



Assignment-Mid Term

NAME: MD. AKHTARUZZAMAN EMON

ID: 18-36074-1

COURSE: DATA WAREHOUSING AND DATA MINING

SECTION: A

SUBMITTED TO- DR. MD. MAHBUB CHOWDHURY MISHU

Introduction:

A data mining function called classification allocates objects in a collection to desired groups or classes. Classification's purpose is to correctly anticipate the target class for each case in the data. In this report I used K-Nearest Neighbor (KNN) procedure. The K-Nearest Neighbor algorithm is an example of a "lazy learner," meaning it does not build a model using the training set until the data set is queried. For this report I used a data set of a bank. There are many data of people in this data set. We use Weka software to find the result.

Data Set:

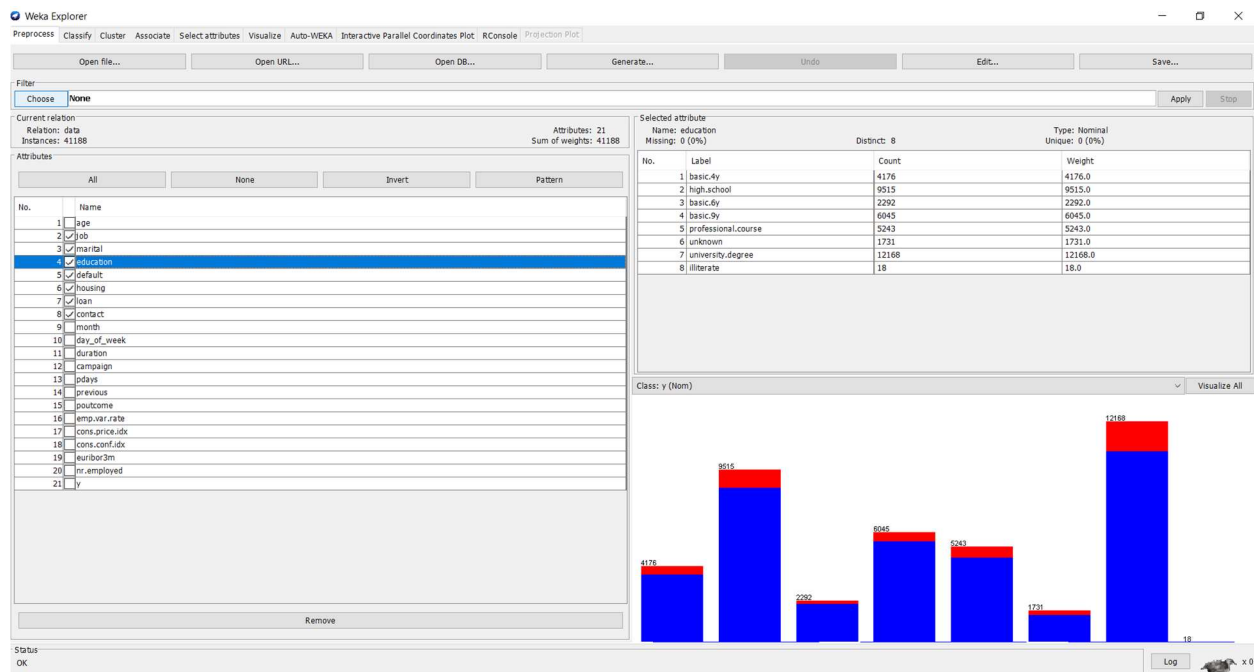
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	outcome	emp.var	r	cons.price	cons.conf	euribor3m	nr.employ	y	
2	56	housemaik	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
3	57	services	married	high.schoc	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
4	37	services	married	high.schoc	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
5	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
6	56	services	married	high.schoc	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
7	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
8	59	admin.	married	profession	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
9	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
10	24	technician	single	profession	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
11	25	services	single	high.schoc	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
12	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
13	25	services	single	high.schoc	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
14	29	blue-collar	single	high.schoc	no	no	yes	telephone	may	mon	137	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
15	57	housemaik	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
16	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
17	54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
18	35	blue-collar	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
19	46	blue-collar	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
20	50	blue-collar	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
21	39	managem	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
22	30	unemploy	married	high.schoc	no	no	no	telephone	may	mon	38	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
23	55	blue-collar	married	basic.4y	unknown	yes	no	telephone	may	mon	262	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
24	55	retired	single	high.schoc	no	yes	no	telephone	may	mon	342	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
25	41	technician	single	high.schoc	no	yes	no	telephone	may	mon	181	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
26	37	admin.	married	high.schoc	no	yes	no	telephone	may	mon	172	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
27	35	technician	married	university	no	no	yes	telephone	may	mon	99	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
28	59	technician	married	unknown	no	yes	no	telephone	may	mon	93	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		
29	39	self-emp	married	basic.9y	unknown	no	no	telephone	may	mon	233	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no		

In this data set there are 21 columns and 41189 rows.

