EXCAVATE ABSTRACT

Group Members:
Arpit Bansal-IIT Kharagpur
Ritwika Chowdhury-IIT Kharagpur
Jitendra Jangid-IIT Kharagpur

Our Objectives:

- To handle **the missing values** in the training data provided to us
- To make an effective **multi-classification predictive model** which can predict the type of glass in terms of **F1 Score**
- To find out the **importance** or **influence** of independent variables in the data towards predicting type of glass

Brief Description of dataset:

- Number of independent variables in training dataset: 9 excluding SN
- Number of response variables: 1 (**categorical** variable) – **'ID of Glass'**
- Total no. of instances in the dataset: **200**

Programming platform used: R programming language

Approach

- Training dataset was imputed by R software
- Missing data were detected and plotted in a histogram using **VIM library and aggr function**
- It was found that there were a total of **6 variables** containing missing values with a **range of 15% to 55%**
- Then missing data were **imputed using mice library in R software**
- In Feature engineering we created new variable **invere_RI** which is inverse of RI variable in train data
- After the imputation of missing values we **splitted out whole training set** in a train data and a test data using 65:35 approach to generate a **robust model**
- Further the splitted train data was trained using **XGBoost Algorithm(Xtreme Gradient Boosting Algorithm)** which was used for our **Multiclassification** purpose
- After that splitted training data was trained using **XGBoost Algorithm** which was used for our **Multiclassification** purpose
- The model was built using an **ensemble** of 3 different parametric values of our **boosting parameters "eta(0.1,0.05,0.01)", "colsample_bytree(0.2,0.4,0.6)" and "subsample(0.4,0.8,1)"** which decided our learning rate and control the overfitting**.** We kept the **"number of rounds(nrounds)"** to 100 and **"max_depth"** to 8
- Using **sapply function** we accounted for our final prediction by taking the **Median** for that particular class for a particular data point.
- Further after getting our predictions for both training and testing data set by Xgboost algorithm we generated the **confusion matrix** table for this multi-classification problem.
- Several different parameters were estimated such as **accuracy of model (95-98 %)**, precision and recall of our predictions and finally the **F1 score** for each class which on an average turned out to be around **95%.**
- So the prepared XGB model could be performed on any other given testing dataset also with same independent variables
- Finally we got the importance of all variables by Xgboost model:

Feature	Importance
Inverse_RI	**0.625**
Mg	**0.077**
Ba	**0.066**
Ca	**0.053**
Al	**0.044**
RI	**0.043**
K	**0.042**
Na	**0.026**
Si	**0.012**