A Machine Learning Approach to Predict Critical Temperature of Superconductors

Group 7- Final Team Project

Mohd Sharik, Mohamed Niaz M, Rishabh Malik

Shiley-Marcos School of Engineering, University of San Diego

AAI-500-A1: Probability and Statistics for Artificial Intelligence

Dr. Ebrahim Tarshizi

June 24, 2024

Abstract

The critical temperature (T_c) is a defining characteristic of superconductors, marking the temperature below which these materials exhibit zero electrical resistance. Accurately predicting T_c is pivotal for advancing the discovery and design of new superconducting materials, which find applications in diverse fields such as energy transmission, medical imaging, and quantum computing. This study explores the application of machine learning and sophisticated feature engineering techniques to forecast T_c based on the elemental properties of superconductors. Through 81 features that are essentially derived from elemental properties, we developed a robust predictive model. Various machine learning algorithms were trained and rigorously evaluated on the engineered feature set, including linear regression, random forest, gradient boosting, and XGBoost. Following meticulous hyperparameter tuning using grid search, the XGBoost model emerged as the top performer, achieving a root mean squared error (RMSE) of 9.75 on the test set and explaining 91.6% of the variance in T_c prediction. These findings demonstrate the effectiveness of our approach in predicting superconductor critical temperatures, which promises to accelerate advancements in this crucial area of research.

Keywords: *machine learning, superconductors, critical temperature, ML in physics.*

INTRODUCTION

Superconductors, which exhibit zero electrical resistance below a certain critical temperature (Tc), hold significant potential for advancing various technological fields. Their unique properties can revolutionize energy transmission by eliminating resistive losses, greatly improving efficiency. In medical imaging, superconducting magnets are pivotal for high-resolution MRI scans, enhancing diagnostic capabilities and patient care. Additionally, superconductors are critical in quantum computing, where they form qubits, the essential units of quantum processors. The critical temperature (Tc) is a crucial parameter that determines the usability of a superconductor. Materials with higher Tc values are preferred because they can operate at more manageable temperatures, minimizing the need for costly and energy-intensive cooling systems. Thus, accurately predicting the Tc of superconductors is vital for discovering and designing new materials with superior properties. Traditional methods for discovering new superconductors typically involve experimental trial and error combined with theoretical calculations. These methods can be laborious and resource-intensive (Wang, 2019). Recently, machine learning has emerged as a powerful tool to expedite material discovery by predicting properties based on chemical composition and structure (Hansen et al., 2015). Despite several studies using machine learning to predict the Tc of superconductors, there remains room for enhancing accuracy and generalizability (Gashmard et al., 2024). This study focuses on developing a reliable and precise model for predicting the Tc of superconductors using machine learning and feature engineering. We utilize a dataset containing elemental properties of various superconductors and apply feature engineering techniques to extract pertinent information (Smith & Brown, 2021). By integrating feature engineering with dimensionality reduction and advanced machine learning algorithms, we aim to achieve superior predictive performance over existing methods. The findings from this research could significantly expedite the discovery and design of new superconducting materials, thereby advancing technologies dependent on superconductivity.

DATA CLEANING AND PREPARATION

The dataset utilized in this study is available from the UCI Machine Learning repository and can be accessed at `https://archive.ics.uci.edu/dataset/464/superconductivty+data`. The dataset used in this study comprises 21,263 superconductors, each characterized by 82 features. These features encompass various elemental properties, including atomic mass, first ionization energy (FIE), atomic

radius, density, electron affinity, fusion heat, thermal conductivity, and valence. For each property, ten statistical measures were calculated: mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, and weighted standard deviation. Additionally, the dataset includes a feature representing the total number of elements in each superconductor. The target variable is the critical temperature (T_c) of each superconductor. There are no missing values in this data as this is a pre-cleaned dataset. For further details, including all relevant files and documents such as the code and the case study presentation, please visit the provided Github repository at GiHub Link.

