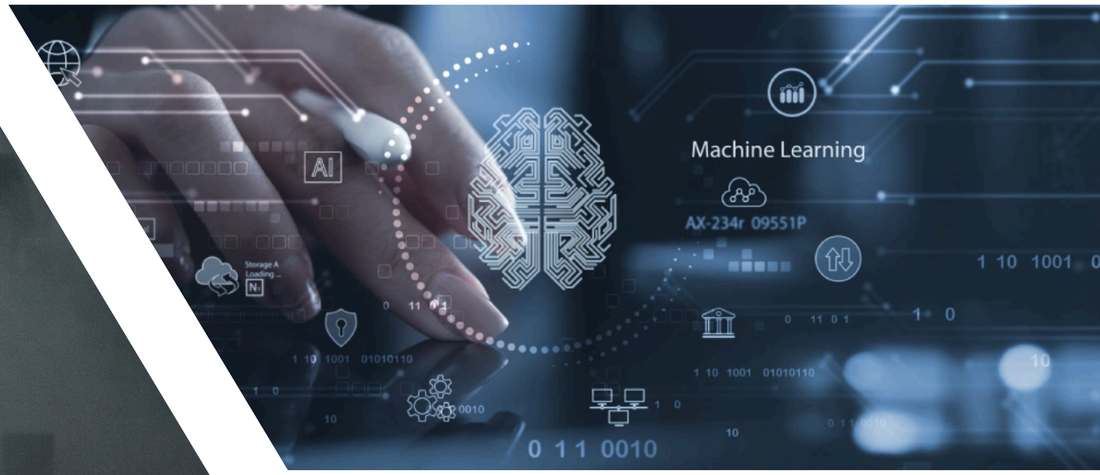
AIR QUALITY AND PUBLIC HEALTH IMPACT

PREDICTING HEALTH IMPACT SCORE

NIROSHAN.K
UKI STU 895

INTRODUCTION

- Air pollution is a major global issue, negatively impacting public health.
- Poor air quality can lead to breathing and heart diseases, more hospital visits, and early deaths.
- This project aims to identify the main causes of these health issues, assess their impact on health, and develop a model to predict the Health Impact Score.



METHODOLOGY

1. DATA ACQUISITION AND INITIAL OBSERVATIONS

- This project utilizes Kaggle's "Air Quality and Health Impact" dataset.
- The dataset comprises 5,811 records with 15 features, including PM10, PM2.5, NO2, SO2, O3 concentrations, as well as temperature, humidity, and wind speed.
- The target variable is the Health Impact Score (0-100).

PM2.5 and PM10 are small particles that can cause diseases.

2. DATA PREPROCESSING

01

HANDLING MISSING VALUES

The dataset was checked for missing values [No missing and duplicate values]

