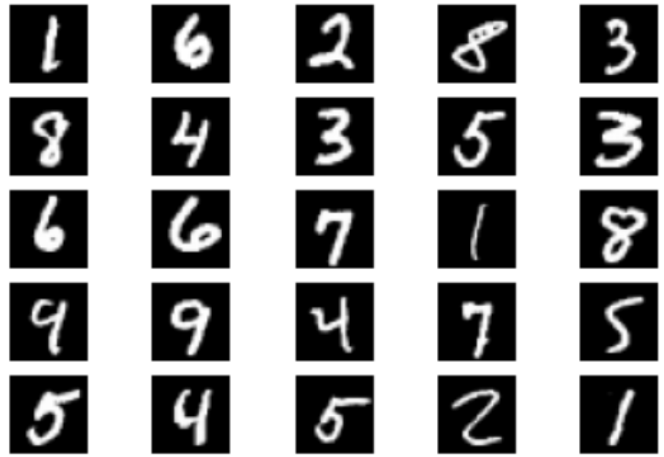
Implementation and evaluation of K-nearest neighbors (KNN) algorithm for handwritten digit recognition

Final Presentation – Data Science 2021

Project 5 Group 2

Nina Gutzeit, Maximilian Hingerl, Emma Kray, Johannes Müller

Recap of 4 Milestones



1. Milestone: implementing data normalization
2. Milestone: implementing PCA
3. Milestone: implementing a classification algorithm
4. Milestone: testing the algorithm

0. Data Cleaning

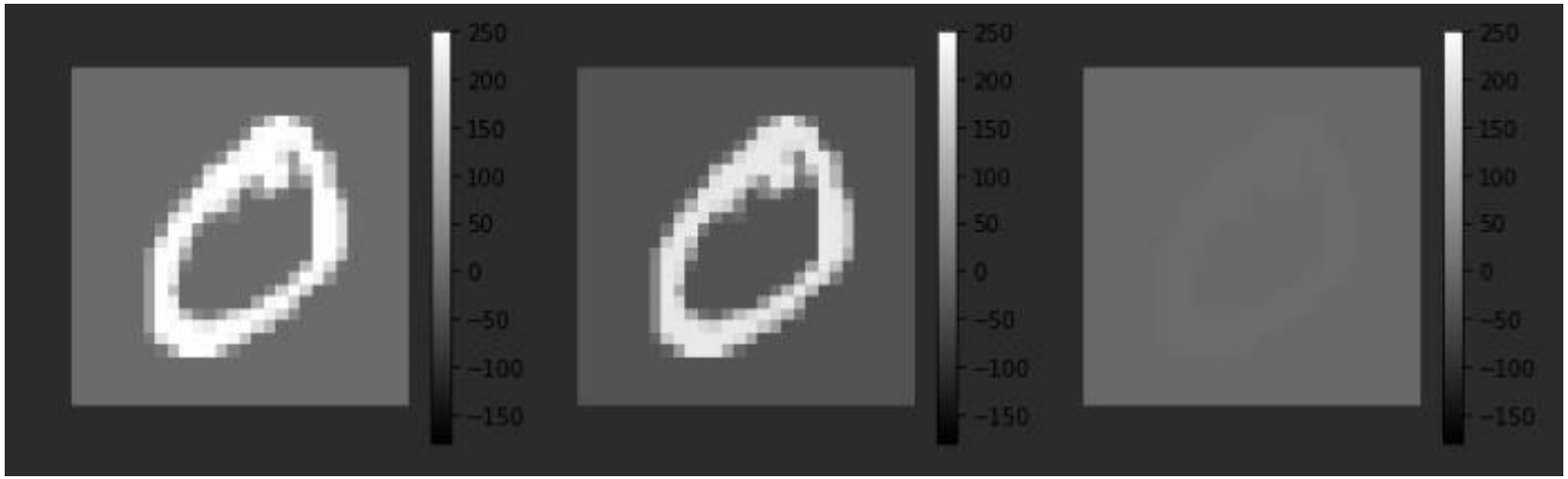
```
# Checking Training data for missing values:
any_na(train_values)
# Checking Test data for missing values:
any_na(test_values)
```

