Lab Number: 4

Title

Use ID3 and J48 algorithms to design decision trees using a primary dataset. Assume the cross validation. Also visualize the tree using the J48 algorithm

Objective

To design decision tree models using the ID3 and J48 (C4.5) algorithms on a primary dataset, evaluate the models using cross-validation, and visualize the final decision tree generated by the J48 algorithm in WEKA.

IDE/Tools Used

Weka 3.8.6

Theory

Primary Dataset: A primary dataset is raw operational data collected directly from the source system. It is unprocessed and may include various categorical or numeric attributes. For this lab, a categorical primary dataset is required since ID3 supports only nominal attributes.

Decision Tree Algorithms: A decision tree is a classification algorithm that splits data into branches based on attribute values, forming a tree-like structure. The goal is to classify the target variable using the simplest possible set of rules.

ID3 Algorithm: ID3 is an early decision tree learning algorithm introduced by Quinlan. It works only with nominal (categorical) attributes. Key features of ID3 algorithm include:

- Uses Information Gain to select the best attribute
- Does not support pruning
- Cannot process numeric attributes directly
- Often produces compact but overfitted trees

J48 Algorithm (C4.5 Implementation): J48 is WEKA's implementation of the C4.5 algorithm and is an improved version of ID3. The key features of J48 algorithms include:

- Supports both nominal and numeric data
- Uses Gain Ratio
- Performs tree pruning to reduce overfitting
- Can handle missing values
- Produces a more generalized model

Cross-Validation: Cross-validation is a model evaluation technique in which the dataset is divided into multiple subsets (or "folds"). The model is trained on some of these subsets and tested on the

remaining ones. This process is repeated several times, each time using a different subset for testing. The average performance across all rounds represents the model's overall accuracy. Cross-validation helps ensure that the evaluation is reliable, less biased, and not dependent on a single train-test split. It can use any number of folds depending on the dataset and requirements.

Implementation

The steps performed in WEKA Explorer are:

1. Conversion of dataset from csv to arff format

Firstly, open the arff viewer from tools and then open the csv file in the viewer. Visualize the dataset and save it as arff format.

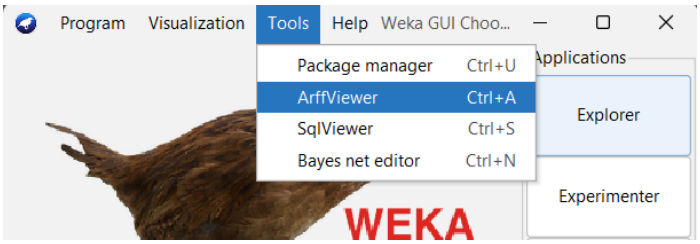


Figure 1: Opening the Arff Viewer

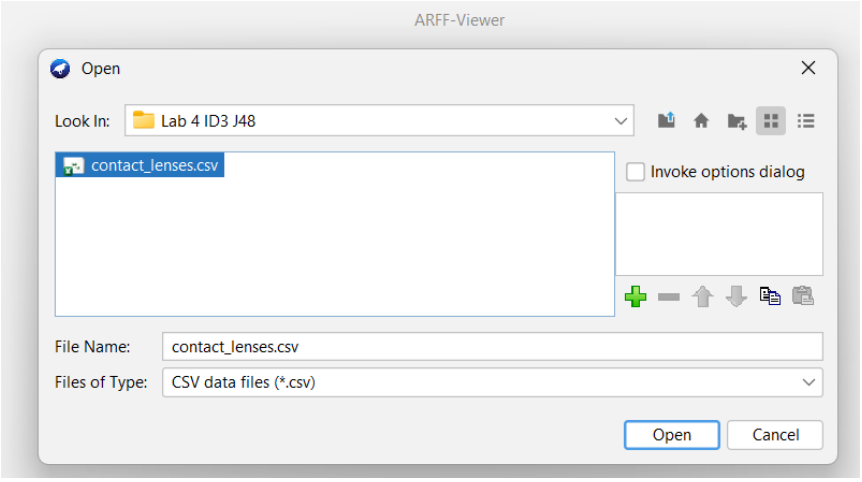
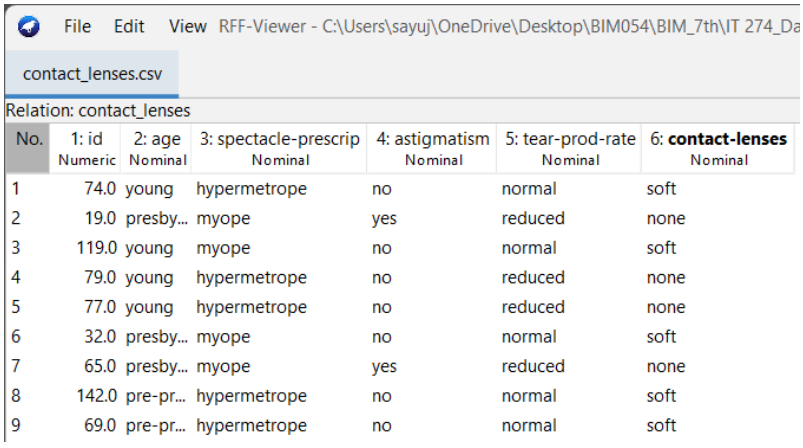


