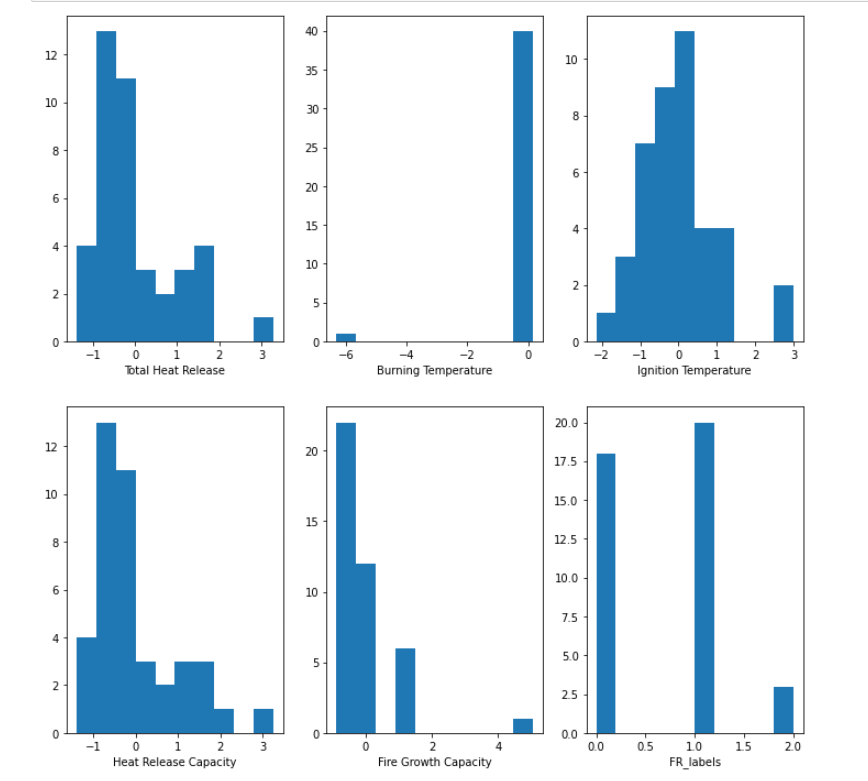
**Baseline Model**

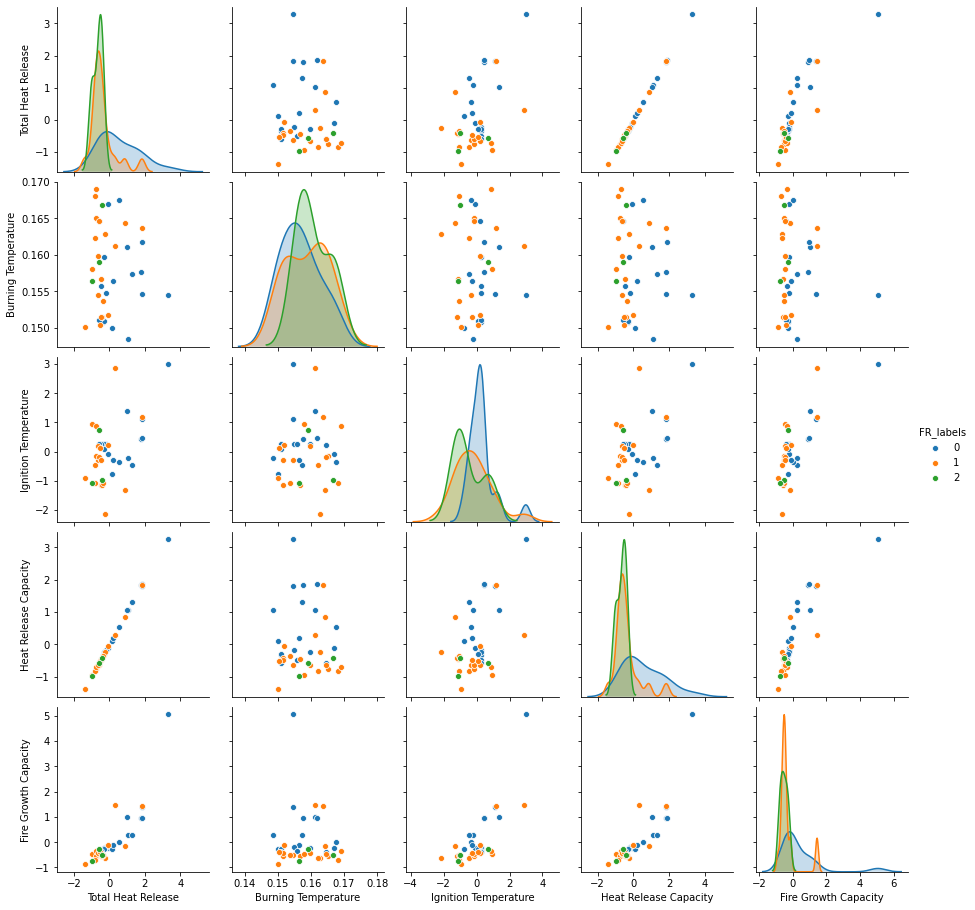
**Data Preprocessing:**

The four features we extracted from original MCC data have very different ranges, and different units. The units of heat release capacity (HRC), Ignition temperature(Ti), Burning temperature (Tb), and Fire growth capacity (FGC) are joule per degree of Celsius, degree of Celsius, degree of Celsius, and Joule per gram per degree of Celsius. For HRC, the feature value ranges from 105 to 103. Ti and Tb range from 25°C to 600°C. It is critical to ensure that these four variables are not weighted more than others. Feature scaling of these four features are performed with standard scaler with help from *sklearn* package. After feature scaling, the four features all showed similar data ranges. The total heat release rate is used to calculated HRC and FGC. Histograms are plotted with each features and FR testing results as shown in Figure 1.



