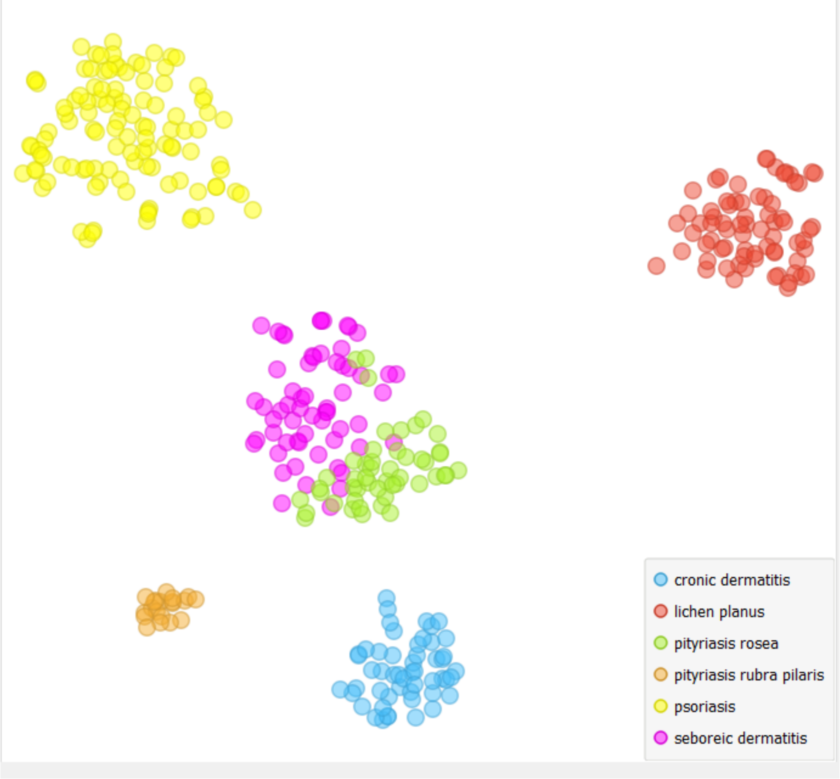
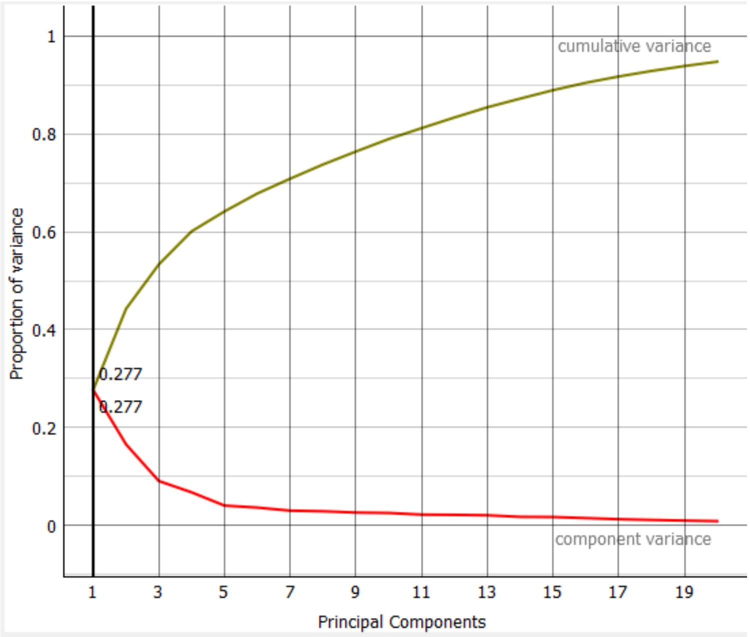
Anastasia Psarou Lab 4. Simple classifiers

