SMOTE-N

SMOTE-N

- **SMOTE-N → Nominal (Categorical) variables, ONLY.**
- Extends the functionality of SMOTE to categorical variables

SMOTE-N vs SMOTE-NC

- SMOTE-NC
  - Continuous
  - Categorical

≠

- SMOTE-N
  - Categorical

SMOTE-N procedure

- Looks only at the minority class examples
- Find the k (usually 5) closest neighbours
- Determine the values of the newly created examples

SMOTE-N procedure

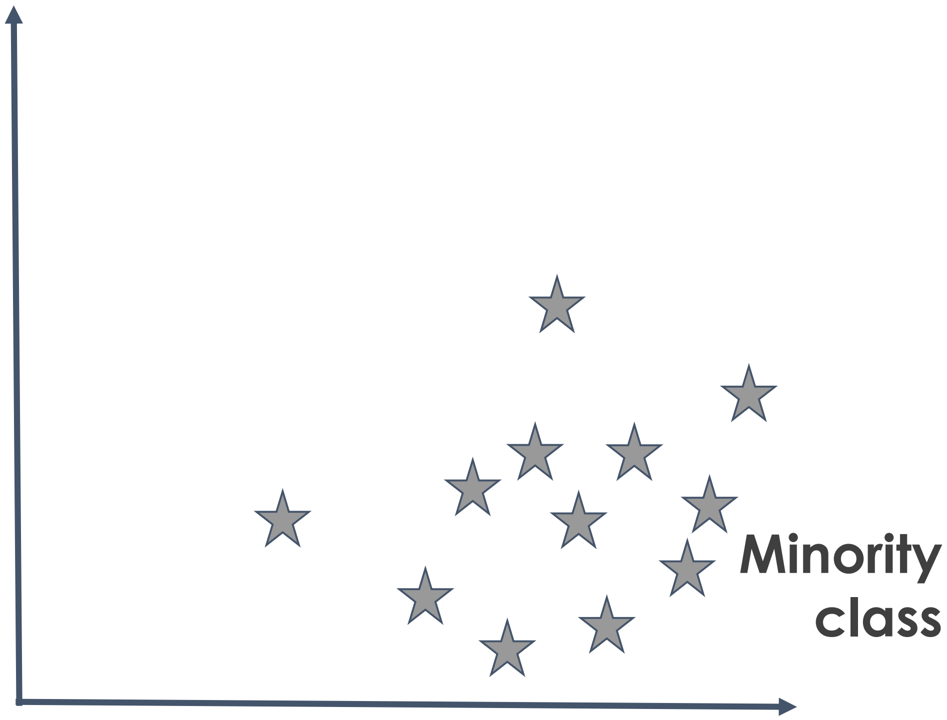
1. Find the k (usually 5) closest neighbours

