Electricity Consumption For Household Appliances

ML PROJECT

GROUP 15

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

1. MOTIVATION

1. **Environmental Imperative:**
   - Rising global energy consumption threatens the environment, necessitating innovative solutions.

2. **Granular Insights for Efficiency:**
   - Conventional methods fall short in tracking household electricity, leading to inefficiencies and higher costs.

3. **Empowering Sustainable Practices:**
   - Our advanced machine-learning techniques, utilizing submeter readings and regression models, provide accurate appliance-specific consumption data, empowering homeowners to optimize energy usage, reduce costs, and contribute to a more sustainable future.

2. LITERATURE REVIEW

1. Artificial neural networks for the prediction of the energy consumption of a passive solar building by S. A. Kalogirou and M. Bojic

- Methodology: Utilized Artificial Neural Networks (ANNs) to predict energy consumption in a passive solar building.
- Model Evaluation: Dynamic thermal building model based on finite volumes and time marching used to assess thermal behavior.
- Accuracy: Trained ANN model with simulated data achieved high accuracy (R2 value of 0.9991) for unknown data.
- Computational Efficiency: The ANN model outperformed dynamic simulation programs, offering a faster tool for modeling building thermal behavior.

2. LITERATURE REVIEW

2. Linear Regression model to forecast electricity consumption in New Zealand

- This study examines the impact of economic and demographic factors on electricity consumption in New Zealand.
- Model Evaluation: This research focuses on GDP, electricity prices, and population as the key variables for predicting electricity use in New Zealand, using multiple linear regression.
- It compare the model's forecasts to national predictions and a Logistic model.
- Computational Efficiency: While the Logistic model accurately describes historical consumption patterns, it tends to be conservative in its forecasts due to saturation limits in the curve.

2. LITERATURE REVIEW

3. Power load forecasting using support vector machine and ant colony optimization by D. Niu, Y. Wang and D. D. Wu

- Technique: Utilized Support Vector Regression with Ant Colony Optimization for power load forecasting.
- ACO for Data Processing: Ant Colony Optimization employed to process and streamline large datasets, reducing training data for SVM and improving processing speed.
- Enhanced Accuracy through Data Mining: Historical daily loading data with similar meteorological features to the forecasting day used to enhance model accuracy.
- Feature Selection Mechanism: Introduced a feature selection mechanism based on ACO to find optimal feature subsets during data reduction.
- Superior Forecasting Accuracy: The proposed method outperforms single SVM and BP neural network models in terms of forecasting accuracy.

DATASET DESCRIPTION

a) DATASET ATTRIBUTES, VISUALIZATIONS AND DETAILS

b) DATASET PREPROCESSING

DATASET ATTRIBUTES

FEATURE	DATATYPE
Date	String
Time	String
Global_active_power	Float
Global_reactive_power	Float
Voltage	Float
Global_intensity	Float
Sub_metering_1	Integer
Sub_metering_2	Integer
Sub_metering_3	Integer

Attribute Information:

1.**Date**: Date in format dd/mm/yyyy

2.**Time**: time in format hh:mm: ss

3.**Global_active_power**: household global minute-averaged active power (in kilowatt)

4.**Global_reactive_power**: household global minute-averaged reactive power (in kilowatt)

5. **Voltage**: minute-averaged voltage (in volt)

6.**Global_intensity**: household global minute-averaged current intensity (in ampere)

FEATURE	DATATYPE
Date	String
Time	String
Global_active_power	Float
Global_reactive_power	Float
Voltage	Float
Global_intensity	Float
Sub_metering_1	Integer
Sub_metering_2	Integer
Sub_metering_3	Integer

7.**Sub_metering_1**: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven, and a microwave (hot plates are not electric but gas powered).

8.**Sub_metering_2**: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing machine, a tumble drier, a refrigerator, and a light.

