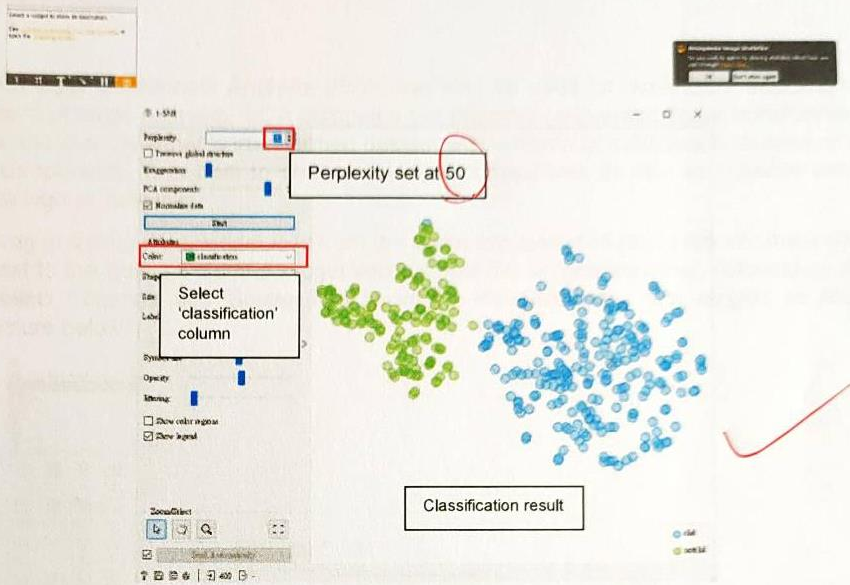
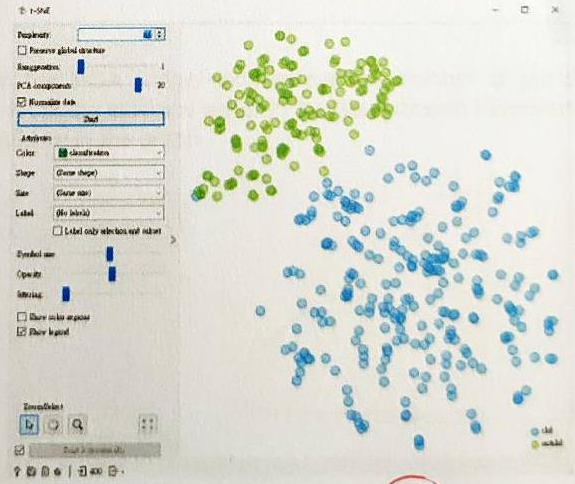
Chronic Kidney Disease data set from Kaggle, *N*= 400, *M* = 24. Blue = ckd, Green = not ckd



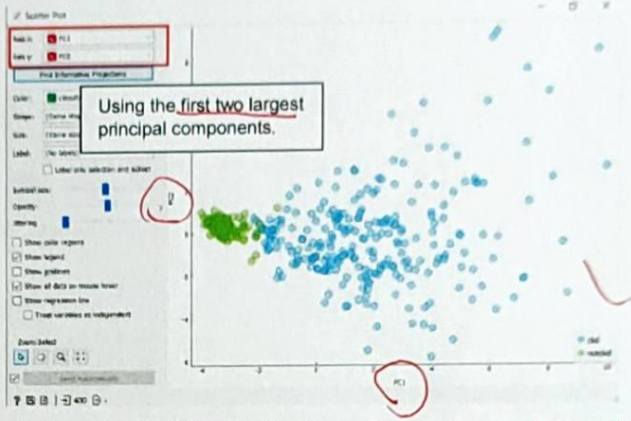
Important t-SNE Parameters for Plot Optimization

1) **Perplexity** can be interpreted as the number of nearest neighbors to distances will be preserved from each point. Using smaller values can reveal small, local clusters, while using large values tends to reveal the broader, global relationships between data points.

