21.01.2025

Final Project: Predicting Thermodynamical Stability of Perovskite Oxides

Submitted by: Umut Tutal

1. Introduction & Goal of the Project

Perovskite structures consist of a large A-site ion (typically a larger cation, such as a rare earth or alkaline earth metal) and a smaller B-site ion (usually a transition metal); distinctly for perovskite oxides (ABO_3), structure is surrounded by oxygen atoms arranged in a framework of BO_6 octahedra.

The structure's flexibility allows for diverse element substitution at A and B sites, enabling tailored electronic, magnetic, and optical properties for specific applications. The BO_6 octahedral network plays a critical role in the overlap of electronic orbitals. Specifically, the d-orbitals of the B-site transition metals overlap with the p-orbitals of oxygen, facilitating strong electron interactions. These electron interactions are essential for controlling properties such as magnetism, electrical conductivity, and catalytic activity (e.g., for oxygen reduction or evolution reactions). The oxygen lattice in perovskites is also highly dynamic and supports the formation and mobility of oxygen vacancies. These vacancies are particularly important in applications like solid oxide fuel cells (where oxygen ions move through the lattice) and catalysis (where oxygen vacancies serve as active sites for chemical reactions). This ability to control oxygen vacancy concentrations further enhances the functionality of perovskite oxides in energy and environmental technologies. In addition, the structure's adaptability allows for tunable polarization and bandgap. This is critical for ferroelectric applications and photovoltaics (e.g., in perovskite solar cells).

Compositional flexibility of perovskite structures is a key factor enabling the properties to be tailored for diverse applications. My research question arises from the need to push this compositional flexibility to its limits, as fully leveraging this tailorability requires a wide exploration of their compositional space. The more stable perovskite structures we can identify, the greater the opportunities for functional materials design. My goal in this direction is to predict the thermodynamic stability of perovskite oxides. Specifically, I am conducting a classification study based on their energy above convex hull values, which serves as a key indicator of stability.

The 2017 study of Wolverton and Emery forms the baseline of my dataset, authors conducted DFT calculations for various properties of perovskite oxides. Additionally, I calculated different features by using the properties of constituent elements, the data for elemental properties are acquired from PubChem Database. My final feature set is presented in Table 1.

Table 1. Complete dataset after additional features.

e_form: Formation Energy (eV)
vpa: Volume per Atom (A^3/atom)
gap pbe: Bandgap (eV) from PBE calculations
Lattice Parameters a,b,c (angstrom)
e_form oxygen: Formation Energy of Oxygen Vacancy (eV)
Weighted Atomic Mass
Electronegativity Difference (A-B)
Electronegativity Ratio (A/B)
Ionization Energy Difference (A-B)
Mean Lattice Parameter
lowest distortion: Local distortion of crystal structure with lowest energy among all considered distortions. Encoded into four binary categorical variables.
(Target variable) **e_hull**: Energy Above Convex Hull (eV)

Several studies from literature have acknowledged that a cutoff of 0.04 eV for e_hull provides a reasonable threshold for distinguishing stable materials. Thus, I have decided to use the same cutoff for stability classification of perovskite oxides; labeling the inputs with e_hull values larger than 0.04 eV as not stable, and the inputs with e_hull values smaller than 0.04 eV as stable.

2. Previous Findings & EDA and Dimensionality Reduction

Features of electronegativity ratios and differences of A and B site elements had contained missing values since elemental data of PubChem is not complete especially for elements later in the periodic table. Missing values constituted approximately 17% of these two features. I used KNN imputation to handle for the missing values and I applied standard scaling prior to imputation. Scaling is crucial before KNN imputation since it relies on calculating distances between data points, features with larger ranges could dominate the calculation without scaling. I chose standard scaling since it is less sensitive to outliers compared to min-max scaling.

