1. **Import to WEKA**

Import selecteddata2\_orig.arff

1. **Filter attribute “mjacg26” for “1” = working**

Filter the attribute related to the person's major activity at the end of reference year

1 = working at job/busi

2 = going to school

3 = retired

4 = other

1. **Removal of attributes from dummy analysis in R**

In the Preprocess tab, select the attributes “immst15” and “mtlswk28” and remove both

1. **Histograms in WEKA to analyze normality**

(Picture of histograms saved in Github)

1. **Transform all attributes from numeric to nominal**

Preprocess – Filter – weka – filters – unsupervised – attribute – “NumericToNominal”

1. **Correlation analysis – qualitative via information gain**

=== Run information ===

Evaluator: weka.attributeSelection.InfoGainAttributeEval

Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: selecteddata2-weka.filters.unsupervised.attribute.Reorder-R1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,19,20,21,22,23,24,25,26,18-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.Remove-R10-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 25564

Attributes: 24

ecage26

ecsex99

marst26

mjacg26

pvreg25

dwtenr25

mortg25

multj28

alhrp28

yrxfte11

clwkr1

prmjb1

fllprt1

nocg2e6

imphwe1

uncoll1

n07c3g10

pubpv10

cqpc42

rppc42

udpd42

wgsal42

hleveg18

penpln1

Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 24 penpln1):

Information Gain Ranking Filter

Ranked attributes:

0.39478 15 imphwe1

0.35364 20 rppc42

0.21221 9 alhrp28

0.21051 22 wgsal42

0.20572 18 pubpv10

0.20509 16 uncoll1

0.18247 17 n07c3g10

0.16615 21 udpd42

0.16329 19 cqpc42

0.13162 11 clwkr1

0.08086 12 prmjb1

0.07193 14 nocg2e6

0.03505 1 ecage26

0.02914 13 fllprt1

0.02712 23 hleveg18

0.02699 10 yrxfte11

0.00582 3 marst26

0.00347 6 dwtenr25

0.00317 5 pvreg25

0.00216 2 ecsex99

0.00183 7 mortg25

0.00157 8 multj28

0 4 mjacg26

Selected attributes: 15,20,9,22,18,16,17,21,19,11,12,14,1,13,23,10,3,6,5,2,7,8,4 : 23

1. **Removal of attributes from correlation analysis**

In the Preprocess tab, select the attribute “mjacg26” remove

1. **Normalize/balanced/imbalanced**

Preprocess – Filter – weka – filters – supervised – instance – “ClassBalancer”

1. **3 algorithms: Decision Tree, Naïve Bayes**

**Decision Tree:**

Decision tree classification was performed on the balanced data set. This was done to reweight the instances in the data so that each class has the same total weight. A J48 decision tree was generated using 10-fold cross validation. The decision tree was generated using penpln1 as the class attribute.

Appendix A shows the true positive rate, false positive rate, precision, and accuracy for the decision tree that was generated from the data. The selected dataset containing all 23 attributes was used and the decision tree correctly classified 83.4% of the instances.

**Naïve Bayes:**

Appendix B

10-fold

82.5% correctly classified

1. **Post-predictive analysis**

Appendix C

Apriori for those with a pension plan