INTERN TRAINING REPORT

Mineral Processing and Multiphase flows Lab, IIT Hyderabad

SUBMITTED BY:

Name: Kadavath Akhil rathod

Gmail: 2021CHB1364@iitrpr.ac.in

University Roll No.: 2021CHB1364@iitrpr.ac.in

Semester/Branch: 6th Semester,Chemical Engineering

SUBMITTED TO:

Dr.Narasimha Mangadoddy

Department of Chemical Engineering

Indian Institute of Technology,Hyderabad

Contents:

Comparative Study of Machine Learning Algorithms in Predicting the Drag Coefficient of Non-Spherical Particles

---

1. Project Overview :

Objective:

To compare various machine learning algorithms in predicting the drag coefficient of non-spherical particles using features such as Reynolds number and shape factor of particles.The drag coefficient is a crucial parameter in fluid dynamics, playing a significant role in understanding the behaviour of particles in fluid flows. For non-spherical particles, predicting this coefficient accurately is challenging due to the complex interactions between particle shape and fluid dynamics.

This study aims to compare various machine learning algorithms in their ability to predict the drag coefficient of non-spherical particles. By leveraging different approaches, we seek to identify the most effective methods for this prediction task.

Significance of the Study:

Accurate prediction of drag coefficients for non-spherical particles has wide-ranging applications in fields such as chemical engineering, environmental science, and aerospace engineering. Improved models can lead to more efficient designs and better understanding of particle-fluid interactions.

2. Introduction :

In this study, various machine learning algorithms were implemented to predict the drag coefficient of non-spherical particles using experimental data. The aim was to determine which algorithm provided the most accurate predictions. This report presents a comparative study of various machine learning algorithms in predicting the drag coefficient of non-spherical particles. The performance of each algorithm was evaluated based on prediction accuracy, and the best-performing models were identified.

The settlement of non-spherical particles, such as propagules of plants and natural sediments, is commonly observed in riverine ecosystems. The settling process is influenced by both particle properties (size, density, and shape) and fluid properties (density and viscosity). Therefore, the drag law of non-spherical particles is a function of both particle Reynolds number and particle shape. Herein, a total of 828 settling data are collected from the literatures, which cover a wide range of particle Reynolds number (0.008–10000). To characterize the influence of particle shapes, sphericity is adopted as the general shape factor, which varies from 0.421 to 1.0.

3. Formal Training Provided :

During the initial phase of the project, I spent a month learning Python programming and the necessary machine learning concepts required for implementing various algorithms.Understood key machine learning concepts including data preprocessing, model training, evaluation, and various regression techniques.

Python Programming and Machine Learning Concepts

- **Duration:** 1 month
- **Activities:**
  1. Learned Python programming necessary for implementing machine learning algorithms.
  2. Understood key machine learning concepts including data preprocessing,model training, evaluation, and various regression techniques.

Data Understanding and Implementing ML

- **Duration:** 1 month
- **Activities:**
  a. **Data Exploration:** Examined the dataset for missing values, outliers, and correlations among features.
  b. **Data Preprocessing:** Implemented techniques for handling missing data, feature scaling, and outlier detection.
  c. **Feature Engineering:** Assessed the importance of features to enhance model performance.
  d. **Implementing ML Algorithms:** The algorithms implemented are Multiple Linear Regression, Polynomial Regression, Support Vector

Regression (SVR) with GridSearchCV, Multiple Linear Regression with Lasso Regression, Polynomial Regression with K-Fold Cross-Validation, Artificial Neural Network (ANN), Radial Basis Function Neural Network (RBFNN), Random Forest Regression, and Decision Tree Regression.

4. Industrial Training :

4.1 Objectives

The primary objective was to implement and compare different machine learning algorithms to predict the drag coefficient of non-spherical particles from given experimental data.

4.2 Tools & Technology Used

- **Programming Language:** Python
    1. Python was chosen for its rich ecosystem of libraries and frameworks suited for data analysis and machine learning.
- **Libraries:** NumPy, pandas, matplotlib, seaborn, scikit-learn
    1. **NumPy:** For numerical computations.
    2. **pandas:** For data manipulation and analysis.
    3. **matplotlib and seaborn:** For data visualization and Plotting.
    4. **scikit-learn:** For providing tools for data preprocessing , implementing machine learning algorithms and optimization techniques.

4.3 Techniques Studied in Different Departments

- **Data Preprocessing:** Handling missing data, detecting outliers,data splitting, feature scaling
- **Machine Learning:** Algorithm implementation, model training, and evaluation

4.4 Software and Tools Used

- **Python IDE:** Google Collaboratory
    1. Google Colab was used for its powerful computational resources and ease of collaboration. It provided an interactive environment for writing and executing Python code, visualizing data, and sharing notebooks with peers.

4.5 Highlights of Training Exposure

Gained hands-on experience in data preprocessing, implementing various machine learning algorithms, and model evaluation.

Practical Implementation of ML Algorithms:

- Gained in-depth knowledge of implementing and fine-tuning various machine learning algorithms, including Multiple Linear Regression, Polynomial Regression, SVR, ANN, RBFNN, Random Forest Regression, and Decision Tree Regression.

Data Analysis and Preprocessing:

- Learned techniques for cleaning, preprocessing, and analyzing data to prepare it for machine learning models.
- Developed skills in detecting and handling outliers and missing data, and understanding feature correlations.

Optimization and Model Evaluation:

- Applied K-Fold Cross-Validation and GridSearchCV to optimize models and improve their performance.
- Evaluated models using various metrics, including R^2, MSE, and others, to compare their effectiveness.

Collaboration and Communication:

1. Worked collaboratively with peers and supervisors, enhancing teamwork and communication skills.
2. Presented findings and results through well-documented reports and visualizations, improving the ability to communicate complex information effectively.

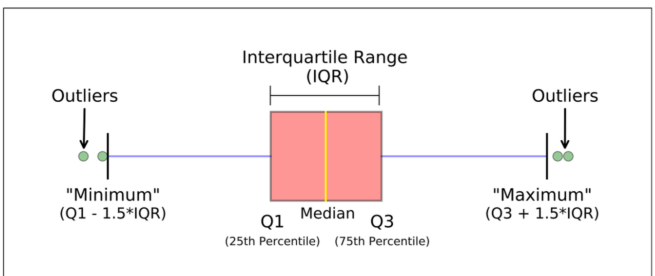
5.Problem Identification/Case Study :

The problem involved predicting the drag coefficient (C_d) of non-spherical particles using input features such as Reynolds number (Re) and shape factor(phi) of particles.

