Final Project

Julia Webb

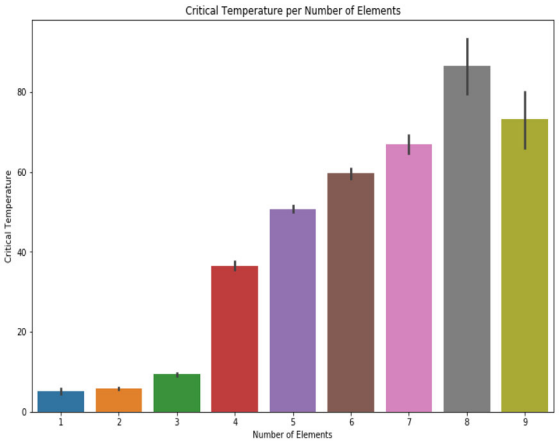
I. Abstract

This report is an analysis of the Superconductivity dataset from the UCI Machine Learning Repository. Specifically, the train.csv file from this data folder will be explored. This file provides values of 81 features, along with *Critical Temperature*, of 21,263 superconductors. Note that *Critical Temperature* is considered the response variable of this dataset, with the other 81 features considered the predictors. Correlation and mutual conditional entropy between these features are explored, and XGBoost is applied to predict the *Critical Temperature* of superconductors based on their feature values. Additionally, hierarchical clustering is performed on the predictors variables, with row labeling by a discretized version of *Critical Temperature*.

II. Analysis of Train Dataset

Each feature belongs to one of nine feature-types within this dataset: Number of Elements, Mass, Fie, Radius, Density, Electron Affinity, Fusion Heat, Thermal Conductivity, and Valence. For each feature type, a correlation heatmap and a mutual conditional entropy heatmap was created, to gain a better understanding of covariance within the data, as well as each feature-type's influence on critical temperature. Upon observing these tables, three-dimensional scatter plots were generated to visualize features that seemed to have a strong relationship with Critical Temperature. Moving forward, each feature group will be explored in subsections of this report.

i) Number of Elements Feature-Type



There is only one variable, *Number of Elements*, in this Feature-Type group. The correlation between *Number of Elements* and *Critical Temperature* is 0.601, and the mutual conditional entropy is 0.653. From observing the barplot in *Fig. 1*, it appears that the highest observation of *Critical Temperature* generally increases as the *Number of Elements increases*.

Figure 1: Number of Elements Barplot

ii) Mass Feature-Type

See *Fig. 2* for a correlation heatmap for the Mass Feature-Type group, with *Critical Temperature* included. Notice that sets of features within this group are highly correlated, resulting in the dark green boxes within the heatmap. Let us first consider the features in the first four rows of the heatmap; these features are all regarding mean atomic mass. It appears that *Weighted Mean Atomic Mass* and *Weighted G. Mean Atomic Mass* both have a stronger correlation with *Critical Temperature*, and with each other, relative to the other two features in this first block. Now, consider these same four variables as seen in *Fig. 3*, which is a heatmap of mutual conditional entropy of this Feature-Type group, with *Critical Temperature* included. *Weighted Mean Atomic Mass* and *Weighted G. Mean Atomic Mass* have a lower mutual-conditional entropy value with correlation, and with each other, relative to the other two features in the first block. This is consistent with the definitions of mutual conditional entropy and correlation. Next, we see that *Entropy Atomic Mass* and *Weighted Entropy Atomic Mass* are highly correlated, with *Weighted Entropy Atomic Mass* having a relatively strong correlation with *Critical Temperature* and a relatively weak mutual conditional entropy. Interestingly, *Range Atomic Mass* and *Weighted Range Atomic Mass* have almost no correlation with each other. Additionally, *Range Atomic Mass* has a positive correlation with *Critical Temperature*, while *Weighted Range Atomic Mass* has a negative correlation with *Critical Temperature*. Finally, *Standard Atomic Mass* and *Weighted Standard Atomic Mass* are highly correlated, with *Standard Atomic Mass* having a slightly higher correlation, but slightly lower mutual conditional entropy, with *Critical Temperature*.

