Iris dataset

1. How many records are there in the dataset? How many features are there? Are there any missing values in the data?
   There are 150 records in total, 5 features. There are no missing values.

2. What's the distribution of the "Outcome" feature? Is this feature skewed? What are the distributions for other features?
   There are 50 Iris-setosa, 50 Iris-versicolor, and 50 Iris-virginica. The feature is not skewed at all. It was evenly distributed. The distribution of other features is not clear.

3. Give the pair plot, What do you perceive after examining the plot? Plot heatmap which shows correlations between four features, What do you perceive after examining the plot?
   The distributions for petal-length and petal-width is very different for each outcome. Petal-length and petal-width are strongly correlated. Sepal-length and petal-length are correlated with less certainty. petal-width and sepal-length are correlated with less certainty.

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

5. Report mean accuracy for different models using K-Fold cross validation.
   KNN 0.966667; DT 0.953333; BNB 0.333333; GNB 0.953333

6. 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.