Report

Knn classification algorithm from scratch using k-fold cross validation has been implemented by using various distance measures like Euclidean distance, Manhattan distance, Chebyshev distance and Minkowski to extend the implementation.

Steps to run code:

1.Run the knn.ipynb in Jupyter Notebook.
2.Currently breast-cancer dataset is used. Replace filename with various dataset to see other results.
3.Please input distance measures as string.

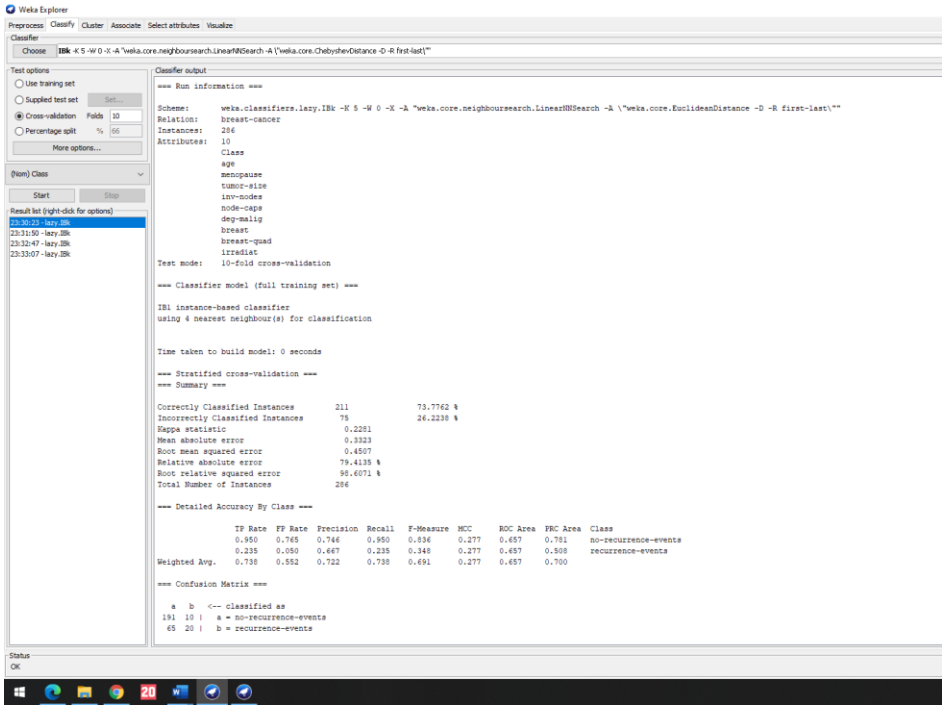
Analyzing data inside Weka:
For Weka to understand the data we need to convert the breast-cancer.data file as well as other datasets to the file format with .arff extension. The arff file has 3sections i.e @relation(which gives information about the file name),@attribute(which is the attribute information where we have feature names, feature values, range of the values and datatypes which can be accessed from breast-cancer.names file if opened as a word document)can have two kinds of features either nominal(string) or numeric and @data (which is accessed from the breast-cancer.data file)where the order should be maintained as the features in the attribute section.

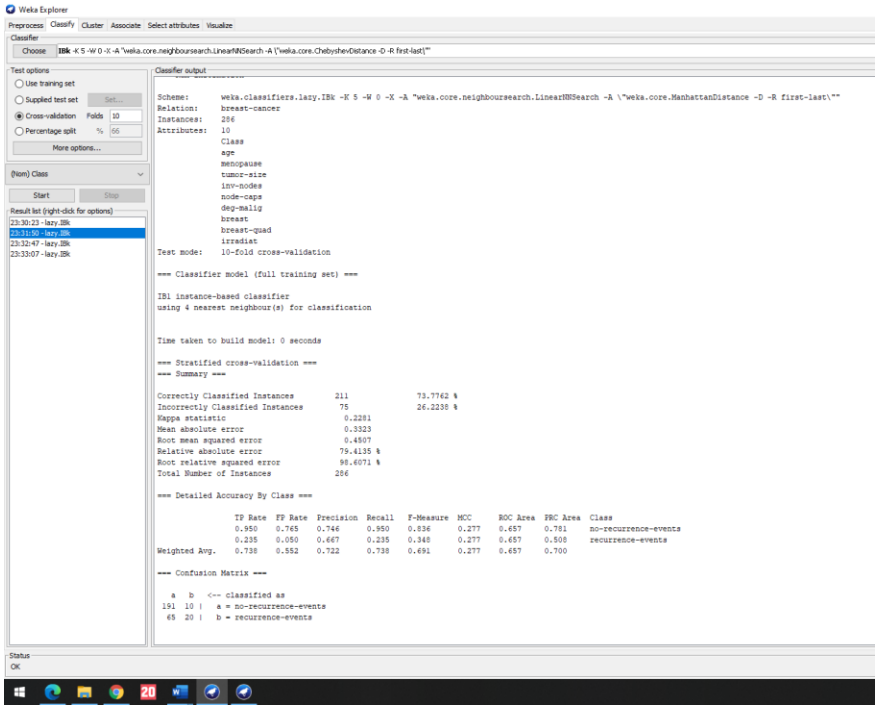
Steps to find accuracy using Weka:

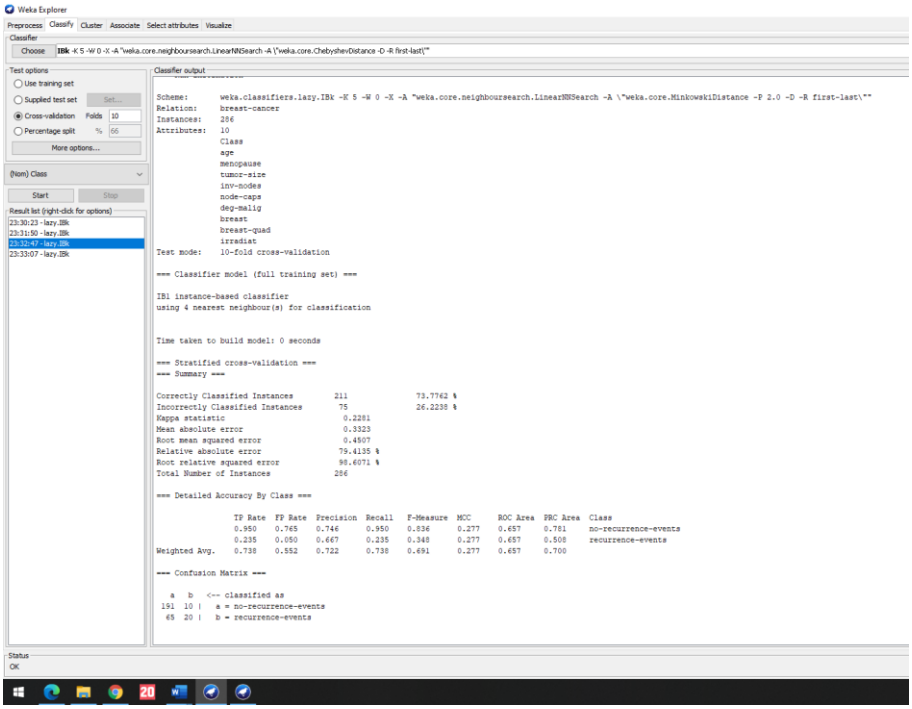
1. Open Weka GUI and choose explorer option.
2. Click open file and select the required .arff file.
3. Click on Classify tab and click choose -> IBK(as it is K nearest neighbor classifier)->click on the textbox adjacent to Choose->a small window pops which helps us choose nearest neighbor search algorithm->KNN enter 10,cross validate as true and select linearNN->Click on the textbox adjacent to Choose again-> Choose Euclidean or any other distance function-> Click Ok-> Start

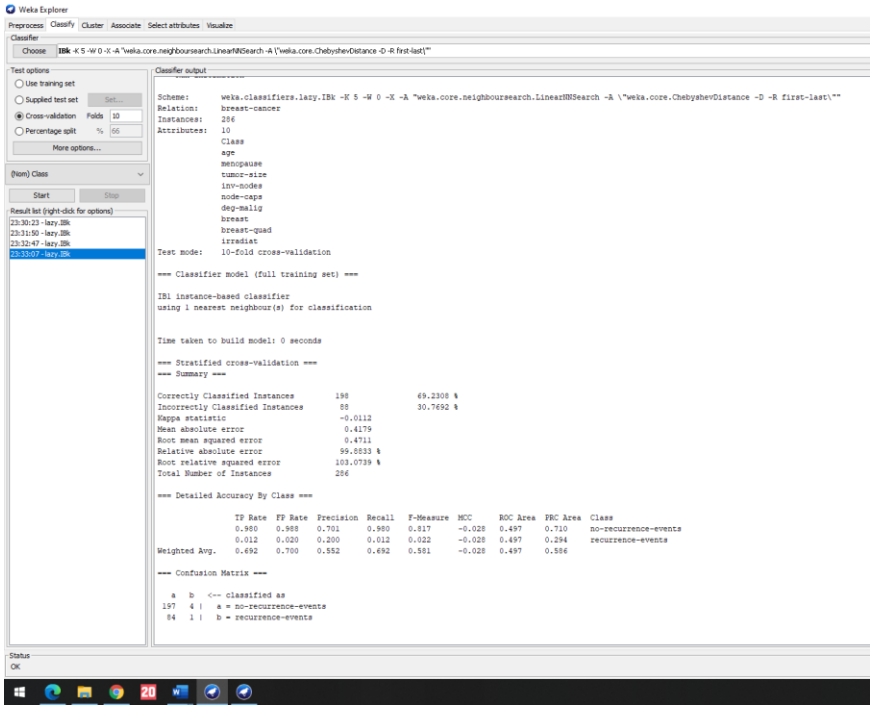
Please find below the comparison of the accuracy of the algorithm using different distance measures with 10-fold cross validation and Weka:
1: Breast cancer dataset

