TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ
HỌC MÁY

Người thực hiện: **HOÀNG NGHĨA ÁI – 21093411**
NGUYỄN TRƯỜNG AN – 21010151
Lớp: **DHKHDL17A**
Khoá**: 17**
Người hướng dẫn: **TS BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ
HỌC MÁY

Người thực hiện: **HOÀNG NGHĨA ÁI – 21093411**
NGUYỄN TRƯỜNG AN – 21026821
Lớp: **DHKHDL17A**
Khoá**: 17**
Người hướng dẫn: **TS BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Chúng em muốn gửi lời cảm ơn chân thành đến giảng viên bộ môn "Máy Học" trong khoa Công nghệ Thông tin - Thầy Bùi Thanh Hùng, người đã cung cấp cho chúng tôi những kiến thức và kỹ năng cơ bản cần thiết để hoàn thành đề tài nghiên cứu này. Tuy nhiên, do kiến thức chuyên ngành của chúng tôi còn hạn chế, nên trong quá trình nghiên cứu đề tài, chúng tôi vẫn còn mắc nhiều thiếu sót trong việc tìm hiểu, đánh giá và trình bày về đề tài. Chúng tôi mong nhận được sự quan tâm và đóng góp ý kiến từ các thầy/cô giảng viên bộ môn để đề tài của chúng tôi được hoàn chỉnh và đầy đủ hơn. Chân thành cảm ơn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

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

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

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm 2024

Tác giả

(ký tên và ghi rõ họ tên)

Hoàng Nghĩa Ái *Nguyễn Trường An*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

TP. Hồ Chí Minh, ngày tháng năm 2024

Giảng viên

(ký tên và ghi rõ họ tên)

TÓM TẮT

Mục tiêu của nghiên cứu này là dự đoán thời tiết dựa trên các phương pháp học máy. Dự đoán thời tiết chính xác có ý nghĩa quan trọng trong nhiều lĩnh vực như nông nghiệp, giao thông và phòng chống thiên tai. Nghiên cứu này sử dụng dữ liệu thời tiết từ các tỉnh khác nhau và áp dụng các kỹ thuật học máy để xây dựng các mô hình dự đoán.

Các hướng tiếp cận chính trong nghiên cứu bao gồm: thu thập và tiền xử lý dữ liệu, phân tích và khám phá dữ liệu, xây dựng mô hình học máy và đánh giá hiệu quả mô hình.

Thu thập và tiền xử lý dữ liệu: Dữ liệu thời tiết từ nhiều tỉnh được thu thập từ nhiều nguồn khác nhau, làm sạch và xử lý để loại bỏ giá trị thiếu và ngoại lệ. Quá trình này bao gồm đọc dữ liệu từ các tệp CSV, loại bỏ các cột không cần thiết và xử lý các giá trị thiếu bằng phương pháp Interquartile Range (IQR) để loại bỏ các giá trị ngoại lệ.

Phân tích và khám phá dữ liệu: Sử dụng các phương pháp thống kê và trực quan hóa dữ liệu như ma trận tương quan và biểu đồ nhiệt để khám phá mối quan hệ giữa các thuộc tính thời tiết. Ma trận tương quan giúp xác định các biến quan trọng, trong khi các biểu đồ giúp trực quan hóa mối quan hệ giữa các biến này.

Xây dựng mô hình học máy: Các mô hình Decision Tree Regression và Random Forest Regression được sử dụng để dự đoán nhiệt độ, độ ẩm và lượng mưa. Dữ liệu được chia thành tập huấn luyện và tập kiểm tra. Các mô hình được huấn luyện trên tập dữ liệu huấn luyện và kiểm tra hiệu quả trên tập kiểm tra.

Đánh giá hiệu quả mô hình: Hiệu quả của các mô hình được đánh giá bằng các chỉ số như Mean Squared Error (MSE) và R-squared. Mean Squared Error đo lường trung bình bình phương của các sai số dự đoán, trong khi R-squared đánh giá mức độ phù hợp của mô hình với dữ liệu quan sát.

