Bansri Rawal
CMSC 435
Assignment 2

2a) What are MAE values for the eight results? You should simply copy he output from the screen.

MAE VALUES:
MAE_01_mean: .0809
MAE_01_mean_conditional: .0796
MAE_01_hd: .0607
MAE_01_hd_conditional: .06
MAE_10_mean: .0832
MAE_10_mean_conditional: .0822
MAE_10_hd: .0657
MAE_10_hd_conditional: .0653

2b) Which of the considered four imputation methods is the most accurate for the *dataset-missing01.csv* dataset? Explain potential reasons why.

The conditional hot deck imputation method is the most accurate for the missing01 dataset. There are many reasons for this better accuracy. The hot deck imputation method allows the calculation of an imputed value from a row that is the most similar and representative of the row that is being compared. In addition, making sure that both rows are from the same class (Y/N) allows the imputed value to be more representative of the actual data value.

2c) Are the results for the same algorithm on the two datasets same/worse/better. Explain why.

The error values of the algorithms for the missing01 dataset are smaller than the error values of the algorithms for the missing10 dataset. This can be attributed to the fact that the algorithms have more data to use for the calculations with the missing01 dataset. With more data, the algorithm can calculate an imputed value that can be more representative of the dataset whereas the imputed value would not be as representative of the entire dataset if 10% of the values are missing because it has less data to work with.

2d) Which of the two unconditional methods is faster, requires fewer computations. Briefly explain why. Give computational complexity of both methods as a function of the number of objects *n*, and use it to support your explanation.

The calculation of the mean method is most certainly faster and requires fewer computations. The mean calculations run at $O(n*m)$ because an outer for loop is necessary to iterate through the rows and an inner for loop is necessary to iterate through each column and compute the mean. For the mean method, comparisons between the dataset are not made which is why it runs faster. The hot deck calculation method is slower and requires many more computations. The hot deck calculations run at $O(n^2*m)$ run-time. The n^2 is because it is necessary to have an outer loop that iterates through the rows as well as a nested loop that iterates through the rows for which Euclidian distance will be calculated and compared. The m is because there is a nested loop to iterate through the columns for the calculation of the Euclidian distance and simply another loop to replace the missing values with imputed values.