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

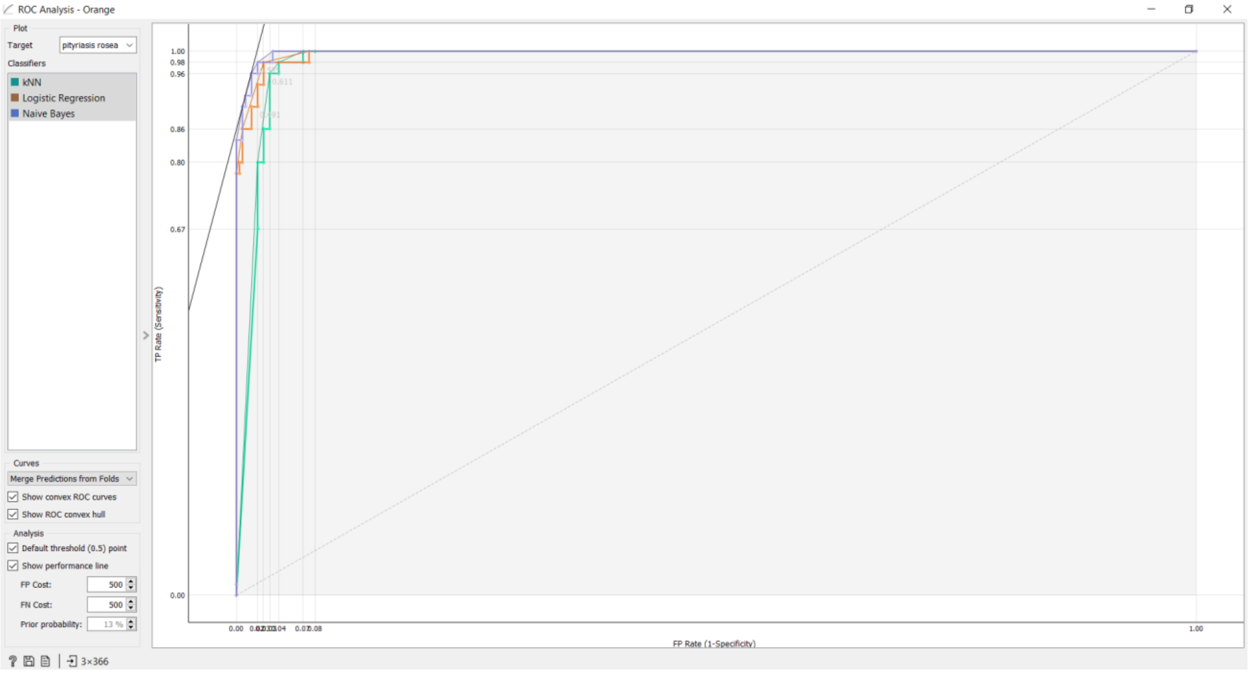


c) Use simple classifiers: Naïve Bayes, Logistic regression and k-NN to build data models. Decide about the values of the metaparameters of the Logistic regression classifier. Which classifier is better?? Compare ROC curves.

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:



ROC Analysis - Orange

Plot

Target: pityriasis rosea

Classifiers

- 🟩 kNN
- 🟫 Logistic Regression
- 🟦 Naive Bayes

Curves

Merge Predictions from Folds

☑ Show convex ROC curves

☑ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

FN Cost: 500

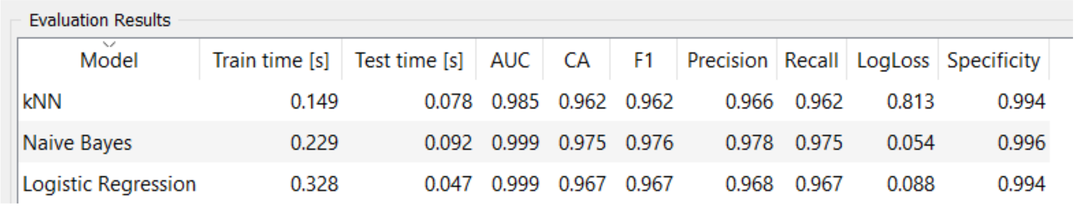
Prior probability: 13 %

3×366

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).