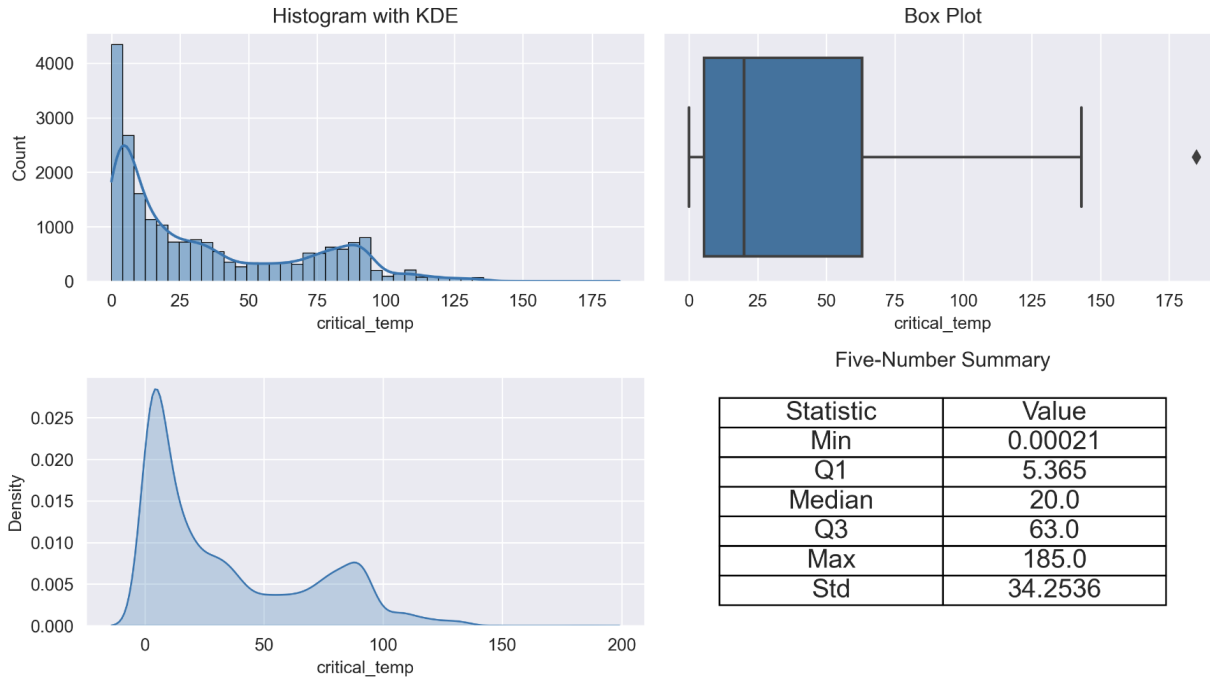
Variable	**Units**	**Description**
Atomic Mass	AMU	Average of the proton and neutron rest masses
First Ionization Energy	kJ/mol	Energy required to remove a valence electron
Atomic Radius	picometer pm	Distance from the nucleus to the outermost electron
Density	kg/m^3	Density at standard temperature and pressure
Electron Affinity	kJ/mol	Energy required to add an electron to a neutral atom
Fusion Heat	kJ/mol	Energy to change from solid to liquid without temperature change
Thermal Conductivity	W/(m × K)	Thermal conductivity coefficient k
Valence	no units	Typical number of chemical bonds formed by the element

Table 1: This table shows the properties of an element which are used for creating features to predict T_c.

EXPLORATORY DATA ANALYSIS

In this section, we analyze the dataset features to inform our feature engineering techniques and model selection. Initially, we explored the critical temperature (T_c) variable using various plots (Figure 1). The plots indicate that the critical temperatures are positively skewed, approximating a gamma distribution. The adjacent figure also provides statistical insights such as the five-number summary and standard deviation, revealing some outliers in the data.

Next, we selected several explanatory variables, particularly mean elemental properties, for univariate analysis. Our observations revealed a variety of distributions among these properties: some followed a Gaussian-Normal distribution (Fischer & Lee, 2004), while others exhibited positive or negative skewness. These findings highlight the non-normal distribution of elemental properties in superconducting materials (see Figure 2).



