**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**



Nhóm 11 gồm các sinh viên:

1. Đỗ Duy Đức – 2151264650
2. Đào Nguyễn Gia Bảo - 2151261233
3. Nguyễn Đỗ Cường - 2151264647
4. Trịnh Đạt - 2151264649

**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN PHÂN TÍCH CHUỖI THỜI GIAN**

ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU NHIỆT ĐỘ TẠI NƯỚC TUSINIA

GVHD: TRẦN ANH ĐẠT

Hà Nội, ngày 15 tháng 6 năm 2024

# MỤC LỤC

Contents

[MỤC LỤC 2](#_Toc169447992)

[Phần I. Giới thiệu chung. 4](#_Toc169447993)

[Phần II. Một số mô hình chuỗi thời gian. 6](#_Toc169447994)

[1. Mô hình Prophet 6](#_Toc169447995)

[2. Mô hình LSTM 7](#_Toc169447996)

[3. Mô hình SARIMA 9](#_Toc169447997)

[Phần III. Xây dựng mô hình 10](#_Toc169447998)

[1. Mô tả dữ liệu 10](#_Toc169447999)

[2. HUẤN LUYỆN MÔ HÌNH 11](#_Toc169448000)

[2.1. Prophet 11](#_Toc169448001)

[2.2 SARIMA 13](#_Toc169448002)

[2.3. LSTM 15](#_Toc169448003)

[Phần IV: Đánh giá mô hình 20](#_Toc169448004)

[Phần V: Kết luận 21](#_Toc169448005)

**LỜI NÓI ĐẦU**

Trong bối cảnh biến đổi khí hậu và những thách thức môi trường toàn cầu, việc nghiên cứu và phân tích dữ liệu nhiệt độ trở nên cực kỳ quan trọng. Nước Tunisia, với vị trí địa lý đặc thù và khí hậu đặc trưng của khu vực Bắc Phi, không nằm ngoài tác động của các hiện tượng thời tiết và biến đổi khí hậu toàn cầu. Việc theo dõi và phân tích dữ liệu nhiệt độ tại Tunisia không chỉ giúp hiểu rõ hơn về các xu hướng thời tiết hiện tại mà còn cung cấp cơ sở để dự báo và ứng phó với những biến đổi trong tương lai.

Đề tài "Phân tích dữ liệu nhiệt độ tại nước Tunisia" nhằm mục đích khám phá và làm sáng tỏ các xu hướng nhiệt độ, những biến động và các yếu tố ảnh hưởng đến nhiệt độ tại quốc gia này. Bằng cách sử dụng các mô hình chuỗi thời gian hiện đại, chúng tôi hy vọng sẽ cung cấp được cái nhìn tổng quan về tình hình nhiệt độ tại Tunisia, đồng thời đưa ra những dự báo và khuyến nghị cho việc ứng phó với biến đổi khí hậu.

Đề tài này sẽ tập trung vào việc thu thập, xử lý và phân tích dữ liệu nhiệt độ từ nguồn đáng tin cậy, bao gồm cả dữ liệu lịch sử và hiện tại. Chúng em sẽ sử dụng các công cụ và phương pháp phân tích dữ liệu tiên tiến để khám phá các mẫu và xu hướng, từ đó đưa ra những kết luận và đề xuất thực tiễn cho các nhà hoạch định chính sách, các tổ chức môi trường và cộng đồng địa phương.

Mong rằng kết quả của nghiên cứu này sẽ góp phần vào việc dự đoán về biến đổi khí hậu tại Tunisia, đồng thời hỗ trợ trong việc xây dựng các chiến lược ứng phó hiệu quả, đảm bảo sự phát triển bền vững cho đất nước và khu vực.

# Phần I. Giới thiệu chung.

1. Dữ liệu chuỗi thời gian.

Chuỗi thời gian (time series) là một chuỗi các điểm dữ liệu xảy ra theo thứ tự liên tiếp trong một khoảng thời gian. Một chuỗi thời gian sẽ theo dõi chuyển động của các điểm dữ liệu đã chọn (chẳng hạn như giá của chứng khoán) trong một khoảng thời gian xác định.

Dữ liệu chuỗi thời gian được ứng dụng phổ biến trong đời sống (theo dõi lượng mưa,nhiệt độ,thị trường chứng khoán…) mà còn được sử dụng để theo dõi và tìm ra xu hướng trong các hoạt động trong danh nghiệp theo thời gian.

Việc phân tích dữ liệu chuỗi thời gian sẽ giúp bạn:

* Hiểu đặc điểm của tệp dữ liệu, sự thay đổi của tệp dữ liệu theo thời gian.
* Xác định được những yếu tố ảnh hưởng đến các biến tại các thời điểm khác nhau.
* Hỗ trợ đưa ra dự đoán giá trị tương lai của các biến trong chuỗi thời gian dựa vào xu hướng của dữ liệu trong quá khứ.

1. Các yếu tố cần lưu ý khi phân tích dữ liệu chuỗi thời gian

Một dữ liệu chuỗi thời gian thường bị ảnh hưởng bởi thành 4 thành phần sau:

* Trend(xu hướng): Đề cập đến hướng di chuyển của dữ liệu qua thời gian, có thể là tăng lên hoặc giảm đi ,thường được nhận biết dựa trên độ dốc của đồ thị biểu diễn dữ liệu.

Một số xu hướng phổ biến :

* Upward Trend:  Có xu hướng tăng theo thời gian.
* Downward Trend:  Có xu hướng giảm theo thời gian.
* Horizontal Trend:  Ổn định hoặc biến đổi rất ít theo thời gian.
* Damped Trend:  Giảm dần theo thời gian, nhưng tốc độ giảm đi chậm dần.
* Non-linear Trend:  Biến đổi không tuân theo một mô hình tuyến tính đơn giản mà có các biến đổi phức tạp hơn.
* Seasonality(tính mùa vụ): Đề cập đến các biến động tăng hoặc giảm mà diễn ra đều đặn và lặp lại trong một khoảng thời gian.

