***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

Integrated Climate Change Impact Analysis

[1]T. Jalaja, [2]Dr. T. Adilakshmi, [3]S. Padma Priya, [4]P. Sivapriya
[1]Assistant Professor, Department of Computer Science and Engineering
Vasavi College of Engineering
Hyderabad, India
e-mail: jalaja.t@staff.vce.ac.in
[2]Professor & HOD, Department of Computer Science and Engineering
Vasavi College of Engineering
Hyderabad, India
e-mail: t_adilakshmi@staff.vce.ac.in
[3]Department of Computer Science and Engineering
Vasavi College of Engineering
Hyderabad, India
e-mail: spadmapriya090902@gmail.com
[4]Department of Computer Science and Engineering
Vasavi College of Engineering
Hyderabad, India
e-mail: sivakrishna.penumarthy@gmail.com

Abstract- The Integrated Climate Change Impact Analysis project aims to address the pressing issue of pollution-induced health hazards by employing a multifaceted approach. By monitoring and recording the concentrations of various pollutants such as sulphur dioxide, ammonia, carbon monoxide, nitrogen dioxide, nitrogen monoxide, benzene, toluene, xylene, PM10, and PM2.5, the project calculates Air Quality Index (AQI) levels to assess the environmental health risk with respect to time series observations. Mapping these pollutants to associated symptoms and subsequent diseases allows for the prediction of health outcomes, enabling proactive measures to mitigate mortality rates. Through classification techniques such as K-means clustering, the project determines the suitability of cities for habitation based on AQI levels, aiding in pollution reduction strategies. Additionally, machine learning algorithms including Random Forest and Gaussian Naive Bayes are employed to predict diseases from symptoms, facilitating early intervention strategies. Mortality rates are calculated using statistical methods, incorporating probability estimates of disease impact on affected populations. We also aim to categorise if a city is safe to survive by analysing the observed AQI levels. Overall, this project serves as a comprehensive tool to assess environmental health risks, guide urban planning decisions, and ultimately foster healthier living environments.

Keywords- Air Quality Index (AQI), K-means clustering, Random Forest, Gaussian Naive Bayes, Machine Learning, Classification, Urban Planning.

I. INTRODUCTION

In today's era, with the ever-increasing concerns surrounding climate change and its adverse effects on public health, there is an urgent call for comprehensive strategies to combat pollution-induced health hazards. The Integrated Climate Change Impact Analysis project stands out as a pioneering initiative aimed at addressing this pressing issue through a multifaceted approach. At its core, the project focuses on meticulously monitoring and analysing the concentrations of key pollutants present in the atmosphere, ranging from harmful substances like sulphur dioxide to fine particulate matter.

Through this meticulous monitoring process, the project endeavours to accurately quantify the environmental health risks faced by urban populations. This involves not only tracking pollutant levels but also delving into the intricate interplay between exposure to pollutants, resulting symptoms, and the development of various diseases. Leveraging sophisticated methodologies such as the calculation of Air Quality Index (AQI) levels, the project employs advanced machine learning algorithms and predictive modelling techniques to unravel these complex relationships. Moreover, the project goes beyond mere data analysis by harnessing the power of data-driven insights. By utilizing techniques such as K-means clustering for city classification based on AQI levels and ensemble learning algorithms for disease prediction, it aims to derive actionable intelligence. This intelligence informs evidence-based urban planning decisions aimed at fostering healthier and more sustainable living environments for urban dwellers.

***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

Through its holistic approach, the Integrated Climate Change Impact Analysis project aspires to pave the way for proactive interventions. By reducing mortality rates attributed to pollution-induced diseases and catalysing meaningful strides towards a cleaner, healthier future for all, the project seeks to create lasting positive impacts on both public health and the environment.

II.RELATED WORK

[1] The authors have studied Environmental Science, focusing on pollution's impact on air and health, analysing sources and effects. Their aim is to understand complex interactions for sustainable solutions.

[2] In the realm of Public Health, the authors meticulously have examined various initiatives aimed at addressing the health repercussions of air pollution, identifying nuanced health risks and proposing targeted interventions to safeguard vulnerable communities while highlighting areas ripe for further exploration.

