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 Deigo

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

Dr. Ebrahim Tarshizi

June 24, 2024

Abstract

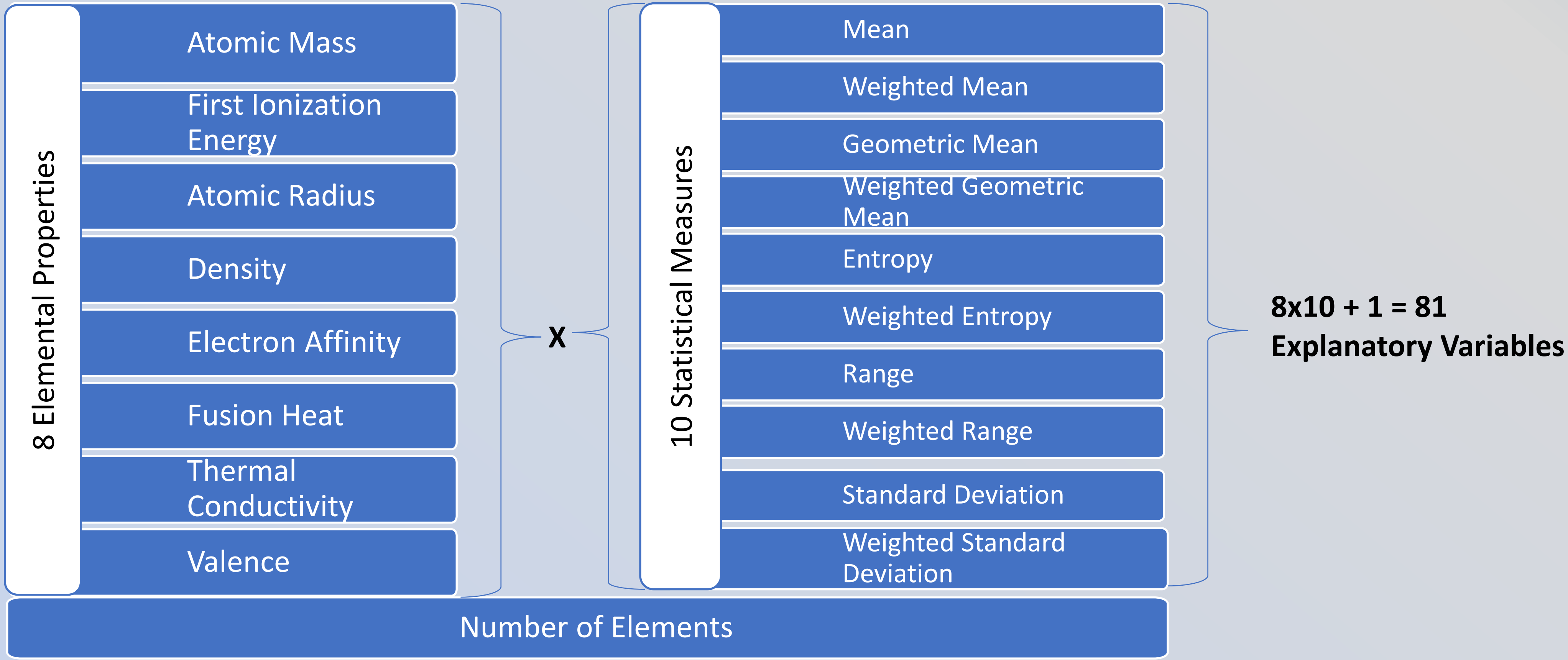
- Superconductor materials exhibit zero electrical resistance below their critical temperature (Tc). Higher Tc is preferred due to manageable temperature requirement.
- Superconductors can increase energy efficiency in applications such as energy transmission, medical imaging, and quantum computing.
- Accurately predicting Tc is enables discovery and design of new superconducting materials with cost-effective material having higher Tc.
- Traditional methods for discovering new superconductor material rely on a combination of theoretical calculations and experimental trial-error method to predict Tc.
- This study explores the use of machine learning and feature engineering to predict Tc of superconductors based on their elemental properties.

Approach

- Utilized dataset (https://archive.ics.uci.edu/dataset/464/superconductivty+dat).
- All documents and code can be accessed at GitHub Link
- Applied machine learning modeling by selecting 30 most influential features using mutual info regression.
- Employed and compared results from multiple regression models such as Linear regression, Random Forest, Gradient Boosting and XGBoost.
- The best results are produced by the XGBoost model, after hyperparameter tuning achieving a RMSE of 9.75 on the test set and explaining 91.6% of the variance in critical temperature.

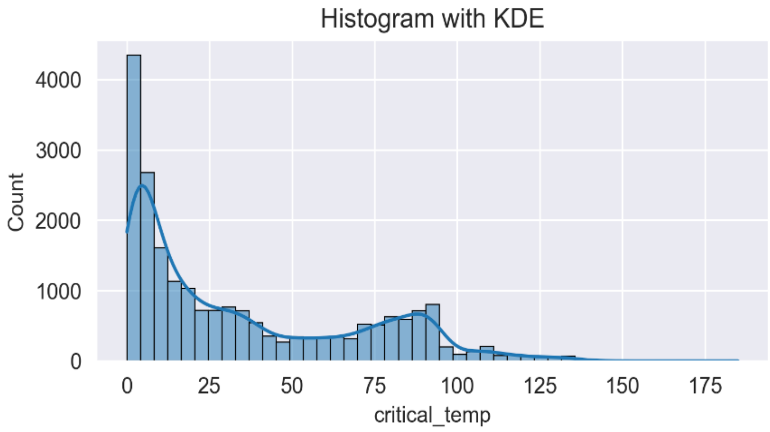
Dataset

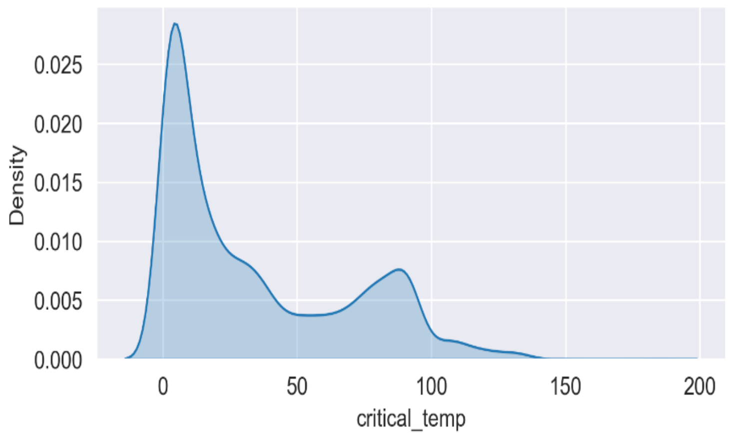
- 21,263 superconductors, each characterized by 82 features.
    - Target variable: Critical Temperature (Tc)
    - 81 Explanatory variables