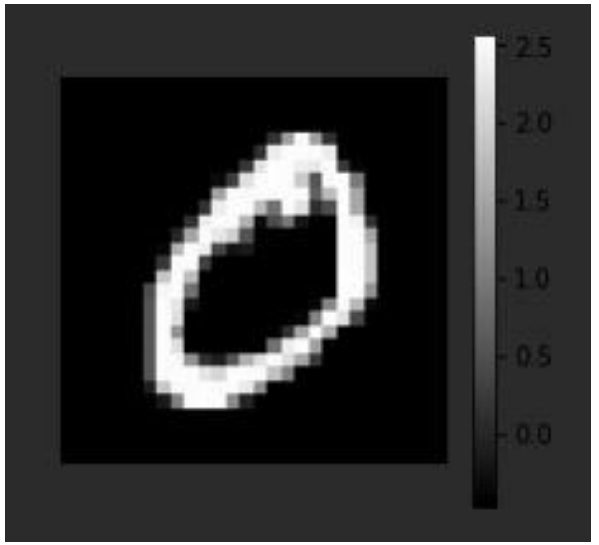
```
There are no missing values in this data.
There are no missing values in this data.
```

```
# Checking data for range
rm_range(train_values)
rm_range(train_labels, upper=10)
rm_range(test_values)
rm_range(test_labels, upper=10)
```

```
No values out of range.
No values out of range.
No values out of range.
No values out of range.
```

1. Standardization





Original image

Centered

Z- transformed

Z-transformed, other colour scale

$$(X_i - \overline{X})$$

$$\frac{(X_i - \overline{X})}{\sigma_i}$$

1. PCA preparation

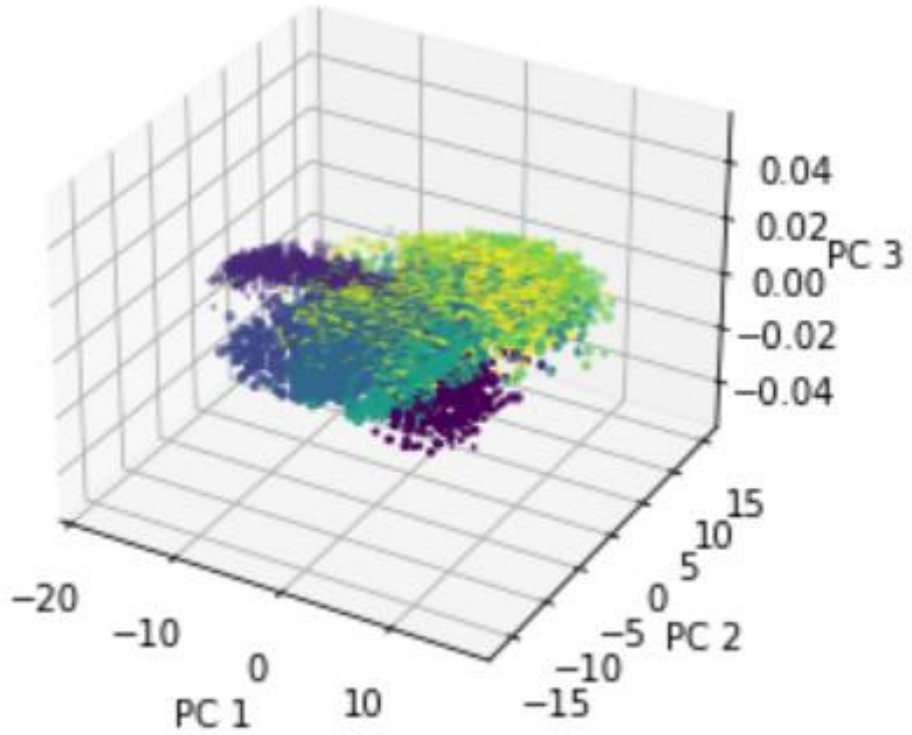
$$\operatorname{corr}(x,y) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \frac{\left(X_i - \overline{X}\right)}{\sigma_x} \frac{\left(Y_i - \overline{Y}\right)}{\sigma_y}$$

