06.01.2020

Applying Classification Algorithms with WEKA

Data Warehousing and Mining (ISE 401)

Selin Karademir

20180301044

Computer Engineering, Engineering Department

INTRODUCTION:

1.1 The classifiers chosen to be used on chosen dataset: KNN, MLP, ID3 J48s

- KNN (k-nearest neighbours) is a supervised learning algorithm and a type of lazy learner; it doesn't learn a discriminative function from the training dataset, instead it memorizes the training dataset.

Due to it being easily implemented by a machine, it is widely preferred for machine learning applications.

In a classification problem, to find out if a given sample belongs to class A or class B, KNN uses a distance metric to find k number of points that are closest to the sample we want to classify in the training dataset.

A voting is made on predicted class labels based on the classes of the k-nearest neighbours. The class label of the sample we want to classify is then determined by a majority vote.

-ID3 (Iterative Dichotomiser 3) is a supervised algorithm that utilizes supervised learning technique. It constructs a decision tree structure (using training data) by predicting a best attribute that yields:

minimum Entropy (uncertainty level in S training dataset)

maximum Information Gain (IG) (the reduced amount of uncertainty after splitting S training dataset into subsets)

In the tree structures,

* Leaves represent classification labels

* Non-leaf nodes are features (attributes)

* Branches represent a decision (rule)

-MLP (Multilayer Perceptron):

It is thought that in the brain, the transactions and memory are distributed concurrently amongst neural networks. While the transactions occur within each neuron in the neural network, memory acts as a link between transactions and neural networks. Thus, a non-linear approach is observed.

We try to implement this knowledge by mimicing biology to distinguish data that cannot be separeted linearly.

MLP algorithm is supervised and is a neural network that is composed of more than one perceptron, meaning it has more than one linear layer, so is widely used for separating non-linear data.

The MLP networks are composed of many functions (layers) that are chained together.

For a three-layer network, it consists of an input layer, output layer and a hidden layer inbetween. We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task.

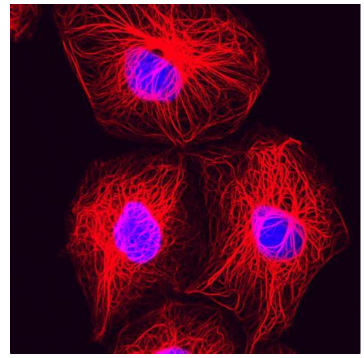
MLP uses backpropagation algorithm to train the model, which consists of:

1. Forward Pass: passing the input to model and multiplying with weights, adding bias at every layer and finding the calculated output of the model.

2. Calculate error or loss: calculating the correctness of predicted data from forward pass using an error function (checking if weights need update). If yes, apply backward pass.

3. Backward pass: if error exists, the weights of the model are updated for the predicted output to be closer to the target output.

1.2 Information about chosen dataset and its attributes:



Title: Breast Cancer Data Set

Breast Cancer Data (Restricted Access), donated in 1988

Sources: Matjaz Zwitter & Milan Soklic (physicians)
Institute of Oncology
University Medical Center
Ljubljana, Yugoslavia

Donors: Ming Tan and Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

Dataset Information: This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Attribute Characteristics	Categorical	No. of instances	286
Associated Tasks	Classification	No. of attributes	9, 1 class

Attribute Information:

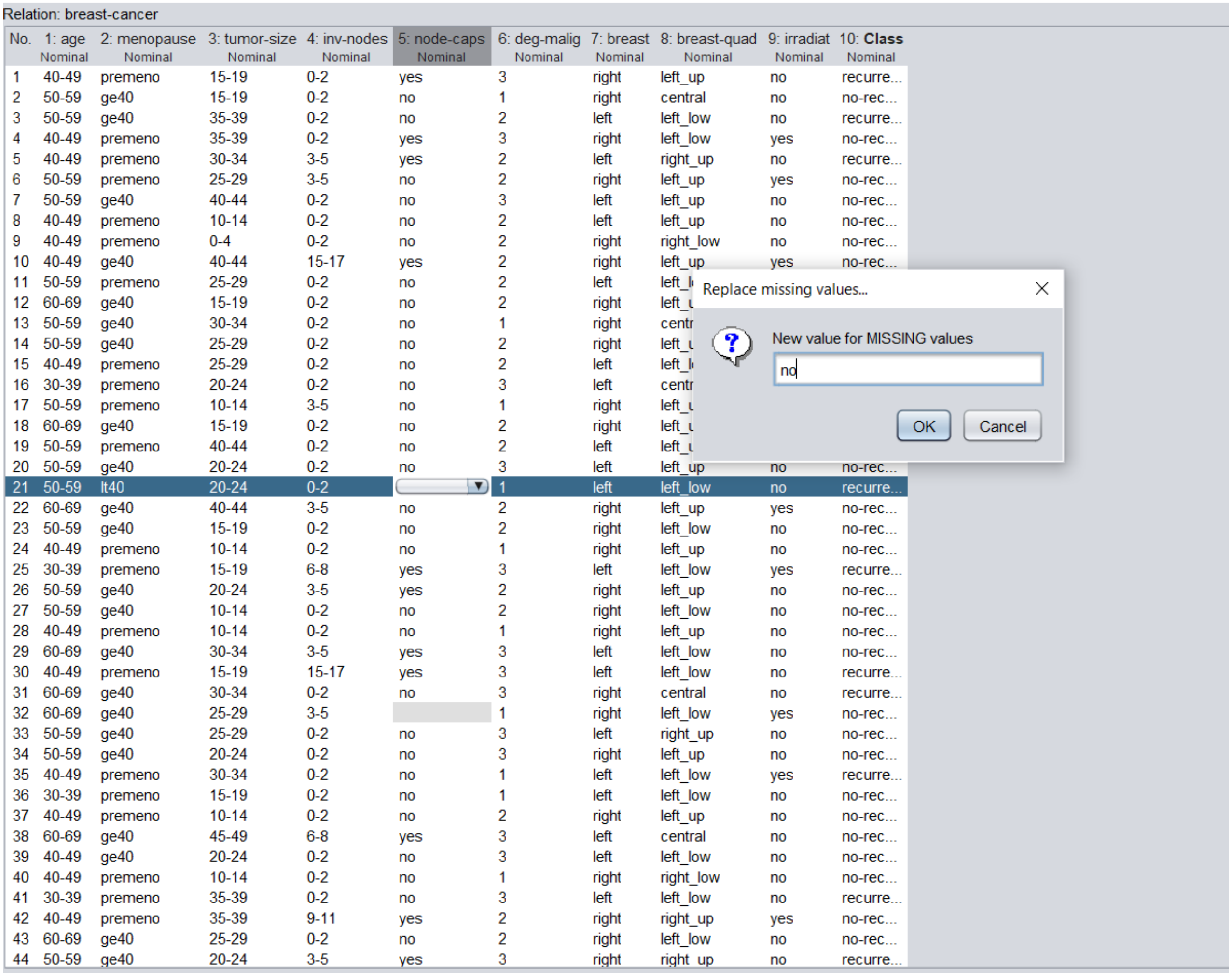
1. **Class:** does the breast cancer occur again in the treated breast? *{no-recurrence-events, recurrence-events}*
2. **age:** age interval for cancer patients *{10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99}.*
3. **menopause:** menopause status of patient at the time diagnosis is considered: (less than 40, greater than or equal to 40, premenopause) *{lt40, ge40, premeno}*
4. **tumor-size:** size of the formed tumor in milimeters. *{0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59}*
5. **inv-nodes:** number interval of invasive cancers in nodes *{0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39}*
6. **node-caps:** have cancer cells reached the lymph nodes? {*yes, no*}
7. **deg-malig:** degree of malignancy of the cell *{1, 2, 3}*
8. **breast:** which breast tissue is the cancer in?*{left, right}*
9. **breast-quad:** which quadrant of the breast is the cancer cell in? *{left-up, left-low, right-up, right-low, central}*
10. **irradiat:** has patient received breast irradiation? *{yes, no}*

This dataset has restricted access. Thus, some attribute value are unknown:

Missing Attribute Values: (denoted by "?") 9 values missing in this dataset: node_caps: 8, breast_quad: 1

1.3 Replacing missing values

For experimental purposes, I've replaced the missing values for the first part of the question with this method. I did this for all attributes with missing values, and saved this arff file as "breast-cancer_replaced.arff"



No.	1: age Nominal	2: menopause Nominal	3: tumor-size Nominal	4: inv-nodes Nominal	5: node-caps Nominal	6: deg-malig Nominal	7: breast Nominal	8: breast-quad Nominal	9: irradiat Nominal	10: Class Nominal
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurre...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-rec...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurre...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-rec...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurre...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-rec...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-rec...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-rec...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-rec...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-rec...
11	50-59	premeno	25-29	0-2	no	2	left	left_l		
12	60-69	ge40	15-19	0-2	no	2	right	left_u		
13	50-59	ge40	30-34	0-2	no	1	right	centr		
14	50-59	ge40	25-29	0-2	no	2	right	left_u		
15	40-49	premeno	25-29	0-2	no	2	left	left_l		
16	30-39	premeno	20-24	0-2	no	3	left	centr		
17	50-59	premeno	10-14	3-5	no	1	right	left_u		
18	60-69	ge40	15-19	0-2	no	2	right	left_u		
19	50-59	premeno	40-44	0-2	no	2	left	left_u		
20	50-59	ge40	20-24	0-2	no	3	left	left_up	no	no-rec...
21	50-59	lt40	20-24	0-2		1	left	left_low	no	recurre...
22	60-69	ge40	40-44	3-5	no	2	right	left_up	yes	no-rec...
23	50-59	ge40	15-19	0-2	no	2	right	left_low	no	no-rec...
24	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-rec...
25	30-39	premeno	15-19	6-8	yes	3	left	left_low	yes	recurre...
26	50-59	ge40	20-24	3-5	yes	2	right	left_up	no	no-rec...
27	50-59	ge40	10-14	0-2	no	2	right	left_low	no	no-rec...
28	40-49	premeno	10-14	0-2	no	1	right	left_up	no	no-rec...
29	60-69	ge40	30-34	3-5	yes	3	left	left_low	no	no-rec...
30	40-49	premeno	15-19	15-17	yes	3	left	left_low	no	recurre...
31	60-69	ge40	30-34	0-2	no	3	right	central	no	recurre...
32	60-69	ge40	25-29	3-5		1	right	left_low	yes	no-rec...
33	50-59	ge40	25-29	0-2	no	3	left	right_up	no	no-rec...
34	50-59	ge40	20-24	0-2	no	3	right	left_up	no	no-rec...
35	40-49	premeno	30-34	0-2	no	1	left	left_low	yes	recurre...
36	30-39	premeno	15-19	0-2	no	1	left	left_low	no	no-rec...
37	40-49	premeno	10-14	0-2	no	2	right	left_up	no	no-rec...
38	60-69	ge40	45-49	6-8	yes	3	left	central	no	no-rec...
39	40-49	ge40	20-24	0-2	no	3	left	left_low	no	no-rec...
40	40-49	premeno	10-14	0-2	no	1	right	right_low	no	no-rec...
41	30-39	premeno	35-39	0-2	no	3	left	left_low	no	recurre...
42	40-49	premeno	35-39	9-11	yes	2	right	right_up	yes	no-rec...
43	60-69	ge40	25-29	0-2	no	2	right	left_low	no	no-rec...
44	50-59	ge40	20-24	3-5	yes	3	right	right_up	no	recurre...

