TRƯỜNG ĐẠI HỌC HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

**NHẬP MÔN TRÍ TUỆ NHÂN TẠO**

NGÀNH: CÔNG NGHỆ THÔNG TIN

***Tên Đồ án:***

**ÁP DỤNG THUẬT TOÁN**

**K – MEANS CLUSTERING VÀO PHÂN CỤM VÀ DỰ BÁO THỜI TIẾT**

**SVTH: NGUYỄN PHÚC NGUYÊN**

**MSSV: 2274802010586**

**GVHD: TRẦN NGỌC VIỆT**

Tp. Hồ Chí Minh – năm 2024

LỜI CẢM ƠN  
-------------

Viết một báo cáo đồ án môn học là một trong những việc khó nhất mà chúng em phải  
hoàn thành trong quá trình học một môn học. Trong quá trình thực hiện đề tài chúng  
em đã gặp rất nhiều khó khăn và bỡ ngỡ. Nếu không có những sự giúp đỡ và lời động  
viên chân thành của nhiều người có lẽ chúng em khó có thể hoàn thành tốt tiểu luận  
này. Đầu tiên chúng em xin gửi lời biết ơn chân thành đến thầy Trần Ngọc Việt,  
người trực tiếp hướng dẫn chúng em hoàn thành dự án này.

Những ý kiến đóng góp của thầy là vô cùng hữu ích, nó giúp chúng em nhận ra các  
khuyết điểm của đồ án. Cảm ơn thầy và các bạn trường Đại học Văn Lang là những  
người đã cùng nhóm em sát cánh và trải nghiệm để hoàn thành đồ án môn học.

Người thực hiện báo cáo

*Nguyễn Phúc Nguyên*

Mục lục

[LỜI CẢM ƠN 2](#_Toc182293800)

[CHƯƠNG 1. MỞ ĐẦU 4](#_Toc182293801)

[1.1 Lí do chọn đề tài 4](#_Toc182293802)

[1.2 Mục tiêu của đề tài 4](#_Toc182293803)

[1.3 Cơ sở lí thuyết 4](#_Toc182293804)

[CHƯƠNG 2. ÁP DỤNG BÀI TOÁN K-MEAN 5](#_Toc182293805)

[2.1 Phát biểu bài toán và input: 5](#_Toc182293806)

[2.2 Hướng giải quyết bài toán 5](#_Toc182293807)

[2.3 Kết quả đầu ra (Output): 6](#_Toc182293808)

[2.4 Mã nguồn 7](#_Toc182293809)

[2.5 Kết luận và nhận xét: 12](#_Toc182293810)

[CHƯƠNG 3. KẾT LUẬN 13](#_Toc182293811)

[3.1. Kết quả đạt được 13](#_Toc182293812)

[3.2. Hạn chế của mô hình 13](#_Toc182293813)

[3.3. Hướng cải thiện trong tương lai 13](#_Toc182293814)

[3.4. Tổng kết 14](#_Toc182293815)

[TÀI LIỆU THAM KHẢO 15](#_Toc182293816)

# CHƯƠNG 1. MỞ ĐẦU

## 1.1 Lí do chọn đề tài

Thời tiết là một yếu tố quan trọng ảnh hưởng đến đời sống và sinh hoạt của con người. Trong bối cảnh biến đổi khí hậu hiện nay, các hiện tượng thời tiết cực đoan như bão, lũ lụt, hay nắng nóng kéo dài đang gia tăng, đe dọa nghiêm trọng đến tài sản và tính mạng con người. Việc dự đoán chính xác các diễn biến của thời tiết có thể giúp chúng ta giảm thiểu thiệt hại và đưa ra các biện pháp ứng phó kịp thời. Chính vì lý do này đã lựa chọn đề tài dự đoán thời tiết nhằm phân tích các yếu tố ảnh hưởng và cải thiện độ chính xác trong việc dự báo.

