Predict the Remaining Useful Life (RUL) of bearings, expressed in either minutes or seconds, using the provided test dataset.

1. Preprocessing Steps

- **Under sampling the Data**: To balance the dataset, the majority class is under-sampled. This involves reducing the number of instances in the majority class to match the number of instances in the minority class, ensuring the model does not bias its predictions toward the dominant class.

- **Feature Scaling**: Standard Scaling is applied to ensure that all features have a mean of 0 and a standard deviation of 1. This step ensures that features with larger ranges do not dominate the training process. The scaler is fit on the training data and applied consistently to both the training and test datasets to avoid data leakage. The feature is calculated using the below formula.

$$X_{New} = X_i - X_{Mean} / (X_{Std})$$

- **Test Data Handling**: For the test data, only feature scaling is applied (no outlier removal or under sampling). The scaling parameters from the training data are used to transform the test data to maintain consistency.

2. Model Design

- **Random Forest Regressor**: A Random Forest model is used for predicting the Remaining Useful Life (RUL) based on the extracted and scaled features.
  It is an ensemble method that creates multiple decision trees during training and aggregates their predictions, making it robust against overfitting and capable of capturing both linear and non-linear relationships.
  It handles high-dimensional data well and automatically evaluates feature importance.

  Key Parameters:

  - n_estimators = 400: Specifies the number of decision trees in the ensemble. A larger number can improve accuracy but increases computation time.

  - random_state = 5: Ensures reproducibility by fixing the randomness in the training process.

- **Model Training**: The Random Forest model is trained using the preprocessed features (features) and target variable (target). Training involves learning patterns from the input features to predict the Remaining Useful Life (RUL).

- **Model Prediction**: The trained model is used to predict RUL values for the test dataset.

3. Experimentation

We had conducted 4 different experiments for the following task. For each experiment R^2 and Root Mean Squared Error (RMSE) scores were calculated to assess the model's performance.

The R-squared metric (R^2) is used to measure how well a model fits data, and how well it can predict future outcomes. It also gives how much of the variation in your data can be explained by your model.

RMSE measures the average difference between values predicted by a model and the actual values.

Random forest regressor was chosen as the primary regression model in all the experiments and the results are tabulated below.

Experiment	Methods Used	R^2	RMSE
1	Scaling + PCA + Random Forest	-0.1282	5525.5588
2	Feature Selection + Scaling + Random Forest	-0.0247	5266.0053
3	Scaling + Random Forest	-0.1186	5501.8327
4	Under Sampling + Scaling + Random Forest	0.2424	4527.5050

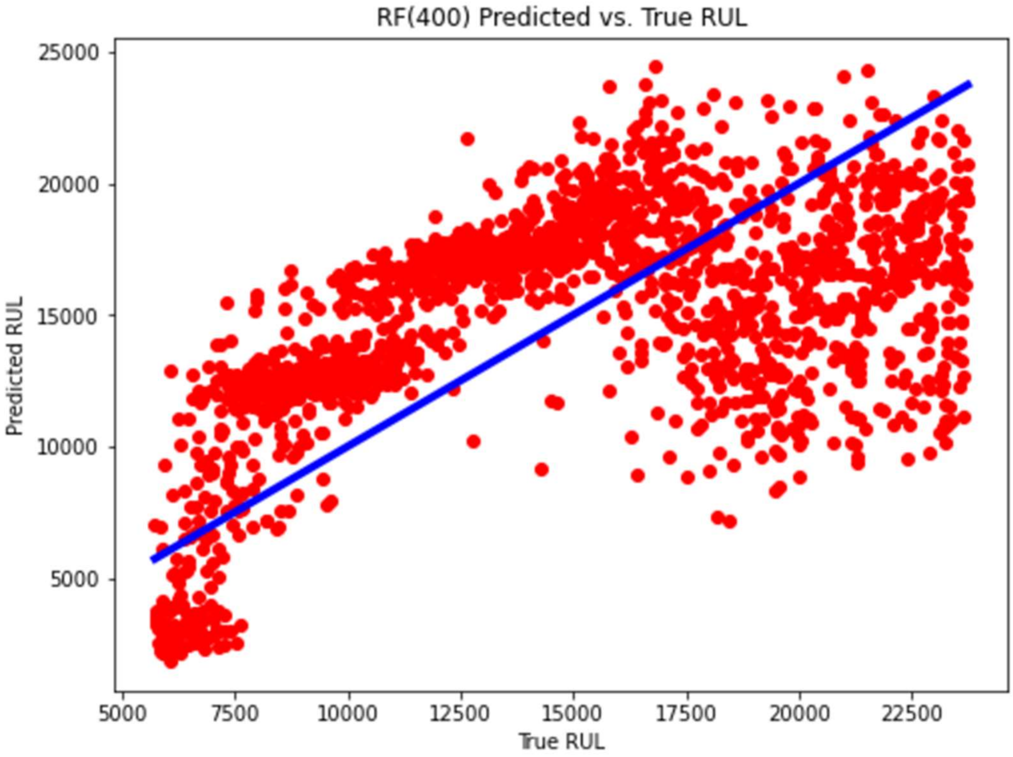
4. Result and Analysis

From the above 4 experiments we can see that the R^2 score for the first 3 experiments is negative. This implies that the model is performing very bad.

In the 4th experiment however it is performing well (indicated by the positive R^2 score of 0.24).

First, we Under Sample the data to balance it and then perform the Feature Scaling. Later we apply the Random Forest Regressor to predict the RUL values.

The output of this model is shown below in the figure.


RF(400) Predicted vs. True RUL
Predicted RUL
True RUL