Predicting Band Gaps in 2D Materials using Machine Learning

Beniam Kumela

1 Introduction

Two-dimensional (2D) materials, composed of atomically thin layers, have attracted widespread interest due to their unique electronic, mechanical, and optical properties. Among these, the electronic band gap is particularly important for practical applications such as transistors, photovoltaics, and optoelectronic devices. The experimental determination of band gaps in 2D materials is often resource intensive and slow, as shown in **Figure 1**. Density functional theory (DFT), particularly with the Perdew-Burke-Ernzerhof (PBE) functional, offers a computational alternative but tends to underestimate band gaps, and more accurate methods are computationally expensive. Machine learning (ML) provides a promising complementary approach by leveraging large datasets to rapidly and accurately predict band gaps, enabling accelerated materials discovery. In this work, we examine various ML techniques—random forest, support vector machine, gradient boosting, lasso regression, and artificial neural networks—to identify the most effective methods for predicting band gaps in 2D materials. To mitigate this, random forests combine multiple trees trained on random data subsets, with the final prediction being the average of all individual trees.

2 Computational Methods

2.1 Machine Learning Theory

Machine learning (ML) techniques efficiently predict band gaps using data from density functional theory (DFT) calculations by learning patterns in large datasets. First, we introduce the random forest regression model, which is based on decision trees. A decision tree splits data into subsets according to input features, forming a structure where each node represents a decision rule and each leaf node a predicted outcome, as shown in **Figure 2**. However, decision trees are prone to overfitting, especially when they grow too deep, capturing noise and reducing generalization to unseen data. To mitigate this, random forests combine multiple trees trained on random data subsets, with the final prediction being the average of all individual trees ($\hat{y}_i$), as shown in **Figure 3a** and **Eqn. 1**. This ensemble approach improves stability and accuracy.

$$\hat{y}_{RF} = \frac{1}{T}\sum_{i=1}^{T}\hat{y}_i \tag{1}$$

Second, we introduce the support vector regression (SVR) model, which aims to fit the data points with minimal error by constructing an optimal hyperplane. Unlike linear regression, SVR focuses on balancing accuracy and robustness by introducing a margin of tolerance called the ϵ-insensitive tube as shown in **Figure 3b**. Data points within this tube are considered correctly predicted, as minor deviations from the hyperplane are tolerated. Only data points outside this tube contribute to the error, quantified using slack variables (ξ_i), which measure how far these points deviated beyond the margin. The goal of SVR is to minimize the sum of these slack variables, enabling its focus on significant deviations and improving its generalization. The optimization problem formulation for the SVR model is shown.

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

Where the first term represents the function smoothness, ϵ specifies the width of the ϵ-tube, and C controls the tradeoff between model complexity and penalty for points outside the ϵ-tube. To use the SVR

model, we need a way to map data points to higher-dimensional feature spaces through a nonlinear function $\Phi(x)$. Instead of computing this function $\Phi(x)$ directly, a kernel function $K(x, x')$ is used to compute the dot product of the transformed vectors $\Phi(x)$, $\Phi(x')$ directly in the feature space. The kernel function that we use in the model is the radial basis function, which is commonly used for noisy data sets and is of the following form:

$$K(x, x') = \exp(-\gamma||x - x'||^2) \tag{3}$$

Third, we introduce the least absolute shrinkage and selection operator (Lasso) regression model. It seeks to minimize the residual sum of squares while keeping the sum of the absolute values of the coefficients less than a fixed value. The formulation of Lasso regression is shown below:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

Where β represents the estimated coefficients from applying Lasso regression, λ dictates the strength of a penalty for the absolute size of coefficients, x are the predictor values, and y are the observed values.

Fourth, we introduce the gradient boosting regressor. It builds a predictive model by combining the outputs of multiple weak models, like decision trees. In a sequential manner, each new tree corrects the errors (residuals) made by the previous ones. For the sake of brevity, we will not go into exactly how the algorithm works but a representative schematic is shown in **Figure 3c**.

Lastly, we introduce the artificial neural network (ANN), a computational model inspired by the human brain, shown in **Figure 3d**. It consists of layers of interconnected neurons that process data sequentially. The input layer receives features from the dataset, which are passed through hidden layers where neurons compute weighted sums of inputs, activated by a ReLU function to introduce non-linearity as shown in **Figure 7**. This allows the network to learn complex patterns. The output layer produces predictions, such as the band gap values. The model then calculates the loss, and through back-propagation, adjusts the weights to minimize the error. This forward and backward propagation continues for multiple iterations until the model converges.

