A Machine Learning approach to Predict the Critical Temperature of a Superconductor

Aseem Tuli

University of Michigan, Ann Arbor, Industrial & Operations Engineering

Abstract

This project aims at using machine learning algorithms to make predictions of the critical temperature of superconductors. Superconductors are substances that exhibit zero electrical resistance below a particular temperature, called Critical Temperature. Each superconductor has its own critical temperature, and this depends on the superconductor's chemical properties, such as thermal conductivity, valence, atomic mass, radii etc. Superconductors find their use in Magnetic Resonance Imagining (MRI), maglev trains, magnetometers, high sensitivity particle detectors and electric generators among other applications. One of the key challenges faced by scientists and engineers is estimating this critical temperature of a superconductor. A superconductor must be cooled to an extremely cold temperature to activate its zero electrical resistance. Reaching temperature as low as 50-60 K is extremely difficult. There has been a data-driven research in the past to predict the critical temperature of the superconductors. This project aims at using an ensemble of models including Random Forest (RF) and Artificial Neural Network (ANN) for predicting the critical temperature on a pruned dataset using correlation matrix and Principal Component Analysis in R studio, and comparing it with the results published by Hamidieh (2018) [1].

1. Introduction

On April 8, 1911, in Netherlands, Physicist Heike Kamerlingh Onnes of Leiden University was studying the resistance of solid mercury at extremely low or cryogenic temperature, using liquid helium as a refrigerant. He observed that at 4.2 K temperature (-452F, -269C), mercury's resistance suddenly disappeared. The Measuring device did not show any resistance [2]. He published his research in a paper titled "On the sudden rate at which the Resistance of Mercury disappears". He was awarded the Noble Prize in Physics for his discovery in 1913.

Following his research, over the years, many scientists have put forward their discoveries that throw light on the anomalous behavior of superconducting elements and alloys at cryogenic temperatures. These include the Meissner effect which explains strong diamagnetism (repulsion of the magnetic field) of the elements when they are exposed to cryogenic conditions. Others include the Josephson effect which explains the flow of electric current between 2 superconductors even if they are separated by an insulator [2].

A superconductor is any metal, alloy or compound that exhibits a characteristic property (superconductivity) which differentiates it from ordinary metallic conductors. In ordinary metallic conductors, as the temperature is lowered, their resistance decreases gradually and eventually (theoretically) disappears at absolute

zero temperature. However, in superconductors, resistance abruptly drops to zero below a characteristic temperature and this is called the superconductor's Critical Temperature. While absolute zero has not been attained as of now, zero resistance has been observed at these critical temperatures of different superconductors.
In 1941, Niobium-Nitride was discovered to exhibit superconductivity at 16 K. In 1953, Vanadium - Silicon was discovered to show superconducting properties at 17.5 K. In 1962, the first commercial superconducting wire consisting of an alloy of Niobium and Titanium was developed by mid 1960s, high energy particle-accelerator electromagnets had started to be manufactured [2].

Technological Applications of superconductors include Magnetic Resonance Imaging or MRI, which is widely used by health professionals for internal body inspection. The production of magnetometers based on SQUIDs which is an instrument capable of detecting even the weakest magnetic fields involves superconductors [3]. They are used in radio frequency and microwave filters (used in the telecom sector). Electric generators and motors also use superconductors. They find their application in railgun magnets, low-loss power cables, high sensitivity particle detectors such as transition edge sensor, kinetic inductance injector and the superconducting nanowire single-photon detector [3].

Over the years, scientists have made attempts to understand what causes superconductivity in some materials. In 1950, Maxwell and Reynolds et al. discovered that the critical temperature of a superconductor depends on the isotopic mass of the constituent material. BCS theory suggested by American physicists John Bardeen, Leon Cooper and John Schrieffer investigates the impact of electron-phonon interaction and cooper pairs on the superconducting properties of the material. Their theory suggested that entropy and thermal conductivity through lattice interaction also influence superconductivity. [4]

In his research paper titled “A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor”, Kam Hamidieh chooses 8 properties based on Conder (2016) [5] and expert judgement including atomic mass, thermal conductivity and valence to predict the critical temperature of the superconductor. He uses 10 statistical parameters including mean, weighted mean and standard deviation and derives new features to study the effect of the properties on critical temperature and make predictions. A total of 80 (8 properties X 10 parameters) features are used for the analysis. An additional feature i.e number of elements is also added.

Table 1 contains the properties chosen by Hamidieh and table 2 contains the parameters to derive features.