Nghiên cứu đã đạt được các kết quả khả quan với các mô hình dự đoán thời tiết, trong đó mô hình Random Forest Regression cho thấy hiệu quả tốt hơn so với Decision Tree. Các mô hình đã thành công trong việc dự đoán các yếu tố thời tiết với độ chính xác cao, giúp cải thiện khả năng dự báo thời tiết trong tương lai.

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC CHỮ VIẾT TẮT

RFR: Random Forest Regression

TP: True Positive - Số trường hợp dương tính đúng

TN: True Negative - Số trường hợp âm tính đúng

FP: False Positive - Số trường hợp dương tính sai (Loại I error)

FN: False Negative - Số trường hợp âm tính sai (Loại II error)

CÁC ĐỘ ĐO SỬ DỤNG VÀ CÔNG THỨC TÍNH

ACC: Accuracy(độ chính xác)

PRE: Precision(độ chính xác dương)

REC: Recall(độ nhạy)

F1: F1-score(Điểm F1)

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

MAE: Mean Absolute Error (Sai số tuyệt đối trung bình).

DANH MỤC CÁC HÌNH VẼ

DỰ BÁO THỜI TIẾT BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY

1. Giới thiệu về bài toán

Bài toán này là bài toán dự đoán thời tiết bằng các phương pháp học máy. Cụ thể, nghiên cứu này áp dụng các kỹ thuật học máy như Decision Tree Regression và Random Forest Regression để dự đoán thời tiết. Bài toán dự đoán thời tiết này sử dụng dữ liệu thu thập từ nhiều tỉnh khác nhau, với mục tiêu xây dựng các mô hình học máy có khả năng dự đoán chính xác các điều kiện thời tiết dựa trên dữ liệu lịch sử. Việc dự đoán này không chỉ giới hạn ở một khu vực cụ thể mà còn có thể mở rộng ra phạm vi rộng hơn, giúp tăng khả năng dự báo và ứng phó với các điều kiện thời tiết khác nhau.

Bài toán dự đoán thời tiết có ý nghĩa vô cùng quan trọng trong nhiều lĩnh vực của cuộc sống và kinh tế. Đầu tiên, trong lĩnh vực nông nghiệp, dự đoán thời tiết chính xác giúp người nông dân có thể lập kế hoạch gieo trồng, tưới tiêu và thu hoạch một cách hiệu quả, từ đó tăng năng suất và giảm thiểu thiệt hại do thời tiết xấu. Thứ hai, trong giao thông vận tải, dự báo thời tiết giúp các cơ quan quản lý và người điều khiển phương tiện có thể chuẩn bị và ứng phó kịp thời với các điều kiện thời tiết khắc nghiệt, giảm thiểu tai nạn và đảm bảo an toàn giao thông. Thứ ba, trong lĩnh vực phòng chống thiên tai, việc dự đoán chính xác các hiện tượng thời tiết cực đoan như bão, lũ lụt, và hạn hán có thể giúp chính quyền và các cơ quan chức năng triển khai các biện pháp phòng ngừa và cứu trợ hiệu quả, giảm thiểu thiệt hại về người và tài sản.

vs2. Phân tích yêu cầu của bài toán

2.1. Yêu cầu của bài toán

Bài toán dự đoán thời tiết bằng các phương pháp học máy có các yêu cầu cụ thể như sau:

+ **Thu thập dữ liệu:** Dữ liệu thời tiết được lấy trên kaggle yêu cầu phải được thu thập từ nhiều tỉnh và các nguồn khác nhau, bao gồm các yếu tố như nhiệt độ, độ ẩm, lượng mưa, tốc độ và hướng gió, áp suất khí quyển, và các yếu tố khí tượng khác.

+ **Tiền xử lý dữ liệu:** Dữ liệu thô cần được làm sạch để loại bỏ các giá trị thiếu và ngoại lệ. Quá trình này bao gồm xử lý các giá trị thiếu bằng các phương pháp như trung bình, trung vị, hoặc loại bỏ các bản ghi không đầy đủ, và loại bỏ các giá trị ngoại lệ bằng phương pháp Interquartile Range (IQR) để đảm bảo tính nhất quán và chất lượng của dữ liệu.