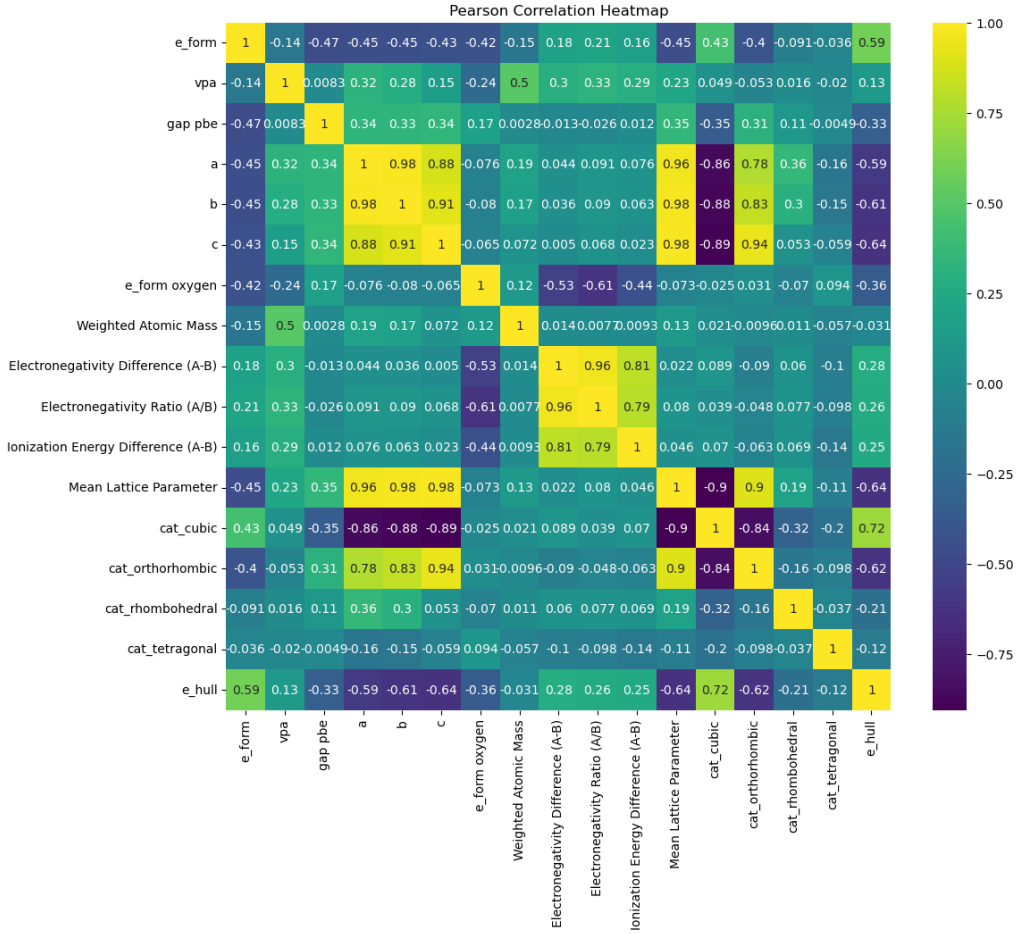

Figure 1. Pearson Correlations of dataset.

Pearson correlation coefficients of my dataset are presented in Figure 1. Some notable deductions can be made using the heatmap. Formation energy of oxygen vacancy has a moderate negative correlation with properties like e_form and e_hull, indicating its importance in stability; as the structure become more stable, energy required to form an oxygen vacancy increases. My target e_hull is influenced by a combination of formation energy, lattice parameters, and electronic properties, with no single feature dominating its behavior. There is a high positive correlation between e_form and e_hull, indicating they are related but not identical. Formation energy measures the stability of a compound relative to its elements, on the other hand, energy above convex hull measures the stability of a compound relative to all other possible phases in the system. Formability of an ABO_3 compound does not necessarily mean that it will form a perovskite structure. This is the reason why e_hull is the stability determining property, and why e_form is an important feature of it. For a perovskite structure, there is a degree of tolerance to symmetrical distortions. A perovskite structure ideally has a cubic symmetry, but orthorhombic, rhombohedral and tetragonal symmetries are also possible. Another important deduction is the negative correlation between lattice parameters and e_hull, this is an expected behavior since smaller unit cells may lead to higher lattice strains, especially if there is poor matching of ionic radii of A and B sites.

