Definition of data mining engineering

Contents

This assignment describes the information excavation procedure utilizing a peculiar information set and a plan called maori hen. At first we use some informations excavation techniques, to preprocess and clean the informations, using filters utilizing maori hen, like take filter and discretization. Then utilizing sort check we apply some basic classifiers to our informations sample, and at the terminal we try to do a anticipation, and reply to the three inquiries of our assignment.

Definition OF DATA Mining

Data excavation refers to “ pull outing or mining cognition from big sums of informations ” . OR, Knowledge Discovery in Databases (KDD) . Alternatively, others view informations excavation as merely an indispensable measure in the procedure of cognition find in databases.

(Review of Business Information Systems – First One-fourth 2008, Ihssan Alkadi)

Weka

Weka is a machine acquisition package, written in Java, and is used for informations analysis, and prognostic mold, inside a graphical user interface. it contains a aggregation of visual image tools and algorithms, and it ‘ s available for free.

Preprocessing

Three preprocessing operations

Three preprocessing operations

Data preprocessing operation is frequently necessary to acquire informations ready for larning. It may besides better the result of the acquisition procedure and lead to more accurate and concise theoretical accounts.

Here is our excel file. (Assignment1_StudentData. xls)

STUDENT DATA EXCEL 1/2

STUDENT DATA EXCEL 2/2

We have to do some alterations, like alteration male with " M " aˆ¦aˆ¦

Changing values, and cleaning our informations. (Assignment1_StudentData file2. xls)

Convert xls to arff format:

First we have to salvage our file as csvaˆ¦ .

(2. Assignment1_StudentData csv)

NOW WE SAVE OUR FILE AS CSV WITH COMMA DELIMITER

Open the file with notepad++ (Csv comma delimiter format)

Relation

The names, the types and the values of the properties are defined by @ property. so following informations types.

The information subdivision of the ARFF file begins with @ informations.

ATTRIBUTE DATA

String with empty infinite, must be enclosed in “ “

Empty infinite must be replaced with “ ? “ , due to preprocessing.

At the terminal we save the file as. arff (student_data. arff)

Runing maori hen.

Click on adventurer

Open the arff file.

Weka arff file is instance sensitive, and has specific format for the variables. If something is n’t in the right format so the file does n’t open.

Case sensitive DATA PRICE (BEng- & gt ; Beng) . We have to replace harmonizing to informations that have been declared on the properties subdivision.

The file is loaded in maori hen. Properties

Remove filter

Now we are traveling to utilize our first filter, in order to take the two properties, given name and surname, because are merely strings and useless for our informations excavation procedure.

On the check preprocess we click on choose to take filter

Choose the remove filter from filters-unsupervised-attribute.

Click on the remove filter

Then we select the figure of properties we want to be removed.

Click on apply

Properties have been removed. Now no1 is gender.

In the arff file we can detect the remove filter is written on the beginning, after the file name.

Discretization

What is discretization?

Some techniques, such as association regulation excavation, can merely be performed on categorical informations. This requires executing discretization on:

Numeral

Continuous properties.

File: 3. Student_Data_removed_discret. arff

After we have choose discretize filter, we enter the index for the property we want to be discretized. We select 3 Numberss of bins.

We do the same for 3 properties and we save it as student_disc and so we open it with notepad++. then we replace some property values with more clear values, utilizing replace on the notepad++ .

WEKA has assigned its ain labels to each of the value ranges for the discretized property. We now need to replace them with more clear values by utilizing Replace. So we replace all “ (-inf-23. 5] ” with the “ 0_23 ”

@ property AGE { “ 0_23 ” , ” 24_27 ” , ” 28_35 ” }

We do the same for the other 2

File: 3. Student_Data_removed_discret. arff

Replace Missing Values Filter

Replacement of Missing Valuess

Now it ‘ s clip to make full in losing values. We do it by choosing from weka-filters-unsupervised-attribute-replace losing values

We can see that Replace losing values filter replaced losing values with some arbitrary values harmonizing to agencies and manners of this property ‘ s dataset.

Valuess replaced.

WEKA Classifiers

Categorization is the processing of happening a set of theoretical accounts (or maps) which describe and distinguish information categories or constructs, for the intents of being able to utilize the theoretical account to foretell the category of objects whose category label is unknown. The derived theoretical account is based on the analysis of a set of preparation informations (i. e. , informations objects whose category label is known) . The derived theoretical account may be represented in assorted signifiers,

such as categorization (IF-THEN) regulations, determination trees, mathematical expression, or nervous webs.

hypertext transfer protocol: //journals. cluteonline. com/index. php/RBIS/article/viewFile/4394/4482

ConjuctiveRule

Choose Conjuctiverule classifier

We cick on the 2nd check on weka – classify. Choose 10 creases transverse proof

The consequences are shown below:

Cases

In (1) we can detect some information like, the strategy (weka. classifiers. rules. ConjunctiveRule) , its parametric quantities, name of the informations file we used, filter ‘ s name. Next we can see the figure of cases, and attributes in the relation.

