PCA and t-SNE comparison and visualization with SOM

Amanda Valtanen

November 2025

1 Introduction

The first task was to compare PCA and t-SNE methods using Bike Sharing Rental dataset and its visualizations. I used an MLP prediction model to predict the total number of rental bikes using the whole dataset and compared its performance with the MLP created using PCA and t-SNE on the dataset.

The second task was to visualize handwritten digits dataset with SOM. Based on the visualization, the method's capability to represent the data structure and its clusters was analyzed.

2 Data and Methodology

Rental Bikes dataset has 17379 samples and 14 features explaining the target variable. I decided to remove two features which represent the number of casual and registered rentals, as their sum directly describes the variable 'count' to be explained. I divided the data into training and test sets and performed standardization. After that, the, the PCA model was created using 6 principal components, which were decided based on the cumulative explained variance plot 1. With 6 principal components, we can explain about 70% of the data variance. I also created a t-SNE model using 2 components. An MLP model was first trained on the original dataset without dimensionality reduction to obtain the R^2 and mean squared error values. After that, PCA and t-SNE were applied as dimensionality reduction techniques, and new MLP models were trained to evaluate their corresponding R^2 and MSE results.

MNIST-784 handwritten digits dataset has 70000 samples and 785 features including the explained variable 'class'. Features present the pixel values and each sample is a handwritten digit. In order to use SOM, the data must be scaled. Because SOM is an unsupervised method, the whole (scaled) dataset is used to train a model. I plotted the U-matrix, which shows the average distances within the neuron weight vectors of the SOM. I added into that figure also the winning digits. I used 1000 iterations for the SOM model.

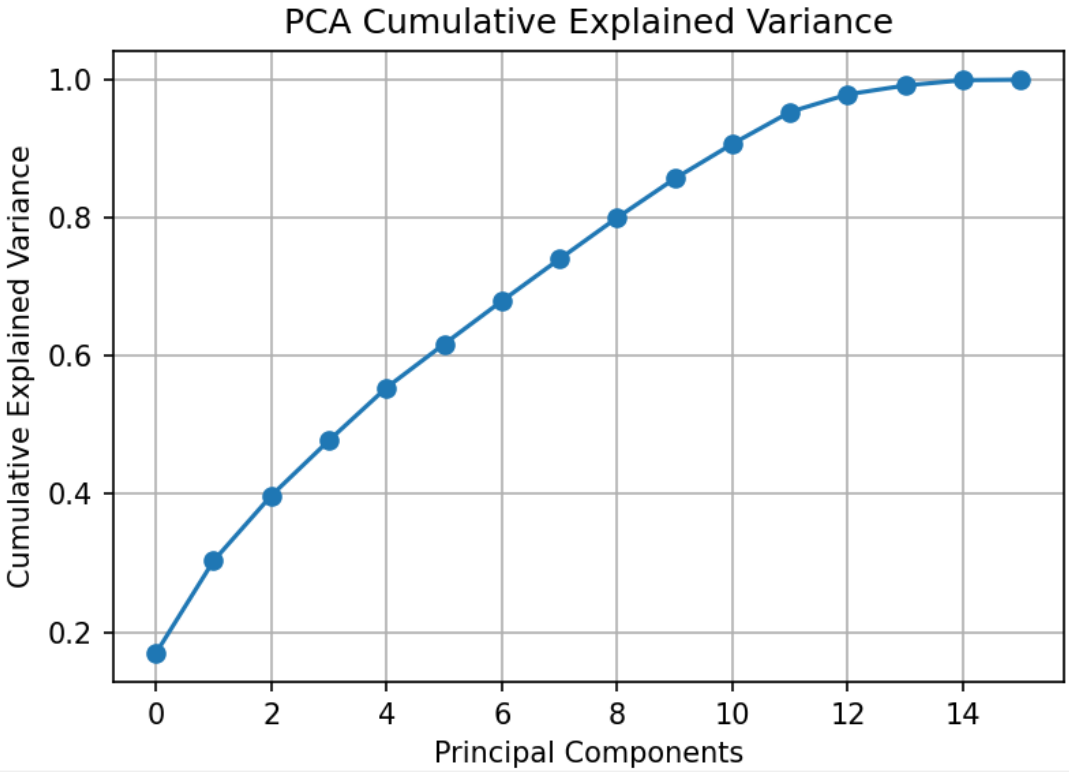

Figure 1: Cumulative explained variance for PCA model.

3 Results

The figure 2 illustrates the results of dimensionality reduction using PCA and t-SNE. In the PCA biplot (left), the first two principal components explain approximately 30% of the total variance. The directions of the loadings indicate the variables that contribute most strongly to each component. For instance, temperature and feeling temperature (temp, feel_temp) are aligned with the first component, whereas seasonal and weather-related variables show stronger loadings toward the second component. The data form a continuous distribution rather than clearly separated clusters. The t-SNE visualization (right) shows a more complex, non-linear structure of the same dataset. Several compact clusters can be observed, which indicates that t-SNE is able to capture local similarities between data points that PCA cannot represent linearly. Overall, PCA provides a linear and interpretable overview of the main directions of variance, while t-SNE offers a more detailed picture of local structure and potential groupings within the data.