Property	Unit	Description
Atomic Mass	Atomic mass unit (AMU)	Total nucleus rest mass
First Ionization Energy	Kilo-Joules per mol (kJ/mol)	Energy required to remove a valence electron
Atomic Radius	Picometer (pm)	Radius of the atom
Density	Kilogram per cubic meter (kg/m^3)	Density at standard temperature and pressure
Electron Affinity	Kilo-Joules per mol (kJ/mol)	Energy required to add an electron to a neutral atom
Fusion Heat	Kilo-Joules per mol (kJ/mol)	Energy to change from solid to liquid without temperature change
Thermal Conductivity	Watts per meter-Kelvin (W/(m x K)	Ability of a material to conduct heat
Valence	No units	Number of chemical bonds formed by the element

Parameter
Mean
Weighted Mean
Geometric Mean
Weighted Geometric Mean
Entropy
Weighted Entropy
Range
Weighted Range
Standard Deviation
Weighted Standard Deviation

Table 1: Properties chosen by Hamidieh [1] to predict the critical temperature of superconductors

Table 2: Parameters used to analyze the effect of elemental properties and derive new features. Weights are calculated on the basis of number of elements in the material. For instance, Re_7Zr_1 has 7 Rhenium atoms to 1 Ziconium atom, hence Re has a weight of 7/8 and Zr, 1/8 [1].

Hamidieh identifies features based on thermal conductivity, valence, electron affinity, atomic radius and atomic mass as the most important features.

Having tried various statistical models, he settles on two: Multiple regression serving as a benchmark model and Gradient Boosted model which is the main tool for making predictions [1].

He decides to use all the features in making predictions instead of using variable selection or dimensionality reduction methods such as Principal Component Analysis. He argues that these methods don't show any benefits of reducing

dimensionality as a large number of Principal Components is needed to capture a significant percentage of the variability in data.

Hamidieh chooses out of sample Root Mean Square Error or RMSE and R squared values are chosen as performance metrics. The comparison has been made on both the performance metrics. Additionally, Mean absolute error or MAE along with the range of critical temperatures of superconductors have been chosen as performance metrics of the prediction model.

Out of sample Root Mean Square error has been estimated by Hamidieh as the root of the mean of 25 iterations of splitting the data into training and testing dataset, fitting the model on training set, making predictions on the testing set and finally calculating the mean squared error of actual vs predicted critical temperature.

Multiple Regression model's out of sample RMSE: +/-17.6 K

Gradient Boosted model's out of sample RMSE: +/- 9.5 K

RMSE, being a popular error metric has its own limitations. As the error values are squared before their mean is calculated, the RMSE gives a relatively high weight to large errors. It increases as the variation in the frequency distribution of error

magnitudes increases [6]. A model when compared with a given benchmark model may have higher RMSE but lower MAE or a better R^2 value.

Therefore an additional metric MAE along with R^2 and RMSE has been chosen.

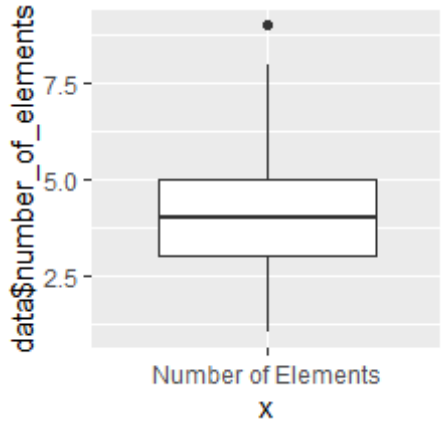
Using more features can sometimes lead to overfitting the data. This means that the trained model, when used to make predictions on unseen data can overfit the data and produce high variance between the actual and observed values.

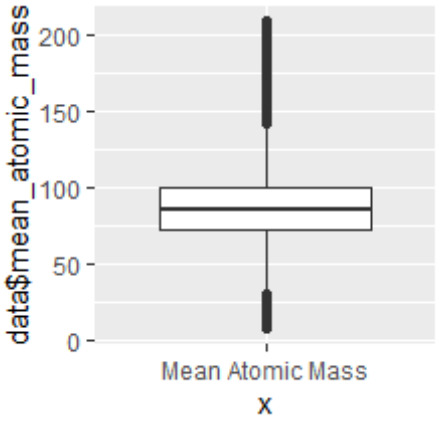
The goal of this project is to use a simplified model, a model that takes a reduced data set as input and makes good predictions. To achieve this, different variable selection methods such as correlation matrix and variable importance graphs and dimensionality reduction methods such as Principal Component Analysis have been used and investigated.

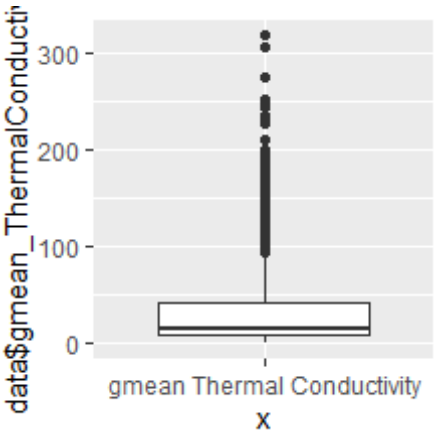
To resolve the dimensionality and accuracy trade off, a +/- 1 K RMSE to 40% reduction in the number of variables has been set as a benchmark. This means that if a model that takes 60% of the total number of features can produce predictions with an RMSE of +/- 10.5 K against Hamidieh's estimate of +/- 9.5 K using all the features, that model will be accepted and considered as a good model.

