TEAM DEANALYZERS

INDEX

1. INTRODUCTION

1.1 PROBLEM STATEMENT

The given data is from a renowned material science laboratory, recording data of 21,263 superconductors with 82 features. These features affect the Critical Temperature of a metal - the temperature at which a material transitions from a non-superconducting state to a superconducting state. Here, we identify the most important parameters affecting Tc and develop a predictive model that can accurately predict Tc based on these parameters. The available files are:

- **Train.txt**: Contains 21,263 samples; each sample records 82 features for each samples.

1.2 DATA DESCRIPTION

The given dataset consists of 21,263 rows and 82 columns. The following features have been provided in different ways in the training data set for every sample:

- No. of elements
- Atomic Mass
- fie
- Atomic Radius
- Density
- Electron Affinity
- Heat of Fusion
- Thermal Conductivity
- Valency

We have been given weighted mean, geometric mean, entropy, range and standard deviation for each of the above properties.

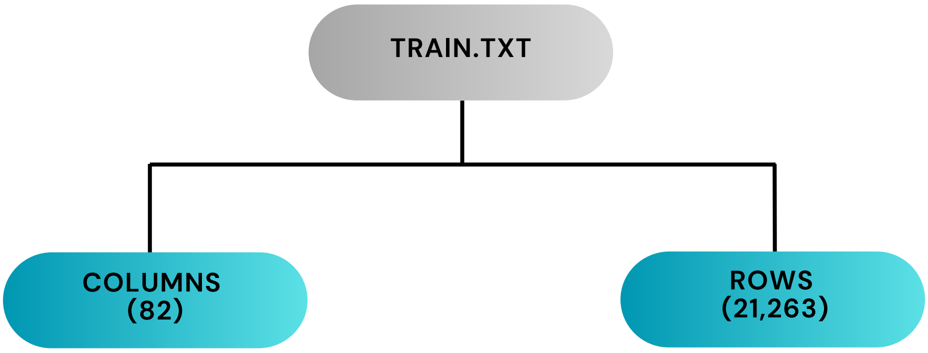


Figure 1: *Flowchart depicting the data distribution*

HIGH LEVEL FLOW

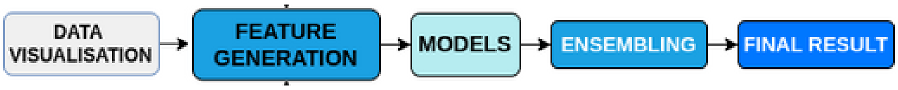


Figure 2: *Approach Flow Chart*

2. DATA ANALYSIS AND VISUALIZATION

This section visualizes the different features affecting Critical Temperature using different charts and plots to identify patterns, trends, and relationships that may not be apparent when looking at raw data.