+ **Xây dựng mô hình:** Sử dụng các kỹ thuật học máy như Decision Tree và Random Forest Regression để xây dựng các mô hình dự đoán thời tiết. Các mô hình này cần được huấn luyện trên tập dữ liệu huấn luyện và được kiểm tra để đảm bảo tính chính xác và khả năng tổng quát hóa.

+ **Đánh giá mô hình:** Sử dụng các chỉ số đánh giá như accuracy, precision, recall, F1-score để đánh giá hiệu quả của các mô hình.

+ **Tối ưu hóa mô hình:** Điều chỉnh các tham số của mô hình để cải thiện độ chính xác dự đoán.

2.2. Các phương pháp giải quyết bài toán

Decision Tree(ID3):

- Nghiên cứu tham khảo: “Decision tree for the weather forecasting”(Rajesh Kumar, Ph.D, 2013)

- Phương pháp giải quyết: Sử dụng phương pháp Decision Tree để dự đoán các hiện tượng thời tiết như sương mù, mưa, và giông bão. Mô hình cây quyết định được xây dựng bằng cách sử dụng các thuộc tính thời tiết như nhiệt độ trung bình, độ ẩm, và áp suất khí quyển.

- Dữ liệu thực nghiệm: Dữ liệu thời tiết được thu thập từ trang web Weather Underground trong một năm, bao gồm 64 bản ghi huấn luyện và 72 bản ghi kiểm tra. Các tham số được sử dụng trong mô hình bao gồm nhiệt độ trung bình, độ ẩm trung bình, và áp suất biển.

- Kết quả đạt được: Mô hình đạt được kappa statistics là 0.0584, với 46 trong số 72 bản ghi kiểm tra được phân loại chính xác. Kết quả cho thấy mô hình có tiềm năng nhưng cần cải thiện thêm bằng cách sử dụng nhiều thuộc tính hơn và tăng kích thước dữ liệu huấn luyện

- Hạn chế:

+ Mô hình dễ bị overfitting khi dữ liệu có nhiễu

+ Hiệu quả phân loại chưa cao cần cải thiện bằng cách tăng kích thước dữ liệu và thêm các thuộc tính dự báo khác

+ Cần áp dụng các kỹ thuật như pre-pruning và post-pruning để giảm thiểu vấn đề overfitting

Random Forest Regression

- Nghiên cứu tham khảo: "Weather Prediction Using Random Forest Machine Learning Model(R. Meenal, Prawin Angel Michael, D. Pamela, E. Rajasekaran, 2021)"

- Phương pháp giải quyết: Bài báo sử dụng mô hình hồi quy rừng ngẫu nhiên (Random Forest Regression) để dự đoán bức xạ mặt trời toàn cầu (Global Solar Radiation - GSR) và tốc độ gió cho bang Tamil Nadu, Ấn Độ. Mô hình này được so sánh với các mô hình hồi quy thống kê và mô hình máy học SVM (Support Vector Machine).

- Dữ liệu thực nghiệm: Dữ liệu thực nghiệm bao gồm các tham số khí tượng như nhiệt độ tối đa, nhiệt độ tối thiểu, áp suất bề mặt, độ ẩm tương đối, cùng với tháng, vĩ độ và kinh độ. Dữ liệu được thu thập từ Cục Khí tượng Ấn Độ (IMD) tại Pune. Dữ liệu được chia thành hai phần: tập huấn luyện (70%) và tập kiểm tra (30%).

- Kết quả đạt được: Độ chính xácc của mô hình hồi quy rừng ngẫu nhiên đạt được giá trị R^2 là 0.97 và MSE là 0.750, cho thấy độ chính xác cao hơn so với các mô hình hồi quy thống kê và mô hình SVM.

