AI Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

| ***Dự đoán chất lượng nước sử dụng kĩ thuật học sâu (GRU + LSTM)*** |
| --- |

**05/08/2025**

Thang

<Đỗ Quyết Thắng SIC251879>

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Data Acquisition

2.2. Training Methodology

2.3. Workflow

2.4. System Diagram

3. Results

3.1. Data Preprocessing

3.2. Exploratory Data Analysis (EDA)

3.3. Modeling

3.4. User Interface

3.5. Testing and Improvements

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

**1.1. Background Information**

**Tên dự án:** Dự đoán chất lượng nước sử dụng học sâu

**Lĩnh vực:** Trí tuệ nhân tạo (AI), Học sâu (Deep Learning), Xử lý dữ liệu

**Công nghệ sử dụng:** Python, TensorFlow/Keras, Pandas, Matplotlib, Flask, Google Colab

**1.2. Motivation and Objective**

**Động lực:** Vấn đề ô nhiễm nguồn nước ngày càng nghiêm trọng, đe dọa đến sức khỏe con người. Việc ứng dụng trí tuệ nhân tạo để tự động phát hiện nước có thể uống được là cần thiết và cấp thiết.

**Mục tiêu:** Xây dựng mô hình học sâu dựa trên dữ liệu các chỉ số hóa học và vật lý của nước để phân loại nước có thể uống được hay không. Đồng thời, cung cấp giao diện giúp người dùng dễ dàng tương tác và dự đoán.

**1.3. Members and Role Assignments**

Đỗ Quyết Thắng

**1.4. Schedule and Milestones**

| **Thời gian** | **Công việc** |
| --- | --- |
| Tuần 1-2 | Thu thập và phân tích dữ liệu |
| Tuần 3 | Tiền xử lý và khám phá dữ liệu |
| Tuần 4 | Huấn luyện mô hình (LSTM và GRU) |
| Tuần 5 | Đánh giá mô hình |
| Tuần 6 | Kiểm thử, cải tiến, viết báo cáo tổng kết |

2. Project Execution

**2.1. Data Acquisition**

**Nguồn:** Dataset từ Kaggle: Water Quality Prediction

**Mô tả:** Bộ dữ liệu Water Potability từ Kaggle (liên kết) bao gồm 3276 mẫu với 10 cột:

* **Đặc trưng**: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity (9 cột, kiểu float64).
* **Mục tiêu:** Potability (0: không uống được, 1: uống được, kiểu int64).
* **Vấn đề**: Có giá trị thiếu ở các cột pH (~491), Sulfate (~781), Trihalomethanes (~162). Dữ liệu không có cột thời gian, cần giả định thứ tự mẫu để áp dụng GRU+LSTM.

**2.2. Training Methodology**

**Mô hình:** Sử dụng hai mô hình mạng học sâu:

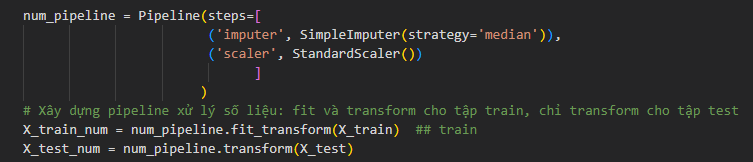
* LSTM (Long Short-Term Memory)
* GRU (Gated Recurrent Unit)

**Tiêu chí đánh giá:** Accuracy, Precision, Recall, F1-score

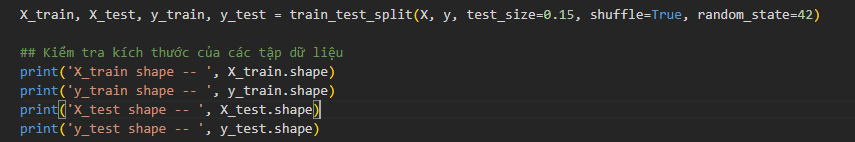
**2.3. Workflow**

1. ***Tiền xử lý dữ liệu***
   1. ***Xử lý giá trị thiếu***

* Phương pháp: Sử dụng SimpleImputer với chiến lược trung vị (median) trong pipeline:



* + Sử dụng SimpleImputer với chiến lược trung vị (median) để điền giá trị thiếu cho các cột ph, Sulfate, Trihalomethanes.
  + Trung vị được chọn vì các đặc trưng có thể không tuân theo phân phối chuẩn, tránh làm lệch dữ liệu.
* **Chuẩn hóa**:
  + Sử dụng StandardScaler để chuyển các đặc trưng về mean=0, std=1.
  + Đảm bảo dữ liệu trên cùng thang đo, phù hợp với mô hình GRU+LSTM.
* **Kết quả**:
  + X\_train\_num: Ma trận (2784, 9) với các giá trị đã được điền và chuẩn hóa.
  + X\_test\_num: Ma trận (492, 9) với các giá trị đã được điền và chuẩn hóa theo cùng tham số từ tập train.
* **Nhận xét**: Pipeline đảm bảo xử lý đồng bộ giữa train và test, tránh rò rỉ dữ liệu.
  1. ***Chia tập dữ liệu:***

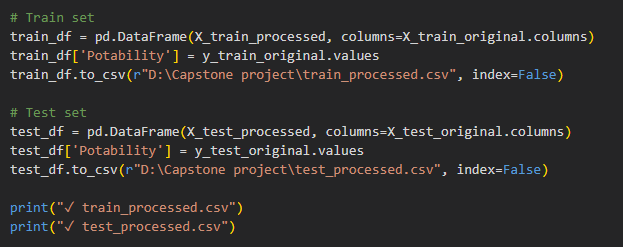
****

**Tỷ lệ:** 85% train, 15% test

**Kích thước**: X\_train (2784, 9), X\_test (492, 9)

**Nhận xét**: Tỷ lệ chia 85:15 là hợp lý, nhưng shuffle=True có thể không phù hợp với dữ liệu chuỗi thời gian, vì GRU+LSTM yêu cầu giữ thứ tự thời gian.

* 1. ***Lưu dữ liệu đã xử lý***

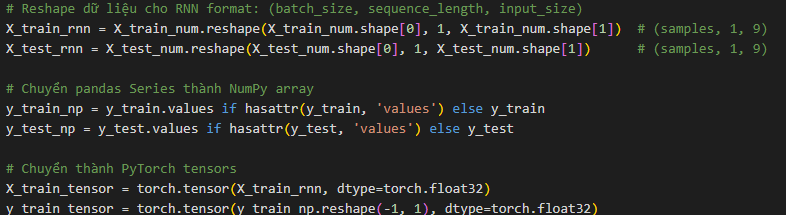
****

* Lưu hai file CSV: **train\_processed**.csv và **test\_processed**.csv.

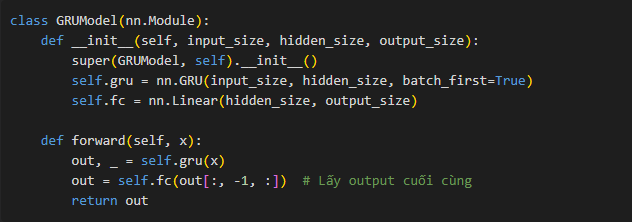
Mỗi file chứa các đặc trưng đã xử lý và cột **Potability**.

1. ***Huấn luyện mô hình GRU và LSTM***
   1. *Mô hình GRU*

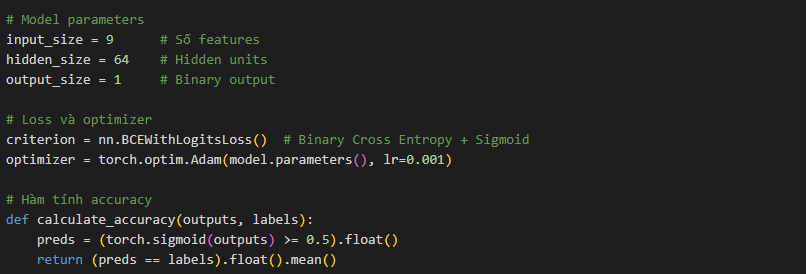
* **Chuẩn bị dữ liệu:**



* Chuyển đổi framework: Từ NumPy sang PyTorch tensors
* Kiểu dữ liệu: float32 để tối ưu bộ nhớ và tốc độ tính toán
* Reshape target: Chuyển từ 1D thành 2D để phù hợp với binary classification
* GPU compatibility: Tensors có thể dễ dàng chuyển lên GPU
* **Định nghĩa kiến trúc mô hình:**