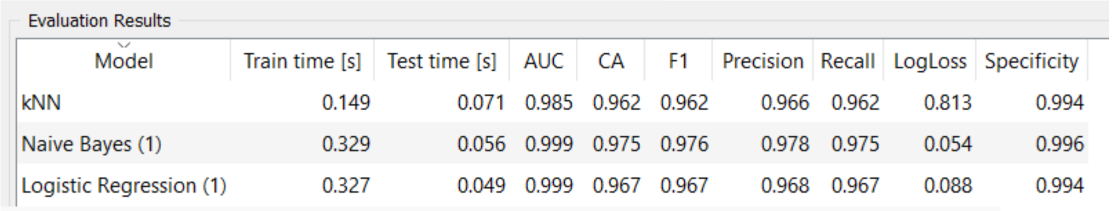
These are the results without the PCA transformation:



Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
kNN	0.149	0.078	0.985	0.962	0.962	0.966	0.962	0.813	0.994
Naive Bayes	0.229	0.092	0.999	0.975	0.976	0.978	0.975	0.054	0.996
Logistic Regression	0.328	0.047	0.999	0.967	0.967	0.968	0.967	0.088	0.994

These are the results after the PCA transformation:



Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
kNN	0.149	0.071	0.985	0.962	0.962	0.966	0.962	0.813	0.994
Naive Bayes (1)	0.329	0.056	0.999	0.975	0.976	0.978	0.975	0.054	0.996
Logistic Regression (1)	0.327	0.049	0.999	0.967	0.967	0.968	0.967	0.088	0.994

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.

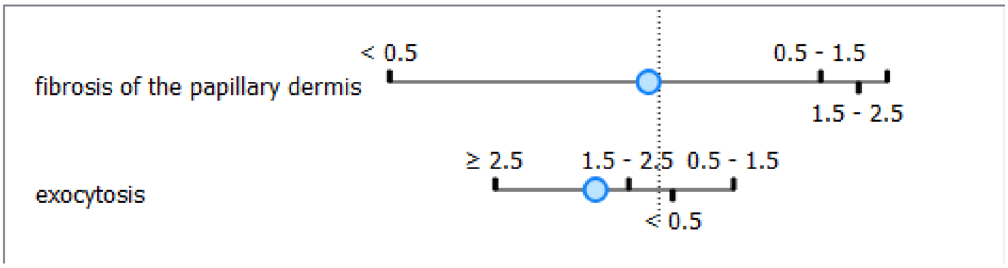
e) Discuss the results. Explain WHY are these results what they are?

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.

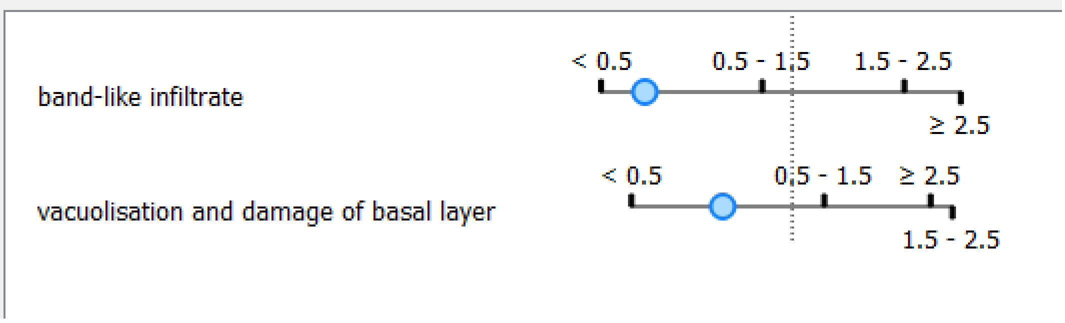
d) Use nomograms (nomogram widget, read the help about this type of visualization) to rank the best features for each class. Remove the best two features explaining each class.

