EDA & FE

EDA -> Exploratory Data Analysis
FE -> Feature engineering

DS -> life cycle

1.] Data collection & Ingestion
↓
EDA [Exploratory Data Analysis]
↓
Pre-Processing the Data
↓
Model building
↓
Evaluate the Model

Statistics

* Statistics is science of collecting the Data & organise, Analyze the data.

* After this we will get some insight from the Data after the Statistics Implementation.

⇒ So, the first point is statistics Analysis in EDA section.

Real time Scenario why we need ML

Example

Sale of Product
↳ Sale is going down

Reasons

Product, paying to customer, leadership, marketing, Competitors

going to collect the dataset to determine the solution. with help of Analysis.

Role of persons (Involved in any kind of above problem)

1) Project Manager
2) Business analyst
3) Data Scientist

→ give the conclusion.

Any kind of data set we have to do Analysis

Famous Data Set's

1) Titanic
2) IRIS
3) Diabetes

According to Data set

* It can be available per any of the Location like web data tools, there will be a remote location to get the data like (SQL, NoSQL)
* It can be different file format CSV, Pdf, XLSX, JSon, XML.

Types of Data

After collecting the data

Generally, Data can be Batch data, (historic Data)
Streaming data (Continuous data)
redancy of

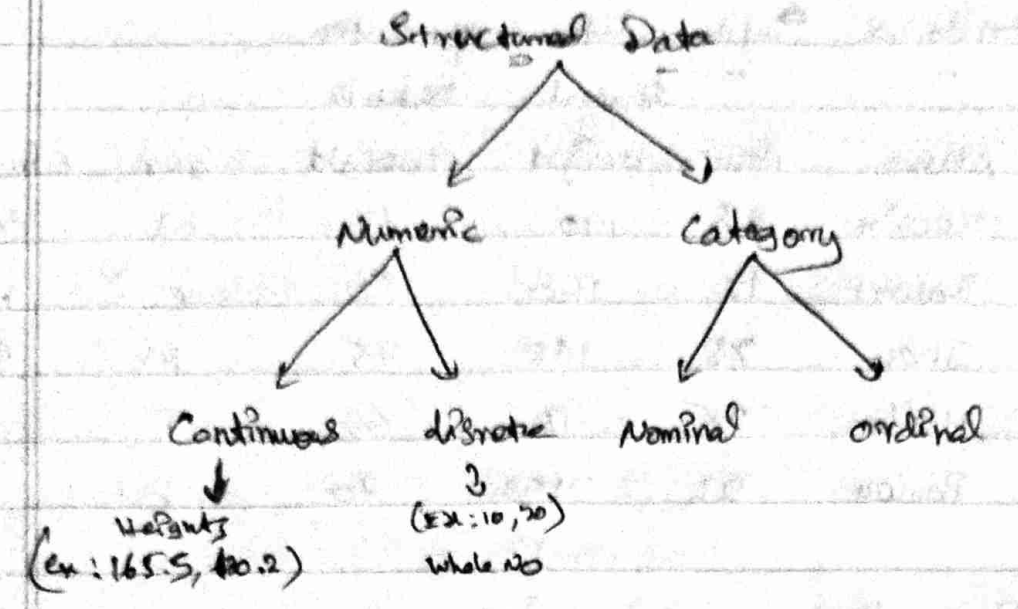
1) Structured data -> (Table) -> ex
2) UnStructured data -> (ex: Img, Vid, Vo
3) Semi Structured data -> (XML, JSon)

EDA + FE

Structured Data

-> feature-1 weight | feature-2 weight | feature-3 BMI

70	170	22
80	180	24
90	190	26
100	200	30
60	160	27



Category [nominal, ordinal]

Male
Female } Category

Blood groups } Category

Nominal

* order doesn't Matter thats called Nominal

Ex: we can have any order of the data can be fine with the Data Set.

Male	Female
Female	Male
Male	Female

This order, It can't be make any diff in Data set

ordinal

* order should be important

Ex: 10th → (1)
12th → (2)
Degree → (3)

Practical Implementation of Data

= Student's details

S.no	Name	Age	Height	weight	sex	education
1	Kevin	28	170	72	M	G
2	Robert	17	140	56	M	Hsc
3	Jhon	38	188	75	M	PhD
4	Hellen	25	160	62	F	UG
5	Bailey	24	156	70	F	PG

First level

* Categorize the features [Colums]

Name	Age	Height	weight	Sex	education
↓	↓	↓	↓	↓	↓
Category	Numeric	Numeric	Numeric	Cate	Cate
↓	↓	↓	(count)	(Nominal)	(Ordinal)
(Nominal)	(continous)	(continous)			

HSC - 1
UG - 2
PG - 3
PHD - 4

Next Analysis

UNIVARIATE → Single Colm
BIVARIATE → two Colm
MULTI → more than two colm

1) If want to check Age's of the students → Univariate Analysis

2) If want to check Height with respective age of the person → Bivariate Analysis

3) If want to check sex with Age & Height → Multivariate Analysis.

Independent / Dependent

1) If we want only —
   Age, Name, Height means it is Independent

2) If we want weight of those people means we depending the people age, Height so it was Dependent.

One Pipeline of ML

1) Data ingestion
2) EDA
3) Pre-Processing (FE, FS)
4) Model Training
5) Evaluating the Model (validation)

→ Data
→ EDA
→ i) missing values
   ii) outliers
   iii) Scaling
→ Change

EDA is

Exploratory Data Analysis is an approach to Analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of Statistical Summary and graphical representations.

EDA

Name	Age	Education	Salary	exp
A	23	IT	20k	2
AB	25	BE	25k	2
BC	22	BSc	18k	1
CD	20	DSE	15k	1

EDA Analysis

1) Profile of the data
2) Statistical analysis
3) Graph base analysis

1) profile of data
   1) # Rows
   2) # Column
   3) # missing value
   4) # numeric value
   5) # Column value
   6) Data types
   7) # duplicates
   8) How much ram occupied

2) Stats based Analysis [univariate, bi-variate, multi]
   1) Variance
   2) Cov
   3) Std
   4) correlation
   5) Chi-Square Test (or) T-Test, (or) Z Test or Anova Test
   6) mean, median, mode

3) Graph base analysis
   1) Box plot
   2) Histogram

3) Scatter plot
4) Pie chart
5) PDF (Normal Density estimation)
6) Count plot
7) heatmap

* Based on EDA we can do a Processing of the Data.

Pre-Processing (FE)

1) Handle the missing values
2) Outlier handle
3) Scaling of data
4) Transformation (log, BoxCox, Square, Cube)
5) Encoding
6) Imbalance Data
7) Feature Selection
8) Feature Transformation (Dimension Reduction (PCA, tSNE))

EDA → FE
missing value (null values) → Handle the missing value
outliers Detection → Handle the outliers
categorical (M, F) → encoding (0, 1)
skewed (High range) → Scaling the data in certain range
Count of each feature → Handle Imbalance & balance Data

If we have subset of feature → Select the particular Some of feature is called feature Selection

Dimension reduction (PCA, tSNE)
two feature f1, f2 → one feature (X)
* Here we reduce the Dimensions.

EDA: Automated tool in Python

1) Pandas Profiling
2) mlpto
3) Dtale
4) Sweetviz
5) Autoviz
6) D-Tale