1.4 The results when we use split ratio option in WEKA (80% 20%) (train 80%, test on 20%)

Correctly classified instances can also be seen on the confusion matrix provided by WEKA as the main diagonal.

(Note that these results are obtained by only changing said table values. Rest of the values are default.)

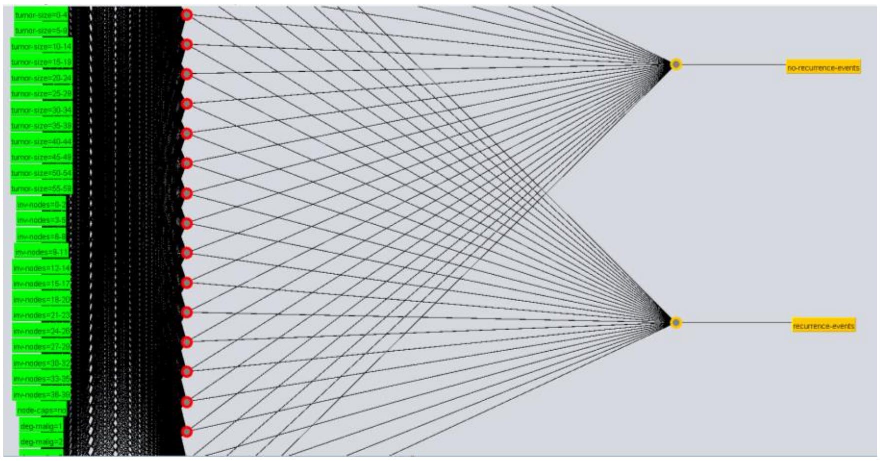
KNN depends on the k parameter:

K value	3	5	11	13
Correctly Classified Instances	38	39	39	38
Incorrectly Classified Instances	19	18	18	19
Accuracy level	66.6667 %	68.4211 %	68.4211 %	66.6667 %

MLP depends on learning rate (n):

n value	0.3	0.5	0.04	0.004
Correctly Classified Instances	37	36	37	39
Incorrectly Classified Instances	20	21	20	18
Accuracy level	64.9123 %	63.1579 %	64.9123 %	68.4211 %

e.g. this is what the neural network looks like for n = 0.3:

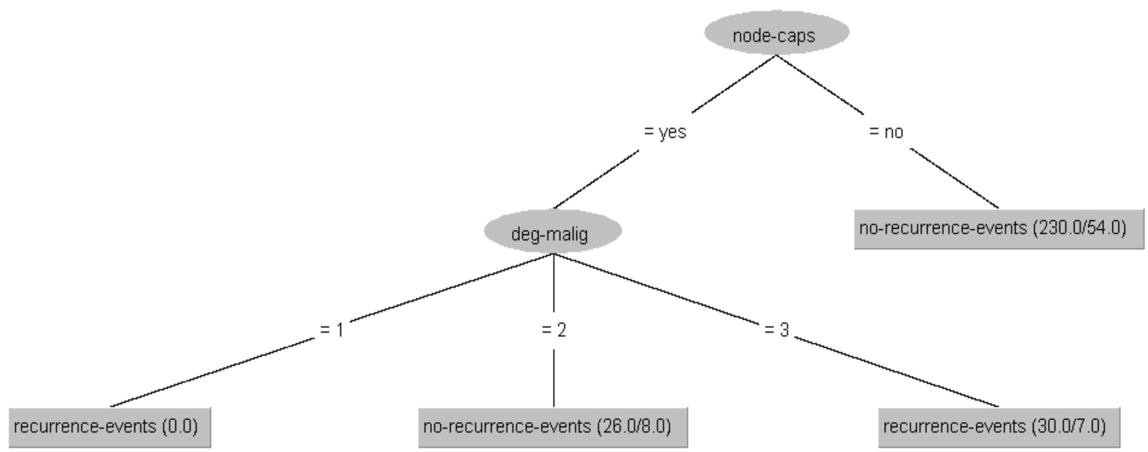


ID3 algorithm using J48 tree

J48 tree	
Correctly Classified Instances	42
Incorrectly Classified Instances	15
Accuracy level	73.6842 %

- **The visual J48 tree model:**

We can see that the node-caps attribute contains the largest gain value, thus it is the main node. It plays the largest role when splitting the other outcomes.