Một số tính mùa vụ phổ biến :

* Holiday Seasonality: Thường được gây ra bởi các sự kiện đặc biệt như ngày lễ, sự kiện đặc biệt .
* Weekly Seasonality: Sự thay đổi lặp lại trong khoảng thời gian 7 ngày.
* Monthly Seasonality: Sự thay đổi lặp lại trong khoảng thời gian 30 hoặc 31 ngày.
* Annual Seasonality: Sự thay đổi lặp lại trong khoảng một năm.
* Cycles(tính chu kỳ): Đề cập đến những biến động lên xuống lặp lại, hoặc những thay đổi định kỳ, có thể kéo dài trong nhiều năm và chuyển từ giai đoạn này qua giai đoạn khác.

Sự khác biệt giữa tính mùa vụ và tính chu kỳ :

* Tính mùa vụ đề cập đến sự thay đổi lặp lại , xảy ra trong một khoảng thời gian cố định,bị ảnh hưởng nhiều yếu tố , có thể dự đoán được.
* Tính chu kỳ không bị giới hạn trong một khoảng thời gian cố định và có thể có diễn ra với tần suất khác nhau,nên việc dự đoán sẽ trở nên khó khăn hơn.
* Iregularity (sự bất thường): Đề cập đến những sự thay đổi bất thường của dữ liệu, xảy ra một cách ngẫu nhiên, có thể trái ngược hoàn toàn với các dữ liệu trong quá khứ, khó có thể giải thích và không dự đoán được trước.

Sự bất thường này có thể do sự sai sót trong đo lường dữ liệu, hoặc những sự kiện bất ngờ diễn ra và có thể vô tình làm ảnh hưởng đến tính chính xác khi đánh giá dữ liệu chuỗi thời gian.

# Phần II. Một số mô hình chuỗi thời gian.

## Mô hình Hidden Markov Models

HMM là quá trình Markov với các tham số không biết trước, nhiệm vụ xác định các tham số ẩn. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp.

Thành phần chính của HMM:

* + **Trạng thái**: HMM có một tập hợp các trạng thái ẩn, được ký hiệu là X ={ X1, X2, X3, ..., Xn},với n là số lượng trạng thái.Mỗi trạng thái có thể đại diện cho một chế độ hoặc một phần khác nhau của hệ thống.
  + **Biểu tượng đầu ra:** Được ký hiệu là O={ O1, O2, O3, ..., Om},với m là số lượng biểu tượng có thể quan sát được như âm thanh, văn bản hoặc hình ảnh.
  + **Xác suất chuyển tiếp:** Được biểu thị bằng ma trận A, trong đó A[i, j] là xác suất chuyển đổi từ trạng thái i sang trạng thái j.
  + **Xác suất đầu ra:** Xác suất phát thải biểu tượng đầu ra từ một trạng thái được biểu thị bằng ma trận B, trong đó B[i, k] là xác suất phát thải biểu tượng k từ trạng thái i.

Công thức tình xác suất chung: Xác suất của chuỗi quan sát  với chuỗi trạng thái ẩn 



Trong đó:  là tập hợp các tham số của Hmm

Xác suất này được tính:





Do đó, xác suất chung  là:



Xác suất của chuỗi quan sát O khi không biết các trạng thái ẩn S:



Hoạt động của HMM: Dự đoán chuỗi trạng thái ẩn nhất định đã tạo ra chuỗi quan sát được nhìn thấy, gồm hai nhiệm vụ chính:

* + Vấn đề đánh giá: Cho một HMM và một chuỗi quan sát, tính toán xác suất của chuỗi quan sát được tạo ra bởi HMM.
  + Vấn đề giải mã: Cho một HMM và một chuỗi quan sát, tìm chuỗi trạng thái ẩn có khả năng cao nhất đã tạo ra chuỗi quan sát.

Ứng dụng của HMM:

* + Nhận dạng tiếng nói: HMM được sử dụng để nhận dạng các từ hoặc cụm từ được nói trong lời nói.
  + Xử lý ngôn ngữ tự nhiên: HMM được sử dụng để phân loại văn bản, phân tích cú pháp và dịch máy.
  + Sinh học: HMM được sử dụng để dự đoán cấu trúc protein, phân tích trình tự DNA và phát hiện gen.
  + Robot: HMM được sử dụng để điều khiển robot và điều hướng trong môi trường.
  + Tài chính: HMM được sử dụng để mô hình hóa giá cổ phiếu và dự đoán xu hướng thị trường.

## Mô hình Factorial HMM

**Giới thiệu:**

Mô hình Factorial hidden Markov là một sự mở rộng của Mô hình Markov Ẩn (Hidden Markov Model - HMM) cho phép mô hình hóa nhiều chuỗi thời gian với sự tương tác của chúng.

HMM truyền thống mô hình hóa một chuỗi thời gian duy nhất với một biến trạng thái ẩn chi phối việc tạo ra biến quan sát tại mỗi bước thời gian. Ngược lại, FHMM mô hình hóa nhiều chuỗi thời gian với nhiều biến trạng thái ẩn tương ứng, những biến này tương tác để tạo ra các biến quan sát.

FHMM giả định rằng mỗi chuỗi thời gian quan sát đều có một trạng thái ẩn duy nhất, nhưng các trạng thái này độc lập có điều kiện khi cho một tập hợp các trạng thái nền tảng chung. Mỗi nhân tố trạng thái ẩn trong FHMM tạo ra một chuỗi thời gian riêng biệt, và các chuỗi này độc lập có điều kiện với nhau khi cho các trạng thái nền tảng chung.

Các trạng thái nền tảng chung trong FHMM hội tụ thông tin từ các nhân tố riêng lẻ để mô hình hóa sự phụ thuộc và tương tác giữa các chuỗi thời gian. Sự hội tụ này đạt được bằng cách tính toán xác suất của mỗi chuỗi thời gian cho từng trạng thái nền tảng và sau đó cân nhắc các xác suất đó để tính toán phân phối xác suất chung trên tập hợp các biến quan sát.

