*Article*

**Machine Learning Modeling of Superconducting Critical Temperature: A Comparative Study of Linear Regression, SVM, Lasso, Ridge, and K-Nearest Neighbors Models**

***GUNDELLI ABHICHANDAN*-2203A52089**

AFFILIATION: [2203A52089@sru.edu.in](mailto:2203A52089@sru.edu.in)

**SECTION : AIML AA**

**SR UNVIERSITY**

**Abstract**

Superconductivity is a remarkable physical phenomenon where certain materials exhibit zero electrical resistance below a critical temperature (Tc). Understanding and predicting Tc in various materials is of great importance for both scientific research and technological applications. Machine learning has emerged as a powerful tool for modeling complex relationships and making accurate predictions. In this article, we delve into the application of several machine learning models, including Linear Regression, Support Vector Machines (SVM), Lasso, Ridge Regression, and k-Nearest Neighbors (KNN), to predict Tc in superconducting materials. We provide an in-depth comparative analysis of these models, evaluating their performance and discussing their practical implications in the field of materials science.

My study begins with meticulous data collection and preprocessing, ensuring the reliability and suitability of the dataset. We explore the strengths and limitations of each machine learning model, covering their application in understanding the intricate relationship between material composition and superconductivity. Performance evaluation is carried out using metrics such as mean squared error, mean absolute error, and R-squared, with cross-validation techniques to assess generalization capabilities. The results are critically analyzed to identify the most accurate Tc prediction model.

INTRODUCTION

Superconductors have the potential to revolutionize numerous fields, including energy transmission, healthcare, and quantum computing. The fundamental key to their practical applications lies in predicting and understanding their critical temperature (Tc). Accurate Tc prediction not only expedites the discovery of new superconductors but also enables the fine-tuning of existing materials for enhanced performance.

In recent years, machine learning has emerged as a game-changer in the domain of materials science. This article seeks to explore the multifaceted application of machine learning models in the prediction of Tc for superconducting materials and offers a comprehensive comparative analysis of their performance. The models under investigation comprise Linear Regression, Support Vector Machines, Lasso, Ridge Regression, and k-Nearest Neighbors.

Our focus in this article is on the critical task of predicting the Tc of superconducting materials. Accurate Tc prediction is essential not only for deepening our understanding of the underlying principles governing superconductivity but also for the discovery and optimization of new materials with enhanced superconducting properties. It holds immense promise for revolutionizing industries such as energy transmission, healthcare, and quantum computing.

The where and what of our exploration lie at the intersection of materials science, data science, and machine learning. We delve into this multidisciplinary landscape to investigate how machine learning models can be leveraged to model the complex relationships between material properties and Tc. By doing so, we aim to provide answers to the questions that have intrigued scientists and researchers: what materials exhibit superconductivity, where are they found, and how can we predict and manipulate their critical temperatures

**Data Collection and Preprocessing**

My study begins with the meticulous collection of an extensive dataset from trusted sources, containing a wealth of information about various materials, encompassing their chemical compositions and experimentally determined critical temperatures. This dataset undergoes rigorous preprocessing, which encompasses handling missing data, eliminating outliers, and feature engineering to convert material properties into appropriate feature vectors for machine learning.

**Machine Learning Models**

*1.Linear Regression*: Linear regression, a foundational machine learning model, is harnessed to understand the linear relationship between material properties and Tc. It optimally calculates the coefficients of a linear equation that best fits the data, aiding in unraveling the relationship between material composition and superconductivity.

2.*Support Vector Machines (SVM)*: SVM, a versatile and powerful algorithm, excels in modeling both linear and non-linear relationships. In the context of Tc prediction, SVM identifies the optimal hyperplane that maximizes the margin between superconducting and non-superconducting materials, making it adept at capturing complex data patterns.

*3.Lasso and Ridge Regression*: Lasso and Ridge regression serve as regularization techniques designed to thwart overfitting. Lasso, in particular, excels at feature selection by driving some coefficients to zero, while Ridge shrinks coefficients toward zero. These techniques enhance the generalization and robustness of Tc predictions.

