Semiconductor Critical Temperature Prediction using Statistical Analysis

Data File: train.csv, unique.csv

The Project is mainly about the semiconductor critical temperature prediction. The critical temperature is a temperature at which the semiconductor conducts with zero resistance. Hence it is very important to predict the if not very correct but at least the nearest temperature for any given semiconductor. The scope of this project is to perform the analysis on the underlying data and finally make predictions on the same data using the statistical methods.

Dataset and Descriptive Analysis

The following dataset consists of 2 components namely train.csv and unique.csv. The train.csv purely consists of the properties and statistics that affect the critical temperature. Finally, the unique.csv corresponds to the chemical formula of the semiconductor material. Both the datasets include 21,263 entries but differ by number of predictor variables in them i.e., the train.csv includes 82 features whereas the unique.csv consists of 88 features. Below attached is the snapshot of the descriptive analysis of the given data. The descriptive analysis includes the count, mean, standard deviation, interquartile range of the dataset.

The first dataset fully includes the numerical values and are pertaining to the physical or the chemical properties of the semiconducting material. Hence the descriptive analysis includes slightly higher values as compared to the second dataset, which mostly talks about the count of the chemical elements present in the final semiconductor material. Fig 1. and Fig 2. majorly describe the data in statistical terms.

	count	mean	std	min	25%	50%	75%	max
number_of_elements	21263.0	4.115224	1.439295	1.000000	3.000000	4.000000	5.000000	9.0000
mean_atomic_mass	21263.0	87.557631	29.676497	6.941000	72.458076	84.922750	100.404410	208.9804
wtd_mean_atomic_mass	21263.0	72.988310	33.490406	6.423452	52.143839	60.696571	86.103540	208.9804
gmean_atomic_mass	21263.0	71.290627	31.030272	5.320573	58.041225	66.361592	78.116681	208.9804
wtd_gmean_atomic_mass	21263.0	58.539916	36.651067	1.960849	35.248990	39.918385	73.113234	208.9804
...	...	...	...	...	...	...	...	...
range_Valence	21263.0	2.041010	1.242345	0.000000	1.000000	2.000000	3.000000	6.0000
wtd_range_Valence	21263.0	1.483007	0.978176	0.000000	0.921454	1.063077	1.918400	6.9922
std_Valence	21263.0	0.839342	0.484676	0.000000	0.451754	0.800000	1.200000	3.0000
wtd_std_Valence	21263.0	0.673987	0.455580	0.000000	0.306892	0.500000	1.020436	3.0000
critical_temp	21263.0	34.421219	34.254362	0.000210	5.365000	20.000000	63.000000	185.0000

Fig 1. Description of the first dataset

Data Analysis

In the previous section, a brief overview of the dataset was given. Going ahead in the direction of data modelling it was very important to thoroughly visualize the data and their relation with the respone variable.

	count	mean	std	min	25%	50%	75%	max
H	21263.0	0.017685	0.267220	0.0	0.0	0.0	0.0	14.0
He	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
Li	21263.0	0.012125	0.129552	0.0	0.0	0.0	0.0	3.0
Be	21263.0	0.034638	0.848541	0.0	0.0	0.0	0.0	40.0
B	21263.0	0.142594	1.044486	0.0	0.0	0.0	0.0	105.0
...	...	...	...	...	...	...	...	...
Pb	21263.0	0.042461	0.274365	0.0	0.0	0.0	0.0	19.0
Bi	21263.0	0.201009	0.655927	0.0	0.0	0.0	0.0	14.0
Po	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
At	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
Rn	21263.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Fig 2. Description of the Second dataset