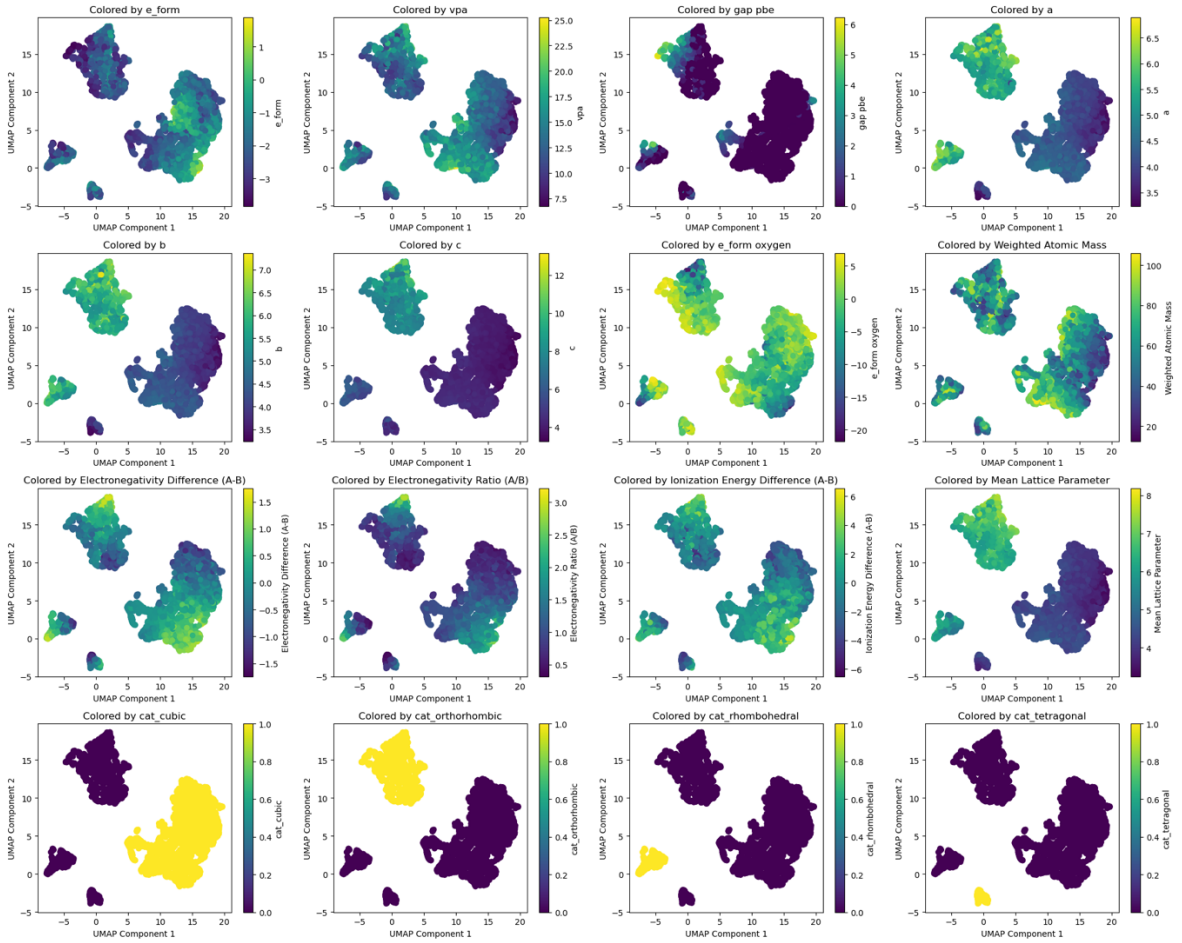

Figure 2. UMAP clustering, colored by each feature.

Clustering results presented in Figure 2. also contain useful information. There are 4 natural clusters in my dataset and highly dominated by the symmetry. Structural properties and electronic properties dominating different clusters. Semiconductor perovskites also belong in their own cluster.

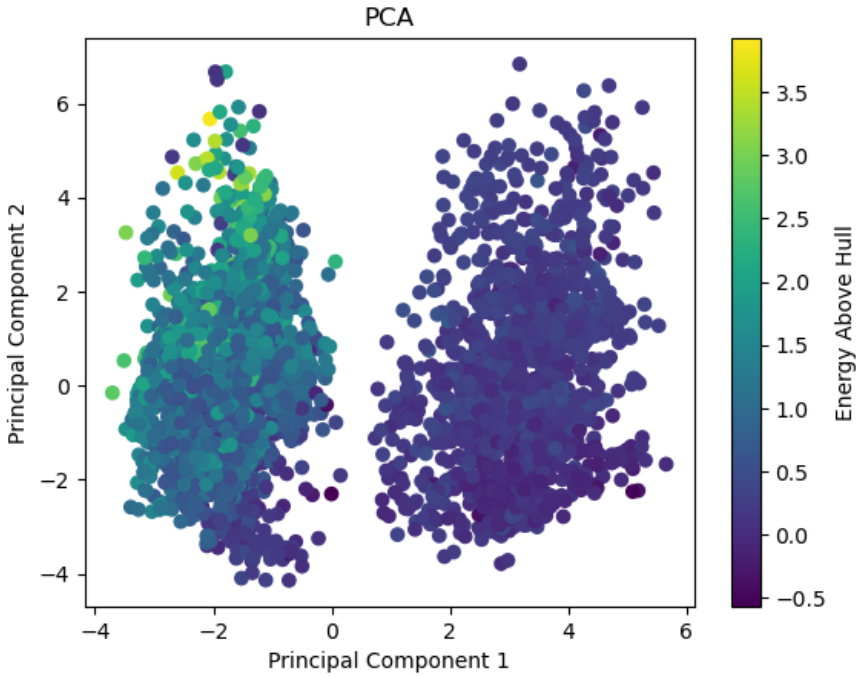

Figure 3. Principal Component Analysis.

PCA results are presented in Figure 3. For visualization purposes, I only used two principal components, while PC1 heavily contains the structural features, PC2 mostly contains the electronic properties. Formation energy is mostly contained in PC1 but a decent amount of it is also represented in PC2. Loadings are clearly presented in the Jupyter notebook of the assignment.