After observing these heatmaps, a three-dimensional scatterplot of *Critical Temperature* against *Weighted Range Atomic Mass* and *Weighted Standard Atomic Mass* was generated (see *Fig. 4*). Higher values of *Critical Temperature* appear to be associated with values of *Weighted Range Atomic Mass* below 50, and values of *Weighted*

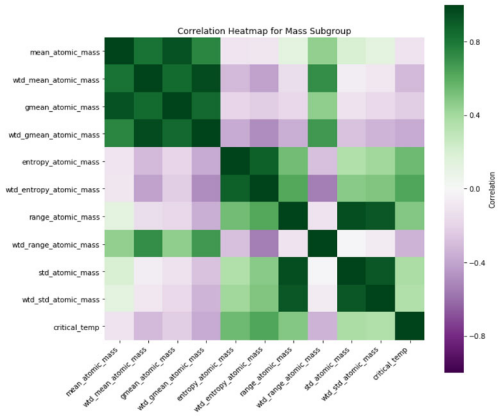


Figure 2: Correlation: Mass

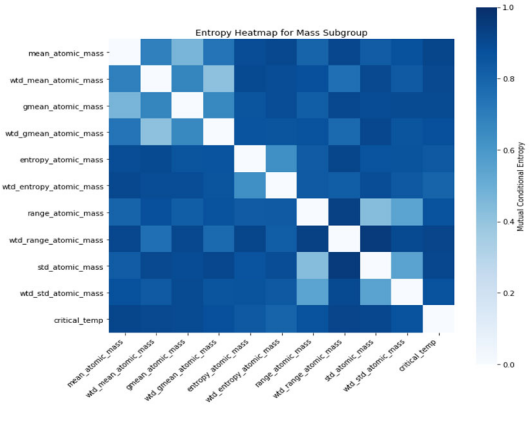


Figure 3: Entropy: Mass

Standard Atomic Mass above 20.

iii) Fie Feature Type

Observe *Fig. 5* for the correlation heatmap of the Fie feature-type subgroup, and *Fig. 6* for the mutual conditional entropy heatmap. Judging from *Fig. 5,* there appears to be less collinearity within this feature-type group, however, analogous sets of variables appear to be correlated with each other, again resulting in the groups of dark green boxes in the heatmap. Out of the first four rows of the correlation heatmap, *Weighted Mean Fie* and *Weighted G. Mean Fie* have a higher correlation with *Critical Temperature*, and a lower mutual conditional entropy with *Critical Temperature* (as seen in *Fig. 6*). *Range Fie* has a slightly higher correlation, and lower entropy, with the response than *Weighted Range Fie*. These two variables did not appear to have a high correlation with each other, however *Range Fie* had a strong correlation with *Standard Fie* and *Weighted Standard Fie*. The opposite is true for their entropies.

See *Fig. 7* for a three-dimensional scatterplot of *Critical Temperature* against *Weighted Mean Fie* and *Weighted Range Fie*. The data points in this plot appear to have an approximate bimodal distribution, with Critical Temperature peaking near *Weighted Mean Fie* values of 1000, and *Weighted Range Fie* Values of 600.

iv) Radius Feature-Type

See *Fig. 8* and *Fig. 9* for a correlation heatmap and entropy heatmap, respectively, of the Radius Feature-Type group. Again, we see sets of variables within this heatmap with a strong correlation. Out of the first four variables, *Weighted Mean Atomic Radius* and *Weighted G. Mean Atomic Radius* have a stronger correlation with *Critical Temperature*, and a very strong correlation with each other. Next consider the set of variables, we will call it set A, such that A = {*Entropy Atomic Radius*, *Weighted Entropy Atomic Radius*, *Range Atomic Radius*}. All of the variables in A have a negative correlation with each other, and negative correlation with set B = {*Weighted* ,