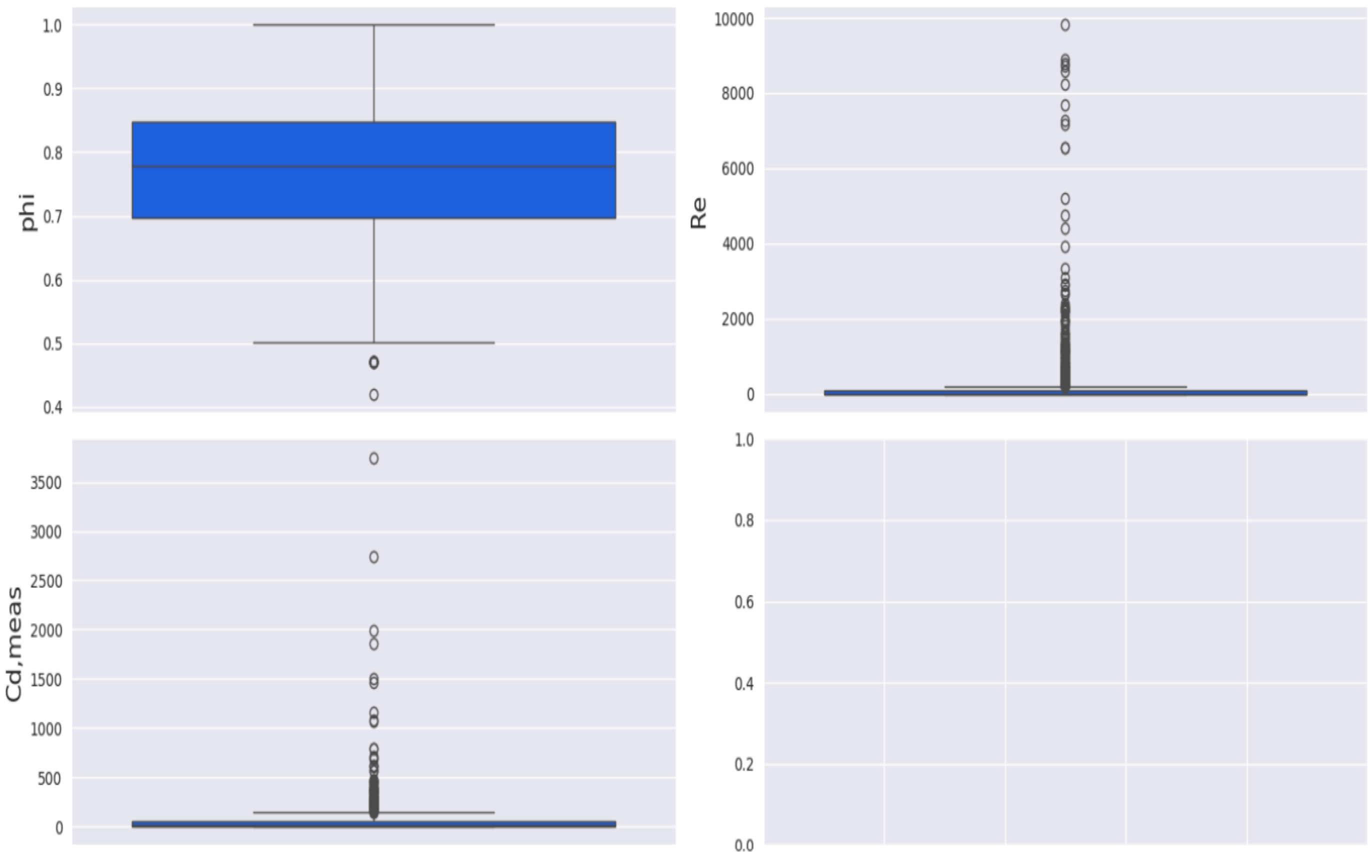
6. Brief on the Project Methodology and Approach:

6.1. Data Analyzation and filtering :

- **Outlier Detection :** By using Interquartile Range Method, we could found the outliers.



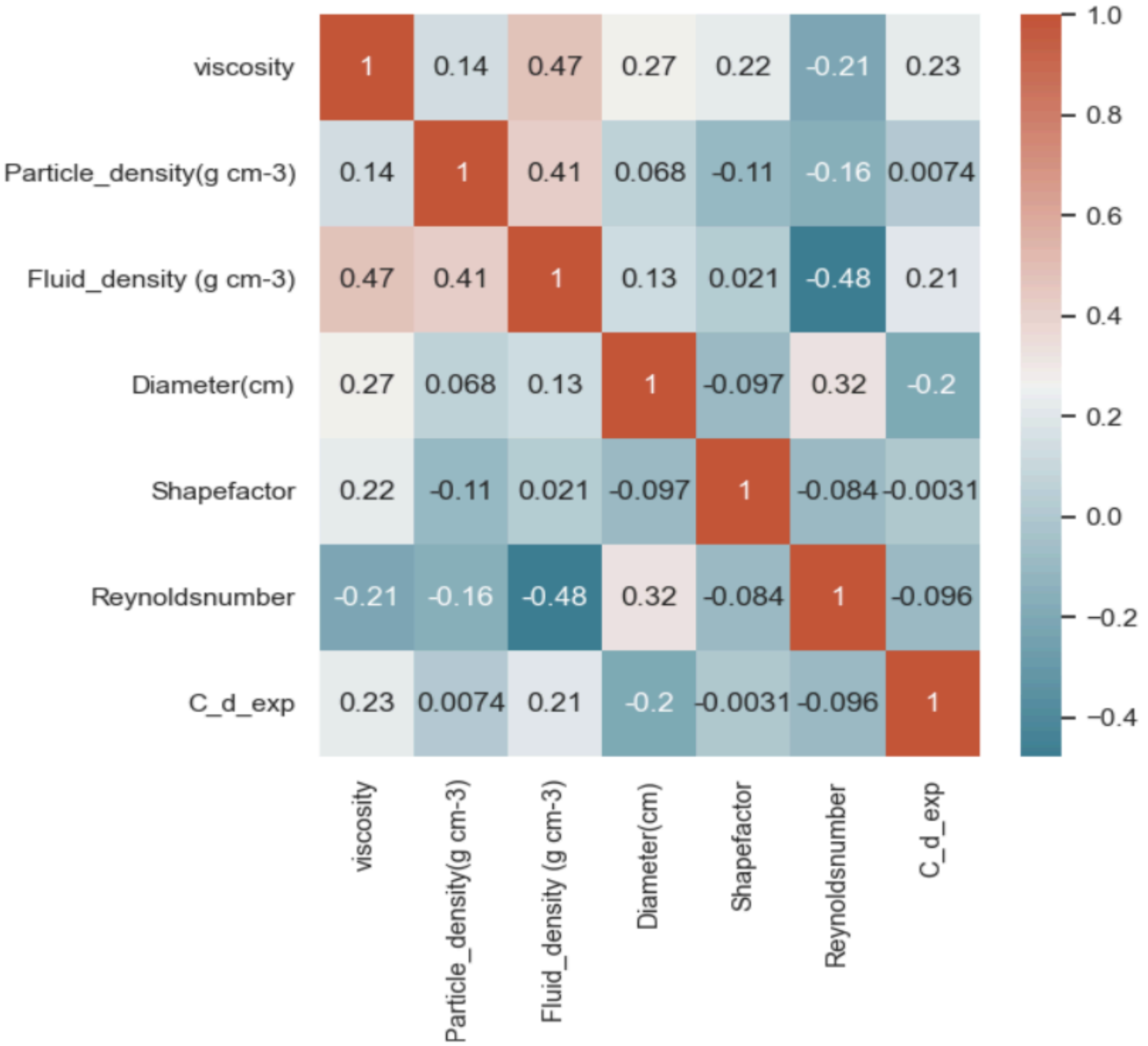
This Method is commonly used and tells that a data point is an outlier,if it is more than above the third quartile or below the first quartile as shown in the above figure.
After Implementing this method to our data, we obtain outliers for each input feature with respect to the Target variable.
Results:



From the above figure we can able to find that there are many more(i.e 296 out of 828) outliers in the Reynolds number and drag coefficient data at maximum interquartile range.
But, we didn’t remove the outliers for the further implementation because the data is experimental and by removing these we might not get the results that are related to experimental data.

- **Correlation among features:**

correlation heatmap is a graphical tool that displays the correlation between multiple variables as a colour-coded matrix. It's like a colour chart that shows us how closely related different variables are.In a correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them. The colour of each cell represents the strength and direction of the correlation, with darker colours indicating stronger correlations.

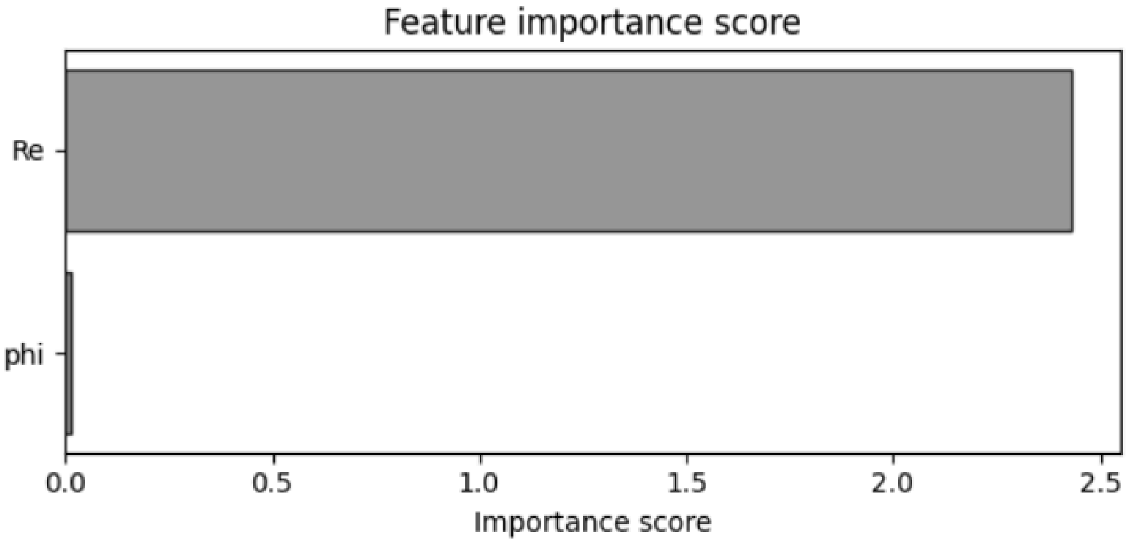


1. Look at the colour of each cell to see the strength and direction of the correlation.
2. Darker colours indicate stronger correlations, while lighter colours indicate weaker correlations.
3. Positive correlations (when one variable increases, the other variable tends to increase) are usually represented by warm colours, such as red or orange.

