Lab2: Using Scikit-Learn

CPSC429/529 Machine Learning

In this lab assignment, you are given a breast cancer dataset (`breast_cancer.csv`) and do classification and dimensionality reduction. Specifically, you will do the followings:

1. Use `GaussianNB` classifier to build the model on the training dataset, predict on the testing dataset, compute the prediction accuracy, and **print out** the prediction accuracy.

2. Use `PCA` to reduce original $\mathbf{X}$'s dimensions into **2** dimensions (`n_components = 2`), add these two transformed column data (two principle components) into your dataframe (which should contain the target column of `diagnosis`). Now, do a scatter plot these two new principle components, separated by `diagnosis`. Your plot should look like the following plot.

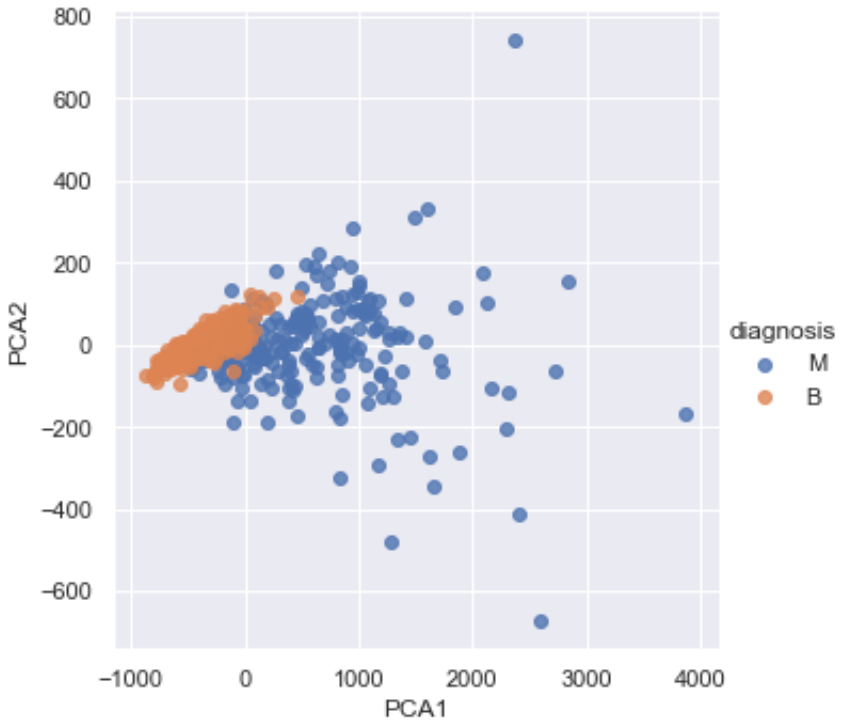

Figure 1: Scatter plot of Lab 1.

The jupyter notebook skeleton of lab 2 (`Lab2.ipynb`) is given to you, so you can complete the remaining parts.