi nhiều yếu tố. Bài toán dự báo nhiệt độ cảm nhận từ nhiệt độ và độ ẩm trong ngày được lấy dữ liệu từ trên website Kaggle, một website trực tuyến dành cho mọi người có thể chia sẻ kinh nghiệm, thực hành học máy.

Bộ dữ liệu bao gồm 96453 bản ghi hàng giờ/hàng ngày cho khu vực thành phố Szeged ở Hungary từ năm 2006 đến năm 2016. Các bản ghi bao gồm các thông tin về:

Time : Thời gian ghi

Summary: Tóm tắt về thời tiết lúc đó (Partly Cloudy, Mostly Cloudy, Foggy, Overcast, …)

precipType: Thời tiết chính xác thời điểm đó (rain, snow, …)

temperature: Nhiệt độ (^oC)

apparentTemperature: Nhiệt độ cảm nhận (^oC)

humidity: Độ ẩm

windspeed: Tốc độ gió (km/h)

windBearing: Hướng gió (độ)

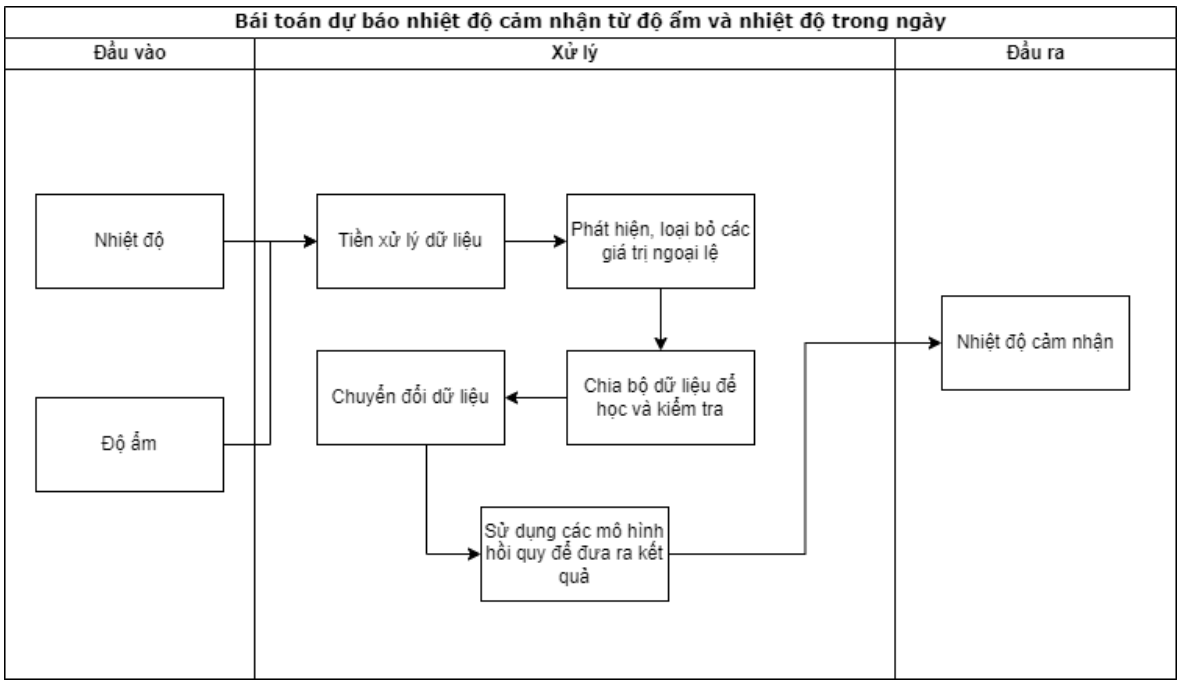
visibility: tầm nhìn

pressure: áp suất (milibars)

Bài toán sẽ sử dụng các kỹ thuật khai phá dữ liệu như Phát hiện ngoại lệ, áp dụng mô hình Hồi quy tuyến tính (Linear Regression), mô hình Hồi quy Cây quyết định (Decision Tree Regression), mô hình Hồi quy Rừng ngẫu nhiên (Random Forest Regression) để dự đoán được nhiệt độ cảm nhận từ giá trị nhiệt độ và độ ẩm đã có rồi đưa ra độ chính xác của dự đoán.

Để đánh giá độ chính xác của dự đoán ta sử dụng hệ số xác định R^2 và sai số toàn phương trung bình (mean squared error). Với R^2 cho ta biết mô hình phù hợp

với dữ liệu bao nhiêu %, sai số toàn phương trung bình là trung bình của bình phương các sai số, cho ta biết sự khác biệt giữa các ước lượng và những gì được đánh giá.



Hình 3.1. Mô hình hóa bài toán dự báo nhiệt độ cảm nhận từ độ ẩm và nhiệt độ trong ngày.

3.2. Thực hiện bài toán trên công cụ jupyter notebook

Jupyter Notebook là một nền tảng tính toán khoa học mã nguồn mở, bạn có thể sử dụng để tạo và chia sẻ các tài liệu có chứa code trực tiếp, phương trình, trực quan hóa dữ liệu và văn bản tường thuật. Jupyter Notebook được viết bằng các ngôn ngữ như Python, R và Julia. Chương trình thực hiện bài toán được thực hiện bằng ngôn ngữ lập trình python trên nền tảng Jupyter Notebook.

3.2.1. Tiền xử lý dữ liệu

Thêm các thư viện cần dùng vào chương trình, hiển thị các file có trong thư mục chạy chương trình:

- Numpy: Thư viện phổ biến và mạnh mẽ cho tính toán khoa học trên Python. Numpy giúp người dùng làm việc hiệu quả với ma trận và mảng.

- Pandas: Thư viện python mã nguồn mở. Thư viện được xây dựng trên Numpy. Pandas cung cấp cấc cấu trúc dữ liệu và phép toán để thao tác với các bảng số và chuỗi thời gian. Thư viện được sử dụng cho các nhiệm vụ khoa học dữ liệu, phân tích dữ liệu và học máy

- Matplotlib: Thư viện vẽ đồ thị giúp trực quan hóa số liệu trên ngôn ngữ lập

trình python.

- Seaborn: Thư viện sử dụng Matplotlib để vẽ đồ thị giúp trực quan hóa các phân phối ngẫu nhiên

- Scipy: Thư viện mã nguồn mở sử dụng cho tính toán khoa học và tính toán kỹ thuật. Scipy cung cấp nhiều chức năng tiện ứng để tối ưu hóa, thống kê và xử lý tín hiệu.

- Scikit-learning: Thư viện máy học miễn phí cho Python. Nó có các thuật toán khác nhau như máy vector hỗ trợ, rừng ngẫu nhiên, k-means, … đồng thời được thiết kế để tương tác với các thư viện như NumPy và SciPy.

```
In [1]: import numpy as np # linear algebra
        import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
        import matplotlib
        import matplotlib.pyplot as plt
        import seaborn as sns
        import scipy
        import scipy.stats as stats

        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import FunctionTransformer
        from sklearn.preprocessing import StandardScaler, MinMaxScaler
        from sklearn.linear_model import LinearRegression
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.preprocessing import PolynomialFeatures
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.metrics import mean_squared_error

        import warnings
        warnings.filterwarnings("ignore")

        import os
        for dirname, _, filenames in os.walk('D:\cao học\Luận văn\chương trình'):
            for filename in filenames:
                print(os.path.join(dirname, filename))

        D:\cao học\Luận văn\chương trình\weather forcecast.ipynb
        D:\cao học\Luận văn\chương trình\weatherHistory.csv
        D:\cao học\Luận văn\chương trình\.ipynb_checkpoints\weather forcecast-checkpoint.ipynb
```

Hình 3.2. Thêm các thư viện cần dùng vào chương trình và hiển thị các file có trong thư mục chạy chương trình

Đọc file csv chứa dự liệu thời tiết, hiển thị kích thước file.

```
In [2]: df = pd.read_csv("D:\cao học\Luận văn\chương trình\weatherHistory.csv")
        print(f"df shape:\t {df.shape}")

        df shape:        (96453, 12)
```

Hình 3.3. Đọc file csv và hiển thị kích thước file

Hiển thị một số mẫu dữ liệu đầu trong file.

```
In [3]: df.head()
```

Out[3]:

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover	Pressure (millibars)	Daily Summary
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.8263	0.0	1015.13	Partly cloudy throughout the day.
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.8263	0.0	1015.63	Partly cloudy throughout the day.
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.9569	0.0	1015.94	Partly cloudy throughout the day.
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.8263	0.0	1016.41	Partly cloudy throughout the day.
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.8263	0.0	1016.51	Partly cloudy throughout the day.

