1. Introduction

This report explains the implementation of a machine learning model to classify and predict the band gap of perovskite materials. The project was part of the **EXCAVATE 2025 Competition**, where the goal was to leverage ML techniques to distinguish between insulators and non-insulators and predict band gaps accurately.

2. Feature Engineering and Data Analytics Insights

Feature engineering played a crucial role in improving the model's predictive power. Here are key transformations and insights derived from post-processing analysis:

a) Key Engineered Features

- **Orbital Gap Difference (orbital_gap_A, orbital_gap_B)**:
    - Captures the energy difference between the highest occupied and lowest unoccupied molecular orbitals.
- **Electronegativity Difference (EA_diff)**:
    - Helps understand charge transfer behavior between A and B cations.
- **Charge Stability Factor (charge_stability)**:
    - Ensures charge neutrality, which is essential for structural stability.
- **Adjusted Tolerance Factor (adjusted_tolerance)**:
    - Refines the stability measure by considering tolerance and octahedral factors.

3. Post-Processing Analysis: Understanding the Role of Features

The post-processing analysis involved interpreting model outputs using **gain-based feature importance, SHAP analysis, and permutation importance**. This helped in identifying the most influential features and understanding their role in band gap prediction.

a) XGBoost Model (Classification)

- **Gain-Based Feature Importance:**
    - charge_stability was the most critical feature for classification.

    - orbital_gap_A, orbital_gap_B, and IE_gap_B also played major roles.
    - Structural properties like mu and tolerance factor had moderate influence.
- **SHAP Analysis:**
    - Features such as IE_gap_B, B_HOMO+, orbital_gap_A, and A_X+ had the highest SHAP values, confirming their global impact.
    - SHAP values indicated that **higher orbital_gap_A values increase the likelihood of a material being an insulator**.
- **Permutation Importance:**
    - orbital_gap_A was ranked the highest, reinforcing the gain-based ranking.
    - B_X+, A_X+, and B_HOMO+ were also influential in classification predictions.
    - EA_diff and charge_stability had lower permutation importance than initially expected, showing their dependency on other correlated features.

b) CatBoost Model (Regression)

- **Feature Importance (Gain-Based):**
    - The most influential features for band gap regression were B', A, Bi, and charge_stability.
    - adjusted_tolerance, B_HOMO+, and B_X+ also had strong contributions.
- **SHAP Summary Plot:**
    - charge_stability, A_X+, and adjusted_tolerance were the most influential in predicting band gap values.
    - The SHAP plots confirmed that **higher values of charge_stability and adjusted_tolerance were linked with lower band gap values**.
    - Certain categorical variables (B', A, Bi) had extreme impacts, emphasizing their role in defining perovskite properties.

c) Key Insights from Post-Processing Analysis

- **Consistency Across Models:**
    - charge_stability was highly influential in both classification and regression, indicating its fundamental role in determining band gap behavior.
    - orbital_gap_A and orbital_gap_B were major drivers in classification but had a **lower impact in regression**, showing that band gap magnitude is influenced by additional factors.
- **Categorical Features Matter:**
    - B', A, and Bi had a disproportionate influence on regression predictions, highlighting the role of **atomic composition**.
- **Feature Dependencies:**
    - The differences in **gain-based, SHAP, and permutation importance rankings** suggest that some features contribute more in direct splits (gain-based), while others influence model behavior through interactions (SHAP values).

4. Model Performance Evaluation

The models were evaluated based on key performance metrics:

a) Classification Model (XGBoost)

- **Accuracy: 94.47%** (High classification performance)
- **Classification Report:**
    - The model successfully separated insulators from non-insulators with strong precision-recall values.

b) Regression Model (CatBoost)

- **Root Mean Squared Error (RMSE): 0.2585** (Low error, strong predictive performance)
- **R^2 Score: 0.9017** (Strong correlation between features and target values)

5. Data Visualization and Insights

Several **data analytics tools** were used to extract insights:

- **Feature Distributions:**
    - Visualizations showed the natural distribution of features like **HOMO, LUMO, and Band Gap**, identifying potential outliers.
- **SHAP Summary Plots:**
    - Confirmed which features had the most impact on model predictions.
    - XGBoost's top SHAP features: IE_gap_B, B_HOMO+, orbital_gap_A.
    - CatBoost's top SHAP features: charge_stability, A_X+, adjusted_tolerance.
- **Feature Importance Plots (XGBoost & CatBoost):**
    - Displayed charge_stability, orbital_gap_A, and IE_gap_B as key influences.
    - CatBoost emphasized B', A, and Bi in the prediction.
- **Permutation Importance Plots:**
    - Confirmed orbital_gap_A was crucial, while EA_diff and charge_stability had a lesser effect than initially expected.

These insights helped refine feature selection, ensuring only the most relevant variables were included.

6. Conclusion

Post-processing analysis provided deeper insights into feature relationships and model behavior. **Feature engineering choices significantly improved both classification and regression models.** Future improvements could focus on:

- Fine-tuning hyperparameters to further optimize performance.
- Exploring deep learning approaches for enhanced predictions.
- Incorporating more domain-specific materials science knowledge into feature engineering.

Name: Radhika Panchal
College: NMIMS MPSTME