SDSC3006
Steel Plates Faults Detection

Ho Tsz Yui 56649810
Chan Yuk Yee 56230549
Cheung Mei Ching 56624041

Date: 23/11/2022

Summary of Contribution

Ho Tsz Yui : Background, Problem Formulation, Conclusion, Discussion

Chan Yuk Yee : Classification Methods, Classification Using PCA, Results

Cheung Mei Ching : Data Preprocessing, Model Selection

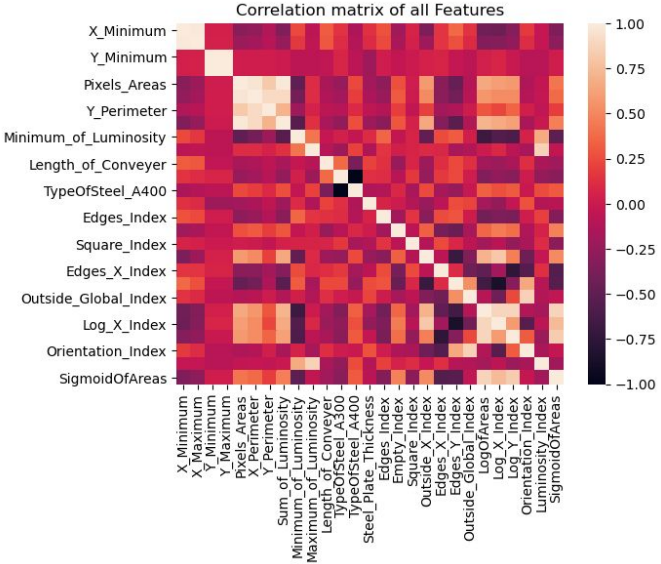
Background

This dataset comes from research by Semeion, Research Center of Sciences of Communication. It has 1941 observations, consists of 27 features, describing geometric shape of fault and 7 class label, indicating the type of fault.

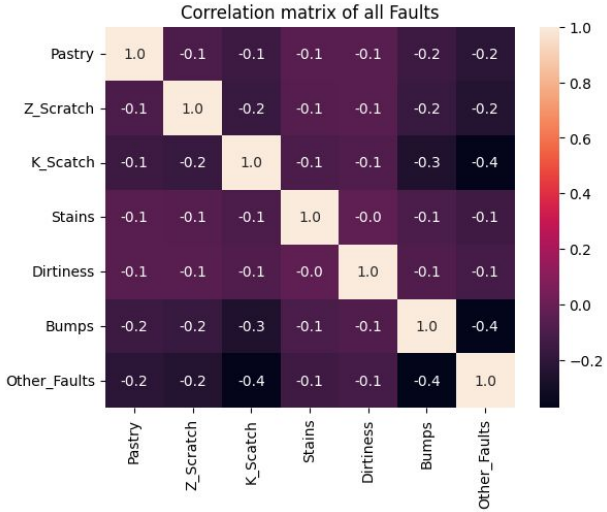
Problem Formulation

Once the fault of steel plate be predicted, the accident that harmful to human could be prevented. Since the response is binary variable, we will train serval classification models to classify the type of fault of steel plates. Further, compare the performance of different models and find out which model has the best results in predicting faults.

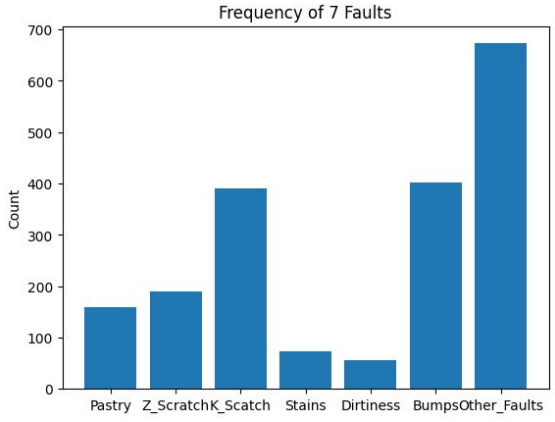
Data analysis and visualization



Lighter color indicates higher correlations between two variables



There are some weak negative correlations between two variables



Other_Faults has overwhelming majority as compared to all other type of faults.

Data Preprocessing

1. Separate the variables into Features(X) as input and Faults(y) as output.
2. Data cleaning: Check and remove the missing value if any
3. Data splitting: 70% into training set and 30% into test data set
4. Data normalization: Min-Max Normalization & **Standardization**
    - Better performance
5. Oversampling

Model Selection - Classification model

Justification:

1. Dataset is consist of multiple categories.
2. The given dataset has labels.

⇒ We believe that if we adopt the regression model instead of the classification model, the accuracy will be relatively lower.