Hình 3.4. Các mẫu dữ liệu đầu trong file

Làm sạch dữ liệu :

- Khử các giá trị trùng lặp.

```
In [4]: df.drop_duplicates(inplace=True)
        df.reset_index(drop=True, inplace=True)
        df.shape
Out[4]: (96429, 12)
```

Hình 3.5. Khử các mẫu dữ liệu trùng lặp

- Tạo một bộ dữ liệu mới chỉ bao gồm các thông số cần sử dụng: “Temperature (C)”, “Apparent Temperature (C)”, “Humidity”.

```
In [5]: df2 = df[['Temperature (C)', 'Apparent Temperature (C)', 'Humidity']]
        df2.head()
```

Out[5]:

	Temperature (C)	Apparent Temperature (C)	Humidity
0	9.472222	7.388889	0.89
1	9.355556	7.227778	0.86
2	9.377778	9.377778	0.89
3	8.288889	5.944444	0.83
4	8.755556	6.977778	0.83

Hình 3.6. Các mẫu dữ liệu đầu trong bộ dữ liệu mới

- Hiển thị thông tin bộ dữ liệu mới.

```
In [6]: print(f"df2 shape:\t {df2.shape}\n")
        df2.info()
```

```
df2 shape:       (96429, 3)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 96429 entries, 0 to 96428
Data columns (total 3 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Temperature (C)             96429 non-null  float64
 1   Apparent Temperature (C)    96429 non-null  float64
 2   Humidity                    96429 non-null  float64
dtypes: float64(3)
memory usage: 2.2 MB
```

Hình 3.7. Hiển thị thông tin bộ dữ liệu mới

- Hiển thị thông tin dữ liệu trong bộ dữ liệu mới.

```
In [7]: df2.describe(include='all')
```

Out[7]:

	Temperature (C)	Apparent Temperature (C)	Humidity
count	96429.000000	96429.000000	96429.000000
mean	11.929692	10.851707	0.734902
std	9.550492	10.695743	0.195466
min	-21.822222	-27.716667	0.000000
25%	4.683333	2.311111	0.600000
50%	12.000000	12.000000	0.780000
75%	18.838889	18.838889	0.890000
max	39.905556	39.344444	1.000000

Hình 3.8. Thông tin dữ liệu trong bộ dữ liệu mới

- Tìm kiếm các giá trị còn thiếu.

```
In [8]: features_na = [features for features in df2.columns if df2[features].isnull().sum() > 0]
if(len(features_na)>0):
    for feature in features_na:
        print("{}: {} %".format(feature, np.round(df2[feature].isnull().mean()*100, 4)))
else:
    print("No any missing value found")

No any missing value found
```

Hình 3.9. Hiển thị các giá trị còn thiếu trong bộ dữ liệu

Sau khi tiền xử lý dữ liệu ta sẽ có một bộ dữ liệu mới đã được "làm sạch" thích hợp với yêu cầu đã đề ra của bài toán.

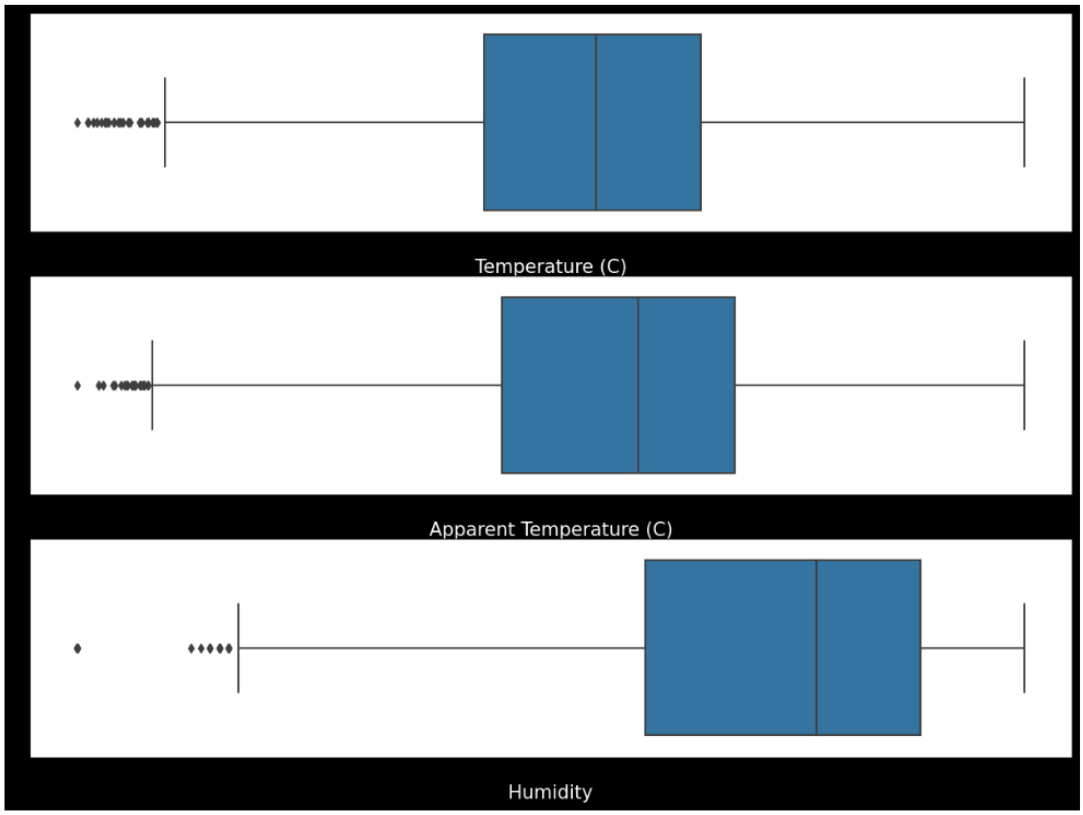
3.2.2. Phát hiện, loại bỏ các dữ liệu ngoại lệ

Vẽ biểu đồ phân bố dữ liệu cho 3 thông số để loại bỏ các dữ liệu ngoại lệ.

```
In [9]: numerical_features = [feature for feature in df2.columns if (df2[feature].dtypes != 'O')]

plt.figure(figsize=(15,45), facecolor='black')
plotnumber =1
for numerical_feature in numerical_features:
    ax = plt.subplot(12,1,plotnumber)
    sns.boxplot(df2[numerical_feature],orient='h')
    plt.xlabel(numerical_feature, color="white", size=15)
    plotnumber += 1
plt.show()
```

Hình 3.10. Hiển thị phân bố dữ liệu cho 3 thông số



Hình 3.11. Biểu đồ phân bố dữ liệu

Tạo bộ dữ liệu mới khử các giá trị ngoại lệ:

- Hiển thị kích cỡ bộ dữ liệu mới, vẽ biểu đồ so sánh dữ liệu Độ ẩm trước và sau khi khử.

