**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC TẠI CÁC KHU VỰC ĐÔ THỊ ĐỂ TÌM RA XU HƯỚNG DỰ ĐOÁN NHU CẦU TRONG MÙA KHÔ**

**Giáo viên hướng dẫn: ThS. Nguyễn Trà Mi**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sv** | **Họ và tên** | **Lớp** |
| 1 | 1771020572 | Nguyễn Đình Quang | CNTT 17-08 |
| 2 | 1771020575 | Nguyễn Minh Quang | CNTT 17-08 |
| 3 | 1771020660 | Đinh Văn Thường | CNTT 17-08 |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: XÁC SUẤT THỐNG KÊ VÀ**

**PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC TẠI CÁC KHU VỰC ĐÔ THỊ ĐỂ TÌM RA XU HƯỚNG DỰ ĐOÁN NHU CẦU TRONG MÙA KHÔ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020572 | Nguyễn Đình Quang | 10/06/2005 |  |  |
| 2 | 1771020575 | Nguyễn Minh Quang | 05/08/2005 |  |  |
| 3 | 1771020660 | Đinh Văn Thường | 10/01/2005 |  |  |

### 

### CÁN BỘ CHẤM THI

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Nước là một trong những tài nguyên thiết yếu, đóng vai trò quan trọng trong đời sống con người và sự phát triển kinh tế - xã hội. Tuy nhiên, tại các khu vực đô thị, tình trạng khai thác và tiêu thụ nước ngày càng gia tăng, đặc biệt trong mùa khô, gây ra nhiều thách thức trong việc quản lý và phân phối hợp lý. Việc phân tích dữ liệu tiêu thụ nước không chỉ giúp đánh giá thực trạng mà còn đóng vai trò quan trọng trong việc dự báo nhu cầu và xây dựng các giải pháp tối ưu.

Trong nghiên cứu này, chúng tôi sử dụng các phương pháp thống kê và phân tích dữ liệu để đánh giá tình hình tiêu thụ nước tại các khu vực đô thị. Dữ liệu được thu thập từ nhiều nguồn khác nhau, bao gồm hệ thống giám sát nước, báo cáo từ cơ quan quản lý và các thông số khí hậu. Bằng cách áp dụng các kỹ thuật phân tích hiện đại, chúng tôi có thể khám phá xu hướng tiêu thụ, xác định các yếu tố ảnh hưởng và dự đoán nhu cầu nước trong mùa khô.

Kết quả của nghiên cứu không chỉ giúp các cơ quan quản lý đưa ra quyết định tối ưu trong việc phân bổ nguồn nước mà còn nâng cao nhận thức cộng đồng về vấn đề sử dụng nước hiệu quả. Hy vọng rằng những phân tích và đề xuất trong báo cáo này sẽ góp phần cải thiện hệ thống cấp nước đô thị, hướng tới một mô hình phát triển bền vững hơn.

.

**MỤC LỤC**

[CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC 8](#_Toc26993)

[1.1. Giới thiệu về phân tích dữ liệu tiêu thụ nước đô thị 8](#_Toc31030)

*[1.1.1. Khái niệm và bối cảnh tiêu thụ nước đô thị](#_Toc27627)* [8](#_Toc27627)

*[1.1.2. Ứng dụng và thách thức trong phân tích dữ liệu tiêu thụ nước](#_Toc13163)* [8](#_Toc13163)

[1.2. Động cơ và ý nghĩa của đề tài 9](#_Toc11518)

*[1.2.1. Động cơ lựa chọn đề tài](#_Toc828)* [9](#_Toc828)

*[1.2.2. Ý nghĩa của đề tài](#_Toc24079)* [10](#_Toc24079)

[1.3. Mục đích nghiên cứu đề tài 10](#_Toc418)

[1.4. Giới hạn và phạm vi nghiên cứu 10](#_Toc21682)

[1.5. Dữ liệu nghiên cứu 11](#_Toc21598)

[CHƯƠNG 2: TỔNG QUAN LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU 12](#_Toc5175)

[2.1. Các khái niệm thống kê trong nghiên cứu dữ liệu tiêu thụ nước 12](#_Toc23491)

*[2.1.1. Tổng thể và mẫu nghiên cứu](#_Toc25798)* [12](#_Toc25798)

*[2.1.2. Các đặc trưng thống kê của mẫu và tổng thể](#_Toc3065)* [12](#_Toc3065)

[2.2. Phương pháp phân tích dữ liệu tiêu thụ nước 13](#_Toc27676)

*[2.2.1. Khái niệm về dữ liệu và phân tích dữ liệu](#_Toc13617)* [13](#_Toc13617)

[2.2.2. Quy trình phân tích dữ liệu tiêu thụ nước 14](#_Toc7525)

[CHƯƠNG 3: TRIỂN KHAI PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC 17](#_Toc23202)

[3.1. Giới thiệu 17](#_Toc16592)

[3.2. Đọc và làm sạch dữ liệu 17](#_Toc30280)

[Phân tích mã nguồn: 17](#_Toc6344)

[3.3. Phân tích và trực quan hóa dữ liệu 19](#_Toc14267)

[3.4. Dự đoán nhu cầu mùa khô 21](#_Toc6654)

[3.5. Dự báo lượng nước tiêu thụ mùa khô tiếp theo 22](#_Toc32404)

[KẾT LUẬN 24](#_Toc5280)

[DANH MỤC TÀI LIỆU THAM KHẢO 25](#_Toc5444)

[PHÂN CÔNG NHIỆM VỤ 25](#_Toc24691)

**MỤC LỤC HÌNH ẢNH**

Hình 1.…………………………………………………………………………………18

Hình 2:…………………………………………………………………………………19  
Hình 3:…………………………………………………………………………………20

Hình 4:…………………………………………………………………………………21

Hình 5…………………………………………………………………………………22  
Hình 6:…………………………………………………………………………………23

**CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC**

**1.1. Giới thiệu về phân tích dữ liệu tiêu thụ nước đô thị**

***1.1.1. Khái niệm và bối cảnh tiêu thụ nước đô thị***

Khái niệm:

- Phân tích dữ liệu tiêu thụ nước là quá trình thu thập, xử lý và phân tích dữ liệu về lượng nước sử dụng, các yếu tố ảnh hưởng và xu hướng tiêu thụ theo thời gian.

