Data science - final project

David Peleg - 313237182

Aviv Arcobi - 209008408

Introduction

In this final project we are going to use a data driven statistical model for predicting the critical temperature of superconductors. The prediction would be based on the features extracted from the superconductor chemical formula. In this project we followed the original paperwork on this subject by Kam Hamidieh.

A superconductor is a material that conducts current with zero resistance and magnetic flux fields are expelled from it. Any superconductor has a characteristic critical temperature below which the resistance drops abruptly to zero and the material becomes a superconductor. Superconductors have many applications. They are used in MRI, SQUID, mass spectrometers and many more.

For many years, simple empirical rules based on experimental results have guided researchers in synthesizing superconducting materials. Here we take a data driven approach to create a statistical model for predicting the critical temperature based on the chemical formula.

Know thy data

There are two files. The first one called “train” contains 81 features extracted from 21263 superconductors along with the critical temperature in the 82nd column. The second one called “unique” contains the chemical formula broken up for all the 21263 superconductors. The numbers in each row represents the chemical proportion of atoms that constitute the superconductor. The last two columns have the critical temperature and chemical formula. In the unique file there are only the first 86 elements, so any superconductor containing an element with an atomic number greater than 86 is excluded.

The 81 features in the train file were extracted in the following way. There are 8 variables/properties which were found to contribute the most for the model prediction. The variables are: thermal conductivity, atomic radius, valence, electron affinity, atomic mass, density, fusion heat and first ionization energy. For each variable 10 features were extracted. The features are: mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, weighted standard deviation. The 81st feature is a numeric variable counting the number of elements in the superconductor.

Basic summaries of the data

In this section we are going to display some figures describing the basic properties of the data.Note: the numbers of the figures correspond to the number of the original figures in the original article.

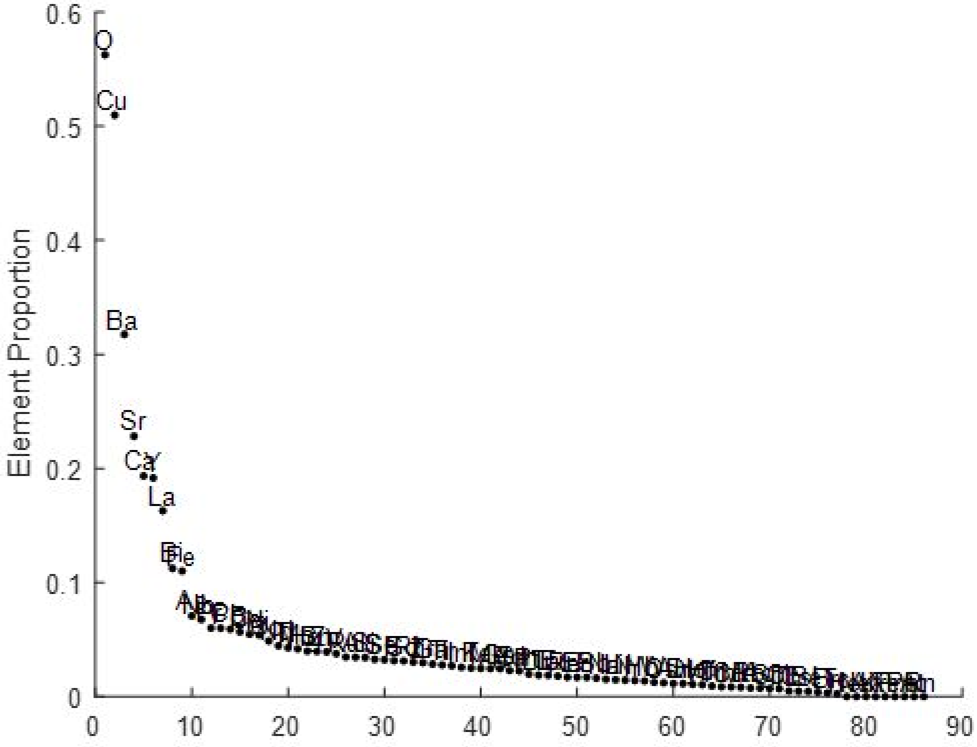


Fig.3 shows the proportions of the superconductors that had each element. For example, oxygen is present in about 56% of the superconductors. Copper, barium, strontium and calcium are the next most abundant elements.