2.1 CORRELATION

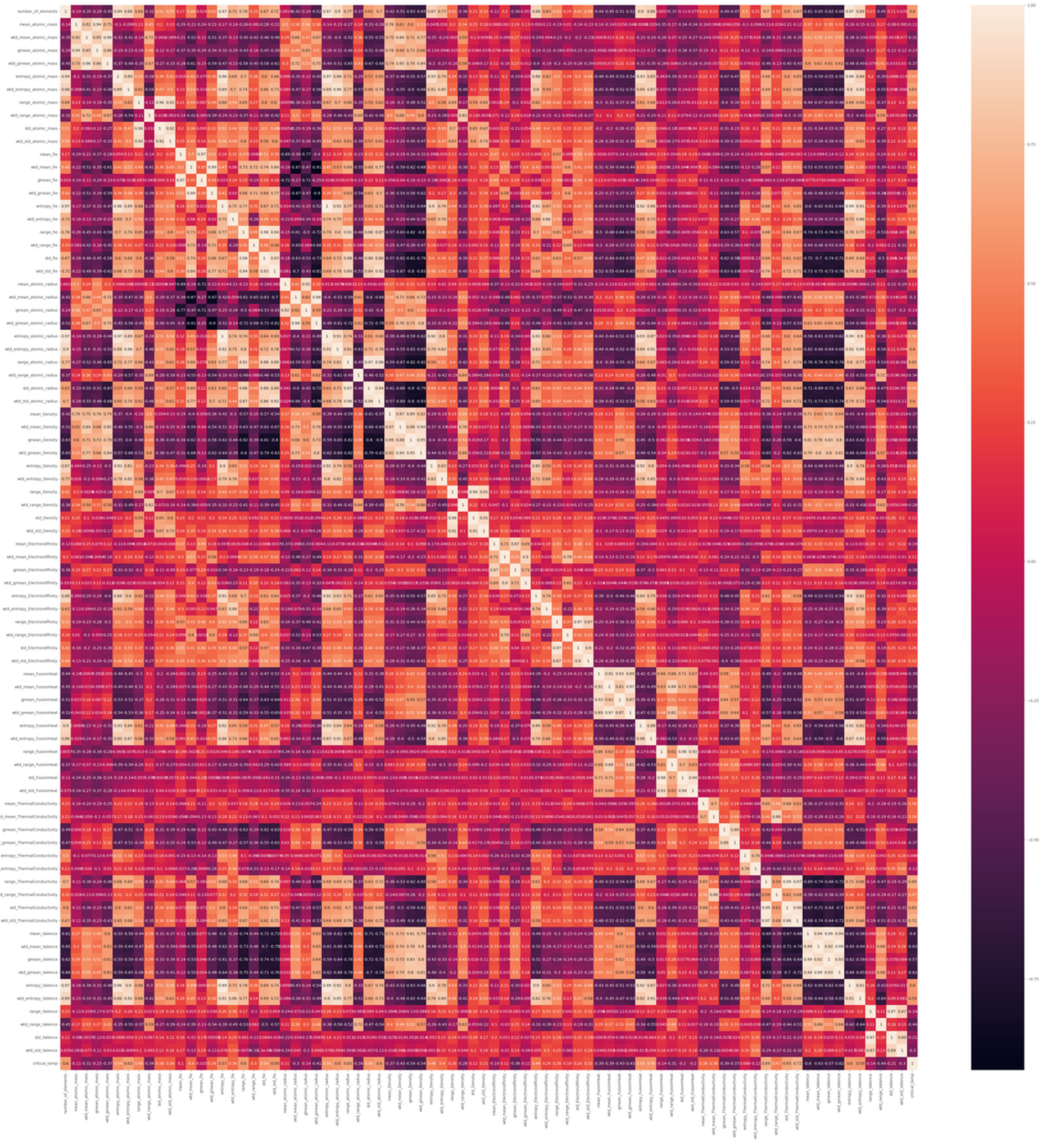


Figure 3: *Heatmap of Correlation of given features*

From the above heatmap we can see Tc has **High Positive Correlatio**n with the following features:

- Number of Elements (0.60)
- Entropy_atomic_mass (0.54)
- wtd_entropy_atomic mass (0.63)
- range_atomic_mass (0.49)
- entropy_fie (0.57)
- range_fie (0.6)
- std_fie (0.54)
- wtd_std_fie (0.58)
- entropy_atomic_radius (0.56)
- wtd_entropy_atomic_radius (0.6)
- range_atomic_radius (0.65)
- std_atomic_radius (0.56)
- wtd_std_atomic_radius (0.60)
- entropy_fusion_heat (0.55)
- wtd_entropy_fusion_heat (0.56)
- range_thermal_conductivity (0.69)
- std_thermal_conductivity (0.65)
- wtd_std_thermal_conductivty (0.72)
- entropy_valence (0.56)
- wtd_entropy_valence (0.6)

And with Following features Tc has **Strong negative Correlation**:

- gmean_density (-0.54)
- wtd_gmean_density (-0.54)
- mean_valence (-0.60)
- wtd_mean_valence (-0.63)
- gmean_valence (-0.57)
- wtd_gmean_valence (-0.62)

Comment: This dataset is highly correlated we might not be having any major problems developing a model for this data!

2.2 DISTRIBUTION PLOTS

Here we try to visualize the range of critical temperature, it's distribution and outliers. Most of the samples have Tc in the range of 0 - 100, with majority towards the lower range (0 - 25).