2.3. Phương pháp đề xuất giải quyết bài toán

- Dự đoán thời tiết là một bài toán yêu cầu phân tích và xử lý một lượng lớn dữ liệu từ nhiều nguồn khác nhau. Để dự đoán các điều kiện thời tiết trong tương lai, việc sử dụng các thuật toán học máy (machine learning) giúp tối ưu hóa quá trình này bằng cách học từ dữ liệu có sẵn và tìm ra các mẫu. Hai thuật toán được lựa chọn để giải quyết bài toán này là ID3 (Iterative Dichotomiser 3) và Random Forest Regression

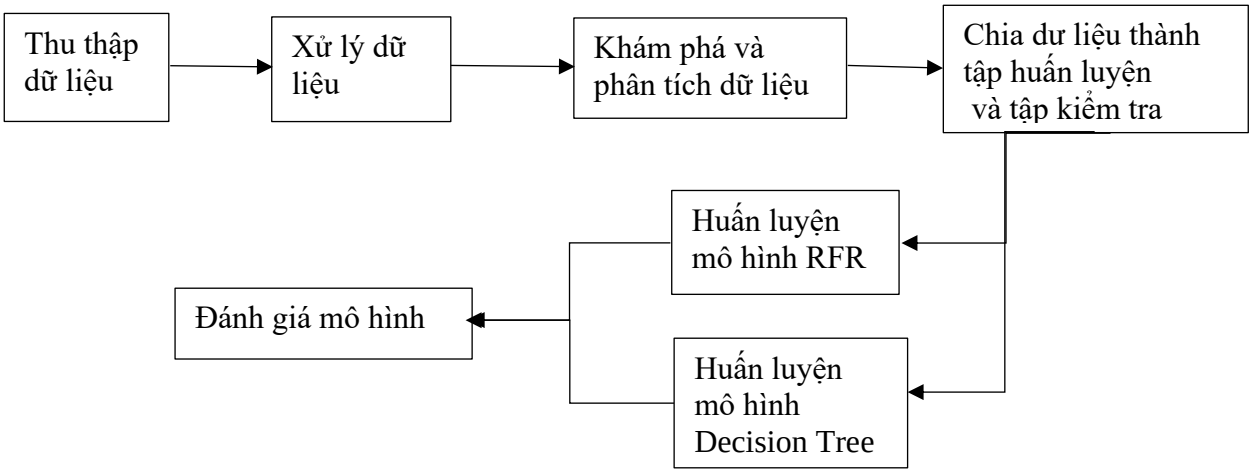
(RFR). Mỗi thuật toán đều có những ưu điểm riêng biệt, giúp nâng cao độ chính xác của dự báo thời tiết.

- Decision Tree là một thuật toán học máy đơn giản nhưng mạnh mẽ, cho phép phân tích dữ liệu và ra quyết định dựa trên các điều kiện phân nhánh. Ưu điểm chính của Decision Tree là khả năng xử lý dữ liệu phi tuyến tính và dễ dàng hiểu được mô hình do cấu trúc cây rõ ràng. Đối với dự đoán thời tiết, Decision Tree có thể giúp xác định các yếu tố quan trọng ảnh hưởng đến thời tiết như nhiệt độ, độ ẩm, áp suất, và gió. Bằng cách phân tích dữ liệu đã có sẵn, Decision Tree có thể thiết lập các quy tắc dựa trên những yếu tố này để dự đoán điều kiện thời tiết trong tương lai.

- Random Forest Regression là một phiên bản mở rộng của Decision Tree, kết hợp nhiều cây quyết định để cải thiện độ chính xác và độ ổn định của dự đoán. Bằng cách tạo ra một "rừng" các cây quyết định từ các mẫu dữ liệu ngẫu nhiên và trung bình kết quả, Random Forest Regressoion giảm thiểu nguy cơ quá khớp (overfitting) và cung cấp dự đoán chính xác hơn. Điều này đặc biệt quan trọng trong dự đoán thời tiết, nơi mà dữ liệu có thể biến động lớn và phức tạp. Random Forest Regression có khả năng xử lý một lượng lớn dữ liệu và khai thác toàn diện các mối quan hệ phức tạp giữa các biến số, từ đó cung cấp dự đoán thời tiết tin cậy hơn.