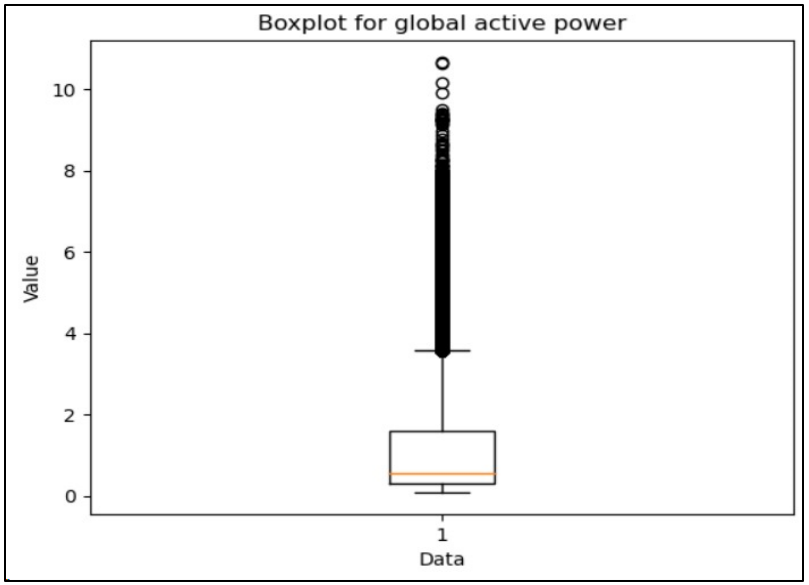
9.**Sub_metering_3**: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water heater and an air conditioner. Gas-powered

DATASET DETAILS

- This dataset tracks six months of detailed household electricity usage from January to June 2007, encompassing various metrics.
- It provides valuable insights with 260,640 measurements.
- The dataset gives electricity usage in one household from three different areas:

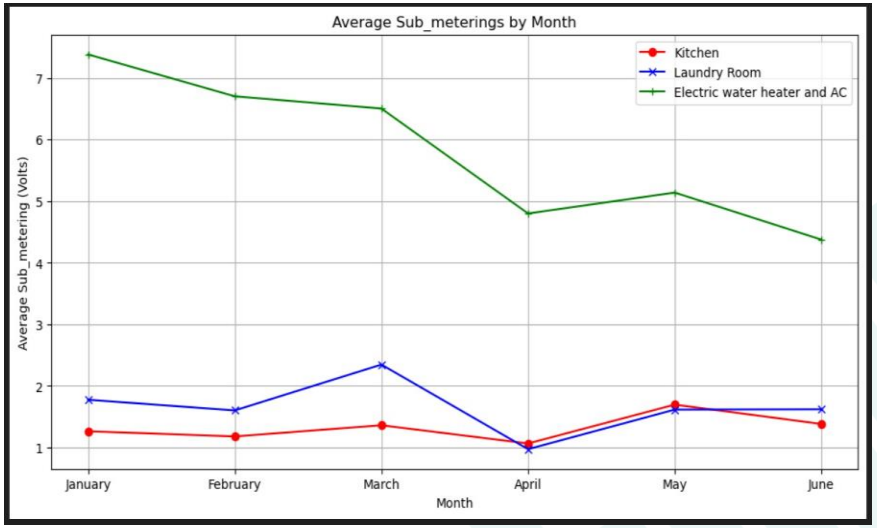
  1. **Submetering1**: Which gives the power usage for the kitchen.
  2. **Submetering2**: Which gives the power usage for the laundry room
  3. **Submetering3**: Which gives the power usage for the Electric water heater and air conditioner.