1.5 The results when we use k-fold cross validation option in WEKA

Test mode: 10-fold (since data isn't too big) cross-validation

KNN depends on the k parameter:

K value	3	5	11	13
Correctly Classified Instances	212	211	208	210
Incorrectly Classified Instances	74	75	78	76
Accuracy level	74.1259 %	73.7762 %	72.7273 %	73.4266 %

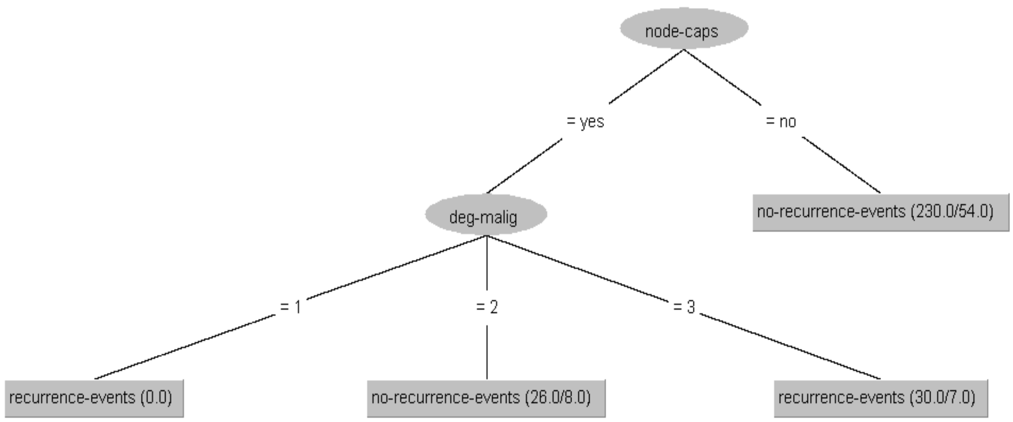
MLP depends on learning rate (n):

n value	0.3	0.5	0.04	0.004
Correctly Classified Instances	183	191	197	209
Incorrectly Classified Instances	103	95	89	77
Accuracy level	63.986 %	66.7832 %	68.8811 %	73.0769 %

ID3 algorithm using J48 tree:

J48 tree	
Correctly Classified Instances	216
Incorrectly Classified Instances	70
Accuracy level	75.5245 %

- The visual J48 tree model:



1.6 Comparison of the highest accuracy levels in both modes for the classifiers

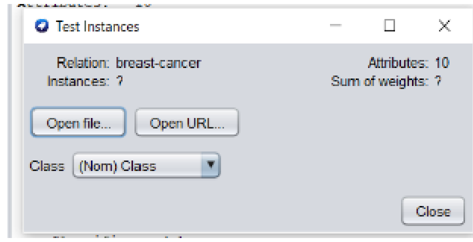
Classification Algorithms	Percentage Split Highest Accuracy Level (%80 %20)		Cross-fold Highest Accuracy Level (10-fold)	
KNN (k-parameter)	for k = 5, k = 11	68.42%	for k = 3	74.13%
MLP (n-parameter)	for n = 0.004	68.42%	for n = 0.004	73.08%
ID3 J48	73.68%		75.52%	

Appearently, cross-fold of 10 test mode helps us achieve higher accuracy results when working with this specific breast cancer dataset.

1.7 Using some items with unknown class labels and determining their classes with Weka

***Test mode:** preferred to use cross-fold since accuracy levels are higher for this specific dataset*

USING KNN: On my “breast-cancer_replaced” arff file, I chose k = 3 and found 74.1259 % which is the highest accuracy level. Now with this accuracy level, I will test my test set. To do this, I loaded my saved “model1_knn” and chose my supplied test set as “breast-cancer” arff file (which has missing values).



We re-evaluate the model and this is the result: (*Since there are 286 instances, I've included some of them for demonstration)*

```
=== Predictions on test set ===

   inst#     actual  predicted error prediction
       1 2:recurrence-events 2:recurrence-events       0.8
       2 1:no-recurrence-events 1:no-recurrence-events       0.999
       3 2:recurrence-events 1:no-recurrence-events   +   0.833
       4 1:no-recurrence-events 1:no-recurrence-events       0.999
       5 2:recurrence-events 1:no-recurrence-events   +   0.5
       6 1:no-recurrence-events 1:no-recurrence-events       0.739
       7 1:no-recurrence-events 1:no-recurrence-events       0.999
       8 1:no-recurrence-events 1:no-recurrence-events       0.833
       9 1:no-recurrence-events 1:no-recurrence-events       0.864
      10 1:no-recurrence-events 1:no-recurrence-events       0.999

                                  . . .
     276 1:no-recurrence-events 1:no-recurrence-events       0.889
     277 1:no-recurrence-events 1:no-recurrence-events       0.75
     278 1:no-recurrence-events 1:no-recurrence-events       0.999
     279 1:no-recurrence-events 1:no-recurrence-events       0.6
     280 1:no-recurrence-events 1:no-recurrence-events       0.999
     281 1:no-recurrence-events 1:no-recurrence-events       0.8
     282 1:no-recurrence-events 1:no-recurrence-events       0.5
     283 1:no-recurrence-events 1:no-recurrence-events       0.666
     284 1:no-recurrence-events 1:no-recurrence-events       0.625
     285 1:no-recurrence-events 1:no-recurrence-events       0.75
     286 1:no-recurrence-events 1:no-recurrence-events       0.999
```

```
=== Summary ===

Correctly Classified Instances         230               80.4196 %
Incorrectly Classified Instances        56               19.5804 %
Kappa statistic                          0.44
Mean absolute error                      0.2703
Root mean squared error                  0.3685
Relative absolute error                 64.6114 %
Root relative squared error             80.6365 %
Total Number of Instances              286
```

In summary, Weka could predict 80.4196% instances correctly when using KNN algorithm, where k = 3 and crossfold of 10 is preferred.

