Relationships Between Critical Temperature Of Superconductors And Their Physical Properties

Zeng Fung Liew (lzfliew@ucdavis.edu)

June 4, 2020

Abstract

The main goal of this analysis is to find the relationships and associations between the critical temperature of superconductors and their physical properties and chemical components. To achieve this goal, we first understand the distributions of all attributes and the general outlook of the data set using histograms and scatter plots. We then move on to analyzing the dual variable associations between all physical properties using heat maps. After obtaining a good idea on how the attributes associate with each other, we produce visualizations of the distribution of superconductors' critical temperature based on its most highly associative physical properties to gain a better understanding of the classification of high and low critical temperatures in superconductors. Lastly, we figure out the best classification method to classify critical temperatures by trying out various machine learning techniques.

1 Introduction

The superconductor data set comes in two files, one of which describes each superconductors weighted and non-weighted physical properties such as atomic mass, Fie, and atomic radius, while the other describes the chemical elements in each superconductor. We wish to understand the distribution of each physical property of superconductors and how they associate with one another. It is also important to then understand how the critical temperatures of superconductors are associated with its respective chemical composition. Making sense of these distributions and associations is critical for us when figuring out the type of superconductors to use in building certain gadgets and understanding the pros and cons that come with them.

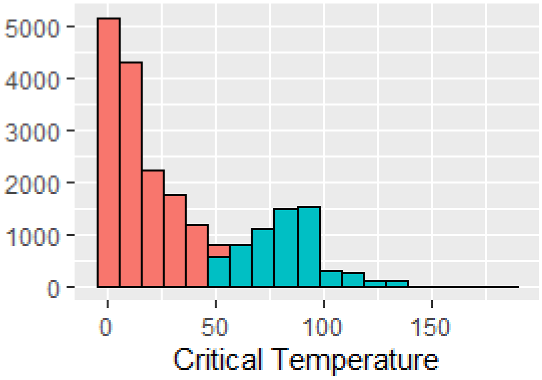


Figure 1: Distribution of Critical Temperature (Red: Low Critical Temperature, Blue: High Critical Temperature)

The next step in analyzing this data is to classify the superconductors based on their respective critical temperature. We will utilize various supervised machine learning techniques to classify the critical temperature of superconductors to observe the difference in results and figure out the best classification method for this data set.

2 Distribution of all Physical Properties of Superconductors

Through our primary analysis, we find that there is an outlier in the critical temperature among superconductors. That superconductor is known to be H_2S, and it has a critical temperature that is at least 43K more than any other superconductors in this data set.

Our next goal is to understand the distributions of all the physical properties of superconductors through histograms. The summary of the distributions is portrayed in Figures 1 to 4.

From the histogram of critical temperature in Figure 1, we find that the there are two main peaks, with the first peak on the far left being significantly higher than the second. This suggests that we can split the critical temperatures into two categories (low and high). For the rest of this data analysis, we will focus on the understanding of each attributes and how it associates with the classification of low and high critical temperatures.

Most of the variables in the data set do not have a distribution that is similar to that of critical temperature, hence their relationships with critical temperature is not as simple as being linear. Some of the variables with the closest association with critical temperature is number of elements. In Figure 2, we can see that number of elements is distributed similar to a normal distribution that is slightly skewed to the left. As we can see, superconductors with high critical temperature tend to fall on the right side of the distribution, which means that superconductors with high critical temperature have a tendency to compose of more elements. This suggests that there is a good association between the number of elements and the critical temperature in superconductors.

While variables like number of elements might provide a "nice" distribution that displays some association with critical temperature, it is not the case for many other variables. A majority of variables are distributed to be similar to a skewed bell curve, as shown in Figure 3.

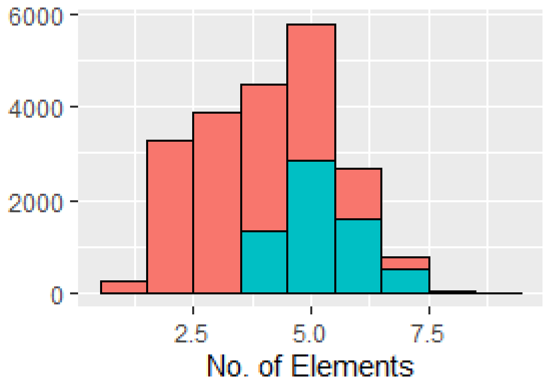
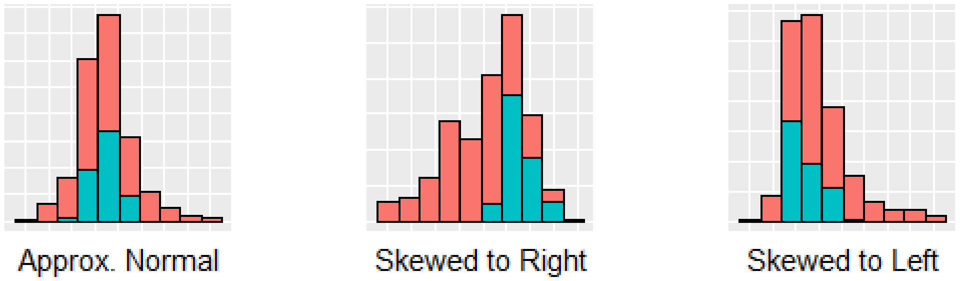