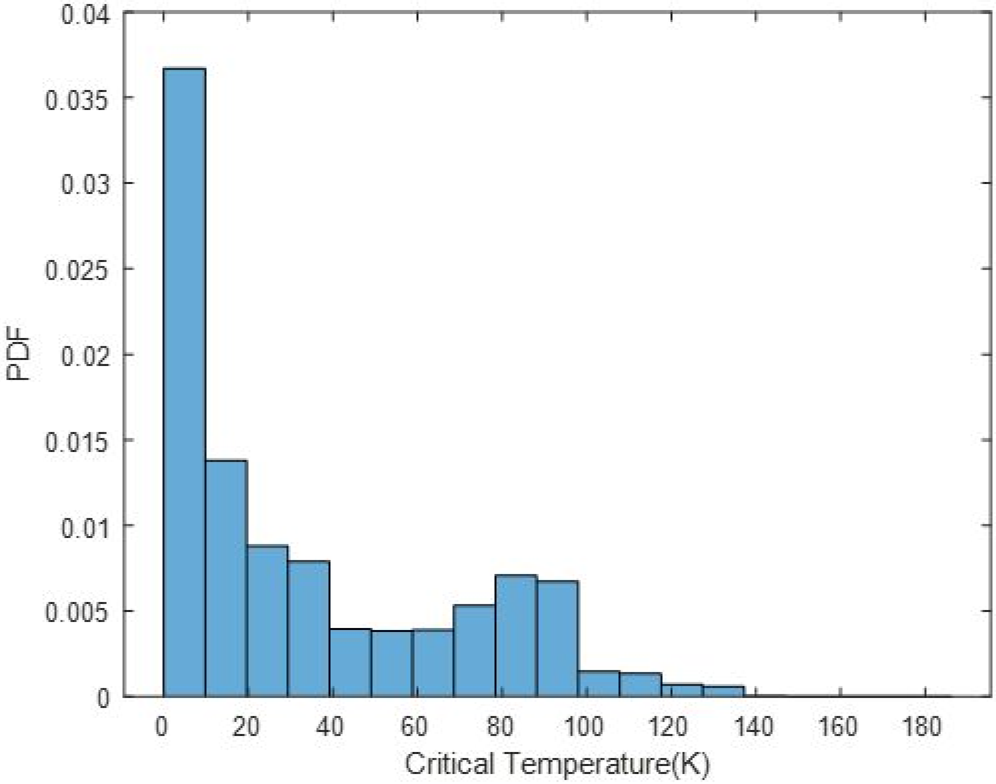


Fig.4 shows the distribution of the superconducting critical temperatures (K) of all 21263 superconductors.

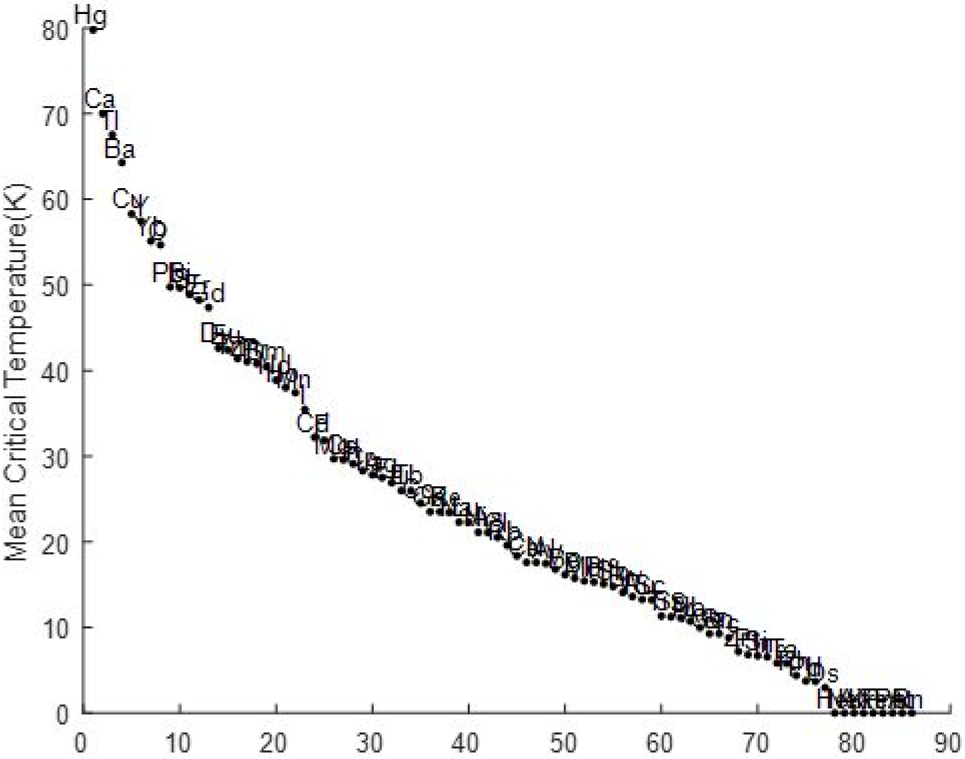


Fig.5 shows the mean superconducting critical temperature grouped by elements. On average, mercury containing materials had the highest superconducting critical temperature followed by calcium and so on.