## 1.2 Mục tiêu của đề tài

Mục tiêu của đề tài là sử dụng phương pháp K-means clustering để phân cụm dữ liệu thời tiết từ một tập dữ liệu có sẵn, và sau đó sử dụng mô hình Random Forest để dự đoán nhiệt độ cho 7 ngày tiếp theo. Kết quả sẽ được hiển thị trực quan qua một ứng dụng web, giúp người dùng dễ dàng theo dõi và hiểu rõ hơn về xu hướng thời tiết.

## A diagram of a number of dots Description automatically generated1.3 Cơ sở lí thuyết

Thuật toán phân cụm đơn giản nhất: K-means

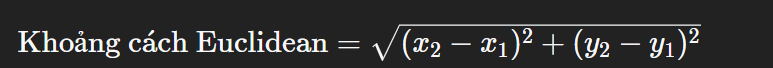
Thuật toán phân cụm K-means

- Khái niệm chung về phân cụm:

Giả sử có K lớp

Biểu diễn mỗi lớp với tâm của nó

Bắt đầu với một số tâm ban đầu ngẫu nhiên

Tìm các tâm tốt nhất cho các lớp, tức là tối ưu hóa các tâm

- Trong giai đoạn phân loại:

Tìm khoảng cách của một mẫu mới từ tất cả các tâm

Tìm tâm có khoảng cách từ mẫu mới là tối thiểu

Phân loại mẫu mới vào lớp có tâm có khoảng cách tối thiểu từ mẫu

# **CHƯƠNG 2. ÁP DỤNG BÀI TOÁN K-MEAN**

## 2.1 Phát biểu bài toán và input:

Mục tiêu

Mục tiêu của bài toán là phân cụm các dữ liệu thời tiết dựa trên nhiệt độ trung bình, tối đa, và tối thiểu, từ đó dự đoán các giá trị này trong tương lai. Kết quả phân cụm có thể giúp phát hiện các mô hình thời tiết và cải thiện dự báo.

Dữ liệu đầu vào (Input): File data\_weather.csv được lấy về từ trang web [Visual Crossing](https://www.visualcrossing.com/): Gồm 7 cột và hơn 9000 dòng dữ liệu.

A screenshot of a computer

Description automatically generated

Dữ liệu thời tiết bao gồm nhiều cột như:

* temp: Nhiệt độ trung bình
* tempmin: Nhiệt độ thấp nhất
* tempmax: Nhiệt độ cao nhất
* precip: Lượng mưa

## 2.2 Hướng giải quyết bài toán

Các bước thực hiện

Tiền xử lý dữ liệu:

* Xử lý các giá trị bị thiếu.
* Chuẩn hóa dữ liệu để đảm bảo tất cả các đặc trưng (feature) nằm trong cùng một khoảng giá trị.

Áp dụng K-means Clustering:

* Chọn số lượng cụm k phù hợp bằng cách thử nghiệm nhiều giá trị k khác nhau và chọn giá trị tối ưu. Sau nhiều lần thửu nghiệm thì chọn k = 3.
* Thực hiện phân cụm các bản ghi thời tiết dựa trên các đặc trưng đã được chuẩn hóa.

Phân tích và đánh giá kết quả phân cụm:

* Đánh giá các cụm đã tạo, kiểm tra xem các cụm có phản ánh các trạng thái thời tiết khác nhau hay không.
* Gán nhãn hoặc mô tả cho từng cụm để nhận biết các mô hình thời tiết phổ.

Sử dụng kết quả phân cụm trong dự báo:

* Sau khi đã có các cụm, tiếp tục theo dõi và gán nhóm cho các ngày mới dựa trên đặc trưng thời tiết.
* So sánh với các dự báo và xác định nhóm gần nhất để dự đoán trạng thái thời tiết.
* Sử dụng mô hình Random Forest để dự đoán

2.3 Kết quả đầu ra (Output):  
A chart of different colored dots

Description automatically generated

Nhận xét:

In ra biểu đồ kết quả đã được phân cụm dựa trên tập dữ liệu được đưa vào huấn luyện mô hình bằng K-mean clustering.