Figure 2: Selecting the csv file



No.	1: id Numeric	2: age Nominal	3: spectacle-prescrip Nominal	4: astigmatism Nominal	5: tear-prod-rate Nominal	6: **contact-lenses** Nominal
1	74.0	young	hypermetrope	no	normal	soft
2	19.0	presby...	myope	yes	reduced	none
3	119.0	young	myope	no	normal	soft
4	79.0	young	hypermetrope	no	reduced	none
5	77.0	young	hypermetrope	no	reduced	none
6	32.0	presby...	myope	no	normal	soft
7	65.0	presby...	myope	yes	reduced	none
8	142.0	pre-pr...	hypermetrope	no	normal	soft
9	69.0	pre-pr...	hypermetrope	no	normal	soft

Figure 3: Visualization of the dataset

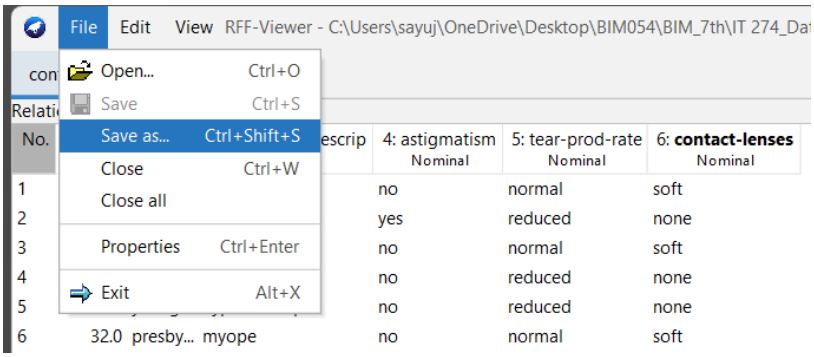


Figure 4: Option to save as arff

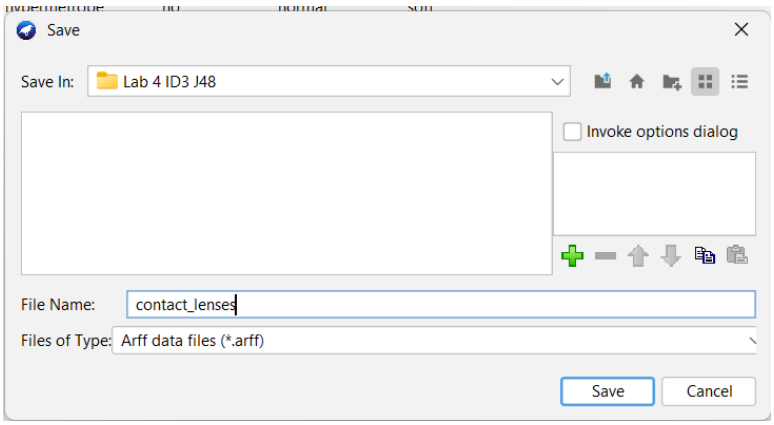


Figure 5: Saving as arff format

2. Loading the dataset

A dataset containing information about patients and whether they should be prescribed contact lenses was used. It includes four categorical predictors: age group, spectacle prescription, astigmatism, and tear production rate. And the target variable indicating the recommended lens type (none, soft, or hard). This data is firstly loaded into the WEKA and is visualized by clicking on edit button. The id column is removed as it has no any relation to the database.

