EDA & Feature engineering

Data Science Lifecycle
1. data ingestion (Project) → (big data tools), remote location (CSQL, NOSQL), some file format CSV, tsv, xml, json, excel, website → HDFS, NOSQL, kafka, spark streaming
2. EDA (analysis) → stats
3. Processing (Pre) (math.)
4. model
5. evalute & validate.

machine learning

Statistics :- Collect, organise, interpretation, analysis → insight

Scientific, healthcare, social problem

Sales of Product → Sales is going down.

Product, Paying to customer, leadership, marketing, Competitor.

dataset → analysis → conclusion.

1. Project manager
2. Busines analyster
3. data scientist

→ domain expert.

Any domain req. EDA + feature eng.

types data :- Batch data, Streaming data → Data

- mini batch data (little more freq)
- historic data (Periodic)
- continues data → live

1. Structure data → table (RxC) — ML
2. unstructure data → video, images, voice, sound, text — DL
3. Semi structure data → xml, json

Deep learning

EDA + FE →

Structure data

- Numeric
  - Continues
  - discreate
- Category
  - Nominal
  - Ordinal

Segregate feature1 & feature2 & feature3

Example

Weight	Height	BMI
70	170	22
80	180	24
90	190	26
100	200	30
60	160	21
Cont	cont.	cont

Continues in nature

Height 160 160.5 160.55

[160 161]

10, 100, 200 → Whole no.

= 1.5 ≠

Category → male, female > cat

black, white > cat

- nominal → order does not matter

male ↓

female ↑

Ordinal → 10th, 12th, Grad., P.g., Phd.

Student Performance

feature

Univariate

bivariate

multivariate

ML

Supervised machine learning

Name	Age	Height	Sex	Weight	Education
Sunny	25	170	male	70	UG
Arijit	30	180	male	80	PG
Priyam	35	160	male	60	UG
Priya	20	150	female	55	Phd
Aditi	27	145	female	58	PG

first-level

Cat, Num, Num, Cat, Num, Cat

25.5

nominal (order ~~matter~~)