=== Classifier theoretical account (full preparation set) ===

Single conjunctive regulation scholar:

— — — — — — — — — — — — — — — —

= & gt ; Degree_Classification = Merit

Class distributions:

Covered by the regulation:

Distinction Merit Pass

0. 1875 0. 5 0. 3125

Probability of categories

Not covered by the regulation:

Distinction Merit Pass

0 0 0

Time taken to construct theoretical account: 0 seconds

Conjuctiverule

“ This category implements a individual conjunction regulation scholar (Merit) that can foretell for numeral and nominal category labels ” . It besides gives us the chance distribution over the categories

differentiation about 19 % ,

merit 50 % ,

base on balls about 31 %

=== Predictions on trial informations ===

inst # , existent, predicted, mistake, chance distribution

1 2: Merit 2: Merit 0. 214 *0. 5 0. 286 CORRECT

2 2: Merit 2: Merit 0. 214 *0. 5 0. 286

3 3: Base on balls 2: Merit + 0. 214 *0. 5 0. 286

1 2: Merit 2: Merit 0. 3 *0. 6 0. 1

2 2: Merit 2: Merit 0. 3 *0. 6 0. 1

3 3: Base on balls 2: Merit + 0. 3 *0. 6 0. 1

1 2: Merit 2: Merit 0. 143 *0. 571 0. 286

2 1: Distinct 2: Merit + 0. 143 *0. 571 0. 286

3 3: Base on balls 2: Merit + 0. 143 *0. 571 0. 286

Here we can see the:

case figure,

existent categorization

predicted categorization

mistake

chance distribution

if the existent and the predicted are different, an mistake is shown with a “ + ” . The chance for a anticipation is shown with “ * ” . As we can see the the Immigration and Naturalization Services # 1 was predicted right.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 12 50 %

Falsely Classified Instances 12 50 %

Kappa statistic 0

Mean absolute mistake 0. 4092

Root mean squared error 0. 4625

Relative absolute mistake 98. 0756 %

Root comparative squared error 101. 0429 %

Entire Number of Instances 24

Type of trying that was used (in our instance the stratified cross-validation) .

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 4 0 | a = Differentiation

0 12 0 | B = Merit

0 8 0 | degree Celsius = Pass

HERE WE HAVE A 3X3 CONFUSION MATRIX, WITH three categories, which shows how many cases have been assigned to each category.

4 Differentiations falsely classified as Merit.

12 Merit right classified as Merit

8 Pass falsely classified as Merit.

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0 0 0 0 0. 188 Differentiation

1 1 0. 5 1 0. 667 0. 51 Merit

0 0 0 0 0 0. 539 Base on balls

Weighted Avg. 0. 5 0. 5 0. 25 0. 5 0. 333 0. 466

True Positive Rate False Positive Rate

Tprate

1st. The first degree (degree_classification= Distinction) , the TP Rate is the ratio of degree_classification = Distinction instances predicted right from the sum of existent ‘ Distinction’. 0 cases right predicted as degree_classification= Distinction, 0+4= 4 cases in all that were really ‘ Distinction. The TP Rate = 12/12 = 1.

2nd TP Rate = 12/12 = 1.

3d TP Rate = 0/8 = 0.

Fprate

1st FP Rate is 0/20= 0.

2nd fprate FP Rate is 12/12= 1.

3d fprate is 0/16= 0.

Now we visualize classifier mistakes on maori hen.

Cases with existent degree_classification = Distinction ; (blue)

Cases with degree_classification = Merit (ruddy)

and cases with degree_classification = Pass (green)

The X grade shows that the predicted and existent degree_classification agree.

A square grade indicates that the anticipation of degree_classification was incorrect.

Decision Table

=== Run information ===

Scheme: weka. classifiers. rules. DecisionTable -X 1 -R -S “ weka. attributeSelection. BestFirst -D 1 -N 5 ”

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

Work_Experience_Years

Work_Experience_Type

Entry_Qualifications

Marital status

Childs

Accomodation

Degree_Classification

Trial manner: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

Decision Table:

Number of developing cases: 24

Number of Rules: 2

Non lucifers covered by Majority category.

Best foremost.