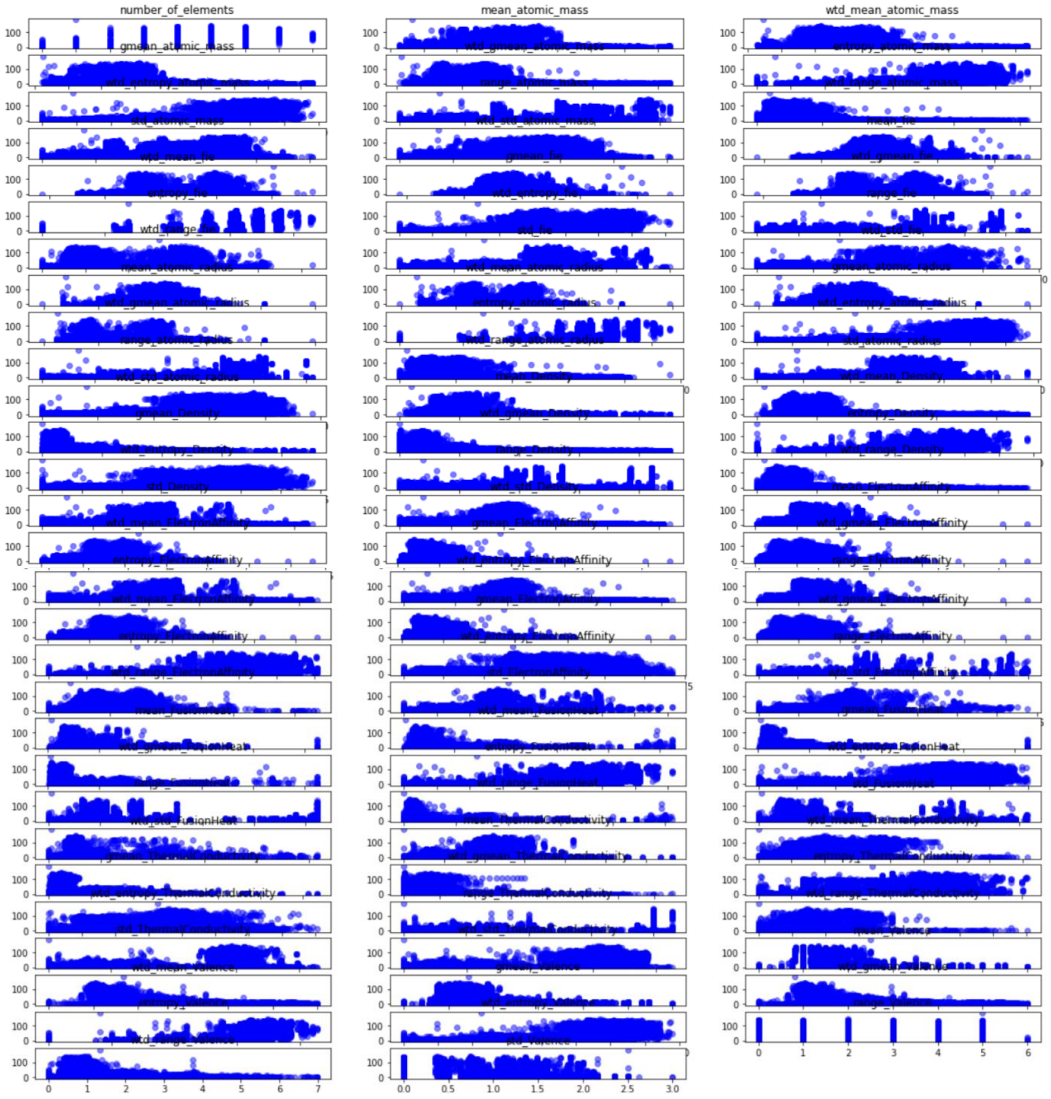

Fig 2. Scatter plot to understand the relation between the predictor variables and the response variable

Various methods were utilised to visualize the data, which include scatter plot shown in the Fig 2. The Fig 2. corresponds to the first dataset and was plotted w.r.t to the predictor variables vs response variables.

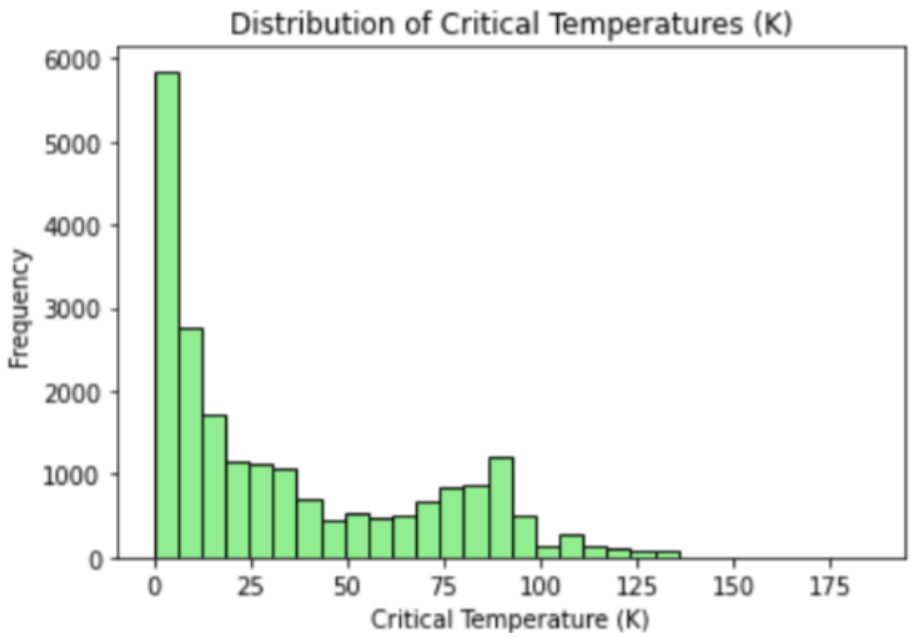

Fig 3. Distribution of the critical temperature using histograms

In the direction of undersatnding the data, it was important to understand the distribution of the critical temperature. From the above Fig 3. It is clear the distribution is more biaed towards the 0K cases as compred to the varied range of critical temperatures. If it were a classification, it would have meant that the model would be more inclined towards the 0K class and hence would be inefficient in predicting other classes.

