**Homework 4 Log:**

**Assumptions:**

* To clean our auto-data we removed every row with an NA value.
* Step 1: We are going to create a classifier table, like page 27 in the book.
* Step 2: We will create a classifier table similar to step 1.
* Step 3: We will not need to clean the titanic data

**Issues:**

* Step 1: Started trying to make a list of multiple lists to create the classifier table, but then decided to use dictionaries.
* Step 1: Instead of building a table exactly like on page 27, we are partitioning each class by the label values.
* Step 1: The standard data type float was not accurate enough, so we started using decimal
* Step 2: Gaussian distribution not completely understood at first
* Step 3: We noticed that the label listing function get\_categories had a bug, and we fixed it
* Step 3: Creating the confusion matrices takes a long time because of how much data there is in the titanic data set, so for testing purposes we used a portion of the data

**Results:**

**===========================================**

**TEST 1: Naive Bayes MPG Classifier**

**===========================================**

**predicted: 4 actual: 5**

**predicted: 8 actual: 9**

**predicted: 7 actual: 7**

**predicted: 2 actual: 2**

**predicted: 7 actual: 7**

**For k=5 Random Subsampling:**

**accuracy: 0.6 , error\_rate: 0.4**

**MPG 1 2 3 4 5 6 7 8 9 10 Total Recognition (%)**

**----- --- --- --- --- --- --- --- --- --- ---- ------- -----------------**

**1 23 2 0 0 0 0 0 0 0 0 25 92**

**2 12 15 1 4 0 0 0 0 0 0 32 46.875**

**3 5 4 10 7 0 0 0 0 0 0 26 38.4615**

**4 1 3 10 29 6 5 0 0 0 0 54 53.7037**

**5 0 0 3 13 20 9 1 0 0 0 46 43.4783**

**6 0 0 0 0 5 11 9 0 0 0 25 44**

**7 0 0 0 0 4 4 17 3 0 0 28 60.7143**

**8 0 0 0 0 1 0 6 8 0 0 15 53.3333**

**9 0 0 0 0 0 0 1 1 0 0 2 0**

**10 0 0 0 0 0 0 0 0 0 0 0 0**

**===========================================**

**TEST 2: Naive Bayes MPG Classifier**

**===========================================**

**predicted: 1 actual: 1**

**predicted: 5 actual: 5**

**predicted: 4 actual: 4**

**predicted: 4 actual: 4**

**predicted: 4 actual: 2**

**For k=5 Random Subsampling:**

**accuracy: 0.8 , error\_rate: 0.2**

**MPG 1 2 3 4 5 6 7 8 9 10 Total Recognition (%)**

**----- --- --- --- --- --- --- --- --- --- ---- ------- -----------------**

**1 23 2 0 0 0 0 0 0 0 0 25 92**

**2 13 14 1 4 0 0 0 0 0 0 32 43.75**

**3 5 4 10 7 0 0 0 0 0 0 26 38.4615**

**4 1 3 5 36 3 4 2 0 0 0 54 66.6667**

**5 0 0 2 18 19 4 2 1 0 0 46 41.3043**

**6 0 0 0 1 7 8 9 0 0 0 25 32**

**7 0 0 0 0 7 5 13 3 0 0 28 46.4286**

**8 0 0 0 0 4 1 4 6 0 0 15 40**

**9 0 0 0 0 1 0 0 1 0 0 2 0**

**10 0 0 0 0 0 0 0 0 0 0 0 0**

**===========================================**

**TEST 3: Naive Bayes Survival Classifier (Titanic)**

**===========================================**

**predicted: yes actual: yes**

**predicted: yes actual: no**

**predicted: yes actual: no**

**predicted: yes actual: yes**

**predicted: no actual: no**

**predicted: yes actual: yes**

**predicted: yes actual: yes**

**predicted: yes actual: yes**

**predicted: yes actual: no**

**predicted: no actual: no**

**For k=10 Random Subsampling:**

**accuracy: 0.7 , error\_rate: 0.3**

**SURVIVED yes no Total Recognition (%)**

**---------- ----- ---- ------- -----------------**

**yes 242 19 261 92.7203**

**no 63 75 138 54.3478**

**accuracy: 0.794486215539 error\_rate: 0.205513784461**

The Stratified k-Fold Nearest Neighbor algorithm takes an excessive amount of time, and will not complete within our lifetimes. So we took a subset of the Titanic data (400 rows) which takes a lot less time to run. These are the results of that.

===========================================

TEST 3: Stratified Survival Classifier (Titanic)

===========================================

SURVIVED yes no Total Recognition (%)

---------- ----- ---- ------- -----------------

yes 270 2340 2610 10.3448

no 30 1350 1380 97.8261

accuracy: 0.406015037594 error\_rate: 0.593984962406