Compulsory Assignment #1

Machine Learning and Deep Learning [KAN-CDSCO2004U]

Examiner:	Somnath Mazumdar
Period	22-02-2024 until 07-03-2024
Students:	Baier, Maurice (167274) Reicheneder, Konstantin Johannes (167304) Reinecke, Nickolas Christian (167574) Tiedchen, Hans Jonathan (167323)
Submission date:	07-03-2024

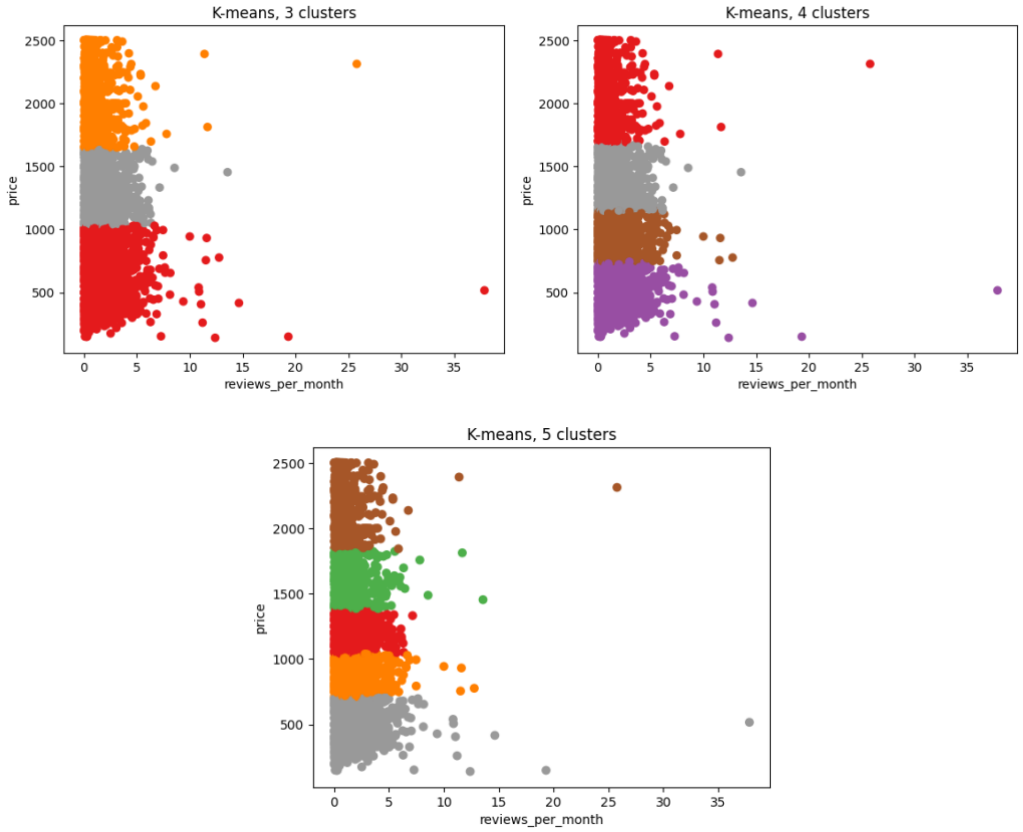
Question 1

1.1

When performing the EDA, it was most significant that the data includes columns containing only empty values. Those were dropped. Then, as we decided to cluster for price and reviews per months in exercise 1.2 we removed rows with missing values in these columns. Moreover, it was observable that the price column contained an outlier which caused skewed results when clustering, as this outlier was detected as an own cluster. Thus, we used the IQR method to remove outliers from the price column.

1.2

We decided to cluster the price and reviews per month columns. By doing that you could possibly find cluster to identify segments in the market, e.g. premium or standard listings with a lot of reviews and thus high demand. We decided for the K-means algorithm as it is easy to implement and efficient. We decided for 3 clusters to ensure interpretable results. However, we changed the cluster size for experimentation and due to the nature of the data adding a cluster only means that there is one more horizontally group is added in the graph.

