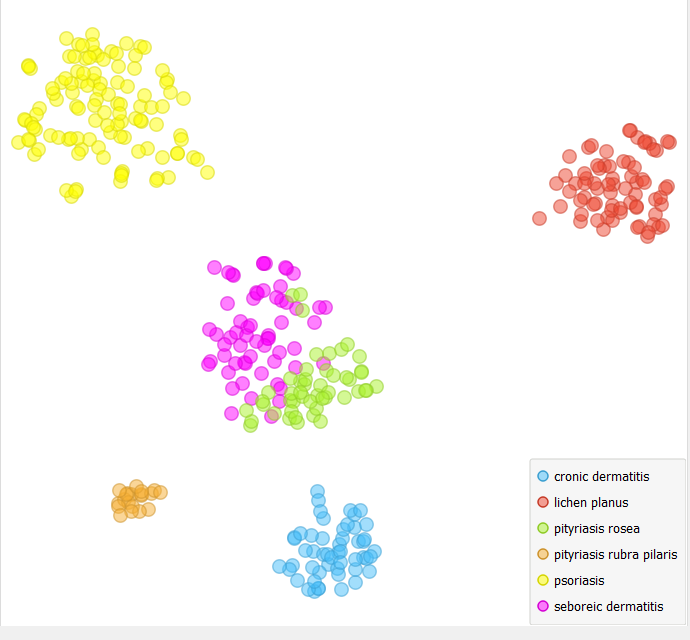
**Anastasia Psarou Lab 4.  Simple classifiers**

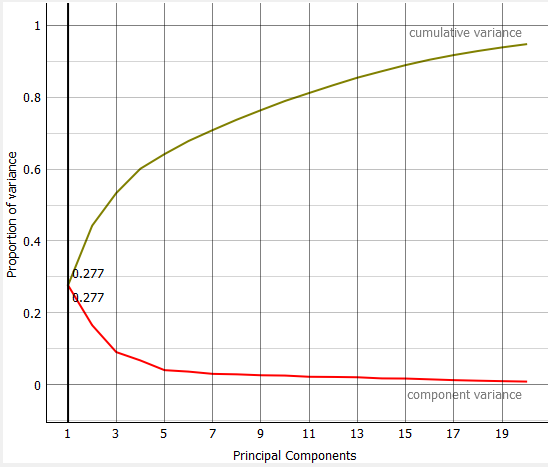
**a) Select the dataset *dermatology*. Read about it and try to understand features.**