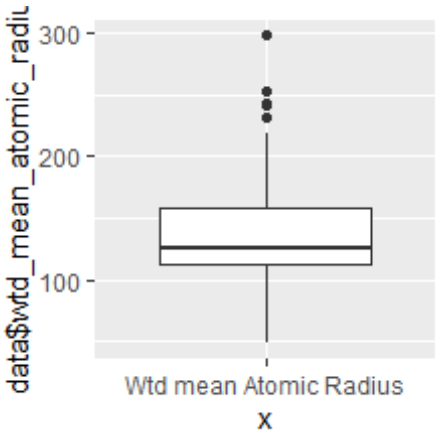
2. Data

The dataset for this project has been obtained from UCI repository https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data. This dataset contains 81 features derived by Hamidieh as explained in the previous section. Some of these features include **number of elements**, **mean atomic mass**, **gmean thermal conductivity, weighted mean atomic radius, entropy density, standard deviation fusion heat (std Fusion Heat)**. Below are the box plots of these features:









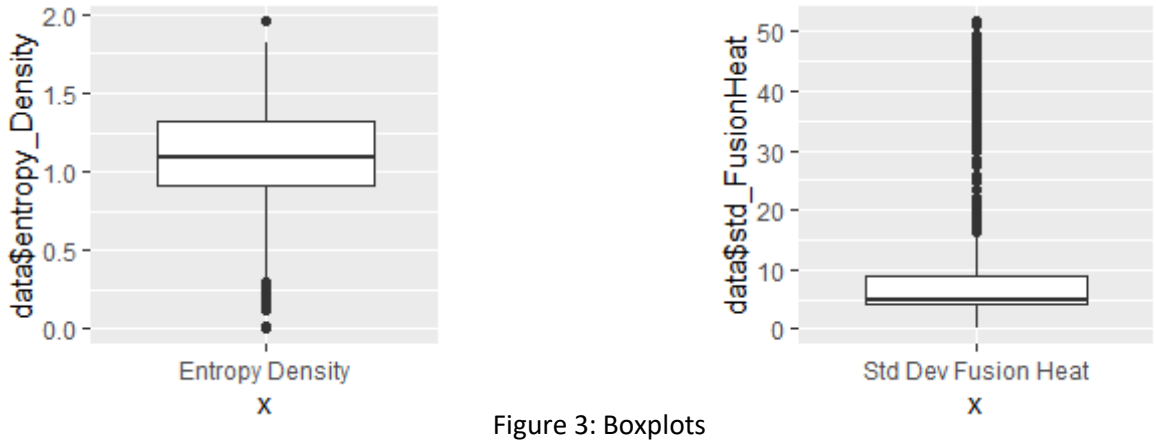


Figure 3: Boxplots

The dataset contains 56% materials that contain Oxygen (O), 50% materials containing Copper (Cu), 32% containing Barium (Ba), 11% containing Iron (Fe) among other elements. Figure 4 shows the proportions of the materials that contain each element:

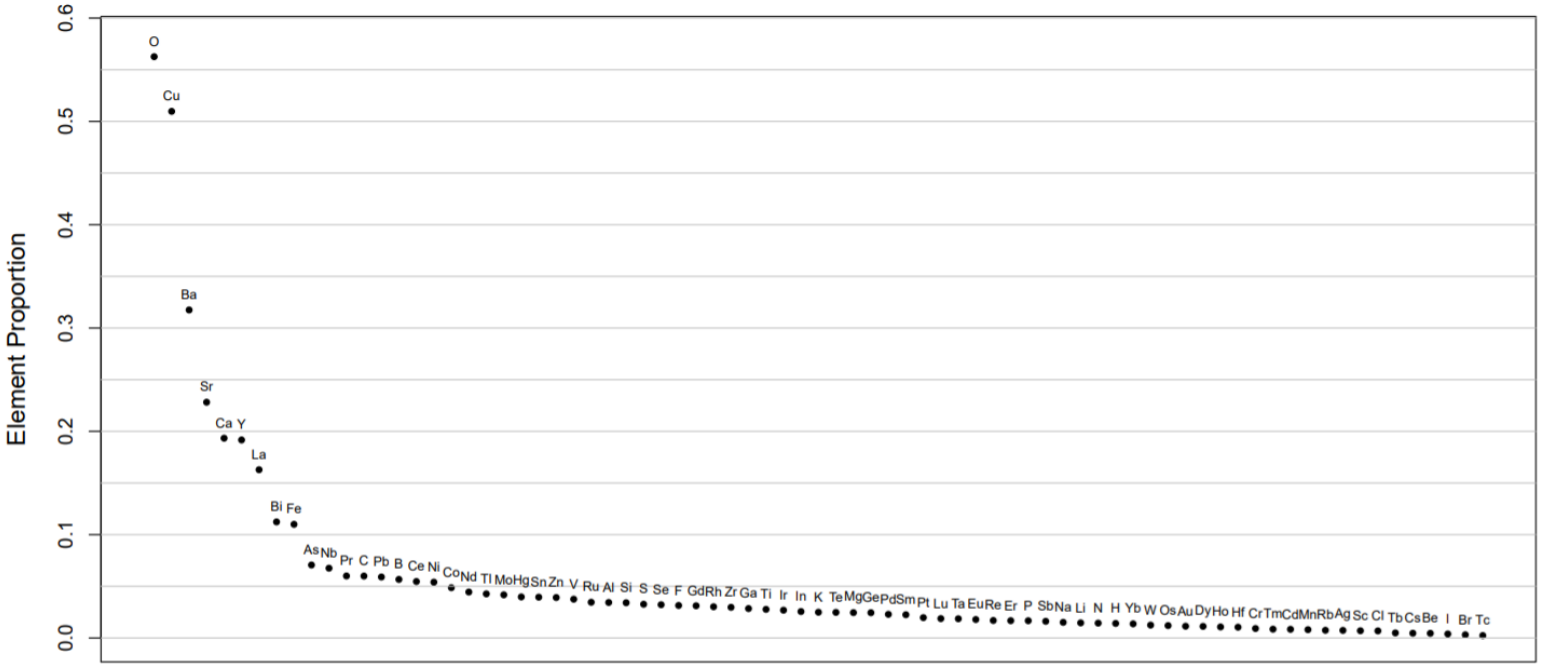


Figure 4: Proportion of materials containing each element

- Iron containing materials have a mean critical temperature of 26.9 +/- 21.4 K
- Materials having copper have a mean critical temperature of 59.9 +/- 31.2 K
- Materials containing Mercury (Hg) have the highest mean critical temperature of 80 K
- Materials containing Silver (Ag) have the highest standard deviation in critical temperature
- It is observed that as the mean critical temperature increases, variability in the critical temperature also increases
- 37% of the superconductors have their critical temperature between 0 and 10 K, 14% have their critical temperature between 10 and 20 K
- There is a small spike in the number of the elements having critical temperature between 80 and 90 K

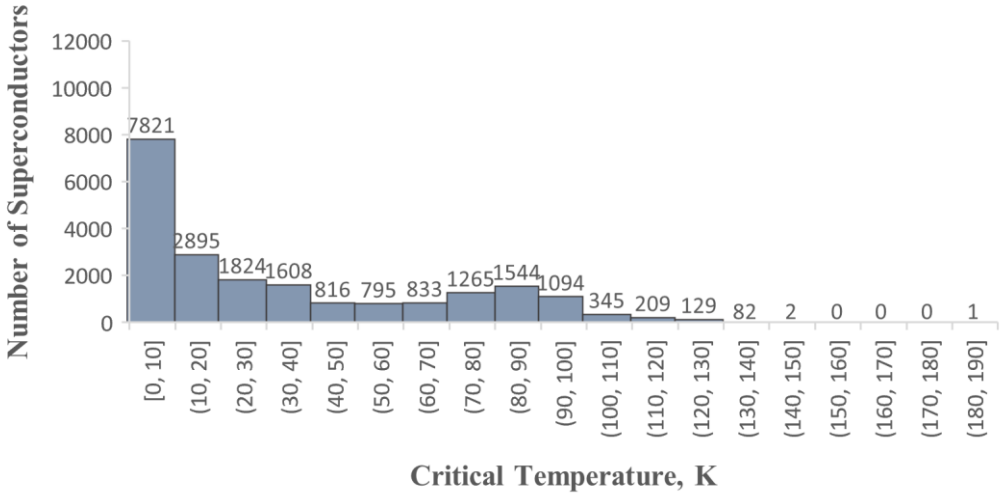


Figure 5: Distribution of Superconductors based on Critical Temperature

3. Model Selection

As of now, there have been two research publications in the world on predicting the critical temperature of superconductors, Valentin et al [7] being the other publication. Since, Valentin et al includes a classification model, therefore results in Hamidieh (2018) become the benchmark.

The following methods and subsequent code have been performed in R studio.

Defining a model that makes accurate predictions is a two-step process that includes:

- Model Selection
- Feature Selection

In terms of model selection, few candidates were selected. These included Generalized Linear Model (GLM), General Additive Model (GAM), Classification and Regression Tree (CART), Random Forest (RF) and Artificial Neural Networks (ANN). These models were trained and tested on a dataset containing 12000 observations and all the features using 20 random holdouts. The idea was to run a screening test on a sample of the original dataset and select or discard the models based on pairwise t-tests and box plot visualizations. Pairwise t-tests are based on hypothesis testing, where:

- Null Hypothesis states that there is no significant difference between error metrics (MAE and MSE) of 2 models in consideration
- Alternative Hypothesis states that there is a significant difference

The hypothesis testing is performed on p-values. If the p-value falls below an alpha value, 0.05 in this case, then null hypothesis is rejected.

If the p-value is more than 0.05 then null hypothesis cannot be rejected.