4. Negative correlations (when one variable increases, the other variable tends to decrease) are usually represented by cool colours, such as blue or green.

- **Permutation feature importance:** Permutation importance assess the significance of each feature independently.By evaluating the impact of individual feature permutations, it calculates importance.

  From our data, the significant features were mainly Reynolds number .But, for accounting the shape of each particle experimentally, we have taken shape factor(phi) as an input feature along with Reynolds number(Re)



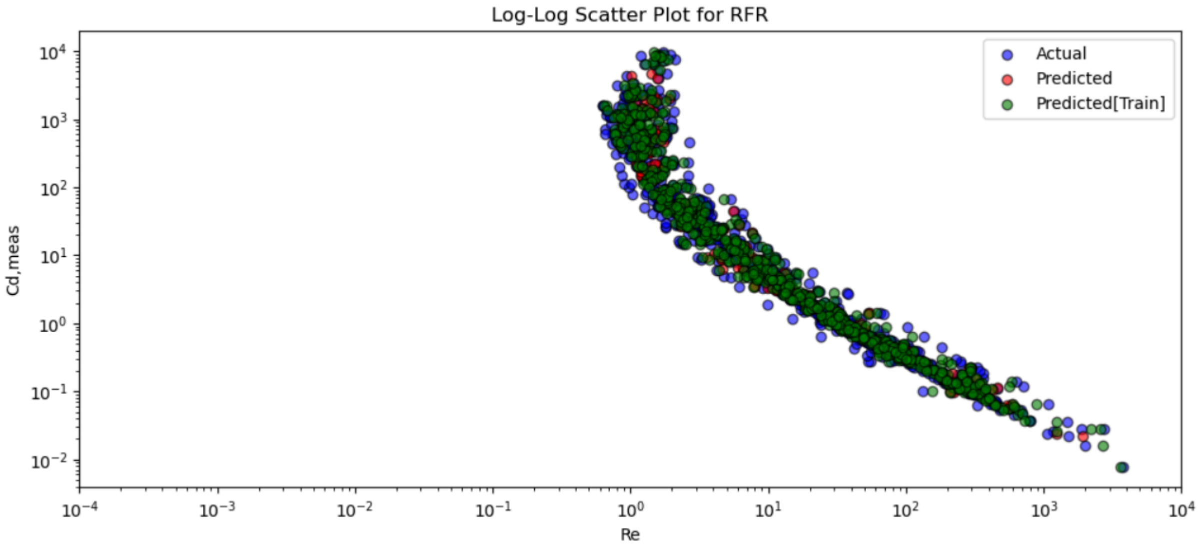
6.2. ML Algorithm Results and plottings:

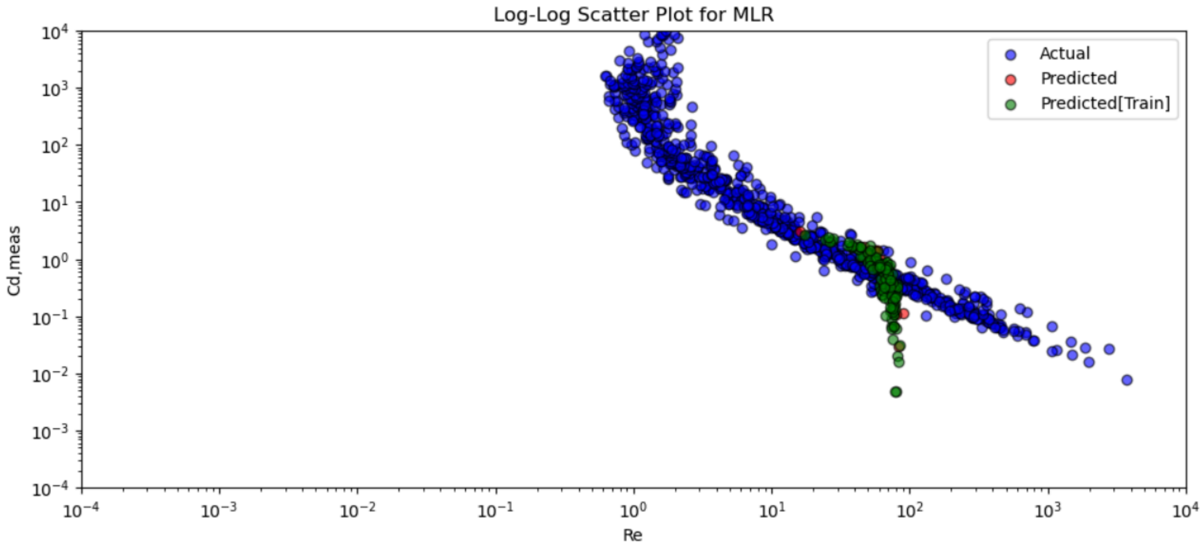
Algorithm	R^2 Score
Multiple Linear Regression + Lasso	-0.0088
MLR + Kfold	-0.000973
Polynomial Regression (deg = 3)	0.0249
Polynomial + Kfold (deg = 3)	0.026

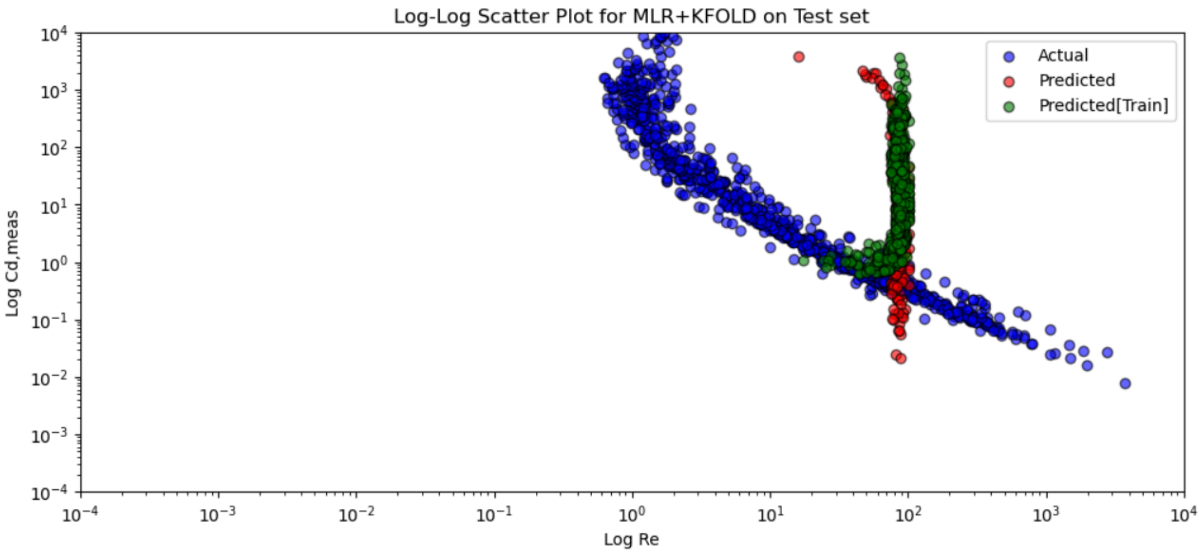
Decision Tree Regression	0.9161
Random Forest Regression	0.9163
Support Vector Regression + GridSearchCV	-0.0091
Artificial Neural Network	0.091
Radial Basis Functional Neural Network	0.321

