Train In Data

SMOTE

SMOTE

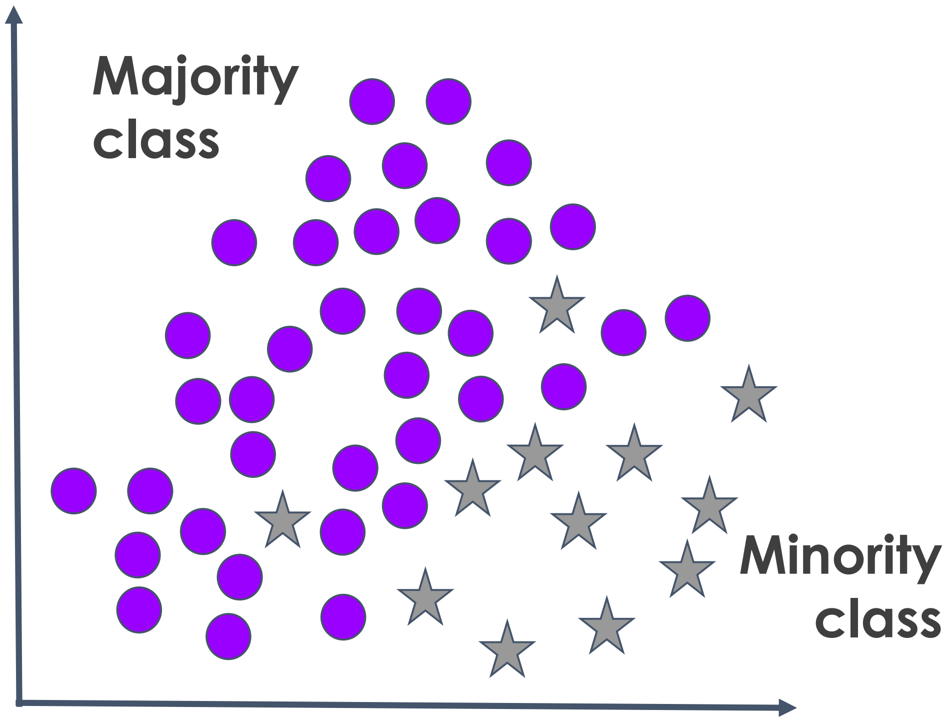
- **S**ynthetic **M**inority **O**ver-sampling **Te**chnique.
- Creates samples by interpolation
- Interpolation is a type of estimation, where we create new data points within the range of known data points

SMOTE

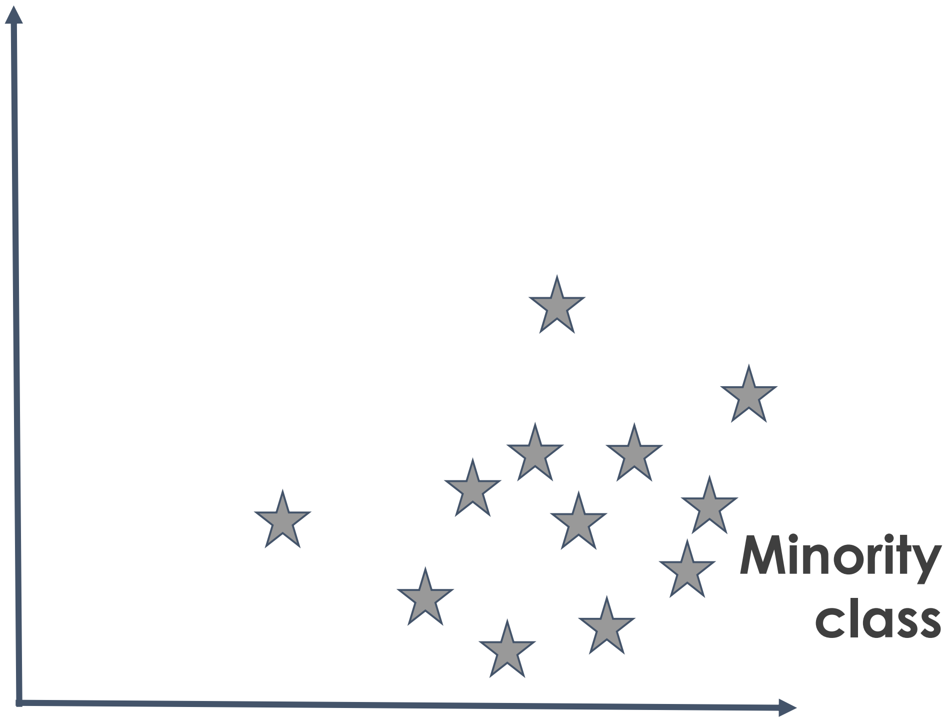
The minority class is “over-sampled” by creating “synthetic examples” instead of extracting data at random.