3. Phương pháp giải quyết bài toán

3.1. Mô hình tổng quát



1. Sơ đồ mô hình tổng quát

3.2. Đặc trưng của mô hình đề xuất

- **Thu thập dữ liệu:** Dữ liệu thời tiết phải được thu thập từ nhiều tỉnh và các nguồn khác nhau. Dữ liệu này bao gồm các thông số như nhiệt độ cao nhất, nhiệt độ thấp nhất, tốc độ gió, hướng gió, lượng mưa, độ ẩm, độ mây, áp suất, ngày tháng, khu vực và tình trạng thời tiết. Dữ liệu này sẽ cung cấp nền tảng cho việc xây dựng và huấn luyện mô hình dự báo thời tiết.

- **Xử lý dữ liệu:** Dữ liệu sau khi thu thập sẽ trải qua các bước xử lý để đảm bảo chất lượng. Trước tiên, dữ liệu ngày tháng được chuyển đổi sang định dạng datetime. Các giá trị ngoại lai trong các cột số được xử lý bằng cách loại bỏ các hàng có giá trị vượt quá phạm vi chấp nhận (dựa trên phương pháp IQR). Các biến phân loại như hướng gió và thời tiết được mã hóa bằng LabelEncoder. Cuối cùng, các đặc trưng ngày tháng như năm, tháng, ngày, và thứ trong tuần được trích xuất từ cột ngày tháng để phục vụ cho việc huấn luyện mô hình.

- **Chia dữ liệu thành tập huấn luyện và tập kiểm tra:** Dữ liệu sau khi xử lý sẽ được chia thành tập huấn luyện và tập kiểm tra. Đối với bài toán hồi quy và phân loại, dữ liệu được chia thành hai phần: tập huấn luyện chiếm 70% và tập kiểm tra chiếm 30%. Quá trình chia này được thực hiện bằng hàm train_test_split từ thư viện scikit-learn, đảm bảo dữ liệu được phân bổ ngẫu nhiên và không bị thiên lệch.

- **Xây dựng và huấn luyện mô hình:** Quá trình xây dựng và huấn luyện mô hình bao gồm việc sử dụng hai thuật toán chính: Rừng ngẫu nhiên (Random Forest) cho hồi quy và Cây quyết định (Decision Tree) cho phân loại.

+ Random Forest Regression:

- Đầu tiên, các đặc trưng và biến mục tiêu được xác định từ dữ liệu đã xử lý. Các đặc trưng bao gồm 'year', 'month', 'day', 'dayofweek', 'wind_d_encoded', và 'weather_encoded'. Các biến mục tiêu là các thuộc tính thời tiết số học như nhiệt độ cao nhất (max), nhiệt độ thấp nhất (min), tốc độ gió (wind), lượng mưa (rain), độ ẩm (humidi), độ mây (cloud), và áp suất (pressure).
- Mô hình hồi quy Rừng ngẫu nhiên (RandomForestRegressor) được huấn luyện trên tập huấn luyện. Mô hình sử dụng 100 cây quyết định

(n_estimators=100) và một giá trị hạt giống ngẫu nhiên (random_state=42) để đảm bảo tính tái lập của kết quả.

- Sau khi huấn luyện, mô hình này sẽ dự đoán trên tập kiểm tra để đánh giá hiệu suất. Các kết quả dự đoán sẽ được so sánh với các giá trị thực tế để tính toán các chỉ số hiệu suất như Mean Squared Error (MSE) và Mean Absolute Error (MAE).

+ Decision Tree:

- Đặc trưng đầu vào (features): Tất cả các cột trong dữ liệu trừ cột 'weather' sẽ được sử dụng làm đặc trưng đầu vào. Các đặc trưng bao gồm các cột số học và các cột phân loại đã được mã hóa như 'province', 'region', 'wind_d', và 'date'. Biến mục tiêu (target): Cột 'weather' được sử dụng làm biến mục tiêu cho phân loại và được mã hóa thành các giá trị số sử dụng LabelEncoder.
- Mô hình phân loại Cây quyết định (DecisionTreeClassifier) được huấn luyện trên tập huấn luyện. Mô hình sử dụng giá trị random_state=42 để đảm bảo tính tái lập của kết quả.
- Sau khi huấn luyện, mô hình sẽ dự đoán trên tập kiểm tra để đánh giá hiệu suất. Các kết quả dự đoán sẽ được so sánh với các giá trị thực tế để tính toán độ chính xác (accuracy) và các chỉ số hiệu suất khác như F1 score, recall, Mean Squared Error (MSE) và Mean Absolute Error (MAE).