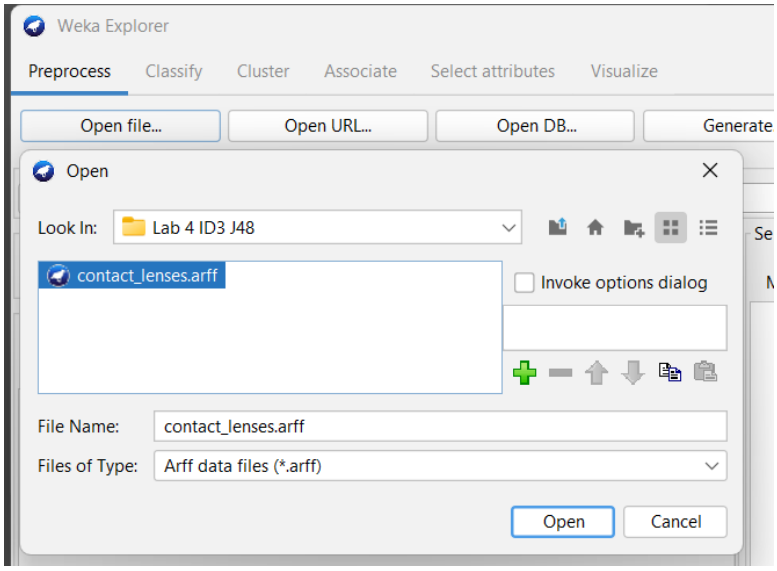


Figure 6: Opening the contact-lenses dataset

Viewer

Relation: contact_lenses

No.	1: id Numeric	2: age Nominal	3: spectacle-prescrip Nominal	4: astigmatism Nominal	5: tear-prod-rate Nominal	6: **contact-lenses** Nominal
1	74.0	young	hypermetrope	no	normal	soft
2	19.0	presby...	myope	yes	reduced	none
3	119.0	young	myope	no	normal	soft
4	79.0	young	hypermetrope	no	reduced	none
5	77.0	young	hypermetrope	no	reduced	none
6	32.0	presby...	myope	no	normal	soft
7	65.0	presby...	myope	yes	reduced	none
8	142.0	pre-pr...	hypermetrope	no	normal	soft
9	69.0	pre-pr...	hypermetrope	no	normal	soft
10	83.0	presby...	myope	no	normal	soft
11	111.0	young	myope	yes	normal	hard
12	13.0	pre-pr...	hypermetrope	no	reduced	none
13	37.0	pre-pr...	hypermetrope	no	reduced	none
14	10.0	pre-pr...	myope	no	normal	soft
15	20.0	presby...	myope	yes	normal	hard
16	57.0	pre-pr...	myope	yes	reduced	none
17	105.0	presby...	myope	yes	normal	hard
18	70.0	young	hypermetrope	no	normal	soft
19	56.0	presby...	hypermetrope	no	normal	soft
20	133.0	presby...	myope	yes	normal	hard
21	30.0	young	myope	yes	normal	hard
22	128.0	young	myope	yes	reduced	none
23	27.0	young	hypermetrope	no	normal	soft
24	129.0	pre-pr...	myope	no	reduced	none