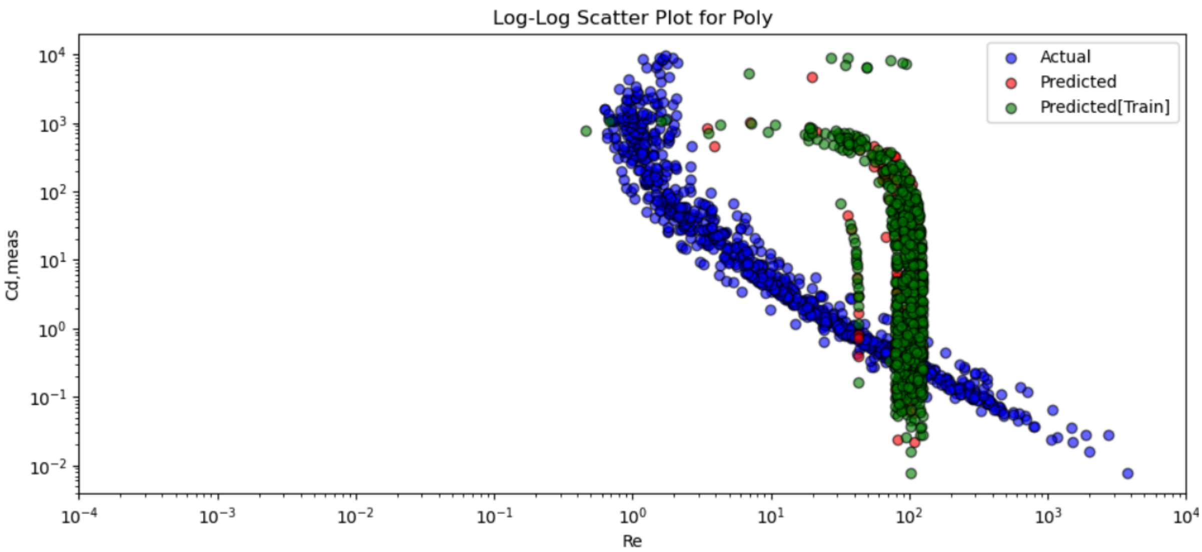
We have Plotted C_d vs Re and Actual vs Predicted both on Log scales in order for better visualization of ideal fit.Below are the pictures related to it:

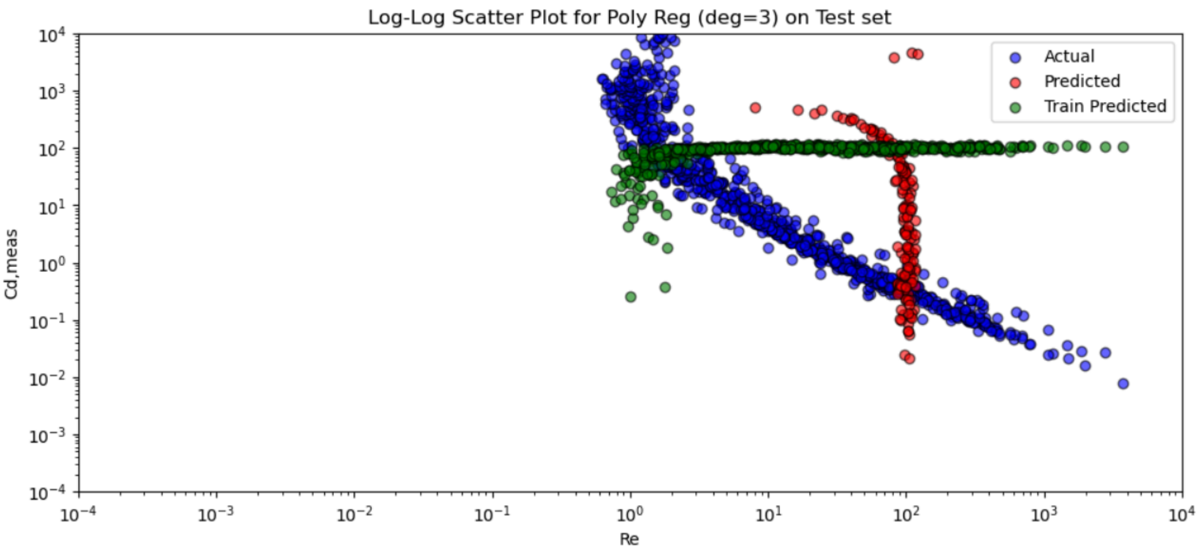
1) Log-Log scatter plot of C_d vs Re for various Algorithms :