The R packages used in model selection include:

- Random Forest
- Rpart
- Stats
- Caret
- NeuralNet
- Mgcv
- Stats

Based on the 20 random holdout tests, the p-values of MAE and MSE of the 5 candidate models are compared using pairwise t-tests and the results are obtained in the tables below:

MAE	GLM	GAM	CART	RF
GAM	0.3134	*NA*	*NA*	*NA*
CART	0.0131	0.0389	*NA*	*NA*
RF	0.0006	0.0009	0.00331	*NA*
NN	0.0008	0.0014	0.00418	0.6555135

Table 3: Pairwise T – Test of p-values of MAE, green signifies no statistical difference between GLM, GAM & CART. Yellow signifies no statistical difference between RF & NN

MSE	GLM	GAM	CART	RF
GAM	0.346018	*NA*	*NA*	*NA*
CART	0.025375	0.045116	*NA*	*NA*
RF	0.000039	0.000460	0.003854	*NA*
NN	0.001032	0.000865	0.004060	0.499062

Table 4: Pairwise T – Test of p-values of MSE, green signifies no statistical difference between GLM, GAM & CART. Yellow signifies no statistical difference between RF & NN

The results of pairwise t-test highlight the following:

- We cannot reject the hypothesis that there is no statistical difference in the MAE and MSE values of GLM and GAM
- We cannot reject the hypothesis that there is no statistical difference in the MAE and MSE values of Random Forest and Artificial Neural Network

The objective is to select the model(s) that give the lowest MAE and MSE values, MAE and MSE being the performance metrics

The following figure contains boxplots of these performance metrics for different models:

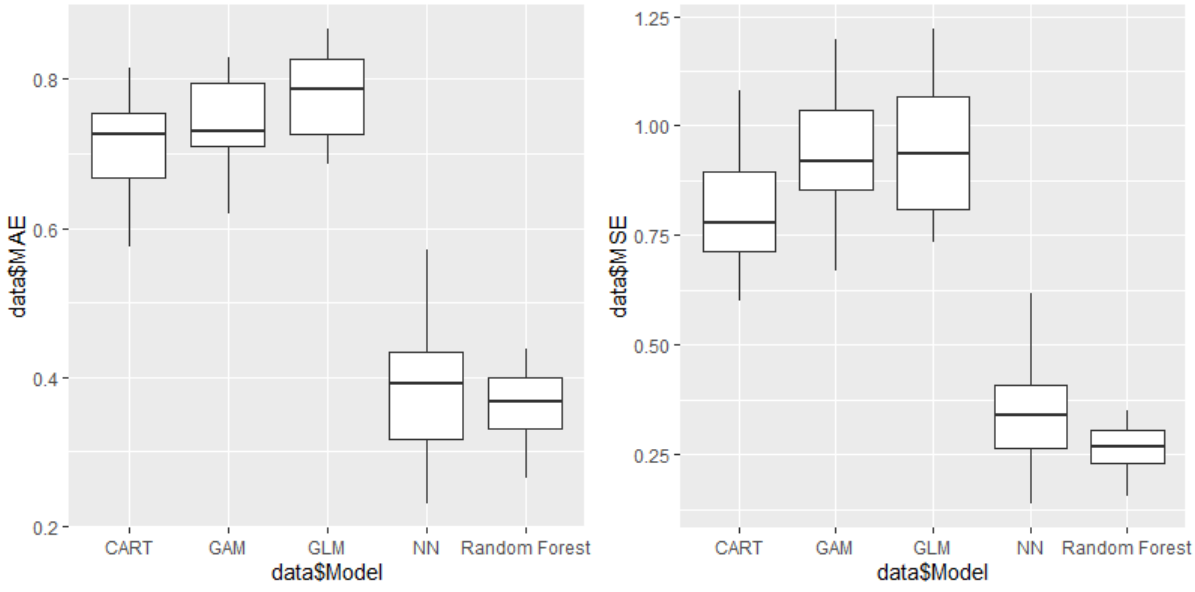


Figure 6: Boxplots of MAE and MSE values of 5 candidate models

- Boxplot analysis shows that Random Forest has the lowest MAE and MSE
- There is no significant difference in the error values of Random Forest and Artificial Neural Networks
- Based on the t-test and boxplots, random forest appears to be the best model
- Since the screening test was performed on 12,000 observations and not the entire dataset, there is a possibility that Artificial Neural Network may make fairly similar predictions if not better
- Random Forest and Artificial Neural Networks have been selected as the final model to make predictions of the critical temperature of superconductors

4. Feature Selection

Feature selection is as important as model selection. The goal is to define a simplified model that makes accurate predictions. As new features are added, the model becomes more complex and even though it may fit the training data really well, there's a high possibility that it will overfit the unseen data. Any anomaly in the unseen or the presence of an outlier will induce high variance in the predictions made by that model.

Reducing extra features makes the model simpler and it becomes less sensitive to any outlier and reduces the variance in the predictions. However, very few features can add a bias in that model. Any model that has a high bias will make predictions that differ from the actual values by a particular degree. If the actual observations can be thought of being represented on a cartesian plane around the origin at (0,0), the bias in the prediction model will change the point of origin and the predictions will be represented around a new origin. The difference in the distance between the original and new point of origin comes from the bias.