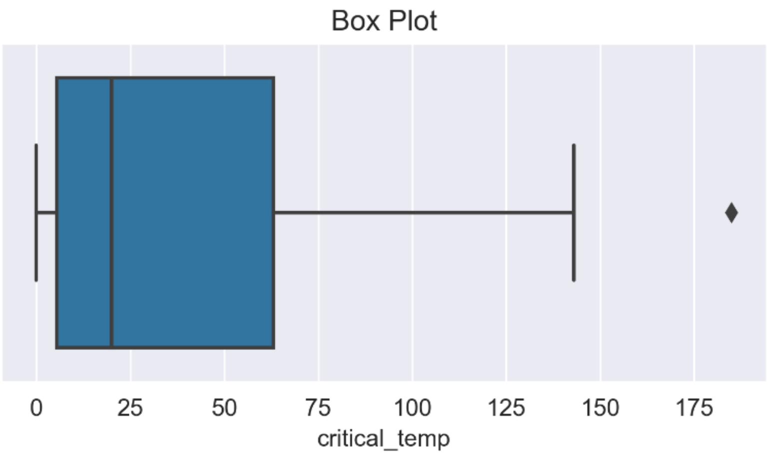


Exploratory Data Analysis

Analysis of Target Variable - Tc





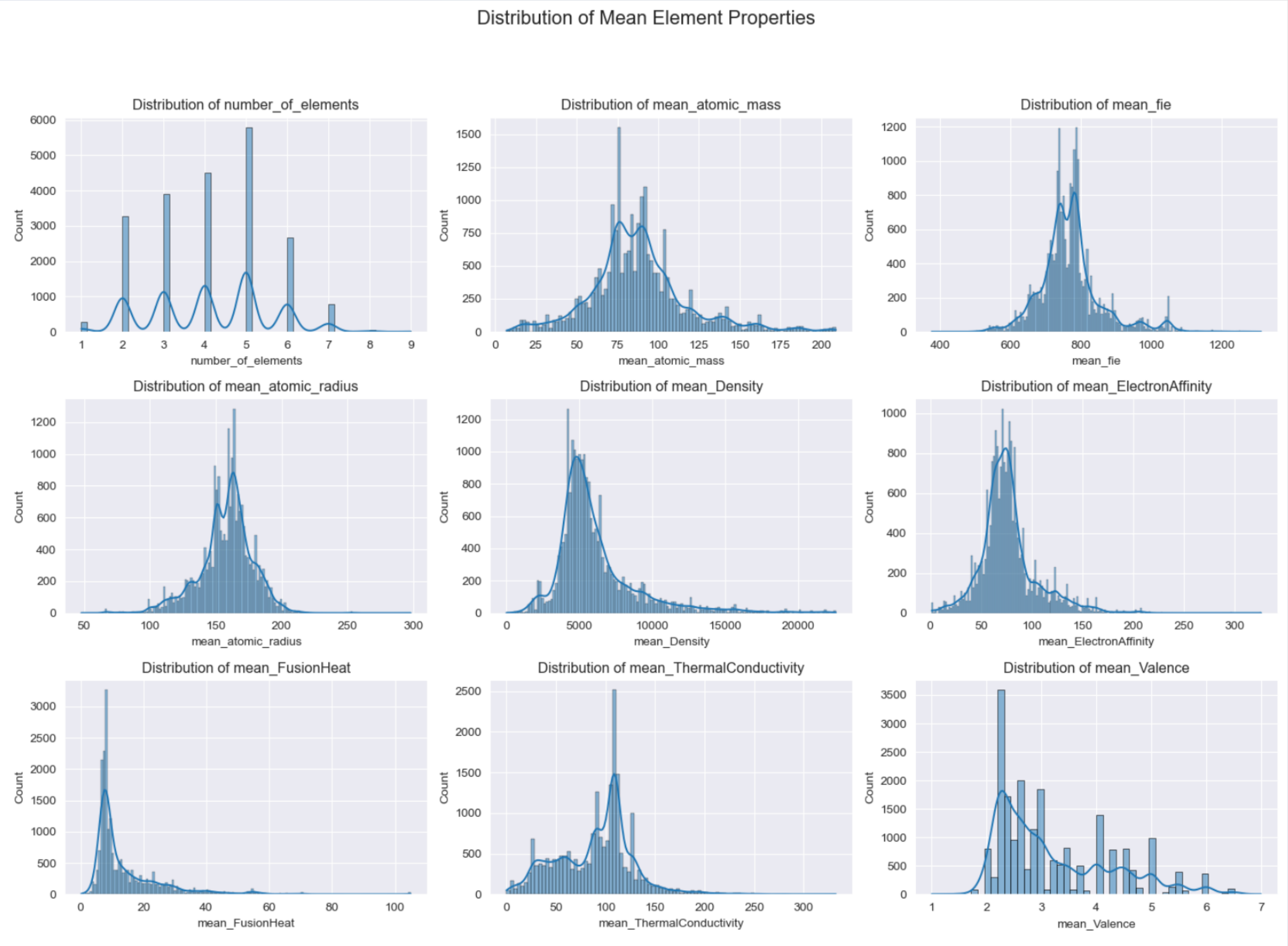


Five-Number Summary

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

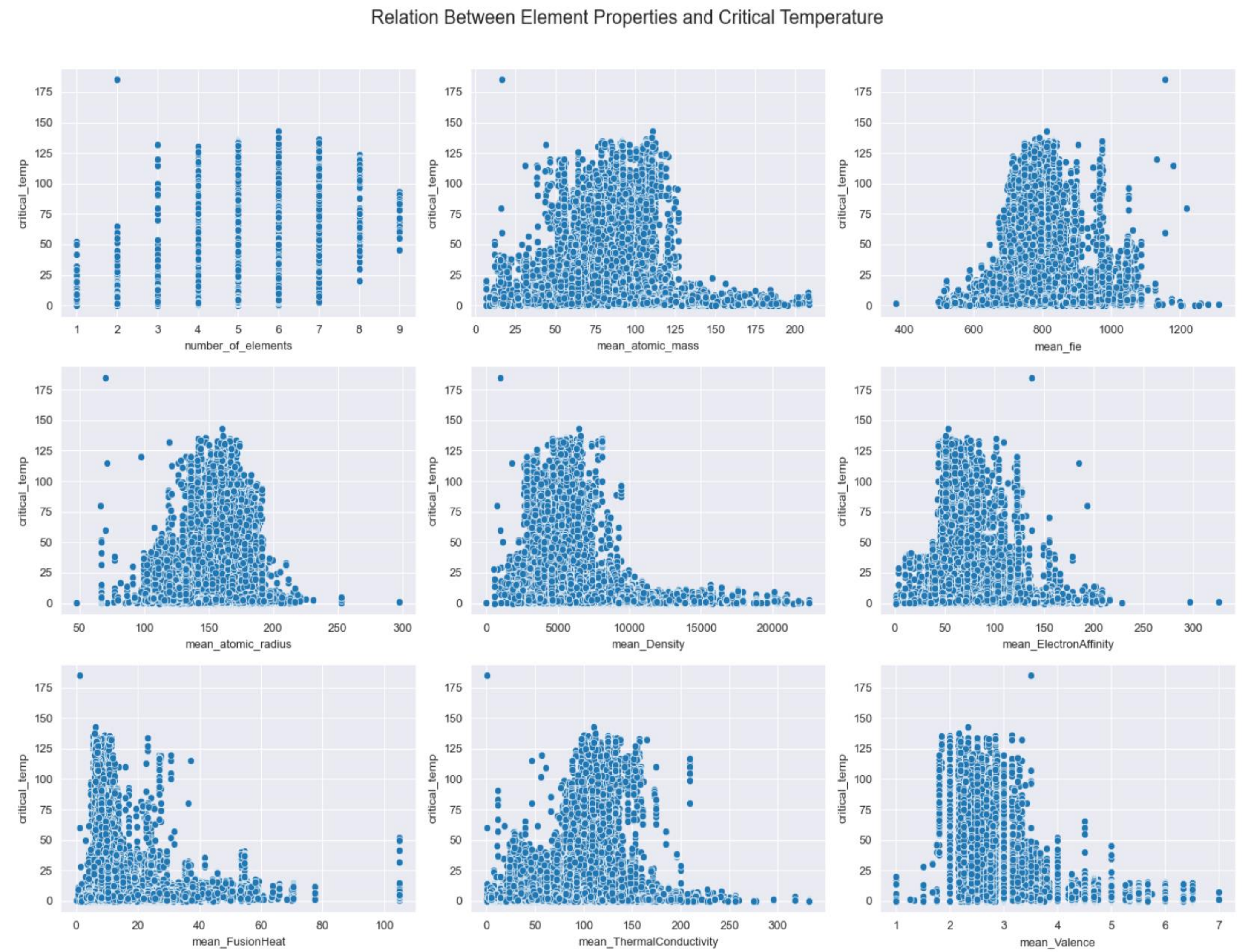
- Analyze the target variable critical temperature (Tc) with histograms, box plots, and KDE plots to understand its distribution.