Figure 2: Distribution of Number of Elements (Red: Low Critical Temperature, Blue: High Critical Temperature)



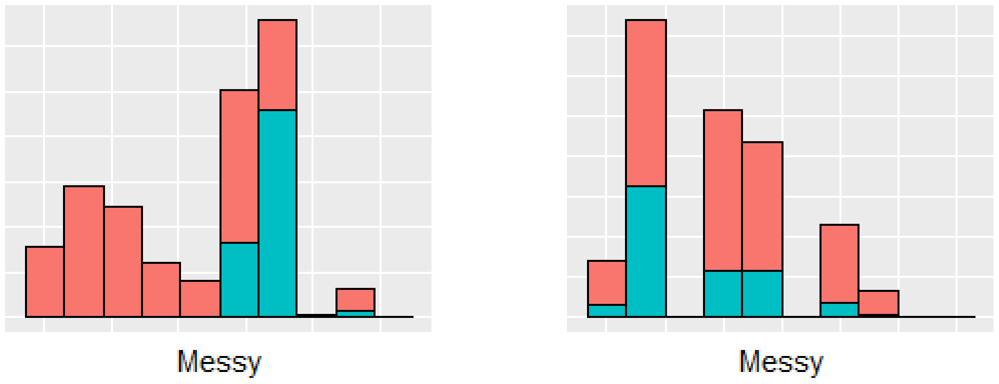
(a) Some measurements of certain properties give out distributions that are approximately normal

(b) Some measurements of certain properties give out distributions that resembles a normal distribution that's skewed to the left

(c) Some measurements of certain properties give out distributions that resembles a normal distribution that's skewed to the right

Figure 3: Distributions of the majority of measurements and properties of superconductors (Red: Low Critical Temperature, Blue: High Critical Temperature)

Additionally, there is also a small number of variables, such as Range Fie and Range Valence, with very messy distributions, as shown in Figure 4. These distributions do not give us much information about how they are related to critical temperature. Therefore, we need to go a step further and analyze how a set of variables combined together can produce a better explanation on the association with critical temperature.



(a) Some measurements of certain properties produce distributions with multiple peaks

(b) Some measurements of certain properties produce distributions with random and awkward gaps

Figure 4: Distributions of certain measurements and properties of superconductors (Red: Low Critical Temperature, Blue: High Critical Temperature)

3 Relationships between Critical Temperature and Physical Properties of Superconductors

After understanding the distributions of the physical properties of superconductors, we want to find out the how the properties associate with each other. For the rest of the section, we will refer the variables by properties and measurements due to the large number of attributes in the data set. Properties refers to the main property of each variable, ie. number of elements, atomic mass, fie, atomic radius etc. Measurement refers to the different variation in measuring each property (except for number of elements and critical temperature), ie. mean, weighted mean, gmean, weighted gmean etc.