As we already understand the importance of variance, variance analysis was performed to understand the variance in the predictor variables. It meant that the predictor variables would contain more information towards prediction as compared to the once with less variance. Fig 4. and Fig 5. Depict the variance present in the dataset.

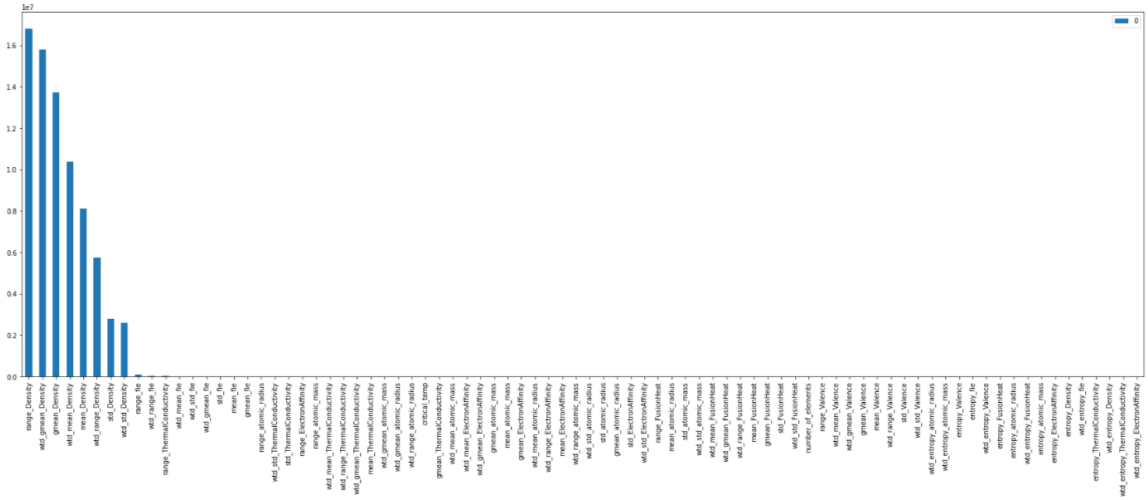

Fig 4. variance in the first dataset

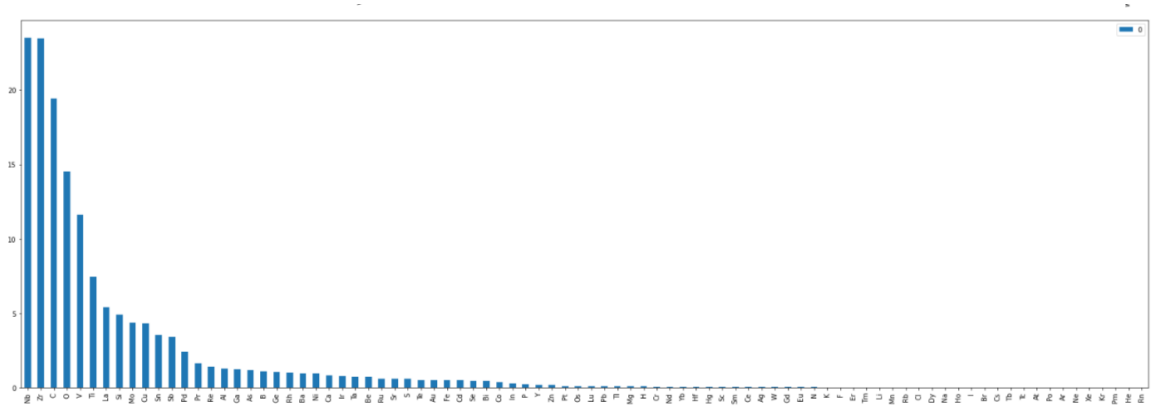

Fig 5.Variance in the second dataset

As we understand the principle components. A simple analysis was done to visualize the principle components in the dataset using the Eigen values and Eigen vectors. The theory states that the components corresponding to the Eigen vector with larger eigen value are nothing but the principle components. The eigen values obtained were normalized to make all the components visible in the graph in the sorted order.

