Diabetes dataset

1. How many records are there in the dataset? How many features are there? What is the "class" label?Are there any missing values in the data? Are there any unexpected outliers?
   There are 768 records in total, 9 features ('Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'], dtype='object')
   The class label is "Outcome". There is no missing values in the data.

   Yes, there are some outliers, or errant values that needs to be excluded from the dataset. There are 35 records BloodPressure == 0, 5 records that Glucose == 0, 11 records diabetes.BMI == 0

2. How many records are there in the dataset after data cleaning?
   There are 724 cases.

3. What's the distribution of the "Outcome" feature? Is this feature skewed?
   The "Outcome" feature is slightly skewed. There are 500 negative cases and 268 positive cases.

4. Plot the histogram of each feature's distribution group by 'Outcome' feature. What do you perceive after examining the plots?
   The distribution of Glucose, and Age features for differnt group is very different. The distributions of other features are different for each group, not as much as Glucose and Age.

5. What is the testing strategy deployed?
   Before training models, we split the data into Train/Test group. Then we train each model (KNN, DT, GNB, BNB), use x_test to yield a prediction (y_pred). At last, we test the model against y_test.

6. Report accuracy for different models using holdout dataset.
   KNN 0.729282; DT 0.756906; GNB 0.734807; BNB 0.657459;

7. Report mean accuracy for different models using K-Fold cross validation. How was it compared with testing against holdout dataset?
   KNN 0.714136; DT 0.696461; GNB 0.754205; BNB 0.656069 ;
   KNN and DT's accuracy dropped a little, BNB and GNB remains same.

8. Intuition developed by running the notebooks?
   Decision Tree's performance is a bit worse than other classifiers in this case. There are many possible reasons. The trees are very sensitive to the noise in input data; the whole model could change if the training set is slightly modified (e.g. remove a feature, add some objects).

   For kNN algorithms, If a dataset has many variables, it is difficult to find the right weights and to determine which features are not important for classification/regression. We could add more weights on glucose and age features accordning to our observation from the plots