MLPR Lab 9

**INSTRUCTIONS**

* Factory Reports are given as a dataset (factoryReports.csv). You need to classify the descriptions given in a dataset as per the category of the description.
* Submit the complete jupyter notebook, the output image shown below and the answers in one PDF.
* Download and install dependencies/libraries as per you required system preferences.

**Step1**: Import libraries

* Numpy
* Pandas
* Train test split
* TFID vectorizer from sklearn
* Random Forest classifier
* Accuracy score, confusion matrix
* Matplotlib
* NLTK (Natural Language tool kit- May be need to be installed in your system)
* Stopwords from nltk

**Step2**: Download “punkt” and “Stopwords” from nltk to do text preprocessing.

**Step3**: Set a random seed to 0 for reproducibility.

**Step4**: Load the csv data file (factoryReports.csv) and do preprocessing and removal of noise. Example- Remove all stop words of English language.

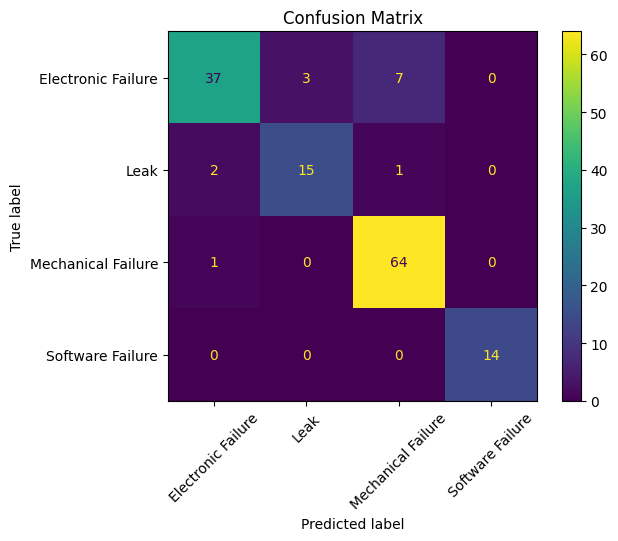
**Step5**: Tokenize all the individual words in given texts and convert them into lower case for the columns ‘Description’ and ‘Category’ columns given in csv file.

**Step6**: Apply TFIDF vectorizer and then split data into x\_train and x\_test.

**Step7**: Train a random forest classifier. Use n\_estimators=100 but you can try different values for estimators and choose the parameter that gives you optimal performance. Use random\_state=42.

**Step8**: Fit the model and make prediction.

**Step9**: Calculate and print the accuracy. Also plot the confusion matrix shown below as an output reference.



**QUESTIONS**

1. Which techniques would you use for keyword normalization in NLP, the process of converting a keyword into its base form?
2. What is TF-IDF in NLP?
3. What are some alternatives for TF-IDF method to parse text data?
4. What advantage(s) does word2vec has over TF-IDF?
5. **Provide two applications of the the above lab assignment problem.**