অধ্যায় ১ঃ পাইথন দিয়ে মেশিন লার্নিং

পাইথন দিয়ে মেশিন লার্নিংয়ের এই বইটিরে আমরা কিভাবে একটি প্রেডেক্টিভ মেশিন লার্নিং মডেল তৈরি করতে হয় তা ধাপে ধাপে শিখব। এই অধ্যায়ে আমরা প্রেডেক্টিভ মেশিন লার্নিং মডেল তৈরির ধাপগুলো নিয়ে সংক্ষিপ্ত আলোচনা করব।

১.১ ভুল পথে পাইথন মেশিন লার্নিং শেখা

আমাদের প্রথমেই যেভাবে মেশিন লার্নিং শেখা উচিত নয় --

1. খুব ভালো পাইথন প্রোগ্রামার হতে হবে।
2. মেশিন লার্নিং এলগরিদম সম্পর্কে গভীর ধারনা থাকতে হবে।
3. একটি মেশিন লার্নিং প্রজেক্ট করতে যে যে টপিক গুলো শিখতে হবে সেগুল ভালো ভাবে না শিখে এড়িয়ে যাওয়া।

১.২ মেশিন লার্নিং মডেল তৈরির ধাপসমূহ

একটি প্রেডেক্টিভ মেশিন লার্নিং মডেল তৈরির জন্য নির্দিষ্ট কিছু ধাপ অনুসরণ করতে হয়। একবার এই ধাপসমূহ অনুসরণ করে কোন মডেল তৈরি সম্পন্ন করতে পারলে তা বার বার অন্যান্য প্রজেক্টে ব্যবহার করা যায়।

একটি প্রেডেক্টিভ মেশিন লার্নিং মডেল তৈরির জন্য ৬টি ধাপ অনুসরণ করতে হয়। ধাপগুলো হচ্ছে

1. প্রবলেম চিহ্নিত করা এবং এর সম্পর্কে খুব ভালো ধারনা নেওয়া
2. সমস্যাটি সমাধান করার জন্য ডেটা কালেক্ট করা এবং ডেটা অ্যানালাইজ করা।
3. মেশিন লার্নিং এলগরিদমের মাধ্যমে মেশিন ভালোভাবে শিখতে পারে তার জন্য ডেটাকে সেভাবে প্রস্তুত করা।
4. সমস্যাটি সমাধান করার জন্য সঠিক এলগরিদম সিলেক্ট করা।
5. মডেল ইভ্যালুয়েট অর্থাৎ মডেলের এক্যুরেসি বাড়ানো।
6. মডেল ফাইনালাইজ করা।

১.৩ মেশিন লার্নিং মডেল তৈরির ধাপ সমূহ নিয়ে আলোচনা

আমরা আগের অনুচ্ছেদে মেশিন লার্নিং মডেল তৈরির ধাপসমূহ কি কি তা জেনেছি । এই অনুচ্ছেদে সেই ধাপসমূহ সম্পর্কে সংক্ষেপে আলোচনা করা হবে। পরবর্তীতে এই প্রতিটা ধাপ নিয়ে একটি করে অধ্যায় তৈরি করা হবে যেখানে ধাপসমূহ নিয়ে বিস্তারিত আলোচনা করা হবে। তাহলে শুরু করা যাক --

১.৩.১ ধাপ ১ঃ ডিফাইন প্রবলেম/ সমস্যা চিহ্নিত করা

আপনি যে সমস্যা সমাধান করবেন তার সম্পর্কে খুব ভালো ধারনা থাকতে হবে। সমস্যাটি কোন ধরনের সমস্যা (ক্লাসিফিক্যাশন/রিগ্রেশন) তা নিশ্চিত হতে হবে। কোন কোন বৈশিষ্ট্য এর কারণে এই সমস্যা হয় তা জানতে হবে। পরবর্তীতে আপনি যে ডেটাসেট সংগ্রহ করবেন বা ডেটাসেট তৈরি করবেন সেখানে ঐ বৈশিষ্ট্য বা ফিচার গুলো আছে কিনা তা নিশ্চিত হতে হবে। সমস্যাটির সম্পর্কে খুব ভালো ধারনা থাকলে পরবর্তীতে ডেটা প্রিপ্রসেসিং এ অনেক উপকারী হয় । এতে করে আপনার মডেলের এক্যুরেসি ভালো হয়।