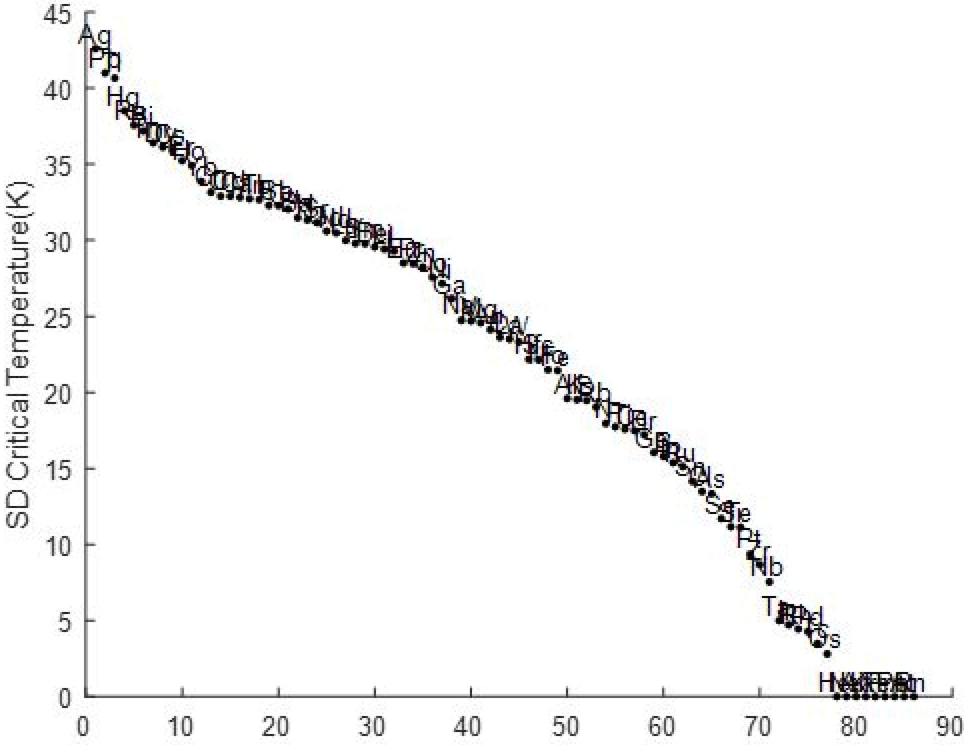


Fig.6 shows the standard deviation (SD) of critical temperature grouped by elements. Silver containing materials had the highest variability followed by lead and so on.

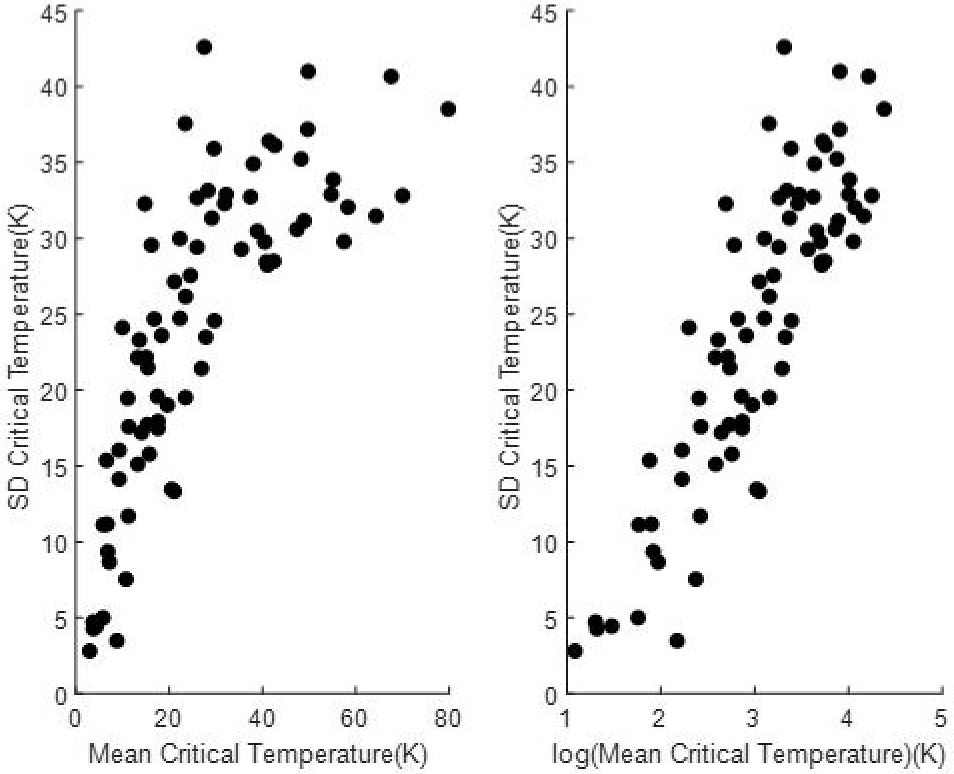
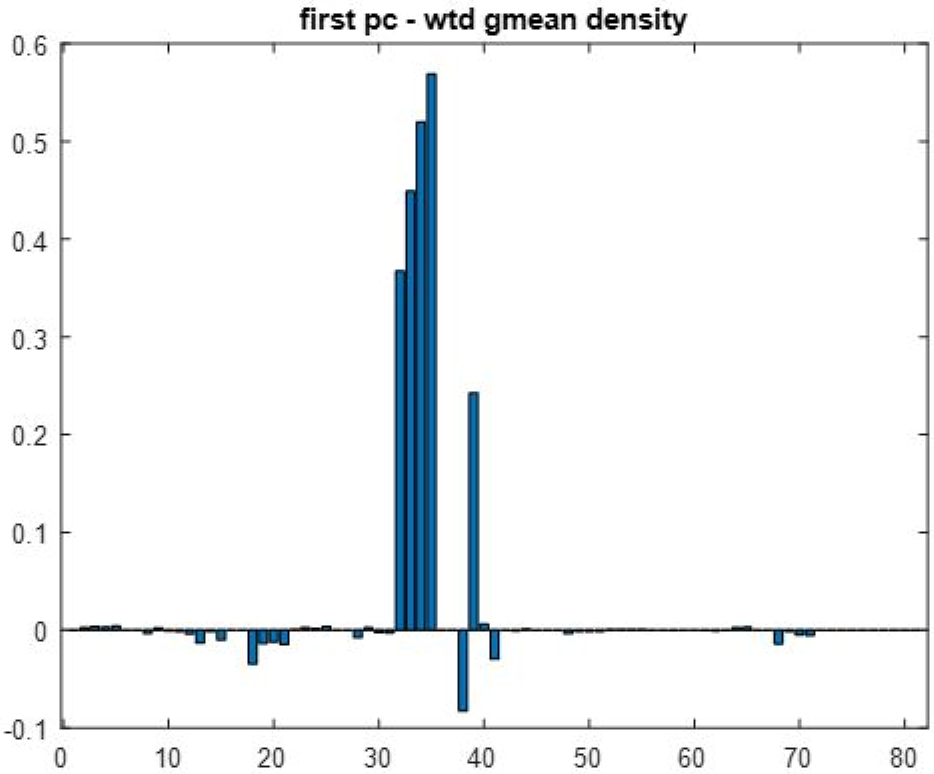


