Predicting the Critical Temperature of a Superconductor

Muskan Kothari
Computer Science and Engineering
PES University
Bangalore, India
muskan.kothari0120@gmail.com

Neha Arun Angadi
Computer Science and Engineering
PES University
Bangalore, India
nehaangadi19@gmail.com

Shreyas Raviprasad
Computer Science and Engineering
PES University
Bangalore, India
shreyasraviprasad457@gmail.com

Abstract—**The purpose of this project is to analyse the various approaches inherited to solve the problem of predicting the critical temperature of superconductors to formulate a model that can best predict the critical temperature. The elemental chemical properties of superconductors are many. Various models aim to narrow it down to the most contributing features that yield the closest prediction. A close look at the correlations between the features and T_c enables a thorough analysis of pitfalls and contributions. Various approaches deal with statistical machine learning models, featureless approach, Bayesian Neural Network approach and the like for nearly optimized predictions.**

Keywords—critical temperature, superconductor, multiple linear regression, regression tree

I. Introduction

Superconductivity of materials is a property by which materials conduct current with zero resistance, which occurs at or below a certain temperature called the critical temperature. Research existing on this property has been prevailing for years. In the field of material science, this allows scientists to synthesise new materials or gauge the optimal temperature of current flow with minimal resistance, whether a certain temperature is feasible and if yes, to what extent must the material be cooled.

Revolutionary new solutions needed to tackle the power grid challenge calls for efficient study and evaluation of superconductivity of materials. Various studies and models have incorporated efficient models that predict this macroscopic property that relates with other chemical properties intricately.

Our paper aims to review and analyse the various existing models, their performance and formulate a model that predicts the critical temperature of materials using a database of recorded superconductivity. This can be used for further research in the telecom industry.

II. Review of literature

A. Predicting Critical Temperature of a Superconductor

This assessment [1] of the NIMS dataset provided a thorough exploratory analysis along with useful insights.

Descriptive analysis was performed which showed that not all distributions were normal and possessed some degree of skewness. This would thus, require data transformation and scaling to be done to have standardized features.

During model development, 14 different models were used which included regression, regularization, instance-based, tree-based, dimensionality reduction and ensemble methods. Since there was data transformation involved, it totaled to 28 models. The MSE was reported for each model along with the standard error, p-value and R^2 values. The calculated p-values allowed the author to extract the most important or significant features in the data. Linear regression was carried out using forward, backward and stepwise selection to get an optimal number of features. Principal component regression performed the worst among all the models while ensemble tree-based algorithms out performed all other models.

Model predictions were plotted and the most important features used for predicting the critical temperature were reported rank wise. Valence and Thermal Conductivity are the two most important features. Future scope includes including more ensemble learning methods on top of the existing models and improving hyperparameter tuning.

The author provides extensive comparison between different models used for the prediction and the shortcomings of each.

B. Featureless approach for predicting Critical Temperature of Superconductors

The authors [2] focused on the accuracies of different ML algorithms applied on the dataset containing the chemical formulas of superconductors. The prediction was made without considering the features from the superconductors chemical formula.

After some analysis in this field, they concluded that though ML, deep learning and AI give good results, there can be more advancement in the research by using a featureless approach in predicting the critical temperature.

The molecule chemical composition formula was used with supervised ML techniques. The chemical formula was sent as an input to the ML techniques to give the critical temperature as the output. The final analysis was done by measuring various errors on different algorithms and the results were recorded.

Upon analysing the results and visualizing them graphically, the authors concluded that featureless analysis using advanced ML algorithms like SVM, SVM with RBF and XGBoost are useful in predicting critical temperature. PCA is also helpful as it helps reduce the data dimension.

C. Machine learning modeling of superconducting critical temperature

In this research [3], several ML schemes were developed to model the critical temperatures of 12,000+ known superconductors using the SuperCon database.

Only coarse-grained features were used. Instead of the heuristic approach of classifying the superconductors based on critical temperature, random forests and simplex fragments were applied on the structural properties data from the AFLOW online repositories.

Magpie was used to convert the information on chemical composition into a meaningful set of attributes which included mean and std deviation of 22 different elemental properties. For the classification, to set the threshold critical temperature, a series of random forest models were trained. Precision, recall and F1 score were also used along with accuracy as metrics for classification.

Their model achieved an R^2 value of 0.88, signifying that the random forest algorithm was a flexible and powerful method. A final accuracy and F1 score of about 92% was achieved.