It prevents duplication. New observations from minority class will not be identical to original ones.

SMOTE: how it works



SMOTE: how it works

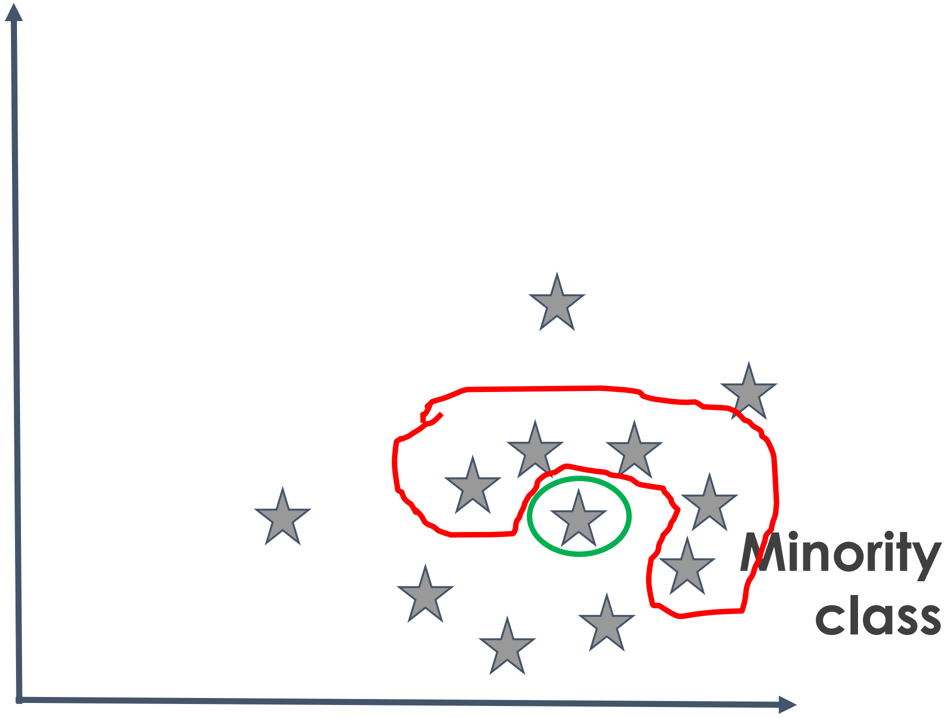


Looks only at the observations from the minority class.