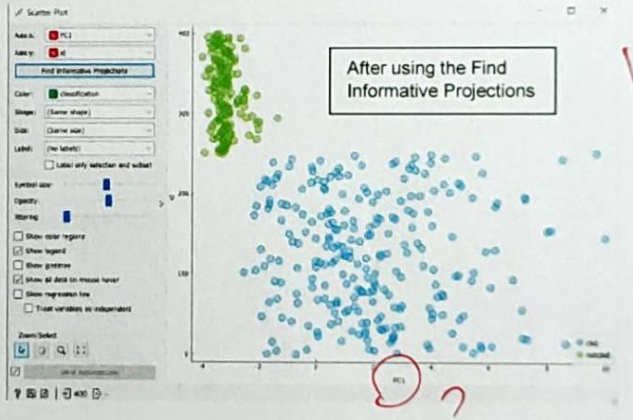


Example of Perplexity set at 70

12. Double-click on the 'Scatter Plot' widget icon to visualize the data with PCA.



we can also click on the Find Informative Projections button. It will show the score plots, which are a list of attribute pairs ordered by average classification accuracy score. Using this information, we can improve the visualization of the dataset with PCA. For example, we select the first rank pair of features in the score plots to visualize. The transformed data in the Scatter Plot show a clearer distinction between classes.



Mobile phone price range data set from Kaggle, *N*= ?, *M* = 21. Price range = 0 to 3

Correlations (1)

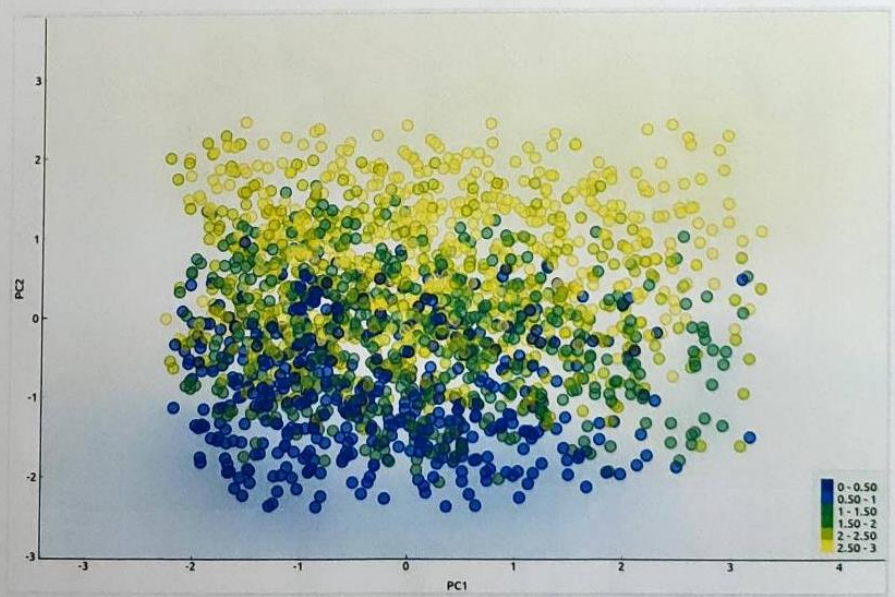
Pearson correlation

price_range

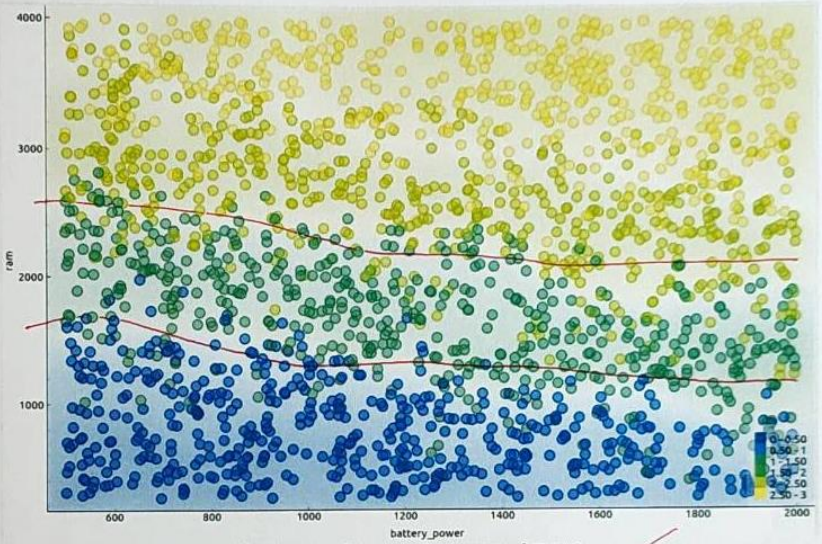
Filter ...

