Regression Analysis of Superconductivity Data

[1]Md Mushtaque Tahmid[1†]

I. Background

The dataset used in this project is a superconductivity dataset introduced by Hamidieh (2018) [1] and made available through the UCI Machine Learning Repository. It contains information on 21,263 superconductors, each represented by 81 derived features that summarize important elemental properties such as atomic radius, valence, thermal conductivity, electron affinity, and atomic mass. These features were calculated as statistical aggregates (mean, standard deviation, range, etc.) based on the chemical composition of each material. The target variable is the critical temperature (Tc), the point at which the material becomes superconducting. The data is multivariate with real-valued attributes, and the main task associated with it is regression. The purpose of this project is to develop predictive models capable of estimating the critical temperature of superconductors from these derived material properties.

II. Analysis Results

A. *Linear Regression*

We first applied Multiple Linear Regression using all 81 derived features from the superconductivity dataset as predictors, and the critical temperature (Tc) as the response variable. The dataset was split into 70% training and 30% testing. The performance metrics are:

- Test RMSE: 17.76
- Test MAE: 13.42
- Test R^2 : 0.728

This indicates that the linear regression model explains around 73% of the variance in Tc, but still leaves substantial error unexplained.

Predicted vs Actual Tc: The scatter plot in Figure 1 shows that predictions roughly follow the diagonal line, but with clear deviations at higher Tc values. The model underestimates higher critical temperatures, indicating limited flexibility.

Residual Plot: Residuals in Figure 2 are spread around zero, but show a funnel shape (variance increases with predicted Tc)

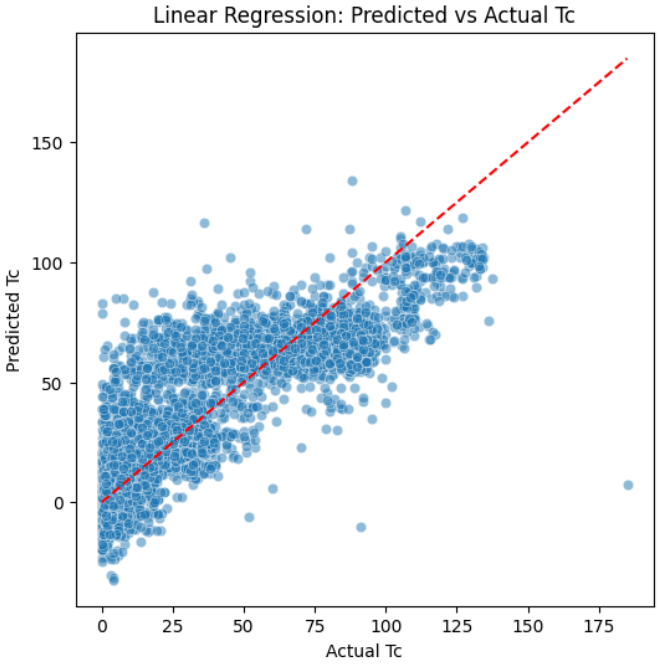

Fig. 1: Linear Regression: Predicted vs Actual Tc

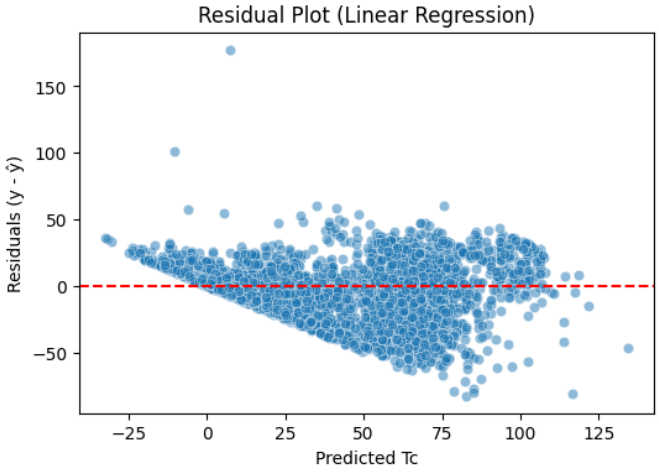

Fig. 2: Residual Plot (Linear Regression

B. *KNN Regression*

We applied a K-Nearest Neighbors (KNN) regression model to predict the superconducting critical temperature (Tc). Since KNN is distance-based, we standardized all predictor variables using z-score scaling. The dataset was split into 70% training and 30% testing, and model performance was evaluated across different values of k. The evaluation identified the optimal number of neighbors as:

- Best k: 2
- Test RMSE: 10.51
- Test MAE: 5.59
- Test R^2 : 0.905

This represents a significant improvement over linear regression, showing that KNN captures nonlinearities in the data more effectively.