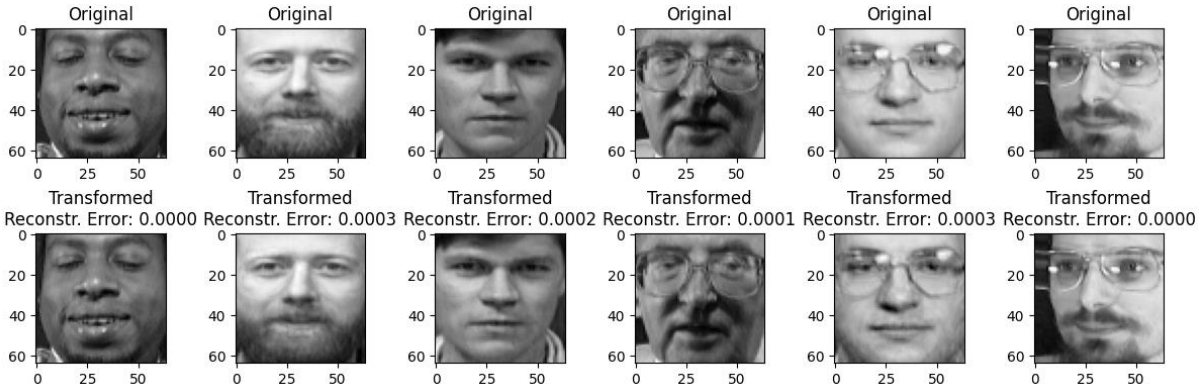


Possibly the outlier detection had such a strong effect on the data, that it is now too evenly distributed among the price category, or the data itself is not perfectly suited for clustering without any further processing.

Question 2

1.1

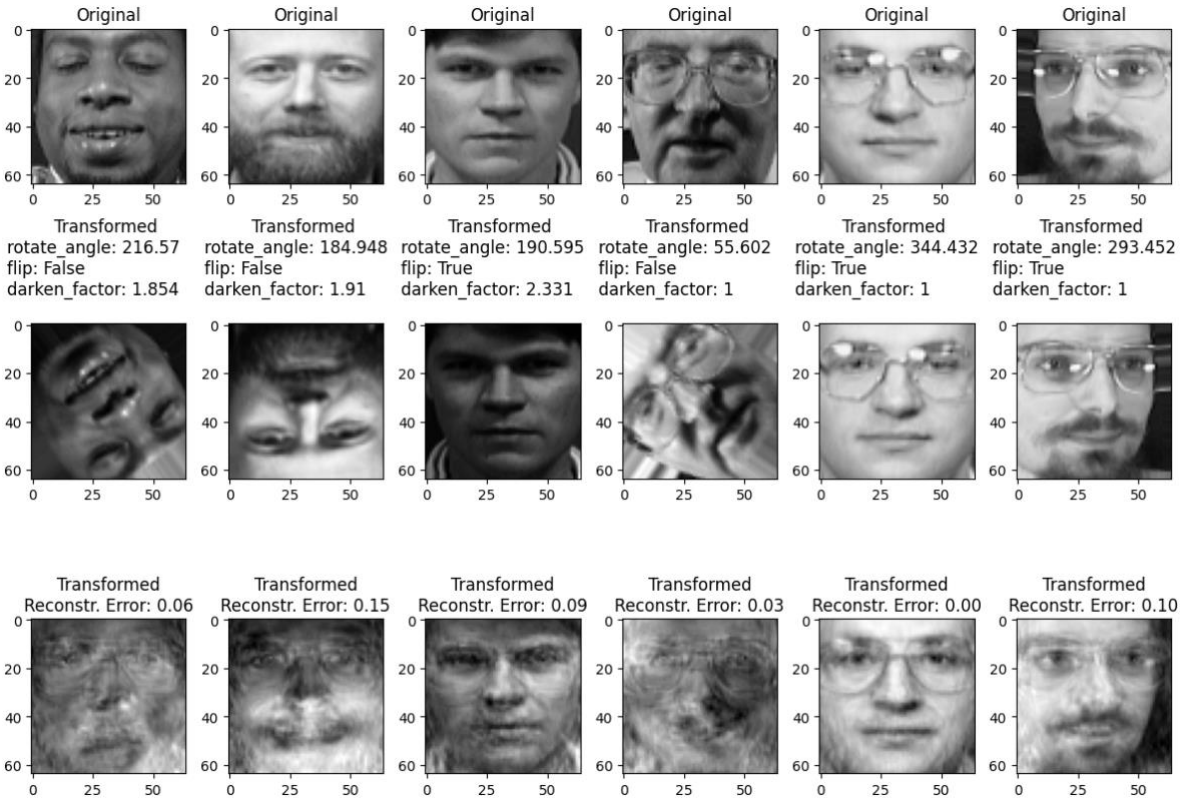
While preserving 99% of the variance those 6 random pictures were reconstructed by using 260 components:



(As pictures are randomly selected each time the code is executed, they might deviate from the pictures from the pre-executed Notebook.)

1.2

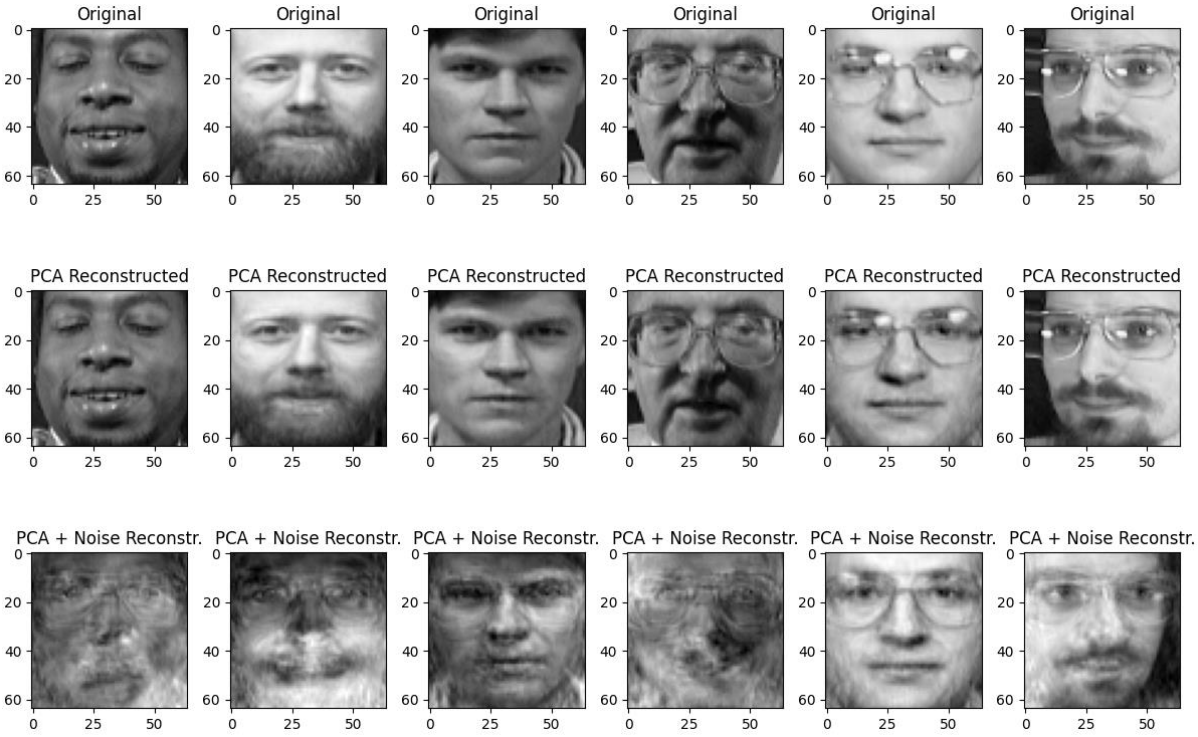
On the same pictures, the three transformations: rotate, flipping and darkening were randomly applied with random magnitude:



Compared to 1.1 when the transformation was not applied, the reconstruction error were notably lower for every image.

