Prakriti Data Analytics (one page report)

Team Name - WAKANDA_FOR3V3R

Ayush Kumar	9661179290	17CS10007	IIT KHARAGPUR
Anshul Choudhary	8918500798	17CS10005	IIT KHARAGPUR
Prakhar Bindal	9593801201	17CS10036	IIT KHARAGPUR
Shubham Raj	9733331647	17CE10054	IIT KHARAGPUR

Methods used

- **Data Cleaning** : The data given contains data of TC(%) and TN(%) as n.d. and 0, as well as certain spectral values are missing. These data examples need to be taken out of consideration while we train and test the model.
- **Feature Engineering** : Data is so sparse, having a lot of features. So first, we did Normalization and then extracted the influential features, who contributed more to the data set using random forest as base model. Influential variables data is shown in the appendix section.
- **Data Splitting** : The data is split into training and testing data sets as 80% of data is chosen randomly to be the training data and the rest is testing data.
- **Model Training :** Since the number of training examples are very low compared to the number of features. We avoid using Deep neural networks. We used different Machine Learning Algorithms like Random Forest, XGBoost, Support Vector Machine to predict the Total Carbon and Total Nitrogen from Elemental, Spectral and combined features. Random Forest Regressor gives promising results with a number of trees equal to 500.
- **Model Testing :** The model is tested on the testing data with the following result shown below in the appendix section.

Appendix

MODEL - Random Forest Regressor

Random forest is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Accuracy

vs	Spectral Data	Elemental Data	Combined Data
TC(%)	93.787	91.157	93.824
TN(%)	75.496	51.373	77.376

RMSE Score

vs	Spectral Data	Elemental Data	Combined Data
TC(%)	1.634	1.949	1.629
TN(%)	0.169	0.239	0.163

No. of influential variables

vs	Spectral Data (out of 2149)	Elemental Data (out of 11)	Combined Data (out of 2160)
TC(%)	135	3	97
TN(%)	356	3	325