Report1\_Dự kiến Dataset

**Báo cáo buổi 1:**

1. **Chọn chủ đề và viết đề cương nộp các thầy cô duyệt .**

* Ứng dụng các mô hình học máy cho bài toán dự báo lũ lụt

1. **Tìm kiếm dataset bão lũ Việt Nam**

Link dataset dự kiến:

<https://github.com/Mandolaro/flood-data/blob/main/Datasets.zip>

Lưu dataset về để trên Drive: [Datasets](https://drive.google.com/drive/folders/1VjvvC3wvKg6F3N-bSzrinhtKX78wiltO?usp=drive_link)

**<thầy hỏi dataset kiếm ở đâu thì em bảo em tìm các bài báo về Machine Learning dự đoán bão lũ. Tìm được bài báo này: -** [**https://www.mdpi.com/2071-1050/14/19/11861**](https://www.mdpi.com/2071-1050/14/19/11861) **, trong bài báo này có giới thiệu bộ dataset>**

Mô tả Dataset:

….

# **Dataset Descriptions**

### **1. Flood damages in Vietnam 1990 - 2022.xlsx**

* **Nội dung**: Ghi nhận thông tin về thiệt hại do lũ lụt tại Việt Nam từ năm 1990 đến 2022. Bao gồm các thông tin về số người bị ảnh hưởng, thiệt hại tài sản và các dữ liệu liên quan khác.
* **Định dạng**: Tệp Excel.

### **2. Land use.csv**

* **Nội dung**: Thống kê về diện tích sử dụng đất ở Việt Nam, bao gồm các loại đất như đất nông nghiệp, đất lâm nghiệp, đất ở và đất đặc biệt. Số liệu được ghi nhận theo tỉnh/thành phố tính đến cuối năm 2018.
* **Định dạng**: CSV.

### **3. Monthly humidity 15 provinces.csv**

* **Nội dung**: Dữ liệu về độ ẩm trung bình hàng tháng tại 15 tỉnh của Việt Nam, với số liệu phân chia theo năm và các tháng.
* **Định dạng**: CSV.

### **4. Monthly rainfall in 15 provinces.csv**

* **Nội dung**: Thống kê lượng mưa hàng tháng tại 15 tỉnh của Việt Nam, được chia theo năm và tháng.
* **Định dạng**: CSV.

### **5. Monthly temperature.csv**

* **Nội dung**: Nhiệt độ không khí trung bình hàng tháng tại 15 tỉnh của Việt Nam, có dữ liệu cho từng tháng trong các năm.
* **Định dạng**: CSV.

### **6. More damages.xls**

* **Nội dung**: Chứa các thông tin bổ sung về thiệt hại do thiên tai. Tệp này có thể gặp lỗi định dạng và cần cài thêm thư viện xlrd để mở.
* **Định dạng**: Tệp Excel.

### **7. Number of people divided into rural and urban region.csv**

* **Nội dung**: Thống kê dân số ở khu vực nông thôn và thành thị của các tỉnh Việt Nam từ năm 1995 đến 2020. Dữ liệu bao gồm cả tổng dân số, dân số thành thị và dân số nông thôn.
* **Định dạng**: CSV.

### **8. Population density.csv**

* **Nội dung**: Mật độ dân số của các tỉnh, thành phố tại Việt Nam từ năm 2011 đến 2020. Tệp này cung cấp số liệu diện tích, dân số trung bình và mật độ dân số theo năm.
* **Định dạng**: CSV.

### **9. Precipitation Monthly 1901 - 2020.csv**

* **Nội dung**: Dữ liệu về lượng mưa trung bình hàng tháng từ năm 1901 đến 2020. Tệp này gặp lỗi định dạng và có thể cần điều chỉnh để đọc được chính xác.
* **Định dạng**: CSV.

### **10. Water flow in major river in Vietnam.csv**

* **Nội dung**: Dữ liệu về lưu lượng nước tại các con sông lớn ở Việt Nam từ năm 2010 đến 2020, bao gồm các giá trị cao nhất và thấp nhất theo từng năm.
* **Định dạng**: CSV.

