Feature Engineering And Exploratory Data Analysis (EDA)

Data Science Life Cycle (ML Life Cycle)

1. Collecting of Data
2. EDA (Analysis)
3. Processing (Pre – Processing)
4. Model
5. Evaluate and Validate Model

Statistics:- Statistics is the science of collecting, organising and analyzing the data.

Collecting, Organise, Interpretation, Analysis ----------> Insight

Examples:- Scientific Problem, Healthcare, Social Problem.

<u>Problems Statements:-</u>

Sales of Product ------> Sales is going down

Analysis ------> Product, Paying to Customer, Leadership, Marketing, Competitor.

Data Set ------> Analysis ------> Conclusion

- Project Manager
- Business Analyst **(Domain Expert)**
- Data Scientists

Only Domain require :- **EDA & FE** ------> Used to get conclusion from the dataset.

Collecting of Data:-

Examples: Big Data Tools, SQL, NO-SQL, Website, Webscraping, Some files format: (CSV, XML, JSON)

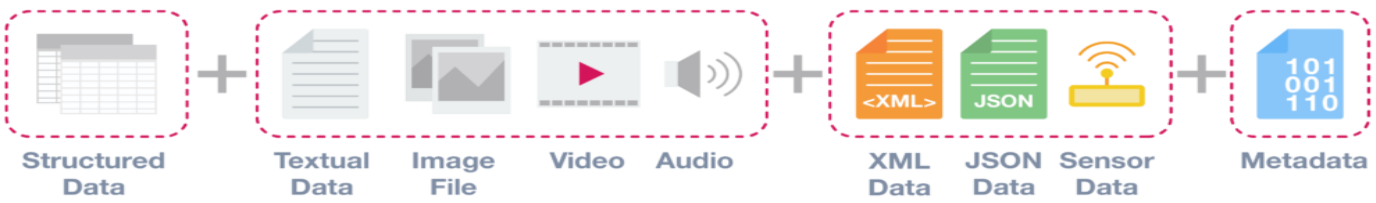
<u>Types Of Data:-</u>

Batch Data: Historical Data, Mini batch Data (Periodic)

Streaming Data: Continuous Data (Live Data)

<u>Data</u>

1) **Structure Data :** Table (Row Column) ------> ML
2) **Unstructure Data:** Video, Images, Voice, Sound, Text -------> DL
3) **Semi Structure Data:** JSON, XML.



Structure Data			
Quantitative (Numerical)		**Qualitative** (Categorical)	
Continuous	**Discrete**	**Ordinal**	**Nominal**
Consists of numerical values that can be measured but not counted.	Consists of numerical values that can be counted.	Consists of text or labels that have a logical order.	Consists of text or labels that have no logical order.
e.g. Weight, Height, Age, Speed {56.06 Kg, 87 Kg}	e.g. whole no. pin code, No. of children {0, 1, 2, 3}	e.g. Beverage size {small, medium, large}, 10^{th}, 12^{th}, Grad, UG, PG, Phd	e.g. Profession {chemist, carpenter}, Male, Female

Students Performance (Dataset)

F --> Feature

Name (F1)	Age (F2)	Height (F3)	Sex (F4)	Weight (F5)	Education (F6)
Nikhil	20	183	Male	65	PG
Prakriti	19	158	Female	-	UG
Categorical	***Numeric***	***Numeric***	***Categorical***	***Numeric***	***Categorical***
Nominal	***Continue***	***Continue***	***Nominal***	***Continue***	***Ordinal***

Used to find types of data:

Univariate: Single Column

Bivariate: Two Column

Multivariate: More than two columns

Independent And Dependent Variable:

Age, (Height, Sex) ----> Independent

Weight ----> Dependent

Q. First EDA is required or FE or Pre-Processing.

Ans. EDA is Required (Analysis of the data based on the given data).

EDA ----> Pre-processing ----> Model

EDA (Analysis):

1. Profile of the Data
2. Statistic Analysis
3. Graph Based Analysis

Profile of the Data:-

Analysis

1) Row
2) Column
3) Missing
4) Category
5) Numeric
6) Duplicate
7) DType
8) RAM (D Size)

Statistics Analysis (Interpretration):-

Univarate, Bivariate, Multivariate

1) Variance
2) Covariance
3) Standard deviation
4) Correlation
5) Chi Square Test
6) T-test
7) Z-test
8) Anova test
9) Mean, Median, Mode

Graph Based Analysis (Plotting):-

1) Box Plot ---> Outlier, Distribution, Statistical Profile.
2) Scatter Plot ---> Outlier, Linear or not
3) Pie Chart
4) Histogram ---> Distribution
5) KDE Plot
6) Count Bar ---> Used for counting (Bar Chart)
7) Heat Map ---> Correlation

Q. Based on EDA, can we do a pre-processing of data? ----> Ans. Yes

Pre-Processing of data:-

1) Missing Value Handel
2) Outlier Handel
3) Scaling of Data
4) Transformation (Log, Box-cox, Square, Cube)
5) Encoding
6) Imbalance Data
7) Feature Selection
8) Dimension Reduction (PCA, LDA, TSNE)
9) Duplicate Value / Duplicate Column
10) Split / Merge / Drop / Add (column)

These are steps of Feature Engineering:

1. Missing Null Values ----> Missing value handel
2. Outlier ----> Handel
3. Categorical ----> Encoding (to change categorical data into numerical data)
4. Skewed Range ----> Scale (Within a certain range)
5. Count of Feature ----> Handel Imbalanced data, Feature Selection, Dimension Reduction.

Some Automated Tools In Python For EDA:-

1. Pandas Profilling
2. Sweet Viz
3. AutoViz
4. D-Tale
5. Mito
6. Knime
7. Dataprep

Ways of Performing Feature Engineering:

1. **Missing Value Handel**
   I. Fill with Random number.
   II. Forward/Backward Filling.
   III. Statistical Approach (Mean, Median, Mode).
   IV. With the help of end of distribution, fill the Missing Values.
   V. Drop the row.
   VI. Impute with KNN (KNN-Imputer)
   VII. ML Algorithm for Missing Values.
   VIII. Build own ML Model to predict Missing Values.

2. **Outlier Handling**
   Detect the Outlier
   I. Z-Score
   II. IQR
   III. Box-Plot
   IV. Scatter Plot
   V. Violin Plot

   Handel Outlier

   I. Drop
   II. Fill with Median/Mean/Mode
   III. Replace / Trimming

3. **Transformation**
   I. Box Cox Transformation
   II. Power Transformation
   III. Log
   IV. Square
   V. Cube
   VI. Yeo Johnson

4. **Scalling of Data**
    I. Standardization
    II. Min Max Scalar
    III. Unit Scaling

5. **Encoding Various Methods**
    I. One Hot Encoding
    II. Label Encoding
    III. Binary Encoding
    IV. Target Guided Encoding
    V. Hash Encoding

6. **Imbalanced Dataset Treatment Various Methods**
    I. Collect More Data
    II. Under Sampling
    III. Over Sampling
    IV. Cluster Based Over Sampling

How to find the best Model Accuracy Various Methods:

a. To increase the accuracy, We need to change the pre-processing technique and use different method or steps from the above.
b. We needs to use each and every pre-processing steps and find the best accuracy.

Q. How do we transform the data?

Import numpy as np

np.log(df)

sns.displot(df)