imbalanced data

Supervised learning

Classification | Regression

→ Pass / fail
→ Diabetes / Non diabetic

1/0
1/0

2 outcomes of target variable (y)
↳ binary classification

Sugar level (f_1)	Cholestrol (f_2)	y (diabetic / not)
250	102	1
300	100	1
400	200	0
—	—	—

diabetic / Non diabetic

{ 90% → Non diabetic
{ 10% → diabetic

imbalanced class
↓
im + balanced

class (y)
Non diabetic | diabetic

Assuming diabetic is 1 (class 1)
non diabetic – 0 (class 0)

non diabetic — 0 (class 0)

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)

	f_1	f_2	f_3	f_4	y		y_{pred}
→					0	→ non diabetic	0
→					1	→ diabetic	0
→					0		0
→					0		0
→					0		0
→	—	—	—	—	?		

overall ⇒ 90% accuracy

Class imbalance
- → Undersampling
- → oversampling
- → Smote

* class imbalance → when the difference in the counts | Percentage of the both the class is huge.

Class imbalance { 80% → class 0 (non diabetic)
20% → class 1 (diabetic)

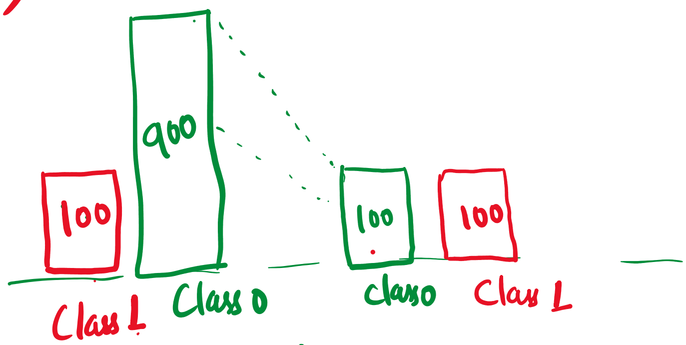
90 — 10%

95 — 5%

99 — 1%

99 - 1%

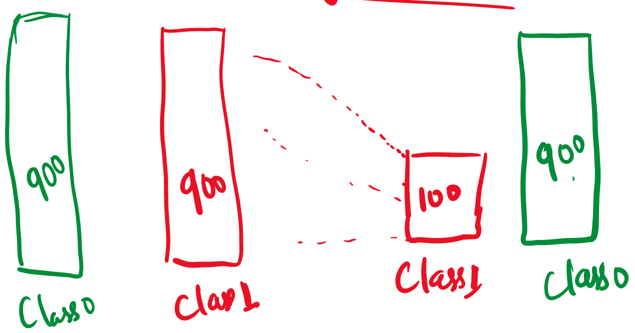
* Undersampling (Down sample)



Original dataset
(class imbalance)

disadvantage
→ loss of data

* Oversampling | Upsample



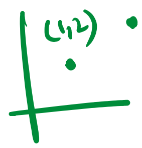
minority class is Upsampled

→ repeat the minority class data.

Class 1 → 100 dPs
→ 100 × 9 = 900

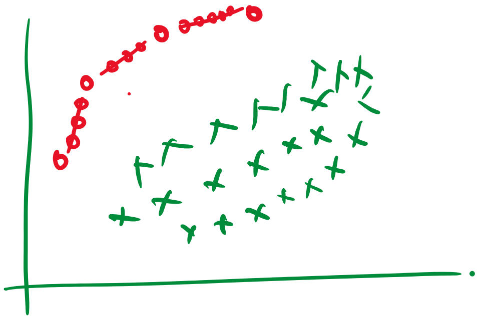
disadvantage
→ No pattern
→ Noise

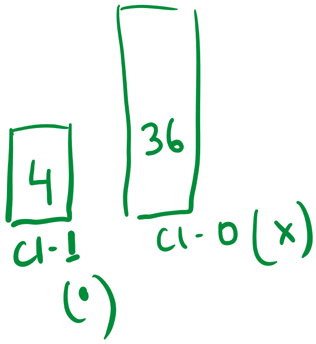
f_1	f_2	f_3
—	—	0
1	2	1
—	—	0
—	—	0
1	2	1
1	2	1
1	2	1



* SMOTE

Synthetic Minority Oversampling technique





How used in industry

900 : 100

↓ ↓

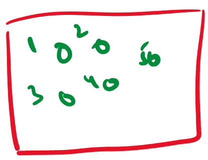
Undersampling SMOTE

↓ ↓

600 : 400

★

with replacement



with replacement	without replacement
S_1 (1,2)	$S_1 = \{1,2\}$
S_2 (2,3)	$S_2 = \{3,4\}$
S_3 (1,3)	$S_3 = \{5\}$
S_4 (4,2)	