ট্রেনিং, টেস্ট ও ভ্যালিডেশন ডেটা (Training, Test and Validation Data)

ধরুন আপনি কোনো পরীক্ষা দেবেন। পরীক্ষা দেওয়ার আগে আপনি অবশ্যই দু-তিন মাস খুব ভালোমতো পড়াশোনা করবেন, নিজেকে প্রস্তুত করবেন, তাই না? কী কী ধরনের প্রশ্ন আসতে পারে, সব পরীক্ষা ধরনের দিতে প্রশ্ন যাচাই বাছাই যাবেন। করে, সমস্ত টপিক ঠিকমতো শেষ করে নিজেকে প্রস্তুত করে তবেই তো নিজেকে এই যে প্রস্তুত করার ব্যাপারটি হলো অনেকটা ট্রেনিং ডেটা দিয়ে নিজেকে ট্রেনিং দেওয়ার মতো।

আর যখন সমস্ত প্রস্তুতি শেষে পরীক্ষা দিতে যাবেন এবং পরীক্ষার রেজাল্ট আপনাকে বলে দেবে আপনার স্কোর কত হয়েছে, সেটাই কিন্তু বলে দেবে আপনার প্রস্তুতি আসলেই কতটুকু ভালো ছিল। পরীক্ষার প্রশ্নই হচ্ছে আপনার টেস্ট ডেটা। আপনি কি পরীক্ষার প্রশ্ন কখনো আগে ভাগে দেখে ফেলতে পারেন? কখনোই না, তাই না? ঠিক সেরকম, টেস্ট ডেটাও কখনো মডেলকে দেখতে দেওয়া হয় না। আগে সে সম্পূর্ণ প্রস্তুত হয়, এর পরে টেস্ট ডেটা হাতে পায়।

যা-ই হোক, এতক্ষণ আমরা দেখলাম ট্রেনিং ডেটা ও টেস্ট ডেটার ধারণাটুকু। এখান থেকে এতটুকু বোঝা যাচ্ছে যে, মডেল একেবারে নিজেকে ট্রেনিং দিয়ে পুরোপুরি তৈরি করে না ফেলা পর্যন্ত টেস্ট ডেটার দেখা পায় না, তাই না?

যে ডেটা দিয়ে আমরা কম্পিউটারকে ট্রেনিং দেব, সে ডেটাকে বলে 'ট্রেনিং ডেটা (Training Data)'। আর যে ডেটা দিয়ে আমরা কম্পিউটারের ট্রেনিং শেষ হওয়ার পর তার পারফরম্যান্স অ্যানালাইসিস করব যে তার কাজে সে কতটুকু ভালো করছে, সেই ডেটাকে আমরা বলব 'টেস্ট ডেটা (Test Data)'।

সবশেষে বলি,ভ্যালিডেশন ডেটার কথা। ভ্যালিডেশন ডেটা আপনার ট্রেনিং ডেটারই একটি অংশ, যেটি আপনার মডেল কতটুকু ভালো হয়েছে,আরো কতটুকু ভালো করা দরকার, মডেলের বিভিন্ন প্যারামিটারগুলো টিউন করার কাজ ইত্যাদি করতে সাহায্য করে।

এটি অনেকটা মূল পরীক্ষার পূর্বে, পরীক্ষার প্রস্তুতি হিসেবে মডেল টেস্ট দেওয়ার মতো। আপনি পড়াশোনা করে নিজেকে প্রস্তুত করলেন মূল পরীক্ষার জন্য (ট্রেনিং ডেটা), এরপর মূল পরীক্ষার আগে নিজের দুর্বলতা ঝালাই করে নেওয়ার জন্য মূল পরীক্ষার প্রশ্নের আদলেই কয়েকটি মডেল টেস্ট দিলেন (ভ্যালিডেশন ডেটা) এবং সেখান থেকে নিজের দুর্বলতা সব বের করে নিজেকে একেবারে প্রস্তুত করে তবেই ফাইনাল পরীক্ষা দিতে গেলেন (টেস্ট ডেটা)।

অর্থাৎ বলা চলে, ভ্যালিডেশন ডেটা হচ্ছে টেস্ট ডেটার মতোই এক ধরনের ডেটা, যা মডেল আগে ভাগে ট্রেনিংয়ের সময় দেখতে পারে না। ট্রেনিং শেষ হওয়ার পরে নিজেকে কিছুটা যাচাই করার জন্য এটি ব্যবহার করে নিজেকে ঠিকমতো 'টিউন' করে নিতে পারে, যাতে সে টেস্ট ডেটার ওপরে ভালো ফলাফল করে। মূলত মডেলকে ভালো পারফরম করতে ঠিকমতো আপডেট করাই হচ্ছে এই ভ্যালিডেশন ডেটার উদ্দেশ্য।

সাধারণত পুরো ডেটাসেটকে দুই ভাগে ভাগ করে একটি ভাগকে ট্রেনিং ডেটা, অন্যটিকে টেস্ট ডেটা এভাবে ব্যবহার করা হয় (Validation Data)

আমাদের গোটা ডেটাসেটকে এখন তিন ভাগে ভাগ করতে পারি-

- ট্রেনিং ডেটা,
- ভ্যালিডেশন ডেটা ও
- টেস্ট ডেটা

ধরুন আপনার কাছে যদি 100টি ডেটা থাকে, আপনি 60টি দিয়ে ট্রেনিং দিলেন, 20টি দিয়ে ভ্যালিডেশন করলেন এবং বাকি 20টি একেবারে সব প্রস্তুতি শেষে টেস্ট করার জন্য রেখে দিলেন। কোনটি ট্রেনিং সেটে যাবে, আর কোনটি টেস্ট সেটে, এগুলো দৈবভাবে বা র‍্যানডমলি নির্বাচন করা হয়।

মোটামুটি এই হচ্ছে ট্রেনিং, ভ্যালিডেশন ও টেস্ট ডেটার ধারণা।