# Introduction

This file is a comparative analysis on Cycon’s ability to perform BernoulliNB classification. This serves as proof that the Cycon page is able to perform BernoulliNB. The following shows BernoulliNB results for various datasets.

|  |  |
| --- | --- |
| **Dataset:** | |
| Shape: 569 x 31  Samples: 357 Benign and 212 Malignant  Classes: B (Benign), M (Malignant)  Purpose: Utilizing gathered medical data, determine if the patient has cancer. | |
| **Comparative Work:**  [**https://www.kaggle.com/code/sandeepchamunni/bernoulli-naive-bayes-classifier/notebook**](https://www.kaggle.com/code/sandeepchamunni/bernoulli-naive-bayes-classifier/notebook) | **Cycon Work:** |
| **Settings:** | |
| Selects the features and adds column titles. (More about this process below)    Utilizes Binarizer to preprocess the data. (More about this process below)    Split the data set    Create and use BrnoulliNB model |  |
| **Results:** | |
|  |  |
| **Any Additional Information:** | |
| Note that there is three versions of dataset for this particular experiment. wdbc, wdbc\_mini, and wdbc\_Preopt.   * wdbc contains the original csv file with header names for each column. The reset contains alterations that was done in the comparative work but saved as a csv to avoid having to do the same alterations every time. * wdbc\_mini contains the csv with only the features columns used in the comparative work. Note that they chose to use Feature\_2 to Feature\_12, removing around 20 columns of data. This is all that remains in wdbc\_mini   Example of this process below.     * wdbc\_Preopt takes the dataset from wdbc\_mini and performs the same preoptimization. Note that the comparative work goes through each column, calculates the mean, then uses the mean as an input for Discretization Binarizer Optimization for the column. This change for all columns was done in wdbc\_Preopt.   Example of this process below.      Note that all these changes to the csv can be done in cycon website by using the Preoptimization selection to remove the columns. Then calculate the mean via an excel sheet and using Binarizer on each column with the corresponding mean.  Note that while the user utilizes random seed, they didn’t do so with a shuffle, as such we can obtain very close but not exactly the same results. | |

## wdbc.csv