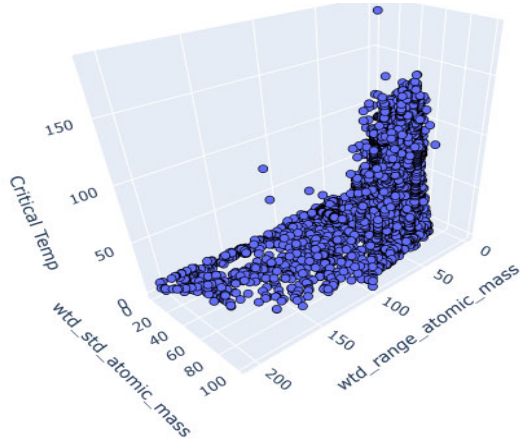


Figure 4: 3-dimensional Scatter Plot - Mass

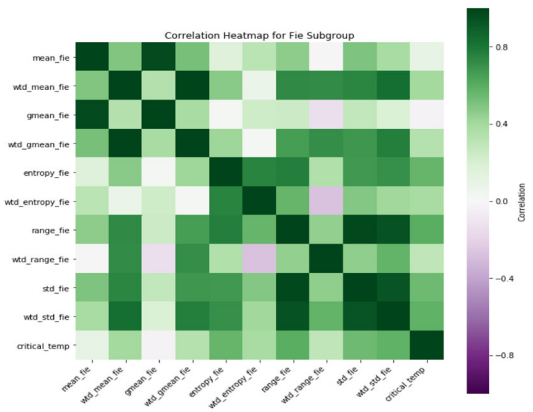


Figure 5: Correlation Heatmap - Fie

Mean Atomic Radius, *G. Mean Atomic Radius*, *Weighted G. Mean Atomic Radius*}, and a negative relationship with set C = {*Standard Atomic Radius*, *Weighted Standard Atomic Radius*, *Critical Temperature*}. These relationships result in the distinct 3 x 3 purple, green, green squares we see through the middle of the heatmap. This same pattern appears down the bottom row, illustrating that the variables in set C are negatively correlated those set A, and positively correlated with each other.

There seems to be less distinct sets in the entropy heatmap (Fig. 9). However, the first four variables appear to have generally lower entropy with each other, as do *Entropy Atomic Radius* and *Weighted Entropy Atomic Radius*. Additionally, there appears to be a low entropy set amongst the last four predictor variables on the heatmap (not including *Critical Temperature*).

To further visualize information contained in the heatmaps, see Fig. 10 for a three dimensional scatter plot of *Critical Temperature* against *Weighted Mean Atomic Radius* and *Weighted Standard Atomic Radius*. The data points appear to have an approximate bimodal distribution, with *Critical Temperature* peaking at near *Weighted Mean Atomic Radius* values and *Weighted Standard Atomic Radius* near (100, 75) and (170, 60).

v) Density Feature-Type

See *Fig. 11* and *Fig. 12* for the correlation heatmap and entropy heatmap, respectively, for the Density subgroup. Let A = {*Mean Density*, *Weighted Mean Density*, *G. Mean Density*, *Weighted G. Mean Density*}, B = {*Entropy Density*, *Weighted Entropy Density*}, and C = {*Range Density*, *Weighted Range Density*, *Standard Density*, *Weighted Standard Density*}. *Fig. 11* shows that variables within set A are strongly positively correlated with each other, as are variables within set B. We also see a strong correlation between the variables within set C, with the

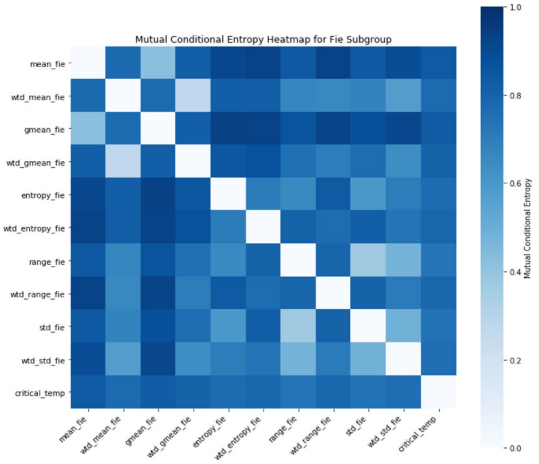


Figure 6: Mutual Conditional Entropy - Fie

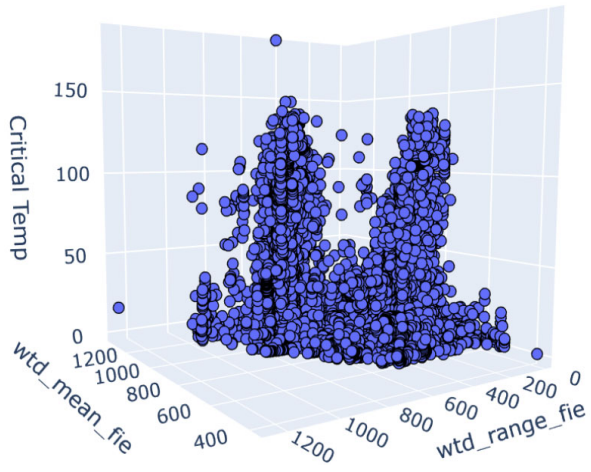


Figure 7: 3-dimensional scatter plot - Fie

exception of *Weighted Range Density*, which has almost no correlation with variables within its set. Observing *Fig. 12*, the variables in each of these sets appear to have low mutual conditional entropy with each other, again with the exception of *Weighted Range Density*, which has high mutual conditional entropy with the other variables in set C.

