U-2 (ML)

DOMS | Page No.
Date / /

Q1 Feature Engineering →

Raw data → set of meaningful insights & features.
↳ improve performance of ML model.

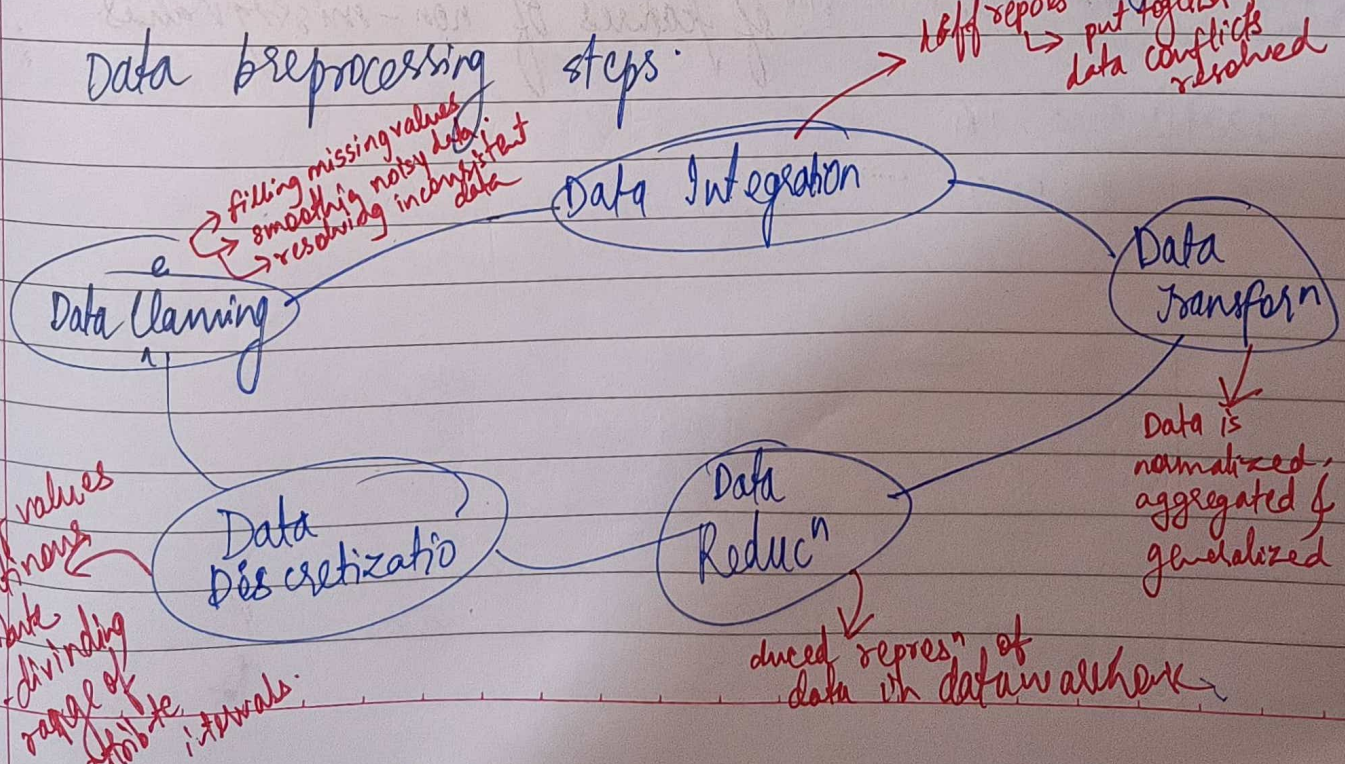
1) Feature Selection :- subset of most relevant features from original set of variables

2) Feature Extraction — creates new features by applying mathematical or statistical transformation.

3) Feature transformation :- • existing feature → more suitable for model
• scaling, normalization, etc.

4) Feature Creation. → • new feature → domain knowledge or prior experience
• specific patterns or rel^n not evident in original data.

Q2 Data preprocessing steps.



DOMS | Page No.
Date / /

Steps → Data Preprocessing -

1. Get dataset
2. Importing Libraries
3. " _ " Dataset
4. Handle missing Data
5. encoding categorical data
6. split data → Train/Test
7. Feature Scaling

Q3 Handle missing data in dataset.

1) Ignore the tuples
2) Fill in missing value manually
3) Use global constant to fill missing value.
4) Use attribute mean to fill missing value
5) Use attribute mean for all samples belongings to same class as given tuple.
6) Use most probable value to fill in missing value -
7) Impute it with mean, median or mode of features of non-missing values.

DOMS | Page No.
Date / /

Q4 feature selection Techniques.

↳ Method of reducing input variable to user model by using only relevant data & rid of noise in data.

- reduce dimensionality of feature space
- speed up a learning algorithm
- improve predictive accuracy
- improve comprehensibility of learning results

Filter Method	Wrapper methods	Embedded Methods.
• generic method ↓ don't incorporate specific ml algo.	• evaluates on specific ml algo	• features → model building process. feature selection → each iteration of model.
• faster than wrapper method	• High computation time	• In b/w wrapper & filter time take.
• less prone to overfitting	• high chance of overfitting	• reduce overfitting by penalizing the coefficient of model.
• ex → chi-square test	• Forward, Backward, Stepwise selection	• ex → LASSO.