**a**

**Cấu trúc cơ bản của FHMM:**

FHMM mở rộng HMM bằng cách sử dụng nhiều chuỗi trạng thái ẩn. Giả sử ta có K chuỗi trạng thái ẩn và mỗi chuỗi có N trạng thái.

Ký hiệu các trạng thái ẩn là **Z**= và các quan sát là 

**Thành phần chính:**

1. **Trạng thái ẩn (Hidden States):**

 đại diện cho trạng thái của chuỗi ẩn thứ k tại thời điểm t

1. **Quan sát (Observations):**

 là quan sát tại thời điểm t

1. **Xác suất chuyển đổi (Transition Probabilities):**

 là ma trận chuyển đổi của chuỗi ẩn thứ k, trong đó 

1. **Xác suất quan sát (Emission probabilities):**

Xác suất quan sát được tính từ các trạng thái ẩn

1. **Xác suất ban đầu (Initial Probabilities)**

**** là xác suất ban đầu của chuỗi ẩn thứ k, trong đó 

**Công thức cụ thể:**

Xác suất của toàn bộ chuỗi quan sát Y được tính bằng cách tích hợp tất cả các xác suất của các trạng thái ẩn:



Trong đó:

* 
* 

**Thuật toán Kỳ vọng – Tối đa (Expectation-Maximization, EM)**

FHMM thường được huấn luyện bằng thuật toán EM, gồm 2 bước chính:

1. Expectation Step (E-step):

* Tính xác suất hậu nghiệm của các trạng thái ẩn dựa trên các quan sát hiện có.
* Sử dụng thuật toán Forward-Backward để tính toán xác suất 

1. Maximization Step (M-step):

* Cập nhật các tham số của mô hình để tối đa hóa hàm khả năng 
* Cập nhật các tham số:

1. Ma trận chuyển đổi: 
2. Xác suất ban đầu: 
3. Xác suất quan sát: Cập nhật dựa trên phân phối xác suất của các quan sát:

**Ưu điểm của mô hình Factorial hidden Markov so với các mô hình khác:**

* Tính linh hoạt của mô hình: FHMMs có thể mô hình hóa chuỗi thời gian với các mẫu phức tạp và các phụ thuộc, vì chúng có khả năng mô hình hóa chuỗi thời gian riêng lẻ cũng như các mối quan hệ giữa chúng. Điều này làm cho FHMMs phù hợp với nhiều loại dữ liệu chuỗi thời gian, bao gồm chuỗi thời gian đa biến, không đồng nhất và đa chiều.
* Ước lượng hiệu quả: FHMMs có thể được ước lượng hiệu quả bằng thuật toán Kỳ vọng-Tối đa hóa (Expectation-Maximization) có thể được thực hiện song song, làm cho nó phù hợp với các chuỗi thời gian lớn và có chiều cao.
* Khả năng diễn giải: FHMMs cho phép dễ dàng diễn giải và trực quan hóa các kết quả. Mô hình này phân tách chuỗi thời gian thành một tập hợp các thành phần nền tảng, mỗi thành phần có thể được mô hình hóa và phân tích riêng biệt. Điều này giúp nhận diện các xu hướng, mẫu và tương quan trong dữ liệu.
* Mô hình hóa sinh: FHMMs cho phép tạo ra dữ liệu chuỗi thời gian mới theo cấu trúc đã học, điều này có thể hữu ích cho việc dự đoán các xu hướng tương lai và lập các dự báo.
* Tính bền vững: FHMMs có khả năng chống chịu với dữ liệu thiếu và có thể xử lý các quan sát không đầy đủ. Khả năng này cho phép FHMMs được sử dụng cho các tác vụ như ước lượng, nơi các giá trị bị thiếu có thể được thay thế bằng các giá trị dự đoán dựa trên mô hình.

## Mô hình Mixture of LDS

MoLDS (Mixture of Linear Dynamical Systems) là một mô hình thống kê mạnh mẽ dùng để mô hình hóa các chuỗi dữ liệu phức tạp, đặc biệt là những chuỗi có cấu trúc thay đổi theo thời gian.

Mô hình này hoạt động bằng cách kết hợp nhiều hệ thống động lực tuyến tính (LDS) khác nhau, mỗi hệ thống đại diện cho một chế độ hoặc trạng thái tiềm ẩn của dữ liệu.

Model Mixture of LDS bao gồm các thành phần chính sau:

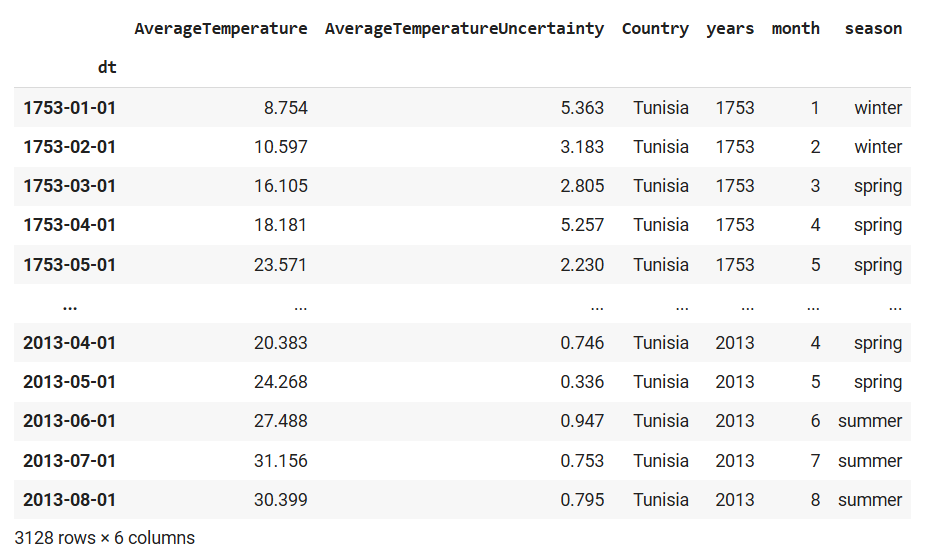
* + Nhiều hệ thống động lực tuyến tính (LDS): Mỗi LDS mô tả mối quan hệ tuyến tính giữa các biến trạng thái tiềm ẩn và các biến quan sát trong một chế độ hoặc trạng thái cụ thể.
  + Biến chuyển đổi: Biến chuyển đổi xác định xác suất chuyển đổi giữa các chế độ hoặc trạng thái khác nhau theo thời gian.
  + Phân phối hỗn hợp: Phân phối hỗn hợp xác định xác suất quan sát mỗi biến quan sát trong từng chế độ hoặc trạng thái.