$$\operatorname{cov}(x,y) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

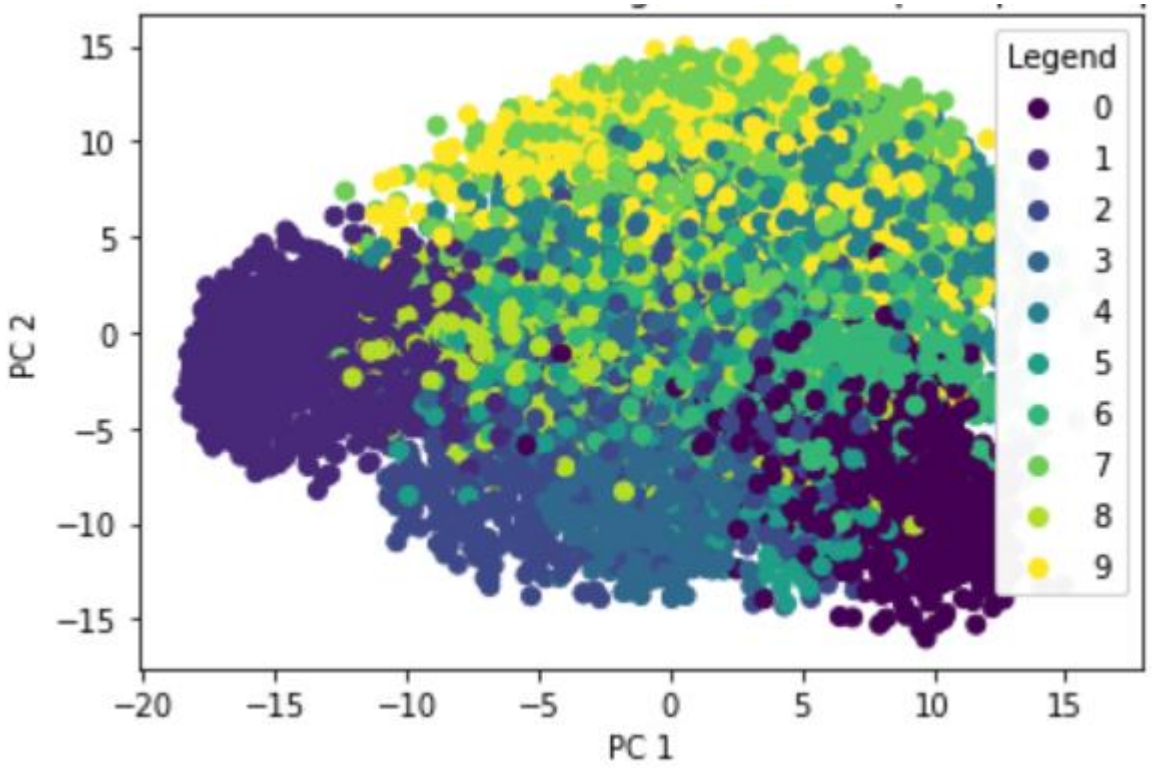
Challenges:
- for loops!
- cannot use numpy functions