Start set: no properties

Search way: forward

Stale hunt after 5 node enlargements

Entire figure of subsets evaluated: 39

Merit of best subset found: 62. 5

Evaluation (for characteristic choice) : CV (leave one out)

Feature set: 1. 9

Rules:

==========

==========

Gender Degree_Classification

==========

==========

M Merit

F Pass

==========

==========

2 regulations produced:

If pupil is a male so “ Merit ”

if a adult female “ Pass ”

Time taken to construct theoretical account: 0. 07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 11 45. 8333 %

Falsely Classified Instances 13 54. 1667 %

Kappa statistic 0. 0127

Mean absolute mistake 0. 4153

Root mean squared error 0. 4741

Relative absolute mistake 99. 5309 %

Root comparative squared error 103. 5729 %

Entire Number of Instances 24

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 2 2 | a = Differentiation

0 9 3 | B = Merit

0 6 2 | degree Celsius = Pass

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0 0 0 0 0. 213 Differentiation

0. 75 0. 667 0. 529 0. 75 0. 621 0. 549 Merit

0. 25 0. 313 0. 286 0. 25 0. 267 0. 434 Base on balls

Weighted Avg. 0. 458 0. 438 0. 36 0. 458 0. 399 0. 454

We besides have 11 right classified cases (about 46 %) and 13 falsely classified cases (54 %) .

Confusion matrix

11 cases were classified right.

4 existent ‘ Distinctions ‘ of which 2 classified as ‘ Merit ‘ and 2 as ‘ Pass ‘ .

12 existent ‘ Merit ‘ of which 9 right classified as ‘ Merit ‘ and 3 classified as ‘ Pass ‘ .

8 existent ‘ Pass ‘ 6 of which classified as ‘ Merit ‘ and 2 right classified as ‘ Pass. ‘

Tp/fp ratio

TP Ratio for Distinction is 0, for Merit is 0. 75 and for Pass is 0. 25.

FP Ratio for Distinction is 0, for Merit 0. 667 and for Pass is 0. 313

J48

=== Run information ===

Scheme: weka. classifiers. trees. J48 -C 0. 25 -M 2

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

Work_Experience_Years

Work_Experience_Type

Entry_Qualifications

Marital status

Childs

Accomodation

Degree_Classification

Trial manner: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

J48 pruned tree

—— —— —— —— —— —— —— —— ——

Gender = F: Pass (6. 0/2. 0)

Gender = M: Merit (18. 0/7. 0)

Number of Leaves: 2

Size of the tree: 3

Time taken to construct theoretical account: 0. 02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 11 45. 8333 %

Falsely Classified Instances 13 54. 1667 %

Kappa statistic 0. 037

Mean absolute mistake 0. 4181

Root mean squared error 0. 5021

Relative absolute mistake 100. 2076 %

Root comparative squared error 109. 6885 %

Entire Number of Instances 24

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 05 0 0 0 0. 069 Differentiation

0. 75 0. 583 0. 563 0. 75 0. 643 0. 514 Merit

0. 25 0. 313 0. 286 0. 25 0. 267 0. 43 Base on balls

Weighted Avg. 0. 458 0. 404 0. 376 0. 458 0. 41 0. 412

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 2 2 | a = Differentiation

0 9 3 | B = Merit

1 5 2 | degree Celsius = Pass

After we have run the j48 classifier tree, we can see the figure of the foliages (2) and the tree size (3) , in text format. The trial manner employed is the 10-cross-fold-validation. If

Gender= male so merit or

if gender = female base on balls.

We besides have 11 right classified cases (about 46 %) and 13 falsely classified cases (54 %) .

confusion matrix

11 cases were classified right.

4 existent ‘ Distinctions ‘ of which 2 classified as ‘ Merit ‘ and 2 as ‘ Pass ‘ .

12 existent ‘ Merit ‘ of which 9 right classified as ‘ Merit ‘ and 3 classified as ‘ Pass ‘ .

8 existent ‘ Pass ‘ , 5 of which classified as ‘ Merit ‘ , 1 classified as ‘ Distinction ‘ and 2 right classified as ‘ Pass. ‘

TP Ratio for Distinction is 0, for Merit is 0. 75 and for Pass is 0. 25. FP Ratio for Distinction is 0. 05, for Merit 0. 583 and for Pass is 0. 313 (explained in ConjuctiveRule) .

Visualize Tree.

NBTREE

=== Run information ===

Scheme: weka. classifiers. trees. NBTree

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

WORK_EXPERIENCE_YEARS

WORK_EXPERIENCE_TYPE

ENTRY_QUALIFICATIONS

Marital status

Child

ACCOMODATION

DEGREE_CLASSIFICATION

Test mode: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

NBTree

— — — — — — — — —

: NB0

Leaf figure: 0 Naive Bayes Classifier

Class

Attribute Distinction Merit Pass

(0. 19) (0. 48) (0. 33)

