Weka (nomRating.csv)

1. Use training set (Kernel Estimator)
    a. Classifier → bayes → NaiveBayes (Use training set)

```
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6384
Mean absolute error                      0.2227
Root mean squared error                  0.3276
Relative absolute error                 54.7875 %
Root relative squared error             72.687  %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.087    0.765      0.712    0.737      0.639    0.906
0.807     High Hospital Quality
                 0.913    0.288    0.889      0.913    0.901      0.639    0.906
0.959     Low Hospital Quality
Weighted Avg.    0.856    0.231    0.854      0.856    0.855      0.639    0.906
0.916

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  211 2224 |    b = Low Hospital Quality
```

2. Use training set (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (Use training set)

```
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6384
Mean absolute error                      0.2227
Root mean squared error                  0.3276
Relative absolute error                 54.7875 %
Root relative squared error             72.687  %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.087    0.765      0.712    0.737      0.639    0.906
0.807     High Hospital Quality
                 0.913    0.288    0.889      0.913    0.901      0.639    0.906
0.959     Low Hospital Quality
Weighted Avg.    0.856    0.231    0.854      0.856    0.855      0.639    0.906
0.916

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  211 2224 |    b = Low Hospital Quality
```

3. Use training set
    a. Classifier → bayes → NaiveBayes (Use training set)

```
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6384
Mean absolute error                      0.2227
Root mean squared error                  0.3276
Relative absolute error                 54.7875 %
Root relative squared error             72.687  %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.087    0.765      0.712    0.737      0.639    0.906
0.807     High Hospital Quality
                 0.913    0.288    0.889      0.913    0.901      0.639    0.906
0.959     Low Hospital Quality
Weighted Avg.    0.856    0.231    0.854      0.856    0.855      0.639    0.906
0.916

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  211 2224 |    b = Low Hospital Quality
```

4. 10-fold CV (Kernel Estimator)
    a. Classifier → bayes → NaiveBayes (10-fold cross-validation)
        i. useKernelEstimator

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2906               85.4957 %
Incorrectly Classified Instances       493               14.5043 %
Kappa statistic                          0.6359
Mean absolute error                      0.225
Root mean squared error                  0.3303
Relative absolute error                 55.3716 %
Root relative squared error             73.2723 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.088    0.761      0.712    0.736      0.637    0.903
0.801     High Hospital Quality
                 0.912    0.288    0.889      0.912    0.900      0.637    0.903
0.958     Low Hospital Quality
Weighted Avg.    0.855    0.232    0.853      0.855    0.853      0.637    0.903
0.913

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  215 2220 |    b = Low Hospital Quality
```

5. 10-fold CV (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (10-fold cross-validation)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2906               85.4957 %
Incorrectly Classified Instances       493               14.5043 %
Kappa statistic                          0.6359
Mean absolute error                      0.225
Root mean squared error                  0.3303
Relative absolute error                 55.3716 %
Root relative squared error             73.2723 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.088    0.761      0.712    0.736      0.637    0.903
0.801     High Hospital Quality
                 0.912    0.288    0.889      0.912    0.900      0.637    0.903
0.958     Low Hospital Quality
Weighted Avg.    0.855    0.232    0.853      0.855    0.853      0.637    0.903
0.913

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  215 2220 |    b = Low Hospital Quality
```

6. 10-fold CV

a. Classifier → bayes → NaiveBayes (10-fold cross-validation)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2906               85.4957 %
Incorrectly Classified Instances       493               14.5043 %
Kappa statistic                          0.6359
Mean absolute error                      0.225
Root mean squared error                  0.3303
Relative absolute error                 55.3716 %
Root relative squared error             73.2723 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.712    0.088    0.761      0.712    0.736      0.637    0.903
0.801     High Hospital Quality
                 0.912    0.288    0.889      0.912    0.900      0.637    0.903
0.958     Low Hospital Quality
Weighted Avg.    0.855    0.232    0.853      0.855    0.853      0.637    0.903
0.913