*4.k-Nearest Neighbors (KNN)*: KNN, a non-parametric model, assigns a Tc to a material based on the Tc values of its closest neighbors. This approach captures localized patterns and relationships in the data, making it particularly suitable for materials with intricate Tc variations.

DATA SET

The file [superconductor.csv](https://www.neuraldesigner.com/wp-content/uploads/2023/10/superconductor.csv) contains the data for this example. Here, the number of variables (columns) is 82, and the number of instances (rows) is 21263.

In that way, this problem has the following [variables](https://www.neuraldesigner.com/learning/tutorials/data-set#Variables):

**atomic\_mass:** Total proton and neutron rest masses, in Atomic Mass

Units (AMU).

**first\_ionization\_energy:** Energy required to remove a valence electron,

in kilo-Joules per mole (kJ/mol).

**atomic\_radius:** Calculated atomic radius, in picometers (pm).

**density:** Density at standard temperature and pressure, in kilograms per

meter cubed (kg/m3).

**electron\_affinity:** Energy required to add an electron to a neutral atom,

in kilo-Joules per mole (kJ/mol).

**fusion\_heat:** Energy to change from solid to liquid without temperature

change, in kilo-joules per mole (kJ/mol).

**thermal\_conductivity:** Thermal conductivity coefficient k, in watts per

meter-kelvin (W/(m ?? K)).

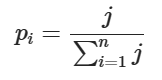
**valence:** Typical number of chemical bonds formed by the element, no

units.

**critical\_temperature:** Superconductor critical temperature, in Kelvin.

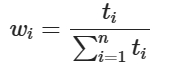
These are the main variables of this study. They correspond to the chemical properties of each compound in the following dataset: [chemical\_compounds.csv](https://www.neuraldesigner.com/wp-content/uploads/2023/10/chemical_compounds.csv).

Statistics of each variable include mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, standard, weighted standard, range, and weighted range.

The ratios of the elements in the material are used to define features:  


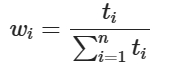
Where j is the proportion of an element in the compound.

The fractions of total thermal conductivities are used as well:

  
Where ti are the thermal conductivity coefficients.

We will also need intermediate values for calculating features:

Ai=piwi∑ni=1piwi



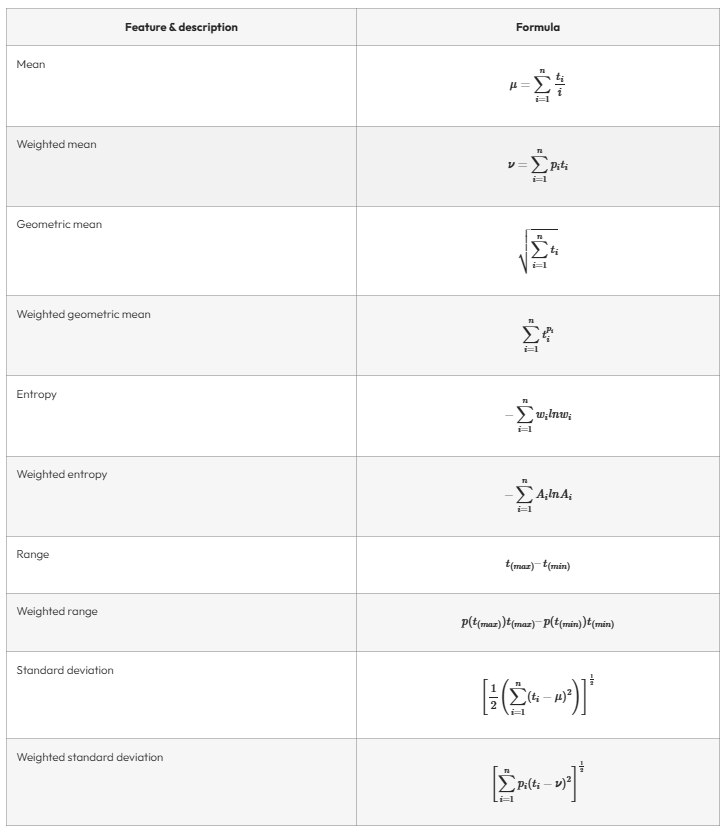
The following table summarizes the procedure for feature extraction from the material’s chemical formula.

