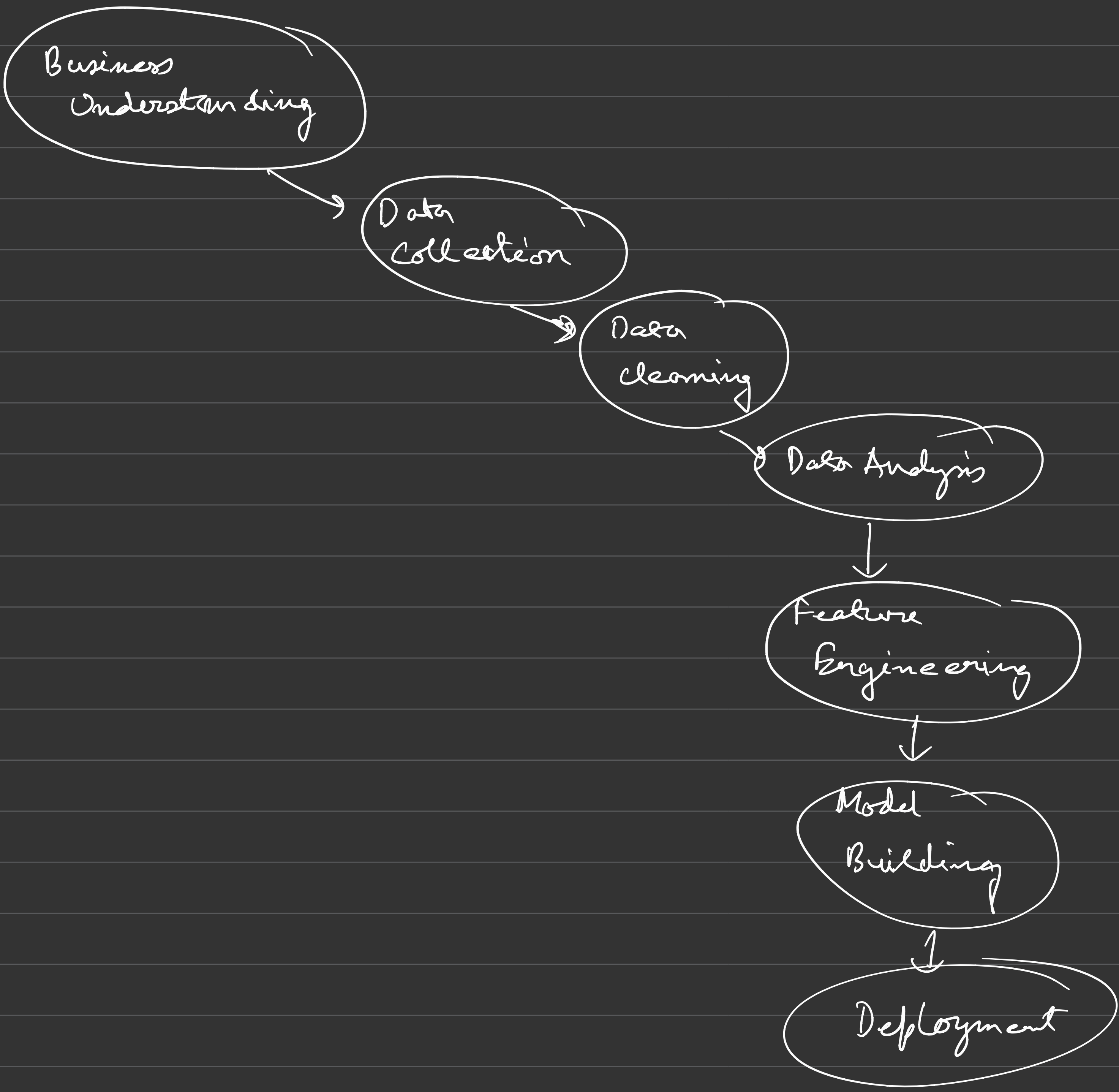

Business Understanding
Data Collection
Data cleaning
Data Analysis
Feature Engineering
Model Building
Deployment

Date: __/__/______

Project →

1) Data Collection
2) Perform Data Cleaning
3) Data Visualisation
4) Perform feature engineering
    1) Feature encoding
    2) Checking outliers
    3) Feature selection
5) Build ML model
6) Automate ML Pipeline
7) Hypertune ML model . along with cross validation

1) Reading data
2) Dealing missing values
    → is null ( )
    → deptime, Arival time, date of journey → datetime
    → date of journey → date, month, year
    → Dep. time, Am → hr min

Analyse time when most flights take off
    → duration → hr min
    → duration 4 price 4 stop
    ~~→ Airline 4 price~~
    → Route
    → Airline 4 price

Date: __/__/____

→ One hot en → $\overset{0}{\text{source}}$, $\overset{1}{\text{dest}}$

→ Label encodin → stops

Delhi – 0
Mum – 1
Kolkata – 2
Bangalore – 3

→ Deleting

→ Outlier detection   IQR → Inter quartile range

$Q_1 = 25\%$ $Q_2 = 50\%$ $Q_3 = 75\%$

$Q_1 - 1.5\ IQR$ $Q_3 + 1.5\ IQR$

→ feature Selection → Dest, Air

mutual info reg

→ RF basic. r2 s.

→ Save pickle.