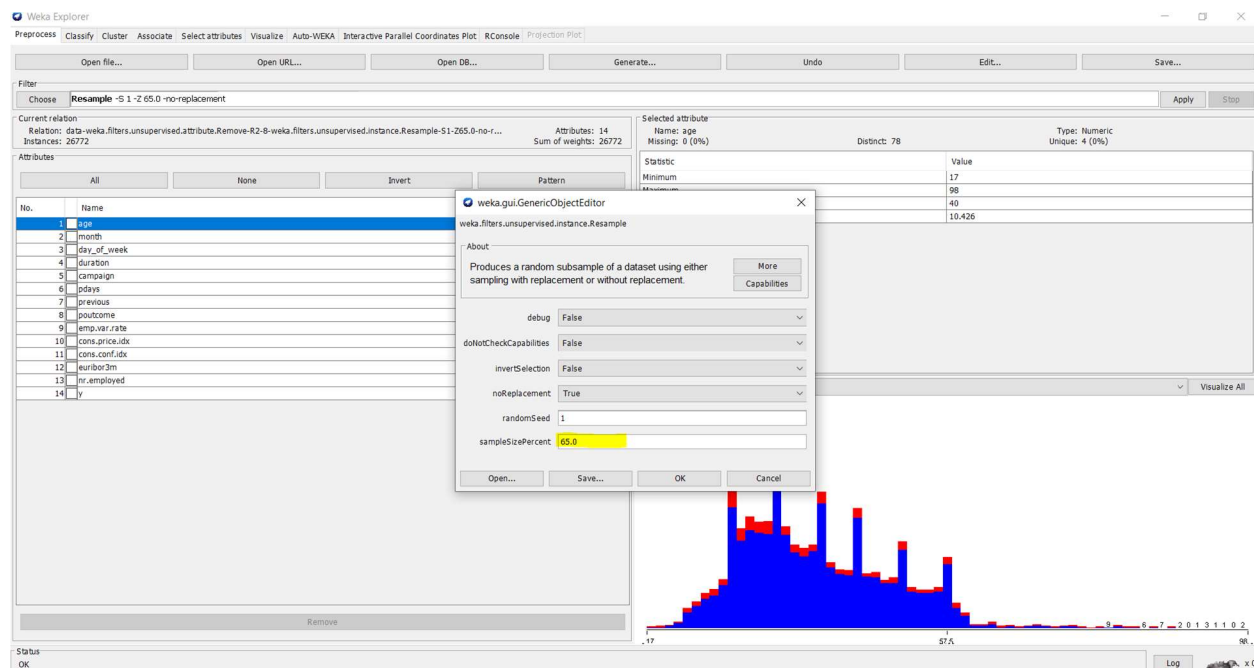
Dataset reference: <https://www.kaggle.com/brijbhushannanda1979/bank-data>

Procedure:

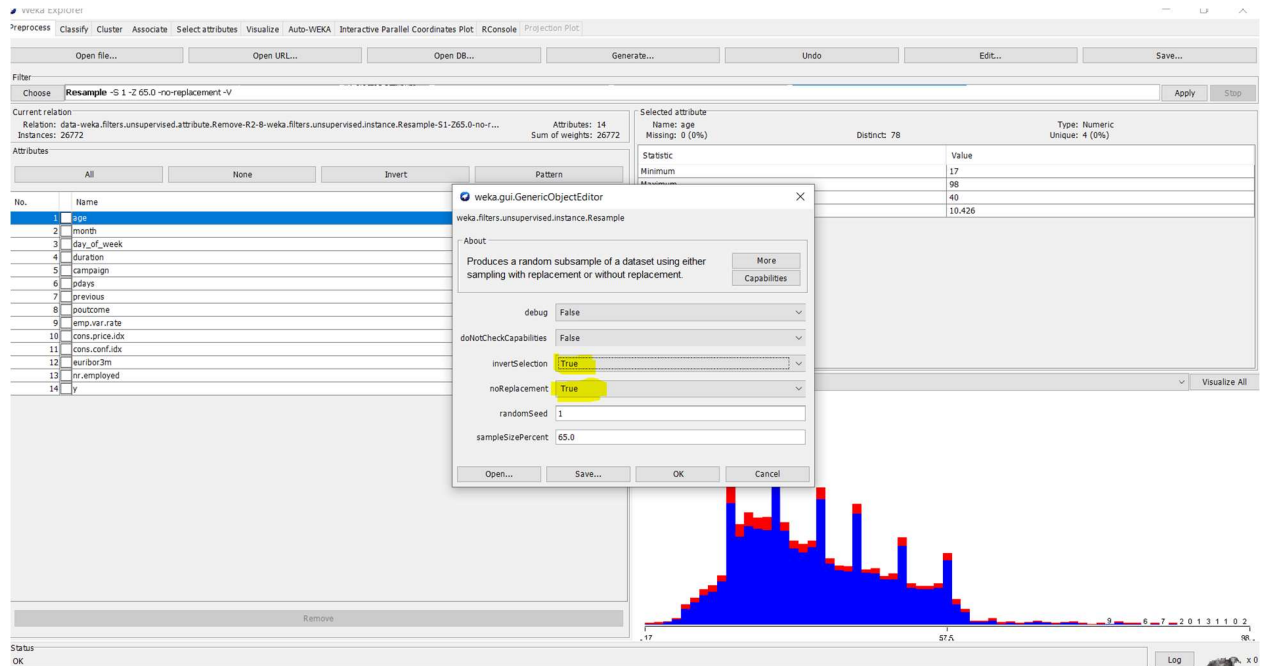
At first, we need to open the data set file in Weka software. Then we need delete unnecessary data from the data set. Then we have to follow the algorithms of K-Nearest Neighbor algorithm, Naïve Bayes.



Then we have to set the data in two set which are training set and test set.



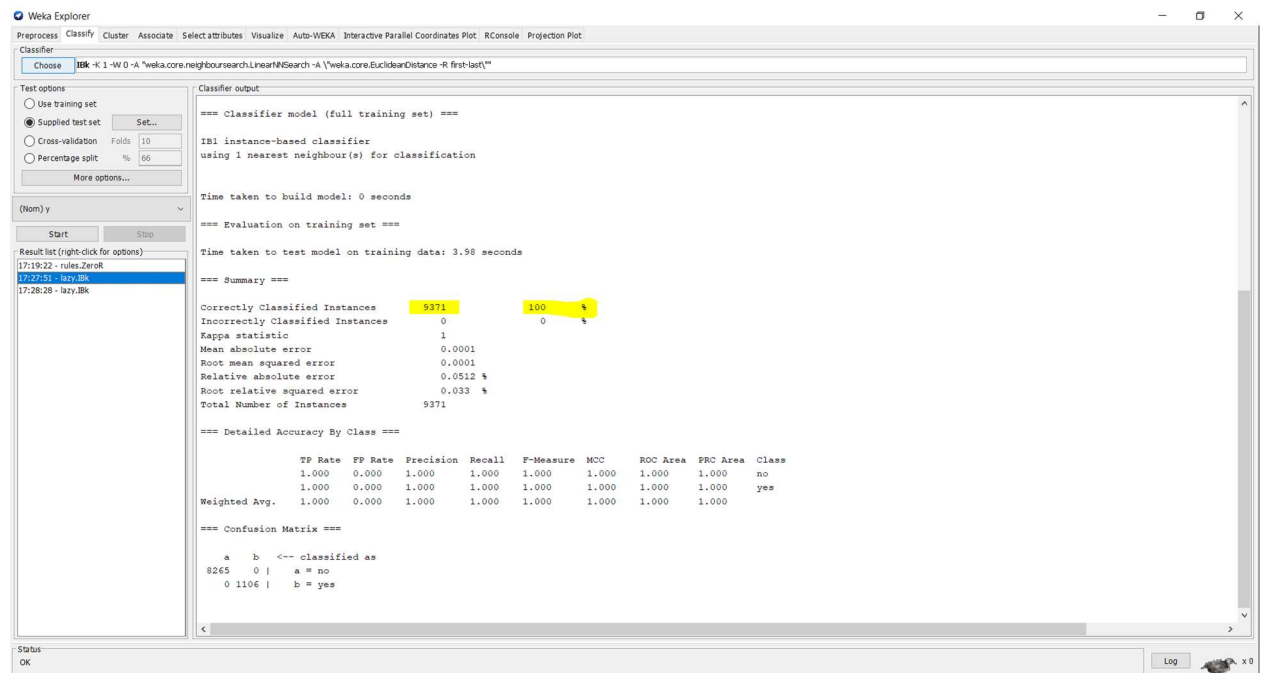
Here we set 65% of data for training set.



