Feature Engineering & EDA

~~One hot encoding~~

Types ~~of Encoding :~~

Data Science Life cycle :-

1) Data Ingestion (Collection of data)
2) EDA
3) Processing
4) Model
5) Evaluate ~~A~~ Validate Model

} Core ML pipelines.

Statistics - It is a science of collecting, organising and analysing data.

Colleet, Organise, Interpretation, Analysis

↓

Insight

Types of Data

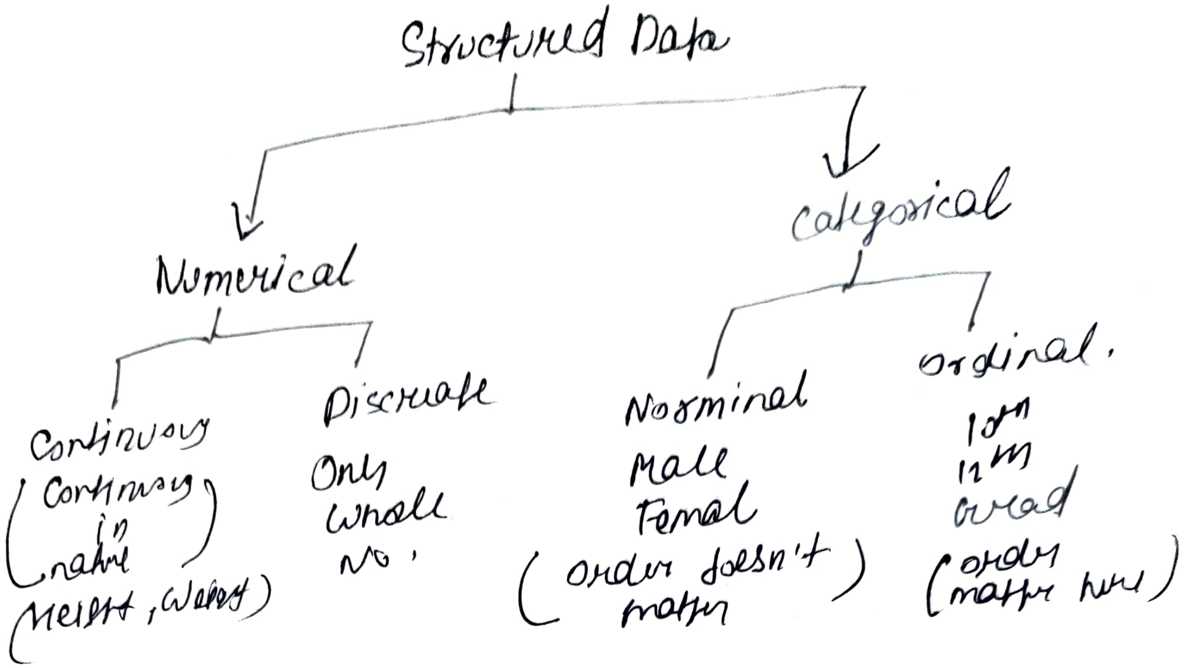
Batch Data : Historical Data, Minibatch Data (Periodic)

Streaming Data : Continuous Data (Live Data)

1) Structured Data : Table (Row x Column) → ML
2) Unstructured " : Videos, Images, Voice, Sound, Text etc.
   ↳ DL
3) Semi-structured : JSON, XML.

Example of Structured Data :-

Feature 1	Feature 2	Feature 3
Weight	Height	BMI



Univariate: Single column.

Bivariate: Two Column.

Multi-variate: More than two column.

Independent & Dependent Variable :-

[Age (Height Sex)] → Independent

Weight → Dependent

Q) First EDA is required or (Preprocessing)

↳ EDA is required.

Order: EDA → Pre-processing ⇒ Models

EDA (Analysis) :-

1) Profiling
2) Statistic Analysis
3) Graph Based Analysis → All types of plots.

