Feature Engineering

Mass Street University

PER EDUCATIONEM PROGRESSUS

What we will cover

- Feature Engineering
- Fine Tuning
- Multiclass
- Multilabel
- Multi-output

Mass Street University

PER EDUCATIONEM PROGRESSUS

Feature Engineering

- Anything that changes our data.
- Very board term but important to learn how to do all facets of it.

Mass Street University

PER EDUCATIONEM PROGRESSUS

JSON Data

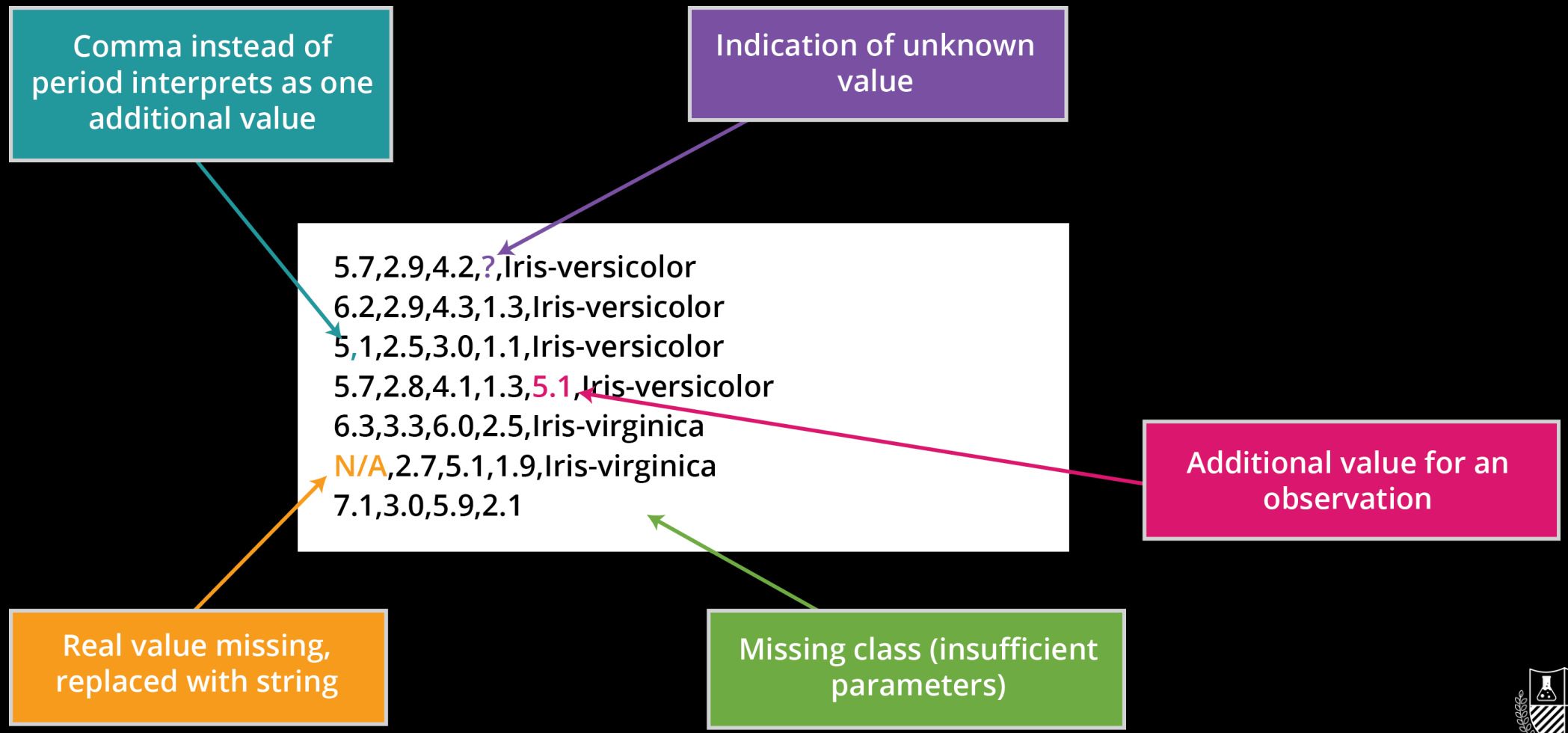
- Heavily nested data.
- Uses dictionaries to hold data.
- Can call on dictionary keys and values.


```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
  "latitude": 40.412086,
  "attributes": {
    "Take-out": true,
    "Wi-Fi": "free",
    "Drive-Thru": true,
    "Good For": {
      "dessert": false,
      "latenight": false,
      "lunch": false,
      "dinner": false,
      "breakfast": false,
      "brunch": false
    },
    "Caters": false,
    "Noise Level": "average",
    "Takes Reservations": false,
    "Delivery": false
```


