01/11/22
5:10pm

Handling Imbalanced Dataset.

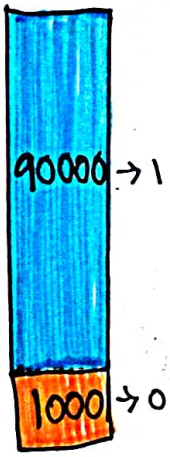
```
df["Virus Present"].value_counts()

Out: 1  200
     0  200
```

— Balanced Data Set

Ex: 50-50%
(or)
45-55%

(+,- = 5%)

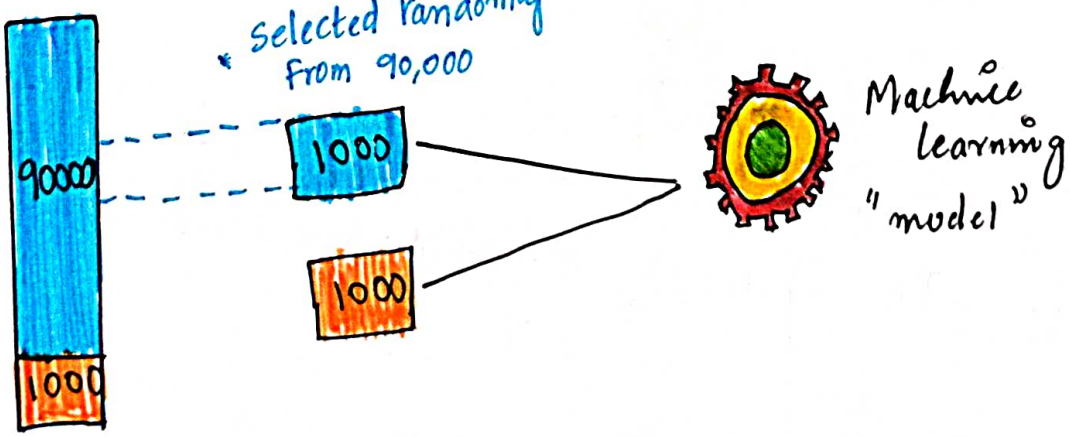
1. Under Sampling major class



Ex:- finance [cibil]
99% [pay]
1% [don't pay]

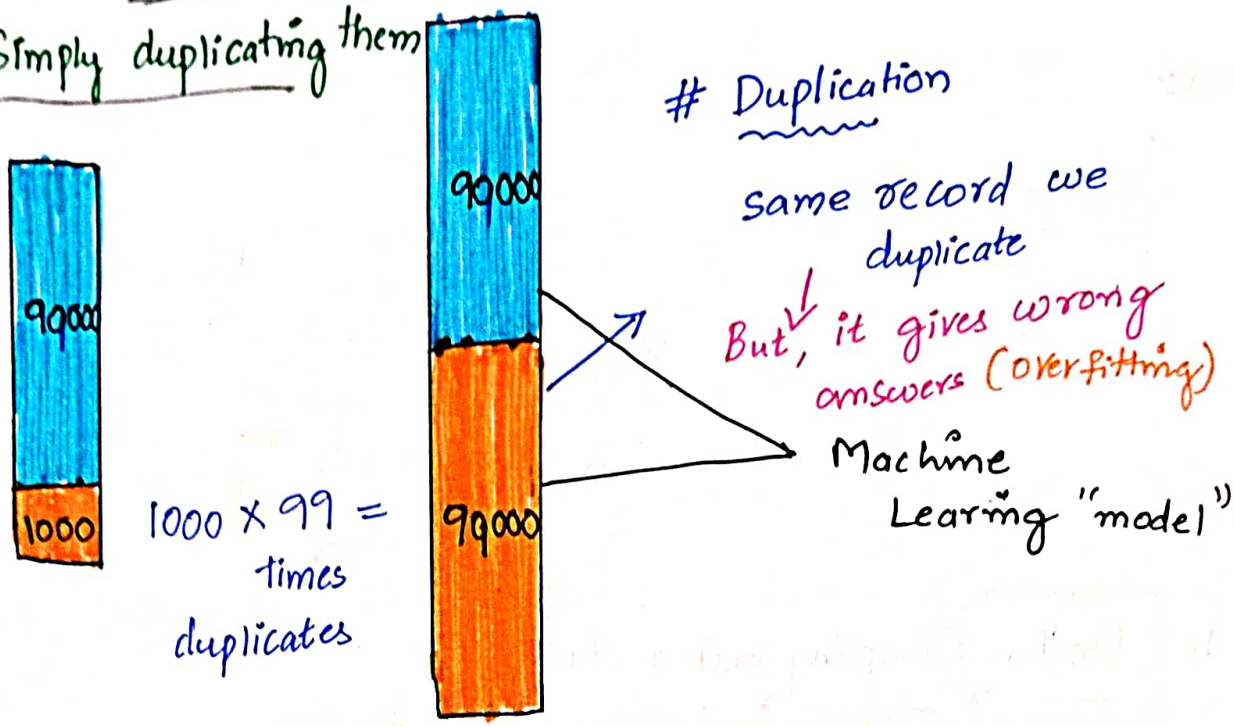
If we apply classification Technique
* it gives always answer as "1"