To better visualize the information contained in the heatmaps, see *Fig. 13* for a three-dimensional scatterplot of *Critical Temperature* against *Weighted Mean Density* and *Weighted Standard Density*. These data points display a large spread, but *Critical Temperature* peaks at *Weighted Mean Density* values near 5,000 and *Weighted Standard Density* values of 4,000.

vi) Electron Affinity Feature-Type

See *Fig. 14* and *Fig. 15* for a correlation heatmap and entropy heatmap, respectively, of *Electron Affinity*. Let A = {*Mean Electron Affinity*, *Weighted Mean Electron Affinity*, *G. Mean Electron Affinity*, *Weighted G. Mean Electron Affinity*}, B = {*Entropy Electron Affinity*, *Weighted Entropy Electron Affinity*}, and C = {*Range Electron Affinity*, *Weighted Range Electron Affinity*, *Standard Electron Affinity*, *Weighted Standard Electron Affinity*}. Features within set A and B display a strong correlation with the other variables in their set. Features within set C also appear to be strongly correlated with each other, with the exception of *Weighted Range Electron Affinity*, which has a weaker correlation with variables within its respective set. Observing the entropy heatmap in *Fig. 15*, variables appear to have lower mutual conditional entropy with other features within their respective set, resulting in the lighter zones on the heatmap. Again, this is with the exception of *Weighted Range Electron Affinity* which appears to have a strong mutual conditional entropy with the other features in set C.

See Fig. 16 for a scatterplot of *Critical Temperature* against *Weighted Range Electron Affinity* and *Weighted Standard Electron Affinity*. There is not much of a pattern in the distribution of these data points. However, there are

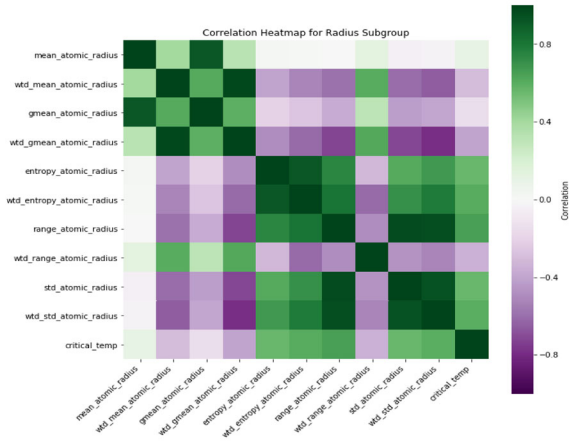


Figure 8: Correlation - Radius

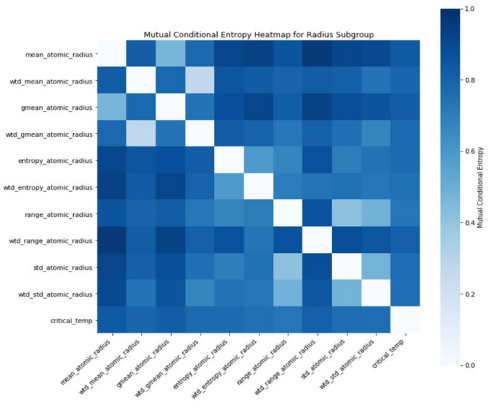


Figure 9: Entropy - Radius

noticeable peaks in Critical Temperature near *Weighted Range Electron Affinity* and *Weighted Standard Electron Affinity* values near 70 and 60, respectively.

vii) Fusion Heat Feature-Type

See *Fig. 17* and *Fig. 18* for a correlation and entropy heatmap, respectively. Let Let A = {*Mean Fusion Heat*, *Weighted Mean Fusion Heat*, *G. Mean Fusion Heat*, *Weighted G. Mean Fusion Heat*}, B = {*Entropy Fusion Heat, Weighted Entropy Fusion Heat*}, and C = {*Range Fusion Heat*, *Weighted Range Fusion Heat*, *Standard Fusion Heat*, *Weighted Standard Fusion Heat*}. The variables within group A have a very strong correlation with each other, as well as the variables within group B and group C. *Weighted Range Fusion Heat* has a lower correlation with the other variables in group C, compared to the rest of the variables in the set. The variables in sets A and C all have a moderate to weak negative correlation with *Critical Temperature*, while the variables in set B have a positive correlation with *Critical Temperature*. It is also interesting to note that set B has a negative correlation with set A and C, while set A and C are positively correlated. These relationships create an interesting symmetry centered in the middle of the correlation heatmap, if you disregard the out-most row and columns that correspond to critical temperature. A similar symmetry can be seen in *Fig. 18*, where variables that had a high correlation show a low mutual conditional entropy.