Figure 7: Visualization of the dataset with id

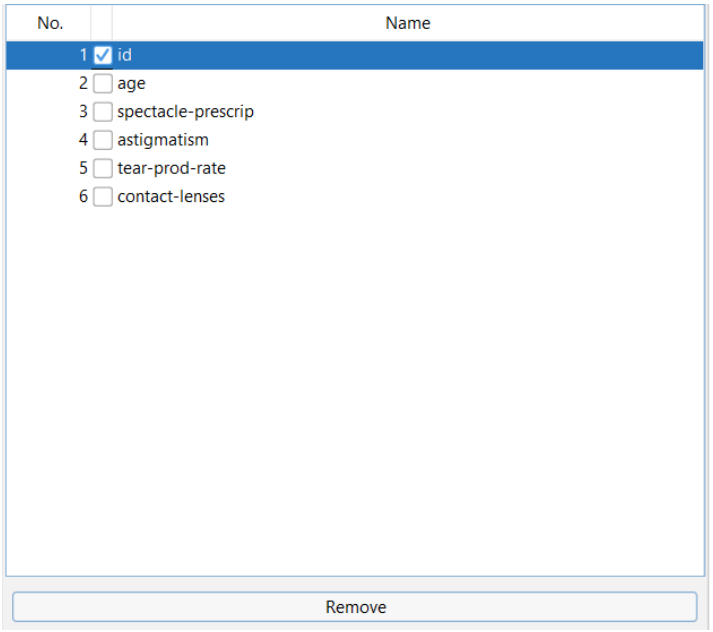


Figure 8: Removing the id from the dataset

Viewer

Relation: contact_lenses-weka.filters.unsupervised.attribute.Remove-R1

No.	1: age Nominal	2: spectacle-prescrip Nominal	3: astigmatism Nominal	4: tear-prod-rate Nominal	5: **contact-lenses** Nominal
1	young	hypermetrope	no	normal	soft
2	presby...	myope	yes	reduced	none
3	young	myope	no	normal	soft
4	young	hypermetrope	no	reduced	none
5	young	hypermetrope	no	reduced	none
6	presby...	myope	no	normal	soft
7	presby...	myope	yes	reduced	none
8	pre-pr...	hypermetrope	no	normal	soft
9	pre-pr...	hypermetrope	no	normal	soft
10	presby...	myope	no	normal	soft
11	young	myope	yes	normal	hard
12	pre-pr...	hypermetrope	no	reduced	none
13	pre-pr...	hypermetrope	no	reduced	none
14	pre-pr...	myope	no	normal	soft
15	presby...	myope	yes	normal	hard
16	pre-pr...	myope	yes	reduced	none
17	presby...	myope	yes	normal	hard
18	young	hypermetrope	no	normal	soft
19	presby...	hypermetrope	no	normal	soft
20	presby...	myope	yes	normal	hard
21	young	myope	yes	normal	hard
22	young	myope	yes	reduced	none
23	young	hypermetrope	no	normal	soft
24	pre-pr...	myope	no	reduced	none

Figure 9: Visualizing the dataset after removing the id

3. Apply ID3 Algorithm

Go to classify and choose ID3 option from the trees section. Set the Cross-validation to 10 folds and click on start. After that, on the right side of the explorer, inside the Classifier output shows our result.

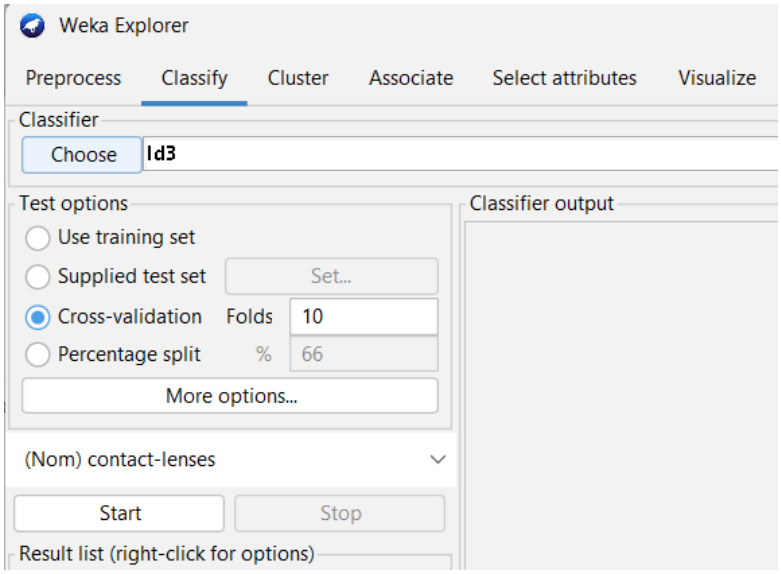


Figure 10: Setting the ID3 and Cross-validation settings

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.Id3
Relation:     contact_lenses-weka.filters.unsupervised.attribute.Remove-R1
Instances:    150
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:    10-fold cross-validation