1	+0.917	price_range	ram
2	+0.201	battery_power	price_range
3	+0.166	price_range	px_width
4	+0.149	price_range	px_height
5	+0.044	int_memory	price_range
6	+0.039	price_range	sc_w

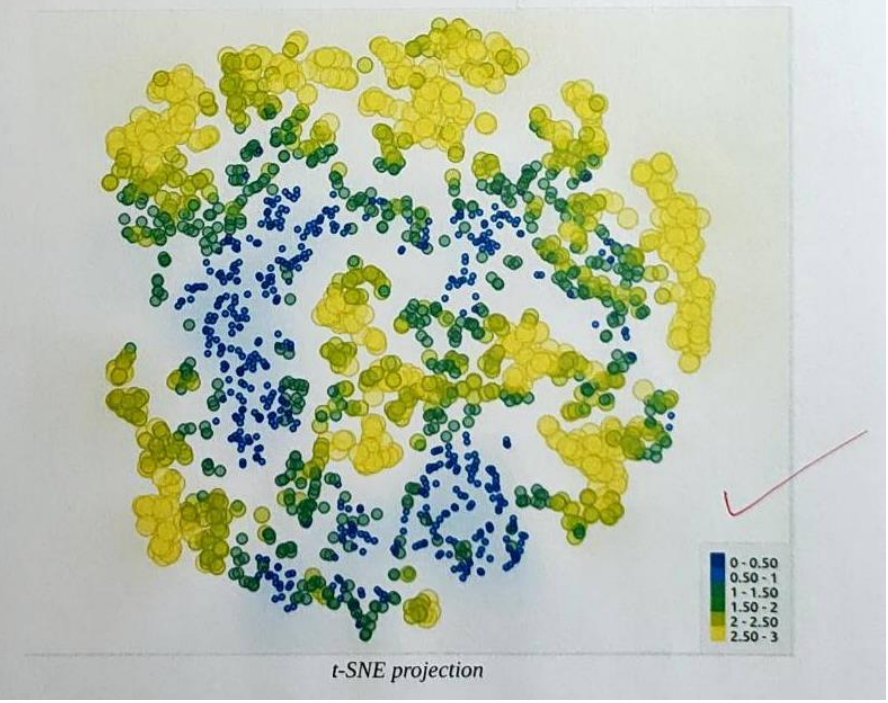
PCA didn't work well on this case, because of the high correlation between predictor and target, as can be seen in the next figures. The Principal Components projections don't show a clearly separation of the classes or targets, but projecting the data onto RAM and Battery Power did a better job, as all the classes can be separated by linear functions. The t-SNE projection also don't show any improvement overall, as all the clusters revolve in a spiral fashion; this could be due to the nature of the dataset.