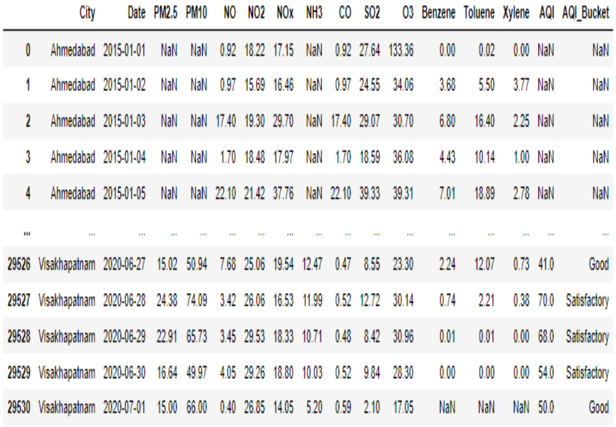
[3] Employing state-of-the-art Data & Machine Learning methodologies, the authors meticulously analyse vast sets of air quality data, harnessing the power of predictive models to foresee emerging patterns and identify regions facing elevated pollution risks. Through this meticulous process, they uncover intricate correlations between pollutants, human health, and environmental factors, enabling the development of targeted interventions and policy frameworks. By proactively addressing these insights, they endeavour to not only mitigate the immediate health consequences of pollution but also catalyse systemic changes that lead toward a future characterized by cleaner air and improved public health outcomes.

III.DATASET

A. Dataset Information

The datasets used are Spanning five years (2015-2020) and 26 Indian cities, with 29530 tuples with respect to time series observations of AQI levels and the harmful gases percentage in the respective cities. It uses gas and AQI data, alongside models predicting disease (based on 121 symptoms for 41 diseases) and mortality rate (considering disease prevalence, age, BMI, and gender). By integrating these, the project aims to understand the complex link between air quality and health outcomes, ultimately informing public health interventions and policy development.

Before proceeding with analysis, the dataset underwent meticulous preprocessing to ensure consistency and accuracy. This involved handling missing data and standardizing features to optimize the performance of machine learning algorithms during training. Such rigorous preparation is essential to ensure the reliability and precision of the subsequent predictive modelling, laying a solid foundation for robust insights into air quality's impact on health.



	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Satisfactory
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Satisfactory
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Satisfactory
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

Fig 1. Dataset during Preprocessing

B. Dataset Splitting

To facilitate effective training and evaluation of our machine learning models, we adopted a standardized approach to dataset partitioning. This involved segregating the dataset into two primary subsets:

Training Set: This training subset was utilized to familiarize models with various data attributes, including pollutant concentrations, Air Quality Index (AQI) measurements, connections between symptoms and diseases, and factors influencing mortality rates. By exposing models to this diverse range of information, they were able to discern underlying patterns and relationships within the dataset. This process enabled the models to learn how different factors interacted and contributed to health outcomes, ultimately enhancing their ability to make accurate predictions.

Test Set: The testing subset, distinct from the training data, served the purpose of assessing the model's performance on unseen data. This allowed for an objective evaluation of how well the model generalized to new instances and predicted outcomes beyond its training scope.

The dataset partitioning procedure guarantees that the trained models exhibit strong generalization capabilities when faced with novel, unseen data, thereby bolstering their reliability and practicality in real-world contexts. The selected ratios for partitioning were fine-tuned to achieve an optimal equilibrium between proficient model training and rigorous evaluation.

Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

IV. PROPOSED METHODOLOGY

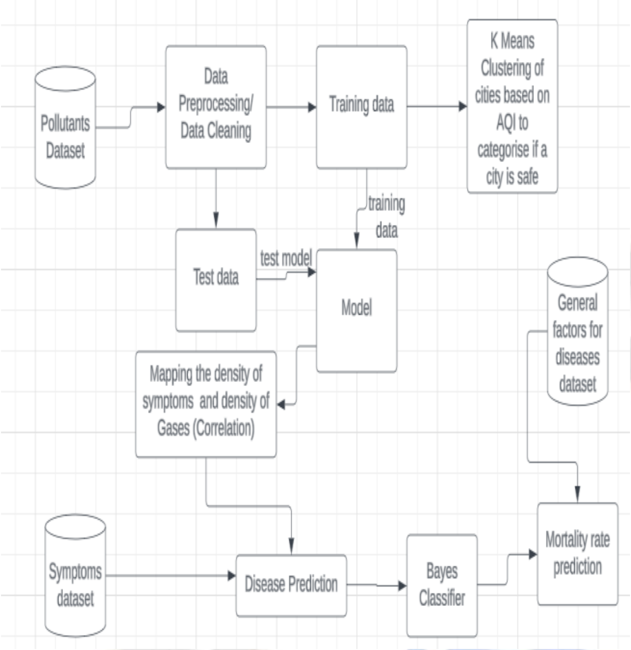

Pollutants Dataset
Data Preprocessing/ Data Cleaning
Training data
K Means Clustering of cities based on AQI to categorise if a city is safe
training data
Test data
test model
Model
General factors for diseases dataset
Mapping the density of symptoms and density of Gases (Correlation)
Symptoms dataset
Disease Prediction
Bayes Classifier
Mortality rate prediction

Fig 2. Proposed methodology architecture

Our proposed methodology delineates a holistic strategy for disease prognosis, control, and clustering leveraging sophisticated machine learning methodologies and a user-centric interface. The methodology is structured into four distinct modules, outlined as follows:

Module 1: Data Preprocessing

The initial module focuses on data preparation. We've obtained a comprehensive dataset comprising details on 41 different diseases and 121 symptoms, comprising over 29,000 records. Subsequent steps involved preprocessing the data to handle missing values, standardize features, and ensure consistency in presentation. Additionally, feature engineering techniques were applied to bolster the dataset's predictive capability.

We've utilized a mapping approach that associates symptoms with their respective severity values relative to the gases responsible for each specific disease, based on the severity of AQI levels. Each of the city is categorised into its AQI Bucket based on the AQI level observed in that city and referring to the categorisation of AQI Levels as given by WHO. These severity values are mapped to the efficacy level of individual symptoms in indicating the presence of a particular ailment.

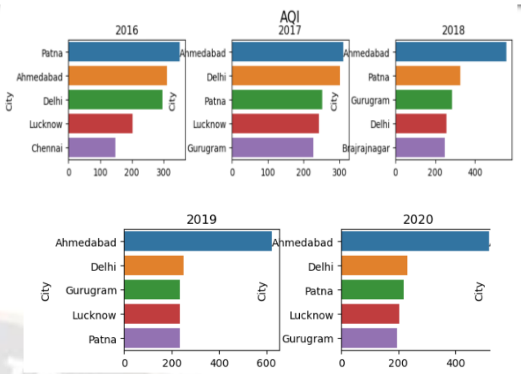


Fig 3. AQI levels of top 5 major cities in India

Module 2 : Model Building using various Algorithms

Further in this module we have used Machine learning algorithms on training dataset to increase the accuracy of the model being developed. The algorithms used include K-Means clustering, Gaussian Naïve Bayes, Random Forest , Bayes Classifier probabilistic approach algorithms.

Basic description of algorithms

1. **K-Means Clustering** : K-means clustering is an unsupervised machine learning algorithm designed to partition data into distinct groups, or clusters, based on similarities among data points. It iteratively assigns data points to clusters with the nearest centroid, aiming to minimize the within-cluster variance. However, K-means requires the number of clusters (k) to be specified in advance and is sensitive to the initial placement of centroids.
2. **Gaussian Naive Bayes** : Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the "naive" assumption of independence between features. It is particularly effective for classification tasks where the features are continuous and assumed to have a Gaussian (normal) distribution. The algorithm calculates the likelihood of a data point belonging to a particular class based on the conditional probabilities of its features given that class. Despite its simplistic assumptions, Gaussian Naive Bayes is often used in practice due to its simplicity, speed, and ability to handle high-dimensional data well, making it suitable for various classification tasks, especially with small to moderate-sized datasets.
3. **Random Forest** : Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. Each tree is built using a random subset of the training data and features, and predictions are aggregated to reduce overfitting and improve performance. Random Forest is robust, handles high-dimensional data well, and provides insights into feature importance, making

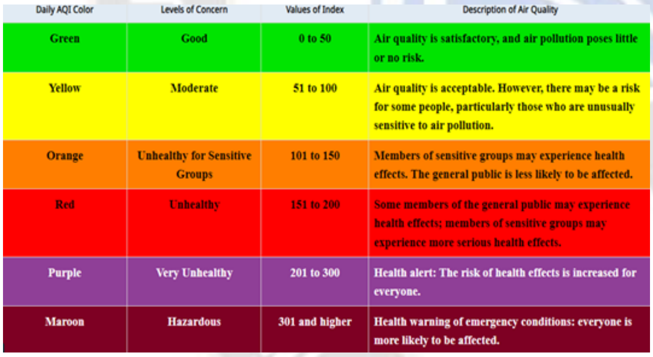
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

it a popular choice for classification and regression tasks across various domains.