Finds its k nearest neighbours

Typically k is 5

SMOTE: how it works

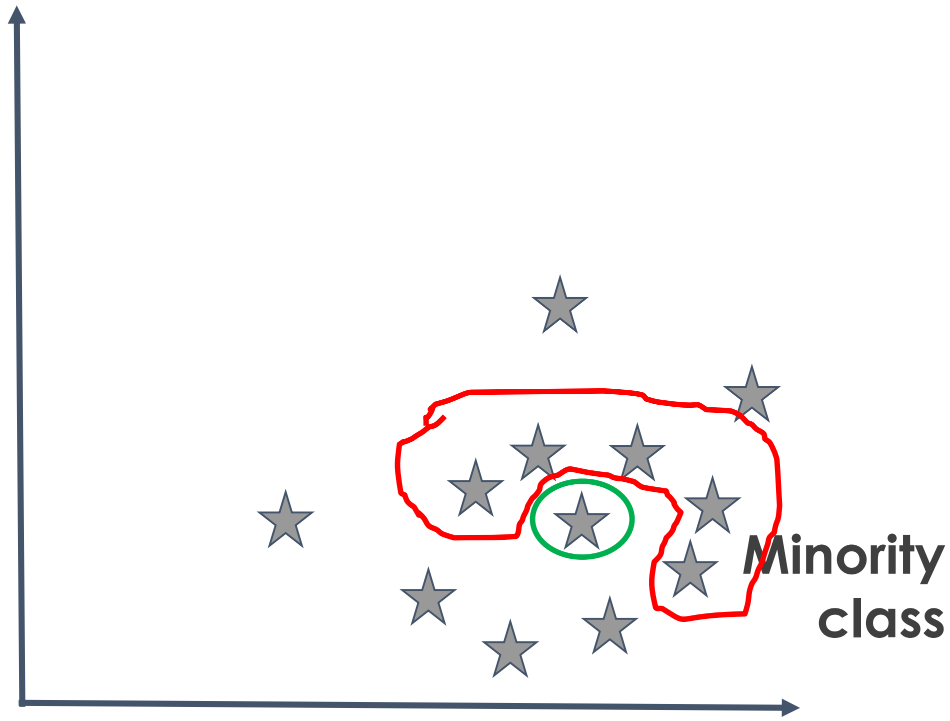


Looks only at the observations from the minority class.