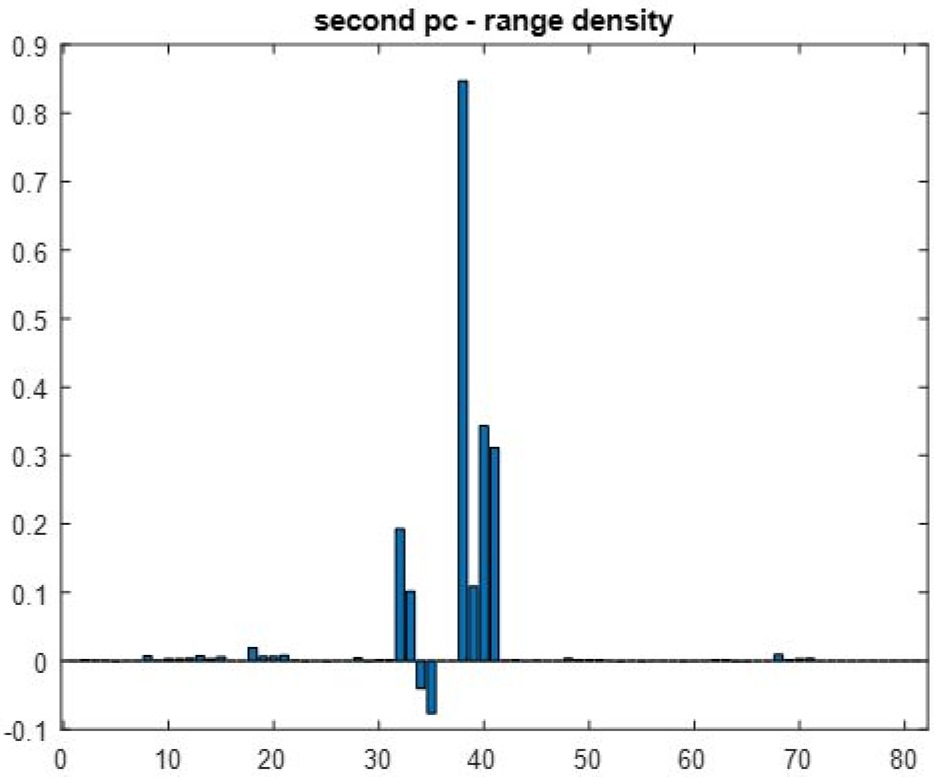
Fig.7 shows the relationship between the mean critical temperature and standard deviation (SD) per element. The second figure shows the logarithm of the mean critical temperature versus SD. On average the higher the mean critical temperature, the higher the variability in critical temperature per element.

PCA analysis

After reviewing the basic properties of the data we calculated the average absolute value of the correlation among the features and found it to be 0.35. This indicates that the features are highly correlated. So, we decided to use PCA analysis in order to reduce the dimensionality of our data/features.We took the first 40 most important features. We will apply this new data later on and see if there is an improvement in the model’s prediction.

We used the matlab function pca.m to perform a pca analysis of the data. Then we plot a bar plots of the coefficients for the first two principal components, in order to see what are the features that contribute the most for each of the principal components.


first pc - wtd gmean density
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0
10
20
30
40
50
60
70
80


second pc - range density
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
0
10
20
30
40
50
60
70
80

We displayed a list of the top 10 most important features. Note: the most important features we've got using pca analysis were different from the most important features the article got using XGBoost gain.