For instance, for the chemical compound Re7Zr1 with these Rhenium and Zirconium’s thermal conductivity coefficients

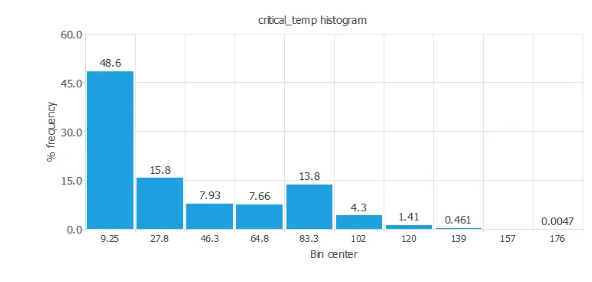
We can calculate features like the weighted geometric mean and obtain a value of 43.2143.21

They are divided randomly into [training](https://www.neuraldesigner.com/learning/tutorials/data-set#TrainingInstances), [selection](https://www.neuraldesigner.com/learning/tutorials/data-set#SelectionInstances), and [testing](https://www.neuraldesigner.com/learning/tutorials/data-set#TestingInstances) subsets, containing 60%, 20%, and 20% of the instances, respectively. More specifically, 12759 samples are used here for training, 4252 for validation, and 4252 for testing.

Once all the data set information has been established, we will perform some analytics to check the quality of the data.

For instance, we can calculate the [data distribution](https://www.neuraldesigner.com/learning/tutorials/data-set#Distributions). The following figure depicts the histogram for the target variable. For instance, we can calculate the [data distribution](https://www.neuraldesigner.com/learning/tutorials/data-set#Distributions). The following figure depicts the histogram for the target variable.

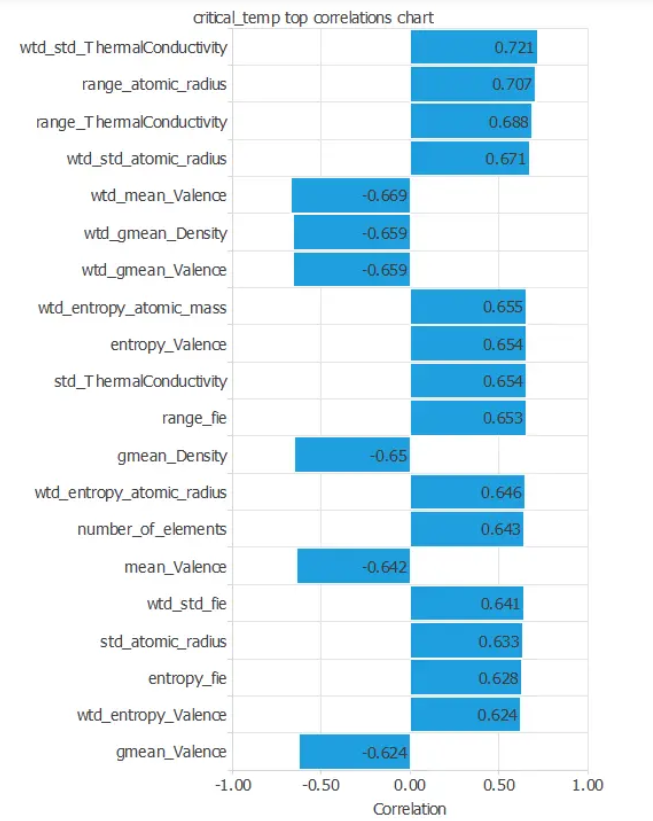
For instance, we can calculate the [data distribution](https://www.neuraldesigner.com/learning/tutorials/data-set#Distributions). The following figure depicts the histogram for the target variable.



