**Machine Learning Pipeline**

Data Handling

Data collection: (data extraction)

Data description

Data Ingestion: (data Loading)

Data transformation: (data quality)

Data Wrangling

Typecasting data, String Manipulations

Missing Values, Outliers

Duplicate values

Filtering data based on condition

EDA (stats, summarization), Visualizations

Feature Preprocessing

Feature Engineering

Feature Scaling

Dimensionality Reduction: Feature selection

Feature projection

Model Building:

Model development

Hyperparameter Tuning & Optimizations

Model evaluation

Model testing

MLOP’s:

CI/CD

Monitoring and maintenance

Model retraining

**Data Wrangling**

**Data cleaning & transforming :**

1. String manipulations, Typecasting data
2. Imputing Missing values
3. Removing Duplicate rows
4. Filtering data based on condittions
5. Handling Outliers

**Exploratory Data Analysis (Stats) :**

**Descriptive Statistics:**

1. Measure of central tendency : Mean, Median, Mode
2. Measure of Dispersion : Variance, Standard deviation
3. Measures of distribution :

* Moment-1:Mean
* Moment-2:standard deviation
* Moment-3:Kurtosis
* Moment-4:Skewness

1. 5 number summary:

* Count
* Range(Minimum - Maximum)
* Quartiles(Q1,Q2:Median,Q3)

1. Different Distributions :

* Normal distriution/Guassian distriution/Standard normal distribution/z-score
* Log normal distribution
* Bernoullis distriution
* Binomial distribution
* Poissons distriution
* Multinomial distribution
* Exponential distribution
* uniform distribution

**Inferential Statistics:**

1. Central limit theorem, Confidence interval
2. TOH concern 1 mean,2 mean : n<30 ( t-test)

n>=30 (Z-test)

1. Test of hypothesis concern : 1 proportion,2 proportion (Z-test)

several proportions (Chi-Squared test)

1. Test of hypothesis concern : 1 Variance (Chi-Squared test)

2 Variance (F-Test)

1. T-test, Z-test, Annova test, Chi-square test, F-test

**Visualizations**

**Feature Preprocessing**

**Feature engineering**

Numerical data:

* Binarization
* Rounding
* Binning
* Interactions
* Transformations

Categorical data:

* Transforming nominal features
* Transforming Ordinal features
* Encoding Categorical features

**Feature Scaling:**

* Standardization (Z-score normalization)
* Min-Max Scaling
* Roust Scaling
* Maximum absolute scaling

**Dimensionality Reduction:**

**1. Feature selection:**

Measures of association:

* Covariance
* Pearson’s corelation & causation
* Spearman's rank correlation coefficient
* Mutual Information

**2. Feature projection:**

* Manifold Learning (t-SNE, UMAP)
* Principal Component Analysis (PCA)

Independent Component Analysis (ICA)

* Singular Value Decomposition (SVD)
* Linear Discriminant Analysis (LDA)

Quadratic Discriminant Analysis

Generalized Discriminant Analysis (GDA)

* Non-negative matrix decomposition

Sequential Non-negative Matrix Factorization (NMF)

**Machine Learning Models**

**Supervised learning**

(Classification, Regression)

1. Naive bayes theorem
2. Linear Regression:
3. SVM/Gaussian Process Regression
4. KNN
5. Decision trees
6. Ensemble methods:

Classification tasks : Takes the majority voting

Regression tasks : Takes average of the results

**Homogeneous ensemble models:**

* Bagging : Random Forest, Extreme Randomized Trees
* Boosting :
* Adaptive Boosting
* Gradient Boosting Machine (GBM)
* Stochastic GBM
* Extreme Gradient Boosting (XGBoost)
* Light GBM
* Histogram-Based GBM
* Categorical GBM

**Heterogeneous ensemble models:**

* Stacking
* Blending