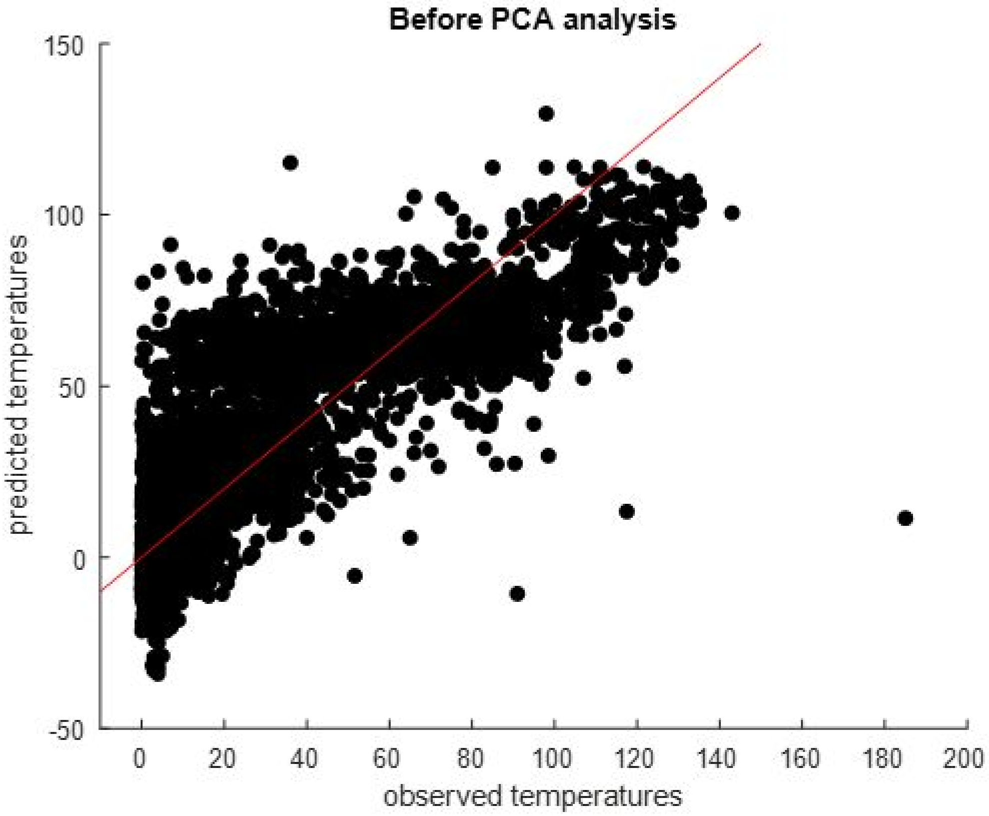
best	2	3	4	5	6	7	8	9	10
wtd_gmean_Density	range_Density	'wtd_range_Density'	'wtd_range_Density'	'mean_Density'	'wtd_std_Density'	'std_Density'	'wtd_gmean_Density'	'range_fie'	'range_ThermalConductivity'

The statistical models

For the next part, we tried to find a machine learning solution to predict the critical temperature of the material based on the features presented earlier. For measuring the accuracy of the model we used 3 methods:

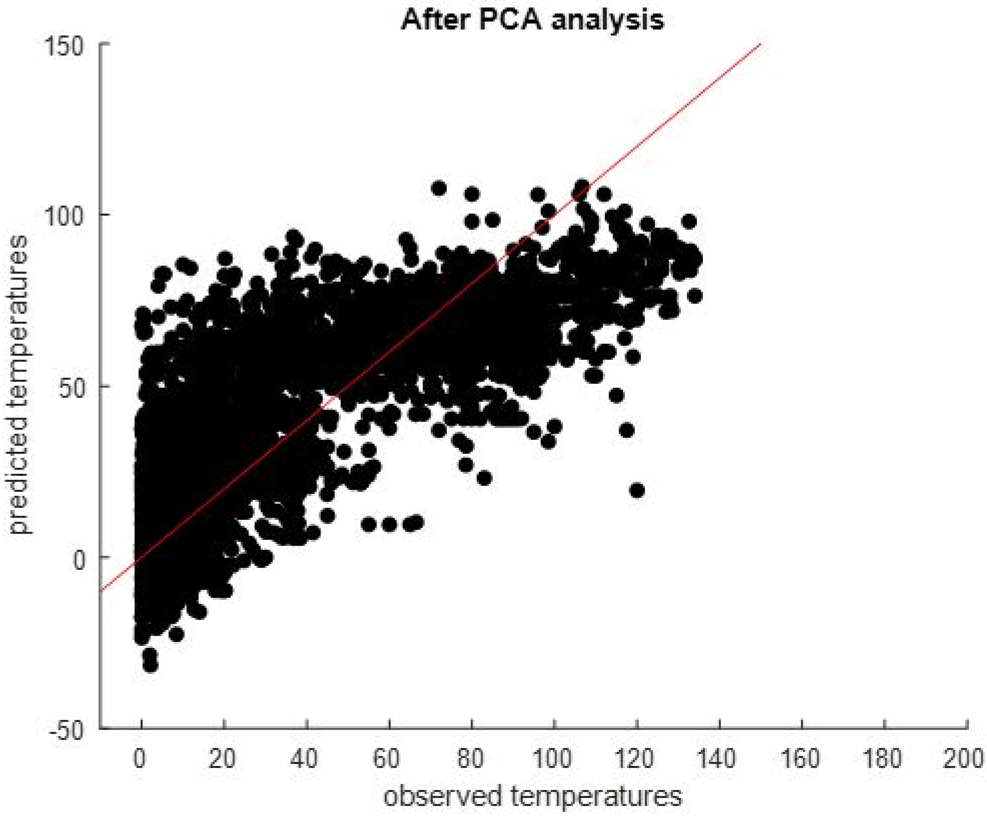
1. Rmse
2. R^2 test
3. Calculating precision by treating the problem as a classification problem and defining every result that is within 7K from the observed temperature as a correct prediction.