Row
column
Missing
Category
Numeric
Duplicate
Dtype
RAM

Variance
Covariance
Std
Correlation
Chisquare test
T-test
Z-test
Anova test
Mean Median Mode

Pre-processiy of Data :-

① Missiy Value Handle
② Outlier Handle
③ Scaliy of Data
④ Transformation -
⑤ Encodiy
⑥ Imbalance Data

⑦ Feature Selection
⑧ Dimension Reduction
⑨ Duplicate value / Duplicate column
⑩ Split / Merge / Drop / Add

There are ~~5~~ steps of feature engi'nr :

① Missiy Null Value (EDA) → Missiy Value Handle (PP)

② Outlier → Handle.

③ Categorical (man, Women) → Encoding.

④ Skewed Range → Scale (within a certain range)

⑤ Count of Feature → { Handle Imbalanced Data, Feature selection, Dimension Reduction (PCA, tsne) }.

Encoding → To change categorical data into numerical data is called encodiy.

Types of Encodiy :-

↳ We are discussiy about categorical var.

Gender < Male, Female

We follow diff encodiy technique to convert it into maths.

Types :-

① Nominal Encoding — For — Nominal cat. var.

② Ordital Encoding — Ordinal cat. var.

↳ Rank

BE
Bcom
Phd
Master

- ① — ① One hot encoding
  - ② One hot encoding with many categorical
  - ③ Mean encoding.
- ② — ④ Label encoding
  - ⑤ Target guided Ordinal Encoding

Dummy variable Trap.

①

State	India	Pak	~~China~~
India	1	0	~~0~~
Pak	0	1	~~0~~
China	0	0	

→ when it is 0 0 then China.

Disadvantage of one hot encoding

In place of country-(pincode)
- So on that time we need to create lots of columns.

↳ For this ④ Label Encoding.

Education	
BE	1
Mas	2
Phd	3

→ We will give rank.

② (One hot encodig with multiple cat)

KDD Orange

Which top 10 categories repeated more

↳ And create 9 columns .

③ Target Guided

f1 ↔ O/p		→ Mean	Classify
A	1		0.73
B	1		0.6
C	0		0.4
D	0		

} Assign rank.

③ Mean Encoding

f1.	O/p	mean
A	1	0.73
B	0	0.6
C	1	0.5
D	0	0.4

Pincode

f1	O/p	
56001	1	0.73
56001	2	0.6

Placed at pincode position.

Why Feature Scaling :-

Features	cm Height	kg Weight	BMI
↳ Magnitude	187	78	
↳ Units	170	84	

→ Scale down this value

① Linear Regression, ② K Means ③ KNN → More feature scaling used.

↳ Global Minima

① Decision tree
② RF
③ xgboost

} No need of feature scaling.

How to handle missing values :—

Cons

① Delete the Rows — Impt data may be deleted.

② Replace the most frequent value → Imbalance data occur.

③ Apply classifier algorithm to predict. → f_2, f_3 & O/p used to predict f_1.

④ Apply unsupervised ML (clustering technique) → Take f_2 & f_3 we start creating into 2 categories

f_1	f_2	f_3	O/p
Male	23	24	Yes
Female	24	25	NO

Handle Categorical Features :—

- Replace each label of the categorical variable by the count.

Cons :- We are losing some kind of info.

	X	count
0	w	150
1	v	80
2	x	72
3	y	75
4		

Handling Ordinal Categories :—
(Ordinal Encoding)

Grade —	A	B	C	Fail
	4	3	2	1

Cons :- Does not add machine learning option.

Sum	Math	
Mon	Bot .	
Tue	Holiday	
Wed	Phy	
Thu	Eng .	
Fri	~~Eng~~ Chem	
Sat	Zoology	

Chem 2 Phy

Sun × Holiday

Eng - Thus
Tue
Eng
Chem ↓ i

Holiday
↑
Phy 1 _ Botany

M_ Phy = Zo _ Eng

Math × Mon ~~x~~

× ~~Thus~~

Math ↓ i

~~Not holiday~~

10	U	
9	.	
8	T	T
7	2	2
6		U
5		
4		
3		
2		
1	S	

R i X 3 R
w

P Q
y 3
W

Mn ANTARCTICA

~~AANAARCEIA~~

NANRATCCI