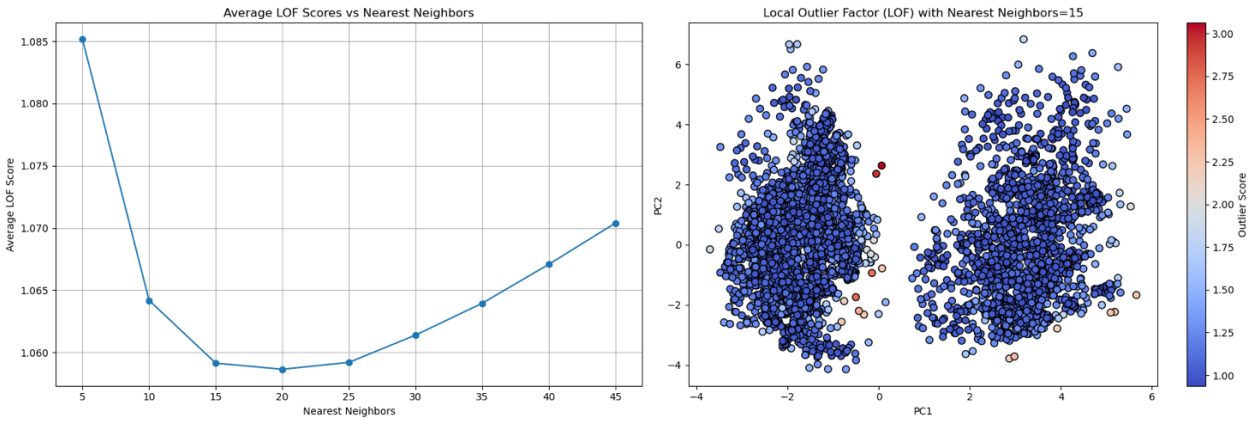

Figure 4. Multivariate Outlier Analysis, Local Outlier Factor

I did a multivariate outlier analysis using LoF and visualized it with first two principal components, presented in Figure 4. I have decided for the number of nearest neighbors by using the elbow method. According to LoF, there are 79 outliers which makes up 1.61% of my dataset. Outliers are perfectly distinguishable in 2D, meaning that their electronic properties (mainly represented in PC2) are

inconsistent with their structural properties (mainly represented in PC1). They will be investigated further while considering their stability; because if they are stable, it means they contain valuable information for my study. I applied Yeo-Johnson transformation to examine the behavior of outliers, but it really did not give a meaningful change. I believe it is due to the absence of skewness in most of my features, except for the bandgaps. Semiconductor perovskites are underrepresented in my dataset and constitute the majority of the outliers. Since I can clearly observe the behavior of these outliers and since I will probably use a tree-based model or a gradient boosting model (which are both robust to outliers) for my classification study, I am leaving the outliers as they are.

3. Model Development and Evaluation

For modelling, I have decided to use my principal components to avoid curse of dimensionality. My original feature set has 16 dimensions which might degrade the model performance.

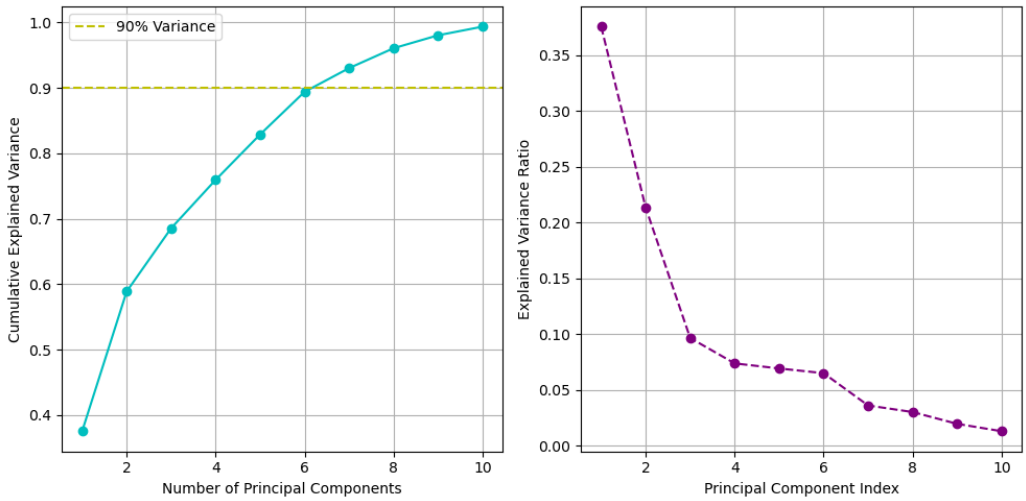

Figure 5. Decision for optimal number of principal components.

For this purpose, I have calculated the number of principal components required to explain at least 90% variance, I also used the elbow method for a better decision. It has been concluded that the optimal number of principal components is 7 and the results are presented in Figure 5. Additionally, I have intended to eliminate -or at least reduce- the potential redundancy by using principal components as features, since redundancy may cause my model to memorize and result with overfitting.