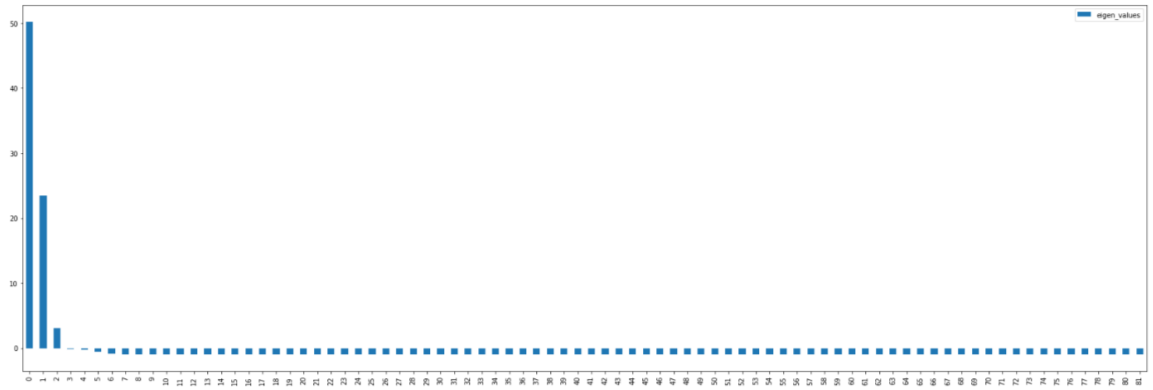

Fig 6. Principle components in the first dataset

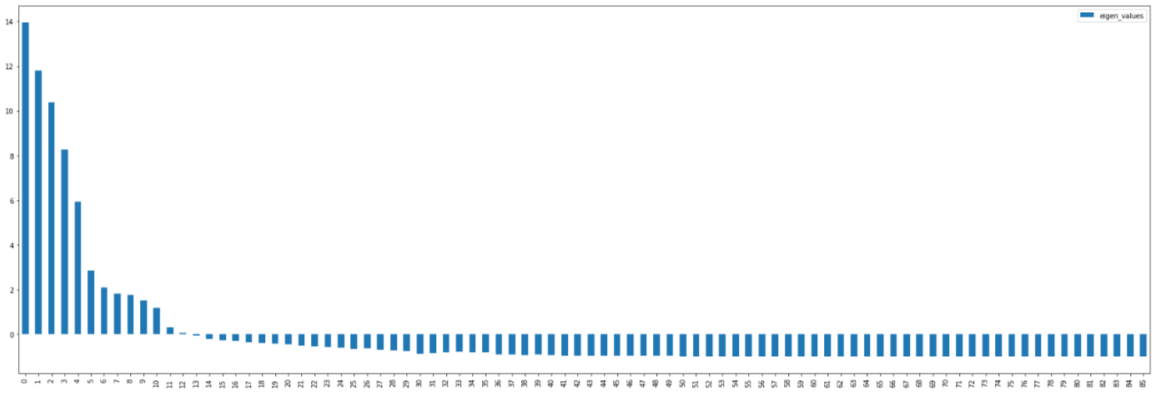

Fig 7. Principle components in the Second dataset

Last but not the least a final visualization was performed to visualize the correlation between the critical temperature and the other predictor variables. This visulization has been provided in the form of doublesided bar plot and heat maps from Fig 8. – Fig 11.

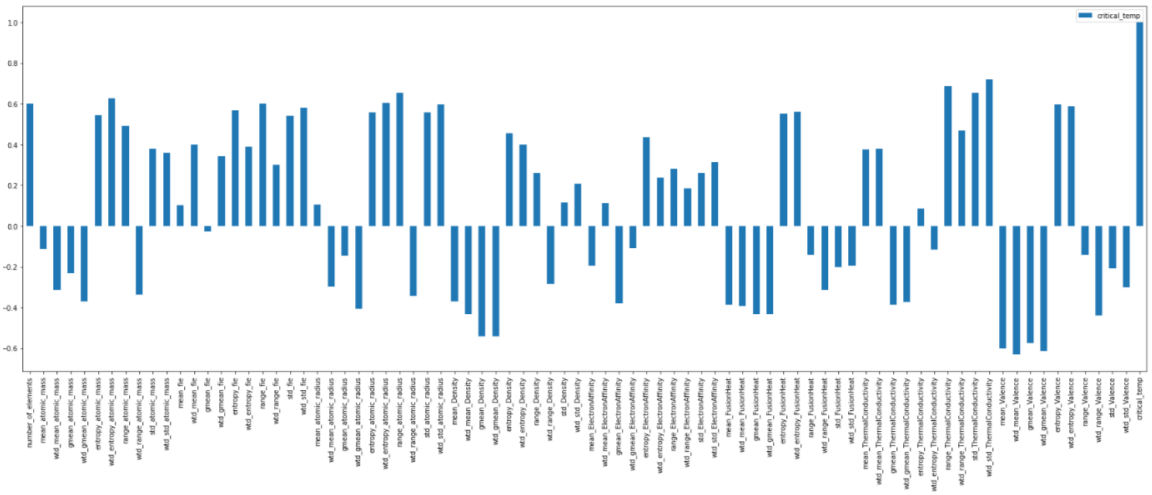

Fig 8. Correlation between the Predictor variables and the response variable (First dataset)

