Lab Number: 5

Title
Use Naïve Bayes' Classifier on a primary dataset

Objective
To apply the Naïve Bayes classification algorithm on a primary dataset using WEKA, evaluate the model using cross-validation, and analyze its classification performance

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.

Naïve Bayes Algorithm: It is a probabilistic classification algorithm based on Bayes' Theorem with an assumption of conditional independence between attributes.

Bayes' Theorem:

$$P(C \mid X) = \frac{P(X \mid C)\, P(C)}{P(X)}$$

Where:
- P(C | X) = probability of class C given predictors X
- P(C) = prior probability of the class
- P(X|C) = likelihood
- P(X) = evidence

Naïve Bayes assumes that every attribute contributes independently to the likelihood, making the classifier:
- Computationally efficient
- Effective on small datasets
- Capable of handling both numeric and nominal attributes
- Robust even with noisy or incomplete data

Examples of where Naïve Bayes is widely used include spam filtering, medical diagnosis, sentiment analysis, and document classification.

Cross-Validation: Cross-validation is an evaluation technique where the dataset is split into multiple parts (folds). The model is repeatedly trained on a portion of the data and tested on the remaining part. The averaged accuracy across these rounds gives a reliable estimate of model performance.

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.

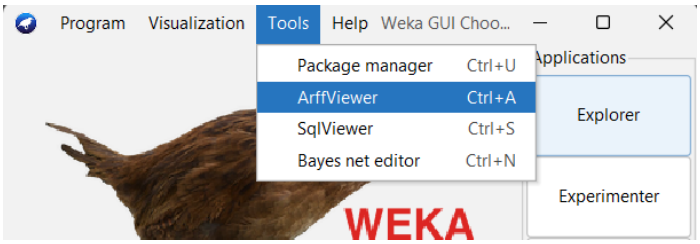


Figure 1: Opening the Arff Viewer

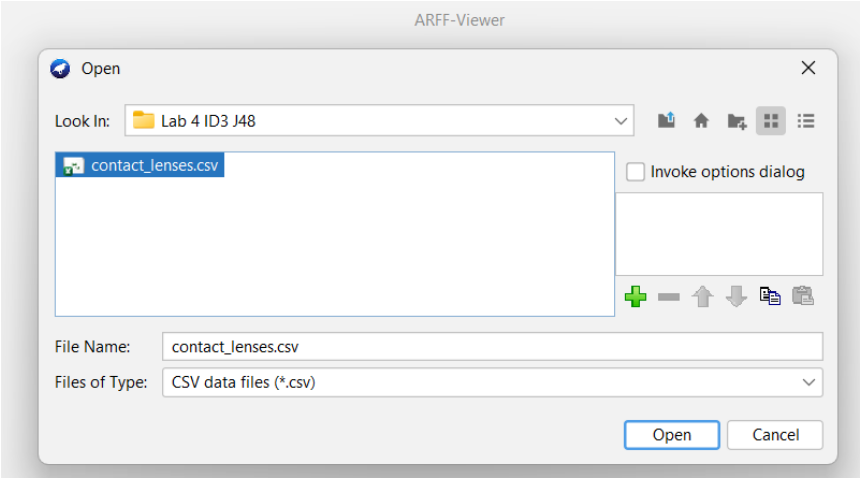
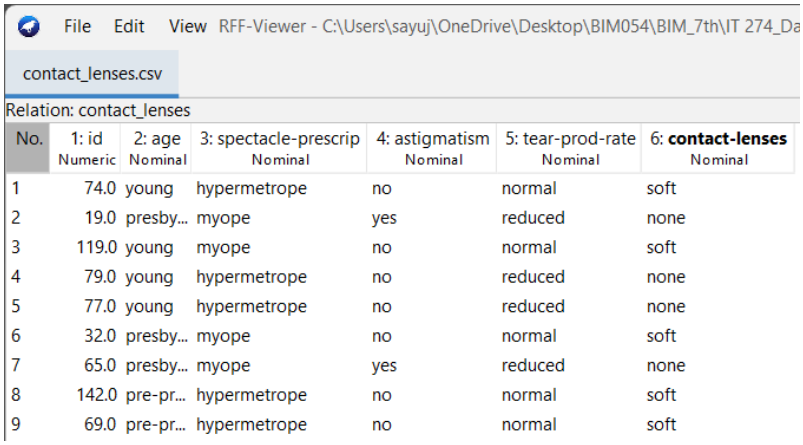