[1] The author is with the Department of Civil Engineering, University of Central Florida, Orlando, USA
[†] Corresponding Author

RMSE vs k curve: The error in Figure 3 decreases sharply from k=1 to k=2, reaching a minimum at k=2. Beyond this point, RMSE increases as k grows, indicating that larger neighborhoods over-smooth the predictions.

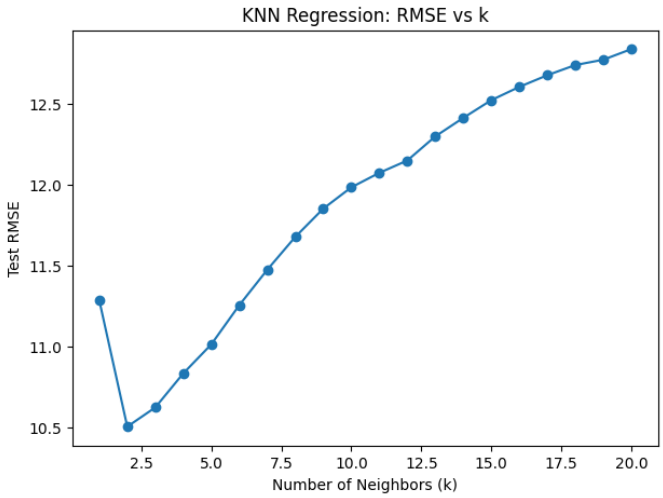

Fig. 3: RMSE vs k

Predicted vs Actual Tc: Predictions align closely with the diagonal line, suggesting good agreement between predicted and actual values. However, some scatter remains, particularly at higher Tc values. It is illustrated in Figure 4.

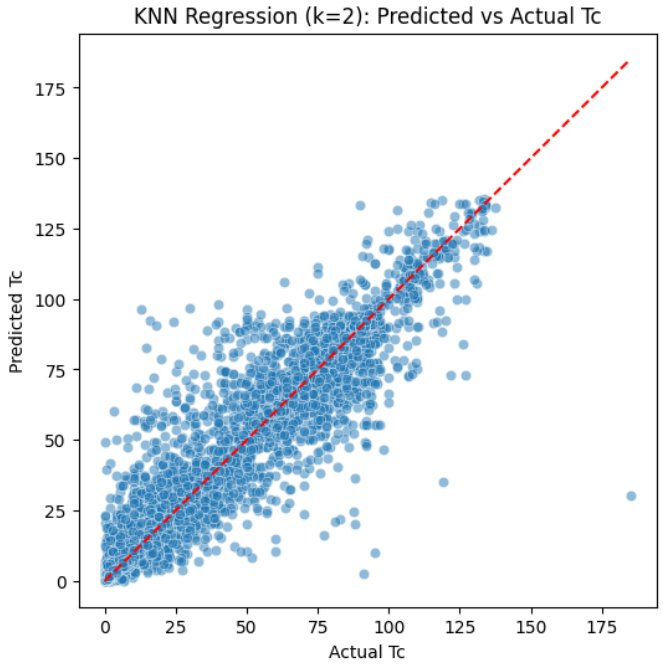

Fig. 4: KNN Regression: Predicted vs Actual Tc

It was seen to outperform linear regression by modeling local patterns. Simple, nonparametric method that adapts well to nonlinear relationships. KNN regression provides a strong performance boost over linear regression, achieving higher accuracy and explaining 90% of the variance in Tc.

C. *Random Forest Regression*

We implemented a Random Forest Regressor, an ensemble method that builds multiple decision trees and averages their outputs. This approach reduces variance, captures nonlinearities, and is well-suited for tabular datasets with many derived features [2]. The model was trained with 200 trees using the default parameters for splitting. Compared to linear regression and KNN, Random Forest provided the lowest prediction error and the highest variance explained. Results:

- Test RMSE: 9.59
- Test MAE: 5.31
- Test R^2 : 0.921

Predicted vs Actual Tc: The scatter plot shows a tight clustering of predictions around the diagonal, indicating strong agreement between predicted and actual values. Only a few deviations are observed at the extremes. It is demonstrated in Figure 5.