### **11. Water velocity in river.csv**

* **Nội dung**: Thống kê về vận tốc nước tại các con sông lớn ở Việt Nam. Dữ liệu bao gồm vận tốc lớn nhất và nhỏ nhất qua các năm từ 2002 đến 2020.
* **Định dạng**: CSV.

—-----------

**Dự kiến tuần sau: Tìm hiểu về Tiền xử lý data**

* **Các file đang ở định dạng khác nhau, cần tiền xử lý đưa về 1 file chung, lọc bỏ các cột không cần thiết, …**

### **1. Khám phá dữ liệu ban đầu**

* Xem qua từng tệp dữ liệu để hiểu rõ cấu trúc và nội dung.
* Kiểm tra các giá trị bị thiếu (missing values) và xem xét cách xử lý: loại bỏ, thay thế hoặc dự đoán.
* Xác định các cột không cần thiết hoặc không liên quan đến mục tiêu phân tích và có thể loại bỏ.

### **2. Xử lý dữ liệu thiếu**

* **Loại bỏ hàng hoặc cột**: Nếu cột/hàng có quá nhiều dữ liệu bị thiếu, cân nhắc loại bỏ.
* **Điền giá trị thiếu**: Điền giá trị trung bình, trung vị hoặc giá trị phổ biến nhất, hoặc có thể sử dụng giá trị dự đoán từ các mô hình nếu dữ liệu có tính dự đoán cao.

### **3. Chuyển đổi dữ liệu**

* **Chuyển đổi loại dữ liệu**: Đảm bảo tất cả các cột có loại dữ liệu phù hợp. Ví dụ, các cột ngày tháng cần được chuyển về định dạng datetime, các cột số cần được chuyển về dạng numeric.
* **Mã hóa dữ liệu phân loại**: Nếu có các biến phân loại, bạn có thể mã hóa chúng thành các giá trị số (sử dụng One-Hot Encoding hoặc Label Encoding).

**…**

Report2,3\_Thay đổi Dataset

**Chuyển sang bộ dataset đơn giản hơn về Flood**

* [Flood Prediction Dataset](https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset/data) - Kaggle

**+, Data trước 11 files, nhiều thông số, cần nhiều công sức trong tiến trình Tiền xử lý dữ liệu, …**

**+, Hỏi thầy thử xem chuyển sang bộ Dataset này được không (không phải Dataset của Việt Nam) cơ mà đầy đủ các thông số cần thiết, dataset với nhiều Features (thuộc tính) cho bài Machine Learning.**

============================================

*```*

*Bộ dataset trước của Việt Nam gồm 11 files .csv, em thử tiền xử lý Data mà thấy khó qua.*

*Em muốn nhắn hỏi xin thầy để đổi sang bộ Dataset này được không ạ.*

*Link Dataset Kaggle: https://www.kaggle.com/datasets/naiyakhalid/flood-prediction-dataset/data*

*www.kaggle.com*

*- Không phải là dataset của Việt Nam*

*- tuy nhiên có nhiều thông số (features, thuộc tính) - cụ thể là 20-21 featuresv. Đủ số lượng nhiều để làm bài Machine Learning dự đoán bão lũ*

*Thầy xem thử giúp em với ạ*

*```*

Dữ liệu "flood.csv" này chứa các thông tin quan trọng cho việc dự đoán lũ lụt, với **50,000 dòng** và **21 cột**. Đây là một tập dữ liệu về phát hiện và dự đoán lũ lụt, được thiết kế để hỗ trợ các cơ quan quản lý và nghiên cứu, giúp cảnh báo và ứng phó kịp thời với các tình huống lũ lụt. Mỗi cột đại diện cho một yếu tố môi trường hoặc kinh tế-xã hội ảnh hưởng đến nguy cơ lũ lụt, và tất cả các cột đều là kiểu số nguyên (int64), giúp phù hợp để áp dụng các mô hình hồi quy mà không cần phải tiền xử lý nhiều.

### **Các Thông Tin Cơ Bản về Bộ Dữ Liệu**

* **Số lượng dòng:** 50,000
* **Số lượng cột:** 21
* **Không có giá trị thiếu** trong các cột, điều này giúp tiết kiệm thời gian làm sạch dữ liệu.
* **Loại biến:** Tất cả các biến đều là dạng số (int64), không có biến phân loại.