3. DATASET VISUALIZATIONS

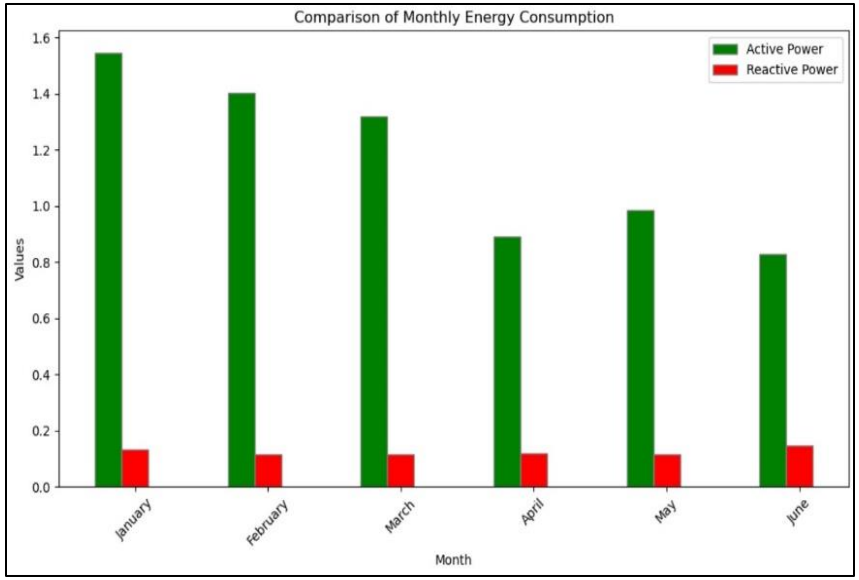


A box plot is a concise data visualization that shows the median, spread, and potential outliers of a dataset, making it easy to compare distributions across different groups. Visualization for Box plot for global active power to check for outliers.

Graph for visualizing Sub metering for every month. Average Voltage consumed by the corresponding appliances with respect to different months



3. DATASET VISUALIZATIONS



• **Active Power:** Represents actual consumed power for useful work.

• **Reactive Power:** Reflects power oscillation between source and load, crucial for maintaining electromagnetic fields but not performing work.

• **Graph:** Illustrates the power difference between active and reactive power in the dataset.

Pre-processing

1. In our dataset, we performed Label encoding to convert the 'Time' column into a numerical format ensuring compatibility with machine learning algorithms. We Implemented One Hot Encoding on 'Month new' to create a binary representation of months.
2. We have performed null checks on our dataset and dropped those entries which had a null or "?" Value.
3. In the date column, the dates were in different formats. We have converted all the dates to a single format

4. METHODOLOGY

- **FEATURE SELECTION**
  - Carefully selected relevant features for the regression task, including 'Time', 'Voltage', 'Sub metering 1', 'Sub metering 2', 'Sub metering 3', 'Month new', 'Global reactive power', 'Global intensity'.

- **TARGET VARIABLE**
  - Designated 'Global active power' as the target variable for regression.

- **DATA SPLITTING**
  - The training set com prised 80 percent of the data, serving as the foundation for model development, while the testing set constituted the remaining 20 percent, providing an independent dataset for rigorous model evaluation.

4. METHODOLOGY

MODELS

- LINEAR REGRESSION
  - Using Linear Regression for its simplicity and interpretability, we aim to model the relationship between variables to predict 'Global active power' based on selected features
- SGD (STOCHASTIC GRADIENT DESCENT) REGRESSION
  - SGD selected for efficient optimization of large-scale models with extensive dataset of over two million measurements.
- RIDGE REGRESSION
  - Ridge regression prevents overfitting in linear models by adding a penalty term. It was applied to enhance performance.
- BAYESIAN REGRESSION
  - Bayesian regression is useful in this dataset as it enables us to incorporate prior knowledge about the relationships between variables, and provides a way to assess uncertainty in parameter estimates, enhancing the reliability of our predictions.
- LASSO REGRESSION
  - Lasso Regression, like Ridge, adds a penalty term to linear regression. It balances simplicity and accuracy, automatically selecting relevant variables.

