**Project Proposal**

**DATA 245 - Machine Learning Technologies**

**Predicting Air Quality for Health Risk Assessment: A Machine Learning Approach**

**Divya Neelamegam, Padhma Cebolu Srinivasan, Poojitha Venkat Ram, Shruti Badrinarayanan, Sourabh Suresh Kumar**

1. **Abstract**

This project endeavors to construct a versatile and powerful hybrid model that combines Time Series Analysis with Health Risk Assessment to comprehensively study air quality trends over the years and their associated health implications. The project encompasses several key phases, beginning with the collection and preprocessing of air quality data. Through Time Series Analysis techniques, we aim to unveil intricate patterns and trends within the air quality index (AQI) data by employing Autoregressive Integrated Moving Average (ARIMA) modeling. It will help gain valuable insights into historical air quality trends. This knowledge will enable informed decision-making and proactive measures to address air quality challenges. Furthermore, Support Vector Machines (SVM) will be leveraged for classification and prediction tasks related to air quality assessment, enhancing our understanding and forecasting of air quality, ultimately contributing to improved public health.

The project also integrates some ensemble methods, specifically XGBoost, which enhances model stability and predictive accuracy. To validate the models' robustness and generalization performance, K-fold Cross validation will be employed, introducing variability in the data splits to ensure their reliability. This analysis will facilitate a deep understanding of historical variations, seasonal influences, and long-term trends in AQI. Following that, we will integrate a Health Risk Assessment module that empowers users to access health risk evaluations derived from the data and prevailing air quality conditions. By harnessing these machine learning models, this project will empower individuals and communities with actionable insights to mitigate health risks associated with air pollution. Robust visualization tools will aid in the interpretation of results, while ensuring ease of usability. This project not only aims to enhance our comprehension of air quality dynamics but also to provide a valuable resource for informed decision-making and improved public health in the face of air pollution challenges. Comprehensive documentation of data sources, analysis methods, and health risk models will ensure the credibility and transparency of this project.

1. **Motivation**

Air quality is a compelling and non-negotiable concern, profoundly influencing human health, ecosystems, and our collective well-being. Poor air quality, laden with harmful pollutants, contributes to a range of severe health issues, from respiratory diseases to premature mortality, while also imposing substantial economic burdens. Additionally, it jeopardizes environmental sustainability, fueling climate change and biodiversity loss. Clean air enhances the quality of life, fosters public awareness of environmental issues, and encourages responsible practices. Monitoring and improving air quality are essential for safeguarding both human health and the planet, compelling us to prioritize cleaner technologies, sustainable energy, and eco-friendly policies for a healthier and more sustainable future. The vigilance and enhancement of air quality are pivotal for the protection of both human health and the Earth itself. Through these actions, we endeavor not only to ensure a healthier future for ourselves but also to safeguard the planet for generations yet to come, illustrating that air quality is not just a motivation—it is an imperative which drives us to prioritize cleaner technologies, advocate for sustainable energy solutions, and champion environmentally conscious policies.

Clean air is not an abstract aspiration; it is a tangible and shared commitment to our well-being and the preservation of our planet for the benefit of generations to come.

1. **Literature Survey**

# K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara and B. Kuhaneswaran (2021), "***Machine Learning Approach for Predicting Air Quality Index***," 2021 International Conference on Decision Aid Sciences and Application (DASA),

# <https://ieeexplore.ieee.org/document/9682221>

# This study employs the Air Quality Index (AQI) as a metric to assess air toxicity. To forecast AQI, machine learning models, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN), have been integrated. The models underwent validation through the K-Fold cross-validation method, with a specified value of K set at 10 to enhance output accuracy compared to using a lower K value. Consequently, the prediction of AQI achieved the highest level of accuracy in this study.

# D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, Worku Abera, "***Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning***",

# <https://doi.org/10.1155/2022/5086622>

# This study anticipates PM2.5 pollutants in the atmospheres of different cities. It employs diverse machine learning models, including linear regression, KNN, ridge and lasso, random forest, Ada Boost, and XGBoost for the prediction task. The evaluation of these models incorporates metrics such as MAE, MAPE, and RMSE, enhancing the reliability of the obtained results.