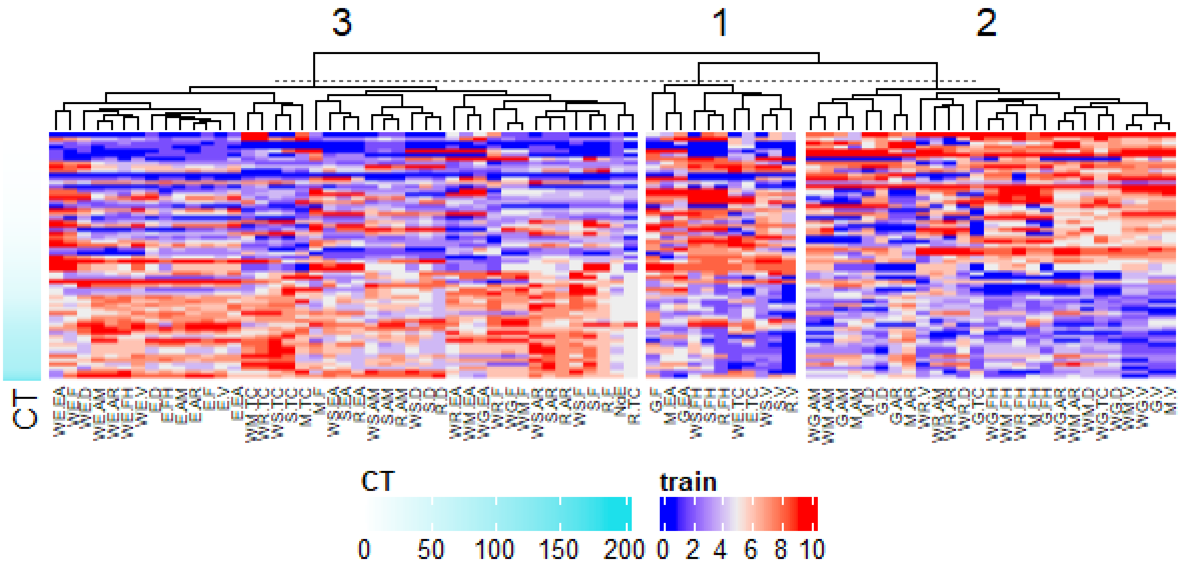


Figure 5: Overall heat map to display the association of each variable with critical temperature

3.1 Relationship summary between Critical Temperature and All Variables

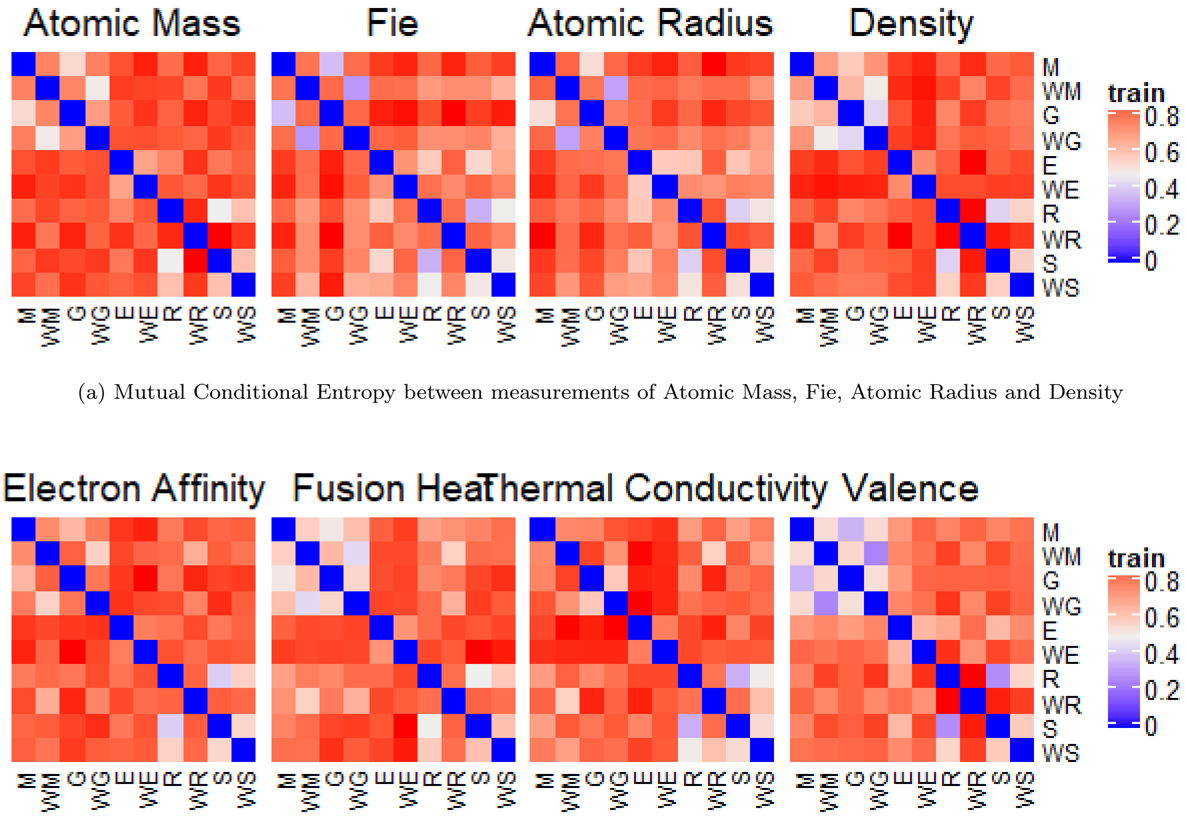
To better understand the association between each variable and critical temperature, a heat map is constructed as shown in Figure 5. As we can see, the distribution of variables in cluster 1 when corresponding to the increasing critical temperature is quite random. In other words, higher critical temperature does not exactly correspond to higher or lower values of those variables. This corresponds to some the distributions like the one in Figure 4b. Variables in this cluster include range valence, entropy themal conductivity, and standard fusion heat.

