**Report on Mini Project Machine Learning -I (DJS23DSL402)**

**AY: 2024-25**

**Band Gap Analysis of Perovskite Materials**

**Soham Kishor Walam : 60009230094**

**Guided By**

**Prof. Mridul Akradi**

**CHAPTER 1: INTRODUCTION**

The rapid development of materials science has brought forward an increasing demand for computational methods to predict material properties. With the ever-expanding repository of synthesized materials, traditional experimental techniques for characterizing electronic properties are no longer feasible on a large scale due to time and cost constraints. Among the various properties, one of the most critical is the electronic band gap, particularly the PBE (Perdew–Burke–Ernzerhof) Band Gap. This property plays a vital role in determining whether a material functions as a conductor, semiconductor, or insulator. Accurate band gap prediction can inform the design of materials for applications in electronics, photovoltaics, sensors, and energy storage systems.

Computational approaches like Density Functional Theory (DFT) have been widely used to estimate band gaps; however, these methods are often computationally intensive and may still suffer from limitations in accuracy, especially for complex compounds. As a result, there is a growing shift towards data-driven methods that can complement and, in some cases, outperform traditional approaches.

Machine learning (ML) has emerged as a powerful alternative due to its ability to analyze large volumes of high-dimensional data and uncover hidden patterns. ML models can learn relationships between material descriptors and target properties, enabling rapid prediction without requiring detailed quantum mechanical simulations. Such models can be retrained and fine-tuned as more data becomes available, improving their performance over time.

In this study, a comprehensive dataset of material descriptors was utilized to explore the application of ML in electronic property prediction. The dataset encompasses a range of chemical, physical, and structural features believed to influence a material’s electronic behaviour. The primary objective of this project is twofold: firstly, to classify materials as insulators or non-insulators based on their PBE Band Gap values using supervised classification techniques; and secondly, to develop a regression model capable of predicting the exact band gap values for those identified as insulators. This dual-pronged approach enables both qualitative and quantitative assessment of materials, making the method versatile and applicable across a broad spectrum of material types.

Ultimately, the integration of machine learning into materials informatics paves the way for accelerated discovery and optimization of novel materials. By automating the prediction of complex properties like the band gap, researchers can prioritize promising candidates for experimental validation, significantly reducing development cycles in materials science.

**CHAPTER 2: DATA DESCRIPTION**

The dataset, named "dataset\_excavate.xlsx", contains various chemical and physical descriptors related to materials. Key features include structural, electronic, and chemical attributes. The target variable, 'PBE Band Gap', is continuous and crucial for determining the insulation properties of a material.

Initial exploration revealed missing values and a mixture of numeric and non-numeric data. The 'PBE Band Gap' column was particularly inconsistent, requiring conversion of non-numeric entries to Nan and subsequent removal of incomplete rows. Additionally, irrelevant columns such as 'functional group' were dropped to maintain numeric consistency in the feature matrix. A new binary target variable 'is\_insulator' was introduced, where 1 indicates a band gap greater than or equal to 0.5 eV, and 0 otherwise.

**CHAPTER 3: DATA ANALYSIS**

To ensure meaningful and accurate predictions from the machine learning models, rigorous data analysis and preprocessing steps were performed. The raw dataset underwent several key transformations aimed at cleaning, preparing, and understanding the structure and distribution of the data.

Preprocessing:

Effective preprocessing is the foundation of any reliable machine learning pipeline. The following steps were implemented:

* Data Type Conversion: Several columns were initially read as non-numeric types due to embedded strings or formatting inconsistencies. These were systematically converted to appropriate numeric data types to enable mathematical operations.
* Missing Value Handling:Missing entries can introduce bias or reduce the training efficiency of machine learning models. All rows with missing values were removed using the .dropna() function, ensuring that only complete and consistent records were retained for training and testing.
* Feature Normalization:The dataset's feature values had varying scales, which could skew the training process. To address this, StandardScaler from the sklearn.preprocessing module was employed. This standardized the dataset by centring the feature values around zero and scaling them to unit variance, thereby improving convergence in algorithms such as Random Forest.
* Train-Test Split: The processed dataset was divided into training and testing sets with an 80-20 split. This approach ensured that the model could learn from the majority of the data while being evaluated on an unseen subset to assess its generalization capabilities. Separate splits were maintained for the classification and regression workflows.

**Exploratory Data Analysis (EDA)**

A thorough initial assessment of the dataset was conducted to

understand the data structure and uncover basic insights:

* Missing Values and Uniqueness: A check for null values across all columns helped verify data completeness post-cleaning. Unique value counts were analysed to understand the variability of input features.
* Data Overview: The .head() function was used to preview the first few rows of the dataset, offering a snapshot of the data. The .info() function provided essential metadata such as column names, non-null counts, and data types, which further informed the preprocessing pipeline.
* Visualization: Scatter plots were generated to visually assess the performance of the regression model. These plots compared predicted values to actual PBE Band Gap values, allowing an intuitive evaluation of the model’s precision and tendency to over- or under-estimate.

Correlation Analysis

Although no explicit heatmaps or correlation plots were presented in the current analysis, an understanding of feature correlation is critical. Strongly correlated features can introduce redundancy, while weak correlations might reduce predictive power. Future iterations of this project could benefit from:

* Pearson or Spearman correlation matrices to identify collinearity.
* Dimensionality reduction techniques such as Principal Component Analysis (PCA) to enhance model efficiency and interpretability.

Overall, the data analysis phase laid a strong groundwork for the modelling stage by ensuring data integrity and revealing initial patterns within the dataset.

**CHAPTER 4: DATA MODELLING**

**Two models were developed using Random Forest algorithms:**

1. **Classification Model:**
   * Target: is\_insulator (binary)
   * Features: All numeric columns excluding the target and categorical fields
   * Model: RandomForestClassifier
   * Accuracy: Reported using accuracy\_score and classification\_report

|  |  |  |
| --- | --- | --- |
| Metrics | Positive | Negative |
| Positive | 671 | 71 |
| Negative | 23 | 266 |

The model successfully differentiated insulators from non-insulators with a high accuracy rate. Classification metrics such as precision, recall, and F1-score were used to assess performance.

1. **Regression Model:**
   * Target: PBE Band Gap (for insulators only)
   * Model: RandomForestRegressor
   * Evaluation: Mean Absolute Error (MAE) = 0.2271
   * Mean Squared Error (MSE) = 0.1252
   * R² Score = 0.7985

Predictions showed good agreement with actual values, with visual confirmation through scatter plots of actual vs predicted values.

Visualization:  
A scatter plot comparing actual vs predicted band gap values confirmed that the model made close predictions. A diagonal line was added to reflect perfect prediction alignment.

**CHAPTER 5: CONCLUSION**

This study effectively demonstrated the potential of machine learning in predicting electronic properties of materials. The classification model accurately identified insulating materials, while the regression model provided reliable band gap predictions for those insulators.

Key takeaways:

* Proper data cleaning and preprocessing are vital for reliable modelling.
* Random Forest models are robust and perform well for both classification and regression tasks.
* Feature selection and imputation methods could be further optimized in future iterations.

Future work could focus on expanding the dataset, exploring more advanced models such as XGBoost or deep learning, and integrating domain-specific knowledge for improved predictions.