1. The goal of this project was to explore machine learning techniques and apply them to the Enron Email dataset. The dataset consists of 146 executives and outlines people identified as POIs (Persons of Interest). A POI was someone who was indicted for fraudulent activity, reached a settlement, or testified for immunity. The data was mostly clean but had some gaps that needed to be handled in order to make processing and analyzing easier. This will be explained below.
2. Features that I was interested in using were messages to and from a POI. I also created the feature "Bonus to Salary Ratio" that explored the relationship between the bonus an employee was given and their respective salary.
3. I ended up favoring the Gaussian Naïve Bayes algorithm. It is an algorithm that is simple to work with and easy to understand. For a beginner in this field, I wanted to choose something that allowed me to easily understand the "behind the scenes" on the data manipulation that was occurring. This will allow me to scale up complexity in the future without feeling overwhelmed. Additionally, processing time for this algorithm was small which allowed me to quickly make changes in my code and see the results.
4. Tuning the parameters of an algorithm means to tweak specific settings to get an output that fits the relationship that you are trying to explore. Tuning can also mean time optimization (i.e. processing time) or modifying parameters to handle outliers or non-standard values.
5. Validation is the process of feeding test data into your algorithm to test how your algorithm responds to real-world data. Validation is important to ensure that your algorithm is behaving as you expect. This plays a larger factor when your data size increases drastically or when you want to add more variables/features to analyze. One mistake is not using a representative sample of data to validate your algorithm. Using a non-representative sample can lead to false positives or false negatives in your analysis of your data.