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

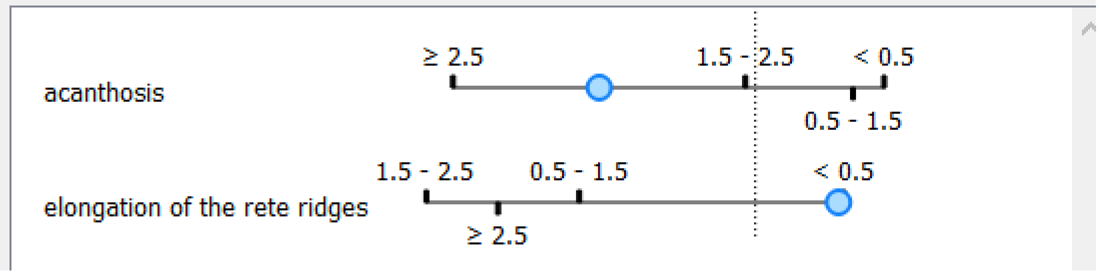
- Cronic dermatitis



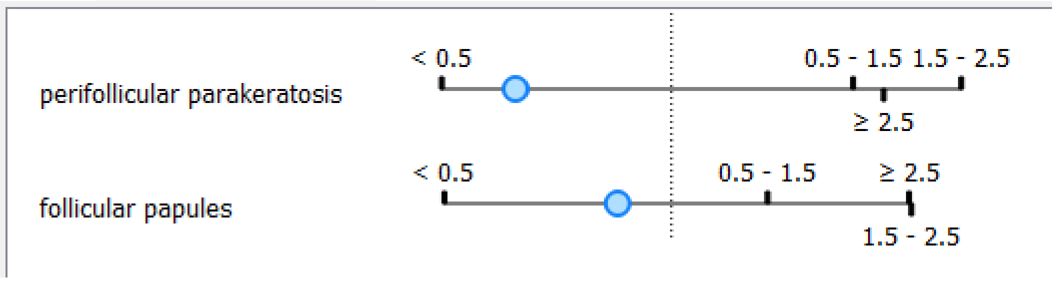
- Litchen planus



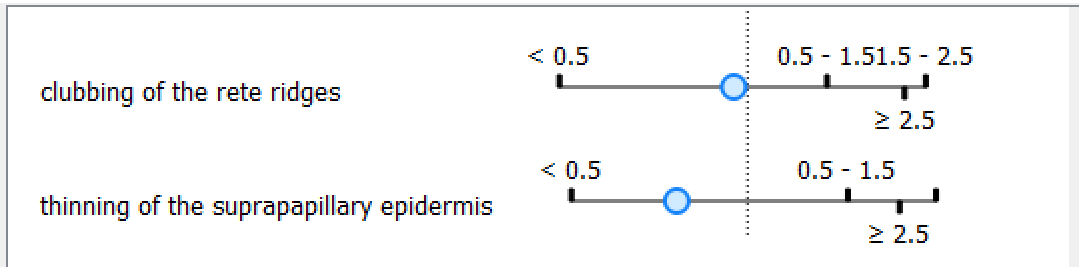
- Pityriasis rosea



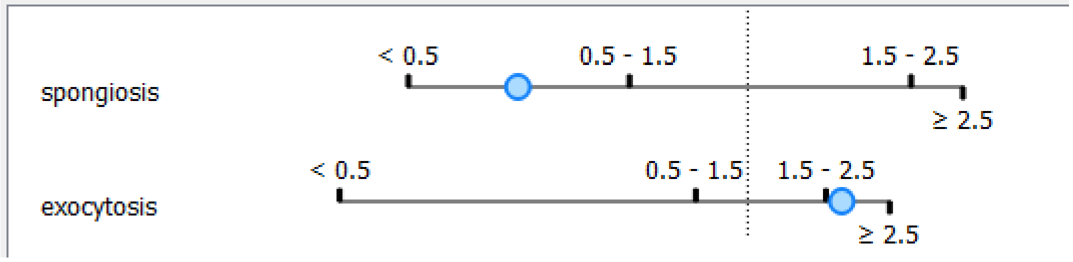
- Pityriasis rubra pilaris



- Psoriasis

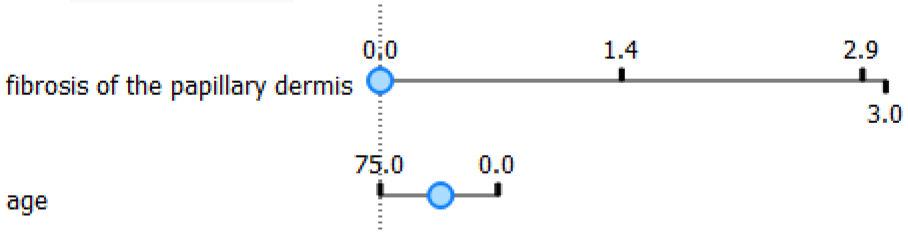


- Seboreic dermatitics

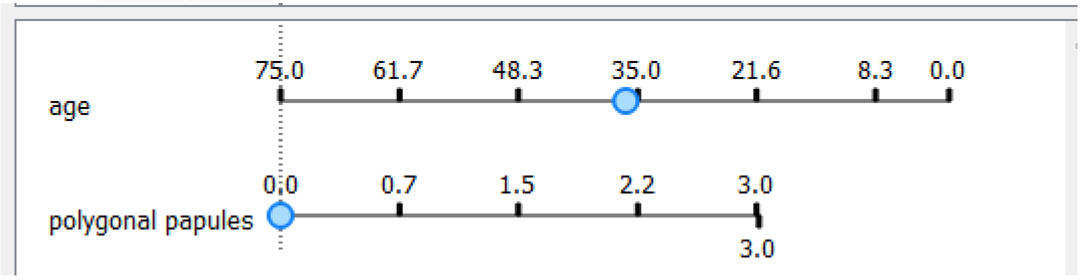


Now we are going to check the best features from Logistic Regression.