Và in ra kết quả dự đoán thời tiết cho 7 ngày tiếp theo.

Dự đoán cho ngày 2024-11-02: Nhiệt độ tối đa = 19.80, Nhiệt độ tối thiểu = 23.10, Nhiệt độ trung bình = 21.45

Dự đoán cho ngày 2024-11-03: Nhiệt độ tối đa = 21.40, Nhiệt độ tối thiểu = 22.70, Nhiệt độ trung bình = 22.05

Dự đoán cho ngày 2024-11-04: Nhiệt độ tối đa = 20.10, Nhiệt độ tối thiểu = 23.00, Nhiệt độ trung bình = 21.55

Dự đoán cho ngày 2024-11-05: Nhiệt độ tối đa = 21.40, Nhiệt độ tối thiểu = 23.30, Nhiệt độ trung bình = 22.35

Dự đoán cho ngày 2024-11-06: Nhiệt độ tối đa = 19.90, Nhiệt độ tối thiểu = 23.50, Nhiệt độ trung bình = 21.70

Dự đoán cho ngày 2024-11-07: Nhiệt độ tối đa = 20.00, Nhiệt độ tối thiểu = 24.00, Nhiệt độ trung bình = 22.00

Dự đoán cho ngày 2024-11-08: Nhiệt độ tối đa = 21.10, Nhiệt độ tối thiểu = 25.90, Nhiệt độ trung bình = 23.50

## 2.4 Mã nguồn

File app.py: Là file chứa chương trình chính xử lí back-end của dự án:

from flask import Flask, render\_template

import pandas as pd

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestRegressor

import matplotlib.pyplot as plt

import numpy as np

app = Flask(\_\_name\_\_)

# Hàm để phân cụm dữ liệu

def cluster\_weather\_data():

    # Tải dữ liệu

    data\_weather = pd.read\_csv('data\_weather.csv')

    # Chuyển đổi cột 'date' thành định dạng datetime

    if 'datetime' in data\_weather.columns:

        data\_weather['datetime'] = pd.to\_datetime(data\_weather['datetime'])

    # Chọn các cột để phân cụm

    # X = data\_weather[['tempmax', 'tempmin', 'humidity']]

    X = data\_weather[['temp']]

    # Chuẩn hóa dữ liệu

    scaler = StandardScaler()

    X\_scaled = scaler.fit\_transform(X)

    # Tạo mô hình K-means

    kmeans = KMeans(*n\_clusters*=3, *random\_state*=0)

    clusters = kmeans.fit\_predict(X\_scaled)

    # Thêm kết quả vào DataFrame

    data\_weather['Cluster'] = clusters

    return data\_weather

# Hàm để vẽ biểu đồ

def plot\_weather\_data(*data*):

    plt.figure(*figsize*=(10, 6))

    plt.scatter(*data*['tempmax'], *data*['tempmin'], *c*=*data*['Cluster'], *cmap*='viridis', *marker*='o')

    plt.title('K-means Clustering of Weather Data')

    plt.colorbar(*label*='Cluster')

    plt.savefig('static/weather\_plot.png')  # Lưu biểu đồ vào thư mục static

    plt.close()  # Đóng biểu đồ để không hiển thị trên máy chủ

# Hàm để dự đoán nhiệt độ cho 7 ngày tiếp theo

def predict\_weather(*data*, *start\_date*='2024-11-01'):

    features = ['tempmax', 'tempmin', 'humidity']

    target\_tempmax = 'tempmax'

    target\_tempmin = 'tempmin'

    # Kiểm tra xem các cột cần thiết có tồn tại hay không

    if not all(col in *data*.columns for col in features + [target\_tempmax, target\_tempmin]):

        print("Dữ liệu không đầy đủ để dự đoán.")

        return []

