MIST | Technology for Advancement

Military Institute of Science and Technology

Department of Computer Science and Engineering

Course Code: CSE - 404

Course Title: Artificial Intelligence Sessional

Course Instructors: Dr. Md Akhtaruzzaman, Dr. Nusrat Sharmin, Raiyan Rahman, and Shahriar Rahman Khan

Final Project Report

of

Correlation of Air Quality and Air-borne Diseases and Prediction of Disease Burden Using Machine Learning

27 June 2023, Tuesday

Members of Group 2

Bijoya Ghosh - 202014009

Ellora Yasi - 202014012

Khaled Hasan Irfan - 202014019

Sheikh Easin Arafat - 202014049

Tariq Hasan Rizu - 202014058

CSE-20 (Section A)

Contents

Abstract

Our project explores the correlation between air quality and disease burden, focusing on predicting disease outcomes using machine learning techniques. By leveraging a data set encompassing key features such as location, year, measure, gender, cause, and PM concentration, we employ advanced machine learning algorithms to analyze the impact of airborne pollutants on disease rates. By converting non-numerical features into numeric values and applying various regression models, we uncover hidden patterns and relationships, enabling the development of predictive models. The project's findings significantly affect public health interventions, policy-making, and resource allocation. By understanding the link between air quality and disease burden, we can design targeted strategies for prevention and intervention, leading to improved public health outcomes. This project contributes to the growing body of knowledge in air quality analysis and provides actionable insights for evidence-based decision-making in public health.

1 Introduction

Our AI project focuses on analyzing the correlation between air quality and disease burden using machine learning techniques. We trained a model using a dataset spanning from 2010 to 2019, consisting of feature columns such as ’Location’, ’Year’, ’Measure’, ’Gender’, ’Cause’, and ’PM Concentration’, with the target variable being the ’Rate Per 100000 Population’. The trained model enables us to predict the effects of airborne diseases, including the number of deaths, disability-adjusted life years (DALY), and years of life lost (YLL), for upcoming years.

By utilizing the power of machine learning algorithms, specifically regression models, we aim to uncover the relationship between air quality factors and their impact on public health. The conversion of non-numerical features into numeric values allows us to effectively analyze the data and derive meaningful insights. The model’s accuracy and performance are evaluated using metrics such as mean squared error, mean absolute error, and R-squared score.

Our project’s findings hold significant implications for public health interventions and policy-making. By understanding the correlation between air quality and disease burden, we can develop targeted strategies to mitigate the adverse effects of airborne diseases. Furthermore, the predictive capabilities of our model provide valuable insights into the potential future trajectory of disease rates, enabling proactive measures for effective resource allocation and intervention planning.

Our project showcases the power of machine learning in exploring the correlation between air quality and disease burden. Moreover, this project seeks to provide valuable insights and predictive capabilities that can drive impactful changes in disease prevention, health policy, and resource allocation. Ultimately, we aspire to create a healthier future for communities by empowering public health authorities with actionable information derived from advanced machine learning techniques.

2 METHODOLOGY

2.1 Data Overview

We mainly collected datasets from two sources. The disease burden data set is collected from Global Health Data Exchange (GHDx) and the Air Exposure data set from World Health Organization (WHO). After that the data sets were filtered and merged to make our own data set for the project.

Table 1: DATASET DETAILS

Column Name	**Description**
Location	Contains the name of the different country
Year	which the data is obtained. This dataset year range is from 2010 to 2019
Measure	Encompasses metrics such as deaths, disability-adjusted life years (DALYs), and years of life lost (YLLs).
Gender	Male and Female categories
Cause	Specific airborne diseases. Here in our dataset only Respiratory infections and tuberculosis, Chronic respiratory diseases, and Cardiovascular diseases are presents.
PM Concentration	PM stands for particulate matter, which refers to tiny particles suspended in the air.
Rate Per 100000 Population	Includes the number of deaths, DALYs, or YLLs per 100,000 individuals.

