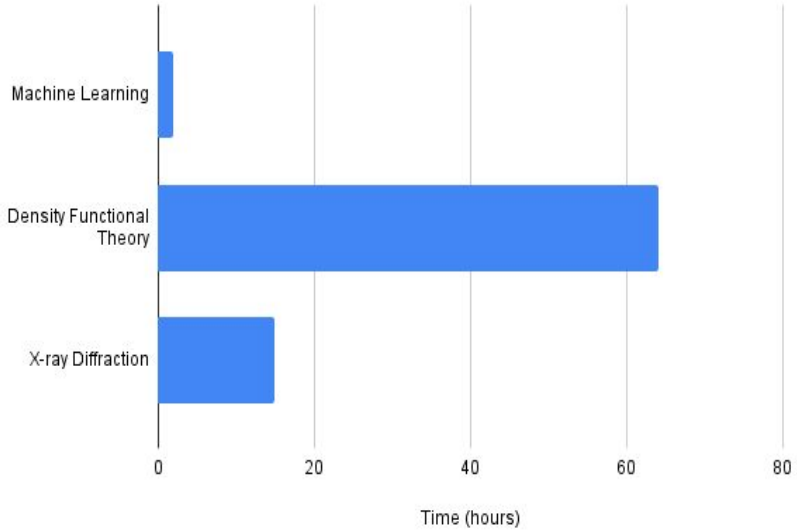
Prediction of Lattice Structure of Perovskites from physical properties using ML

- Group 22: Sunil S Raja (SPS, Batch 20), Kaling Vikram Singh (SPS, Batch 20).
  Guide: Dr Subhankar Mishra

- Perovskites are versatile materials that have a similar crystal structure to calcium titanium used in a variety of industries, including solar panels.
- Their varying A and B compositions result in different structures, such as cubic, monoclinic, or orthorhombic. Good oxide ion conductivity is necessary for fuel cell applications, and a cubic perovskite's reduced distortion can provide this property.
- DFT calculations and XRD techniques are currently used to determine the materials' crystalline structure, but ML models can potentially reduce the cost and energy consumption of these processes.



Relevant works

Machine learning models were explored to predict perovskite crystal structures with promising results. **Santosh and Taher's** work achieved accuracies ranging from 62.8% to 80.3% using XgBoost, SVM, Light BGM, and Random Forest algorithms. However, their dataset was limited and biased towards orthorhombic structures, and they accounted for the tolerance factor. **Jarin et al.** achieved an accuracy of 95% using genetic algorithm support vector regression and various neural networks without oversampling. They removed less important features but did consider the tolerance factor. Overall, these models offer potential for cost and energy savings, but limitations should be taken into account.

Baselines Implemented

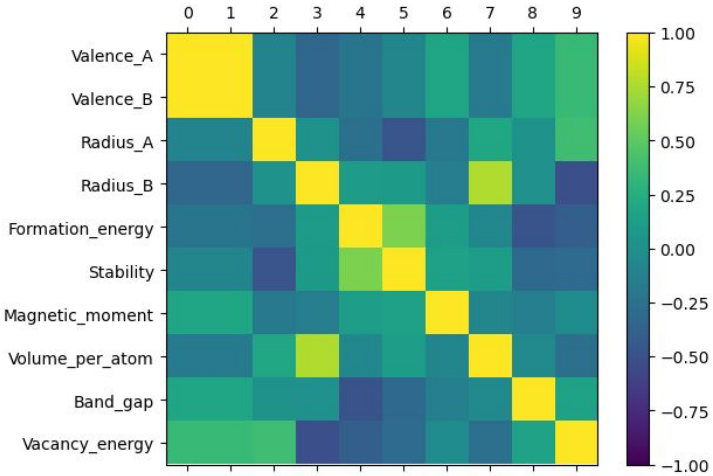
- Database Collection
- Feature Selection and Data Processing
- Model Selection (Light BGM, XgBoost, SVMs)
- Hyper-Parameter Optimization
- Testing for Accuracy

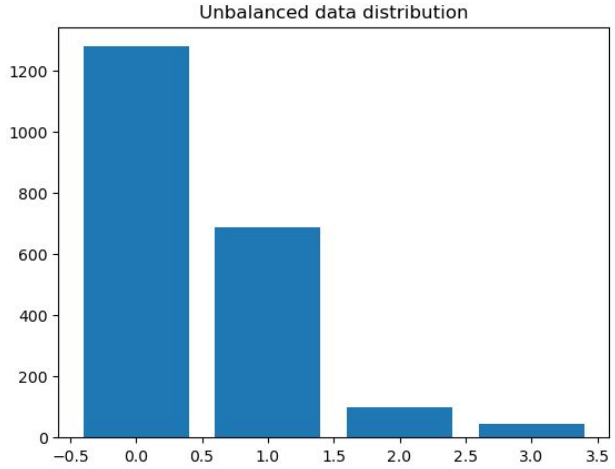
Notes:

- Boosting algorithms (Light BGM, XgBoost) work on decision trees to improve performance and correct errors
- SVMs use support vectors and kernel functions for maximum accuracy
- Classifiers are used for feature importance matrix
- Main aim is to correctly classify perovskites