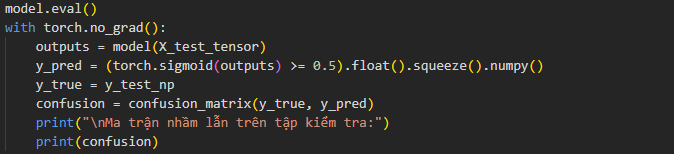


* GRU Layer: 9 inputs → 64 hidden units
* Linear Layer: 64 hidden → 1 output (binary classification)
* out[:, -1, :]: Chỉ lấy output của timestep cuối
* **Cấu hình training**



**Mục đích:** Huấn luyện mô hình GRU bằng tối ưu hóa Adam và hàm mất mát BCEWithLogitsLoss.

* Huấn luyện trong 100 epoch, mỗi epoch tính loss và accuracy.
* Sử dụng sigmoid để chuyển đầu ra về xác suất trước khi phân loại.
* Theo dõi quá trình huấn luyện bằng cách in loss và accuracy mỗi epoch.
* **Đánh giá mô hình trên tập kiểm tra:**

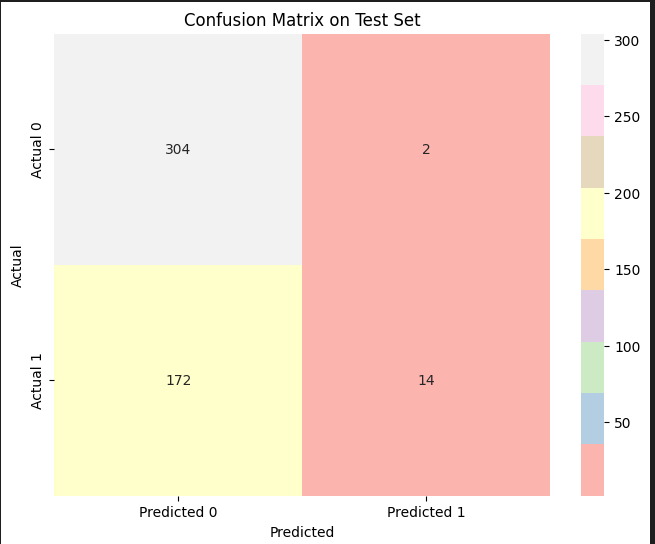


**Kết quả đánh giá trên tập kiểm tra:**

* Accuracy: 64.63%
* Precision: 87.50%
* Recall: 7.53%
* F1-score: 13.86%

=> Precision cao nhưng Recall thấp, cho thấy mô hình rất thận trọng khi dự đoán nước uống được. Phần lớn mẫu dương tính bị bỏ sót.

Đánh giá hiệu năng của mô hình GRU bằng confusion matrix.



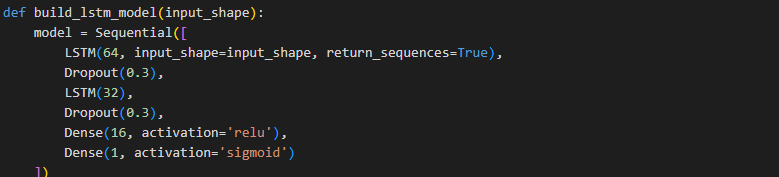
* 1. *Mô hình LSTM*
* **Chuẩn bị dữ liệu:**



