REPORT

UNVEILING BAND GAPS IN PEROVSKITE OXIDES FOR NEXT-GEN ELECTRONICS

Problem Statement :

Perovskite oxides are a class of materials with tunable electronic properties, making them valuable for solar cells, LEDs, semiconductors and optoelectronic devices. A key property that determines their usefulness is the band gap (Eg), which defines whether a material is an insulator, semiconductor, or conductor. We aim to use machine learning (ML) models to predict band gaps quickly and efficiently.

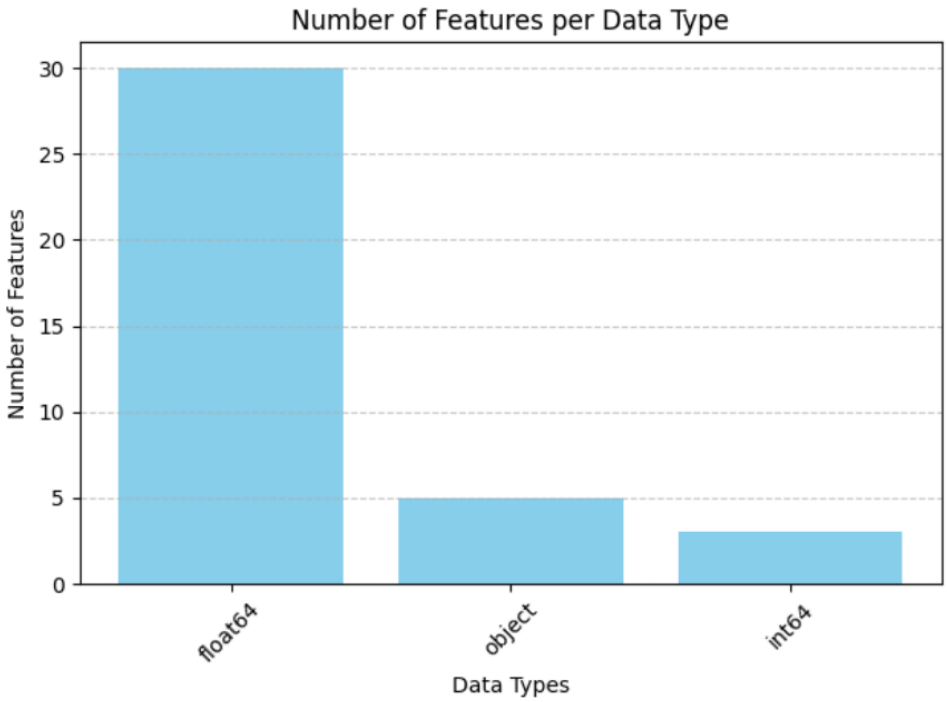
The Solution to analyze, preprocess, and train ML models using a real-world materials science dataset and compete to achieve the most accurate predictions.

The Dataset :

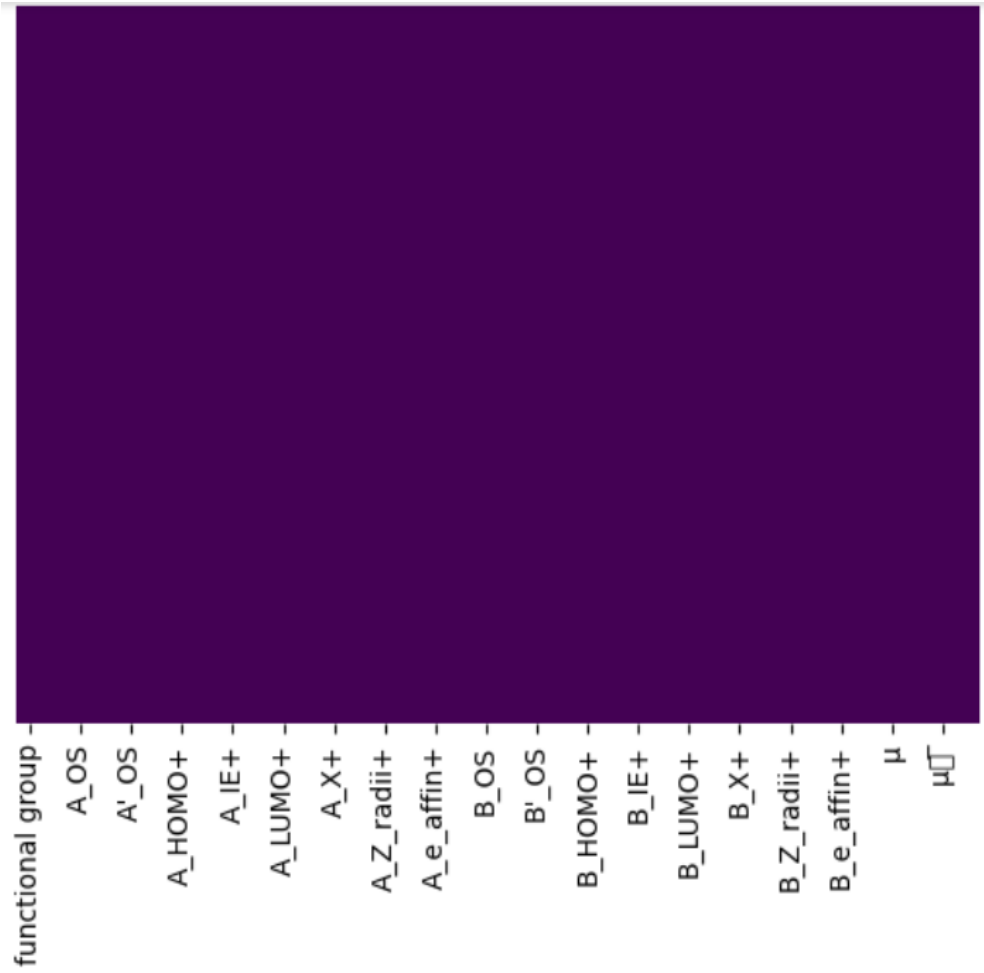
The dataset provided is an excel file containing various compounds of the Perovskite class with varying band gap and other features.