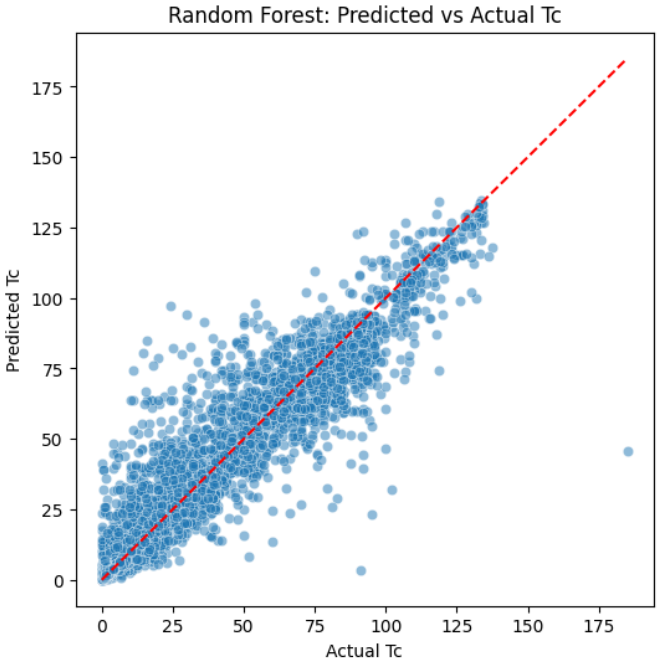

Fig. 5: Random Forest: Predicted vs Actual Tc

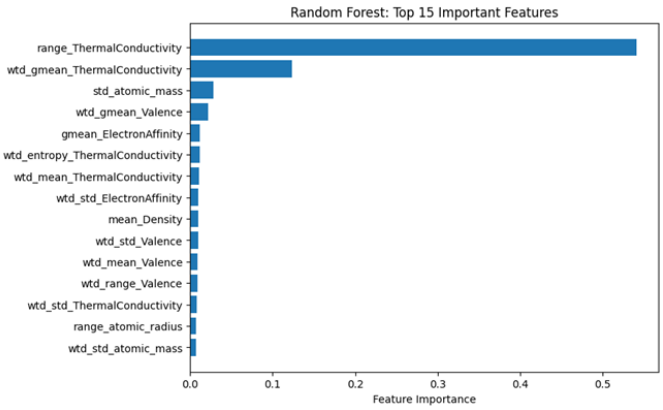

Fig. 6: Feature Importance

Feature Importance: The Random Forest identifies range of thermal conductivity, geometric mean of thermal conduc-

tivity, and standard deviation of atomic mass as the most influential features for predicting Tc.

Random Forest delivered the strongest predictive performance in this study, with an R^2 of 0.921 and the lowest RMSE and MAE. The model not only predicts Tc accurately but also highlights key material properties influencing superconductivity. This makes it a powerful choice for regression tasks on the dataset.

III. Conclusion

This study applied three different regression models: Linear Regression, K-Nearest Neighbors, and Random Forest to predict the critical temperature of superconductors using the dataset. Linear Regression provided a simple and interpretable baseline, explaining about 73% of the variance in Tc but failing to capture nonlinear relationships and underestimating higher Tc values. KNN regression significantly improved predictive performance, achieving an R^2 of 0.905, as it was able to adapt to local patterns in the feature space. Random Forest regression produced the best results overall, with an R^2 of 0.921 and the lowest RMSE and MAE values, demonstrating its strength in modeling nonlinearities and interactions between features. In addition to strong predictive accuracy, Random Forest also offered insights into which features most strongly influenced Tc, such as thermal conductivity and atomic mass [2]. Taken together, these findings indicate that ensemble methods such as Random Forest are particularly effective for this type of structured, tabular dataset, while simpler models remain valuable as benchmarks. The analysis reinforces the importance of using flexible models to capture the complex structure–property relationships that underlie superconductivity.

IV. Appendix

Google Colab Notebook

References

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

[2] G. A. Noman, M. M. Tahmid, M. A. Raihan, and M. S. Hoque, "Artificial intelligence-based perceived motorcycle risk prediction in bangladesh's urban driving environment," in *International Conference on Advances in Civil Infrastructure and Construction Materials*. Springer Nature Switzerland Cham, 2023, pp. 379–388.