2.2 Data Preparation and Workflow

The dataset was taken from 2D Materials Encyclopedia where DFT calculations were performed in the Vienna Ab Initio Simulation Package (VASP) software on 6,351 2D materials with PBE approximation for the exchange-correlation functional and projector augmented wave (PAW) method for electron-ion interaction [7]. We compute 136 elemental and compositional features using the Materials Agnostic Platform for Informatics and Exploration (Magpie) as implemented by the matminer.featurizers python package [8, 9]. For all models, we use an 80-20 train-test split when training and validating data as well as a 3-fold cross-validation and R^2 scoring when performing hyperparameter tuning. Exhaustive randomized and grid search hyperparameter tuning was performed for all regression models and a Bayesian optimization scheme across a parameter grid was used for the neural network.

3 Results and Discussion

As shown in **Figures 5a-b**, the dataset has a large range of band gap values heavily skewed towards values less than 2 eV (metallic and semiconductor materials). The correlation matrix, shown in **Figure 5c**, highlights that some of the features, like different statistical properties of the same elemental feature, are highly correlated. This correlation may have an adverse effect during model development and training.

We trained several regression models, including random forest (RFR), support vector (SVR), gradient boosting (GBR), and artificial neural network (ANN) models. The performance of these models, based on the coefficient of determination (R^2) and root mean squared error (RMSE) is shown in **Table 1**.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

Among them, the GBR model outperformed the others, while the Lasso regressor demonstrated the poorest performance. For noisy, tabular datasets, the GBR is often reported as a "silver bullet" for predictive accuracy aligning to these results [10]. Band gap values most likely have a non-linear relationship with the features which is probably why Lasso performed so poorly. To contextualize the model accuracies with existing literature, a study predicting E_g from a dataset of over 4,000 DFT-PBE band gaps for binary compounds achieved a model accuracy of 0.81 with a GBR model [14]. Another study, which used a dataset of 3,896 experimental non-zero band gaps of inorganic solids, reported a model accuracy of 0.90 with a support vector classification (SVC) and SVR workflow [15].

Table 1: Ranking of model accuracies using R^2 and RMSE metrics.

Model	R^2	**RMSE**
GBR	0.7133	0.7676
RFR	0.7097	0.7725
ANN	0.6674	0.8268
SVR	0.6571	0.8395
Lasso	0.4462	1.0669
Linear	0.4311	1.0885

We also identified the top five features influencing E_g predictions as shown in **Table 2**. The top three features of the RFR and GBR models were similar, emphasizing melting temperature (T_m) and electronegativity (χ). In contrast, the SVR and ANN models highlighted features such as oxidation state and valence electrons. A closer look at the high-performing models reveals some physical intuition behind these feature rankings. Experimental studies suggest that T_m is significant because, as a material's size decreases, local bond strength may increase due to broken bonds, leading to a rise in E_g. The reduction in bond count also impacts cohesive energy, which lowers T_m. Additionally, non-linear empirical relationships between χ and E_g have been reported in several experimental studies [12, 13]. Supporting this, a machine learning study predicting E_g from binary compounds also ranked χ among its top 5 features [14].

Table 2: Ranking of top 5 most important features for the different models tested. Equivalent features between RFR and GBR colored red, between SVR and ANN colored blue. Adapted from **Figure 7a-d**

Importance	**RFR**	**SVR**	**GBR**	**ANN**
1	Mean T_m	σ ox. state	Mean T_m	σ ox. state
2	Avg. Dev χ	Range NdValence	Mode T_m	Mean NpValence
3	Mode T_m	Avg. Dev NValence	Mean χ	Mean Covalent Rad.
4	Range Column	Mean NpValence	Mean NUnfilled	Min. Gsvol_pa
5	Min T_m	Mode NsUnfilled	Avg. Dev Column	Mode Gsvol_pa

In future studies with greater computational resources, we would aim to conduct a more comprehensive hyperparameter tuning process. Additionally, we would focus on using features with lower correlations to streamline the data featurization stage. A promising avenue could involve encoding atomic position data. For instance, analyzing the RFR model regression in **Figure 6** reveals several incorrect predictions of zero band gap values. This issue might be addressed by first classifying materials as metallic or non-metallic and then applying regression to the non-metallic subset, potentially improving the accuracy of these predictions. Furthermore, we would consider using a different dataset specifically tailored for 2D materials. It is well-known that DFT-PBE often underestimates the electronic band gaps of semiconductors and insulators. The PBE functional's inability to fully account for strong electron-electron interactions can introduce discrepancies between the features and the predicted band gaps. Despite these limitations, the current study provides a reasonably accurate model that captures key elemental and compositional features.