==========

==========

Gender

F 2. 0 2. 0 5. 0

M 4. 0 12. 0 5. 0

[entire] 6. 0 14. 0 10. 0

Age

0_23 3. 0 4. 0 5. 0

24_27 3. 0 5. 0 2. 0

28_35 1. 0 6. 0 4. 0

[entire] 7. 0 15. 0 11. 0

WORK_EXPERIENCE_YEARS

0_6 2. 0 3. 0 5. 0

7_13 2. 0 6. 0 4. 0

14_20 3. 0 6. 0 2. 0

[entire] 7. 0 15. 0 11. 0

WORK_EXPERIENCE_TYPE

Yes 4. 0 10. 0 8. 0

No 2. 0 4. 0 2. 0

[entire] 6. 0 14. 0 10. 0

ENTRY_QUALIFICATIONS

BSc 3. 0 7. 0 3. 0

BEng 2. 0 3. 0 5. 0

Other 2. 0 5. 0 3. 0

[entire] 7. 0 15. 0 11. 0

Marital status

Married 2. 0 5. 0 3. 0

Single 4. 0 8. 0 7. 0

Widowed 1. 0 2. 0 1. 0

Divorced 1. 0 1. 0 1. 0

[entire] 8. 0 16. 0 12. 0

Child

0 5. 0 10. 0 7. 0

1 1. 0 2. 0 2. 0

2_5 1. 0 3. 0 2. 0

[entire] 7. 0 15. 0 11. 0

ACCOMODATION

Medway 2. 0 9. 0 4. 0

London 3. 0 3. 0 3. 0

Other 2. 0 3. 0 4. 0

[entire] 7. 0 15. 0 11. 0

Number of Leaves: 1

Size of the tree: 1

Time taken to construct theoretical account: 0. 02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8 33. 3333 %

Falsely Classified Instances 16 66. 6667 %

Kappa statistic -0. 1566

Mean absolute mistake 0. 4353

Root mean squared error 0. 5196

Relative absolute mistake 104. 3251 %

Root comparative squared error 113. 5122 %

Entire Number of Instances 24

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 1 0 0 0 0 Differentiation

0. 583 0. 667 0. 467 0. 583 0. 519 0. 486 Merit

0. 125 0. 375 0. 143 0. 125 0. 133 0. 43 Base on balls

Weighted Avg. 0. 333 0. 475 0. 281 0. 333 0. 304 0. 386

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 3 1 | a = Differentiation

0 7 5 | B = Merit

2 5 1 | degree Celsius = Pass

We notice that we have 8 right classified cases (33 %) and 16 falsely classified cases (about 67 %) .

BayesNet

=== Run information ===

Scheme: weka. classifiers. bayes. BayesNet -D -Q weka. classifiers. bayes. net. search. local. K2 — -P 1 -S BAYES -E weka. classifiers. bayes. net. estimate. SimpleEstimator — -A 0. 5

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

Work_Experience_Years

Work_Experience_Type

Entry_Qualifications

Marital status

Childs

Accomodation

Degree_Classification

Trial manner: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

Bayes Network Classifier

non utilizing ADTree

attributes= 9 # classindex= 8

•

Network construction (nodes followed by parents)

Gender (2) : Degree_Classification

Age (3) : Degree_Classification

Work_Experience_Years (3) : Degree_Classification

Work_Experience_Type (2) : Degree_Classification web construction

Entry_Qualifications (3) : Degree_Classification

Marital_Status (4) : Degree_Classification cardinality of the variable.

Children (3) : Degree_Classification

Accomodation (3) : Degree_Classification

Degree_Classification (3) :

Each variable is followed by a list of parents.

LogScore Bayes: -228. 45858687285016

LogScore BDeu: -310. 9532615667094

LogScore MDL: -299. 3766047456865

LogScore ENTROPY: -224. 6923397325098

LogScore AIC: -271. 69233973250977

Here the logarithmic mark of the web construction for assorted methods of marking is listed.

Time taken to construct theoretical account: 0 seconds

=== Predictions on trial informations ===

inst # , existent, predicted, mistake, chance distribution

1 2: Merit 2: Merit 0. 272 *0. 691 0. 037

2 2: Merit 3: Pass + 0. 047 0. 077 *0. 877

3 3: Base on balls 1: Distinct + *0. 509 0. 182 0. 309

Predictions on the information.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8 33. 3333 %

Falsely Classified Instances 16 66. 6667 %

Percentage of the correctly, and falsely classified cases.

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 1 0 0 0 0. 025 Differentiation

0. 583 0. 583 0. 5 0. 583 0. 538 0. 472 Merit

0. 125 0. 438 0. 125 0. 125 0. 125 0. 383 Base on balls

Weighted Avg. 0. 333 0. 454 0. 292 0. 333 0. 311 0. 368

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 2 2 | a = Differentiation

0 7 5 | B = Merit

2 5 1 | degree Celsius = Pass

Confusion matrix.

In order to demo the graphical construction in the consequence list we right snap the BayesNet. And from the pop-up bill of fare we select Visualize Graph.