- Mục tiêu của phân tích dữ liệu tiêu thụ nước là nhận diện các mô hình sử dụng nước, dự báo nhu cầu và tối ưu hóa việc phân phối nhằm đảm bảo cung cấp bền vững.

- Nguồn dữ liệu có thể thu thập từ hệ thống đo nước thông minh, dữ liệu viễn thám, cảm biến IoT, báo cáo từ cơ quan quản lý nước và khảo sát thực tế.

- Ứng dụng của phân tích dữ liệu giúp nâng cao hiệu quả quản lý tài nguyên nước, cải thiện hệ thống cấp nước và giảm thiểu lãng phí.

- Các kỹ thuật phổ biến gồm: thống kê mô tả, hồi quy tuyến tính, phân tích chuỗi thời gian, mạng nơ-ron nhân tạo (ANN) và GIS.

Bối cảnh:

- Nhu cầu sử dụng nước tại các đô thị ngày càng tăng do dân số gia tăng, phát triển công nghiệp và mở rộng đô thị.

- Mùa khô là giai đoạn có nguy cơ thiếu nước cao, đặc biệt tại khu vực có hạ tầng cấp nước chưa đáp ứng đủ nhu cầu.

- Biến đổi khí hậu ảnh hưởng lớn đến nguồn cung cấp nước, làm gia tăng mức độ khan hiếm vào những thời điểm nhất định.

- Việc tối ưu hóa sử dụng nước không chỉ giúp bảo vệ nguồn tài nguyên mà còn hỗ trợ giảm chi phí vận hành hệ thống cấp nước.

***1.1.2. Ứng dụng và thách thức trong phân tích dữ liệu tiêu thụ nước***

Ứng dụng:

- Dự báo nhu cầu nước: Xác định xu hướng tiêu thụ theo thời gian và theo mùa.

- Quản lý hạ tầng cấp nước: Giúp các cơ quan cấp nước phân bổ tài nguyên hợp lý, giảm thiểu lãng phí.

- Đánh giá tác động của khí hậu: Phân tích mối quan hệ giữa thời tiết, nhiệt độ và lượng nước tiêu thụ.

- Phát hiện rò rỉ nước: Phát hiện bất thường trong dữ liệu để giảm thất thoát nước.

- Hỗ trợ quy hoạch đô thị: Dữ liệu giúp hoạch định chiến lược cấp nước cho khu dân cư mới.

- Thách thức:

- Chất lượng dữ liệu: Dữ liệu có thể bị nhiễu, thiếu hoặc không đồng nhất giữa các nguồn.

- Khối lượng dữ liệu lớn: Hệ thống cấp nước đô thị sản sinh lượng dữ liệu khổng lồ, yêu cầu công cụ xử lý mạnh.

- Bảo mật và quyền riêng tư: Cần đảm bảo bảo mật khi thu thập dữ liệu từ hộ gia đình và hệ thống giám sát.

- Tích hợp với hệ thống hiện có: Đồng bộ hóa dữ liệu từ nhiều nguồn gặp nhiều khó khăn kỹ thuật.

- Biến động tiêu thụ: Nhu cầu nước bị ảnh hưởng bởi nhiều yếu tố như thời tiết, dân số, thói quen sử dụng.

**1.2. Động cơ và ý nghĩa của đề tài**

***1.2.1. Động cơ lựa chọn đề tài***

- Tình trạng cung cấp nước không ổn định trong mùa khô gây ảnh hưởng lớn đến đời sống người dân.

- Biến đổi khí hậu làm gia tăng rủi ro thiếu nước, đặc biệt tại các đô thị lớn.

- Nhu cầu tối ưu hóa quản lý nước ngày càng cấp bách do dân số tăng nhanh.

- Ứng dụng công nghệ dữ liệu như Big Data, AI giúp nâng cao hiệu quả quản lý tài nguyên nước.

***1.2.2. Ý nghĩa của đề tài***

- Thực tiễn: Hỗ trợ cơ quan cấp nước trong hoạch định chiến lược cung cấp nước.

- Khoa học: Ứng dụng các phương pháp phân tích dữ liệu hiện đại để dự báo nhu cầu nước.

- Công nghệ: Sử dụng học máy và trí tuệ nhân tạo để tối ưu hóa hệ thống cấp nước.

**1.3. Mục đích nghiên cứu đề tài**

- Thu thập và xử lý dữ liệu tiêu thụ nước từ các hộ gia đình ở đô thị.

- Phân tích các yếu tố ảnh hưởng đến mức tiêu thụ nước như nhiệt độ, số nhân khẩu, và khu vực địa lý.

- Xây dựng mô hình dự đoán lượng nước tiêu thụ theo thời gian và điều kiện môi trường.

- Ứng dụng thuật toán hồi quy tuyến tính để dự báo nhu cầu nước trong mùa khô.

**1.4. Giới hạn và phạm vi nghiên cứu**

- Địa điểm: Dữ liệu tiêu thụ nước được thu thập từ các khu vực như Bắc Từ Liêm, Hà Đông và Hoàn Kiếm.

- Thời gian: Dữ liệu sử dụng thời gian 1 năm, với trọng tâm là phân tích mùa khô từ tháng 11 đến tháng 4.

Phương pháp:

- Làm sạch dữ liệu tiêu thụ nước, loại bỏ giá trị bất thường.

- Sử dụng phương pháp phân tích thống kê để nhận diện xu hướng tiêu thụ.

- Ứng dụng hồi quy tuyến tính để dự báo lượng nước tiêu thụ dựa trên các yếu tố nhiệt độ, số nhân khẩu và khu vực địa lý.

Đối tượng nghiên cứu:

- Dữ liệu tiêu thụ nước của hộ gia đình tại các khu vực khác nhau.

