```r
### Load libraries
# Vizualization
library(ggplot2)
library(corrplot)

# Models
library(Rtsne) # t-Distributed Stochastic Neighbor Embedding
library(e1071) # SVM model

# Evaluation
library(caret) # Cross-Validation
library(Metrics) # RMSE metric
```

1 Data Preparation

In this section, we will load the data and explore the formats. The dataset contains two files:

- `train.csv` - containing 81 features extracted from 21263 superconductors together with their own critical temperature
- `unique_m.csv` - the chemical formula for each one of the 21263 conductor

This analysis will use the data contained in the `train.csv`.

```r
original_data <- read.csv('Data/train.csv')
dim(original_data)

## [1] 21263    82
```

The structure of the data is of the form 81 features, with the 82th column as the target variable, and 21263 instances. The first feature column represents the number of elements the materials have. Then the rest of the 80 columns are based on the following properties with their corresponding units:

- **Atomic Mass** - Atomic Mass Unit (AMU) - Total proton and neutron rest masses
- **First Ionization Energy** - kilo-Joules per mole ($\frac{kJ}{mol}$) - Energy required to remove the a valence electron
- **Atomic Radius** - Picometer (pm) - Calculated atomic radius
- **Density** - Kilograms per meters cubes ($\frac{kg}{m^3}$) - Density at standard temperature and pressure
- **Electron Affinity** - kilo-Joules per mole ($\frac{kJ}{mol}$) - Energy required to add an electron to a neutral atom

- **Fusion Heat** - kilo-Joules per mole ($\frac{kJ}{mol}$) - Energy to change from solid to liquid without temperature change
- **Thermal Conductivity** - Watts per meter Kelvin ($\frac{W}{mK}$) - Thermal conductivity coefficient κ
- **Valence** - # - Typical number of chemical bonds formed by the element

For each one of the enumerated properties, the following measures make up the 80 columns of the data:

- **Mean** $= \mu = \frac{\sum e_i}{n}$
- **Weighted Mean** $= v = \sum p_i \times e_i$
- **Geometric Mean** $= \sqrt[n]{\prod e_i}$
- **Weighted Geometric Mean** $= \prod e_i^{p_i}$
- **Entropy** $= -\sum w_i ln(w_i)$
- **Weighted Entropy** $= -\sum A_i ln(A_i)$
- **Range** $= e_{\text{max}} - e_{\text{min}}$
- **Weighted Range** $= p_{\text{max}} e_{\text{max}} - p_{\text{min}} e_{\text{min}}$
- **Standard Deviation** $= \sqrt{\frac{\sum (e_i - \mu)^2}{n}}$
- **Weighted Standard Deviation** $= \sqrt{\sum p_i (e_i - v)^2}$

Where e_i is the property value for the element i of the material, n is the number of elements in the material, p_i is the ratio of elements in the material, w_i is the fraction of the total property value, and $A_i = \frac{p_i w_i}{\sum p_j w_j}$

With further analysis, we can see that there are no missing data and that the data type of all of the instances is numerical (3 integers, and 79 continuous). Note that the 3 integer variables will not be transformed into categorical variables. So, no changes will be made and no column will be removed as all of them will be providing information on the target variable.

2 Exploratory Data Analysis

In this section, we will try to understand some of the relations between the target variable and the predictors.

2.1 Target variable

It is critical to consider the distribution of the target variable, namely the critical temperature. In Figure 1, the values for the critical temperature is left-skewed, with the data trending towards 0K, meaning the temperatures are mostly low. Yet with a tail beyond 175K indicating the existence of elements with critical temperature more accessible in practical applications. Now, we will look to see if the entry with highest temeprature is worth keeping for our analysis.