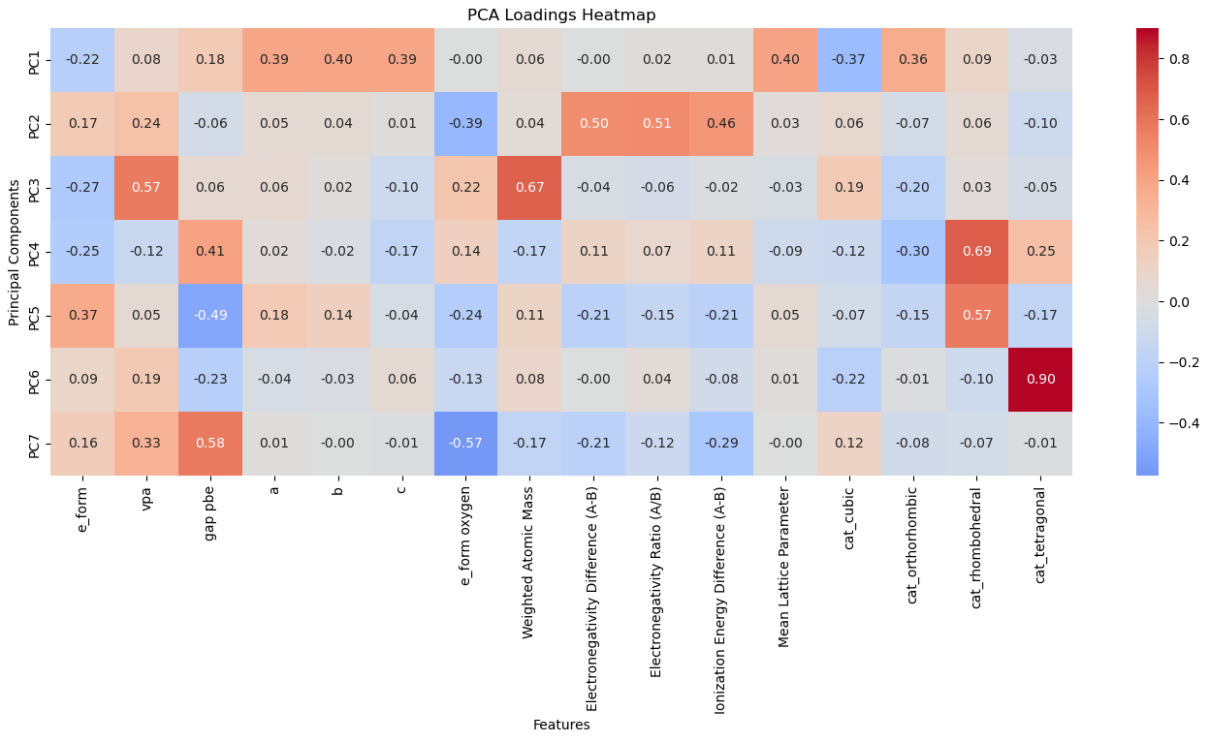

Figure 6. Loadings of seven principal components.

A heatmap is plotted for better visualization of the loadings of principal components and presented in Figure 6.

I have applied a stratified train-test-validation splitting with a ratio of 0.70:0.15:0.15. I have decided to use a validation set to use early stopping methods after my first experience with modelling. At my first attempt, I noticed that any model that I used is memorizing the data, so I have decided to proceed with early stopping (along with regularization) to prevent overfitting.

I have conducted a baseline evaluation for 4 classification algorithms: Logistic Regression, Random Forest, Support Vector Classifier and XGBoost. I used all four models simultaneously without any hyperparameter tuning to get an idea about how each model performs.

