Machine Learning Roadmap.

Python language (most preferred by companies)

↓

exploratory data analysis (explore the data gathered, like the nature/features in data)

↓

feature engineering (⟶ (i) exploratory Data analysis
(ii) handling missing value
(iii) Handling outliers
(iv) Categorical encoding
(v) Normalizing & Standardization

↓

feature Selection (selecting the appropriate feature accord to model)

ML Algorithms - Regression & classification, clustering

Regression → if continous return da
classification → if discrete —11—
clustering → if natude is not known

(i) Corelation
(ii) Forward Elimination
(iii) Backward Elimination

ML

(iv) Univariate Selection
(v) Random Forest importance
(vi) Feature selection with Decision Tree.

Machin learning model is chosen like Linear, logistic Regression, Decision Tree, Random Forest, kmeans,

↓

Hyper paramter tuning — Improving the model, by Improving accuracy.

↓

To Improve model still more we use Grid search, RandomisedSearch, hyperopt, Genetic Algorithms.

↓

Dockers & kubernetes to store model

↓

Deploy the model

↓

End to end ML projects.

classification of machine Learning

(i) Supervised Learning → future prediction
(ii) unsupervised learning → classification task
(iii) Reinforcement learning → used in Games, when played against computer.

Advantages

(i) to identify trends & patterns
(ii) human intervention not needed.
(iii) Handling multi-dimensional & multi-variety data.

Disadvantages

(i) Large amount of data & time & Resources req'd.
(ii) high error & susceptibility.

Areas of usage

Automatic language transition
Self driving cars
Product recomendations
traffic predictions
Speech Recognition
Image Recognition
stock market trading
Onlin fraud Detection
Virtual personal Assitant
Email Span & malware filtering
Medical Diagnosis

Libraries in python to Learn

Numpy → To solve numerical problems.

matplotlib / seaborn → To analyse the data & have it in graphical structure, to see nature of data.

scikitLearn → To build machine learning model

scipy → it provides constants.

TensorFlow → To perform Deep learning

NLTK / opencv → if data is in image or textual format.

Data types in Machine Learning

types of variables.

- (Numbers) (int & float) Numericals
  - Discrete (Always repeat) (there are only few perfect numbers) like 1231 1222 3 12 333
  - Continous (in a given range there are many numbers) 1.1, 1.2 1.3 1.4 1.5
- Categorical (string type data)
  - Ordinal → there is a order flow like M, S, XS, L, XL, XXL
  - ~~Nominal~~ Nominal. → Randomly any value Data is not connected between them i.e rows.
- Dates & Times
- Mixed

Categorical is a object data

Mixed variable is a object data.

Date time variables

like Date of joining & birth. contains date & time.

Mixed variables

it is a mixed mixture of both Numerical & categorical.

Data cleaning

It is the process of preparing data for analysis / ML / DL by removing & modifying data that is incorrect, incomplete, irrelevant, duplicated, & improperly formatted.

Steps in Data cleaning

(i) Handling Missing Data
(ii) Outliers Detection & Handling
(iii) Data Scaling & Transformation
(iv) Encoding Categorical variables
(v) Handling Duplicates
(vi) Dealing with inconsistent Data.

What is missing value

→ having blank value in the Data set,
→ Algorithms dont work on missing value. Coz it works on math problem. for what _ x 2 = ?, we wont come to know.
→ If we want to remove that data or fill any value based on our purpose
→ ~~want~~ first identify missing value.
∴ Then using pandas we would remove.