File explanation:

CSV files from external source

* **H-1B\_Disclosure\_Data\_FY2019** - Original Dataset from US government
* **SOC** - Standard Occupational Classification
* **f500** - list of fortune 500 companies
* **USregion** - US states separated into region

CSV files generated from H-1B\_Disclosure\_Data\_FY2019 to be used for ML

* **2019H1BNumericNew1** - data with all features
* **bestchi20** - 20 features selected using chi-squared feature selection
* **mutual\_10BestSelection** - 10 features selected using mutual information feature selection

Jupyter Notebook files

* **DataPreparation** – Clean and prepare dataset to be used for ML. Generates 2019H1BNumericNew1, chi25BestSelection and mutual\_10BestSelection files
* **KNN validation** – performance measurement with figures with KNN
* **Logistic\_regression validation** - performance measurement with figures with Logistic Regression
* **Random Forest validation** - performance measurement with figures with Random Forest
* **SVMs validation** - performance measurement with figures with Random Forest
* **Random Forest** – train Random Forest
* **SVMs** – train SVM

Installation: To run these files, we recommend you to set up Anaconda (version 4.8.1) and use Jupyter Notebook environment.

Execution order:

1. Make sure all the excel files and notebook files in a same folder (for the directory). In the beginning you need four excel files: H-1B\_Disclosure\_Data\_FY2019, SOC, f500 and USregion
2. Run DataPreparation code to generate necessary files (caution: since dataset is huge more than 700,000 rows, this may quite long time (around 5 to 10 min)
3. In feature selection part in DataPrepartion, make sure you alternate between the chi2 and mutual\_info\_classif to use both feature selection
4. Upon successful file generation, you can run the rest of codes to train and validate with graphs and measurements