4. METHODOLOGY

• MODELS

- K – Nearest Neighbors
  - KNN is most useful when labeled data is too expensive or impossible to obtain, and it can achieve high accuracy in a wide variety of prediction-type problems.
- MLP Classifier
  - MLP is preferred over other types of neural networks because of its ability to model complex non-linear relationships between the inputs and outputs.
- Decision Trees
  - Decision Tree is a tree structure with nodes representing features, edges indicating feature values, and leaf nodes depicting the final outcome. They prove effective for regression problems, making them a suitable choice for our project.
- Random Forest Regressor
  - Random forests are a collection of several decision trees trained in similar-sized datasets made from the same original dataset using random sampling.

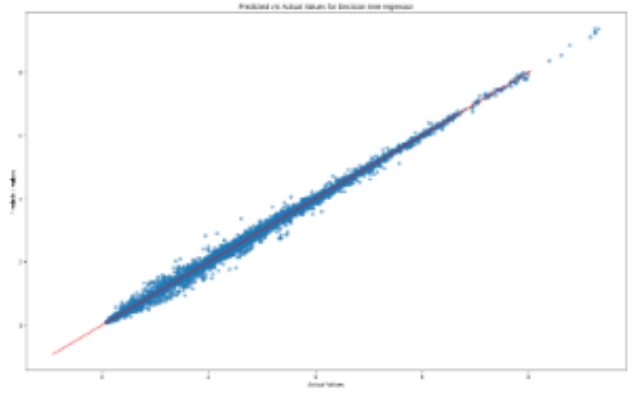
5. RESULT AND ANALYSIS

- We assessed various regression models and collected their results.
- Comparative analysis was performed utilizing standard metrics such as MSE, RMSE, and R2 Score.
- Among the array of machine learning models employed, Linear Regression emerged as the top-performing choice, a conclusion supported by both the visual representations and our calculated metrics.
- The specific metric values for Regressions model:

Model	MSE	RMSE	R2 Score
Linear Regression	0.00204	0.04518	0.99854
SGD Regressor	0.00148	0.03856	0.99851
Ridge Regression	0.00204	0.04518	0.99854
Bayesian Ridge Regression	0.00204	0.04518	0.99854
Lasso Regression	0.00237	0.0487	0.9983
Decision Tree Regressor	0.00154	0.03934	0.99889
Random Forest Regressor	0.00083	0.02894	0.9994
MLP Regressor	0.02206	0.14852	0.98423
KNN Regressor	0.01626	0.12753	0.98837

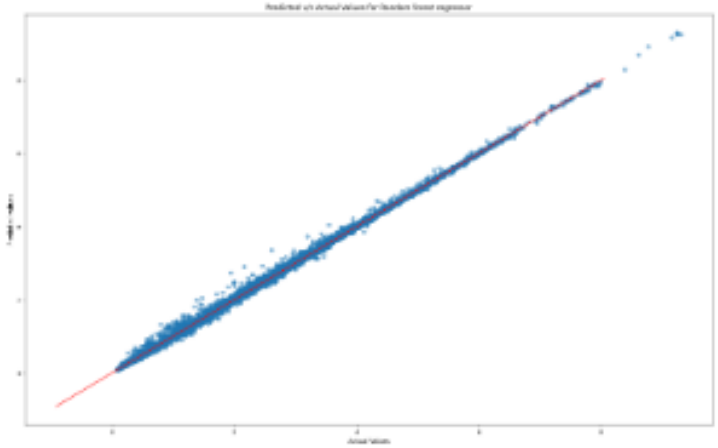
5. RESULT AND ANALYSIS

To showcase the models' performance, we've prepared the following visualizations.