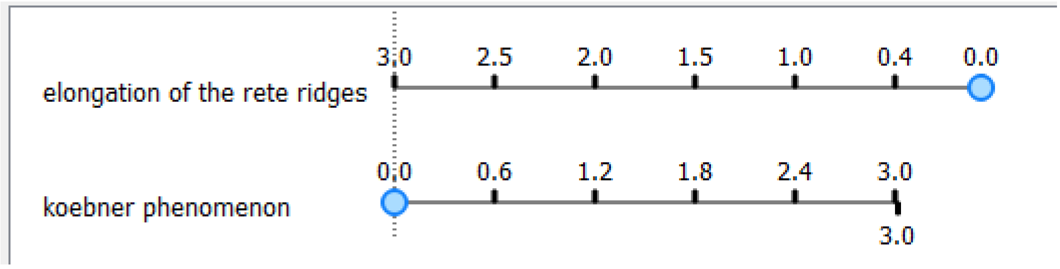
- Cronic dermatitis



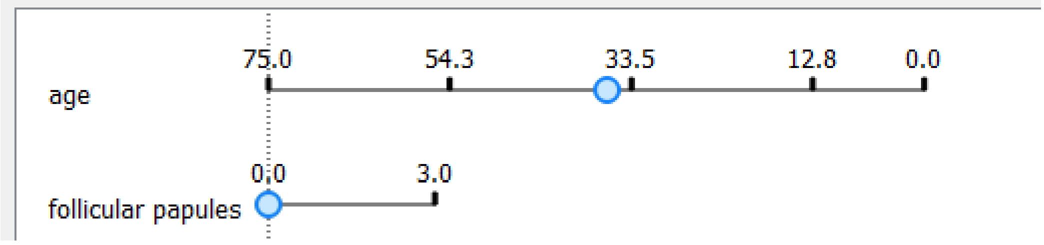
- Lichen planus



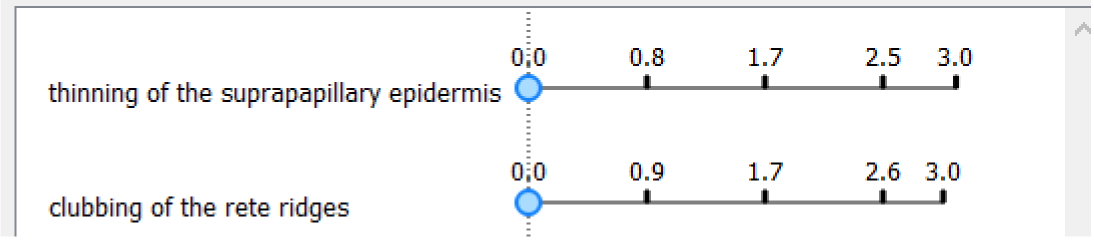
- Pityriasis rosea



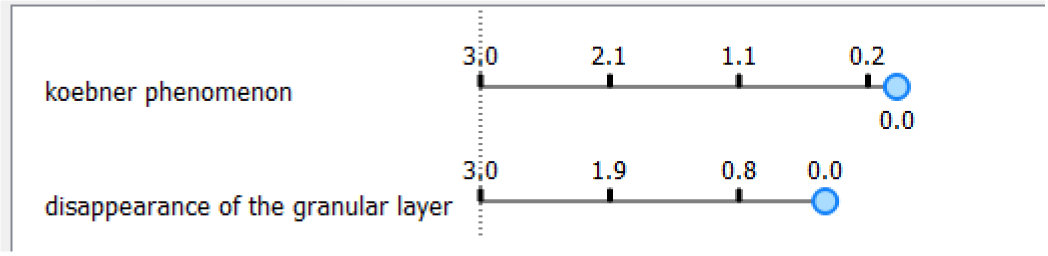
- Pityriasis rubra pilaris



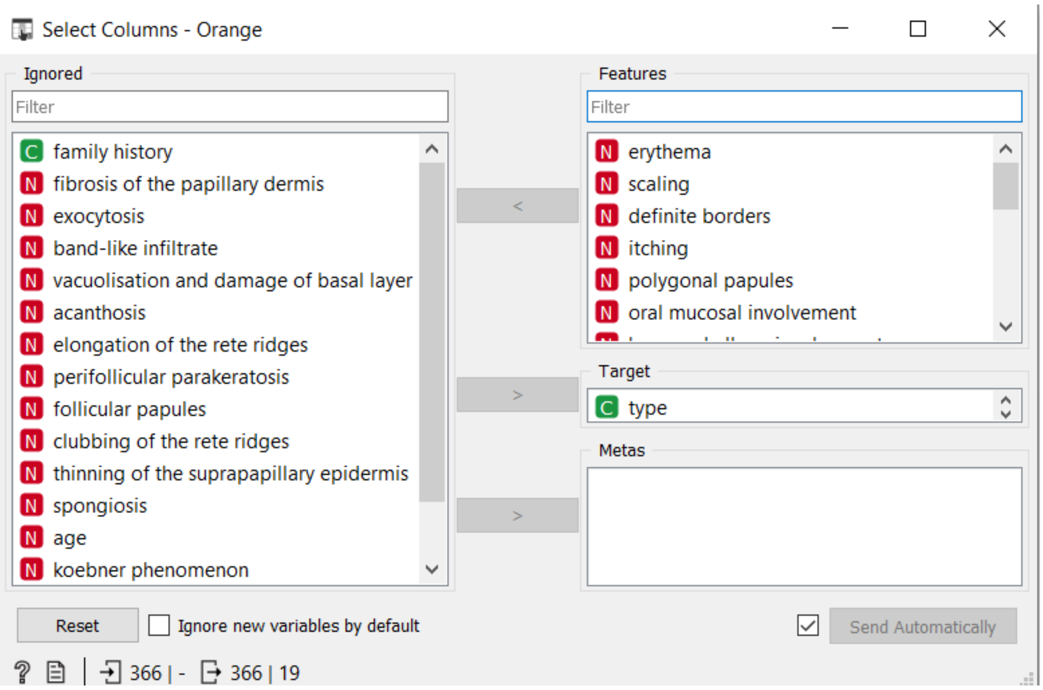
- Psoriasis


thinning of the suprapapillary epidermis
0.0
0.8
1.7
2.5
3.0
clubbing of the rete ridges
0.0
0.9
1.7
2.6
3.0

- Seboreic dermatitis



So, I removed these features from Select Columns widget.



Select Columns - Orange

Ignored

Filter

- C family history
- N fibrosis of the papillary dermis
- N exocytosis
- N band-like infiltrate
- N vacuolisation and damage of basal layer
- N acanthosis
- N elongation of the rete ridges
- N perifollicular parakeratosis
- N follicular papules
- N clubbing of the rete ridges
- N thinning of the suprapapillary epidermis
- N spongiosis
- N age
- N koebner phenomenon

Features

Filter

- N erythema
- N scaling
- N definite borders
- N itching
- N polygonal papules
- N oral mucosal involvement

