Implementation of Major Classification Algorithms

Aim :

Implement Major Classification Algorithms in WEKA.

1. Naïve Bayes Classifier

In statistics, Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.

Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Implementation of Naïve Bayes Classifier :

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     labor-neg-data
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
```

contribution-to-dental-plan
bereavement-assistance
contribution-to-health-plan
class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naïve Bayes Classifier

Attribute	Class bad (0.36)	good (0.64)
duration		
mean	2	2.25
std. dev.	0.7071	0.6821
weight sum	20	36
precision	1	1
wage-increase-first-year		
mean	2.6563	4.3837
std. dev.	0.8643	1.1773
weight sum	20	36
precision	0.3125	0.3125
wage-increase-second-year		
mean	2.9524	4.447
std. dev.	0.8193	0.9805
weight sum	15	31
precision	0.3571	0.3571
wage-increase-third-year		
mean	2.0344	4.5795
std. dev.	0.1678	0.7893
weight sum	4	11
precision	0.3875	0.3875
cost-of-living-adjustment		
none	10.0	14.0

tcf	2.0	8.0
tc	6.0	3.0
[total]	18.0	25.0
working-hours		
mean	39.4887	37.5491
std. dev.	1.8903	2.9266
weight sum	19	32
precision	1.8571	1.8571
pension		
none	12.0	1.0
ret_allw	3.0	3.0
empl_contr	6.0	8.0
[total]	21.0	12.0
standby-pay		
mean	2.5	11.2
std. dev.	0.866	2.0396
weight sum	4	5
precision	2	2
shift-differential		
mean	2.4691	5.6818
std. dev.	1.5738	5.0584
weight sum	9	22
precision	2.7778	2.7778
education-allowance		
yes	4.0	8.0
no	10.0	4.0
[total]	14.0	12.0
statutory-holidays		
mean	10.2	11.4182
std. dev.	0.805	1.2224
weight sum	20	33
precision	1.2	1.2

```
vacation
  below_average                    12.0     8.0
  average                           8.0    11.0
  generous                          3.0    15.0
  [total]                          23.0    34.0
longterm-disability-assistance
  yes                               6.0    16.0
  no                                9.0     1.0
  [total]                          15.0    17.0
contribution-to-dental-plan
  none                              8.0     3.0
  half                              8.0     9.0
  full                              1.0    14.0
  [total]                          17.0    26.0
bereavement-assistance
  yes                              10.0    19.0
  no                                4.0     1.0
  [total]                          14.0    20.0
contribution-to-health-plan
  none                              9.0     1.0
  half                              3.0     8.0
  full                              7.0    15.0
  [total]                          19.0    24.0
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          51               89.4737 %
Incorrectly Classified Instances         6               10.5263 %
Kappa statistic                          0.7741
Mean absolute error                      0.1042
Root mean squared error                  0.2637
Relative absolute error                 22.7763 %
```

```
Root relative squared error                55.2266 %
Total Number of Instances                  57
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
ROC Area  PRC Area  Class
                 0.900    0.108    0.818      0.900    0.857      0.776
0.965     0.926     bad
                 0.892    0.100    0.943      0.892    0.917      0.776
0.965     0.983     good
Weighted Avg.    0.895    0.103    0.899      0.895    0.896      0.776
0.965     0.963
=== Confusion Matrix ===
  a  b   ← classified as
 18  2 |  a = bad
  4 33 |  b = good
```

2. Decision Trees

Decision tree learning or induction of decision trees is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's target value. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values are called regression trees.

Implementation of Decision Trees :

```
=== Run information ===
Scheme:       weka.classifiers.trees.DecisionStump
Relation:     labor-neg-data
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
```

```
               pension
               standby-pay
               shift-differential
               education-allowance
               statutory-holidays
               vacation
               longterm-disability-assistance
               contribution-to-dental-plan
               bereavement-assistance
               contribution-to-health-plan
               class
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Decision Stump
Classifications
pension = none : bad
pension != none : good
pension is missing : good
Class distributions
pension = none
bad   good
1.0   0.0
pension != none
bad   good
0.4375     0.5625
pension is missing
bad   good
0.06666666666666667   0.9333333333333333
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          46               80.7018 %
```

