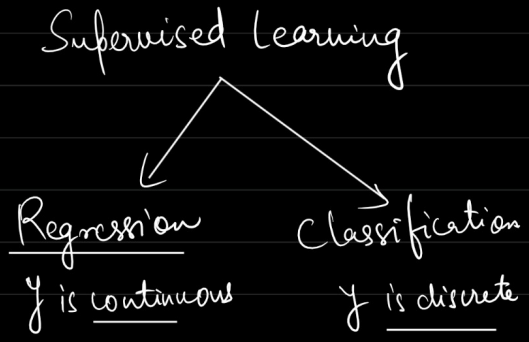
Imbalanced data



y { → Pass | Fail
  { → Diabetic | Not diabetic

* 2 outcomes of target variable (y)
  → binary classification problem.
  (bi)
  ↳ two.

Sugar level (f1)	Cholestrol (f2)	y (diabetic/Not)
250	102	1 → diabetic
300	100	1
400	200	0 → non diabetic
—	—	—

0 → 90% — 0 (non-dia)
10% — 1 (diabetic)

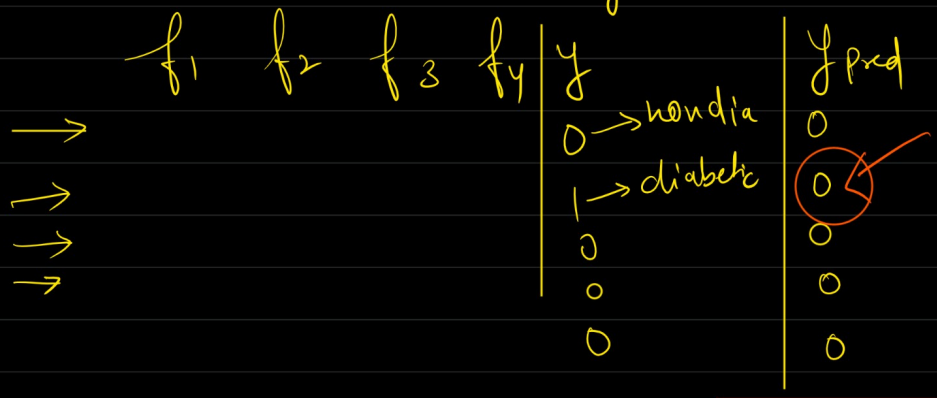
* Assuming you don't build any model → you predict (guess) every thing as class 0
  ⟹ 90% of accuracy.

* Class imbalance
  ↓
  im + balanced class.

→ When one class has very high percentage as compared to other class, this is class imbalance.

90% class 0 ⟶ majority class (non diabetic)

10% class 1 ⟶ minority class (diabetic)



* you don't want your model to learn only the data pattern which is in majority ⟹ you need to deal with class imbalance.

Class Imbalance :- When the difference in the counts / Percentages of both the class is huge.

Class imbalance | 80% → Class 0
20% — Class 1

90% - 10%

95% - 5%

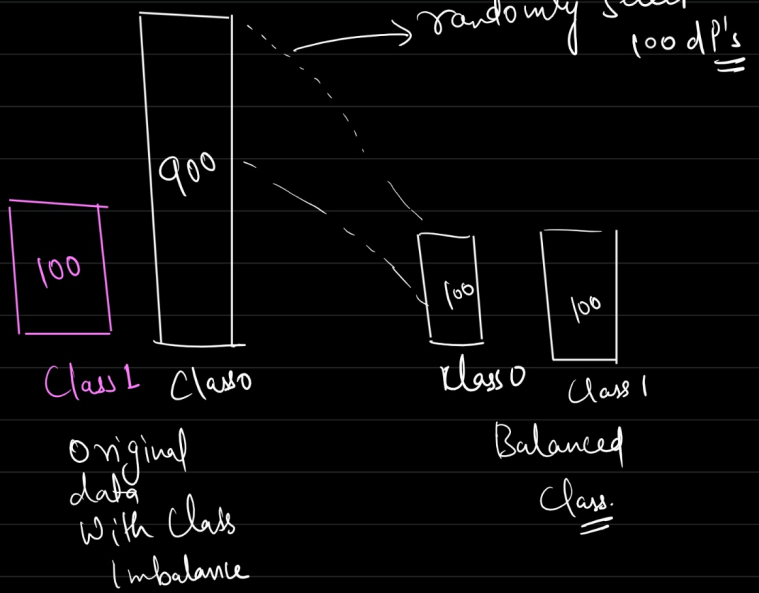
99 - 1%

* Solution to class imbalance

1. Undersampling (down sampling)
2. Oversampling (Upsampling)
3. SMOTE

① Undersampling (Down Sample)