Ứng dụng của MoLDS:

* Xử lý tín hiệu: Phát hiện và theo dõi các đối tượng trong video, nhận dạng giọng nói…
* Phân tích chuỗi thời gian: Dự báo giá chứng khoán, phát hiện bất thường trong dữ liệu cảm biến...
* Robot học: Điều khiển robot di chuyển trong môi trường phức tạp, học các nhiệm vụ mới…
* Nhận dạng ngôn ngữ: Phân biệt ngôn ngữ, dịch máy.

# Phần III. Xây dựng mô hình

## Mô tả dữ liệu



Bộ dữ liệu về nhiệt độ tại nước Tusinia được thu thập từ năm 1753 đến 2013.

Bộ dữ liệu của nhóm em gồm 7 cột:  
- dt: Cột dữ liệu về chuỗi thời gian

- AverageTemperature: Cột nhiệt độ trung bình

- AverageTemperatureUncertainty: Cột nhiệt độ trung bình không xác định

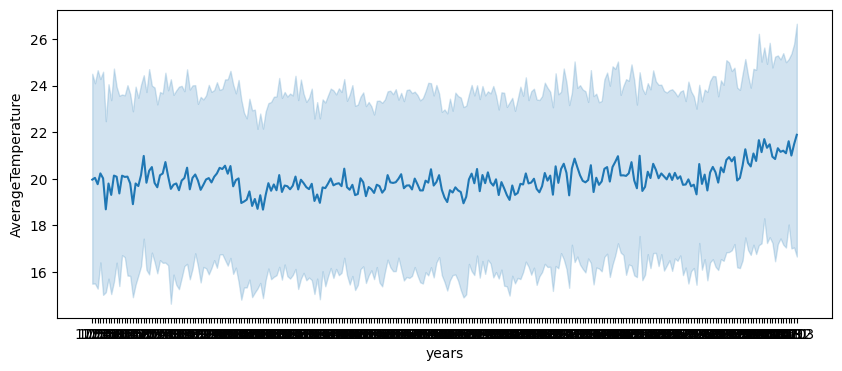
- Country: Quốc gia

- years: năm

- month: tháng

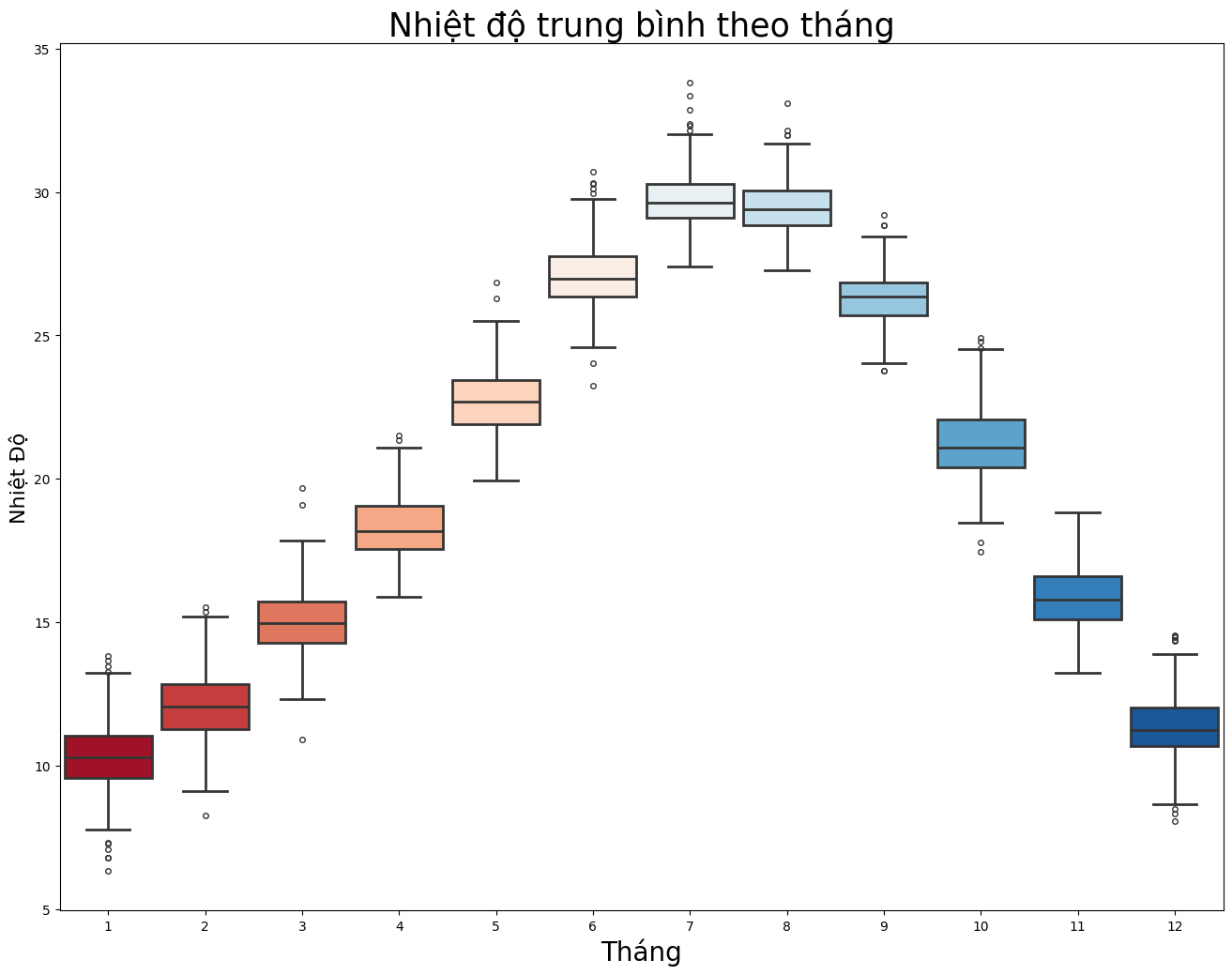
- season: mùa

1. **Phân tích dữ liệu**



Hình 3.1: Nhiệt độ trung bình theo các năm

Từ hình 3.1 ta có thể thấy rằng nhiệt độ qua các năm tại Tusinia có xu hướng tăng nhẹ vào khoảng thời gian trong các năm gần đây.



Hình 3.2: Tính mùa vụ trong dữ liệu về nhiệt độ

Từ hình 3.2 ta thấy rằng bộ dữ liệu nhiệt độ tại nước Tusinia các tháng trong năm có tính mùa vụ do gần như trong mọi năm, các tháng mùa đông đều có nhiệt độ thấp và các tháng mùa hè đều có nhiệt độ cao.