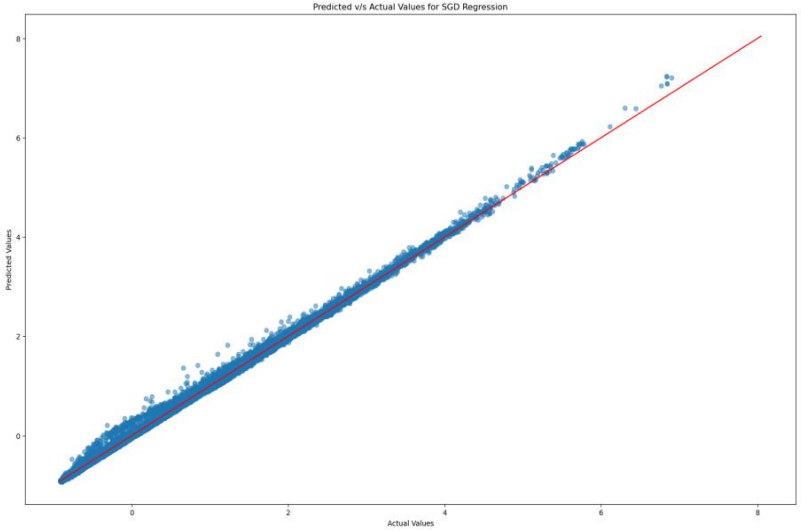
Decision Trees Regression Performance (left)

Random Forest Regression (below)



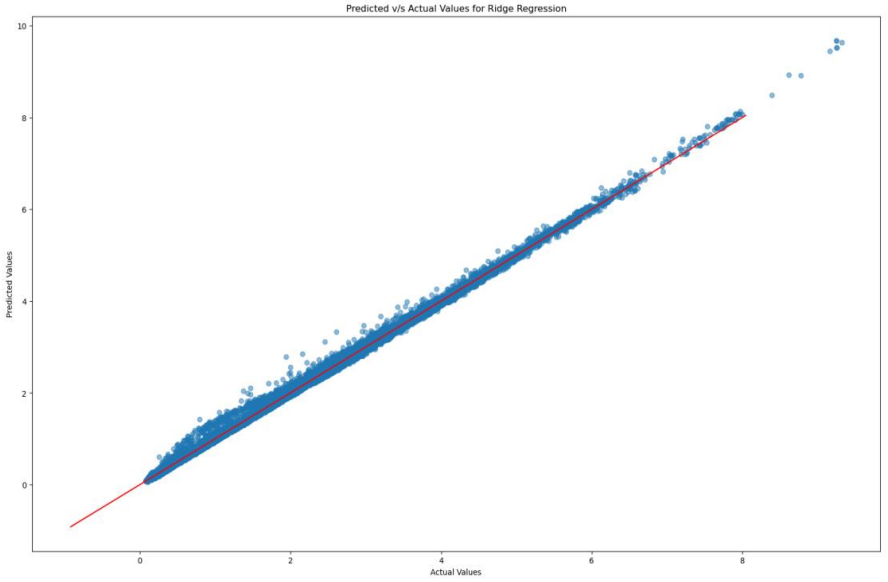
The visualizations clearly indicate a close alignment between **Predicted and Actual values for Decision trees and Random forests Regressor.** These results affirm the high efficiency of our model.

