ABSTRACT

Problem Statement

The aim is to predict “The formation energy” and “The band-gap energy” in order to facilitate the discovery of new transparent conductors so that the computational cost gets reduced.

Data description

The dataset consists of 11 different features of transparent conducting materials and the goal is to predict two properties i.e., the formation energy and band gap energy to facilitate the discovery of new transparent conductors.

Data Exploration

- The dataset contains 1900 rows and 14 columns, the last two columns being the target variables. It does not contain any null values and hence no data imputation was needed.
- There are two categorical variables “*spacegroup*” & “*number_of_total_atoms*” and the rest predictor variables are continuous variables.

Model Selection and Fitting

Next, the training dataset was split into two parts :- 70 % for train and 30 % for test. A 5-Fold Cross Validation was performed on these parts in order to get a better idea of how it would perform on new input data.

Different models were trained on the above mentioned dataset, namely, Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Kernel Ridge Regression, Gradient Boosting Classifier and XGBoost Classifier.

The highest accuracy was given by **Gradient Boosting Classifier** for both the target variables. This was derived by looking at the RMSE values of the trained models.

RMSE for ***'formation_energy_ev_natom'*** : 0.043897
RMSE for ***'bandgap_energy_ev'*** : 0.242477