**READ ME-Maple Squad**

Please read this file till the end.

**Team:** Maple Squad (Jaskiran Kaur, Shubham Deshmukh, Sohini Sarkar)

**Name of the competition:** Toxicity prediction

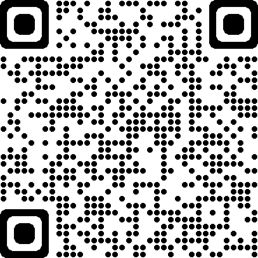
Using machine learning algorithms to make prediction about the chemical whether it is toxic or not.

**The Project is uploaded to a github repository.**

**Link*:*** [***https://github.com/shubham10deshmukh/Chemical-Toxicity-Prediction-ML-Kaggle***](https://github.com/shubham10deshmukh/Chemical-Toxicity-Prediction-ML-Kaggle)

**or**

**Scan this QR code to open the GitHub Repository.**



The repository can be cloned, if having problem in it the files are given in a zip file as well, below are some details to open it.

The **Zip file** contains the following:

1. The Final Report in pdf format.
2. Python notebook (code file) of the project mapleSquadCode.ipynb
3. Four Datasets in .CSV format.
4. The Read Me file.
5. The actual submission file.

**Prerequisites:**

* Anaconda

You can download and install from

[**https://www.anaconda.com/products/individual#**](https://www.anaconda.com/products/individual)

For more information see

[**https://www.geeksforgeeks.org/how-to-install-anaconda-on-windows/**](https://www.geeksforgeeks.org/how-to-install-anaconda-on-windows/)

The Assignment is built on Python 3 using Anaconda IDE on a Jupyter notebook, however, the python notebook file can be opened and run in any Python supported environment. However, Jupyter is recommended.

* Key Dependencies: Python Libraries

Following are the specific versions for the various packages. You can enter this in the terminal on your computer.

Numpy 1.19.2

pip install numpy

Pandas 1.1.3

pip install pandas

Scikit-learn 0.23.2

pip install sklearn

Matplotlib 3.3.2

pip install matplotlib

Imbalanced-learn 0.8.0

pip install imblearn

* Jupyter Notebook for Python 3 and supported web browser.

For help how to launch the jupyter notebook on anaconda navigator click on <https://nickmccullum.com/python-course/how-to-install-anaconda/>

* Microsoft Excel for viewing CSV files.

**Implementation:**

KAGGLE FINAL SUBMISSION-PY.py ## This is the code file to train and test the model.

Four URL variables are mentioned at the start of the code, in which the path of the CSV files is mentioned. The relative paths are already set for all files. The dataset file "feamat.csv", "test.csv", "train.csv" and “features\_id\_name\_mappings.csv” are the default version of the datasets available on the Kaggle Competition page. All the files are included in the zip folder.

**url\_train\_datafile** = 'train.csv'

**url\_test\_datafile** = 'test.csv'

**url\_feamat\_datafile** = 'feamat.csv'

**url\_mappings\_datafile** = 'features\_id\_name\_mappings.csv'

**train.csv**## This contains ID column and expected column for train set.

**test.csv**## This contains ID column.

**feamat.csv** ## This contains the 1075 columns regarded as features.

**newMergedDF** ##newMergedDF is our new merged Dataset after data preparation part by handling missing data.

**X\_train and X\_test** ## The data is divided into two datasets of the same size for feature selection.

**train\_X, val\_X, train\_y, val\_y** ## The merged training dataset is split into 80:20 ratio with 80% training data and 20% testing data for the modeling.

**Y\_prediction** ## This contains the final prediction of the model.

**Maple\_Squad-1.csv** ## the final prediction result is exported to a csv file.

A new path might need to be set in any other computer/desktop for a local file and code to run and execute. For help see <https://www.computerhope.com/issues/ch000549.htm>

**Technical Details**

**Please Run the code step by step, print statements signify that the code block is run successfully, F1-Macro score is also printed.**