1. N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, G. (2023) Arulkumaran, "***Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis"***, <https://doi.org/10.1155/2023/4916267>

The objective of this study is to develop a reliable prediction model for the Air Quality Index (AQI) to support climate control efforts. The research employs the Synthetic Minority Oversampling Technique (SMOTE) and incorporates three different approaches: Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR) for AQI determination. The outcomes reveal that CatBoost and Random Forest, in conjunction with SMOTE, yield highly accurate predictions for air quality.

1. Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, Leonardo Vanneschi, (2020) ***"A Machine Learning Approach to Predict Air Quality in California"***,

<https://doi.org/10.1155/2020/8049504>

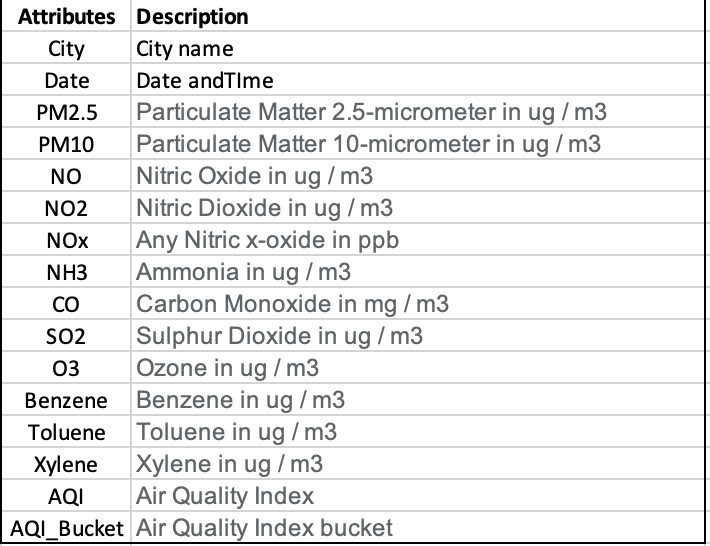
This paper asserts that predicting air quality is a challenging endeavor owing to its dynamic nature. Among the top-performing models for air quality prediction in California, support vector regression (SVR) is employed to forecast pollutants and predict the Air Quality Index (AQI). In contrast to previous studies, this paper introduces the use of radial base function, enabling SVR to achieve the most precise predictions.

1. **Methodologies**

i) Data Collection:

The data has been made publicly available by the Central Pollution Control Board, https://cpcb.nic.in/ which is the official portal of Government of India. The data was gathered between 15 January 2015 and 1 July 2020 and includes the hourly and daily AQI and air quality values for various Indian cities. There are 737408 rows of instances for each of the 16 attributes in this dataset. The dataset includes elements that are both internal and external to this investigation. These features include city, date, time, the concentration of PM2.5, PM10, NO, NO2, NOx, NH3, SO2, O3, CO, Benzene, Toluene, Xylene, AQI and AQI\_BUCKET. Except for NOx and CO, which were measured in ppb and mg/m3, the dataset measured PM levels and other pollutant data in micrograms per cubic meter (g/m3).

***Figure 1: Metadata***



ii) Data Cleaning and Feature Selection:

To assure the quality and dependability of the dataset, data preprocessing will comprise dealing with missing values, using imputation techniques when appropriate, and identifying and treating outliers.

In addition, based on an improvement in performance, Principal Component Analysis (PCA) or analogous feature selection techniques will be used.

iii) Proposed Models:

1. ARIMA - To gain valuable insights into historical air quality trends, enabling informed decision-making and proactive measures to address air quality challenges, Autoregressive Integrated Moving Average (ARIMA) modeling technique will be employed, a well-established and widely used method for time series analysis. ARIMA models will enable us to assess the presence of trends, seasonality, and other temporal patterns in air quality data.
2. Support Vector Machines (SVM)- SVM will be leveraged for classification and prediction tasks for air quality assessment. By harnessing SVM's discriminative power, this research aims to enhance our understanding and forecasting of air quality, contributing to improved public health.
3. XGBoost- XGBoost, an advanced machine learning algorithm, will serve as the cornerstone of our project aimed at classifying and predicting air quality.By leveraging XGBoost, we intend to develop a robust model for classifying air quality into meaningful categories and making accurate predictions.

iv) Validation Techniques:

K-fold Cross validation will be employed to assess the models’ generalization performance, introducing variability in the data splits to validate their robustness.