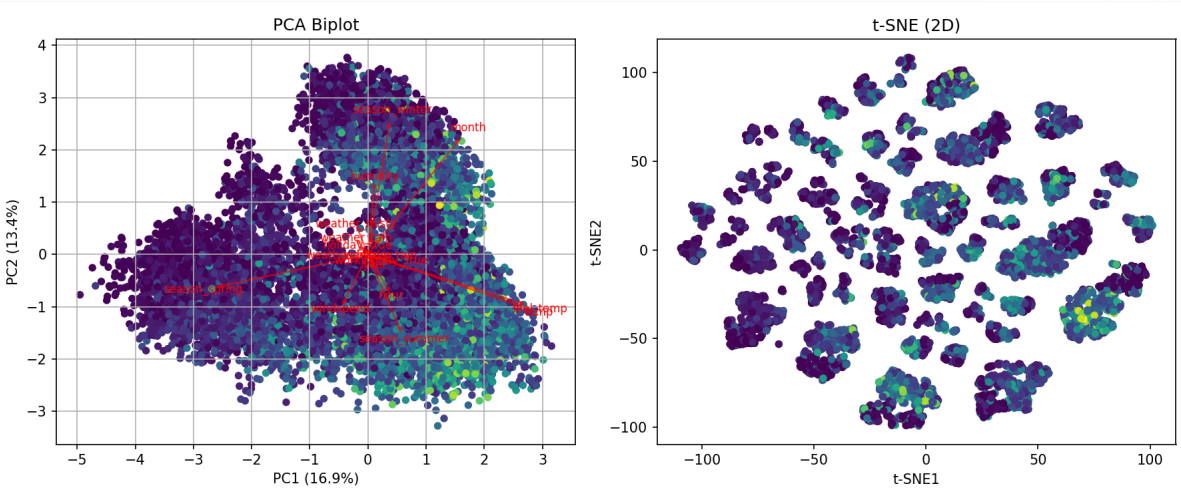

Figure 2: Comparison of PCA and t-SNE visualizations of the dataset.

When MLP was trained with the original dataset (without two removed columns), the model achieved an R^2 value of 0.7455 and an MSE of 8057.73, indicating a very good fit between the predicted and actual values. In contrast, applying dimensionality reduction significantly decreased the model's predictive accuracy. With PCA-reduced data (6 components), the R^2 value dropped to 0.4298 and the MSE increased to 18055.23, while using t-SNE resulted in an even lower R^2 of 0.1962 and an MSE of 25453.5. These results suggest that, although dimensionality reduction techniques can be useful for visualization and exploratory analysis, compressing the data into only two components leads to significant information loss and thus poorer regression performance for this dataset.

The figure 3 shows SOM U-Matrix visualization. In the figure, the colored dots represent the winning neurons for input samples in the dataset. Each dot is placed on the neuron that best matches a given digit. In the figure, the weight vectors of neurons in light gray areas are similar, meaning that they represent data points that are identical or very close to each other, while dark areas represent large distances between neurons.

The figure shows that the colorful dots and light colored areas form clusters. Similar numbers, such as 3 and 8 are close to each other on the map, while different numbers, such as 1 and 6, are far apart. SOM parameters and model randomness had a significant impact on the results, but visible clustering was still noticeable.

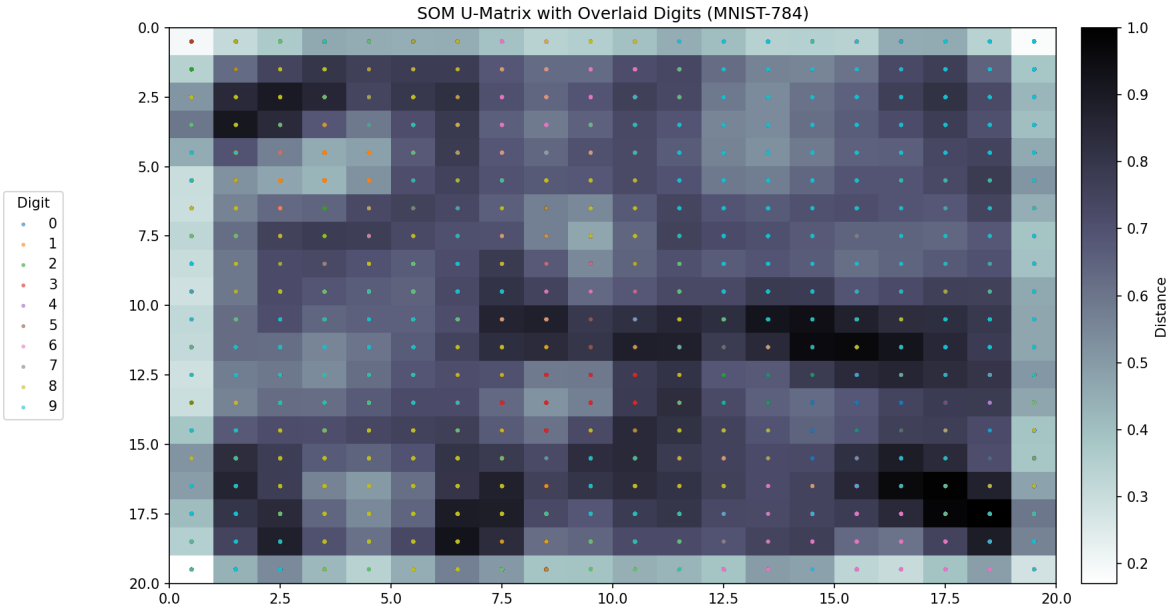

Figure 3: U-matrix of the SOM, colored points indicate the winning neurons.