2.2 System Architecture

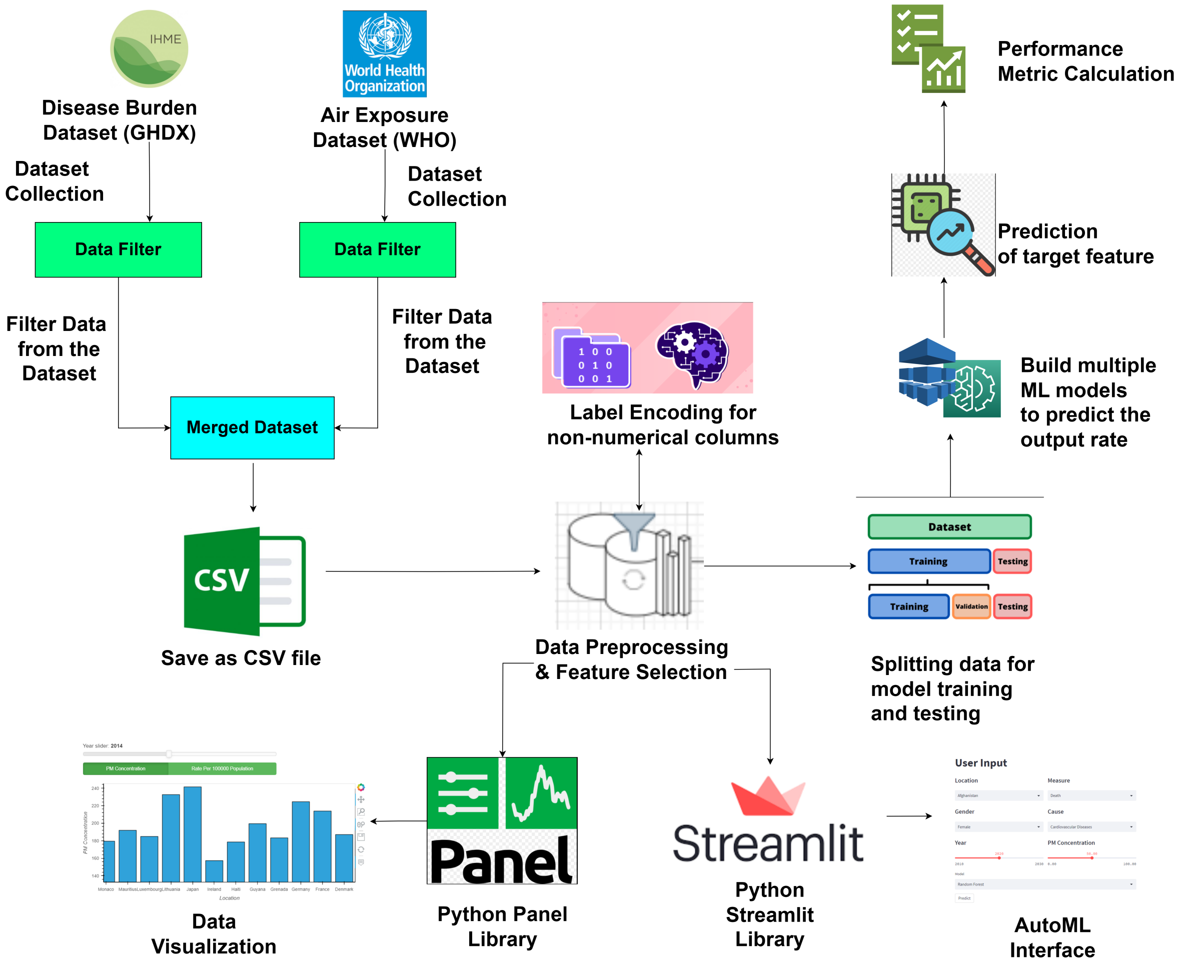

IHME
Disease Burden Dataset (GHDX)
World Health Organization
Air Exposure Dataset (WHO)
Dataset Collection
Dataset Collection
Data Filter
Data Filter
Filter Data from the Dataset
Filter Data from the Dataset
Merged Dataset
CSV
Save as CSV file
Label Encoding for non-numerical columns
Data Preprocessing & Feature Selection
Performance Metric Calculation
Prediction of target feature
Build multiple ML models to predict the output rate
Dataset
Training
Testing
Training
Validation
Testing
Splitting data for model training and testing
Data Visualization
Panel
Python Panel Library
Streamlit
Python Streamlit Library
User Input
Location
Measure
Gender
Cause
Year
PM Concentration
Predict
AutoML Interface

Figure 1: System Architecture of our system

Our system architecture is designed to support the correlation analysis of air quality and disease burden using machine learning techniques. The architecture consists of several key components that enable data preprocessing, model training, prediction, and evaluation.

1. **Data Input:**

   - The system takes input from a dataset that contains relevant information such as location, year, measure, gender, cause, and PM concentration.
   - This dataset serves as the foundation for our analysis and prediction tasks.

2. **Data Preprocessing:**

   - Before feeding the data into machine learning models, preprocessing steps are applied.

- This includes handling missing values, encoding categorical variables, and scaling numerical features.
- Preprocessing ensures the data is in a suitable format for training and testing the models.

3. **Model Training:**

   - Various machine learning models are employed to analyze the correlation between air quality factors and disease burden.
   - These models include decision tree regressors, gradient boosting regressors, ridge regression, neural network regression, support vector regression, and polynomial regression.
   - Each model is trained using the preprocessed data to learn the underlying patterns and relationships.