These features not only exhibit meaningful correlations with band gaps but also align with established physical intuition, offering valuable insights for further research.

4 Appendices

4.1 Code

The following code is included:

1. eda.ipynb → exploratory data analysis and dataset featurization
2. model.ipynb → model development and valiadation

4.2 Supplemental Figures

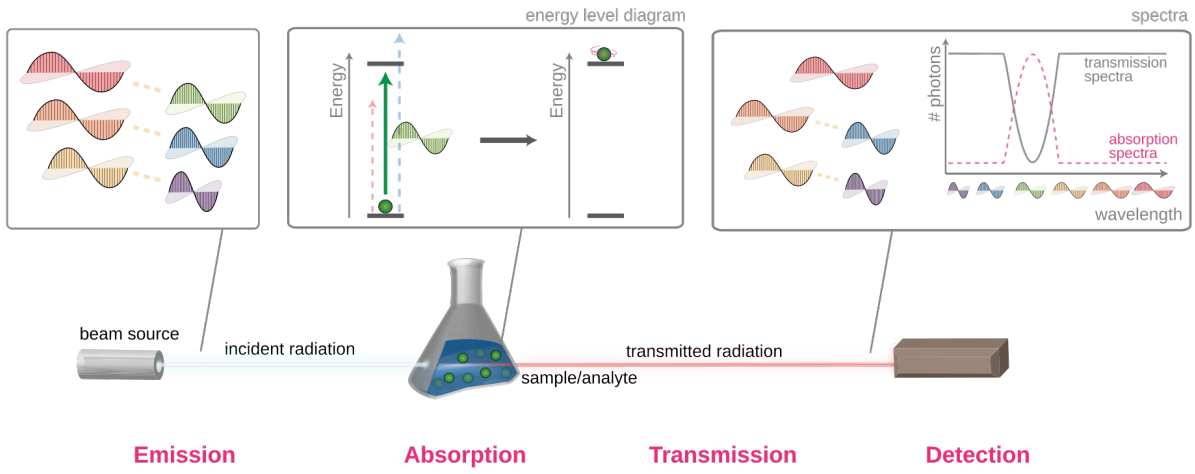


Figure 1: Overview of optical absorption spectroscopy for band gap determination [1]. A white light source emits a spectrum of wavelengths onto a sample, where photons matching the material's energy gap are absorbed, exciting the molecules. By analyzing the transmitted light and comparing it to the incident beam, the absorption spectrum and band gap are obtained.

4.3 Bibliography

References

[1] Absorption Spectroscopy. https://en.wikipedia.org/wiki/Absorption_spectroscopy (accessed 2024-12-01).

[2] Decision Tree. https://www.smartdraw.com/decision-tree/ (accessed 2024-12-01).

[3] Chaya. Random Forest Regression. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84 (accessed 2024-12-01).

[4] Support Vector Regression (SVR). https://www.youtube.com/watch?v=kPw1IGUAoY8 (accessed 2024-12-01).

[5] GeeksforGeeks. Gradient Boosting in ML. https://www.geeksforgeeks.org/ml-gradient-boosting/ (accessed 2024-12-01).