The models were combined into a single pipeline and then searched the entire ICSD dataset for the possibility of new superconductors.

D. Prediction of critical temperature and new superconducting materials

This paper [4] applied ML models to speculate the critical temperature using the same NIMS dataset. The change here is how the dataset is preprocessed. It was found that multiple critical temperature values were reported for the same or extremely similar compounds. It is speculated that the data gathered came from a variety of laboratories which might have caused the discrepancy.

Three different datasets were extracted from this. The logic behind dividing into three separate sets is that critical temperature highly depends on sample size as well as the number of defects; therefore, the same material may have multiple values for critical temperature.

Outliers were removed by keeping only those samples whose values remained within 3 standard deviations from the mean of each feature.

Random forest regression was used and it performed well giving an average R^2 value of 0.9. In order to further improve the model's accuracy, the data was broken down based on the quantity of chemical elements as well as the ratios. The model performed noticeably better this time around with an average R^2 of 0.94. After further model evaluation, it was observed that outlier removal did not benefit the quality and sc_mean data provided the best output. While sc_mean did provide better results, sc_min was used for alloys particularly for their relatively low critical temperatures. The model does not comment on the presence of superconductivity, rather the critical temperature of a compound that does. The authors further recommend using linear regression models and nonlinear dependencies for prediction of critical temperatures.

E. Data-driven statistical model for critical temperature prediction

The approach that was accepted here [5] was entirely data-driven. A statistical model was constructed with 21, 263 rows of superconductors after data processing. In terms of percentage, this was 67% of the original dataset.

One of the crucial keys to note is that this model does not predict whether a material is a superconductor, rather gives predictions for superconductors.

A total of 81 features were extracted from each superconductor and one 1 additional column of the observed T_c values. After a thorough analysis with various statistical models, two models were considered in this approach: A multiple regression model that served as the benchmark and a gradient boosted model as the key predicting model.

The gradient boosted models work closely with trees which account for the points which are difficult to predict and encounter the various intricate interactions between features. Associated with the data preparation process was the attempt to reduce the dimensionality. On prior analysis, it was found that a certain number of features resulted in high correlation. This observation motivated dimensionality reduction using Principal Component Analysis (PCA), but the returns in terms of improvement and benefits were not appreciable. This was due to the fact that a large number of principal components were required for a substantial percentage of data variation and hence, the PCA approach was abandoned.

The latest improvement called XGBoost was used to improve on the performance further and returned an out-of-sample prediction of 9.5K based on RMSE and an out-of-sample R2 values of 0.92 for one out of the 750 trees generated by the XGBoost model, which is considerably well.

Feature importance was evaluated using the XGBoost model which did so using the gain obtained for each feature. The result obtained listed the top 20 features that contribute most to the T_c prediction.

F. Variational Bayesian Neural Network approach

The SuperCon database was obtained from NMIS here as well. This study [6] examined and evaluated an approach for prediction in twofold and made use of the Bayesian Neural Network which is a generative machine-learning framework and was the focus of the approach.

The prediction was based on superconductors, chemical elements and formulas. The dataset followed a 70-30 split for train and test set without the use of validation set to explore the effect of VBNN after overcoming the overfitting challenge.

The experiment converged with 1000 epochs, batch size of 10 and 100 hidden layers. The performance of the model was shown to have the R^2 value very close to the best model (0.94), with the RMSE value of 3.83 K.

III. Dataset

For the purpose of predicting the critical temperature of a superconductor, we have chosen the SuperCon database [7] maintained by Japan's National Institute for Materials Science (NIMS).

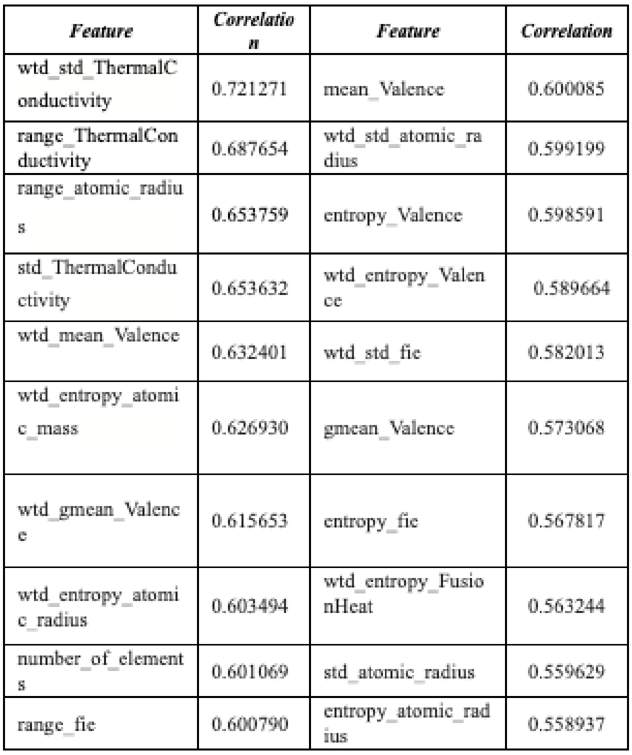
Upon analysing the dataset, it was found that though there was no missing data, duplicates had to be taken care of. The dataset consists of 81 continuous valued features along with 21263 recorded critical temperatures. The 82nd feature is the target column consisting of all critical temperatures.