Statistic	Value
Min	0.00021
Q1	5.365
Median	20.0
Q3	63.0
Max	185.0
Std	34.2536

Figure 1: Exploratory data analysis of critical temperature

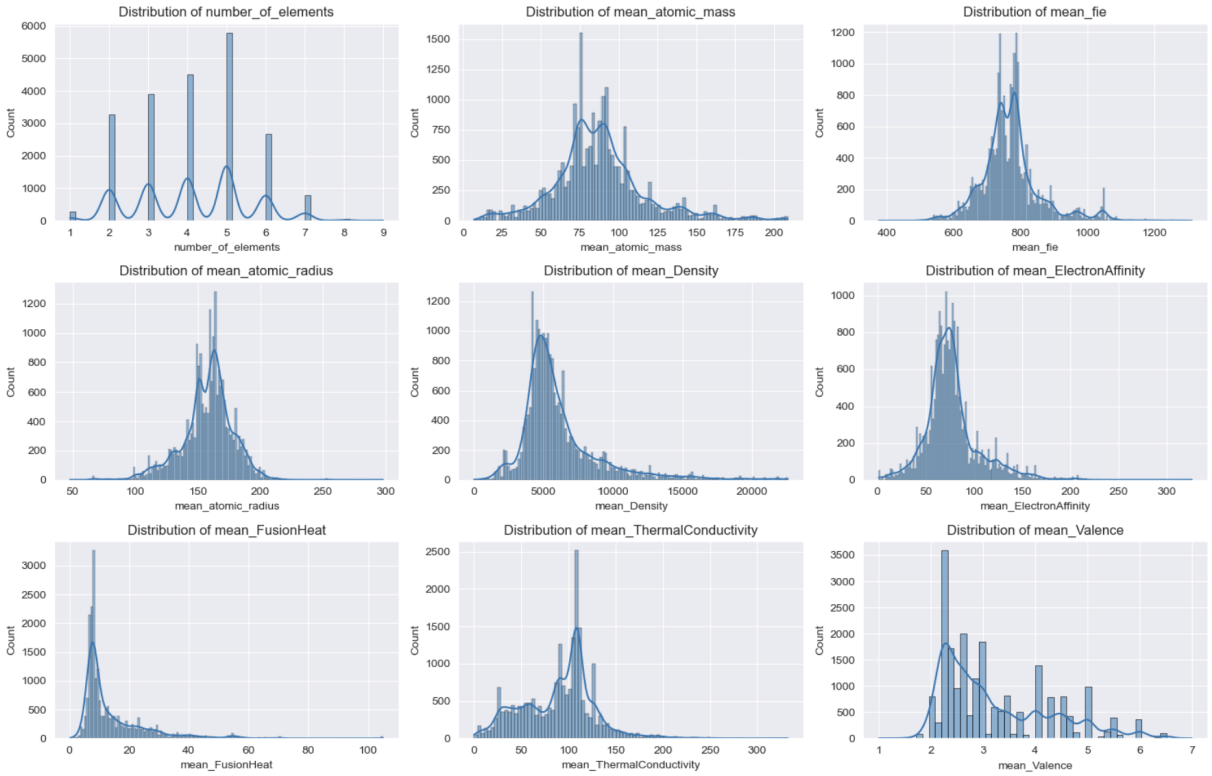


Figure 2: Distribution of Mean Properties

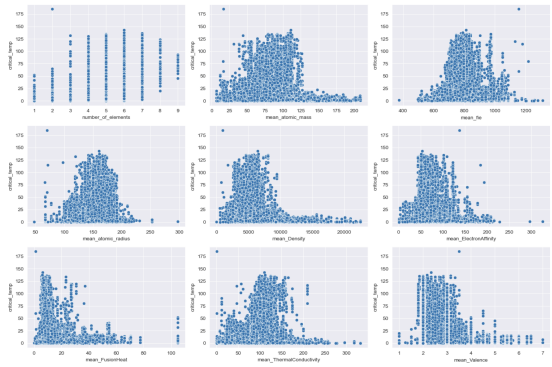
BI-VARIATE EDA AND ASSUMPTIONS VALIDATION

Statistical and machine-learning approaches often rely on several assumptions about the underlying data for effective modeling. In this phase of our research, we rigorously validated and confirmed these assumptions particularly relevant to linear models. The key assumptions examined include (a) a linear relationship between the target variable and explanatory variables (Qu, 2024), (b) the absence of multi-collinearity among explanatory variables (Upendra et al., 2023), (c) normality of errors (Zhang, 2023), and (d) homoscedasticity (Wu & Drton, 2023). These assumptions are crucial as they ensure the reliability and interpretability of the model outputs.

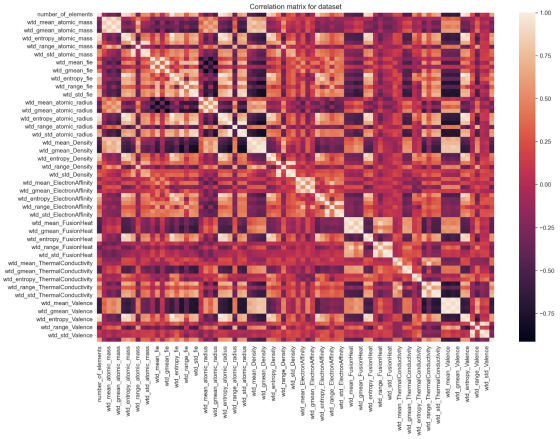
To validate these assumptions, various diagnostic plots and statistical tests were employed. Figure 3 illustrates these validations, showcasing scatterplots for assessing linearity, correlation matrix for multi-collinearity assessment, Q-Q plots for error normality, and residual plots for homoscedasticity.