Data Processing

- Data points were replaced with conventions (Cubic=>0,Magnetic Moment(-)=>0,etc)
- Few features and instance were dropped; such as lattice angles as from this the lattice structures can be determined without ML and some instances didn't have vacancy energy.
- The data distribution was skewed with more than 60% of the data set from the cubic structure and 2% for tetragonal, to balance the data we had to Oversample the data





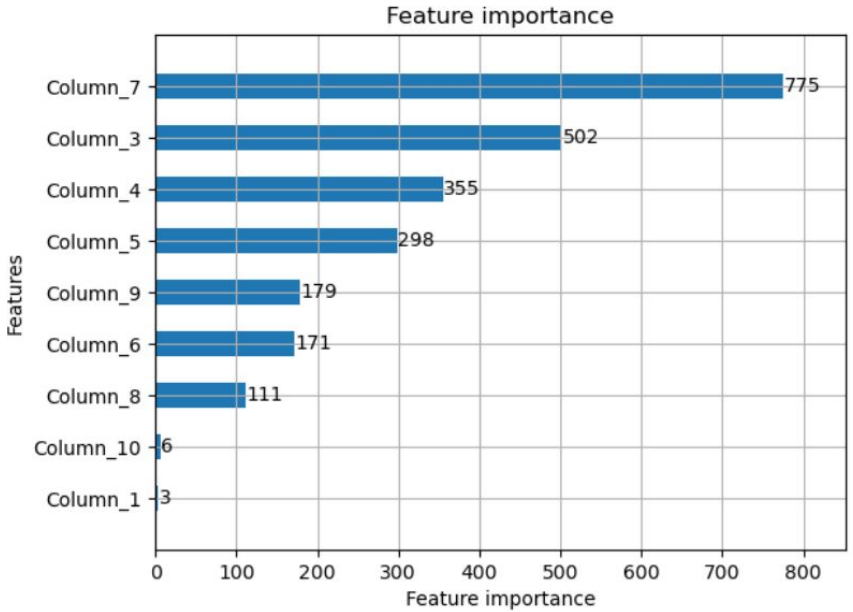
Feature Selection

- Oversampling was carried out with SMOTE algorithm to get a balanced data set. This was also done in **Santosh and Taher's** work.

- Features such as compound name(as it didn't have any impact) and lattice angles(as ML is not required if we know these parameters) were dropped from the database.

- Radius of A was obtained to be the the prominent feature followed by valency.

- Classification report and Confusion matrix were output for each instance of the hyperparameter to find the better model.



SVM

- Hyperparameters: Penalty Parameter(C), Kernel (linear,RBF,sigmoid, polynomial) and gamma($1/\sigma^2$). (Degree also in case of polynomial kernel)
- Tuning using Grid Search
- Result: Penalty Parameter obtained was 219.5,with RBF kernel and gamma as 0.005.
- Accuracy: 91.33%

Actual \ Predicted	Cubic	Orthorhombic	Rhombohedral	Tetragonal
Cubic	312	6	8	1
Orthorhombic	3	289	28	15
Rhombohedral	1	41	272	4
Tetragonal	0	4	1	307

Weighted SVM

- Hyperparameters: Penalty Parameter(C), Kernel (linear,RBF,sigmoid, polynomial),weights for each instance and gamma($1/\sigma^2$). (Degree also in case of polynomial kernel)
- Result: Penalty Parameter was 208, RBF kernel, weights are 5 for naturally occurring,1 for artificial and 0.5 for SMOTE generated,gamma is 0.01.
- Accuracy: 92.03%

Actual \ Predicted	Cubic	Orthorhombic	Rhombohedral	Tetragonal
Cubic	312	5	9	1
Orthorhombic	3	295	24	13
Rhombohedral	1	38	275	4
Tetragonal	0	4	1	307

Light Gradient-Boosting Machine(LGBM)

- Hyperparameters: height of decision tree, learning rate.
- Results: Optimum height was obtained to be 7 and learning rate was obtained to be 0.08.
- Accuracy:94.32%

Actual \ Predicted	Cubic	Orthorhombic	Rhombohedral	Tetragonal
Cubic	328	6	3	2
Orthorhombic	0	279	23	9
Rhombohedral	2	17	279	4
Tetragonal	0	1	3	323

Extreme Gradient Boosting(XGBoost)

- Hyperparameters: height of decision tree, learning rate, and number of epochs.
- Results after hypertuning: No of epochs was obtained as 20, learning rate as 1.25 and tree height 2.
- Accuracy:94.13%

Actual \ Predicted	Cubic	Orthorhombic	Rhombohedral	Tetragonal
Cubic	327	6	4	2
Orthorhombic	0	278	25	8
Rhombohedral	2	26	268	6
Tetragonal	1	2	2	322

Summary and Future Goals

- The Baseline was successfully built and implemented for oxide perovskites.
- Light GBM was obtained to be most accurate classifier with 94.32% accuracy.
- The future implementation includes implementation of the model for halide perovskites also, we expect that in case of halides we would have to more feature importance to electronegativity of the instance.
- This will lead to a changes in the hyper parameters and weight functions.

Accuracy vs Model	
Model	Accuracy
SVM	91.33%
Weighted SVM	92.03%
Light GBM	94.32%
XGBoost	94.13%