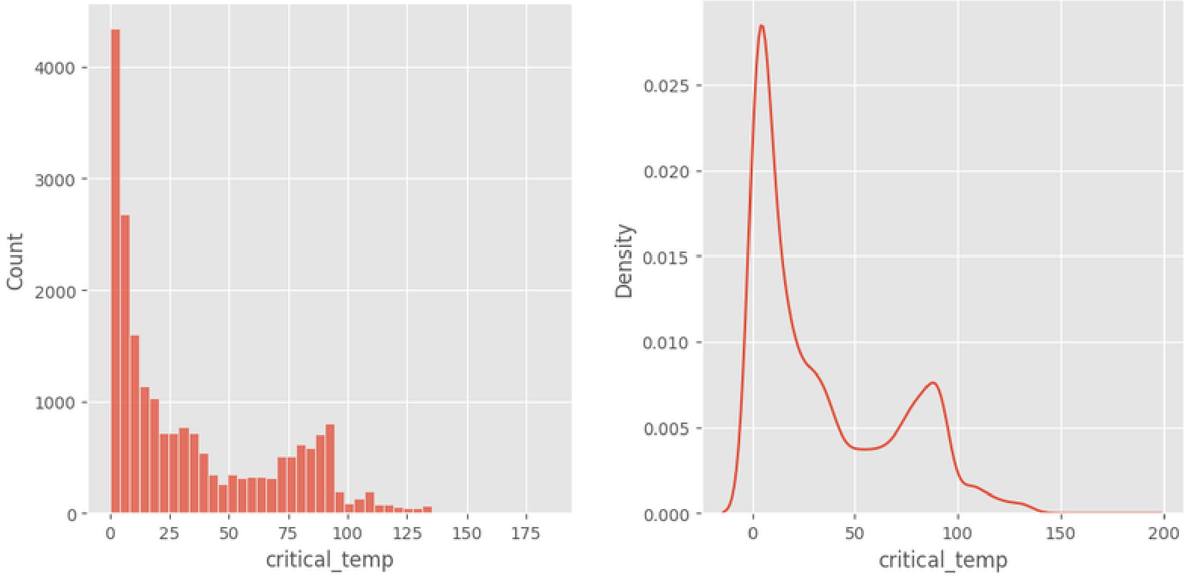


Figure 4: *Distribution Plots*

2.3 SCATTER PLOT

The scatter plot gives the distribution of sample's temperature with indices. The plot is denser for indices < 10,000 and Tc < 100.

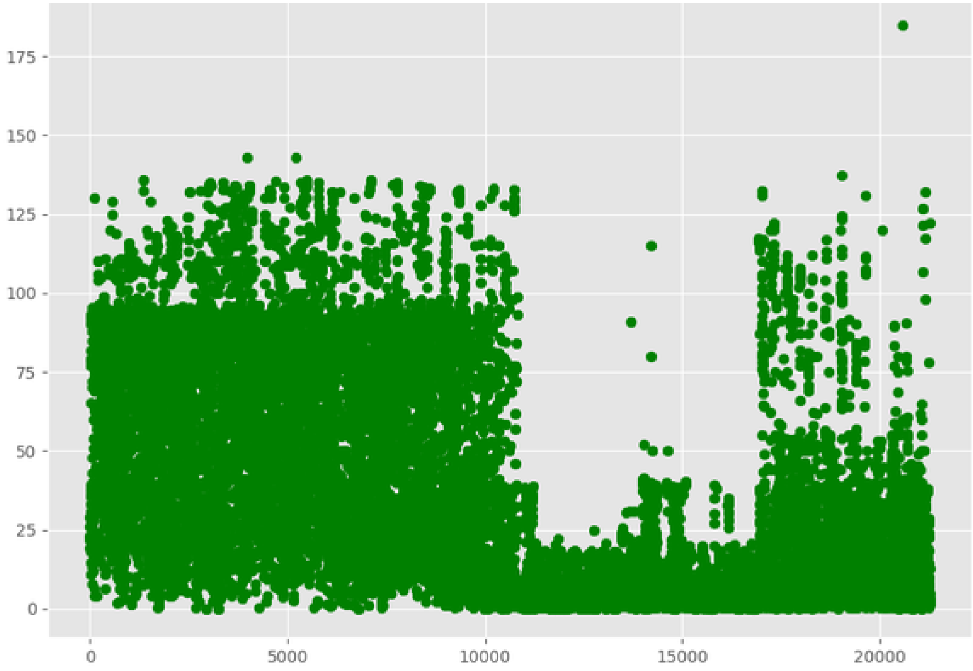


Figure 5: *Scatter Plot for sample temperature*

2.4 2D VISUALIZATION OF DATASET

We reduce the dimension of original dataset in TWO dimensions and analyze the scatter plot for the same.

This looks like Sample with higher Tc values tends to have low values for V1 (first principal component) and V2 (second principal component) with exceptions. Let's check this inference with **regplots**.