* Under Sampling



2. Over Sampling minority class by duplication

* Generate new Sample from Current Sample by Simply duplicating them

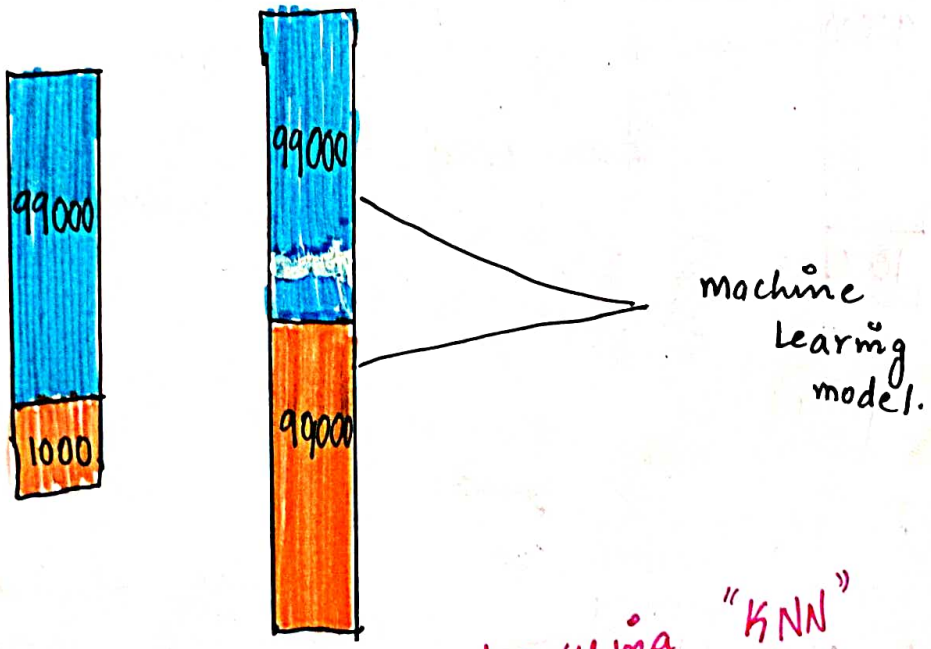


Duplication

same record we duplicate

But, it gives wrong answers (overfitting)

3. Over Sampling minority class using SMOTE



* Generate Synthetic examples using "KNN"

* SMOTE :- Synthetic minority Over Sampling Technique.

Synthetic data

Ex:- Synthetic data.

age : 22

next record, we create

age : 22.1 (which is almost close to 22. But not same data)

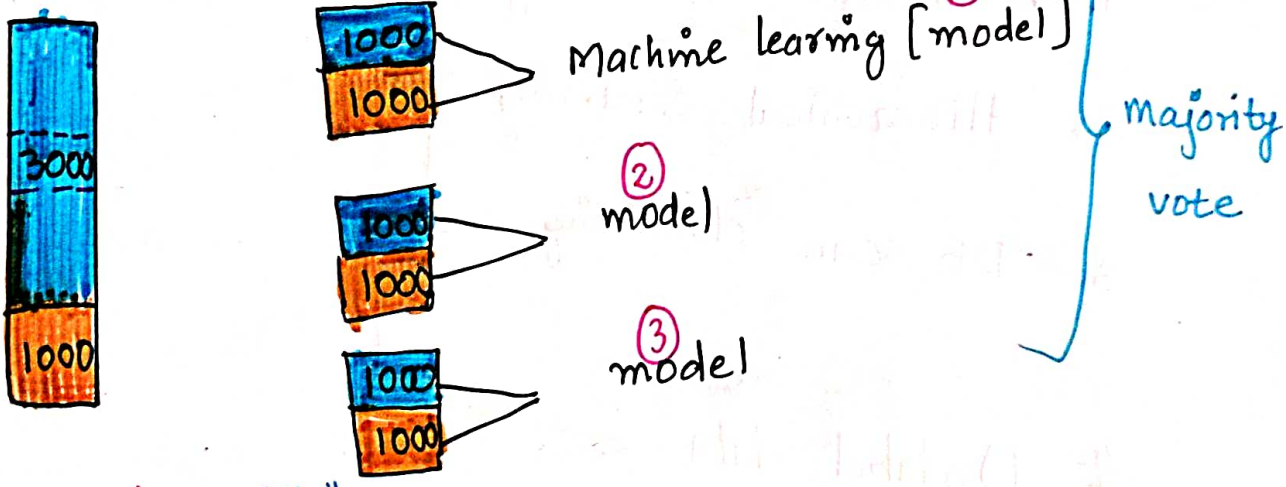
Ex:2

22
21
24
22
28

Avg : 23.4

If we want to create new data = 23.4 (Synthetic data.)

4. Ensemble Method



Ex:- 3100 records, 1000 records → 1100 " 1000 " 1000 records ±5%.

* 3000, divided into three parts.
* Then will make model, by Each part 1, 2, 3 ↓ divided by "minority class"
* And we Take Majority vote.
* Applying of cross validation.

30/4/22 6:00pm