- Đánh giá mô hình: Hiệu suất của các mô hình được đánh giá trên tập kiểm tra bằng các chỉ số như Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) cho mô hình hồi quy và accuracy cho mô hình phân loại. Đối với mô hình phân loại, độ chính xác được tính toán riêng cho từng thuộc tính và trung bình để đánh giá tổng quan. Hiệu suất của mô hình hồi quy được đánh giá bằng cách so sánh giá trị dự đoán với giá trị thực tế và tính toán tỷ lệ phần trăm chính xác cho từng thuộc tính.

4.Thực nghiệm

4.1 Dữ liệu

- Dữ liệu thời tiết này được lấy từ nguồn Kaggle, một nền tảng chia sẻ và phân tích dữ liệu mở rộng rãi trong cộng đồng khoa học dữ liệu. Tập dữ liệu chứa thông tin thời tiết của các tỉnh thành Việt Nam qua nhiều năm, gồm 181,960 dòng và 12 cột thông tin. Các cột bao gồm:

+ province: Tên tỉnh/thành phố
+ max: Nhiệt độ cao nhất trong ngày (°C)
+ min: Nhiệt độ thấp nhất trong ngày (°C)
+ wind: Tốc độ gió (Km/h)
+ wind_d: Hướng gió
+ rain: Lượng mưa (mm)
+ humidi: Độ ẩm (%)
+ cloud: Độ che phủ mây (%)
+ pressure: Áp suất khí quyển (hPa)
+ date: Ngày ghi nhận dữ liệu
+ region: Khu vực địa lý (Nam, Bắc, Trung)
+ weather: Mô tả thời tiết (nắng, mưa, gió, mây)

4.2 Xử lý dữ liệu

- Phải tiền xử lý dữ liệu bởi vì tiền xử lý dữ liệu là một bước quan trọng giúp làm sạch, biến đổi và định dạng lại dữ liệu để phù hợp với yêu cầu của các thuật toán học máy. Nếu không tiền xử lý dữ liệu đúng cách, kết quả dự đoán có thể bị sai lệch hoặc không chính xác do ảnh hưởng của các giá trị thiếu, dữ liệu nhiễu hoặc định dạng không phù hợp.
- Có thể xử lý dữ liệu bằng những cách sau:

+ Kiểm tra và xử lý giá trị thiếu (Missing Values): kiểm tra dữ liệu để phát hiện các giá trị bị thiếu. Sau đó thay thế các giá trị đó bằng các phương pháp như trung bình, trung vị, mode hoặc có thể loại bỏ giá trị đó.
+ Mã hóa biến phân **loại:** Chuyển đổi các biến phân loại (như province, wind_d, region, weather) thành dạng số để các thuật toán có thể xử lý. Có thể sử dụng các kỹ thuật như One-Hot Encoding hoặc Label Encoding.
+ Loại bỏ hoặc điều chỉnh dữ liệu nhiễu (Outliers): Phát hiện và xử lý các giá trị ngoại lai (outliers) để đảm bảo rằng dữ liệu không bị méo mó.

4.3 Công nghệ sử dụng

- Bài toán dự đoán thời tiết được hiện thực bằng ngôn ngữ lập trình Python. Các thư viện sử dụng bao gồm pandas để xử lý và quản lý dữ liệu, scikit-learn để thực hiện các thuật toán học máy như Decision Tree và Random Forest Regression, mã hóa dữ liệu như Label Encoder.

