## Comparing classifiers: Divorce classification/prediction

This dataset is about participants who completed the personal information form and a divorce predictors scale. The data is a modified version of the publicly available at https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set (by injecting noise so you will not get the exactly same results as on UCI website). The dataset marriage.csv is contained in the homework folder. There are 170 participants and 54 attributes (or predictor variables) that are all real-valued. The last column of the CSV file is label y (1 means "divorce", 0 means "no divorce"). Each column is for one feature (predictor variable), and each row is a sample (participant). A detailed explanation for each feature (predictor variable) can be found at the website link above. Our goal is to build a classifier using training data, such that given a test sample, we can classify (or essentially predict) whether its label is 0 ("no divorce") or 1 ("divorce").

For this problem, first the data was split into training and testing sets (80/20 split) and then Naive Bayes, Logistic Regression, and KNN models were fit using the training feature and response data. Afterwards, predictions were created using the fitted models on the test attribute data, and the accuracy on the test response data for each model was reported. These can be seen below:

- Naive Bayes: Test accuracy = 0.9411764705882353
- Logistic Regression: Test accuracy = 0.9411764705882353
- KNN: Test accuracy = 0.9411764705882353

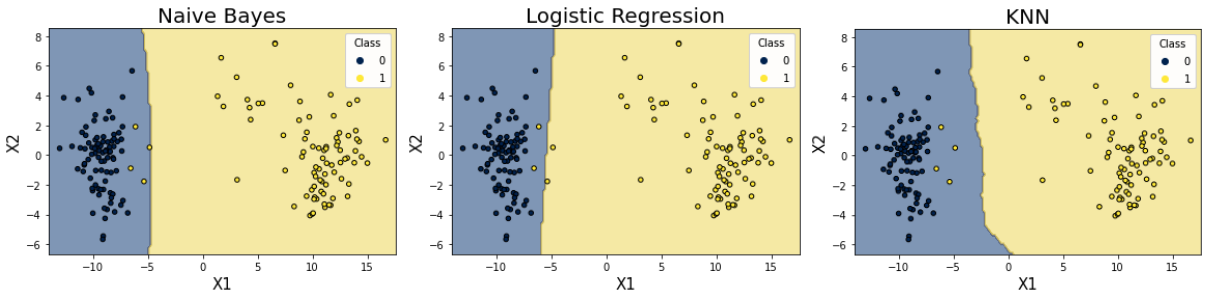
Notably, all these accuracies are quite high and the same. The fact that all three accuracies are high suggests that the classes are easily separable. The fact that there all the accuracies are the same, however, shows that there may be a few data points which are more difficult to classify that all of the models struggle with. To further examine which model performs the best, the training accuracies for each model were also examined. These can be seen below:

- Naive Bayes: Train accuracy = 0.9852941176470589
- Logistic Regression: Train accuracy = 1.0
- KNN: Train accuracy = 0.9852941176470589

From this, it can be seen that the accuracy of the logistic regression model is higher than that of the Naïve Bayes and KNN models, which still have the same accuracy as each other. This shows that the logistic regression model may perform a little better than the Naïve Bayes and KNN models, though overfitting is certainly possible.

When considering the assumptions of each model, some conclusions can be drawn about why logistic regression may perform the best in this setting. The Gaussian Bayes classifier assumes a Gaussian distribution of the data, KNN takes a geometric interpretation of the data, and logistic regression uses a linear decision boundary. Given that logistic regression performed slightly better than the other two models, it can be assumed that the data may be linearly separable, which is why the linear decision boundary used in logistic regression performed slightly better than the non-linear decision boundaries of the other two models.

After performing PCA on the data and projecting it into a two-dimension space, the models were refit using the two-dimensional PCA results on the training set. Plots were created of the decision boundary, along with the correct label for each data point in the same color. This allowed for the accuracy of each model to be examined, as it can be seen which potentially troublesome points were causing the test accuracies to be the same. These plots can be seen below:




From these plots, it can be seen that there are two distinct clusters of classes 0 and 1, but with four data points in class 1 which are more closely clustered with class 0. These distinct clusters show why each model's accuracy was so high as it easy to separate the classes. Additionally, the four data points in class 1 which are more closely clustered with class 0 show why each model had the same test accuracy, as none of the models were able to successful classify these points. Furthermore, the decision boundary of each model can be seen, and given that the data are mostly linearly separable, it can be seen why logistic regression may have performed marginally better than the other two models without linear decision boundaries.