Our first approach was using a simple linear regression. For that we used matlab's fitlm.m function on a split data set with 30% test and 70% train. From looking at a scatter plot of the predicted results as a function of the observed ones, one might think that the classifier did a pretty good job.



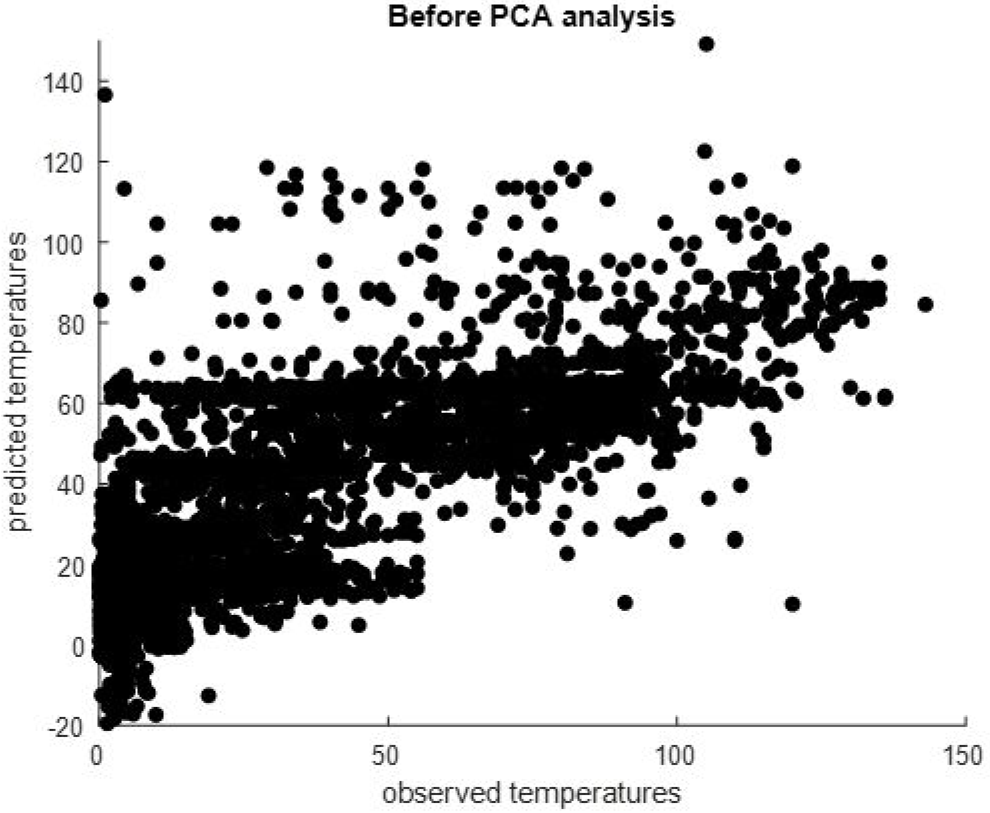
However, the results were: RMSE = 18.56K, R^2 = 0.723, precision = 40% suggesting that the data is not linearly separable.

We also applied the reduced data we have using PCA.

