Homework 3
Zack Spears
Data Mining

1. I chose to use chi squared test and pearson correlation for my feature selection methods.

   For Chi-Squared, you need ranges to work with since these are real valued variables for sepal and petal length and width. I will determine ranges which work well and then test those ranges at the groups.
   Once ranges are found, we make the table of values from the chi-squared example. From the table we multiply the totals from each matching and divide by the overall total to get the expected value. We then subtract the observed from the expected, square that and divide by the expected. Finally we sum up all of the values for each feature. Since we are assuming there is no relationship, the larger value the better, since this makes the null hypothesis less likely to be true. We choose the 2 features then with the highest chi-squared value as our features to work with.

   For the pearson correlation, the data needs to be in the form of numbers, so that the calculation can be done. Since everything but the final classification is real valued, this is simple to do by assigning the three classes the values of 0,1 and 2 respectively. From there we find the mean and standard deviation of each feature and of the classifications and the plug into the formula for the coefficient. The values closest in absolute value to 1 are the best, so we will take the 2 with absolute values closest to 1.

2. I did the work on the attached Excel sheet. On sheet 1 is the chi-squared and sheet 2 has the pearson correlation coefficient. Both the tests suggest the the petal length and width are better predictors than the sepal length and width.