Mass Street University

PER EDUCATIONEM PROGRESSUS

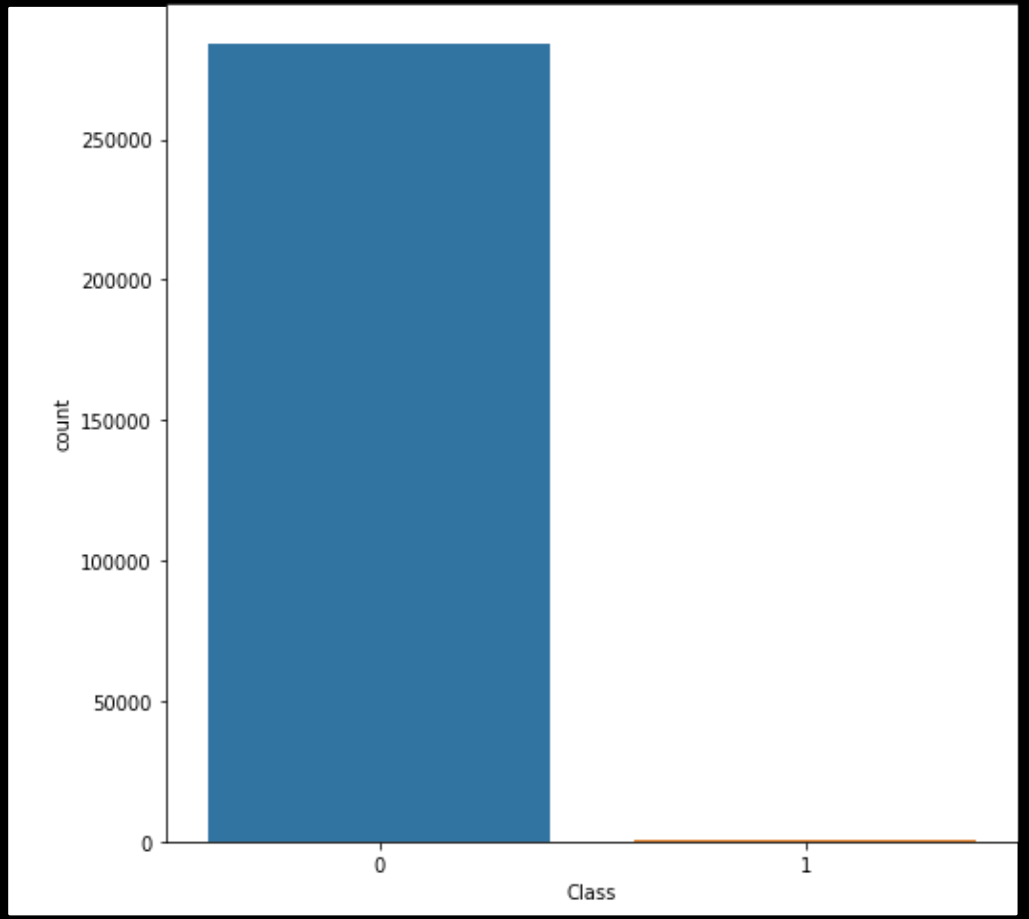
CSV Data

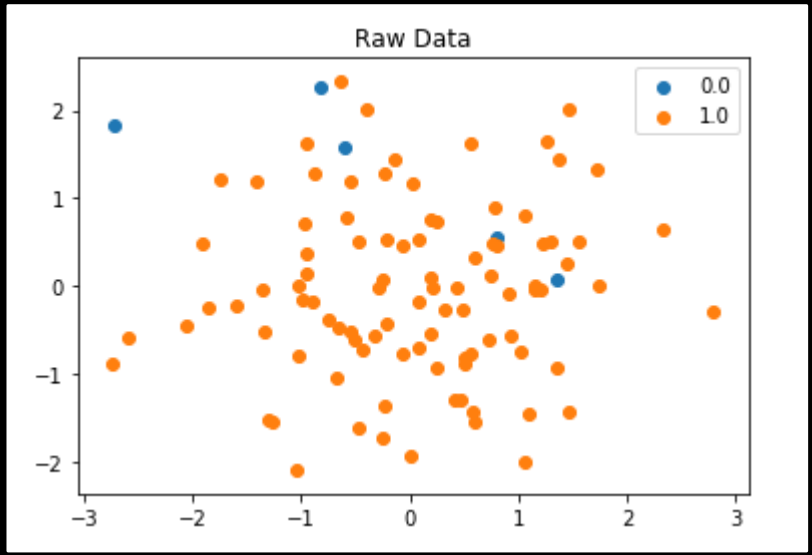


Mass Street University