Figure 2: Selecting the csv file



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

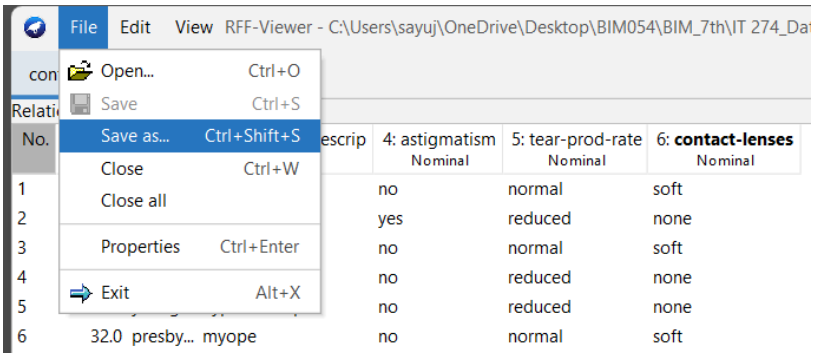


Figure 4: Option to save as arff

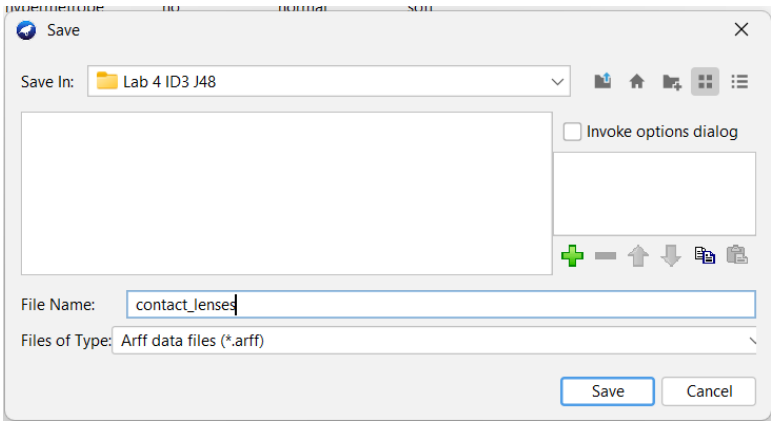


Figure 5: Saving as arff format

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.