Finds its k nearest neighbours

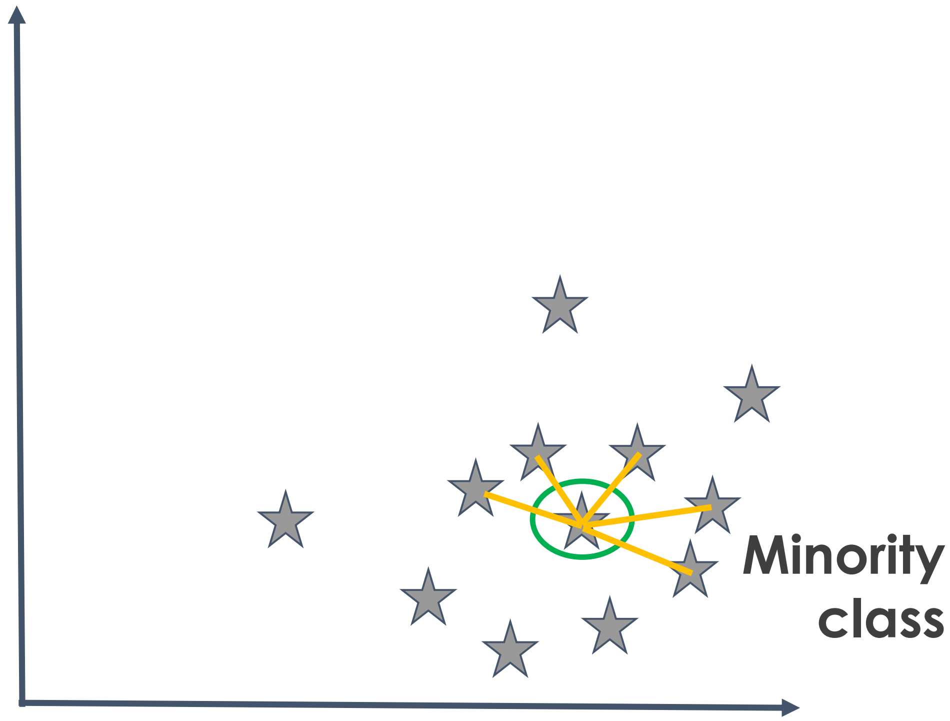
Typically k is 5

SMOTE: how it works



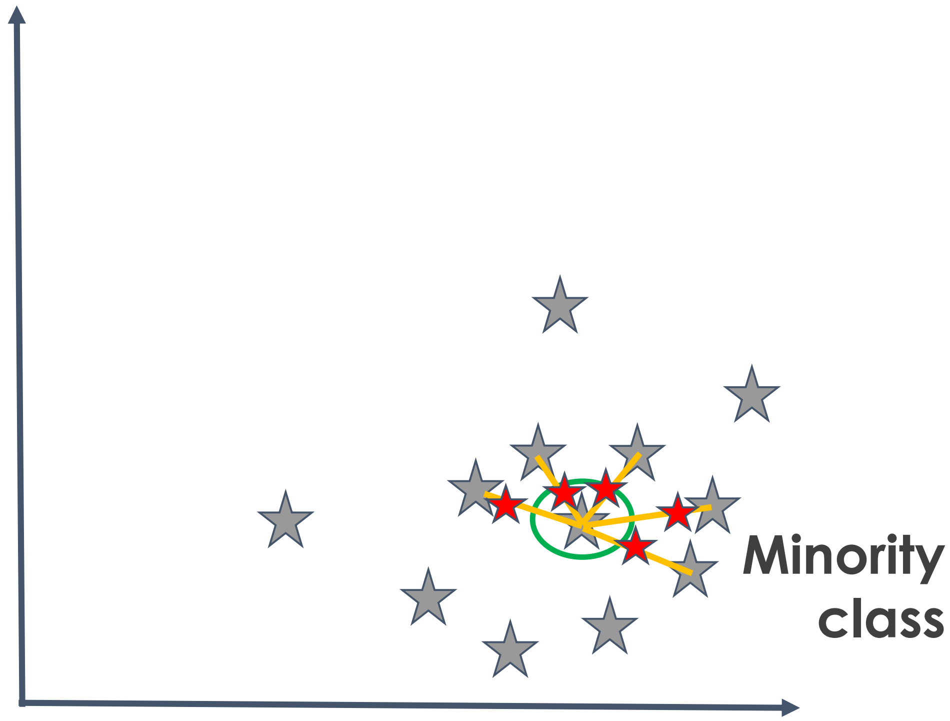
Determines the **distance** between the neighbours and the sample we want to generate a new observation from

SMOTE: how it works



Determines the **distance** between the neighbours and the sample we want to generate a new observation from

SMOTE: how it works

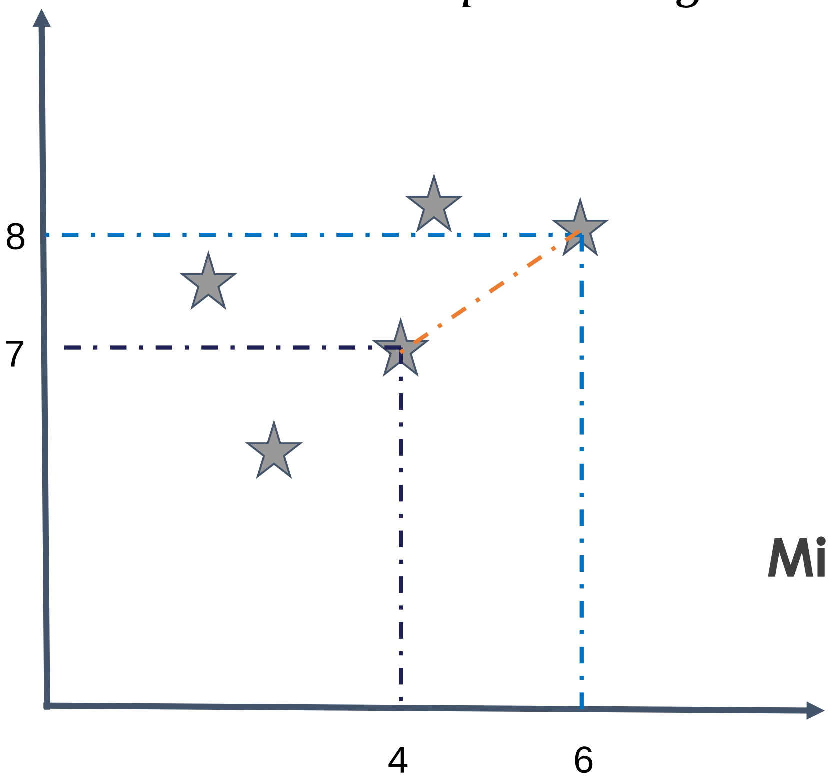


Multiplies that distance by a **random number** and adds it to the original sample to place the new observation in the dataset