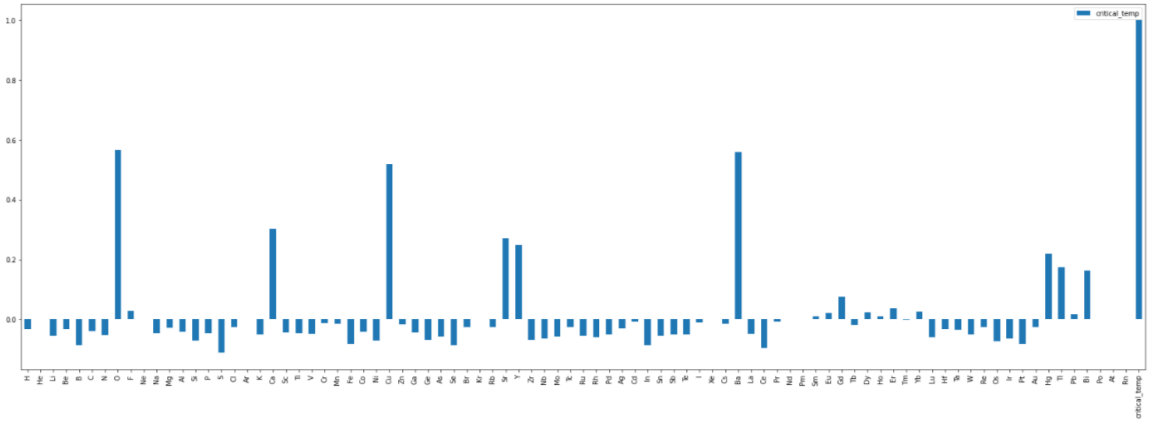

Fig 9. Correlation between the predictor variables and the response variables (second Dataset)

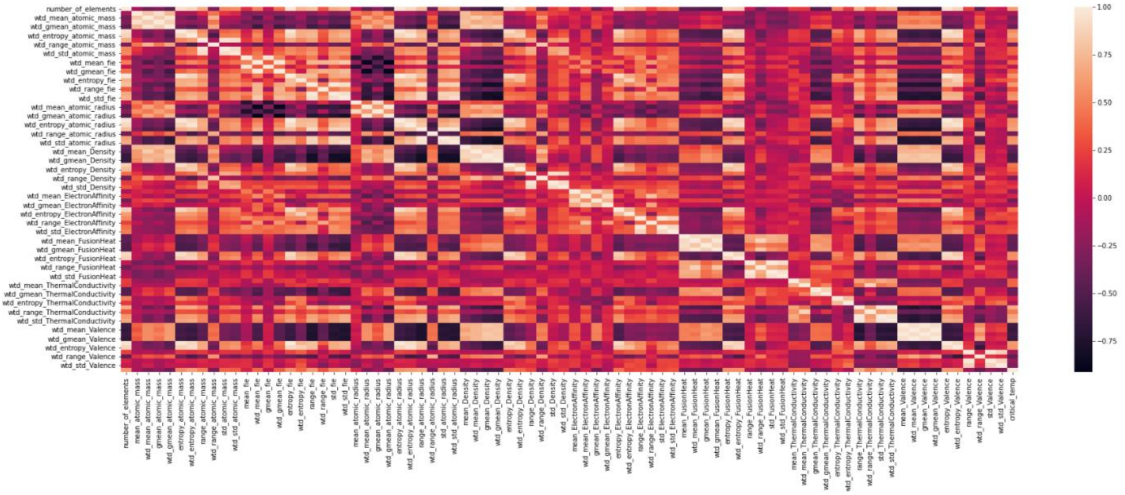

Fig 10. Heatmap to understand the correlation

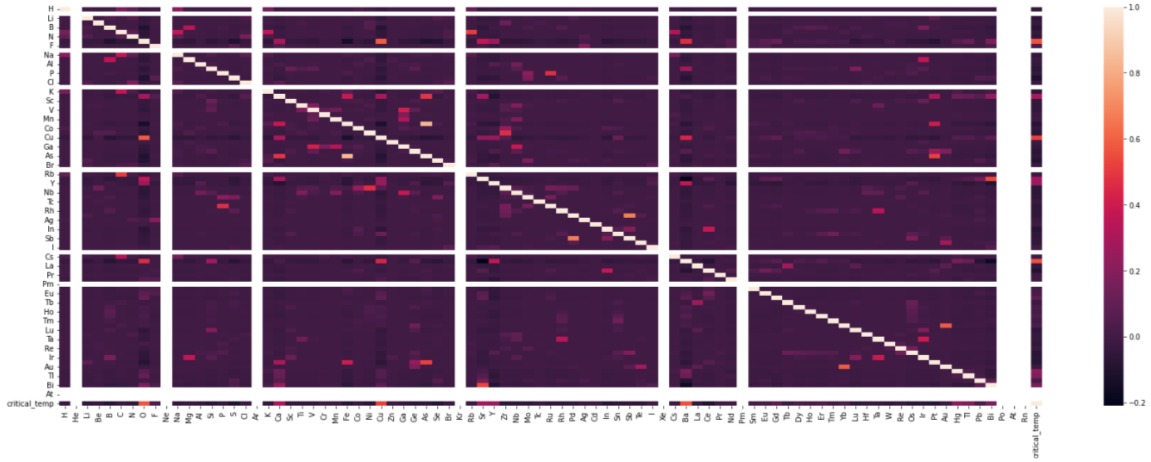

Fig 11. HeatMap to understand the correlation

Data Modelling Methodology

In this section we will mainly talk about the modelling methodologies required to predict the accurate critical temperature. In the previous section we have purely discussed about the relation between the various predictor variables and the response and also most visualized the data. Now that we understand the data thoroughly, we will focus on training the models using Linear Regression and Random Forest Regressor.