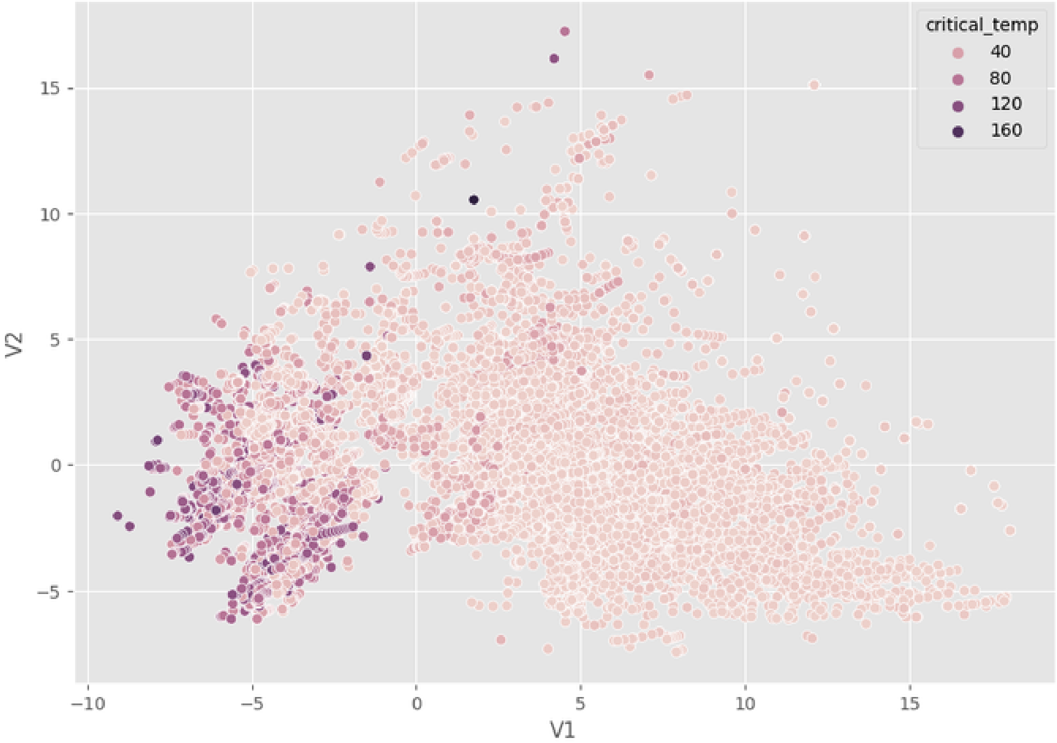


Figure 6: *Scatter Plot with reduced dimension*

This gives a better visualization of Tc with V1 and V2 (the new dimensions in which data was reduced).

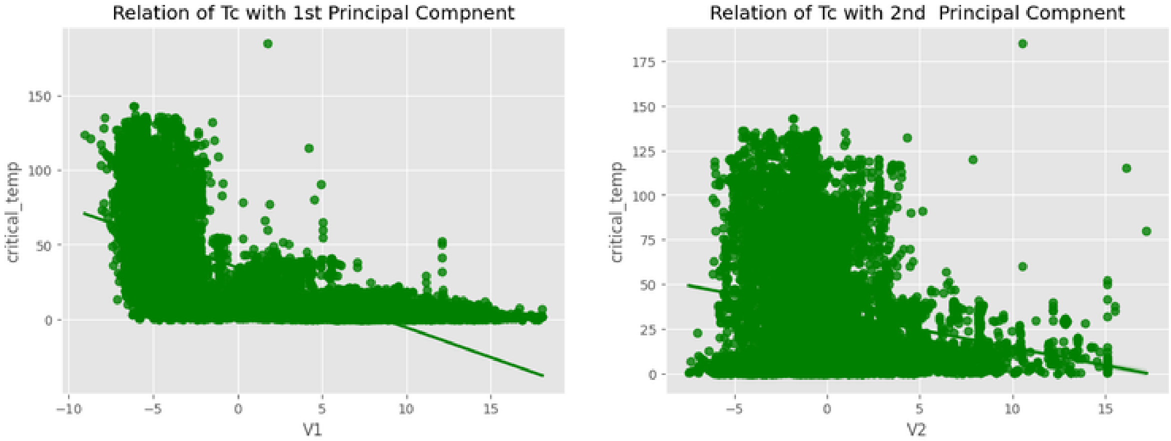
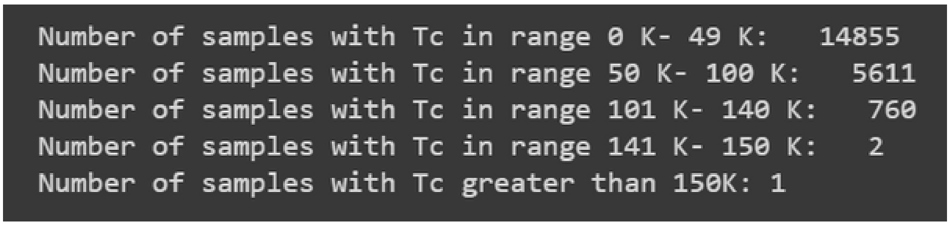