Here we set 55% of data for test set.

Result of K-Nearest Neighbor (KNN):

Training set:



Test set:

The screenshot shows the Weka Explorer interface with the 1-Nearest Neighbour classifier selected. The test set is 'lazy.BK'. The classifier output shows a summary of performance metrics and a detailed accuracy by class table.

Classifier output

```
=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Evaluation on test set ===
Time taken to test model on supplied test set: 2.33 seconds

=== Summary ===
Correctly Classified Instances      9371      100 %
Incorrectly Classified Instances      0      0 %
Kappa statistic              1
Mean absolute error          0.0001
Root mean squared error      0.0001
Relative absolute error       0.0512 %
Root relative squared error   0.033 %
Total Number of Instances      9371

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	no
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	yes
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

Confusion Matrix

```

a  b  <-- classified as
8265  0 |  a = no
0 1106 |  b = yes
```

Result of Naïve Bayes:

Training set:

The screenshot shows the Weka Explorer interface with the Naïve Bayes classifier selected. The training set is 'lazy.BK'. The classifier output shows a summary of performance metrics and a detailed accuracy by class table.

Classifier output

```
mean      5179.2452 5093.8252
std. dev.  65.6272 86.9747
weight sum  8265    1106
precision  26.45   26.45

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
Time taken to test model on training data: 0.07 seconds

=== Summary ===
Correctly Classified Instances      8173      87.2152 %
Incorrectly Classified Instances    1198      12.7841 %
Kappa statistic              0.4462
Mean absolute error          0.1414
Root mean squared error      0.3317
Relative absolute error       67.9116 %
Root relative squared error   102.8028 %
Total Number of Instances      9371

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.911	0.416	0.942	0.911	0.926	0.450	0.872	0.977	no
	0.584	0.089	0.467	0.584	0.519	0.450	0.872	0.498	yes
Weighted Avg.	0.872	0.377	0.886	0.872	0.878	0.450	0.872	0.920	

Confusion Matrix

```

a  b  <-- classified as
7527 738 |  a = no
460 646 |  b = yes
```

Test set:

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section on the left has 'Supplied test set' selected. The 'Classifier output' pane on the right displays the following results:

```
mean      5175.2452 5093.8252
std. dev.  65.6272  86.9747
weight sum  8265    1106
precision  26.45   26.45

Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances      8173      87.222 %
Incorrectly Classified Instances    1198
Kappa statistic                    0.4462
Mean absolute error                 0.1414
Root mean squared error             0.3317
Relative absolute error             67.9116 %
Root relative squared error         102.8028 %
Total Number of Instances          9371

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      ----
0.511   0.416   0.942    0.911   0.926     0.450   0.872    0.977    no
0.584   0.089   0.467    0.584   0.515     0.450   0.872    0.498    yes
Weighted Avg.   0.872   0.377   0.886    0.872   0.878     0.450   0.872    0.920

=== Confusion Matrix ===

      a   b   <-- classified as
7527  738 |   a = no
 460  646 |   b = yes
```

Results:

By using these particular algorithms, we got some different results. We used K-Nearest Neighbor algorithm, Naïve Bayes procedure. From these K-Nearest Neighbor (KNN) procedure is the best algorithm for this problem. I got 100% accuracy of KNN in training set and also got 100% in test set. On the other hand, in Naïve Bayes procedure I got 87.22% accuracy in training set and test set. So, I prefer the K-Nearest Neighbor procedure.

