Joaquim JUSSEAU

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.

Joaquim JUSSEAU

Contents:

Joaquim JUSSEAU

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:

Joaquim JUSSEAU

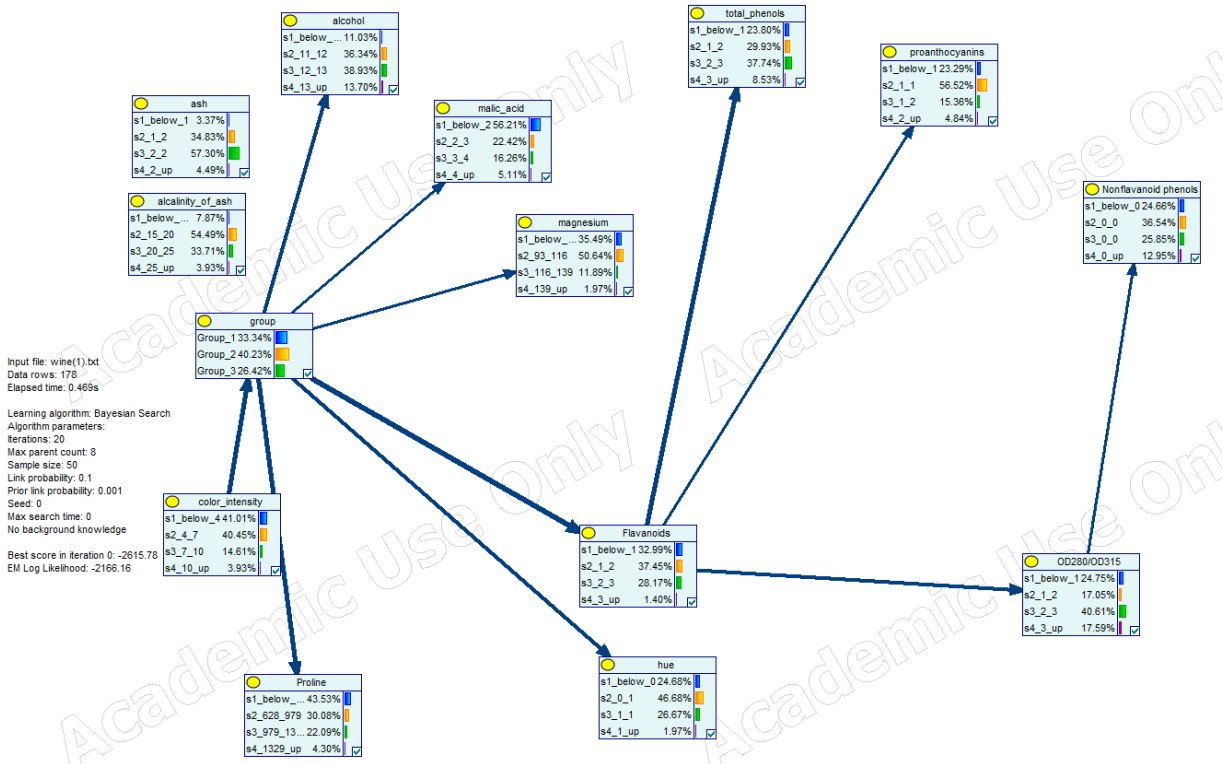
Basic statistics | Correlation matrix

	Mean	Variance	StdDev	Min	Max	Count
group	1.9382	0.600679	0.775035	1	3	178
alcohol	13.0006	0.659062	0.811827	11.03	14.83	178
malic_acid	2.33635	1.24802	1.11715	0.74	5.8	178
ash	2.36652	0.0752646	0.274344	1.36	3.23	178
alcalinity_of_ash	19.4949	11.1527	3.33956	10.6	30	178
magnesium	99.7416	203.989	14.2825	70	162	178
total_phenols	2.29511	0.39169	0.625851	0.98	3.88	178
Flavanoids	2.02927	0.997719	0.998859	0.34	5.08	178
Nonflavanoid phenols	0.361854	0.0154886	0.124453	0.13	0.66	178
proanthocyanins	1.5909	0.327595	0.572359	0.41	3.58	178
color_intensity	5.05809	5.37445	2.31829	1.28	13	178
hue	0.957449	0.052245	0.228572	0.48	1.71	178
OD280/OD315	2.61169	0.504086	0.70999	1.27	4	178
Proline	746.893	99166.7	314.907	278	1680	178

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.

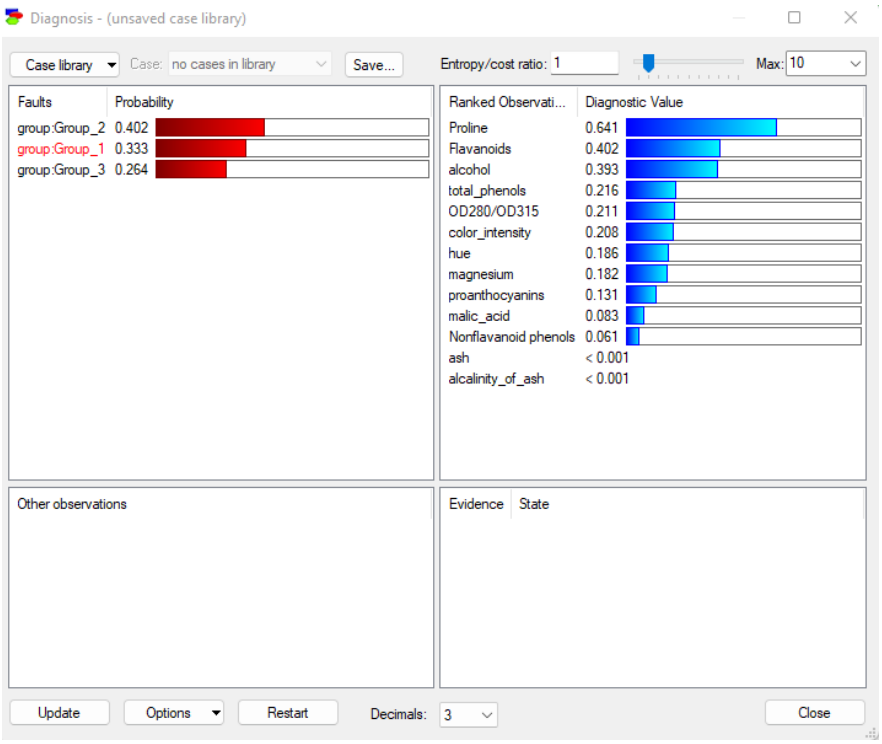
I/ Bayesian Search Model:



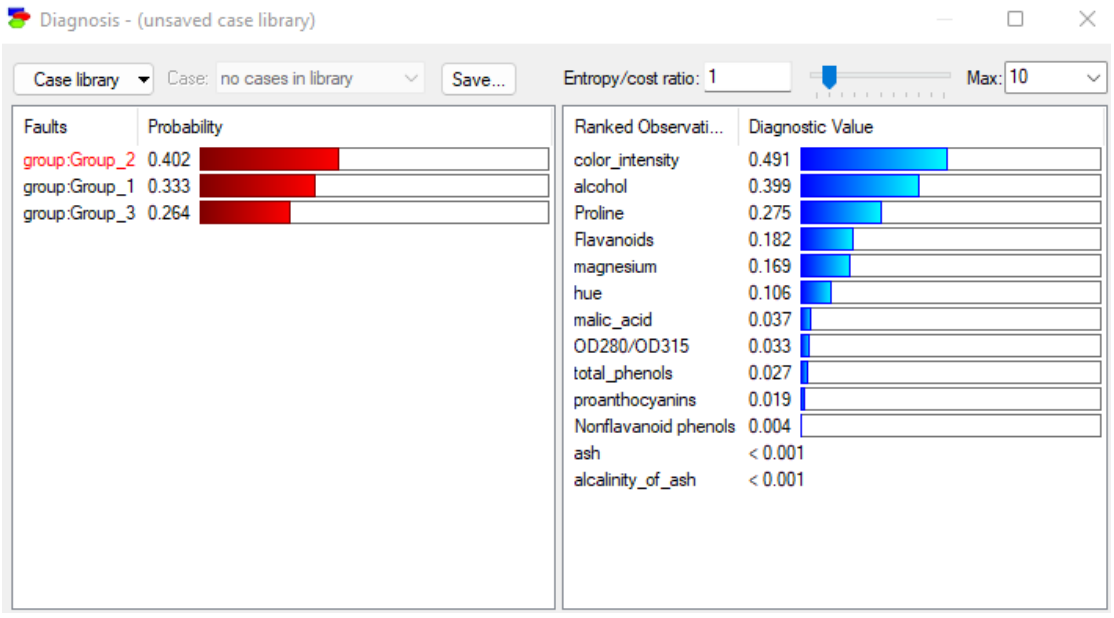
Joaquim JUSSEAU

1/ Diagnosis for each group

Group 1:

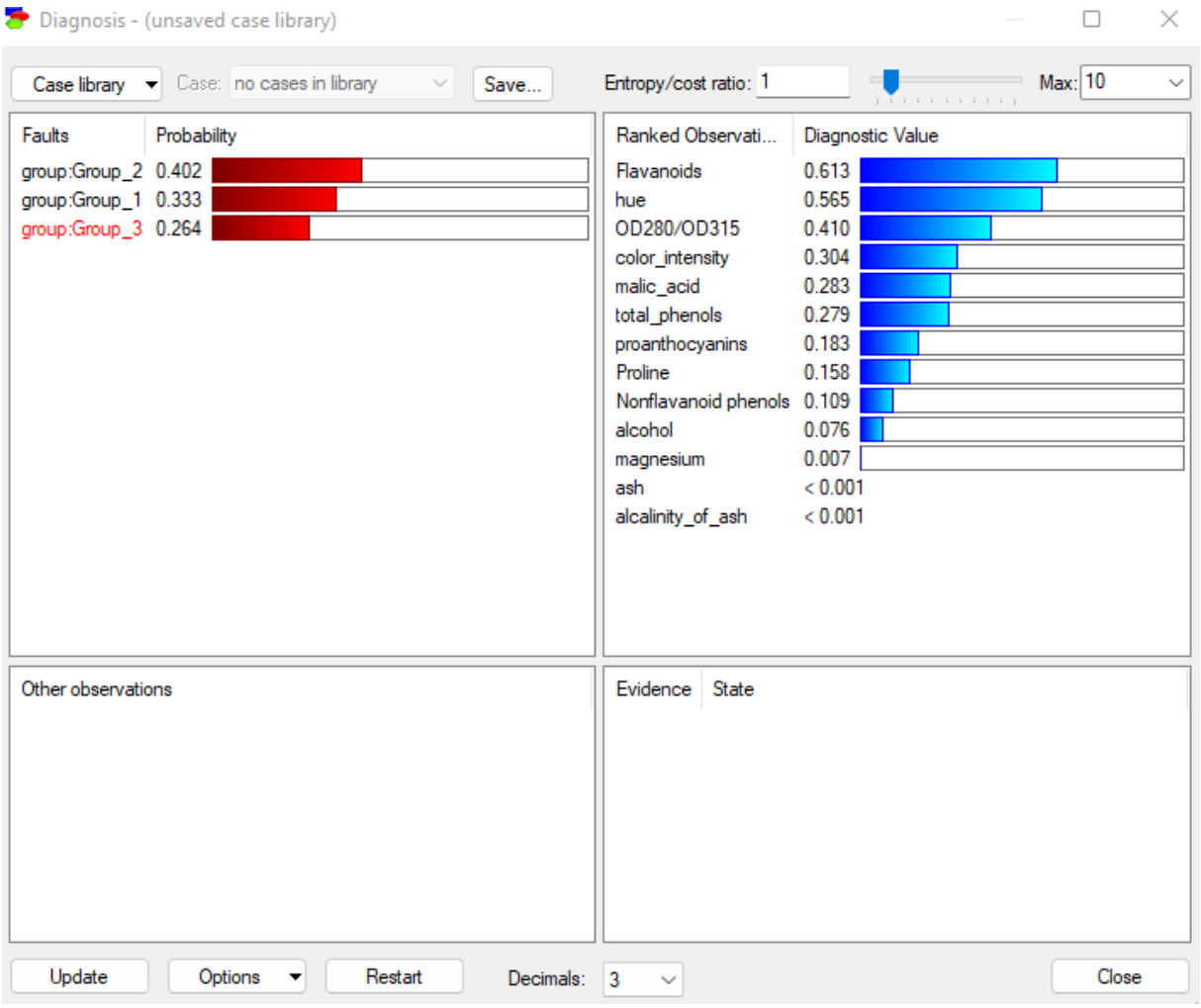


Group 2:



Joaquim JUSSEAU

Group 3:



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.

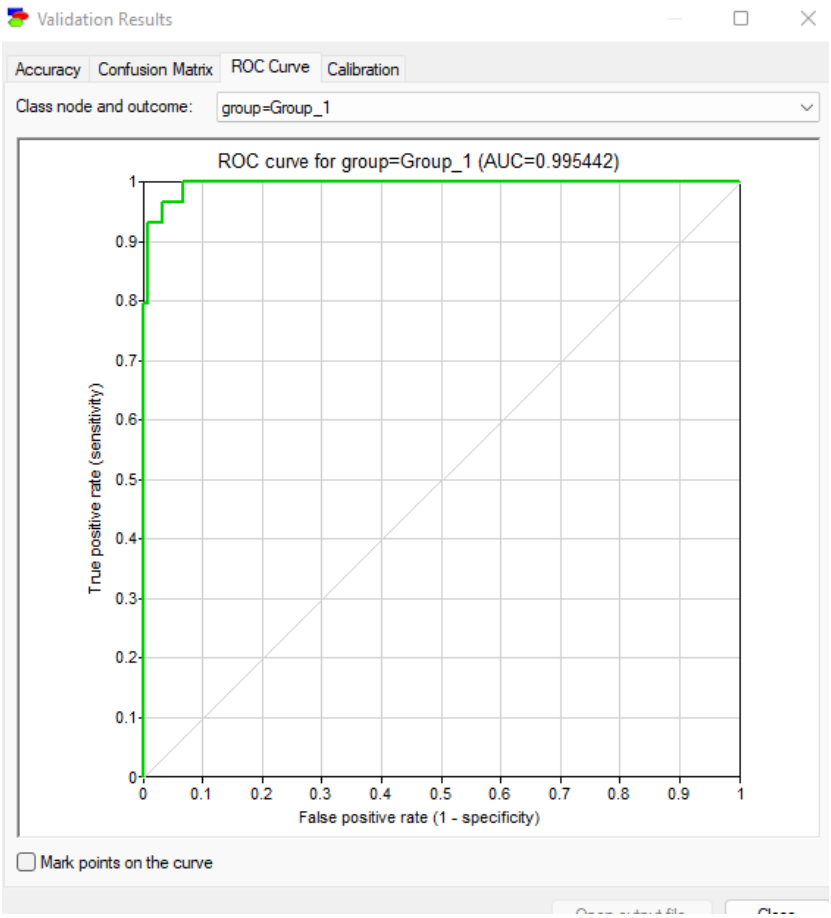
2/ Model Validation

group = 0.960674 (171/178)
Group_1 = 0.932203 (55/59)
Group_2 = 0.957746 (68/71)
Group_3 = 1 (48/48)

Confusion Matrix:

		Predicted		
		Group_1	Group_2	Group_3
Actual	Group_1	**55**	4	0
	Group_2	1	**68**	2
	Group_3	0	0	**48**

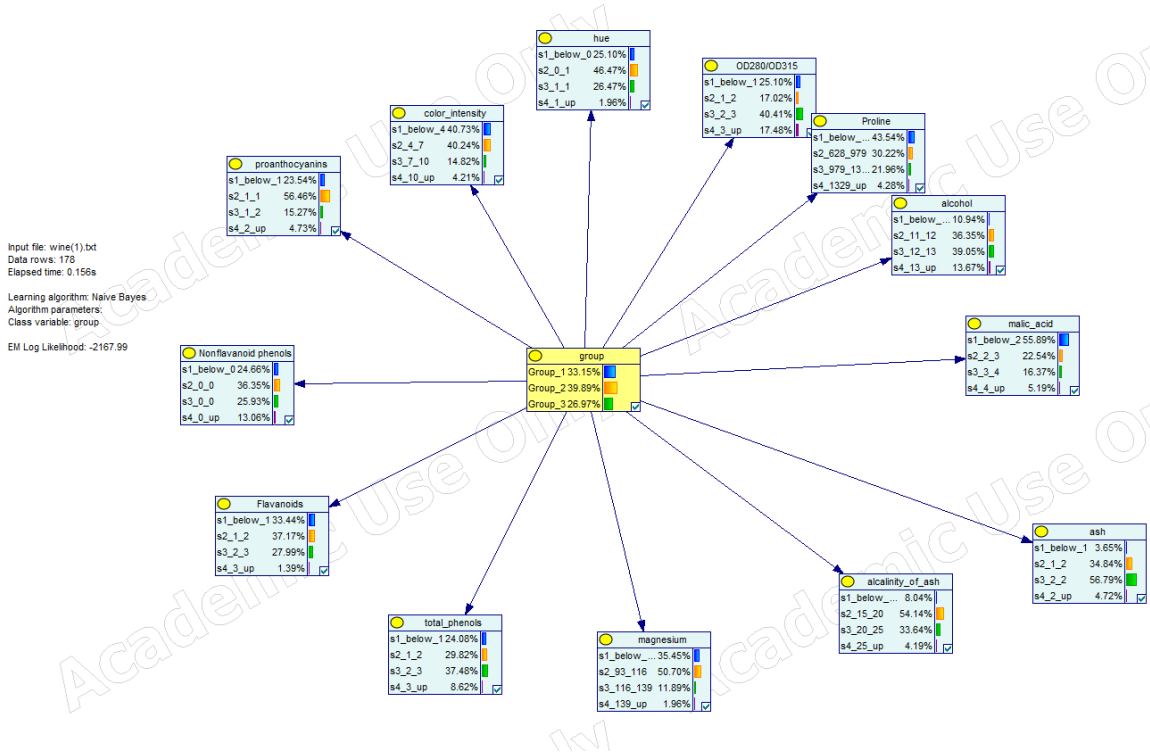
ROC Curve:



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

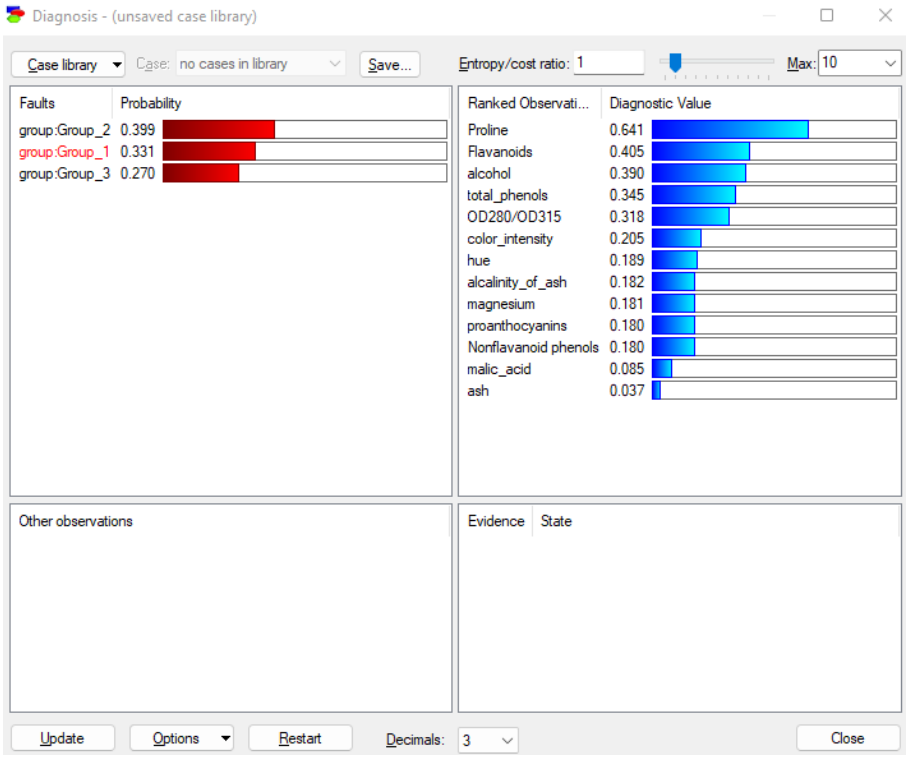
II/ Naive Bayes Model:

Joaquim JUSSEAU



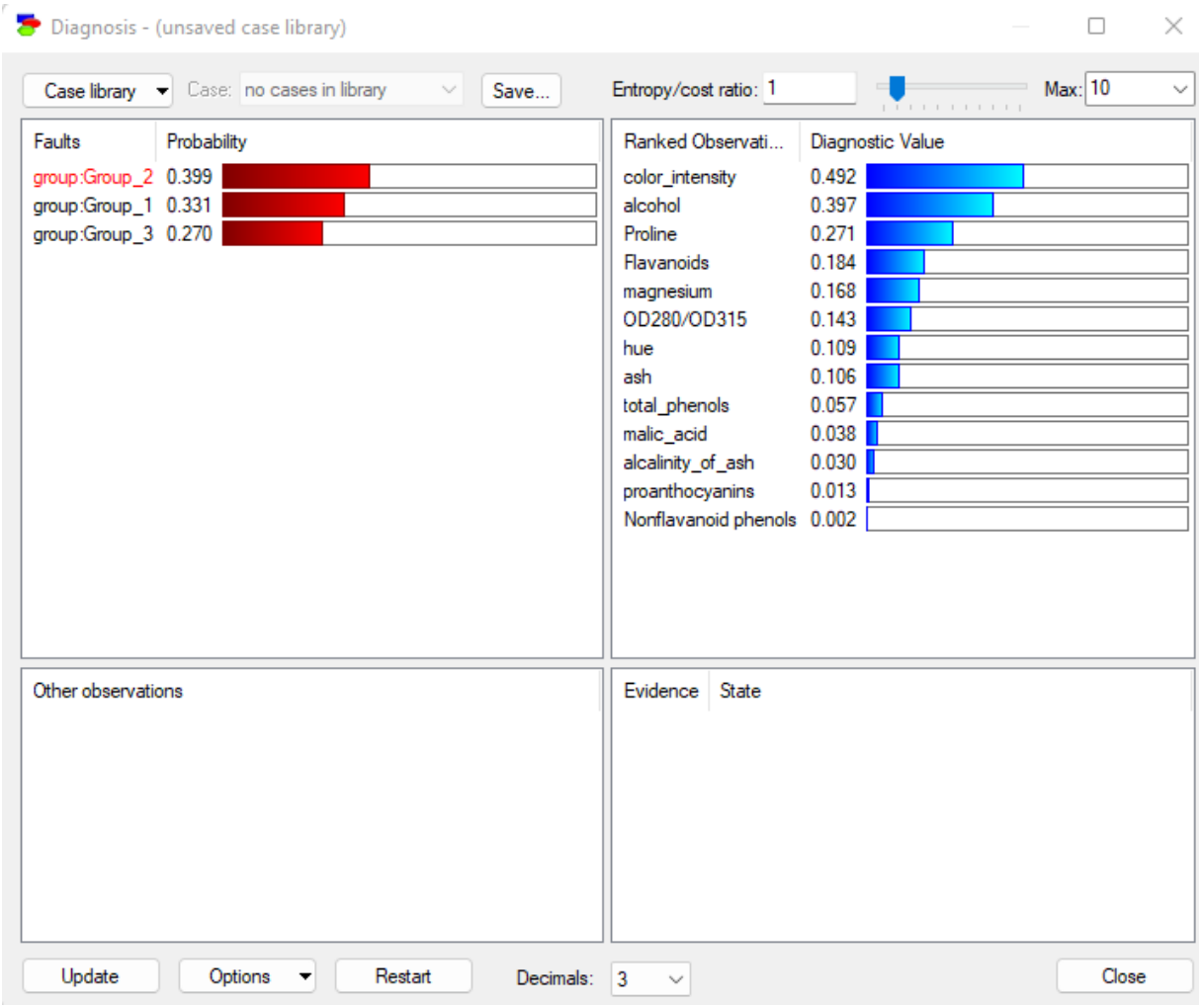
1/ Diagnosis for each group

Group 1:

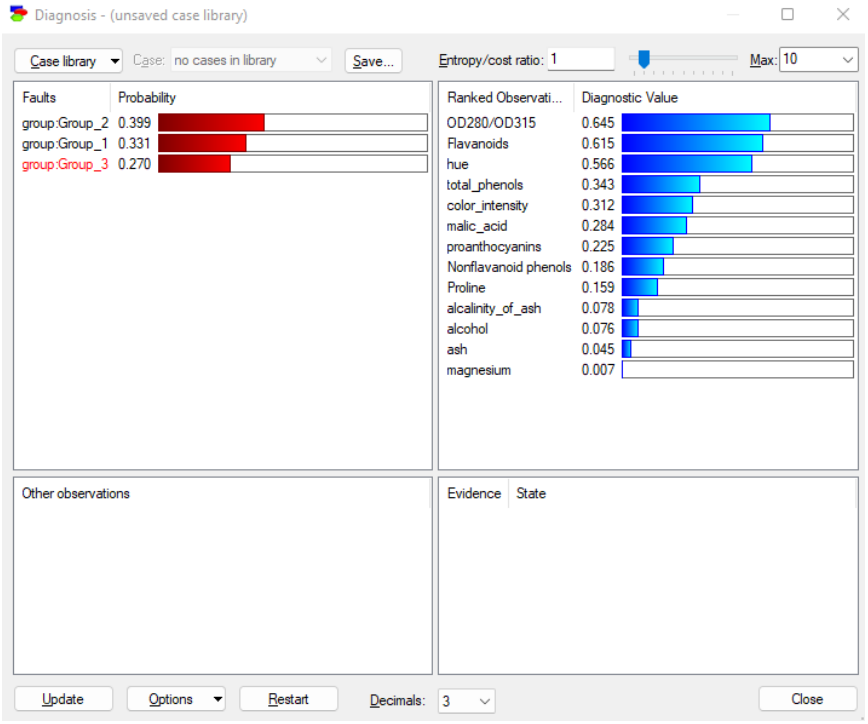


Joaquim JUSSEAU

Group 2 :



Group 3 :