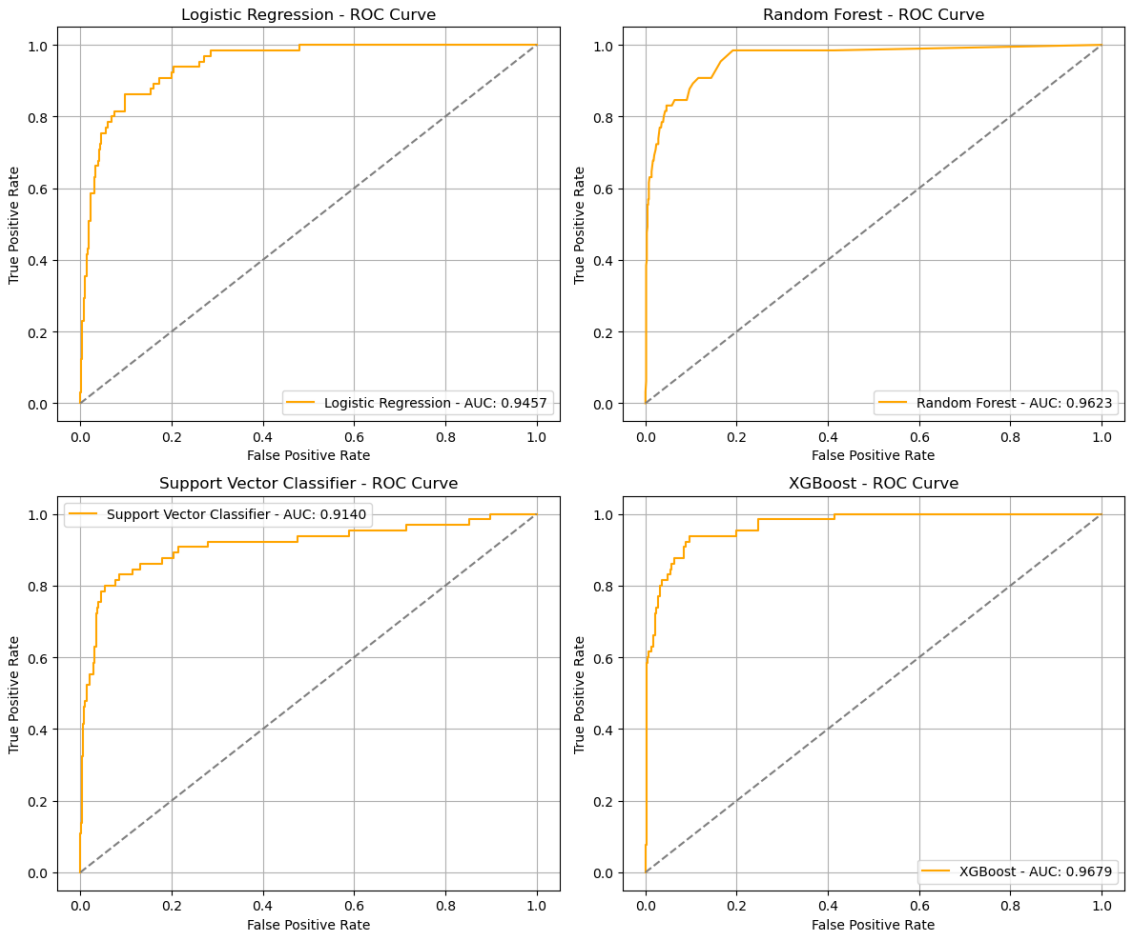

Figure 7. ROC curves from baseline evaluation.

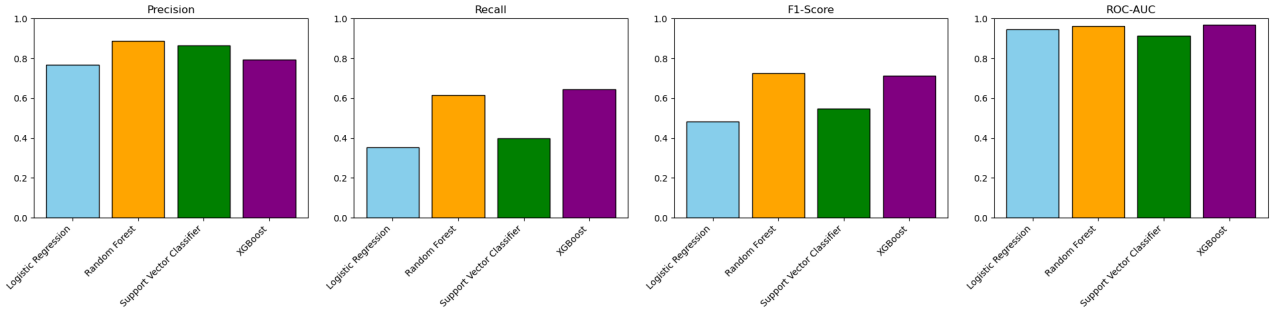

Figure 8. Bar plots for Presicion, Recall, F1-Score, ROC-AUC metrics from baseline evaluation.

The results of the baseline evaluation are presented in Figure 7. and Figure 8. It can be observed that the Random Forest and XGBoost algorithms are the best performing, especially considering recall values. Recall and precision are equally important metrics for my study, since false positives (precision) may result with waste of time and experimental resources and false negatives (recall) may result with missing a perovskite forming composition. Thus, F1-Score is selected as the determining evaluation metric.

As I proceed, I have applied hyperparameter tuning only for the qualified models. I have performed a grid search with 5-fold cross validation to obtain the optimal hyperparameters. After hyperparameter tuning with grid search and early stopping, it has been concluded that XGBoost is the best performing model and the most suitable for my dataset. Thus, only XGBoost algorithm will be discussed in the continuation of this assignment, necessary information about the hyperparameter tuning of Random Forest is available in the Jupyter notebook.