PER EDUCATIONEM PROGRESSUS

Imbalanced Data

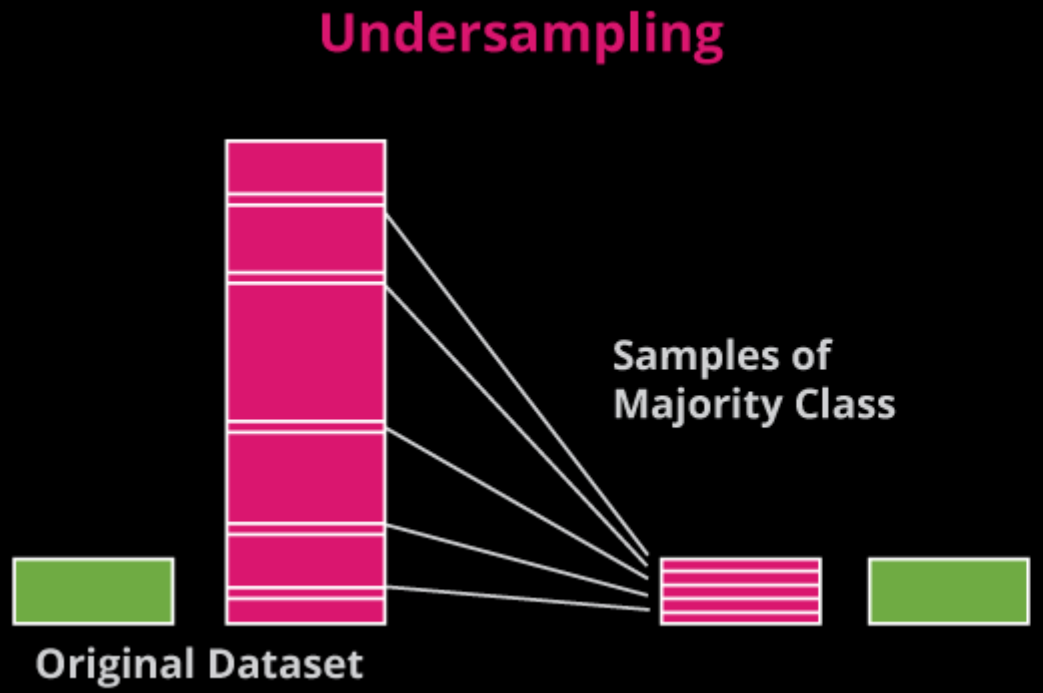
- Data that is heavily skewed to one class or one class having much less data than the other classes.