    # Tạo mô hình hồi quy rừng ngẫu nhiên cho tempmax và tempmin

    model\_tempmax = RandomForestRegressor(*n\_estimators*=100, *random\_state*=42)

    model\_tempmin = RandomForestRegressor(*n\_estimators*=100, *random\_state*=42)

    # Chia dữ liệu thành tập huấn luyện

    X = *data*[features]

    y\_tempmax = *data*[target\_tempmax]

    y\_tempmin = *data*[target\_tempmin]

    # Huấn luyện mô hình

    model\_tempmax.fit(X, y\_tempmax)

    model\_tempmin.fit(X, y\_tempmin)

    # Dự đoán cho 7 ngày tiếp theo

    predictions = []

    # Kiểm tra kiểu dữ liệu cột datetime

    if *data*['datetime'].dtype == 'float64':

*data*['datetime'] = pd.to\_datetime(*data*['datetime'], *unit*='s', *origin*='unix')

    elif *data*['datetime'].dtype != 'datetime64[ns]':

        print("Cột datetime không có kiểu dữ liệu hợp lệ.")

        return []

    last\_row = *data*.iloc[-1]  # Lấy dòng cuối cùng của dữ liệu

    # Thay đổi ngày bắt đầu

*start\_date* = pd.to\_datetime(*start\_date*)

    for i in range(1, 8):  # Dự đoán cho 7 ngày tiếp theo

        next\_date = *start\_date* + pd.Timedelta(*days*=i)

        # Tạo dữ liệu mới để dự đoán với một chút nhiễu

        new\_data = pd.DataFrame({

            'tempmax': [last\_row['tempmax'] + np.random.uniform(-2, 2)],

            'tempmin': [last\_row['tempmin'] + np.random.uniform(-2, 2)],

            'humidity': [last\_row['humidity'] + np.random.uniform(-10, 10)]

        })

        # Dự đoán nhiệt độ tối đa và tối thiểu

        predicted\_tempmax = model\_tempmax.predict(new\_data)[0]

        predicted\_tempmin = model\_tempmin.predict(new\_data)[0]

        # Tính nhiệt độ trung bình

        predicted\_temp = (predicted\_tempmax + predicted\_tempmin) / 2

        predictions.append((next\_date.strftime('%Y-%m-%d'), predicted\_tempmax, predicted\_tempmin, predicted\_temp))

        # In ra kết quả dự đoán chi tiết

        print(f"Dự đoán cho ngày {next\_date.strftime('%Y-%m-%d')}: "

              f"Nhiệt độ tối đa = {predicted\_tempmax:.2f}, "

              f"Nhiệt độ tối thiểu = {predicted\_tempmin:.2f}, "

              f"Nhiệt độ trung bình = {predicted\_temp:.2f}")

    print("Kết quả dự đoán:", predictions)

    return predictions

@app.route('/')

def index():

    # Phân cụm dữ liệu

    data\_weather = cluster\_weather\_data()

    # Vẽ biểu đồ

    plot\_weather\_data(data\_weather)

    # Dự đoán thời tiết

    predictions = predict\_weather(data\_weather)

    # Trả về template với dữ liệu dự đoán

    return render\_template('index.html', *predictions*=predictions)

if \_\_name\_\_ == '\_\_main\_\_':

    app.run(*debug*=True)

Bên cạnh đó dự án còn được hiển thị trên website local để có thể trực quan được kết quả thu được:

File index.html

<!DOCTYPE html>

<html *lang*="vi">

<head>

    <meta *charset*="UTF-8">

    <meta *name*="viewport" *content*="width=device-width, initial-scale=1.0">

    <title>Dự Đoán Thời Tiết</title>

    <style>

        body {

            font-family: Arial, sans-serif;

            margin: 20px;

            background-color: #f4f4f4;

        }

        h1, h2 {

            color: #333;

        }

        table {

            width: 100%;

            border-collapse: collapse;

            margin-top: 20px;

            background-color: white;

        }

        th, td {

            border: 1px solid #ddd;

            padding: 8px;

            text-align: center;

