Types of Machine Learning: Supervised, Unsupervised

মেশিন লার্নিং – এই টার্মটি সবার আগে ব্যবহার করেন আর্থার স্যামুয়েল (Arthur Samuel, 1901-1990) নামের একজন আমেরিকান বিজ্ঞানী। তাঁকে আর্টিফিশিয়াল ইন্টেলিজেন্স ও কম্পিউটার গেমিংয়ের একজন পথিকৃৎ বলা চলে।

মেশিন লার্নিংঅ্যালগরিদমগুলোকে সাধারণত দুটি ভাগে ভাগ করা যায় :

- Supervised Learning
- Unsupervised Learning

Supervised Learning

আমরা এখন একটু সুপারভাইজড লার্নিং নিয়ে কথা বলি। সুপারভাইজ মানে হচ্ছে কাউকে কোনোকিছু করতে শিখিয়ে দেওয়া, দেখিয়ে দেওয়া- যাতে সে পরবর্তী সময়ে নিজে নিজে কাজটিকরতে পারে। মেশিন লার্নিংয়ের ক্ষেত্রেও, সুপারভাইজড লার্নিং অ্যালগরিদম হচ্ছে সেই সব অ্যালগরিদম, যেগুলো ব্যবহার করে আমরা কম্পিউটারকে মানুষের মত চিন্তা করা শেখানোর জন্য ট্রেনিং দেই। আর ট্রেনিং দেওয়ার জন্য আমাদের যেটি করতে হবে সেটি হচ্ছে কম্পিউটারকে পর্যাপ্ত সংখ্যক উদাহরণ দিতে হবে।

এর দুই ধরনের প্রবলেম দেখা যায়-

ক্লাসিফিকেশন প্রবলেম (Classification Problem)।

রিগ্রেশন প্রবলেম (Regression Problem)।

ক্লাসিফিকেশন প্রবলেম (Classification Problem)।

প্রথমে জিরাফ দেখিয়ে আর জিরাফের বৈশিষ্ট্য বলে আপনার মস্তিষ্ককে ট্রেনিং দেওয়া হলো জিরাফ চেনার জন্য। এরপর আপনাকে যত প্রাণীর ছবিই দেওয়া হোক না কেন, আপনি আলাদা করতে পারবেন যে কোনটি জিরাফ আর কোনটি জিরাফ নয়। এ ধরনের সমস্যাগুলোকে আমরা বলি ক্লাসিফিকেশন প্রবলেম (Classification Problem)।

রিগ্রেশন প্রবলেম (Regression Problem)।

প্রথমে বিভিন্ন পিৎজার সাইজ ও তাদের দাম বলে দেওয়া হলো, আর তারপরে সেই তথ্যের ওপরে ভিত্তি করে আপনাকে আন্দাজ করতে বলা হলো অন্য আরেক সাইজের পিৎজার দাম যেটি আপনার অজানা। এ ধরনের সমস্যাগুলোকে আমরা বলি রিগ্রেশন প্রবলেম (Regression Problem)।

Unsupervised Learning

সুপারভাইজড লার্নিংয়ে আমরা যেরকম ডেটার পাশাপাশি সঠিক উত্তরটিও দিয়ে দিতাম, আনসুপারভাইজড লার্নিংয়ে সেরকম ধরনের কোনো ট্রেনিং দিয়ে নেওয়া হবে না কম্পিউটারকে। তাকে ট্রেনিং-এর জন্য শুধু ডেটা দিয়ে দেওয়া হবে, সঠিক উত্তর নয়।

বেশ কিছু ধরনের সমস্যা আছে আনসুপারভাইজড লার্নিংয়ের ক্ষেত্রে-

ক্লাস্টারিং প্রবলেম (Clustering Problem)

প্রিন্সিপাল কম্পোনেন্ট অ্যানালাইসিস (Principal Component Analysis - PCA).

ক্লাস্টারিং প্রবলেম (Clustering Problem)

নিজের মতো করে চেষ্টা করবে ওই ডেটার মধ্যে প্যাটার্ন বা বিভিন্ন ডেটার মধ্যে সামঞ্জস্য খুঁজে বের করতে। সেই সামঞ্জস্যের ওপরে ভিত্তি করে সে সবগুলো ডেটা বিভিন্ন গ্রুপে সাজাবে। এই ধরনের সমস্যাগুলোকে আমরা বলি ক্লাস্টারিং প্রবলেম (Clustering Problem),

প্রিন্সিপাল কম্পোনেন্ট অ্যানালাইসিস (Principal Component Analysis - PCA)

ধরুন,কোন ডেটাসেটের জন্য, ফিচারের সংখ্যা অনেক বেশি (> 10,000) এবং সেই তুলনায় ট্রেনিং ডেটা পর্যাপ্ত নেই। এ ধরনের ক্ষেত্রে হিসাবনিকাশ অনেক বেশি জটিল হয়ে যায় এবং Costing-ও অনেক বেড়ে যায় (সময়, মেমোরি বেশি লাগে ইত্যাদি)। এসব ক্ষেত্রে, এত এত ফিচারের মধ্যে খুব ভালোমতো যদি ডেটা অ্যানালাইসিস করা যায়, তাহলে দেখা যাবে মাত্র কিছুসংখ্যক ফিচার আমাদের দরকার,বাকিগুলো না হলেও হবে। তখন, আমরা অপ্রয়োজনীয় ফিচারগুলো বাদ দিয়ে দিই, যাকে বলে ডাইমেশন রিডাকশন (Dimension Reduction) 1 এক ধরনের অ্যালগরিদম আছে, যেগুলো ডেটার এই ডাইমেশন রিডাকশনের কাজ করে দেয়, এগুলোকে বলে ডাইমেনশনালিটি রিডাকশন অ্যালগরিদম (Dimensionality Reduction Algorithm)। এই ধরনের কোনো একটি অ্যালগরিদম ব্যবহার করে আমরা ডেটার ফিচারের সংখ্যা কমিয়ে নিয়ে আসি। এই ধরনেরই একটি অ্যালগরিদম হলো প্রিন্সিপাল কম্পোনেন্ট অ্যানালাইসিস (Principal Component Analysis) বা পিসিএ (PCA)। নাম থেকেই বোঝা যাচ্ছে, কোথাও অনেকগুলো কম্পোনেন্ট আছে, আমাদের সেখান থেকে যে কম্পোনেন্ট/কম্পোনেন্টগুলো সবচেয়ে বেশি গুরুত্বপূর্ণ সেগুলো রেখে বাকিগুলো বাদ দিয়ে দিতে হবে। ব্যাপারটি আসলেই অনেকটা এরকম। মেশিন লার্নিংয়ের ক্ষেত্রে, এই কম্পোনেন্ট বলতে বোঝায় ফিচার ডেটা।