1. The Enron dataset is popular worldwide, and it was used in a legal investigation after the Enron’s collapse to determine people of interest using the available data. The data provided was either financial features or email features, and using machine learning science and applying the proper algorithms along with fine tuning, it was possible to gain a better insight of people of interest inside the corporation, using AI to grab a common sense of what groups of features an Enron employee responsible of Enron’s collapse may lie within.

   Total number of data points were 146 and number of features are 17. It happens to be outliers in the dataset represented in the ‘TOTAL’ and for more efficient implementation ‘THE TRAVEL AGENCY IN THE PARK’ was included as an outlier, and both were visualized and removed from the dataset before the training process. There were some features with much missing data and consequently they were removed from the dataset and not considered in further processes.

2. Features used: salary, total payments, bonus, total stock value, expenses, exercised stock options, from-poi percentage, to-poi percentage (2 latter features were calculated from original email features), shared receipt with poi and mean-top-2000-80%-tfidf-sums (calculated). The features were selected upon 2D visualizations and the ones with the least missing data. Feature scaling was implemented since several algorithms were tried at first (requiring scaling and normalization of features to ensure efficient assessment of an algorithm), although the ‘DecisionTree’ algorithm used at the end do not necessarily require feature scaling. From-poi and to-poi email percentages were calculated from email features, which upon it was believed a percentage of the latter calculation would be directly proportional to the probability of a person being in interest or corrupt. Mean-top-2000-80%-tfidf-sum feature was calculated to grab an insight of how emails from a poi would be constructed and it was implemented using term frequency – inverse document frequency.

   Feature importance;
   salary, total payments, bonus, total stock value, expenses, exercised stock options, from-poi percentage, to-poi percentage, shared receipt with poi and mean-top-2000-80%-tfidf-sums: 0.1245, 0.3777, 0.158, 0, 0.0435, 0, 0.0831, 0 0.2132 and 0, respectively.

   Using grid search CV, the following results were obtained:
   {‘DT__criterion’: ‘gini’, DT__min_samples_split’: 2, ‘Kbest__k’: 10, ‘pca_ _n_components’: 10}

3. Multiple algorithms were tried and the final one used was decision tree. The following results were obtained.

DecisionTree:

```
Accuracy: 0.84087Precision: 0.39160     Recall: 0.34950  F1: 0.36935
      F2: 0.35718
Total predictions: 15000     True positives:  699   False positives:
1086  False negatives: 1301  True negatives: 11914
```

SVC:

```
Accuracy: 0.86307Precision: 0.15385     Recall: 0.00600  F1: 0.01155
      F2: 0.00743
Total predictions: 15000     True positives:   12   False positives:
66    False negatives: 1988  True negatives: 12934
```

KNeighborsClassifier:

```
Accuracy: 0.85827Precision: 0.00781     Recall: 0.00050  F1: 0.00094
      F2: 0.00062
Total predictions: 15000     True positives:    1   False positives:
127   False negatives: 1999  True negatives: 12873
```

Pipeline was used along with grid search and several parameters were used to be tuned:

1. PCA (was selected later manually to be ‘all’ since it resulted to a better performance)
2. Kbest (was selected later manually to be ‘all’ since it resulted to a better performance)
3. DT criterion
4. DT min samples split

Parameter tuning is essential since different values to various parameters might put specific weights or certain specifics to process an algorithm leading to different performances.

4. Parameter tuning is performed by choosing specific values for the various options to process an algorithm indicating certain weights or/and techniques so that the program behavior shall vary accordingly. If it happens that bad tuning has been performed the program might behave in an opaque manner that would lead to bad performance. Parameter tuning was performed via grid search CV and few parameters were chosen manually, after trying multiple values. Algorithms parameter being tuned:

    - DT criterion
    - DT minimum split

5. Validation is the process of an algorithm being evaluated using a testing set. Things might go wrong during validation is; having unshuffled training/testing set or training the algorithm using really small dataset which isn't sufficient to train a program into a high-bias system. Validation method used in this project was a training/testing dataset split, using the testing set for evaluation, and cross validation was performed using gridsearchCV, as well the use of PCA to choose the principle components (although it was chosen -manually- to be 'all' at the final version of the program).

6. Evaluation metrics used were precision, recall and F1 score.

```
Accuracy: 0.84087Precision: 0.39160     Recall: 0.34950  F1: 0.36935
     F2: 0.35718
Total predictions: 15000     True positives:  699   False positives:
1086  False negatives: 1301  True negatives: 11914
```

Precision = true positives / (true positives + false positives)

Recall = true positives / (true positives + false negatives)

F1 score = (2 * precision * recall) / (precision + recall)

Precision is the fraction of the data classified correctly as positive among all data classified as positives whether being true or false. Recall is the fraction of the data classified correctly as positives among all originally true data. F1 score is an evaluation method that grabs a sense of both precision and recall.