Following steps were performed while running Linear Regression on the data

1. It was noticed that for few of the features, the values are sufficiently apart, hence to them all in the same range, Standardization was performed.
2. Next step was to understand if the principal components found in the data analysis are representative of the dataset. Hence PCA was performed with varied number of components.
3. Once the features and the labels were made available, the following step was to train the model using Linear regression. The model was trained mainly trained using 5-fold cross validation approach.
4. Finally, the model was evaluated using the R2 statistic and the Mean-squared-error.

The above procedure was majorly run on both the datasets i.e., the train.csv and the unique.csv. As mentioned earlier, the PCA was performed with varied number of components namely 20,40,80. As a part of the outcome, the following details were output from the code

1. MSE for cross validated method
2. MSE for the full data
3. R2 statistic for the cross validated method
4. R2 statistic for the full data
5. Number of features used in training the model.
6. Intercept of the line
7. Regression coefficients

As referenced in the paper, the Linear Regressor was purely used as a baseline model. Below attached is the output pertaining to each of the dataset.

```
Intercept: 34.42121913535255
Coefficients: [-4.00308703 -1.88257164  2.22514372 -1.28459591  2.67613648  3.41569257
 -0.82045703 -0.5083912  -1.94797304  0.02395542 -3.43070843 -2.54650312
  0.40376027  0.46758994  3.90768216 -0.15424801 -0.4146399  -4.93821507
  2.98711561  7.41691014]
             cols         val
0              R2    0.621274
1           R2_CV    0.587646
2             MSE  444.361342
3          MSE_CV  483.817487
4  No. of Features   20.000000

Intercept: 34.42121913535251
Coefficients: [-4.00308703 -1.88257164  2.22514372 -1.28459591  2.67613649  3.41569259
 -0.82045704 -0.50839116 -1.94797245  0.02395715 -3.43070338 -2.54650138
  0.40377462  0.46756131  3.90764439 -0.15440494 -0.41449035 -4.93925794
  2.98928271  7.41961881  2.24687263  2.92170377 12.63163063 -1.24943855
 11.38268453 -6.39244712  5.14327771  0.85414669 -3.5431655  -1.55993487
 -8.55816832 -0.78475671 -5.50465639 -2.43432315 -9.77306822 -2.92339642
 -3.9351684  -0.16481957 -1.43002581  8.58763883]
             cols         val
0              R2    0.691978
1           R2_CV    0.652940
2             MSE  361.404202
3          MSE_CV  407.208061
4  No. of Features   40.000000

Intercept: 34.42121913535249
Coefficients: [-4.00308703e+00 -1.88257164e+00  2.22514372e+00 -1.28459591e+00
  2.67613649e+00  3.41569259e+00 -8.20457042e-01 -5.08391156e-01
 -1.94797245e+00  2.39571468e-02 -3.43070338e+00 -2.54650138e+00
  4.03774617e-01  4.67561304e-01  3.90764440e+00 -1.54404918e-01
 -4.14490343e-01 -4.93925786e+00  2.98928310e+00  7.41961911e+00
  2.24687349e+00  2.92170316e+00  1.26316340e+01 -1.24943953e+00
  1.13826963e+01 -6.39247167e+00  5.14325569e+00  8.54333200e-01
 -3.54295702e+00 -1.56012036e+00 -8.55802263e+00 -7.84899262e-01
 -5.50474193e+00 -2.42893252e+00 -9.77948760e+00 -2.92544659e+00
 -3.93386978e+00 -1.27432751e-01 -1.43827324e+00  8.67528353e+00
  8.76628025e+00 -1.21725673e+01 -2.01995900e+00 -8.62160234e+00
  8.22030528e+00  1.65806415e+00 -5.07464354e+00  4.93498383e+00
  4.37215507e+00 -9.75162013e-01  2.34466111e+00 -1.01441427e+00
  2.22222203e+01  9.52059019e+00  4.55279719e+00  1.23365036e+01
  2.80329021e+01  2.57200350e+01  2.98245910e+01 -5.20684998e+00
 -5.41803715e+00 -2.19195844e+01 -1.55377585e+01 -9.94109705e+00
  2.23396537e+01  3.39206736e+01 -4.04438062e+01 -1.43838637e+01
 -2.34067080e+01 -1.89682971e+01 -5.19318527e+00 -2.30129466e+01
 -1.75278362e+01 -5.64704541e+00  2.08928572e+01 -6.56860968e+01
  7.39759340e+01  6.11092847e+01  1.04573290e+02  9.79912322e+01]
             cols         val
0              R2    0.735130
1           R2_CV    0.692217
2             MSE  310.773692
3          MSE_CV  361.123173
4  No. of Features   80.000000
```