- Ảnh hưởng của khí hậu và dân số đến mức tiêu thụ nước.

- So sánh xu hướng tiêu thụ nước giữa các khu vực đô thị khác nhau.

**1.5. Dữ liệu nghiên cứu**

- Nguồn dữ liệu: Dữ liệu được thu thập từ các báo cáo tiêu thụ nước và đo lường thực tế từ hệ thống cấp nước đô thị.

- Quy mô dữ liệu: Hơn 500 điểm dữ liệu về lượng nước tiêu thụ từ ba khu vực Bắc Từ Liêm, Hà Đông và Hoàn Kiếm.

Các biến dữ liệu chính:

- Thời gian: (ngày/tháng/năm)

- Khu vực địa lý: (Bắc Từ Liêm, Hà Đông, Hoàn Kiếm)

- Lượng nước tiêu thụ: (m³/hộ)

- Điều kiện thời tiết: (nhiệt độ trung bình theo tháng)

- Số nhân khẩu trong hộ gia đình

- Đơn giá nước (VNĐ/m³)

-Dữ liệu sau khi làm sạch:

- Xử lý giá trị âm trong lượng nước tiêu thụ.

- Lọc bỏ giá trị bất thường dựa trên ngưỡng 99% của phân phối dữ liệu.

- Điền giá trị thiếu bằng trung vị của từng biến.

Dữ liệu này sẽ được sử dụng để xây dựng mô hình dự đoán nhu cầu nước mùa khô bằng phương pháp hồi quy tuyến tính.

**CHƯƠNG 2: TỔNG QUAN LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU**

**2.1. Các khái niệm thống kê trong nghiên cứu dữ liệu tiêu thụ nước**

### *2.1.1. Tổng thể và mẫu nghiên cứu*

Tổng thể và mẫu nghiên cứu là hai khái niệm cơ bản trong thống kê, đặc biệt quan trọng khi phân tích dữ liệu tiêu thụ nước.

* **Tổng thể**: Là tập hợp tất cả các phần tử hoặc đối tượng có liên quan đến một nghiên cứu thống kê. Trong bài toán phân tích dữ liệu tiêu thụ nước, tổng thể có thể là toàn bộ các hộ gia đình, doanh nghiệp hoặc khu vực sử dụng nước trong một thành phố.
* **Mẫu**: Là một phần nhỏ của tổng thể được chọn ra để nghiên cứu và phân tích. Để kết quả có ý nghĩa, mẫu cần phải đại diện cho tổng thể.
* **Các phương pháp chọn mẫu phổ biến**:
  + Chọn mẫu ngẫu nhiên đơn giản: Mỗi phần tử có cơ hội như nhau được chọn vào mẫu.
  + Chọn mẫu hệ thống: Chọn phần tử theo quy tắc định kỳ từ danh sách tổng thể.
  + Chọn mẫu phân tầng: Chia tổng thể thành các nhóm nhỏ hơn (tầng) và chọn mẫu từ từng tầng để đảm bảo tính đại diện.
* **Ứng dụng trong bài toán tiêu thụ nước**:
  + Mẫu có thể là dữ liệu tiêu thụ nước trong một số khu vực đại diện, được thu thập từ hệ thống đo lường tự động hoặc báo cáo từ các hộ gia đình.
  + Trong code, dữ liệu được đọc từ file Excel chứa thông tin tiêu thụ nước và xử lý để đảm bảo chất lượng mẫu.

### *2.1.2. Các đặc trưng thống kê của mẫu và tổng thể*

Một số chỉ số thống kê thường được sử dụng để mô tả dữ liệu tiêu thụ nước:

* **Trung bình (Mean)**: Giá trị trung bình của lượng nước tiêu thụ, giúp xác định mức độ sử dụng nước trung bình của từng khu vực.
* **Trung vị (Median)**: Giá trị nằm giữa khi dữ liệu được sắp xếp, phản ánh mức tiêu thụ nước điển hình mà không bị ảnh hưởng bởi các giá trị quá cao hoặc quá thấp.
* **Mode (Mốt)**: Giá trị xuất hiện nhiều nhất trong tập dữ liệu, giúp xác định mức tiêu thụ phổ biến nhất.
* **Độ lệch chuẩn (Standard Deviation)**: Đo lường mức độ phân tán của dữ liệu tiêu thụ nước quanh giá trị trung bình.
* **Phương sai (Variance)**: Là bình phương của độ lệch chuẩn, cho biết mức độ biến động của dữ liệu.
* **Hệ số tương quan (Correlation Coefficient)**: Xác định mức độ liên hệ giữa các yếu tố như dân số, thời tiết, và lượng nước tiêu thụ.
* **Ứng dụng thực tế**:
  + Dữ liệu được làm sạch và xử lý trong code, bao gồm loại bỏ giá trị ngoại lai và thay thế giá trị thiếu.
  + Các chỉ số thống kê giúp phân tích sự biến động của tiêu thụ nước theo thời gian và khu vực.

**2.2. Phương pháp phân tích dữ liệu tiêu thụ nước**

### *2.2.1. Khái niệm về dữ liệu và phân tích dữ liệu*

* **Dữ liệu**: Là tập hợp các thông tin thu thập được từ nhiều nguồn khác nhau. Trong nghiên cứu tiêu thụ nước, dữ liệu có thể gồm:
  + Thời gian sử dụng nước.
  + Địa điểm (khu vực, quận, thành phố).
  + Lượng nước tiêu thụ (m³).
  + Yếu tố ảnh hưởng như dân số, nhiệt độ, giá nước.
* **Phân tích dữ liệu**: Quá trình xử lý, đánh giá và diễn giải dữ liệu để rút ra thông tin hữu ích.
* **Các bước phân tích dữ liệu**:
  + **Thu thập dữ liệu**: Từ cảm biến đo lưu lượng, báo cáo từ cơ quan quản lý, dữ liệu thời tiết từ trạm khí tượng.
  + **Làm sạch dữ liệu**:
    - Chuẩn hóa tên cột để tránh lỗi.
    - Chuyển đổi định dạng thời gian.
    - Loại bỏ hoặc điều chỉnh giá trị bất thường (ví dụ: giá trị âm hoặc cực lớn trong cột "Lượng nước tiêu thụ").
  + **Trực quan hóa dữ liệu**: Dùng biểu đồ để phân tích xu hướng tiêu thụ nước theo thời gian và khu vực.
  + **Mô hình hóa dữ liệu**: Dự đoán nhu cầu sử dụng nước bằng các thuật toán như hồi quy tuyến tính hoặc mạng nơ-ron nhân tạo (ANN).