- The table to summarized the data.

Distribution of Mean Element Properties



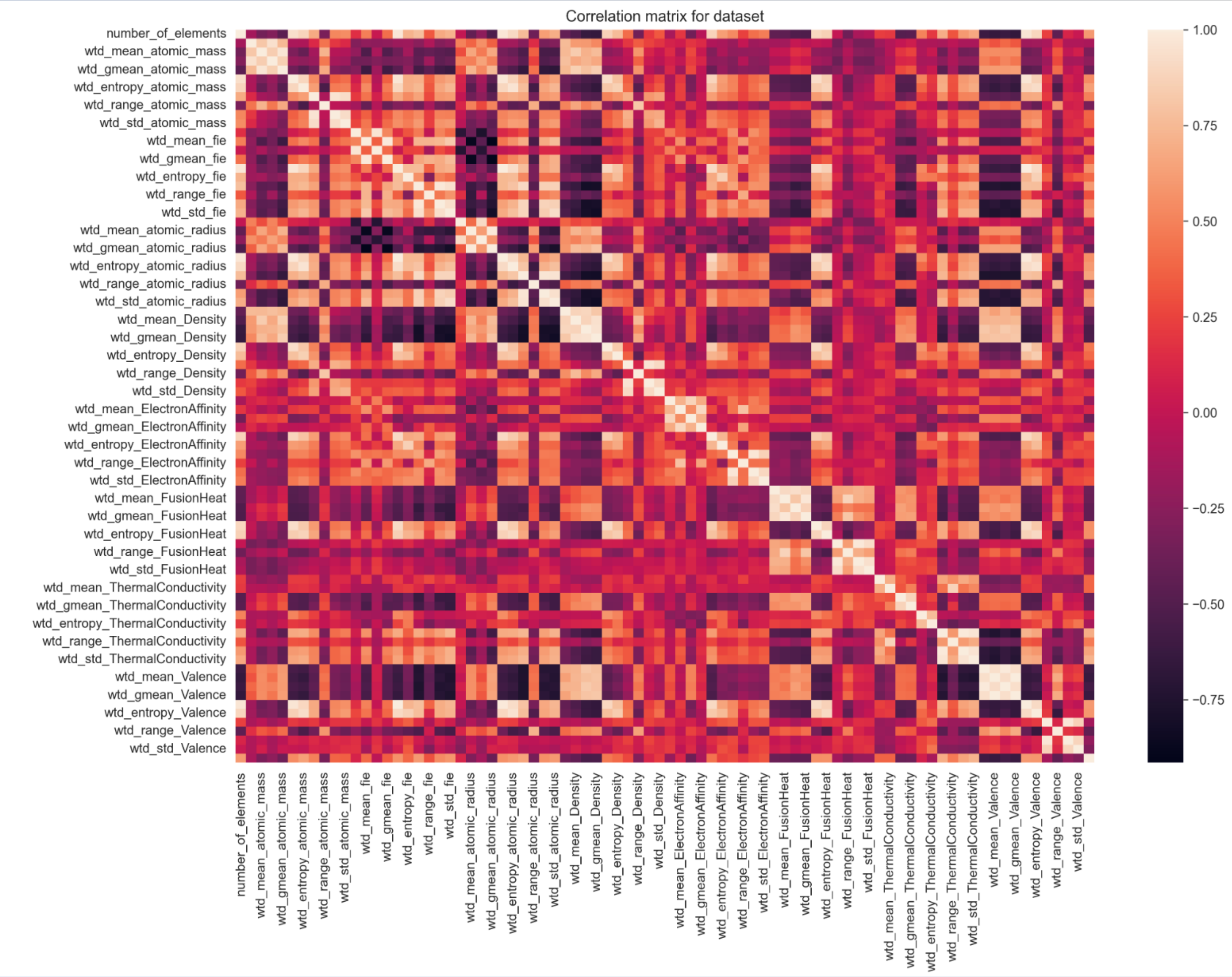
- Distribution of the mean elemental characteristics are analyzed for normal distribution.
- Number of elements and mean thermal conductivity are skewed to the left.
- Mean density and mean valence are skewed to the right.
- Other variables follow almost normal distribution.