   Distance: Value Difference Metric

2. Determine the values of the newly created examples

   Majority Vote

SMOTE-N: how it works

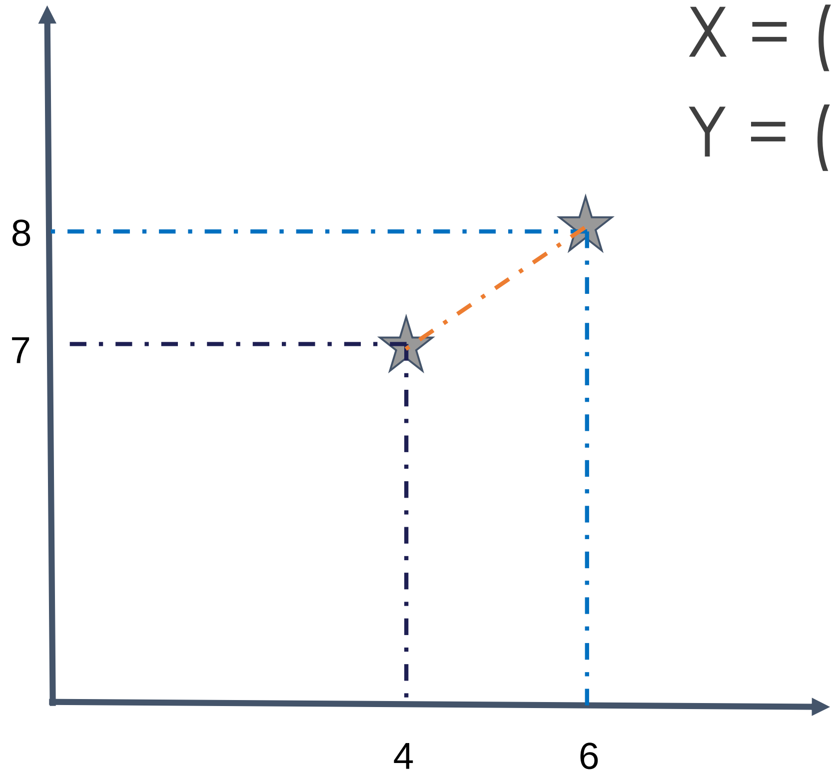


Looks **only** at observations from the minority class.

Finds its k (typically 5) nearest neighbours

The neighbours are found based on distances

Distance in numerical vectors



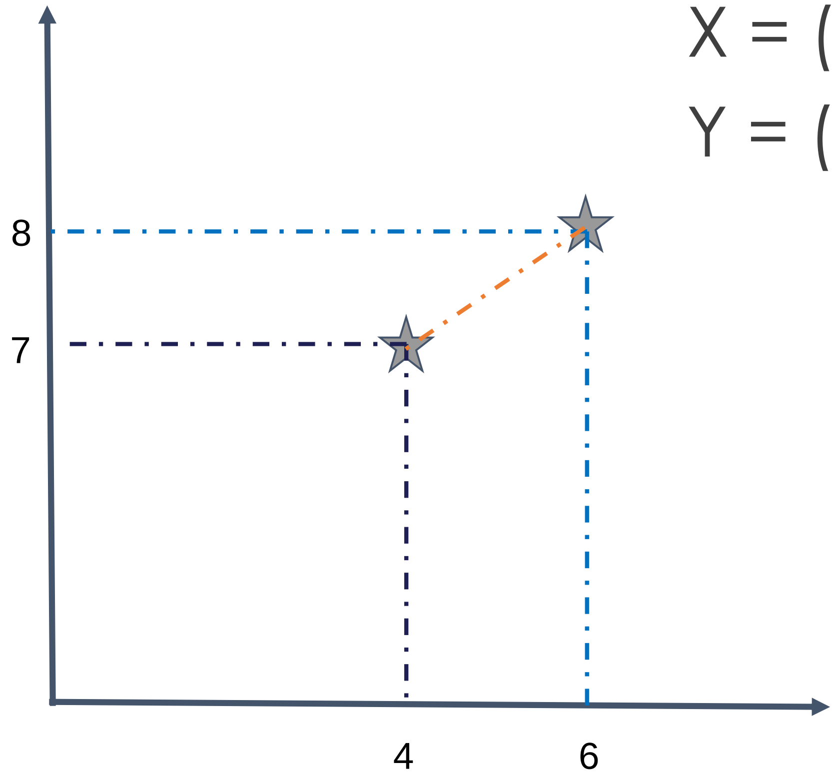
X = (4,7)

Y = (6, 8)

$$L1 = \sum |X - Y|$$

$$L2 = \sqrt{\sum (X - Y)^2}$$

Distance in numerical vectors



X = (4,7)

Y = (6, 8)