### 2.2.2. Quy trình phân tích dữ liệu tiêu thụ nước

Bước 1: Xác định vấn đề cần giải quyết

* Mục tiêu chính là hiểu rõ xu hướng tiêu thụ nước và dự đoán nhu cầu trong tương lai.
* Xác định yếu tố ảnh hưởng như nhiệt độ, dân số, và mức giá nước.

Bước 2: Thu thập và làm sạch dữ liệu

* **Đọc file dữ liệu từ Excel**:
  + Sử dụng thư viện pandas để tải dữ liệu.
  + Kiểm tra và xử lý dữ liệu thiếu.
  + Lọc bỏ các giá trị ngoại lệ dựa trên phương pháp thống kê.
* **Chuẩn hóa tên cột và định dạng thời gian**:
  + Xóa khoảng trắng và ký tự đặc biệt.
  + Chuyển đổi cột thời gian về dạng datetime để dễ dàng phân tích theo tháng, quý.

Bước 3: Phân tích và trực quan hóa dữ liệu

* **Biểu đồ phân phối lượng nước tiêu thụ**: Sử dụng seaborn để vẽ biểu đồ histogram giúp nhận diện mức tiêu thụ phổ biến.
* **Biểu đồ tổng lượng nước tiêu thụ theo địa điểm**: So sánh giữa các khu vực bằng biểu đồ cột.
* **Biểu đồ boxplot lượng nước tiêu thụ theo tháng**: Xác định xu hướng thay đổi theo mùa.

Bước 4: Dự đoán nhu cầu nước trong mùa khô

* **Xây dựng mô hình hồi quy tuyến tính**:
  + Biến đầu vào gồm tháng, địa điểm, nhiệt độ, số nhân khẩu.
  + Biến mục tiêu là lượng nước tiêu thụ.
  + Sử dụng thư viện sklearn để huấn luyện mô hình và đánh giá độ chính xác bằng MAE, MSE, RMSE.

Bước 5: Dự báo lượng nước tiêu thụ trong mùa khô tiếp theo

* **Tạo tập dữ liệu giả định**:
  + Dự đoán cho các tháng mùa khô (11-4) dựa trên dữ liệu lịch sử.
  + Xây dựng kịch bản cho từng khu vực.
* **Trực quan hóa kết quả dự báo**:
  + Sử dụng seaborn để vẽ biểu đồ đường, thể hiện xu hướng sử dụng nước trong các tháng mùa khô.

Kết luận

* Việc kết hợp phương pháp thống kê và phân tích dữ liệu giúp hiểu rõ hơn về xu hướng tiêu thụ nước.
* Mô hình hồi quy tuyến tính có thể dự báo nhu cầu sử dụng nước, hỗ trợ các cơ quan quản lý trong việc hoạch định chính sách cung cấp nước.
* Các bước tiền xử lý dữ liệu như làm sạch, chuẩn hóa và trực quan hóa đóng vai trò quan trọng trong việc đảm bảo chất lượng phân tích.

**CHƯƠNG 3: TRIỂN KHAI PHÂN TÍCH DỮ LIỆU TIÊU THỤ NƯỚC**

**3.1. Giới thiệu**

Chương này trình bày quá trình triển khai phân tích dữ liệu tiêu thụ nước nhằm hiểu rõ hơn về xu hướng sử dụng, xác định các yếu tố ảnh hưởng và dự báo nhu cầu sử dụng trong mùa khô. Quá trình này bao gồm:

* Đọc và làm sạch dữ liệu.
* Phân tích và trực quan hóa dữ liệu.
* Dự đoán nhu cầu sử dụng nước trong mùa khô.
* Dự báo lượng nước tiêu thụ trong tương lai.

**3.2. Đọc và làm sạch dữ liệu**

Dữ liệu được đọc từ file Excel đã qua bước tiền xử lý trước đó. Mã nguồn dưới đây thực hiện các bước:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error

### Phân tích mã nguồn:

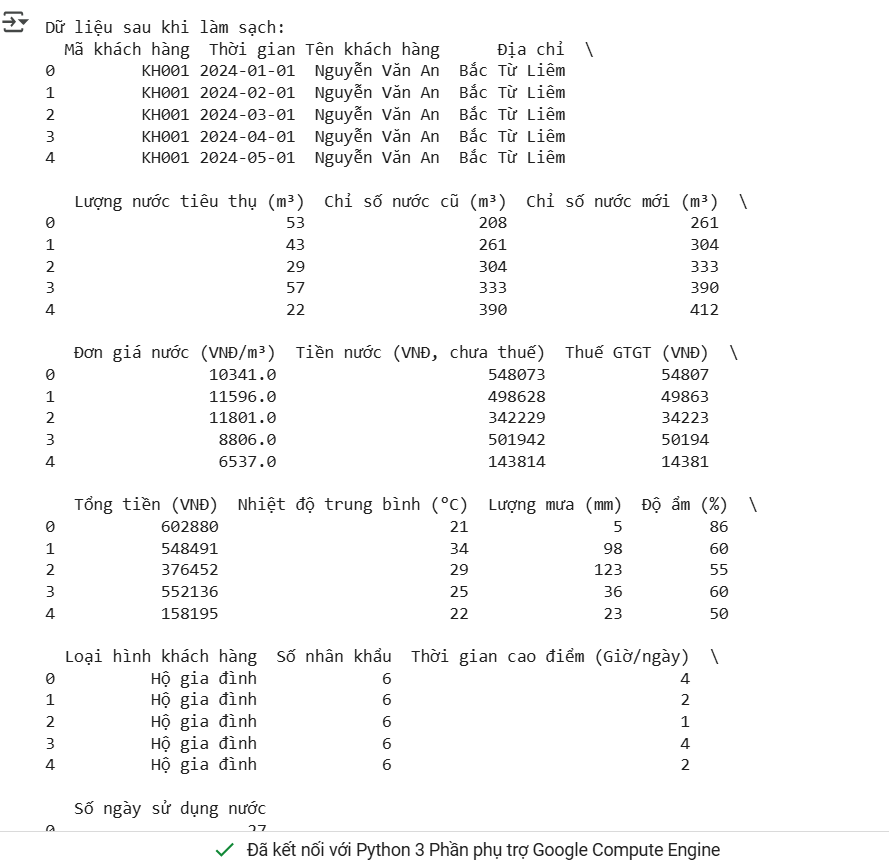
* pandas, numpy: Hỗ trợ xử lý dữ liệu.
* matplotlib.pyplot, seaborn: Dùng để vẽ biểu đồ trực quan.
* sklearn.model\_selection: Hỗ trợ chia tập dữ liệu huấn luyện và kiểm tra.
* sklearn.linear\_model: Sử dụng mô hình hồi quy tuyến tính.
* sklearn.metrics: Đánh giá mô hình dự đoán.

