EDA & Feature Engineering

Data science life cycle:-

1. data ingestion (project) ⇒ (data) (big data tool, remote location, cssl, NOSql) some file format, json xml, excel, website
2. EDA (analysis)
3. Processing (Pre Processing)
4. Model ⇒ machine learning
5. evaluated and Validate

Statistics:- collect, organize, interpretation, analysis

Insight

Scientific, healtcare, social problem

Example:-

Sales of product → sales is going down

* Product, Pricing to customer, leadership, marketing competetor

↳ dataset → analysis → conclusion

1. project manager
2. Business analyst
3. data scientist

} domain Expert

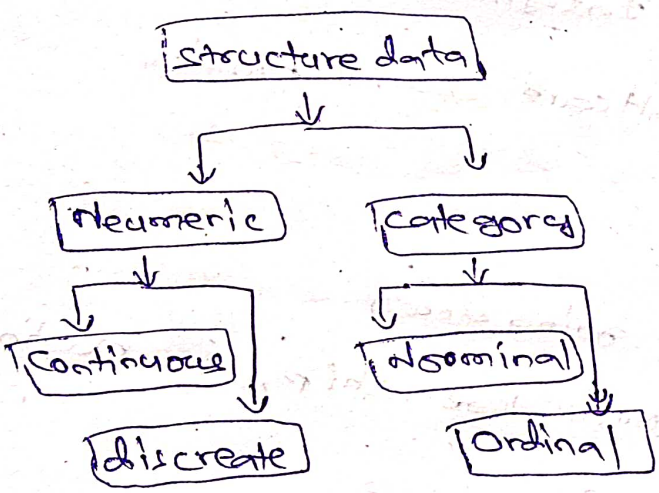
only domain required EDA + feature Engineering

types data :- Batch data , streaming data

Batch data → historic data (periodic)
Batch data → minibatch data (little more frequency)
streaming data → continuous data → live

1. structure data → table (Row x colums) ML
2. unstructure data ⇒ video, images, audio, text ⇒ DL ↳ deep learning
3. semistructure data ⇒ xml , JSon

* EDA + FE



segregate features

weight	Height	Bmi
70	170	22
80	150	24
90	180	25
75	190	26
85	155	27
65	165	28
60	170	29
55		

Continuous Continuous

Continuous :- Height 160, 165, 160.55
weight

discreate :- No. of Bank acco {1, 2, 3, }
childers {2, 2, 1, 4}

Categorey :- → male, female > categorey
black, white > categorey

Nominal :- order doesnot matter
male
female

ordinal :- order matters

10th
12th
Degree
PG
PhD

Example table

Student performance

Name	Age	Height	Sex	weight	education
Ram	24	180	male	50	UG
Arun	22	170	male	60	PG
Priya	23	166	Female	70	UG
maya	25	150	Female	80	Phd
Gaya	26	160	Female	65	PG

Cat → Nominal
Num → Continuous
Num → Conti
Cate → Nominal
Num → Conti
Cate → ordinal

**

Univariate :- single column

Bivariate :- two column

multivariate :- more than 2 column

⇒ more than 2 columns

**

Independent :- column doesnot depend on other colums

Dependent :- column depends on other colums

dependent
weight
Age, height, Sex
Independent

Core ML Pipline →
1. Data ingestion
2. EDA - analysis
3. Pre Processing
4. model building
5. evaluation or validation of model

first EDA is req

Datat analysis

Feature ~~Eng~~ion/Column
1. missing value
2. outlier
3. scaling

EDA → Preprocessing ⇒ model
impact

Example of real life

Biriyani (model)

Raw
1. chicken
2. rice
3. onion
4. oil
5. species

→ EDA →
1kg
1kg
1kg
1lit
500gm

Preprocessing
→ clean
mix
cut

model → Biriyani

test → validation

Name	Age	Edaction	Salary	exp
Sai	25	UG	25K	2
raghu	30	PG	30K	1
ravi	31	UG	40lc	5
keerthi	40	phd	50K	10
Thalu	45	UG	35K	3

1. EDA (analysis)
2. Profile of the data
3. Statical analysis
4. Graph based analysis

Profile of the data

① Raw
② col
③ missing
④ cat
⑤ num
⑥ duplicate
⑦ Dtype
⑧ RAM

} analysis

stats based

① Var
② cov
③ std
④ correlation
⑤ chisquare
⑥ t-test
⑦ z-test
⑧ anova test
⑨ mean | median | mode

} univar
bivar
multivar.

Graph Based analysis

① Box plot → outliers, distribution, skewed, parts
② scatterplot → outlier, linear
③ pie
④ histogram → distribution
⑤ KDE
⑥ count bar → R, C (base)
⑦ heatmap → corr

using univariate, bivariate, multivariate

dashboard, plotting, data analysis

Preprocessing of data

1. missing value handle
2. outliers handle
3. scaling of data
4. transformation (log, Boxes, Square, cube)
5. encoding
6. imbalance data
7. feature selection
8. Dim reduction (PCA, TSNE)

Core ml pipeline

1. Data collected
2. EDA (analysis)
3. Pre processing or FE
4. model building
5. evaluation matrix or validation

EDA

1. Profile
2. stats based analysis
3. Graph based analysis

Preprocessing

1. missing value
2. Outlier value
3. Scale
4. Transformation
5. encoding
6. handle imbalanced
7. feature selection
8. dimentionality reduction (PCA, TSNE)
9. duplicate value / duplicate col
10. Split / merge / drop label

EDA ⇒ Preprocessing ⇒ model

Ways of performing feature Engineering :-

① missing value handle
   ① Random
   ② forword filling / back ward filling
   ③ statical approach
      mean, median, mode
   ④ end of distribution
   ⑤ drop the row
   ⑥ KNN - imputer
   ⑦

② Outlier

detect	handling
Z-score	drop
IQR	median
boxplot	replace trim ming.
scatter plot	
Violin plot	

③ transformation :-
   box plot
   Power transformation
   log
   Square
   Cube

④ scaling
   standardization
   min-max
   unit scaling

⑤ Encoding
   One hot
   label encoding
   Binary encoding
   target guided encoding
   Hash encoding

⑥ imbalanced
   ① under sampling
   ② Over sampling
   ③ cluster based over sampling