PRAKRITI 2020
Data Analytics Report

Team Name

INTANGIBLE

Team Members

Saurav Kumar Nishant

Saw Sachin Azad

Tusar Mohandass

Saurabh Kumar Pandey

Steps followed in building the model and Selecting Features:

- Data has been imported in the Google Colaboratory.
- Null value has been checked (there aren't any NULL values) and data types of each feature has been checked.
- Two columns (**TC (%) and TN (%)**) which were supposed to be of **float type** were in **object type** so we checked the string value and removed it.
- Finally when **string (n.a.)** was removed from these two columns then type was converted from object to float by typecasting.
- Data set was splitted into **X (Input) and y1 and y2 (outputs)**.
- Then Input data X was modelled into three categories **X1 (Only elemental data(highlighted in green)), X2 (Only Spectral data(not highlighted)) and X3 (Combination of elemental and spectral data).**
- **Pearson's correlations coefficient** were checked for each of the three situations and if two features were found to be correlated one from those two were removed.
- **MinMax Scaling** was done on both Input and Output in all the three situations.
- Then on all the three given sets of data three models were trained i.e. **Linear Regression model, Decision Tree Regressor and Gradient Boosting Regressor** on X_scaled_corr.
- **X_scaled_corr** is the **best set of features** obtained from Pearson's Correlation Coefficient after it has been scaled.
- **Root mean squared errors** were calculated in each case to select the best Model.

Results:

Number of pairs of correlated features (based on Pearson's Correlation coefficient) :

X1-Only elemental data(highlighted in green) >0.85	X2-Only Spectral data(not highlighted) >0.9	X3-Combination of elemental and spectral data. >0.9
4	1886	1887

Root Mean Squared Error (RMSE) for TC(%) (y1):

	X1-Only elemental data(highlighted in green)	X2-Only Spectral data(not highlighted)	X3-Combination of elemental and spectral data.
Linear Regression	**0.21422**	**0.09985**	**0.09792**
Decision Tree Regressor	**0.09777**	**0.11456**	**0.09259**
Gradient Boosting Regressor	**0.08214**	**0.13700**	**0.09728**

Root Mean Squared Error (RMSE) for TN(%) (y2):

	X1-Only elemental data(highlighted in green)	X2-Only Spectral data(not highlighted)	X3-Combination of elemental and spectral data.
Linear Regression	**0.00783**	**0.00446**	**0.00434**
Decision Tree Regressor	**0.00442**	**0.00373**	**0.00456**
Gradient Boosting Regressor	**0.00468**	**0.00470**	**0.00528**