Each superconductor had 8 main chemical properties which further had 10 more features extracted using a multitude of statistical transformations such as weighted mean, geometric mean, entropy and so on. Another dataset includes the chemical composition of each known superconductor in the NIMS dataset.

IV. Initial Insights

From the EDA performed, the following 20 features showed the highest correlation with the critical temperature.

TABLE I. Top 20 correlations



Feature	*Correlation*	*Feature*	*Correlation*
wtd_std_ThermalConductivity	0.721271	mean_Valence	0.600085
range_ThermalConductivity	0.687654	wtd_std_atomic_radius	0.599199
range_atomic_radius	0.653759	entropy_Valence	0.598591
std_ThermalConductivity	0.653632	wtd_entropy_Valence	0.589664
wtd_mean_Valence	0.632401	wtd_std_fie	0.582013
wtd_entropy_atomic_mass	0.626930	gmean_Valence	0.573068
wtd_gmean_Valence	0.615653	entropy_fie	0.567817
wtd_entropy_atomic_radius	0.603494	wtd_entropy_FusionHeat	0.563244
number_of_elements	0.601069	std_atomic_radius	0.559629
range_fie	0.600790	entropy_atomic_radius	0.558937

V. Problem Statement and Proposed Solution

The goal of this project is to develop a model which predicts the critical temperature of a superconductor with acceptable accuracy. The model would be able to predict temperatures for unseen compounds as well.

Our approach differs from current work on the topic where we plan to make use of both, the chemical composition and the properties of the material. As far as our knowledge goes, models have only taken into consideration either only the chemical properties or only the chemical composition and not both.

Our initial approach consists of combining both datasets to form a 168 feature table from which multiple models such as regression, decision trees, PCA and so on will be trained. We also plan to use some sort of ensemble method where tree-based algorithms are trained on both datasets separately. The model giving the best performance will be considered for predictions.

First, we explored various models without combining the datasets. Few basic models were trained on the dataset, like linear regression, ridge regression, decision tree and gradient boosting regressor. The same set of models were trained with PCA to gauge the range of values obtained. Along with the aforementioned models, SVR was trained but the performance was far worse without standardisation. In order to improve the performance, we applied min max scaling and standardisation.

TABLE II. Performance without combining datasets

Model	*MSE*	
	Without PCA	*With PCA*
Linear Regression	313.6330452225	382.456054492577
Ridge Regression	314.3548454210	382.456052536200
Decision Tree Regressor	153.7706196165	173.146755351791
SVR	619.7967703667	---
Gradient Boosting Regressor	211.7878190974	287.20209054409

Applying our planned novel approach, models were trained after concatenating the dataset of all features and formulas of superconductors.

TABLE III. Performance with combining datasets

Model	***MSE***	
	Without PCA	***With PCA***
Linear Regression	284.4167808756412	295.13052744966564
Ridge Regression	284.0009980627075	295.132424330563
Decision Tree Regressor	140.58162798406	188.26223188434565
SVR	257.13038106166226	---
Gradient Boosting Regressor	197.33213026416945	266.9759761098954
Random Forest	---	151.6991765037624

The best performance was obtained when the features for the first dataset were narrowed down to the top 20 features which were highly correlated with the critical temperature combined with all of the features in the second dataset.

TABLE IV. Performance with top 20 features

Model	***MSE***
Linear Regression	322.8400354768235
Ridge Regression	322.77551489878005
Decision Tree Regressor	137.267794040878
Gradient Boosting Regressor	195.64679389008222
SVR	603.9938419467876

VI. Results and Conclusion

To summarize the performances, the following models gave the best results:

TABLE V. top 3 models

Model	***MSE***
Decision Tree Regressor (Top 20 features)	137.267794040878
Decision Tree Regressor (Combining datasets)	140.58162798406
Random Forest (Combining datasets)	151.69917650376246

Decision tree proved to be the most suitable model irrespective of the approach chosen. The MSE values were also closely related.