Naivebayes

=== Classifier theoretical account (full preparation set) ===

Naive Bayes Classifier

Class

Attribute Distinction Merit Pass

(0. 19) (0. 48) (0. 33)

==========

==========

Gender

F 2. 0 2. 0 5. 0

M 4. 0 12. 0 5. 0

[entire] 6. 0 14. 0 10. 0

Age

0_23 3. 0 4. 0 5. 0

24_27 3. 0 5. 0 2. 0

28_35 1. 0 6. 0 4. 0

[entire] 7. 0 15. 0 11. 0

Work_Experience_Years

0_6 2. 0 3. 0 5. 0

7_13 2. 0 6. 0 4. 0

14_20 3. 0 6. 0 2. 0

[entire] 7. 0 15. 0 11. 0

Work_Experience_Type

Yes 4. 0 10. 0 8. 0

No 2. 0 4. 0 2. 0

[entire] 6. 0 14. 0 10. 0

Entry_Qualifications

BSc 3. 0 7. 0 3. 0

BEng 2. 0 3. 0 5. 0

Other 2. 0 5. 0 3. 0

[entire] 7. 0 15. 0 11. 0

Marital status

Married 2. 0 5. 0 3. 0

Single 4. 0 8. 0 7. 0

Widowed 1. 0 2. 0 1. 0

Divorced 1. 0 1. 0 1. 0

[entire] 8. 0 16. 0 12. 0

Childs

0 5. 0 10. 0 7. 0

1 1. 0 2. 0 2. 0

2_5 1. 0 3. 0 2. 0

[entire] 7. 0 15. 0 11. 0

Accomodation

Medway 2. 0 9. 0 4. 0

London 3. 0 3. 0 3. 0

Other 2. 0 3. 0 4. 0

[entire] 7. 0 15. 0 11. 0

With this classifier the end product is a bit different.

Now alternatively of naming chances when degree_classification='Distinction ' or degree_classification= ' Merit ' or ' Pass ' Weka lists a distinct calculator based on the figure of happenings of each property e. g when degree_classification = Distinction the distinct calculator of mentality has the undermentioned counts:

Gender= F: 2, Age= 0_23: 3, Work_Experience_Years= 0_6: 3 and so on.

If we look closer the dataset we can see that these Numberss are greater by 1 than the existent 1s e. g there is merely 1 cases holding degree_classification= Distinction and gender= F but our count is 1+1= 2.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8 33. 3333 %

Falsely Classified Instances 16 66. 6667 %

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 1 0 0 0 0 Differentiation

0. 583 0. 667 0. 467 0. 583 0. 519 0. 486 Merit

0. 125 0. 375 0. 143 0. 125 0. 133 0. 43 Base on balls

Weighted Avg. 0. 333 0. 475 0. 281 0. 333 0. 304 0. 386

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 3 1 | a = Differentiation

0 7 5 | B = Merit

2 5 1 | degree Celsius = Pass

confusion matrix the natural Numberss are shown, with Distinction, Merit and Pass stand foring the category labels. Here there were 24 cases, so the per centums and natural Numberss add up to 3 + 1 = 4, 7 + 5 = 12 and

2+5+1= 8. And we can besides see how the cases were classified from these algorithm.

RandomTree

=== Run information ===

Scheme: weka. classifiers. trees. RandomTree -K 0 -M 1. 0 -S 1

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

WORK_EXPERIENCE_YEARS

WORK_EXPERIENCE_TYPE

ENTRY_QUALIFICATIONS

Marital status

Child

ACCOMODATION

DEGREE_CLASSIFICATION

Test mode: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

RandomTree

==========

AGE = 0_23

| ENTRY_QUALIFICATIONS = BSc

| | ACCOMODATION = Medway: Merit (1/0)

| | ACCOMODATION = London: Distinction (1/0)

| | ACCOMODATION = Other

| | | MARITAL_STATUS = Married: Pass (1/0)

| | | MARITAL_STATUS = Single: Differentiation (1/0)

| | | MARITAL_STATUS = Widowed: Differentiation (0/0)

| | | MARITAL_STATUS = Divorced: Differentiation (0/0)

| ENTRY_QUALIFICATIONS = BEng: Pass (2/0)

| ENTRY_QUALIFICATIONS = Other

| | MARITAL_STATUS = Married: Differentiation (0/0)

| | MARITAL_STATUS = Single

| | | WORK_EXPERIENCE_TYPE = Yes: Pass (1/0)

| | | WORK_EXPERIENCE_TYPE = No: Merit (1/0)

| | MARITAL_STATUS = Widowed: Merit (1/0)

| | MARITAL_STATUS = Divorced: Differentiation (0/0)

AGE = 24_27

| WORK_EXPERIENCE_YEARS = 0_6

| | GENDER = F: Pass (1/0)

| | GENDER = M: Differentiation (1/0)

| WORK_EXPERIENCE_YEARS = 7_13: Merit (2/0)

| WORK_EXPERIENCE_YEARS = 14_20

| | ENTRY_QUALIFICATIONS = BSc: Merit (1/0)

| | ENTRY_QUALIFICATIONS = BEng: Merit (1/0)

| | ENTRY_QUALIFICATIONS = Other: Differentiation (1/0)