Understanding and confirming these assumptions are imperative for ensuring the robustness of our model predictions and interpretations (Montgomery et al., 2012; Hair et al., 2019; Gelman et al., 2013). By adhering to these assumptions, we can enhance the accuracy and reliability of our predictive models, thereby advancing the application of statistical and machine-learning techniques in predicting critical temperatures of superconductors.

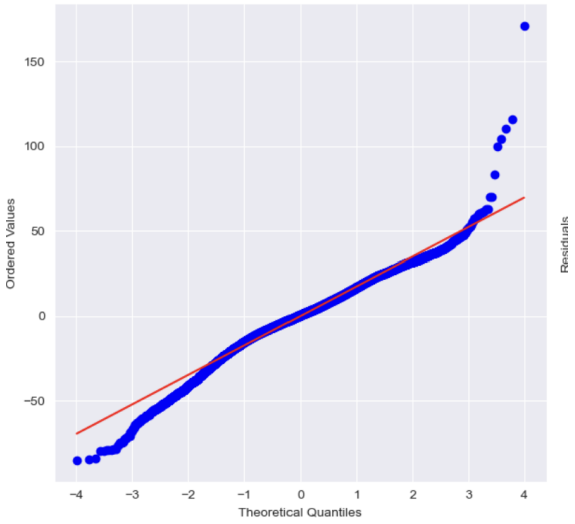
The preceding analysis indicates that a model such as linear regression may not be suitable due to the failure of almost all essential assumptions. We will substantiate this assertion in the forthcoming section. In the subsequent phase of our analysis, we explored the relationship between the mean critical temperature and the number of elements in superconductors. Our findings demonstrated significant variations in the mean critical temperatures across different groups of elements. Figure 4(a) visually depicts these discrepancies. To further validate these observations statistically, we conducted a one-way ANOVA test (Chatzi & Doody, 2023), which yielded an F-statistic of 2185.2 with an extremely small p-value of 0.0. This F-statistic indicates a large ratio of variability between groups compared to within groups. The extremely small p-value suggests that the observed data is highly unlikely under the assumption that there is no difference between groups. Therefore, the study concludes that there is a significant difference among the groups being compared. This implies that at least one group's mean is significantly different from the others, indicating that the factor being tested has a notable impact on the dependent variable. Further analyses or post-hoc tests may be necessary to determine which specific groups differ significantly from each other.



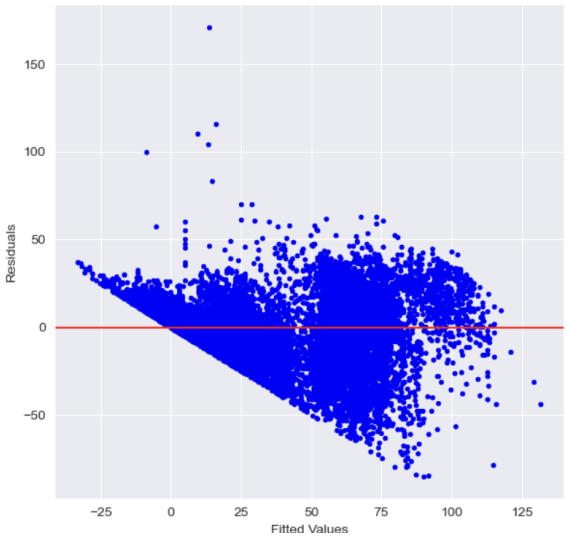
(a) Scatter Plot on subset features for Linearity assumption.