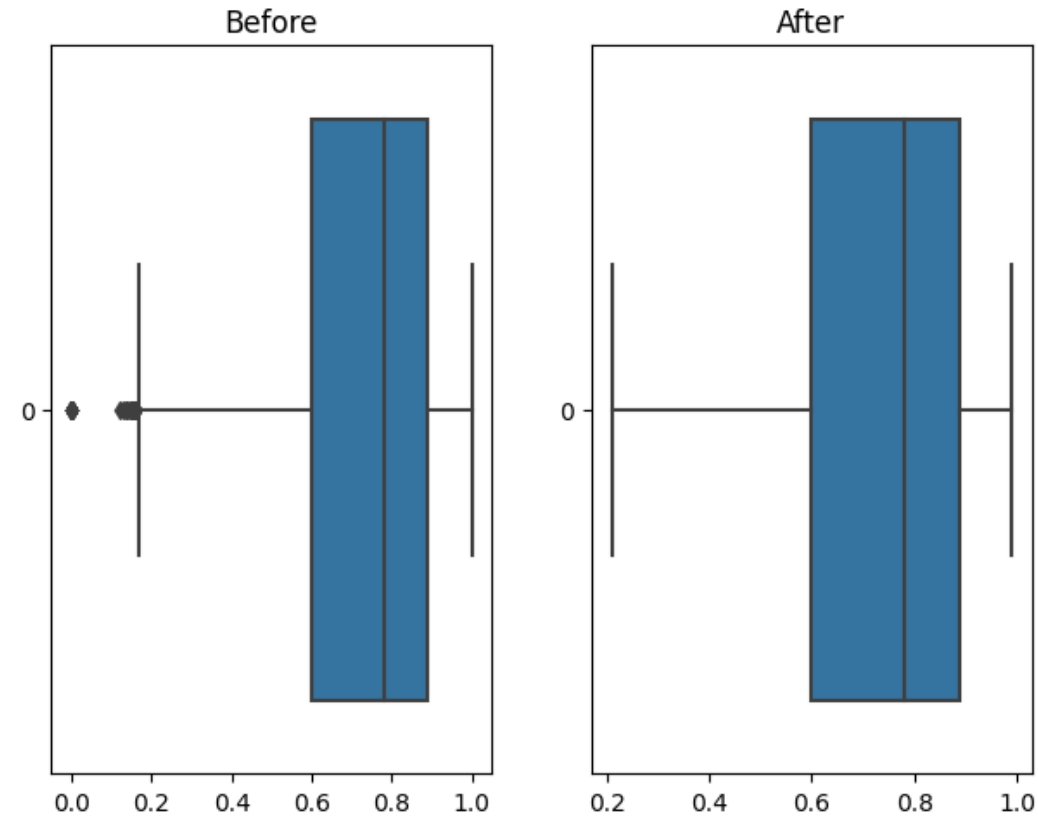
```
In [10]: df3 = df2.copy()

In [11]: def draw_boxplots(feature,min_val,max_val):
             fig, axes = plt.subplots(1,2)
             plt.tight_layout()
             plt.figure(figsize=(15,10), facecolor='black')
             sns.boxplot(df2[feature],orient='h',ax=axes[0])
             axes[0].title.set_text("Before")
             sns.boxplot(df3[feature],orient='h',ax=axes[1])
             axes[1].title.set_text("After")
             plt.show()

In [12]: feature = 'Humidity'
         min_val = 0.2
         max_val = 1

         df3 = df3[(df3[feature]>min_val) & (df3[feature]<max_val)]
         print('Shape: ',df3.shape)
         draw_boxplots(feature, min_val, max_val)
```

Hình 3.12. Hiển thị kích cỡ bộ dữ liệu độ ẩm mới




```
<Figure size 1500x1000 with 0 Axes>
```

Hình 3.13. Biểu đồ so sánh dữ liệu Độ ẩm trước và sau khi khử

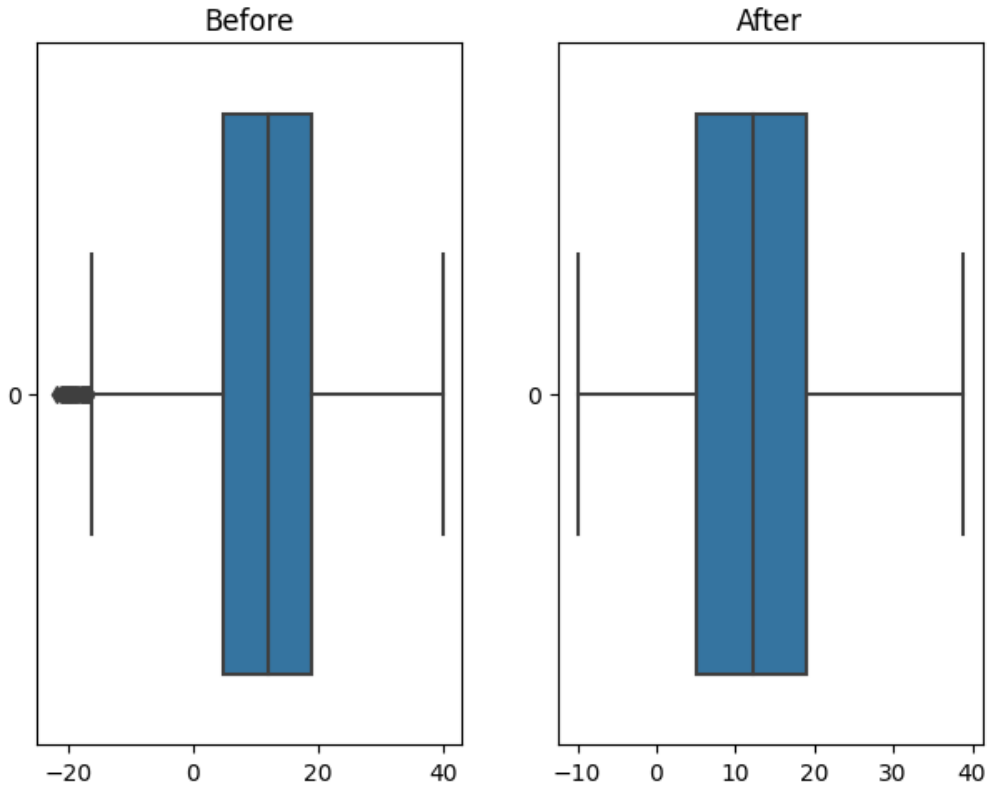
- Hiển thị kích cỡ bộ dữ liệu mới, vẽ biểu đồ so sánh dữ liệu Nhiệt độ trước

và sau khi khử.

```
In [13]: feature = 'Temperature (C)'
         min_val = -10
         max_val = 40
         df3 = df3[(df3[feature]>min_val) & (df3[feature]<max_val)]
         print('Shape: ',df3.shape)
         draw_boxplots(feature, min_val, max_val)

         Shape:  (92899, 3)
```

Hình 3.14. Hiển thị kích cỡ bộ dữ liệu Nhiệt độ mới




```
<Figure size 1500x1000 with 0 Axes>
```