$$L1 = \sum |X - Y|$$

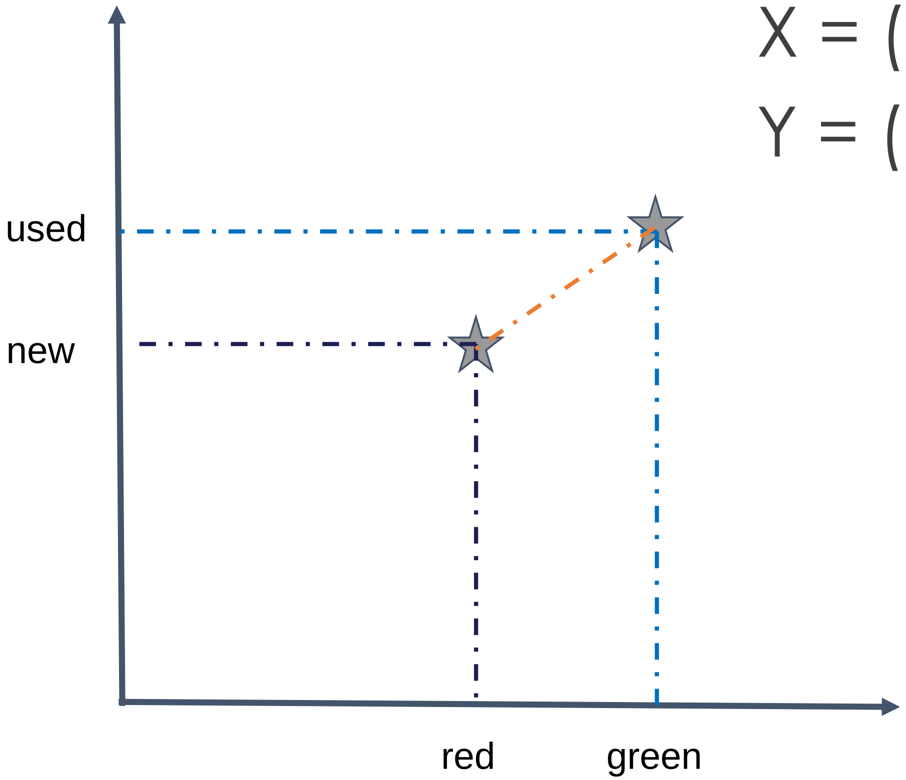
$$|4 - 6| + |7 - 8| = 2 + 1 = 3$$

$$L2 = \sqrt{\sum (X - Y)^2}$$

$$((4-6)^2 + (7-8)^2)^{1/2} = (4 + 1)^{1/2} = 2.24$$

Train In Data

Distance in categorical vectors



X = (green,used)

Y = (red, new)

$$L1 = \sum |X - Y|$$

|green - red| + |used - new| = ?

Value Difference Metric (VDM)

- $N_{a,x}$ is the number of examples in the training set that have value x for variable a;
- $N_{a,x,c}$ is the number of examples that have value x for feature a given class c (conditional probability);
- C is the number of classes;
- q is a constant, usually 1 or 2;

$$L1 = \sum |X - Y|$$

|green - red| + |used - new| = ?

$$vdm_a(x,y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

Distance between values

Colour	Target
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	0
green	0
red	0
red	0
red	0
red	1
red	1
red	1
red	1
red	1
red	1
red	1
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	1

$$= \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

Distance between values

Colour	Target
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	0
green	0
red	0
red	0
red	0
red	1
red	1
red	1
red	1
red	1
red	1
red	1
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	1

	0 - Na,x,c	1 - Na,x,c	Col - Na,x
	0	1	
green	2	8	10
red	3	7	10
blue	9	1	10

$$= \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

Distance between values

Colour	Target
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	0
green	0
red	0
red	0
red	0
red	1
red	1
red	1
red	1
red	1
red	1
red	1
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	1

	0 - Na,x,c	1 - Na,x,c	Col - Na,x
	0	1	
green	2	8	10
red	3	7	10
blue	9	1	10

	Conditional probability		
	0	1	
green	0.20	0.80	0.33
red	0.30	0.70	0.33
blue	0.90	0.10	0.33

$$= \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

Distance between values

Colour	Target
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	1
green	0
green	0
red	0
red	0
red	0
red	1
red	1
red	1
red	1
red	1
red	1
red	1
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	0
blue	1

	0 - Na,x,c	1 - Na,x,c	Col - Na,x
	0	1	
green	2	8	10
red	3	7	10
blue	9	1	10

	Conditional probability		
	0	1	
green	0.20	0.80	0.33
red	0.30	0.70	0.33
blue	0.90	0.10	0.33