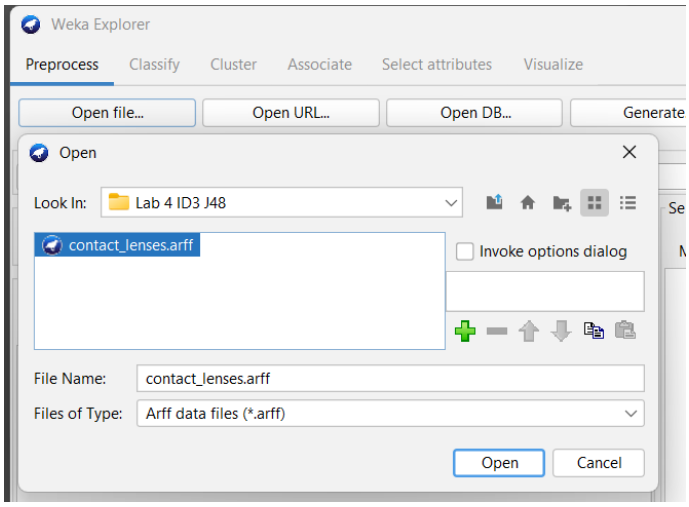


Figure 6: Opening the contact-lenses dataset

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

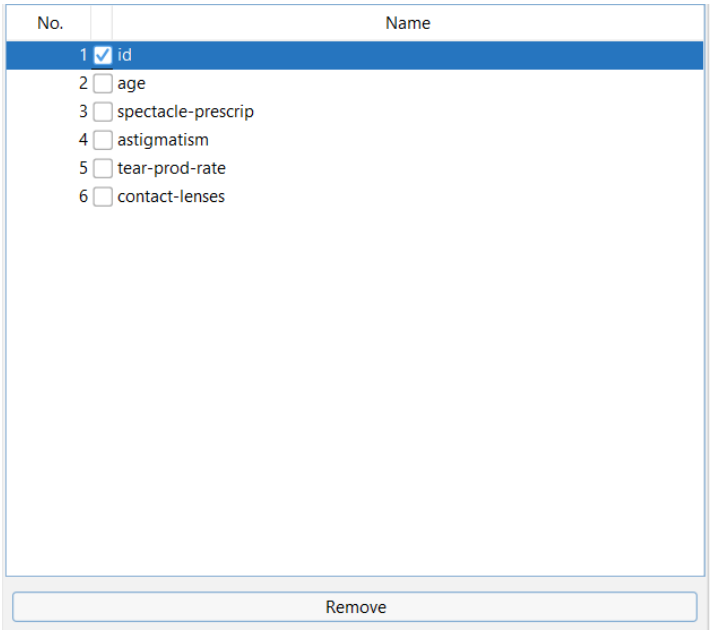
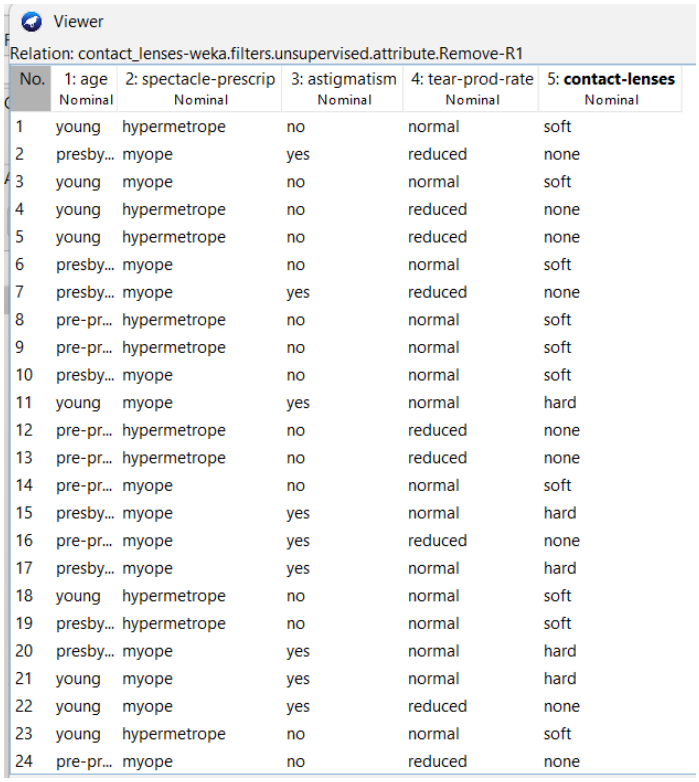


Figure 8: Removing the id from the dataset



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 Naïve Bayes Classifier

Go to Classify tab and Choose NaiveBayes from bayes tab. Select Cross-validation under Test Options to 10 and run the classifier by clicking start. WEKA will then generate the result in Classifier output section that will contain: Model summary, class probabilities, correct and incorrect classifications, confusion matrix, and detailed accuracy by class.

