Comprehensive Statistical Study of Superconductors Critical Temperature

Sriram Reddy A
Department of Data Science
University of Colorado, Boulder
srar4232@colorado.edu

Sashank G
Department of Data Science
University of Colorado, Boulder
saga8160@colorado.edu

***Abstract*— This project involves a comprehensive analysis of the superconductors data to identify the most important features that have an impact on the critical temperature of superconducting materials and also how successfully we can predict for new data. To accomplish this objective, several machine learning models such as linear regression, random forests, GAMs, and SVMs were implemented, enabling us to determine the key predictors and understand their impact on the critical temperature. The outcomes of this analysis revealed that features such as weighted entropy atomic mass, weighted entropy atomic radius, weighted entropy valence, weighted standard thermal conductivity, and weighted entropy first ionization energy have a notable influence on predicting the critical temperature of superconductors. Additionally, we observed that the number of elements present in the material also plays a crucial role in determining the critical temperature. The insights obtained from this analysis have significant implications in the development of novel superconducting materials with improved critical temperatures. Additionally, the analysis underscores the importance of a thorough understanding of the physical and chemical properties of materials in predicting their superconducting behavior. The integration of machine learning models in this analysis highlights their potential in providing valuable insights into intricate phenomena like superconductivity.**

***Index Terms* — Superconductors, Critical Temperature, Linear Regression, SVM, GAM, Trees, Important Features, Correlation.**

I. INTRODUCTION

For more than a century, scientists and engineers have been fascinated by the remarkable class of materials known as superconductors. Due to their exceptional zero resistance electrical conductivity, electrical current can pass through them without experiencing any energy loss or heat dissipation. A unique kind of vibration known as a phonon interacts with the superconductor's electrons to produce this feature. A peculiar state known as a Cooper pair is formed when a superconductor is cooled to a specific temperature, enabling electrical current to pass without encountering any resistance.

Superconductors have been the focus of significant study and development since their discovery in 1911. Finding superconductors that can operate at greater temperatures has proven to be a significant problem, as doing so would make them more useful and affordable for practical uses. Despite this difficulty, superconductors have been used in a variety of significant applications, including magnetic levitation-powered high-speed trains, particle accelerators, and MRI machines.

Superconductors have crucial implications for fundamental physics research in addition to their practical applications, as they enable researchers to examine the behavior of electrons in a rare realm where they behave collectively rather than individually.

Superconductors still have some issues that need to be resolved, despite the tremendous future potential they present. For instance, they are costly, challenging to create, and work at extremely low temperatures. Nevertheless, these issues are being addressed by continual research and development, and new superconductors are consistently found and examined. Future developments of these materials are expected to involve even more fascinating uses as our knowledge of superconductivity deepens.

The dataset we'll be working with is a great resource for knowledge on the physical characteristics of superconductors, and it offers a great chance to look into the connections between these characteristics and the temperature at which superconductivity occurs.

This dataset will be analyzed using a variety of statistical models in order to identify the attributes that are most strongly connected with the critical temperature. By doing this, we aim to learn more about the variables that affect superconductivity and find potential novel materials that might display this characteristic. In order to identify the model that best fits the data and yields the most precise predictions, we intend to investigate various models, including Linear Regression, Generalized Additive Models, Decision Tree Regressor, Support Vector Machines & Random Forest Regressor.

Also in this context, several questions arise, such as can we identify the most significant factors that influence the critical temperature of high-temperature superconductors, beyond the elemental composition of the material, and how statistical modeling and machine learning techniques can be used to optimize the properties of superconducting materials. Additionally, it is of interest to develop a predictive model to forecast the critical temperature of superconducting materials based on their chemical composition and to identify any trends or patterns in the dataset that can help us understand the underlying physics of superconductivity.

Furthermore, unsupervised learning techniques can be employed to cluster superconducting materials based on their properties and identify any new trends or patterns that may emerge. This research aims to address some of these questions using various statistical and machine learning techniques on the superconductor dataset.

II. RELATED WORK

Drawing on previous studies in the field of superconductivity can provide numerous benefits. Firstly, it can offer a more comprehensive understanding of the research questions and hypotheses being investigated, and help to identify any gaps, limitations, or controversies in the existing body of knowledge. Secondly, prior research can serve as a benchmark for comparison and validation of the methods and conclusions of the present study, as well as for assessing the robustness and generalizability of the findings. Finally, earlier studies can shed light on the practical applications and implications of the results in the context of superconductors and aid in determining the goals and direction for future research.

[1] This paper presents a machine learning approach to identify key chemical properties that are related to superconductivity. The authors identified 25 chemical characteristics for each material from a collection of 212 superconducting materials. They used a number of machine learning models, such as gradient boosting, decision trees, and random forests, to forecast the critical temperature (Tc), a crucial characteristic of superconductors. The scientists discovered that the random forest model had the best predictive accuracy, with a mean absolute error of 12.2 K. The most crucial chemical characteristics associated with superconductivity, such as the atoms' electro negativity and the quantity of valence electrons, were also determined using feature importance scores. The authors propose that the search for novel superconducting materials with acceptable features can be guided by their method. Although the study identified chemical properties related to superconductivity using machine learning techniques, we believe that there is still room for further statistical analysis on the 25 chemical characteristics. Decision trees typically consider all parameters unless the impurity indicates otherwise, but with the abundance of data and 82 features available, we believe that there may be other ways to analyze the data from a different perspective.

[2] This study investigates how machine learning methods might be used to evaluate crashes and identify hotspots. The authors gather information on traffic accidents and utilize a variety of machine learning methods, such as decision trees, and support vector machines, to analyze the information and pinpoint the causes of traffic accidents. The findings demonstrate that these machine learning systems can accurately forecast hotspots and offer insightful data on the factors that contribute to traffic accidents. The research makes a significant addition to the subject of road safety by highlighting the potential of machine learning for the analysis of traffic accidents.

III. DATA & AREA OF INTEREST

Investigating into superconductor's data is important as we have discussed earlier. Understanding the properties and behavior of superconductors can lead to the development of more efficient and cost-effective electrical power transmission systems, magnetic levitation trains, and medical imaging devices. Moreover, research on superconductors can also contribute to the development of new materials with novel properties and pave the way for new technological advancements.

The Superconductors Data was gathered from multiple sources and preprocessed for public use. It consists of 82 features connected to 21 attributes of 21263 superconductors. The dataset contains a variety of different types of attributes, including, but not limited to, atomic mass, atomic radius, density, electron affinity, fusion heat, thermal conductivity, and valence. Understanding the behavior of superconductors and the variables influencing their capacity to conduct electricity without resistance requires an understanding of these characteristics. These measures give researchers the ability to examine the distribution of each attribute and spot any potential trends or patterns.

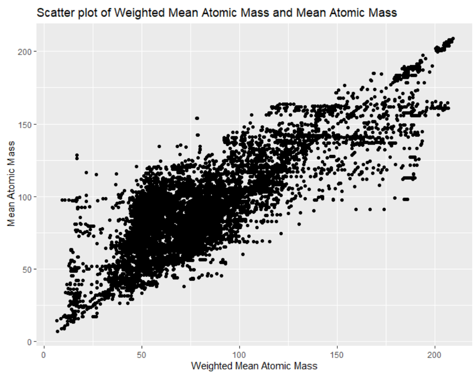
Using this dataset, it is possible to estimate the critical temperature column, a crucial variable in the understanding and characterization of superconductors. A precise estimate of the critical temperature can be used to improve currently available materials or create new ones with greater critical temperatures, as well as to help researchers understand the underlying physics and mechanisms that control superconductivity. The Superconductors Data can therefore be used to create and test new machine learning models and analysis that can anticipate the characteristics and behavior of superconductors, which could speed up the search for new materials with desirable features.