$$= \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

green – green = |0.2 – 0.2| + |0.8 – 0.8| = **0**

green – red = |0.2 – 0.3| + |0.8 – 0.7| = **0.2**

green – blue = |0.2 – 0.9| + |0.8 – 0.1| = **1.4**

red – blue = |0.3 – 0.9| + |0.7 – 0.1| = **1.2**

Distance between values

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

	0 - Na,x,c	1 - Na,x,c	Col - Na,x
	0	1	Col
new	5	11	16
used	9	5	14

	Conditional probability		
	0	1	Col
new	0.31	0.69	0.53
used	0.64	0.36	0.47

$$= \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q$$

new – new = |0.31 – 0.31| + |0.69 – 0.69| = **0**

new – used = |0.31 – 0.64| + |0.69 – 0.36| = **0.66**

Distance between observations

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

green – green = **0**

green – red = **0.2**

green – blue = **1.4**

red – blue = **1.2**

new – new = **0**

new – used = **0.66**

$$\Delta(X,Y) = \sum_{f=1}^{F} \delta(X_f, Y_f)^r$$

where f is features (variables) and r is typically 1 or 2

Distance between observations

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

green – green = **0**

green – red = **0.2**

green – blue = **1.4**

red – blue = **1.2**

new – new = **0**

new – used = **0.66**

$$\Delta(X,Y) = \sum_{f=1}^{F} \delta(X_f, Y_f)^r$$

where f is features (variables) and r is typically 1 or 2

Δ([green; used], [green;used]) = (green – green)^2 + (used – used)^2 = **0**

Δ([green; used], [green;used]) = 0 + 0 = **0**

Distance between observations

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

green – green = **0**

green – red = **0.2**

green – blue = **1.4**

red – blue = **1.2**

new – new = **0**

new – used = **0.66**

$$\Delta(X,Y) = \sum_{f=1}^{F} \delta(X_f, Y_f)^r$$

where f is features (variables) and r is typically 1 or 2

Δ([green; used], [green;new]) = (green – green)^2 + (used – new)^2 = **0**

Δ([green; used], [green;new]) = 0 + 0.66^2 = **0.436**

Distance between observations

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

green – green = **0**

new – new = **0**

green – red = **0.2**

new – used = **0.66**

green – blue = **1.4**

red – blue = **1.2**

$$\Delta(X,Y) = \sum_{f=1}^{F} \delta(X_f, Y_f)^r$$

where f is features (variables) and r is typically 1 or 2

Δ([green; used], [red; used]) = (green – red)^2 + (used – used)^2 = **0**

Δ([green; used], [red; used]) = 0.2^2 + 0 = **0.04**

Distance between observations

Colour	Cond	Target
green	used	1
green	new	1
green	used	1
green	new	1
green	new	1
green	used	1
green	used	1
green	new	1
green	new	0
green	new	0
red	used	0
red	used	0
red	used	0
red	used	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
red	new	1
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	used	0
blue	new	0
blue	new	0
blue	new	0
blue	new	1

green – green = **0**

green – red = **0.2**

green – blue = **1.4**

red – blue = **1.2**

new – new = **0**

new – used = **0.66**

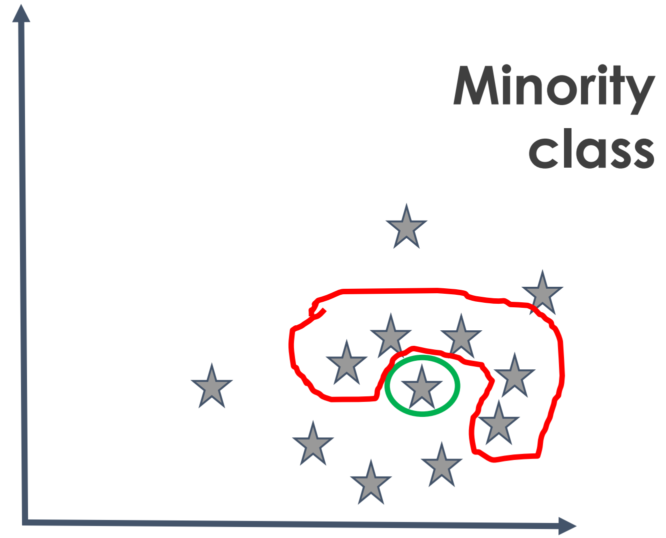
$$\Delta(X,Y) = \sum_{f=1}^{F} \delta(X_f, Y_f)^r$$

