Thyroid dataset

1. How many records are there in the dataset? How many features are there? Are there any missing values in the data? What is the distribution of 'class' label?
   There are 215 records in total, 6 features ('T3_resin', 'Serum_thyroxin', 'Serum_triiodothyronine', 'Basal_TSH', 'Abs_diff_TSH', 'Outcome').

   There are no missing values.

   For outcome feature, there are 150's 1, 35's 2, and 30's 3. Outcome = 1 is the most frequent case.

2. Give the histograms of each feature, What do you perceive after examining the plot?
   Serum_thyroxin, Serum_triiodothyronine, Basal_TSH, and Abs_diff_TSH are all rigth-skewed.

3. What is the testing strategy deployed?
   We use K-Fold cross validation.

4. Report mean accuracy for different models using K-Fold cross validation.
   KNN 0.925758; DT 0.953463; GNB 0.967749; BNB 0.734848;

5. Intuition developed by running the notebooks?
   Bernoulli Naive Bayes assumes that all our features are binary such that they take only two values. It shows the worst accuracy in iris_dataset, since all of four features = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width'] are continuous instead of binary.

   Gaussian Naive Bayes is used in cases when all our features are continuous. For example in Iris dataset features are sepal width, petal width, sepal length, petal length. So its features can have different values in data set as width and length can vary. We can’t represent features in terms of their occurrences.