AGE = 28_35

| WORK_EXPERIENCE_YEARS = 0_6

| | ACCOMODATION = Medway: Pass (1/0)

| | ACCOMODATION = London: Merit (1/0)

| | ACCOMODATION = Other: Pass (1/0)

| WORK_EXPERIENCE_YEARS = 7_13: Merit (1/0)

| WORK_EXPERIENCE_YEARS = 14_20

| | MARITAL_STATUS = Married: Merit (3/0)

| | MARITAL_STATUS = Single: Pass (1/0)

| | MARITAL_STATUS = Widowed: Differentiation (0/0)

| | MARITAL_STATUS = Divorced: Differentiation (0/0)

Size of the tree: 38

Time taken to construct theoretical account: 0 seconds

=== Predictions on trial informations ===

inst # , existent, predicted, mistake, chance distribution

1 2: Merit 2: Merit 0 *1 0

2 2: Merit 1: Distinct + *1 0 0

3 3: Base on balls 3: Base on balls 0. 25 0 *0. 75

1 2: Merit 2: Merit 0 *1 0

2 2: Merit 3: Pass + 0 0 *1

3 3: Base on balls 1: Distinct + *0. 5 0. 5 0

1 2: Merit 3: Pass + 0 0 *1

2 1: Distinct 3: Pass + 0 0 *1

3 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 1: Distinct 3: Pass + 0 0 *1

3 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 1: Distinct 3: Pass + 0 0 *1

1 2: Merit 3: Pass + 0 0 *1

2 1: Distinct 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 3: Pass + 0. 286 0. 286 *0. 429

2 3: Base on balls 3: Base on balls 0. 286 0. 286 *0. 429

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 9 37. 5 %

Falsely Classified Instances 15 62. 5 %

Kappa statistic -0. 0588

K & A ; B Relative Info Score 30. 3172 %

K & A ; B Information Score 0. 4544 spots 0. 0189 bits/instance

Class complexness | order 0 36. 2284 spots 1. 5095 bits/instance

Class complexness | strategy 15039. 4448 spots 626. 6435 bits/instance

Complexity betterment (Sf) -15003. 2164 spots -625. 134 bits/instance

Mean absolute mistake 0. 4315

Root mean squared error 0. 6334

Relative absolute mistake 103. 4335 %

Root comparative squared error 138. 3867 %

Entire Number of Instances 24

The size of the tree is 38. We have:

9 right classified cases (about 38 %)

15 falsely classified cases (about 63 %) .

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 1 0 0 0 0. 375 Differentiation

0. 583 0. 5 0. 538 0. 583 0. 56 0. 524 Merit

0. 25 0. 438 0. 222 0. 25 0. 235 0. 363 Base on balls

Weighted Avg. 0. 375 0. 413 0. 343 0. 375 0. 358 0. 446

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 1 3 | a = Differentiation

1 7 4 | B = Merit

1 5 2 | degree Celsius = Pass

Confusion matrix

9 cases were classified right.

4 existent ‘ Distinctions ‘ of which 1 classified as ‘ Merit ‘ and 3 as ‘ Pass ‘ .

12 existent ‘ Merit ‘ of which7 right classified as ‘ Merit ‘ , 4 classified as ‘ Pass ‘ and 1classified as ‘ Distinction ‘ .

From the 8 existent ‘ Pass ‘ , 5 of them are classified as ‘ Merit ‘ , 1 classified as ‘ Distinction ‘ and2 right classified as ‘ Pass. ‘

TREE VISUALIZE

NaiveBayesSimple

=== Classifier theoretical account (full preparation set) ===

Naive Bayes (simple)