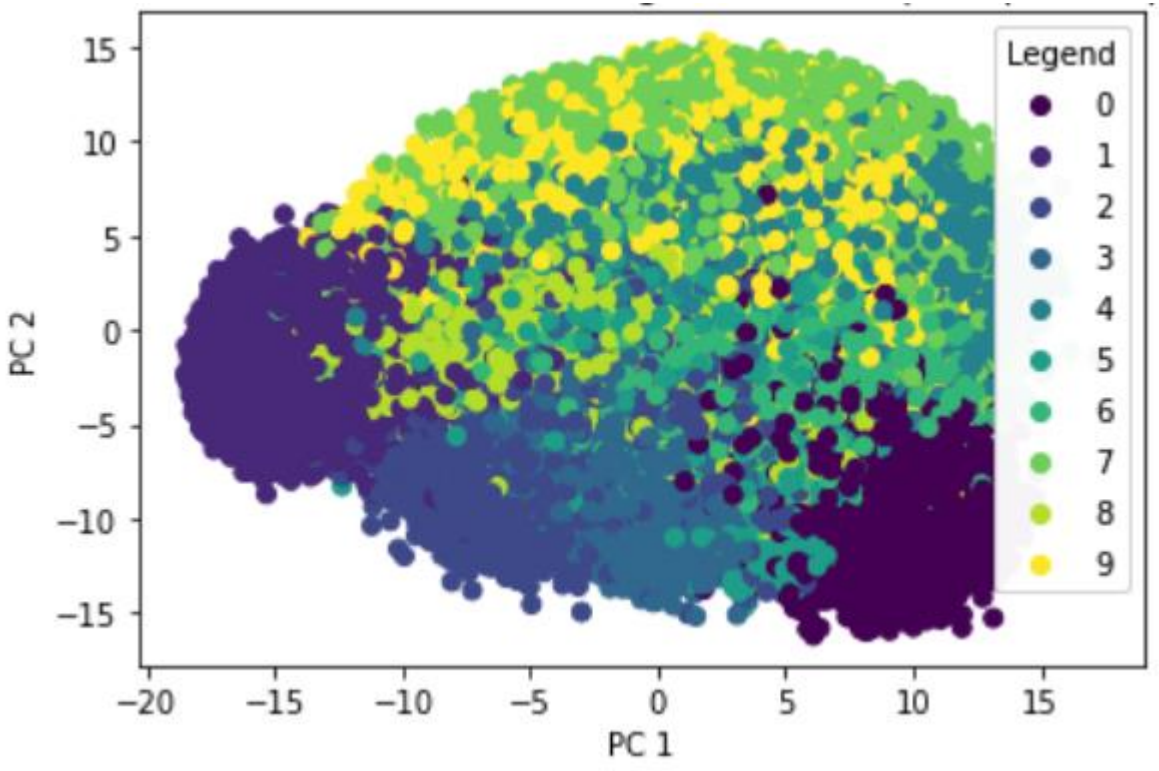
2. Principal Component Analysis

- Three best principal components
- 3D plot

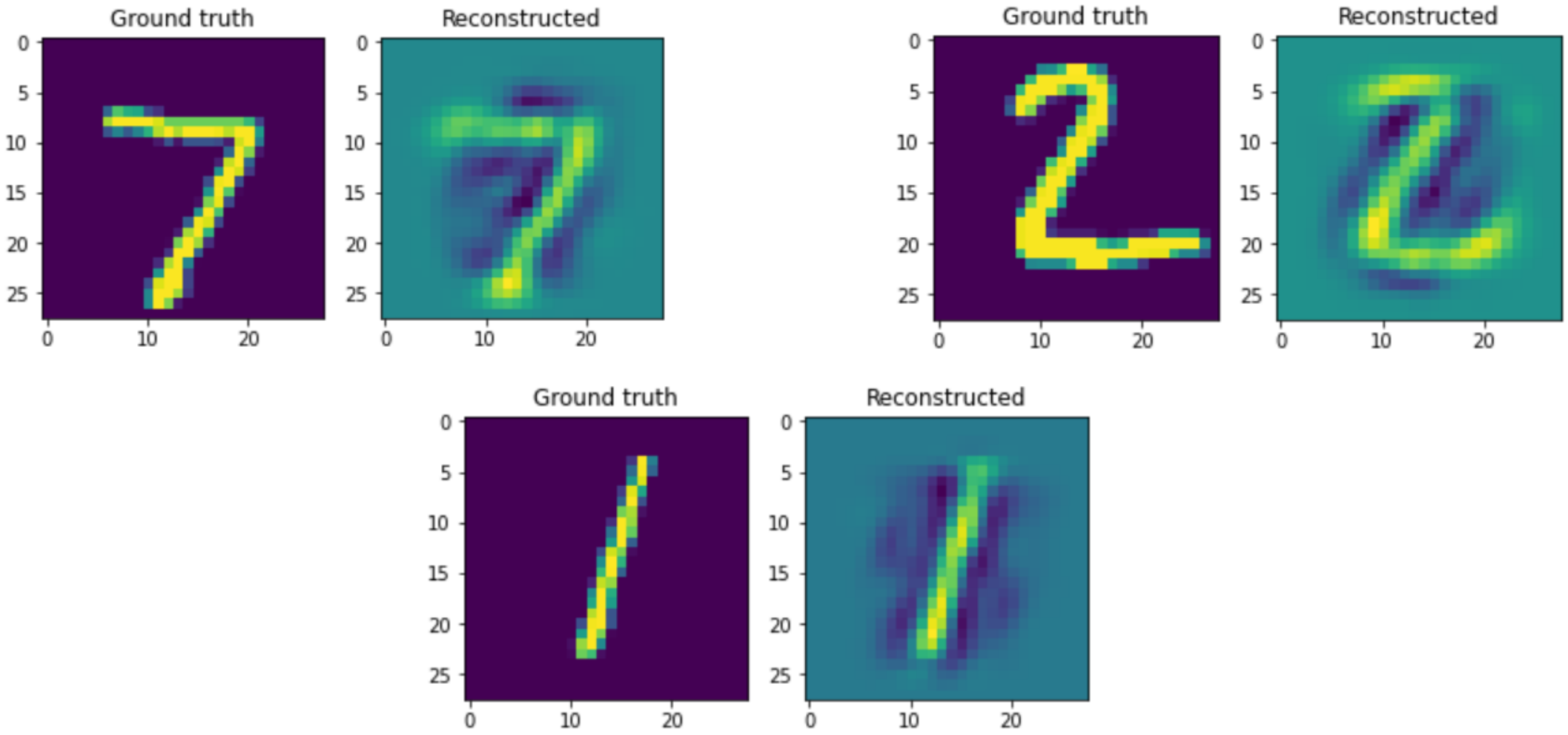


Comparison of test and training values

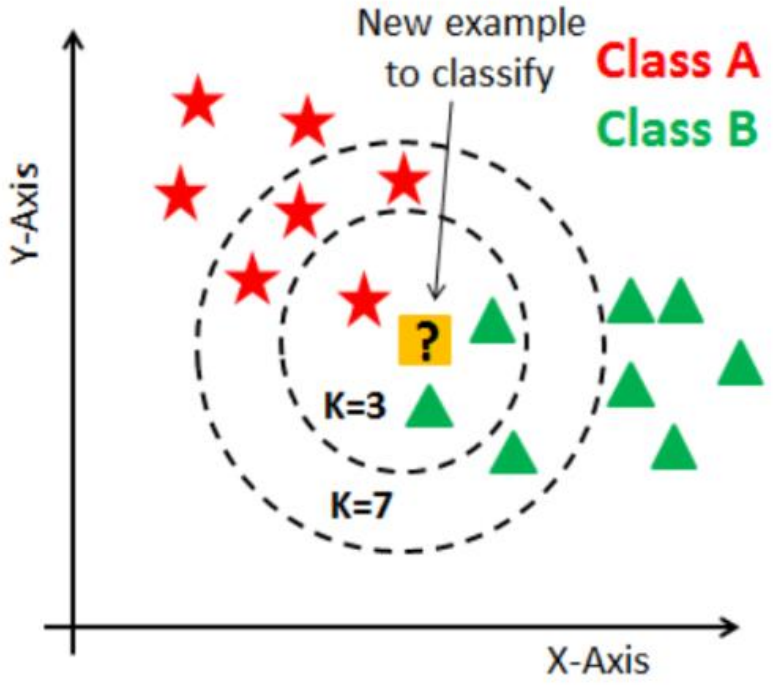




Comparison before and after PCA



3.1 KNN – the algorithm



3.2 KNN – distance methods

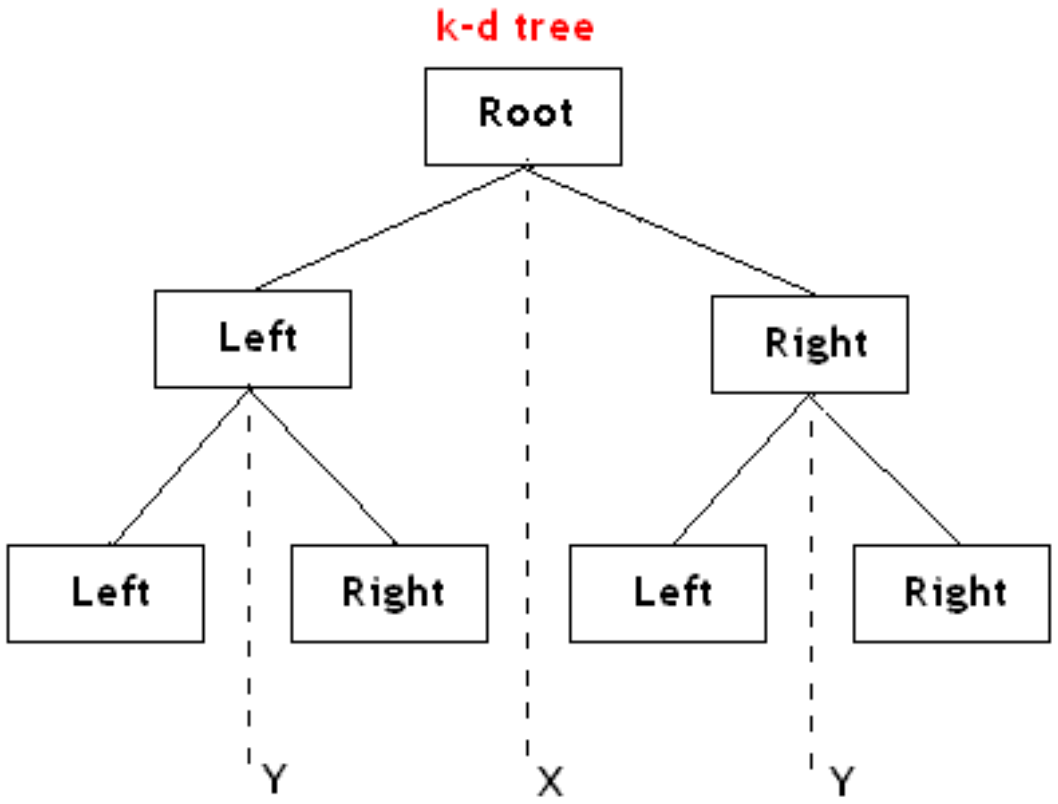
- Manhattan
- Euclidean
- Euclidean performed better
- Formulas:

```
correctly classified vs wrongly classified numbers using  the euclidean distance: 9803 vs 197
correctly classified vs wrongly classified numbers using the manhattan distance: 9786 vs 214
```

3.3 KNN – multiprocessing

- The concept
  - More threads → more processes at the same time
- Time improvement, but still to slow