```
Incorrectly Classified Instances        11               19.2982 %
Kappa statistic                          0.5393
Mean absolute error                      0.2102
Root mean squared error                  0.3358
Relative absolute error                 45.9597 %
Root relative squared error             70.3345 %
Total Number of Instances               57
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
ROC Area  PRC Area  Class
                 0.550    0.054    0.846      0.550    0.667      0.564
0.835     0.815     bad
                 0.946    0.450    0.795      0.946    0.864      0.564
0.835     0.851     good
Weighted Avg.    0.807    0.311    0.813      0.807    0.795      0.564
0.835     0.838
=== Confusion Matrix ===
  a  b   <-- classified as
 11  9 |  a = bad
  2 35 |  b = good.
```

3. Classification and Regression Trees

There are two main types of decision trees:

- Classification tree analysis is when the predicted outcome is the class (discrete) to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number.

The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. in 1984. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.

Implementation of Classification and Regression Trees :

```
=== Run information ===
Scheme:         weka.classifiers.meta.ClassificationViaRegression -W
weka.classifiers.trees.M5P -- -M 4.0
```

```
Relation:     labor-neg-data
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Classification via Regression
Classifier for class with index 0:
M5 pruned model tree:
(using smoothed linear models)
wage-increase-first-year <= 4.55 :
|   pension=none <= 0.5 :
|   |   working-hours <= 36.5 : LM1 (9/0%)
|   |   working-hours >  36.5 :
|   |   |   shift-differential <= 3.5 : LM2 (5/0%)
|   |   |   shift-differential >  3.5 :
```

```
|   |   |   |   wage-increase-first-year <= 2.75 : LM3 (5/83.814%)
|   |   |   |   wage-increase-first-year >  2.75 : LM4 (14/0%)
|   pension=none >  0.5 : LM5 (11/0%)
wage-increase-first-year >  4.55 : LM6 (13/0%)
LM num: 1
class =
      -0.0515 * duration
      - 0.1851 * wage-increase-first-year
      + 0.0443 * working-hours
      + 0.236 * pension=none
      - 0.0225 * shift-differential
      - 0.5762
LM num: 2
class =
      -0.1125 * duration
      - 0.2172 * wage-increase-first-year
      + 0.0364 * working-hours
      + 0.236 * pension=none
      - 0.0261 * shift-differential
      + 0.1224
LM num: 3
class =
      -0.1156 * duration
      - 0.2331 * wage-increase-first-year
      + 0.0364 * working-hours
      + 0.236 * pension=none
      - 0.023 * shift-differential
      + 0.1288
LM num: 4
class =
      -0.1068 * duration
      - 0.2195 * wage-increase-first-year
```

```
      + 0.0364 * working-hours
      + 0.236 * pension=none
      - 0.023 * shift-differential
      + 0.0143
LM num: 5
class =
      -0.0767 * duration
      - 0.1349 * wage-increase-first-year
      + 0.0341 * working-hours
      + 0.3259 * pension=none
      - 0.0183 * shift-differential
      - 0.0512
LM num: 6
class =
      -0.0461 * duration
      - 0.0867 * wage-increase-first-year
      + 0.0238 * working-hours
      + 0.2735 * pension=none
      - 0.0109 * shift-differential
      - 0.2876
Number of Rules : 6
Classifier for class with index 1:
M5 pruned model tree:
(using smoothed linear models)
wage-increase-first-year <= 4.55 :
|   pension=ret_allw,empl_contr <= 0.5 : LM1 (11/0%)
|   pension=ret_allw,empl_contr >  0.5 :
|   |   working-hours <= 36.5 : LM2 (9/0%)
|   |   working-hours >  36.5 :
|   |   |   shift-differential <= 3.5 : LM3 (5/0%)
|   |   |   shift-differential >  3.5 :
|   |   |   |   wage-increase-first-year <= 2.75 : LM4 (5/83.814%)
```