Class Differentiation: P (C) = 0. 18518519

Attribute Gender

F M

0. 33333333 0. 66666667

Attribute Age

0_23 24_27 28_35

0. 42857143 0. 42857143 0. 14285714

Attribute Work_Experience_Years

0_6 7_13 14_20

0. 28571429 0. 28571429 0. 42857143

Attribute Work_Experience_Type

Yes No

0. 66666667 0. 33333333

Attribute Entry_Qualifications

BSc BEng Other

0. 42857143 0. 28571429 0. 28571429

Attribute Marital_Status

Married Single Widowed Divorced

0. 25 0. 5 0. 125 0. 125

Attribute Children

0 1 2_5

0. 71428571 0. 14285714 0. 14285714

Attribute Accomodation

Medway London Other

0. 28571429 0. 42857143 0. 28571429

Class Merit: P (C) = 0. 48148148

Attribute Gender

F M

0. 14285714 0. 85714286

Attribute Age

0_23 24_27 28_35

0. 26666667 0. 33333333 0. 4

Attribute Work_Experience_Years

0_6 7_13 14_20

0. 2 0. 4 0. 4

Attribute Work_Experience_Type

Yes No

0. 71428571 0. 28571429

Attribute Entry_Qualifications

BSc BEng Other

0. 46666667 0. 2 0. 33333333

Attribute Marital_Status

Married Single Widowed Divorced

0. 3125 0. 5 0. 125 0. 0625

Attribute Children

0 1 2_5

0. 66666667 0. 13333333 0. 2

Attribute Accomodation

Medway London Other

0. 6 0. 2 0. 2

Class Base on balls: P (C) = 0. 33333333

Attribute Gender

F M

0. 5 0. 5

Attribute Age

0_23 24_27 28_35

0. 45454545 0. 18181818 0. 36363636

Attribute Work_Experience_Years

0_6 7_13 14_20

0. 45454545 0. 36363636 0. 18181818

Attribute Work_Experience_Type

Yes No

0. 8 0. 2

Attribute Entry_Qualifications

BSc BEng Other

0. 27272727 0. 45454545 0. 27272727

Attribute Marital_Status

Married Single Widowed Divorced

0. 25 0. 58333333 0. 08333333 0. 08333333

Attribute Children

0 1 2_5

0. 63636364 0. 18181818 0. 18181818

Attribute Accomodation

Medway London Other

0. 36363636 0. 27272727 0. 36363636

Here we see the categorization theoretical account that was produced from NaiveBayesSimple algorithm. We see the listing of the categories and properties and their chances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 8 33. 3333 %

Falsely Classified Instances 16 66. 6667 %

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 3 1 | a = Differentiation

0 7 5 | B = Merit

2 5 1 | degree Celsius = Pass

Random Tree

=== Run information ===

Scheme: weka. classifiers. trees. RandomTree -K 0 -M 1. 0 -S 1

Relation: student_data-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Remove-R1-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R2-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R3-weka. filters. unsupervised. attribute. Discretize-F-B3-M-1. 0-R7-weka. filters. unsupervised. attribute. ReplaceMissingValues

Cases: 24

Properties: 9

Gender

Age

WORK_EXPERIENCE_YEARS

WORK_EXPERIENCE_TYPE

ENTRY_QUALIFICATIONS

Marital status

Child

ACCOMODATION

DEGREE_CLASSIFICATION

Test mode: 10-fold cross-validation

=== Classifier theoretical account (full preparation set) ===

RandomTree

==========

AGE = 0_23

| ENTRY_QUALIFICATIONS = BSc

| | ACCOMODATION = Medway: Merit (1/0)

| | ACCOMODATION = London: Distinction (1/0)

| | ACCOMODATION = Other

| | | MARITAL_STATUS = Married: Pass (1/0)

| | | MARITAL_STATUS = Single: Differentiation (1/0)

| | | MARITAL_STATUS = Widowed: Differentiation (0/0)

| | | MARITAL_STATUS = Divorced: Differentiation (0/0)

| ENTRY_QUALIFICATIONS = BEng: Pass (2/0)

| ENTRY_QUALIFICATIONS = Other

| | MARITAL_STATUS = Married: Differentiation (0/0)

| | MARITAL_STATUS = Single

| | | WORK_EXPERIENCE_TYPE = Yes: Pass (1/0)

| | | WORK_EXPERIENCE_TYPE = No: Merit (1/0)

| | MARITAL_STATUS = Widowed: Merit (1/0)

| | MARITAL_STATUS = Divorced: Differentiation (0/0)

AGE = 24_27

| WORK_EXPERIENCE_YEARS = 0_6

| | GENDER = F: Pass (1/0)

| | GENDER = M: Differentiation (1/0)

| WORK_EXPERIENCE_YEARS = 7_13: Merit (2/0)

| WORK_EXPERIENCE_YEARS = 14_20

| | ENTRY_QUALIFICATIONS = BSc: Merit (1/0)

| | ENTRY_QUALIFICATIONS = BEng: Merit (1/0)

| | ENTRY_QUALIFICATIONS = Other: Differentiation (1/0)

AGE = 28_35

| WORK_EXPERIENCE_YEARS = 0_6

| | ACCOMODATION = Medway: Pass (1/0)

| | ACCOMODATION = London: Merit (1/0)

| | ACCOMODATION = Other: Pass (1/0)

| WORK_EXPERIENCE_YEARS = 7_13: Merit (1/0)

| WORK_EXPERIENCE_YEARS = 14_20

| | MARITAL_STATUS = Married: Merit (3/0)

| | MARITAL_STATUS = Single: Pass (1/0)

| | MARITAL_STATUS = Widowed: Differentiation (0/0)

| | MARITAL_STATUS = Divorced: Differentiation (0/0)

Size of the tree: 38

Time taken to construct theoretical account: 0 seconds

=== Predictions on trial informations ===

inst # , existent, predicted, mistake, chance distribution

1 2: Merit 2: Merit 0 *1 0

2 2: Merit 1: Distinct + *1 0 0

3 3: Base on balls 3: Base on balls 0. 25 0 *0. 75

1 2: Merit 2: Merit 0 *1 0

2 2: Merit 3: Pass + 0 0 *1

3 3: Base on balls 1: Distinct + *0. 5 0. 5 0

1 2: Merit 3: Pass + 0 0 *1

2 1: Distinct 3: Pass + 0 0 *1

3 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 1: Distinct 3: Pass + 0 0 *1