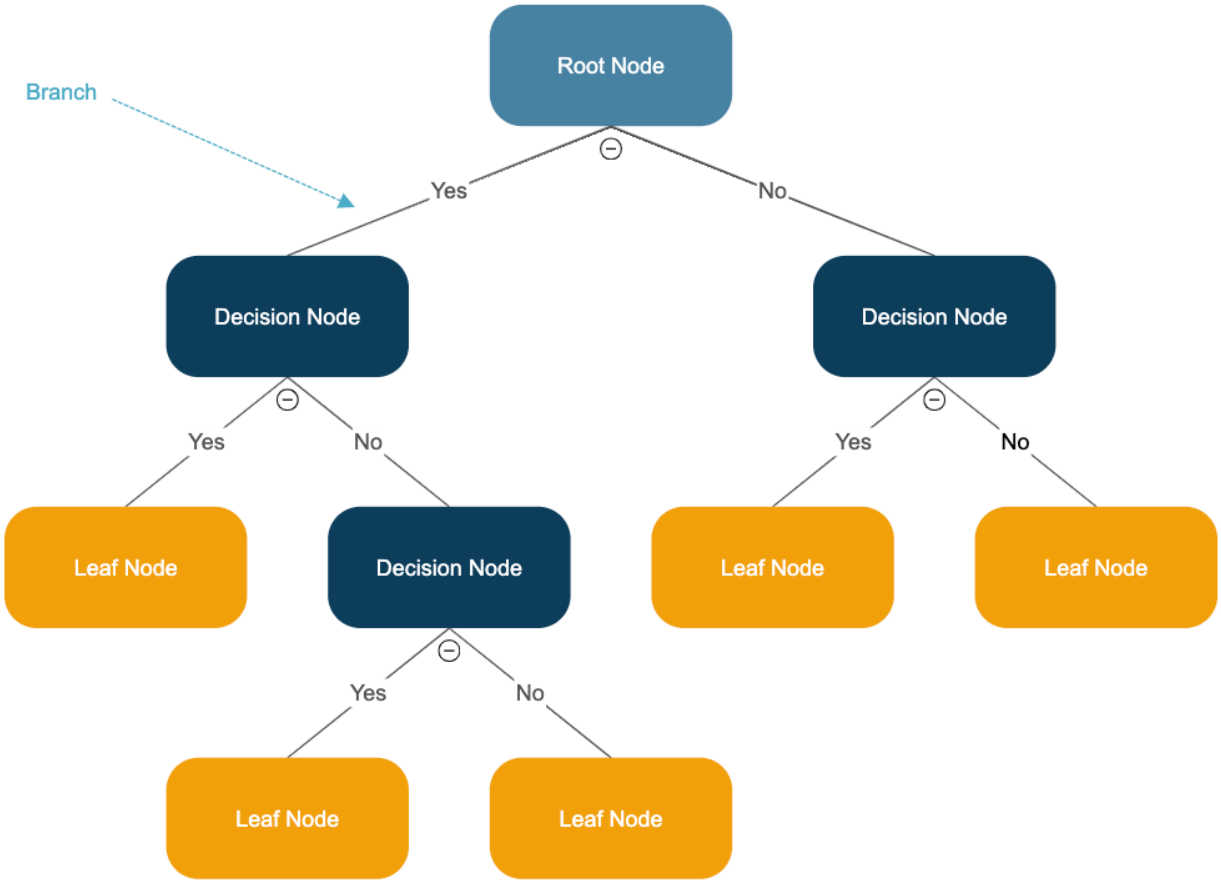


Figure 2: Schematic representation of decsion trees [2].

[6] Brownlee, J. A Gentle Introduction to the Rectified Linear Unit (ReLU). https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (accessed 2024-12-01).

[7] 2D Matpedia. http://www.2dmatpedia.org/ (accessed 2024-12-01).

[8] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. https://www.nature.com/articles/npjcompumats201628 (accessed 2024-12-01).

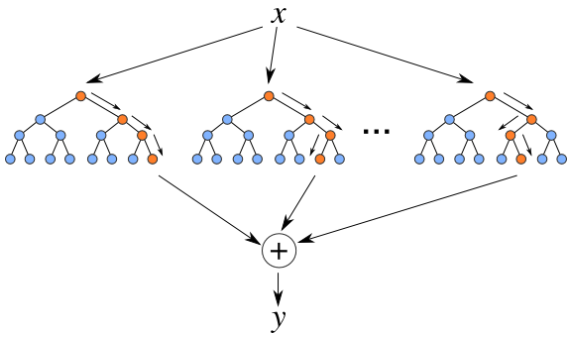
[9] Bandstructure¶. https://hackingmaterials.lbl.gov/matminer/featurizer$_s ummary.html$ (accessed 2024-12-01).

[10] Burba, D. Gradient Boosting: A Silver Bullet in Forecasting. https://towardsdatascience.com/gradient-boosting-a-silver-bullet-in-forecasting-5820ba7182fd (accessed 2024-12-01).