```
|   |   |   |   wage-increase-first-year >  2.75 : LM5 (14/0%)
wage-increase-first-year >  4.55 : LM6 (13/0%)
LM num: 1
class =
      0.0767 * duration
      + 0.1349 * wage-increase-first-year
      - 0.0341 * working-hours
      + 0.3259 * pension=ret_allw,empl_contr
      + 0.0183 * shift-differential
      + 0.7253
LM num: 2
class =
      0.0515 * duration
      + 0.1851 * wage-increase-first-year
      - 0.0443 * working-hours
      + 0.236 * pension=ret_allw,empl_contr
      + 0.0225 * shift-differential
      + 1.3402
LM num: 3
class =
      0.1125 * duration
      + 0.2172 * wage-increase-first-year
      - 0.0364 * working-hours
      + 0.236 * pension=ret_allw,empl_contr
      + 0.0261 * shift-differential
      + 0.6416
LM num: 4
class =
      0.1156 * duration
      + 0.2331 * wage-increase-first-year
      - 0.0364 * working-hours
      + 0.236 * pension=ret_allw,empl_contr
```

```
      + 0.023 * shift-differential
      + 0.6352
LM num: 5
class =
      0.1068 * duration
      + 0.2195 * wage-increase-first-year
      - 0.0364 * working-hours
      + 0.236 * pension=ret_allw,empl_contr
      + 0.023 * shift-differential
      + 0.7497
LM num: 6
class =
      0.0461 * duration
      + 0.0867 * wage-increase-first-year
      - 0.0238 * working-hours
      + 0.2735 * pension=ret_allw,empl_contr
      + 0.0109 * shift-differential
      + 1.0142
Number of Rules : 6
Time taken to build model: 0.19 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          47               82.4561 %
Incorrectly Classified Instances        10               17.5439 %
Kappa statistic                          0.6149
Mean absolute error                      0.2313
Root mean squared error                  0.3283
Relative absolute error                 50.5579 %
Root relative squared error             68.7574 %
Total Number of Instances               57
=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
ROC Area  PRC Area  Class
```

```
                 0.750    0.135    0.750      0.750    0.750      0.615
0.918     0.880      bad
                 0.865    0.250    0.865      0.865    0.865      0.615
0.918     0.951      good
Weighted Avg.    0.825    0.210    0.825      0.825    0.825      0.615
0.918     0.926

=== Confusion Matrix ===

  a  b   <-- classified as

 15  5 |  a = bad

  5 32 |  b = good
```

4. Support Vector Machines (SVMs)

In machine learning, support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

Implementation of Support Vector Machines (SVMs) :

```
=== RUN INFORMATION ===

SCHEME:      WEKA.CLASSIFIERS.FUNCTIONS.SMO -C 1.0 -L 0.001 -P 1.0E-12
-N 0 -V -1 -W 1 -K "WEKA.CLASSIFIERS.FUNCTIONS.SUPPORTVECTOR.POLYKERNEL
-E 1.0 -C 250007" -CALIBRATOR "WEKA.CLASSIFIERS.FUNCTIONS.LOGISTIC -R
1.0E-8 -M -1 -NUM-DECIMAL-PLACES 4"