5. RESULT AND ANALYSIS

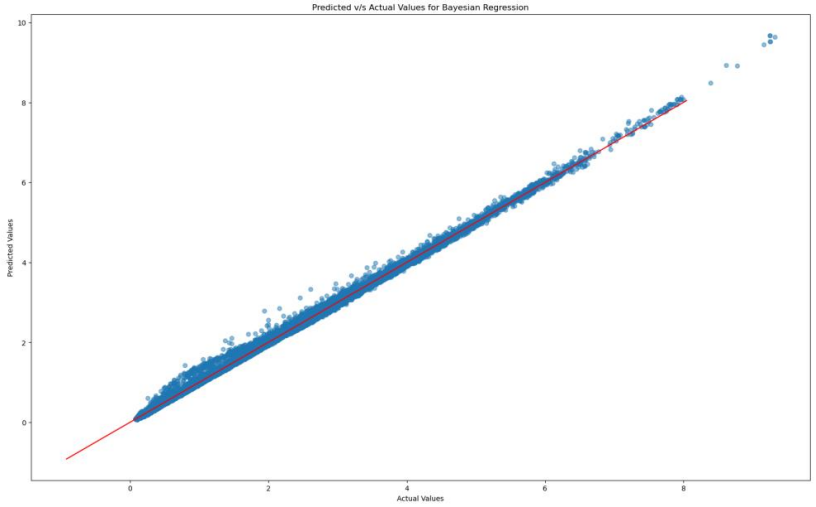


Predicted VS Actual values for SGD

Predicted VS Actual values for Ridge Regression

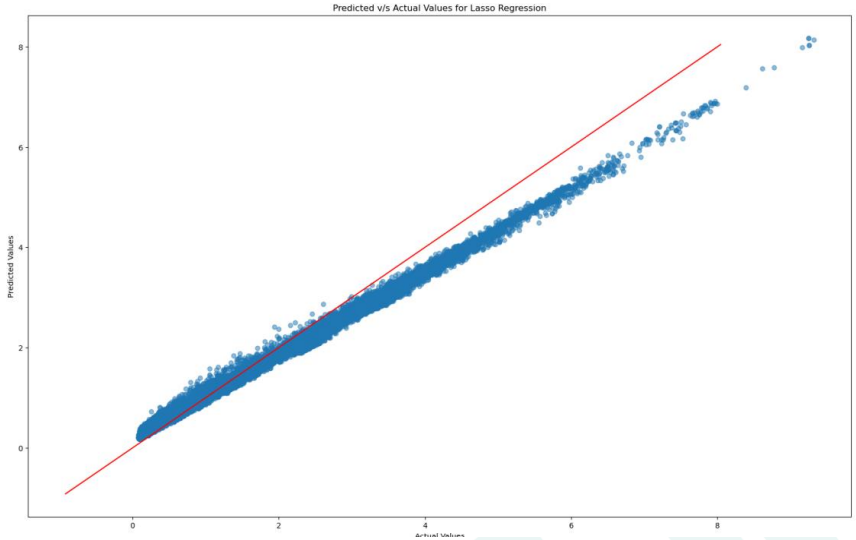


5. RESULT AND ANALYSIS

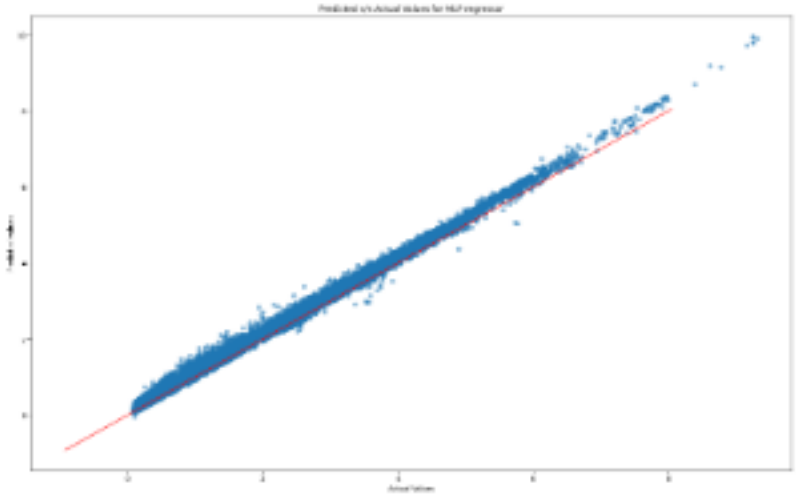


Predicted VS Actual values for Bayesian Ridge Regression

Predicted VS Actual values for Lasso Regression

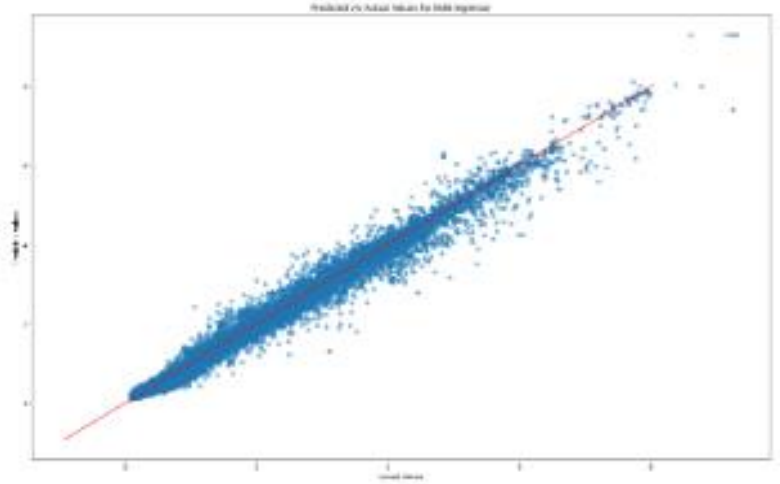


5. RESULT AND ANALYSIS

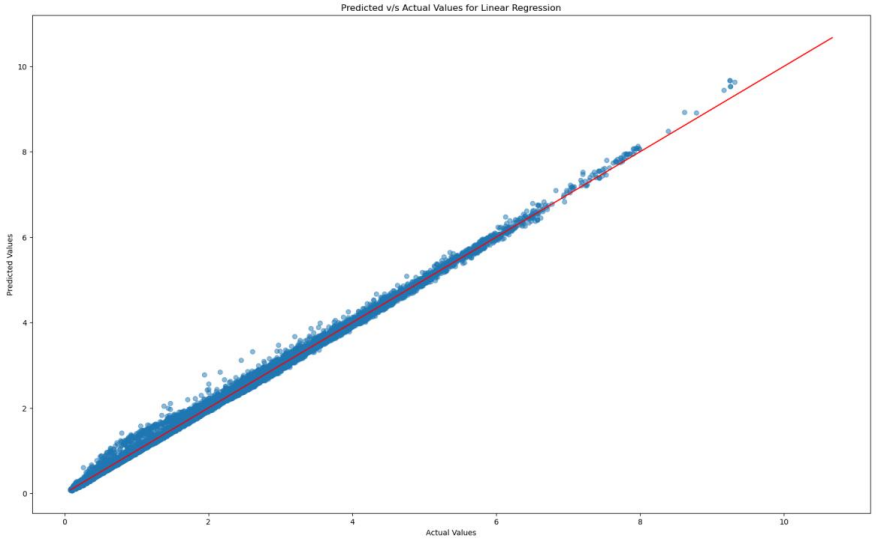


Predicted VS Actual values for MLP Regressor

Predicted VS Actual values for KNN Regressor



5. RESULT AND ANALYSIS



Predicted VS Actual values for Linear Regression

6. CONCLUSION

- Concluding from our findings, it's evident that all the regression models deliver strong performance.
- However, currently, since we need to choose the best model, we can go with either decision trees or random forests.
- As of now, our top-performing model are Decision Trees and random forests.
- If complexity is not a problem, we should go with random forests. Otherwise, we can proceed with decision trees as the difference between their performance is not that significant.

8. CONTRIBUTION

- Shivansh Mittal – Data Preprocessing, Visualization, Ridge Regression, Hyperparameter tuning, MLP and Report.
- Niharika Singh – Data Preprocessing, Data Visualization, Lasso Regression, Hyperparameter tuning, MLP and Report.
- Rishabh Sharma – Linear Regression, SGD Regression Random Forest , Decision Tree and PPT.
- Tushar Suredia – Bayesian Ridge Regression, Decision Tree, KNN, metric calculation and PPT.

But overall, it was a team effort.

THANK YOU