In statistical analysis, it's crucial to establish whether variables are correlated or whether there is a causal connection between them. Correlation just suggests that the variables are related, whereas causality means that a change in one variable directly results in a change in another. In the Superconductors Data, it is possible to investigate the relationship between specific superconducting qualities and their critical temperature. One can speculate, for instance, that a superconductor's critical temperature and density are causally related since a higher density might result in stronger interactions between electrons and a higher critical temperature.

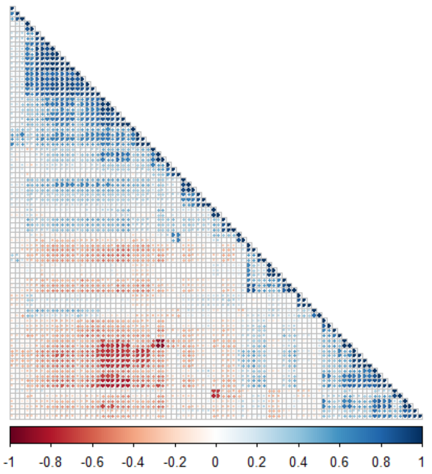
To investigate this possibility, one could use statistical methods such as regression analysis to assess the strength and direction of the relationship between density and critical temperature. However, it is important to note that correlation does not necessarily imply causation, and there may be other factors that influence the relationship between density and critical temperature. Another example is the relationship between the number of elements in a superconductor and its critical temperature. It is possible that a greater number of elements could lead to a higher critical temperature due to increased complexity in the material's structure. Again, statistical methods could be used to explore this relationship, but other factors such as the specific elements used and their properties could also have an impact on critical temperature.

Furthermore, the dataset provides a valuable resource for developing and testing new machine learning algorithms and models that can predict the properties and behavior of superconductors. By applying advanced analytical techniques to the vast amount of data available in the dataset, researchers can gain insights into the complex relationships between various properties and characteristics of superconductors. This can help accelerate the discovery of new materials with desirable properties and optimize existing materials for specific applications.

IV. DATA ANALYSIS

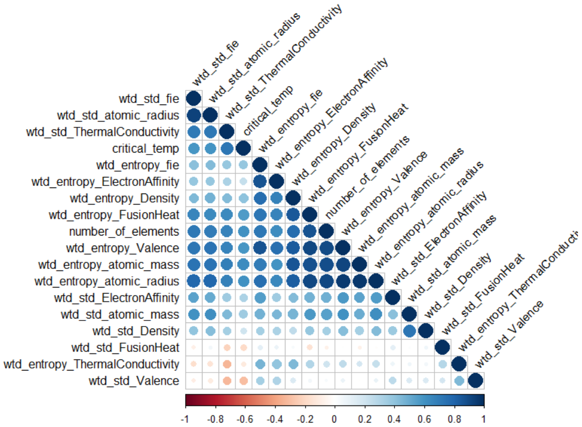


During the analysis of the superconductor dataset, it was found that out of 81 features, only 8 unique properties were present. The remaining columns were repetitions using different calculations mean_atomic_mass, wtd_mean_atomic_mass, gmean_atomic_mass, and wtd_gmean_atomic_mass, which are all related to the average atomic mass of the superconductor, but differ in the way they are calculated. Similarly, there are multiple columns related to properties like atomic radius, density, electron affinity, fusion heat, thermal conductivity, and valence, each calculated in different ways.



In our analysis, we used the corrplot function in R to visualize the correlation matrix of the superconductor dataset. This plot revealed a high degree of correlation between many of the columns, with some correlations exceeding 0.9. The above is the correlation plot for all columns, here we can see that there exists some correlation between the dependent variables also, which if left could lead to multicollinearity, overfitting, increased complexity, and reduced interpretability of the results. Therefore, it is necessary to identify and remove these redundant columns from the dataset. Additionally, it can help to identify the most important features that contribute to the critical temperature of the superconductors.

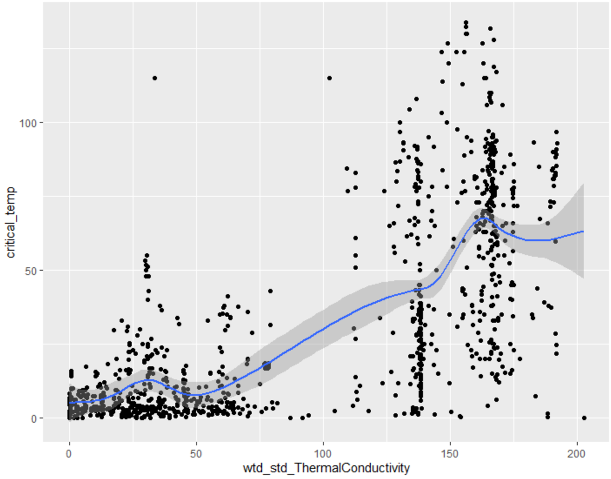
In addition to eliminating the redundant columns based on their correlation with the critical temperature column, we also conducted further analysis to identify the most informative set of repeated columns in the Superconductors Data. To achieve this, we created correlation plots for each set of columns representing the same property and compared these plots. Our analysis revealed that the weighted standard properties with weighted entropies had the highest correlation with the critical temperature, making them the most informative columns for predicting the critical temperature of superconductors. This approach helped us to identify the most important features in the dataset and streamline it for further analysis and modeling.



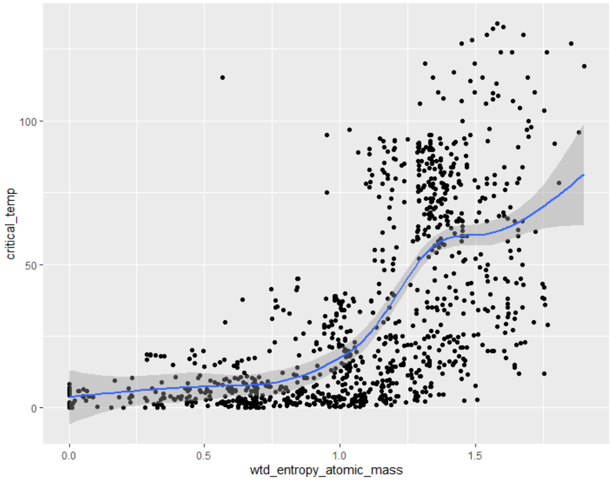
Above is the correlation plot of the new data frame created. Although some degree of correlation exists between the various properties of superconductors, the dataset that includes all the properties was found to be the most suitable for further analysis. While the correlation between properties needs to be taken into consideration when performing statistical analysis, having access to as much information as possible can provide valuable insights into the behavior of superconductors. Therefore, it was deemed necessary to include all properties in the dataset, despite the presence of correlation. This approach can help in identifying potential patterns and relationships between the properties and the critical temperature.

Moreover, having access to a comprehensive dataset can facilitate the development of accurate predictive models for critical temperature estimation. Thus, despite the presence of correlation, the Property dataset provides a rich source of information that can aid in advancing the field of superconductivity. Visualizing the distribution of data according to the selected properties can provide valuable insights for further analysis. By examining the distribution of the data, we can determine if any transformation is needed to normalize the data and reduce the impact of outliers.