→ auto make ml pipeline → Training, prediction, r2score

mean absolute error, mean square error, mean absolute % error

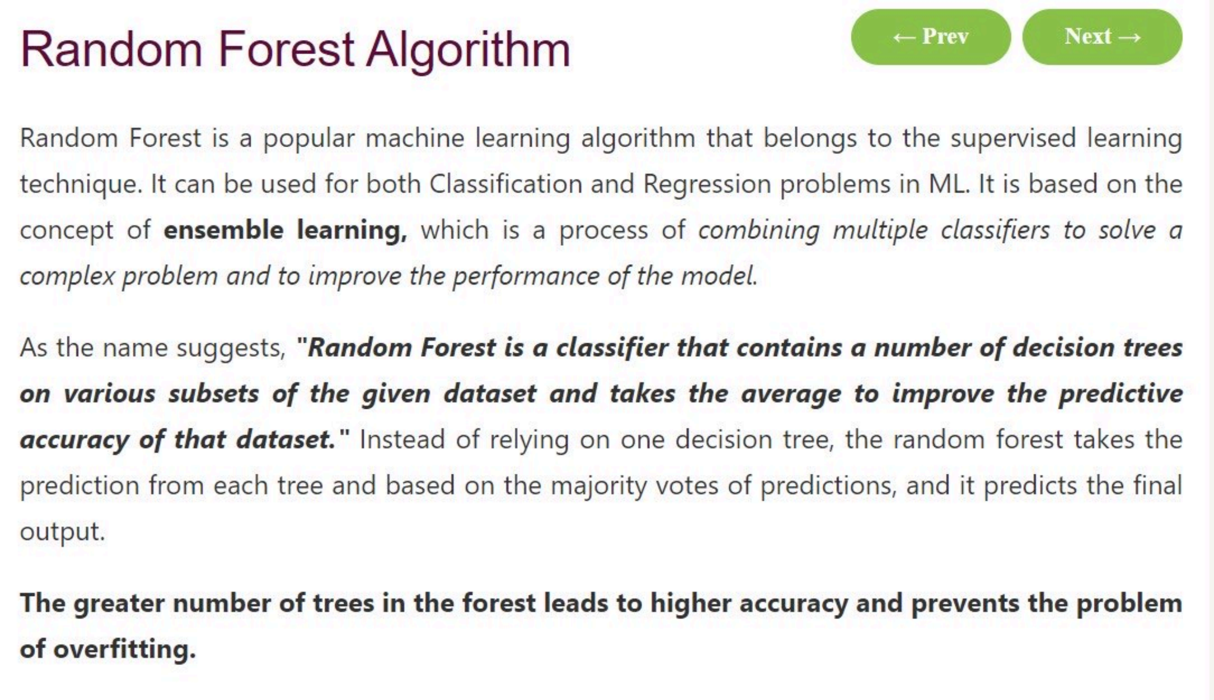
→ DT RF

→ Hypertune.

- R2 score ⇒ $R^2 = 1 - \frac{SS_{res}}{SS_{Tot}}$

where SS_{res} → sum of squares of residual error.

SS_{Tot} → Total sum of errors.



Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

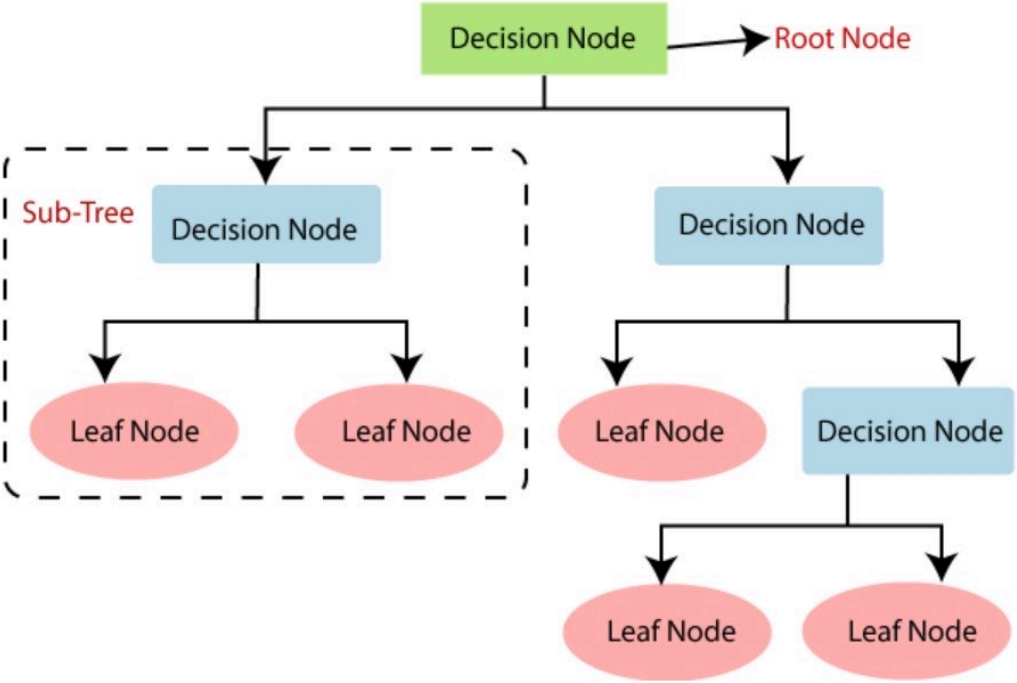
As the name suggests, **"*Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.*"** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Decision Tree Classification Algorithm

← Prev Next →

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**


Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node

Date: _/_/___

Encoding Techniques to convert Categorial data

1) data not in order → One Hot encoder
2) data in order → Label Encoder

Hyperpmater

Randomized ser
Grid search

Decision Tree

$IG = E - \{\text{Weigh} * \text{Entropy}]$

$E = -p^{+} \log_2 p^{+} - p^{-} \log_2 p^{-}$.

IG = measurement of changes in entropy

E = measure impurity in given attribute. It specifies randomness in data.

GI = measure impurity used while creating a dT in the CART Al

$GI = 1 - \sum_j p_j^2$

Hyperpar