02

REDUCE SKEWNESS

Box-Cox transformation was used to correct skewness

03

REMOVING OUTLIERS

Outliers were identified and removed using the Interquartile Range (IQR) method

04

ADDRESSING CLASS IMBALANCE

Oversampling was used to balance the class distribution

05

FEATURE SCALING

Standard scaling was applied to normalize the features

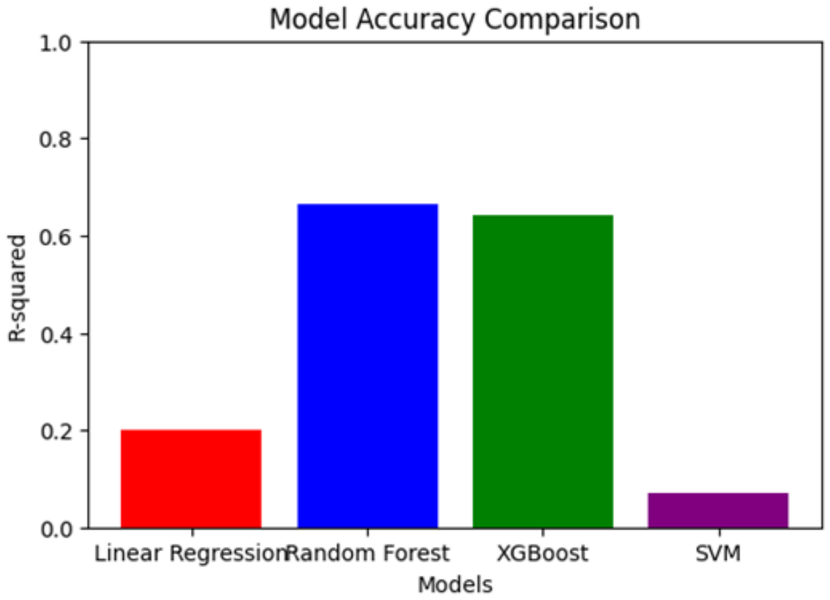
06

FEATURE SELECTION

The top 8 features were selected using the Recursive Feature Elimination (RFE)

3. MODEL TRAINING AND EVALUATION

- Various machine learning models, including Linear Regression, Random Forest, XGBoost, and SVM, were trained and tested.
- The models were compared based on metrics such as Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE).

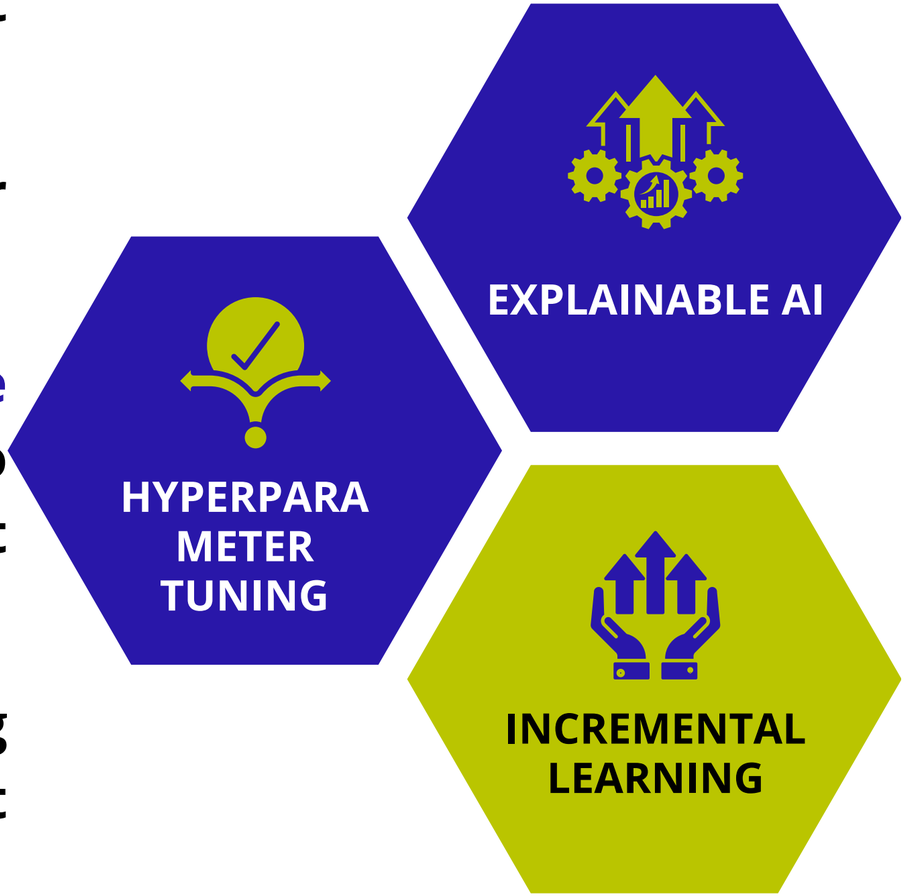


MODEL	R^2	MAE	MSE
RANDOM FOREST	0.6637	6.65	94.68
XG-BOOST	0.6416	6.68	100.9
SVM	0.0711	11.10	261.8
LINEAR REGRESSION	0.2012	12.15	225.1

Random Forest gave the best results with an R^2 of 0.6637 because it can handle complex patterns in the data.

4. OPTIMISATIONS

- The Random Forest model was chosen as the best based on its performance.
- RandomizedSearchCV was used for hyperparameter tuning.
- Explainable AI was explored using Partial Dependence Plots (PDPs) and Feature Importance analysis to visualize how different features affect the target variable.
- Additionally, incremental learning was explored using the SGDRegressor model, allowing the model to adapt to new data and improve over time.