PCA Projection for Mobile Phones Classification



Projecting data into battery and RAM



t-SNE projection

Glass type classification data set from Kaggle, *N*= ?, *M* = 10. type = 1 to 7

Pearson correlation

Type

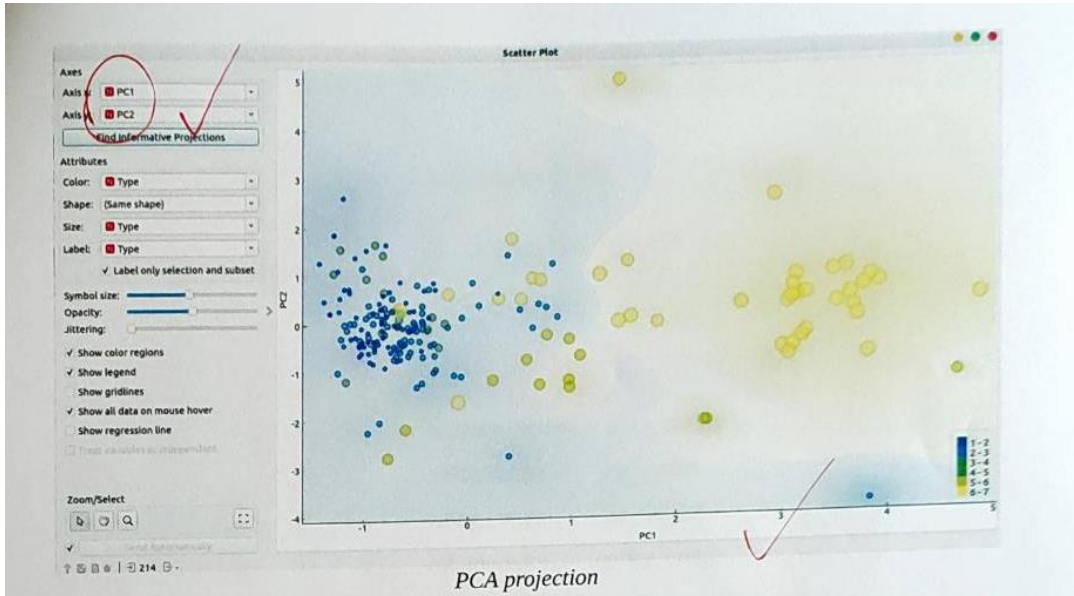
Filter ...

1	-0.745	Mg	Type
2	+0.599	Al	Type
3	+0.575	Ba	Type
4	+0.503	Na	Type
5	-0.188	Fe	Type
6	-0.164	RI	Type
7	+0.152	Si	Type
8	-0.010	K	Type
9	+0.001	Ca	Type