(b) Correlation matrix



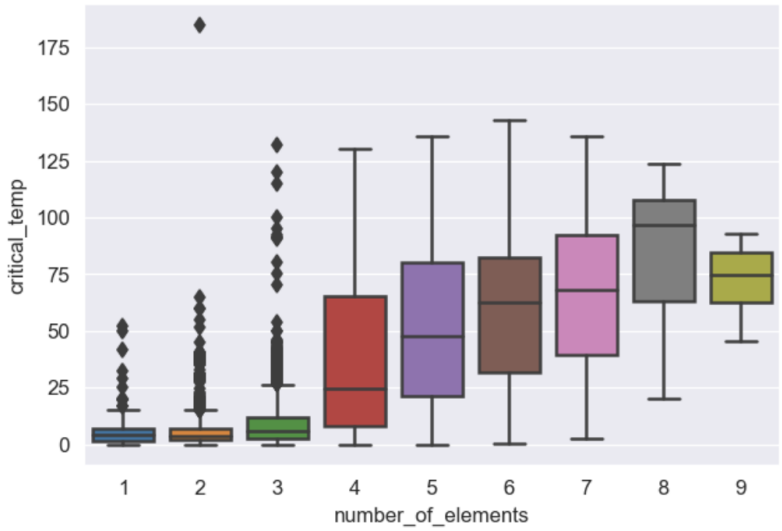
(c) QQ-Plot for Gaussian Distribution of errors



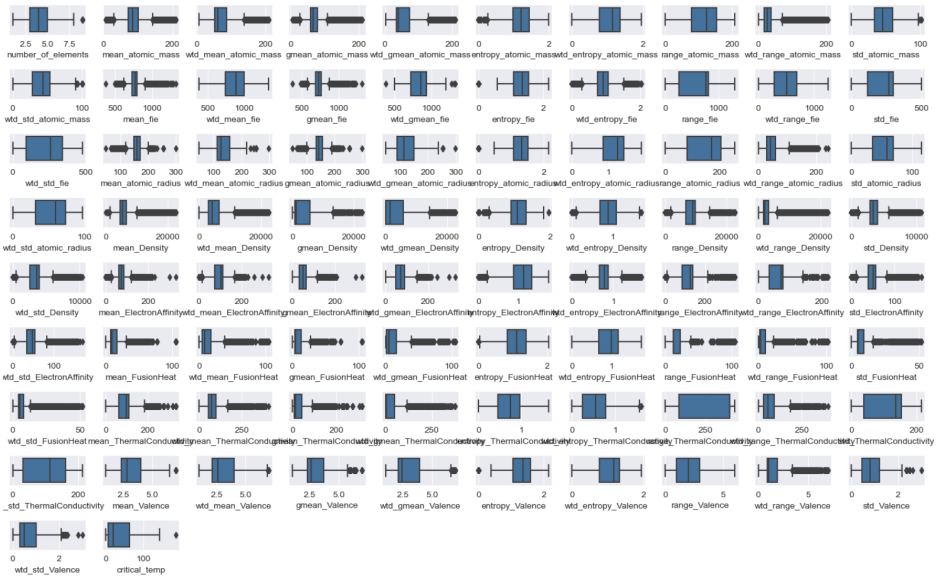
(d) Residual plot for Homoscedasticity

Figure 3: Validation of linear regression assumptions: These images indicate that none of the essential assumptions of linear regression are validated.

In the last part of our analysis, we used box plots to identify outliers within the dataset. Figure 4(b) illustrates the presence of outliers in our data.



(a) Box Plot Showing Mean Critical Temp. against the Number of Elements



(b) Box Plot for Each Feature to Assess Outliers

Figure 4: (a) Box Plot Showing Mean Critical Temperature Against the Number of Elements. (b) Box Plot for Each Feature to Assess Outliers, Evidencing Their Presence in the Data.

Subsequently, we partitioned the data into training and testing sets. we allocated 90% of the data, amounting to 19,136 data points, for training, and 10% of the data, totaling 2,127 data points, for testing. This separation is essential to prevent data leakage during the phases of data engineering and model training.

METHOD

In this section, we explore methodologies crucial for predicting superconductors' critical temperature. We cover outlier removal to enhance data integrity, feature selection to identify impactful variables, evaluation metrics to gauge model performance, and Model selection to ensure the chosen approach aligns with predictive accuracy and computational efficiency, providing a robust framework for superconductivity research.

OUTLIER REMOVAL

In the previous section, we identified outliers within the dataset. In this subsection, our focus is on effectively detecting and removing these outliers. Handling outliers in high-dimensional data poses challenges because what may appear as an outlier in one feature could be a normal data point in another feature. To tackle this issue, we utilized a machine learning technique known as Local Outlier Factor (LOF) (Srinarayani et al., 2024). LOF employs a tree-based algorithm to detect outliers using two key parameters: n neighbors (number of neighbors) and contamination (percentage of data considered as outliers). Specifically, we set n neighbors = 20 and contamination = 0.005. Below is the mathematical formulation illustrating its operation:

$$\text{LOF}(p) = \frac{\sum_{q \in \text{neighbors}(p)} \frac{\text{reach\_dist}(q,p)}{\text{reach\_dist}(p)}}{\text{n\_neighbors}}$$

where:

- neighbors(p) denotes the set of *n_neighbors* closest data points to p,
- reach_dist(q, p) is the reachability distance from q to p,
- reach_dist(p) is the average reachability distance from p to its *n_neighbors*.