=== Confusion Matrix ===

    a    b   <-- classified as
  686  278 |    a = High Hospital Quality
  215 2220 |    b = Low Hospital Quality
```

7. 3-fold CV (Kernel Estimator)
    a. Classifier → bayes → NaiveBayes (3-fold cross-validation)
        i. useKernelEstimator

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6377
Mean absolute error                      0.2252
Root mean squared error                  0.3308
Relative absolute error                 55.4173 %
Root relative squared error             73.3894 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.709    0.085    0.767      0.709    0.736      0.639    0.901
0.799     High Hospital Quality
                 0.915    0.291    0.888      0.915    0.901      0.639    0.901
0.957     Low Hospital Quality
Weighted Avg.    0.856    0.233    0.854      0.856    0.854      0.639    0.901
0.912

=== Confusion Matrix ===

    a    b   <-- classified as
  683  281 |    a = High Hospital Quality
  208 2227 |    b = Low Hospital Quality
```

8. 3-fold CV (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (3-fold cross-validation)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6377
Mean absolute error                      0.2252
Root mean squared error                  0.3308
Relative absolute error                 55.4173 %
Root relative squared error             73.3894 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.709    0.085    0.767      0.709    0.736      0.639    0.901
0.799     High Hospital Quality
                 0.915    0.291    0.888      0.915    0.901      0.639    0.901
0.957     Low Hospital Quality
Weighted Avg.    0.856    0.233    0.854      0.856    0.854      0.639    0.901
0.912

=== Confusion Matrix ===

    a    b   <-- classified as
  683  281 |    a = High Hospital Quality
  208 2227 |    b = Low Hospital Quality
```

9. 3-fold CV

a. Classifier → bayes → NaiveBayes (3-fold cross-validation)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2910               85.6134 %
Incorrectly Classified Instances       489               14.3866 %
Kappa statistic                          0.6377
Mean absolute error                      0.2252
Root mean squared error                  0.3308
Relative absolute error                 55.4173 %
Root relative squared error             73.3894 %
Total Number of Instances             3399

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.709    0.085    0.767      0.709    0.736      0.639    0.901
0.799     High Hospital Quality
                 0.915    0.291    0.888      0.915    0.901      0.639    0.901
0.957     Low Hospital Quality
Weighted Avg.    0.856    0.233    0.854      0.856    0.854      0.639    0.901
0.912

=== Confusion Matrix ===

    a    b   <-- classified as
  683  281 |    a = High Hospital Quality
  208 2227 |    b = Low Hospital Quality
```

10. 90% split (Kernel Estimator)

a. Classifier → bayes → NaiveBayes (Percentage split 90%)

i. useKernelEstimator

```
=== Summary ===

Correctly Classified Instances         297               87.3529 %
Incorrectly Classified Instances        43               12.6471 %
Kappa statistic                          0.6813
Mean absolute error                      0.2223
Root mean squared error                  0.3289
Relative absolute error                 54.3198 %
Root relative squared error             72.385  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.717    0.062    0.826      0.717    0.768      0.684    0.902
0.810     High Hospital Quality
                 0.938    0.283    0.890      0.938    0.913      0.684    0.902
0.954     Low Hospital Quality
Weighted Avg.    0.874    0.219    0.871      0.874    0.871      0.684    0.902
0.912

=== Confusion Matrix ===

   a   b   <-- classified as
  71  28 |   a = High Hospital Quality
  15 226 |   b = Low Hospital Quality
```

11. 90% split (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (Percentage split 90%)
        i. useSupervisedDiscretization

```
=== Summary ===

Correctly Classified Instances         297               87.3529 %
Incorrectly Classified Instances        43               12.6471 %
Kappa statistic                          0.6813
Mean absolute error                      0.2223
Root mean squared error                  0.3289
Relative absolute error                 54.3198 %
Root relative squared error             72.385  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.717    0.062    0.826      0.717    0.768      0.684    0.902
0.810     High Hospital Quality
                 0.938    0.283    0.890      0.938    0.913      0.684    0.902
0.954     Low Hospital Quality
Weighted Avg.    0.874    0.219    0.871      0.874    0.871      0.684    0.902
0.912

=== Confusion Matrix ===

   a   b   <-- classified as
  71  28 |   a = High Hospital Quality
  15 226 |   b = Low Hospital Quality
```