4. **Bayes Classifier :** A Bayes classifier is a statistical classification model based on Bayes' theorem, which describes the probability of an event occurring based on prior knowledge of conditions related to the event. In the context of classification, the Bayes classifier calculates the probability of a data point belonging to a particular class given its features. It then assigns the data point to the class with the highest probability.The classifier makes its decisions using conditional probabilities estimated from the training data. It assumes that the features are conditionally independent given the class label, although this assumption may not always hold true in practice. The Bayes classifier is often referred to as the optimal classifier when the underlying assumptions are met, meaning it minimizes the misclassification rate.

Module 3 : Group cities using K-Means Clustering

We aim to group cities based on their air quality, represented by the Air Quality Index (AQI), using K-means clustering. The goal is to predict the AQI bucket for each city, which categorizes cities into different levels of air quality. The below is the AQI levels chart given by WHO :



Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Fig 4. AQI Levels classification as given by WHO

In this module,, we tend to group cities based on their air quality using K-means clustering to predict the Air Quality Index (AQI) bucket for each city. Here's a brief outline of the process:

1. **Data Preparation**: Organize AQI data for different cities over time.
2. **Feature Selection**: Choose relevant features such as pollutant concentrations.
3. **Normalization**: Scale features to ensure comparability.
4. **Determining K**: Decide on the number of clusters using methods like the elbow method.
5. **Clustering**: Apply K-means to group cities based on similar air quality.
6. **Interpretation**: Analyze clusters to understand air quality profiles.
7. **Predicting AQI Buckets**: Assign AQI buckets to cities based on their clusters.
8. **Evaluation**: Assess clustering performance and AQI bucket predictions.

This process helps in understanding air quality patterns across cities and aids in effective air quality management.

Module 4 : Associate the AQI levels and Gases with the disease it causes

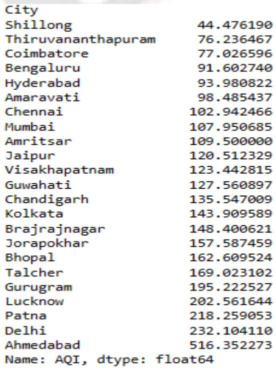
In this module we have applied classification to the AQI levels and the gases based on the densities of AQI and the density of symptoms of the respective disease with respect to the percentage of gases we are making probabilistic and statistically significant approach to know how they are associated with each other and utilizing their correlational values to understand this.

Module 5 : Associate the disease with the mortality rate

In the final module we tend to have a probabilistic approach using Naive Bayes Algorithm to understand the extent to which the disease affects one's health. Through this we are able to know the mortality rate a specific disease can cause. Furthermore, we use the confusion matrix to achieve accuracy.

V. RESULTS AND ANALYSIS

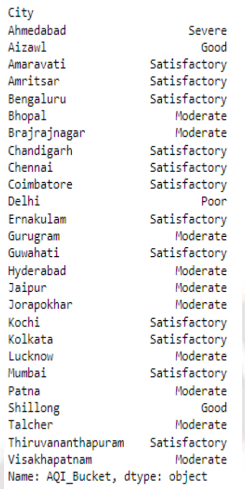
Following meticulous implementation and thorough evaluation, our system showcased commendable performance across various aspects, affirming the efficacy of our integrated methodology for city's AQI classification, disease prediction, and mortality rate. Our findings strongly validate the experience of people living in highly polluted cities being affected with so many diseases, bolstered by compelling quantitative performance metrics. These outcomes firmly establish our system as a notable and indispensable contribution to the advancing domain of reduction of pollution in India.



```
City
Shillong                44.476190
Thiruvananthapuram      76.236467
Coimbatore              77.026596
Bengaluru               91.602740
Hyderabad               93.980822
Amaravati               98.485437
Chennai                102.942466
Mumbai                 107.950685
Amritsar               109.500000
Jaipur                 120.512329
Visakhapatnam          123.442815
Guwahati               127.560897
Chandigarh             135.547009
Kolkata                143.909589
Brajrajnagar           148.400621
Jorapokhar             157.587459
Bhopal                 162.609524
Talcher                169.023102
Gurugram               195.222527
Lucknow                202.561644
Patna                  218.259053
Delhi                  232.104110
Ahmedabad              516.352273
Name: AQI, dtype: float64
```

