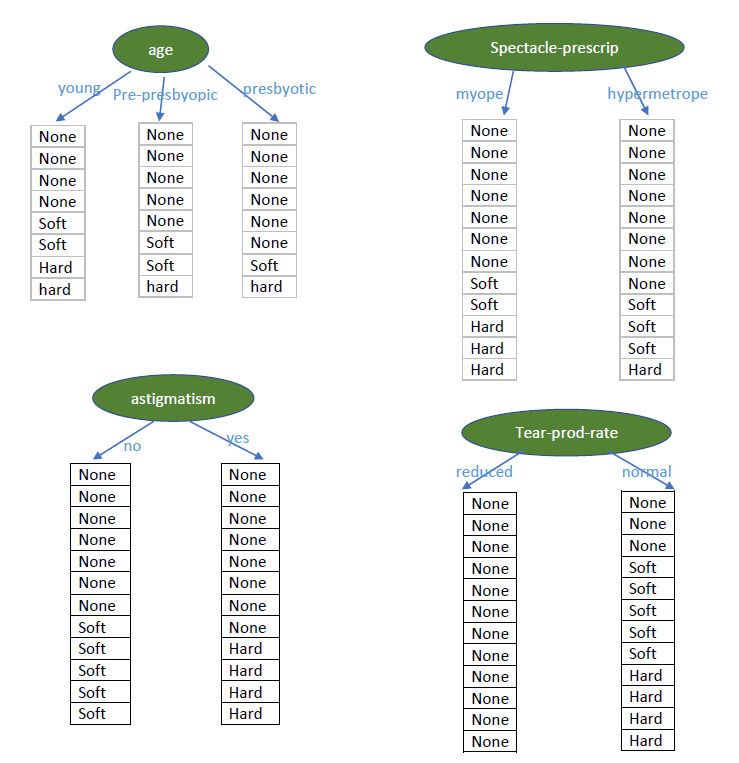
|  |  |  |  |
| --- | --- | --- | --- |
| SENG 474 / CSC 578D Data Mining – Assignment 1 |  |  |  |

Problem 1:

The visual summary of the Contact Lenses data



Contact Lenses Stat – First Tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | | Frequency | | | |
| Attribute item | Category | None | Soft | Hard | subtotal |
| Age | young | 4 | 2 | 2 | 8 |
|  | pre-presbyopic | 5 | 2 | 1 | 8 |
|  | presbyopic | 6 | 1 | 1 | 8 |
| Spectacle\_prescription | myope | 7 | 2 | 3 | 12 |
|  | hypermetrope | 8 | 3 | 1 | 12 |
| Astigmatism | yes | 8 | 0 | 4 | 12 |
|  | no | 7 | 5 | 0 | 12 |
| Tear-prod-rate | normal | 3 | 5 | 4 | 12 |
|  | reduced | 12 | 0 | 0 | 12 |
| Total |  | 60 | 20 | 16 |  |

Entropy Calculation for the First Tree

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Attribute | | Entropy | | | |  |
|  |  | None | Soft | Hard | Total | Grand total |
| Age | young | 0.5 | 0.5 | 0.5 | 1.5 | 1.287 |
|  | pre-presbyopic | 0.424 | 0.5 | 0.375 | 1.299 |
|  | presbyopic | 0.311 | 0.375 | 0.375 | 1.061 |
| Spectacle\_prescription | myope | 0.454 | 0.431 | 0.5 | 1.384 | 1.286577 |
|  | hypermetrope | 0.389975 | 0.5 | 0.298747 | 1.188722 |
| Astigmatism | yes | 0.389975 | 0 | 0.528321 | 0.918296 | 0.949082 |
|  | no | 0.453604 | 0.526264 | 0 | 0.979869 |
| Tear-prod-rate | normal | 0.5 | 0.526264 | 0.528321 | 1.554585 | 0.777293 |
|  | reduced | 0 | 0 | 0 | 0 |

Calculation details:

Smallest

Age – young

Info(4, 2, 2) = entropy(4/8, 2/8, 2/8) = -(4/8)\* (4/8)-(2/8)\* (2/8) )-(2/8)\* (2/8) = 1.5

Age – pre-presbyopic

Info(5, 2, 1) = entropy(5/8, 2/8, 1/8) = -(5/8)\* (5/8)-(2/8)\*(2/8)-(1/8)\*(1/8)= 1.299

Age – presbyopic

Info(6,1,1) = entropy(6/8, 1/8, 1/8) = -(6/8)\* (6/8)-(1/8)\*(1/8)- )-(1/8)\*(1/8)= 1.061

Total entropy for Age = (8/24)\*1.5+ (8/24)\*1.299+(8/24)\*1.061 = 1.286691

Spectacle-Prescription – myope

Info(7, 2, 3) = entropy(7/12, 2/12, 3/12) = -(7/12)\*(7/12)-(2/12)\*(2/12)-(3/12)\*(3/12)= 0.692216

Spectacle-Prescription – hypermetrope

Info(8, 3, 1) = entropy(8/12, 3/12, 1/12) = -(8/12)\*(8/12)-(3/12)\*(3/12)-(1/12)\*(1/12)= 0.594361

Total entropy for Spectacle-Prescription = 0.692216+0.594361 = 1.286577

Astigmatism – yes

Info(8, 0 , 4) = entropy(8/12, 0/12, 4/12) = -(8/12)\*(8/12)-0\*(0)-(4/12)\*(4/12)= 0.459148

Astigmatism – no

Info(7, 5, 0) = entropy(7/12, 5/12, 0/12) = -(7/12)\*(7/12)-(5/12)\*(5/12)-0\*(0)= 0.489934

Total entropy for Astigmatism = 0.459148 + 0.489934 = 0.949082

Tear-prod-rate - normal

info(3, 5, 4) = entropy(3/12, 5/12, 4/12) = -(3/12)\*(3/12)-(5/12)(5/12)-(4/12)\*(4/12)= 0.777293

info(0, 0, 0) = entropy(0/12, 0/12, 0/12) = 0

Total entropy for Tear-prob-rate = 0.777293

Contact Lenses Stat – First Tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Attribute Second Tree | | |  |  |
|  |  | None | Soft | Hard | Subtotal |
| Age | young | 0 | 2 | 2 | 4 |
|  | pre-presbyopic | 1 | 2 | 1 | 4 |
|  | presbyopic | 2 | 1 | 1 | 4 |
| Spectacle\_prescription | myope | 1 | 2 | 3 | 6 |
|  | hypermetrope | 2 | 3 | 1 | 6 |
| Astigmatism | yes | 2 | 0 | 4 | 6 |
|  | no | 1 | 5 | 0 | 6 |

Entropy Calculation for the Child node of First Tree:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Attributes | | Frequency | | | |  | |
|  |  | None | Soft | Hard | total | Grand Total | |
| Age | young |  | 0 | 0.5 | 0.5 | 1 |  |
|  | pre-presbyopic | | 0.5 | 0.5 | 0.5 | 1.5 |  |
|  | presbyopic | | 0.5 | 0.5 | 0.5 | 1.5 | 1.333333 |
| Spectacle\_prescription | myope |  | 0.430827 | 0.528321 | 0.5 | 1.459148 |  |
|  | hypermetrope | | 0.528321 | 0.5 | 0.430827 | 1.459148 | 1.459148 |
| Astigmatism | yes |  | 0.528321 | 0 | 0.389975 | 0.918296 |  |
|  | no |  | 0.430827 | 0.219195 | 0 | 0.650022 | 0.784159 |

Smallest

The smallest entropy value is 0.784159, Astigmatism = No. The first decision tree node looks as below.