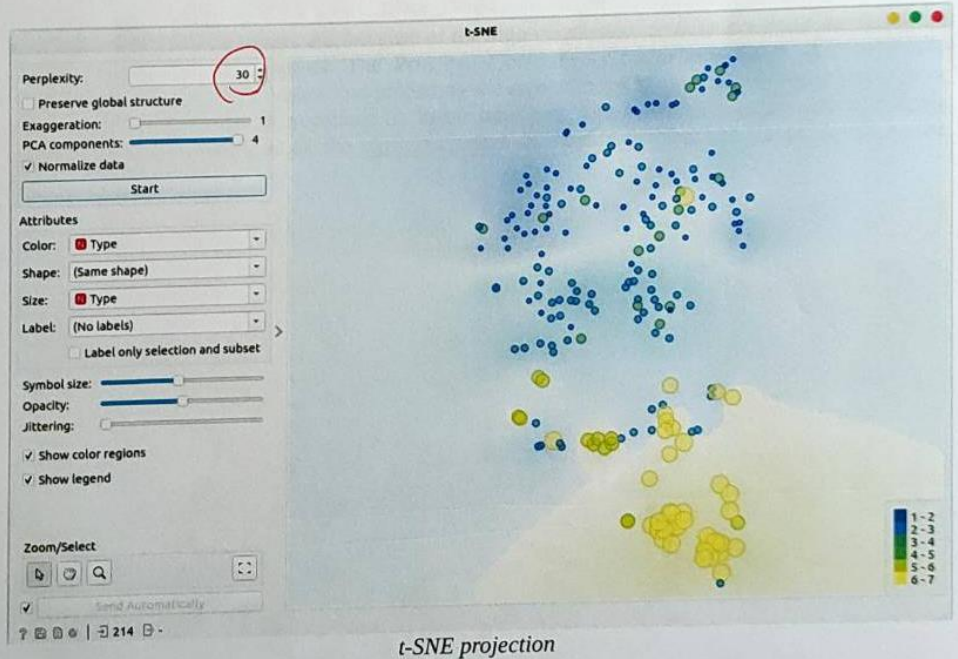
Finished

214 214

4. The correlation shows that the best estimator for the output is the Mg column, with a negative correlation of .745, followed by Aluminum and Barium. The features used as predictors where those above 0.5, either negative or positive. After correlation test, a module '*Select columns*' was added to filter in the desired columns.
5. For dimension reduction, two approaches can be used:
   1. PCA (Principal Component Analysis): It's based in the projection of the features to the components or axes that provide the most variance through eigenvectors. In other words, it decorrelates the data, in a linear way. The components of the transformation are orthogonal (dot product is zero, or they're perpendicular), and form a basis (a vector can be represented as the linear combination of a basis).
   2. t-SNE: t-distributed Stochastic Neighbor Embedding is a statistical method to visualize high dimensional data by projecting the data into a two or three dimensional map. It's non-linear, and it models each high-dimensional point in such a way that similar objects are modeled by nearby (or neighbor) points, and dissimilar objects are modeled by distant points with high probability. It requires the tuning of hyperparameters, such as perplexity (from information theory, which is nothing more than the exponentiation of the entropy or uncertainty of a random variable).



