Advanced databases and data warehouses  
Data analysis with Bayesian network models

Project report

Abstract:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

**Contents:**

I/ Basic statistics of the dataset

II/ Bayesian Search Model

III/ Naïve Bayes Model

IV/ TAN Model

V/ Conclusion

The attributes are:

1) Alcohol

2) Malic acid

3) Ash

4) Alcalinity of ash

5) Magnesium

6) Total phenols

7) Flavanoids

8) Nonflavanoid phenols

9) Proanthocyanins

10)Color intensity

11)Hue

12)OD280/OD315 of diluted wines

13)Proline

All attributes are continuous, so discretization was necessary, I chose an unform widths discretization for all variables.

The goal here is to find the best model which is able to predict the group of the wine.

Concerning the model validation, I used the K-fold crossvalidation with 10 folds.

Here some basic statistics of this dataset:

Table

Description automatically generated

It’s important to see that there is no missing value and all of them are coherent.

In the following parts of this report, you will see 3 different models: Bayesian Search model, naive Bayes model and TAN model.

**I/ Bayesian Search Model:**

Diagram

Description automatically generated

1/ Diagnosis for each group

Group 1:

Graphical user interface, chart

Description automatically generated

Group 2:

Graphical user interface, application

Description automatically generated

Group 3:

Chart

Description automatically generated

We can see that in each group the strongest variable is not the same; Proline for group 1, color\_intensity for group 2 and finally Flavanoids for the group 3.

2/ Model Validation

Text

Description automatically generated

Confusion Matrix:

Table

Description automatically generated

ROC Curve:

Chart, line chart

Description automatically generated

We can already see that this first model is quite good with an accuracy of 96%.

II/ Naive Bayes Model:

Diagram, schematic

Description automatically generated

1/ Diagnosis for each group

Graphical user interface, chart

Description automatically generatedGroup 1:

**Graphical user interface, application

Description automatically generatedGroup 2:**

Graphical user interface, chart

Description automatically generated

**Group 3:**

2/ Model Validation

Confusion Matrix:

Table

Description automatically generated

Text

Description automatically generated

ROC Curve:

Chart, line chart

Description automatically generated

IV/ TAN Model

Diagram

Description automatically generated

1/ Diagnosis for each group

Group 1:

Graphical user interface

Description automatically generated

Group 2:

Graphical user interface, chart

Description automatically generated

Group 3:

Graphical user interface, chart

Description automatically generated

2/ Model Validation

Confusion Matrix:

Text

Description automatically generated

Table

Description automatically generated

ROC Curve:

Chart, line chart

Description automatically generated

**V/ Conclusion**

In conclusion we can say that each of the 3 models are good and accurate thanks to the number of records, netherless the TAN model is the best, with an accuracy of 96.63% and a ROC Curve almost perfect.