Relationship: Tc vs. Mean Element Properties



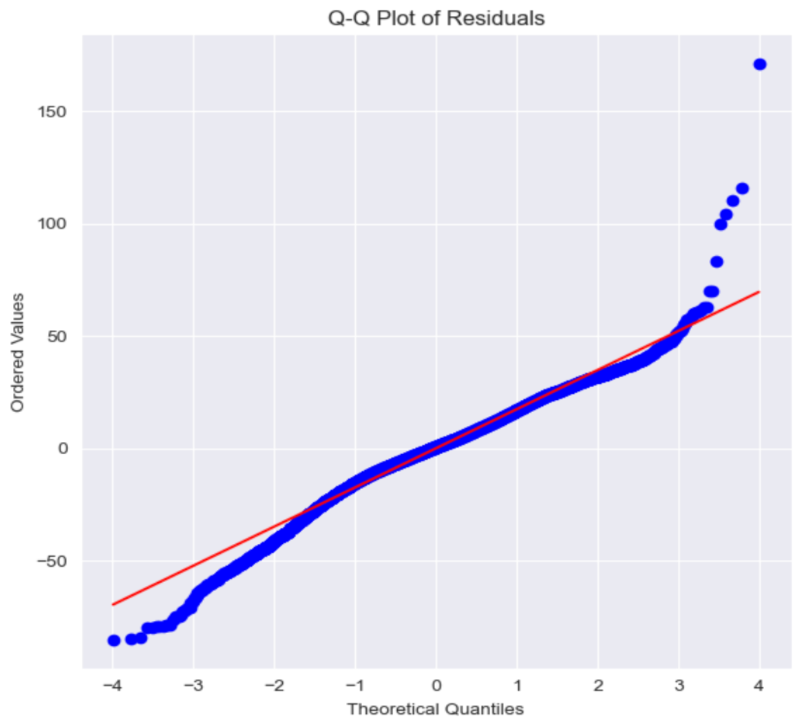
- Relationship between the target variable and the mean elemental properties was checked with the scatterplot.
- The scatter plot indicates moderate non-linear relationship existence between Tc and many of the mean elemental properties.
- This hints that linear regression by itself wouldn't be able to offer an effective model for predicting the critical temperature of superconductor.

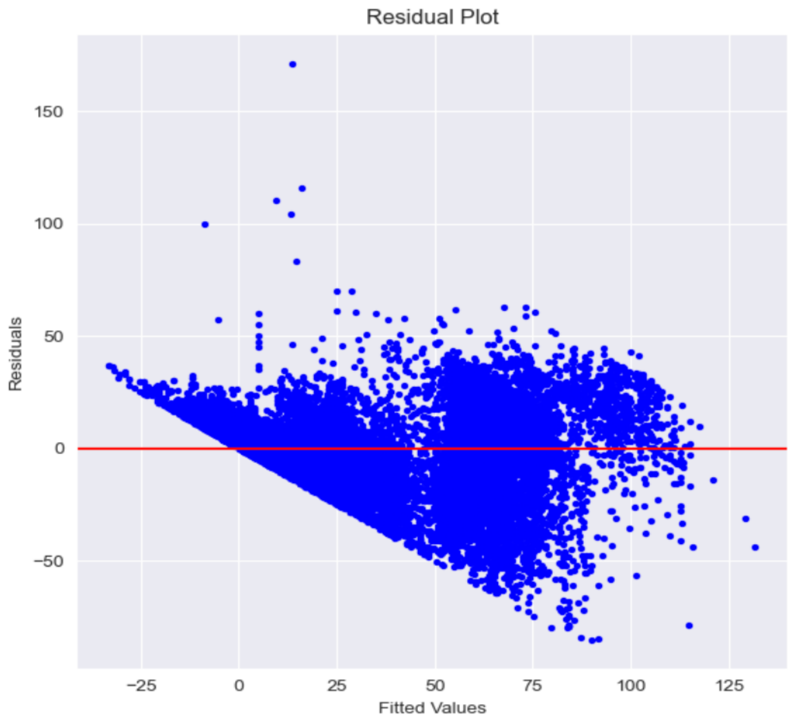
Multicollinearity Assumption Validation



- The correlation matrix with heatmap is used to see the relationships among the explanatory variables.
- This plot can highlight the level of collinearity existing among the variables.
- In view of the multicollinearity, principal component assessment was planned to be tested for effective modeling.