**b) Visualize the dataset, using PCA and t-SNE.**

t-SNE display



PCA display

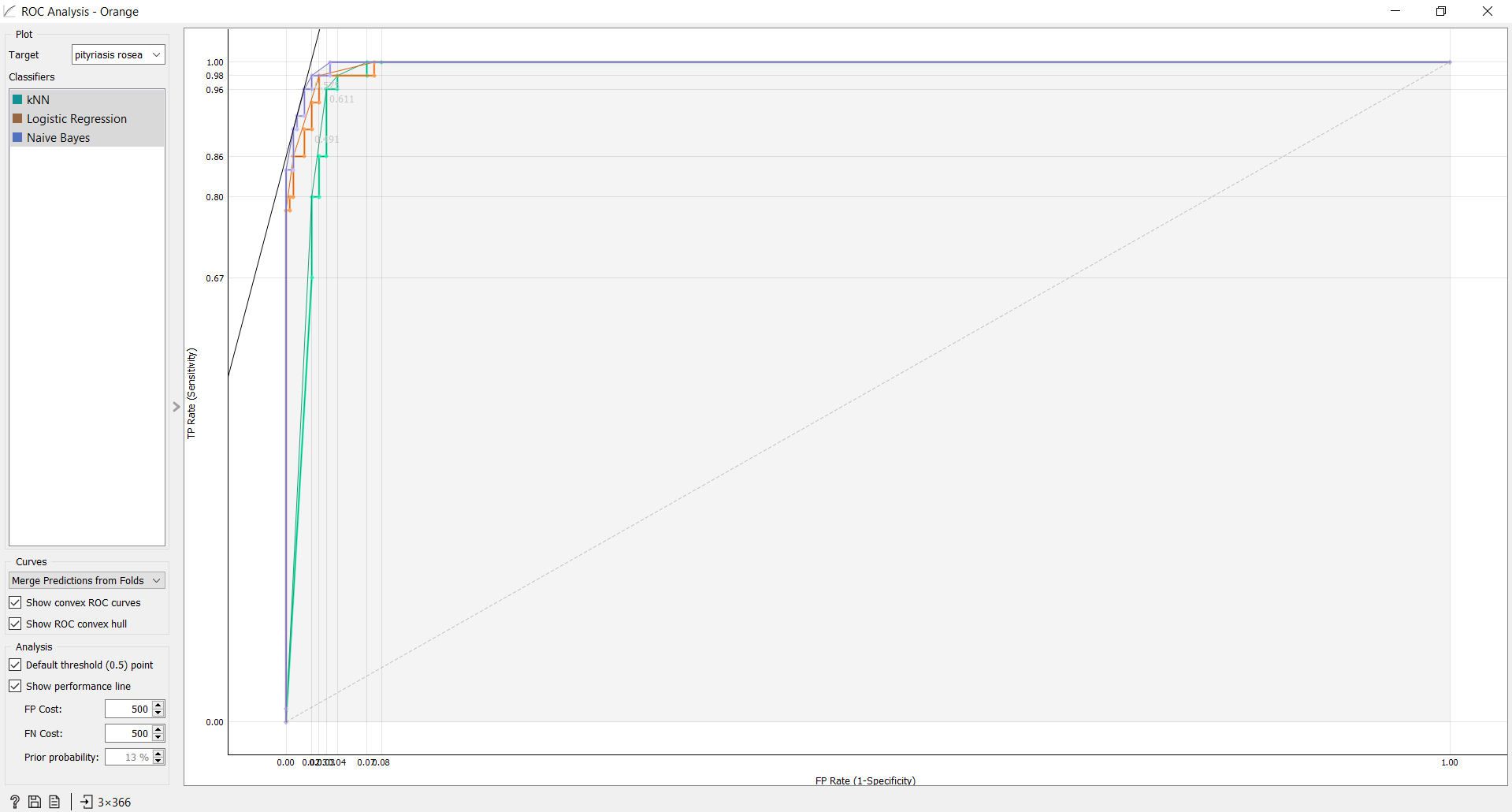


**c) Use simple classifiers: Naïve Bayes, Logistic regression and k-NN to build data models. Decide about the values of the metaparameters of the Logistic regression classifier. Which classifier is better?? Compare ROC curves.**

I choose the Ridge L2 metaparameter at Logistic Regression and C = 0.5.

I used as target the pityriasis rosea.

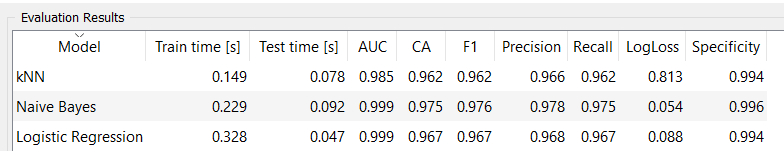
So, these are ROC analysis:



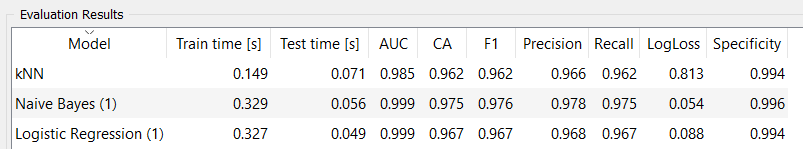
From this photo it is clear that Naïve Bayes is the best classifier. This happens as its line in the ROC analysis has smaller gradient compared to the other classifiers.

**d) Compare the efficiency of all classifiers for raw data and for data transformed employing PCA (far all dimensions).**

These are the results without the PCA transformation:



These are the results after the PCA transformation:



We can see that as far as the kNN is concerned the only thing that has a little different value before and after the PCA transformation is Test Time.

As far as Naïve Bayes is concerned only Train and Test Time have different values without and after PCA transformation. This happens also with Logistic Regression classifier.

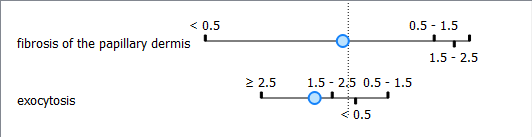
**e) Discuss the results. Explain WHY are these results what they are?**

PCA can be used to simplify visualizations of large datasets. So, PCA does a transformation of data but that does not affect the Accuracy, Precision etc. So, we excepted to do not see a difference in the values of these parameters.

