2022 3rd International Conference for Emerging Technology (INCET)
Belgaum, India. May 27-29, 2022

Tree Based Models for Critical Temperature Prediction of Superconductors

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

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

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

2022 3rd International Conference for Emerging Technology (INCET) | 978-1-6654-9499-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INCET54531.2022.9825251

***Abstract*—The purpose here is to analyze the various approaches inherited to solve the problem of predicting the critical temperature of superconductors and formulate a model that can best predict the critical temperature of a given compound and for unseen compounds as well. In the field of material science, this allows scientists to synthesize 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. Various models aim to narrow it down to the most contributing features that yield the closest prediction. In this paper, the approach is to implement various models that consider both the chemical composition as well as chemical properties of a compound. Finally, comparative results of tree based models are shown which give a considerable improvement in MSE by 23.72%.**

Keywords—critical temperature, superconductor, multiple linear regression, decision tree regressor, random forest, normalization

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. Practically speaking, it requires cooling the compounds to extremely low temperatures. This requires time, consistent experimenting, domain expertise and intuition, cost, equipment, and many trial and errors until a solution or a concrete result is obtained. Sometimes, it may not be feasible to conduct these methods. It may also require time to gauge the chemical properties and composition of certain unseen compounds. It has to be known before an appropriate environment can be created or a prediction can be made on whether the compound can exhibit superconductivity. The amount of compounds that may show this property usually turns out to be <5% from a sample of compounds.

Revolutionary new solutions needed to tackle the power grid challenge calls for efficient study and evaluation of superconductivity of materials. Various studies deal with statistical machine learning models, Bayesian Neural Network approach and the like for nearly optimized predictions. One of the novel approaches talked about a featureless method to predict the critical temperature.

This paper explains the attempt at formulating a robust model and exploring the performance of the same that predicts the critical temperature of materials using a database of a combination of chemical properties as well as the chemical composition of various compounds. This can be used for further research in the telecom industry.

II. Related works

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 R2 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 analyzing 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 R2 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 R2 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 R2 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 the prediction.

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.

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 Tc values. 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. The result obtained listed the top 20 features that contribute most to the Tc 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 performance of the model was shown to have the R2 value very close to the best model (0.94), with the RMSE value of 3.83 K.

III. METHODS

For the purpose of predicting the critical temperature of a superconductor, the SuperCon database [10] maintained by Japan's National Institute for Materials Science (NIMS) is chosen. Upon analyzing 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.

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

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.

The proposed approach differs from current work where both the chemical composition and the properties of the material are used. For the previous work done in this field, models have only taken into consideration either only the chemical properties or only the chemical composition and not both in.

The 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. There is also a 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.

Combining the two datasets could show potential improvement in predicting the critical temperature. The weighted significance of each property and the importance of each atom towards the final prediction can be analyzed through the proposed approach. This shows the interdependence between the features of the two datasets comprising the chemical formulas and properties of the superconductors.

First, various models were explored 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 standardization. In order to improve the performance, min max scaling and standardization were applied.

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 the discussed 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

IV. RESULTS AND DISCUSSION

To summarize the performances, following are the top 3 models that 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.