The variables in this feature-type group show a high mutual conditional entropy, and weak correlation, with *Critical Temperature*. Due to this fact, there were no three dimensional scatter plots (with *Critical Temperature* on the z-axis), that showed a noticeable distribution.

viii) Thermal Conductivity Feature-Type

Observe Fig. 19 for the correlation heatmap of the Thermal Conductivity group. Let A = {*Mean Thermal Conductivity*, *Weighted Mean Thermal Conductivity*, *G. Mean Thermal Conductivity*,

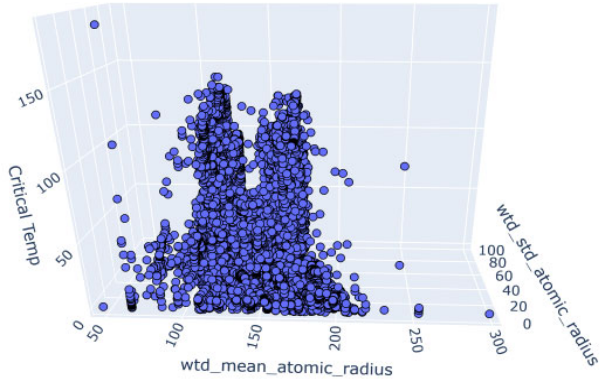


Figure 10: 3-dimensional Scatter Plot - Radius

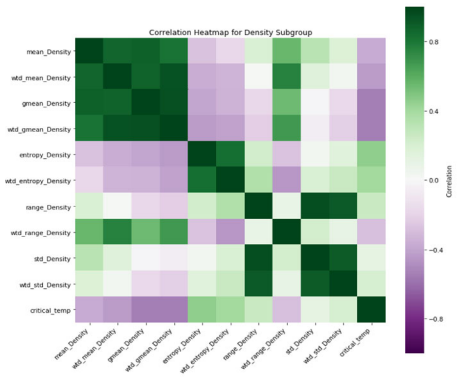


Figure 11: Correlation - Density

Weighted G. Mean Thermal Conductivity}, B = {*Entropy Thermal Conductivity*, *Weighted Entropy Thermal Conductivity*}, and C = {*Range Thermal Conductivity*, *Weighted Range Thermal Conductivity*, *Standard Thermal Conductivity*, *Weighted Standard Thermal Conductivity*}. Variables within set A, B, and C have a positive correlation with each other. Set C has a positive correlation with *Critical Temperature*, while B has a negative correlation with the response. Within set A, the variables regarding the mean Thermal Conductivity have a weak positive correlation with *Critical Temperature*, while the variables regarding the gmean Thermal Conductivity show a weak negative correlation with *Critical Temperature*.

See Fig. 20 for a heatmap of mutual conditional entropy within the Thermal Conductivity feature-type group. With each respective set (A, B, C), features appear to have relatively low mutual conditional entropy with other variables in their set. All variables appear to have a relatively high mutual conditional entropy with *Critical Temperature*. Three-dimensional scatter plots were generated of *Critical Temperature* against various features from the Thermal Conductivity feature-type group, however none of the scatterplots showed a significant pattern, so they are not included in this report.

ix) Valence Feature-Type

Let A = {*Mean Valence*, *Weighted Mean Valence*, *G. Mean Valence*, *Weighted G. Mean Valence*}, B = {*Entropy Valence*, *Weighted Entropy Valence*}, and C = {*Range Valence*, *Weighted Range Valence*, *Standard Valence*, *Weighted Standard Valence*}. Judging from the correlation heatmap in *Fig. 21*, variables within set A are very highly correlated with each other, as well as variables in set B. Additionally, variables set A have a strong negative correlation with variables in set B. With the exception of *Weighted Range Valence*, variables in set C have strong positive correlation with each other, but close to zero correlation with variables outside of their set. On the other hand, *Weighted Range Valence* has very low correlation with set C, but a positive correlation with variables in