USING MLP: Again, I will do my test for the highest accuracy value I got before. Thus, I choose **n = 0.004,** which gave the result of 73.0769 % accuracy level.

After re-evaluating the model:

```
Predictions on test set ===

inst#     actual  predicted error prediction
    1 2:recurrence-events 1:no-recurrence-events   +   0.684
    2 1:no-recurrence-events 1:no-recurrence-events       0.939
    3 2:recurrence-events 1:no-recurrence-events   +   0.92
    4 1:no-recurrence-events 1:no-recurrence-events       0.56
    5 2:recurrence-events 2:recurrence-events       0.666
    6 1:no-recurrence-events 1:no-recurrence-events       0.779
    7 1:no-recurrence-events 1:no-recurrence-events       0.843
    8 1:no-recurrence-events 1:no-recurrence-events       0.917
    9 1:no-recurrence-events 1:no-recurrence-events       0.909
   10 1:no-recurrence-events 1:no-recurrence-events       0.884

                            ...

  276 1:no-recurrence-events 1:no-recurrence-events       0.912
  277 1:no-recurrence-events 1:no-recurrence-events       0.819
  278 1:no-recurrence-events 1:no-recurrence-events       0.864
  279 1:no-recurrence-events 2:recurrence-events   +   0.598
  280 1:no-recurrence-events 1:no-recurrence-events       0.82
  281 1:no-recurrence-events 1:no-recurrence-events       0.572
  282 1:no-recurrence-events 1:no-recurrence-events       0.736
  283 1:no-recurrence-events 1:no-recurrence-events       0.522
  284 1:no-recurrence-events 2:recurrence-events   +   0.621
  285 1:no-recurrence-events 1:no-recurrence-events       0.887
  286 1:no-recurrence-events 1:no-recurrence-events       0.672
```

In summary, Weka could predict 79.3706% intances correctly where cross-fold 10 and MLP n=0.004.

```
=== Summary ===

Correctly Classified Instances         227               79.3706 %
Incorrectly Classified Instances        59               20.6294 %
Kappa statistic                          0.4522
Mean absolute error                      0.3099
Root mean squared error                  0.3947
Relative absolute error                 74.0773 %
Root relative squared error             86.3622 %
Total Number of Instances              286
```

USING ID3 J48: We achieved a higher accuracy level when cross-fold of 10 was used (75.5245 %) hence we'll repeat this for our test model.

After re-evaluating the model:

```
== Predictions on test set ===

   inst#     actual  predicted error prediction
       1 2:recurrence-events 2:recurrence-events       0.767
       2 1:no-recurrence-events 1:no-recurrence-events       0.765
       3 2:recurrence-events 1:no-recurrence-events   +   0.765
       4 1:no-recurrence-events 2:recurrence-events   +   0.767
       5 2:recurrence-events 1:no-recurrence-events   +   0.692
       6 1:no-recurrence-events 1:no-recurrence-events       0.765
       7 1:no-recurrence-events 1:no-recurrence-events       0.765
       8 1:no-recurrence-events 1:no-recurrence-events       0.765
       9 1:no-recurrence-events 1:no-recurrence-events       0.765
      10 1:no-recurrence-events 1:no-recurrence-events       0.692

                              ...

  276 1:no-recurrence-events 1:no-recurrence-events       0.765
  277 1:no-recurrence-events 1:no-recurrence-events       0.765
  278 1:no-recurrence-events 1:no-recurrence-events       0.765
  279 1:no-recurrence-events 1:no-recurrence-events       0.692
  280 1:no-recurrence-events 1:no-recurrence-events       0.765
  281 1:no-recurrence-events 1:no-recurrence-events       0.765
  282 1:no-recurrence-events 1:no-recurrence-events       0.692
  283 1:no-recurrence-events 1:no-recurrence-events       0.692
  284 1:no-recurrence-events 1:no-recurrence-events       0.692
  285 1:no-recurrence-events 1:no-recurrence-events       0.765
  286 1:no-recurrence-events 1:no-recurrence-events       0.765
```

In summary, Weka could predict 75.8741% of intances correctly where ID3 J48 classification is used with cross-fold of 10.

```
=== Summary ===

Correctly Classified Instances         217               75.8741 %
Incorrectly Classified Instances        69               24.1259 %
Kappa statistic                          0.2899
Mean absolute error                      0.3658
Root mean squared error                  0.427
Relative absolute error                 87.4414 %
Root relative squared error             93.4284 %
Total Number of Instances              286
```

2.1 Apriori algorithm informaiton

$$X \rightarrow Y$$

Apriori Algorithm: is an algorithm that works with categorical data and can be used for both supervised and unsupervised conditions. It is named "a priori" because it uses prior knowledge of frequent itemset properties. **Association rules** are used to identify the set of items or attributes that occur together in a table. An itemset consists of two or more items. To see if the rules are useful, we use three metrics:

-Support: Fraction of transactions that contain an itemset. (Number of Transactions Containing X) / (Total Number of Transactions)

-Confidence: This says how likely item Y is present when item X is present, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appearss.