Fig 12. Outcome of Linear regression on the first dataset

```
Intercept: 34.42121913535242
Coefficients: [13.69857413  3.34557648  1.07289128  3.36876443  1.40781633  0.10380594
  1.18066876 -2.56110235  1.60278175  1.83987222 -1.284922    0.88456677
  2.04703503  1.84911742 -0.50917273  0.34528524 -0.32412685 -2.81360657
  1.1700641   0.89281499]
              cols         val
0               R2    0.527662
1            R2_CV    0.429117
2              MSE  554.196867
3           MSE_CV  669.820138
4  No. of Features   20.000000

Intercept: 34.42121913535251
Coefficients: [ 1.36984082e+01  3.34885638e+00  1.11441600e+00  3.30054799e+00
  1.58436821e+00  5.78601716e-03  7.57579315e-01 -2.53543957e+00
  1.87326299e+00  1.72802256e+00 -1.12465831e+00  1.74058830e+00
 -1.89421675e+00 -9.12149177e-01  1.76183938e+00  1.04361745e+00
 -3.99407500e-02 -1.47409060e+00 -9.11241444e-01  1.30702890e+00
 -1.80594636e+00 -2.56531600e-01 -3.07118176e-01 -8.69804795e-01
  4.28751473e-01 -3.25248465e-01  1.52026355e+00  1.09691468e+00
 -3.31798816e-01 -1.09516249e+00 -6.25431873e-01  3.77031996e-01
 -4.44503119e-01 -3.49584860e-01  5.21922893e-01 -5.66155566e-01
  3.54831900e-01 -6.00115051e-02  4.18887274e-01 -7.12442277e-01]
              cols         val
0               R2    0.535768
1            R2_CV    0.444736
2              MSE  544.686439
3           MSE_CV  651.494157
4  No. of Features   40.000000

Intercept: 34.421203908731506
Coefficients: [ 1.36979789e+01  3.34746290e+00  1.10742926e+00  3.31435592e+00
  1.55799736e+00  7.99881670e-02  7.79259230e-01 -2.52472866e+00
  1.83597353e+00  1.68531090e+00 -1.06461477e+00  1.79968436e+00
 -1.72445135e+00 -4.96665518e-01  1.84865901e+00  6.85564518e-01
  3.70726793e-02  1.14399047e+00 -1.63550797e+00  2.65916699e+00
 -4.78891921e-01  8.20720639e-01 -4.98773424e-01 -8.73713611e-01
 -3.67642867e-02  7.37595406e-01  5.46725121e-01 -7.72587180e-01
 -5.91950685e-01 -5.99343691e-01  1.04641257e+00  6.48318316e-01
 -5.25822923e-01  2.34696972e-01 -4.44616049e-01 -9.05908213e-02
 -3.53593423e-01  3.25699070e-01 -3.89022378e-01  5.55818247e-01
  6.28965831e-01  2.35812887e-01  4.68518473e-01 -9.74675910e-01
 -1.64640622e-01 -4.97037706e-01  2.07438826e-01  5.36789419e-01
  3.02188492e-01 -3.12294827e-01 -4.14391833e-01 -6.48111960e-01
  5.63883920e-01  5.01501477e-01 -6.21542716e-01  7.30034973e-01
 -1.81935061e+00 -4.64347948e-01 -1.39218512e+00  4.39561171e-01
 -1.00320196e+00  1.63608140e+00  5.64079011e-02 -3.35489083e+00
  1.29444147e+00 -1.05737934e+00  4.47314044e+00  1.28426445e+00
  1.38363238e+00  4.29012058e-02  1.36393846e-01  1.98248313e+00
  4.82869891e-01  3.98993955e+00 -1.81722934e+00  1.78012132e+01
  6.14772548e+00 -3.50233657e+13  4.67990232e+12 -2.08559098e+12
  2.65249725e+11  7.43739425e+11 -5.20097500e+11 -8.84165750e+11
 -3.96314295e+12  1.72932419e+11]
              cols         val
0               R2    0.616394
1            R2_CV    0.387672
2              MSE  450.087472
3           MSE_CV  718.448478
4  No. of Features   86.000000
```

Fig 13. Outcome of Linear regression on the second dataset

Outcomes from the above experiment

1. Model predicted more accurately when it was trained on the entire feature set rather than few top principle components
2. It was also noticed that the first dataset was more related to the critical temperature rather than the second dataset.

Second step in the data modelling was to train a slightly more flexible model. Hence Random Forest Regressor was chosen. As we already inferenced from the previous model, PCA components was dropped and so only the 5-fold cross validation was performed before training the Random Forest.

