1) Mixed Variables আসলে কী

একটা কলামের ডেটা টাইপ যদি একরকম না হয়, বা একই তথ্যের ভিতরে একসাথে সংখ্যাও থাকে আবার ক্যাটেগরিও থাকে, তখন তাকে Mixed Variable বলা হয়।

মেশিন লার্নিং মডেল সাধারণত চায় প্রতিটা feature একটি নির্দিষ্ট টাইপের হোক
সংখ্যা হলে numeric, আর শ্রেণি হলে categorical
Mixed থাকলে model confusion, encoding সমস্যা, এবং preprocessing pipeline ভেঙে যায়
তাই আগে আলাদা করা জরুরি

2) Mixed Data দুইভাবে আসে

Case 1: Single cell এর ভিতর সংখ্যা এবং অক্ষর একসাথে

উদাহরণ Titanic dataset-এর Cabin কলাম
C85, B123, E46

এখানে দুটি আলাদা তথ্য একসাথে আছে
C বা B বা E হল ক্যাটেগরি, যেমন কেবিন ব্লক বা সেকশন
85 বা 123 বা 46 হল কেবিন নম্বর, numeric অংশ

সমস্যা
এই কলামটাকে সরাসরি numeric বানানো যাবে না
আবার এটাকে সরাসরি categorical ধরলেও হাজার হাজার unique value হয়ে যাবে, যেমন C85, C86, C87 সব আলাদা category
এটা মডেলের জন্য খারাপ, কারণ high cardinality তৈরি হয়

সমাধান
Regex ব্যবহার করে একে দুইটি নতুন কলামে ভাগ করা

1. CabinLetter বা CabinGroup: শুধু অক্ষর অংশ, যেমন C, B, E
2. CabinNumber: শুধু সংখ্যা অংশ, যেমন 85, 123, 46

এতে লাভ কী
Category কমে যায়
আগে হাজারটা unique cabin ছিল, এখন letter অংশ হয়তো 5-10টা group
এবং numeric অংশ আলাদা হওয়ায় তুমি চাইলে এটাকে numeric feature হিসেবে ব্যবহার বা binning করতে পারবে
সাথে visualization এবং analysis সহজ হয়

ক্লাসরুম ধারণা
তুমি মূলত একটি compound feature কে দুইটি atomic feature এ ভেঙে দিচ্ছো
এটাই feature engineering

Case 2: Same column এর ভিন্ন ভিন্ন **row** তে ভিন্ন টাইপের ডেটা

উদাহরণ
একই কলামে কিছু row: 1, 2, 3
আর কিছু row: Alone, WithFamily, Unknown টাইপ টেক্সট

সমস্যা
এই কলামটাকে numeric ধরলে text-এর কারণে error হবে
categorical ধরলে numeric 1,2,3 category হয়ে যাবে, যা অর্থবোধক নাও হতে পারে
অর্থাৎ একই কলামের ভিতরে mixed schema থাকার কারণে processing ভেঙে যায়

সমাধান
এখানেও দুইটা নতুন কলাম বানাতে হয়

1. Numeric column
   শুধু সংখ্যাগুলো রাখা হবে
   pd.to_numeric(errors='coerce') ব্যবহার করলে যে মান numeric না, সেটা NaN হয়ে যাবে
   মানে text গুলো numeric কলামে missing হয়ে যাবে
2. Categorical column
   শুধু text গুলো রাখা হবে
   যেখানে সংখ্যা ছিল সেখানে NaN বসবে

এতে লাভ কী
এখন দুটো কলামই clean
একটা পুরো numeric, অন্যটা পুরো categorical
এখন pipeline অনুযায়ী numeric scaling, missing value impute, categorical encoding সব সহজে করা যাবে

3) যে টুলগুলো কাজে লাগে

pd.to_numeric(errors='coerce')

এটা এমন একটি কৌশল যা বলে দেয়
যেটা numeric না, সেটাকে force করে NaN বানিয়ে দাও
ফলে numeric অংশ cleanভাবে আলাদা করা যায়

ব্যবহারিক যুক্তি
Mixed row সমস্যায় numeric আলাদা করতে এটা সবচেয়ে সহজ এবং robust

Regex

Regex দিয়ে তুমি pattern ধরে অংশ আলাদা করো
যেমন letters-only, digits-only
Single cell mixed সমস্যায় এটা সবচেয়ে উপকারী

ব্যবহারিক যুক্তি
Compound string থেকে structured data বের করার standard টুল হল regex

4) কেন **Mixed Variables** আলাদা করা এত গুরুত্বপূর্ণ

1. Unique category কমে যায়
   Cabin এর মতো কলামে C85, C86, C87 সব আলাদা category হলে high cardinality হয়
   এর বদলে শুধু C, B, D ধরলে category সংখ্যা কমে এবং model stable হয়
2. Model-ready feature তৈরি হয়
   Model এক feature এ এক ধরনের meaning চায়
   একসাথে letter এবং number থাকলে meaning মিশে যায়
   ভাঙলে meaning পরিষ্কার হয়
3. Analysis এবং visualization সহজ হয়
   CabinLetter দিয়ে countplot, survival rate তুলনা
   CabinNumber দিয়ে distribution, outlier analysis

5) এক লাইনে মূল মন্ত্র

Mixed variable মানে এক জায়গায় একাধিক অর্থ বা টাইপ
তাই নিয়ম হলো
ভেতর থেকে category অংশ এবং numeric অংশ টেনে বের করে দুইটি আলাদা কলামে সাজানো