where f is features (variables) and r is typically 1 or 2

Δ([green; used], [red; new]) = (green – red)^2 + (used – new)^2 = **0**

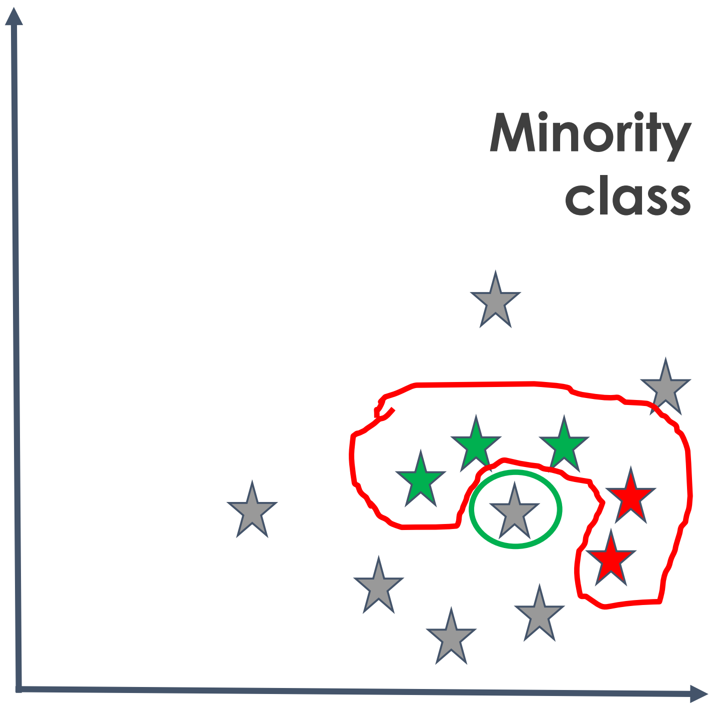
Δ([green; used], [red; used]) = $0.2^2 + 0.66^2$ = **0.477**

SMOTE-N: how it works



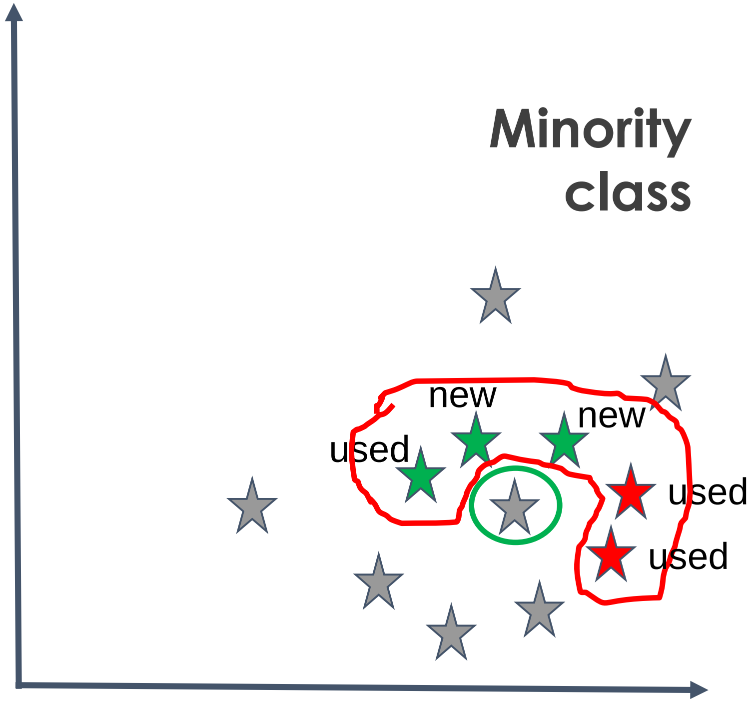
- With the VDM we determine distances
- With distances, we can train a KNN.
- We find the k nearest neighbours of each observation from the minority
- Values of the **new examples** are those shown by the majority of the neighbours

SMOTE-N: how it works



- Values of the **new examples** are those shown by the majority of the neighbours
- In this example, the new observation is green for the variable "colour"

SMOTE-N: how it works



- Values of the **new examples** are those shown by the majority of the neighbours
- In this example, the new observation is green for the variable “colour” and used for the variable “condition.

Imbalanced-learn: SMOTE-N

```
sampler = SMOTEN(
    sampling_strategy='auto', # samples only the minority class
    random_state=0,  # for reproducibility
    k_neighbors=5,
    n_jobs=4,
)

X_res, y_res = sampler.fit_resample(X, y)
```

THANK YOU

www.trainindata.com