Fig 5. AQI Levels of the Indian Cities

Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*



```
City
Ahmedabad                    Severe
Aizawl                         Good
Amaravati              Satisfactory
Amritsar               Satisfactory
Bengaluru              Satisfactory
Bhopal                     Moderate
Brajrajnagar               Moderate
Chandigarh             Satisfactory
Chennai                Satisfactory
Coimbatore             Satisfactory
Delhi                          Poor
Ernakulam              Satisfactory
Gurugram                   Moderate
Guwahati               Satisfactory
Hyderabad                  Moderate
Jaipur                     Moderate
Jorapokhar                 Moderate
Kochi                  Satisfactory
Kolkata                Satisfactory
Lucknow                    Moderate
Mumbai                 Satisfactory
Patna                      Moderate
Shillong                       Good
Talcher                    Moderate
Thiruvananthapuram     Satisfactory
Visakhapatnam              Moderate
Name: AQI_Bucket, dtype: object
```

Fig 6 . Classification of a city into its AQI_Bucket

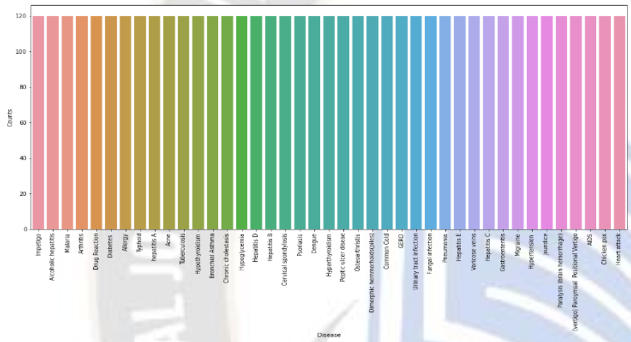


Fig 7. Disease vs Symptoms count for 120 samples (Balanced dataset)