Target

C type

Metas

Reset ☐ Ignore new variables by default

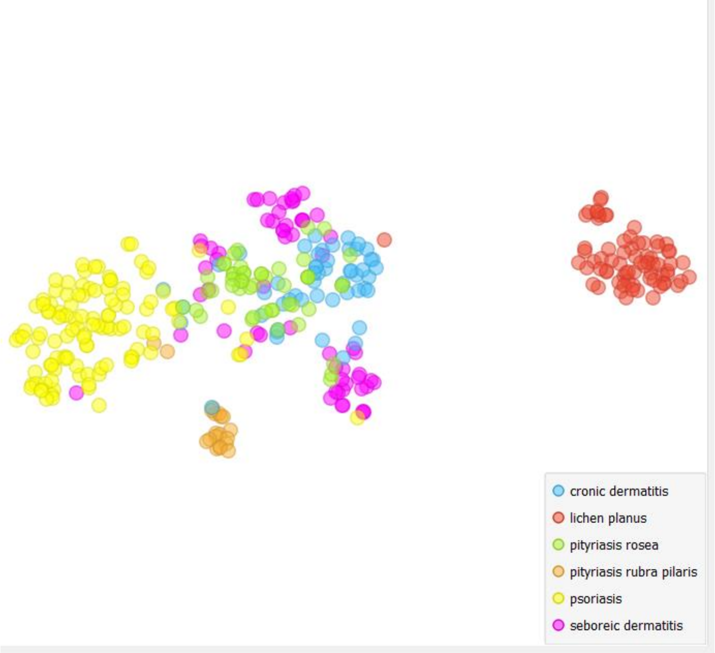
☑ Send Automatically

366 | - 366 | 19

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

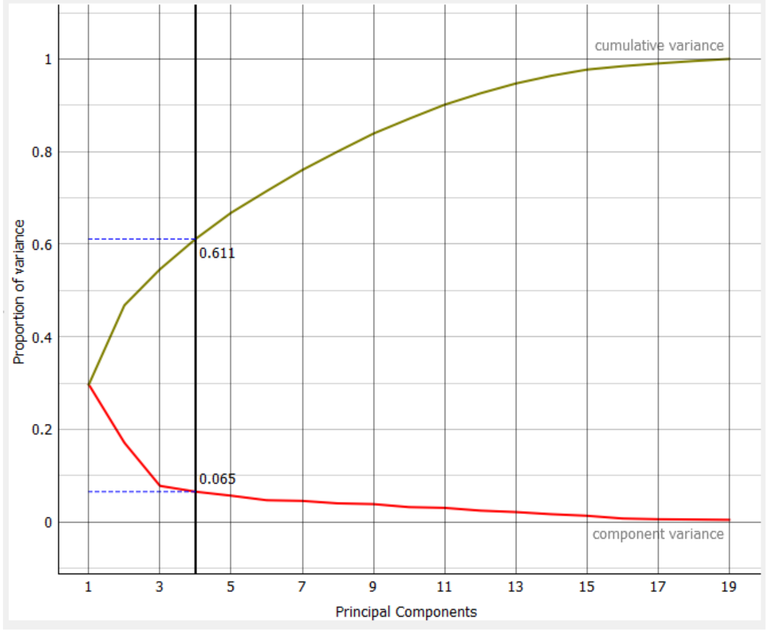
b) Visualize the dataset, using PCA and t-SNE.