4.4 Cách đánh giá

- Các độ đo được sử dụng để đánh giá bao gồm: Accuracy, Precision, Recall, F1-Score, MSE

- Công thức tính:

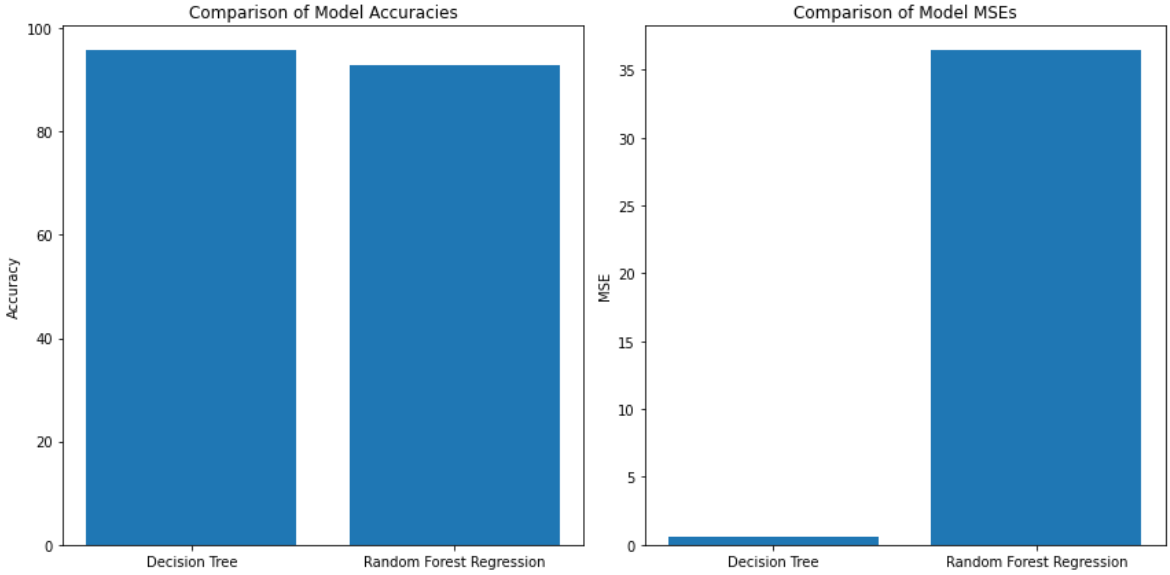
- Accuracy: (TP + TN)/(TP + TN + FP + FN)
- Precision: TP/(TP + FP)
- Recall: TP/(TP + FN)
- F1-Score: 2*(Precision * Recall)/(Precision + Recall)
- MSE: $(1/n)*\Sigma(y\_true - y\_pred)^2$

5. Kết quả đạt được

- Kết quả đánh giá:

- **Random Forest Regression:** Độ chính xác tổng thể: 92-93%, MSE = 28.63

- **Decison Tree:** Độ chính xác: 95.66 %, MSE = 0.563



Hình 5.1. Biểu đồ so sánh Accuracy và MSE của 2 mô hình

- Giải thích sự khác biệt: Sự khác biệt về kết quả giữa Decision Tree và Random Forest Regression xuất phát từ cách chúng xử lý dữ liệu và giảm thiểu lỗi dự đoán. Random Forest với kỹ thuật ensemble, giúp giảm overfitting và cải thiện độ chính xác tổng thể, mặc dù MSE có thể cao hơn trong một số trường hợp. Điều này cho thấy Random Forest có khả năng tổng quát hóa tốt hơn và hoạt động ổn định hơn trên dữ liệu mới so với Decision Tree. Trong khi Decision Tree dễ bị overfitting do nó tối ưu hóa trên toàn bộ tập dữ liệu bằng cách phân chia nó thành các node nhỏ hơn. Điều này dẫn đến mô hình học quá chi tiết từ dữ liệu huấn luyện, và khi gặp dữ liệu mới, mô hình có thể hoạt động kém

6. Kết luận

Trong nghiên cứu của bạn, mô hình Random Forest Regression (RFR) đã cho thấy độ chính xác cao hơn so với mô hình Decision Tree (DT). Tuy nhiên, về mặt Mean Squared Error (MSE), mô hình Decision Tree lại cho thấy kết quả tốt hơn, tức là có sai số thấp hơn so với RFR. Điều này chỉ ra rằng trong khi RFR có khả năng phân loại tốt hơn, Decision Tree lại có khả năng dự đoán giá trị cụ thể của các tham số thời tiết với sai số thấp hơn.

