**Blast Furnace Productivity Solution Approach**

Please refer Jupyter Notebook, detailed comments are provided.

Step 1: All the required libraries are loaded.

Step 2: Both train and test data is read from csv to pandas dataframes.

Step 3: Both train and test data is concatenated to one dataframe for analysis. Column S.NO is dropped.

Step 4: Data Visualization Seaborn Pairplots:

1. From these plots it can be seen that there are many outliers and also negative values in the target Productivity which doesn’t makes sense. Either these has to be discarded or made positive since productivity can’t be negative.
2. It is also seen that parameters CO2Exhaust and Burden are linearly dependent on each other that is correlated.

Step 5: Check for Null values

We can see that 65% of the values of OreSource are null. From Seaborn Violin plots we can see that few OreSource has effect on Productivity but these are less frequent. Here OreSource is dropped.

Step 6: Seaborn Heatmap (check for Correlation) From heatmap it is seen that CO2Exhaust and Burden are highly correlated. Of the two the one which has low correlation with Productivity(target) will be dropped i.e CO2Exhaust.

Step 7: Seaborn Pointplots:

Pointplots of Average productivity across Fe\_Content and SinterQuality are plotted. Also pointplots of Average productivity across BlastAirTemp and Al2O3 are plotted on data without Outliers.

Feature Engineering

Step 8: Creating Dummies for Fe\_Content and SinterQuality

Step 9: Separating Train and Test

Step 10: Removal of Outliers from Train data

Productivity of more than 5 are considered as outliers. We can try with more than 4 and 3 but we might lose some information about other parameters.

Step 11: Weights of different parameters on Productivity are measured using Permutation Importance.

Step 12: Partial Dependence Plots To get the insights of different parameters and their effect on productivity to find optimum parameters.

Step 13: SHAP values

These values shows the contribution of different parameters on Productivity from the baseline for each prediction.

Step 14: Distplot of Productivity shows that it follows Normal Distribution.

Step 15: Scaling

Scaling is used to avoid any bias in the data.

Step 16: Machine Learning Models

Different machine learning models are fitted using Root Mean Squared Logarithmic Error as the evaluation criteria.

We have trained using RandomForest and GradientBoosting , Since data is very less other algorithms like XGBoost will overfit.