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

Id3


tear-prod-rate = normal
|  astigmatism = no
|  |  age = young: soft
|  |  age = presbyopic
|  |  |  spectacle-prescrip = hypermetrope: soft
|  |  |  spectacle-prescrip = myope: soft
|  |  age = pre-presbyopic: soft
|  astigmatism = yes
|  |  spectacle-prescrip = hypermetrope
|  |  |  age = young: hard
|  |  |  age = presbyopic: none
|  |  |  age = pre-presbyopic: none
|  |  spectacle-prescrip = myope: hard
tear-prod-rate = reduced: none

Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         149               99.3333 %
Incorrectly Classified Instances         1                0.6667 %
Kappa statistic                          0.9887
Mean absolute error                      0.0084
Root mean squared error                  0.0719
Relative absolute error                  2.1425 %
Root relative squared error             16.216  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.009    0.972      1.000    0.986      0.982    0.995     0.968     soft
                 0.988    0.000    1.000      0.988    0.994      0.987    0.994     0.995     none
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     hard
Weighted Avg.    0.993    0.002    0.994      0.993    0.993      0.988    0.995     0.990

=== Confusion Matrix ===

  a  b  c   <-- classified as
 35  0  0 |  a = soft
  1 83  0 |  b = none
  0  0 31 |  c = hard
```

Figure 11: Output for ID3 algorithm

4. Apply the J48 Algorithm

Go to classify and choose J48 option from the trees section. Set the Cross-validation to 10 folds and click on start. After that, on the right side of the explorer, inside the Classifier output shows our result.

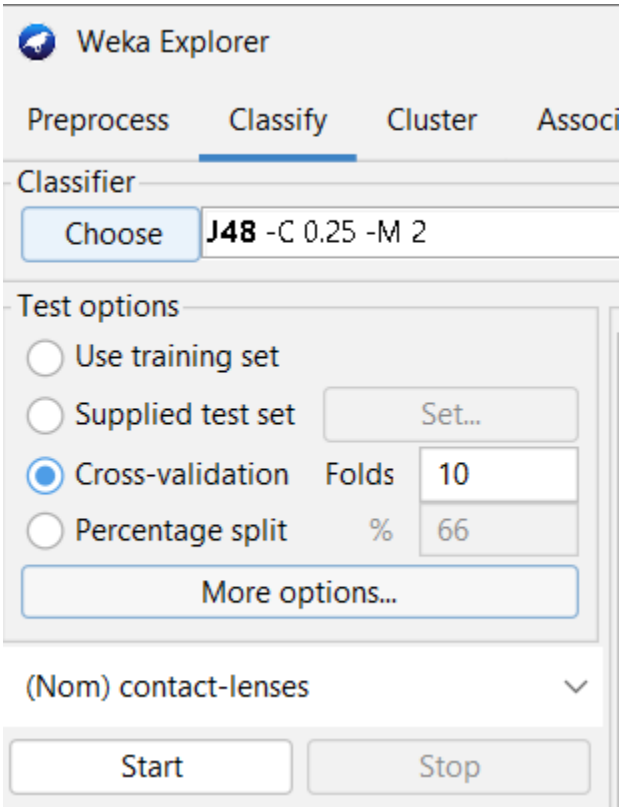


Figure 12: Setting the J48 and Cross-validation settings

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     contact_lenses-weka.filters.unsupervised.attribute.Remove-R1
Instances:    150
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:    10-fold cross-validation

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

J48 pruned tree
------------------

tear-prod-rate = normal
|   astigmatism = no: soft (36.0/1.0)
|   astigmatism = yes
|   |   spectacle-prescrip = hypermetrope
|   |   |   age = young: hard (6.0)
|   |   |   age = presbyopic: none (9.0)
|   |   |   age = pre-presbyopic: none (7.0)
|   |   spectacle-prescrip = myope: hard (25.0)
tear-prod-rate = reduced: none (67.0)

Number of Leaves  :     6

Size of the tree :      10


Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         149               99.3333 %
Incorrectly Classified Instances         1                0.6667 %
Kappa statistic                          0.9887
Mean absolute error                      0.0087
Root mean squared error                  0.0676
Relative absolute error                  2.204  %
Root relative squared error             15.254  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.009    0.972      1.000    0.986      0.982    0.992     0.943     soft
                 0.988    0.000    1.000      0.988    0.994      0.987    0.991     0.995     none
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     hard
Weighted Avg.    0.993    0.002    0.994      0.993    0.993      0.988    0.993     0.984

=== Confusion Matrix ===

  a  b  c   <-- classified as
 35  0  0 |  a = soft
  1 83  0 |  b = none
  0  0 31 |  c = hard
```