Figure 7: *Regression Plots in new dimensions*

2.5 FURTHER ANALYSIS

We run some pandas queries and get some information as:

- A majority of the samples in our data set has very low Tc (14855 samples to be precise).
- 5611 samples have Tc between 50K - 100K .
- 760 samples have Tc between 101K - 140 K.
- 2 samples have Tc between 141K - 150K.
- only 1 sample has Tc greater than 150K.



```
Number of samples with Tc in range 0 K- 49 K:   14855
Number of samples with Tc in range 50 K- 100 K:   5611
Number of samples with Tc in range 101 K- 140 K:   760
Number of samples with Tc in range 141 K- 150 K:   2
Number of samples with Tc greater than 150K: 1
```

Figure 7: *Pandas Queries*

3. MAKING PREDICTIONS

5.1 MACHINE LEARNING MODEL STATISTICS

Model	Root Mean Square Error
Logistic Regression	17.38
Decision Tree Regressor	11.53
Gradient Boosting Regressor	12.33
Random Forest Regressor	8.97
Bagging Regressor	9.51
MLP Regressor	10.07

5.2 ENSEMBLING

Bagging Regressor + Random Forest Regressor

RMSE = 9.1134

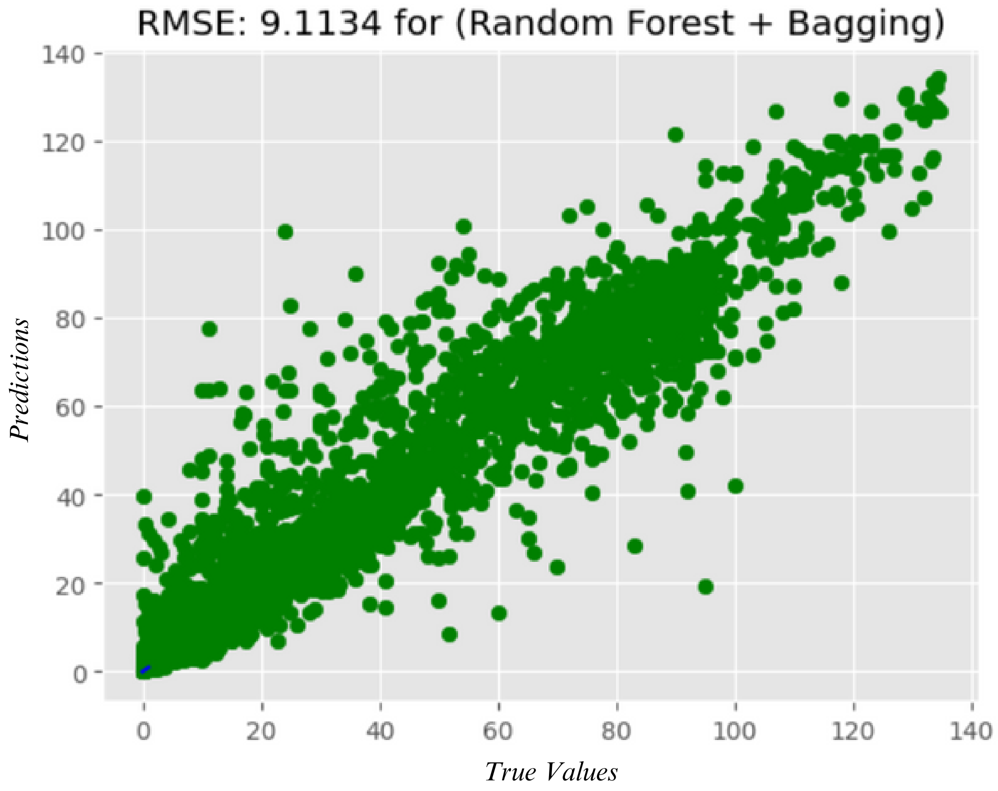


Figure 7: *Predictions of Ensembled Model*

4. CONCLUSION

- The goal of the competition is to build a model will predict a **critical temperature** for any sample based on the given features.

- It was observed from exploratory data analysis that there were no errors or empty values in the dataset. Hence, we were able to use maximum amount of available information in the prediction.

- As per the results obtained, minimum root mean square error of **8.97** (mse = 80.46) was achieved using Random Forest Regressor. On performing **ensembling** of Random Forest and Bagging Regressor, an accuracy of **~9.11 rmse** was achieved.

Link to the Notebook:
https://drive.google.com/file/d/1N-dZhl9IjTP5BqDPZG5sYGCRXe9cUicS/view?usp=sharing