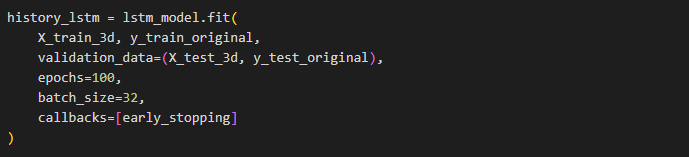
* **Thiết lập Early Stopping**

****

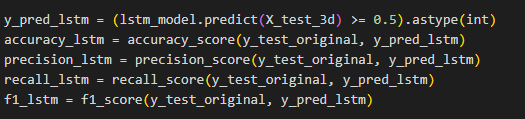
* Ngăn mô hình overfitting bằng cách dừng sớm nếu val\_loss không cải thiện.
* Nếu sau 10 epoch liên tiếp không giảm val\_loss, sẽ dừng và khôi phục trọng số tốt nhất.
* **Xây dựng kiến trúc LSTM model:**

****

* Xây dựng mô hình LSTM sâu với nhiều lớp nhằm tăng khả năng học đặc trưng.
* **Huấn luyện mô hình:**

****

* Huấn luyện mô hình với dữ liệu training, đồng thời đánh giá trên validation set.
* Cài đặt: 100 epoch, batch size = 32, kết hợp EarlyStopping.
* Output: Biến history\_lstm lưu lại quá trình huấn luyện để trực quan hóa nếu cần.
* **Hàm đánh giá:**



Kết quả đánh giá trên tập kiểm tra:

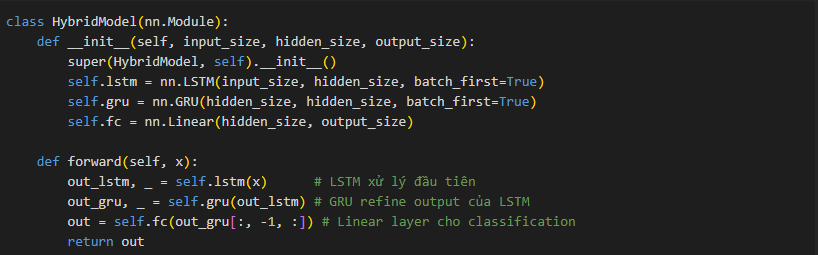
* Accuracy: 68.70%
* Precision: 63.56%
* Recall: 40.32%
* F1-score: 49.34%

=> Mô hình cân bằng hơn giữa Precision và Recall, cho kết quả tổng thể tốt hơn GRU.

* 1. *Mô hình Hybird*
* **Chuẩn bị dữ liệu**

****

* **Kiến trúc model Hybird**

****

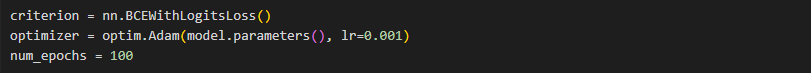
* Kết hợp 2 loại mạng RNN là LSTM và GRU để tận dụng ưu điểm của cả hai.

**Cấu trúc mô hình:**

* LSTM: học đặc trưng chuỗi đầu vào.
* GRU: tiếp tục xử lý đầu ra từ LSTM để nắm bắt các mối quan hệ sâu hơn.
* Linear: lớp đầu ra 1 node cho bài toán phân loại nhị phân.

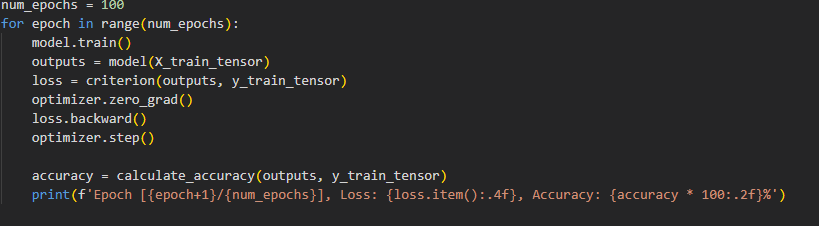
**Ý nghĩa:** LSTM có khả năng ghi nhớ dài hạn tốt, GRU thì nhẹ và nhanh hơn. Kết hợp này giúp cân bằng giữa hiệu quả và hiệu suất.