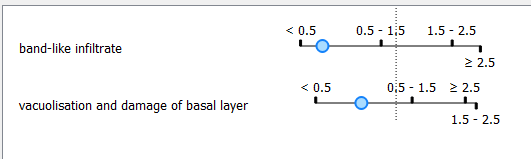
**d)  Use nomograms (nomogram widget, read the help about this type of visualization) to rank the best features for each class. Remove the best two features explaining each class.**

Firstly, we are going to check the best features from Naïve Bayes.

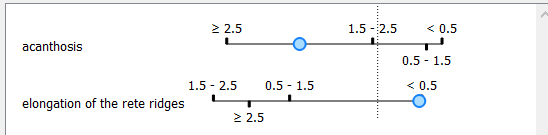
* Cronic dermatitis



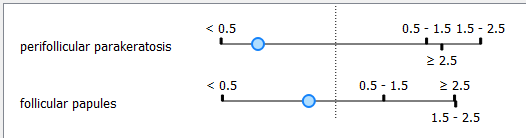
* Litchen planus



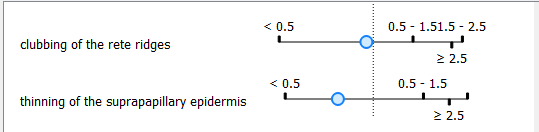
* Pityriasis rosea



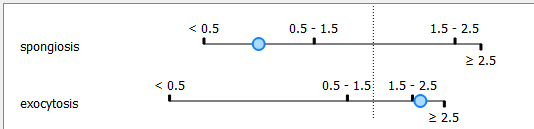
* Pityriasis rubra pilaris



* Psoriasis

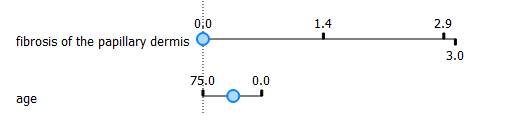


* Seboreic dermatitics

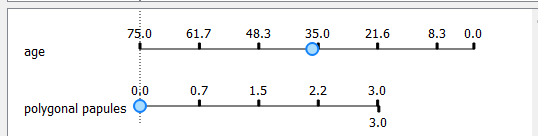


Now we are going to check the best features from Logistic Regression.