12. 90% split

a. Classifier → bayes → NaiveBayes (Percentage split 90%)

```
=== Summary ===

Correctly Classified Instances         297               87.3529 %
Incorrectly Classified Instances        43               12.6471 %
Kappa statistic                          0.6813
Mean absolute error                      0.2223
Root mean squared error                  0.3289
Relative absolute error                 54.3198 %
Root relative squared error             72.385  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.717    0.062    0.826      0.717    0.768      0.684    0.902
0.810     High Hospital Quality
                 0.938    0.283    0.890      0.938    0.913      0.684    0.902
0.954     Low Hospital Quality
Weighted Avg.    0.874    0.219    0.871      0.874    0.871      0.684    0.902
0.912

=== Confusion Matrix ===

   a   b   <-- classified as
  71  28 |   a = High Hospital Quality
  15 226 |   b = Low Hospital Quality
```

13. 80% split (Kernel Estimator)
    a. Classifier → bayes → NaiveBayes (Percentage split 80%)
        i. useKernelEstimator

```
=== Summary ===

Correctly Classified Instances         591               86.9118 %
Incorrectly Classified Instances        89               13.0882 %
Kappa statistic                          0.6791
Mean absolute error                      0.2223
Root mean squared error                  0.3251
Relative absolute error                 53.8446 %
Root relative squared error             70.6404 %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.723    0.068    0.823      0.723    0.770      0.682    0.912
0.831     High Hospital Quality
                 0.932    0.277    0.886      0.932    0.909      0.682    0.912
0.958     Low Hospital Quality
Weighted Avg.    0.869    0.213    0.867      0.869    0.867      0.682    0.912
0.920

=== Confusion Matrix ===

   a   b   <-- classified as
 149  57 |   a = High Hospital Quality
  32 442 |   b = Low Hospital Quality
```

14. 80% split (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (Percentage Split 80%)
        i. useSupervisedDiscretization

```
=== Summary ===

Correctly Classified Instances         591               86.9118 %
Incorrectly Classified Instances        89               13.0882 %
Kappa statistic                          0.6791
Mean absolute error                      0.2223
Root mean squared error                  0.3251
Relative absolute error                 53.8446 %
Root relative squared error             70.6404 %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.723    0.068    0.823      0.723    0.770      0.682    0.912
0.831     High Hospital Quality
                 0.932    0.277    0.886      0.932    0.909      0.682    0.912
0.958     Low Hospital Quality
Weighted Avg.    0.869    0.213    0.867      0.869    0.867      0.682    0.912
0.920

=== Confusion Matrix ===

   a   b   <-- classified as
 149  57 |   a = High Hospital Quality
  32 442 |   b = Low Hospital Quality
```

15. 80% split

a. Classifier → bayes → NaiveBayes (Percentage split 80%)

```
=== Summary ===

Correctly Classified Instances         591               86.9118 %
Incorrectly Classified Instances        89               13.0882 %
Kappa statistic                          0.6791
Mean absolute error                      0.2223
Root mean squared error                  0.3251
Relative absolute error                 53.8446 %
Root relative squared error             70.6404 %
Total Number of Instances              680

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.723    0.068    0.823      0.723    0.770      0.682    0.912
0.831     High Hospital Quality
                 0.932    0.277    0.886      0.932    0.909      0.682    0.912
0.958     Low Hospital Quality
Weighted Avg.    0.869    0.213    0.867      0.869    0.867      0.682    0.912
0.920

=== Confusion Matrix ===

   a   b   <-- classified as
 149  57 |   a = High Hospital Quality
  32 442 |   b = Low Hospital Quality
```

16. 70% split (Kernel Estimator)

a. Classifier → bayes → NaiveBayes (Percentage split 70%)

i. useKernelEstimator

```
=== Summary ===

Correctly Classified Instances         888               87.0588 %
Incorrectly Classified Instances       132               12.9412 %
Kappa statistic                          0.6844
Mean absolute error                      0.2218
Root mean squared error                  0.3225
Relative absolute error                 53.8215 %
Root relative squared error             69.903  %
Total Number of Instances             1020

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.730    0.068    0.825      0.730    0.775      0.687    0.914
0.836     High Hospital Quality
                 0.932    0.270    0.887      0.932    0.909      0.687    0.914
0.959     Low Hospital Quality
Weighted Avg.    0.871    0.208    0.868      0.871    0.868      0.687    0.914
0.921

=== Confusion Matrix ===

   a   b   <-- classified as
 227  84 |   a = High Hospital Quality
  48 661 |   b = Low Hospital Quality
```

