Data mining: Supervised Machine learning

machine learning

Supervised

↓

requires data Sceintist to train the algorithms with Label input with desried outputs

More accurate

unsupervised

↓

do not required any superviser not Label or any thing It will identit the hidden patterns it usefful for Anglysig data.

less accurate.

workflow form Data to decision.

first we collect data may be it will stored in sql table or no sql

data

↓

after importing data we are gonna find insight of the data means like pattern, what kind of data, what should we predict the (y variable)

~~Feature~~

insights

↓

Next we are going to find some features (what the important features) features selection and extraction we can called as feature engeeneering or we can use various dimension reduction technique like PCA, TSNE etc.

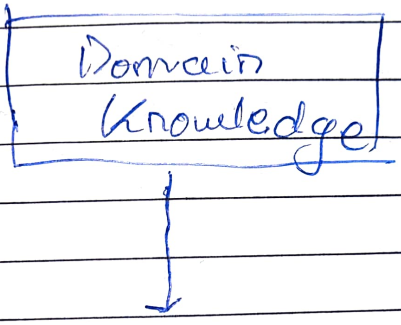
Features

↓

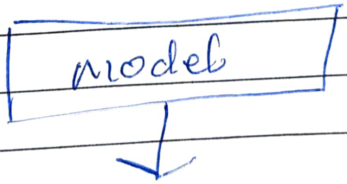
PAGE No.

DATE / /

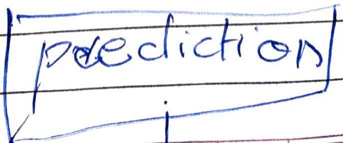
by domain Knowledge we can know which are important feature and which are not.



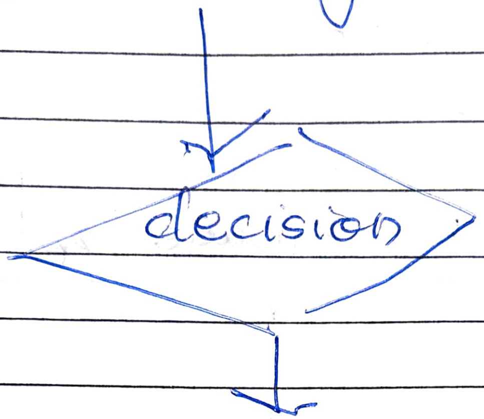
Next we should select the right model (keras, sklearn, pytorch) if it is supervised learning we should go for sklearn and if it is Neural Network we should go for sklearn and pytorch.



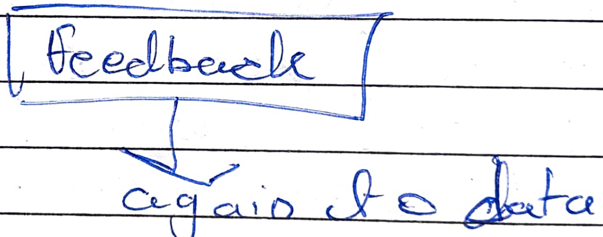
after that we should get the prediction out of it. (by domain expert we can know if prediction is proper or not).



then take decision if prediction are not proper do it again or not



take feedback from domain expert



and these process ~~are~~ may be co
tinew because today prediction are giving accurate tommorow is not

Data Nuances

★ Feature Noise

- value may not be accurate it may be Extreme value, Random Anomaly

★ how to remove Feature Noise

⇒ by Binning.

★ but what is binning

binning is used to deal with Noisy data we should smooth the data.

there are ~~Four~~ three technique of binning

① equal partioned bin

② Bin mean

- we will refore equal partioned bin to take out mean and we insert mean in Bins for example

Bin$_1$ - 9,9,9

Bin$_2$ - 22,22,22

Bin$_3$ - 29,29,29.

3) Bin Boundaries.

we will take min max values ~~and~~ thoes (Boundaries values) and thoes who are in middles we will change values thoes who are closest to th boundaries. for example (we will refe do equal partitioned Bin).

Bin 1 : 4, 4, 15

Bin 2 : 21, 21, 24

Bin 3 : 25, 25, 34

★ missing features

⇒ missing values

★ Non - Normal feature Distribution

⇒ highly skewed or kurtosis.

★ Heterogeeous features.

⇒ ranges, scales, distribution

example Age, temperature, Blood press

★ multimodality
For to solve this we will use standardization.

★ multi modality issue features
mix of numeric, symbdic, series, text.

log transformation
⇒ it reduces or remove the skewness of our original data there are two more method two remove skewness is
① log Square transformation
② box cox transformation