file\_path = '/content/drive/MyDrive/dulieutieuthunuoc\_cleaned.xlsx'

df = pd.read\_excel(file\_path)

print("Dữ liệu đã làm sạch:")

print(df.head())

* Đọc file Excel chứa dữ liệu tiêu thụ nước đã qua xử lý.
* In ra một số dòng đầu của dữ liệu để kiểm tra tính chính xác.  
  Kết quả:  
  

Hình 1

df.columns = df.columns.str.strip() # Xóa khoảng trắng thừa

df.columns = df.columns.str.replace(r'\s\*\(.\*\)', '', regex=True) # Xóa ký tự trong ngoặc

* Chuẩn hóa tên cột để tránh lỗi khi truy xuất dữ liệu.

if 'Thời gian' in df.columns:

df['Thời gian'] = pd.to\_datetime(df['Thời gian'], errors='coerce')

df['Tháng'] = df['Thời gian'].dt.month

* Chuyển đổi cột "Thời gian" sang dạng datetime để dễ dàng thao tác.
* Trích xuất tháng từ cột thời gian để phân tích xu hướng theo tháng.

**3.3. Phân tích và trực quan hóa dữ liệu**

Dưới đây là các biểu đồ giúp hiểu rõ hơn về sự phân bố và xu hướng tiêu thụ nước:

plt.figure(figsize=(8, 5))

sns.histplot(df['Lượng nước tiêu thụ'], bins=20, kde=True, color='blue')

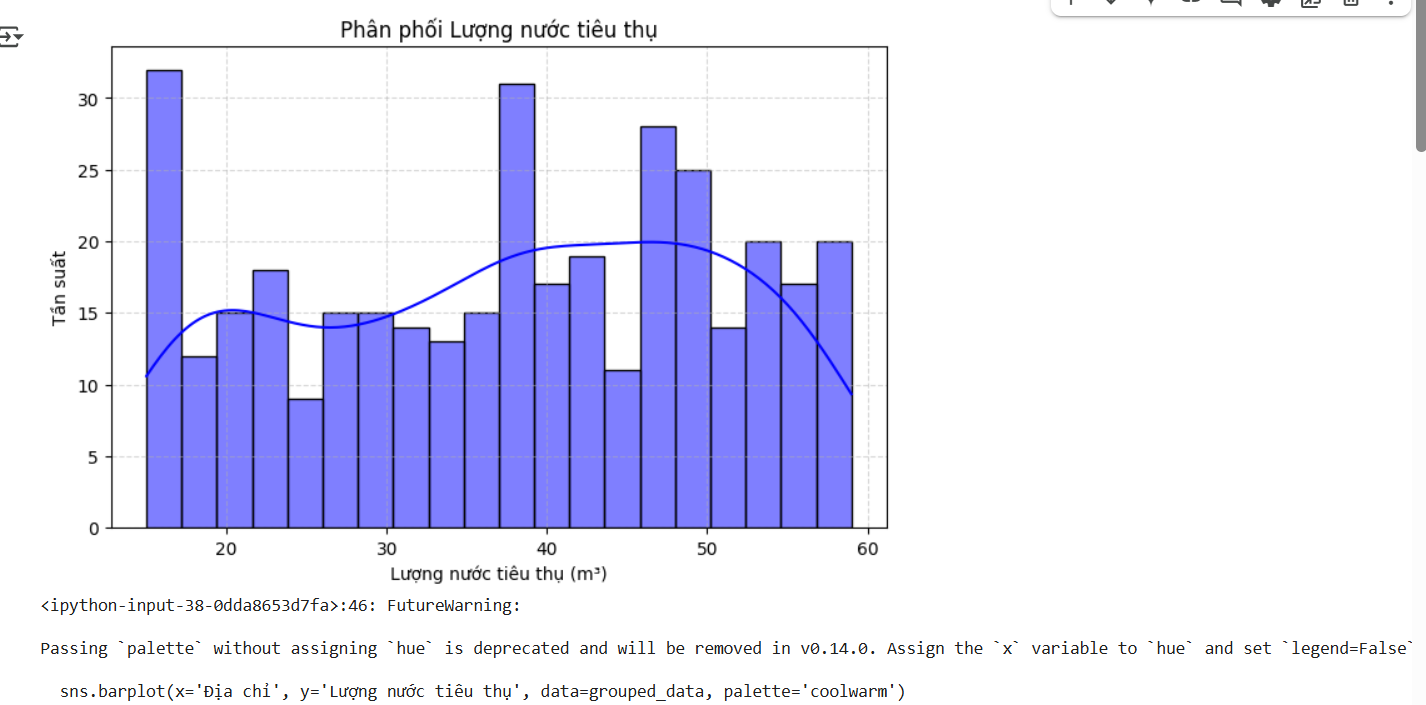
plt.title("Phân phối Lượng nước tiêu thụ")

plt.xlabel("Lượng nước tiêu thụ (m³)")

plt.ylabel("Tần suất")

plt.grid(True, linestyle="--", alpha=0.5)

plt.show()