Key Insights from the dataset :

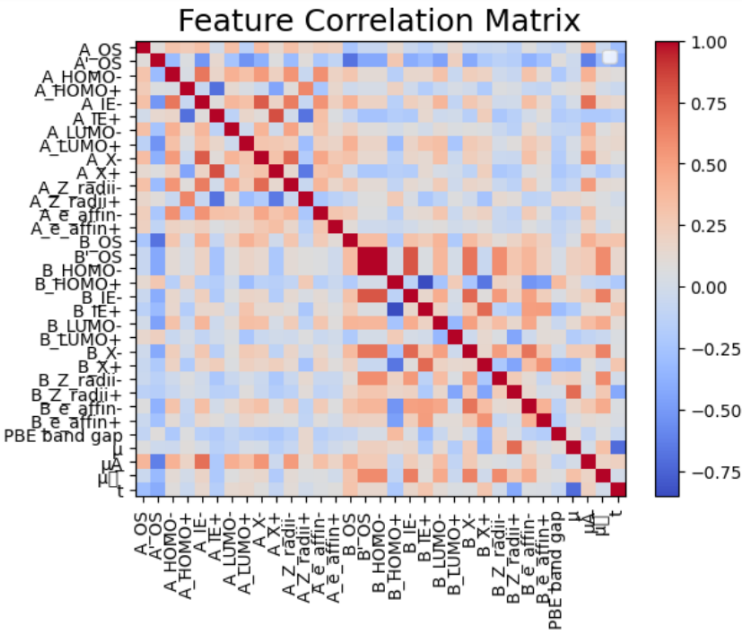
1. The dataset consists of 5151 rows and 38 columns.
2. Data types Distribution:
    a. Float64 : 30
    b. Int64 : 3
    c. Object : 5



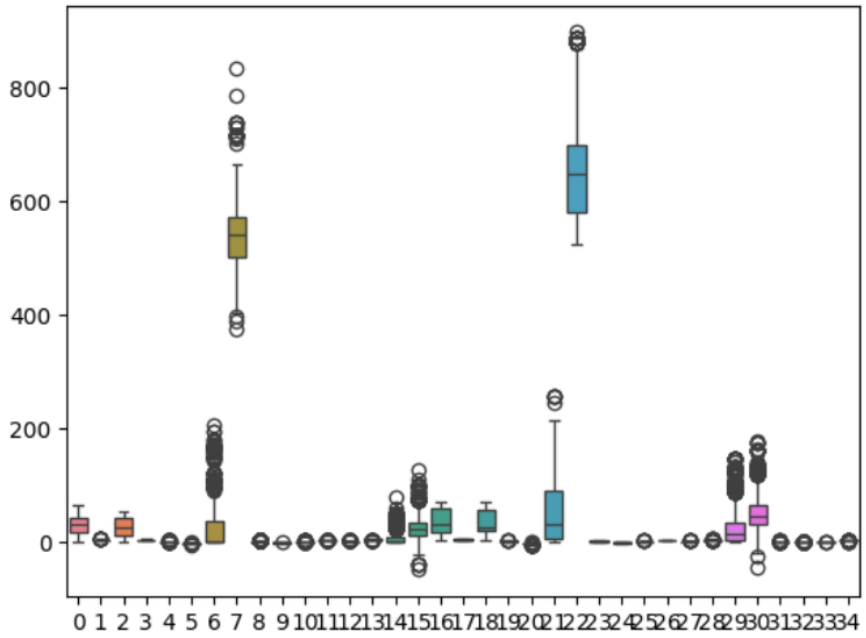
3. There are no null or missing values in the dataset.



4. The correlation between the features is as follows:



5. The distribution of values of the features is as follows:



The Machine Learning Workflow:

The provided machine learning solution has the following workflow:

1. Importing the libraries
2. Importing the dataset
3. Feature Engineering the dataset
4. Model building
5. Evaluating the model with various accuracy metrics.

The Classification Model :

XGBoost (Extreme Gradient Boosting) is an advanced machine learning algorithm based on **decision tree ensembles**. It's designed to be **high-performance**, **efficient**, and **accurate.** It is used to accurately predict whether a given input is an insulator or not.

XGBoost uses Gradient Boosting, where:

- Each tree is built to minimize a loss function (like squared error or log loss).
- It uses the gradients (slopes) of the loss function to decide how to improve.

Hyperparameters Tuned in the model:

1. n_estimators : **10**
2. criterion : **entropy**
3. random state : **42**
4. max_cat_threshold : **20**

Performance metrics: The model achieved an accuracy score of **0.9992**

The Regression Model

Random Forest Regression is an **ensemble learning method** that builds **multiple decision trees** and combines their outputs to produce **more accurate and stable predictions**.
It's an extension of the **decision tree algorithm**, aiming to reduce **overfitting** and **improve generalization**.

Hyperparameters Tuned in the model:

1. n estimators : **250**
2. random state : **42**
3. maximum depth : **8**

Performance Metrics:

- Mean Squared Error : **0.155**
- R-Squared Error : **0.751**