This formula quantifies the degree of outlier status for each data point p, helping in the identification and subsequent removal of outliers. Before the removal of outliers, we had 19,136 data points in training and after the removal of some outliers, we reduced this number to 19,040.

FEATURE SELECTION

In statistical analysis and machine learning, the influence of variables in predicting the target variable varies, and unnecessary variables can lead to significant issues such as computational resource constraints, poor

performance, and increased computation time (Jolliffe, 2002).

In this study, Mutual information regression evaluates the dependency between variables by measuring the amount of information obtained about one variable through another, facilitating the identification of relevant features. Mutual information regression calculates the mutual information score between variables, quantifying the amount of information gained about one variable by knowing the value of another. It is particularly useful for feature selection as it identifies relationships that might not be linear and can capture complex dependencies between variables.

The mutual information $I(X;Y)$ between variables X and Y is calculated using the following formula:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

where:

- $p(x,y)$ is the joint probability distribution of X and Y,
- $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

This formula quantifies the amount of information shared between X and Y, which is crucial for identifying relevant features in mutual information regression.

EVALUATION METRICS

In this section, we explore methodologies crucial for predicting the critical temperature of superconductors. We begin with outlier removal to enhance data integrity by eliminating anomalous data points. Feature selection follows, identifying influential variables to streamline models and optimize performance. Evaluation metrics such as mean squared error (MSE), coefficient of determination (R-squared), and adjusted R^2 score quantify model accuracy and goodness-of-fit, respectively. Finally, model selection ensures the chosen approach aligns with both predictive accuracy and computational efficiency, providing a robust framework for superconductivity research.

EVALUATION METRICS

Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i are the actual values and $\hat{y}_i$ are the predicted values.

Coefficient of Determination (R-squared) assesses the proportion of the variance in the dependent variable that is predictable from the independent variables, given by:

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

where $\bar{y}$ is the mean of the observed data.

Adjusted Coefficient of Determination (R^2_{adj}) adjusts the R^2 score for the number of predictors in the model, given by:

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where n is the number of observations and k is the number of predictors.

MODEL SELECTION AND TRAINING

This section of the research focuses on machine learning model selection and training-testing results. As discussed earlier, when the assumptions of a statistical technique or machine learning model are not met, the model tends to perform poorly. We demonstrated this with the linear regression model. Additionally, we explored non-parametric machine learning models that do not assume specific distributions in the data. Below are detailed explanations of each model:

- **Linear Regression**:
  - **Description**: Linear Regression models the relationship between the dependent variable y and independent variables $\mathbf{x}$ using a linear approach. It estimates coefficients $\beta_0, \beta_1, \ldots, \beta_n$ to minimize the sum of squared residuals $\sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in}))^2$.

  - **Mathematics**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

    where β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are coefficients, $x_1, x_2, \ldots, x_n$ are predictors, and ϵ is the error term.

- **Random Forest**:

  - **Description**: Random Forest is an ensemble learning method that constructs multiple decision trees during training. For regression tasks, predictions are made by averaging the outputs of individual trees. This ensemble approach helps reduce overfitting and improves robustness against noise and outliers.

  - **Mathematics**:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$$

    where $f_i(\mathbf{x})$ represents the prediction of the i-th tree and N is the number of trees.

- **Gradient Boosting Regressor**:

  - **Description**: Gradient Boosting builds models sequentially, where each new model corrects errors made by the previous models. For regression, it combines weak learners (typically decision trees) to improve prediction accuracy. It uses gradient descent optimization to minimize the overall prediction error.

  - **Mathematics**:

$$\hat{y} = \sum_{i=1}^{M} \gamma_i h_i(\mathbf{x})$$

    where M is the number of trees, $h_i(\mathbf{x})$ is the prediction of the i-th tree, and γ_i is the weight applied to each tree's prediction.

- **XGBoost**:

  - **Description**: XGBoost is an optimized distributed gradient boosting library that uses an objective function to minimize errors and regularization terms to prevent overfitting. It applies a tree-boosting algorithm that sequentially adds new models to correct errors made by existing models.