-Lift: This says how likely item Y is present when item X is present, while controlling for how popular item Y is. If **lift value >** 1 , this means that item Y is *likely* to be present if item X is present, while a value < 1 means that item Y is *unlikely* to be present if item X is present.

Thus, if the lift of { X -> Y } equals 1, it implies no association between items.

2.2 Dataset and attribute information

Labor Relations Data Set

Final settlements in labor negotitions in Canadian industry

Source: *Collective Barganing Review, montly publication, Labour Canada, Industrial Relations Information Service, Ottawa, Ontario, K1A 0J2, Canada, (819) 997-3117*

Dataset Information: The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc) in Canada in 87 and first quarter of 88.

Data was used to test 2 tier approach with learning from positive and negative examples.

Attribute Characteristics	Nominal	No. of instances	57
Associated Tasks	Classification	No. of attributes	16, 1 class attribute

Note that the dataset contains some missing values.

Attribute Information: for the original dataset

1. **dur:** duration of work agreement *[1..7]*

2. **wage1.wage** : wage increase in first year of contract *[2.0 .. 7.0]*

3. **wage2.wage** : wage increase in second year of contract *[2.0 .. 7.0]*

4. **wage3.wage** : wage increase in third year of contract *[2.0 .. 7.0]*

5. **cola** : cost of living allowance *[none, tcf, tc]* (treating customers fairly, too costly)

6. **hours.hrs** : number of working hours during week *[35 .. 40]*

7. **pension** : employer contributions to pension plan *[none, ret_allw, empl_contr]*

8. **stby_pay** : standby pay: additional pay for employees required to be immediately available for duty *[2 .. 25]*

9. **shift_diff** : shift differencial : supplement for working on extra shifts *[1 .. 25]*

10. **educ_allw.boolean** : is education allowance given? *[true false]*

11. **holidays** : number of statutory (public) holidays. depending on your eligibility and employer, you are entitled to a day off without losing pay on these dates. *[9 .. 15]*

12. **vacation** : amount of paid vacation days *[ba, avg, gnr]* (bad amount, average, generous)

13. **lngtrm_disabil.boolean** : employer's help during employee's longterm disability *[true , false]*

14. **dntl_ins** : employers contribution towards the dental plan *[none, half, full]*

15. **bereavement.boolean** : employer's financial contribution towards the covering the costs of bereavement *[true , false]*

16. **empl_hplan** : employer's contribution towards the health plan *[none, half, full]*

2.3 Numeric to nominal conversion and handling missing data

Before continuing, the numeric values should be converted to nominal values since Apriori only works with categorical data. Weka discretes data to separate numeric data into bins, and converts them to nominal. Afterwards, with a manual manner, we change the labels into more understandable names:

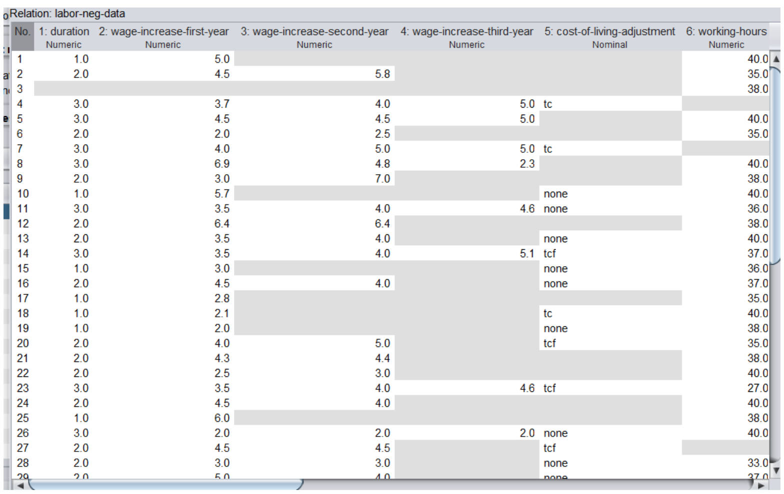
```
@relation labor-neg-data-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-R1-precision6-weka.filters.

@attribute duration {SHORT_hrs,LONG_dur}
@attribute wage-increase-first-year {LOW_wage,AVERAGE,HIGH_wage}
@attribute wage-increase-second-year {LOW_wage,AVERAGE,HIGH_wage}
@attribute wage-increase-third-year {LOW_wage_wage3,AVERAGE_wage3,HIGH_wage_wage3}
@attribute cost-of-living-adjustment {none,tcf,tc}
@attribute working-hours {SHORT_hrs,AVERAGE_hrs,LONG_hrs}
@attribute pension {none,ret_allw,empl_contr}
@attribute standby-pay {LOW_stby_pay,AVERAGE_stby_pay,HIGH_stby_pay}
@attribute shift-differential {LOW_wage_shiftdiff,AVERAGE_shiftdiff,HIGH_wage_shiftdiff}
@attribute education-alSHORT_hrsance {yes,no}
@attribute statutory-holidays {LOW_no_hldys,HIGH_no_hldys}
@attribute vacation {LOW_no_vacation,AVERAGE,HIGH_no_vacation}
@attribute LONG_durterm-disability-assistance {yes,no}
@attribute contribution-to-dental-plan {none,half,full}
@attribute bereavement-assistance {yes,no}
@attribute contribution-to-health-plan {none,half,full}
@attribute class {bad,good}