A model that has high bias is said to make precise but less accurate predictions. On the contrary, a model that has high variance is said to make more accurate but less precise predictions.

Hamidieh has included all 81features to train the models and make predictions. In this section, 2 different feature selection / dimensionality reduction methods have been used:

- The first method involves a correlation matrix. Using cor() function in R, a correlation matrix of 81 x 81 size is produced. Features that have a correlation of more than +/- 90% are redundant and one of them can be discarded. This is achieved by analyzing the importance of the features. varImp() function of caret package in R has been used to identify important variables. Using this method, the dataset is reduced to 46 predictor features, instead of 81. The following table contains the 20 most important features:

Feature	Importance
wtd_gmean_ThermalConductivity	914.767941
wtd_mean_ThermalConductivity	644.174973
wtd_gmean_Valence	494.444033
wtd_std_ElectronAffinity	468.485126
wtd_std_Valence	304.765695
wtd_gmean_atomic_radius	289.835857
wtd_entropy_ThermalConductivity	286.781652
std_atomic_mass	284.099164
wtd_entropy_Valence	230.742741
wtd_entropy_atomic_mass	227.488621
wtd_std_ThermalConductivity	225.539295
std_Density	221.275243
wtd_range_Valence	187.118432
wtd_mean_FusionHeat	168.659708
wtd_range_atomic_mass	160.585289
wtd_std_atomic_mass	159.579497
gmean_ElectronAffinity	154.912321
wtd_std_Density	150.203574
wtd_std_atomic_radius	130.449018
wtd_range_atomic_radius	117.987447

Thermal Conductivity, Valence, Electron Affinity, Atomic Radii and mass are ranked in the 6 most important features as identified by Hamidieh. Therefore, the list of important features obtained by this method is consistent with the findings of Hamidieh using all 81features

- Principal Component Analysis is a dimensionality reduction method in which the most important features of the variables are captured in principal components based on their eigen vectors and eigen values. Principal Components are ordered in nature. This means that the first principal component explains the maximum variability in the data, followed by the next principal component and so on. These principal components or PC help in reducing the dimensionality of the features by capturing their essence in a reduced dimensional space. For instance, if 90% of the variability in data is explained by 65-70 variables in their feature space, the same variability can be explained by just 10-15 principal components in reduced dimensions. \ This doesn't mean that the actual number of variables will be reduced. Rather the same variability in data will be now explained by 15 transformed variables instead of the original 70. These transformed variables contain features of all those 70 variables based on their relative importance in explaining the variability in data. For instance, if variable number 25 can explain 50% of variation in data alone, the first principal component will

contain maximum features or characteristics of this variable, followed by the next important variable and so on.

PCs are selected based on the desired percentage of variation in the dataset.

A total of 11 PCs have been selected that cumulatively explain 90% of the variation

Thus, there are two models and two feature selection methods for fitting the data and making predictions.

5. Analysis

In this section, out of sample MAE, RMSE and R^2 values of superconductors are estimated. A total of 20,000 observations have been chosen for 30 random holdout testing. The remaining 1263 observations will be used for testing the predictive accuracy of the final model.

Each model goes through the following procedure:

1. **When Correlation is used:**

- Randomly split the data into 80% training set and 20% testing set
- Scale the training data, saving the mean and standard deviation
- Fit the model on training data

- Scale the testing data using mean and standard deviation of the training data
- Predict the critical temperature of the testing data
- Un-scale the predicted critical temperature and actual critical temperature of the testing set
- Estimate the performance metrics (MAE, RMSE and R^2)

  MAE is calculated as mean of absolute difference of predicted and actual value

  RMSE is calculated as root of mean of squared difference of predicted and actual value

  R^2 is calculated as 1 – ratio of sum of squares of the difference of actual and predicted value and sum of squares of the difference of actual and the mean of the actual value
- Repeat the process 30 times
- Take mean of the error metrics

2. When PCA is used:

- Randomly split the data into 80% training set and 20% testing set
- Scale the training data, saving the mean and standard deviation
- Perform principal component analysis on the training data using prcomp function in R
- Save the eigen vectors

- Select the principal components that cumulatively explain 90% variation in the data, based on their eigen values
- Transform the observations of training data with these principal components
- Fit the model using this transformed training data
- Scale the testing data using mean and standard deviation of the training data
- Multiply the testing data set with the saved eigen vectors
- Select the same principal components that are in the training set
- Predict the critical temperature of the transformed data set
- Un-scale the predicted critical temperature and actual critical temperature of the testing set
- Estimate the performance metrics (MAE, RMSE and R^2)

  MAE is calculated as mean of absolute difference of predicted and actual value

  RMSE is calculated as root of mean of squared difference of predicted and actual value

  R^2 is calculated as 1 – ratio of sum of squares of the difference of actual and predicted value and sum of squares of the difference of actual and the mean of the actual value
- Repeat the process 30 times
- Take mean of the error metrics

6. Results

Number of trees in Random Forest was kept at 500, number of variables at each split was chosen to be 6.