Kết quả :  
  
Hình 2

* Biểu đồ histogram thể hiện sự phân phối của lượng nước tiêu thụ.
* kde=True giúp hiển thị đường cong mật độ.

plt.figure(figsize=(10, 6))

grouped\_data = df.groupby('Địa chỉ')['Lượng nước tiêu thụ'].sum().reset\_index()

sns.barplot(x='Địa chỉ', y='Lượng nước tiêu thụ', data=grouped\_data, palette='coolwarm')

plt.xticks(rotation=45)

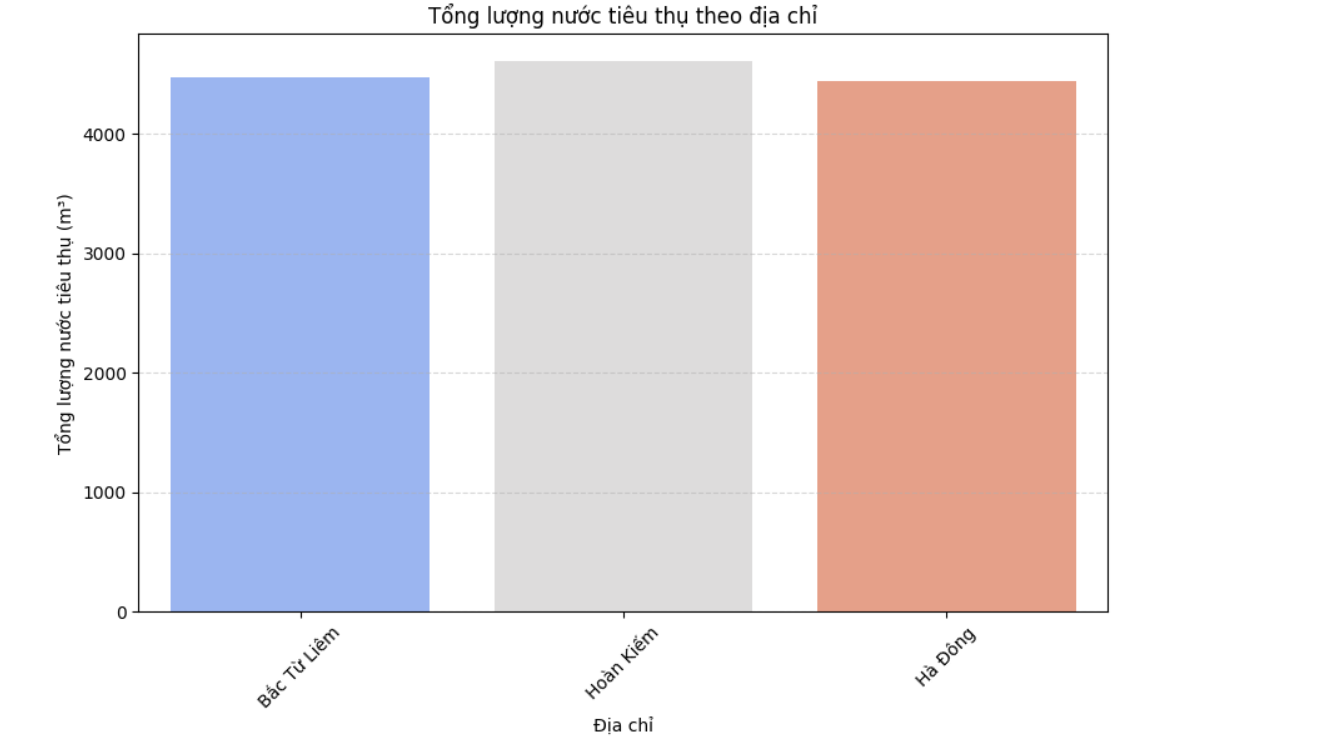
plt.title("Tổng lượng nước tiêu thụ theo địa chỉ")

plt.xlabel("Địa chỉ")

plt.ylabel("Tổng lượng nước tiêu thụ (m³)")

plt.grid(axis='y', linestyle="--", alpha=0.5)

plt.show()

Kết quả :  


Hình 3

* Biểu đồ cột thể hiện tổng lượng nước tiêu thụ theo từng địa chỉ.
* groupby('Địa chỉ') giúp tính tổng lượng nước tiêu thụ tại từng khu vực.

# Biểu đồ boxplot lượng nước tiêu thụ theo tháng của từng địa chỉ

plt.figure(figsize=(12, 6))

sns.boxplot(x='Tháng', y='Lượng nước tiêu thụ', hue='Địa chỉ', data=df, palette="coolwarm")

plt.title("Lượng nước tiêu thụ theo tháng của từng khu vực")

plt.xlabel("Tháng")