```

Also handling the missing data with Weka GUI:

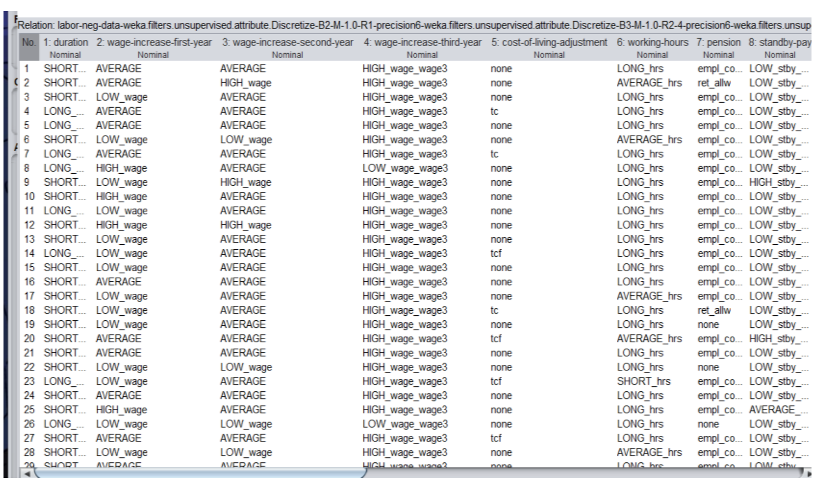
Before:



Relation: labor-neg-data

No.	1: duration Numeric	2: wage-increase-first-year Numeric	3: wage-increase-second-year Numeric	4: wage-increase-third-year Numeric	5: cost-of-living-adjustment Nominal	6: working-hours Numeric
1	1.0	5.0				40.0
2	2.0	4.5	5.8			35.0
3						38.0
4	3.0	3.7	4.0	5.0	tc	
5	3.0	4.5	4.5	5.0		40.0
6	2.0	2.0	2.5			35.0
7	3.0	4.0	5.0	5.0	tc	
8	3.0	6.9	4.8	2.3		40.0
9	2.0	3.0	7.0			38.0
10	1.0	5.7			none	40.0
11	3.0	3.5	4.0	4.6	none	36.0
12	2.0	6.4	6.4			38.0
13	2.0	3.5	4.0		none	40.0
14	3.0	3.5	4.0	5.1	tcf	37.0
15	1.0	3.0			none	36.0
16	2.0	4.5	4.0		none	37.0
17	1.0	2.8				35.0
18	1.0	2.1			tc	40.0
19	1.0	2.0			none	38.0
20	2.0	4.0	5.0		tcf	35.0
21	2.0	4.3	4.4			38.0
22	2.0	2.5	3.0			40.0
23	3.0	3.5	4.0	4.6	tcf	27.0
24	2.0	4.5	4.0			40.0
25	1.0	6.0				38.0
26	3.0	2.0	2.0	2.0	none	40.0
27	2.0	4.5	4.5		tcf	
28	2.0	3.0	3.0		none	33.0
29	2.0	5.0	4.0		none	37.0

After:



Relation: labor-neg-data-weka.filters.unsupervised.attribute.Discretize-B2-M-1.0-R1-precision6-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-4-precision6-weka.filters.unsup

No.	1: duration Nominal	2: wage-increase-first-year Nominal	3: wage-increase-second-year Nominal	4: wage-increase-third-year Nominal	5: cost-of-living-adjustment Nominal	6: working-hours Nominal	7: pension Nominal	8: standby-pay Nominal
1	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
2	SHORT...	AVERAGE	HIGH_wage	HIGH_wage_wage3	none	AVERAGE_hrs	ret_allw	LOW_stby_...
3	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
4	LONG_...	AVERAGE	AVERAGE	HIGH_wage_wage3	tc	LONG_hrs	empl_co...	LOW_stby_...
5	LONG_...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
6	SHORT...	LOW_wage	LOW_wage	HIGH_wage_wage3	none	AVERAGE_hrs	empl_co...	LOW_stby_...
7	LONG_...	AVERAGE	AVERAGE	HIGH_wage_wage3	tc	LONG_hrs	empl_co...	LOW_stby_...
8	LONG_...	HIGH_wage	AVERAGE	LOW_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
9	SHORT...	LOW_wage	HIGH_wage	HIGH_wage_wage3	none	LONG_hrs	empl_co...	HIGH_stby_...
10	SHORT...	HIGH_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
11	LONG_...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
12	SHORT...	HIGH_wage	HIGH_wage	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
13	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
14	LONG_...	LOW_wage	AVERAGE	HIGH_wage_wage3	tcf	LONG_hrs	empl_co...	LOW_stby_...
15	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
16	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
17	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	AVERAGE_hrs	empl_co...	LOW_stby_...
18	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	tc	LONG_hrs	ret_allw	LOW_stby_...
19	SHORT...	LOW_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	none	LOW_stby_...
20	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	tcf	AVERAGE_hrs	empl_co...	HIGH_stby_...
21	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
22	SHORT...	LOW_wage	LOW_wage	HIGH_wage_wage3	none	LONG_hrs	none	LOW_stby_...
23	LONG_...	LOW_wage	AVERAGE	HIGH_wage_wage3	tcf	SHORT_hrs	empl_co...	LOW_stby_...
24	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...
25	SHORT...	HIGH_wage	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	AVERAGE_...
26	LONG_...	LOW_wage	LOW_wage	LOW_wage_wage3	none	LONG_hrs	none	LOW_stby_...
27	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	tcf	LONG_hrs	empl_co...	LOW_stby_...
28	SHORT...	LOW_wage	LOW_wage	HIGH_wage_wage3	none	AVERAGE_hrs	empl_co...	LOW_stby_...
29	SHORT...	AVERAGE	AVERAGE	HIGH_wage_wage3	none	LONG_hrs	empl_co...	LOW_stby_...