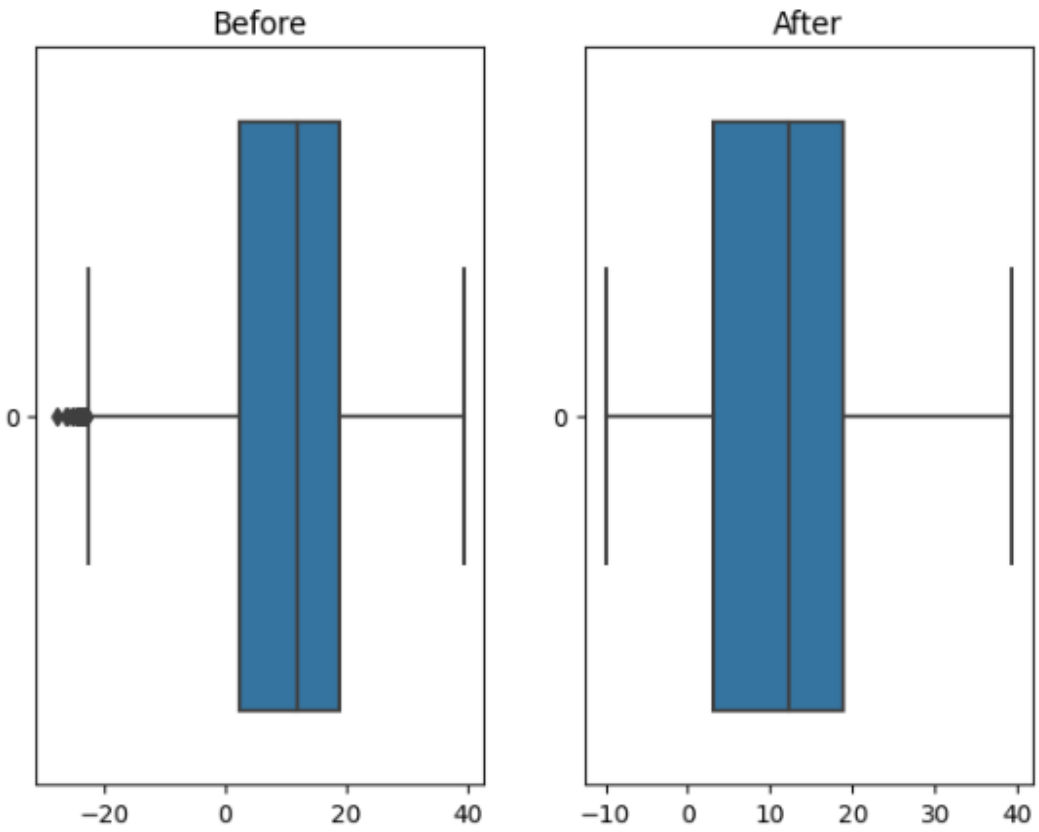
Hình 3.15. Biểu đồ so sánh dữ liệu Nhiệt độ trước và sau khi khử

- Hiển thị kích cỡ bộ dữ liệu mới, vẽ biểu đồ so sánh dữ liệu Nhiệt độ cảm nhận trước và sau khi khử.

```
In [14]: feature = 'Apparent Temperature (C)'
         min_val = -10
         max_val = 40
         df3 = df3[(df3[feature]>min_val) & (df3[feature]<max_val)]
         print('Shape: ',df3.shape)
         draw_boxplots(feature, min_val, max_val)

         Shape:  (91882, 3)
```

Hình 3.16. Hiển thị kích cỡ bộ dữ liệu Nhiệt độ cảm nhận mới




```
<Figure size 1500x1000 with 0 Axes>
```

Hình 3.17. Biểu đồ so sánh dữ liệu Nhiệt độ cảm nhận trước và sau khi khử

Các dữ liệu ngoại lệ có thể làm sai lệch các dự báo trong mô hình, tác động tiêu cực đến chất lượng và độ chính xác của kết quả bài toán. Vì vậy, ở bước này ta cần phải phát hiện và loại bỏ chúng.

3.2.3. Chia bộ dữ liệu để học và kiểm tra

Hiển thị kích cỡ bộ dữ liệu.

```
In [15]: X,y = df3.iloc[:,[0,2]], df3.iloc[:, [1]]
         print(f"X shape: {X.shape}\ny shape: {y.shape}")

         X shape: (91882, 2)
         y shape: (91882, 1)
```

Hình 3.18. Hiển thị kích cỡ bộ dữ liệu

Chia bộ dữ liệu ra 70% để học, 30% để kiểm tra:

```
In [16]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

         X_train.reset_index(inplace=True)
         X_train.drop(['index'], axis=1, inplace=True)
         X_test.reset_index(inplace=True)
         X_test.drop(['index'], axis=1, inplace=True)
         y_train.reset_index(inplace=True)
         y_train.drop(['index'], axis=1, inplace=True)
         y_test.reset_index(inplace=True)
         y_test.drop(['index'], axis=1, inplace=True)

         print(f"X_train shape: {X_train.shape}\tX_test shape: {X_test.shape}")
         print(f"y_train shape: {y_train.shape}\ty_test shape: {y_test.shape}")

         X_train shape: (64317, 2)       X_test shape: (27565, 2)
         y_train shape: (64317, 1)       y_test shape: (27565, 1)
```

Hình 3.19. Kết quả sau khi chia bộ dữ liệu

3.2.4. Chuyển đổi dữ liệu

- Hiển thị độ lệch không thiên vị trên trục của dữ liệu Nhiệt độ và dữ liệu độ ẩm.

```
In [17]: X_train.skew()

Out[17]: Temperature (C)    0.147866
         Humidity          -0.698541
         dtype: float64
```

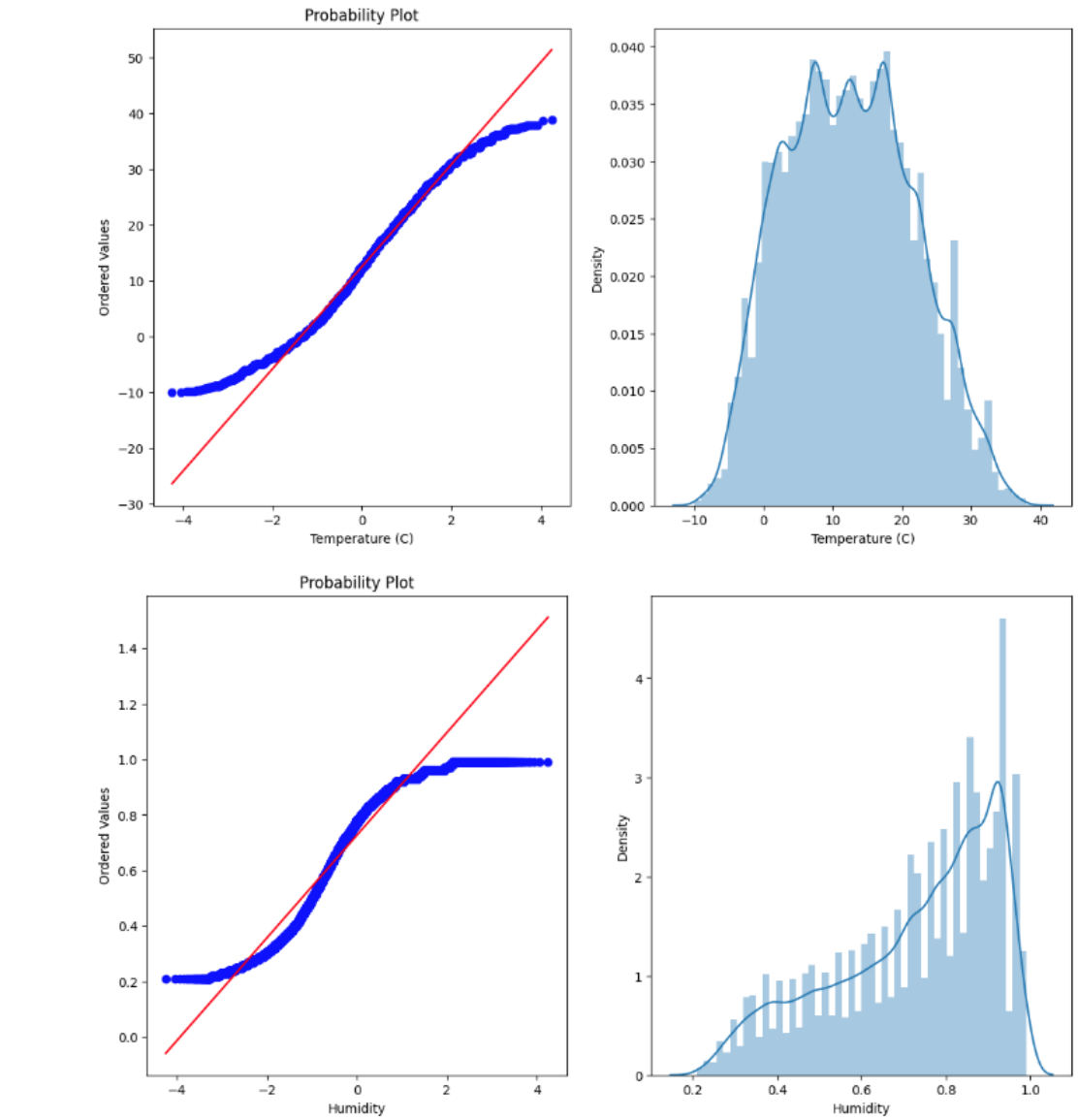
Hình 3.20. Độ lệch không thiên vị trên trục của dữ liệu nhiệt độ và dữ liệu độ ẩm

