Train In Data

Neighbourhood Cleaning Rule

Neighbourhood Cleaning Rule (NCL)

Remove samples from the majority class that are closest to the boundary (with the other classes).

Expands on ENN, by "cleaning" examples from the majority class that are neighbours to the minority.

Enhance the separation of the classes, remove noise.

Neighbourhood Cleaning Rule (NCL)

- Cleaning
- Final dataset shape varies
- Removes hard cases

NCL: Procedure

Step 1:

1) Trains a 3 KNN on entire dataset.
2) Finds each observation’s 3 closest neighbours (for majority classes only).
3) Keeps or removes observation based on neighbours agreement with its class (uses majority vote).

So far, ENN.

NCL: Procedure

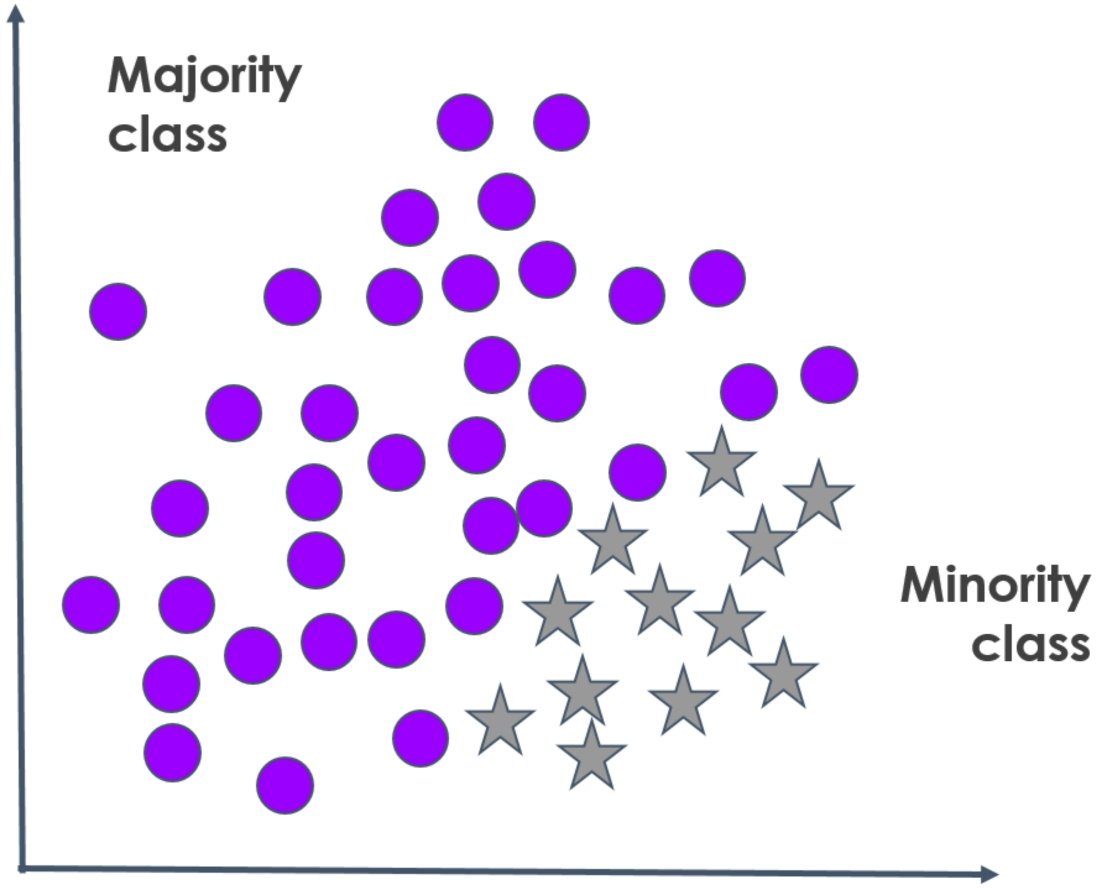
Step 2:

Now clean further:

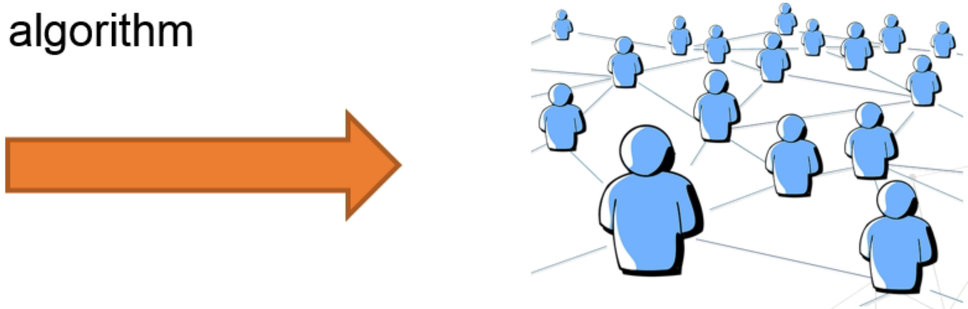
4) Find the 3 neighbours of each observation from the minority class.
5) If all neighbours or most neighbours disagree with the minority class, remove them.

Except: if the neighbours belong to a class with few samples. In the original article, they would only remove a neighbour if it belongs to a class with at least half as many observations as those in the minority.