t-SNE display

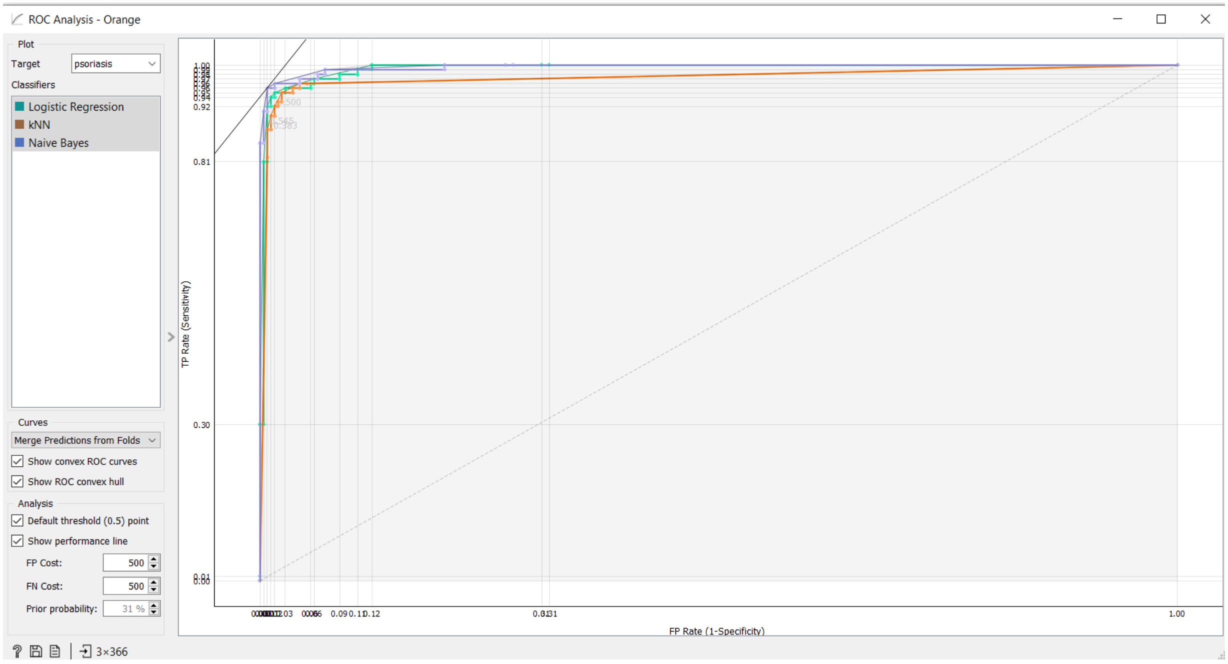


We can see that classes are not as discrete as they were before removing the best features of each class.

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.



ROC Analysis - Orange

Plot

Target: psoriasis

Classifiers

- 🟩 Logistic Regression
- 🟫 kNN
- 🟦 Naive Bayes

Curves

Merge Predictions from Folds

☑ Show convex ROC curves

☑ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

FP Cost: 500

FN Cost: 500

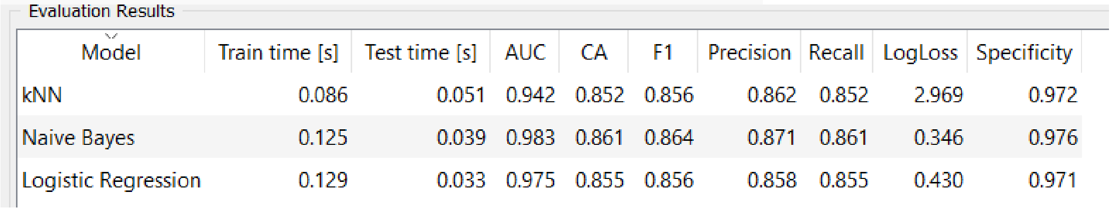
Prior probability: 31 %

3×366

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).

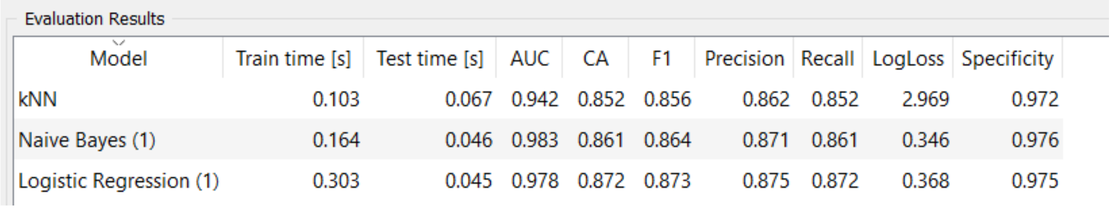
These are the results without the PCA transformation:



Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
kNN	0.086	0.051	0.942	0.852	0.856	0.862	0.852	2.969	0.972
Naive Bayes	0.125	0.039	0.983	0.861	0.864	0.871	0.861	0.346	0.976
Logistic Regression	0.129	0.033	0.975	0.855	0.856	0.858	0.855	0.430	0.971

These are the results after the PCA transformation:



Evaluation Results

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
kNN	0.103	0.067	0.942	0.852	0.856	0.862	0.852	2.969	0.972
Naive Bayes (1)	0.164	0.046	0.983	0.861	0.864	0.871	0.861	0.346	0.976
Logistic Regression (1)	0.303	0.045	0.978	0.872	0.873	0.875	0.872	0.368	0.975

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.