– **Mathematics**:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $\mathcal{L}(\phi)$ is the objective function, l is the loss function measuring prediction error, $\hat{y}_i$ is the predicted value, and $\Omega(f_k)$ is the regularization term penalizing complexity.

RESULTS

In our study, we compared these models using training and testing data. The results are summarized in the following table:

Model	**RMSE**	**MAE**	**R2 Score**	**Adjusted R2 Score**
Linear Model (GLM)	20.00	15.26	0.66	0.66
Random Forest (RF)	14.09	9.61	0.83	0.83
Gradient Boosting (GB)	4.20	1.35	0.98	0.98
XGBoost (XGB)	4.56	2.01	0.98	0.98

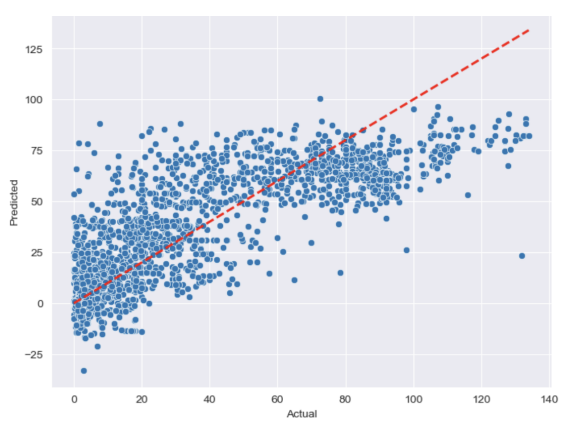
Table 2: Training Results of Machine Learning Models

Model	**RMSE**	**MAE**	**R2 Score**	**Adjusted R2 Score**
Linear Regression (LR)	20.01	15.24	0.65	0.64
Random Forest (RF)	14.32	9.75	0.82	0.82
Gradient Boosting (GB)	11.04	5.77	0.89	0.89
XGBoost (XGB)	9.75	5.30	0.92	0.92

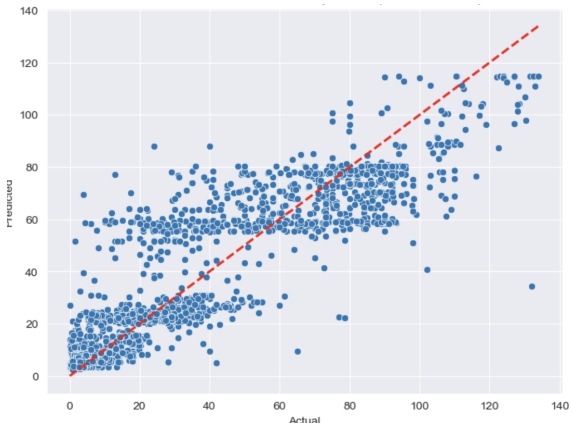
Table 3: Testing Results of Machine Learning Models

we also performed multiple grid searches for hyperparameter training and trained up to 1000 estimators for XGBoost, as it proved to be the best base model. After running various combinations, we obtained the following best parameters:

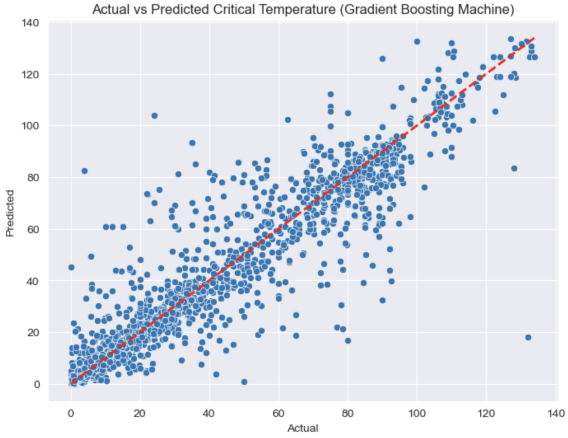
```
objective='reg:squarederror', n_estimators=400, learning_rate=0.1,
  max_depth=10, reg_lambda=0.6, reg_alpha=0.05, random_state=42
```