4. **Model Evaluation:**

   - Once the models are trained, they are evaluated using performance metrics such as mean squared error (MSE), mean absolute error (MAE), and R2 score.
   - These metrics provide insights into the accuracy and reliability of the models in predicting disease burden based on air quality factors.

5. **Prediction:**

   - The trained models are then used to make predictions on new or unseen data.
   - By providing inputs related to location, year, measure, gender, cause, and PM concentration, the models can estimate disease burden, including the number of deaths, DALY, and YLL.
   - These predictions help in understanding the potential impact of air quality on disease outcomes.

6. **Result Analysis:**

   - The predicted results are analyzed and interpreted to gain insights into the correlation between air quality and disease burden.
   - The findings can be visualized through graphs, charts, or other graphical representations to facilitate understanding and decision-making.

2.3 Description of used Machine Learning Models

Random Forest Regressor: Random Forest Regressor Model is an ensemble learning model that builds multiple decision trees and averages their predictions. This makes it more robust to overfitting and can often provide better predictions than a single decision tree.

Decision Tree Regressor: Decision Tree Regressor Model is a simple and interpretable model that builds a tree-like structure to predict the target variable. It is relatively easy to train and can be used to solve a wide variety of regression problems.

Light GBM : Light GBM is a gradient-boosting model that is known for its speed and efficiency. It can be used to solve a wide variety of regression problems, including those with a large number of features.

Gradient Boosting Regression: Gradient Boosting Regression Model is another ensemble learning model that builds multiple decision trees in a sequential manner. Each tree is trained to correct the errors of the previous tree, which can help to improve the overall accuracy of the model.