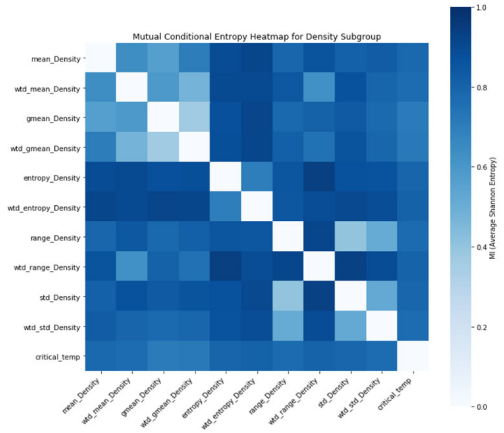


Figure 12: Entropy - Density

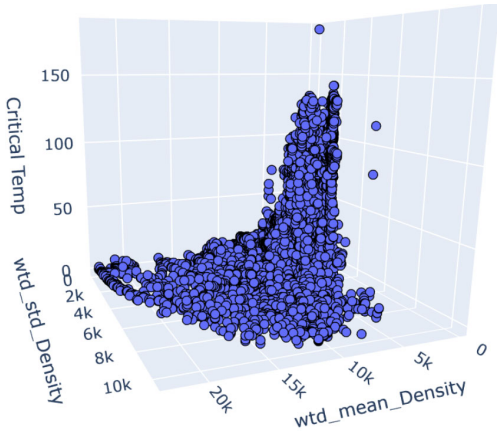


Figure 13: 3-dimensional Scatter Plot - Density

set A and a negative correlation with variables in set B. Set A has a negative correlation with *Critical Temperature*, while set B has a negative correlation with the response. Set C has a weak negative correlation with *Critical Temperature*. The entropy heatmap (*Fig. 22*) shows a reverse pattern from the correlation heatmap . However, the pattern is less distinct than in *Fig. 21*.

See *Fig. 23* for a three-dimensional scatterplot of *Critical Temperature* against *Weighted Mean Valence* and *Range Valence*. *Range Valence* only takes on discrete integer values, but it is interesting to note that the distribution of *Critical Temperature* against *Weighted Mean Valence* is relatively similar for every value of *Range Valence*. The data points are skewed towards larger values of *Weighted Mean Valence* with *Critical Temperature* peaking near *Weighted Mean Valence* values of 2.

III. Prediction

The train dataset is large and complicated, with high levels of correlation between the predictors, and a seemingly non-linear relationship between the response variable, *Critical Temperature*, and the predictors. To make predictions with such a complex dataset, XGBoost (eXtreme Gradient Boosting) was applied. The resulting classifier had a cross validation score of 0.91 with standard deviation of 0.00248 when validating the training set of predictors and critical temperatures. Furthermore, it had a cross validation score of 0.88 and standard deviation of 0.00828 when validating the test set of predictors and critical temperatures. The Mean Squared Error was 102.42.

IV Classification

Critical Temperature was discretized into four groups, based on its quantiles. The grouping is as follows:

Group 1: (-0.00079, 5.365] *Group 2: (5.365, 20.0]* *Group 3: (20.0, 63.0]* *Group 4: (63.0, 185.0]*

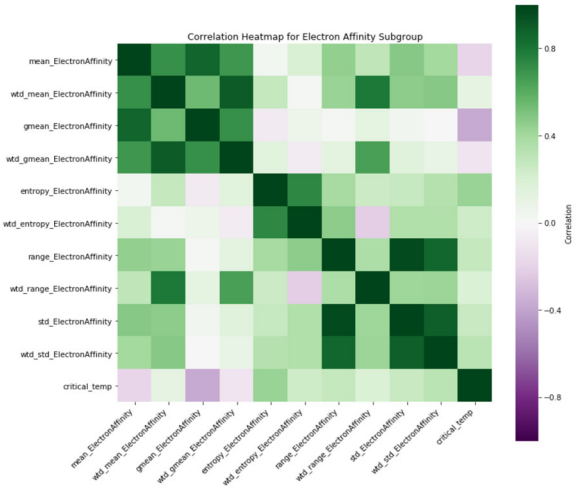


Figure 14: Correlation - Electron Affinity

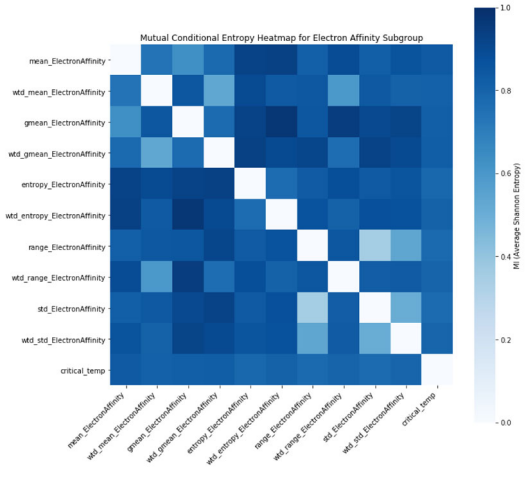


Figure 15: Entropy - Electron Affinity

Next, hierarchical clustering was performed with each feature-type group (*Fig. 24*). After the clustering process was completed, the rows were labeled by their corresponding *Critical Temperatur*e group. In every clustering, *Group 1* and *Group 3* tended to cluster together, and *Group 2* and *Group 4* tended to cluster together. Thus, rows belonging to *Group 1* and *Group 3* are color coded with light green and dark green, respectively. Rows belonging to *Group 2* and *Group 4* are color coded with light blue and dark blue, respectively.

