UE20CS461A – Capstone Project Phase – 2

Project Progress Review #1

Project Title : Monitoring the concentration of air pollutants and its health hazards using Machine Learning models.

Project ID : 102

Project Guide : Prof. Saritha

Project Team :

Aditi Jain	PES2UG20CS021
Aditya R Shenoy	PES2UG20CS025
Ananya Adiga	PES2UG20CS043
Anirudha Anekal	PES2UG20CS051

Outline

- Abstract and Scope of the Project
- Capstone Project Phase – 1
  - Summary of work
  - Suggestions
  - Improvements
- Approach using simulated imitated data
- Expected Deliverables
- Contributions
- Technologies
- Gantt chart
- References

Abstract

- Introducing a continuous air quality monitoring system that tracks local air quality to anticipate higher chances of lung cancer. A blend of Adaptive LSTM and ARIMA ML models will be employed and hosted on a cloud platform.

- The system's focus is to constantly gauge PM2.5, PM10, NO2, and CO levels, offering users real-time insights into their environmental air quality.

- Utilizing this data, the system will evaluate potential lung cancer risks, proactively alerting users to changing air quality and potential health concerns via IoT sensors.

- Our vision is to provide users with timely information, helping them make informed decisions for their well-being, driven by cutting-edge technology and real-world impact.

Scope

- We are proposing a system that continuously monitors air quality in the user's area, predicting and alerting about increased lung cancer risk. Using a hybrid of Adaptive LSTM and ARIMA ML models, this will be hosted on a cloud platform.

- Our focus is on tracking PM2.5, PM10, NO2, and CO levels in real-time, offering users insights into local air quality.

- By analyzing this data, the system will gauge lung cancer risk, proactively notifying users via IoT sensors if air quality changes.

- Our aim is to empower users with timely information for informed choices, employing cutting-edge tech to make a real-world impact.

Summary of Work Done in Capstone Project Phase - 1

Work done in Phase 1:

- Performed literature review.
- Finalized the ML model.
  - bidirectional adaptive LSTM + ARIMA
- Created architectural and structural diagrams.
- Chosen cloud platforms
  - ThingSpeak
  - Hugging face, Streamlit
- Exploratory Data Analysis

Suggestions given:

- Find datasets for our city
- Make visualisation better
- Stay on track of the proposed timeline

Improvements:

- Tackle the dataset issue
- Fine tuned the modules for the IoT sensor station
- Improved data visualisation & data pre-processing

Approach using simulated imitated data

- **Approach if data is scarce:**
  - Addressing the issue without actual datasets by outlining the methodology to solve it effectively.
- **Methodology Presentation:**
  - Illustrating step-by-step algorithm for solving the issue using a well-structured approach.
- **Simulated Environment:**
  - Generating data in a controlled, simulated environment for demonstration purposes.
  - Data will be generated randomly to showcase the methodology.

This approach is a conceptual approach we will take in case proper datasets are not available.

- **Simulation Tool or Manual Creation:**
  - Utilizing a simulation tool to generate data or manually creating representative data points.
  - Ensuring the simulated data mimics real-world scenarios.
- **Focused Presentation:**
  - Focus is on showcasing the thought process, algorithmic solutions, and data generation approach.
  - Emphasizing methodology's feasibility and potential real-world impact.

Expected Deliverables

- Sourcing data
- Data preprocessing
- Data visualisation
- Data input
- Model training + Optimisation
- Generating results
- Real time data acquisition via IoT
- Deployment to cloud

Expected Deliverables

Sourcing data

Sourced original lung cancer dataset from Harvard website
Tried to source dataset from National Cancer Institute

Data pre-processing

Feature Selection
Cleaned the data for NULL values and duplicate entries
Data transformation

Expected Deliverables

Data visualisation

Checking outliers, for each column and removing them
Checking the data range via graphs
Histograms, count plots, skew graphs

Data input

Automated Data Collection via IoT sensors
File Upload .csv file for the already acquired data
Data Streaming for continous data via IoT sensors