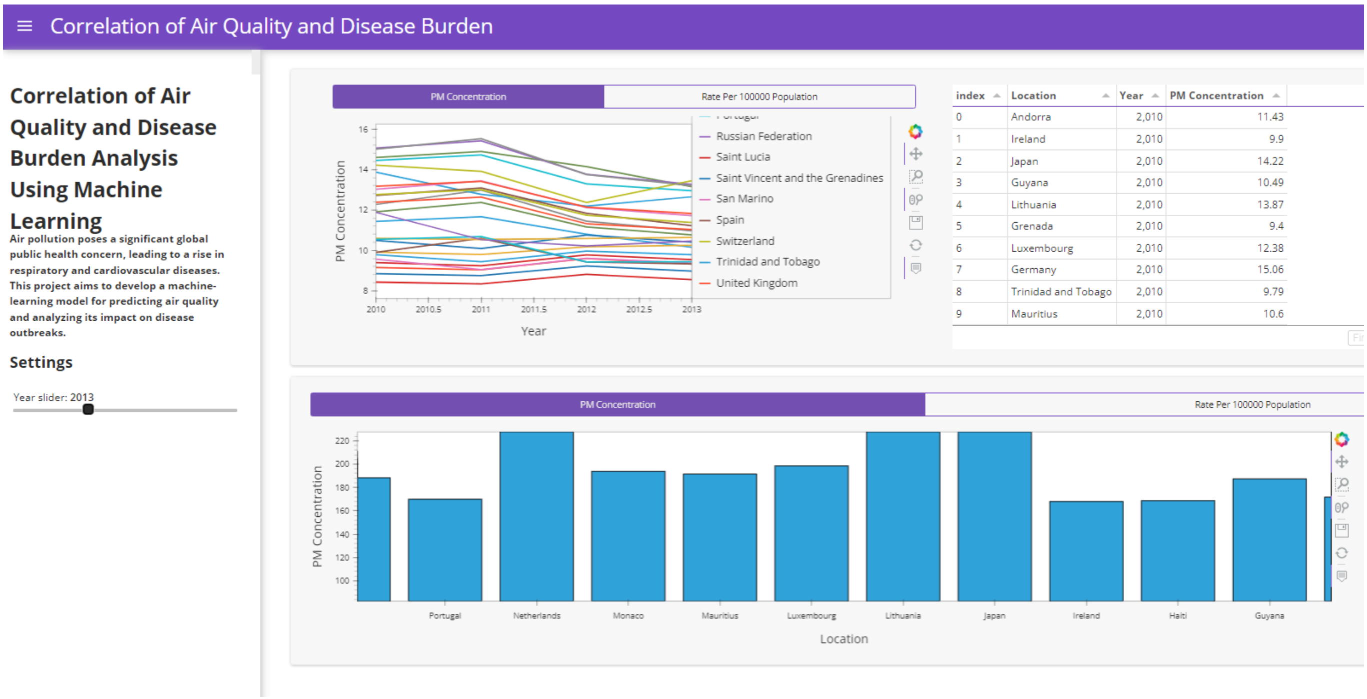
MLPRegressor: MLPRegressor is a neural network model that can be used for regression tasks. It is a more complex model than the other models mentioned, but it can often provide better predictions.

Polynomial Regression: Polynomial Regression Model is a linear regression model that uses a polynomial function to model the relationship between the independent and dependent variables. It can be used to fit non-linear relationships, but it can be less interpretable than other models.

2.4 Implementation

2.4.1 Panel dashboard

A panel dashboard is a visualization tool that helps with the visual display of key information, data, and metrics that are relevant to a specific domain or context. It provides a consolidated view of various metrics, allowing users to monitor and analyze data in a concise and accessible manner. By using the Python panel library, we created this dashboard to visualize our dataset.In Figure: 5, First graph is Line plot for the PM Concentration on different countries at each year. Second showing Data of annual PM Concentration for different countries of a selected year. and The last shows Bar chart for the PM Concentration of relevant countries in the selected year.



≡ Correlation of Air Quality and Disease Burden

Correlation of Air Quality and Disease Burden Analysis Using Machine Learning

Air pollution poses a significant global public health concern, leading to a rise in respiratory and cardiovascular diseases. This project aims to develop a machine-learning model for predicting air quality and analyzing its impact on disease outbreaks.