1.3

The reconstructed images compared to the original look the following:



Observable is that some kind of transformation have a higher effect of the reconstruction output than others. Looking at the images 1, 2, 4 versus images 3, 5, 6 it can be seen that a high degree of rotation has a significant effect on the ability to reconstruct the image.

Question 3

Data Preprocessing

Machine learning applications in all technology fields and applied in real-life problems have increasingly continued to diversify and increase exponentially (Maharana et al., 2022). In today's world which is characterized through big data and increasingly large datasets, data preprocessing techniques have become essential for knowledge discovery. Despite being less known than other steps in the data analysis process, data preprocessing is a crucial step that often involves significantly more effort and time than subsequent steps (Ramirez-Gallego et al., 2017). According to estimates data preprocessing can take up between 50% to 80% of the entire

time spent (Pyle, 1999; Kadhim, 2018). These estimates show the significance of this vital step in the knowledge discovery process.

Big data has led to the abundance of raw data at an ever-increasing pace which needs to be analyzed to extract their underlying value (Mayer-Schnberger & Cukier, 2013; Garcia et al., 2016). However, this raw data is highly vulnerable to missing data, noise, outliers, and inconsistency because of their huge size, multiple resources, and their gathering methods (Alasadi & Bhaya, 2017; Ramirez-Gallego et al., 2017). The goal of data preprocessing is to reduce the complexity in these datasets so that they can be easily processed by data mining applications for knowledge discovery.

Maharana et al. (2022) cluster the problems with this raw data in three categories: too much data, too little data, and fractured data.

Too much data can become a problem due to irrelevant data or noisy data in the dataset. Substantial data sizes can also take up a lot of computation space and therefore significantly increase processing time (Beynon et al., 2001; Donoho, 2000).

If the available data includes an insufficient amount of data of all kinds, the reliability of the knowledge gained from the data may be questionable and possible cannot be used for generating data insights (Graham, 2009).

Fractured data can become an issue when deriving data from several groups or different platforms. This incompatibility in goals, depth, and standard can lead to problems during the modeling (Maharana et al., 2022).

Data preprocessing techniques aims to address these problems. These techniques include data cleaning, integration, transformation, and reduction (Alasadi & Bhaya, 2017; Sutha & Tamilselvi, 2015; Garcia et al., 2016). Figure 1 displays these techniques for data preprocessing. These techniques also help to improve the data quality for better and more efficient performance of the adjacent knowledge discovery steps and are thus essential for efficient value creation of big data (Razavi et al., 2006). Ultimately, the decision which techniques to use always depends on the unique characteristics of the data available as well as its sources (Bolón-Canedo et al., 2015).

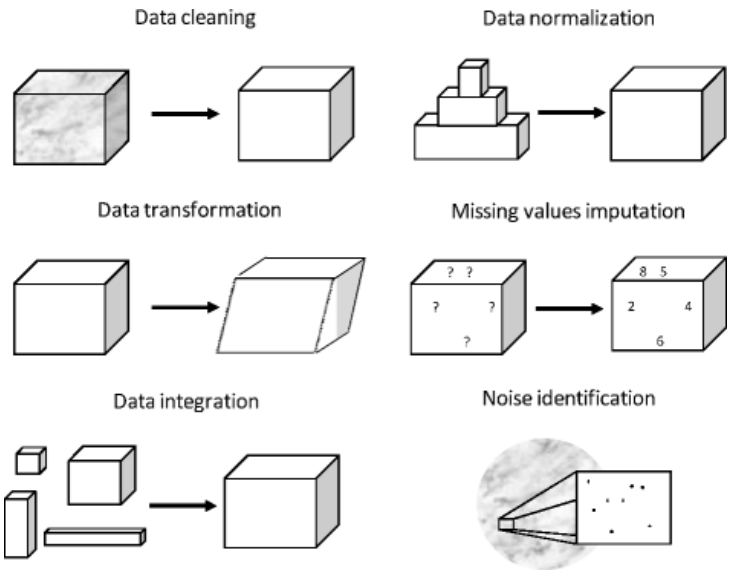


Figure 1 (Source: Garcia et al., 2016)