In cluster 2, most of the variables seem to be inversely proportional to critical temperature, ie. high critical temperature corresponds to low values of those variables. Some of the most notable ones are the weight and non-weighted mean and gmean measurements of valence.

In cluster 3, most of the variables seem to be proportional to critical temperature, ie. high critical temperature corresponds to high values of those variables. Interestingly, this includes most of the properties with entropy and weighted entropy measurements.

3.2 Relationship between Measurements among Properties

To find the association between any two variables, we calculate its Mutual Conditional Entropy. When this is done on every possible set of two variables, we will be able to form a matrix M to contain all the Mutual Conditional Entropy values. Note that the element m_{ij} in M corresponds to the Mutual Conditional Entropy between attribute i and attribute j.



(a) Mutual Conditional Entropy between measurements of Atomic Mass, Fie, Atomic Radius and Density

(b) Mutual Conditional Entropy between measurements of Electron Affinity, Fusion Heat, Thermal Conductivity and Valence

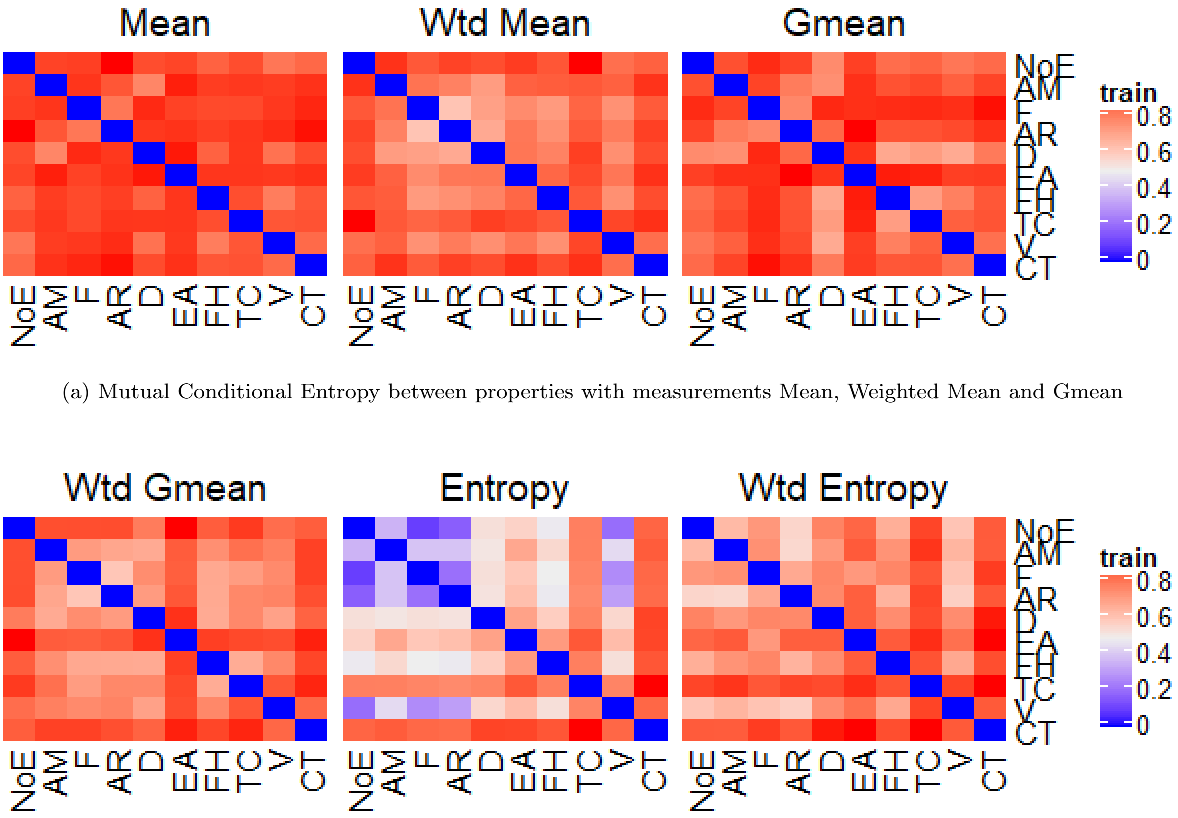
Figure 6: Mutual Conditional Entropy Heatmaps between measurements within each property

We then display the resulting matrix M in the form of heat maps. Figure 6 shows the heat maps of Mutual Conditional Entropy within each level 1 attribute (except number of elements and critical temperature).