Residual Analysis – Normality & Homoscedasticity



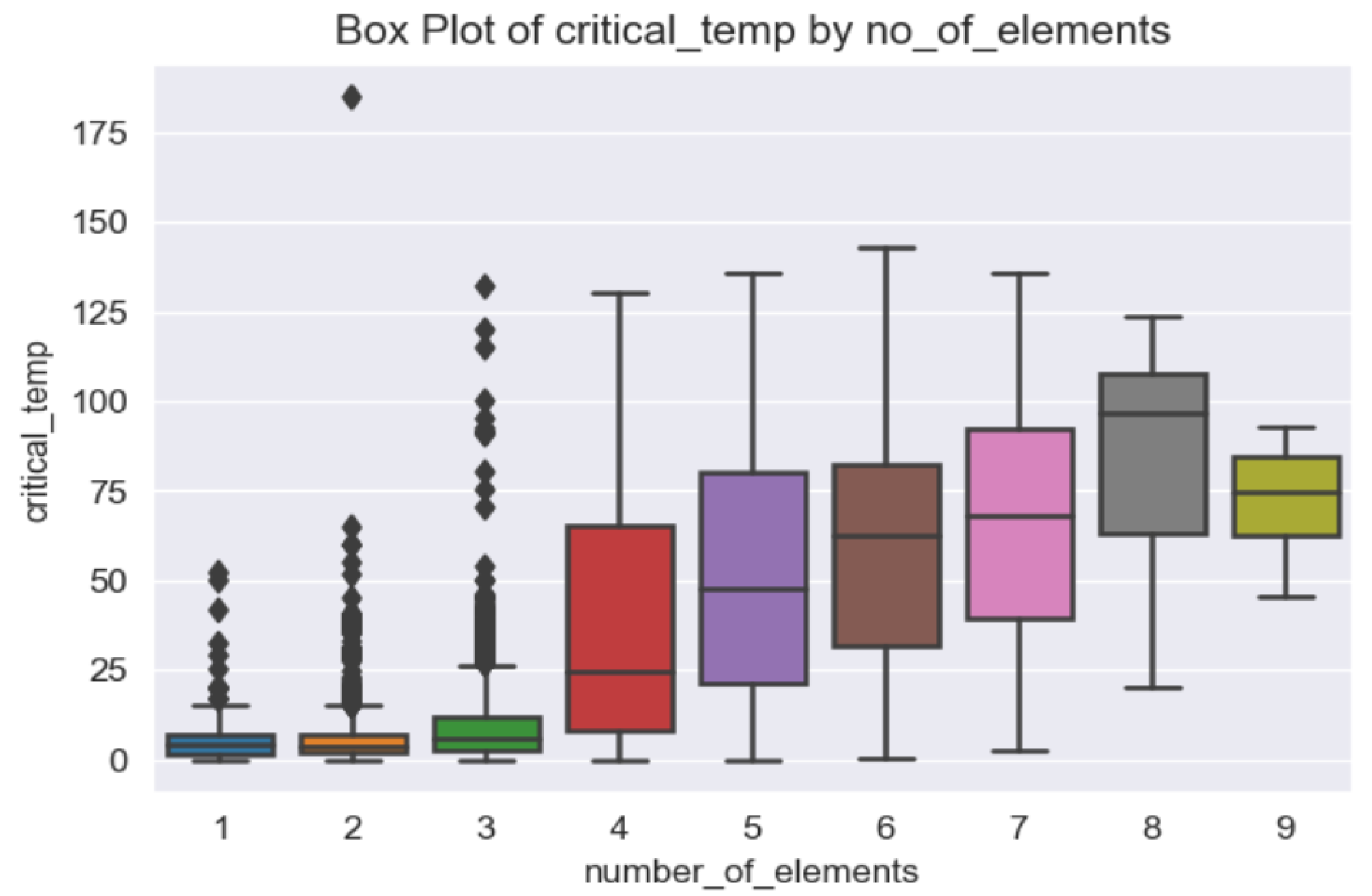



```
For Normalization Assumption (Kolmogorov-Smirnov Test) -->
Kolmogorov-Smirnov Test Statistic: 0.047788290698783126
Kolmogorov-Smirnov Test p-value: 1.2262993797650276e-42

 * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
For Homoscedasticity Validation (Breusch-Pagan Test Results) -->
Lagrange multiplier statistic: 3345.6746894394723
p-value: 0.0
f-value: 48.82834149692371
f p-value: 0.0
```

- The Q-Q plot explains that residuals are failing the normality assumption.
- The residual plot indicated that heteroscedasticity exists in error, failing the homoscedasticity assumption.

Critical temperature vs. Num of Elements




```
ANOVA F-statistic: 2185.2637675998517
ANOVA p-value: 0.0
The means of critical_temp are significantly different across groups defined by no_of_elements.
```

- ANOVA statistical test to compare means across number of elements in superconductors.
- We found that it does have a significant effect.

Outlier Analysis

- Distribution of features analyzed using box plots.
- Outliers may cause problems in analysis. Few outliers are removed.

```
Shape before outlier removal:
(19136, 82)

Shape after outlier removal:
(19040, 82)
```

Feature Selection

- To Reduce dimensionality and achieving effective performance with employed mutual info regression score to select top 30 features.

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

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.

std_fie	0.936162
gmean_Density	0.929497
std_ThermalConductivity	0.919061
entropy_atomic_mass	0.917424
entropy_atomic_radius	0.909855
range_fie	0.900306
range_ElectronAffinity	0.895783
entropy_Density	0.890029
entropy_FusionHeat	0.883159
entropy_ElectronAffinity	0.880617
std_ElectronAffinity	0.874442
range_Density	0.868894
wtd_gmean_Valence	0.866866
gmean_FusionHeat	0.865918
mean_ThermalConductivity	0.864272
wtd_mean_Valence	0.861474
std_atomic_radius	0.847472
gmean_ElectronAffinity	0.837540
gmean_ThermalConductivity	0.831210
wtd_gmean_Density	0.826616
range_atomic_mass	0.823123
entropy_Valence	0.821986
entropy_fie	0.819133
mean_FusionHeat	0.817270
mean_ElectronAffinity	0.811273
gmean_atomic_mass	0.811124
range_atomic_radius	0.805622
wtd_gmean_FusionHeat	0.800456
mean_Density	0.797578
wtd_std_ThermalConductivity	0.794076