## HUẤN LUYỆN MÔ HÌNH

### 3.1. HMM

# Định nghĩa và huấn luyện HMM

model = hmm.GaussianHMM(n\_components=4, covariance\_type="diag", n\_iter=100)

model.fit(X\_scaled)

# Dự đoán các trạng thái ẩn

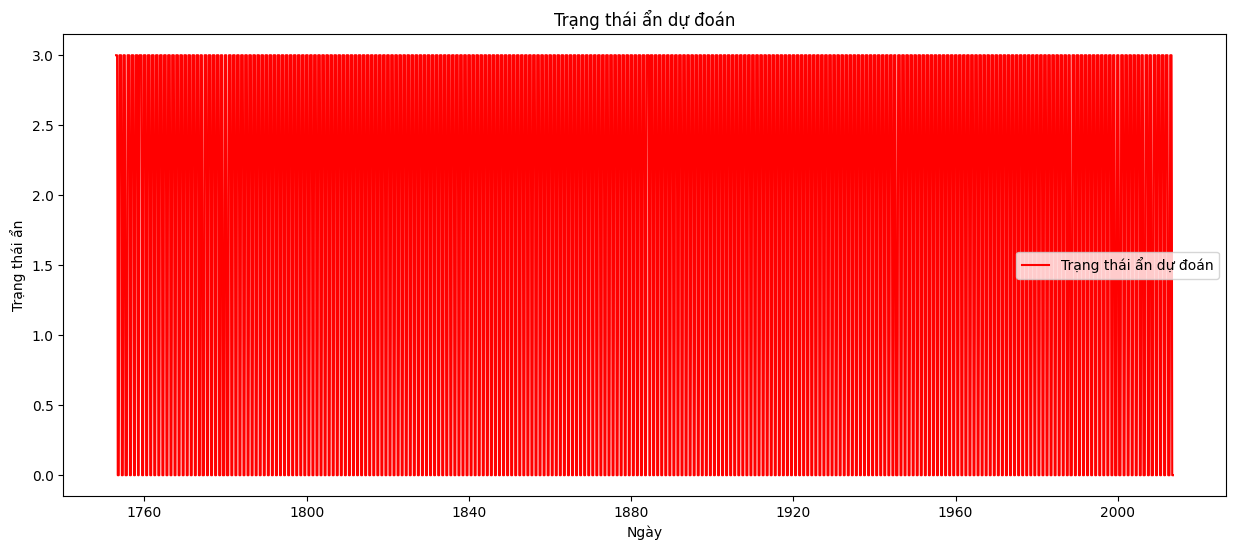
hidden\_states = model.predict(X\_scaled)

Các tham số được sử dụng:

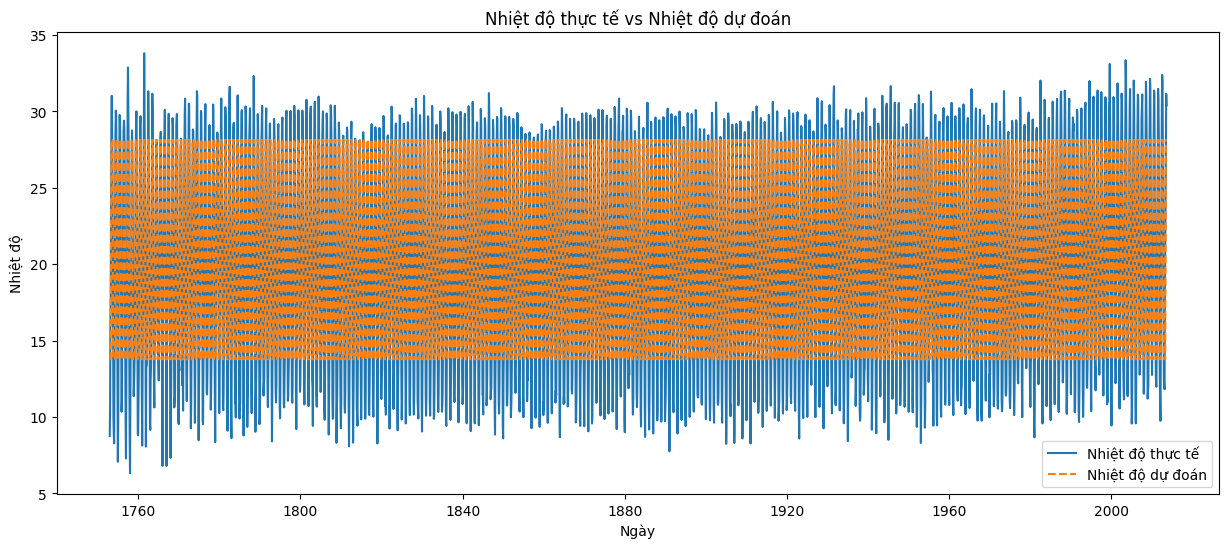
* n\_components=4: Số lượng trạng thái ẩn trong mô hình HMM. Nhóm chọn 4 trạng thái ẩn.
* covariance\_type="diag": Loại ma trận hiệp phương sai được sử dụng trong mô hình HMM. Các tùy chọn cho covariance\_type bao gồm:
* "full": Ma trận hiệp phương sai đầy đủ (không bị hạn chế).
* "tied": Một ma trận hiệp phương sai chung cho tất cả các thành phần.
* "diag": Chỉ có các phần tử trên đường chéo của ma trận hiệp phương sai (ma trận hiệp phương sai đường chéo).
* "spherical": Ma trận hiệp phương sai là một giá trị vô hướng (các biến số là độc lập và có cùng phương sai).
* n\_iter=100: Số lần lặp tối đa cho thuật toán EM (Expectation-Maximization) được sử dụng để huấn luyện mô hình HMM. Số lượng này xác định số lần lặp lại tối đa của thuật toán EM để hội tụ.