2.4 Association rules and comments

Proceeding with applying the Apriori algorithm using default values. The metric is based on confidence level:

```
Apriori
=======

Minimum support: 0.85 (48 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 9

Best rules found:

 1. LONG_durterm-disability-assistance=yes 49 ==> bereavement-assistance=yes 49    <conf:(1)> lift:(1.06) lev:(0.05) [2] conv:(2.58)
 2. standby-pay=LOW_stby_pay 52 ==> shift-differential=LOW_wage_shiftdiff 50    <conf:(0.96)> lift:(1.01) lev:(0.01) [0] conv:(0.91)
 3. wage-increase-third-year=HIGH_wage_wage3 51 ==> shift-differential=LOW_wage_shiftdiff 49    <conf:(0.96)> lift:(1.01) lev:(0.01) [0] c
 4. bereavement-assistance=yes 54 ==> shift-differential=LOW_wage_shiftdiff 51    <conf:(0.94)> lift:(1) lev:(-0) [0] conv:(0.71)
 5. shift-differential=LOW_wage_shiftdiff 54 ==> bereavement-assistance=yes 51    <conf:(0.94)> lift:(1) lev:(-0) [0] conv:(0.71)
 6. standby-pay=LOW_stby_pay 52 ==> bereavement-assistance=yes 49    <conf:(0.94)> lift:(0.99) lev:(-0) [0] conv:(0.68)
 7. statutory-holidays=LOW_no_hldys 52 ==> shift-differential=LOW_wage_shiftdiff 49    <conf:(0.94)> lift:(0.99) lev:(-0) [0] conv:(0.68)
 8. statutory-holidays=LOW_no_hldys 52 ==> bereavement-assistance=yes 49    <conf:(0.94)> lift:(0.99) lev:(-0) [0] conv:(0.68)
 9. wage-increase-third-year=HIGH_wage_wage3 51 ==> bereavement-assistance=yes 48    <conf:(0.94)> lift:(0.99) lev:(-0.01) [0] conv:(0.67)
10. shift-differential=LOW_wage_shiftdiff 54 ==> standby-pay=LOW_stby_pay 50    <conf:(0.93)> lift:(1.01) lev:(0.01) [0] conv:(0.95)
```

COMMENTS ABOUT THE DEPICTED RULES: **deciding the likeliness, unlikeliness or non-association with lift values**

1. *Likely to occur:* if long term disability assistance is provided for 49 people, bereavement assistance is also provided for 49 people
2. *Likely to occur:* if stand by pay is low for 52 people, shift differential wage is also low for the 50 of them.
3. *Unlikely to occur:* if the third year wage increasement is high for 51 people, shift diff. wage is low for the 49 of them.
4. *There's no association between:* bereavement assistance being provided for 54 people, shift diff. wage is low for the 51 of them.
5. *There's no association between:* if shift differential is low for 54 people, bereavement asisstance is provided for 51 of them.
6. *Unlikely to occur:* if the stand by pay is low for 52 people, bereavement assitance is provided for 49 of them.
7. *Unlikely to occur:* if 52 people experience low amounts of statutory holidays, the shift diff. wage is low for 49 of them.
8. *Unlikely to occur:* if 52 people experience low amounts of statutory holidays, bereavement assitance is provided for 49 of them.
9. *Unlikely to occur:* if the third year wage increasement is high for 51 people, bereavement assitance is provided for 48 of them.
10. *Likely to occur:* if the shift diff. wage is low for 54 people, stand by pay is also low for 50 people.

Resources:

MIT OpenCourseWare: 10. Introduction to Learning, Nearest Neighbors:
https://www.youtube.com/watch?v=09mb78oiPkA

https://www.kdnuggets.com/2016/01/implementing-your-own-knn-using-python.html

https://www.r-bloggers.com/2015/04/id3-classification-using-data-tree/

https://iq.opengenus.org/id3-algorithm/

https://storage.googleapis.com/supplemental_media/udacityu/5414400946/ID3%20Algorithm%20for%20Decision%20Trees.pdf

https://becominghuman.ai/understanding-neural-networks-1-the-concept-of-neurons-287be36d40f

https://medium.com/@AI_with_Kain/understanding-of-multilayer-perceptron-mlp-8f179c4a135f

https://www.geeksforgeeks.org/apriori-algorithm/

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

http://archive.ics.uci.edu/ml/datasets/Labor+Relations

https://www.macmillan.org.uk/cancer-information-and-support/breast-cancer/breast-cancer-recurrence

https://training.seer.cancer.gov/breast/anatomy/quadrants.html

https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html