



**ANALYSIS REPORT**

**Regression Analysis Report**

1. Overview   
The purpose of this analysis is to use regression to predict a target variable. The Global Risk Report   
  
A range of features are included in the dataset to predict vulnerability. Data preprocessing, EDAs, model development, cross-validation, hyperparameter tuning, and feature selection are all included in this analysis.

**2. Dataset Description**

**Dataset Name: Air pullulation dataset**

**Source:** Kaggle

3. Preprocessing Data   
Data cleaning involved dropping missing values. Since there weren't many, they were dropped.   
  
absent values. IQR was used to handle outliers.   
Feature Transformation: Label encoder was used to encode categorical variables while Standard Scaler was used to standardize numerical features.

**3. Analysis of Exploratory Data (EDA)   
To comprehend the distribution of features and identify patterns, histograms for every column, box plots, bar charts, line graphs, scatter plots, and heat maps were made.   
  
Findings: The correlation heatmap demonstrated a high association between exposure and air pollution.**

**4. Model Building**

**Linear Regression from scratch**

**KNeighborsRegressor**

**5. Model Evaluation**

**Metrics of performance:   
R-squared: Indicates the percentage of the target variable's variance that the model can account for.   
The gap between expected and actual values is measured by the mean absolute error, which has a lower   
  
values that show improved model performance.   
The average squared difference between the actual and anticipated values is measured by the mean squared error, or MSE.**

Linear Regression:

Train Metrics (sklearn):

Mean Squared Error: 80.6993

Mean Absolute Error: 4.8432

R-Squared: 0.9742

Test Metrics (sklearn):

Mean Squared Error: 76.9413

Mean Absolute Error: 4.7763

R-Squared: 0.9759

KNeighborsRegressor

Train Metrics (KNN):

Mean Squared Error: 15.4685

Mean Absolute Error: 1.1690

R-Squared: 0.9951

Test Metrics (KNN):

Mean Squared Error: 24.5022

Mean Absolute Error: 1.5066

R-Squared: 0.9923

Final model:

Final Model: KNN Regressor with Selected Features

Mean Absolute Error: 0.7635485474820655

Mean Squared Error: 12.568648341501524

R-squared: 0.9960558342264716

6. Adjusting Hyperparameters   
The parameters of Ridge Regression and Linear Regression were adjusted using GridSearchCv.

Limitation:

Limitations include the potential for overfitting with KNN due to its sensitivity to the choice of k and distance metrics. Both models (Linear Regression and KNN) assume linear relationships and might struggle with complex, non-linear data. Additionally, feature selection and preprocessing steps may need further refinement for optimal results.

Future work

Future work could involve exploring advanced models like Random Forests or Gradient Boosting to improve accuracy. Additionally, hyperparameter tuning, cross-validation, and incorporating more feature engineering techniques can further enhance model performance. Implementing model interpretability tools and deploying the model for real-time predictions could be valuable next steps.