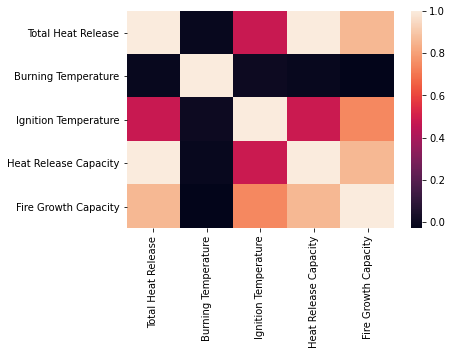
**Figure 1**. Histograms of features and FR testing labels after feature scaling.

One notable change to the data set was the removal of a single outlier that severely skewed the ‘Burning Temperature’ feature shown in Figure 1. Seaborn’s *pairplot* function is then used to plot out pairs of features to show pairwise relationships in the current MCC dataset between features and FR testing labels as shown in Figure 2.



**Figure 2.** Pair plot for each features and FR testing label after dropping outlier.

An additional assessment of the features was done using the correlation matrix shown in Figure 3. to check for collinearity between features. As seen in Figure 2. and emphasized in Figure 3., the features ‘Total Heat Release’ and ‘Heat Release Capacity’ are extremely correlated, and it would be possible to remove one of these features for future models.



**Figure 3.** Heatmap of Correlation Matrix

**Model Training Process:**

The general procedure of constructing the baseline mode was shown in Figure 4. First, we preprocess our data to extract the feature matrix, which is a 41x6 matrix. Afterwards, we used standard scaling to scale all the features in the matrix. Next, we decided to do a train test split to our processed data with 10% of the data left for validation set. Before modeling, we do cross validation to the data.

Five different regression models have been chosen for the project:

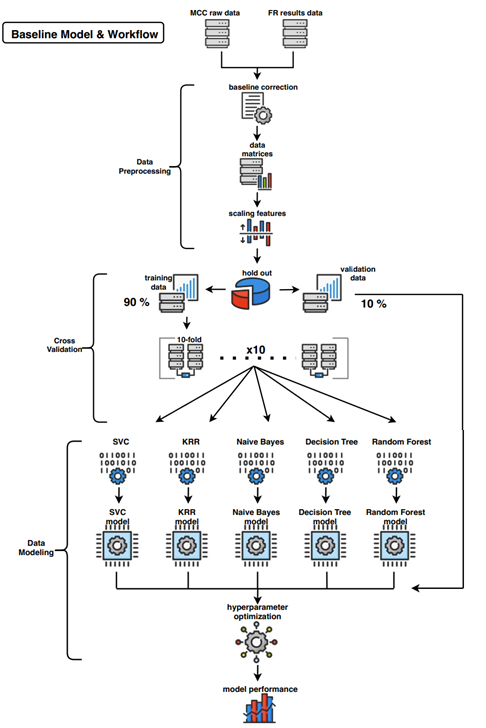
1.SVC.

2.Kernel ridge regression (KRR).

3.Naive Bayes

4.Decision tree

5.Random forest.



**Figure 4**. Pipeline diagram for building baseline model and workflow

According to the pipeline diagram in our description file, a 90% train-test split will be performed to the scaled dataset. Support Vector Classifier (SVC) is selected as our baseline model. GridSearchCV is used for determining the optimal hyperparameters with a 3-fold cross-validation. At the end, the accuracy, precision and recall scores are printed as benchmark metrics with other four potential models.

**Paths forward:**

Thus far, a complete pipeline connecting feature matrix to results has been built up. The SVC baseline model, itself, provides a decent but not perfect performance on the validation set. It's still worth trying to improve the based model and do further analysis in the future. The possible paths will include but not limited to:

***(1)*** Try other classification models such as random forest, neural network, etc. Compare the performance of each model.

***(2)*** Extract more features from raw data using principle component analysis (PCA). Compare the performance among different feature combinations. For example, calculate and compare the metrics of models using (a) features with physical meaning alone, (b) features extracted by PCA alone, and (c) combination of both.

***(3)*** There are three classes for the output. Only two materials fall in the category of class 2 (hard to determine), whereas the other two classes (class 0 & 1) distribute almost evenly. Resampling methods and generative methods such as Gaussian Naive Bayes method can be applied to fix the problem.

***(4)*** Due to the small size of our dataset, the fitting scores will largely depend on the train-test split. It might be improved by generating or obtaining new data.