From the eight heat maps, we can see that the top left and bottom right corners of each heat map tends to have a lighter shade of red or purple, whereas the rest of the heat map tend to have a darker shade of red. This means that the Mutual Conditional Entropy values between any two of mean, weighted mean, gmean and weighted gmean, and any two of range, standard and weighted standard is small, which suggest higher associativity between them.

On the other hand, we see extreme dark red bands at the bottom right of all the heat maps at weighted range against range, standard and weighted standard. Another set of darker red bands can be observed in the density, fusion heat and thermal conductivity heat maps between the set of entropy and weighted entropy and the set of mean, weighted mean, gmean and weighted gmean. This suggests little associativity between those variables. In general, we see little associativity between weighted range and range or both standard variables. Additionally, there is also a relatively weak relationship between both entropy and weighted entropy and all four types of means.

3.3 Relationship between Properties with the same Measurements



(a) Mutual Conditional Entropy between properties with measurements Mean, Weighted Mean and Gmean

(b) Mutual Conditional Entropy between properties with measurements Weighted Gmean, Entropy and Weighted Entropy

Figure 7: Mutual Conditional Entropy Heatmaps of all properties with the same measurements (Part 1)

Just like the way we analyzed the relationship between measurements for each property, we produce heat maps of Mutual Conditional Entropy between Properties with the same measurements. However, unlike the case in the previous subsection, we do not observe any straightforward patterns in all heat maps, and it is quite clear that for most measurements, the properties are not very associative with each other. The only exceptions are shown in the entropy and weighted entropy heat maps in Figure 7b. In the entropy heatmap, we can see that there is an extremely high association (low mutual condition entropy) between number of elements, atomic mass, Fie and atomic radius. A similar pattern can be observed in the weighted entropy and range heat maps respectively, although not as obvious.

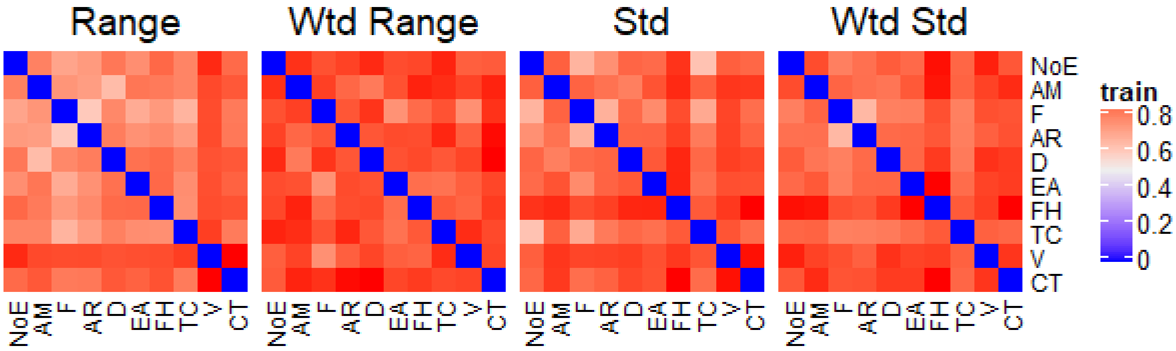


Figure 8: Mutual Conditional Entropy between properties with measurements Range, Weighted Range, Standard and Weighted Standard (Part 2)

On the other hand, it is important to note that the relatively low set of Mutual Conditional Entropy values with critical temperature seem to appear mostly in the range heat map in Figure 8. This suggests that a big portion of range attributes can be chosen to classify critical temperature in the next subsection.

3.4 Distribution of Critical Temperature based on Highly Associated Variables

We now move on to observing the distribution of critical temperature with respect to its most highly associative attributes. In Figure 7, we built two heat maps that display the classification of superconductors by low and high critical temperatures based on the distribution of the set of variables highly associated with critical temperature. Note that when selecting the set of variables, only one attribute is selected from each level 1 physical property. For example, if Weighted Mean Density has been selected, Gmean Density cannot be selected for the set.