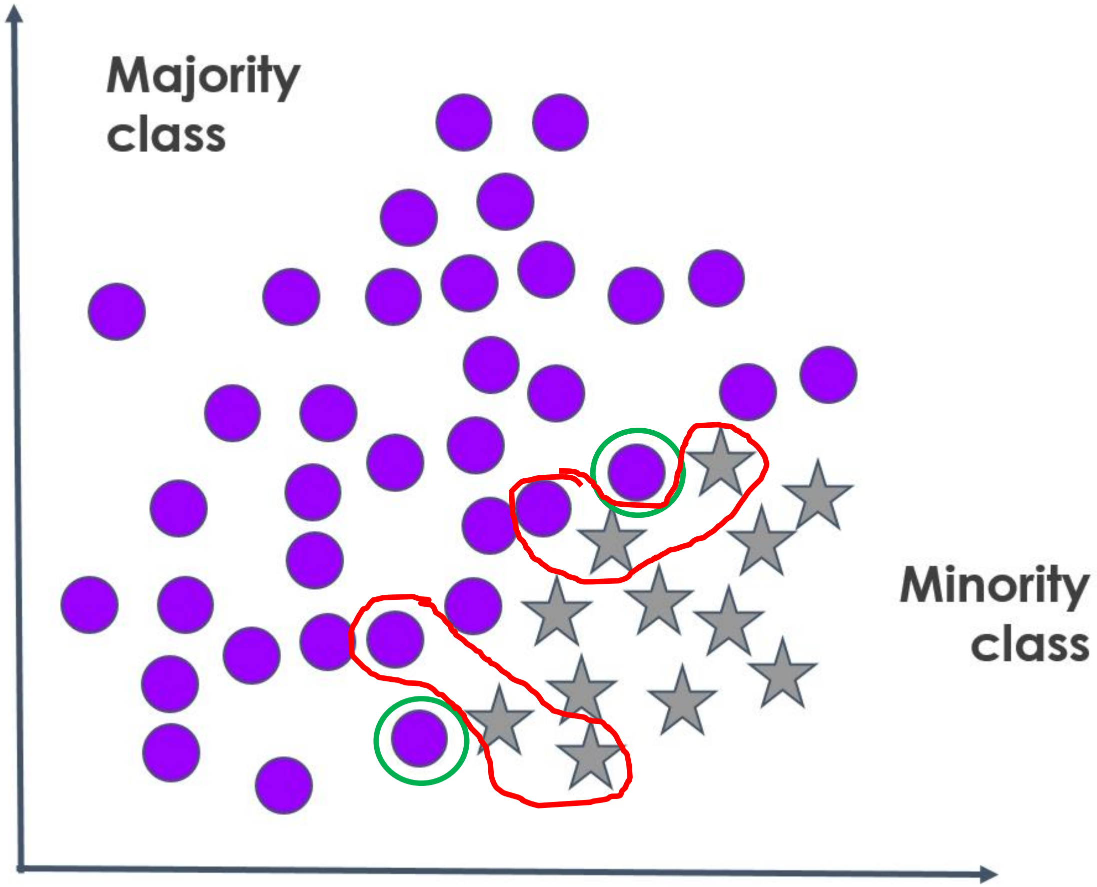
NCL: first step



Train a 3 KNN algorithm

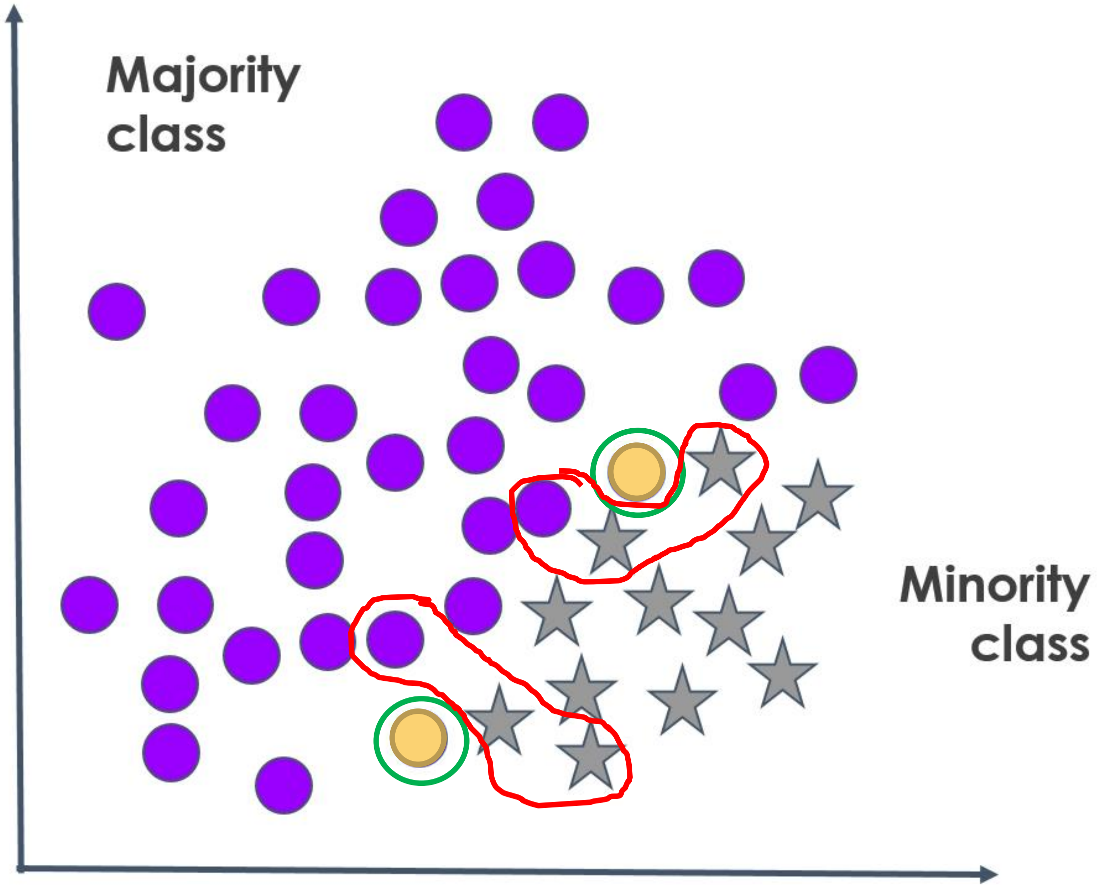


NCL: first step



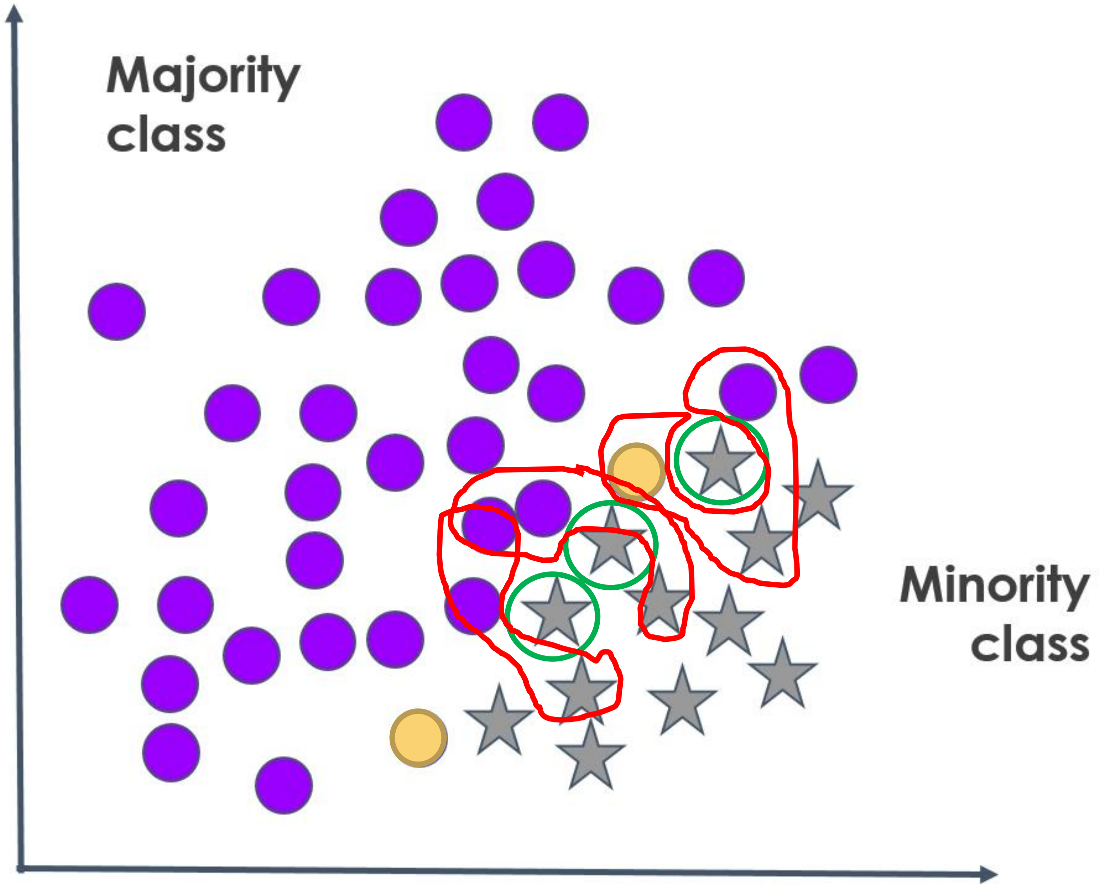
- Look at 3 neighbours from majority classes.
- If most neighbours disagree, flag observation for removal.

NCL: first step