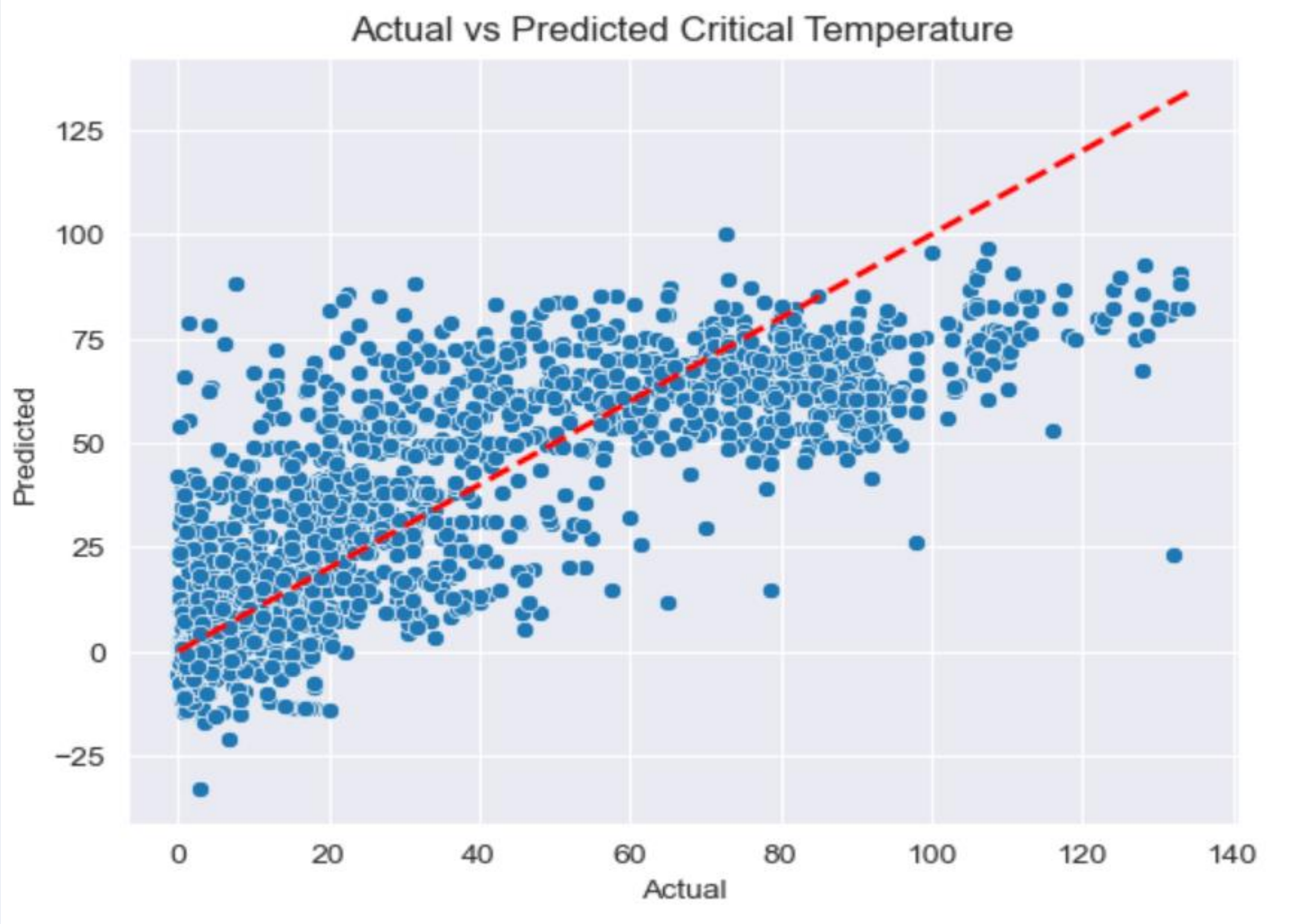
Model Selection

Simple Linear Regression

- Basic linear model that assumes a linear relationship between the features and the target variable has been fit.
- R2 results are not sufficient due to the non-linear relationship between target and explanatory variables.

```
Training results:
Training RMSE: 20.00363
Training MAE: 15.25545
Training R2 score: 0.65934
Training Adjusted R2 score: 0.65881

Testing results:
Testing RMSE: 20.01363
Testing MAE: 15.24122
Testing R2 score: 0.64744
Testing Adjusted R2 score: 0.64239
```