১.৩.২ ধাপ ২ঃ ডেটা অ্যানালাসিস/ ডেটা বিশ্লেষণ

আপনি যে ডেটাসেট সংগ্রহ করেছেন সেখানে কতটি ফিচার আছে, কতটি অব্জারভেশন আছে, ফিচারগুলোর মধ্যকার সম্পর্ক কি ইত্যাদি হচ্ছে ডেটা অ্যানালাসিস/ ডেটা বিশ্লেষণ। মেশিন লার্নিং মডেল তৈরির এই ধাপটি খুব গুরুত্বপূর্ন কেননা এই ধাপেই আপনি সিদ্বান্ত নিবেন আপনাকে ডেটার কোথায় এবং কি উপায়ে ডেটাকে প্রসেস করে মেশিন লার্নিং এলগরিদমের জন্য প্রস্তুত করবেন। ডেটার ভেতরকার খবর জানার জন্য ডেটা অ্যানালাসিসের জুরি নেই। এই অংশেই আমরা জানতে পারব ডেটা কিভাবে কথা বলে। ডেটার এই ভেতরকার খরব দুই ভাবে জানতে পারি।

ক) বর্ননামূলক পরিসংখ্যানের মাধ্যমে (Descriptive statistics)

খ) ডেটা ভিজ্যুয়ালাইজেশনের মাধ্যমে (Data visualization)

ডেটা অ্যানালাসিস অধ্যায়ে এই সম্পর্কে বিস্তারিত আলোচনা করা হবে।

১.৩.৩ ধাপ ৩ঃ প্রিপ্রেয়ার ডেটা/ ডেটা প্রস্তুত করা

মেশিন লার্নিং এলগরিদমের মাধ্যমে মেশিন ডেটা থেকে শেখে। তাই ডেটাকে এমনভাবে প্রস্তুত করতে হবে যেনো এলগরিদমগুলো সেই ডেটাসেট থেকে ডেটাগুলোর মধ্যে একটা প্যাটার্ন বের করতে পারে।

তিন ভাবে ডেটা প্রস্তুতের কাজটি করা যায়--

ক) প্রিপ্রসেস ডেটা (Preprocess Data)

খ) ফিচার সিলেকশন (Feature selection)

গ) ফিচার ইঞ্জিনিয়ারিং (Feature Engineering)

ক) প্রিপ্রসেস ডেটা (Preprocess Data):

ডেটাসেটের ডেটাগুলো বিভিন্ন স্কেলে থাকলে মেশিন লার্নিং এলগরিদম ডেটা থেকে ভালোভাবে শিখতে পারে না। তাই ডেটাগুলোকে একটি নির্দিষ্ট স্কেলে আনতে হয়। ডেটা নরমাল ডিস্ট্রিবিউশন হলে মেশিন ভালোভাবে শিখতে পারে তাই ডেটার ডিস্ট্রিবিউশন যদি নরমাল ডিস্ট্রিবিউশনের কাছাকাছি হয় তাহলে সেই ফিচারের ডেটাগুলোকে নরমাল ডিস্ট্রিবিউশনে পরিবর্তন করার মাধ্যমে মেশিনকে আরো ভালো ভাবে শেখানো যায়। ডেটা প্রিপ্রসেসের কিছু টেকনিক হচ্ছে --

১) Rescale data/Normalize data

২) Standardize data

৩) Binarize data

খ) ফিচার সিলেকশন (Feature selection):

মেশিনকে ভালভাবে শেখানোর জন্য ফিচার অনেক গুরুত্বপূর্ন। তবে অনেক বেশি ফিচার হলে মেশিন ভালোভাবে শিখতে পারে না। কারণ ফিচারগুলোর মধ্যে অনেক ফিচার রয়েছে যেগুলো মেশিনকে শিখতে কনফিউজ করে দেয়। তাই সঠিক ফিচার সিলেকশন করা খুবই গুরুত্বপূর্ন। চারভাবে ফিচার সিলেকশন করা যায়।

১) Univariate Selection.

২) Recursive Feature Elimination.

৩) Principle Component Analysis.