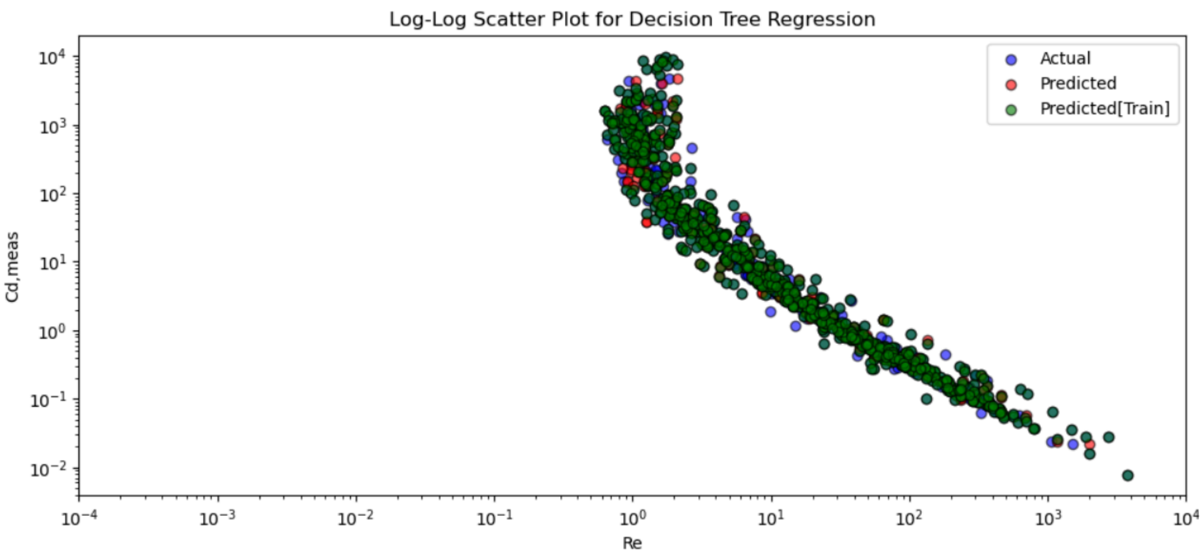

Log-Log Scatter Plot for RFR
Actual
Predicted
Predicted[Train]
Cd,meas
Re

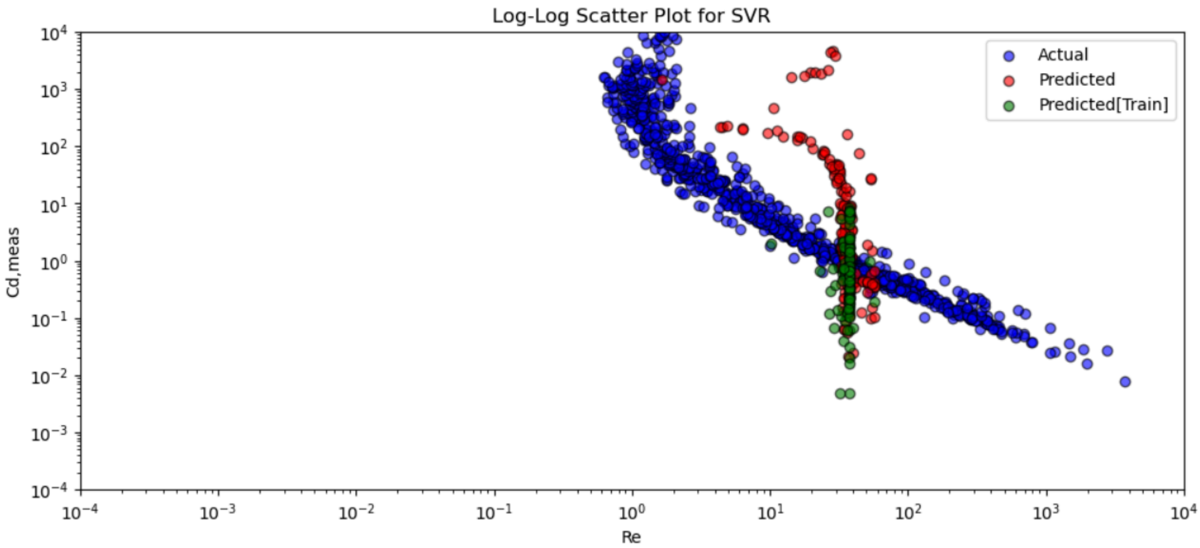

Log-Log Scatter Plot for MLR
Actual
Predicted
Predicted[Train]
Cd,meas
Re

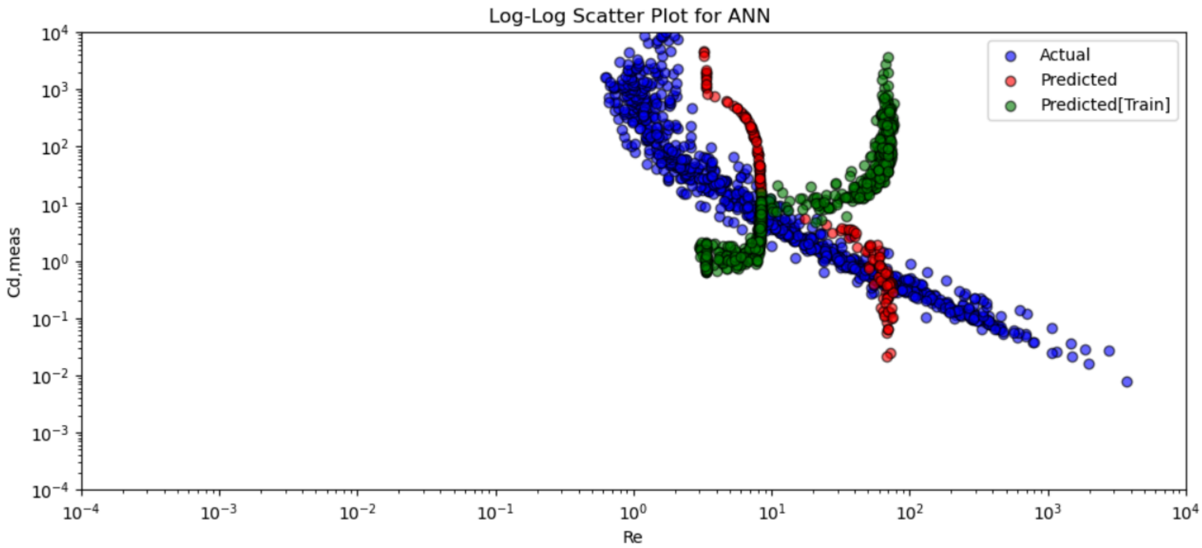

Log-Log Scatter Plot for MLR+KFOLD on Test set
Actual
Predicted
Predicted[Train]
Log Cd,meas
Log Re

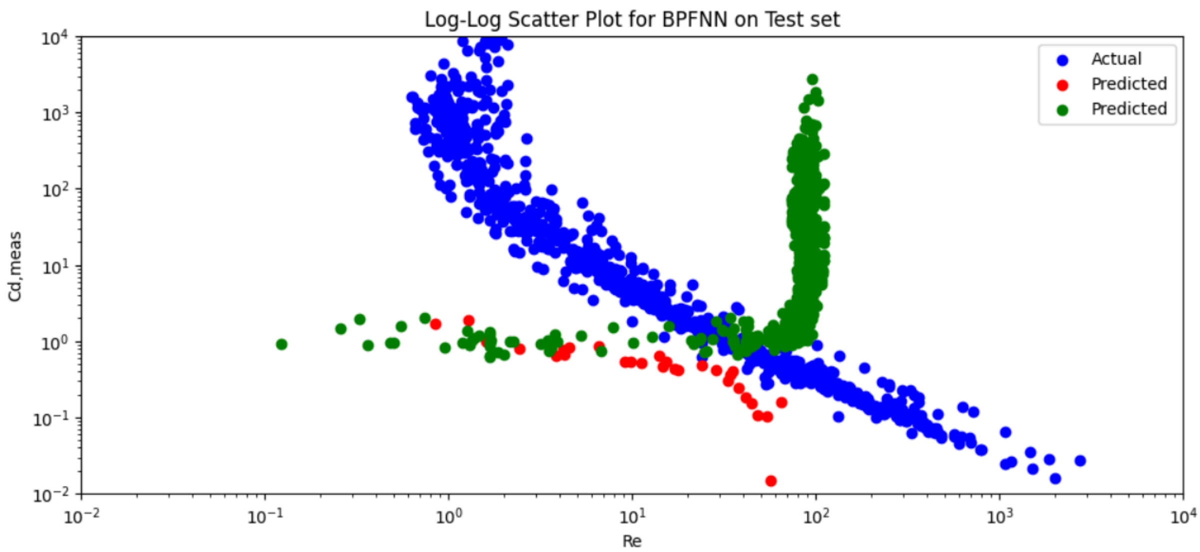

Log-Log Scatter Plot for Poly
Actual
Predicted
Predicted[Train]
Cd,meas
Re

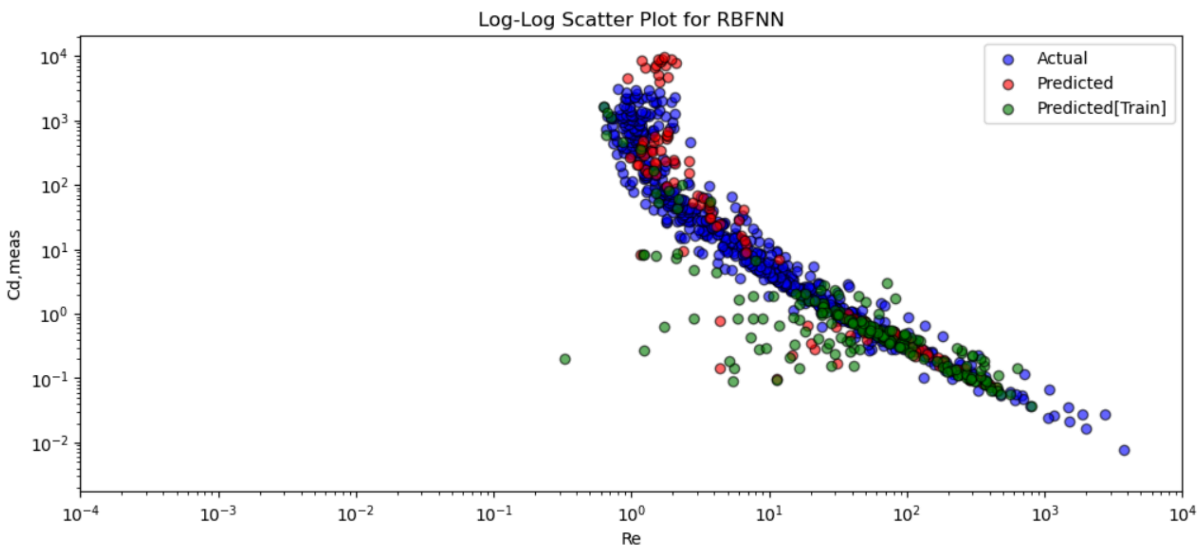

Log-Log Scatter Plot for Poly Reg (deg=3) on Test set
Actual
Predicted
Train Predicted
Cd,meas
Re


Log-Log Scatter Plot for Decision Tree Regression
Actual
Predicted
Predicted[Train]
Cd,meas
Re