On the left side of the figure, we classify the critical temperature based on variables with the highest correlation with critical temperature, which are:

- Weighted Standard Thermal Conductivity (WS.TC)
- Range Atomic Radius(R.AR)
- Weighted Entropy Atomic Mass (WE.AM)
- Number of Elements (NoE)
- Range Fie (R.F)
- Entropy Valence (E.V)
- Weighted Entropy Fusion Heat (WE.FH)
- Entropy Density (E.D)
- Entropy Electron Affinity (E.EA)

On the right side of the figure, we classify the critical temperature based on variables with the smallest Mutual Conditional Entropy values with critical temperature, which are:

- Gmean Density (G.D)
- Range Thermal Conductivity (R.TC)
- Range Fie (R.F)
- Weighted Mean Valence (WM.V)
- Weighted Entropy Fusion Heat (WE.FH)
- Weighted Entropy Atomic Mass (WE.AM)
- Number of Elements (NoE)
- Range Atomic Radius (R.AR)
- Range Electron Affinity (R.EA)

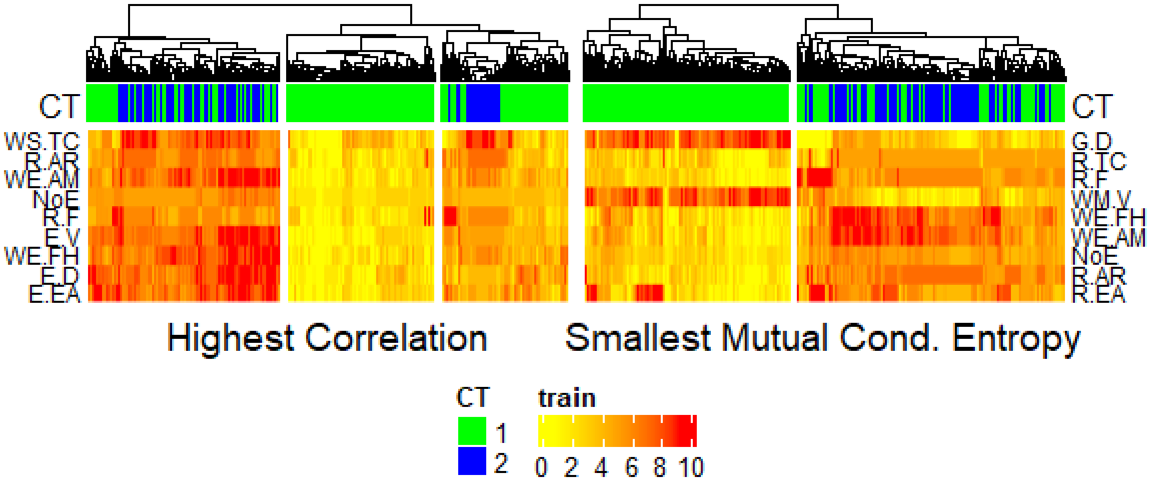


Figure 9: Distribution of Critical Temperature based on Highly Associated Variables

In the results shown in Figure 9, we can see that the classification of critical temperature using the set of variables that has the highest correlation with it gives us 3 clusters. Cluster 1 consists mostly of superconductors of high critical temperature. This is a cluster where all of the variables are presented at a higher set of values. In cluster 2, we see that all the superconductors have low critical temperature, and their corresponding variables are also presented at a lowest set of values. As for cluster 3, which consists of about $\frac{1}{3}$ of high critical temperature superconductors, all the corresponding variables tend to be presented at a mid-range set of values.

As for the classification of critical temperature using the set of variables with the smallest values of Mutual Conditional Entropy with critical temperature, we observe that the results can be split into two clusters. Cluster 1 corresponds to superconductors with high gmean density and range Fie values, with relatively small values on the other variables. This is a cluster where we can classify superconductors with low critical temperature with 100% accuracy. On the other hand, Cluster 2 corresponds to the exact opposite of Cluster 1 in terms of the values of the set of variables, and it yields about 70% of superconductors with high critical temperature.

4 Relationship between Critical Temperature and Composition of Elements of Superconductors

In Figure 10, we constructed a heat map that describes the element composition of some of the most abundant elements among superconductors, arranged by critical temperature. The goal of this heat map is to tell us the elements that are prevalent in superconductors of high and low critical temperatures.