- Look at 3 neighbours from majority classes.
- If most neighbours disagree, flag observation for removal.

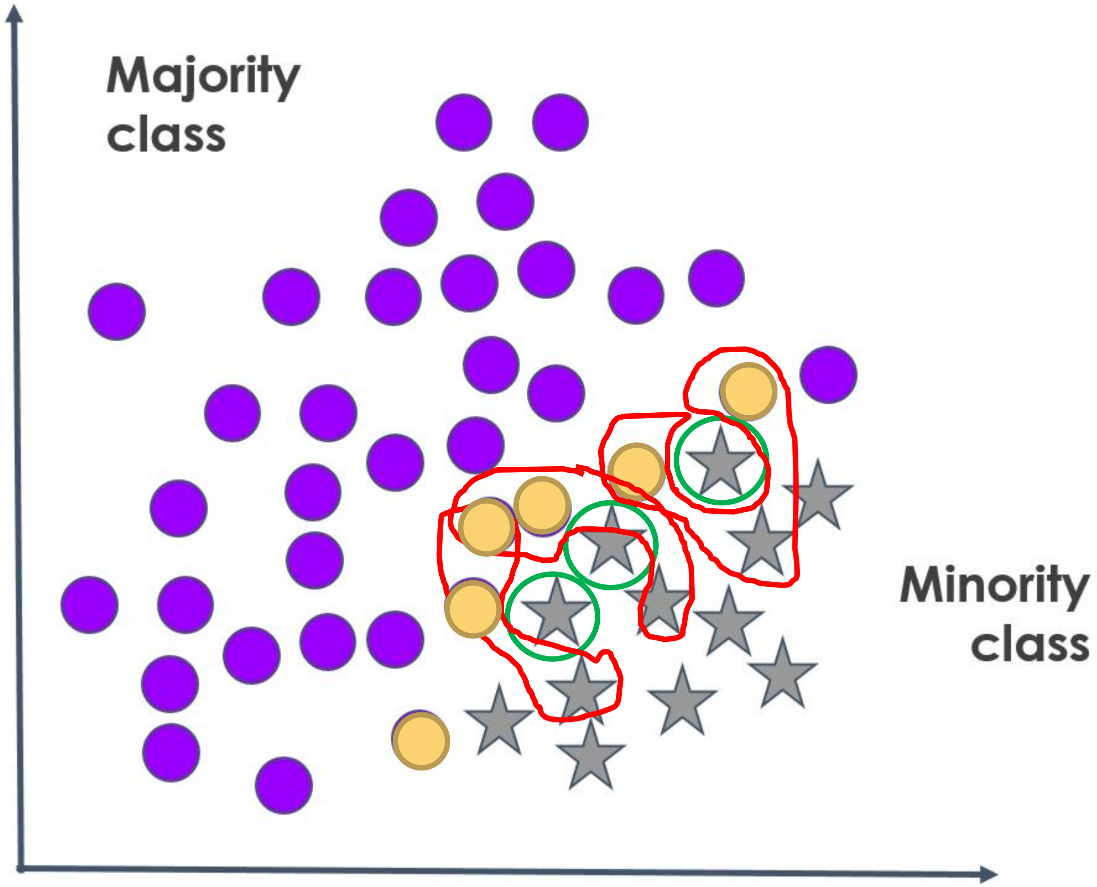
NCL: second step



- Looks at 3 neighbours from each minority class
- If **most /all** neighbours disagree with the minority and belong to the majority:

➢ Flag for removal

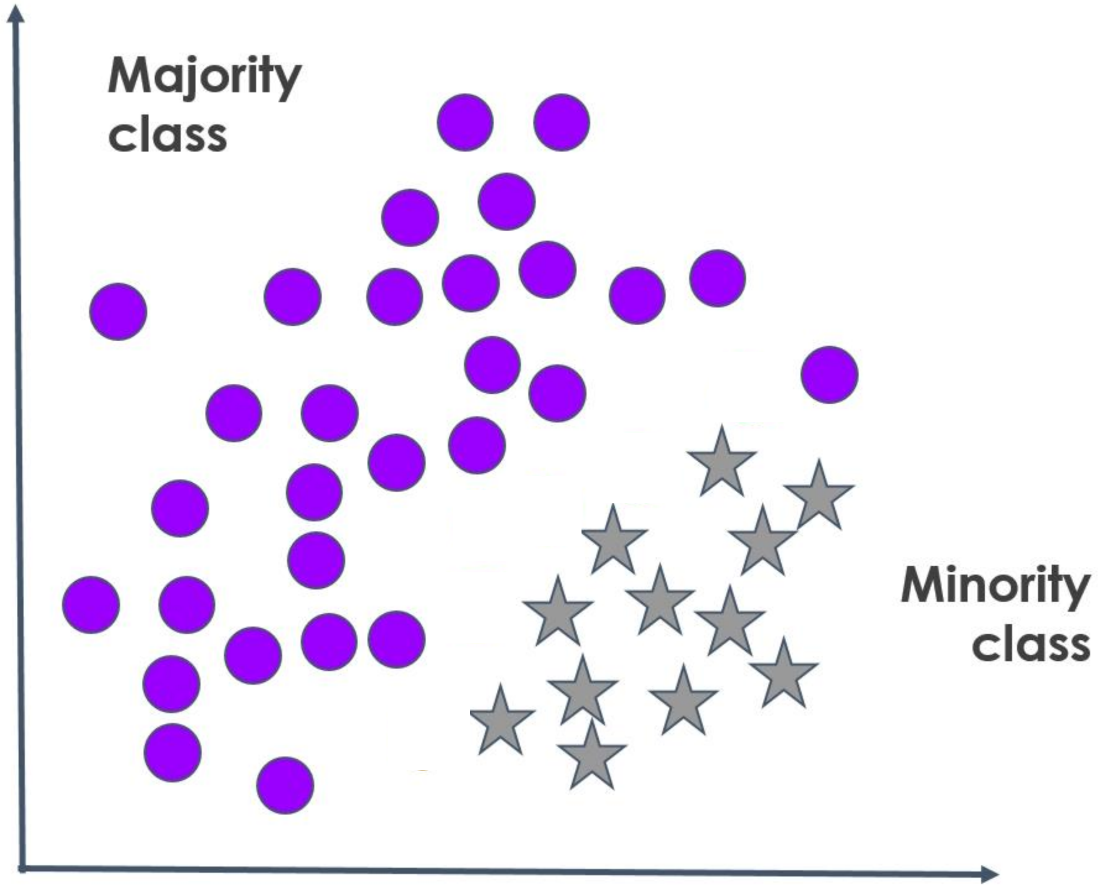
NCL: second step



- Looks at 3 neighbours from each minority class
- If most neighbours disagree with the minority and belong to the majority:

➢ Flag for removal

NCL: final dataset

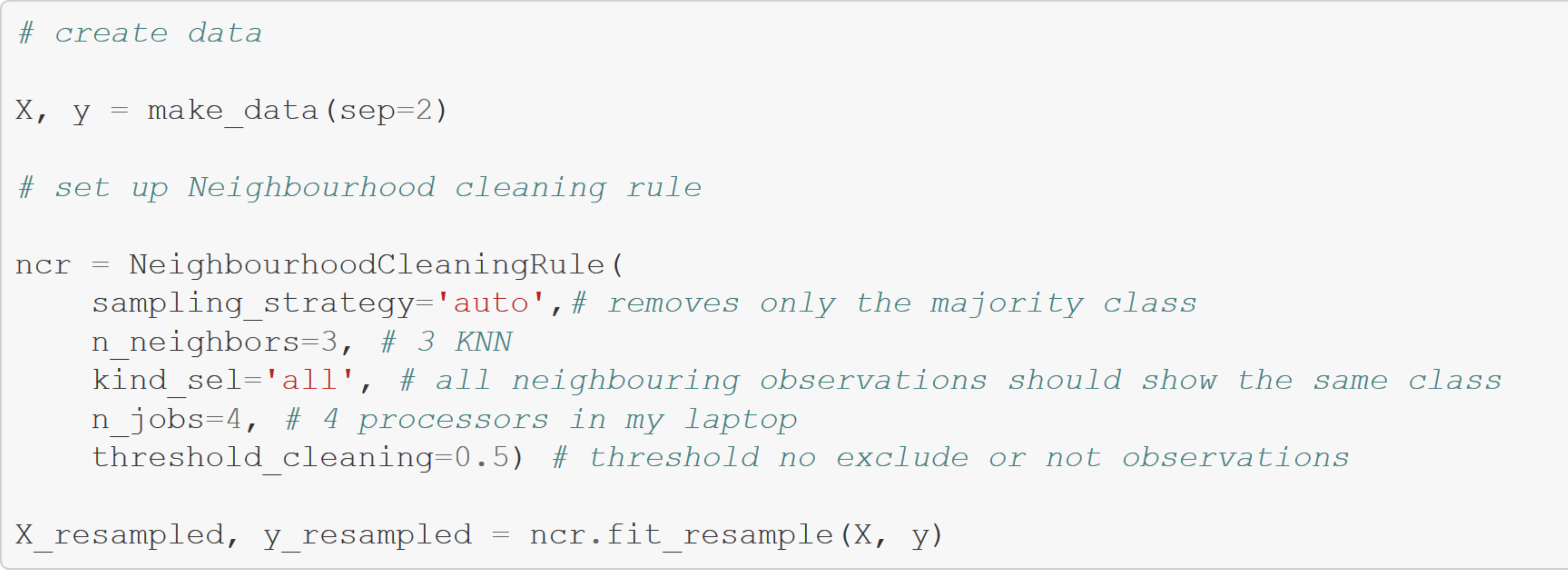


Remove all flagged observations

Final dataset:

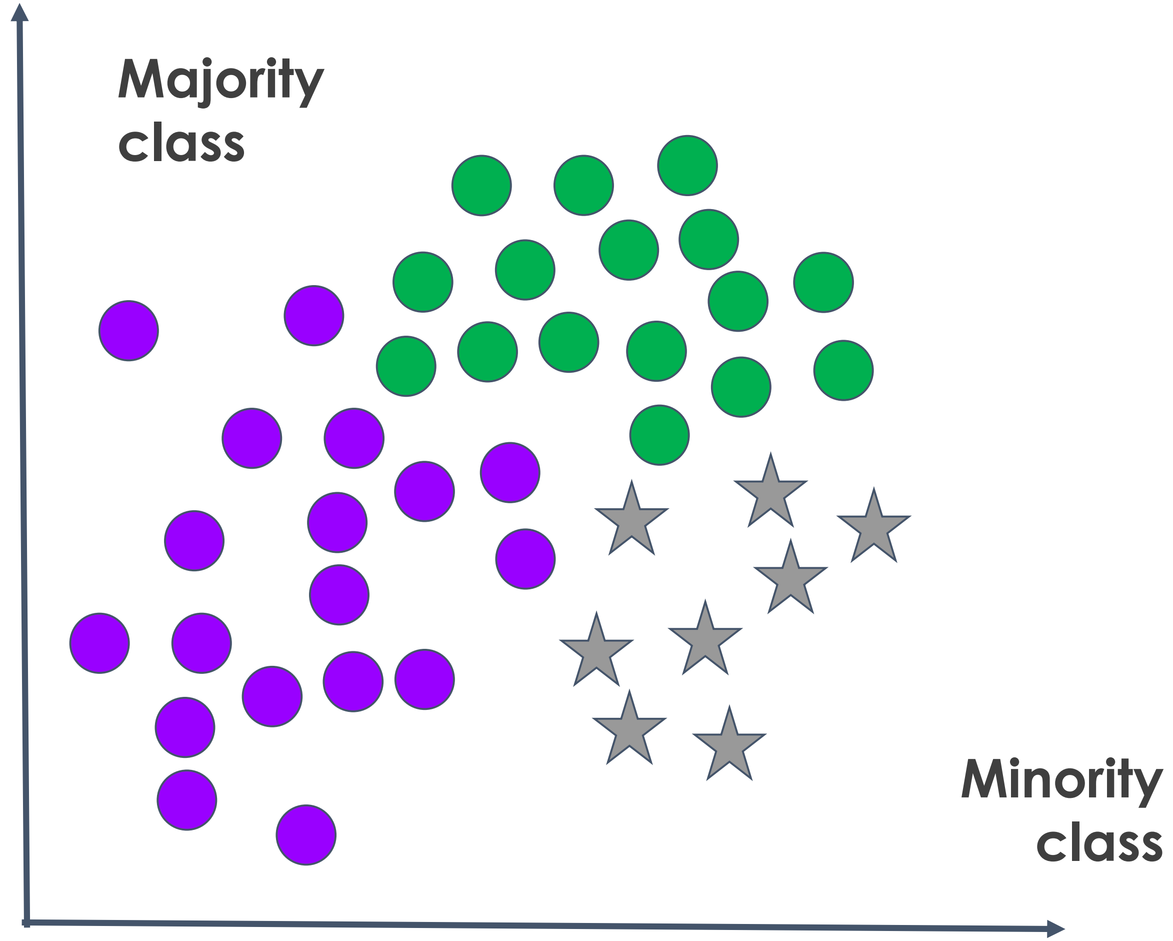
- Retains all observations from minority
- Removes some from majority

Imbalanced-learn: NCL



```
# create data

X, y = make_data(sep=2)

# set up Neighbourhood cleaning rule

ncr = NeighbourhoodCleaningRule(
    sampling_strategy='auto',# removes only the majority class
    n_neighbors=3, # 3 KNN
    kind_sel='all', # all neighbouring observations should show the same class
    n_jobs=4, # 4 processors in my laptop
    threshold_cleaning=0.5) # threshold no exclude or not observations

X_resampled, y_resampled = ncr.fit_resample(X, y)
```