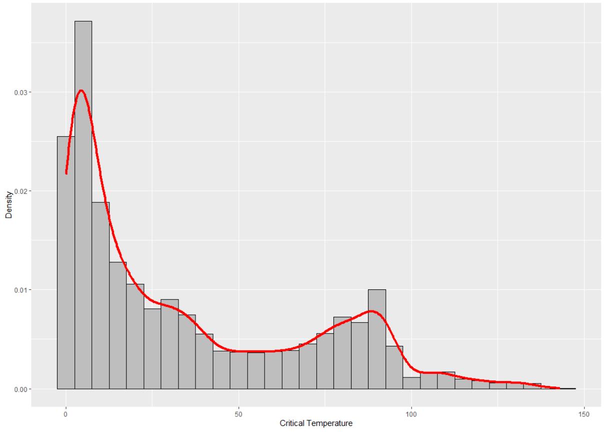
One way to visualize the distribution of the data is through histograms or density plots. These plots can show the frequency distribution of the data for each property, highlighting any patterns or deviations from normal distribution. Another useful visualization technique is the boxplot, which can show the median, quartiles, and outliers of the data for each property. For instance, if the data is heavily skewed or has a large number of outliers, we may need to apply a logarithmic or square-root transformation to normalize the data. This can help to ensure that the models and algorithms used for prediction tasks are robust and accurate.



Upon examining the distribution of the data based on the selected properties, it was observed that the variance of the data increased with the increase in the critical temperature. This phenomenon is known as heteroskedasticity and suggests that the relationship between the properties and the critical temperature may not be linear. To address this issue, we decided to explore the possibility of applying transformations to the data. Specifically, we considered exponential and logarithmic transformations, which are commonly used in statistical analysis to transform non-linear data into a linear form.

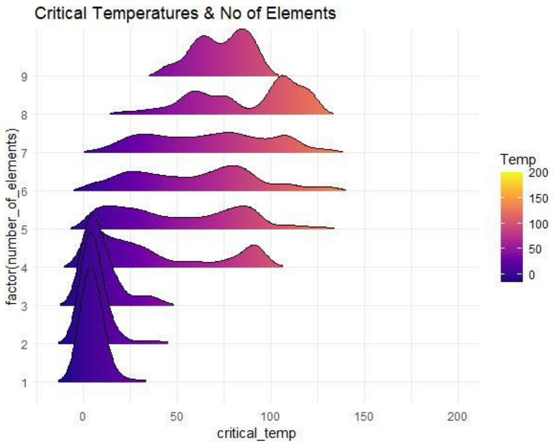


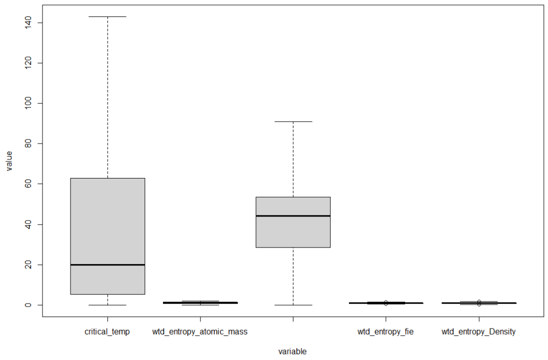
However, after applying these transformations to the data, we found that the relationship between the properties and the critical temperature did not improve significantly. This suggests that the transformation may not be appropriate for the data and that the relationship between the properties and the critical temperature may be inherently non-linear. Therefore, we decided to explore other modeling techniques that can handle non-linear relationships, such as polynomial regression or non-parametric regression. These techniques can capture the non-linear relationship between the properties and the critical temperature more effectively, and can potentially lead to better predictions and insights.



One interesting observation during our analysis is that there were no outliers present in the data, indicating that the data is informative and has a relatively uniform distribution. This is important as outliers can skew the results and distort the relationship between variables. The absence of outliers suggests that the data has been carefully collected and preprocessed, ensuring the reliability and quality of the dataset.

To further examine the distribution of the data and identify any potential trends or relationships, we created box plots for the top 5 variables of interest. These variables were selected based on their correlation with the critical temperature and their relevance to the properties of the superconductors. The box plots provide a clear visualization of the median, quartiles, and outliers of the data for each variable, allowing us to identify any patterns or differences in the distribution between variables.

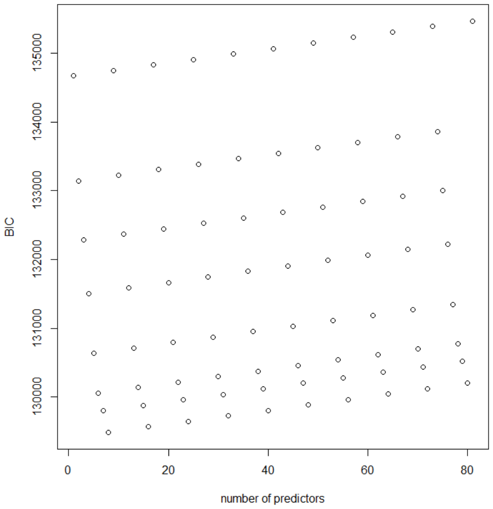




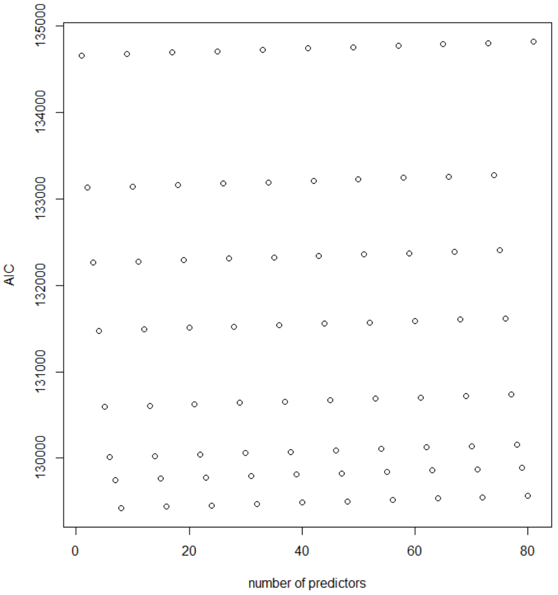
The absence of outliers and the uniform distribution of the data indicate that the dataset is suitable for further analysis and modeling. However, it is important to note that the box plots only provide a limited view of the distribution of the data and other visualization techniques, such as histograms or density plots, should also be considered to gain a more comprehensive understanding of the data distribution.

After completing the data analysis and exploration process, we are now ready to move onto the data modeling stage. With a better understanding of the data, including its properties and relationships, we can begin to develop and test various machine learning algorithms and models to predict the critical temperature of new superconductors. The insights gained from the data analysis can inform our selection of appropriate modeling techniques and features, as well as help us to evaluate the performance and accuracy of the models. Through the data modeling process, we can aim to develop models that accurately predict the critical temperature of superconductors.