৪) Feature Importance

গ) ফিচার ইঞ্জিনিয়ারিং (Feature Engineering):

মেশিন লার্নিংয়ের সবচেয়ে গুরুত্বপূর্ন অংশ হচ্ছে ফিচার ইঞ্জিনিয়ারিং। ফিচার ইঞ্জিনিয়ারং হচ্ছে মেশিন লার্নিং এলগরিদমকে আরো ভালোভাবে কাজ করানোর জন্য ডোমেইন এক্সপার্টিজ ব্যবহার করে নতুন ফিচার তৈরি করার পদ্বতি। ডেটাসেটের দুই বা ততোধিক ফিচার যুক্ত করে নতুন ফিচার তৈরি করা হয় যা মেশিনকে শিখতে সাহায্য করে। এক্ষেত্রে যে ডোমেইনে কাজ করতেছেন সেই ডোমেইনের সম্পর্কে ভালো ধারনা থাকা জরুরী।

ফিচার ইঞ্জিনিয়ারিং এ যে যে কাজগুলো করা হয়ে থাকে --

i) ফিচারগুলোর মধ্যেকার সম্পর্ক জেনে দুই বা ততোধিক ফিচারকে যুক্ত করে নতুন ফিচার তৈরি করা।

ii) মিসিং ডেটা হ্যান্ডেল করা।

iii) ফিচারের ডেটা টাইপ স্ট্রিং টাইপ থাকলে তাদেরকে নিউম্যারিকেল টাইপে নিয়ে আনা।

iv) ক্যাটাগরিক্যাল ভ্যারিয়েবল হ্যান্ডেল করা।

১.৩.৪ ধাপ ৪ঃ ইভ্যালুয়েট এলগরিদম(Evaluate algorithm)

ডেটাসেটকে দুই ভাগে ভাগ করে এক ভাগ ডেটা দিয়ে মেশিনকে ট্রেইন করাতে হয় । তারপর আরেক ভাগ দিয়ে টেস্ট করে দেখতে হয় মেশিন ঠিক ঠাক মতো কাজ করছে কিনা। মেশিনের পারফর্মেন্স কেমন?? বিভিন্ন টেকনিক ব্যবহার করে মেশিনের পারফর্মেনস বাড়ানো?? কোন মেশিন লার্নিং এলগরিদম দিয়ে মেশিন ভালো ভাবে শিখতে পারছে তা চিহ্নিত করা ইত্যাদি কাজ এই ধাপে করা হয়। এই ধাপের কাজগুলো হচ্ছে --

ক) Resampling Data

খ) Algorithm Evalution matrics

গ) Spot - check Algorithm

ঘ) Model selection

ঙ) Pipelines

ক) Resampling Data:

মেশিনকে যে ডেটা দিয়ে ট্রেইন করানো হয় আবার সেই ডেটা দিয়ে টেস্ট করানো হলে মডেল ওভারফিটিং হয়ে যায়। অর্থাৎ মেশিন দেখা ডেটার উপর ভালো পারফর্মেন্স করে কিন্তু না দেখা ডেটা উপর ভালো কাজ করে না। তাই পুরো ডেটা সেটকে দুই ভাগে ভাগ করে এক ভাগকে দিয়ে মেশিনকে ট্রেইন করানো আরেক ভাগ দিয়ে মেশিনকে টেস্ট করাতে হয়। তারপর চেক করতে হয় মেশিন কেমন কাজ করছে। অনেক ভাবেই এই কাজ করা যায় আমরা চারভাবে এই কাজটি করা শিখব --

১) Train and Test split

২) k-fold cross validation

3) leave one cut cross validation

4) Repeated Random Test-Train split

খ) Algorithm Evalution matrics:

আমাদের ট্রেইন করানো মেশিনটি কেমন পারফর্মেন্স করছে তা জানা দরকার, তা না হলে আমরা বুঝব কিভাবে আমাদের মেশিনটি ঠিক ঠাক মতো কাজ করছে কিনা। মেশিনকে কি আরো ভালো ভাবে শিখাতে হবে ?? কোন অংশে আরো ভালোভাবে নজর দিতে হবে ইত্যাদি জানা যাবে Algorithm Evalution matrics এর মাধ্যমে।

ক্লাসিফিকেশন প্রবলেম শলভ করার জন্য যে মডেল তৈরি করা হয়েছে তার পারফরমেন্স জানা যায় পাঁচটি টেকনিকের মাধ্যমে --

- Classification Accuracy.
- ROC Curve.
- Confusion Matrix.
- Classification Report
- Logarithmic Loss

Regression প্রবলেম সল্ভ করার জন্য যে মডেল তৈরি করা হয়েছে তার পারফরমেন্স জানা যায় তিনটি টেকনিকের মাধ্যমে--