Joaquim JUSSEAU

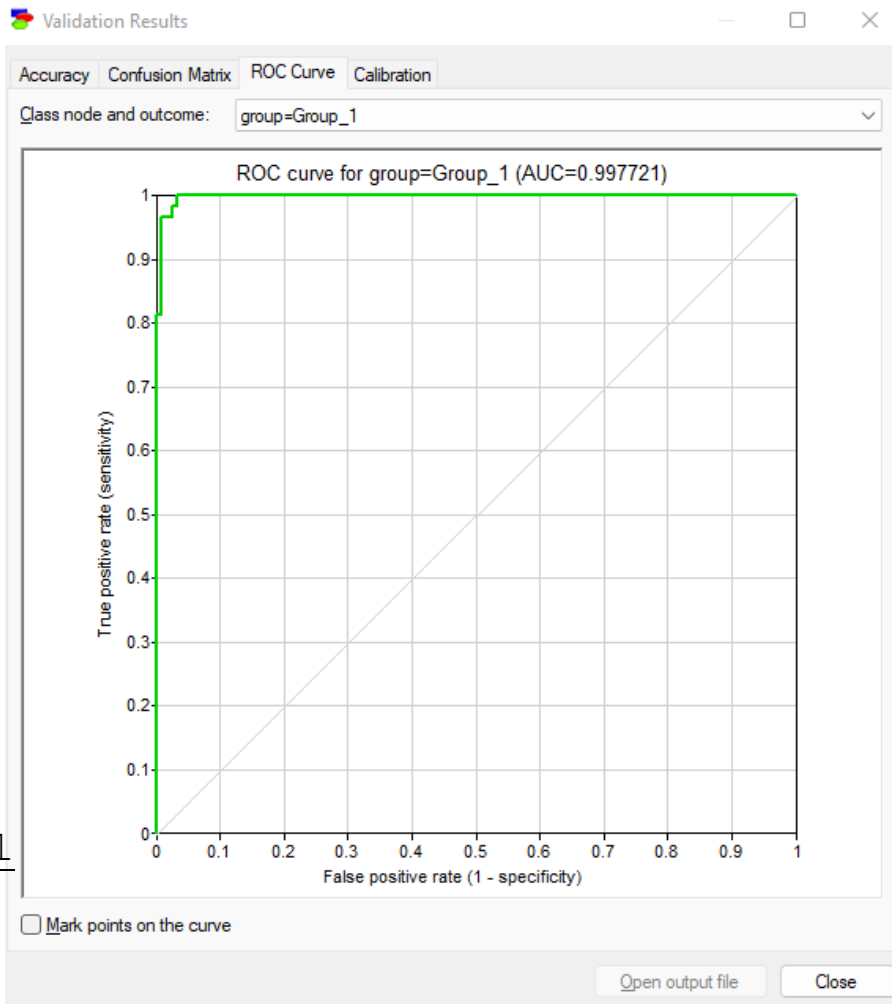
2/ Model Validation

Confusion Matrix:

		Predicted		
		Group_1	Group_2	Group_3
Actual	Group_1	**57**	2	0
	Group_2	1	**66**	4
	Group_3	0	0	**48**

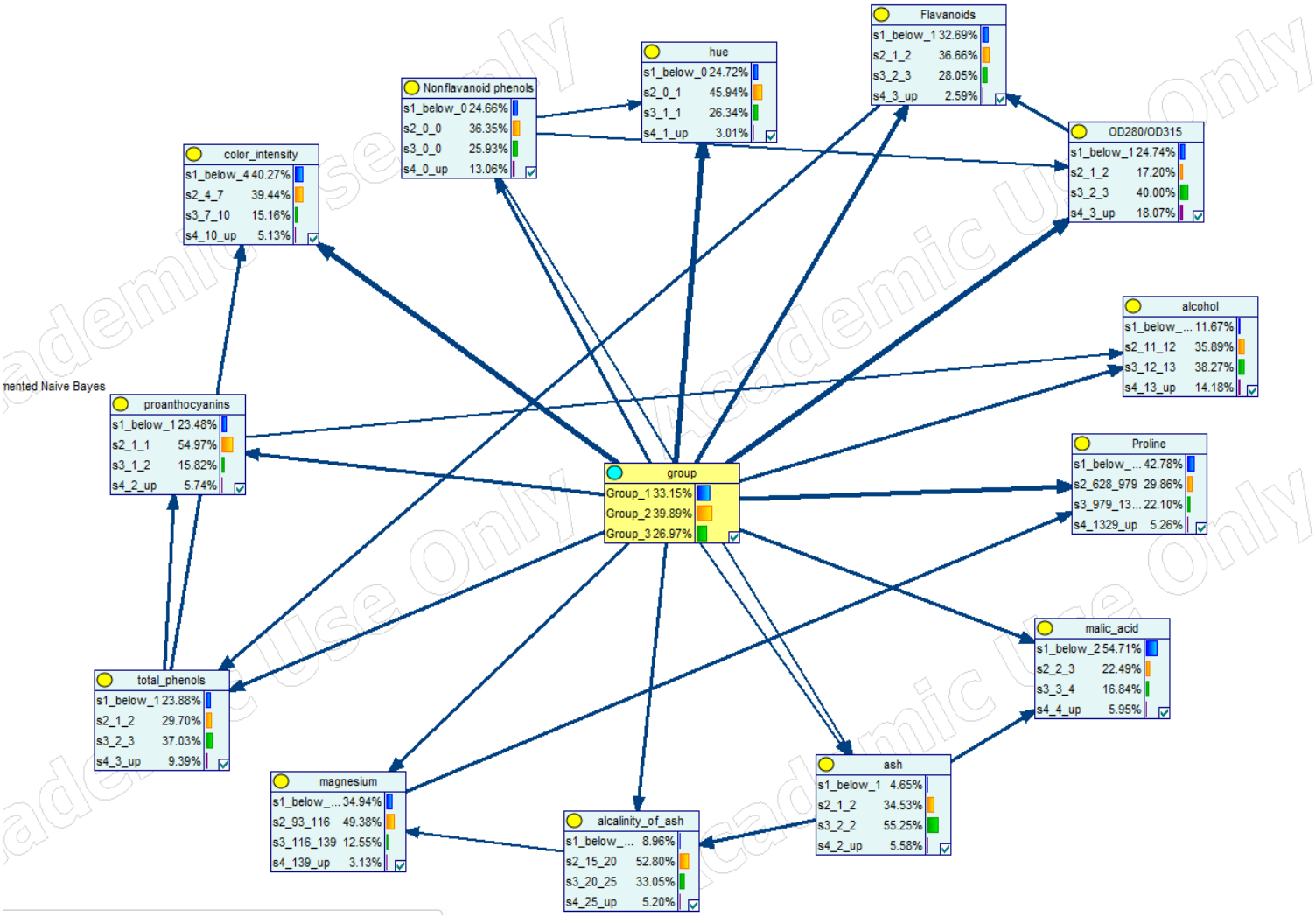
group = 0.960674 (171/178)
Group_1 = 0.966102 (57/59)
Group_2 = 0.929577 (66/71)
Group_3 = 1 (48/48)

ROC Curve:



IV/ TAN Model

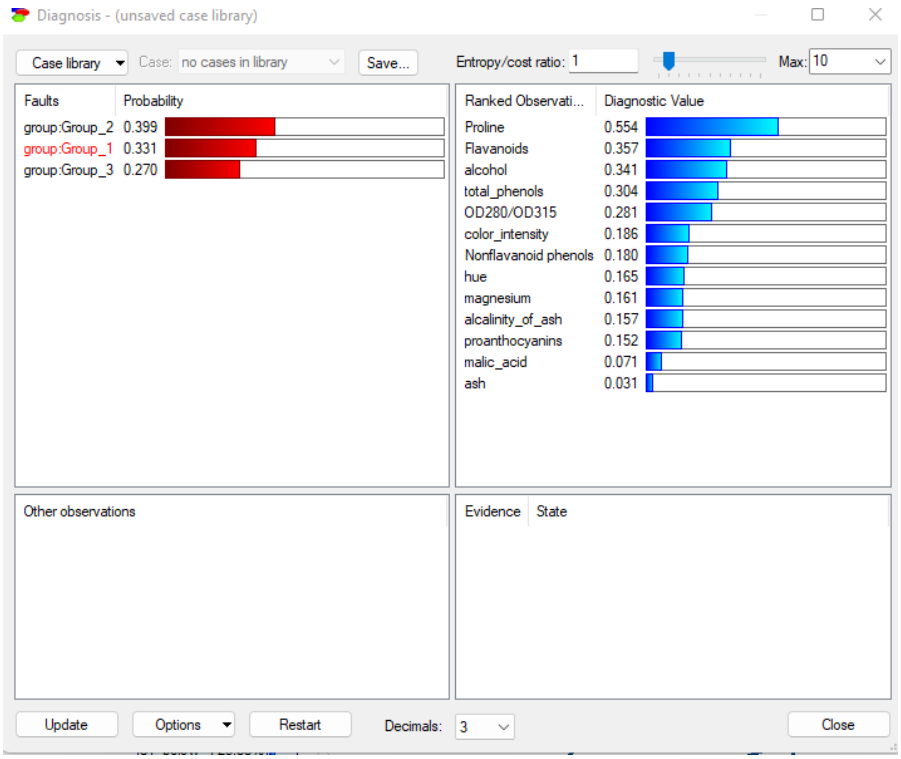
Joaquim JUSSEAU



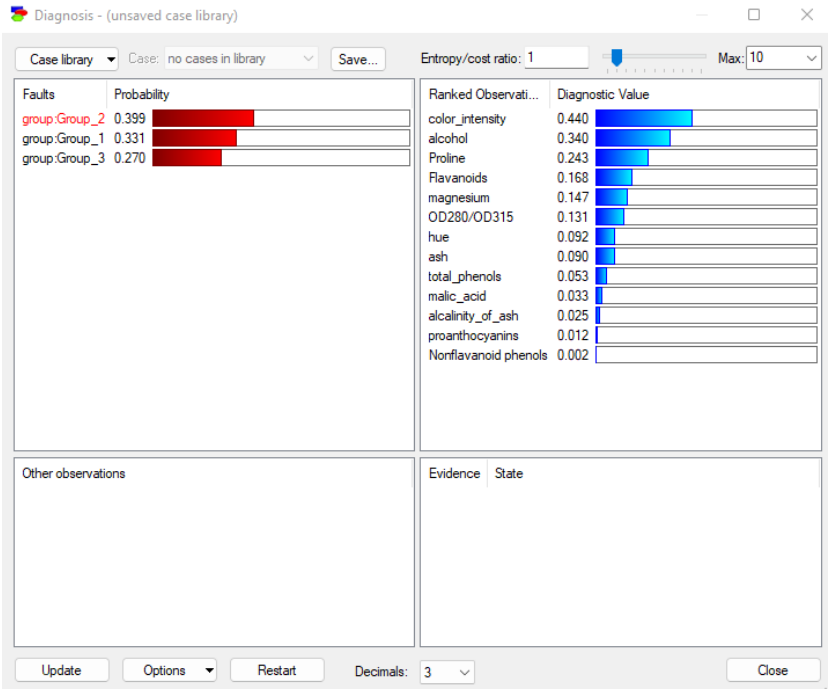
1/ Diagnosis for each group

Group 1:

Joaquim JUSSEAU

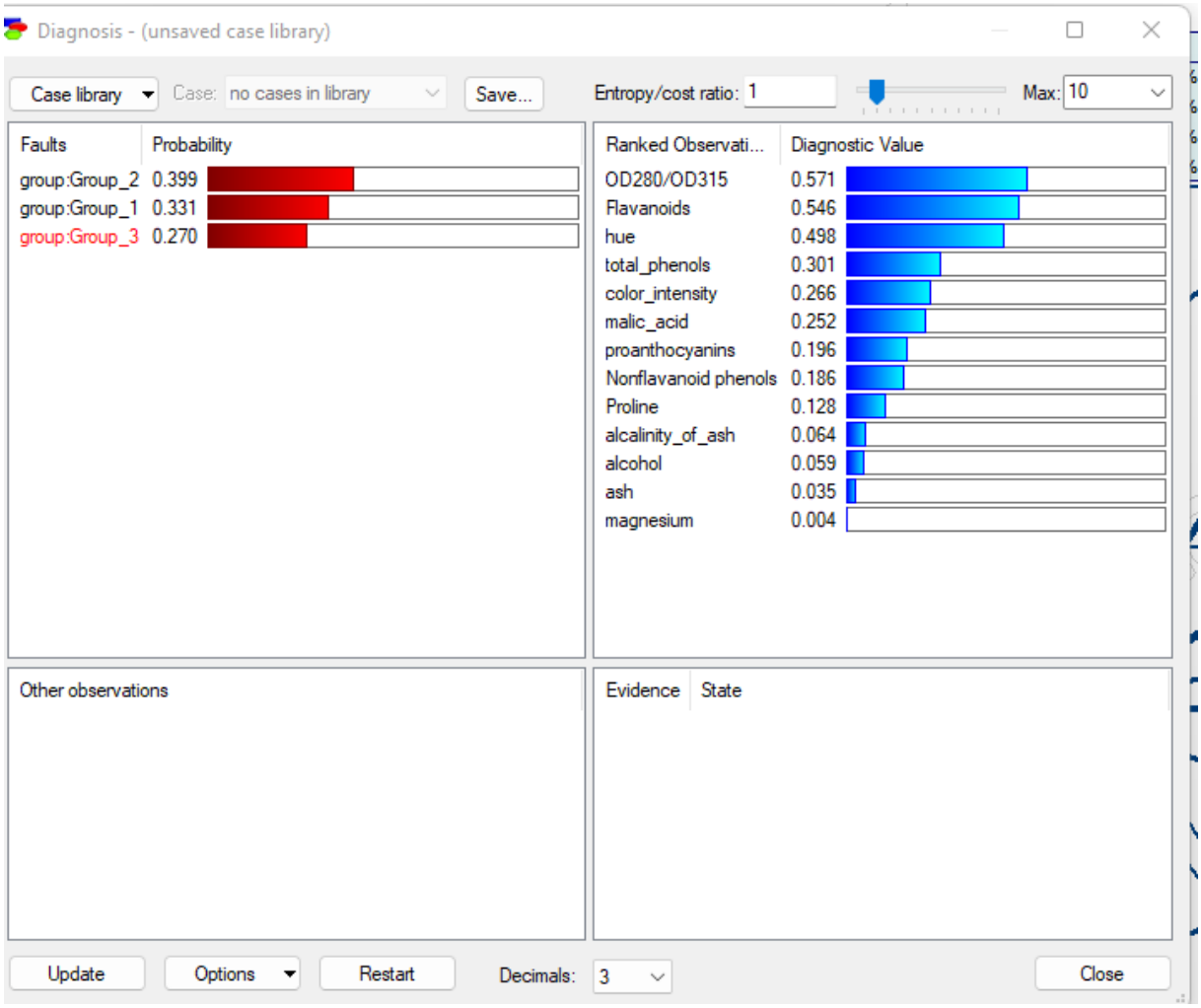


Group 2:



Group 3:

Joaquim JUSSEAU



2/ Model Validation

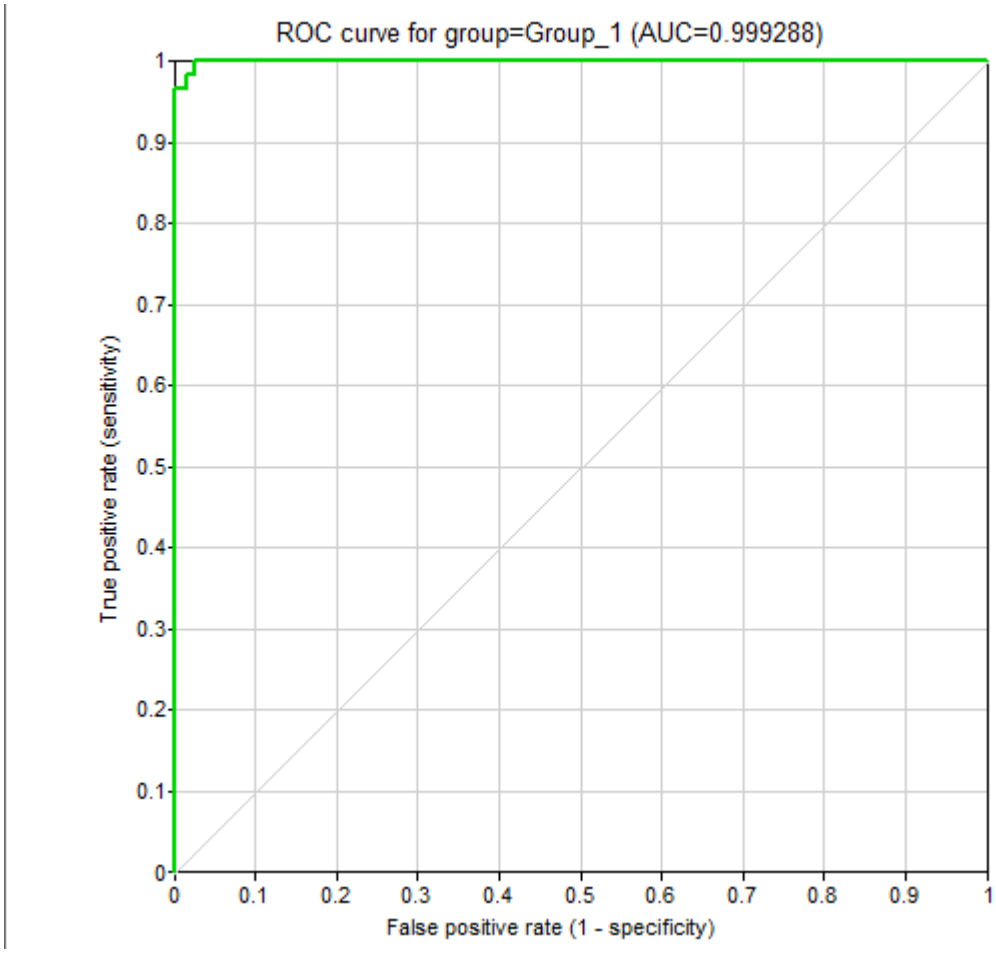
Confusion Matrix:

Accuracy:

group = 0.966292 (172/178)
Group_1 = 0.983051 (58/59)
Group_2 = 0.929577 (66/71)
Group_3 = 1 (48/48)

		Predicted		
		Group_1	Group_2	Group_3
Actual	Group_1	**58**	1	0
	Group_2	2	**66**	3
	Group_3	0	0	**48**

ROC Curve:



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.