**Final Project Proposal – Group 14**

108061806 簡婉軒, 108061892 Shreya Upadhyay, 109061804 邱信豪, 109061520 陳俊宇

**Title:** Expert system for Fuel desulfurization warning prediction based on Machine Learning Techniques

**Methods:**

The dataset contains 25 measurement and sampling variables recorded from fuel gas desulfurization equipment (FGD), which is an equipment operating process data with desulfurization results as label. This Fuel desulfurization warning prediction dataset has more than 20k samples for training; with 5 different labels. To choose the model, we first did some data analysis to get a brief idea of the data distribution. The below plot shows the data distribution with respect to fuel desulfurization classes (5 labels).

*Fig1: data distribution over 5 classes*

In this plot, we can see that the data is hugely unbalanced; to reduce the biases we will apply some up-sampling techniques after normalization [6] methods on the training data to make data balanced and also make all features appear in the same range. We also run some base machine learning model [2] as the baseline model to see if the data is suitable for which structure of algorithm such as rule- based, tree-based, lazy learning, bayes theorem based etc.

During analysis and also seeing the results from the base models, we saw that some features can have more predictive power than others. To avoid the curse of dimensionality or the overfitting of the data we are going to use some feature ranking or selecting important features [1] based on some wrapper or filter-based methods.

|  |  |  |
| --- | --- | --- |
| Models | Accuracy | Log Loss |
| Random Forest | 71.8 | 1.026 |
| Naïve bayes | 45.8 | 1.654 |
| SVM | 47.1 | 1.253 |

*Table1: baseline machine learning models performance*

With the initial base model performances and based on number of observations, speed of training, linearity and number of features; we can think of decision tree-based algorithms, random forest, kernel SVM [4], KNN and boosting algorithms. Peculiarly, we are going to use more than one classifier in stacking or voting form to conclude the prediction from results provided by different chosen best classifiers. Also, we are going to do some experiments with neural networks. As the neural networks mostly works well for the audio, video, image or text data which has the continuous information with its time and space information; There are some work [3], where the researchers have tried artificial neural networks in terms of predicting low sulphur content fuel oil consumptions. Specifically, we will try the Deep neural network (DNN) and feed forward artificial neural network[4] as Multi-Layer Perceptron (MLP) to see the prediction performance with 1-d features recorded from equipment for better fuel desulfurization warning prediction.

**References:**

1. Miao, Jianyu, and Lingfeng Niu. "A survey on feature selection." *Procedia Computer Science* 91 (2016): 919-926.
2. Tamascelli, Nicola, Nicola Paltrinieri, and Valerio Cozzani. "Predicting chattering alarms: A machine Learning approach." *Computers & Chemical Engineering* 143 (2020): 107122.
3. Tran, Tien Anh. "Design the prediction model of low-sulfur-content fuel oil consumption for M/V NORD VENUS 80,000 DWT sailing on emission control areas by artificial neural networks." *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment* 233.1 (2019): 345-362.
4. Saldana, Diego Alonso, et al. "Prediction of density and viscosity of biofuel compounds using machine learning methods." *Energy & fuels* 26.4 (2012): 2416-2426.
5. Patro, S., and Kishore Kumar Sahu. "Normalization: A preprocessing stage." *arXiv preprint arXiv:1503.06462* (2015).