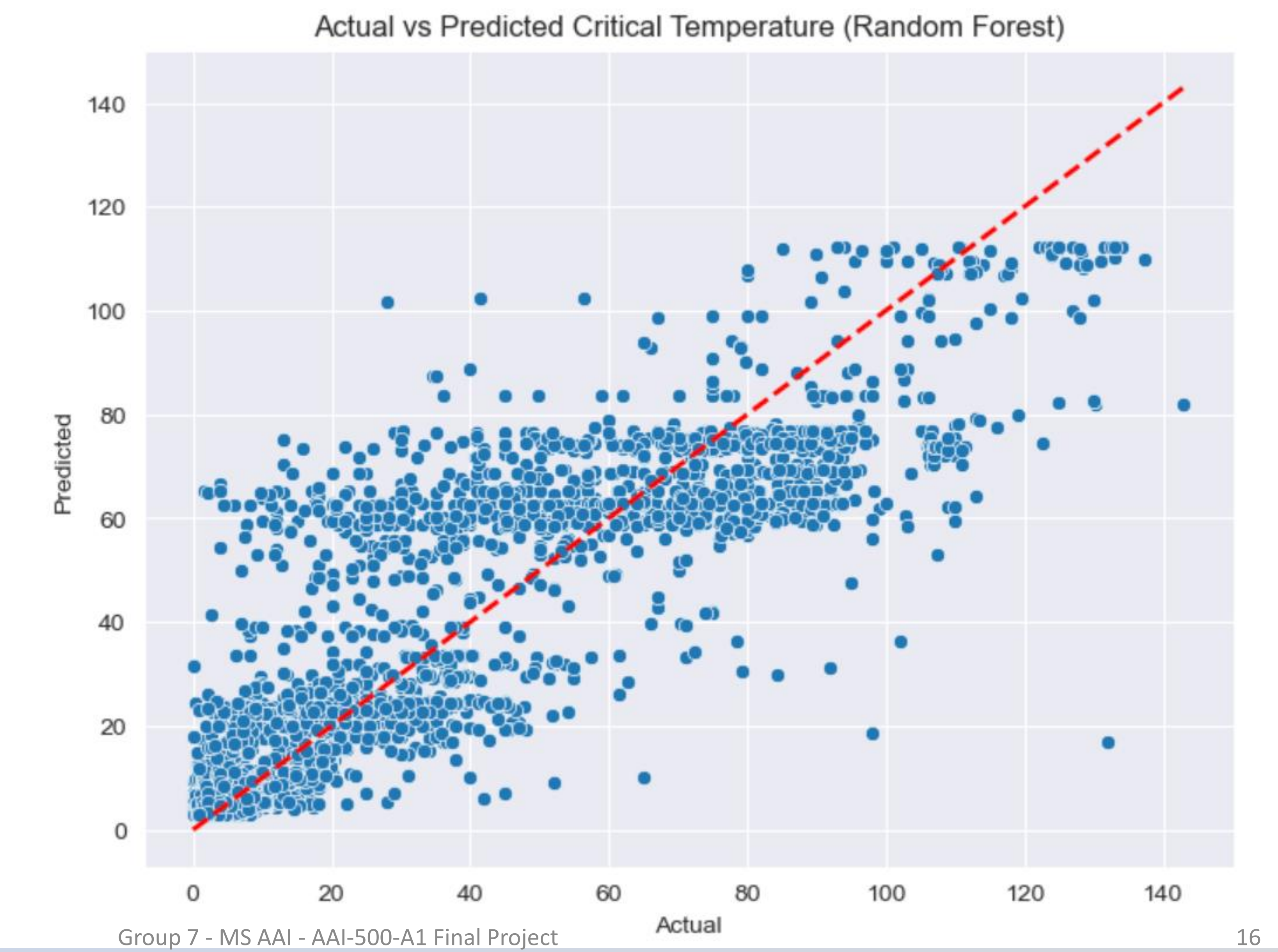


Random Forest

- Ensemble learning method that constructs a multitude of decision trees and outputs the mean prediction of the individual trees.

```
Training results:
Training RMSE: 14.30270
Training MAE: 9.50299
Training R2 score: 0.82653
Training Adjusted R2 score: 0.82621
```

```
Testing results:
Testing RMSE: 15.56559
Testing MAE: 10.51397
Testing R2 score: 0.79272
Testing Adjusted R2 score: 0.79075
```

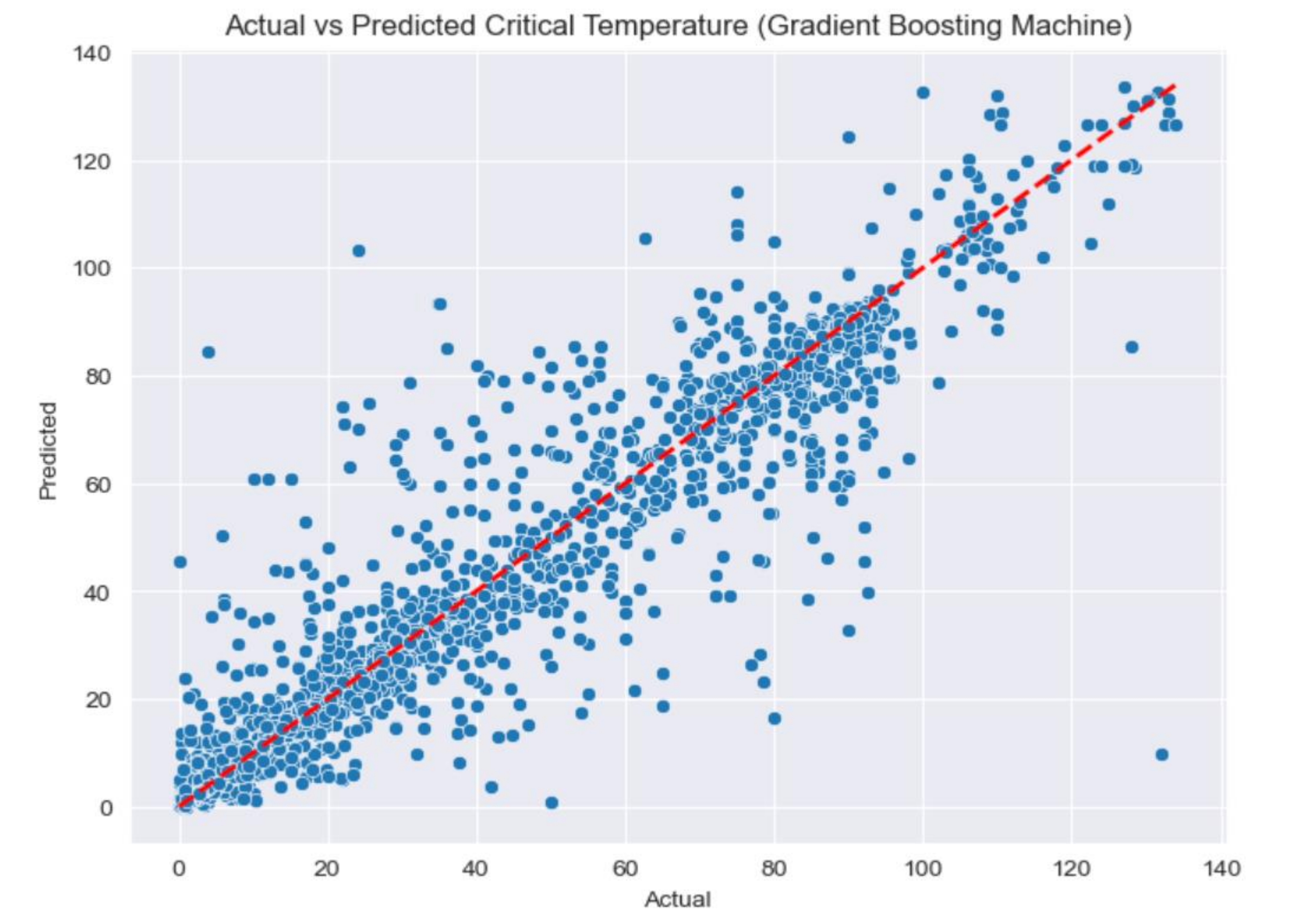


Gradient Boosting Regressor

- Gradient boosting algorithm known for its speed and performance
- Often used in winning solutions for machine learning competitions.

```
Training results:
Training RMSE: 4.20498
Training MAE: 1.35191
Training R2 score: 0.98495
Training Adjusted R2 score: 0.98492
```

```
Testing results:
Testing RMSE: 11.04418
Testing MAE: 5.77022
Testing R2 score: 0.89264
Testing Adjusted R2 score: 0.89110
```