Expected Deliverables

Model training + Optimisation

Scaling data and fitting it for the model
Apply the fitted data to the model
Hyperparameter Tuning to increase optimisation
Model Evaluation

Generating results

Evaluate Performance
Visualize Results as different graphs
Handle Edge Cases
Deploy the model

Expected Deliverables

Real time data acquisition via IoT

Sensor data generation (like MQ9/7, SharpGP2Y10, etc.)
Data Retention and Archival
Feedback Loop and Improvements

Deployment to cloud

Connecting the IoT sensors to ThingSpeak
Deployment of ML model to Streamlit/Hugging Face

Contribution

Modules	Members	Development
Sourcing data	Everyone	Time - 1 week
Data pre-processing	Aditi Jain Ananya Adiga	Time - 3 days Lines of code - 40 lines
Data visualisation	Aditi Jain Ananya Adiga	Time - 3 days Lines of code - 50 lines

Contribution

Modules	Members	Development
Building the IoT sensor station	Anirudha Anekal Aditya R. Shenoy	Time - 2 weeks
Calibrating and code for the IoT sensors	Anirudha Anekal Aditya R. Shenoy	Time - 4 days Lines of code - 37 lines
Uploading IoT sensor readings to ThingSpeak	Anirudha Anekal Aditya R. Shenoy	Time - 3 days

