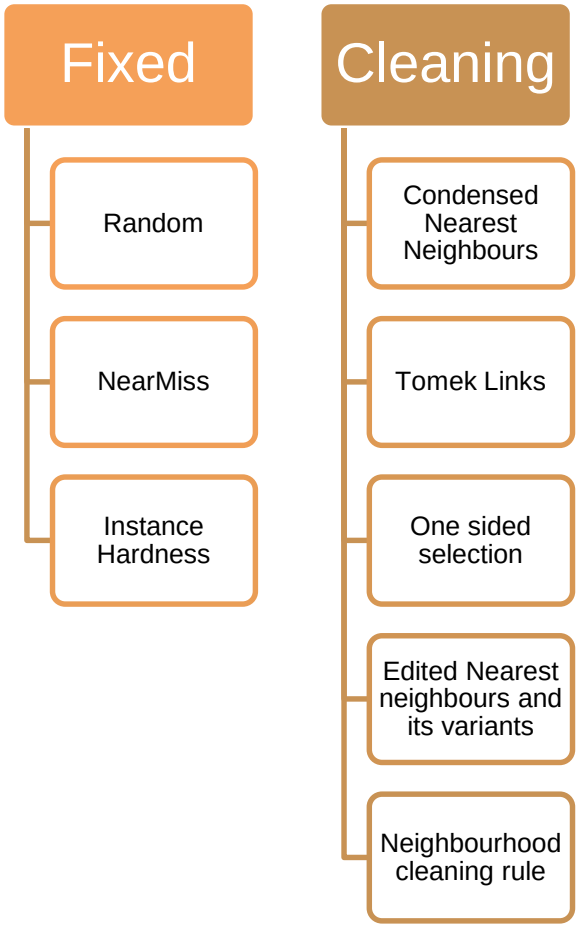

# Wrap-up
# Under-sampling

# In summary

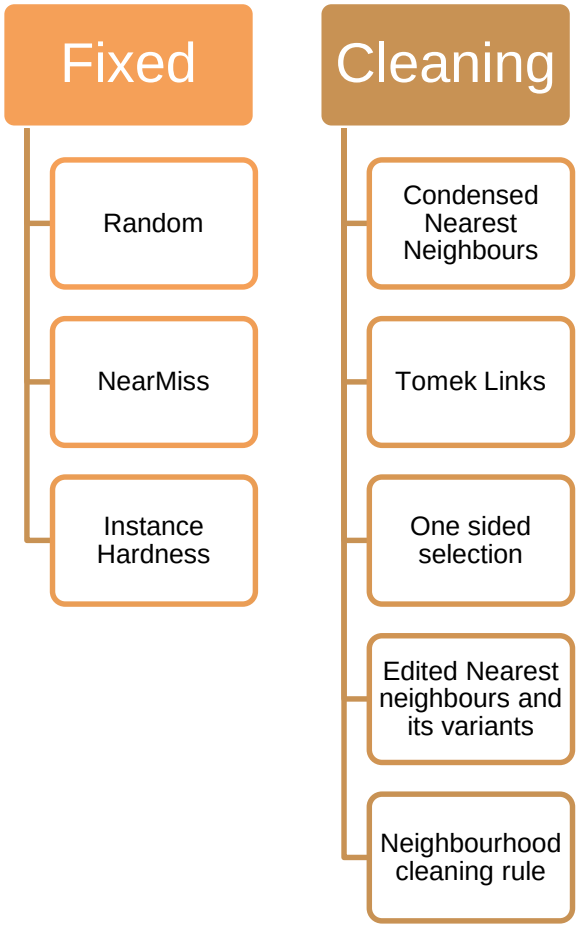
- There is no consensus in the community regarding which technique should be used with imbalanced datasets
- No rule of thumb on which technique should be applied on what type of dataset
- Trial and test

# Fixed vs Cleaning under- sampling



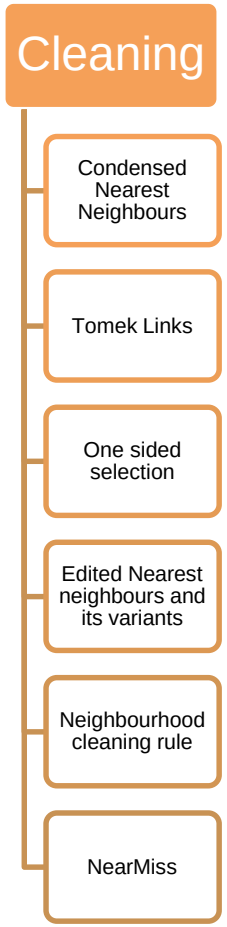

- Do I want a fixed size dataset?
- Do I want to reduce datasize a lot?

# Under- sampling categorical variables



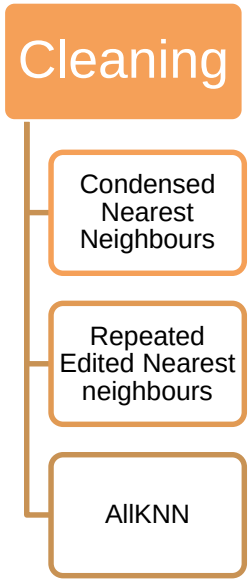

- Only **Random under-sampling** handles categorical variables out of the box
- For all the rest, we need to encode the variables first

# Cleaning methods rely on KNN




- KNN is distance based ➔ scale the variables
- For categorical and discrete variables the traditional distance metrics (i.e., Euclidean, Manhattan) are not suitable, consider using alternative metrics, or alternative under-sampling methods

# Big datasets and Cross-Validation




- Some Cleaning methods involve training several KNNs
- KNN algorithms do not scale well
- High training times if using cross-validation or very big datasets

Train In Data

# THANK YOU

www.trainindata.com