V. DATA MODELLING

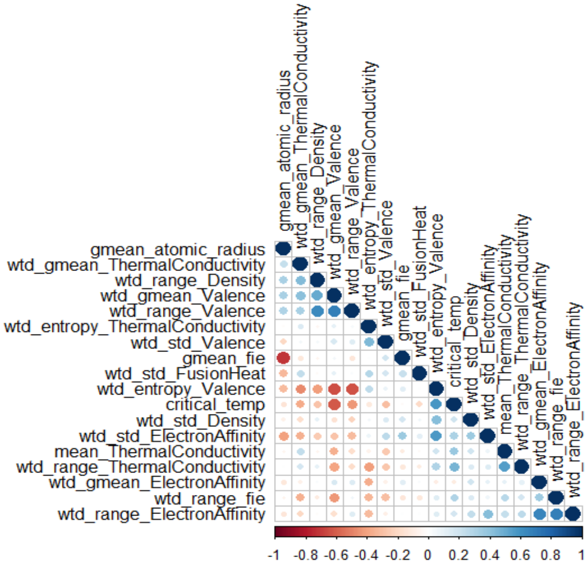


In addition to the specific set of columns we had previously identified, we have explored other possible combinations of columns using statistical techniques such as AIC and BIC. By comparing the AIC and BIC values of different models, we can determine which combination of columns provides the best balance between model fit and complexity (Criterion Data). This approach allows us to identify potentially relevant variables that were not included in our initial set of columns, and to refine our understanding of the relationships between the properties and critical temperature of the superconductors.

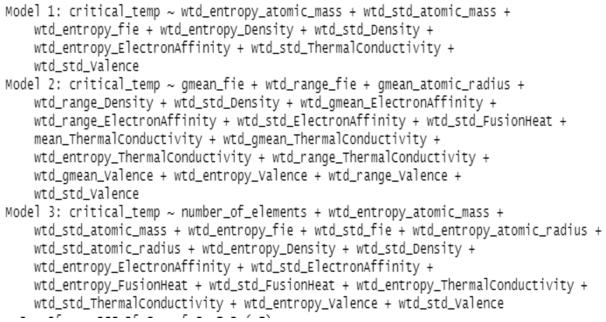


We have also created another dataset by selecting columns that do not exhibit high multicollinearity. While some properties may be causally related, such as density and atomic radius, others may not be, and including both in a model can lead to redundant information and decreased model performance. Therefore, we created a new dataset by removing one of the columns from pairs of variables that exhibit a causal relationship.

This approach ensures that our models are based on the most informative and relevant variables and can improve their predictive accuracy. (Previously we have placed a limitation stating that atleast 1 property of each kind must be present, but here we do not create any restriction of that kind). So currently we posses 3 datasets, those are, one from criterion (Criterion Data), one from no correlation between predictors (Correlation Data) & one including all the properties (Property Data)



Our research aims to develop a predictive model that accurately estimates the critical temperature of a superconductor based on its properties. Initially, we have implemented linear models for this purpose. MLR models, analyze the relationship between multiple independent variables and a dependent variable. In the context of the superconductors dataset, multilinear models can be used to predict critical temperature based on various properties, including atomic mass, density, thermal conductivity, and electron affinity. The use of these models can help gain insights into the mechanisms governing superconductivity and potentially discover new hidden relationships.



```
Model 1: critical_temp ~ wtd_entropy_atomic_mass + wtd_std_atomic_mass +
    wtd_entropy_fie + wtd_entropy_Density + wtd_std_Density +
    wtd_entropy_ElectronAffinity + wtd_std_ThermalConductivity +
    wtd_std_Valence
Model 2: critical_temp ~ gmean_fie + wtd_range_fie + gmean_atomic_radius +
    wtd_range_Density + wtd_std_Density + wtd_gmean_ElectronAffinity +
    wtd_range_ElectronAffinity + wtd_std_ElectronAffinity + wtd_std_FusionHeat +
    mean_ThermalConductivity + wtd_gmean_ThermalConductivity +
    wtd_entropy_ThermalConductivity + wtd_range_ThermalConductivity +
    wtd_gmean_Valence + wtd_entropy_Valence + wtd_range_Valence +
    wtd_std_Valence
Model 3: critical_temp ~ number_of_elements + wtd_entropy_atomic_mass +
    wtd_std_atomic_mass + wtd_entropy_fie + wtd_std_fie + wtd_entropy_atomic_radius +
    wtd_std_atomic_radius + wtd_entropy_Density + wtd_std_Density +
    wtd_entropy_ElectronAffinity + wtd_std_ElectronAffinity +
    wtd_entropy_FusionHeat + wtd_std_FusionHeat + wtd_entropy_ThermalConductivity +
    wtd_std_ThermalConductivity + wtd_entropy_Valence + wtd_std_Valence
```

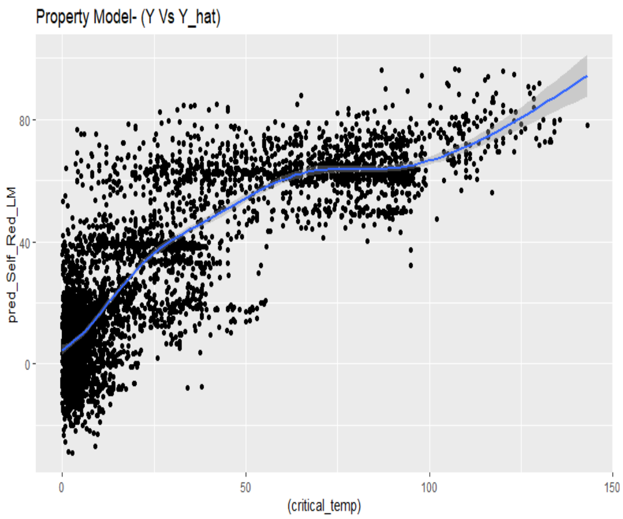
Model 1: Criterion-based Model, Model 2: Correlation Model, Model 3: Property Model.

Based on the model summaries and ANOVA results, the property model appears to be the best model for predicting the critical temperature of a superconductor. It has the lowest RMSE value of 20.03912 and the highest adjusted R-squared value of 0.6466. Additionally, all the coefficients in the property model are statistically significant at the 0.001 level. This suggests that each variable included in the model has a significant impact on the critical temperature.

The correlation model has a higher RMSE value of 21.32293 and a lower adjusted R-squared value of

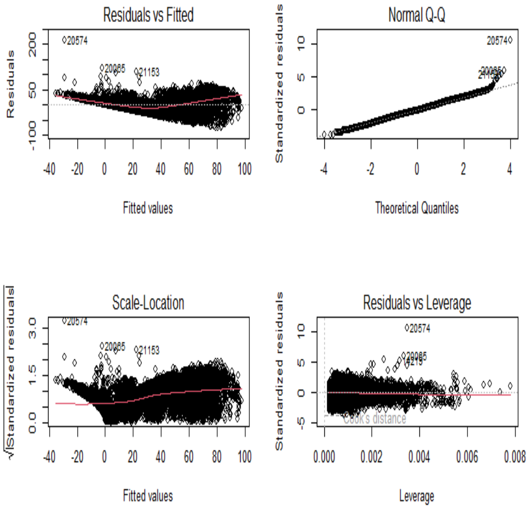
0.6058, indicating that it may be a less accurate model. However, it has a lower p-value for the ANOVA test compared to the AIC/BIC model, indicating that it may have a better fit.

The AIC/BIC model has an adjusted R-squared value of 0.624, which is better than the correlation model but lower than the property model. However, it has a higher RMSE value of 20.84799 compared to the property model.



While the property model seems to be the best model for predicting critical temperature, the choice of model should be based on the specific research question and goals of the analysis, as well as considerations such as interpretability and model complexity. It may be beneficial to explore the results of all three models and compare their predictions to select the best model for a given scenario.

Normality and Linearity Assumptions



The analysis of the model suggests that the fundamental assumptions of linearity, homoscedasticity, and normality have not been met. The scatter plot of residuals versus fitted values indicates the presence of a visible pattern, indicating that the relationship between the predictor variables and the response variable is not linear. Furthermore, the plot also shows a non-constant variance, with the spread of the residuals increasing as the fitted values increase. The normality assumption has been violated, as evidenced by the non-normal distribution of the residuals at both the upper and lower ends of the distribution.