Mô hình sẽ được huấn luyện với dữ liệu đã được chuẩn hóa, sau đó mô hình sẽ tiến hành dự đoán các trạng thái ẩn.

Kết quả:



Hình 3.3: Trạng thái ẩn dự đoán của HMM



Hình 3.4: Biểu đồ giá trị thực tế với dự đoán

Từ biểu đồ ta có thể thấy rằng có một số sai lệch giữa hai đường. Mô hình chưa dự đoán được xu hướng của nhiệt độ và giá trị dự đoán đa phần nhỏ hơn so với giá trị nhiệt độ thực tế.

### 3.2 FHMM

|  |
| --- |
| # Số lượng trạng thái và biến quan sát  num\_factors = 2 # số lượng HMM con  num\_states\_per\_factor = 4 # số lượng trạng thái mỗi HMM con  num\_obs = 1 # chỉ sử dụng một biến quan sát  # Khởi tạo các HMM con  models = [hmm.GaussianHMM(n\_components=num\_states\_per\_factor, covariance\_type='diag', n\_iter=100) for \_ in range(num\_factors)]  # Huấn luyện từng HMM con với cùng một dữ liệu đã chuẩn hóa  for model in models:  model.fit(observations\_scaled, lengths)  print("All individual HMMs trained")  class FactorialHMM:  def \_\_init\_\_(self, models):  self.models = models  self.num\_factors = len(models)  self.num\_states = np.prod([model.n\_components for model in models])  self.num\_obs = models[0].means\_.shape[1]  def sample(self, n\_samples):  factor\_samples = [model.sample(n\_samples)[0] for model in self.models]  combined\_samples = np.sum(factor\_samples, axis=0)  factor\_states = [model.sample(n\_samples)[1] for model in self.models]  combined\_states = np.stack(factor\_states, axis=1)  return combined\_samples, combined\_states  def predict(self, observations):  factor\_states = [model.predict(observations) for model in self.models]  combined\_states = np.stack(factor\_states, axis=1)  return combined\_states  # Khởi tạo FHMM với các HMM con đã huấn luyện  fhmm = FactorialHMM(models)  # Dự đoán chuỗi trạng thái  state\_sequence = fhmm.predict(observations\_scaled)  # Tạo các quan sát mới dựa trên mô hình đã học  new\_obs\_scaled, new\_state\_sequence = fhmm.sample(len(observations\_scaled)) |

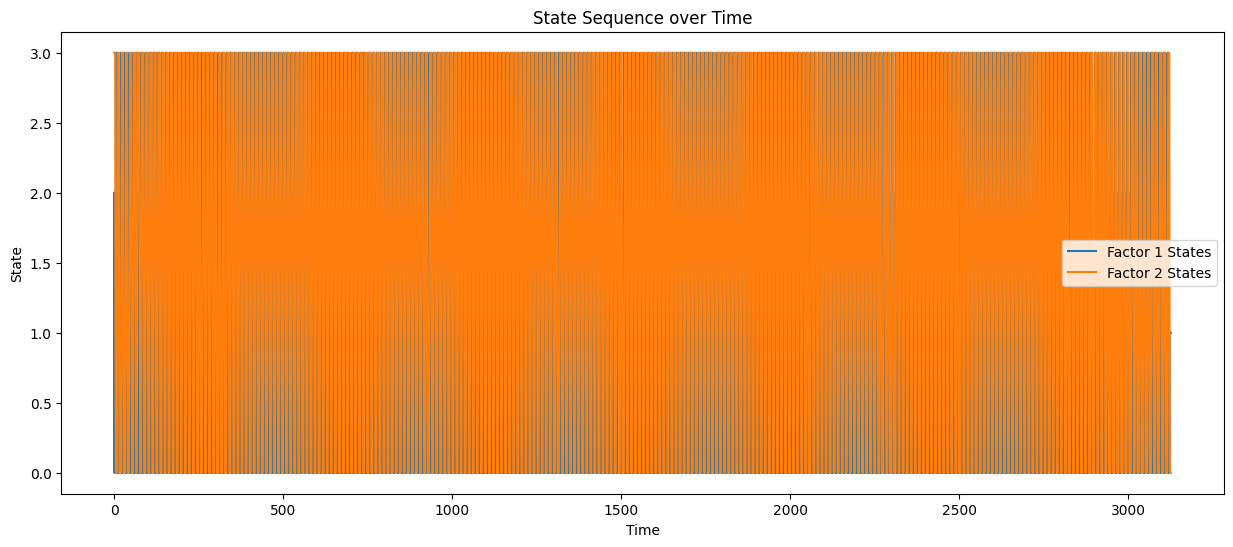
- num\_factors: Số lượng mô hình HMM con được sử dụng trong mô hình tổ hợp FHMM.

