Machine Learning Preprocessing Summary (Viva Ready)

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
1	**Missing Data**	Dataset-এর যেই জায়গায় value নেই, সেটা handle করা	Model missing value বুঝতে পারে না	Age: 20, 25, NaN → NaN fill/remove করা
	Methods	Imputation, Deletion, MICE		
	Imputation	Missing value অনুমান করে বসানো	Data loss কম হয়	Salary missing → average salary বসানো
	Deletion	Missing row/column remove করা	যদি missing অনেক কম হয়	1000 row-এর মধ্যে 5টা missing row delete
	MICE	অন্য feature-এর relationship দেখে missing value predict করা	Complex dataset-এর জন্য	Age + Experience দেখে missing salary predict

---

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
2	**Data Cleaning**	Dataset-এর ভুল data পরিষ্কার করা	Data quality improve করার জন্য	Duplicate, wrong format, inconsistent value remove
	Duplicate	একই data বারবার থাকলে remove	একই data model-কে বেশি গুরুত্ব দেওয়া থেকে বাঁচায়	Same network record 2 বার থাকলে 1টা রাখা
	Noise	ভুল/অপ্রয়োজনীয় data	Model ভুল pattern শিখতে পারে	Age = 500 (ভুল value)
	Inconsistency	একই data বিভিন্ন format-এ থাকা	Data uniform করা	Dhaka, DHAKA, dhaka → Dhaka

---

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
3	**Outlier**	অস্বাভাবিক data খুঁজে বের করা	Extreme value model-কে confuse করতে পারে	Salary: 30k,40k,50k,500000
	IQR	Data range দেখে unusual value detect	Simple statistical method	500000 detect করা
	Z-score	Mean থেকে কত দূরে আছে দেখা	Extreme value detect	Z > 3 হলে outlier
	Isolation Forest	ML দিয়ে abnormal sample খোঁজা	Large dataset-এর জন্য	Network attack detection

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
4	**Sampling**	বড় dataset থেকে কিছু data নেওয়া	Training দ্রুত করার জন্য	10 লাখ data থেকে 1 লাখ নেওয়া
	Under Sampling	বেশি class কমানো	Class balance করা	Normal 9000 → 1000
	Over Sampling	কম class বাড়ানো	Minority class শেখানো	Attack 1000 → 9000
	Stratified Sampling	Class ratio একই রাখা	Fair train/test split	Train এবং test-এ Normal/Attack ratio একই

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
5	**Imbalance**	Class-এর data সংখ্যা সমান না হওয়া	Model যেন শুধু বড় class না শেখে	Normal=9000, Attack=1000
	SMOTE	নতুন minority sample তৈরি করে	Minority বাড়ানো	নতুন Attack sample তৈরি
	ADASYN	কঠিন minority sample বেশি তৈরি করে	Difficult case শেখানো	Hard attack বেশি generate

	Class Weight	কম class-কে বেশি importance দেওয়া	Data না বদলে model balance করা	Attack ভুল করলে বেশি penalty

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
6	**Transformation**	Data-এর distribution পরিবর্তন করা	Skewed data ঠিক করার জন্য	100,200,10000
	Log Transformation	বড় value ছোট করা	Extreme value effect কমানো	10000 → smaller scale
	Box-Cox	Data normal করার চেষ্টা	Positive data-এর জন্য	Income data
	Yeo-Johnson	Positive + negative দুইটাই handle করে	Flexible transformation	Temperature data

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
7	**Scaling**	Feature-এর range একই করা	বড় value যেন বেশি influence না করে	Age=20, Salary=50000
	Standard Scaling	Mean=0, SD=1	Logistic Regression/SVM-এর জন্য	Value normalize
	Min-Max	0 থেকে 1 range	Fixed range দরকার হলে	Pixel values
	Robust Scaling	Median/IQR ব্যবহার করে	Outlier থাকলে ভালো	Salary dataset

No	Preprocessing Family	সহজ ভাষায় কী?	কেন দরকার?	Example
8	**Encoding**	Text data-কে number করা	ML model text বুঝে না	TCP → number
	One-Hot Encoding	প্রতিটা category-এর আলাদা column	False order তৈরি করে না	TCP, UDP, HTTP
	Ordinal Encoding	Order থাকলে number দেওয়া	Ranking data-এর জন্য	Low=1 Medium=2 High=3

	Target Encoding	Target অনুযায়ী category value	High category data	City অনুযায়ী average

No	Preprocessing Family	সহজ ভাষায় কী?	Example
9	**Feature Engineering**	নতুন useful feature তৈরি করা	Existing data থেকে extra information বের করা
	Aggregation	অনেক feature combine করা	Source bytes + Destination bytes = Total bytes
	Interaction	Feature-এর relationship বের করা	Bytes/Packet
	Temporal Feature	Time থেকে feature বানানো	Timestamp → Night/Day

No	Preprocessing Family	সহজ ভাষায় কী?	Example
10	**Feature Selection**	Important feature রাখা, unnecessary বাদ দেওয়া	50 feature → best 10 feature
	MI	Feature কত useful measure করা	Packet count important কিনা
	RFE	ধাপে ধাপে feature remove	20 → 10 feature
	LASSO	কম important feature-এর weight zero	Feature remove

No	Preprocessing Family	সহজ ভাষায় কী?	Example
11	**Simulation/Augmentation**	নতুন data তৈরি করা	Data কম হলে বাড়ানো
	Monte Carlo	Random simulation	Traffic simulation
	GAN	Existing data দেখে নতুন realistic data তৈরি	New leaf image
	VAE	Data pattern শিখে generate	Synthetic sample

	Diffusion	Noise থেকে data তৈরি	AI image generation

No	Preprocessing Family	সহজ ভাষায় কী?	Example
12	**Statistical Analysis**	Data-এর relationship analysis করা	Feature important কিনা দেখা
	t-test	দুই group compare	Normal vs Attack
	ANOVA	তিন বা বেশি group compare	DoS vs DDoS vs Scan
	Chi-square	Categorical relationship	Protocol vs Attack
	Mann-Whitney	Non-normal দুই group compare	Traffic comparison
	p-value	Difference significant কিনা	$p < 0.05$
	Effect Size	Difference কত বড়	Practical importance

Missing → value নেই
Cleaning → ভুল data ঠিক
Outlier → অস্বাভাবিক data
Sampling → data কমানো
Imbalance → class balance
Transformation → data shape ঠিক
Scaling → range equal
Encoding → text to number
Feature Engineering → নতুন feature
Feature Selection → ভালো feature রাখা
Augmentation → data বাড়ানো
Statistics → relationship খোঁজা