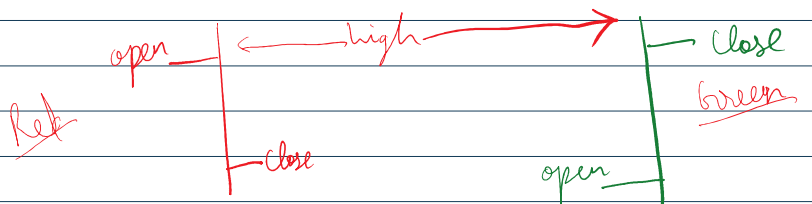


## Day 2 MCQ Quiz

Q1 Target Variable  $\rightarrow$  Usage of  $H_2O \rightarrow$  Real Value  
 $\downarrow$   
 Regression variable

Q2 Unsupervised learning

Q3  $\left( \begin{array}{l} \text{Red} \rightarrow \text{loss} \\ \text{Green} \rightarrow \text{gains} \end{array} \right)$   $\left( \begin{array}{l} \text{opening} + \text{closing} \\ \text{price} \quad \text{price} \end{array} \right)$



Q4 logistic Reg, Random Tree Classifier

\* Corpus = Collection

Q5 bagging is part Random Forest

Q6 PCA

Data Preparation: Numerical, Categorical, Text & Image Data

Roadmap:

① Read Data via Pandas & do EDA

$df = pd.read_csv('path')$   
 $\uparrow$  stored in RAM  $\uparrow$  In HDD

size of CSV < RAM  
always

② Data Preparation  $\downarrow$

(i) Identify the target variable

(ii) segregate I/p & O/p

(iii) Split the data into test & train  $df \rightarrow X, y \rightarrow \boxed{\text{Split}} \rightarrow X_{\text{-train}}, X_{\text{-test}}$   
 $y_{\text{-train}}, y_{\text{-test}}$

(iv) Pass training data through data preparation pipeline  
 $\hookrightarrow$  cleaning,

(iv) Pass training data through data preparation pipeline  
 ↳ cleaning  
 ↳ transformation.

Never do (mix) it! Not right

$X_{\text{train}} \rightarrow \text{Data Prep} \rightarrow X_{\text{train-transform}}$

$X_{\text{test}} \rightarrow \text{Data Prep} \rightarrow X_{\text{test-transform}}$

↓

$X_{\text{train-transform}}$        $y_{\text{train}}$

$X_{\text{test-transform}}$        $y_{\text{test}}$

③ Building the model

Training Data  $\rightarrow$  Algo  $\rightarrow$  model (learning of patterns/Relationship b/w I/p & o/p in the data)

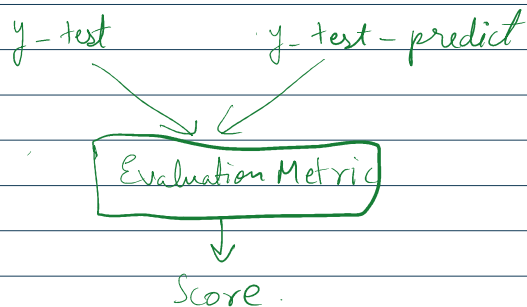


④ Evaluate the data

(i) Use model to predict on the unseen data

$X_{\text{test-transform}} \rightarrow \text{Model} \rightarrow y_{\text{test-predict}}$

(ii) Use an evaluation matrix to check the performance



Data Preparation