**Các Biến Quan Trọng trong Dữ Liệu**

Mỗi biến trong bộ dữ liệu này cung cấp thông tin về các yếu tố môi trường và nhân tố con người có khả năng ảnh hưởng đến nguy cơ và mức độ nghiêm trọng của lũ lụt. Các biến chính gồm:

1. **MonsoonIntensity:** Cường độ mùa mưa; lượng mưa lớn làm tăng nguy cơ lũ lụt.
2. **TopographyDrainage:** Khả năng thoát nước của địa hình; địa hình tốt giúp thoát nước hiệu quả, giảm nguy cơ lũ.
3. **RiverManagement:** Chất lượng và hiệu quả của các biện pháp quản lý sông, như nạo vét và duy tu bờ sông, giúp ngăn ngừa lũ.
4. **Deforestation:** Tình trạng phá rừng; giảm khả năng hấp thụ nước của đất, tăng dòng chảy bề mặt và nguy cơ lũ.
5. **Urbanization:** Mức độ đô thị hóa; các bề mặt không thấm nước làm giảm khả năng thấm nước, tăng nguy cơ lũ.
6. **ClimateChange:** Ảnh hưởng của biến đổi khí hậu; dẫn đến các hình thái mưa bất thường và có thể gây lũ lụt.
7. **DamsQuality:** Chất lượng và bảo trì đập nước; các đập bị hỏng có thể gây lũ lớn.
8. **Siltation:** Tình trạng bồi lắng sông ngòi; tích tụ phù sa giảm khả năng thoát nước, tăng nguy cơ lũ.
9. **AgriculturalPractices:** Các thực hành nông nghiệp không bền vững có thể gây ra sự suy thoái đất, làm tăng nguy cơ lũ.
10. **Encroachments:** Mức độ lấn chiếm vùng lũ và các lối thoát nước tự nhiên, gây cản trở dòng chảy.
11. **IneffectiveDisasterPreparedness:** Thiếu kế hoạch và hệ thống cảnh báo khẩn cấp làm tăng thiệt hại khi có lũ.
12. **DrainageSystems:** Hệ thống thoát nước tốt giúp giảm nguy cơ ngập lụt.
13. **CoastalVulnerability:** Khu vực ven biển dễ bị ảnh hưởng bởi triều cường và nước biển dâng.
14. **Landslides:** Độ dốc cao và đất không ổn định dễ dẫn đến sạt lở.
15. **Watersheds:** Ảnh hưởng của lưu vực nước có thể làm tăng hoặc giảm nguy cơ lũ.
16. **DeterioratingInfrastructure:** Cơ sở hạ tầng kém làm tăng nguy cơ ngập lụt.
17. **PopulationScore:** Khu vực đông dân cư chịu thiệt hại nhiều hơn.
18. **WetlandLoss:** Mất đi các vùng đầm lầy giảm khả năng hấp thụ nước tự nhiên, làm tăng nguy cơ lũ.
19. **InadequatePlanning:** Quy hoạch đô thị kém làm tăng tính dễ tổn thương trước lũ lụt.
20. **PoliticalFactors:** Các yếu tố chính trị như tham nhũng và thiếu ý chí đầu tư vào cơ sở hạ tầng thoát nước.
21. **FloodProbability:** Xác suất xảy ra lũ lụt trong khu vực. Đây là **biến mục tiêu** của bài toán dự báo.

### **Biến Mục Tiêu**

* **FloodProbability:** Là biến mục tiêu cần dự đoán, đại diện cho xác suất xảy ra lũ trong khu vực.