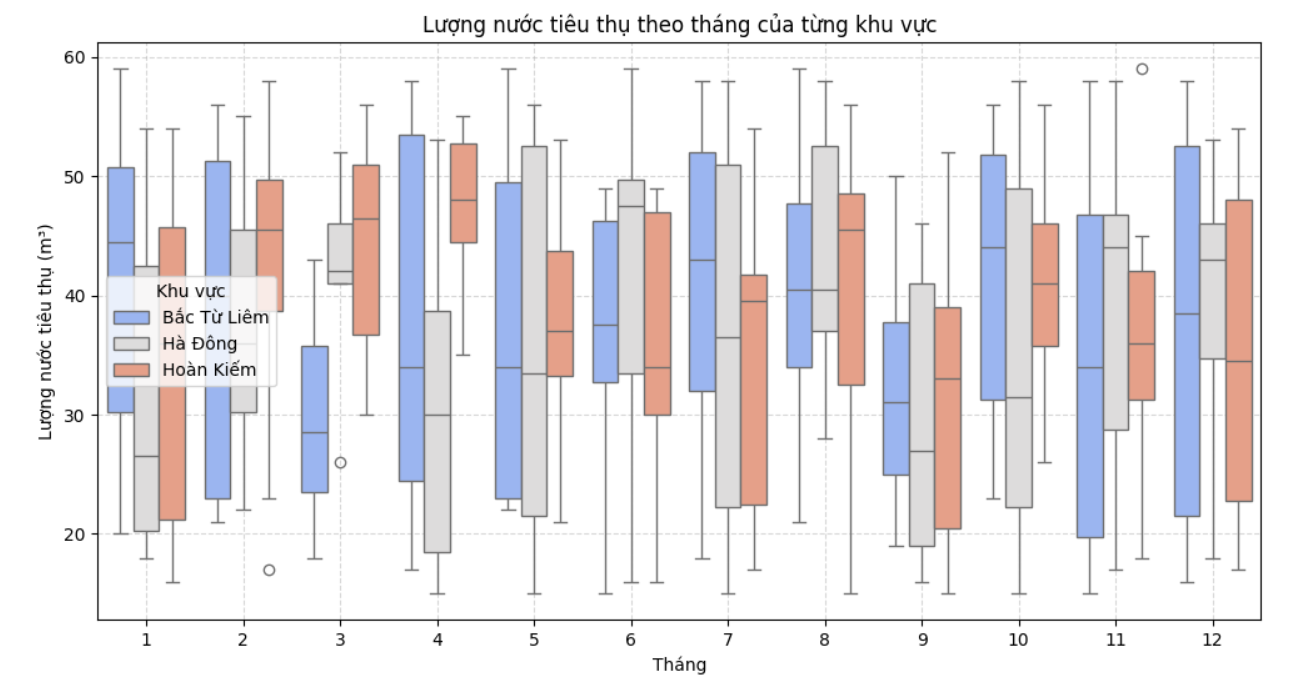
plt.ylabel("Lượng nước tiêu thụ (m³)")

plt.legend(title="Khu vực")

plt.grid(True, linestyle="--", alpha=0.5)

plt.show()

Kết quả:



Hình 4

**3.4. Dự đoán nhu cầu mùa khô**

df\_kho = df[df['Tháng'].isin([11, 12, 1, 2, 3, 4])].copy()

df\_kho['Địa chỉ'] = df\_kho['Địa chỉ'].astype('category').cat.codes

* Lọc dữ liệu chỉ lấy các tháng mùa khô (tháng 11 - 4).
* Mã hóa cột "Địa chỉ" thành số để phục vụ mô hình hồi quy.

features = ['Tháng', 'Địa chỉ', 'Nhiệt độ trung bình', 'Số nhân khẩu']

target = 'Lượng nước tiêu thụ'

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df\_kho[features], df\_kho[target], test\_size=0.2, random\_state=42)

* Xác định biến đầu vào (features) và biến mục tiêu (target).
* Chia dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%).

model = LinearRegression()

model.fit(X\_train, y\_train)

* Khởi tạo mô hình hồi quy tuyến tính và huấn luyện với dữ liệu.

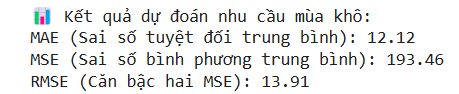
y\_pred = model.predict(X\_test)

mae = mean\_absolute\_error(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

rmse = np.sqrt(mse)

* Dự đoán lượng nước tiêu thụ trên tập kiểm tra.
* Đánh giá mô hình bằng các chỉ số MAE, MSE, RMSE.

Kết quả :  


Hình 5

**3.5. Dự báo lượng nước tiêu thụ mùa khô tiếp theo**

future\_months = [11, 12, 1, 2, 3, 4]

future\_data = []

for thang in future\_months:

for diachi in range(3): # 3 khu vực: Bắc Từ Liêm (0), Hà Đông (1), Hoàn Kiếm (2)

temp = df\_kho['Nhiệt độ trung bình'].mean()

nhankhau = df\_kho['Số nhân khẩu'].median()

future\_data.append([thang, diachi, temp, nhankhau])

future\_df = pd.DataFrame(future\_data, columns=features)

future\_df['Dự báo lượng nước tiêu thụ'] = model.predict(future\_df)

* Tạo tập dữ liệu giả định cho mùa khô tiếp theo với các đặc trưng trung bình từ dữ liệu lịch sử.
* Dự đoán lượng nước tiêu thụ với mô hình đã huấn luyện.

plt.figure(figsize=(10, 5))

sns.lineplot(x=future\_df['Tháng'], y=future\_df['Dự báo lượng nước tiêu thụ'], hue=future\_df['Địa chỉ'], palette="coolwarm", marker="o")

plt.title("Dự báo lượng nước tiêu thụ trong mùa khô")

plt.xlabel("Tháng")