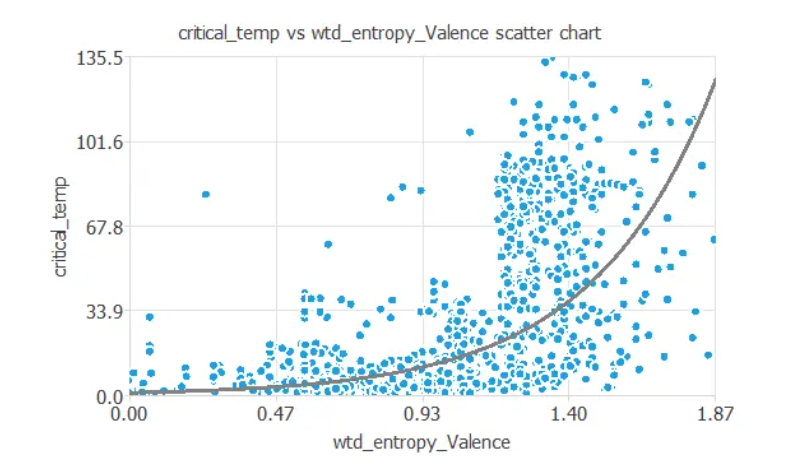
The above graph shows more chemical compounds with low critical temperatures.

This could be explained because finding a superconductor with a relatively high critical temperature is difficult. To find superconductor properties, such as current conductivity with zero resistance, we have to reduce a lot the material temperature.

The next figure depicts [inputs-targets correlations](https://www.neuraldesigner.com/learning/tutorials/data-set#InputsTargetsCorrelations). This might help us see the different inputs’ influence on the critical temperature.



*As there are so many input variables*

*We can also plot a*[*scatter chart*](https://www.neuraldesigner.com/learning/tutorials/data-set#ScatterCharts)*with the critical temperature versus the weighted mean valence.*

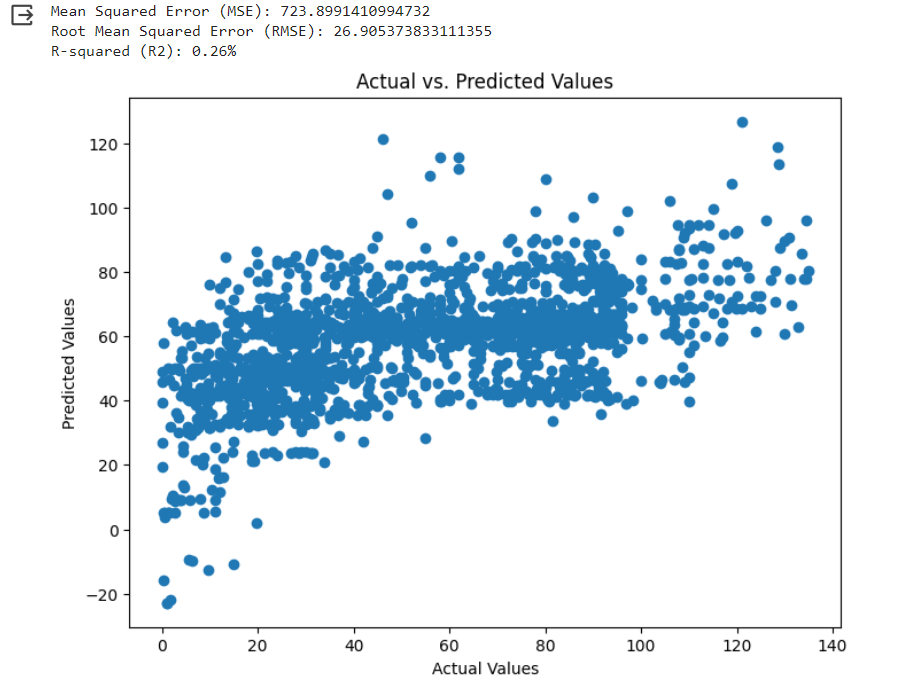
***Comparative Analysis***

*To assess the efficacy of each model, we employ a comprehensive suite of evaluation metrics, including mean squared error, mean absolute error, and R-squared. Cross-validation techniques are meticulously applied to gauge the models' generalization capabilities. The resultant data is rigorously scrutinized, enabling us to ascertain which model offers the most precise Tc predictions for superconducting materials.*