1. **Deliverables**
2. Data Acquisition (October 5): Have all necessary data, cleaned, and ready for analysis.
3. Time Series Model Development (October 10): Implement ARIMA for time series forecasting of air quality parameters.
4. Model Development (October 15): Develop robust machine learning models to select and incorporate relevant features.
5. Combined Model Integration (October 20): Integrate the ARIMA and chosen machine learning model to generate a comprehensive risk assessment model.
6. Intermediate Status Report (November 2)
7. Health Risk Assessment (November 3): Classify health risk indices based on the combined air quality predictions.
8. Conclusion, Results and Visualization (November 5): Document the project's findings through visualizations and analysis.
9. Slides, Report/Documentation and Technical Paper (October 10 - November 11): Prepare the project report and a technical paper summarizing the project for potential submission to a journal or conference. (Continuously updated from the beginning of the project)
10. Journal Submission (Post-project)
11. Healthy Team Collaboration: Throughout the project, we will document our work thoroughly, maintain a clean and organized codebase, and conduct regular meetings or checkpoints to review progress and address any challenges.
12. **Team Members and Roles**
13. Data Acquisition (October 5):
    1. Team Members: All
    2. Responsibility: Collect and clean all necessary data, ensuring it is ready for analysis.
14. Time Series Model Development (October 10):
    1. Team Members: Divya Neelamegam
    2. Responsibility: Implement the ARIMA model for time series forecasting of air quality parameters.
15. Model Development (October 15):
    1. Team Members: Padhma Cebolu Srinivasan
    2. Responsibility: Develop a robust machine learning model to select and incorporate relevant features from the data.
16. Combined Model Integration (October 20):
    1. Team Members: Poojitha Venkat Ram
    2. Responsibility: Integrate the ARIMA model with the machine learning model to create a comprehensive risk assessment model.
17. Intermediate Status Report (November 2):
    1. Team Members: All
    2. Responsibility: Collaboratively prepare an intermediate status report outlining the project's progress and key findings.
18. Health Risk Assessment (November 3):
    1. Team Members: Shruti Badrinarayanan
    2. Responsibility: Implement the classification of health risk indices based on the combined air quality predictions.
19. Conclusion, Results, and Visualization (November 5):
    1. Team Members: All
    2. Responsibility: Collaboratively document the project's findings through visualizations and in-depth analysis.
20. Slides, Report/Documentation, and Technical Paper (October 10 - November 11):
    1. Team Members: All (Continuously updated from the beginning of the project)
    2. Responsibility: Collaboratively work on the project report, documentation, and technical paper summarizing the project for potential submission to a journal or conference.
21. Journal Submission (Post-project):
    1. Team Members: Sourabh Suresh Kumar
    2. Responsibility: Lead the effort to prepare and submit the project's technical paper to a relevant journal.
22. **Rubric Criteria**

The conditions of the rubrics were met in the following ways:

1. Abstract: Briefly explained the project's goals, ML approaches, and projected impact
2. Motivation: Presented a strong argument for the importance of researching air quality with the goal to safeguard human health and the environment, providing a crucial context.
3. Literature Review: Covered related work using ML models like XGBoost, SVR for forecasting air quality. Provided an insightful historical context.
4. Methodology: Logical steps were outlined, such as model validation, model validation using SVM, ARIMA modeling, and data cleaning. Technical choices that make sense.
5. Deliverables: Project is divided into manageable phases with deadlines and it promotes accountability.
6. Team Roles: To promote teamwork, each team member is assigned specific roles during various phases.
7. Relevance: Use and application of time series, classification, ensembling techniques to a real-world problem
8. Technical Difficulties - Big data preprocessing, parameter tuning, and interpreting complex models could pose challenges.
9. Novelty: Uniquely integrated ARIMA with SVM, XGBoost for air quality forecasting and health risk modeling.
10. Impact: Models could inform health policy, urban planning, raise awareness on air quality issues.

**8. Others**

As we progress through the project, we may adapt and refine our approach to ensure optimal results. This includes an ongoing evaluation of our chosen methodologies and techniques.

Additionally, if time permits, we would also like to explore the development of respiratory diseases caused by poor air quality, seeking correlations between air pollution and health outcomes.

Furthermore, we plan to leverage advanced machine learning models (LLMs) to enhance our project's predictive capabilities and gain deeper insights from the data.