**VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEMS**

Logo

Description automatically generated

**PROJECT SUMMARY**

**DATA MINING SUBJECTS**

**PREDICTION AND CLASSIFICATION OF AIR QUALITY USING POLLUTION DATA  
MEMBERS:**

**Bui Dinh Trieu 21521576**

**Cao Pham Tien Anh 20520380**

**Instructors**

**TS. Cao Thi Nhan**

**Nguyen Thi Viet Huong**

**Ho Chi Minh City, November 2023**

|  |
| --- |
| **VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY**  **UNIVERSITY OF INFORMATION TECHNOLOGY**  **FACULTY OF INFORMATION SYSTEMS**    Logo  Description automatically generated  **PROJECT SUMMARY**  **DATA MINING SUBJECTS**  **PREDICTION AND CLASSIFICATION OF AIR QUALITY USING POLLUTION DATA MEMBERS:**  **Bui Dinh Trieu 21521576**  **Cao Pham Tien Anh 20520380**  **Instructors**  **TS. Cao Thi Nhan**  **Nguyen Thi Viet Huong**  **Ho Chi Minh City, November 2023** |

**TABLE OF CONTENTS**

[Prologue 1](#_Toc151317824)

[CHAPTER 1: Problem description and dataset: Description of the problem and the dataset used. 2](#_Toc151317825)

[**1.1** **Predicting future air quality** 2](#_Toc151317826)

[**1.4** **Classification of air quality states** 2](#_Toc151317827)

[**1.4** **Classification of air quality states** 2](#_Toc151317828)

[**1.4** **Dataset** 2](#_Toc151317829)

[CHAPTER 2: Algorithms used: Description of the algorithms used 3](#_Toc151317830)

[**2.1.** **Time Series Forecasting with LSTM (Long Short-Term Memory):** 3](#_Toc151317831)

[**2.2.** **ARIMA (AutoRegressive Integrated Moving Average):** 4](#_Toc151317832)

[**2.3.** **Random Forest Classifier:** 4](#_Toc151317833)

[**2.4.** **Support Vector Machine (SVM) Classifier:** 4](#_Toc151317834)

[**2.5.** **Correlation Analysis:** 4](#_Toc151317835)

[**2.6.** **Principal Component Analysis (PCA):** 4](#_Toc151317836)

[CHAPTER 3: Expected results: Description of the expected results of the problem 4](#_Toc151317837)

[**3.1.** **Expected results:** **Predicting Future Air Quality** 5](#_Toc151317838)

[**3.2.** **Expected results:** **Finding Correlations Between Types of Toxins** 5](#_Toc151317839)

[**3.3.** **Expected results:** **Air Quality State Classification** 5](#_Toc151317840)

[References 6](#_Toc151317841)

# CHAPTER 1: Problem description and dataset: Description of the problem and the dataset used.

* 1. **Predicting future air quality**

This problem involves using attributes (such as CO, NO2, O3, SO2 Mean) to predict future air quality based on historical data samples. This can be useful for predicting bad air quality situations and giving early warning.

1. 1. **Classification of air quality states**

Classification model to determine if air quality at a particular point in time falls into states such as "good," "poor," "hazardous," "normal".

* 1. **Finding correlations between types of toxins:**

Mine data to understand the correlation between toxins such as CO, NO2, SO2 and O3. This problem helps us to better understand how these factors interact and can lead to measures to improve air quality.

* 1. **Dataset**

Link dataset:

[US Pollution 2000-2021 (kaggle.com)](https://www.kaggle.com/datasets/alpacanonymous/us-pollution-20002021)

Date: Dates, usually sorted in the format YYYY-MM-DD and used to specify specific times for data types.

Year: The year represents the year of the measurement data, allowing you to view information by year.

Month: The month represents the month of the measurement data.

Day: The date represents the date of the measurement data.

Address: The specific address or location where the measurement data was collected.

State: The state in the United States where the measurement data was collected.

Country: The country or territory where the measurement data was collected.

City: The specific city or region where the measurement data was collected.

O3 Mean: The average value of ozone (O3) in the air at a specific location and specific time.

O3 1st Max Value: The highest measured value for ozone (O3) for a specific period.

O3 Max Hour: The hour of the day when O3 reaches its highest value.

O3 AQI: Air Quality Index for ozone (O3), calculated based on measured values.

CO Mean: The average value of CO (carbon monoxide) in the air at specific location and specific time.

CO 1st Max Value: The highest measured value for CO (carbon monoxide) over a specific period.