Report2,3\_Model Dự Kiến

**LINEAR REGRESSION MODEL - KNN MODEL**

**Dự kiến triển khai thuật toán:**

* Sử dụng hai thuật toán học máy cơ bản là **Linear Regression** và **K-Nearest Neighbors (KNN)**.
* Mục tiêu: Dự đoán xác suất xảy ra lũ lụt - **FloodProbability** dựa trên các yếu tố:
  + **MonsoonIntensity:** Cường độ mưa
  + **TopographyDrainage:** Khả năng thoát nước
  + **RiverManagement:** Quản lý sông
  + **Deforestation:** Phá rừng
  + **Urbanization:** Đô thị hóa
  + **ClimateChange:** Biến đổi khí hậu
  + **DamsQuality:** Chất lượng đập
  + **Siltation:** Bồi lắng
  + **DrainageSystems:** Hệ thống thoát nước
  + **…**

### **1. Thuật toán Linear Regression (Hồi quy tuyến tính)**

* **Định nghĩa đơn giản:**
  + Linear Regression là phương pháp dự đoán dựa trên mối quan hệ tuyến tính (mối quan hệ dạng đường thẳng) giữa các yếu tố. Nếu có hai yếu tố mà khi một yếu tố thay đổi thì yếu tố kia cũng thay đổi, Linear Regression sẽ cố gắng vẽ một đường thẳng đi qua dữ liệu đó, từ đó giúp dự đoán giá trị tương lai.
* **Ví dụ đơn giản:**
  + Giả sử các bạn học sinh để ý thấy rằng càng nhiều thời gian học tập, điểm số có xu hướng càng cao. Linear Regression sẽ cố gắng vẽ một đường thẳng qua các điểm số này để thấy rằng nếu chúng ta dành nhiều thời gian học hơn, điểm số dự đoán sẽ tăng lên.
* **Ứng dụng vào dự đoán xác suất lũ lụt:**
  + Khi áp dụng vào bài toán dự đoán xác suất lũ, Linear Regression sẽ phân tích mối quan hệ giữa các yếu tố như lượng mưa, khả năng thoát nước, và quản lý sông ngòi. Giả sử, nếu lượng mưa tăng, đường thẳng dự đoán cho thấy khả năng lũ lụt cũng sẽ tăng lên. Khi đó, dự đoán của chúng ta có thể giúp cảnh báo trước về nguy cơ lũ.

### **2. Thuật toán K-Nearest Neighbors (KNN) (K láng giềng gần nhất)**

* **Định nghĩa đơn giản:**
  + KNN là thuật toán dựa vào các ví dụ "gần nhất" trong dữ liệu đã có. Khi cần dự đoán cho một trường hợp mới, KNN sẽ tìm k trường hợp trước đó có các yếu tố tương tự và xem kết quả là gì. Dựa trên kết quả này, KNN đưa ra dự đoán cho trường hợp mới.
* **Ví dụ đơn giản:**
  + Giả sử một học sinh muốn dự đoán điểm số của mình trong lần thi sắp tới. Em ấy có thể tìm điểm số của những lần thi gần nhất mà điều kiện học tập của em ấy tương tự (thời gian học, số lần ôn tập, vv). Dựa vào các điểm số trước đó, em ấy có thể ước lượng điểm thi sắp tới của mình.
* **Ứng dụng vào dự đoán xác suất lũ lụt:**
  + Trong bài toán dự đoán lũ, KNN sẽ xem xét những ngày trong quá khứ có điều kiện thời tiết, cường độ mưa, và các yếu tố tương tự. Ví dụ, nếu có một ngày nào đó có lượng mưa và độ ẩm giống với những ngày trước đó đã từng xảy ra lũ, KNN sẽ dự đoán rằng ngày mới này cũng có nguy cơ cao xảy ra lũ lụt.

### **Tóm lại:**

* **Linear Regression** sẽ dự đoán xác suất lũ dựa trên mối quan hệ tuyến tính giữa các yếu tố.
* **KNN** sẽ dự đoán xác suất lũ bằng cách tìm các ngày có điều kiện gần giống với ngày cần dự đoán.

**==============**

*```*

*1. Hỏi thầy xem Dataset*

*2. Là đọc kỹ report về 2 thuật toán*

*3. Slide thì có 3 phần chính đã up lên*

*- Mục tiêu chọn đè tài*

*- Dataset (có up lên slide rồi, cơ mà đợi thầy duyệt dataset )*

*- Model: (đã up cơ bản nhất - như report tuần 3) --- về sau sang tuần 4 sẽ update thêm sau.*

*```*