Problem Statement 2 – Statistics and ML

Remaining Useful Life (RUL) is the estimated remaining operational lifespan of an engine before it reaches a point of failure or maintenance. The main objective of this problem statement is to estimate the RUL of the engine using multiple time series data. This requires in-depth analysis of the data from different engines, each operating in different operation settings until the engine fails. The main challenge is to develop a predictive model that uses time series data to efficiently predict the number of continuous operations or cycles before the engine fails.

The dataset consists of multiple engines with their own set of sensor readings recorded over time when operating under three different settings until failure occurred. Initially, the provided dataset is loaded into the notebook, and the required preprocessing and exploration are done, which involves labelling the columns, dropping the missing values from the dataset, and verifying the null values in each column. The histogram plot for analysing the distribution of cycles depicts a right skewness, which suggests that a significant number of engines experience early failures or have shorter lifespans.

Remaining Useful Life (RUL) is calculated by subtracting an engine's current cycle duration from the overall maximum duration expected before failure. In the dataset utilized for this project, the variable **'cycle'** is used to calculate RUL. This variable was chosen because it directly reflects the wear and tear suffered by the engine over time. By comparing this with the total cycles expected before failure, we can predict when failure occurs.

For predicting RUL, the machine learning algorithm **Random Forest Regressor** is used. The model is trained on the engineered features using the training data set, and their performance is assessed using the evaluation metric **Root Mean Squared Error (RMSE)**, which reduces large errors more significantly. During the training phase, the model achieved an RMSE of 41.52, indicating the average difference between predicted and actual RUL values in the training dataset. Upon applying the model to predict RUL for test data, the RMSE improved to 33.37, showing the model's ability to generalize well.

Fine tuning indeed optimizes the performance of the machine learning model. A tuning method such as Grid search, which completely searches through a manually entered hyperparameter space, followed by cross-validation, helps in assessing model performance more quickly and reduces overfitting. Performance can also be improved by reducing dimensionality by transforming features into principal components. Additionally, ensemble methods which combine predictions from multiple models, can also be used to improve accuracy.

Flowchart