V. Works Cited

DataTechNotes.com https://www.datatechnotes.com/2019/06/regression-example-with-xgbregressor-in.html
*STA 160 Discussion Slide*s: Unsupervised Learning, Trees and Ensemble Learning

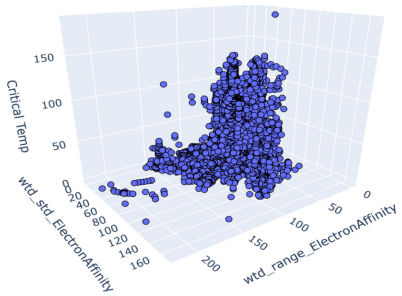


Figure 16: 3-dimensional scatter plot - Electron Affinity

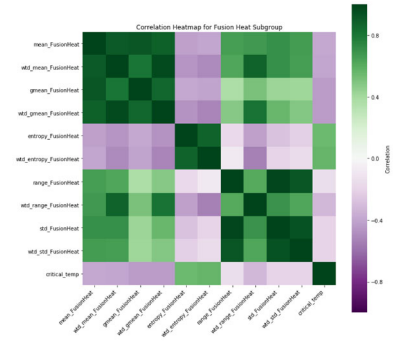


Figure 17: Correlation - Fusion Heat

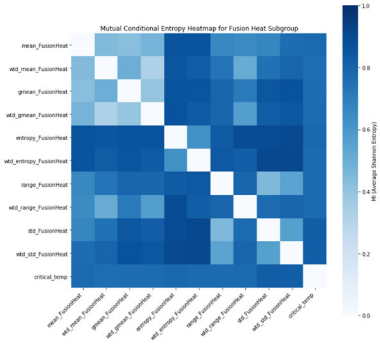


Figure 18: Entropy - Fusion Heat

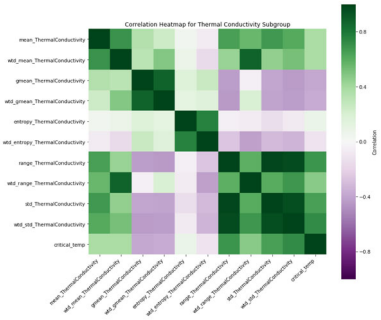


Figure 19: Correlation - Thermal Conductivity

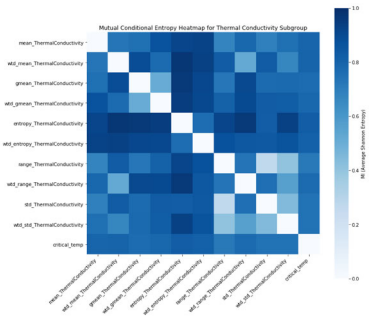


Figure 20: Entropy - Thermal Conductivity

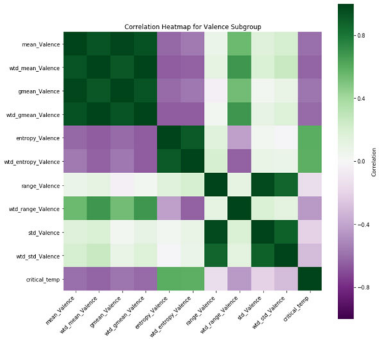