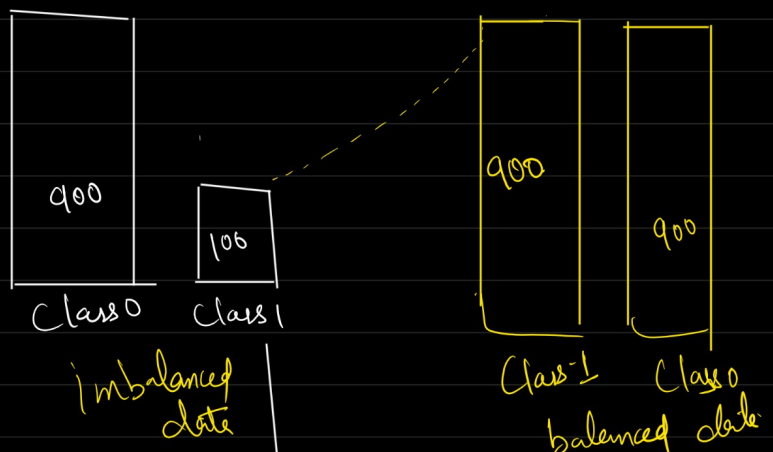
disadvantage — Loss of data.



② Oversampling | Upsample

→ minority class is Upsample

→ repeat the minority class data.



disadvantage

* ML is about learning Pattern in data

→ data is just repeating in case of oversampling

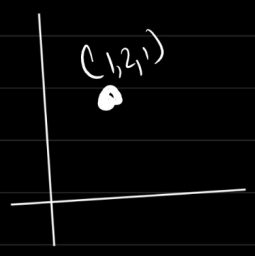
→ No Pattern

→ Noise

Class - 100 dP's

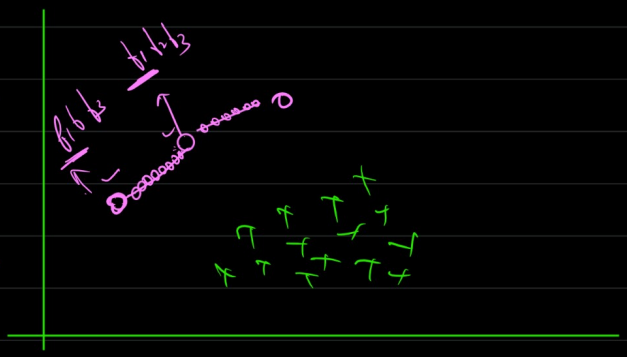
$100 \times 9 = 900.$

	f_1	f_2	f_3	y
→	1	2	1	1
→	1	2	1	1
→	1	1	2	1



③ SMOTE (Synthetic Minority Oversampling technique)

* Problem of No pattern is solved as seen in Oversampling



X → class 0
O → class 1

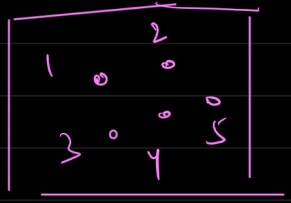
→ Like datapoints (class 0) will always be together this is the reason smote is effective.

How to use in industry

900 : 100
↓ ↓
Undersampling SMOTE
↓ ↓
600 : 400

* In practical implementation of class imbalance.

Population.



With replacement (data point can be repeated)
S_1 (1,2)
S_2 (1,3)
S_3 (2,3)
S_4 (4,2)

without replacement
S_1 (1,2)
S_2 (3,4)
S_3 (5)

↓ data will not be repeated.