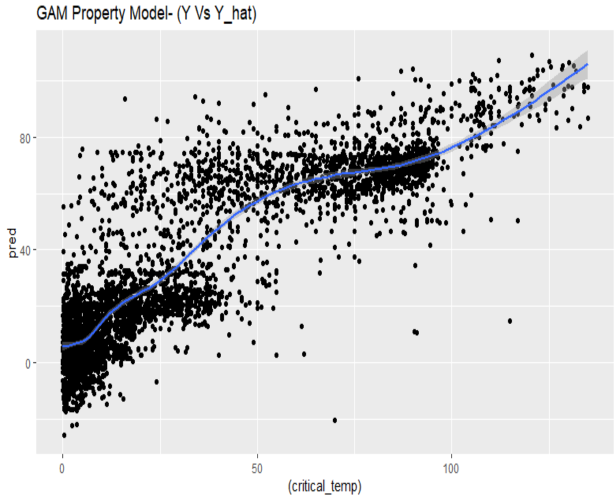
These violations suggest that the current linear model may not be an accurate representation of the relationship between the predictor variables and the response variable. In an attempt to address these issues, we applied transformations to the data, but unfortunately, they did not result in any significant improvement in the linearity, homoscedasticity, or normality assumptions.

Overall, the violations of these fundamental assumptions suggest that the current linear model may not provide an accurate representation of the relationship between the predictor variables and the response variable. It may be necessary to explore alternative modeling approaches or to consider additional data cleaning and preprocessing techniques to improve the fit and accuracy of the model.

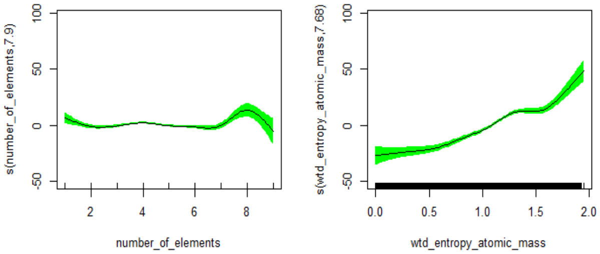
Generalized additive model

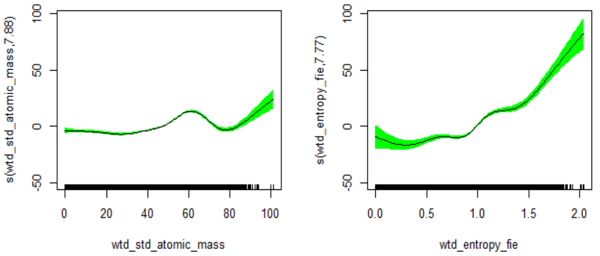
Generalized Additive Models (GAMs) are a class of statistical models that extend the capabilities of linear models by allowing for non-linear relationships between predictor variables and the response variable. As we saw from our analysis of the linear model, there were several violations of assumptions that suggest a non-linear model may be more appropriate. GAMs are a natural extension to address these issues and allow for more flexibility in modeling complex relationships.

One of the main benefits of GAMs is their ability to capture non-linear relationships between variables without specifying the exact functional form of the relationship. This allows for a more accurate representation of complex relationships in the data that may not be captured by linear models. Additionally, GAMs can be useful in situations where there are high-dimensional predictor variables, as they can handle large amounts of data without the need for feature selection or dimensionality reduction.

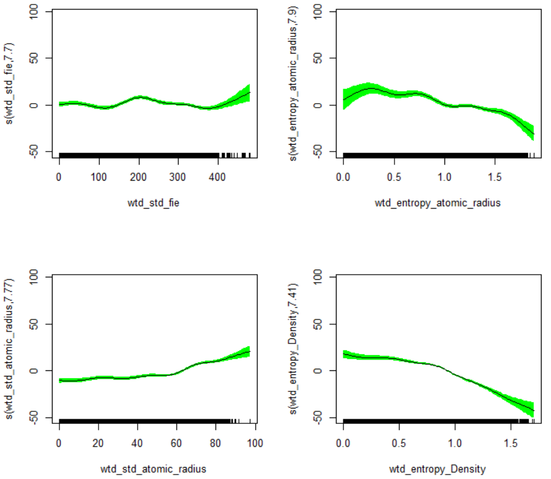


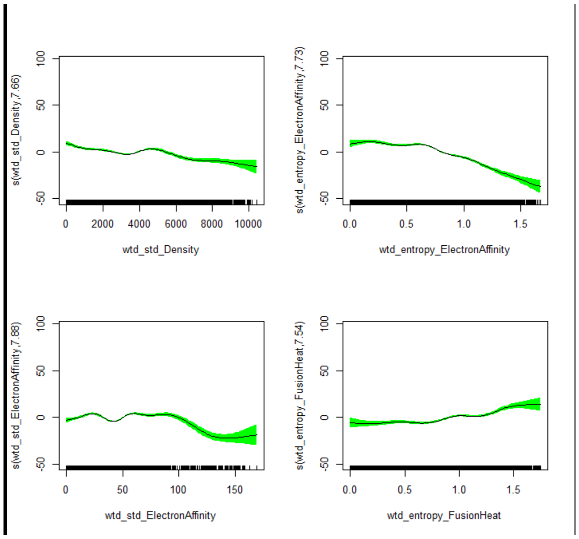
In our dataset, GAMs could be useful in identifying non-linear relationships between the superconductor properties and critical temperature that may not be captured by linear models. This could help to improve the accuracy and interpretability of the model and provide more insights into the factors that affect the critical temperature of superconductors.



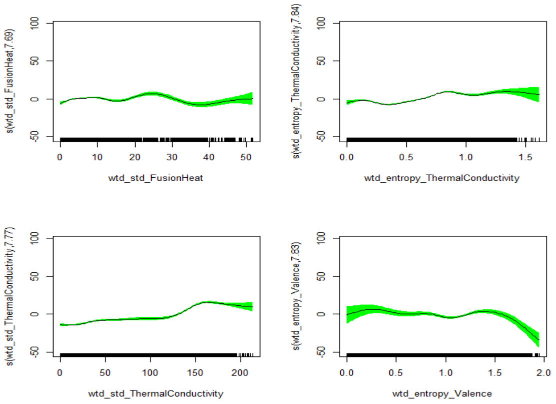


In this case, the partial effects plots suggest that the relationship between the predictor variables and the critical temperature is non-linear. This is important because it implies that a simple linear relationship between the variables may not be sufficient to capture the full complexity of the relationship. By using GAM models with smooth functions, we can account for the non-linear relationship and improve the accuracy of our predictions.





The use of GAM models allows for non-linear relationships between the predictor variables and the response variable, potentially improving the accuracy of the model, as we can see in the graph, the line is much more linear compared to Linear models.