- num\_states\_per\_factor: Số lượng trạng thái cho mỗi HMM con.

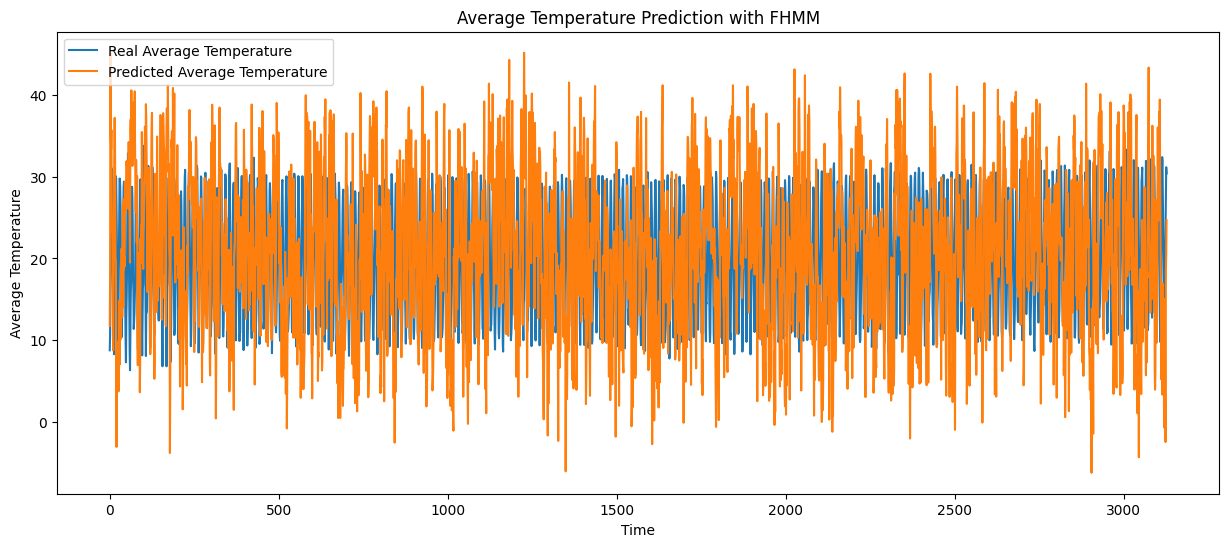
- num\_obs: Số lượng biến quan sát (chỉ sử dụng một biến quan sát là cột nhiệt độ).

Mô hình FHMM sẽ huấn luyện nhiều mô hình HMM con sau đó kết hợp chúng lại.

Kết quả:



Hình 3.5: Kết quả trạng thái dự đoán của FHMM



Hình 3.6: Giá trị dự đoán với giá trị thực tế của mô hình FHMM

Từ biểu đồ ta có thể thấy mô hình có phần dự đoán xu hướng có phần đa dạng hơn so với mô hình HMM. Tuy vậy, vẫn còn một sự chênh lệch nhất định giữa giá trị nhiệt độ thực tế với giá trị dự đoán, khi mà mô hình dự đoán ra các điểm nhiệt độ với độ lớn trên 40. Nhiệt độ như vậy chưa phù hợp với thực tế nước Tusinia.

### Mixture of LDS

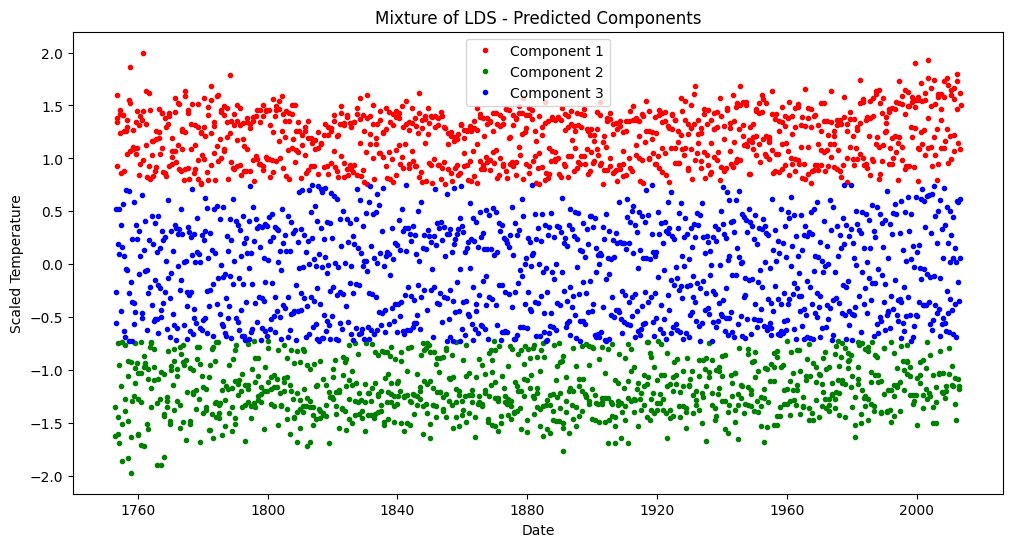
|  |
| --- |
| # Xác định số thành phần của mô hình Mixture of LDS  n\_components = 3  # Xây dựng và huấn luyện mô hình Mixture of LDS  model = BayesianGaussianMixture(n\_components=n\_components, random\_state=0)  model.fit(X\_scaled)  # Dự đoán nhãn của dữ liệu  labels = model.predict(X\_scaled) |

Tham số mô hình sử dụng:

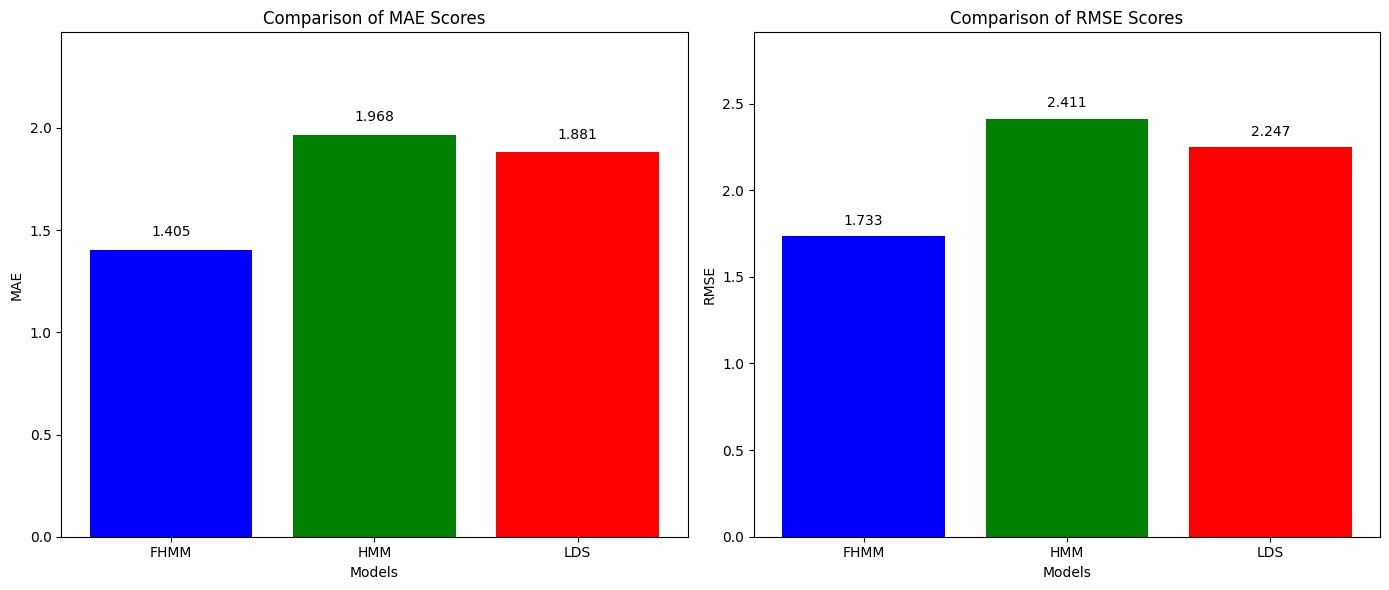
- n\_components = 3: Xác định số lượng thành phần của mô hình Mixture

Mô hình sẽ được huấn luyện với dữ liệu đã được chuẩn hóa, sau đó mô hình sẽ tiến hành dự đoán sự phân tách của các thành phần trên dữ liệu nhiệt độ theo thời gian.

Kết quả:



# Phần IV: Đánh giá mô hình



Từ bảng độ đo đánh giá trên, ta có thể thấy mô hình FHMM là mô hình có chỉ số độ đo tốt nhất.

# Phần V: Kết luận

Phần V: Kết luận

Trong bài tập lớn này, chúng em đã thực hiện phân tích dữ liệu nhiệt độ tại đất nước Tunisia bằng cách sử dụng một số các công cụ và kỹ thuật phân tích dữ liệu. Đây là một tóm tắt về những điều chính mà chúng tôi đã thực hiện và những kết quả đạt được:

Thu thập và khám phá dữ liệu: Đầu tiên, nhóm đã thu thập dữ liệu về nhiệt độ trung bình tại Tunisia. Sau đó khám phá dữ liệu để hiểu cấu trúc và các đặc điểm quan trọng của nó như các giá trị thiếu, phân bố của dữ liệu, và xu hướng thời gian.

Tiền xử lý dữ liệu: Chúng em đã thực hiện các bước tiền xử lý dữ liệu như loại bỏ giá trị thiếu, chuẩn hóa dữ liệu để chuẩn bị cho các bước phân tích và mô hình hóa sau này.

Phân tích thống kê đơn giản: Chúng em đã sử dụng các kỹ thuật thống kê đơn giản như biểu đồ dạng đường để phân tích xu hướng, mùa vụ và các biến động trong dữ liệu nhiệt độ.

Xây dựng mô hình dự đoán bằng HMM, FHMM: Để dự đoán nhiệt độ trong tương lai, chúng tôi đã sử dụng mô hình HMM, FHMM. Mô hình này đã được huấn luyện trên dữ liệu lịch sử và sau đó được sử dụng để dự đoán các trạng thái ẩn và nhiệt độ trong tương lai. Chúng tôi đã sử dụng các kỹ thuật như chuẩn hóa dữ liệu, chia dữ liệu thành tập huấn luyện và kiểm tra, và đánh giá hiệu suất của mô hình bằng các thước đo như Mean Squared Error (MSE).

Kết quả và nhận định: Mô hình FHMM đã đạt được kết quả chỉ số đánh giá tốt nhất trong việc dự đoán nhiệt độ tại Tunisia. Qua đó, chúng tôi có thể đưa ra những nhận định về xu hướng thời tiết, dự báo cho các hoạt động và các ứng dụng khác liên quan đến nhiệt độ.

Tóm lại, bài tập lớn này không chỉ giúp chúng tôi áp dụng và thực hành các kỹ năng phân tích dữ liệu mà còn cung cấp một cái nhìn sâu sắc hơn về biến động của nhiệt độ tại Tunisia và khả năng dự đoán thông qua các mô hình học máy tiên tiến như mô hình Markov ẩn. Đây là một bước đầu tiên trong việc áp dụng khoa học dữ liệu vào việc hiểu và dự báo các hiện tượng tự nhiên quan trọng.