Normal

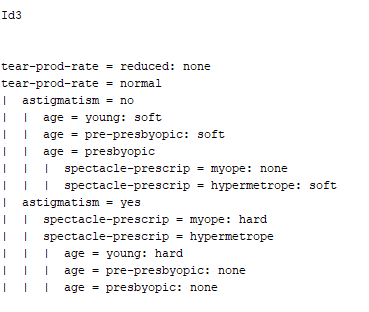
Reduced

none

yes

no

Weka result:



Problem 2:

***Rule 1: If the outlook = overcast, play = yes***

t: total number of instances covered by rule –

p: positive (correct) examples of the class covered by rule

Step 1: Find the highest p/t rate for the first rule.

Overcase: p/t = 4/4 = 1 (highest among all attributes)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | | Frequency | | | Rate |
|  |  | play - yes | play - no | subtotal | p/t |
| outlook | sunny | 2 | 3 | 5 | 0.4 |
|  | overcast | 4 | 0 | 4 | 1 |
|  | rainy | 3 | 2 | 5 | 0.6 |
| temperature | cool | 3 | 1 | 4 | 0.75 |
|  | mild | 4 | 2 | 6 | 0.666667 |
|  | hot | 2 | 2 | 4 | 0.5 |
| humidity | normal | 6 | 1 | 7 | 0.857143 |
|  | high | 3 | 4 | 7 | 0.428571 |
| windy | TRUE | 3 | 3 | 6 | 0.5 |
|  | FALSE | 6 | 2 | 8 | 0.75 |

***Rule 2: If humidity = normal and windy = false, play = yes***

Step 1: Choose the second highest p/t rate, humidity = normal and filter other instances:

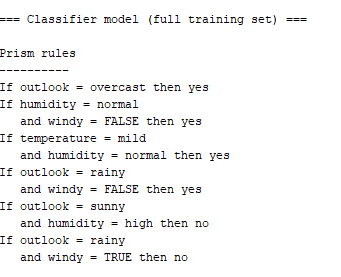
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| if humidity = normal | | |  |  |  |
| 4 | rainy | cool | normal | FALSE | yes |
| 6 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | cool | normal | FALSE | yes |
| 9 | rainy | mild | normal | FALSE | yes |
| 10 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | hot | normal | FALSE | yes |
| 5 | rainy | cool | normal | TRUE | no |

Note: Cover rate = total play=yes frequency / total play=yes in each attribute (outlook, temperature, windy)

Cover rate ( ‘windy’ = false) = 4/6 = 0.666667

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | play - yes | play - no | subtotal | p/t | cover rate |
| outlook | sunny | 2 | 0 | 2 | 1 | 0.333333 |
|  | overcast | 2 | 0 | 2 | 1 | 0.333333 |
|  | rainy | 2 | 1 | 3 | 0.666667 | 0.333333 |
| temperature | cool | 3 | 1 | 4 | 0.75 | 0.5 |
|  | mild | 2 | 0 | 2 | 1 | 0.333333 |
|  | hot | 1 | 0 | 1 | 1 | 0.166667 |
| windy | TRUE | 2 | 1 | 3 | 0.666667 | 0.333333 |
|  | FALSE | 4 | 0 | 4 | 1 | 0.666667 |
|  |  |  |  |  |  |  |
| total 'yes' |  | 6 |  |  |  |  |
| total 'no' |  | 1 |  |  |  |  |

Weka results:



Problem 3:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Attriutes | | | | | |  |
| Class | Age | young | | pre-presbyopic | | presbyopic |  |  |
| Total Instances | Frequency | Prob | Frequency | Prob | Frequency | Prob | Total % |
| None | 15 | 4 | 0.28 | 5 | 0.33 | 6 | 0.39 | 1.00 |
| Soft | 5 | 2 | 0.38 | 2 | 0.38 | 1 | 0.25 | 1.00 |
| Hard | 4 | 2 | 0.43 | 1 | 0.29 | 1 | 0.29 | 1.00 |
| Class | spectacle\_prescription | myope | | hypermetrope | |  |  |  |
| Total Instances | Frequency | Prob | Frequency | Prob | Total % |  |  |
| None | 15 | 7 | 0.47 | 8 | 0.53 | 1.00 |  |  |
| Soft | 5 | 3 | 0.57 | 2 | 0.43 | 1.00 |  |  |
| Hard | 4 | 3 | 0.67 | 1 | 0.33 | 1.00 |  |  |
| Class | astigmatism | yes | | no | |  |
| Total Instances | Frequency | Prob | Frequency | Prob | Total % |
| None | 15 | 8 | 0.53 | 7 | 0.47 | 1.00 |
| Soft | 5 | 0 | 0.14 | 5 | 0.86 | 1.00 |
| Hard | 4 | 4 | 0.83 | 0 | 0.17 | 1.00 |
| Class | tear-prob-rate | normal | | reduced | |  |
| Total Instances | Frequency | Prob | Frequency | Prob | Total % |
| None | 15 | 3 | 0.24 | 12 | 0.76 | 1.00 |
| Soft | 5 | 5 | 0.86 | 0 | 0.14 | 1.00 |
| Hard | 4 | 4 | 0.83 | 0 | 0.17 | 1.00 |

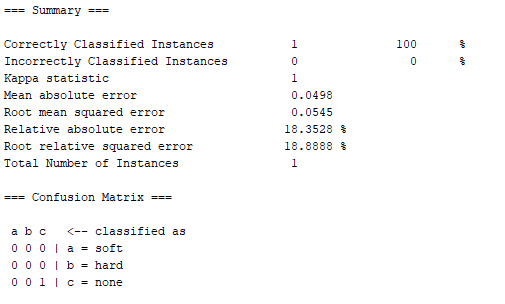
Probability for each instance is calculated as:

Note: is not included in the calculation because it is a scalar and it does not affect the final classification.

RESULT:

Because is the highest value, ( is classified as ‘None’.

Weka result:



APPENDIX

Original Data:

Contac lenses

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Sectacle\_prescription | Astigmatism | Tear-prod-rate | Contact-lenses |
| young | hypermetrope | yes | normal | hard |
| young | myope | yes | normal | hard |
| pre-presbyopic | myope | yes | normal | hard |
| presbyopic | myope | yes | normal | hard |
| presbyopic | myope | no | normal | none |
| pre-presbyopic | hypermetrope | yes | normal | none |
| presbyopic | hypermetrope | yes | normal | none |
| young | hypermetrope | no | reduced | none |
| pre-presbyopic | hypermetrope | no | reduced | none |
| presbyopic | hypermetrope | no | reduced | none |
| young | myope | no | reduced | none |
| pre-presbyopic | myope | no | reduced | none |
| presbyopic | myope | no | reduced | none |
| young | hypermetrope | yes | reduced | none |
| pre-presbyopic | hypermetrope | yes | reduced | none |
| presbyopic | hypermetrope | yes | reduced | none |
| young | myope | yes | reduced | none |
| pre-presbyopic | myope | yes | reduced | none |
| presbyopic | myope | yes | reduced | none |
| young | hypermetrope | no | normal | soft |
| pre-presbyopic | hypermetrope | no | normal | soft |
| presbyopic | hypermetrope | no | normal | soft |
| young | myope | no | normal | soft |
| pre-presbyopic | myope | no | normal | soft |

Weather - Nominal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Outlook | Temperature | Humidity | Windy | Play |
| 0 | sunny | hot | high | FALSE | no |
| 1 | sunny | hot | high | TRUE | no |
| 2 | overcast | hot | high | FALSE | yes |
| 3 | rainy | mild | high | FALSE | yes |
| 4 | rainy | cool | normal | FALSE | yes |
| 5 | rainy | cool | normal | TRUE | no |
| 6 | overcast | cool | normal | TRUE | yes |
| 7 | sunny | mild | high | FALSE | no |
| 8 | sunny | cool | normal | FALSE | yes |
| 9 | rainy | mild | normal | FALSE | yes |
| 10 | sunny | mild | normal | TRUE | yes |
| 11 | overcast | mild | high | TRUE | yes |
| 12 | overcast | hot | normal | FALSE | yes |
| 13 | rainy | mild | high | TRUE | no​ |