Data Analysis of Superconductivity Dataset

Rugilė Pečiulytė, 2025-12-16

Physical meaning of critical temperature for superconductors is their temperature, at which the electrical resistance of a material drops to abruptly zero. Superconductivity dataset, donated to UCI machine learning repository in 2018 (DOI: 10.24432/C53P47), was analysed in this project. The first part of this dataset (train.csv) was selected to analyze 82 extracted physical features of over 21 000 superconductors in predicting of the critical temperature.

To start off the project, exploratory data analysis was performed. One of them, the correlation heatmap showed some of the features having strong correlations with each other, but upon a closer look, it seems that the correlating features were just different statistical descriptors of the same feature (i.e."wtd_mean_atomic_radius" and "gmean_atomic_radius"). Overall, these plots provided a summarized overview of the data, but did not give conclusive results, so they are not included.

Principal component analysis (PCA) employs transformation of highly dimensional data onto a new coordinate system. During this analysis, PCA plot showed that materials with higher critical temperatures occupy a specific region in the reduced feature space (Fig. 1). Even though it could not be said which parameters are most important in predicting critical temperature, it shows that combination of some specific features is prominent.

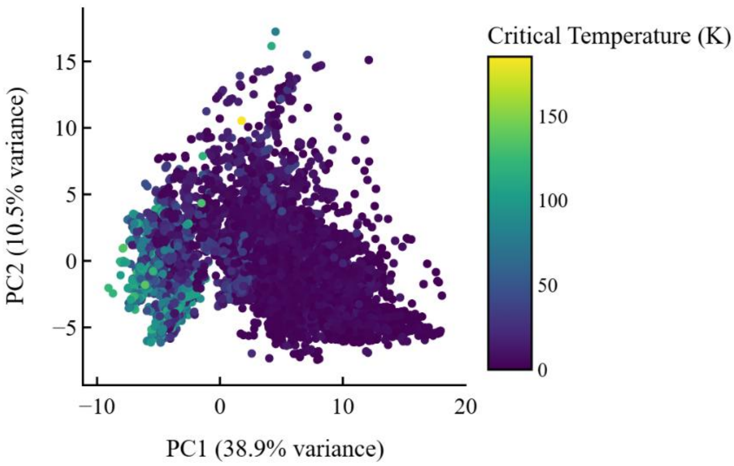

Fig. 1. PCA projection of superconductivity features colored by critical temperature.

A random forest model was trained to predict the target critical temperature. In this case, prediction of critical temperature with an R^2 value of 0.93 and an RMSE of approximately 9 K was made (Fig. 2). The close agreement between predicted and measured values indicates that the physical feature descriptors could be used to predict critical temperature of superconductors with high confidence. Feature importance analysis, extracted from the random forest, reveals which descriptors are the most influential in predicting critical temperature, in particular, the range and weighted geometric mean of thermal conductivity (Fig. 3). This suggests that heat transfer related properties are of most importance when predicting critical temperature.

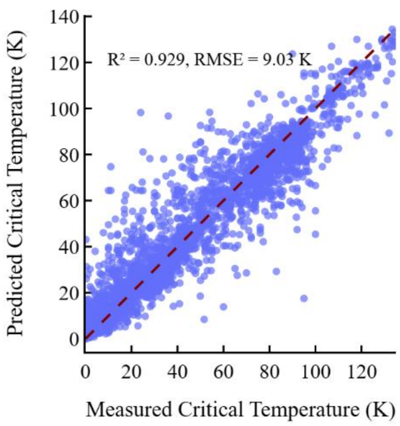

Fig. 2. Predicted versus measured critical temperature using a random forest regressor. The dashed line indicates perfect agreement.

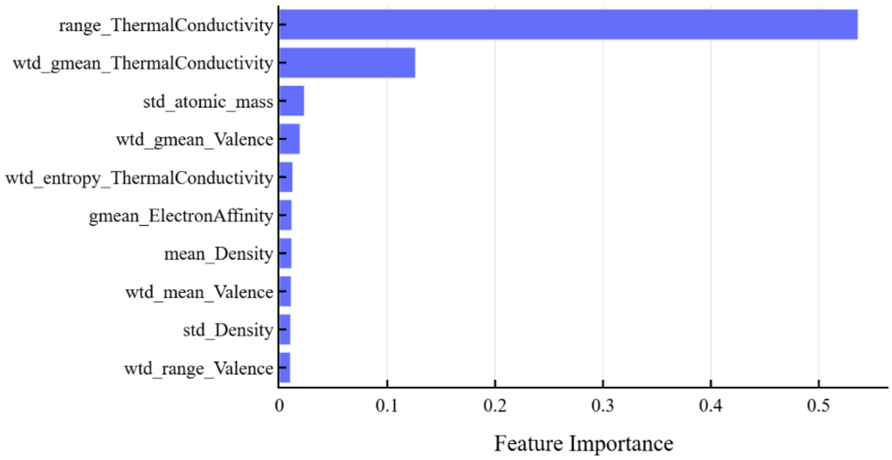

Fig. 3. Top compositional descriptors ranked by feature importance.