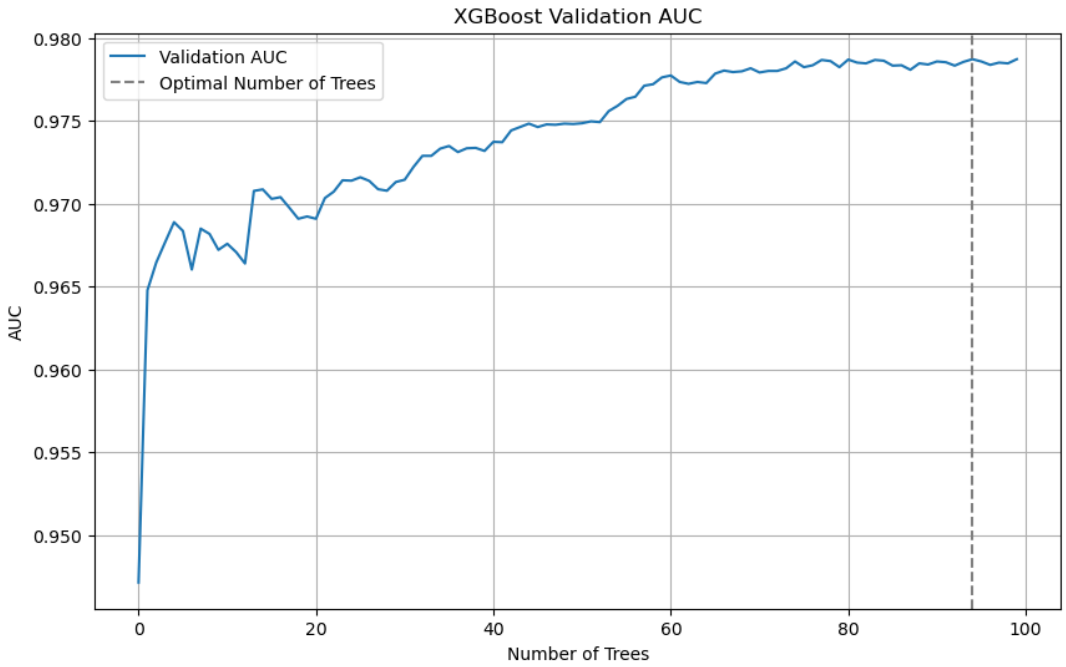

Figure 9. Early stopping of XGBoost Model.

After hyperparameter tuning of XGBoost model, I have used an early stopping method (with optimal hyperparameters) to decide for the final hyperparameter, number of trees used. The results are presented in Figure 9. The algorithms simultaneously calculating the area-under-curve using the validation set while displaying the progress. After reaching a plateau where AUC is not increasing with increasing the number of trees, the algorithm concludes the training.

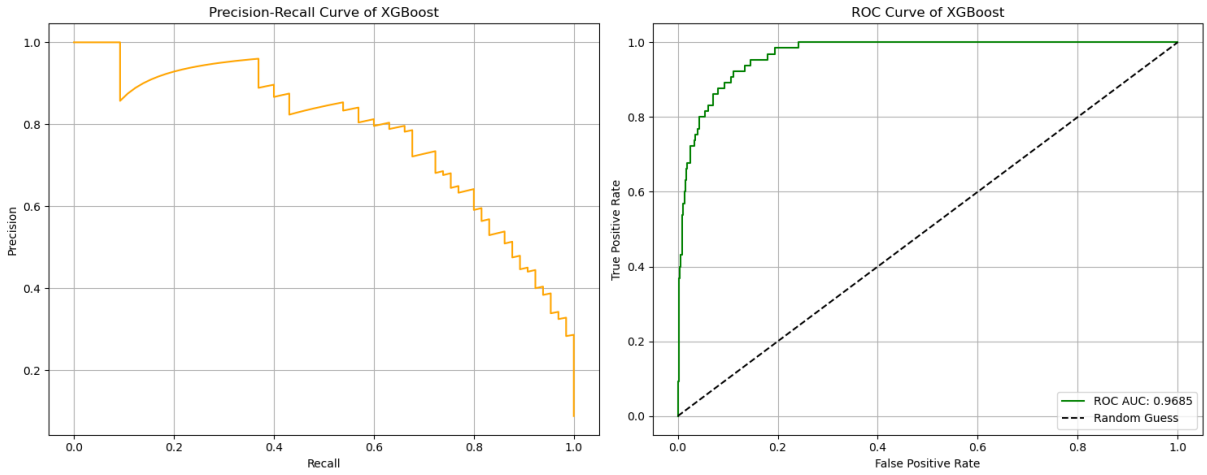

Figure 10. Precision-Recall and ROC curves for XGBoost Model.

With using the optimal number of trees I have obtained by early stopping, I have trained the final XGBoost model for my project. The evaluation metrics of the model is presented in Figure 10. I suspected from an overfitting again, so I checked for the evaluation metrics of the training set too, (in order to see if they are all equal to 1 which may indicate overfitting) however, the results are very similar with the test set, meaning that early stopping was successful for my case. All in all, these metrics are still too high and indicating a redundancy in features, if not overfitting.