Figure 13: Output for J48 algorithm

5. Visualizing the J48 Tree

After successfully running the J48 algorithm, right click on the result list and click on the Visualize tree option to get the full graphical decision tree displayed.

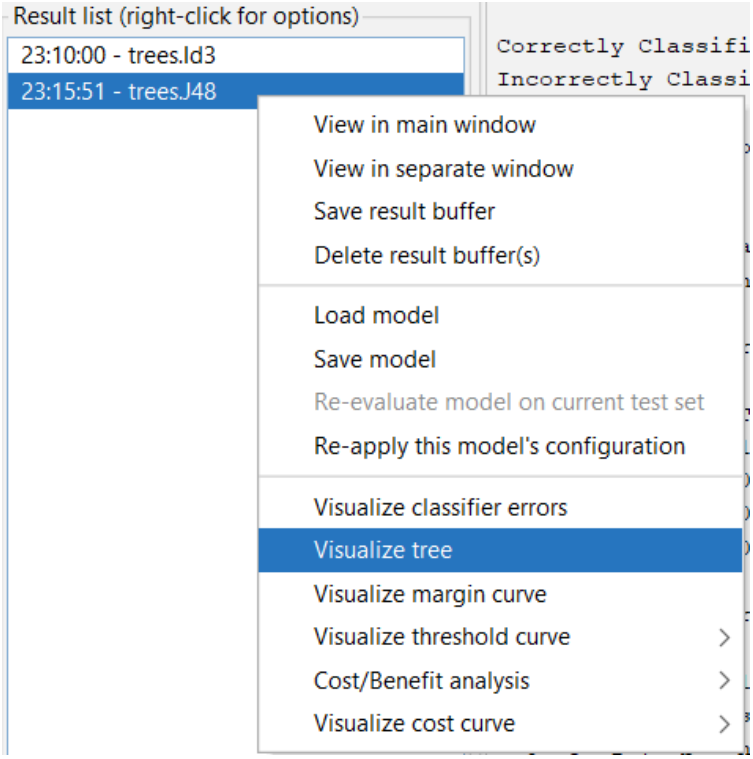


Figure 14: Option to visualize tree from the result list

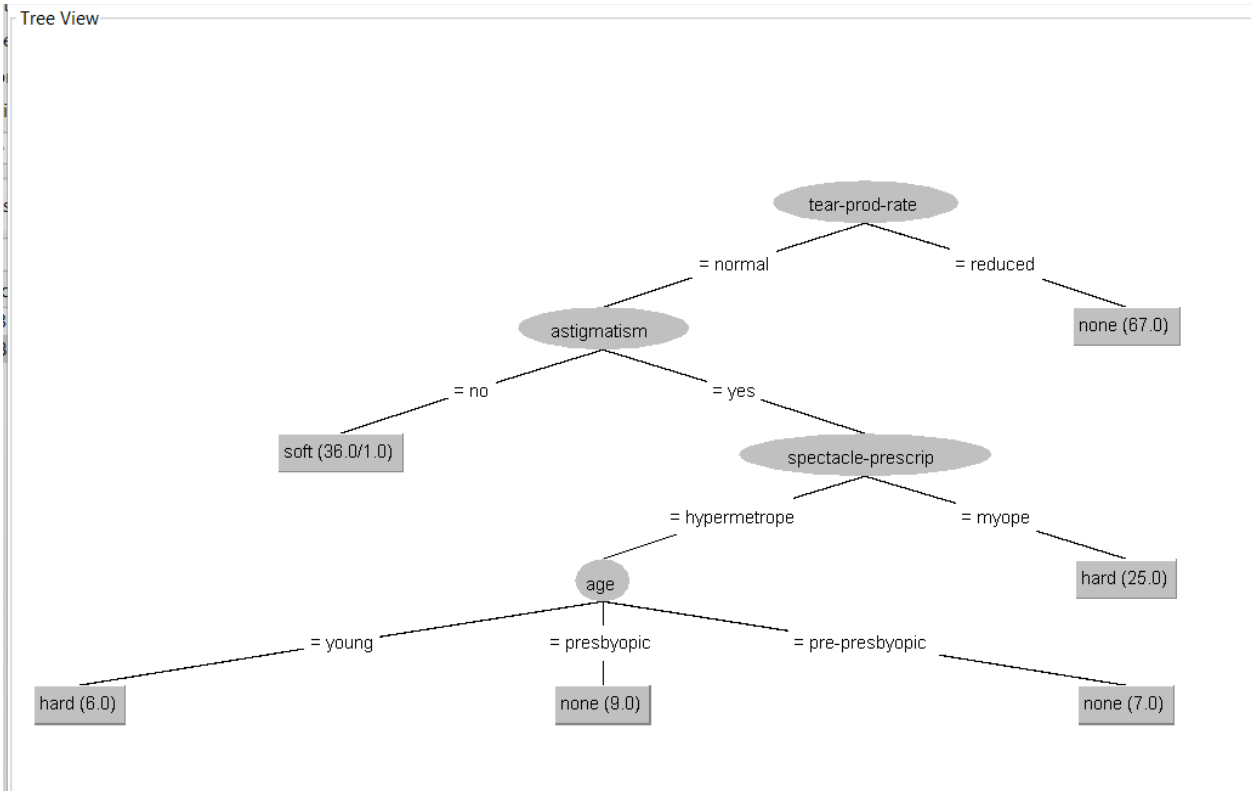


Figure 15: Full graphical decision tree of the result

Discussion

The performance evaluation of the ID3 and J48 decision tree algorithms showed that both models performed exceptionally well on the primary dataset. Using 10-fold cross-validation, each algorithm achieved 99.33% classification accuracy, with only one instance misclassified out of 150. The high kappa statistic (0.9887) further indicates strong agreement between predicted and actual class labels beyond random chance.

Per-class performance metrics also demonstrated strong predictive ability. For both models, the hard class achieved perfect precision, recall, and F-measure (1.000), meaning all hard instances were correctly identified with no false positives. The soft class showed a recall of 1.000 and F-measure of 0.986, although its precision was slightly lower (0.972) due to occasional misclassification. The none class also performed strongly, with near-perfect recall (0.988) and perfect precision (1.000), indicating very few false negatives. The weighted ROC areas—0.995 for ID3 and 0.993 for J48—suggest excellent overall class separability.

Although both models achieved identical accuracy, structural differences help explain their behavior. ID3 generated a more complex, fully expanded tree, as it does not implement pruning and relies solely on information gain, which can favor attributes with many distinct values. In contrast, J48 produced a more concise and generalized tree, due to pruning and the use of gain ratio to reduce attribute selection bias. This leads to better interpretability and potentially stronger generalization on unseen data, even though accuracy differences were negligible in this dataset.

Conclusion

Both ID3 and J48 demonstrated excellent performance on the primary dataset, each achieving 99.33% accuracy with strong precision, recall, and F-measure scores across all classes. While both algorithms delivered highly reliable predictions, J48 holds a practical advantage due to its pruning mechanism and use of gain ratio, resulting in a simpler and more generalized decision tree compared to the more complex, unpruned ID3 model. Overall, the experiment confirms that decision tree classifiers are effective for this classification task, with J48 being the more suitable choice for real-world deployment and long-term prediction stability.