- Tạo hàm, vẽ đồ thị lượng tử cho dữ liệu nhiệt độ và dữ liệu độ ẩm.

```
In [18]: def draw_qq_hist(feature):
             plt.figure(figsize=(20,80), facecolor='white')
             ax = plt.subplot(10,3,1)
             stats.probplot(X_train[feature], dist="norm", plot=plt)
             plt.xlabel(feature)
             ax = plt.subplot(10,3,2)
         #     ax.set_title("Hist")
             sns.distplot(X_train[feature])
             plt.xlabel(feature)
             plt.show()

In [19]: for feature in X_train.columns:
             draw_qq_hist(feature)
```

Hình 3.21. Tạo hàm, vẽ đồ thị lượng tử cho dữ liệu nhiệt độ và dữ liệu độ ẩm



Hình 3.22. Đồ thị lượng tử cho dữ liệu nhiệt độ, độ ẩm

- Chuyển đổi dữ liệu Độ ẩm, hiển thị lại độ lệch không thiên vị trên trục của dữ liệu nhiệt độ và dữ liệu độ ẩm và vẽ đồ thị lượng tử của chúng.

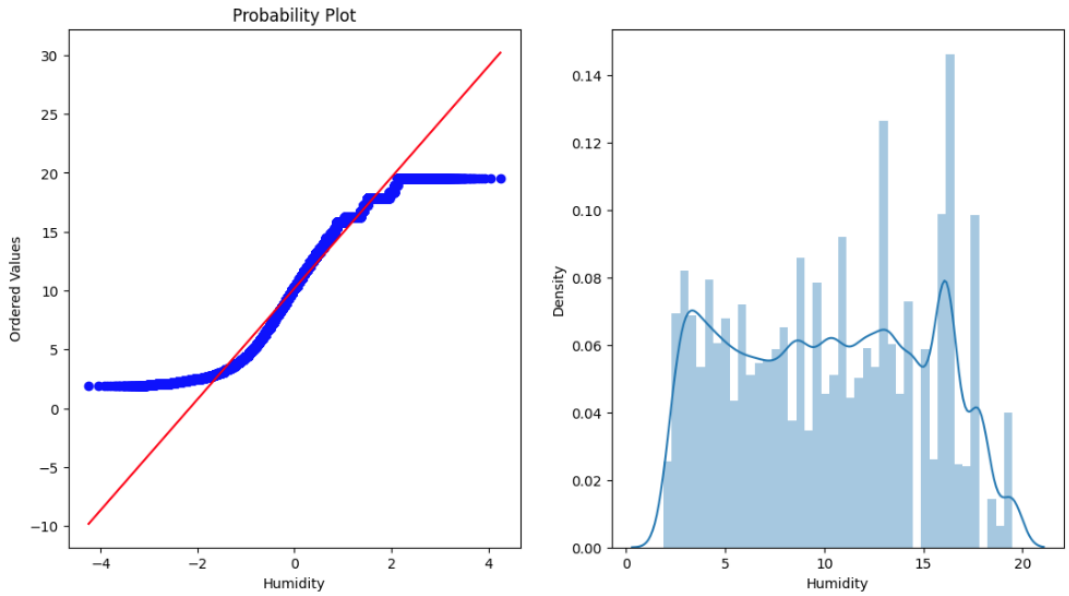
```
In [20]: columns = ['Humidity']
         exponential_transformer = FunctionTransformer(lambda x: np.exp(x*3), validate=True)
         data_new = exponential_transformer.transform(X_train[columns])
         df_new = pd.DataFrame(data_new, columns=columns)
         X_train['Humidity']=df_new['Humidity']

         X_train.skew()

Out[20]: Temperature (C)    0.147866
         Humidity           0.033180
         dtype: float64
```

Hình 3.23. Hiển thị lại độ lệch không thiên vị trên trục của dữ liệu Nhiệt độ, độ ẩm draw_qq_hist('Humidity')

```
In [21]: draw_qq_hist('Humidity')
```



Hình 3.24. Đồ thị lượng tử cho dữ liệu độ ẩm

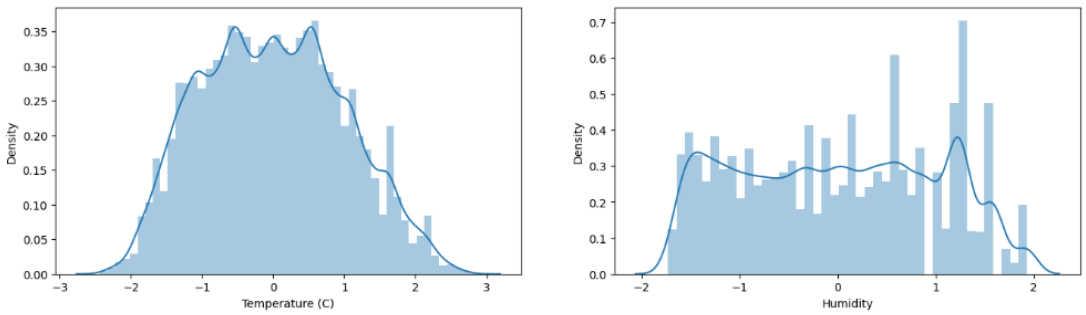
- Co dữ liệu.

```
In [22]: scaler_x = StandardScaler().fit(X_train)
         scaler_y = StandardScaler().fit(y_train)
```

```
In [23]: X_train_scale = scaler_x.transform(X_train)
         X_test_scale = scaler_x.transform(X_test)
         y_train_scale = scaler_y.transform(y_train)
         y_test_scale = scaler_y.transform(y_test)

         X_train = pd.DataFrame(X_train_scale, columns=[X_train.columns])
         X_test = pd.DataFrame(X_test_scale, columns=[X_test.columns])
         y_train = pd.DataFrame(y_train_scale, columns=['Apparent Temperature (C)'])
         y_test = pd.DataFrame(y_test_scale, columns=['Apparent Temperature (C)'])
```

```
In [24]: plt.figure(figsize=(25,50), facecolor='white')
         plotnumber =1
         for feature in ['Temperature (C)', 'Humidity']:
             ax = plt.subplot(10,3,plotnumber)
             sns.distplot(X_train[feature])
             plt.xlabel(feature)
             plotnumber+=1
         plt.show()
```



Hình 3.25. Co dữ liệu

Chuyển đổi dữ liệu giúp dữ liệu dễ tương thích với các mô hình thuật toán để có thể đưa ra kết quả bài toán một cách tốt nhất.