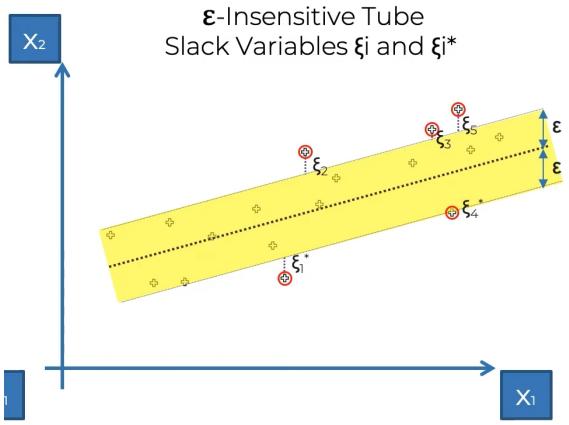
[11] https://doi.org/10.1063/1.4931571 (accessed 2024-12-01).

[12] Di Quarto, F. https://doi.org/10.1557/PROC-654-AA4.8.1 (accessed 2024-12-01).

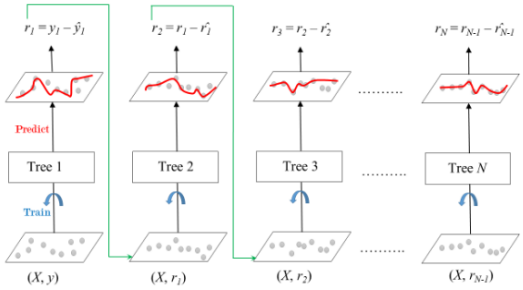
[13] Dagenais, K. Modeling Energy Band Gap as a Function of Optical Electronegativity for Binary Oxides. https://static1.squarespace.com/static/5443d7c7e4b06e8b47de9a55/t/59b00a92be42d6107983e966/1504709268156/Mod Energy-Band-Gap-as-a-Function-of-Optical-Electronegativity-for-Binary-Oxides.pdf (accessed 2024-12-02).



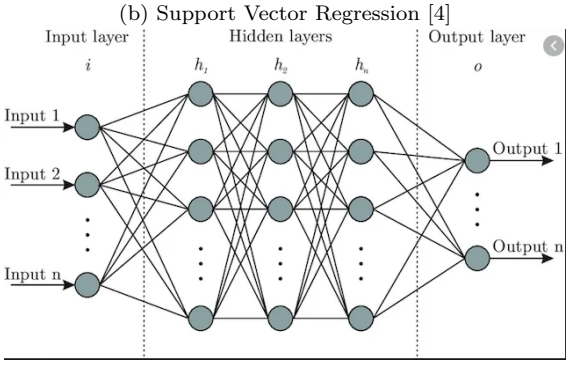
(a) Random Forest [3]



(b) Support Vector Regression [4]



(c) Gradient Boosting Regression [5]



(d) Artificial Neural Network

Figure 3: Schematic representations of different machine learning models.

[14] Guo, M.; Xu, X.; Xie, H. Predicting the Band Gap of Binary Compounds from Machine-Learning Regression Methods. https://doi.org/10.26434/chemrxiv-2021-jhg7b (accessed 2024-12-01).

[15] Geron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems; O'Reilly Media, 2019.

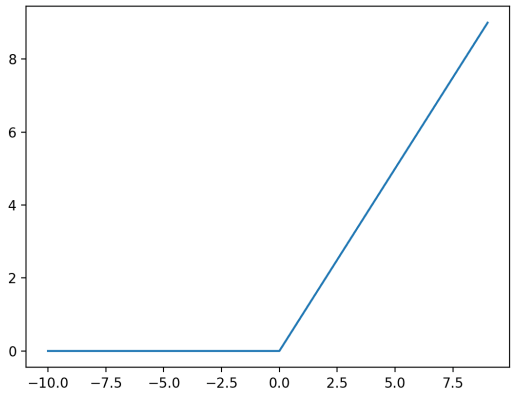
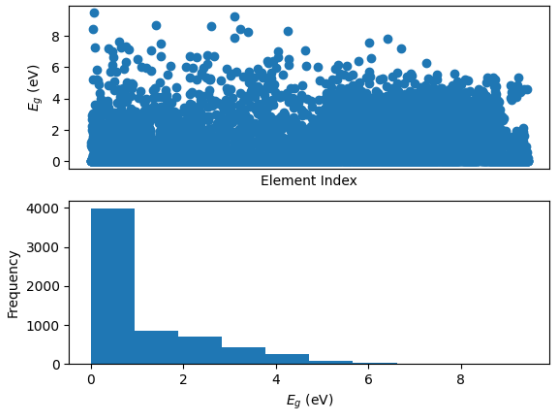
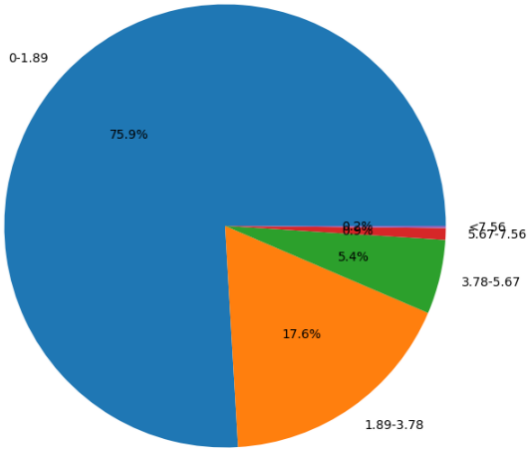