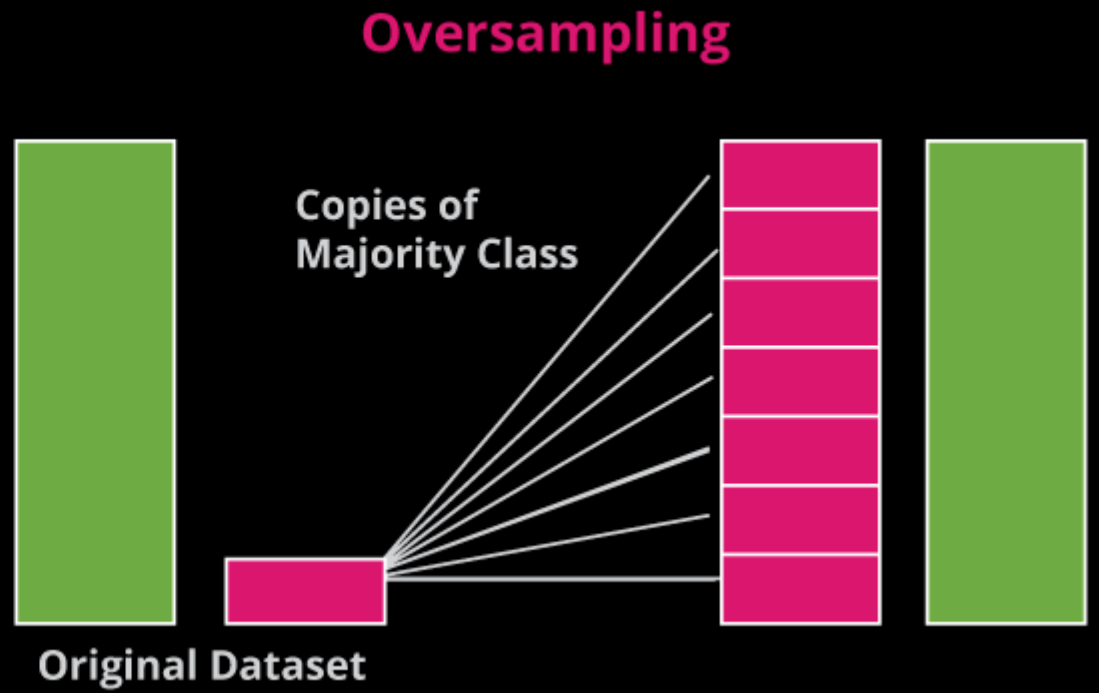




Mass Street University
PER EDUCATIONEM PROGRESSUS

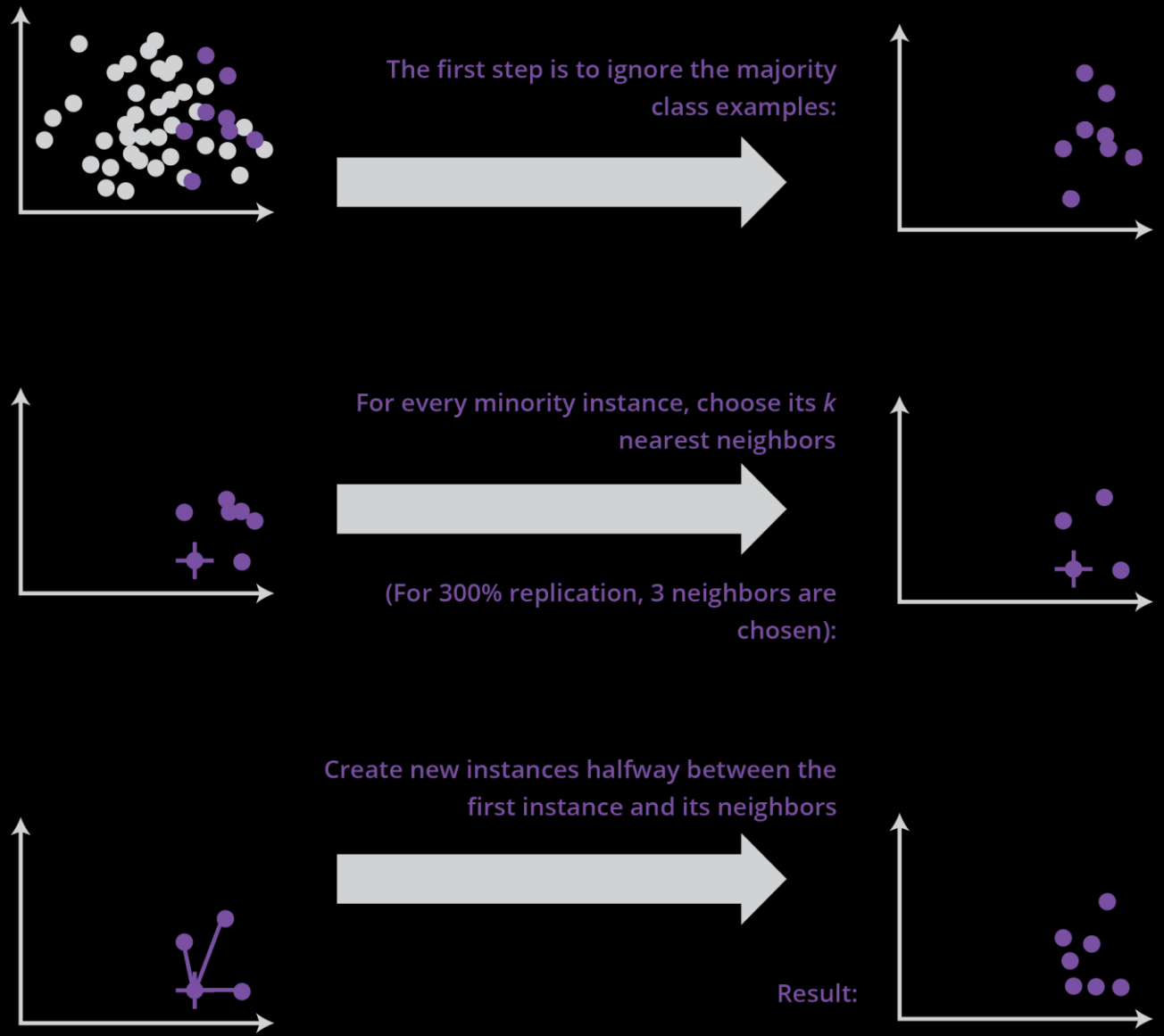
Under and Over Sampling

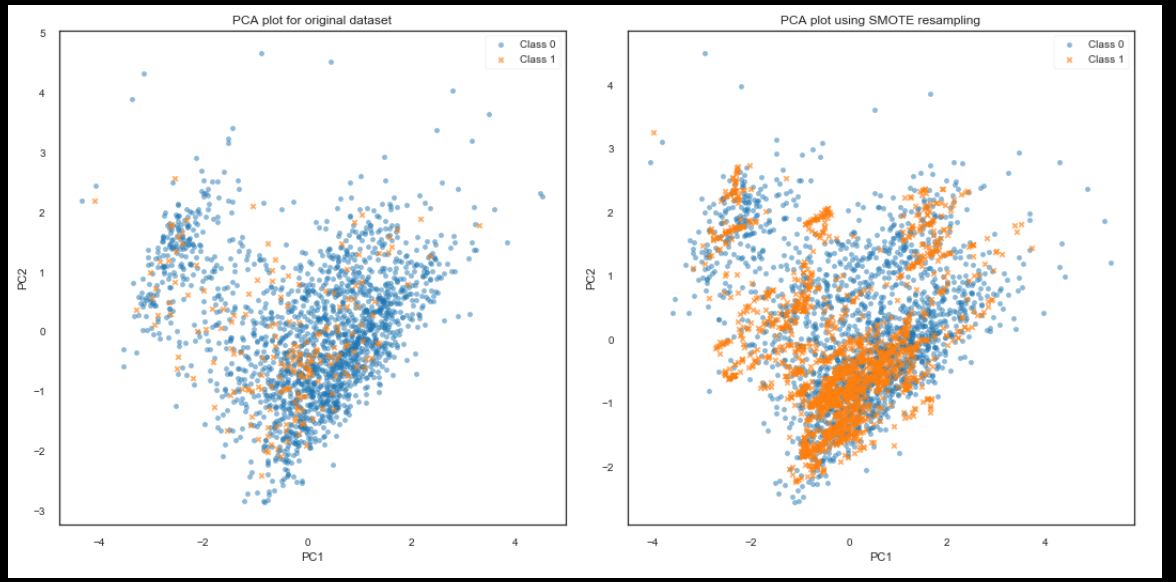




Mass Street University
PER EDUCATIONEM PROGRESSUS

SMOTE





Mass Street University
PER EDUCATIONEM PROGRESSUS

Class weighting (feature weighting)

- Gives importance to certain classes.
- Penalizes our models more for mistakes on classes that are important.
- Is usually built in to SciKit Learn and Keras' models.