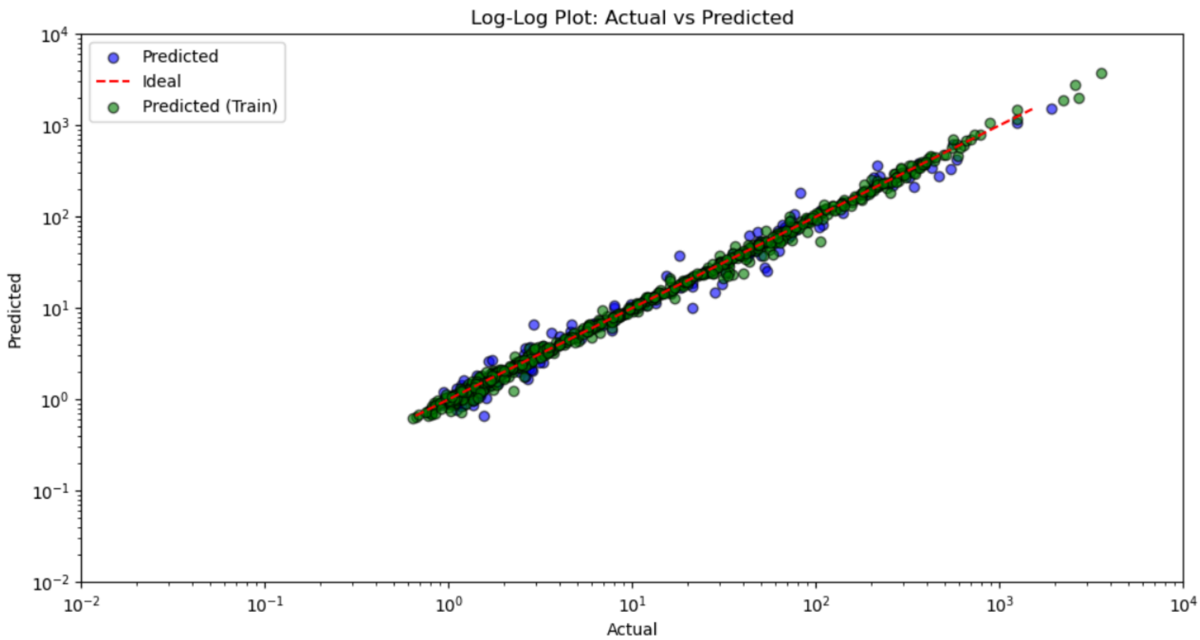

Log-Log Scatter Plot for SVR
Actual
Predicted
Predicted[Train]
Cd,meas
Re


Log-Log Scatter Plot for ANN
Actual
Predicted
Predicted[Train]
Cd,meas
Re

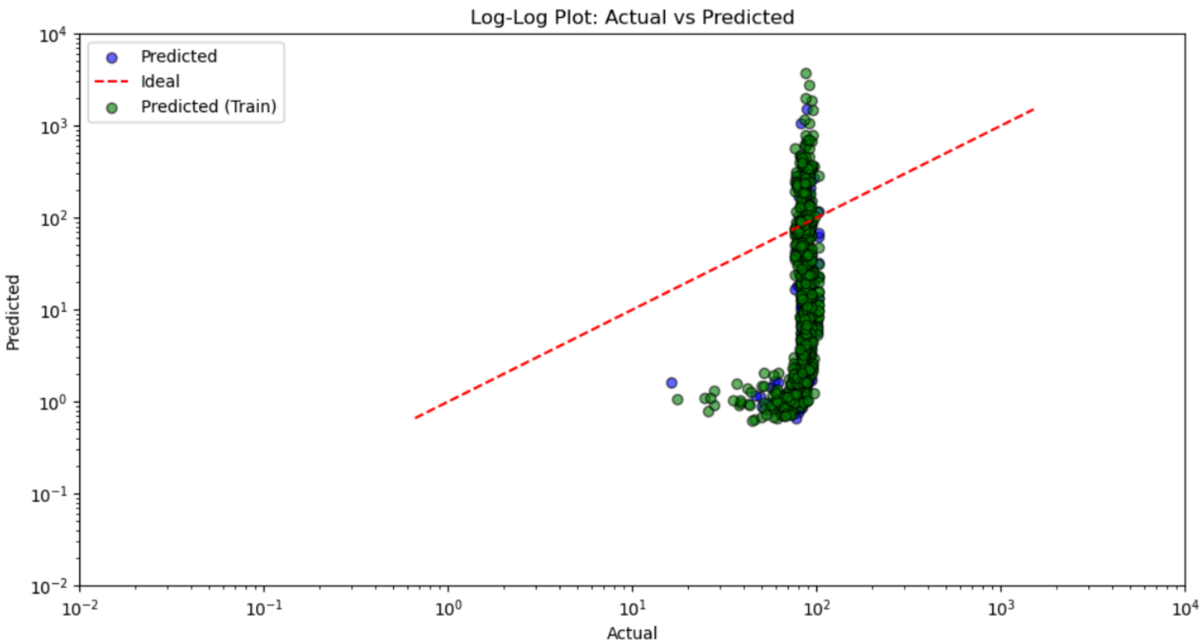




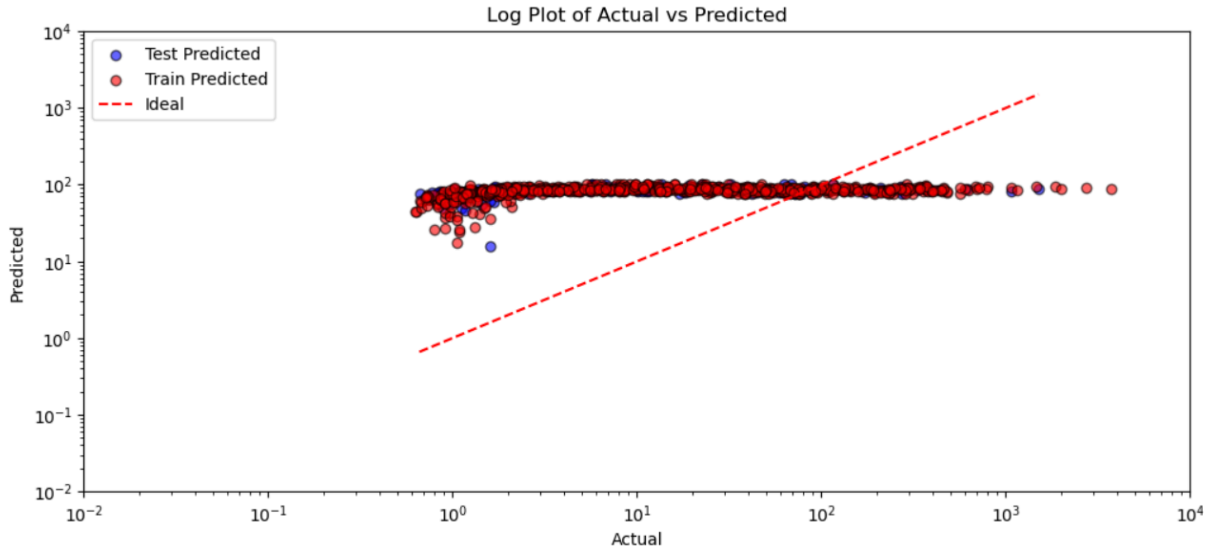
2) <u>Scatter Plot of Actual vs Predicted on Log-Log scale</u>:

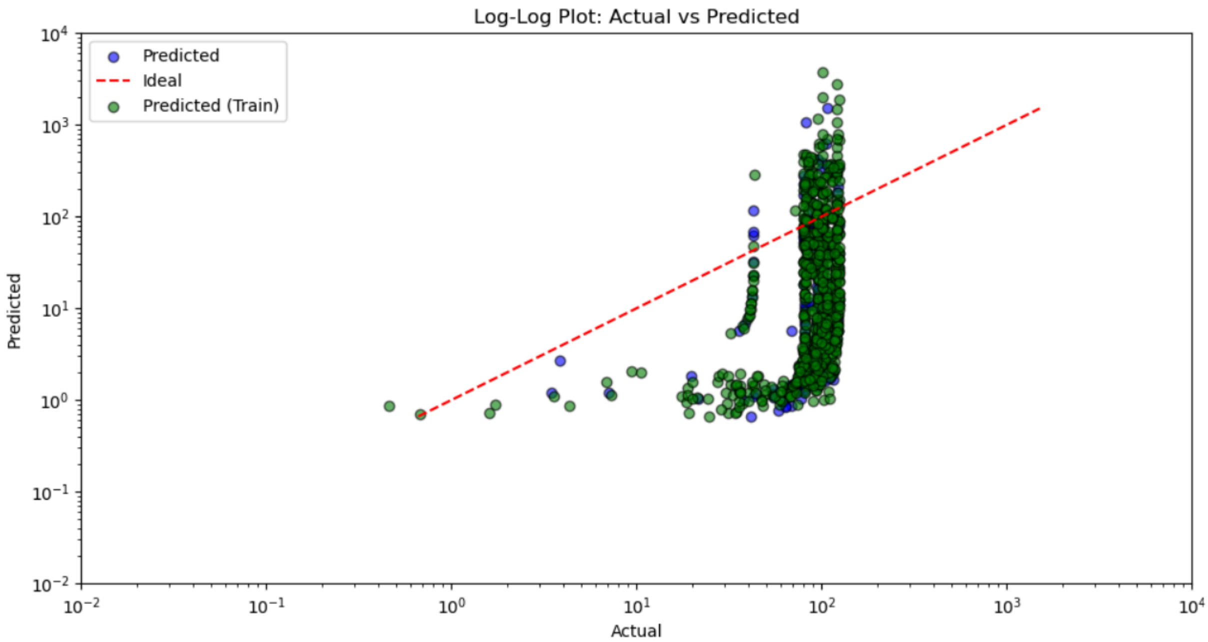


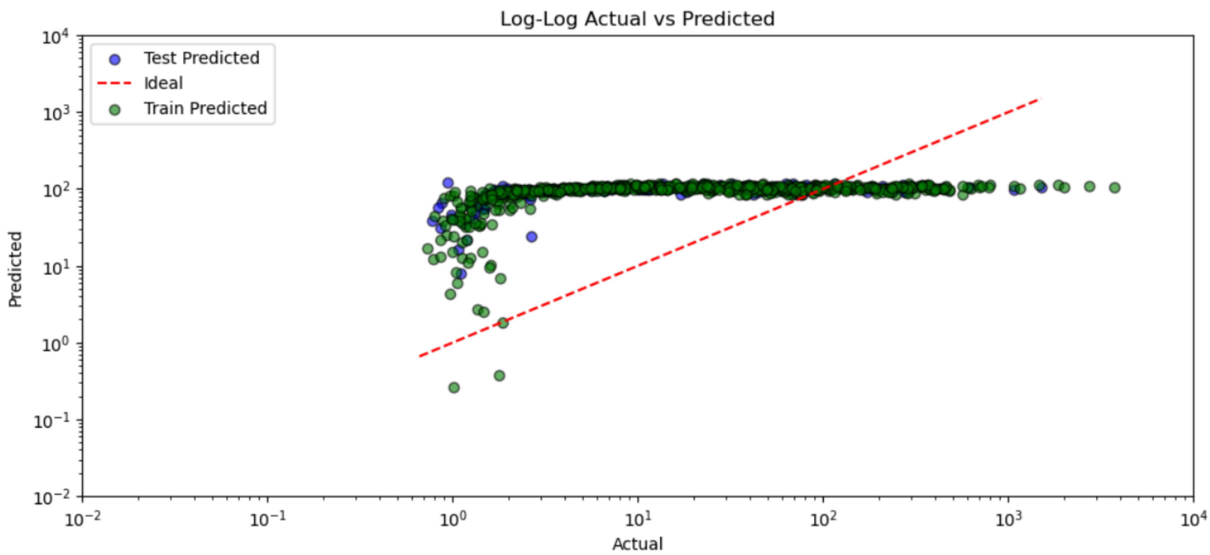
Log-Log Plot: Actual vs Predicted
Predicted
Ideal
Predicted (Train)
Actual

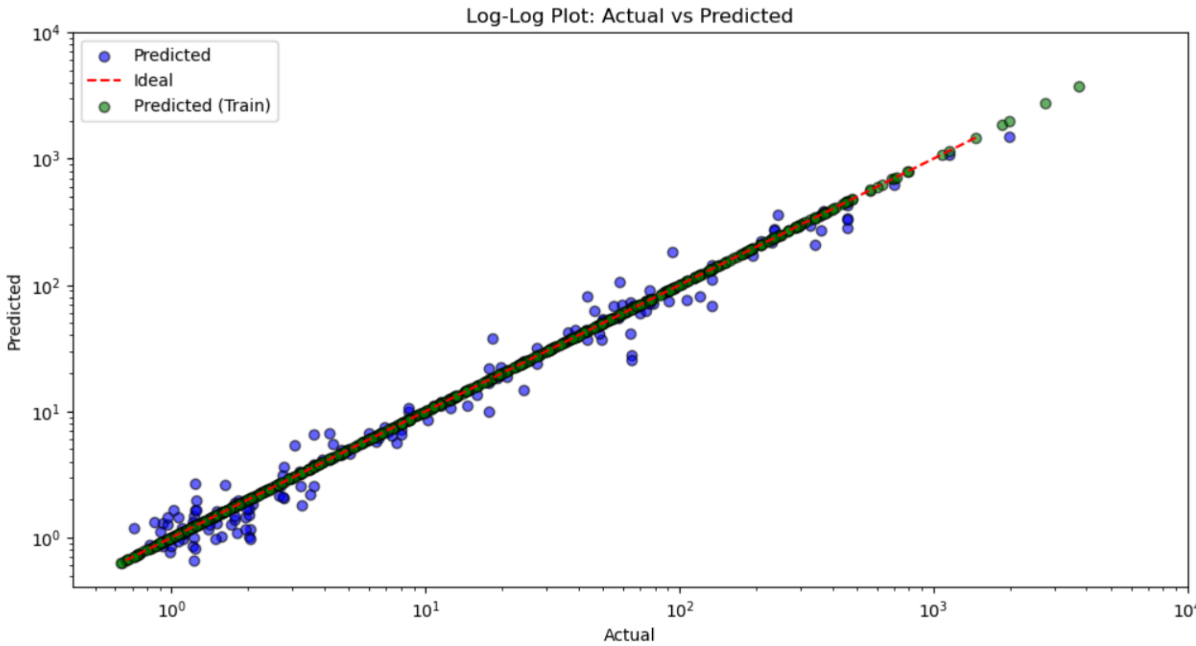


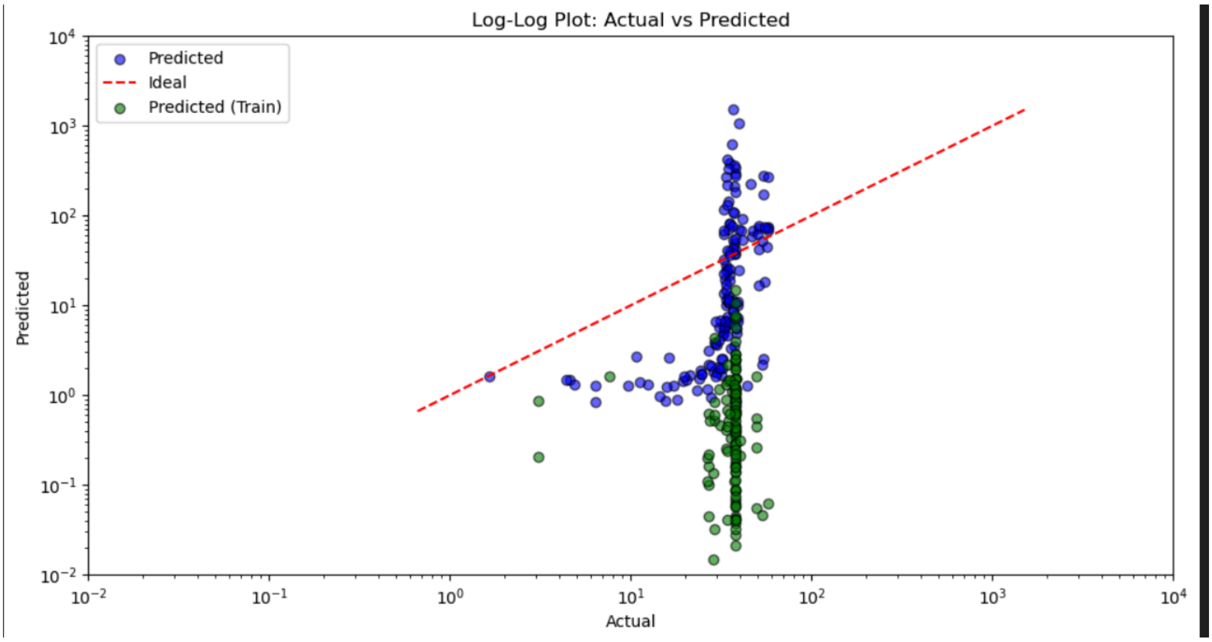
Log-Log Plot: Actual vs Predicted
Predicted
Ideal
Predicted (Train)
Actual