DOMS | Page No.
Date / /

Q5 Statistical Measures → Feature Engineering

selecting & transforming variables/features in dataset for creating predictive model.

- Count based feature selection → 1) count of individual values within column

2) idea of distribution & range of values

1) Mean → Mean of dataset is avg of all data value.

Sample Mean = $\bar{x}$ = Sum of values of the observation / total no. of observⁿ in sample = $\frac{\sum x_i}{n}$

Popl^n Mean = M = Sum of value of N observⁿ / No. of observⁿ in popln = $\frac{\sum x_i}{n}$

2) Median → data set is value in midden when data items are arranged in ascending order.

3) Mode → value that occurs is greatest frequency

DOMS | Page No.
Date / /

Q6 PCA -

- Principal Component Analysis
- unsupervised ml algo
- used for dimensionality reduction.
- converts observation of correlated fuetion into set of linearly uncorrelated features.
- new transformed features → principal component.
- popular tool → EDA & predictive modelling.
- technique to draw strong patterns from reduced variance.
- Appln → image preprocessing, movie reccomendation system, optimizing power allocation.
- contains imp variables & deletes/drops least important variable.

Q7 Multi Dimensional Scaling (MDS)

- non-linear technique → embedding data in lower-dimensional space.
- map points residing in MDS to LDS.
- dimensionality reduction → input data → not linearly arranged
- visual representation of distances or dissimilarities b/w of high dimensional data.
- iterative & minimize diff b/w pair of points in original data.
- preprocessing step → dimensionality redn → classifn & regression problem

MDS → Metric MDS / classical MDS → preserve pairwise distance/dissimilate measure as much as possible.

↓
Non Metric MDS
↳ ranks of dissimilarity metric is known.

Q8 Matrix Factorization -

- decomposing matrix in product of other matrix.

appln → collaborative filtering

- matrix → smaller matrix → uncover hidden patterns, simplify computations, etc.

- used in recommendation system.
  ↳ decomposes user item interaction data into user & item matrices to generate personalized recommendations based on latent features.

Q9 feature scaling - • every feature in same footing without any upfront importance.

- algorithm performance improvement.
- preventing numeric instability
- each characteristics → same consideration.

1) Min-Max Scaling → find min & max value of column
   ↳ subtract min from entry & divide result by diff b/w max & min.

values are shifted & rescaled so their range can vary from 0 to 1.

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2) simple feature scaling → divided each value by max. value for feature.

DOMS | Page No.
Date / /

Q9 Normalization

↳ data preparation technique
↳ transforming columns in a dataset to same scale
↳ features more consistent with each other → model → accurate output
↳ make data homogenous over all records & fields.
↳ rescaling real-valued numeric attributes into 0 to 1 range.
↳ algo ⇒ KNN, SVM, Neural Networks, principle networks

Types → 1) Z-score or standard score.

• values are normalized based on mean & standard deviation of data.

$$X_{new} = \frac{X_{old} - \mu_a}{\sigma_A}$$

μ_a - standard mean
σ_A - standard deviation

Q10 - PCA helps in dimensionality reduction by -
• reduce no. of features or variable in dataset.

1) Data Compression -
2) Variance Maximization
3) Dimension Ranking
4) Noise Reduction.
5) Interpretability
6) Improved Model Performance
7) Visualization
8) Feature Engineering
1) Trade off → amt of variance retained & no. of dimensions reduced.

DOMS | Page No.
Date / /

Q11- feature extraction & types → Kernel PCA.
Local Binary Pattern.

- dimensionality reduc^n technique
- transform high dimensional data into lower dimensional data while preserving essential info & patterns.

Kernel PCA -

- extension of standard PCA.
- Kernel method capture non-linear relⁿship in data
- by mapping into high dimensional data.
- result is nonlinear feature → analysis of modelg.

Applⁿ → PCA insuffient → Kernel PCA helps.
. in OpenCV, bio informatics &, etc.

Local Binary Pattern (LBP).

- Text Descriptor → used in img analysis & computer vis
- quantifies local texture patterns in image by comparing neighbouring pixels.
- texture analysis, object recognition, face recognition, c
- can capture intricate texture details.

DOMS | Page No.
Date / /

Q12 Backward & Forward Selection Process.

- Used for feature selection in ML.
- choose subset of relevant features from original.
- improve model performance, reduce overfitting, etc.

* Forward selection → • Iterative Process.
• starts with empty set of features & gradually adds one feature at a time to build final feature subset

Steps → 1) empty set 2) evaluate feature 3) add selected feature
4) Iterate 5) Final Model → has selected features.

* Backward selection –
starts with all available features & progressively removes one feature at a time to determine final feature subset.

steps → 1) all feature - empty list / track of demerated features
2) evaluate all feature
3) Remove Least significant
4) Iterate
5) Model → has required subset of features.