$$New\ sample = original\ sample\ - factor\ * (original\ sample\ - neighbour)$$

SMOTE: numerical example

$$New\ sample = original\ sample\ - factor\ * (original\ sample\ - neighbour)$$



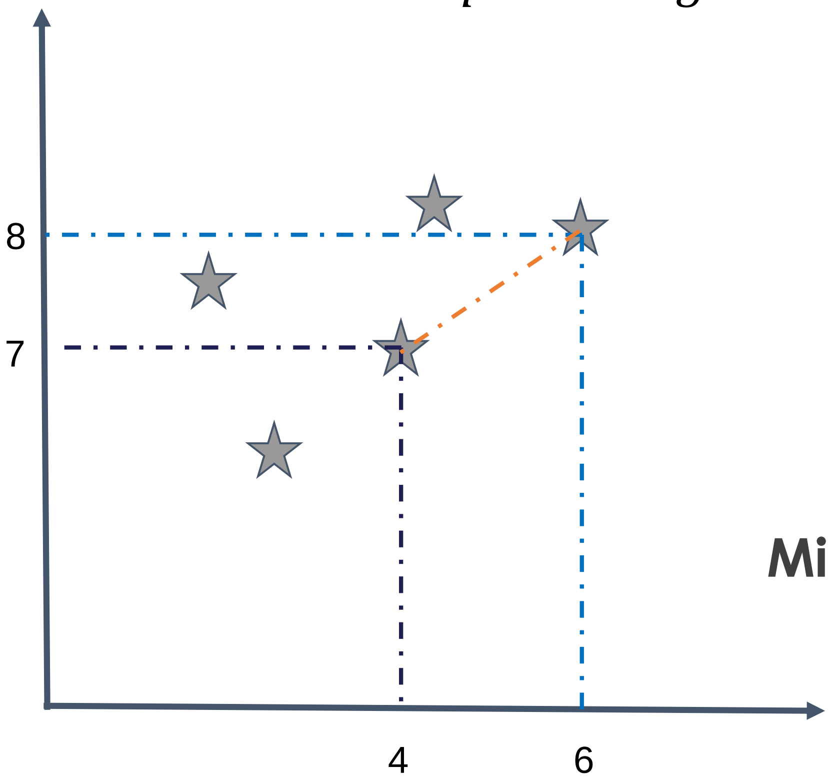
Minority class

X_{ori} = (4,7)

X_{neig} = (6, 8)

SMOTE: numerical example

$$New\ sample = original\ sample\ - factor\ * (original\ sample\ - neighbour)$$



Minority class

X_{ori} = (4,7)

X_{neig} = (6, 8)

$$New\ sample = (4{,}7) - 0.8\ * (\ (4{,}7) - (6{,}8))$$

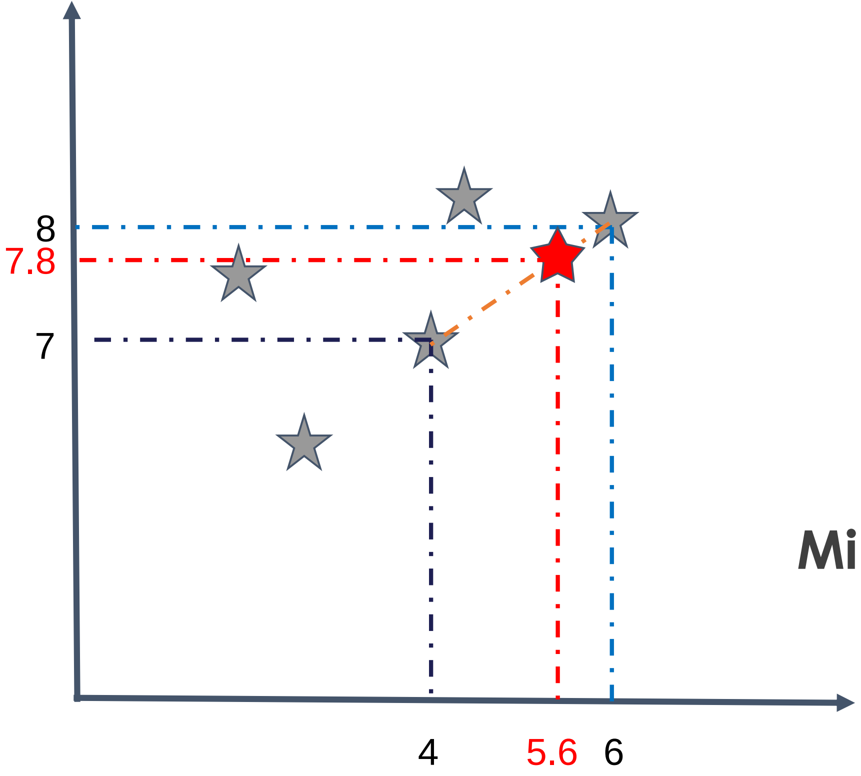
$$New\ sample = (4{,}7) - 0.8\ * (\ (-2, -1))$$

$$New\ sample = (4{,}7) - \big(\ (-1.6\ , -0.8)\big)$$

$$New\ sample = (5.6, 7.8)$$

SMOTE: numerical example

$$New\ sample = original\ sample\ - factor\ * (original\ sample\ - neighbour)$$



Minority class

X_{ori} = (4,7)