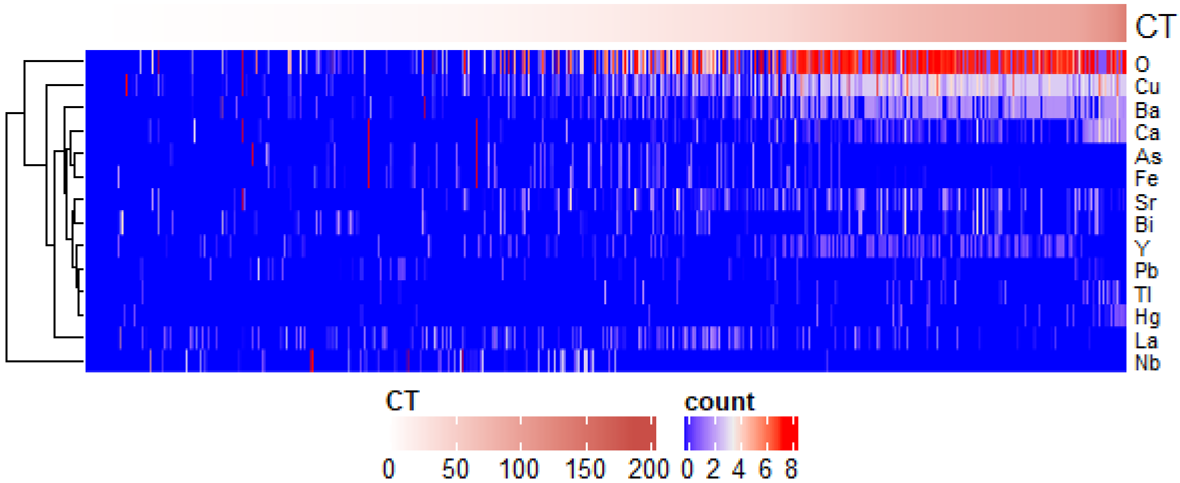
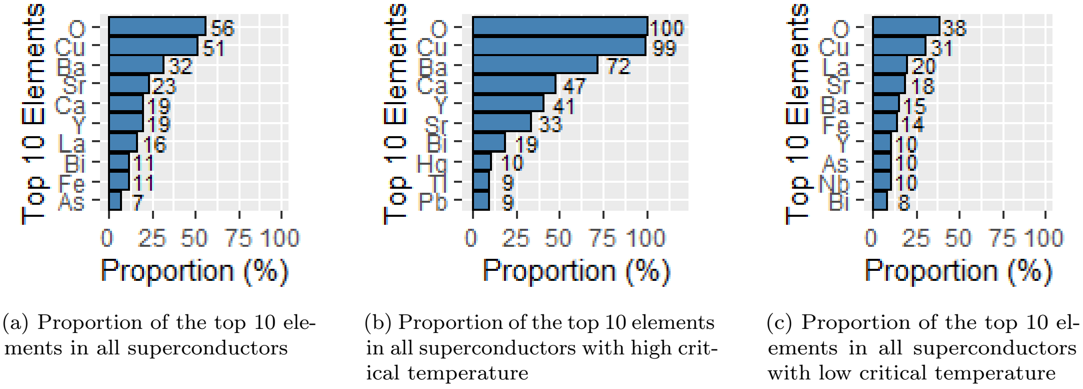


Figure 10: Heat map of some of the most abundant elements among superconductors

From the figure, we can see that Oxygen (O) is the most prevalent element, and it appears in large quantities especially in superconductors of high critical temperatures, where many contain a large number of oxygen molecules.

The next set of prevalent elements in superconductors with high critical temperatures are (in decreasing order) Copper (Cu), Barium (Ba), Calcium (Ca), and Strontium (Sr). Just like Oxygen, Copper also exists in large quantities among superconductors with high critical temperature. In this list, notice that they are all metals, with copper being a transition metal and the rest being alkali earth metals (or Group 2 elements). This could be due to the strong and stable ionic bonds that are present between these metals and Oxygen. However, it is interesting that all Group 1 elements (or alkali metals) and some Group 2 elements such as Lithium (Li), Beryllium (Be), Sodium (Na) and Magnesium (Mg) are not part of the compositions of most superconductors. It is possible that compounds formed with these elements are not very stable as compared to some other metals due to their high reactivity with Oxygen in the atmosphere.

On the flip side, we observed that no superconductor contains any noble gases (Group 18 elements), and only a very selected few of them contain halogens such as Fluorine (F), Chlorine (Cl), and Bromine (Br).



(a) Proportion of the top 10 elements in all superconductors

(b) Proportion of the top 10 elements in all superconductors with high critical temperature

(c) Proportion of the top 10 elements in all superconductors with low critical temperature

Figure 11: Prevalent elements in superconductors

To further understand the composition of elements in superconductors, we produce bar charts as shown in Figure 11 to display the top 10 elements among superconductors. As we can see in all three charts, Oxygen (O) and Copper (Cu) are the most prevalent elements regardless of the critical temperature of a superconductor, although it is especially true for superconductors with high critical temperature where almost all of them contain these two elements.