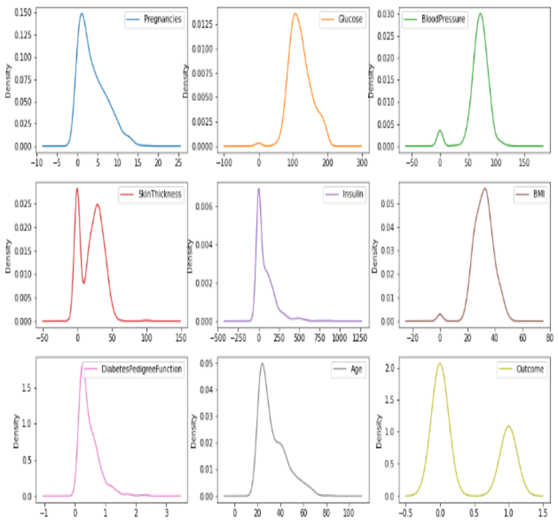


Fig 8. Density vs AQI for the Indian city Hyderabad

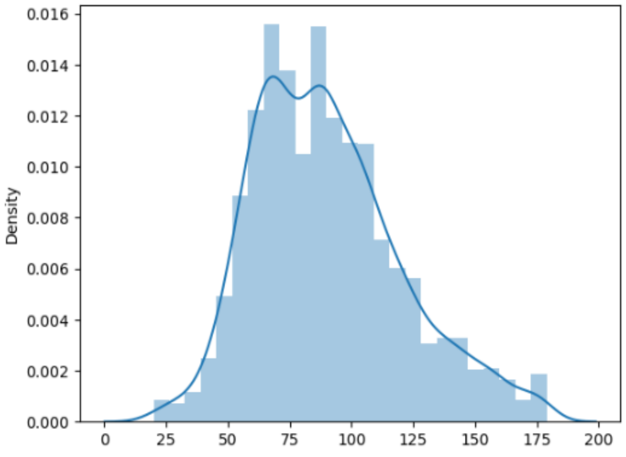


Fig 9. Density levels of a few Symptoms plotted as a graph

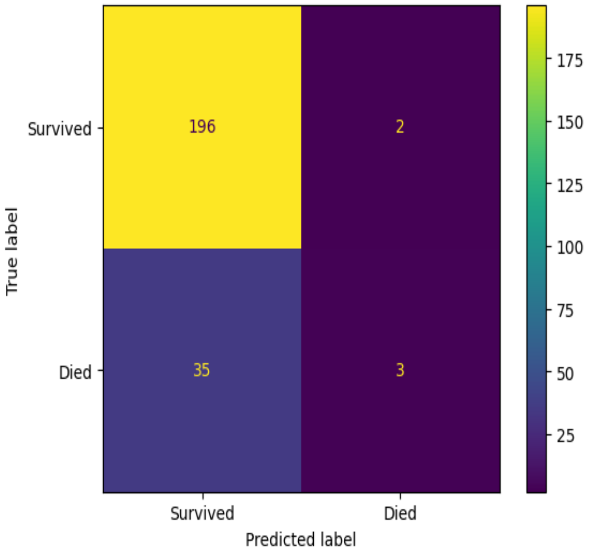


Fig 10. Confusion Matrix data for the city Hyderabad for Cardiovascular disease Mortality rates

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, while our project has successfully showcased the effectiveness of an integrated approach to climate change impact analysis and disease prediction, it's essential to acknowledge certain limitations and outline future directions for improvement. One limitation is the availability and quality of data. Despite leveraging various machine learning techniques, our models may have been constrained by the scope and quality of the datasets used. Future research could focus on incorporating more diverse and comprehensive datasets to enhance the robustness and generalizability of the models. Additionally, the complexity of environmental and health interactions presents challenges in accurately capturing all relevant factors.

Future efforts could explore more sophisticated modeling techniques or interdisciplinary approaches to better account for these complexities. Moreover, while our project highlights the importance of proactive interventions for pollution mitigation and public health management, the practical implementation of such interventions may face barriers such as regulatory constraints or resource limitations. Collaboration with stakeholders and policymakers will be crucial in overcoming these challenges and translating research findings into actionable policies.

Looking ahead, future research could also explore the integration of emerging technologies such as Internet of Things (IoT) devices or satellite imagery for real-time monitoring of environmental parameters, thus enabling more timely and targeted interventions. Overall, while our project lays a solid foundation for integrated climate change impact analysis and pollution mitigation, there is still ample room for improvement and further innovation to address the pressing challenges posed by environmental pollution and its associated health impacts.

REFERENCES

[1] Md. Fahim, Md. Ekthiar Uddin, Rizve Ahmed, Md. Rashedul Islam, Nadeem Ahmded (2022) “A Machine Learning Based Analysis Between Climate Change and Human Health”.

[2] Arup Mohanty (2021) ”Impact of climate change on human health and agriculture in recent years”.

[3] L.R. Akshaya Deepa, N. Praveen, (2011) ‘’ Impact of Climate change and adaptation to green technology in India”

[4] Yi Ge, Yiwei Xu, Jing Shi, Bingjie Li, Zesesn Li, Xiaoyan Hu, ‘’A Review of the Impact of Long-term Climate Change”, 2023

[5] M S Bhumika, Niyam Momaya, Rohit Nandan, K Suhas, Shikha Tripathi " Effect of Climate Change using Predictive Models with Remote Sensing Data " 2023

[6] William W. Kellogg, Stephen H. Schneider "Global Air Pollution and Climate Change", 1978

[7] Jagan Mohan Reddy Danda, Kumar Priyansh, Hossain Shahriar, Hisham Haddad “ Predicting Mortality Rate Based On Comprehensive Features of Intensive Care Unit Patients”, 2022.

[8] Vikram Bali, Deepthi Aggrawal , Sumit Singh, Arpit Shukla “ Life Expectancy Prediction and Analysis using ML”, 2021