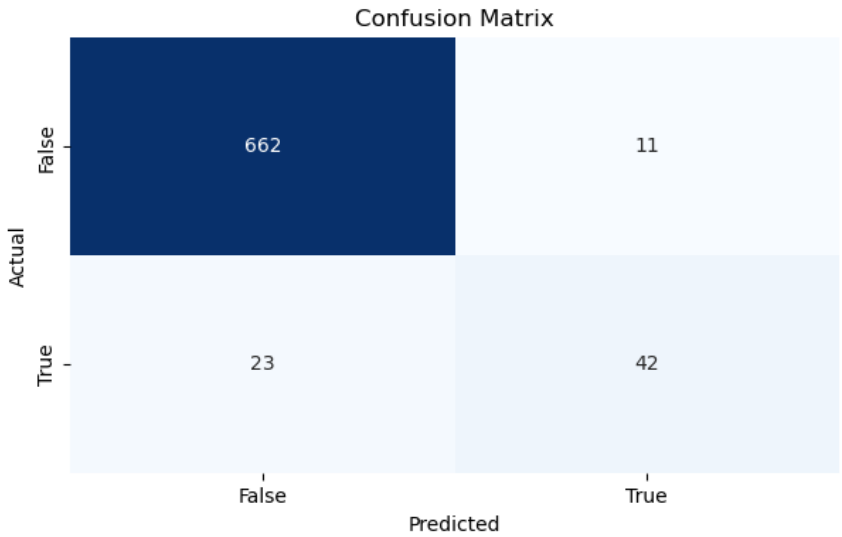

Figure 11. Confusion Matrix of XGBoost Model.

Confusion matrix of the final XGBoost model is presented in Figure 11. It demonstrates the distribution of actual positives, actual negatives, false positives and false negatives in the test set.

4. Model Interpretability and Feature Importance

Figure 12. Shapley Analysis.

The results of Shapley analysis on features are shown in Figure 12. Since I have used principal components as my features, a Shapley analysis is not as informative as it should be, but it is still possible to interpret the results when examined in accordance with loadings presented in Figure 6.

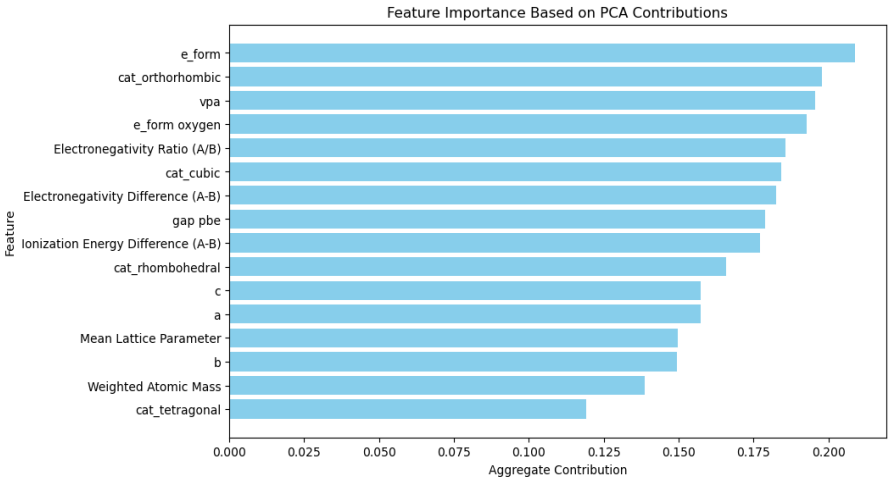

Figure 13. Feature Importance based on PCA contribution.

I have also tried to calculate feature importance based on their contribution to the explained variance. For this purpose, I did a matrix multiplication such that:

$$\text{Feature Importance} = [\text{Loadings}] \times [\text{Explained Variance}]$$

This results in a vector of size equal to the number of original features, where each entry quantifies the overall contribution of that feature to the variance captured by the PCs.

5. Reflection and Future Work

In my opinion, the most crucial deficit of my project is the lack of ionic radii of A and B site elements, since I have learned through my research that they are highly important features. Although it is not possible for me to obtain the exact results, it was possible to approximate them using Shannon Radii Table. The table approximates ionic radius for each element by their coordination number and oxidation state in a certain structure. I have been trying to obtain the exact CNs and oxidation states of A and B sites for my perovskite oxides by using Pymatgen modules, however, I was able to calculate only for small fractions of my dataset and even those are computationally expensive.

Another future work may be mapping a stability space while highlighting practically important features such as bandgap for photovoltaic applications or oxygen vacancy formation energy for catalytic applications. Such a mapping would be helpful for tailoring and designing for specific applications.

The most challenging part, on the other hand, was to prevent overfitting. I had to make a lot of research about hyperparameter tuning and regularization parameters, also I had to use an early stopping algorithm with a validation set which I didn’t use at my initial attempts.

References to Data Sources:

1. Emery, A., Wolverton, C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites. Sci Data 4, 170153 (2017). https://doi.org/10.1038/sdata.2017.153

2. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2025). PubChem 2025 update. Nucleic Acids Res., 53(D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059