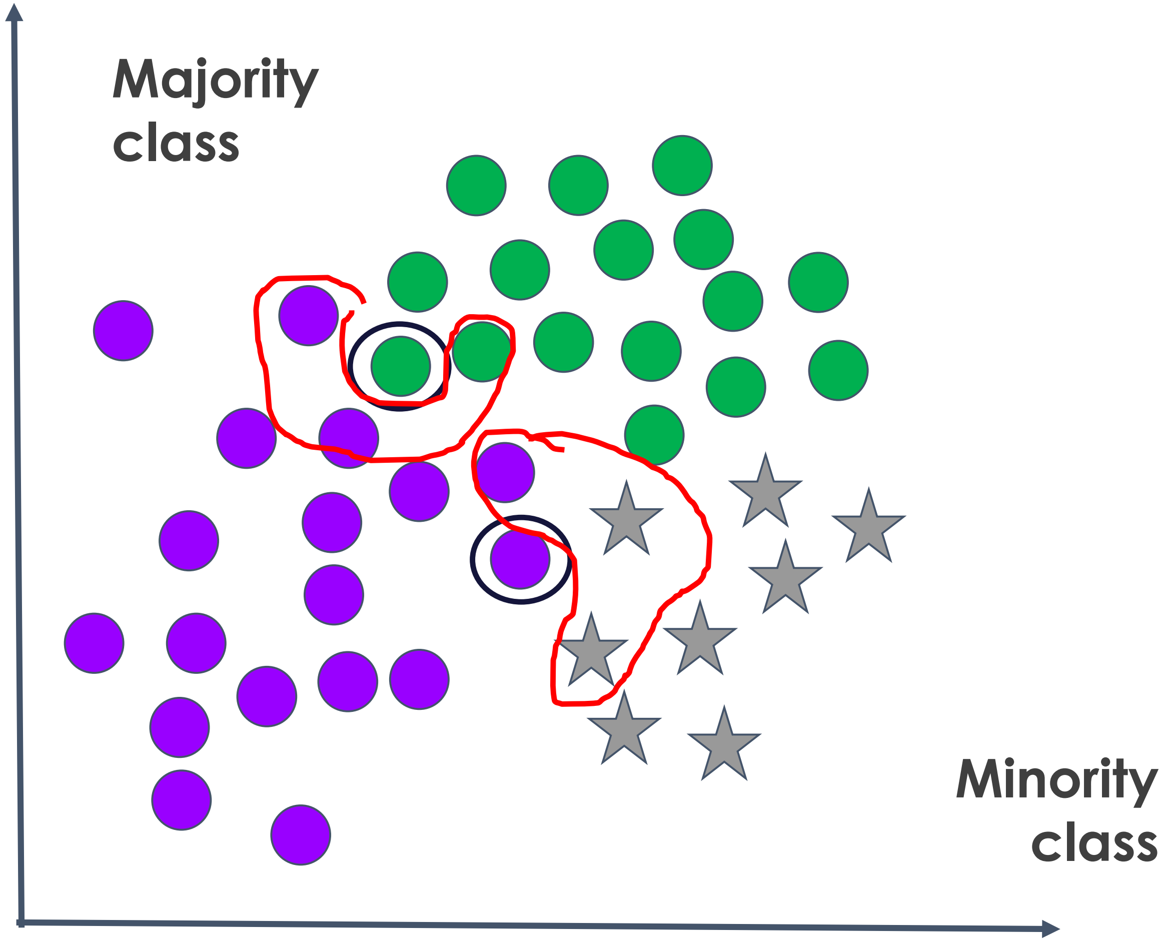
Multi-class



One vs Rest

Only majority classes are undersampled.

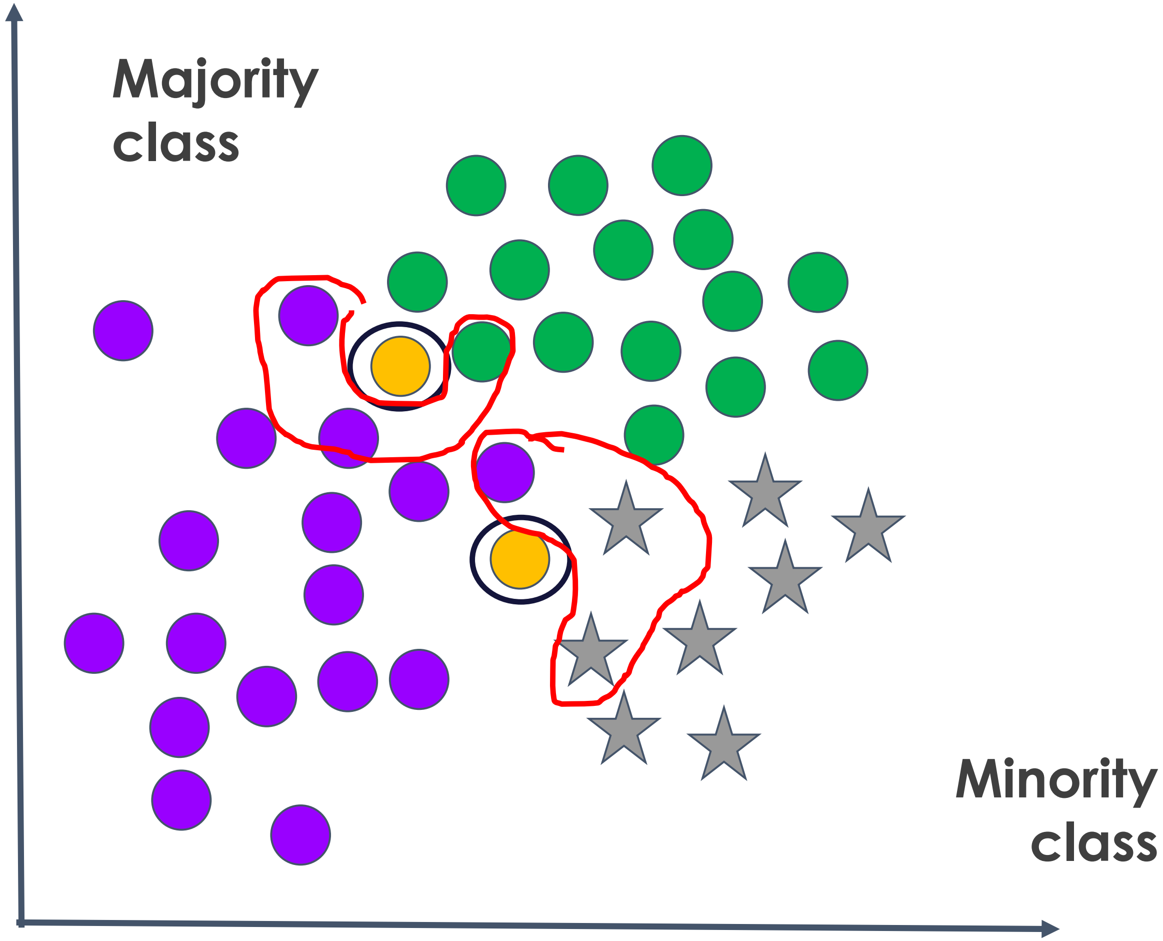
Multi-class: ENN



One vs Rest

When most neighbours disagree, flag the observation.