Settings

Year slider: 2013

index	Location	Year	PM Concentration
0	Andorra	2,010	11.43
1	Ireland	2,010	9.9
2	Japan	2,010	14.22
3	Guyana	2,010	10.49
4	Lithuania	2,010	13.87
5	Grenada	2,010	9.4
6	Luxembourg	2,010	12.38
7	Germany	2,010	15.06
8	Trinidad and Tobago	2,010	9.79
9	Mauritius	2,010	10.6

Figure 2: Panel dashboard of our system

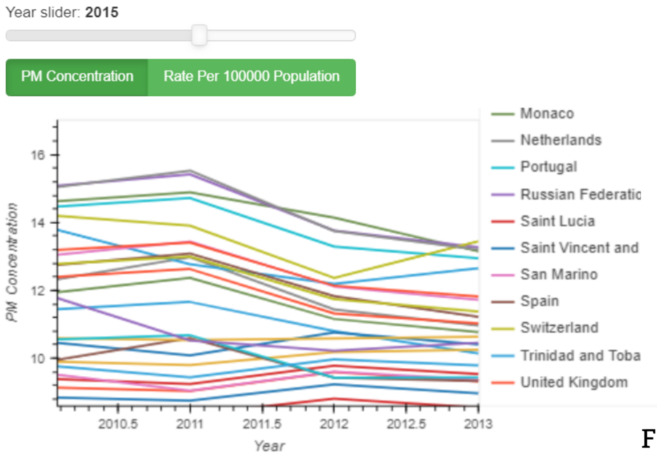


Figure 3: Line plot for the PM Concentration on different countries each year

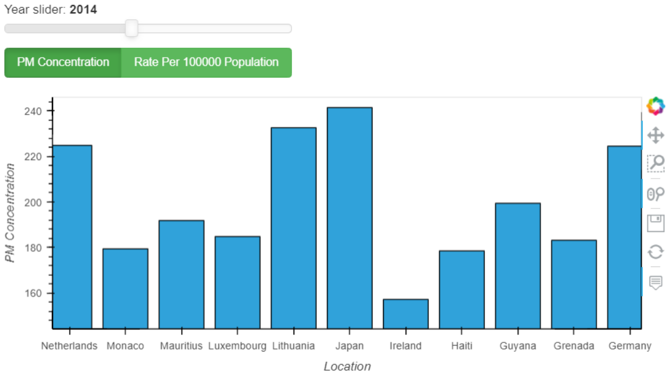
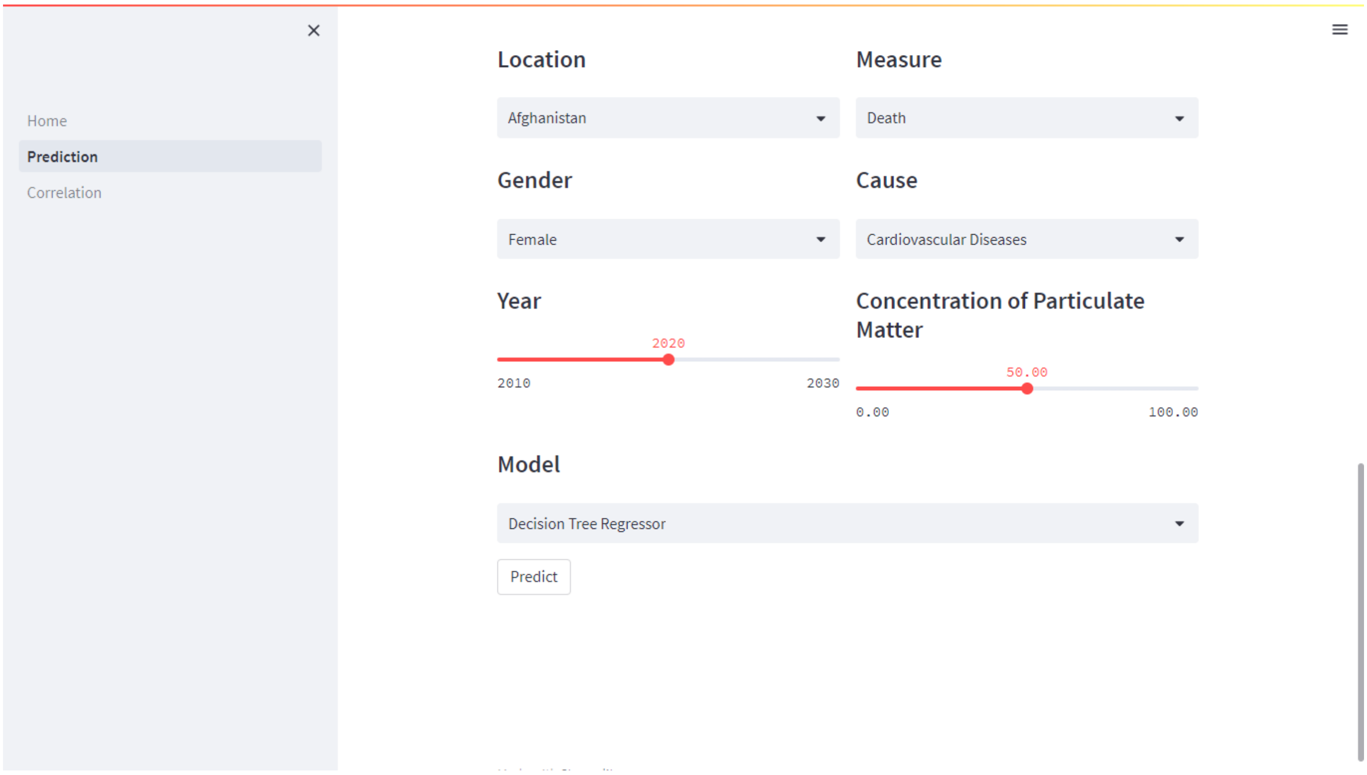


