14-09-2025

Agenda:

→ FE - 6
→ EDA - 1

→ Data : house price) FE (1 to 5)
: diabetes dataset (FE - 6)

Imbalanced Data ?

→ classification prob, one class has more sample than other.

F :

company →

↓ M : Majority

survey

↓

biased

↓ ↘ female

Male → ML → Male

conv prediction

1000 people → 90 → ML
→ 910

→ Solution : resampling

910 90

↓ ↑

non-C C

Techniques to handle Imbalance:

- Random Under-Sample (RUS)
- Random Over-Sample (ROS)
- SMOTE (Synthetic Minority Oversampling Technique)
- ADASYN (Adaptive Synthic Sampling)

→ RUS

Male : 100

Female : 30

→ Reduce Majority class by randomly removing sample.

Majority → Male → 30 ⟩ Dataset

Female → 30 ⟩

→ Disadvantage → we are throwing useful data.

→ ROS

→ Duplicate samples from minority class.

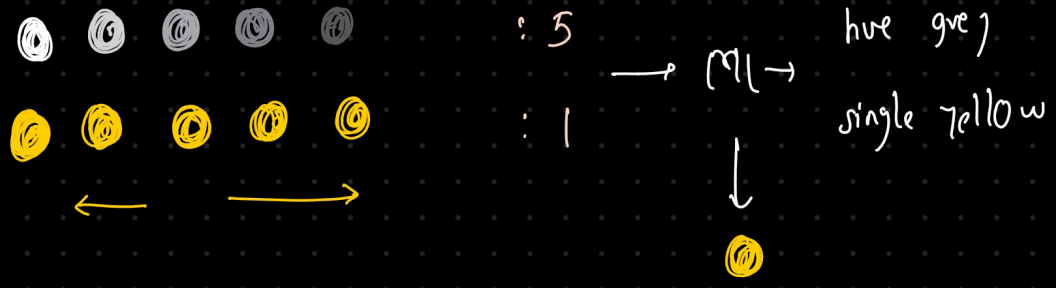
Male : 100

Female : 30 → ROS → Female : 30 + 70

70 ← by duplication

→ Can cause overfitting

↓ ↓ ↓



→ SMOTE

→ instead of duplicating data, create synthic data by interpolating between minority sample.

	Age	Sugar	
A →	50	200	
B →	60	220	
C →	55	210	→ within A & B

Better than ROS

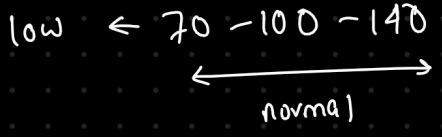
→ works only with numerical data

→ ADASYN

→ similar to SMOTE, but smarter.

→ generate more synthetic samples near boundaries

→



100 people → 70% → 70 - 140

15 → > 140

15 → < 70

→ help generate data of under represented categories

def fun ( )

M : 100

F : 30

50 → smoke

<50 → ADASTN

EDA - I

Dataset → set of analysis → pattern/relationship

data → info → how many colms are numerical?
→ categorical?

read about columns

dt.head ()

categorical → count plot → fix data imbalanced?

→ Analysis → visualization

↓

statistical method

Data → EDA → Feature - Engineering → ML → Deployment

Data → NLP → ML
↓

A → i don't like this topic → negative
→ i like this topic → positive
→ Hello → neutral
→ topic is okay → neutral

Summarize and visualize what the data looks like

- Central tendency (mean, median, mode)
- Variability (variance, std, IQR)
- Correlation & Associations
- Frequency counts. histogram, boxplot, plots...
- handling missing values, outliers

ASK : What is happening in the dataset?

Make inferences or generalizations about the population from the sample

- Hypothesis Test
- Confidence interval
- Effect size
- Regression based inference

ASK : Can we claim this effect exist in the larger population or is it just by chance?

NLP → language

nltk → nlp task → english / french

corpus - collection of text / data

stop words : a, an, the, she, it, on, at
(common words which doesn't add any significant meaning)

n-grams :
sequence of n words that appear together in a text:

n=1, Unigram : movie, good, bad, moral, Hello

n=2, Bigram : great movie, not good, Hello Alok

n=3, Trigram : waste of time, best movie ever

unigram → not, good → not good
not → -ve
good → +ve

Bigram : not good → -ve

Trigram : not good direction

} useful