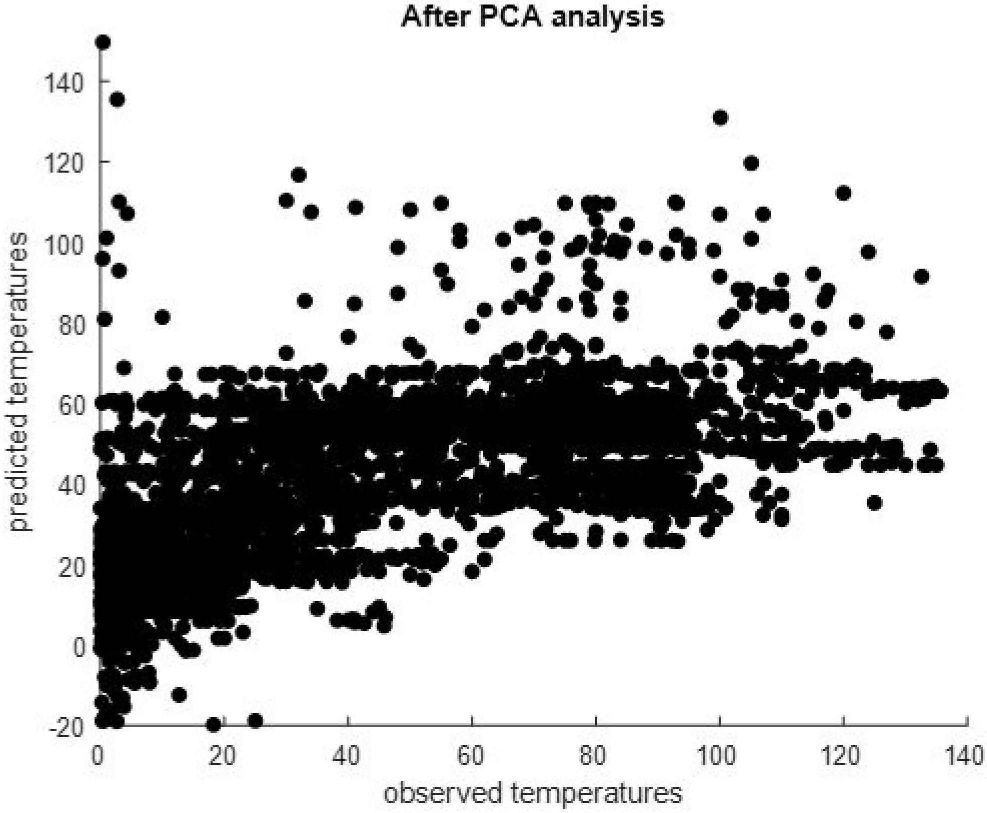


We only got a slight deviation from the original data set indicating that pca analysis was indeed useful. with: RMSE = 19.36K, R^2 = 0.68, precision = 30%

Results were even worse when trying to create a classifier from the chemical formula of the materials. RMSE = 22.03K, R^2 = 0.615, precision = 30%.



We also applied the reduced data we have using PCA.



We only got a slight deviation from the original data set indicating that pca analysis was indeed useful. with: RMSE = 25.97K, R^2 = 0.42, precision = 20%

Our second approach was to create a neural network classifier. For the analysis based on material features we used matlab's feed forward network, which has a simple linear decision function at its final layer. The first layer had 30 nodes and a ReLU transfer function while the second layer had 20 nodes and a softmax transfer function. We noticed that the loss function chosen to train the model greatly affected the results we got. We started with SGD with momentum, but the loss function quickly diverged. To try and even out the divergence we tried using momentum and variable learning rate, but it ended up with approximately the same results we got from the linear regression model.

After a few more attempts we settled for the Bayesian Regularization loss function. We once again split the data into 70% training and 30% test and we noticed that at around ~30 epochs we had difference between the training and testing performance which suggests overfitting. At the end, our neural network classifier yielded much better results than the linear classifier with RMSE = 12K, R^2 = 0.89, precision = 0.64. for reference, the state of the art results achieved in the article were an RMSE of 9.5K.

We then tried to filter out redundant features from the data set examples using pca analysis. the results were fairly close to the ones we got from the original data set with: RMSE = 14K, R^2 = 0.8319, precision = 0.57. these results further validate our assumption that the data is highly correlated.

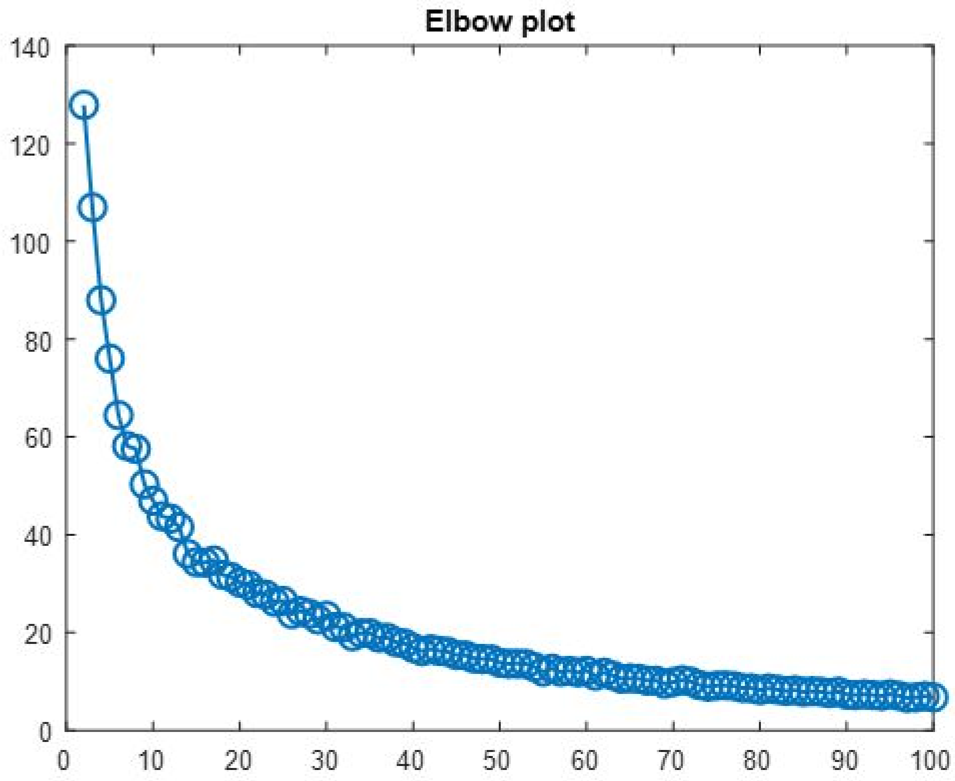
For the analysis based on chemical formula we used a similar network architecture with the resilient gradient loss function and noticed overfitting at ~2600 epochs.

