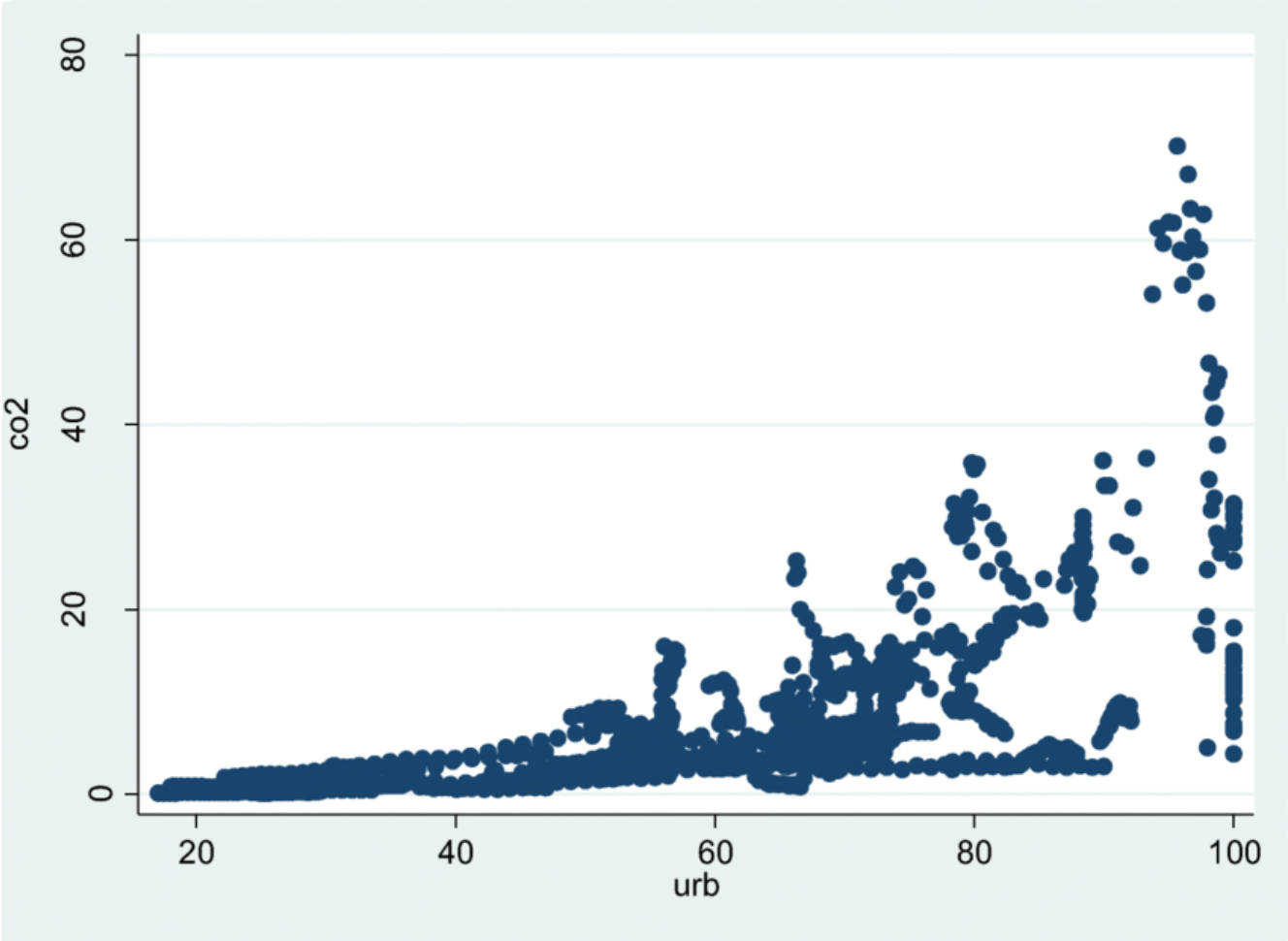
REPORT OF GHG EMISSION -2023

MAY 18

Imarticus Learning
Guided By :Mr. Arun U
Presented by:Mrs S.A.NOOHA NASHEERIN

OUR SOURCE DATA .

- **X coordinates (km)**
- **Y coordinates (km)**



- **Measured Depth (m)** –Measuring of c02 emission
- **Deviation (deg)** –measuring of gas which deviated
- **Abandoned (True/False)** –Abandoned oil and gas which threatening lives
- **Surface-Casing Weight (kg/m)** –The greenhouse gas emission increase the heat temperature of earth surface
- **Production-Casing Size (mm)** –The inner casing that is placed all the way down to the bottom of the well,thus seprating the producing zone from the other formation layer
- **Cumulative GAS Prod. (e3m3)** – The total amount of oil and gas recovered from a reservoir in a particular time
- **Month Well Spudded** –The process to drill a well in the oil and gas industry

- **Classification** –Global warming often described as the most recent Climatic changes.
- **Emission Rate (m3/day)** –The total amount

Here I discuss my machine learning capstone project I applied various algorithms classification to identify the most reliable algorithm that depicts the highest performance on both training and test data and that can be considered for the future dataset.

Dataset

The data file "GHG_Emission.csv" has been retrieved from AER website; where the locations of the wells have been changed, and some key properties are generated synthetically or are greatly manipulated for confidential reasons.

Regression and Classification

Gathering Data

First, the dataset was imported and read using pandas. The data was shuffled and then random. seed (42) was used to save the state of a random function. The index of the data was reset.

Data Processing

Stratified sampling was performed for even distribution of data. The test and training data were split based on that. The outliers were removed for instances out of the range of $\mu \pm 2.5\sigma$, imputation (with median) was performed, text handling using one-hot encoding and standardization.

CLASSIFICATION

Model Training for Classification

Binary classification was applied using the following Machine Learning models below. • Dummy Classifier • Stochastic Gradient Descent • Logistic Regression • Support Vector Machine: Linear • Support Vector Machine: Polynomial Kernel • Decision Trees • Random Forest • Adaptive Boosting with Linear SVM • Adaptive Boosting • Hard and Soft Voting • Shallow Neural Network (with 3 layers) • Deep Neural Network (with 6 layers)

The hyperparameters were fine-tuned using RandomizedSearchCV based on accuracy. The optimized parameters were used to predict accuracy. K-fold cross-validation with 5-folds (cv=5) was applied and then the mean of 5 accuracies for each classifier was calculated. These optimized hyper-parameters for all the above-mentioned algorithms were used for finding the performance on the test dataset as well.

Model Performance for Classification

Random Forest should be used for future datasets as it gives the best performance on both testing and training data.

of pollutant emitted