While research has clearly shown the necessity of data preprocessing in knowledge discovery (e.g., Ramirez-Gallego et al., 2017; Bolón-Canedo et al., 2015), there are also some limitations to data preprocessing which need to be carefully evaluated. These limitations include overfitting risks, loss of information, computational overhead, and domain specificity.

Overfitting is significant problem especially in machine learning. It occurs due to noise, the limited size of the training set, as well as the complexity of classifiers (Ying, 2019). Reducing this effect is crucial to guarantee a model's performance when dealing with real-world issues by feature selection (Rice et al., 2020; Ying, 2019).

Too much data cleaning while preprocessing data can also result in the loss of relevant information for the knowledge discovery thus generating an inaccurate image of the data available (Garcia et al., 2012).

Additionally, some preprocessing techniques can be computationally expensive, especially given the large amounts of data they need to handle. Assessing the computational cost of preprocessing steps and optimizing them is necessary to maintain efficiency in model training and deployment (López et al., 2012).

Lastly, preprocessing methods often need to be tailored to specific domains. Understanding the domain and characteristics of the data is crucial to choosing appropriate preprocessing techniques that align with the specific requirements and challenges of the given application (Isensee et al., 2018; Bolón-Canedo et al., 2015).

In conclusion, while data preprocessing plays a significant role in the knowledge discovery process of data, ensuring its effectiveness, its limitations must be carefully evaluated to strike

a balance between enhancing model performance and preserving the integrity and generalizability of the underlying data.

References

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

Beynon, M. D., Kurc, T., Catalyurek, U., Chang, C., Sussman, A., & Saltz, J. (2001). Distributed processing of very large datasets with DataCutter. Parallel Computing, 27(11), 1457-1478.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-based systems, 86, 33-45.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture, 1(2000), 32.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 1-22.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems, 25(1), 13-21.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Isensee, F., Jaeger, P. F., Full, P. M., Wolf, I., Engelhardt, S., & Maier-Hein, K. H. (2018). Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8 (pp. 120-129). Springer International Publishing.

Kadhim, A. I. (2018). An evaluation of preprocessing techniques for text classification. International Journal of Computer Science and Information Security (IJCSIS), 16(6), 22-32.

López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanc

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91-99.

Mayer-Schnberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Choice Rev. Online, 50, 50-6804.

Pyle, D. (1999). Data preparation for data mining. Morgan Kaufmann.

Razavi, A. R., Gill, H., Åhlfeldt, H., & Shahsavar, N. (2005). A data pre-processing method to increase efficiency and accuracy in data mining. In Artificial Intelligence in Medicine: 10th Conference on Artificial Intelligence in Medicine, AIME 2005, Aberdeen, UK, July 23-27, 2005. Proceedings 10 (pp. 434-443). Springer Berlin Heidelberg.

Rice, L., Wong, E., & Kolter, Z. (2020, November). Overfitting in adversarially robust deep learning. In International Conference on Machine Learning (pp. 8093-8104). PMLR.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 239, 39-57.

Razavi, A. R., Gill, H., Åhlfeldt, H., & Shahsavar, N. (2005). A data pre-processing method to increase efficiency and accuracy in data mining. In Artificial Intelligence in Medicine: 10th Conference on Artificial Intelligence in Medicine, AIME 2005, Aberdeen, UK, July 23-27, 2005. Proceedings 10 (pp. 434-443). Springer Berlin Heidelberg.

Sutha, K., & Tamilselvi, J. J. (2015). A review of feature selection algorithms for data mining techniques. International Journal on Computer Science and Engineering, 7(6), 63.

Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.