Here we once again received much better results than the ones we got from the linear classifier. However, our classifier was still not accurate enough on the test set to be used for prediction. We got RMSE = 20K, R^2 = 0.6, precision = 42%.

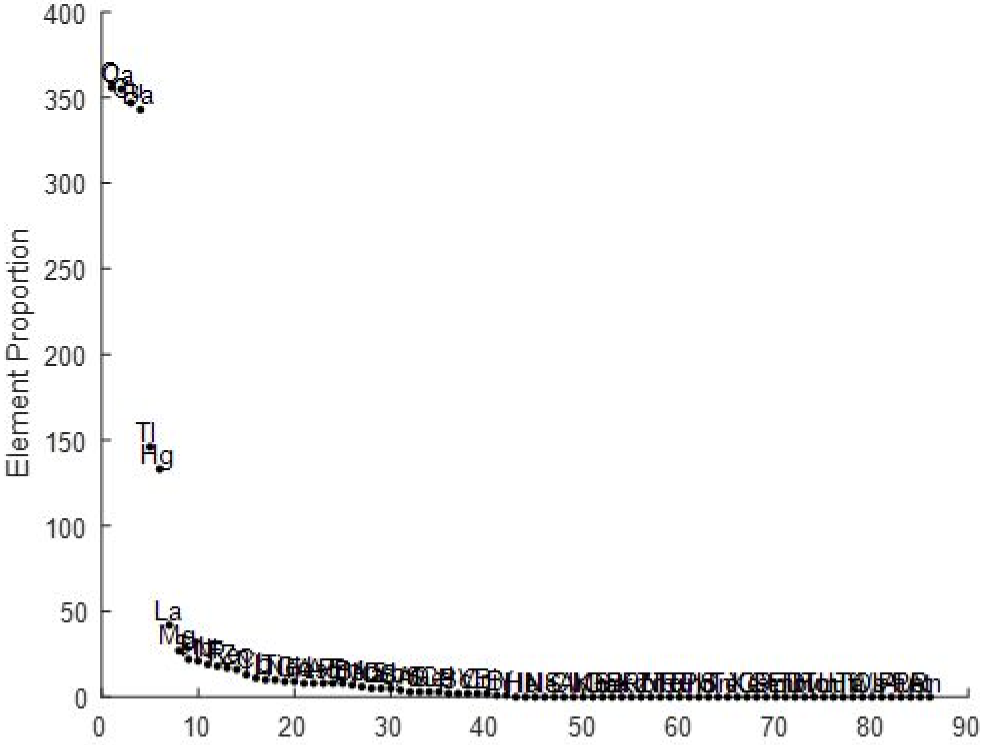
We then once again tried to filter out redundant features using pca analysis. The results we got this time were far worse with: RMSE = 38K, R^2 = -1.05, precision = 0.37. We think that the reason for the major difference in result is that removing features means changing the entire chemical formula of the material rather than ignoring some of its properties like we did in the previous case.

Clustering

As a final remark, we wanted to see if there are certain elements that causes the critical temperature to increase. This information can be helpful if one wants to synthesize new superconductor with high critical temperature. For that we cluster the superconductors into groups. First, in order to find the optimal number of clusters we use the elbow method. We found the optimal number of clusters to be approximately 65.



Then, we checked which cluster has the highest critical temperature. (we made sure that it’s passes some threshold). Finally, we examined which elements are the most abundant in that cluster. We found that oxygen, calcium, copper and barium are the most common ones.



We can compare this to the results of the most abundant element on fig.3. This way we can find an element that is used in superconductors but shouldn’t be. For example, the element SR is the fourth most abundant element in the overall data set (fig. 3) and is not in the top 10 most abundant elements of the high temperature group. indicating it might not be wise to use it for superconductor creation

Summary and conclusions

In this project we first represented some basic properties of our data. We also were able to predict pretty well the critical temperature of a superconductor based on its chemical formula. We made the prediction using two models: linear regression model and a neural network model. We compared these results before and after applying a PCA analysis. And as a final step, we saw which elements are responsible for a high critical temperature using k means approach. This can be helpful for researchers interested in finding high temperature superconductors.

Bibliography

A data-driven statistical model for predicting the critical temperature of a superconductor - Kam Hamidieh - https://www.sciencedirect.com/science/article/pii/S0927025618304877