PCA projection

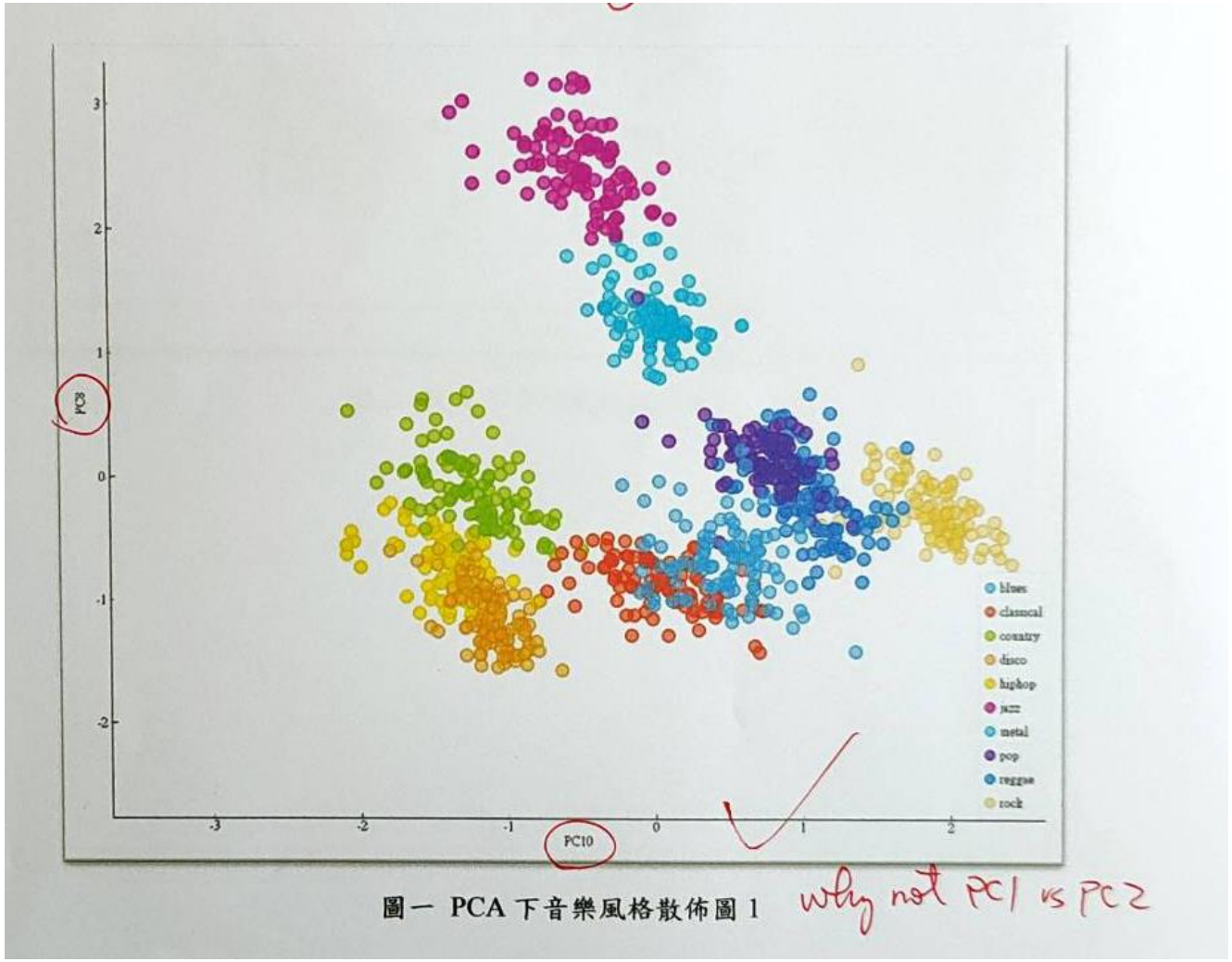


t-SNE projection

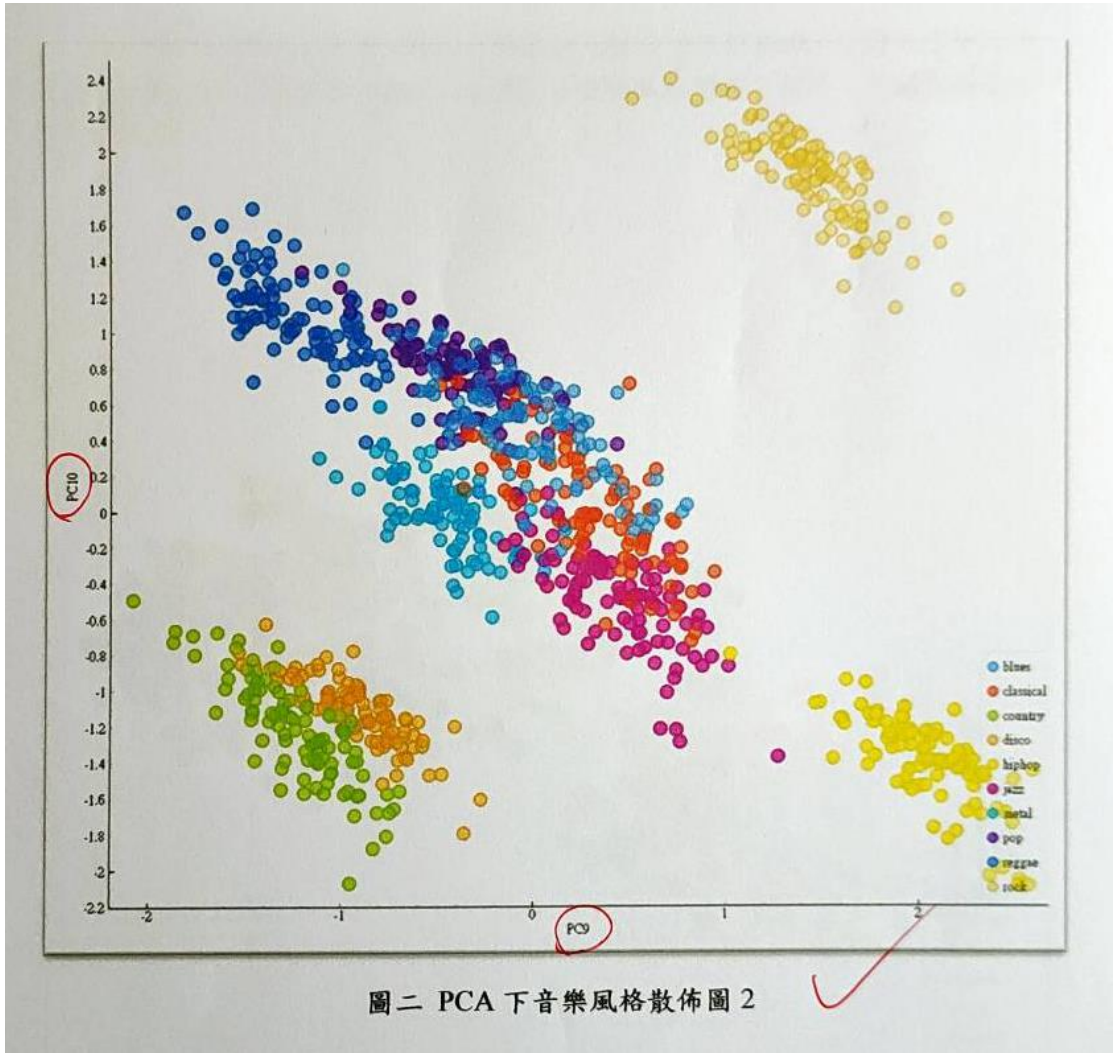
As can be noticed, with this approach, both projections show that most of the data can be classified only in three categories of glass, with most of the data shown in color blue and yellow.