17. 70% split (Supervised Discretization)
    a. Classifier → bayes → NaiveBayes (Percentage split 70%)
        i. useSupervisedDiscretization

```
=== Summary ===

Correctly Classified Instances         888               87.0588 %
Incorrectly Classified Instances       132               12.9412 %
Kappa statistic                          0.6844
Mean absolute error                      0.2218
Root mean squared error                  0.3225
Relative absolute error                 53.8215 %
Root relative squared error             69.903  %
Total Number of Instances             1020

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.730    0.068    0.825      0.730    0.775      0.687    0.914
0.836     High Hospital Quality
                 0.932    0.270    0.887      0.932    0.909      0.687    0.914
0.959     Low Hospital Quality
Weighted Avg.    0.871    0.208    0.868      0.871    0.868      0.687    0.914
0.921

=== Confusion Matrix ===

   a   b   <-- classified as
 227  84 |   a = High Hospital Quality
  48 661 |   b = Low Hospital Quality
```

18. 70% split

a. Classifier → bayes → NaiveBayes (Percentage split 70%)

```
=== Summary ===

Correctly Classified Instances         888               87.0588 %
Incorrectly Classified Instances       132               12.9412 %
Kappa statistic                          0.6844
Mean absolute error                      0.2218
Root mean squared error                  0.3225
Relative absolute error                 53.8215 %
Root relative squared error             69.903  %
Total Number of Instances             1020

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
PRC Area  Class
                 0.730    0.068    0.825      0.730    0.775      0.687    0.914
0.836     High Hospital Quality
                 0.932    0.270    0.887      0.932    0.909      0.687    0.914
0.959     Low Hospital Quality
Weighted Avg.    0.871    0.208    0.868      0.871    0.868      0.687    0.914
0.921

=== Confusion Matrix ===

   a   b   <-- classified as
 227  84 |   a = High Hospital Quality
  48 661 |   b = Low Hospital Quality
```

Notes

Naive Bayes cannot handle numerical attributes, in which case it is treated by (1) discretizing the numeric variables or (2) using probability density functions. Discretizing the continuous variable would be a simple solution; however, it is subjective which causes loss of information, but it is still used as a quick way to prepare the data before applying Naive Bayes classification.

By default, a Gaussian distribution is assumed for each numerical attributes. In order to handle numerical attributes, we will automatically convert them to nominal attributes with the ***useSupervisedDiscretization*** parameter set as TRUE.

The algorithm was changed to use a kernel estimator with the ***useKernelEstimator*** argument which can be used to better match the actual distribution of the attributes in our dataset.

useKernelEstimator and ***useSupervisedDiscretization*** parameters are mutually exclusive.

What "Use training set" does:

(1) Weka takes 3399 labeled data.
(2) It applies an algorithm to build a classifier from these 3399 data.
(3) It applies that classifier AGAIN on these 3399 data.
(4) It provides us with the performance of the classifier (applied to the same 3399 data from which it was developed).

What "Use 10 fold CV" does:

(1) Weka takes 3399 labeled data.
(2) It produces 10 equal sized sets. Each set is divided into two groups: (90%) 3059 labeled data are used for training and (10%) 340 labeled data are used for testing.
(3) It produces a classifier with an algorithm from 3059 labeled data and applies that on the 340 testing data for set 1.
(4) It does the same thing for set 2 to 10 and produces 9 more classifiers.
(5) It averages the performance of the 10 classifiers produced from 10 equal sized (3059 training and 340 testing) sets.

Results

Naive Bayes assumes that the input values are nominal (numerical inputs are supported by assuming a distribution).

Naive Bayes uses a simple implementation of Bayes Theorem (hence naive) where the prior probability for each class is calculated from the training data and assumed to be independent of each other (conditionally independent), which is an unrealistic assumption because we expect the variables to interact and be dependent. However, we should keep in mind that this assumption makes the probabilities fast and easy to calculate.

From classification using Weka, the correctly classified instances are as follows:

- 90% split (87.3529 %)
- 70% split (87.0588 %)
- 80% split (86.9118 %)
- 3-fold CV (85.6134 %)
- Use training set (85.6134 %)
- 10-fold CV (85.4957 %)

Training set should not be used because it will have generalization error (i.e., using only one data set, we could achieve perfect accuracy by simply learning this particular set, but not the general concept). In addition, using training/test sets and cross-validation are conceptually the same thing. Cross-validation simply takes a more rigorous approach by averaging over the entire data set.