	Fixed vs Cleaning	Under-sampling criteria	Categorical variables	Distance based	Number of KNN
Random Undersampling	Fixed	Random	yes	No	-
Condensed Nearest Neighbours	Cleaning	● Train 1-KNN on minority class ● Predict class of observation in majority ● If class matches real class, take another observation and do the same ● If class does not match, put that observation together with the minority group and train another 1-KNN ● Proceed until all observations from majority were examined **● The final dataset consists of the minority class + all observations from the majority that were subsiquently miss-classified by the KNNs**	No	Yes	Several
Tomek Links	Cleaning	● If 2 observations from different class are closest neighbours to each other, they are Tomek Links ● Remove the observation from the majority in the Tomek Link, or the entire Tomek Link from the final data **● The final dataset is the original one minus the Tomek Links (or one of the observations in the Tomek Link)**	No	Yes	1
One Sided Selection	Cleaning	● Train 1-KNN on minority class ● Predict the class of ALL observation in majority group ● If class matches the real class, remove the observations from the dataset ● Finally, carry out Tomek Links on the remaining observations. **● The final dataset consists of the minority + all observations from the majority that were miss-classified minus the Tomek Links**	No	Yes	2
Edited Nearest Neighbours	Cleaning	● Find the 3 closest neighbours to each observation from the majority class ● If all or most neighbours belong to a different class, remove observation from the dataset **● The final dataset is the original minus all observations from the majority, for which their 3 closest neighbours disagree with its class**	No	Yes	1
Repeated ENN	Cleaning	● Repeats Edited Nearest Neighbours several times. ● Stops at after a number of iterations entered by user, or when no more observations are removed, whatever happens first.	No	Yes	Several
All KNN	Cleaning	● Repeats Edited Nearest Neighbours several times. ● In the first iteration always examines the 1 closest neighbour of each observation from the majority ● Increases the number of neighbours examined at each iteration by 1 ● Stops at the round corresponding to the number of neighbours determined by the user, or when one of the majority classes becomes the minority, whatever happens first	No	Yes	Several
Neighbourhood Cleaning Rule	Cleaning	**Step 1:** ● Find the 3 closest neighbours of each majority class observation ● If most neighbours disagree with the class, flag that observation for removal **Step 2:** ● Find the 3 closest neighbours for each observation of the minority class ● If the neighbours disagree, flag the neighbours for removal **● The final dataset is the original minus the observations from the majority that were flagged in steps 1 and 2**	No	Yes	1
Near Miss	Fixed	**Version 1:** ● Determine the mean distance of each observation from the majority to its **K closest** neighbours from the minority ● Retain the observations from the majority with the smallest average distance **Version 2:** ● Determine the mean distance of each observation from the majority to its **K furthest neighbours** from the minority ● Retain the observations from the majority with the smallest average distance **Version 3:** ● Find the 3 closest neighbours of each minority class that belong to the majority class ● Remove all observations from majority that are not a closest neighbour as above ● For the remaining observations of the majority, determine the average distance to its K closest neighbours from the minority ● Retain those obs from the majority for which the average distance is the largest	No	Yes	1 or 2
Instance Hardness Threshold	Fixed-ish	● Instance hardness measures how hard an observation is to classify correctly. ● For a sample of class 1, the instance hardness is 1 - p(1), where p is the probabilty. ● IH depends on the algorithm used for the classification ● This method removes observations where p of its class is below a certain threshold ● The threshold is determined automatically to achieve a certain number of observations in the final dataset **● The final dataset contains all observations from minority class + the observations from the majority with the highest probabilities of their class**	No	No	-