	Weka	k=10
Euclidean Distance	73. 776%	73.214%
Manhattan Distance	73. 776%	73.214%
Chebyshev Distance	69.230%	73.571%
Minkowski Distance	73. 776%	76.071%
Hamming Distance	-	74.643%
Jaccard Distance	-	74.643%


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start Stop
Result list (right-click for options)
23:30:23 - lazy.IBk
23:31:50 - lazy.IBk
23:32:47 - lazy.IBk
23:33:07 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -D -R first-last\""
Relation: breast-cancer
Instances: 286
Attributes: 10
Class
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 4 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 211 73.7762 %
Incorrectly Classified Instances 75 26.2238 %
Kappa statistic 0.2281
Mean absolute error 0.3323
Root mean squared error 0.4507
Relative absolute error 79.4135 %
Root relative squared error 98.6071 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.950 0.765 0.746 0.950 0.836 0.277 0.657 0.781 no-recurrence-events
0.235 0.050 0.667 0.235 0.348 0.277 0.657 0.508 recurrence-events
Weighted Avg. 0.738 0.552 0.722 0.738 0.691 0.277 0.657 0.700
=== Confusion Matrix ===
a b <-- classified as
191 10 | a = no-recurrence-events
65 20 | b = recurrence-events
Status
OK

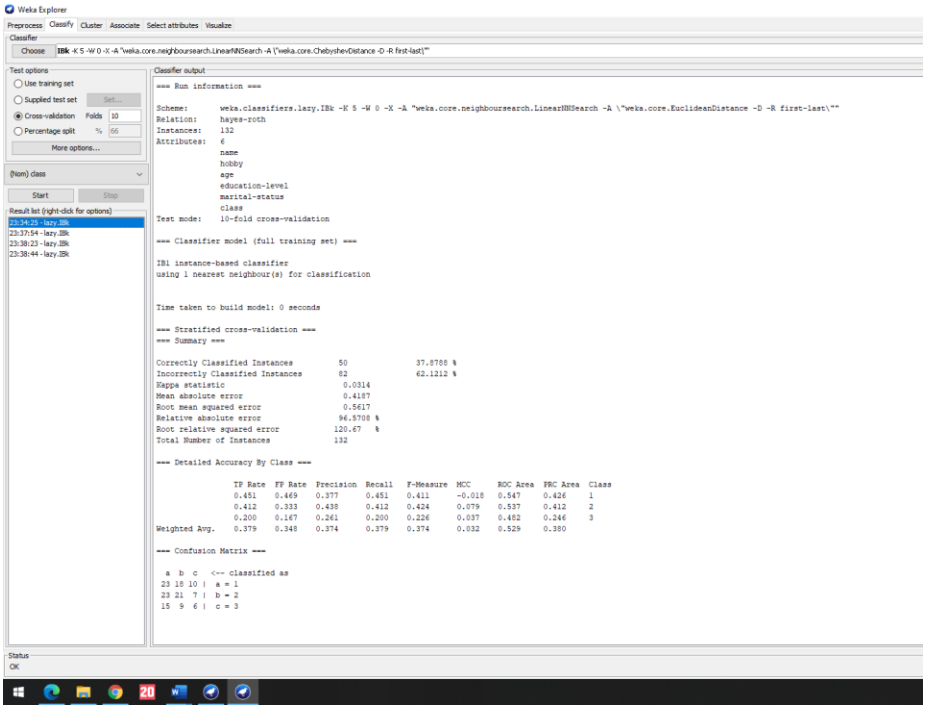

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start Stop
Result list (right-click for options)
23:30:23 - lazy.IBk
23:31:50 - lazy.IBk
23:32:47 - lazy.IBk
23:33:07 - lazy.IBk
Classifier output
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ManhattanDistance -D -R first-last\""
Relation: breast-cancer
Instances: 286
Attributes: 10
Class
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 4 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 211 73.7762 %
Incorrectly Classified Instances 75 26.2238 %
Kappa statistic 0.2281
Mean absolute error 0.3323
Root mean squared error 0.4507
Relative absolute error 79.4135 %
Root relative squared error 98.6071 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.950 0.765 0.746 0.950 0.836 0.277 0.657 0.781 no-recurrence-events
0.235 0.050 0.667 0.235 0.348 0.277 0.657 0.508 recurrence-events
Weighted Avg. 0.738 0.552 0.722 0.738 0.691 0.277 0.657 0.700
=== Confusion Matrix ===
a b <-- classified as
191 10 | a = no-recurrence-events
65 20 | b = recurrence-events
Status
OK

