How to use GEOML Program

This program developed in python simply finds the all files containing drill data(by searching for all .snd extensions) and laboratory reports(by searching for all .xlsx or .xlsm extensions) and performs file parsing, data extraction, data matching, data merging as well as model building using Machine learning(Lightgbm) and model evaluation.

In order to use, all the python libraries/frameworks listed below must be installed on the computer to be used.

"https://pypi.org/project/regex/" >Regex library , "https://pypi.org/project/openpyxl/" >Openpyxl Library, "https://pypi.org/project/numpy/" >Numpy Library, "https://pypi.org/project/pandas/" > Pandas Library, "https://pypi.org/project/os-win/" > OS Library, "https://pypi.org/project/scikit-learn/" > Sci-kit Learn Library, "https://pypi.org/project/lightgbm/" > Lightgbm Library, "https://stackoverflow.com/questions/47722353/how-to-install-warnings-package-in-python" > Warnings Library, "https://docs.python.org/3/library/time.html" > Time Library

The python file must be run in any python compatible Integrated Development Environment (IDE) such as VScode, Pycharm, Spyder etc. The python file should also be in the same folder where the data files(drill data and laboratory) subfolder is. If this isn't done, the path to the data file folder must be specified in the code. e.g. "GEOML-1" may be replaced with "downloads/GEOML-1" or "downloads\GEOML-1" on windows if the data files are in the GEOML-1 folder in the downloads folder.

```python
for pair in datapairlist("GEOML-1"):
    try:
        # read SNDs in datapairlist as comma separated values
        snd_table = pd.read_csv(pair[1], delimiter = "\r\n", header = None, sep = " ", names = "v")
        snd_table_lst = []
```

The program takes about 15-20 seconds to process each file pair (snd/workbook pair). This is mostly due to parsing of each spreadsheet in a workbook.

If new soil types are added in the data files, they must be added to the soil_type dictionary. Here is a list of recorded soil types:

1. LEIRE - CLAY
2. KVIKKLEIRE - QUICK CLAY
3. TØRRSKORPELEIRE - WEATHERED CLAY
4. SILT - SILT
5. TØRRSKORPESILT - WEATHERED SILT
6. SAND - SAND
7. GRUS - GRAVEL
8. GYTJE - GYTJA

9. ORG. MATR. - ORG. MAT.
10. MATJORD - TOPSOIL
11. DY - DY
12. MATERIALE - MATERIAL
13. FYLLMASSE - FILL SOIL

To add or remove a soil type, add/remove both the Norwegian name and the English name alongside its left comma. The names should be enclosed in quotation marks ("" or '').

```
#Dictionary containing Soil type translation from Norwegian to English
soil_type = {'LEIRE': 'CLAY', 'KVIKKLEIRE': 'QUICK CLAY', 'TØRRSKORPELEIRE': 'WEATHERED CLAY',
             'SILT': 'SILT', 'TØRRSKORPESILT': 'WEATHERED SILT', 'SAND': 'SAND',
             'GRUS': 'GRAVEL', 'TORV': 'PEAT', 'GYTJE': 'GYTJA',
             'ORG. MATR.': 'ORG. MAT.', 'MATJORD': 'TOPSOIL', 'DY': 'DY',
             'MATERIALE': 'MATERIAL', 'FYLLMASSE': 'FILL SOIL'}
```

To use on a single soil sample (one worksheet and snd file) or set of samples, a new folder should be created in the same directory of the program file where the data files would be added. The name of the new folder would then be put in the datapairlist function. e.g. datapairlist("GEOML-1") to datapairlist("new-folder-name").

```
for pair in datapairlist("GEOML-1"):
    try:
        # read SNDs in datapairlist as comma separated values
        snd_table = pd.read_csv(pair[1], delimiter = "\r\n", header = None, sep = " ", names = "v")
        snd_table_lst = []
```