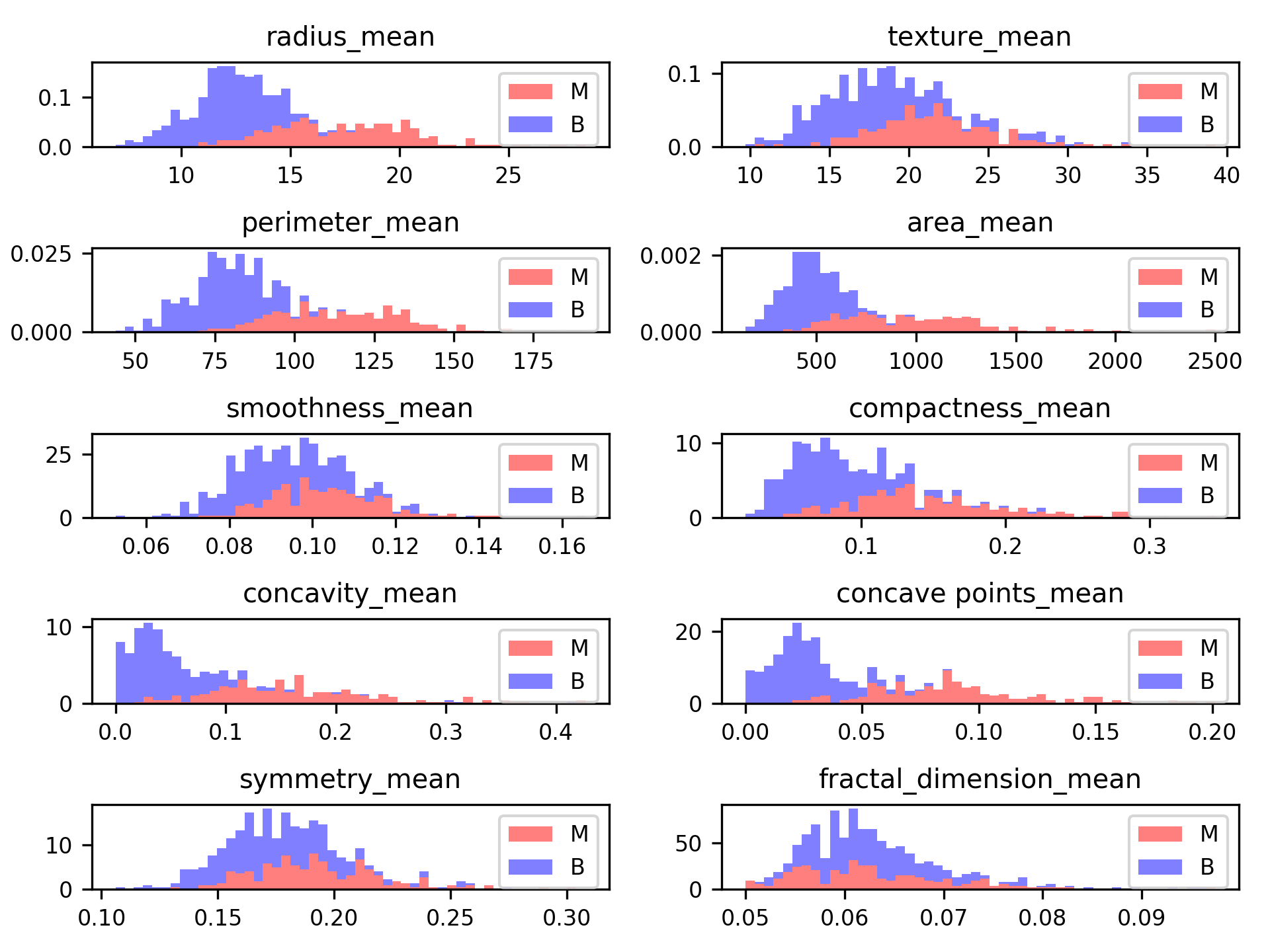
Luke Dosen

Machine Learning HW 3

**Dataset and Preprocessing:**

For this homework, I decided to use the breast cancer Wisconsin data set retrieved from this address: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>. The dataset includes information about the mean, standard deviation, and “worst case” for the following 10 features. The following histograms reveal the features that matter the most in malignant classification (radius, perimeter, area, compactness, concavity, and concave points) It is evident through these graphs that there are also no outliers that need to be dealt with in the data set. I applied a standard scalar to the feature sets to standardize the data before training time.

**Classifier and Parameter Tweaking:**

After the data was cleaned and normalized, an SVM model was fit to the training data using the default parameters. This yielded an accuracy score of ~96%. First the C value was tweaked; increasing the C value to 5.0 increased the accuracy score to ~98%, somewhat counter to my expectations. Next, I tried using a linear kernel as opposed to the default rbf kernel. This dropped the accuracy to around ~95%. Thus, the optimal settings for the SVM were a C value of 5.0 using an rbf kernel. The KNN model I implemented was the most accurate when only the closest neighbor was examined. When tested on the test set, the KNN model with k = 1 had an accuracy of ~95% while the SVM with the above settings achieved 100% accuracy.