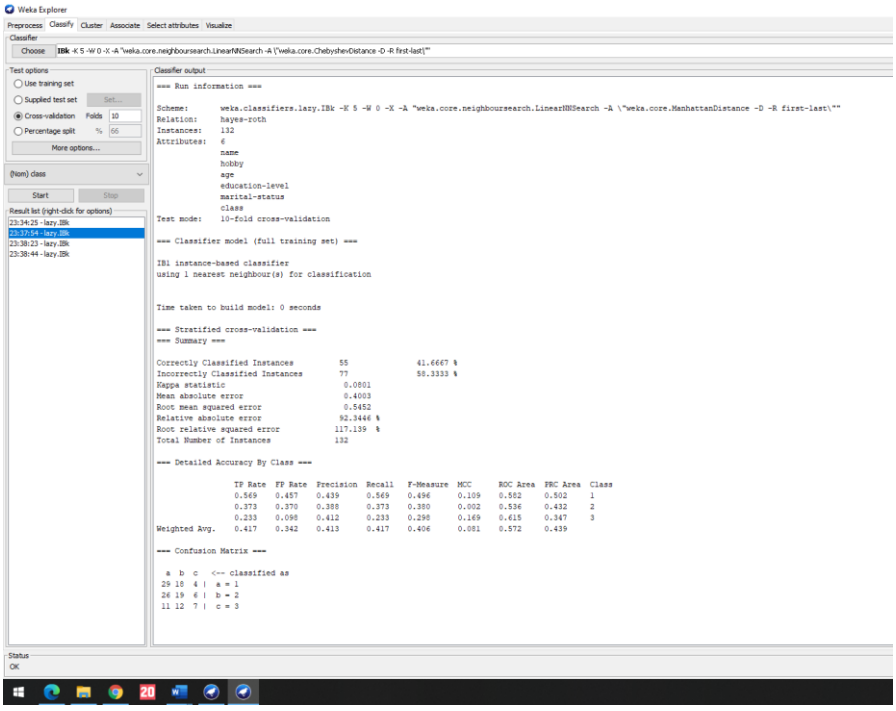

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start Stop
Result list (right-click for options)
23:30:23 - lazy.IBk
23:31:50 - lazy.IBk
23:32:47 - lazy.IBk
23:33:07 - lazy.IBk
Classifier output
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.MinkowskiDistance -P 2.0 -D -R first-last\""
Relation: breast-cancer
Instances: 286
Attributes: 10
Class
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 4 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 211 73.7762 %
Incorrectly Classified Instances 75 26.2238 %
Kappa statistic 0.2281
Mean absolute error 0.3323
Root mean squared error 0.4907
Relative absolute error 79.4135 %
Root relative squared error 98.6071 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.950 0.765 0.746 0.950 0.836 0.277 0.657 0.781 no-recurrence-events
0.235 0.050 0.667 0.235 0.348 0.277 0.657 0.508 recurrence-events
Weighted Avg. 0.738 0.552 0.722 0.738 0.691 0.277 0.657 0.700
=== Confusion Matrix ===
a b <-- classified as
191 10 | a = no-recurrence-events
65 20 | b = recurrence-events
Status
OK