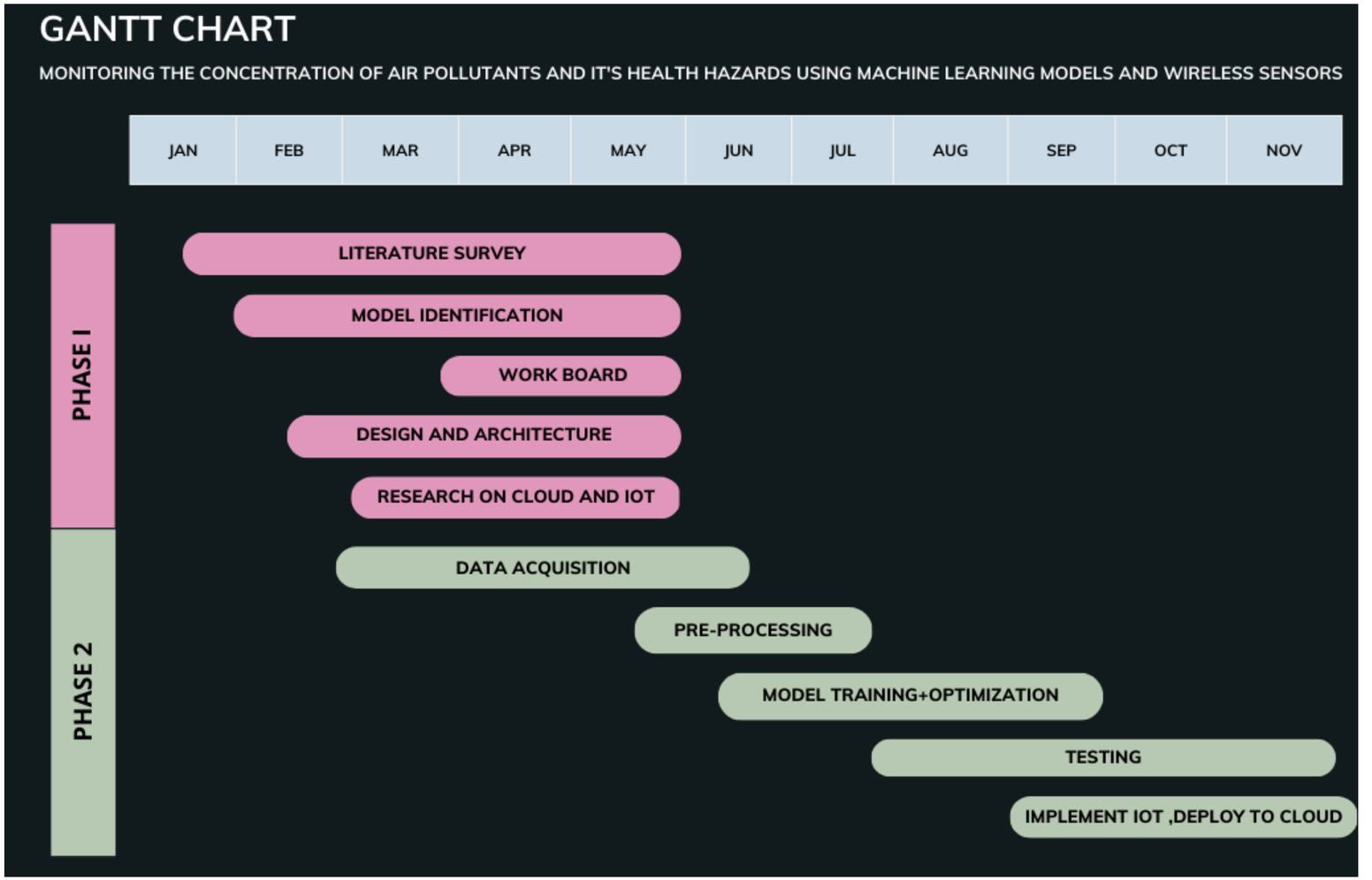
Contribution

Modules	Members	Development
Coding the model	Aditi Jain Ananya Adiga	Time - 3 weeks Lines of code - 58 lines
Hyperparameter Tuning to increase optimisation Model Evaluation	Everyone	Time - 3 days Lines of code - 20 lines

Technologies Used so far

- Google Colab
- ARIMA + enhanced LSTM
- Sensors: MQ7/9(Carbon Monoxide), GP2Y1010F (Dust)
- ESP8266
- ThingSpeak
- Hugging Face / Streamlit

Gantt Chart


GANTT CHART
MONITORING THE CONCENTRATION OF AIR POLLUTANTS AND IT'S HEALTH HAZARDS USING MACHINE LEARNING MODELS AND WIRELESS SENSORS
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
PHASE I
PHASE 2
LITERATURE SURVEY
MODEL IDENTIFICATION
WORK BOARD
DESIGN AND ARCHITECTURE
RESEARCH ON CLOUD AND IOT
DATA ACQUISITION
PRE-PROCESSING
MODEL TRAINING+OPTIMIZATION
TESTING
IMPLEMENT IOT ,DEPLOY TO CLOUD

References

[1] An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities
By: Muhammad Taha Jilani, Husna Gul A.Wahab
[https://ieeexplore.ieee.org/document/8981707]

[2] How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking
By: Kung-Min Wang, Kun-Huang Chen, Shieh-Hsen Tseng
[https://www.mdpi.com/1660-4601/19/14/8445]

[3] Assessment of indoor air quality in academic buildings usng IOT and deep learnings
By: Mohammad Marzouk and Mohammad Atef
[https://www.mdpi.com/1667822]

References

[4] Household Ventilation May Reduce Effects of Indoor Air Pollutants for Prevention of Lung Cancer: A Case-Control Study in a Chinese Population.
By: Jin Z-Y, Wu M, Han R-Q, Zhang X-F, Wang X-S, et al.
[https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102685]

[5]Determination of Air Quality Life Index (AQLI) in Medinipur City of West Bengal(India) During 2019 To 2020 : A contextual Study
By: Samiran Rana
[https://www.researchgate.net/publication/360622768_Determination_of_Air_Quality_Life_Index_Aqli_in_Medinipur_City_of_West_BengalIndia_During_2019_To_2020_A_contextual_Study]

[6] The nexus between COVID-19 deaths, air pollution and economic growth in New York state: Evidence from Deep Machine Learning
By: Cosimo Magazzino , Marco Mele , Samuel Asumadu Sarkodie
[https://www.sciencedirect.com/science/article/pii/S0301479721003030]

Thank You

We open the floor for questions.