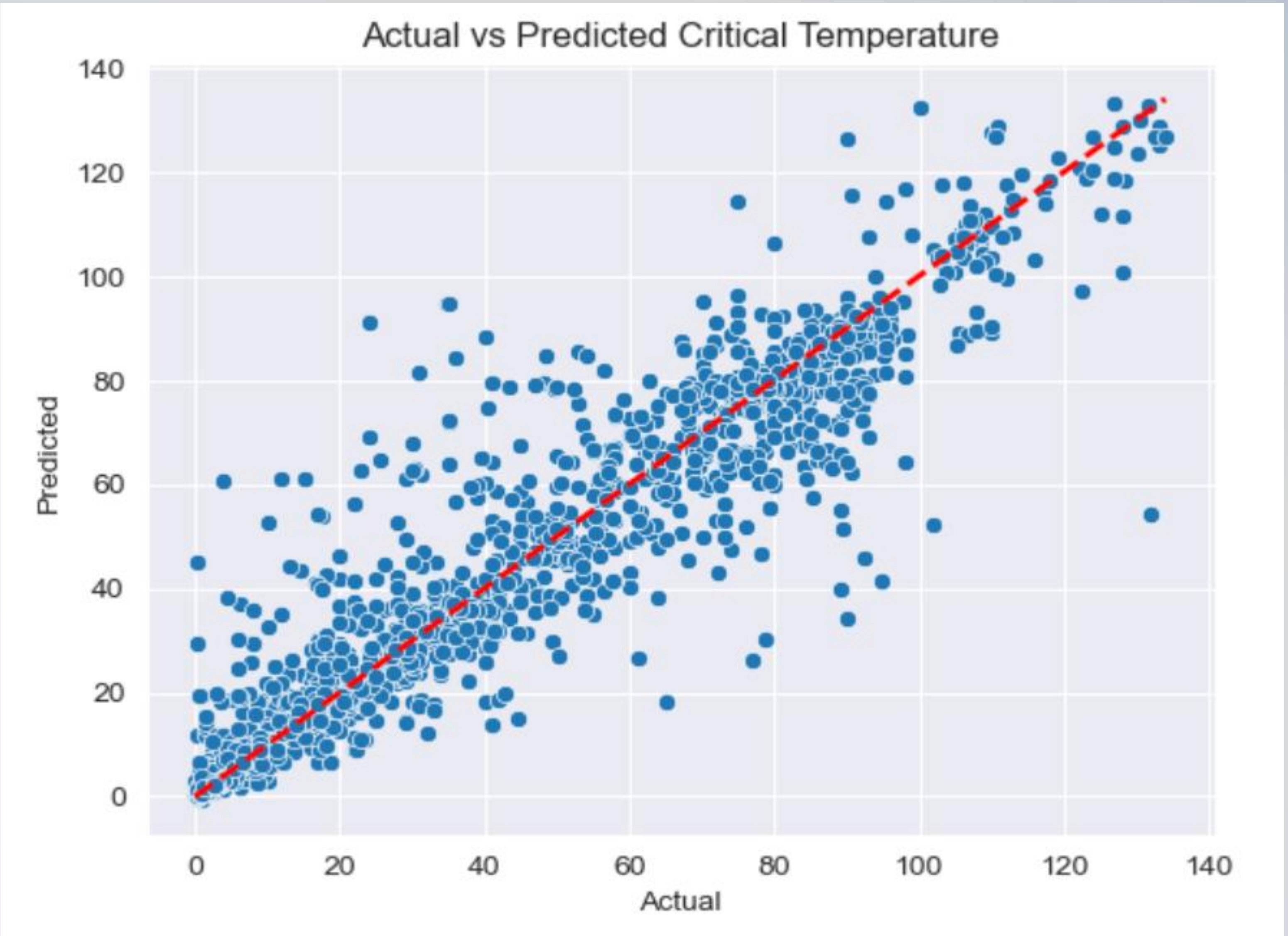


XGBoost Model- Winner of The Race

- **Selection of optimal parameters for the ML Model**
  - **learning_rate**: Step size shrinkage used to prevent overfitting. Range: 0.001 to 0.25.
  - **max_depth**: Maximum depth of each tree. Range: 6 to 20.
  - **n_estimators**: Number of trees in the ensemble. Range: 700 and 1500.
  - **subsample**: Fraction of the training data used for each tree. Values: 0.25, 0.5, and 0.75.
  - **reg_lambda**: L2 regularization term on weights Values: 0.1, 0.5, 1.0, and 1.5.

```
Training results:
Training RMSE: 4.56446
Training MAE: 2.00625
Training R2 score: 0.98226
Training Adjusted R2 score: 0.98224

Testing results:
Testing RMSE: 9.74952
Testing MAE: 5.29571
Testing R2 score: 0.91633
Testing Adjusted R2 score: 0.91514
```



Challenges

- Linear regression and GLM models being parametric had upper cap in explainable variability.
- Nonparametric ensemble models offered relatively better performance without any assumptions on the underlying data.
- Computational resource limitations remain a challenge, as improving model performance often requires computationally expensive hyperparameter tuning processes.

Conclusion

- Temperature control (Tc) of the conductor is the key challenge in application of Superconductor. This model can help the scientists to synthesize superconductor materials that suit Tc achievable at target environment.

- XGBoost Machine learning with GridSearchCV hyperparameter tuning predicted the critical temperature with R2 score of 91.63%

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

Thank You