CO Max Hour: The hour of the day when CO reaches its highest value.

CO AQI: Air Quality Index for CO (carbon monoxide) calculated based on measured values.

SO2 Mean: The average value of SO2 (sulfur dioxide) in the air at specific location and specific time.

SO2 1st Max Value: The highest measured value for SO2 (sulfur dioxide) over a specific period.

SO2 Max Hour: The hour of the day when SO2 reaches its highest value.

SO2 AQI: Air Quality Index for SO2 (sulfur dioxide), calculated based on measured values.

NO2 Mean: The average value of ozone NO2 (nitrogen dioxide) in the air at a specific location and specific time.

NO2 1st Max Value: The highest value measured for NO2 (nitrogen dioxide) over a specific period.

NO2 Max Hour: The hour of the day when NO2 reaches its highest value.

NO2 AQI: Air Quality Index NO2 (nitrogen dioxide), calculated based on measured values.

# CHAPTER 2: Algorithms used: Description of the algorithms used

* 1. **Time Series Forecasting with LSTM (Long Short-Term Memory):**

LSTM is an RNN (Recurrent Neural Network) neural network architecture specifically designed to process time series data. It is capable of learning complex patterns and can be used to predict future air quality based on historical attributes such as CO, NO2, O3, SO2 Mean.

* 1. **ARIMA (AutoRegressive Integrated Moving Average):**

ARIMA is a method of statistical time series analysis. It can be used to predict trends and cycles in data, helping to predict future air quality based on time attributes.

* 1. **Random Forest Classifier:**

Random Forest is an Ensemble Learning algorithm that uses multiple decision trees to classify data. It can be used to determine air quality status into groups such as "good," "poor," "hazardous," "normal" based on attributes such as CO, NO2, O3, SO2 Mean.

* 1. **Support Vector Machine (SVM) Classifier:**

SVM is a powerful classification algorithm, which can be used to determine air quality status based on the spatial characteristics of the data.

* 1. **Correlation Analysis:**

Use correlation analysis to measure the relationship between toxins such as CO, NO2, SO2 and O3. This helps us to better understand how they interact and can affect air quality.

* 1. **Principal Component Analysis (PCA):**

PCA is used to reduce data dimension and identify key components that can help understand the correlation between independent variables.

# CHAPTER 3: Expected results: Description of the expected results of the problem

* 1. **Expected results:** **Predicting Future Air Quality**

LSTM and ARIMA models are trained on historical datasets.

Predictions of future air quality are generated for each sample of data in the test set.

A comparison chart between actual and predicted values is created to evaluate the performance of the model.

Evaluate performance using measures such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

* 1. **Expected results:** **Finding Correlations Between Types of Toxins**

Correlation matrices between toxins such as CO, NO2, O3 and SO2 are generated from the data.

The heatmap chart clearly shows the correlation between toxins.

Detailed analysis of the correlation between pairs of toxins to better understand the interaction between them.

Provide conclusions and suggest measures to improve air quality based on the correlation detected.

* 1. **Expected results:** **Air Quality State Classification**

The Random Forest model and SVM Classifier were trained on a dataset with labeled air quality status.

Evaluate your model's performance using test sets and compare the predicted results with actual labels.

The Confusion Matrix and measurements such as Precision, Recall, and F1-Score are calculated to evaluate the gradability of the model.

# References

* [1] [What is LSTM? Introduction to Long Short-Term Memory (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/#:~:text=LSTM%20(Long%20Short%2DTerm%20Memory,ideal%20for%20sequence%20prediction%20tasks.)
* [2] [Random Forest algorithm — Machine Learning cho dữ liệu dạng bảng (machinelearningcoban.com)](https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html)
* [3] [US Pollution 2000-2021 (kaggle.com)](https://www.kaggle.com/datasets/alpacanonymous/us-pollution-20002021)
* [4] [Download Files | AirData | US EPA](https://aqs.epa.gov/aqsweb/airdata/download_files.html)
* [5] [Data Mining: 6 Kỹ thuật quan trọng trong khai phá dữ liệu (isb.edu.vn)](https://insight.isb.edu.vn/6-ky-thuat-quan-trong-trong-khai-pha-du-lieu/)
* [6] [A systematic review of data mining and machine learning for air pollution epidemiology | BMC Public Health | Full Text (biomedcentral.com)](https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-017-4914-3)
* [7] [Application of data mining combined with power data in assessment and prevention of regional atmospheric pollution - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2352484723001579)