Brief Description:

This directory contains the following files:
- `Assignment 2.pdf` : Description of problem statement
- `Train_B.csv` : Contains the data used for training the Naive-Bayes Classifier
- `data_processor.py` : Contains the helper functions related to the processing of the data
- `main.py` : Contains the solution to problems provided in the `Assignment 2.pdf`
- `model.py` : Contains all the necessary functions and implementation of Naive-Bayes Algorithm for Classification problems
- `requirements.txt` : Contains all the necessary dependencies and their versions
- `simulation.txt` : Sample simulation output on entire data (it is advisable to train on a subset of input data to obtain results in less time) - `utils.py` : Contains all the helper functions used by the above files (if any)
- `variance_ratio_cumulative_sum.png` : The plot of variance ratio cumulative sum **vs** number of princial components
- `variance_ratio_PCA.png` : The plot of proportion of variance explained **vs** Principal Component

Directions to use the code

1. Download this directory into your local machine
2. Copy the file `Train_B.csv` to the directory where the code resides
3. Ensure all the necessary dependencies with required version and latest version of Python3 are available (verify with `requirements.txt`)
   ```
   pip3 install -r requirements.txt
   ```
4. Run specific functions with the aid of `main.py`

For giving the input fraction (the fraction of dataset to be used for the model)

- Using the default full dataset
  ```
  python3 main.py
  ```
- Giving input fraction (say 0.1, that is 10% of the dataset randomly chosen) -- fraction should be between 0 and 1
  ```
  python3 main.py --frac 0.1
  ```
- For more help regarding the arguments
  ```
  python3 main.py --help
  ```

For giving the input outlier threshold

- If the number of outliers in a sample is more than threshold then the sample will be dropped
- Using the default threshold
  ```
  python3 main.py
  ```
- Default threshold is the maximum value of the outlier in the dataset (that is 3 for the given dataset)
- Giving input outlier threshold
  - (say 2, that is samples having number of outliers greater than or equal to 2 will be dropped)
  - outlier should be positive integer

  ```
  python3 main.py --outlier 2
  ```
- For more help regarding the arguments
  ```
  python3 main.py --help
  ```