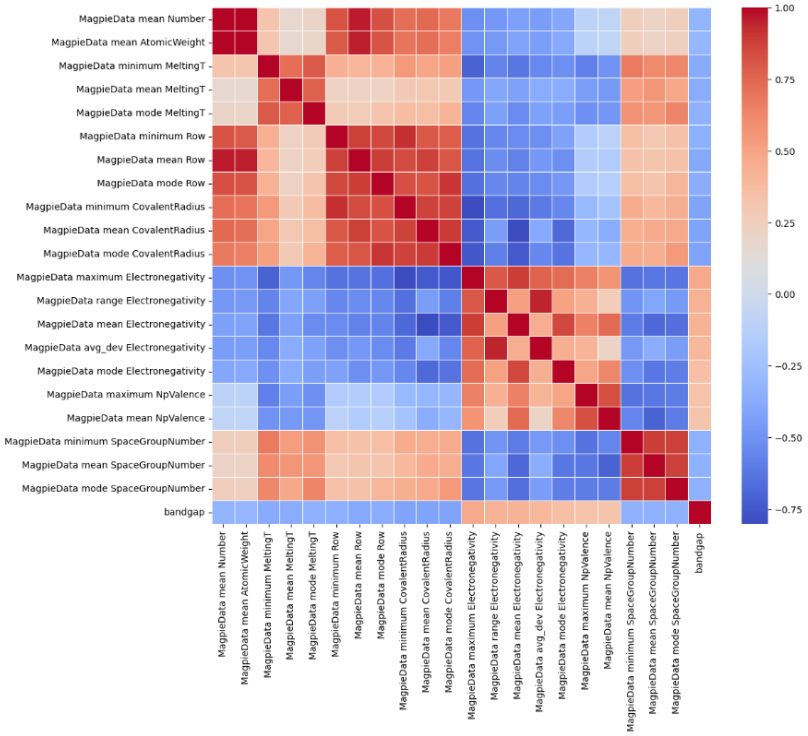
Figure 4: Rectified linear unit (ReLU) activation function [6].



(a) Top: Scatter plot of band gaps for the different materials in the dataset. Bottom: Frequency of band gap values across the dataset.



(b) Pie chart distribution of band gap values for different ranges (eV).



(c) Filtered (highest values) correlation coefficients between dataset features.

Figure 5: Exploratory Data Analysis (EDA) of the dataset.

