**Big Data Analytics & Management**

**Project #5**

**Matthew S. Hartstein**

Introduction:

For this assignment, I decided to use three pre-made datasets that are included in the ‘sklearn’ Python library: (1) a wine dataset, (2) a numerical digits dataset, and (3) a breast cancer dataset. The reason why I decided to do this is because I was having some trouble pre-processing the three datasets that I had chosen from Kaggle.com. In summary, I was able to successfully complete this assignment and showed proper results for each of the three datasets that were used.

Results:

A screenshot of a cell phone

Description automatically generated

**Figure #1A: Wine Dataset Description**

A screen shot of a computer

Description automatically generated

**Figure #1B: Wine Dataset Results**

A screenshot of a cell phone

Description automatically generated

**Figure #2A: Digits Dataset Description**

A picture containing window, table, rain, large

Description automatically generated

**Figure #2B: Digits Dataset Results**

A screenshot of a cell phone

Description automatically generated

**Figure #3A: Breast Cancer Dataset Description**

A screenshot of a cell phone

Description automatically generated

**Figure #3B: Breast Cancer Dataset Results**

Conclusion:

After examining my source code, it is important to note that the user must comment out 2 out of 3 datasets in order to compile and return each of the three results; my code does not do all of this at once. I chose this design decision because it made it easier to implement and to debug the code/datasets to ensure high quality conclusions. Each declaration is located on lines 7, 8, and 9.

You will also notice that my code successfully compiles for each of the three pre-define datasets; there should be no issue running my code for each of the datasets; there are no errors for all three datasets that were used for this assignment. As a side note, Python 3.x was used for my program.

In total, I feel very proud and accomplished after implementing the source code for this assignment. I feel very confident in my abilities to understand and implement a Naïve Bayesian Classification Algorithm to analyze various datasets and how important and useful this algorithm is to formulate models and to make expected results, as well as making predictions from datasets.