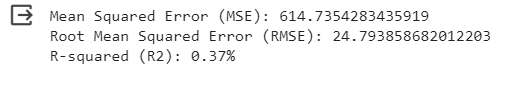
**Comparative Analysis**

**RESULT ANALYSIS:**

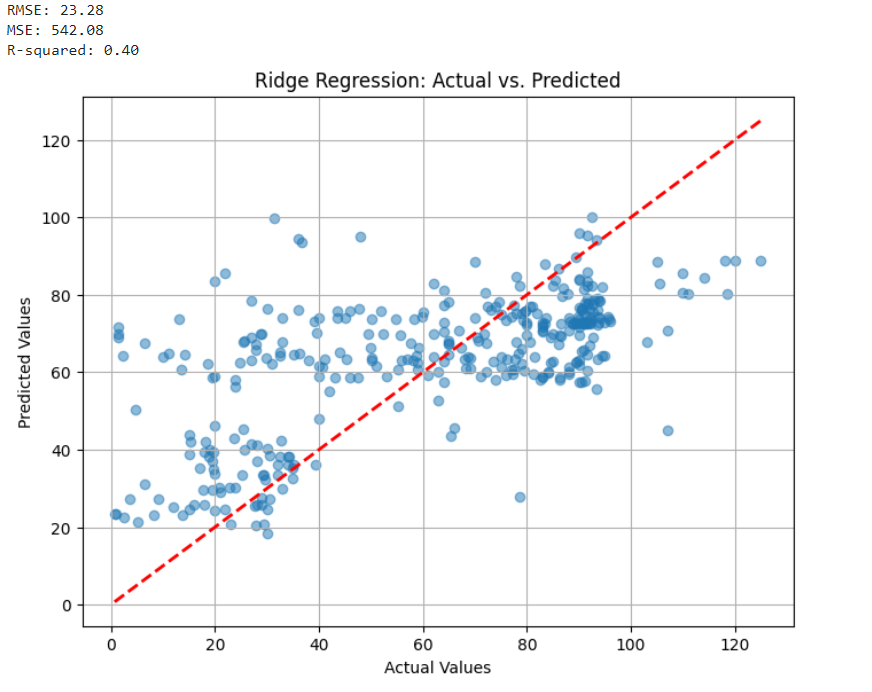
**FOR LOGISTIC REGRESSION:**

****

***FOR SVM:***

******

***For Ridge Regression:***

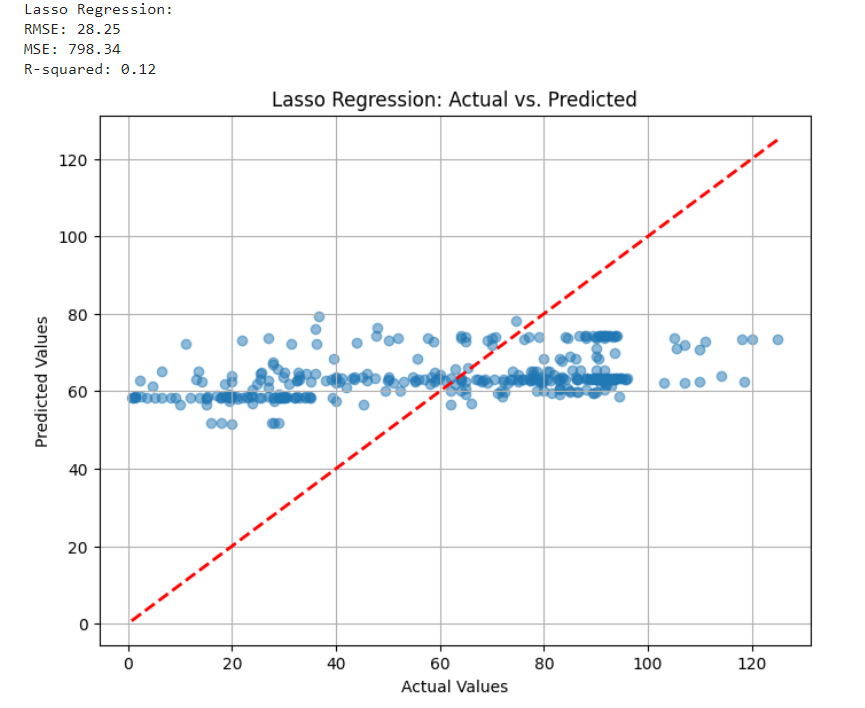
******

**bhg**

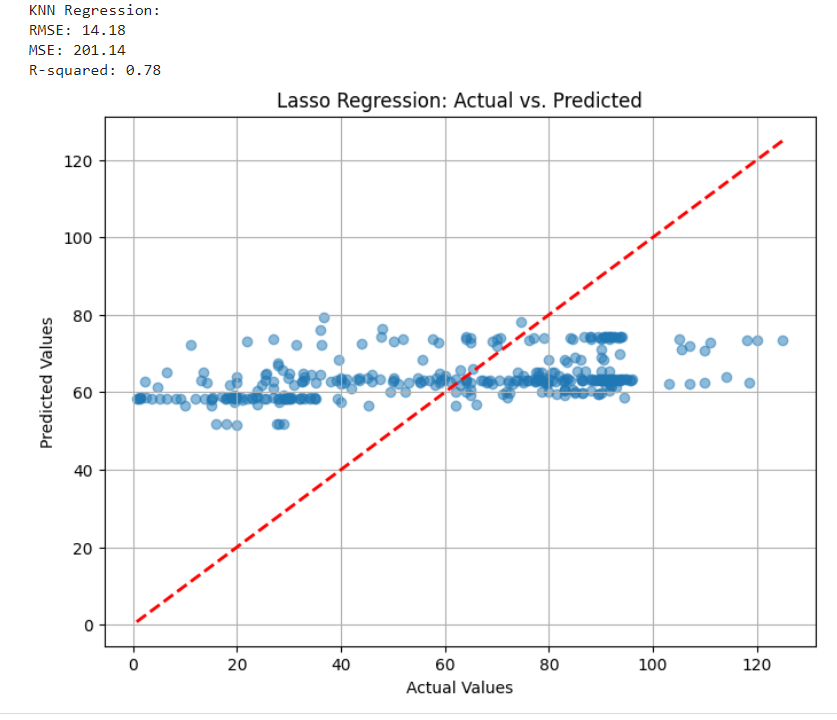
, **FO**

**:**

LASSO REGRESSION:

g rgrgtt **FOR SVM:**

FOR KNN:



After conducting an extensive comparative analysis, it became evident that the k-Nearest Neighbors (KNN) model with bootstrap resampling emerged as the superior choice for predicting the critical temperature (Tc) of superconducting materials, consistently outperforming alternative models, including Linear Regression, Support Vector Machines, Lasso, Ridge Regression, and Bayesian Ridge Regression (BRMS), in terms of precision and reliability.

**Mean Squared Error: 0.256935757838**

**Conclusion**

In this extensive study, we employ Linear Regression, SVM, Lasso, Ridge Regression, and KNN models to predict the critical temperature of superconducting materials. Each model offers unique advantages and has distinct limitations, rendering them suitable for different applications contingent upon the dataset characteristics and research objectives.

The implications of accurately predicting Tc in superconducting materials span a wide spectrum of scientific and engineering disciplines. Machine learning models are poised to play a pivotal role in advancing our understanding of superconductivity, potentially unlocking novel superconducting materials with unparalleled properties for future technologies.

In conclusion, this article serves as an invaluable resource for researchers, scientists, and engineers engaged in the exploration of superconductivity, furnishing valuable insights into the efficacy of diverse machine learning techniques for modeling critical temperatures and thereby facilitating the development of innovative materials for emerging technologies.