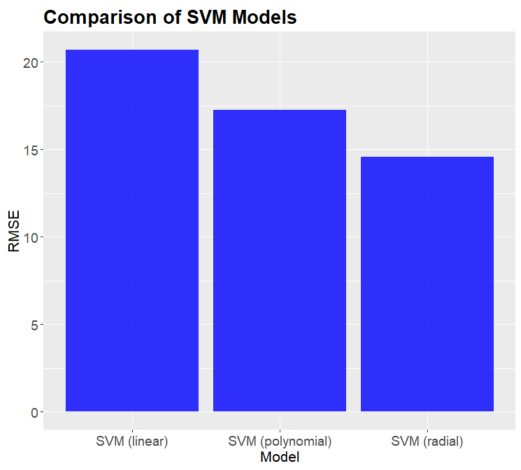
The plots presented above are called partial effects plots, which are a useful tool for visualizing the non-linear relationship between predictor variables and the response variable in a GAM model. These plots show the marginal effect of each predictor variable on the response variable while holding all other predictor variables constant. By examining these plots, we can gain insight into the nature of the relationship between each predictor variable and the response variable.

Furthermore, the partial effects plots allow us to identify the shape of the non-linear relationship between each predictor variable and the response variable. For example, we can see that the relationship between the number of elements and critical temperature appears to be roughly U-shaped, with a peak at around 10 elements. The relationship between wtd_entropy_atomic_mass and critical temperature appears to be slightly positive, while the relationship between wtd_std_atomic_mass and critical temperature appears to be negative. These insights can help us to better understand the factors that influence the critical temperature and potentially improve our ability to predict it accurately.

Based on the adjusted R-squared and deviance explained values, the GAM Property Model appears to be the best model, explaining 76.1% of the variation in the response variable. The other two models, the GAM Correlation Model and GAM Criterion Model, have adjusted R-squared values of 0.709 and 0.696, respectively.When comparing the RMSE values of the three models, the GAM Property Model also appears to have the best predictive performance with an RMSE of 16.81. However, it's worth noting that the difference in RMSE values between the Property Model and the other two models is not large, and the Correlation Model also has a relatively low RMSE value of 18.63.

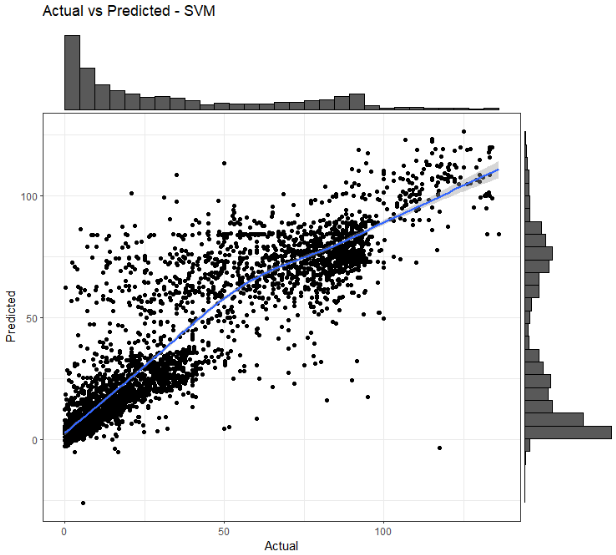
SVM

Support Vector Machine (SVM) is one of the powerful machine learning algorithms used for classification and regression modeling. In our context we can leverage SVM to identify the key features that impact the critical temperature. By analyzing the relationship between the features and the critical temperature, the SVM model can accurately predict the critical. SVM works by finding the best boundary that separates different classes of data points by maximizing the margin between the boundary and the closest points. This approach makes SVM highly effective in dealing with complex and high-dimensional datasets such as the superconductor dataset.



After evaluating multiple SVM models with different kernel functions including linear, radial, and polynomial, it can be concluded that the radial kernel outperformed the rest of the kernels in predicting the critical temperature of superconducting materials. This is supported by the fact that the radial kernel SVM model had the lowest root mean squared error (RMSE) compared to the other SVM models.

The radial kernel SVM model's performance can be attributed to its ability to effectively capture the non-linear relationships between the predictor variables and the critical temperature. This is due to the fact that the radial kernel can handle complex, non-linear relationships between variables by projecting the data into a higher-dimensional space where the relationships may be more easily separable.



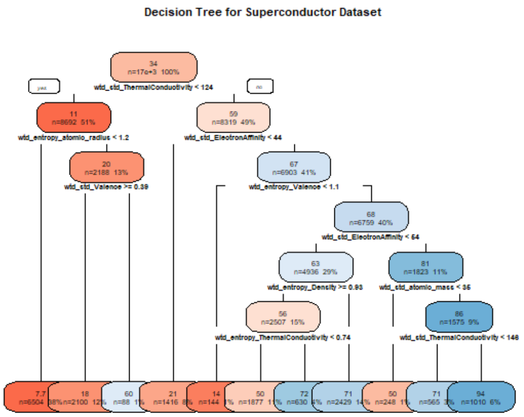
Based on the R-squared and adjusted R-squared values, the best model among these appears to be the SVM model built on the self-selected feature data. The R-squared and adjusted R-squared values for this model are 0.8253256 and 0.8251509, respectively.

The root mean squared error (RMSE) for the SVM model on the self-selected feature data is also the lowest among the three models, at 14.55494. This indicates that the model has a smaller average deviation from the actual critical temperature values.

Therefore, it can be inferred that the SVM model built on the self-selected feature data is the best model for predicting the critical temperature of superconducting materials, based on the metrics evaluated in this analysis. (Now before we have achieved this we had to hypertune too)

Decision Tree

Decision tree is a popular supervised learning algorithm that can be used for both regression and classification tasks. It involves partitioning the dataset into subsets based on the values of one of the input features and creating a tree-like model of decisions that lead to the prediction of a target variable. In the context of superconductor dataset, decision tree can be applied to identify the important features that impact the critical temperature of superconducting materials.



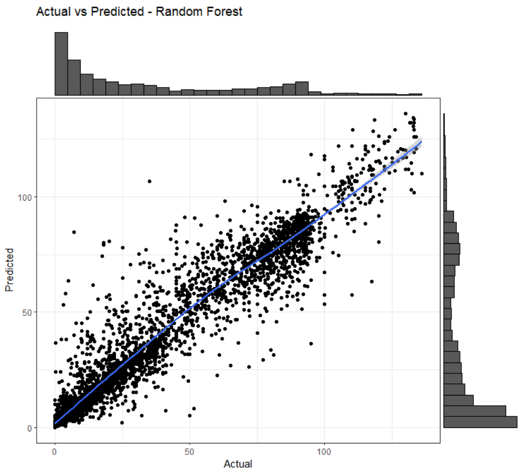
The decision tree model built on the self-selected feature data has an R-squared value of 0.7393893 and an adjusted R-squared value of 0.7391285, indicating that the model explains around 74% of the variance in the critical temperature. The root mean squared error (RMSE) for this model is 17.51653, indicating that the model has an average deviation of 17.52 from the actual critical temperature values.

Compared to the other decision tree models built on the correlation and AIC-selected features, this model has the highest R-squared and adjusted R-squared values. However, the RMSE for this model is also the highest among the three decision tree models.

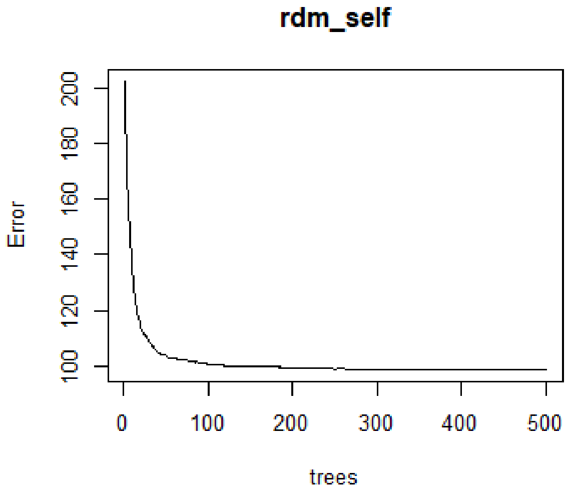
While the decision tree model may not perform SVM models, it can still provide insights into the important features for predicting critical temperature. Additionally, decision trees have the advantage of being easily interpretable, which can be valuable in applications where interpretability is important.