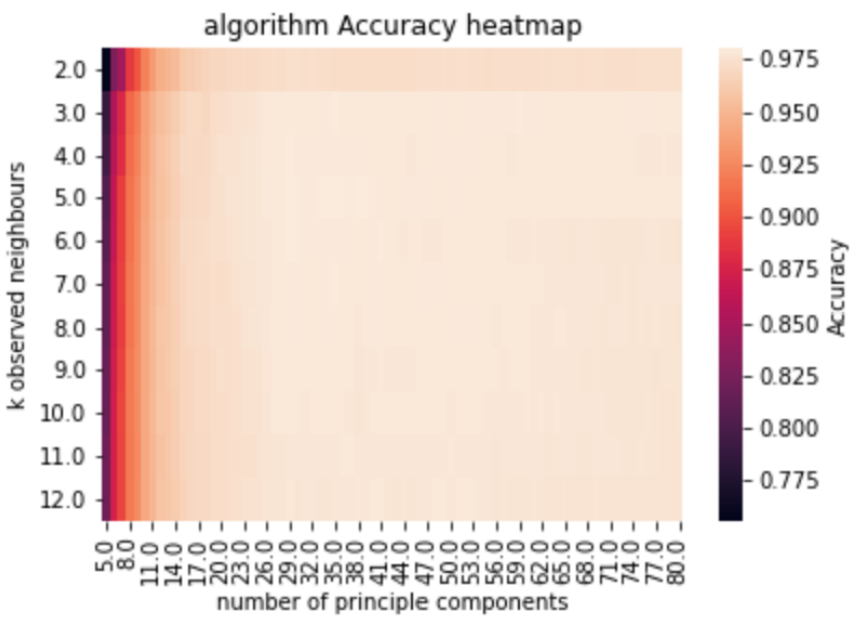
3.4 KNN – kd-trees

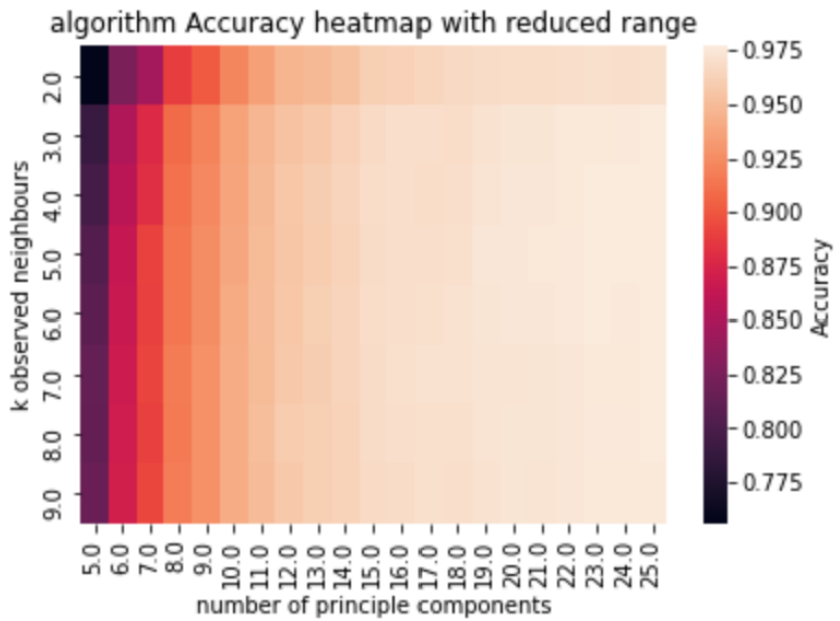


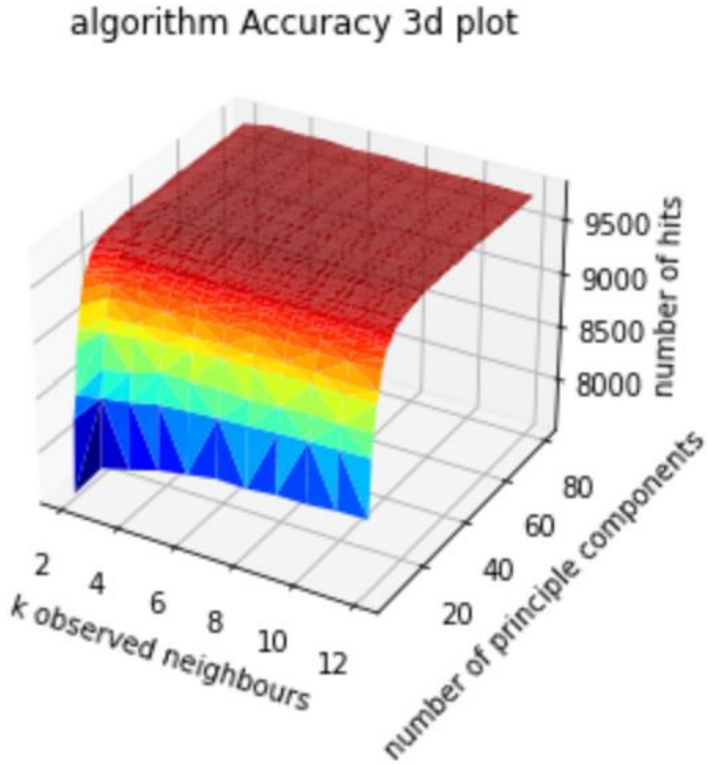
- Space partioning method
- Splitting training data along the median for each dimension