3.2.5. Sử dụng mô hình Hồi quy tuyến tính (Linear Regression) giải quyết bài toán

Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Nó mô hình toán học biến không xác định hoặc phụ thuộc và biến đã biết hoặc độc lập như một phương trình tuyến tính. Ví dụ, giả sử rằng bạn có dữ liệu về chi phí và thu nhập của bạn trong năm ngoái. Kỹ thuật hồi quy tuyến tính phân tích dữ liệu này và xác định rằng chi phí của bạn là một nửa thu nhập của bạn. Sau đó, họ tính toán một chi phí trong tương lai không rõ bằng cách giảm một nửa thu nhập được biết đến trong tương lai.

Chạy mô hình Hồi quy tuyến tính và hiển thị một số giá trị nhiệt độ cảm nhận dự đoán.

```
In [25]: model = LinearRegression()
         model.fit(X_train, y_train)
```

```
Out[25]: LinearRegression
         LinearRegression()
```

```
In [26]: predictions = model.predict(X_test)
```

```
In [27]: y_hat = pd.DataFrame(predictions, columns=["predicted"])
         y_hat.head()
```

Out[27]:

	predicted
0	0.202825
1	-1.661548
2	-0.558778
3	0.260922
4	-0.608209

Hình 3.26. Hiển thị một số giá trị dữ đoán dựa trên mô hình hồi quy tuyến tính

Sử dụng hệ số xác định R^2 và sai số toàn phương trung bình (mean squared error) để đánh giá độ chính xác của việc sử dụng mô hình Hồi quy tuyến tính vào bài toán.

```
In [28]: R_Square_Score = model.score(X_test, y_test)
         R_Square_Score

Out[28]: 0.9806145061145068

In [29]: mse_error = mean_squared_error(y_test, y_hat)
         mse_error

Out[29]: 0.019372854190451972
```

Hình 3.27. Hệ số xác định R2 và sai số toàn phương trung bình khi áp dụng mô hình hồi quy tuyến tính vào bài toán

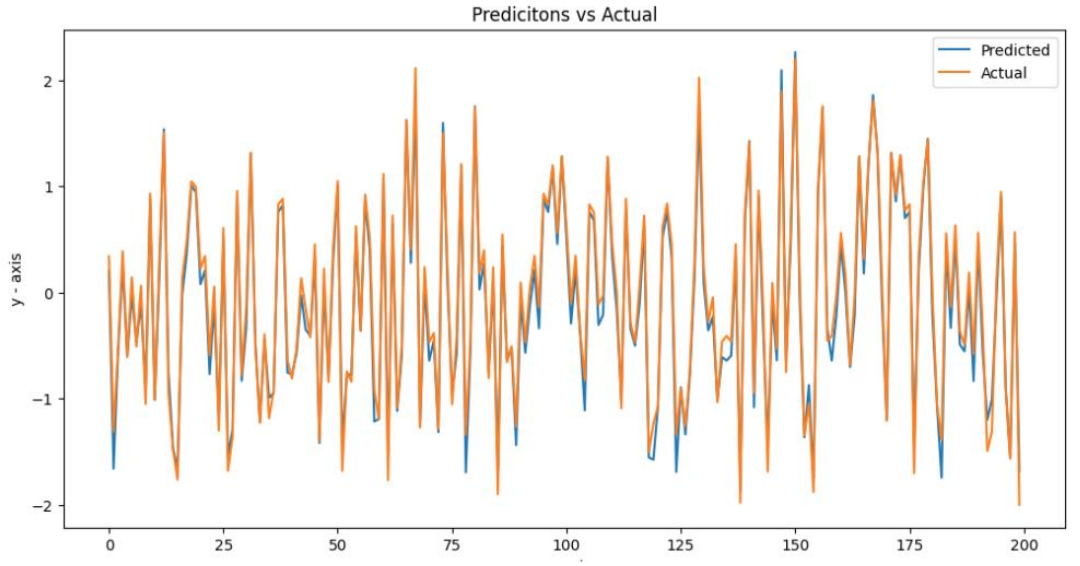
Vẽ biểu đồ so sánh giá trị dự đoán và giá trị thực tế (sử dụng 200 giá trị đầu) với đường cam là đường biểu diễn giá trị thực tế và đường xanh là đường biểu diễn giá trị dự đoán được.

```
In [30]: plt.figure(figsize=(12, 6))

         plt.plot(y_hat[:200], label = "Predicted")
         plt.plot(y_test[:200], label = "Actual")

         plt.xlabel('x - axis')
         plt.ylabel('y - axis')
         plt.title('Predicitons vs Actual')
         plt.legend()

         plt.show()
```



Hình 3.28. Biểu đồ so sánh giá trị dự đoán và thực tế khi áp dụng mô hình hồi quy tuyến tính

3.2.6. Sử dụng mô hình Hồi quy Cây quyết định (Decision Tree Regression) giải quyết bài toán

Cây quyết định là một công cụ ra quyết định sử dụng cấu trúc cây giống như lưu đồ hoặc là một mô hình về các quyết định và tất cả các kết quả có thể có của chúng, bao gồm cả kết quả, chi phí đầu vào và tiện ích.

Hồi quy cây quyết định quan sát các đặc điểm của một đối tượng và đào tạo

một mô hình trong cấu trúc của cây để dự đoán dữ liệu trong tương lai nhằm tạo ra đầu ra liên tục có ý nghĩa. Đầu ra liên tục có nghĩa là đầu ra/kết quả không rời rạc, tức là nó không được biểu diễn chỉ bằng một tập hợp số hoặc giá trị rời rạc, đã biết

Chạy mô hình Hồi quy Cây quyết định và hiển thị một số giá trị nhiệt độ cảm nhận dự đoán.

```
In [31]: model = DecisionTreeRegressor()
         model.fit(X_train, y_train)

Out[31]: DecisionTreeRegressor
         DecisionTreeRegressor()

In [32]: predictions = model.predict(X_test)

In [33]: y_hat = pd.DataFrame(predictions, columns=["predicted"])
         y_hat.head()

Out[33]:
```

	predicted
0	0.337451
1	-1.220690
2	-0.669298
3	0.387087
4	-0.631532

Hình 3.29. Hiển thị một số giá trị dự đoán dựa trên mô hình hồi quy cây quyết định

Sử dụng hệ số xác định R^2 và sai số toàn phương trung bình (mean squared error) để đánh giá độ chính xác của việc sử dụng mô hình Hồi quy cây quyết định vào bài toán.

```
In [34]: R_Square_Score = model.score(X_test, y_test)
         R_Square_Score

Out[34]: 0.9765345222128011

In [35]: mse_error = mean_squared_error(y_test, y_hat)
         mse_error

Out[35]: 0.02345017786835338
```

Hình 3.30. Hệ số xác định R2 và sai số toàn phương trung bình khi áp dụng mô hình hồi quy cây quyết định vào bài toán