Continues

Univariate → single col

bivariate → two col

multivariate → more than 2

Data

col/feat

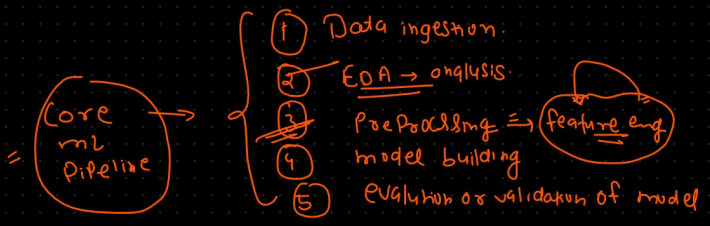
Uni, bi, multi

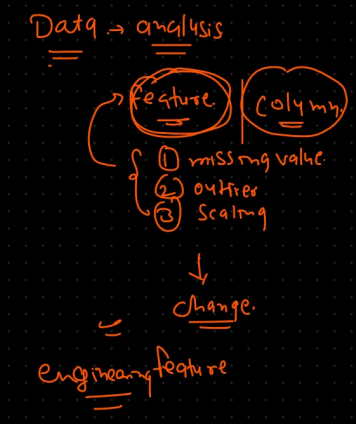
Independent | Dependent

Age, height, Sex → Weight

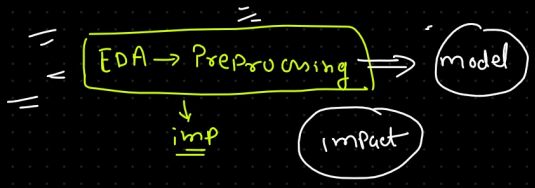
Dependent

independent

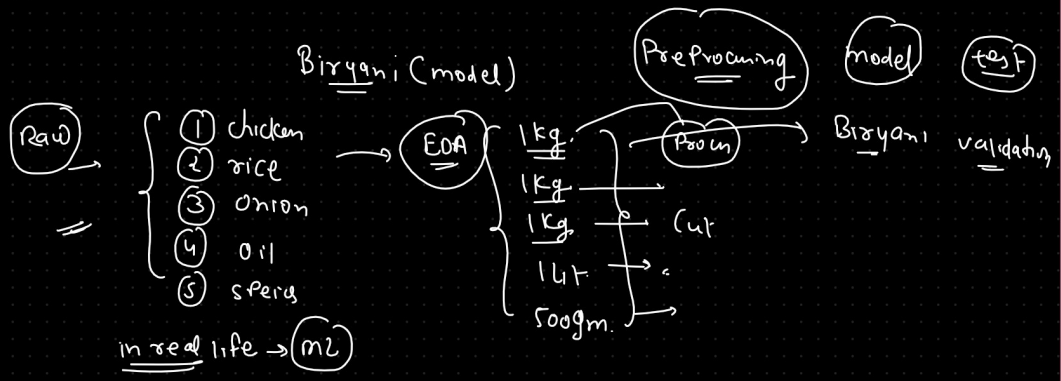




First EDA is req. or FEUTP? → EDA



EDA
PreProcessing



1. EDA
2. PreProcessing

break till → 4:30

Univariate.
bivariate.
{ Q. EDA
Q. Preprocessing or FE

① EDA ② Preprocess.

ID feature
↓
Conclusion

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

NAME	AGE	Education	Salary	Exp
Sunny	25	UG	25K	2
deepak	30	PG	30K	3
rushi	40	UG	40K	5
Aman	50	Phd	50K	10
Shalini	20	UG	35K	1

{ Sunny
25
DS
BLR }

Profile of the data.

① Row
② col
③ missing
④ Cat
⑤ num
⑥ duplicate
⑦ Dtype
⑧ RAM

= Analysis

= Stats based (Interpretation)

uni, bi, multi

① Var
② Cov
③ Std
④ Correlation
⑤ Chisqure
⑥ t-test
⑦ z-test
⑧ anova test
⑨ mean / median / mode

uni, bi, multi

dashboarding
data analysis
Plotting

Graph based analysis

① Box Plot → outlier, distribution Statical, Profile
② Scatter Plot → Outlier, linear
③ Pie
④ histogram → distribution
⑤ KDE
⑥ Countbar R, C (bar)
⑦ heatmap → Corr

EDA
= Observation.
Conclusion

⇒ Based on a EDA can be do a Procning of the Data ?
analysis

no. of way

Preprocning of data

① missing value handle
② Outlier handle.
③ Scaling of data.
④ transformation (log, Boxcox, squre, cube)
⑤ encoding
⑥ Imbalance data.
⑦ feature selection
⑧ Dim reduction (PCA, tSNE)

Feature eng.

missing null value → missing value handle

EDA → PP

Outlier → handle

1000 → Subset → 200

Cat (man, woman) → encoding

100

Skewed range. → Scale (with in a certain range)

1) 10
2) 40
3) 20
4) 10
5) 20

(5) ← Count of feature → handle imbalance

→ feature selection

→ dim red (PCA, tSnE) LDA

male → 1

female → 0

$y = mx + c$

$y = m(male) + c$

$0 \times male$

$y = mx1 + c$

int

Sex. → Encoding → Numerical. → ML model.

m
F
m
F
m
m
F

- Practical.

EDA
1) Profile.
2) State.
3) Graph.

Data.

Preprocessing - Past research

1) missing → n
2) outlier → n
3) Scale → n
4) transform → n
5) encode → n
6) Imbalance → n
7) drop | duplication → n
8) feature selection
9) dim red

automated in Python

EDA

Pandas Profiling

→ auto

Knime

= to only 1 with respect to = atleast 3 automated tool maxs

Dataset (10) → 1 → atleast 3 automate
max 5 tools

EDA + feq. Eng → 10 in depth in your github repo

analysis
FE

automate.

Pandas Profiling

↓
data
↓
entire eda.

pp → again i will visit

- automated

Pick any
complet

tomorrow by
12 PM IST

Sunny.savita@neuron.ai

Krish.naik@neuron.ai