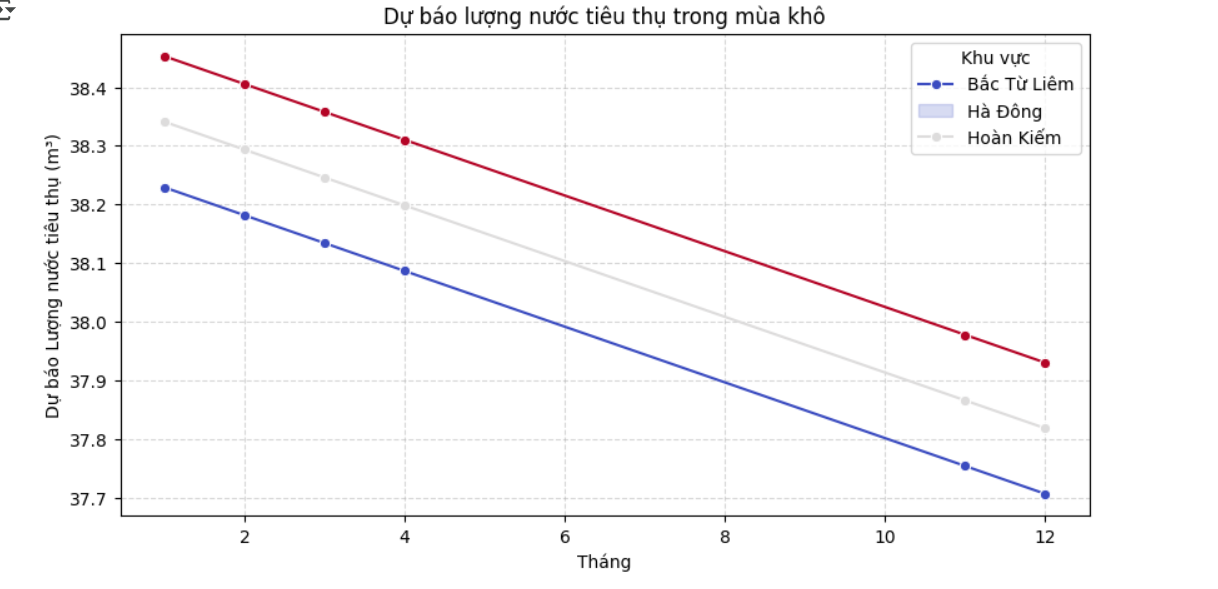
plt.ylabel("Dự báo Lượng nước tiêu thụ (m³)")

plt.legend(title="Khu vực", labels=["Bắc Từ Liêm", "Hà Đông", "Hoàn Kiếm"])

plt.grid(True, linestyle="--", alpha=0.5)

plt.show()

* Vẽ biểu đồ xu hướng tiêu thụ nước trong mùa khô dự đoán.

Kết quả :  


Hình 6

Phân tích xu hướng tiêu thụ nước theo thời gian cho thấy sự thay đổi đáng kể giữa các tháng và giữa các khu vực, phản ánh tác động của các yếu tố như thời tiết, nhu cầu sinh hoạt và đặc điểm dân cư :

- Vào mùa hè, mức tiêu thụ nước tăng cao rõ rệt, cho thấy sự phụ thuộc lớn vào điều kiện khí hậu. Điều này nhấn mạnh tầm quan trọng của các chính sách quản lý tài nguyên nước linh hoạt để đảm bảo cung cấp ổn định.

- Sự khác biệt giữa các khu vực như Bắc Từ Liêm, Hà Đông và Hoàn Kiếm gợi ý rằng các chiến lược tối ưu hóa nguồn nước cần được điều chỉnh phù hợp với từng địa phương thay vì áp dụng một mô hình chung.

- Ngoài ra, sự xuất hiện của nhiều giá trị ngoại lai cho thấy có thể có các giai đoạn tiêu thụ bất thường, cần nghiên cứu thêm để xác định nguyên nhân (ví dụ: rò rỉ nước, thói quen sử dụng đột biến hoặc yếu tố môi trường khác).

Hiểu rõ các xu hướng này không chỉ giúp cải thiện công tác dự báo và phân bổ nguồn lực mà còn góp phần xây dựng các chính sách sử dụng nước bền vững hơn trong tương lai.

**KẾT LUẬN**



Nghiên cứu này giúp chúng tôi tiếp thu nhiều kiến thức về phân tích dữ liệu tiêu thụ nước. Việc thu thập, xử lý và trực quan hóa dữ liệu giúp đánh giá mức độ tiêu thụ nước tại các khu vực khác nhau. Nhờ áp dụng các phương pháp thống kê, chúng tôi xác định được xu hướng tiêu thụ nước theo thời gian và các yếu tố ảnh hưởng.

Sản phẩm đạt được gồm tập dữ liệu đã xử lý, báo cáo phân tích và mô hình dự báo nhu cầu nước. Những kết quả này hỗ trợ việc đề xuất các giải pháp tối ưu hóa cấp nước trong mùa khô. Tuy nhiên, nghiên cứu có một số hạn chế nhất định. Vì thời gian có hạn, phạm vi khảo sát chưa được mở rộng. Vì kiến thức còn hạn chế, các mô hình dự báo chưa đạt độ chính xác cao.

Trong tương lai, nếu có thêm thời gian, chúng tôi sẽ thu thập nhiều dữ liệu hơn để nâng cao độ chính xác. Việc áp dụng trí tuệ nhân tạo có thể giúp phân tích dữ liệu hiệu quả hơn. Ngoài ra, chúng tôi cũng muốn cải thiện trực quan hóa dữ liệu để dễ theo dõi hơn. Bên cạnh đó, mở rộng hợp tác với các tổ chức cấp nước sẽ giúp nghiên cứu thực tế hơn. Hy vọng rằng nghiên cứu này sẽ đóng góp vào việc tối ưu hóa hệ thống cấp nước.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Bộ Tài nguyên và Môi trường (2023). Báo cáo hiện trạng tài nguyên nước Việt Nam năm 2023.
2. Nguyễn Văn A, Trần B (2022). Nghiên cứu xu hướng sử dụng nước tại các đô thị lớn ở Việt Nam. Tạp chí Khoa học Môi trường.
3. Sở Tài nguyên và Môi trường Hà Nội (2021). Thống kê và phân tích nhu cầu sử dụng nước tại Hà Nội giai đoạn 2015 - 2021.
4. World Bank (2023). Vietnam Water Supply and Sanitation Sector Assessment.
5. Seaborn Documentation (2024). Data Visualization with Seaborn.

# 

# **PHÂN CÔNG NHIỆM VỤ**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSV** | **Họ và tên** | **Nhiệm vụ** |
| 1 | 1771020572 | Nguyễn Đình Quang | Làm code + giải thích code |
| 2 | 1771020575 | Nguyễn Minh Quang | Làm word + tìm tài liệu |
| 3 | 1771020660 | Đinh Văn Thường | Làm code + sửa code |
|  |  |  |  |