        }

        th {

            background-color: #4CAF50;

            color: white;

        }

        img {

            max-width: 100%;

            height: auto;

        }

        img {

            display: block;

            margin: auto;

            max-width: 100%;

            height: auto;

        }

    </style>

</head>

<body>

    <h1>Dự Đoán Thời Tiết cho 7 Ngày Tới</h1>

    <h2>Biểu Đồ Phân Cụm</h2>

    <img *src*="{{ url\_for('static', filename='weather\_plot.png') }}" *alt*="Biểu đồ phân cụm">

    <h2>Kết Quả Dự Đoán</h2>

    <table>

        <tr>

            <th>Ngày</th>

            <th>Nhiệt Độ Tối Thiểu (°C)</th>

            <th>Nhiệt Độ Tối Đa (°C)</th>

            <th>Nhiệt Độ Trung Bình (°C)</th>

        </tr>

        {% for prediction in predictions %}

        <tr>

            <td>{{ prediction[0] }}</td>

            <td>{{ prediction[1] | round(2) }}</td>

            <td>{{ prediction[2] | round(2) }}</td>

            <td>{{ prediction[3] | round(2) }}</td>

        </tr>

        {% endfor %}

    </table>

</body>

</html>

## 2.5 Kết luận và nhận xét:

**Kết luận:**

Phương pháp K-means Clustering giúp phân loại và nhóm các ngày có trạng thái thời tiết tương tự dựa trên các đặc trưng thời tiết. Đây là một cách tiếp cận hiệu quả trong việc xác định các mẫu thời tiết mà không cần sử dụng đến các mô hình dự đoán phức tạp. Kết quả từ phân cụm có thể hỗ trợ trong phân tích, nhận dạng mẫu thời tiết, và cung cấp những dự đoán tổng quát dựa trên các trạng thái cụm.

**Nhận xét:**

Ưu điểm:

Đơn giản và hiệu quả: K-means là thuật toán đơn giản, dễ hiểu và triển khai, giúp tiết kiệm tài nguyên tính toán.

Phân loại nhanh chóng: K-means có thể nhanh chóng nhóm dữ liệu thời tiết vào các cụm, giúp xác định các mẫu thời tiết một cách nhanh chóng.

Khả năng mở rộng: Có thể áp dụng cho dữ liệu lớn và mở rộng thêm các biến số thời tiết mới để nhóm chính xác hơn.

Nhược điểm:

Phụ thuộc vào số cụm k: Việc lựa chọn số lượng cụm k thích hợp có thể khó khăn và yêu cầu thử nghiệm nhiều lần, nếu chọn không đúng có thể dẫn đến phân nhóm không chính xác.

Thiếu khả năng dự báo chi tiết: K-means chỉ tạo ra các nhóm chung chung, không thể dự báo cụ thể cho từng yếu tố thời tiết như nhiệt độ hoặc lượng mưa theo thời gian.\

# **CHƯƠNG 3. KẾT LUẬN**

Dự án đã thành công trong việc ứng dụng các thuật toán K-means và Random Forest để phân tích và dự báo thời tiết dựa trên bộ dữ liệu thời tiết lịch sử. Quá trình thực hiện bao gồm các bước thu thập, tiền xử lý dữ liệu và áp dụng các mô hình học máy nhằm đạt được mục tiêu dự báo các yếu tố thời tiết như nhiệt độ trung bình (tavg), nhiệt độ tối thiểu (tmin), nhiệt độ tối đa (tmax).

## **3.1. Kết quả đạt được**

Áp dụng K-means: Phương pháp K-means đã được sử dụng để phân cụm dữ liệu thời tiết nhằm phát hiện các xu hướng và mẫu hình ẩn trong dữ liệu. Việc phân nhóm giúp xác định các đặc điểm tương đồng giữa các ngày có điều kiện thời tiết tương tự, hỗ trợ cải thiện độ chính xác của mô hình dự báo.