Human activity recognition data set from Kaggle, *N*= 7352, *M* = 563 (?), Type = 6 activities (walk, walk upstairs, walk down stairs, laying, sit, stand)

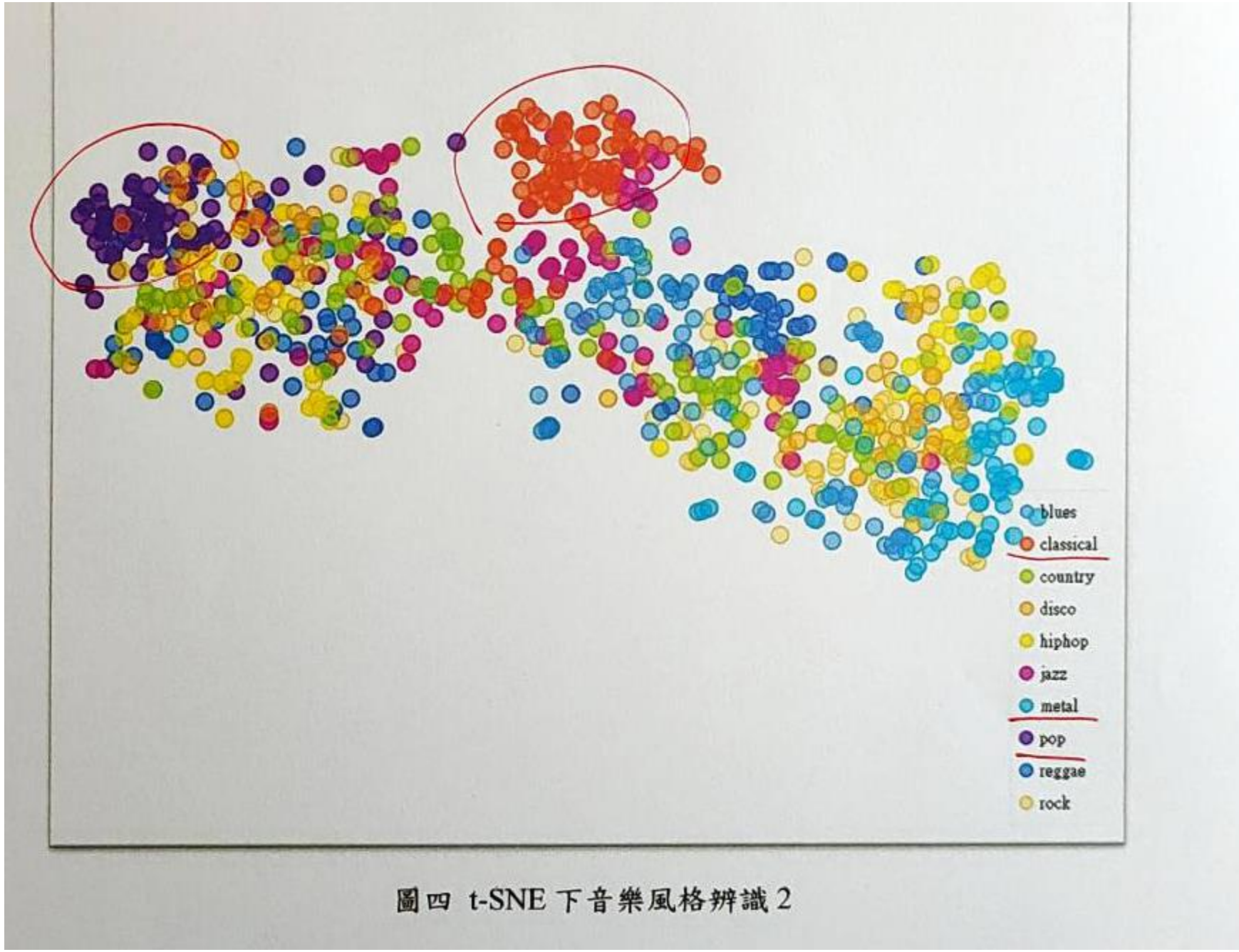
Music style data set from Github, *N*= 1000, *M* = 30 (?), type = 10 music styles
(blue, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock)