Figure 4: Bar chart for the PM Concentration of relevant countries in the selected year

2.4.2 Streamlit



Location

Afghanistan

Measure

Death

Gender

Female

Cause

Cardiovascular Diseases

Year

2020

Concentration of Particulate Matter

50.00

Model

Decision Tree Regressor

Figure 5: Streamlit User Interface of our system

Streamlit Implementation

Streamlit is an open-source Python library that allows us to build interactive web applications for data exploration and visualization. In our project, we have utilized Streamlit to create a user-friendly graphical interface for deploying our machine learning models and providing a seamless user experience.

1. **Installation:** Start by installing the Streamlit library using pip or conda package managers. Streamlit can be easily installed in your Python environment.
2. **Importing Libraries:** Import the necessary libraries, including Streamlit and other dependencies, to your Python script.
3. **Data Loading and Preprocessing:** Load the dataset into the Streamlit application and perform any required preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features. This ensures that the data is in a suitable format for training the machine learning models.
4. **Model Training and Testing:** Train your machine learning models using the preprocessed dataset. Streamlit provides a streamlined way to incorporate model training and testing code within the

application. Define functions to train the models, set hyperparameters, and evaluate their performance using various metrics.

5. **Creating User Interfaces:** Streamlit allows you to create interactive user interfaces with minimal code. Design user-friendly input forms where users can select options, input values, or upload files. These interfaces enable users to provide input parameters related to location, year, measure, gender, cause, and PM concentration for predicting disease burden.
6. **Displaying Results:** Streamlit provides options to display the results of the machine learning models. Showcase the predicted disease burden, including the number of deaths, DALY, and YLL based on the user inputs. Additionally, present the results in graphical form, such as charts or graphs, to provide visual representations of the data and predictions.
7. **Deployment:** Once the application is developed, it can be deployed on a server or cloud platform to make it accessible to users. Streamlit provides built-in functionality for deployment, making it easy to share the application with others.

