# Data Understanding

## Data Exploration

## Outliers

# Algorithm Selection

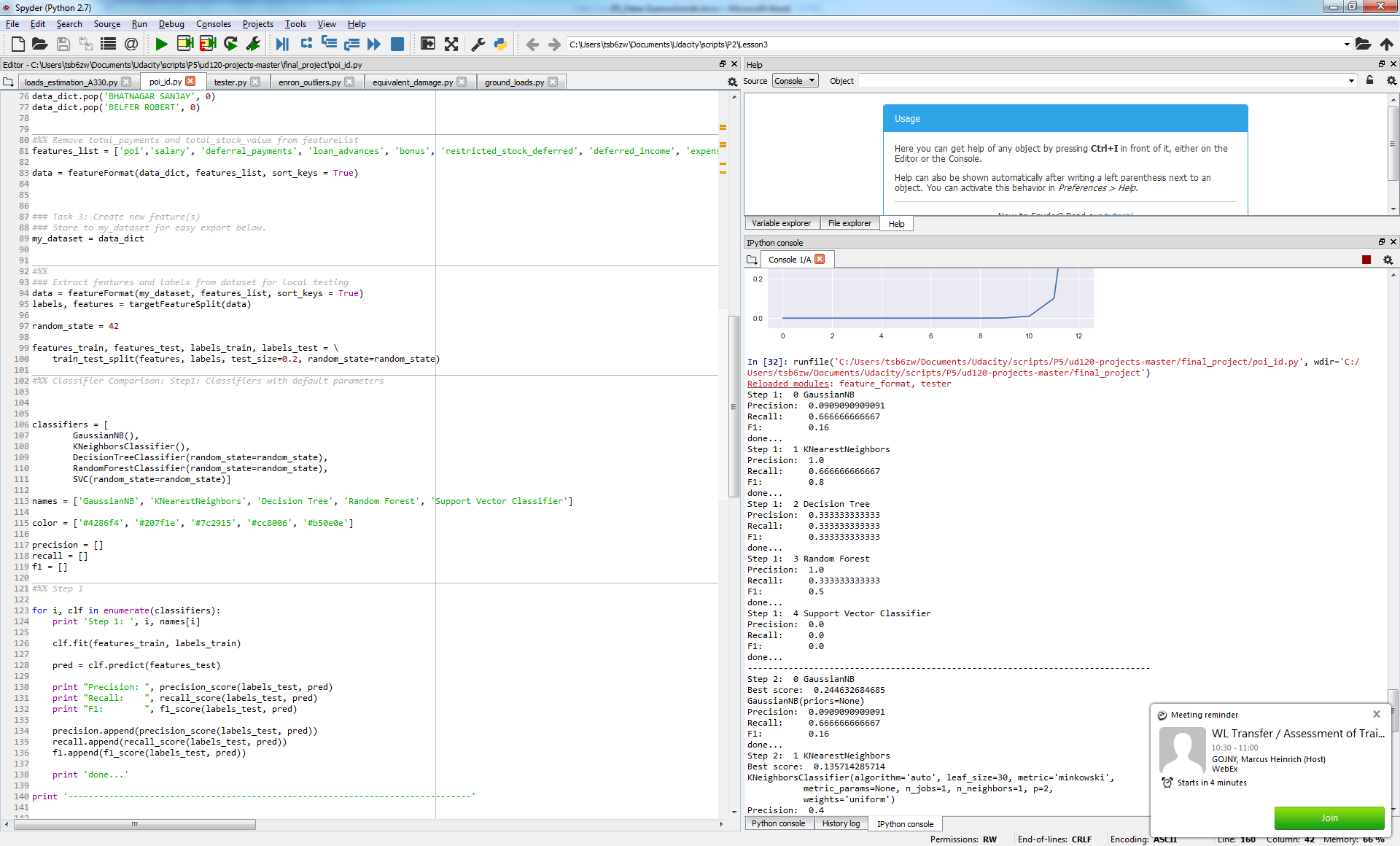
Algorithm selection is done in several steps:

1. Split dataset into a training and a test set. Here, this is done with train\_test\_split; the size of the test set is .2
2. Train classifiers with default parameters (no tuning) on the training set
3. Calculate precision, recall, and f1 score based on test set
4. No feature selection at this point

Here 5 different algorithms are selected:

1. Gaussian Naives Bayes
2. K Nearest Neighbors
3. Decision Tree
4. Random Forest
5. Support Vector Classifier

The results of this first run are shown in the following figure:



# Feature Selection

# Final Validation

In the previous section it was seen that the results in terms of precision and recall differ when simply using train\_test\_split and when using StratifiedShuffleSplit.

This shows the importance of the validation, as the performance may depend a lot on the selected test set.

The metrics used to evaluate the performance are precision and recall. What this means for this project is as follows:

* Low Recall: Number of False Negatives is too high, i.e. a lot of employees are predicted not to be a POI whereas in reality they were
* Low Precision: Number of False Positive is too high, i.e. a lot of employed are predicted to be a POI whereas in reality they were not.
* In both cases the number of True Positives may also be too low, i.e. the number of correctly identified POIs

It is arguably which one is more important but I would tend to achieve a better recall than precision, as it seems better to falsely identify someone as POI (and to exonerate them later) than to miss a potential POI. On the other hand, … So f1