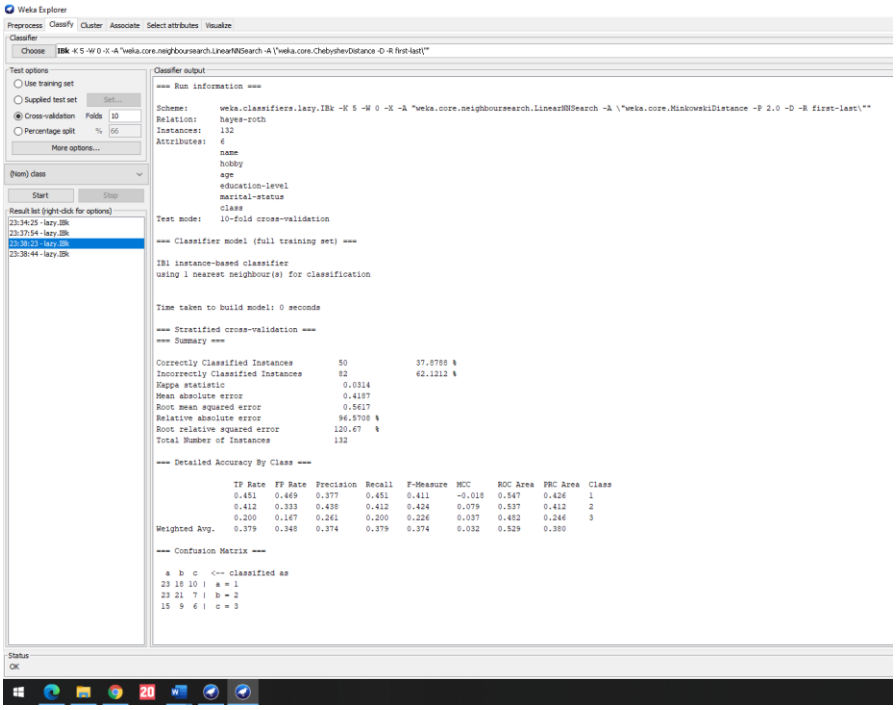

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start Stop
Result list (right-click for options)
23:30:23 - lazy.IBk
23:31:50 - lazy.IBk
23:32:47 - lazy.IBk
23:33:07 - lazy.IBk
Classifier output
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Relation: breast-cancer
Instances: 286
Attributes: 10
Class
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 198 69.2308 %
Incorrectly Classified Instances 88 30.7692 %
Kappa statistic -0.0112
Mean absolute error 0.4179
Root mean squared error 0.4711
Relative absolute error 99.8833 %
Root relative squared error 103.0739 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.980 0.988 0.701 0.980 0.817 -0.028 0.497 0.710 no-recurrence-events
0.012 0.020 0.200 0.012 0.022 -0.028 0.497 0.294 recurrence-events
Weighted Avg. 0.692 0.700 0.552 0.692 0.581 -0.028 0.497 0.586
=== Confusion Matrix ===
a b <-- classified as
197 4 | a = no-recurrence-events
84 1 | b = recurrence-events
Status
OK

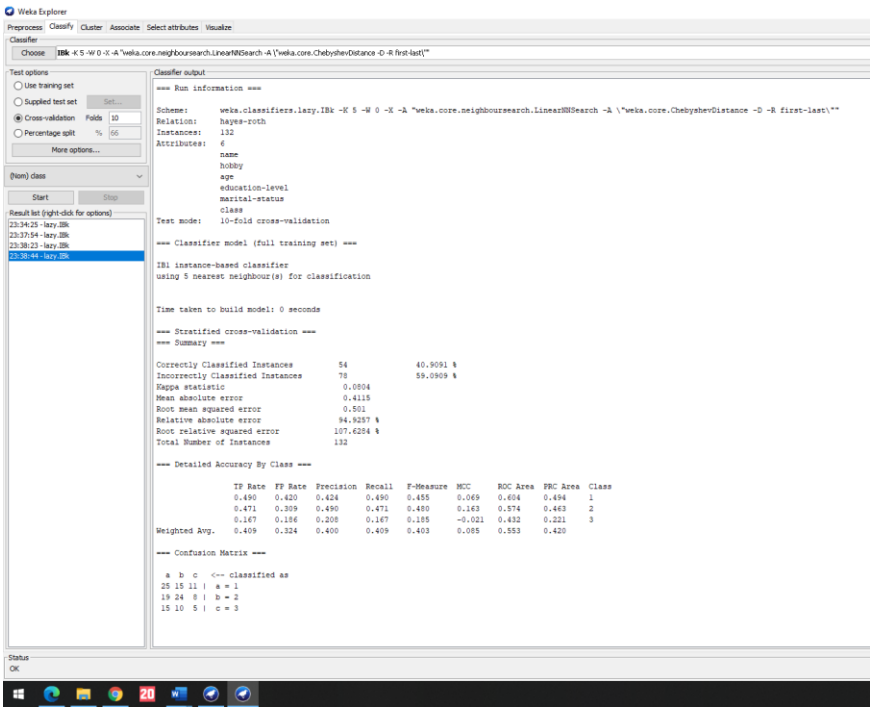
2: Hayes-roth dataset

	Weka	k=10
Euclidean Distance	37.878%	36.154%
Manhattan Distance	41.667%	40.0%
Chebyshev Distance	40.909%	26.923%
Minkowski Distance	37.878%	33.846%
Hamming Distance	-	61.538%
Jaccard Distance	-	63.077%



