In recent years, Explainable Artificial Intelligence (XAI) has become a valuable addition to the toolbox of decision-makers, as it can provide helpful insights into real-world situations. However, explainable models, such as decision trees, typically require more training data to arrive at reliable results than black-box models, such as gradient boosting machines. The need for extensive datasets severely limits the applicability of some XAI methods when gathering data is a challenging and lengthy process, e.g., in medicine and social sciences.

One promising approach to obtain a well-performing explainable model from limited data makes use of dataset augmentation [1]. In the case of tabular data, dataset augmentation consists in training a generator on a dataset and then producing data that is plausibly sampled from the original distribution [2]. The new data is either labeled by the generator itself or by a black-box model trained on the original dataset. An explainable model is then trained on the augmented dataset. If any, the resulting performance gain is used as a proxy for augmentation quality.

Most existing generators can only generate numerical data. Hence in datasets with mixed attributes, e.g., containing both categorical and numerical attributes, categorical attributes have to be encoded, i.e., transformed into numerical attributes. However, no extensive comparison of augmentation pipelines in their complete form exists in the literature — below the arrows is an example of pipeline. This thesis aims to design and conduct a comparison study to inform the end-user about which augmentation technique best suits their XAI model of choice.

Tasks of this thesis include:

- Literature review on tabular data augmentation techniques with focus on mixed data
- Proposal of a workflow to benchmark complete augmentation pipelines
- Ranking of complete augmentation pipelines
- Analysis of the results

The student working on this assignment should possess:

- Working knowledge of English
- Confidentiality with Python or motivation to learn it
- Curiosity and ability to defend novel approaches

To help you in the research process, we offer thorough mentoring and regular meetings, access to our Institute’s computing infrastructure.

Throughout this work, you will acquire expertise in hot research areas such as Explainable Artificial Intelligence (XAI) and Generative Adversarial Networks (GANs), strengthen your Python coding skills and make a relevant contribution to the topic of data augmentation.