Following is the graph of the actual (orange) and predicted values (blue) of the critical temperatures for the decision tree regressor, linear regression and SVR respectively, after combining the tables for features and formulas and extracting only the 20 features that gave the best correlation with critical temperature.

FIGURE I. decision tree regressor predictions

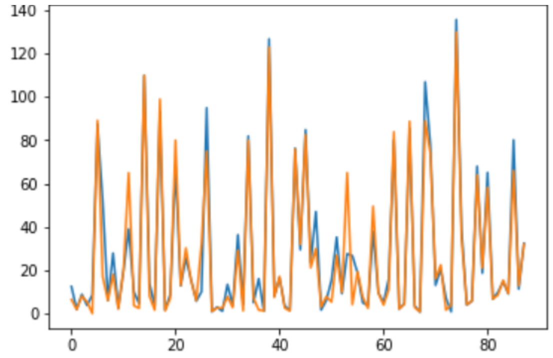


FIGURE II. Linear regression predictions

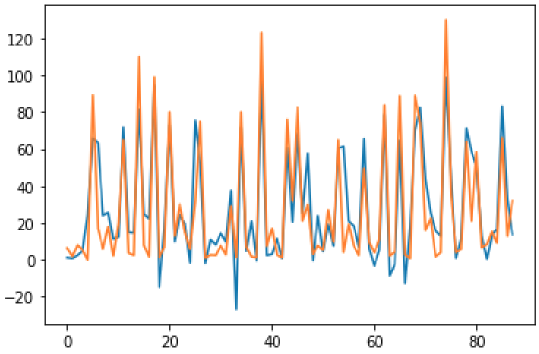
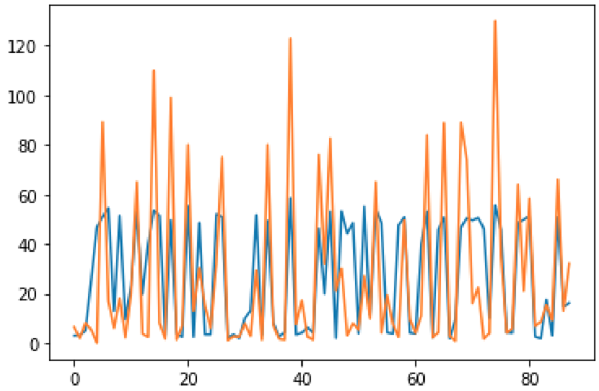


FIGURE III. svr predictions



As observed in figures 1, 2 and 3, we can conclude that Decision Tree Regressor gives the least error when compared to the other generally used models.

Our approach of making use of both datasets showed promising results as all models showed improvement while decision tree regressor showed an approximate improvement of 8.57% in MSE.

The decision tree regressor showed an approximate improvement of 10.73% in MSE.

The Decision tree regressor can be used in predicting the superconductor's critical temperature within reasonable margins. The applications vary from MRI machines to 6G telecommunication systems.

VII. Acknowledgment

We would like to express our profound gratitude to Dr. Gowri Srinivasa and the entire team of TAs of the course, for encouraging and providing us with this opportunity to get hands-on experience in the field, and guiding us along the way. We would also like to thank the Computer Science and Engineering department at PES University, for always inspiring us to conduct frequent research and inculcating a problem-solving discipline in us.

VIII. References

[1] https://github.com/robertvici/Predicting-the-Critical-Temperature-of-a-Superconductor/blob/master/28243447_FIT5149_Ass1.pdf

[2] M. Gaikwad and A. R. Doke, "Featureless approach for predicting Critical Temperature of Superconductors," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225447.

[3] Stanev, V., Oses, C., Kusne, A.G. *et al.* Machine learning modeling of superconducting critical temperature. *npj Comput Mater* 4, 29 (2018). https://doi.org/10.1038/s41524-018-0085-8

[4] Matasov, A., Krasavina, V. Prediction of critical temperature and new superconducting materials. *SN Appl. Sci.* 2, 1482 (2020). https://doi.org/10.1007/s42452-020-03266-0

[5] Hamidieh, K. (2018). A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor. *arXiv: Applications*.

[6] https://arxiv.org/abs/2002.04977

[7] https://supercon.nims.go.jp