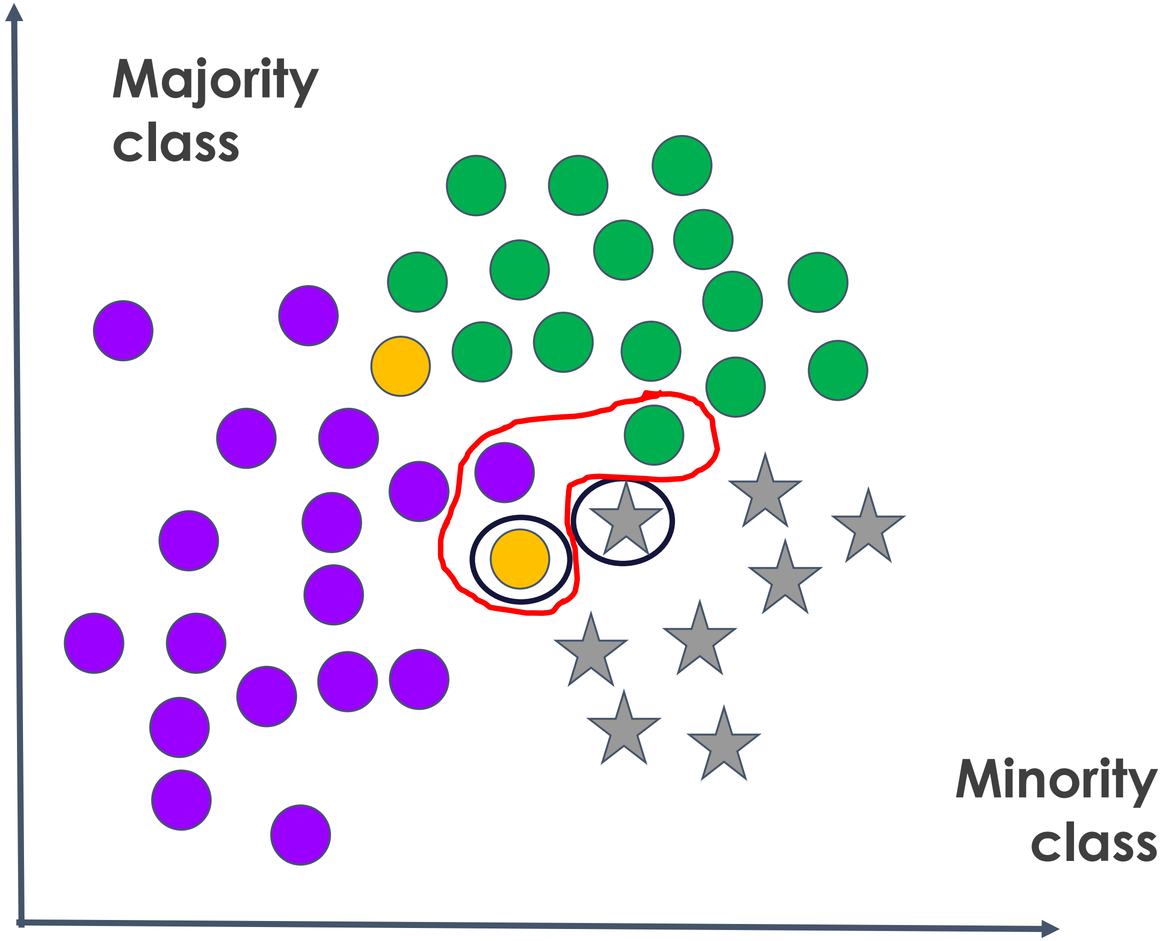
Multi-class: ENN



One vs Rest

When most neighbours disagree, flag the observation.