Figure 21: Correlation - Valence

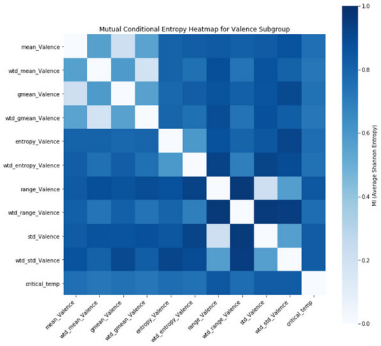


Figure 22: Entropy - Valence

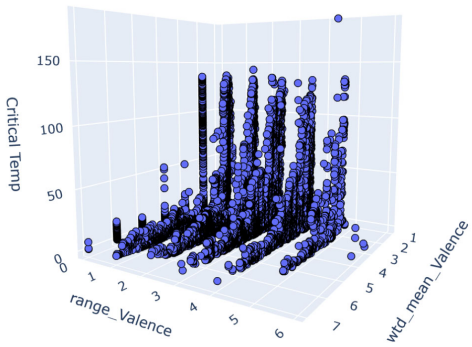
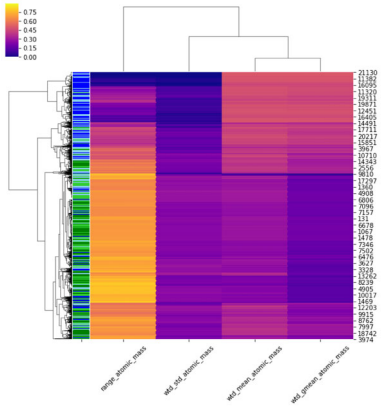
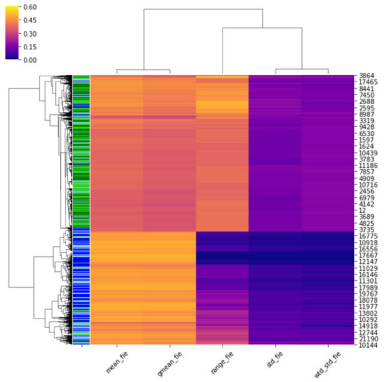


Figure 23: 3-dimensional scatter plot Valence

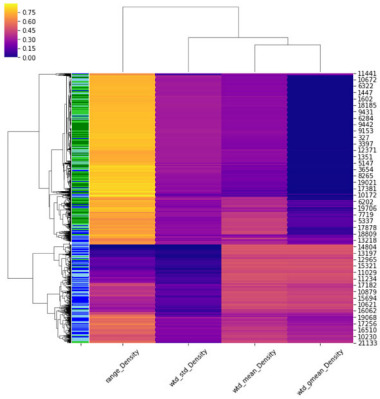
Figure 24: Hierarchical Clusterings Per Feature Type:



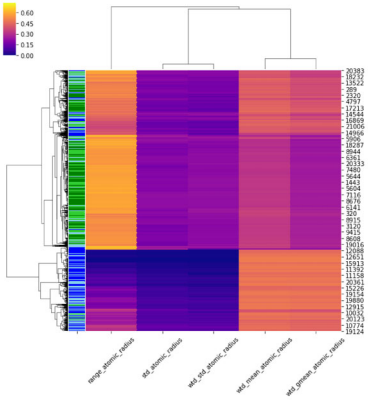
Mass



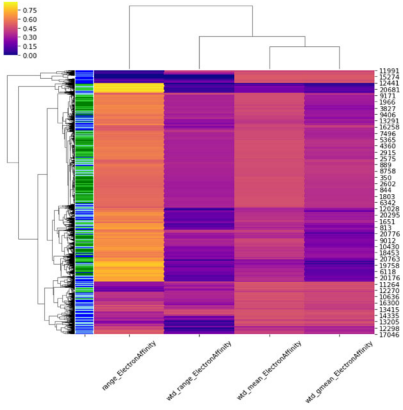
Fie



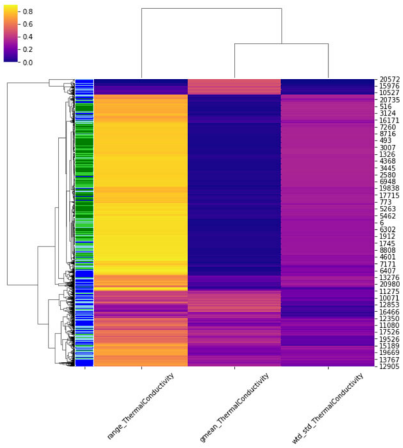
Density



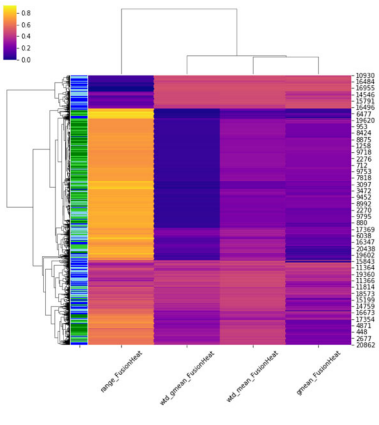
Radius



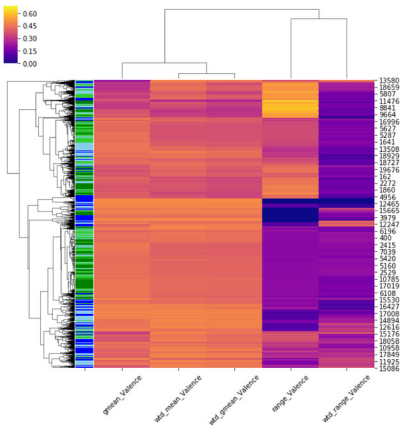
Electron Affinity



Thermal Conductivity



Fusion Heat



Valence