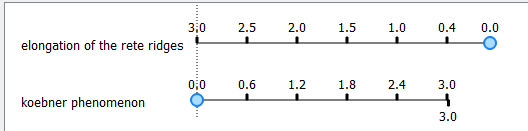
* Cronic dermatitis



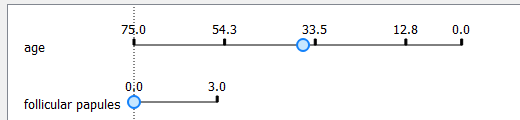
* Lichen planus



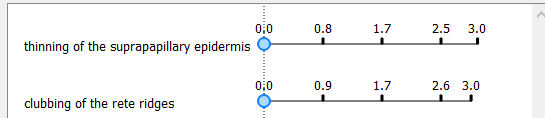
* Pityriasis rosea



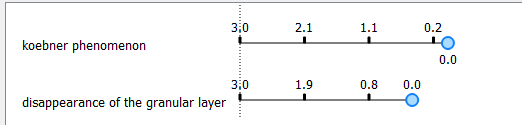
* Pityriasis rubra pilaris



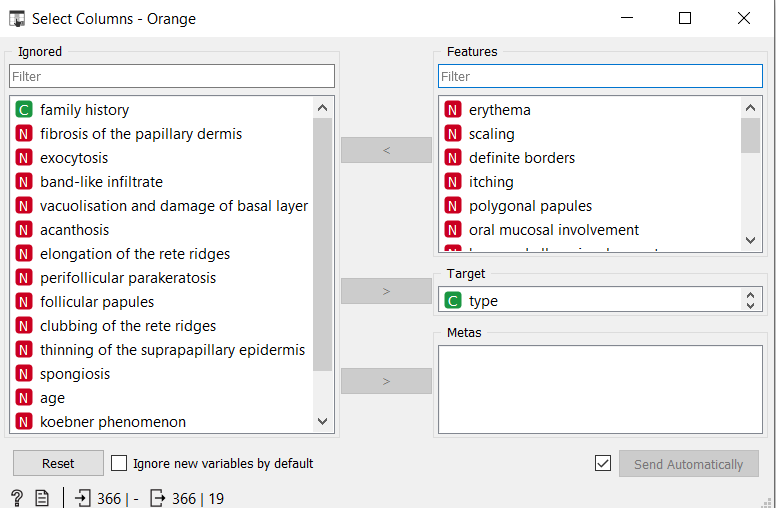
* Psoriasis



* Seboreic dermatitis



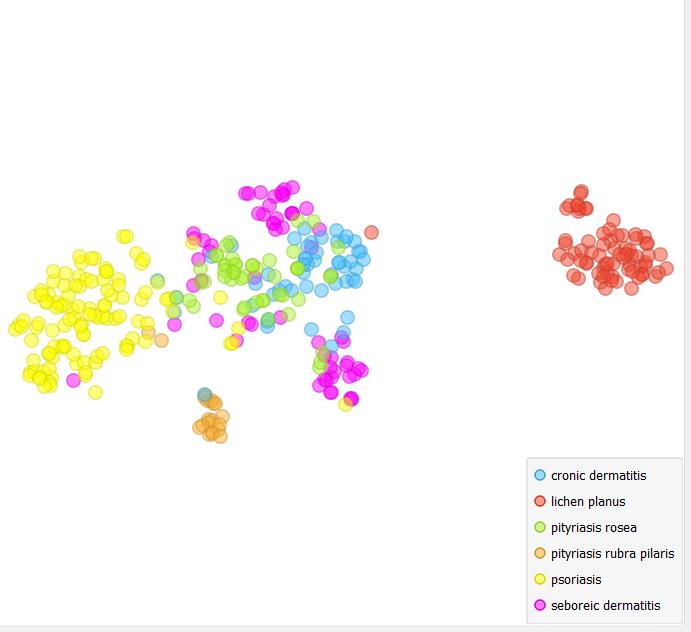
So, I removed these features from Select Columns widget.



**e) Use the remaining features for b) to e).**

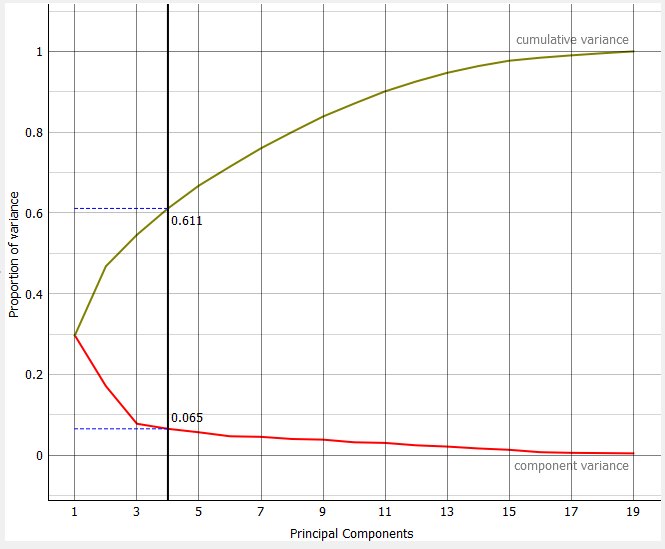
**b) Visualize the dataset, using PCA and t-SNE.**