https://www.codeproject.com/Articles/18113/KD-Tree-Searching-in-N-dimensions-Part-I

4. testing



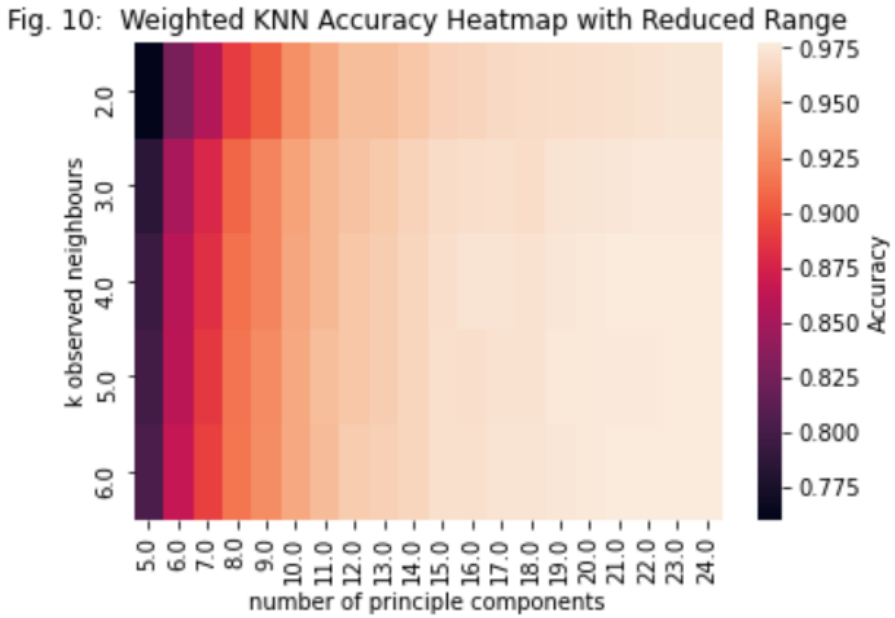




- Optimal number of k and pca:

Weighted knn

- Inverse weight of distance between labels
- Addition to normal KNN
- More accurate
- Much slower



Main project performance and different approaches

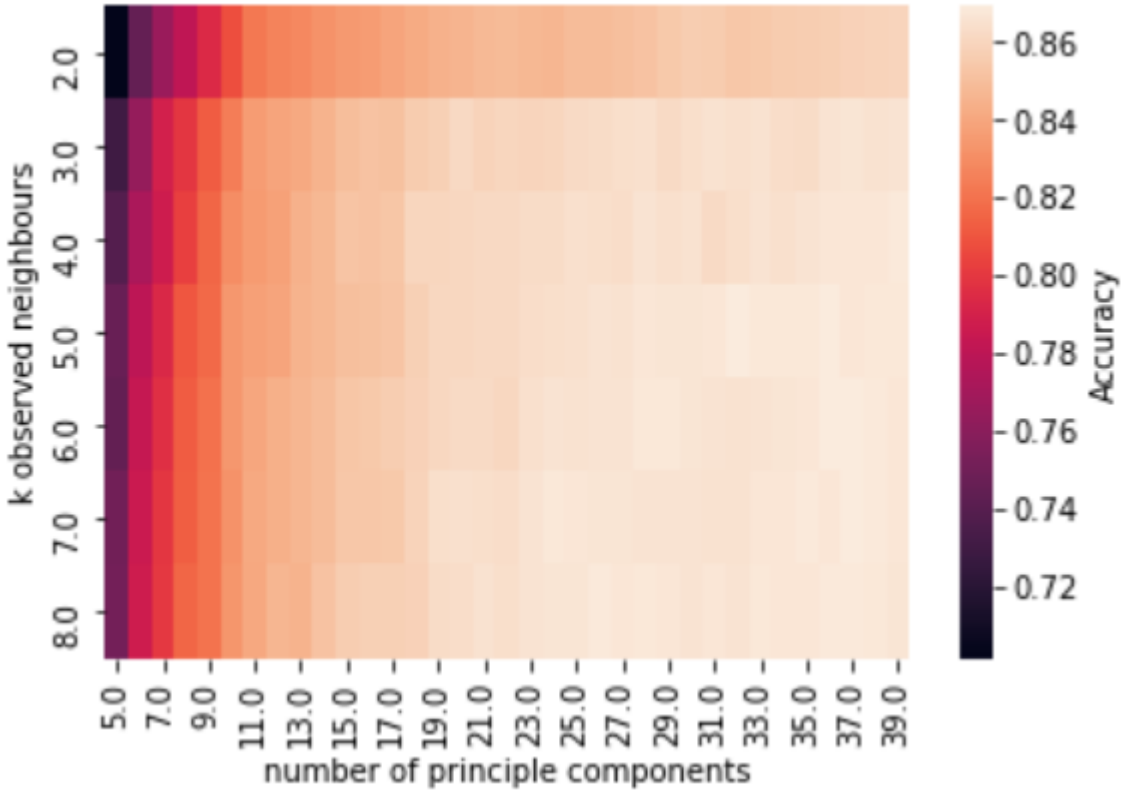
- Hier vor allem wunderschöne plots, die performance darstellen
- Auf nächster Folie oder so die anderen approaches