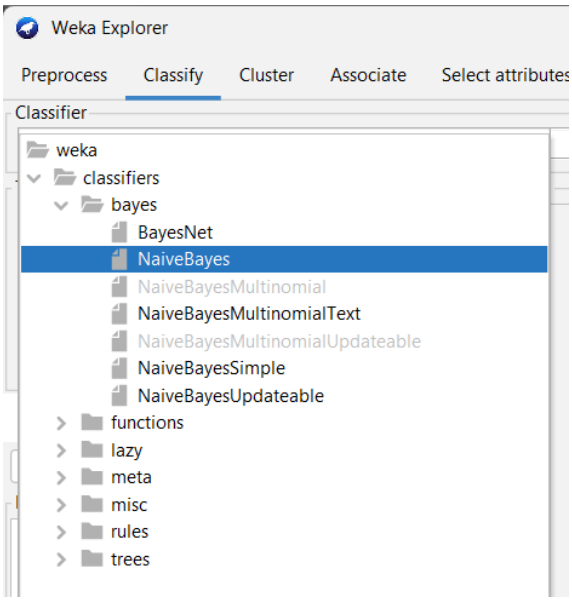


Figure 10: Choosing the Naive Bayes classifier option

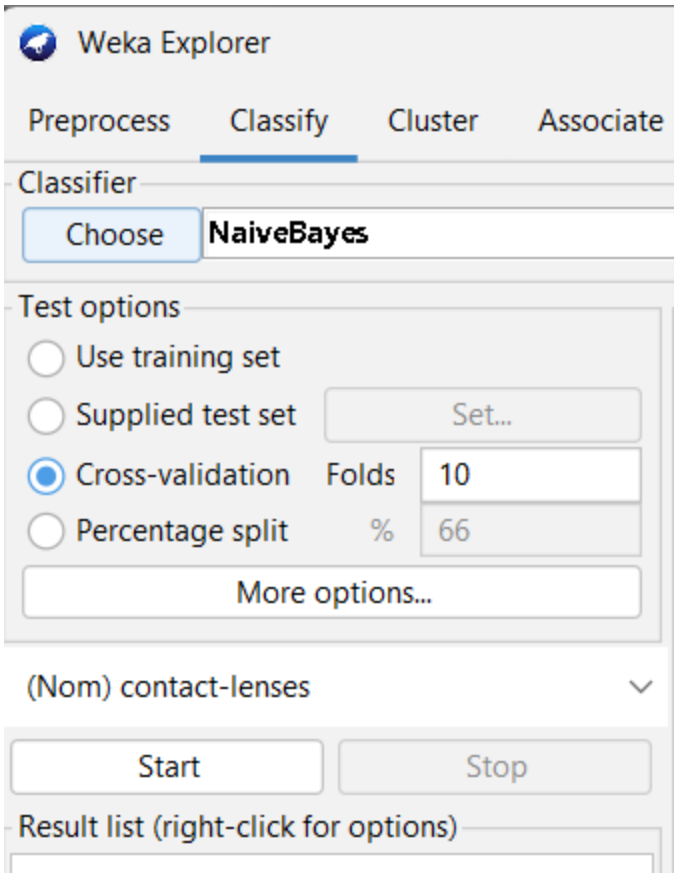


Figure 11: Setting cross-fold value

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
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) ===
Naive Bayes Classifier

                     Class
Attribute              soft   none   hard
                     (0.24) (0.56) (0.21)
=============================================
age
  young                12.0   23.0   14.0
  presbyopic           13.0   30.0   13.0
  pre-presbyopic       13.0   34.0    7.0
  [total]              38.0   87.0   34.0

spectacle-prescrip
  hypermetrope         22.0   47.0    7.0
  myope                15.0   39.0   26.0
  [total]              37.0   86.0   33.0

astigmatism
  no                   36.0   33.0    1.0
  yes                   1.0   53.0   32.0
  [total]              37.0   86.0   33.0

tear-prod-rate
  normal               36.0   18.0   32.0
  reduced               1.0   68.0    1.0
  [total]              37.0   86.0   33.0



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         140               93.3333 %
Incorrectly Classified Instances        10                6.6667 %
Kappa statistic                          0.8912
Mean absolute error                      0.1166
Root mean squared error                  0.1896
Relative absolute error                 29.5934 %
Root relative squared error             42.7684 %
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    0.997     0.987     soft
                 0.881    0.000    1.000      0.881    0.937      0.875    0.991     0.994     none
                 1.000    0.076    0.775      1.000    0.873      0.846    0.996     0.986     hard
Weighted Avg.    0.933    0.018    0.947      0.933    0.935      0.894    0.993     0.991

=== Confusion Matrix ===

  a  b  c   <-- classified as
 35  0  0 |  a = soft
  1 74  9 |  b = none
  0  0 31 |  c = hard
```

Figure 12: Output of the algorithm

Discussion

The Naïve Bayes classifier performed strongly on the primary dataset, as reflected in the evaluation metrics. The model achieved a weighted average precision of 0.947, recall of 0.933, and an F-measure of 0.935, indicating a high level of reliability in its predictions. The per-class results showed particularly strong performance for the soft class, with a perfect true positive rate (1.000) and an F-measure of 0.986. The none class achieved perfect precision (1.000), although its recall was slightly lower at 0.881, suggesting a few instances were misclassified into other categories. The hard class achieved perfect recall as well but showed a lower precision of 0.775, indicating some false positives where other classes were predicted as hard.

Despite these minor imbalances, the classifier demonstrated excellent overall discrimination ability, with a weighted ROC area of 0.993, very close to a perfect score. This indicates that the model is highly effective at separating the classes. The results confirm that Naïve Bayes, despite its simplicity and independence assumptions, can deliver very strong performance on categorical primary datasets.

Conclusion

The Naïve Bayes classifier was successfully applied to the primary dataset, and the evaluation using cross-validation showed that the model performed with high accuracy and reliability. The classifier demonstrated strong precision, recall, and overall F-measure, supported by an excellent ROC area. These results highlight Naïve Bayes as an effective and efficient method for classification tasks on small to medium-sized datasets. The experiment confirms the practical usefulness of Naïve Bayes and its suitability as a baseline model for supervised learning.