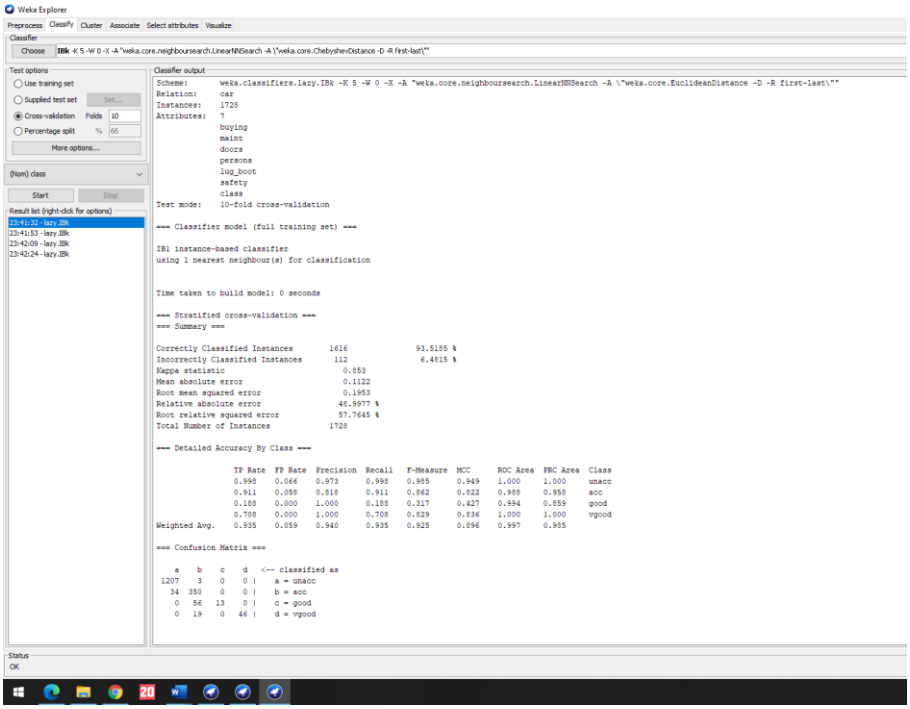

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start Stop
Result list (right-click for options)
23:34:25 - lazy.IBk
23:37:54 - lazy.IBk
23:38:23 - lazy.IBk
23:38:44 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ManhattanDistance -D -R first-last\""
Relation: hayes-roth
Instances: 132
Attributes: 6
name
hobby
age
education-level
marital-status
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 55 41.6667 %
Incorrectly Classified Instances 77 58.3333 %
Kappa statistic 0.0801
Mean absolute error 0.4003
Root mean squared error 0.5452
Relative absolute error 92.3446 %
Root relative squared error 117.139 %
Total Number of Instances 132
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.569 0.457 0.439 0.569 0.496 0.109 0.582 0.502 1
0.373 0.370 0.388 0.373 0.380 0.002 0.536 0.432 2
0.233 0.098 0.412 0.233 0.298 0.169 0.615 0.347 3
Weighted Avg. 0.417 0.342 0.413 0.417 0.406 0.081 0.572 0.439
=== Confusion Matrix ===
a b c <-- classified as
29 18 4 | a = 1
26 19 6 | b = 2
11 12 7 | c = 3
Status
OK