Another finding that produces the same conclusion as before is the prevalence of Group 2 elements, except that this time we find that they are not only prevalent among superconductors with high critical temperatures but for all superconductors.

A key difference that separates superconductors with high and low critical temperatures is the prevalence of Period 6 elements. Period 6 elements such as Bismuth (Bi), Mercury (Hg), Thallium (Tl) and Lead (Pb) are among the top 10 most prevalent elements in high critical temperature superconductors. On the other hand, with the exception of Copper, Yttrium (Y) and Bismuth (Bi), low critical temperature superconductors tend to contain various transition metals and Group 15 elements unlike superconductors with high critical temperature.

In short, Oxygen and Copper are the most prevalent elements in all superconductors, regardless of critical temperature. Then, we will find that there is a large proportion of high critical temperature superconductors that contain Group 2 elements or Period 6 elements, as compared to low critical temperature superconductors of which a large proportion contains transition elements or Group 15 elements.

5 Classification of Superconductors based by Critical Temperature

In this section, we classify the critical temperature of superconductors with various supervised machine learning classification methods. A random subset of 80% of the data is selected from the full data set as training data. The rest of the data will be considered as our testing data. We first "train" our training data by fitting it with some models such as a linear or polynomial fit. Then, we use this fitted model to predict the classes of our testing data and analyze its accuracy. The supervised learning techniques that will be used are described below:

Linear Regression

In Linear Regression, we fit the full model of the form $Y = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n$ on the training data. Then, we predict the critical temperatures of the superconductors in the testing data by keying in the variables needed. Since this method gives us a continuous response, we classify each response less than 50 Kelvin to be of low critical temperature, and high critical temperature for responses more than 50 Kelvin.

Linear Discriminant Analysis (LDA)

In LDA, we want to find a hyperplane that splits between the two classes such that each class is at opposite sides of the hyperplane. Once we have achieved this hyperplane, we predict the class of our testing data by checking the side of the hyperplane that each data point is at.

Support Vector Machines (SVM)

Just like LDA, we want to find the best hyperplane that splits between the two classes, except that we are implementing a different model to solve this problem.

Random Forest Classification

The random forest classification method utilizes a large number of decision trees from the various variables in the data set to produce classifications that best separate the low and high critical temperature among superconductors. After obtaining the best classification of the training set, we use the random forest model to predict our testing data.

5.1 Results

The results from the classification methods used are shown in Table 1. The calculation of error rate is done using the formula:

$$\frac{\text{False Positive} + \text{False Negative}}{\text{Total Data}}$$

Table 1: Confusion Matrix and Error Rates of all Classification Methods

a: Linear Regression

Actual \ Predicted	Low	High
Low	2639	338
High	42	1234

Error rate: $\frac{42+338}{4253} = 8.93\%$

b: Linear Discriminant Analysis (LDA)

Actual \ Predicted	Low	High
Low	2654	323
High	46	1230

Error rate: $\frac{46+323}{4253} = 8.67\%$

c: Support Vector Machines (SVM)

Actual \ Predicted	Low	High
Low	2869	108
High	78	1198

Error rate: $\frac{66+249}{4253} = 7.41\%$

d: Random Forest Classification

Actual \ Predicted	Low	High
Low	2869	108
High	78	1198

Error rate: $\frac{78+108}{4253} = 4.37\%$

5.2 Discussion

As we can see from this section, the accuracy of all the supervised learning methods are relatively low (below 10%). The random forest classification method can be considered as the best classification method with its error rate of less than 5%. However, the trade off to the better accuracy of Support Vector Machines and Random Forest Classification is the longer time taken to complete the classification. Additionally, while these methods produce some relatively accurate classification results, the same methods such as Linear Regression and Random Forest might not work as well under regression predictions.

6 Conclusion

The lack of association between many of these properties is what makes this data complex. There are some variables which highly associate with one another, some of which include the dual relationship between number of elements, atomic mass, fie, and atomic radius at the entropy factor, but they don't necessarily associate well with critical temperature..

When it comes to the association with critical temperature, both sets of predictor variables from correlation and Mutual Conditional Entropy gave really good classifications where we can quite distinctly observe the distribution of critical temperature and its association with each physical property. In addition to that, we found that Oxygen and Copper are the most abundant elements among superconductors, especially those with higher critical temperature. Also, Period 6 elements are more abundant among superconductors with high critical temperature, as compared to transition elements and Group 15 elements being more abundant among superconductors with low critical temperature.

Lastly, we experimented the classification of low and high critical temperature with some supervised machine learning techniques. In this data set, all the classification methods used have fared very well, especially the random forest classifier.