RELATION:     LABOR-NEG-DATA

INSTANCES:    57

ATTRIBUTES:   17

              DURATION

              WAGE-INCREASE-FIRST-YEAR

              WAGE-INCREASE-SECOND-YEAR

              WAGE-INCREASE-THIRD-YEAR

              COST-OF-LIVING-ADJUSTMENT

              WORKING-HOURS

              PENSION

              STANDBY-PAY
```

```
            SHIFT-DIFFERENTIAL
            EDUCATION-ALLOWANCE
            STATUTORY-HOLIDAYS
            VACATION
            LONGTERM-DISABILITY-ASSISTANCE
            CONTRIBUTION-TO-DENTAL-PLAN
            BEREAVEMENT-ASSISTANCE
            CONTRIBUTION-TO-HEALTH-PLAN
            CLASS
TEST MODE:   10-FOLD CROSS-VALIDATION
=== CLASSIFIER MODEL (FULL TRAINING SET) ===
SMO
KERNEL USED:
  LINEAR KERNEL: K(X,Y) = <X,Y>
CLASSIFIER FOR CLASSES: BAD, GOOD
BINARYSMO
MACHINE LINEAR: SHOWING ATTRIBUTE WEIGHTS, NOT SUPPORT VECTORS.
       0.0754 * (NORMALIZED) DURATION
 +     0.7894 * (NORMALIZED) WAGE-INCREASE-FIRST-YEAR
 +     0.8109 * (NORMALIZED) WAGE-INCREASE-SECOND-YEAR
 +     0.339  * (NORMALIZED) WAGE-INCREASE-THIRD-YEAR
 +    -0.0216 * (NORMALIZED) COST-OF-LIVING-ADJUSTMENT=NONE
 +     0.2843 * (NORMALIZED) COST-OF-LIVING-ADJUSTMENT=TCF
 +    -0.2628 * (NORMALIZED) COST-OF-LIVING-ADJUSTMENT=TC
 +    -0.5644 * (NORMALIZED) WORKING-HOURS
 +    -0.8    * (NORMALIZED) PENSION=NONE
 +     0.2033 * (NORMALIZED) PENSION=RET_ALLW
 +     0.5968 * (NORMALIZED) PENSION=EMPL_CONTR
 +     0.3396 * (NORMALIZED) STANDBY-PAY
 +    -0.0055 * (NORMALIZED) SHIFT-DIFFERENTIAL
 +    -0.5502 * (NORMALIZED) EDUCATION-ALLOWANCE=NO
 +     0.6464 * (NORMALIZED) STATUTORY-HOLIDAYS
```

```
 +      -0.2443 * (NORMALIZED) VACATION=BELOW_AVERAGE
 +      -0.0503 * (NORMALIZED) VACATION=AVERAGE
 +       0.2946 * (NORMALIZED) VACATION=GENEROUS
 +      -1.2183 * (NORMALIZED) LONGTERM-DISABILITY-ASSISTANCE=NO
 +      -0.2628 * (NORMALIZED) CONTRIBUTION-TO-DENTAL-PLAN=NONE
 +      -0.0485 * (NORMALIZED) CONTRIBUTION-TO-DENTAL-PLAN=HALF
 +       0.3113 * (NORMALIZED) CONTRIBUTION-TO-DENTAL-PLAN=FULL
 +      -0.6222 * (NORMALIZED) CONTRIBUTION-TO-HEALTH-PLAN=NONE
 +       0.2688 * (NORMALIZED) CONTRIBUTION-TO-HEALTH-PLAN=HALF
 +       0.3534 * (NORMALIZED) CONTRIBUTION-TO-HEALTH-PLAN=FULL
 -       0.2873

NUMBER OF KERNEL EVALUATIONS: 1055 (93.756% CACHED)

TIME TAKEN TO BUILD MODEL: 0.01 SECONDS

=== STRATIFIED CROSS-VALIDATION ===

=== SUMMARY ===

CORRECTLY CLASSIFIED INSTANCES          51               89.4737 %
INCORRECTLY CLASSIFIED INSTANCES         6               10.5263 %
KAPPA STATISTIC                          0.7635
MEAN ABSOLUTE ERROR                      0.1053
ROOT MEAN SQUARED ERROR                  0.3244
RELATIVE ABSOLUTE ERROR                 23.0111 %
ROOT RELATIVE SQUARED ERROR             67.9505 %
TOTAL NUMBER OF INSTANCES               57