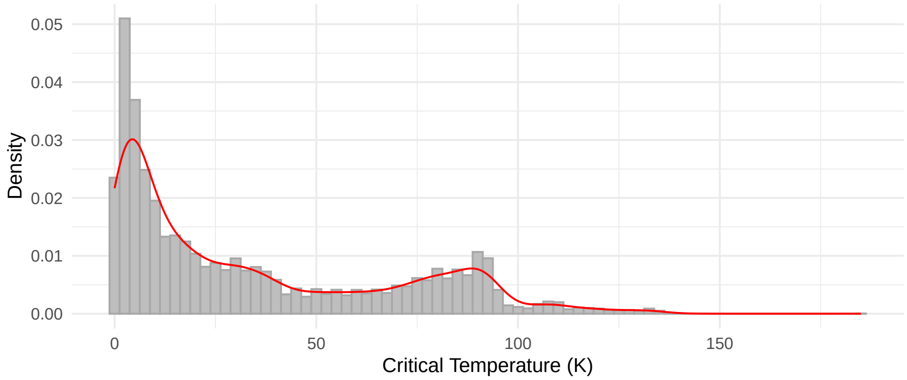

Figure 1: Ditribution of Critical Temperature values

The $1.5 \cdot IQR$ rule will be used to determine if the entry with the critical temperature 185K is an outlier.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
##   0.00021    5.36500   20.00000   34.42122   63.00000 185.00000
```

$$Q_3 + 1.5 \cdot IQR = Q_3 + 1.5 \cdot (Q_3 - Q_1) = 63 + 1.5 \cdot (63 - 5.365) = 149.4525$$

Conform the $1.5 \cdot IQR$ rule, the entry with the critical temperature 185K is an outlier. Although, looking for the material, it turns out to be hydrogen sulphide, which is one of the superconductors with the highest critical temperature. Additionally, the purpose of the analysis is to give the posibility to discover warmer superconductors. Therefore, the entry will not be removed.

2.2 Feature Engineering

Due to the large number of features, it is worth applying dimension reduction algorithms and looking further into the data structure. The algorithms we will use are:

1. t-Distributed Stochastic Neighbour Embedding (t-SNE)

2. Principal Component Analysis (PCA)
3. Corrolation Clustering
4. Hierarchical Clustering
5. Random Forest
6. Gradient Boosting

2.2.1 t-Distributed Stochastic Neighbour Embedding

The first method to use for dimension reduction is t-distributed stochastic neighbour embedding or t-SNE. This algorithm will take the high-dimensional data points and reduce them to a 2D dimension for better visualisation. This means that it will transform the 81 features of the data into 2 features so the relationship between the entries is preserved.

The main steps to the t-SNE algorithm are:

1. Compute High-Dimensional Similarities

   The high-dimensional similarity refers to the calculation of the probability of an entry choosing another entry as its neighbour. The probability is calculated between each two points using the Gaussian Distribution as such:

   $$p_{i|j} = \frac{exp(-\frac{||x_i - x_j||^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(-\frac{||x_i - x_k||^2}{2\sigma_i^2})}$$

   The Gaussian distribution is centered around each x_i and the variance σ_i is determinated by the perplexity parameter. The perplexity is a hyperparameter that influences the balance between preserving local and global relationships in data. A perplexity is more or less a target number of neighbours for the central point x_i. A higher perplexity results in a higher varience and an emphasis on the global structure. While a lower perplexity focuses on the local structure.

   On the other hand, $||x_i - x_j||$ represents the Euclidean distance between the points x_i and x_j in the high-dimensional space.

2. Compute Low-Dimensional Similarities

   The low dimension is the new space build using the pairwise similarity using the Student's t-distribution with a single degree of freedom:

   $$q_{i|j} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||y_i - y_k||^2)^{-1}}$$

3. Optimize the Embedding

- t-SNE - What does it do? -Small introduction and why I am applying it

3 Comments

For further studies, I believe it is worth doing analysis on readings at different pressures. As it is seen for the case of hydrogen sulphide (H_2S) reaching superconductivity properties at the critical temperature of 203K when the pressure is 1.5 million bar. Therefore, following this path will save up resources in finding materials that will present superconductivity properties at temperatures closer to room temperature. - High acuracy is not a target as we want to understand what makes an element with the desired superconductive critical temperature. ->Focus on the transparency and