Mặc dù RFR cho độ chính xác cao, nhưng lại yêu cầu nhiều tài nguyên tính toán hơn và thời gian huấn luyện lâu hơn so với Decision Tree. Decision Tree dù có MSE thấp nhưng lại kém linh hoạt hơn trong việc tổng quát hóa dữ liệu mới hoặc dữ liệu nhiễu. Cả hai mô hình đều có thể gặp khó khăn trong việc xử lý các biến động lớn của dữ liệu môi trường.

Để cải thiện hiệu suất và khả năng dự đoán chính xác của các mô hình, có thể xem xét áp dụng các mô hình học máy khác như Gradient Boosting hoặc Neural Networks để cải thiện độ chính xác và giảm sai số. Việc kết hợp các kỹ thuật feature engineering và feature selection cũng sẽ hỗ trợ giảm thiểu nhiễu và tăng cường độ chính xác của dự đoán. Ngoài ra, việc áp dụng các kỹ thuật học sâu có thể khám phá các mối quan hệ phức tạp hơn trong dữ liệu thời tiết, từ đó cải thiện hiệu quả của mô hình.

LÀM VIỆC NHÓM

Cách Thức Làm Việc Nhóm

- Nhóm của chúng tôi làm việc chủ yếu qua các nền tảng trực tuyến như Messenger, Zalo để trao đổi thông tin hàng ngày và sử dụng Zoom để tổ chức các cuộc họp chính thức. Mỗi cuộc họp trực tuyến qua Zoom thường kéo dài khoảng 2 giờ, được tổ chức hai buổi 1 tuần. Trong các cuộc họp, chúng tôi thảo luận về tiến trình công việc, phân công nhiệm vụ, và giải quyết các vấn đề phát sinh. Việc sử dụng các công cụ trực tuyến giúp tăng cường sự linh hoạt và thuận tiện cho tất cả thành viên, đặc biệt là khi cần phản hồi nhanh chóng hoặc điều chỉnh kế hoạch công việc.

- Tổng Số Lần Gặp Nhau

+ Tổng số buổi: 14 buổi (7 tuần, mỗi tuần 2 buổi).

+ Tổng Thời Gian Gặp Nhau

+ Tổng thời gian: 28 giờ (mỗi buổi 2 giờ).

Trong suốt quá trình làm việc, nhóm đã duy trì sự gắn kết và hợp tác chặt chẽ, đảm bảo rằng mọi thành viên đều có đủ nguồn lực và thông tin cần thiết để hoàn thành nhiệm vụ được giao. Việc phân chia công việc rõ ràng và có trách nhiệm đã giúp tăng hiệu quả làm việc và đạt được mục tiêu của dự án một cách thành công.

TỰ ĐÁNH GIÁ (Bài nhóm)

STT	Nội dung	Điểm chuẩn	Tự chấm	Ghi chú
1 (8.5đ)	**1.1 Giới thiệu về bài toán**	0.5	0.5	
	1.2 Phân tích yêu cầu của bài toán	1	1	
	1.3 Phương pháp giải quyết bài toán	1.5	1.5	
	1.4 Thực nghiệm	4	4	
	1.5 Kết quả đạt được	1	0.5	
	1.6 Kết luận	0.5	0.5	
2 (1đ)	**Báo cáo (**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)**	1đ	1	
3 (0.5đ)	**Điểm nhóm** (chú ý trả lời các câu hỏi trong mục làm việc nhóm)	0.5đ	0.5	
Tổng điểm			9.5	

TỰ ĐÁNH GIÁ (Bài cá nhân)

STT	Nội dung	Điểm chuẩn	Tự chấm	Ghi chú
1 (9đ)	**1.1 Giới thiệu về bài toán**	0.5	0.5	
	1.2 Phân tích yêu cầu của bài toán	1	1	
	1.3 Phương pháp giải quyết bài toán	1.5	1.5	
	1.4 Thực nghiệm	4.5	4.5	
	1.5 Kết quả đạt được	1	0.5	
	1.6 Kết luận	0.5	0.5	
2 (1đ)	**Báo cáo (**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)**	1đ	1	
Tổng điểm			9.5	