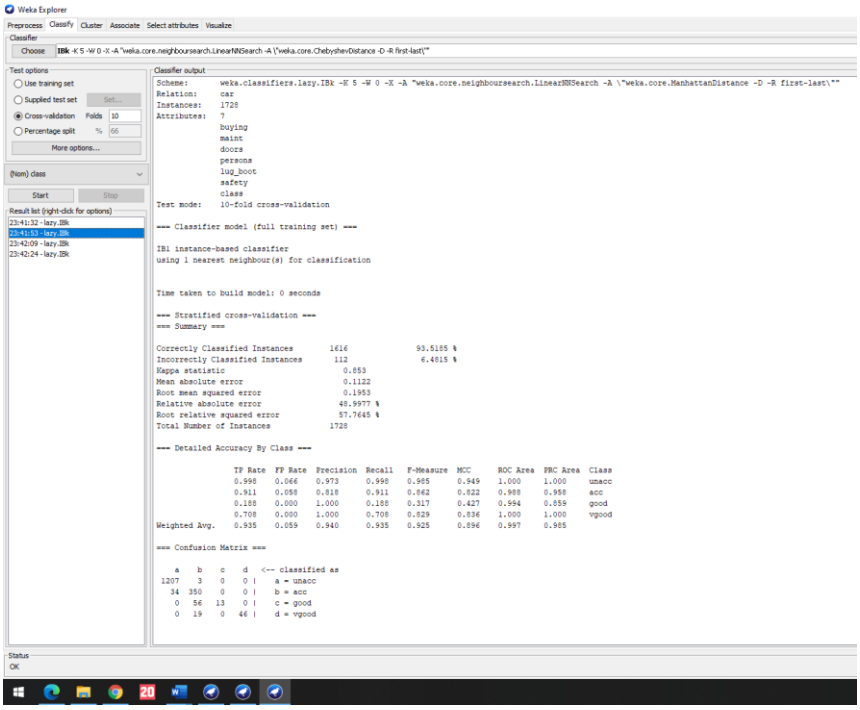

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start Stop
Result list (right-click for options)
23:34:25 - lazy.IBk
23:37:54 - lazy.IBk
23:38:23 - lazy.IBk
23:38:44 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.MinkowskiDistance -P 2.0 -D -R first-last\""
Relation: hayes-roth
Instances: 132
Attributes: 6
name
hobby
age
education-level
marital-status
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 50 37.8788 %
Incorrectly Classified Instances 82 62.1212 %
Kappa statistic 0.0314
Mean absolute error 0.4187
Root mean squared error 0.5617
Relative absolute error 96.5708 %
Root relative squared error 120.67 %
Total Number of Instances 132
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.451 0.469 0.377 0.451 0.411 -0.018 0.547 0.426 1
0.412 0.333 0.438 0.412 0.424 0.079 0.537 0.412 2
0.200 0.167 0.261 0.200 0.226 0.037 0.482 0.246 3
Weighted Avg. 0.379 0.348 0.374 0.379 0.374 0.032 0.529 0.380
=== Confusion Matrix ===
a b c <-- classified as
23 18 10 | a = 1
23 21 7 | b = 2
15 9 6 | c = 3
Status
OK