- Mean Absolute Error.
- Mean Squared Error.
- R2

গ) spot-check algorithm:

বিভিন্ন এলগরিদম ব্যবহার করে মডেল তৈরি করে দেখা হয় কোন এলগরিদম কেমন কাজ করছে এই প্রক্রিয়াকে spot-check algorithm বলে। যেহেতু আমাদের বেশিভাগ সময়ই supervised সমস্যা নিয়ে কাজ করতে হয় এবং এই supervised সমস্যাগুলো দুই ধরনের হয় classification প্রবলেম এবং regression প্রবলেম । এদের সমাধান করার জন্য আবার আলাদা আলাদা এলগরিদম রয়েছে। সেগুলো হচ্ছে --

classification algorithm: সাধারণত ৬টি classification algorithm বহুল ব্যবহৃত হয় ।

Regression algorithm: Regression algorithm গুলো হচ্ছে --

ঘ) Model selection

spot - check algorithm এর মাধ্যমে আমরা জানতে পেরেছি কোন এলগরিদম ভালো কাজ করেছে। যে এলগরিদম সবচেয়ে ভালো কাজ করেছে তাকে সিলেক্ট করে পরবর্তিতে এই মডেলকে ইম্প্রুভ করে আরো ভালো মডেল তৈরি করতে হবে। মডেল সিলেকশন করার সময় আমরা বক্সপ্লট করে কোন এলগরিদম সবচেয়ে ভালো কাজ করছে তা জানতে পারি।

ঙ) Pipelines

মেশিন লার্নিং এ কিছু স্ট্যান্ডার্ড ওয়ারফ্লো রয়েছে (একটা নির্দিষ্ট ধাপের কাজ করার পর পরবর্তি ধাপের কাজ) যা আমরা অটোমেট করতে পারি। মেশিন লার্নিং পাইপলাইনের মাধ্যমে আমরা এই কাজ করতে পারি।

১.৩.৫ ধাপ ৫ঃ ইম্প্রুভ রেজাল্ট(Improve result)

এই ধাপে কাজ করা হবে মডেলের পারফরমেন্স কিভাবে বাড়ানো যায় । আমরা দুইভাবে মেশিনের পারফর্মেন্স ভালো কোরানো শিখব ।

ক) Algorithm parameter tunning.
খ) Ensemble Methods

ক) Algorithm parameter tunning

scikit learn এ algorithm গুলো ক্লাস হিসেবে থাকে। ক্লাসগুলো প্যারামিটার হিসেবে সংখ্যা নেয়। এই সংখ্যাগুলো পরিবর্তন করে মডেলের পারফর্মেন্স বাড়ানোর পদ্বতি হচ্ছে Algorithm parameter tunning।

যেমনঃ `knn = KNeighborsclassifier(n_neighbors = 3)`

এখানে `n_neighbors` এর সংখ্যা বাড়িয়ে কমিয়ে মডেলের এক্যুরেসি বাড়ানো যায়।

খ) Ensemble Methods:

অনেকগুলো এলগরিদম একসাথে করে একটি এলগরিদম তৈরি করাকে Ensemble Methods বলে। তিনটি বহুল ব্যবহৃত এলগরিদম হচ্ছে --

১) Bagging algorithm

2) Bosting algorithm

3) Voting algorithm

১) Bagging algorithm: Bagging algorithm কে bootstrap aggregation বলা হয়। তিনটি bagging algorithm হচ্ছে --

i) Bagged Decession Tree

ii)Random Forest

iii) Extra Tress

2) Bosting algorithm: দুটি জনপ্রিয় boosting algorithm হচ্ছে --

```
i) Ada Boost.

ii) Stochastic Gradient Boosting
```

৩) Voting algorithm

দুই বা ততোধিক মেশিন লার্নিং এলগরিদমের প্রেডিকশনকে কম্বাইন করে voting algorithm. Voting algorithm সম্পর্কে বিস্তারিত আমরা অধ্যায়ে জানব

১.৩.৩ ধাপ৬ঃ প্রেজেন্ট রেজাল্ট (Present Result)

দুই ভাবে মেশিন লার্নিং প্রজেক্টকে ফাইনালাইজ করা যায়।

1. Finalize model with pickle
2. Finalize model with joblib.

ধাপগুলো সম্পর্কে অধ্যায়ে বিস্তারিত আলোচনা করা হবে।

১.৪ মেশিন লার্নিং ওয়ার্কফ্লো