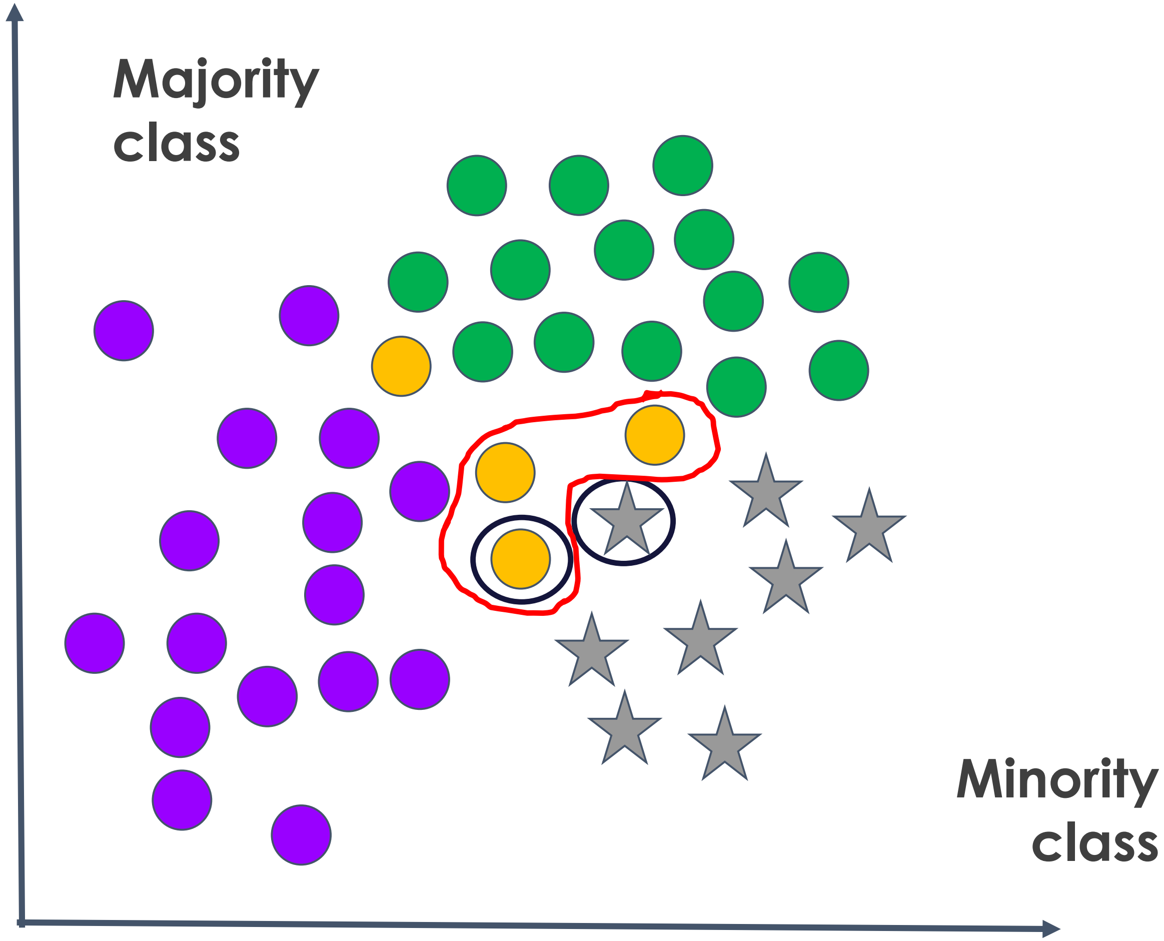
Multi-class: cleaning



One vs Rest

When all or most neighbours disagree, flag the observation.

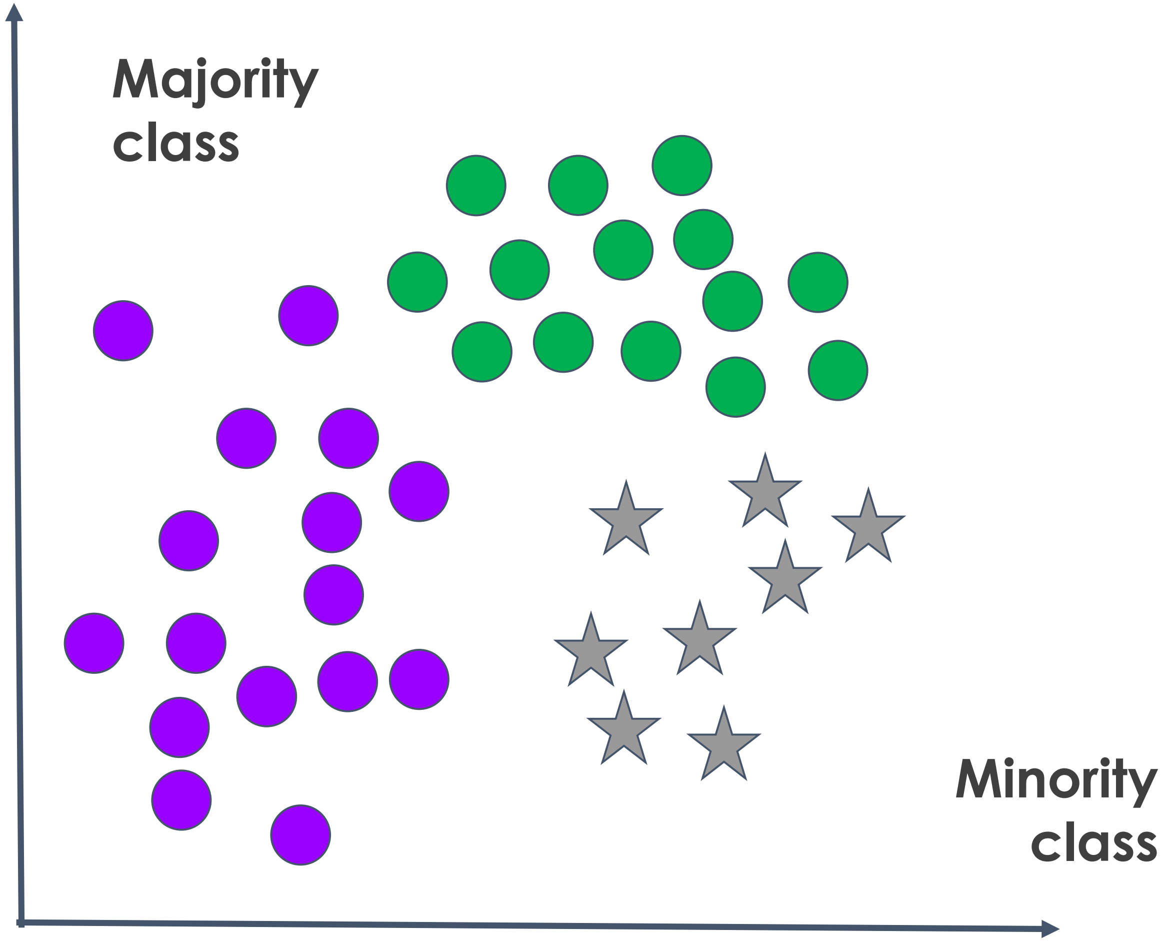
Multi-class: cleaning



One vs Rest

When all or most neighbours disagree, flag the observation.

Multi-class: final dataset



One vs Rest

Final dataset = original minus flagged observations

THANK YOU

www.trainindata.com