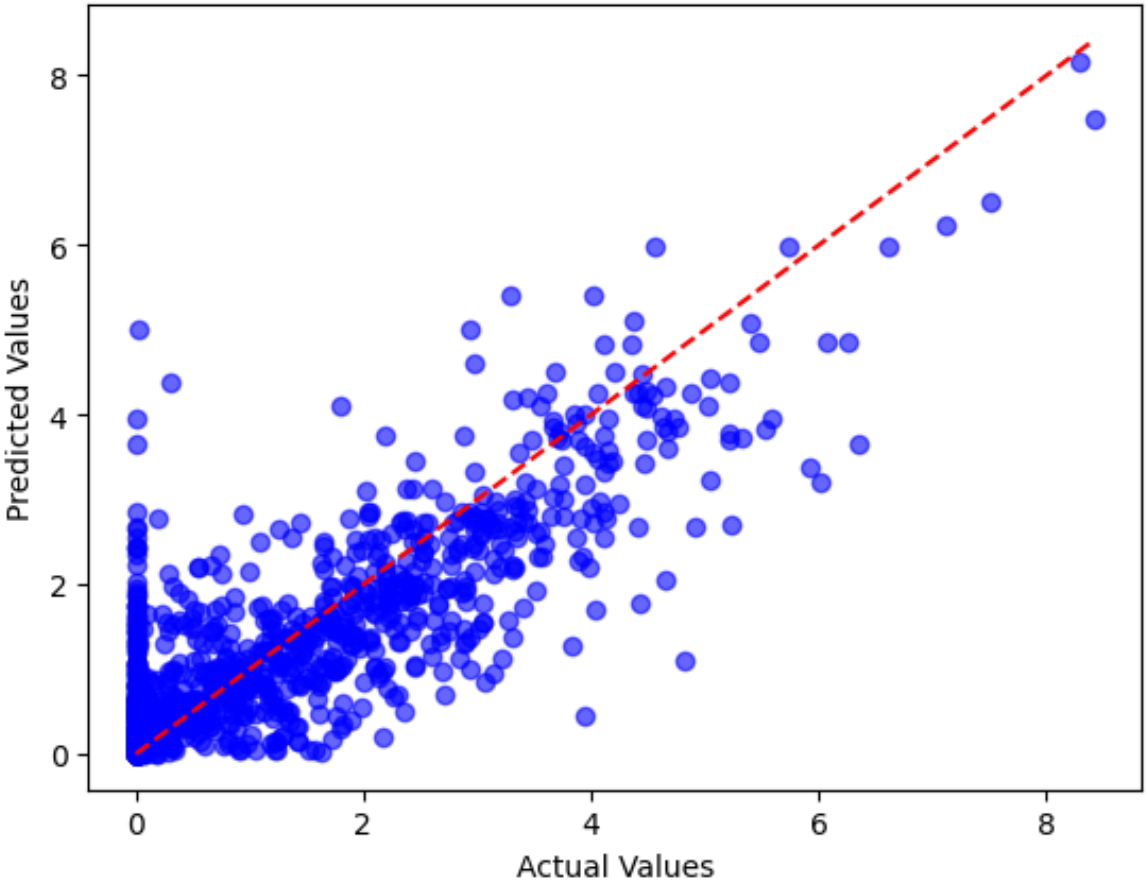
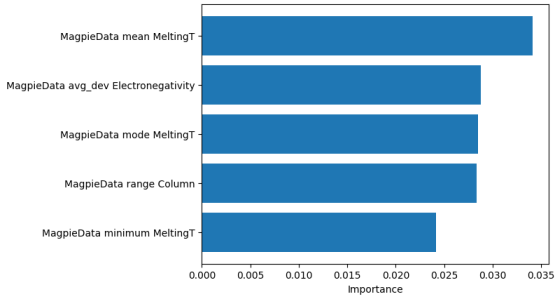
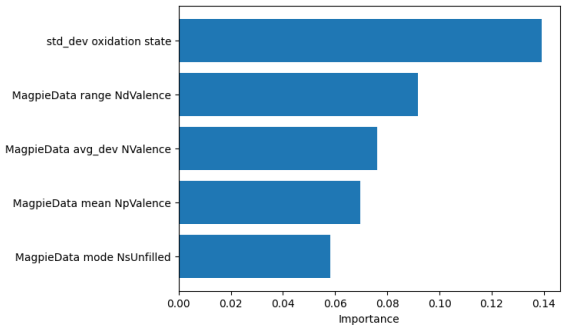


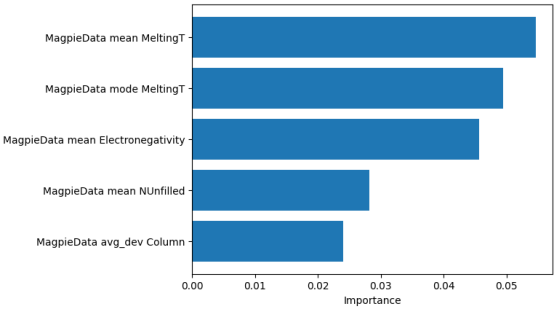
Figure 6: Predicted band gaps as a function of the actual values with the line of best fit for the random forest regression model.



(a) Random forest



(b) Support vector



(c) Gradient Boosting

(d) Artificial neural network

Figure 7: Feature importance plots for different models