Mô hình Random Forest: Sau khi phân cụm dữ liệu, mô hình Random Forest được áp dụng để dự đoán các yếu tố thời tiết. Kết quả cho thấy mô hình có khả năng dự báo khá chính xác nhiệt độ trung bình (tavg), nhiệt độ tối thiểu (tmin), nhiệt độ tối đa (tmax) trong khoảng thời gian 7 ngày.

## **3.2. Hạn chế của mô hình**

Mặc dù đạt được những kết quả khả quan, dự án vẫn tồn tại một số hạn chế cần khắc phục:

Độ chính xác dự báo dài hạn: Mô hình hiện tại chỉ hoạt động tốt trong dự báo ngắn hạn. Khi dự báo cho khoảng thời gian dài hơn, độ chính xác bị giảm do mô hình không thể nắm bắt được những biến động thời tiết phức tạp và bất ngờ.

Yêu cầu dữ liệu lớn: Để mô hình hoạt động hiệu quả, cần có một lượng lớn dữ liệu thời tiết lịch sử với độ chi tiết cao. Tuy nhiên, dữ liệu thu thập có thể không đầy đủ hoặc thiếu sót ở một số khu vực, dẫn đến sai lệch trong dự báo.

## **3.3. Hướng cải thiện trong tương lai**

Để nâng cao độ chính xác và hiệu quả của mô hình, cần thực hiện một số cải tiến như sau:

Tích hợp thêm các yếu tố thời tiết: Ngoài nhiệt độ và lượng mưa, có thể thu thập thêm các yếu tố như độ ẩm, tốc độ gió, áp suất khí quyển, và thời gian nắng để làm phong phú thêm dữ liệu đầu vào.

Sử dụng các mô hình học sâu (Deep Learning): Các mô hình mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN), đặc biệt là LSTM và GRU, có khả năng nắm bắt tốt hơn các quan hệ thời gian và có thể cải thiện dự báo dài hạn.

Kết hợp nhiều mô hình (Ensemble Learning): Sử dụng kỹ thuật học máy tích hợp, kết hợp các mô hình khác nhau để tận dụng ưu điểm của từng mô hình, từ đó cải thiện độ chính xác và tính ổn định của dự báo.

## 3.4. Tổng kết

Dự án đã hoàn thành mục tiêu ban đầu là phân tích và dự báo thời tiết ngắn hạn, có khả năng ứng dụng các mô hình học máy vào các bài toán thời tiết thực tế. Tuy nhiên, vẫn cần tiếp tục cải tiến mô hình để đạt được độ chính xác cao hơn, đặc biệt là trong dự báo dài hạn. Những cải tiến và hướng nghiên cứu tiếp theo được đề xuất sẽ giúp nâng cao chất lượng dự báo và mang lại nhiều ứng dụng thực tiễn trong tương lai.

Qua dự án này, chúng tôi đã có thêm kinh nghiệm trong việc xử lý dữ liệu lớn, áp dụng mô hình học máy và tối ưu hóa dự báo thời tiết, đồng thời đóng góp một phần vào việc hiểu rõ hơn về những yếu tố tác động đến khí hậu và môi trường.

# **TÀI LIỆU THAM KHẢO**

[1] B. Gupta, R. Choudhary, and N. Saxena, “Weather Forecasting Using K-Means and ANN,” International Journal of Computer Applications, vol. 182, no. 45, pp. 1–4,2018.  
[2] Dominik Olszewski, “Asymmetric k-Means Clustering of the AsymmetricSelf-Organizing Map”, 2015.

[3] P. R. Kumari and G. L. Kumar, “Random Forest Algorithm for Predictive Data Analysis in Weather Forecasting,” International Journal of Engineering and Advanced Technology, vol. 8, no. 5, pp. 1122–1125, 2019.

[4] R. Meenal1, Prawin Angel Michael, D. Pamela and E. Rajasekaran, “Weather prediction using random forest machine learningmodel,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 22, no. 2, pp. 1208-1215, 2021