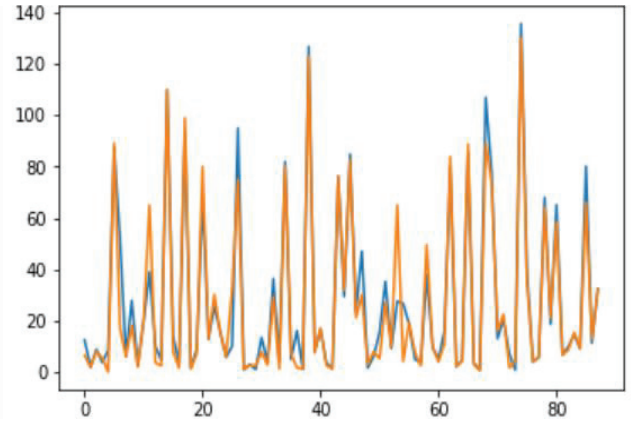

Fig. 1. Decision tree regressor predictions

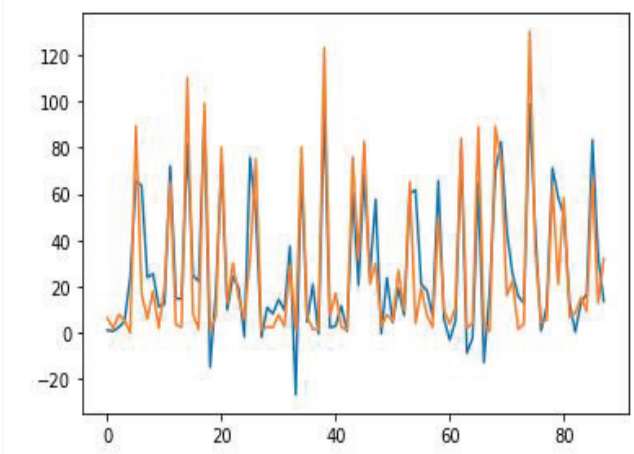

Fig. 2. Linear regression predictions

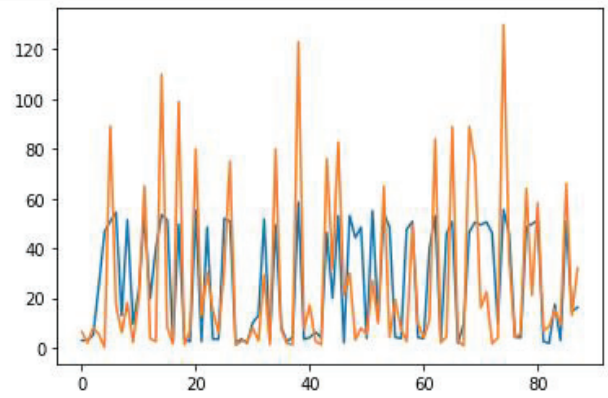

Fig. 3. SVR predictions

As observed in figures 1, 2 and 3, it can be concluded that Decision Tree Regressor gave the least error when compared to the other generally used models.

These models were trained with a train-test split of 67-33 and a random state of 42. To further optimize the model parameters, specifically for the decision tree regressor for methods of combining composition and properties as well as top 20 features, lower train-test splits and higher random states were tried. Finally, an even better MSE score was obtained for both the decision tree models at a split of 0.22 and random state of 50. Below table summarizes the results of top two models with new values:

TABLE I. TOP 2 MODELS WITH HYPERPARAMETER TUNING

Model	*MSE*
Decision Tree Regressor (Top 20 features)	117.29802796276137
Decision Tree Regressor (Combining datasets)	118.06722874833548

The proposed approach gives predictions for only the known superconductors due to the lack of data on chemical properties for new or undiscovered superconductors. However, a close approximation of chemical properties for new compounds can be done using algorithms like K-Means by comparing their chemical formulas. This is where the holistic aspect of this approach can be used for future work on unseen compounds.

The capabilities of superconductors in generating magnetic fields and their recent usage in building quantum computers can bring in advancements in the field. The applications can vary from MRI machines to 6G telecommunication systems.

V. CONCLUSION

It can be seen that the same model for different approaches eventually converges to nearly the same result. The approach of making use of both datasets showed promising results as all models showed improvement while decision tree regressor showed an approximate improvement of 23.72% in MSE compared to the MSE obtained without combining both features and without PCA applied. Random

forest performed the second best. Thus, tree based models in comparison to the other models performed significantly better on the proposed approach.

The Decision tree regressor can be used in predicting the superconductor's critical temperature within reasonable margins.

ACKNOWLEDGMENT

Expressing profound gratitude to Dr. Gowri Srinivasa and the entire team of TAs of the course, for encouraging and providing with this opportunity to get hands-on experience in the field, and guiding us along the way. It is also worth expressing gratitude to the Computer Science and Engineering department at PES University, for conducting frequent research and inculcating problem-solving disciplines.

REFERENCES

[1] Robert, "Predicting the critical temperature of a superconductor", unpublished.

[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

[3] Stanev, V., Oses, C., Kusne, A.G., Rodriguez, E., Paglione, J.,

[4] Curtarolo, S. and Takeuchi, I., 2018. "Machine learning modeling of superconducting critical temperature". npj Computational Materials, 4(1), pp.1-4.

[5] Matasov, A., Krasavina, V., "Prediction of critical temperature and new superconducting materials". SN Appl. Sci. 2, 1482 (2020).

[6] Hamidieh, K., 2018. "A data-driven statistical model for predicting the critical temperature of a superconductor". Computational Materials Science, 154, pp.346-354.

[7] Le, T.D., Noumeir R., Quach H., "Critical temperature prediction for a superconductor: a variational bayesian neural network approach"

[8] Owolabi, T., Akande, A., and Olatunji, S. (2014). "Prediction of superconducting transition temperatures for fe-based superconductors using support vector machine". 35, 12–26.

[9] J. Drugowitsch, "Variational bayesian inference for linear and logistic regression," arXiv preprint arXiv:1310.5438, 2013.

[10] B. Kailkhura et al., "Reliable and explainable machine learning methods for accelerated material discovery," arXiv preprint arXiv:1901.02717, 2019.

[11] Matthias, B. T. (1955). "Empirical relation between superconductivity and the number of electrons per atom". Phys. Rev., 97, 74–76.

[12] Japan's National Institute for Materials Science, Superconducting

[13] Material Database (SuperCon), 2019.

[14] Konno, T., Kurokawa, H., Nabeshima, F., Sakishita, Y., Ogawa, R.,

[15] Hosako, I. and Maeda, A., 2018. Deep Learning Model for Finding New Superconductors. arXiv preprint arXiv:1812.01995.