Works also for other data sets (very short insight in fashion mnist and performance)

- Requirement:
  - Same format 28x28

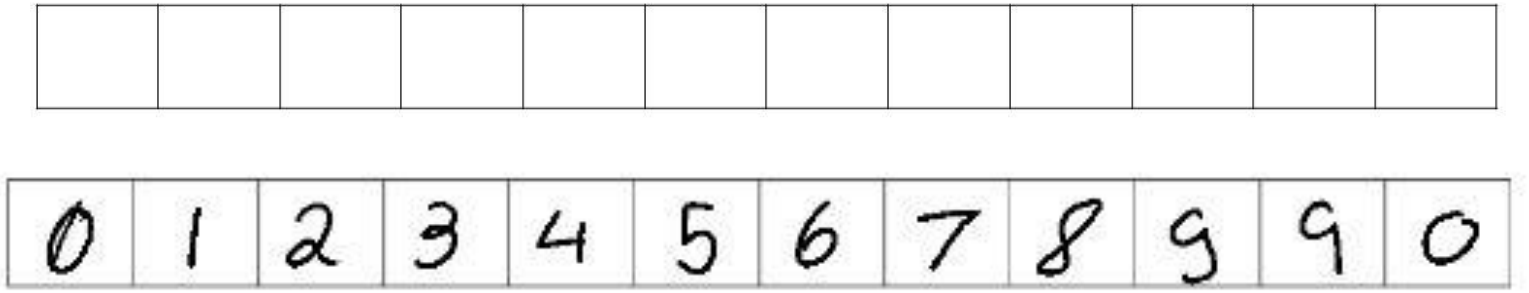
Fashion Mnist performance

Fig. 12: KNN Accuracy Heatmap with Reduced Range for Fashion Mnist



Phone numbers

Please write your phone number in **bold** legible handwriting. For better results write your number in the centre of each box using a digital ballpoint pen. Write in the large without drawing beyond the border.



Phone numbers limitations

vs

- unique handwriting
- number variants: crossing 7s, capping 1s

→ Is 60000 images enough?

Our teamwork

Thank you for your attention!

- Beardmore, A. (2020, October 12). Uncovering the Environmental Impact of Cloud Computing. Earth.Org. https://earth.org/environmental-impact-of-cloud-computing/.
- Cook, G. (2012, April). How Clean is Your Cloud? greenpeace.org. https://www.greenpeace.org/static/planet4-international-stateless/2012/04/e7c8ff21-howcleanisyourcloud.pdf.s
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). K-nearest neighbors. In An introduction to statistical learning: with applications in R (p. 163). essay, Springer.
- Lucivero F. (2020). Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. Science and engineering ethics, 26(2), 1009–1030. https://doi.org/10.1007/s11948-019-00171-7 (Lucivero, 2020)
- Sattiraju, N. (2020, April 2). Secret Cost of Google's Data Centers: Billions of Gallons of Water. Time. https://time.com/5814276/google-data-centers-water/. (Sattiraju, 2020)
- Walsh, B. (2014, April 2). New Greenpeace Report Shows the Environmental Impact of the Internet. Time. https://time.com/46777/your-data-is-dirty-the-carbon-price-of-cloud-computing/. (Walsh, 2014)