Experiment No. 4

Aim: Implement Decision Tree classifier models to perform supervised classification and evaluate model performance.

Software tools: Google Colab, Python Libraries(Pandas,Scikit-learn,Matplotlib,Seaborn)

Theory:

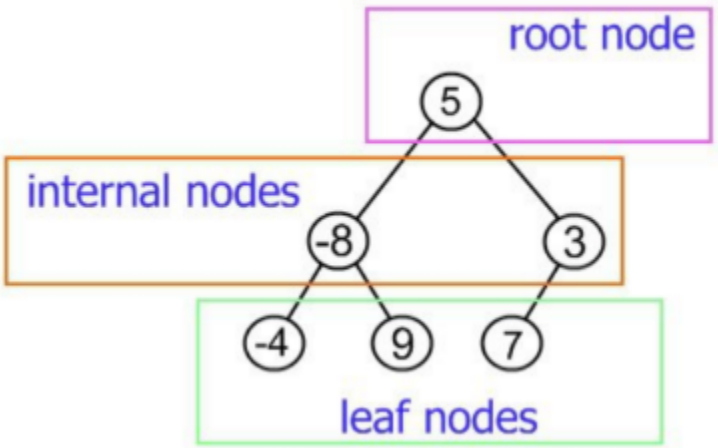
Decision Tree is a supervised learning algorithm that is used for both classification and regression tasks. It is a versatile and intuitive model that resembles a flowchart-like structure, where each internal node represents a feature, each branch represents a decision rule, and each leaf node represents an outcome or prediction.
Here are the key characteristics and components of a Decision Tree:

Nodes:

- Root Node: The topmost node in the tree, representing the feature that best splits the dataset into subsets based on a certain criterion.
- Internal Nodes: Intermediate nodes that represent subsequent feature splits.
- Leaf Nodes (Terminal Nodes): End nodes that provide the final prediction or classification.

Edges: Branches connecting nodes, indicating the decision rules or conditions based on feature values.

Criterion: A decision tree uses a criterion (also called impurity or loss function) to determine how to split the data at each internal node. Common criteria include Gini



impurity and entropy for classification tasks and mean squared error (MSE) for regression tasks.
Splitting: The process of dividing the dataset into subsets at each internal node based on a chosen criterion. The goal is to minimize the impurity or error in the resulting subsets.
Pruning: Decision trees can become overly complex and prone to overfitting (fitting noise in the data). Pruning is a technique to reduce the size of the tree by removing branches that do not significantly improve predictive performance.
Predictions:

- Classification: For classification tasks, a decision tree assigns the majority class in a leaf node as the prediction.
- Regression: For regression tasks, a decision tree assigns the mean (or another measure of central tendency) of the target values in a leaf node as the prediction.

Interpretability: Decision trees are highly interpretable models, making it easy to understand and explain the reasoning behind predictions.

Applications :

- Classification: Spam email detection, disease diagnosis, customer churn prediction.
- Regression: Predicting house prices, demand forecasting, quality control.

Advantages :

- Simplicity and interpretability.
- Can handle both categorical and numerical features.
- No need for feature scaling.
- Works well with both small and large datasets.
- Nonlinear relationships can be captured.

Limitations :

- Prone to overfitting (high variance) if not pruned.
- May not capture complex relationships effectively.
- Sensitive to small variations in the data.
- Lack of smoothness in predictions for regression tasks.

Conclusion :

We have implemented the Decision Tree algorithm for supervised classification tasks, and the model was able to learn patterns from the training dataset and classify new data effectively.