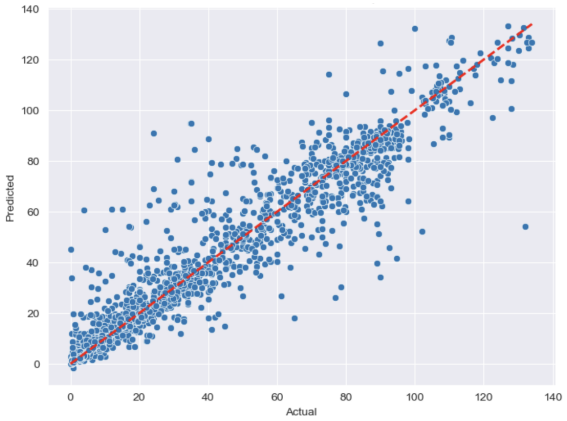
(a) Linear Regression



(b) Random Forest



(c) CGradiant Boosting



(d) XGBoost

Figure 5: Scatter Plot Between Predicted vs Actual Values

CONCLUSION

Our study demonstrates the effectiveness of machine learning techniques in predicting the critical temperature (T_c) of superconductors. We conducted an extensive exploratory data analysis, revealing significant variations in critical temperatures among different groups of elements. Applying feature engineering and dimensionality reduction techniques enabled us to extract pertinent information from the dataset, resulting in improved model performance. Advanced machine learning models, particularly Gradient Boosting and XGBoost, exhibited superior predictive accuracy compared to traditional linear regression.

As shown in the table above during the testing phase, the XGBoost model achieved the lowest R2 Score of 0.92, MAE of 5.30, and RMSE of 9.75. Similarly, the Gradient Boosting (GB) model displayed comparable performance with an RMSE of 11.04, MAE of 5.77, and an R2 score of 0.89. In contrast, Linear Regression (LR) and Random Forest (RF) models showed relatively higher errors and lower R2 scores, indicating lesser effectiveness in capturing the complex relationships within the data.

REFERENCES

References

[1] Chatzi, A., & Doody, O. (2023). The one-way ANOVA test explained. *Nurse Researcher, 31*. doi:10.7748/nr.2023.e1885

[2] Drucker, H. (1997). Improving Regressors Using Boosting Techniques. *Proceedings of the 14th International Conference on Machine Learning*.

[3] Elango, S., Natarajan, E., Varadaraju, K., Ezra, M., Durairaj, R. R., Mohanraj, K., & Osman, M. (2022). Extreme Gradient Boosting Regressor Solution for Defy in Drilling of Materials. *Advances in Materials Science and Engineering, 2022*, 1–8. doi:10.1155/2022/8330144

[4] Gashmard, H., Shakeripour, H., & Alaei, M. (2024). Predicting superconducting transition temperature through advanced machine learning and innovative feature engineering. *Scientific Reports, 14*(1), 3965. doi:10.1038/s41598-024-54440-y

[5] Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K.-R., & Tkatchenko, A. (2015). Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters, 6*(12), 2326–2331. doi:10.1021/acs.jpclett.5b00831

[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Linear Regression. In *An Introduction to Statistical Learning* (pp. 59–128). doi.org/10.1007/978-1-0716-1418-1_3

[7] Qu, K. (2024). Research on linear regression algorithm. *MATEC Web of Conferences, 395*, 01046. doi:10.1051/matecconf/202439501046

[8] Santibanez, S., Kloft, M., & Lakes, T. (2015). Performance Analysis of Machine Learning Algorithms for Regression of Spatial Variables. A Case Study in the Real Estate Industry.

[9] Srinarayani, K., Reddy, K., Reddy, C., & Pranav, C. (2024). Consistent Robust Analytical Approach for Outlier Detection in Multivariate Data using Isolation Forest and Local Outlier Factor. *International Journal of Innovative Science and Research Technology (IJISRT)*, 132–136. doi:10.38124/ijisrt/IJISRT24MAY229

[10] Suleiman, M., & Labadin, J. (2016). Improved Feature Selection Based on Mutual Information for Regression Tasks (IFSMIR). *Journal of IT in Asia, 6*. doi:10.33736/jita.330.2016

[11] Upendra, S., Abbaiah, D., & Balasiddamuni, D. (2023). Multicollinearity in Multiple Linear Regression: Detection, Consequences, and Remedies. *International Journal for Research in Applied Science and Engineering Technology, 11*, 1047–1061. doi:10.22214/ijraset.2023.55786

[12] Wu, J., & Drton, M. (2023). Partial Homoscedasticity in Causal Discovery with Linear Models.

[13] Zhang, Y. (2023). Asymptotic Normality of M-Estimator in Linear Regression Model with Asymptotically Almost Negatively Associated Errors. *Mathematics, 11*, 3858. doi:10.3390/math11183858