圖一 PCA 下音樂風格散佈圖 1



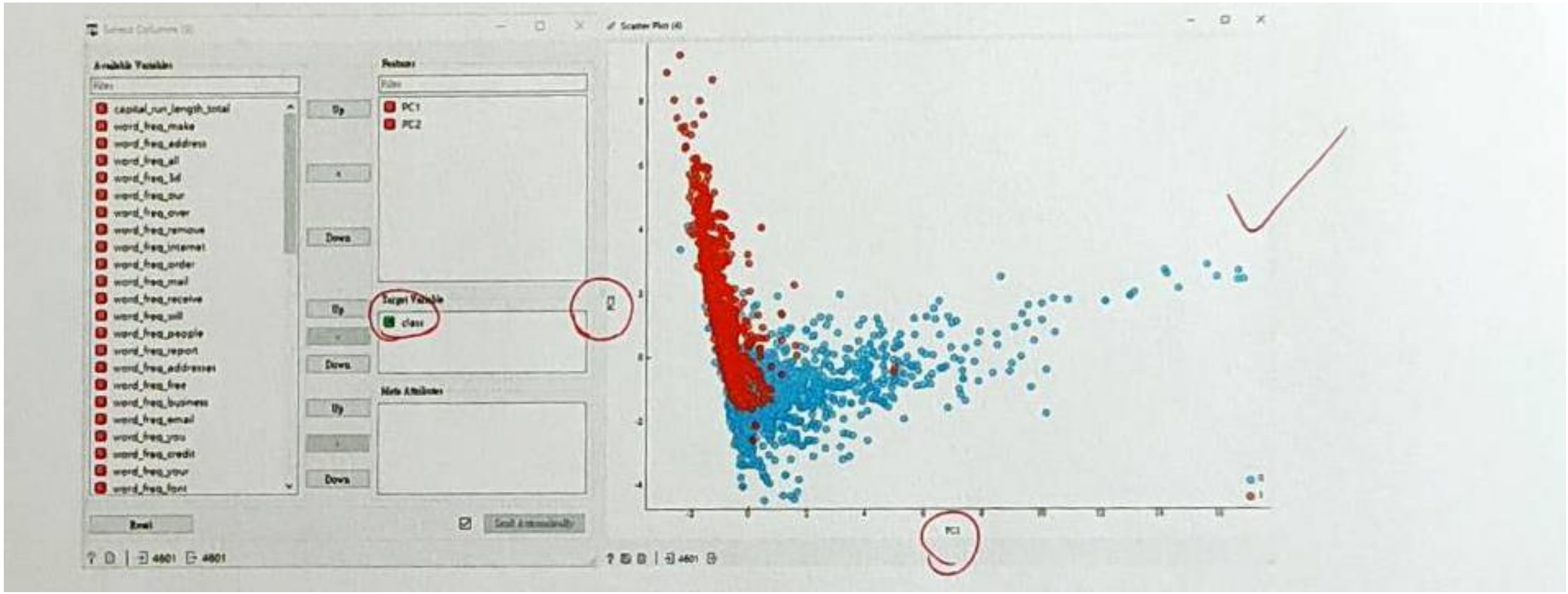
圖二 PCA 下音樂風格散佈圖 2



圖四 t-SNE 下音樂風格辨識 2

Abalone data set from Github, *N*= ?, *M* = 9 (1 categorical), type = age (category)

Spam email classification set from Kaggle, *N*= 4601, *M* = 58, type = 1 spam, 0 not



Customer transaction data set from ?, *N*= ?, *M* = 11?, type = high/low credit risk