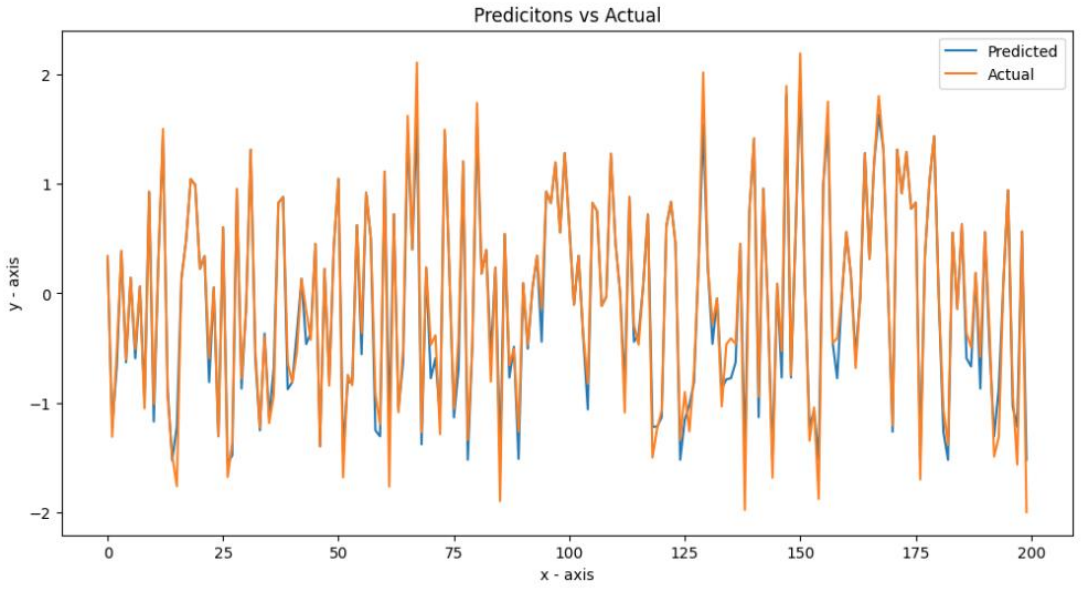
Vẽ biểu đồ so sánh giá trị dự đoán và giá trị thực tế (sử dụng 200 giá trị đầu) với đường cam là đường biểu diễn giá trị thực tế và đường xanh là đường biểu diễn giá trị dự đoán được.

```
In [36]: plt.figure(figsize=(12, 6))

         plt.plot(y_hat[:200], label = "Predicted")
         plt.plot(y_test[:200], label = "Actual")

         plt.xlabel('x - axis')
         plt.ylabel('y - axis')
         plt.title('Predicitons vs Actual')
         plt.legend()

         plt.show()
```



Hình 3.31. Biểu đồ so sánh giá trị dự đoán và thực tế khi áp dụng mô hình hồi quy cây quyết định

3.2.7. Sử dụng mô hình Hồi quy Rừng ngẫu nhiên (Random Forest Regression) giải quyết bài toán

Rừng ngẫu nhiên là một kỹ thuật tổng hợp có khả năng thực hiện cả nhiệm vụ hồi quy và phân loại với việc sử dụng nhiều cây quyết định. Ý tưởng cơ bản đằng sau điều này là kết hợp nhiều cây quyết định để xác định đầu ra cuối cùng thay vì dựa vào các cây quyết định riêng lẻ.

Hồi quy rừng ngẫu nhiên là một thuật toán học có giám sát sử dụng phương pháp học đồng bộ để hồi quy. Phương pháp học tập đồng bộ là một kỹ thuật kết hợp các dự đoán từ nhiều thuật toán học máy để đưa ra dự đoán chính xác hơn so với một mô hình duy nhất.

Chạy mô hình Hồi quy Rừng ngẫu nhiên và hiển thị một số giá trị nhiệt độ cảm nhận dự đoán.

```
In [37]: model = RandomForestRegressor()
         model.fit(X_train, y_train)

Out[37]: RandomForestRegressor
         RandomForestRegressor()

In [38]: predictions = model.predict(X_test)

In [39]: y_hat = pd.DataFrame(predictions, columns=["predicted"])
         y_hat.head()

Out[39]:
```

	predicted
0	0.337451
1	-1.543681
2	-0.584426
3	0.387087
4	-0.681038

Hình 3.32. Hiển thị một số giá trị dữ đoán dựa trên mô hình hồi quy rừng ngẫu nhiên

Sử dụng hệ số xác định R^2 và sai số toàn phương trung bình (mean squared error) để đánh giá độ chính xác của việc sử dụng mô hình Hồi quy cây quyết định vào bài toán.

```
In [40]: R_Square_Score = model.score(X_test, y_test)
         R_Square_Score

Out[40]: 0.984122695215193

In [41]: mse_error = mean_squared_error(y_test, y_hat)
         mse_error

Out[41]: 0.015866952492946763
```

Hình 3.33. Hệ số xác định R2 và sai số toàn phương trung bình khi áp dụng mô hình hồi quy rừng ngẫu nhiên vào bài toán

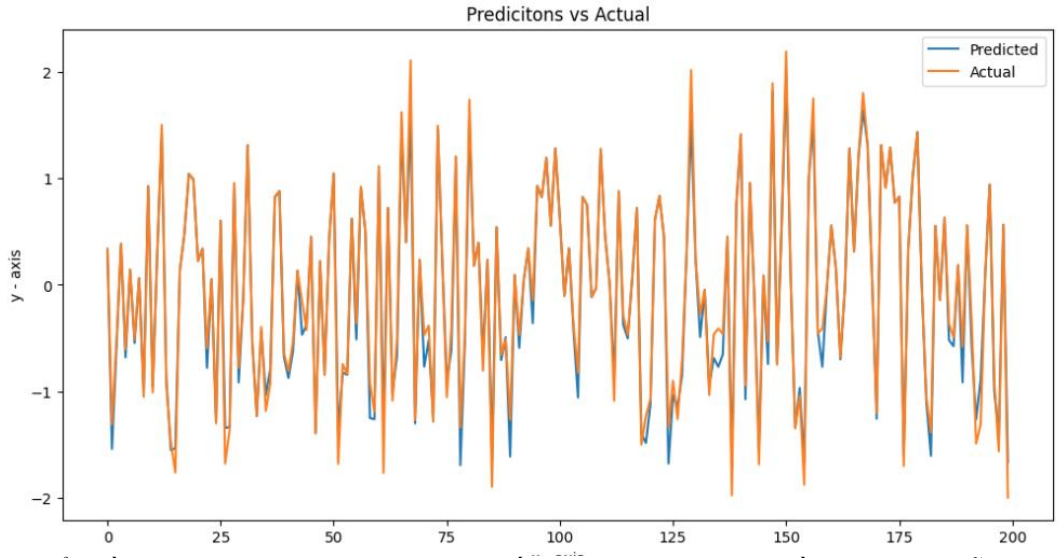
Vẽ biểu đồ so sánh giá trị dự đoán và giá trị thực tế (sử dụng 200 giá trị đầu) với đường cam là đường biểu diễn giá trị thực tế và đường xanh là đường biểu diễn giá trị dự đoán được.

```
In [42]: plt.figure(figsize=(12, 6))

         plt.plot(y_hat[:200], label = "Predicted")
         plt.plot(y_test[:200], label = "Actual")

         plt.xlabel('x - axis')
         plt.ylabel('y - axis')
         plt.title('Predicitons vs Actual')
         plt.legend()

         plt.show()
```



Hình 3.34. Biểu đồ so sánh giá trị dự đoán và thực tế khi áp dụng mô hình hồi quy rừng ngẫu nhiên

3.2.8. Đánh giá kết quả

Từ kết quả đạt được ở trên ta có bảng sau:

Bảng 3.1. Bảng so sánh các mô hình hồi quy áp dụng vào bài toán

Tên mô hình	Hệ số xác định R^2 (R Square)	Sai số toàn phương trung bình (Mean squared error)
Mô hình hồi quy tuyến tính	0.9806145061145068	0.019372854190451972
Mô hình hồi quy cây quyết định	0.9765345222128011	0.02345017786835338
Mô hình hồi quy rừng ngẫu nhiên	0.984122695215193	0.015866952492946763

Từ bảng trên, ta có thể thấy mô hình hồi quy tuyến tính, hồi quy cây quyết định và hồi quy rừng ngẫu nhiên áp dụng vào bài toán dự báo nhiệt độ cảm nhận từ nhiệt độ và độ ẩm đều có hệ số xác định R^2 cao và sai số toàn phương trung bình rất nhỏ. Điều này chứng tỏ các mô hình có độ phù hợp cao với dữ liệu và sự khác biệt giữa các giá trị ước lượng và giá trị thực tế là rất nhỏ.