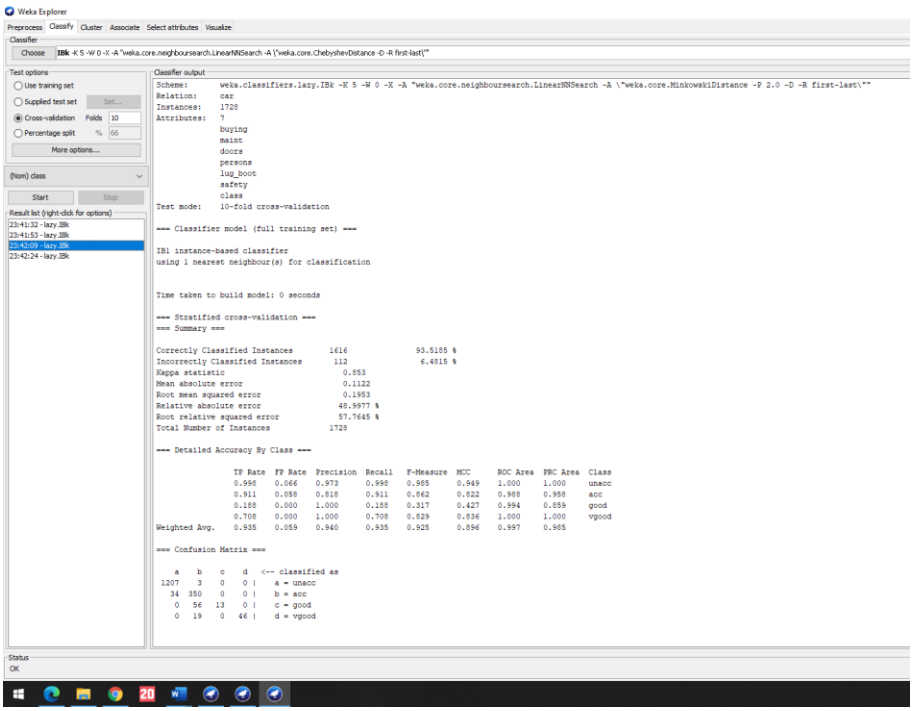


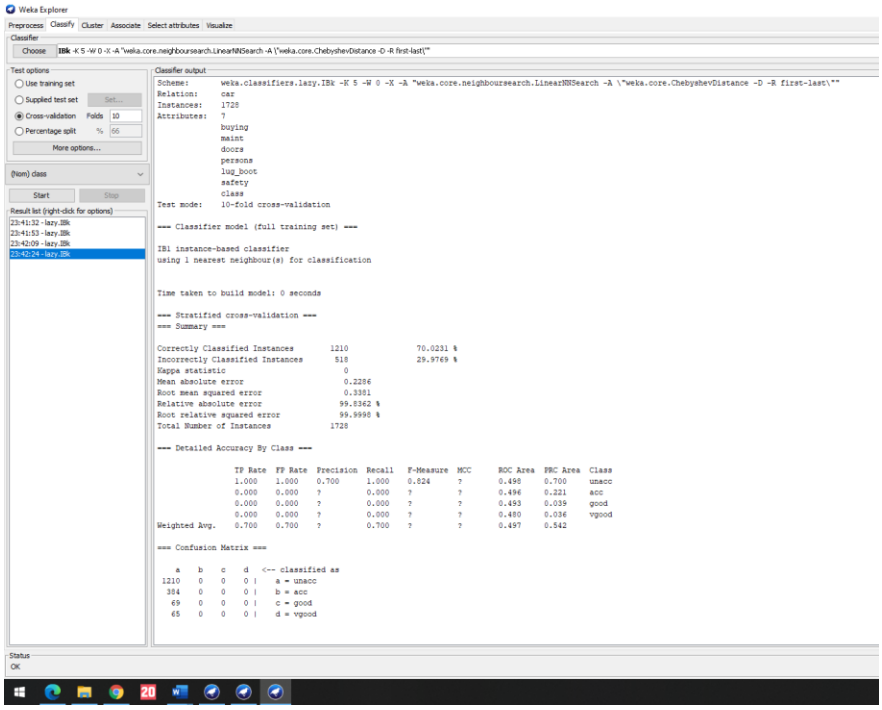
3: Car dataset

	Weka	k=10
Euclidean Distance	93.518%	82.5%
Manhattan Distance	93.518%	90.29%
Chebyshev Distance	70.023%	67.616%
Minkowski Distance	93.518%	89.535%
Hamming Distance	-	88.256%
Jaccard Distance	-	64.593%


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
23:41:32 - lazy.IBk
23:41:53 - lazy.IBk
23:42:09 - lazy.IBk
23:42:24 - lazy.IBk
Classifier output
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -D -R first-last\""
Relation: car
Instances: 1728
Attributes: 7
buying
maint
doors
persons
lug_boot
safety
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1616 93.5185 %
Incorrectly Classified Instances 112 6.4815 %
Kappa statistic 0.853
Mean absolute error 0.1122
Root mean squared error 0.1953
Relative absolute error 48.9977 %
Root relative squared error 57.7645 %
Total Number of Instances 1728
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.998 0.066 0.973 0.998 0.985 0.949 1.000 1.000 unacc
0.911 0.058 0.818 0.911 0.862 0.822 0.988 0.958 acc
0.188 0.000 1.000 0.188 0.317 0.427 0.994 0.859 good
0.708 0.000 1.000 0.708 0.829 0.836 1.000 1.000 vgood
Weighted Avg. 0.935 0.059 0.940 0.935 0.925 0.896 0.997 0.985
=== Confusion Matrix ===
a b c d <-- classified as
1207 3 0 0 | a = unacc
34 350 0 0 | b = acc
0 56 13 0 | c = good
0 19 0 46 | d = vgood
Status
OK


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ChebyshevDistance -D -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
23:41:32 - lazy.IBk
23:41:53 - lazy.IBk
23:42:09 - lazy.IBk
23:42:24 - lazy.IBk
Classifier output
Scheme: weka.classifiers.lazy.IBk -K 5 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.ManhattanDistance -D -R first-last\""
Relation: car
Instances: 1728
Attributes: 7
buying
maint
doors
persons
lug_boot
safety
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1616 93.5185 %
Incorrectly Classified Instances 112 6.4815 %
Kappa statistic 0.853
Mean absolute error 0.1122
Root mean squared error 0.1953
Relative absolute error 48.9977 %
Root relative squared error 57.7645 %
Total Number of Instances 1728
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.998 0.066 0.973 0.998 0.985 0.949 1.000 1.000 unacc
0.911 0.058 0.818 0.911 0.862 0.822 0.988 0.958 acc
0.188 0.000 1.000 0.188 0.317 0.427 0.994 0.859 good
0.708 0.000 1.000 0.708 0.829 0.836 1.000 1.000 vgood
Weighted Avg. 0.935 0.059 0.940 0.935 0.925 0.896 0.997 0.985
=== Confusion Matrix ===
a b c d <-- classified as
1207 3 0 0 | a = unacc
34 350 0 0 | b = acc
0 56 13 0 | c = good
0 19 0 46 | d = vgood
Status
OK





References:

1. https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/
2. https://machinelearningmastery.com/k-fold-cross-validation/
3. https://machinelearningmastery.com/distance-measures-for-machine-learning/
4. https://aiaspirant.com/distance-similarity-measures-in-machine-learning/

Dataset Links:

https://archive.ics.uci.edu/ml/datasets/Hayes-Roth
https://archive.ics.uci.edu/ml/datasets/car+evaluation
https://archive.ics.uci.edu/ml/datasets/breast+cancer