Through the use of Streamlit, we have created an interactive and intuitive web-based interface for our project, enabling users to explore the dataset, interact with the machine learning models, and visualize the results. This streamlines the process of data analysis, prediction, and result interpretation, making it accessible to a wider audience and facilitating informed decision-making in the context of air quality and disease burden correlation.

By harnessing the power of Streamlit, we have enhanced the user experience, enabling seamless interaction with our project and promoting the utilization of machine learning models in the domain of public health research and intervention strategies.

3 Evaluation metrics

As the target column of our project is a continuous value, for these reason we have used these below three performance metrics for evaluation.

R2 score: R2 score is a measure of how well the model fits the data. It is calculated by squaring the correlation coefficient between the predicted values and the actual values. A perfect model would have an R2 score of 1, while a model that does not fit the data at all would have an R2 score of 0.

Mean Squared Error: MSE is a measure of the average squared error between the predicted values and the actual values. It is calculated by taking the sum of the squared differences between the predicted values and the actual values, and then dividing by the number of samples. A lower MSE indicates a better fit between the model and the data.

Mean Absolute Error: MAE is a measure of the average absolute error between the predicted values and the actual values. It is calculated by taking the sum of the absolute differences between the predicted values and the actual values, and then dividing by the number of samples. A lower MAE also indicates a better fit between the model and the data.

Table 2: Performance Metrics in Machine Learning Models

Model Name	R2_score	Mean Squared Error	Mean Absolute Error
Random Forest Regressor Model	0.9624	286035.27	179.87
Decision Tree Regressor Model	0.9484	393319.50	139.01
Light GBM	0.8337	1295016.35	669.57
GradientBoostingRegresson	0.5577	3372184.01	1054.37
MLPRegressor	0.5015	4262390.08	1158.53
PolynomialRegression	0.5003	3809641.68	1075.67
Linear regression Model	0.0711	7082953.73	1787.75

According to the table 2 , the Random Forest Regressor Model has the highest R2 score, followed by the Decision Tree Regressor Model which describes that we have got the best result from the Random Forest Regressor model. The Linear Regression model has the lowest R2 score.

4 Conclusion

In conclusion, our AI project on the correlation of air quality and disease burden analysis using machine learning techniques has been a success. Through the utilization of a Random Forest Regression model, we have gained valuable insights into the effects of airborne diseases, specifically the number of deaths, disability-adjusted life years (DALY), and years of life lost (YLL), based on air quality factors.

By preprocessing the dataset and transforming non-numerical features into numeric values, we were able to effectively train the model using the 'Location', 'Year', 'Measure', 'Gender', 'Cause', and 'PM Concentration' features. The inclusion of data from 2010 to 2019 provided a comprehensive temporal perspective, enhancing the reliability and relevance of our analysis.

The trained model demonstrated solid performance, as indicated by metrics such as mean squared error (MSE), mean absolute error (MAE), and R-squared score (R2). This performance showcases the model's ability to capture complex relationships between air quality and disease burden outcomes.

Our project's findings hold significant implications for public health planning and resource allocation. The accurate prediction of disease burden metrics for upcoming years can guide proactive interventions and policies to mitigate the adverse effects of airborne diseases. By considering air quality factors, decision-makers can make informed choices to improve population health outcomes.

Moreover, our project contributes to the broader field of air quality and disease burden analysis. By employing machine learning techniques, we have unveiled hidden patterns and associations between air quality variables and disease outcomes. This data-driven approach enhances our understanding of the complex dynamics between air quality and disease burden. In summary, our project demonstrates the potential of machine learning in analyzing the correlation between air quality and disease burden. The insights gained from this project provide a foundation for evidence-based decision-making, public health interventions, and policy formulation. Ultimately, our work aims to improve the overall health and well-being of communities by addressing the challenges posed by airborne diseases and fostering a cleaner and healthier environment.