After multiple iterations, the learning rate of Neural Network was kept at 0.0001 and threshold at 0.01. The stepmax value of Neural Network was 1e+08. Number of hidden layers was 2. Number of replications was kept at 5. Activation Function was chosen to be sigmoid and algorithm was chosen to be back propagation.

The out of sample MAE, RMSE and R^2 were estimated as the mean value of 30 holdout tests. Table 6 contains the performance metric values for Random Forest across 30 holdouts using correlation and variable importance for feature selection:

Sno	Model	Holdout	MAE	MSE	RMSE	R. squared
1	RF	1	5.59743	96.88645	9.843091	0.916515936
2	RF	2	5.40666	92.28703	9.606614	0.9212524
3	RF	3	5.259349	86.59642	9.30572	0.925011453
4	RF	4	5.320153	83.22473	9.122759	0.930459419
5	RF	5	5.22001	86.01694	9.274532	0.925106005
6	RF	6	5.492969	91.78077	9.580228	0.922220254
7	RF	7	5.348838	88.47095	9.4059	0.924117384
8	RF	8	5.398703	88.7989	9.423317	0.924323723
9	RF	9	5.524645	91.11806	9.545578	0.923230625
10	RF	10	5.195707	86.78586	9.315893	0.926224518
11	RF	11	5.301917	87.4793	9.353037	0.925941042
12	RF	12	5.387969	97.66885	9.882755	0.917242723
13	RF	13	5.130269	85.90439	9.268462	0.9249246
14	RF	14	5.28242	87.16261	9.336092	0.925074206
15	RF	15	5.427087	93.1167	9.649699	0.922443591
16	RF	16	5.292177	89.48116	9.459448	0.925093096
17	RF	17	5.446752	92.17698	9.600884	0.919829574
18	RF	18	5.259094	84.14787	9.173215	0.926737725
19	RF	19	5.211879	85.18201	9.22941	0.930247426
20	RF	20	5.247545	89.1551	9.442198	0.924089436
21	RF	21	5.2189	83.90097	9.159747	0.929185421
22	RF	22	5.329409	92.46139	9.615685	0.919998478
23	RF	23	5.365	86.65294	9.308756	0.926015355
24	RF	24	5.160972	83.31332	9.127613	0.926723646

25	RF	25	5.391839	90.90016	9.534157	0.922554502
26	RF	26	5.299694	90.07633	9.490855	0.92326022
27	RF	27	5.363642	90.88076	9.53314	0.920812226
28	RF	28	5.358638	90.35311	9.505425	0.922197332
29	RF	29	5.488603	92.24428	9.604388	0.921005179
30	RF	30	5.542455	93.18219	9.653092	0.920743945

Table 6: MAE, RMSE and R^2 values for Random Forest using correlation for feature selection

Table 7 contains the performance metric values for Artificial Neural Network across 30 holdouts using correlation and variable importance for feature selection:

Sno	Model	Holdout	MAE	MSE	RMSE	R. squared
1	NN	1	5.685443	106.9919	10.34369	0.89681358
2	NN	2	6.010728	122.4686	11.06655	0.886450676
3	NN	3	6.925476	97.04146	9.850962	0.875083925
4	NN	4	5.505512	116.8831	10.81125	0.899723372
5	NN	5	7.33363	92.97096	9.642145	0.884907221
6	NN	6	6.581546	104.2135	10.2085	0.917343744
7	NN	7	7.69923	114.0078	10.67744	0.881685545
8	NN	8	6.723709	99.25503	9.962682	0.920512539
9	NN	9	5.85476	113.1336	10.63643	0.919439227
10	NN	10	6.599412	121.767	11.03481	0.924185001
11	NN	11	6.811309	120.6429	10.98376	0.867644741
12	NN	12	6.273645	127.1879	11.27776	0.86394554
13	NN	13	6.826197	114.1394	10.6836	0.887124121
14	NN	14	6.524118	97.62033	9.8803	0.917433653
15	NN	15	5.74449	104.3184	10.21364	0.892615566
16	NN	16	7.720027	110.0366	10.48983	0.87515898
17	NN	17	6.099562	103.0706	10.15237	0.906038788
18	NN	18	6.018527	109.9189	10.48422	0.884422279
19	NN	19	7.784447	100.031	10.00155	0.897291666
20	NN	20	5.549013	107.6079	10.37342	0.884389455
21	NN	21	5.974556	97.38617	9.868443	0.913629841
22	NN	22	6.013318	99.83166	9.991579	0.867125613
23	NN	23	6.193592	110.6369	10.51841	0.890108006
24	NN	24	6.849484	98.3759	9.918463	0.925892319
25	NN	25	6.068127	129.8462	11.39501	0.903135006
26	NN	26	6.170878	103.8067	10.18856	0.898369866
27	NN	27	6.022862	102.9797	10.14789	0.915337436
28	NN	28	5.976427	102.4392	10.12122	0.914313549
29	NN	29	5.820453	105.782	10.28504	0.870512313
30	NN	30	6.051686	132.9651	11.53105	0.883231883