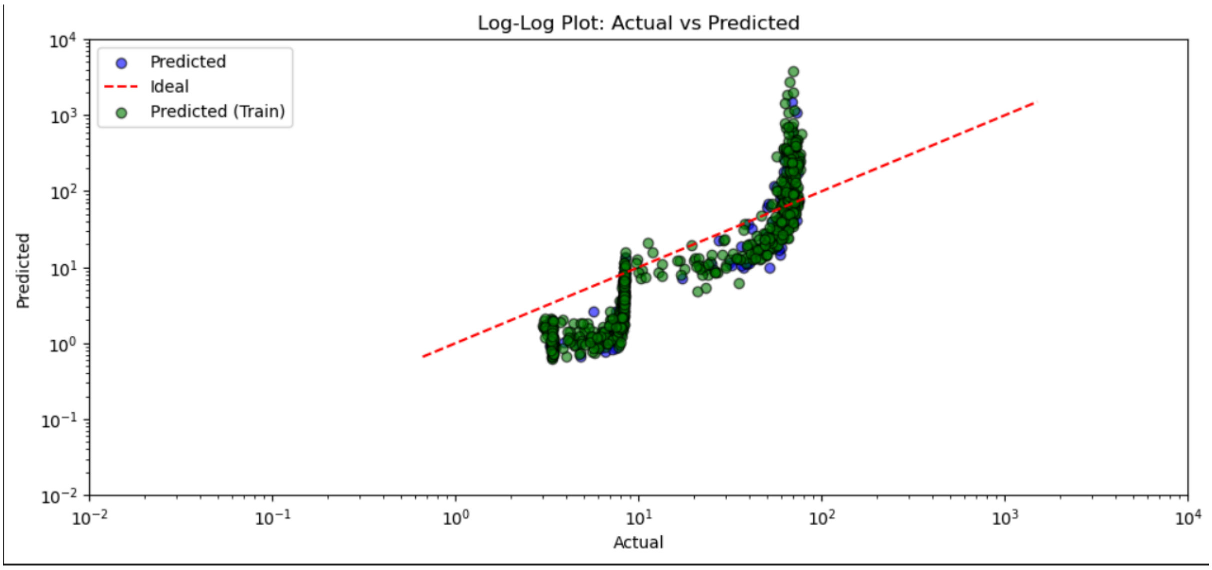

Log Plot of Actual vs Predicted
Test Predicted
Train Predicted
Ideal
Predicted
Actual

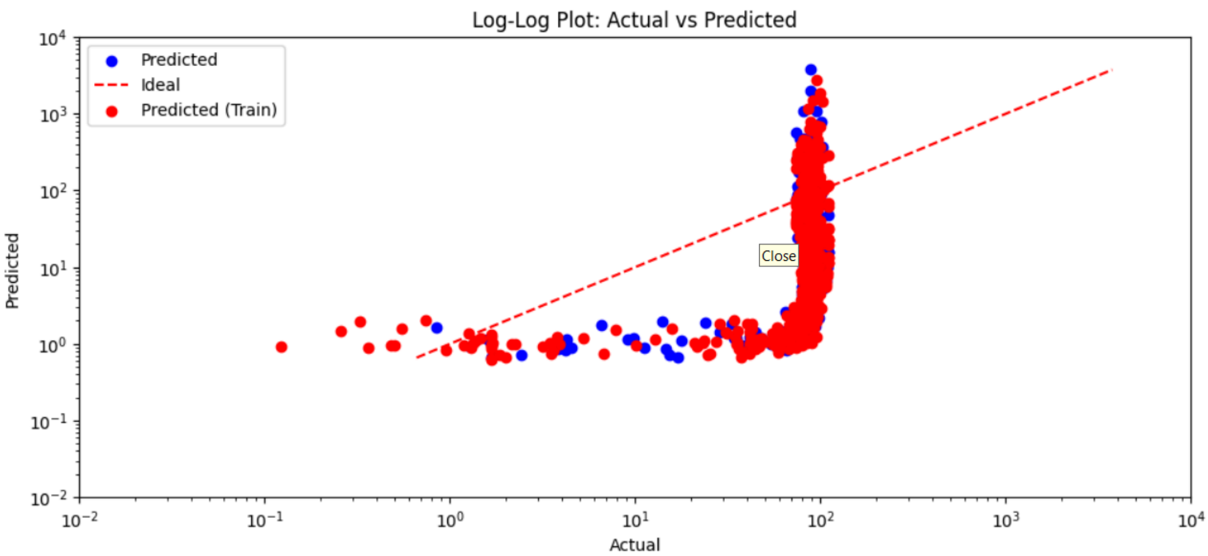

Log-Log Plot: Actual vs Predicted
Predicted
Ideal
Predicted (Train)
Predicted
Actual

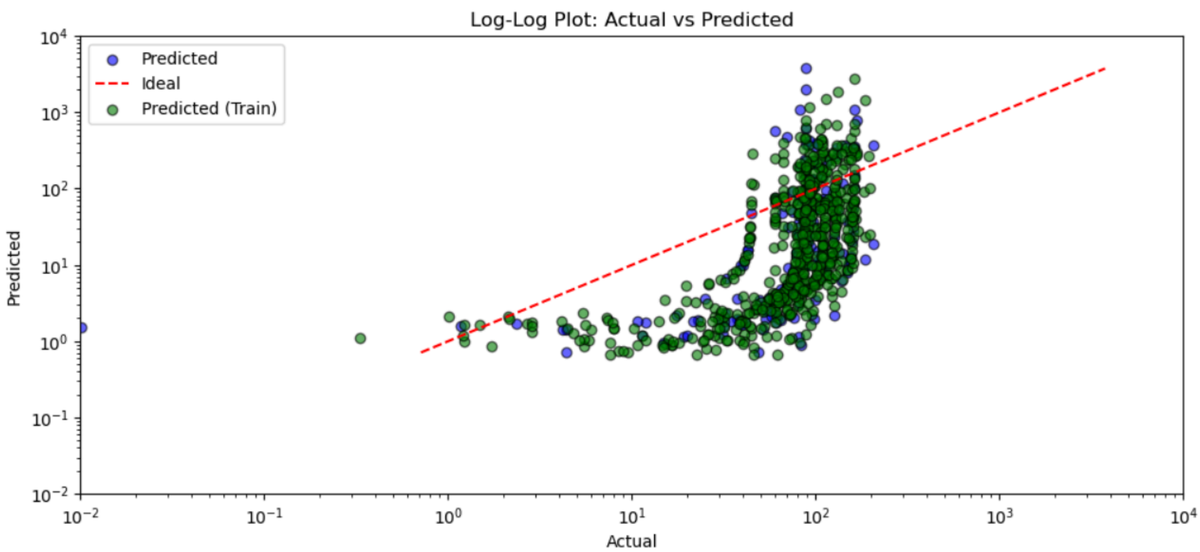

Log-Log Actual vs Predicted
Test Predicted
Ideal
Train Predicted
Predicted
Actual


Log-Log Plot: Actual vs Predicted
Predicted
Ideal
Predicted (Train)
Predicted
Actual


Log-Log Plot: Actual vs Predicted
Predicted
Ideal
Predicted (Train)
Predicted
Actual







Based on the above Plottings , the best plots of C_d vs Re are Random Forest Regression , Decision Tree Regression and Radial basis Functional Nueral Network.

7. Future Work:

7.1 Potential Improvements

- Incorporate more advanced feature engineering techniques
- Explore deep learning architectures for improved performance
- Investigate the impact of different particle shape descriptors

7.2 Extension to Other Particle Shapes Future studies could extend this work to a broader range of non-spherical particle shapes and flow conditions.

8 . Conclusion:

This comparative study highlights the effectiveness of various machine learning algorithms in predicting the drag coefficient of non-spherical particles. RBFNN, Decision Tree Regression, and Random Forest Regression emerged as the most reliable models, demonstrating their capability to handle the complexity of the data. Future work may involve further tuning of these models and exploring additional features to enhance prediction accuracy.