=== DETAILED ACCURACY BY CLASS ===

               TP RATE   FP RATE   PRECISION   RECALL  F-MEASURE  MCC      ROC AREA  PRC AREA  CLASS
                 0.800     0.054      0.889      0.800     0.842       0.766    0.873     0.781     BAD
                 0.946     0.200      0.897      0.946     0.921       0.766    0.873     0.884     GOOD
WEIGHTED AVG.    0.895     0.149      0.894      0.895     0.893       0.766    0.873     0.848

=== CONFUSION MATRIX ===

  A  B   <-- CLASSIFIED AS
```

```
 16  4 |  A = BAD
  2 35 |  B = GOOD
```

5. *k*-Nearest Neighbors Algorithm (*k*-NN)

In statistics, the *k*-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the *k* closest training examples in the dataset. The output depends on whether *k*-NN is used for classification or regression.

Implementation of k-Nearest Neighbors Algorithm (k-NN) :

```
=== Run information ===

Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-
12        -N        0        -V        -1        -W        1        -K
"weka.classifiers.functions.supportVector.PolyKernel   -E   1.0   -C
250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -
M -1 -num-decimal-places 4"

Relation:     labor-neg-data

Instances:    57

Attributes:   17

              duration

              wage-increase-first-year

              wage-increase-second-year

              wage-increase-third-year

              cost-of-living-adjustment

              working-hours

              pension

              standby-pay

              shift-differential

              education-allowance

              statutory-holidays

              vacation

              longterm-disability-assistance

              contribution-to-dental-plan

              bereavement-assistance

              contribution-to-health-plan
```

```
              class
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>
Classifier for classes: bad, good
BinarySMO
Machine linear: showing attribute weights, not support vectors.
         0.0754 * (normalized) duration
 +       0.7894 * (normalized) wage-increase-first-year
 +       0.8109 * (normalized) wage-increase-second-year
 +       0.339  * (normalized) wage-increase-third-year
 +      -0.0216 * (normalized) cost-of-living-adjustment=none
 +       0.2843 * (normalized) cost-of-living-adjustment=tcf
 +      -0.2628 * (normalized) cost-of-living-adjustment=tc
 +      -0.5644 * (normalized) working-hours
 +      -0.8    * (normalized) pension=none
 +       0.2033 * (normalized) pension=ret_allw
 +       0.5968 * (normalized) pension=empl_contr
 +       0.3396 * (normalized) standby-pay
 +      -0.0055 * (normalized) shift-differential
 +      -0.5502 * (normalized) education-allowance=no
 +       0.6464 * (normalized) statutory-holidays
 +      -0.2443 * (normalized) vacation=below_average
 +      -0.0503 * (normalized) vacation=average
 +       0.2946 * (normalized) vacation=generous
 +      -1.2183 * (normalized) longterm-disability-assistance=no
 +      -0.2628 * (normalized) contribution-to-dental-plan=none
 +      -0.0485 * (normalized) contribution-to-dental-plan=half
 +       0.3113 * (normalized) contribution-to-dental-plan=full
 +      -0.6222 * (normalized) contribution-to-health-plan=none
```

```
 +       0.2688 * (normalized) contribution-to-health-plan=half
 +       0.3534 * (normalized) contribution-to-health-plan=full
 -       0.2873

Number of kernel evaluations: 1055 (93.756% cached)

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances          51               89.4737 %
Incorrectly Classified Instances         6               10.5263 %
Kappa statistic                          0.7635
Mean absolute error                      0.1053
Root mean squared error                  0.3244
Relative absolute error                 23.0111 %
Root relative squared error             67.9505 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.800    0.054    0.889      0.800    0.842      0.766    0.873     0.781     bad
                 0.946    0.200    0.897      0.946    0.921      0.766    0.873     0.884     good
Weighted Avg.    0.895    0.149    0.894      0.895    0.893      0.766    0.873     0.848

=== Confusion Matrix ===

  a  b   <-- classified as
 16  4 |  a = bad
  2 35 |  b = good
```