Metrics

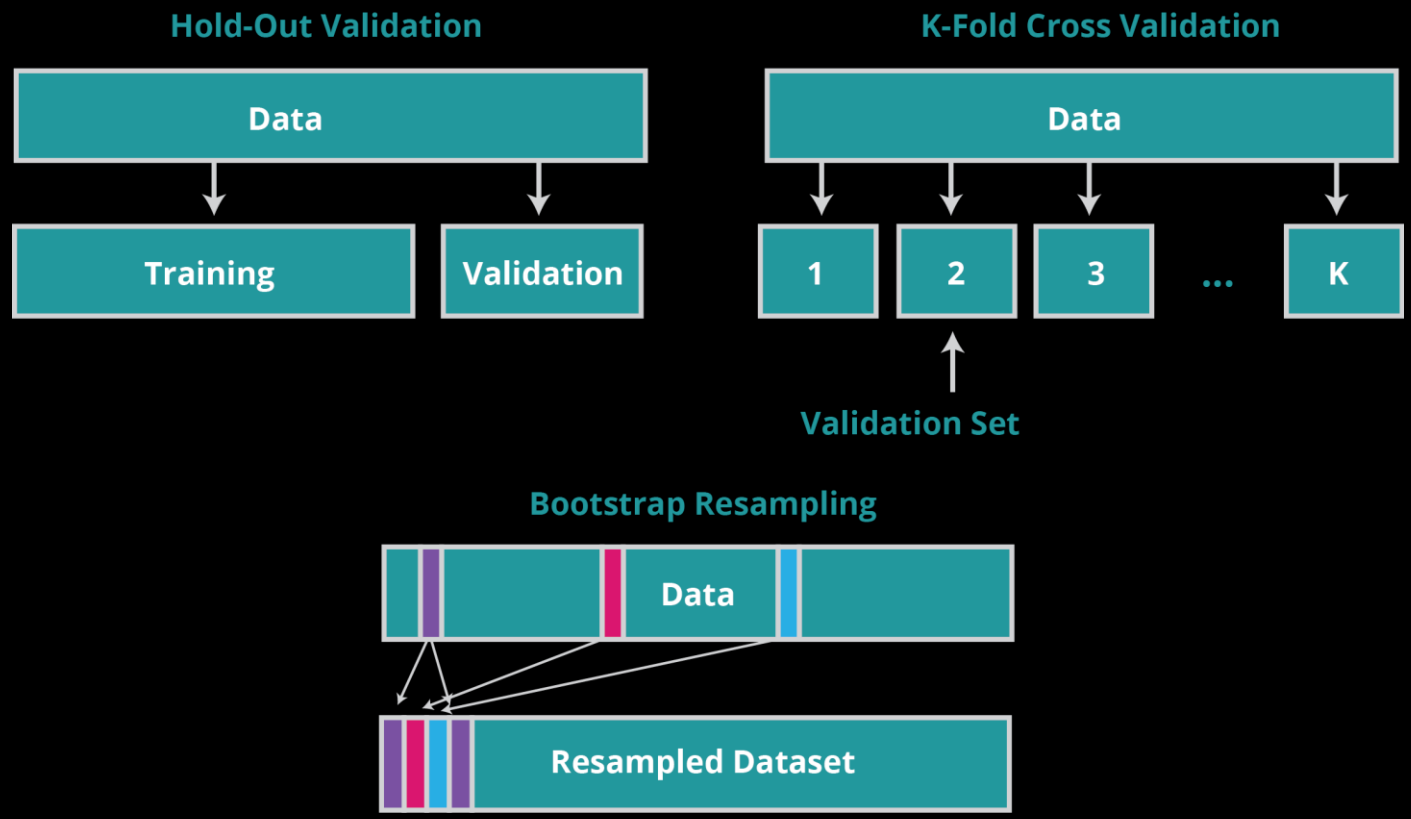
Metric	Formula
True Positive Rate, Recall	$\frac{TP}{TP+TN}$
False Positive Rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-Measure	$\frac{2 \times Precision \times Recall}{Precision + Recall}$

Mass Street University

PER EDUCATIONEM PROGRESSUS

Ensemble sampling (Bootstrapping)

- Creates many smaller datasets to work with to get a good distribution of the data.
- You can either do K-Fold or Hold testing on this method.
- Works best a majority of the time.



Mass Street University
PER EDUCATIONEM PROGRESSUS