apsara

Date:

- **Decision Tree**

It is a supervised ML algorithm used for classification and regression.

→ **Working Mechanism**

Select best feature from dataset
Split data based on this feature
Repeat the process recursively for each subset.
Stop when
- ↳ All samples belong to same class
- ↳ Max depth is reached
- ↳ Min samples per node condition is met.

→ **DT for classification**

a) Ginni Index

$$\{ Gini = 1 - \sum P_i^2 \}$$

Lower Gini - Better split.

b) Entropy

$$\{ Entropy : - \sum P_i \log_2 P_i \}$$

Measure randomness in data, 0 → Pure node

Information Gain

$$\{ IG = Entropy(parent) - \sum \frac{n_i}{n} Entropy(child_i) \}$$

Page No.:

apsara

Date:

- **DT for Regression**

Criteria → MSE = $\frac{1}{n} \sum (y - \bar{y})^2$

split chosen that minimize MSE

- **Some challenges of DT**

Prone to overfitting, High variance (How much model change with data change).
↓
small change in data → total different tree

* **Random forest**

→ It is an ensemble ML algo that builds multiple decision tree and aggregates their predictions.

It follows Bagging approach → Idea → train models independently on different random samples and avg their prediction to get final one, it reduce variance & overfitting.

Random forest uses,

- **Bootstrap sampling** → technique in which multiple datasets are created from the original dataset with replacement.
Each sample is selected randomly, means some data appear multiple times and some not at all purely random.

- **Feature Sampling** → At each split, consider only subset of feature.

Page No.:

## * Working of RF

Create multiple bootstrap dataset
↓
Train one decision tree on each dataset
↓
Make prediction
↳ Classification → Majority voting
↳ Regression → Avg of all predictions made by diff DT.

## * Key Points

- High accurate as make prediction by ensemble boosting.
- Reduce variance & overfitting
- Handle large datasets well.

- Higher computation cost.
- Requires more memory.