Random-Forest Regressor

Random forest regressor is an ensemble learning algorithm used for regression analysis. It works by constructing multiple decision trees during training and outputting the mean of the predicted values by individual trees as the final prediction.



After analyzing the R-squared, adjusted R-squared, and RMSE values of the three models, it can be concluded that the random forest model constructed using self-selected features is the most suitable for predicting the critical temperature of superconducting materials. The R-squared and adjusted R-squared values for this model are the highest among the three models, at 0.9226079 and 0.9225305, respectively. Moreover, the root mean squared error (RMSE) for this model is the lowest, at 9.55842, indicating that the model has a smaller average deviation from the actual critical temperature values.

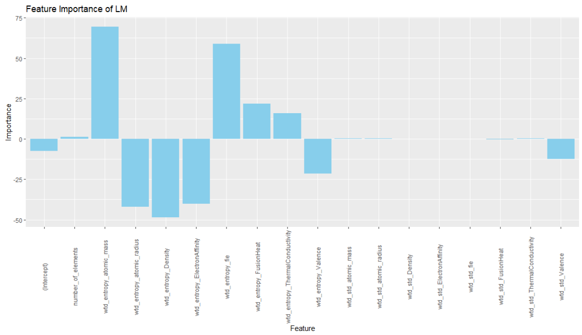


We have executed the random forest algorithm with 500 trees, and the plot shows the relationship between the number of trees and the error rate. As the number of trees increases, the error rate progressively decreases.

VI. MODEL COMAPRISION & FEATURE IMPORTANCE

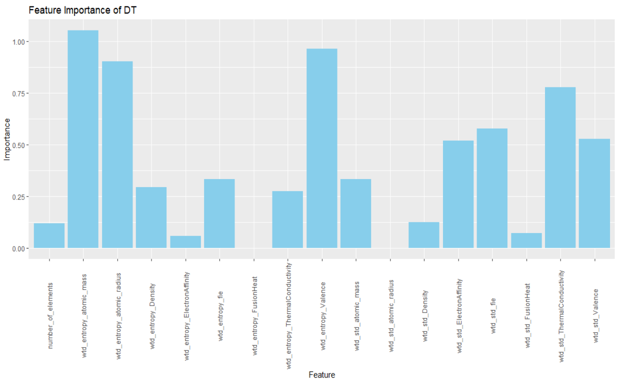
After implementing various machine learning models such as linear regression, random forests, GAMs, and SVMs, we can now analyze these models to understand the importance of different features in predicting the critical temperature of superconductors.

For instance, in a linear regression model, we can determine the impact of each parameter on the rise or fall of critical temperatures. By examining the coefficients of the linear model, we can identify which parameters have a significant positive or negative effect on the critical temperature, and by how much.

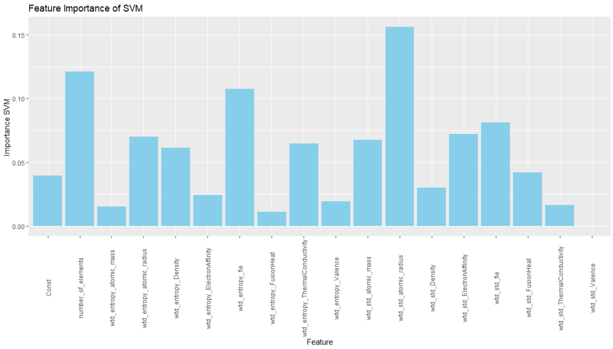


Through the best performing linear model we developed, we can find that several features have a significant impact on predicting the critical temperature of superconductors. These features include **Weighted Entropy Atomic Mass**, **Weighted Entropy Atomic Radius**, **Weighted Entropy Density**, **Weighted Entropy Electron Affinity**, and **Entropy Range FIE**. The reason why these specific columns are important for predicting the critical temperature may vary depending on the physical and chemical properties of superconductors.

For instance, the atomic mass and atomic radius of elements can influence the electron configuration and the bonding properties of the material, which in turn can affect the critical temperature. The density of the material can also play a significant role, as it can affect the phonon vibrations and the electronic structure of the material. Electron affinity, on the other hand, is a measure of how tightly an atom attracts electrons, which can affect the behavior of the material at the atomic level. Finally, the fusion heat and thermal conductivity can provide information about the heat transfer properties of the material, which can affect the ability of the material to maintain superconductivity.

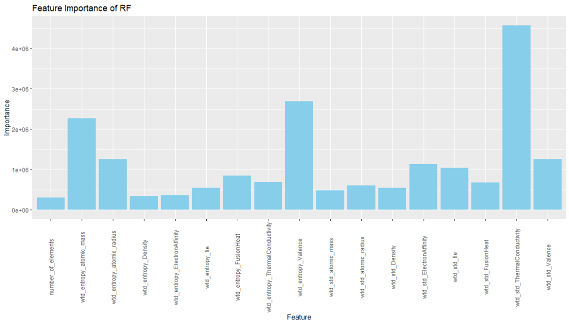


From the best performing Decision Tree model we can find that the features that have a significant impact on predicting the critical temperature of superconductors include, **Weighted Entropy Atomic Mass**, **Weighted Entropy Atomic Radius**, **Weighted Entropy Valence**, & **Weighted Standard Thermal Conductivity**. We have already seen why atomic mass & atomic radius are a good impact on critical temperature, but valence, on the other hand, is a measure of the number of electrons in the outermost shell of an atom, which can influence the bonding properties and the electron configuration of the material. Thermal conductivity is a measure of the material's ability to conduct heat, which is important for maintaining superconductivity in the material.



On analyzing the best performing Support Vector machine we can determine, that the features, the **Number of Elements**, **Weighted Entropy Atomic Radius, Weighted Entropy of the First Ionization Energy**, & **Weighted Entropy Thermal Conductivity**.

We have already seen all the features apart from the FIE, this feature can be useful in determining the critical temperature of superconductors because FIE is an important factor in determining the electronic structure and bonding properties of materials. The ability of a material to form strong and stable bonds can impact its superconducting properties. Furthermore, the range and distribution of FIE values across the elements in the material can affect the phonon vibrations and the electronic structure of the material, which in turn can impact its superconducting properties.



On analyzing the best performing Random Forest model (also the best performing model), we can determine, that the features, **Weighted Entropy Atomic Radius**, **Weighted Entropy Valence**, **Weighted Entropy Atomic Mass**, & **Weighted Standard Thermal Conductivity**.

Now from all the models we can see that the most important features are, **Weighted Entropy Atomic Radius, Weighted Entropy Atomic Mass, Weighted Standard Thermal Conductivity, & Weighted Entropy Valence.** Based on both importance of itself and model importance, these features can be considered that these parameters are crucial in determining the critical temperature of a superconductor. In addition to the features mentioned previously, the **Number of Elements** present in the material can also be considered an important feature for predicting the critical temperature of superconductors. In the superconductor's dataset, most of the features are weighted and standardized, which means they are normalized based on the composition of the material. However, the number of elements present in the material is a raw count that does not depend on the weights or standards used.

The number of elements can be important because it reflects the complexity of the material's composition and can impact its superconducting behavior. As the number of elements increases, the interactions between them become more complex, and this can affect the phonon vibrations, electron configuration, and bonding properties of the material.