t-SNE display

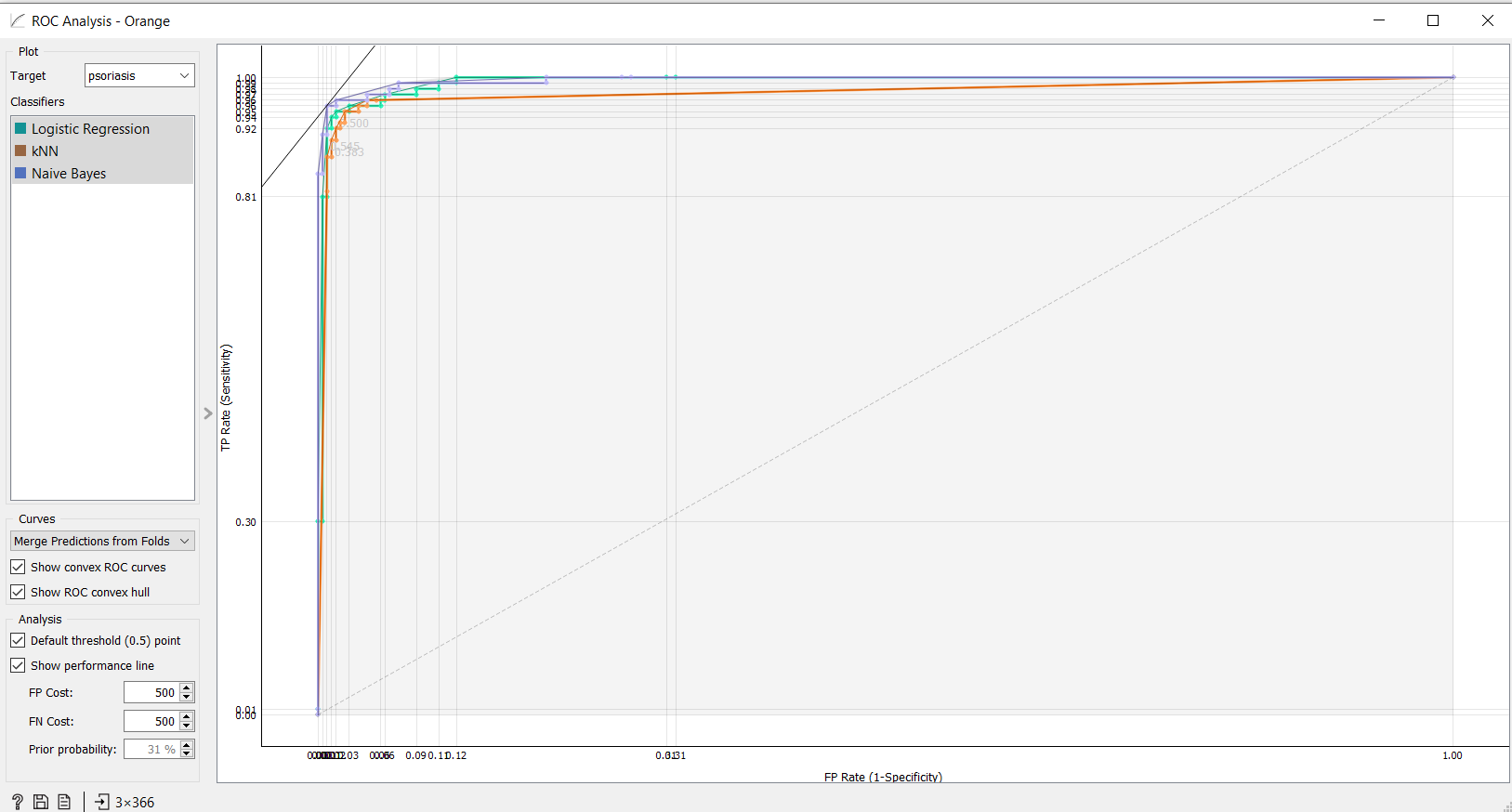


We can see that classes are not as discrete as they were before removing the best features of each class.

PCA display



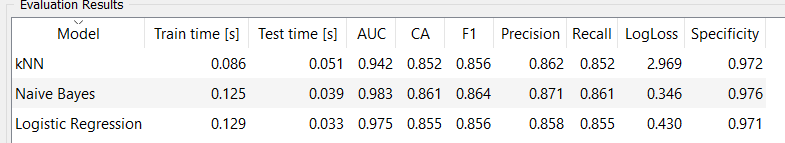
**c) Use simple classifiers: Naïve Bayes, Logistic regression and k-NN to build data models. Decide about the values of the metaparameters of the Logistic regression classifier. Which classifier is better?? Compare ROC curves.**



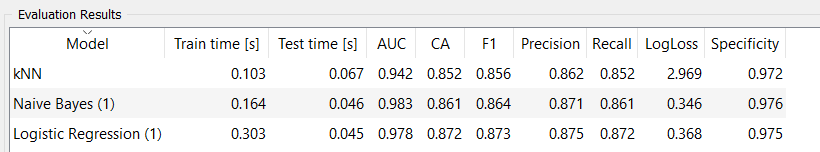
From the above graph it is clear that the best classifier is Naïve Bayes. We came to this conclusion and in the above question.

**d) Compare the efficiency of all classifiers for raw data and for data transformed employing PCA (far all dimensions).**

These are the results without the PCA transformation:



These are the results after the PCA transformation:



Naïve Bayes and kNN do not appear any changes except from Train and Test Time as was noticed before.

Only in Logistic Regression we can see some differences in the values but they are very small.

**e) Discuss the results. Explain WHY are these results what they are?**

The changes in the Logistic Regression are appeared perhaps due to the absence of the best features and the unstabilization that this action caused. As far as kNN and Naïve Bayes are concerned they perhaps are more adaptive and do not need best features to draw right results.

**f) Formulate conclusions.**

Without the best features the set is not as nice. This can also be seen by the t-SNE and PCA displays, where the classes are not so discrete. Also, from ROC analysis we can see that even if we eliminate some of the best features the best classifier is not affected.