X_{neig} = (6, 8)

$$New\ sample = (4{,}7) - 0.8\ * (\ (4{,}7) - (6{,}8))$$

$$New\ sample = (4{,}7) - 0.8\ * (\ (-2, -1))$$

$$New\ sample = (4{,}7) - \left(\ (-1.6\ , -0.8)\right)$$

$$New\ sample = (5.6, 7.8)$$

SMOTE: Python implementation

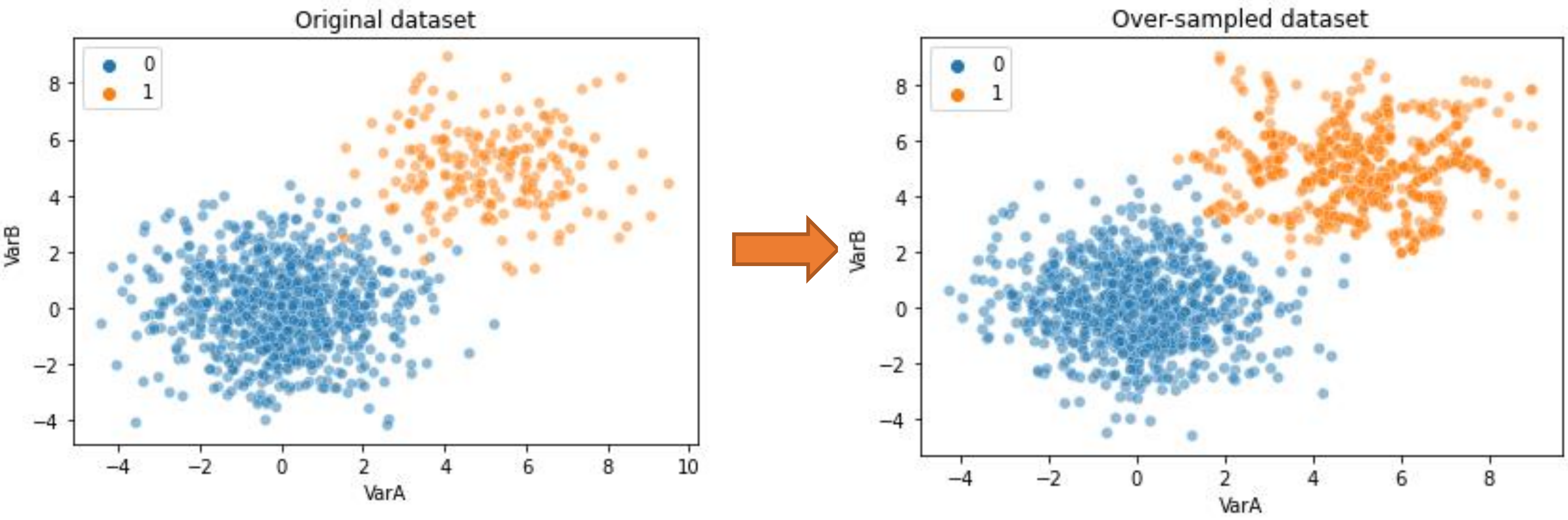
1. Isolates minority class samples
2. Trains KNN and finds K nearest neighbours to each sample of minority class
3. Determines how many new samples need to be generated
4. Selects from which samples a new sample will be generated (random)
5. Selects the neighbour that will be used to extrapolate the sample (random)
6. Finds a random factor
7. Creates the new sample

Imbalanced-learn: SMOTE

```
ros = SMOTE(strategy = ‘auto’,
                random_state = 29,
                 k_neighbours = 5)

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

Imbalanced-learn: SMOTE



Original dataset

Over-sampled dataset

Train In Data

THANK YOU

www.trainindata.com