Từ các lý thuyết về các mô hình hồi quy cùng với các giá trị của hệ số xác định R^2 và Sai số toàn phương trung bình của các mô hình hồi quy, có thể thấy Mô

hình hồi quy rừng ngẫu nhiên là mô hình cho các kết quả tốt nhất. Điều này có thể được giải thích bởi đường hồi quy của các mô hình. Từ đó, có thể thấy điểm mạnh, điểm yếu của các mô hình hồi quy như sau :

- Mô hình hồi quy tuyến tính: Do đường hồi quy của mô hình này là đường thẳng nên các giá trị biểu thị mối quan hệ giữa nhiệt độ, độ ẩm và nhiệt độ cảm nhận luôn ở xung quanh đường thẳng tuyến tính đó không có quá nhiều sự biến đổi, giúp cho các giá trị để đánh giá mô hình trên khá tốt.

- Mô hình hồi quy cây quyết định: Do đường hồi quy của mô hình này là một đường cong giống với hình sin nên các giá trị biểu thị mối quan hệ giữa nhiệt độ, độ ẩm và nhiệt độ cảm nhận sẽ ở xung quanh đường đó làm gây ra sự biến thiên lớn ở giá trị, từ đó các giá trị để đánh giá mô hình bị kém hơn so với mô hình hồi quy tuyến tính.

- Mô hình hồi quy rừng ngẫu nhiên: Do mô hình thực hiện sử dụng nhiều cây quyết định để bổ trợ cho nhau, cùng với đó không bị giới hạn bởi một đường hồi quy tuyến tính nên các giá trị dự đoán được sẽ sát và thực tế và các giá trị đánh giá mô hình sẽ là tốt nhất trong các mô hình hồi quy.

Việc sử dụng các mô hình hồi quy vào bài toán giúp tăng độ chính xác của việc dự đoán nhiệt độ cảm nhận.

KẾT LUẬN VÀ KIẾN NGHỊ

KẾT LUẬN

Luận văn đã hoàn thành được các công việc như sau :

1. Đã đưa ra được nhưng lý thuyết về khai phá dữ liệu, các kỹ thuật khai phá dữ liệu cùng với dự báo thời tiết

2. Đã tìm hiểu, nghiên cứu được các kỹ thuật khai phá dữ liệu ứng dụng trong lĩnh vực dự báo thời tiết.

3. Từ các tài liệu tham khảo, đã đưa ra được một số so sánh khi ứng dụng các kỹ thuật khai phá dữ liệu vào bài toán cụ thể.

4. Ứng dụng được một số kỹ thuật khai phá dữ liệu như phát hiện, loại bỏ các dữ liệu ngoại lệ, các mô hình hồi quy vào một bài toán dự báo thông số khí quyển cụ thể.

5. Kỹ thuật phát hiện, loại bỏ các dữ liệu ngoại lệ giúp bộ dữ liệu được "sạch" hơn giúp tăng độ chính xác của dự đoán.

6. Từ các giá trị hệ số xác định R^2 và sai số toàn phương trung bình khi ứng dụng các mô hình hồi quy vào giải quyết bài toán, đánh giá được độ phù hợp của mô hình với dữ liệu và sự khác biệt giữa các giá trị ước lượng và giá trị thực tế.

7. Với ba mô hình hồi quy được sử dụng để đưa ra dự đoán, mô hình hồi quy rừng ngẫu nhiên là cho ra kết quả phù hợp nhất vì nó đã kết hợp các dự đoán từ nhiều thuật toán học máy

KIẾN NGHỊ

Từ những nội dung luận văn đã trình bày và những kết quả đã đặt được, học viên mong muốn:

1. Tiếp tục tìm hiểu thêm và sâu hơn việc ứng dụng các kỹ thuật khai phá dữ liệu trong dự báo thời tiết.

2. Có thể thu thập, xây dựng một bộ số liệu thời tiết cho một khu vực cụ thể tại Hà Nội.

3. Từ bộ số liệu thu thập được, ứng dung các kỹ thuật khai phá dữ liệu để đưa ra các dự báo về một số thông số khí quyển, cũng như về thời tiết nhưng ngày tới tại khụ vực đó.

TÀI LIỆU THAM KHẢO

[1]. Nguyễn Đức Nam và nnk. Ứng dụng đồng hóa dữ liệu dự báo các trường khí tượng độ phân giải cao cho khu vực Than Uyên (Lai Châu). TẠP CHÍ KHÍ TƯỢNG THỦY VĂN THÁNG 4 NĂM 2021, 59-71

[2] Han J., Kamber M., Jian Pei, 2012, *Data Mining: Concepts & Techniques Third Edition*, Morgan Kaufmann Publishers is an imprint of Elsevier, USA.

[3] Harsha Dessai, Siddhi Naik, 2021, Weather Forecasting Using Data Mining.

[4] P.Kalaiyarasi, Mrs.A.Kalaiselvi, 2018, Data Mining Techniques Using To Weather Prediction.

[5] Elia Georgiana Petre, 2009, A Decision Tree for Weather Prediction, *Buletinul UniversităŃii Petrol. – Gaze din Ploieşti*, vol.LXI, No.1, 77-82

[6] P.Hemalatha, March 2013, Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System, *International Journal Of Computational Engineering Research (ijceronline.com)*, Vol. 3, Issue. 3.

[7] Soo-Yeon Ji, S.Sharma, B.Yu, Dong Hyun Jeong, 2012, Designing a Rule-Based Hourly Rainfall Prediction Model, *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*.

[8] Gaurav J.Sawale, Dr. Sunil R.Gupta, 2013, Use of Artificial Neural Network in Data Mining For Weather Forecasting, *International Journal Of Computer Science And Applications*, Vol. 6, No.2,

[9] A.R.W.M.M.S.C.B. Amarakoon, 2010, Effectiveness of Using Data Mining for Predicting Climate Change in Sri Lanka.

[10] M.Kalyankar, S.Alaspurkar, 2013, Data Mining Technique to Analyse the Metrological Data, *International Journal of Advanced Research in Computer Science and Software Engineering 3(2)*, 114-118.

[11] K.Pabreja, 2012, Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst, *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 3 (1) , 2996 - 2999.

[12] P.Dutta, H.Tahbilder, 2014, Prediction Of Rainfall Using Data mining Technique Over Assam, *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol. 5, No.2.

[13] N.Khandelwal, R.Davey, 2012, Climatic Assessment Of Rajasthans Region For Drought With Concern Of Data Mining Techniques, *International Journal Of Engineering Research and Applications (IJERA)*, Vol. 2, Issue 5, 1695-1697.

[14] S.Kannan , S.Ghosh, 2010, Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output, *Springer-Verlag*.

[15] Olaiya Folorunsho and A.B.Adeyemo, 2012, Application of Data Mining Techniques in Weather Prediction and Climate Change Studies, *International Journal of Information Engineering and Electronic Business.*, vol. 1, no. 1, 51–59.

[16] Divya Chauhan, Jawahar Thakur, 2014, Data Mining Techniques for Weather Prediction: A Review, *Int. J. Eng. Res. Gen. Sci*., Vol. 2 Issue. 8, 2184–2189.