Table 7: MAE, RMSE and R^2 values for Neural Network using correlation for feature selection

The following table summarizes the mean of 30 MAE, RMSE and R^2 values for Random Forest and Artificial Neural Network using correlation as shown above:

	Random Forest	Artificial Neural Network
MAE	5.3423	6.3804
MSE	89.25	108.911
RMSE	9.445	10.425
R^2	0.923	0.895

Table 8: Mean of 30 MAE, MSE, RMSE & R^2 for RF and NN using correlation

Similarly, the following table summarizes the mean of 30 MAE, RMSE and R^2 values for Random Forest and Artificial Neural Network using PCA

	Random Forest	Artificial Neural Network
MAE	7.192	8.564
MSE	108.25	127.691
RMSE	10.404	11.300
R^2	0.837	0.782

Table 9: Mean of 30 MAE, MSE, RMSE & R^2 for RF and NN using PCA

7. Discussion

Based on out of sample error and R^2 values, Random Forest on a reduced dataset using correlation and variable importance outperforms the Gradient Boosting Method used by Hamidieh in his research. Artificial Neural Network is also accepted as a good model for prediction as the RMSE values are fairly similar at the expense of 40% reduction in the number of features. It is well within the threshold limits mentioned earlier in this paper. Correlation and variable

importance are preferred over Principal Component Analysis as both the models perform better using the former method of feature selection.

We have observed that reducing the features has a positive impact on the accuracy of the model. Not only does it increase the predictive accuracy of the model, but it also simplifies it and makes it less sensitive to any anomaly in the unseen data.

Although both Neural Network and Random Forest are completely different, the out of sample results suggest that their predictions of critical temperature on the testing data can be comparable.

An ensemble of random forest and artificial neural network is used for final prediction. Here the critical temperature of 1263 superconductors are predicted, and the accuracy of the models is tested by comparing its value with the observed values.

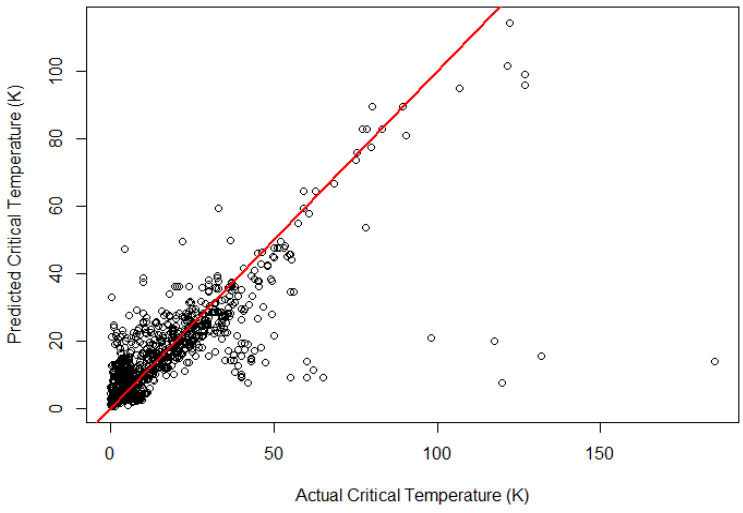


Figure 7: Actual vs Predicted Critical Temperature of 1263 superconductors using an ensemble of NN and RF

The plot above shows actual vs predicted critical temperature of 1263 superconductors using an ensemble of neural network and random forest on the reduced feature space. The following table summarizes the error statistics of the ensemble.

Metric	Value
MAE	5.207
MSE	96.578
RMSE	9.827
R^2	0.906

Table 10: Error statistics of the ensemble of RF and NN

The predictive accuracy is comparable to the out of sample error values. This indicates that the model ensemble is accurate and reliable for future predictions.

8. Conclusion

Correlation and Variable Importance have been effective feature selection methods in improving the prediction of critical temperature of the superconductors. Random forest is found to be performing better than Gradient Boosting model on this dataset. An ensemble of random forest and neural network can be used to make future predictions. Therefore, it can be concluded that machine learning approach using elemental properties of superconducting materials can effectively predict the critical temperature.

9. Acknowledgment

I would like to thank Dr Seth Guikema, Professor at the Department of Industrial and Operations Engineering, University of Michigan for his teachings and his mentorship.

10. References

[1] https://www.semanticscholar.org/paper/A-Data-Driven-Statistical-Model-for-Predicting-the-Hamidieh/b3bea0ac481f0869cb746f3b44d5689bf1a9b924

[2] http://www.superconductors.org/History.htm

[3] http://www.superconductors.org/History.htm

[4]https://qudev.phys.ethz.ch/static/content/courses/phys4/studentspresentations/supercond/Ford_The_rise_of_SC_6_7.pdf

[5] https://journals.aps.org/prb/abstract/10.1103/PhysRevB.93.220504

[6] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[7] https://arxiv. org/abs/1709.02727