* **Định nghĩa hàm Loss và Optimizer**

****

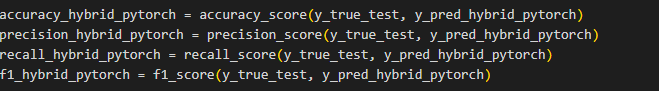
Cài đặt hàm mất mát và thuật toán tối ưu để huấn luyện mô hình.

* BCEWithLogitsLoss: hàm kết hợp sigmoid + binary cross-entropy cho bài toán phân loại nhị phân.
* Adam: thuật toán tối ưu hóa với tốc độ hội tụ tốt.
* **Huấn luyện mô hình**

****

Tối ưu trọng số qua nhiều vòng lặp (epoch).

* Mỗi epoch gồm các bước: forward pass, tính loss, backward pass (gradient), update trọng số.
* Theo dõi loss và accuracy mỗi epoch để kiểm tra tiến trình huấn luyện.
* **Hàm đánh giá:**

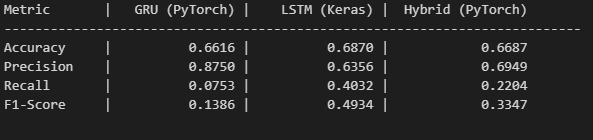
****

Kết quả đánh giá trên tập kiểm tra:

* Accuracy: 66.87%
* Precision: 69.49%
* Recall: 22.04%
* F1-score: 33.47%

=> Mô hình hybrid đạt được sự cân bằng giữa GRU và LSTM, với Recall cao hơn GRU và Precision cao hơn LSTM, nhưng vẫn còn hạn chế trong việc phát hiện các mẫu dương tính.

1. ***Đánh giá kết quả***



=> LSTM vẫn là mô hình thể hiện tốt nhất về tổng thể. GRU có độ chính xác cao với mẫu dương nhưng bỏ sót nhiều mẫu thực sự dương tính. Hybrid có hiệu năng trung gian, cho thấy tiềm năng nếu được tinh chỉnh thêm.

**2.4. System Design**

3. Results

**3.1. Data Preprocessing**

* Điền giá trị thiếu bằng trung bình hoặc phương pháp thống kê
* Chuẩn hóa dữ liệu bằng MinMaxScaler
* Chuyển đổi định dạng phù hợp cho mô hình LSTM/GRU

**3.2. Exploratory Data Analysis (EDA)**

* Quan sát phân bố các đặc trưng như pH, hardness, solids,...
* Kiểm tra tương quan giữa các đặc trưng
* Tìm hiểu sự khác biệt giữa hai lớp dữ liệu (uống được / không)

**3.3. Modeling**

* **GRU:** Đạt độ chính xác ~64%, F1-score ~0.1386
* **LSTM:** Đạt độ chính xác ~68%, F1-score ~0.4934
* Đánh giá: Mô hình LSTM hoạt động tốt hơn trong trường hợp này

**3.4. User Interface**

**3.5. Testing and Improvements**

* So sánh các mô hình truyền thống (Logistic Regression, Random Forest) với GRU/LSTM
* Tối ưu hyperparameter: số tầng, learning rate, số neuron
* Cải tiến giao diện và kiểm thử người dùng



4. Projected Impact

**4.1. Accomplishments and Benefits**

* Cung cấp công cụ dự đoán chất lượng nước nhanh chóng, tiết kiệm thời gian và chi phí
* Tăng khả năng phát hiện nước ô nhiễm ở vùng xa, hỗ trợ quyết định kịp thời
* Ứng dụng tiềm năng trong hệ thống giám sát chất lượng nước tự động

**4.2. Future Improvements**

* Thu thập dữ liệu thực tế đa dạng hơn từ các khu vực khác nhau
* Tích hợp IoT để đo đạc thông số nước theo thời gian thực
* Phát triển mô hình sâu hơn (Transformer, Attention) để cải thiện độ chính xác