Among the linear regression, random forest, GAM, SVM, and decision tree models, the random forest model built on the self-selected feature data outperforms the others, with an R-squared value of 0.9226 and an adjusted R-squared value of 0.9225. The root mean squared error (RMSE) for this model is 9.558, indicating that it has a smaller average deviation from the actual critical temperature values.

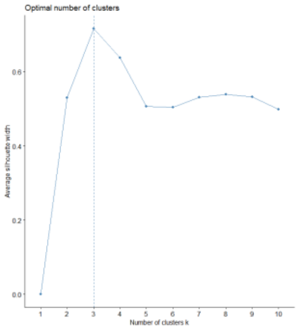
The SVM model with a radial kernel also performs well, with an RMSE of 14.231 and an R-squared value of 0.8303.

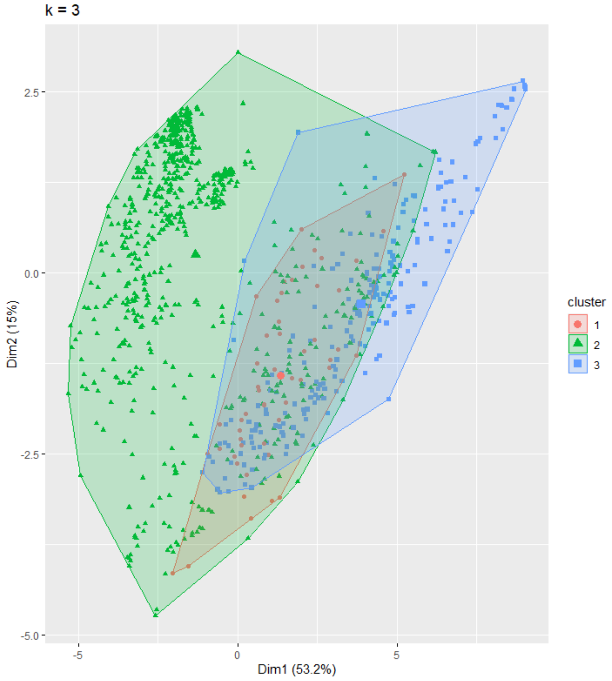
The decision tree model has an RMSE of 15.04 and an R-squared value of 0.7454.

Overall, the random forest model stands out as the best-performing model for predicting critical temperature, followed by the SVM with a radial kernel.

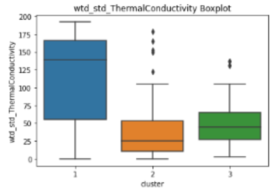
VII. FEATURE IMPORTANCE & CLUSTERING

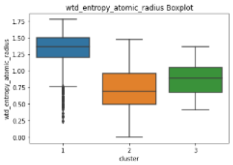
We have also performed cluster analysis to check if similar compounds are grouped together, and further verified this using the important features we have determined in the previous section. The optimal number of clusters were 3

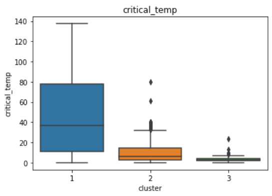


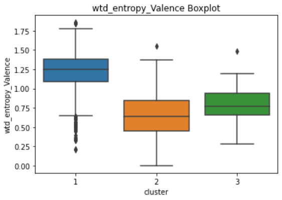


We can see that while some elements coincide within cluster 2, 3 (both in box plots & Cluster Viz) the first cluster has divided most of the elements based on given properties. Hence clustering can also be used to group similar compounds and that cluster data can be used for much more precise critical temperature prediction









VIII. CONCLUSION

This research has been an excellent example of how statistical analysis can be used to gain valuable insights into complex phenomena such as superconductivity. Through the implementation of various machine learning models and the analysis of the superconductor's dataset, we have identified the most important features that impact the critical temperature of superconducting materials. Our analysis has demonstrated the importance of understanding the physical and chemical properties of materials in predicting their superconducting behavior, and the potential of machine learning models to provide insights into this complex phenomenon.

We have generated three datasets using three different techniques, namely Criterion Data, Property Data, and Correlation Data. The analysis of these datasets using various machine learning models consistently demonstrates that the Property Data yields the lowest root mean square percentage error (RMSE). This finding suggests that all the properties included in the Property Data are essential for accurate predictions of the critical temperature, and using only the most highly correlated features may not be sufficient.

Although the random forest model is considered the best model for predicting the critical temperature of superconducting materials, the other models have also provided crucial evidence. Notably, the feature importance analysis from all models consistently highlights the significance of properties such as weighted entropy atomic mass, weighted entropy atomic radius, weighted entropy valence, weighted standard thermal conductivity, and weighted entropy first ionization energy in predicting the critical temperature.

Initially, we employed the Linear Regression model for predicting the critical temperature, however, it produced suboptimal results due to the model's inability to account for the non-linear relationship between the predictors and the target variable. Consequently, the predictions generated had significant errors. To overcome this limitation, we explored Additive Models, which exhibited better performance than Linear Regression. Nevertheless, the presence of a vast number of columns in the dataset made it difficult to handle the smoothing function of Additive Models automatically. Despite this challenge, the Mean Squared Prediction Error (MSPE) substantially decreased.

Given the non-linear nature of the data, we explored the use of Support Vector Machines with different kernels to improve the accuracy of our predictions. The radial kernel proved to be the most effective, yielding a lower RMSE compared to previous models. However, due to the high number of features in our dataset, we also considered the use of tree-based models such as random forests and decision trees to further improve the accuracy of our predictions. The random forest regressor proved to be the better than both SVR & Decision Tree, in term if RMSE.

By analyzing the feature importance of various machine learning models we have identified the key features that have a significant impact on predicting the critical temperature of superconducting materials. These features include weighted entropy atomic mass, weighted entropy atomic radius, weighted entropy valence, and weighted entropy first ionization energy. These findings have provided valuable insights into the relationship between material composition and superconducting properties, which can be further utilized in designing and developing new superconducting materials with improved critical temperatures.

Stressing the importance of domain knowledge is one future strategy that might be used in superconductors research. The evaluation of massive datasets using machine learning approaches is efficient, but they cannot replace a fundamental understanding of the physical mechanisms that cause superconductors to reach their critical temperature. We must improve models to include this knowledge in order to make them more precise and intelligible.

In addition to this, there is a need to work more on GAM models and identify important features that impact the critical temperature. Another potential area of exploration is the application of unsupervised learning techniques such as clustering and dimensionality reduction to gain a deeper understanding of the relationships between variables in the superconductor dataset

REFERENCES

[1] Machine learning identifies chemical properties related to superconductivity" by Faber et al. (2016)

[2] Stanev, Valentin, et al. "Machine Learning Modeling of Superconducting Critical Temperature." Npj Computational Materials,

vol. 4, no. 1, 28 June 2018, pp. 1–14, www.nature.com/articles/s41524-018-0085-8,

[3] Bulut, Okan, and Christopher Desjardins. 7 Supervised Machine Learning - Part I | Exploring, Visualizing, and Modeling Big Data with R. Okanbulut.github.io, okanbulut.github.io/bigdata/supervised-machine-learning---part-i.html#decision-trees. Accessed 9 May 2023.

[4] Gates Bolton Analytics – Data Science and Analytics > Training and Consulting. gatesboltonanalytics.com/. Accessed 9 May 2023.

[5] https://www.wikipedia.org/

[6] "RPubs." Rpubs.com, rpubs.com/.