3 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 1: Distinct 3: Pass + 0 0 *1

1 2: Merit 3: Pass + 0 0 *1

2 1: Distinct 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 2: Merit 0 *1 0

2 3: Base on balls 2: Merit + 0 *1 0

1 2: Merit 3: Pass + 0. 286 0. 286 *0. 429

2 3: Base on balls 3: Base on balls 0. 286 0. 286 *0. 429

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 9 37. 5 %

Falsely Classified Instances 15 62. 5 %

Kappa statistic -0. 0588

K & A ; B Relative Info Score 30. 3172 %

K & A ; B Information Score 0. 4544 spots 0. 0189 bits/instance

Class complexness | order 0 36. 2284 spots 1. 5095 bits/instance

Class complexness | strategy 15039. 4448 spots 626. 6435 bits/instance

Complexity betterment (Sf) -15003. 2164 spots -625. 134 bits/instance

Mean absolute mistake 0. 4315

Root mean squared error 0. 6334

Relative absolute mistake 103. 4335 %

Root comparative squared error 138. 3867 %

Entire Number of Instances 24

The size of the tree is 38. We have:

9 right classified cases (about 38 %)

15 falsely classified cases (about 63 %) .

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0 0. 1 0 0 0 0. 375 Differentiation

0. 583 0. 5 0. 538 0. 583 0. 56 0. 524 Merit

0. 25 0. 438 0. 222 0. 25 0. 235 0. 363 Base on balls

Weighted Avg. 0. 375 0. 413 0. 343 0. 375 0. 358 0. 446

=== Confusion Matrix ===

a B degree Celsius & lt ; — classified as

0 1 3 | a = Differentiation

1 7 4 | B = Merit

1 5 2 | degree Celsius = Pass

Confusion matrix

9 cases were classified right.

4 existent ‘ Distinctions ‘ of which 1 classified as ‘ Merit ‘ and 3 as ‘ Pass ‘ .

12 existent ‘ Merit ‘ of which7 right classified as ‘ Merit ‘ , 4 classified as ‘ Pass ‘ and 1classified as ‘ Distinction ‘ .

From the 8 existent ‘ Pass ‘ , 5 of them are classified as ‘ Merit ‘ , 1 classified as ‘ Distinction ‘ and2 right classified as ‘ Pass. ‘

TREE VISUALIZE

Prediction

Answer 1. To reply the inquiries we are traveling to see our conjuctiveRule theoretical account.

=== Classifier theoretical account (full preparation set) ===

Single conjunctive regulation scholar:

— — — — — — — — — — — — — — — —

= & gt ; Degree_Classification = Merit

Class distributions:

Covered by the regulation:

Distinction Merit Pass

0. 1875 0. 5 0. 3125

So here we can see that the degree categorization that a pupil will likely acquire is a virtue.

On the other theoretical account (determination tabular array)

==========

Gender Degree_Classification

==========

M Merit

F Pass

==========

The regulation is:

So for a peculiar pupil who is male, harmonizing to the regulations he will likely acquire a degree categorization = virtue. And for a female grade categorization = base on balls.

Answer 2.

In order to reply to the inquiry we have to follow some specific stairss on maori hen. First of all we have to make a new arff file with the degree categorization informations of the pupil we want to happen (26 old ages old individual male pupils with a BSc) . We will go forth empty the grade categorization (?) information in every row on the arff.

We have two categories of ages.

A pupil over 26 old ages old belongs to two categories: 24-27 and 28-35.

Our. arff file will be:

Then we are traveling to run a trial set

Now we are traveling to see our file.

Click on

Save as prediction. arff

Our prediction. arff file:

In our new file, a new property has been created: predicted grade. The value of the predicted grade is virtue. So the mature pupil with BSc will likely acquire a virtue.

Answer 3

We have to look on J48 theoretical account.

=== Classifier theoretical account (full preparation set) ===

J48 pruned tree

— — — — — — — — — —

Gender = F: Pass (6. 0/2. 0)

Gender = M: Merit (18. 0/7. 0)

We can detect that a pupil with base on balls is a female. So the reply is that the pupil has to be a female in order to acquire a base on balls.

Decision

This assignment had as is object to do clear to the pupils what information excavation is, inside the methodological analysis was used, inside the weka plan. We can see now why preprocessing is needed in order to clean our informations, replace losing values or, discretize our numeral properties. We besides can see the differences between categorization and anticipation. We can utilize categorization, and weka classifiers to foretell something. In our instance we predicted what categorization grade would possible take a pupil with specific features. Following the information excavation and categorization procedure, we have seen with our eyes, what information excavation is.