Pre-processing

- The range of KLHL17 is 10.11507928; the range of HES4 is 10.812213.
- Yes, it's necessary to perform data standardization because the variance of the data is great.

PCA Analysis

- The number of features of this dataset is too great. PCA can help to identify the most principal columns.
- The two first principal components explain 0.0294+0.0131=0.0425=4.25% of the dataset
- Not so much, because the two first principal components does not explain a lot of the dataset.
- From the scatter we plot, we can see that different types are pretty much separated by those 2 principal features. However, the type in blue is not separated clearly. More features need to be added to get a better performance.
- We then randomly choose two features and draw another plot. We can see we can barely separated different types of cells. Therefore, PCA does its job to some extent.

Categorization using Gradient Descent

- As we can see, the error rate of the predict result is 0.327. Therefore, we may need a better model.
- We may use multilayer perceptron, which is better at handling high-dimensional and poor-linear data.