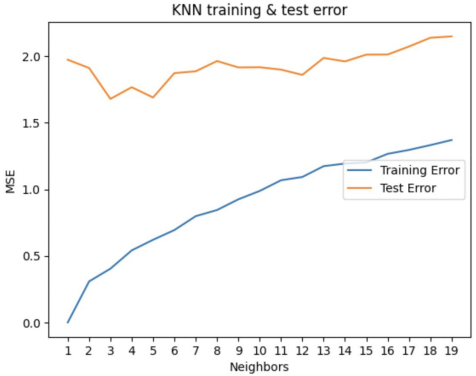
Methods

1. Four classification methods were used. i.e., K Nearest Neighbor, Logistic Regression, Random Forest, and XGBoost.
2. PCA was used to reduce the dimensions to 19.
3. All approaches are implemented with and without PCA.

Classification Methods

K Nearest Neighbor(k=3): Density-based classification algorithm that uses proximity to classify or predict groupings of individual data.

⇒ The smaller k value, the smaller the test error.



Logistic Regression: Linear classification algorithm. The classifier assumes that the class attribute is linear in the coefficients of the predicted attribute, which are used to predict the probability of happening.

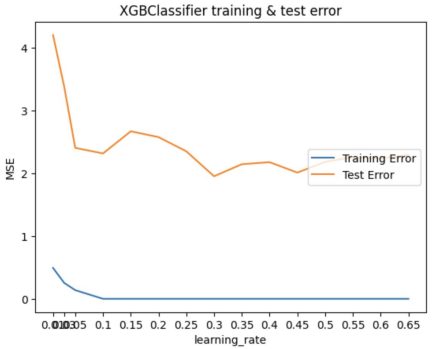
Classification Methods

Random Forest(max_depth=20): Decision-tree-based classification algorithm consisting of many decision trees. In a random forest, a subset of features is randomly selected and the best predictors are filtered out.

⇒ The deeper max-depth, the smaller the test error.

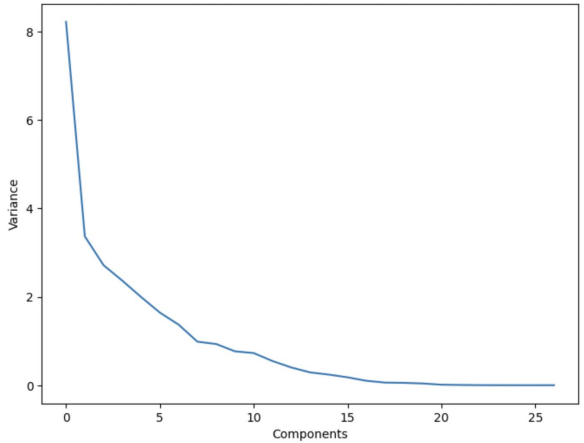
XGBoost(learning_rate=0.3) : Decision-tree-based ensemble classification algorithm that uses a gradient-boosting framework.

⇒ The optimal learning rate is 0.3.



Classification Using PCA

- Minimize the dimensionality of a dataset consisting of many correlation variable, while preserving the variation present in the dataset.
- Avoid testing and training overfitting.
- The feature vector of 10-20 principal components can be represented.
- PCA was used to reduce the dimensions to 19.



Results - Comparsion

	K Nearest Neighbor	Logistic Regression	Random Forest	XGBoost
Simple Accuracy	73.1%	68.3%	**76.0%**	75.5%
PCA Accuracy	73.1%	68.3%	**76.7%**	73.8%

The fault diagnosis performance of the above model is expressed using statistical accuracy.

Model without PCA: Random Forest performs the best with an accuracy of 76.0%. KNN and XGBoost perform above 70%.

Model with PCA: Random Forest performs the best, with an accuracy of 76.7%. KNN and XGBoost perform above 70%.

Conclusion

In this report, we have trained four classification models to predict the type of faults of steel plates. Three of them are learned from class including K-Nearest Neighbour, Random Forest, and Logistic Regression. One extra model not covered in class is Extreme Gradient Boosting. We have split the data into a train set and a test set. Dimension Reduction is performed by PCA, it reduces the number of predictors from 27 to 19. After our modeling with and without PCA, we found that the random forest with PCA is the best model among those models we trained since it has the highest accuracy (76.7%). Moreover, we noticed that without PCA, the random forest still has higher performance than other models.

Discussion

We adopt some simple classification models only in this report. In order to obtain higher performance in prediction, we can train more complex and advanced supervised learning models. It may provide better accuracy in high-dimensional data. One limitation is that we have not performed cross-validation. We do not know whether our result varies or not. Our result may suffer from high variance, further cause overfitting. Also, we have not implemented parameter tuning in each model. It may increase the prediction accuracy in our modeling.