KEY FINDINGS

1. PM2.5 is the biggest contributor to health issues (32%), followed by O_3 (25%) and NO_2 (18%), emphasizing their impact on public health.

2. Higher humidity and lower wind speed worsen pollution effects by limiting pollutant dispersion.

3. Reducing PM2.5 and O_3 emissions through pollution controls can significantly lower the Health Impact Score (HIS).

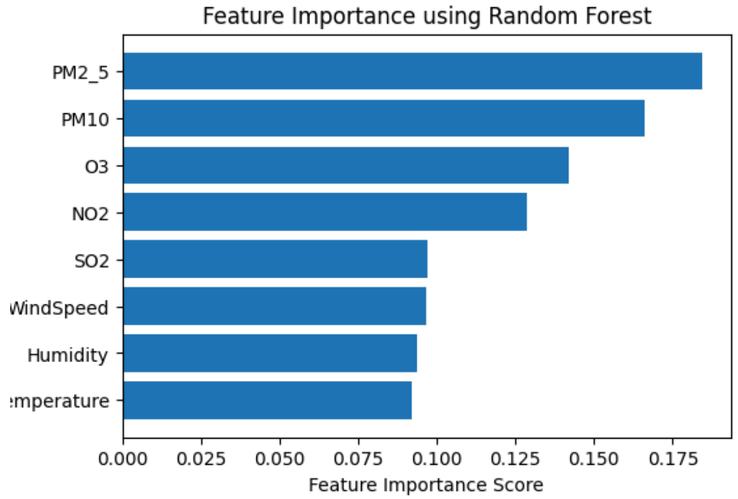
RESULTS AND DISCUSSION

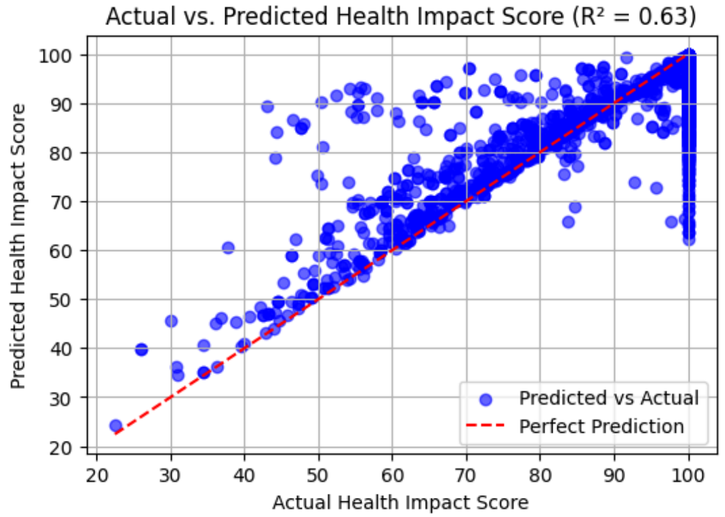
The Random Forest model with tuned hyperparameters achieved the best performance, with an R-squared score of [0.6827].

The PDPs and Feature Importance analysis revealed the complex relationships between air quality, weather conditions, and health impacts.

The incremental learning approach showed promising results in adapting the model to new data

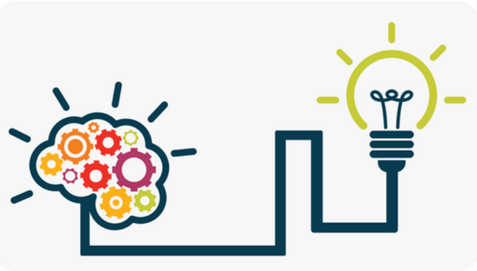
This model can be deployed for real-time air quality monitoring





CONCLUTION

This study successfully used data science techniques to predict health impacts from air quality data. The analysis identified PM2.5 as the dominant factor, meaning reducing it can significantly improve public health.



Future work includes integrating advanced techniques and developing a real-time dashboard for pollution alerts and health advisories.

This model can be deployed for real-time air quality monitoring, providing policymakers and health professionals with valuable data-driven insights to mitigate air pollution risks.

THANK YOU

Thank you for your time and attention today.