Outcome of the above experiment

1. Random forest significantly outperformed Linear Regression
2. Top 20 important features were visualized from the trained ensemble-based model
3. RMSE was calculated for the model trained on each of the dataset.

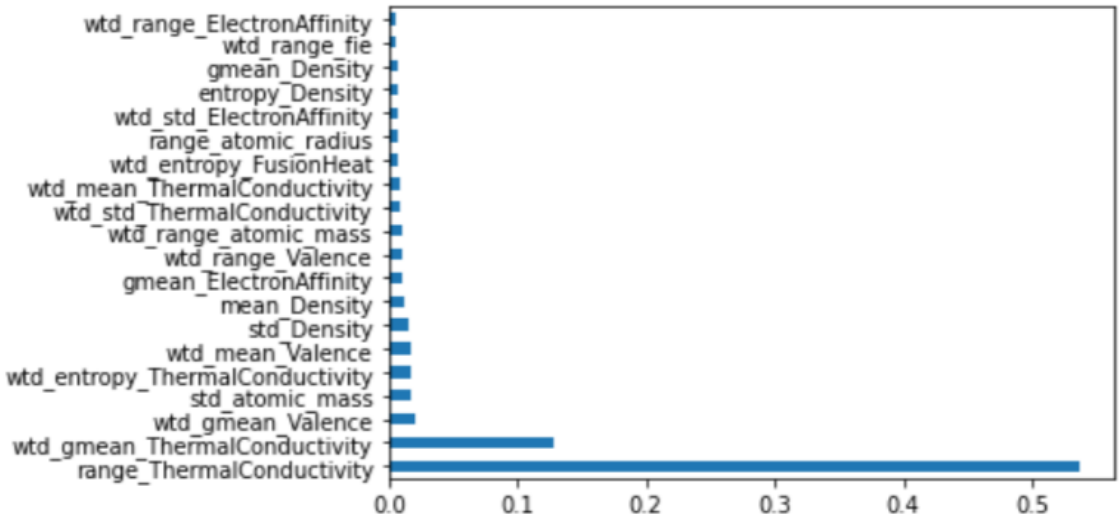

Fig 14. Important Features w.r.t the first dataset

As can be the seen in the above Fig 14. It can be said that the predictor variables namely

1. Range_ThermalConductivity
2. Wtd_gmean_ThermalConductivity
3. Wtd_gmean_valence

Form the important features in the first dataset,if we were to consider the top 3.And in terms of the second dataset

1. Cu
2. Ca
3. O
4. Ba

Form the important features.

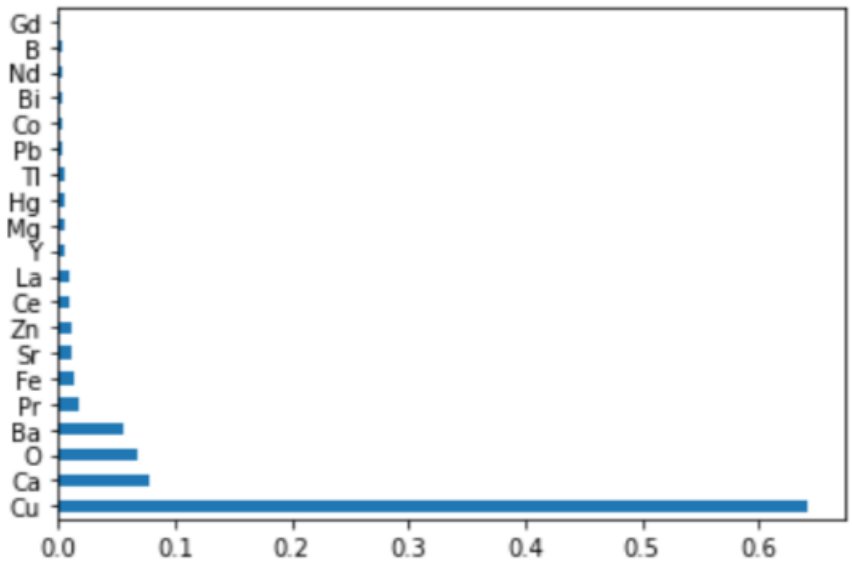

Fig 15. Important Features w.r.t the second dataset

Future scopes of this experiement

1. As referenced in the paper, XGboost model can be trained and utilised to predict the critical temperature more accuarctely.
2. Curve fitting can be utilised to visualize the underestimation or the underestimation happening in the underlying mode.

Conclusion

A thorough data analysis was performed on both the dataset. Finally Linear Regressor and RandomForest Regreesors were trained in varied conditions. It was conclude in the end that Random Forest performed better as compared to the Linear Regressor on the First dataset.

References

- Hamidieh, Kam, A data-driven statistical model for predicting the critical temperature of a
superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346-
354

Appendix

The code that was utilized perform the various analysis and the modelling is attached in the form of link below.

https://www.kaggle.com/sanjayshan/semiconductor-critical-temperature-prediction