Experiment No: 5

Aim: Implement Decision tree classification algorithm

Theory:

- Decision trees are very strong and most suitable tools for classification and prediction.
- The attractiveness of decision tree is due to the fact that in contrast to neural networks, decision tree represent rules.
- Rules are represented using linguistic variables so that user interpretability may be achieved.
- By comparing the records with the rules one can easily find a particular category to which the record belongs to.
- Decision tree method is mainly used for the tasks that possess the following properties:

i) The tasks or the problems in which the records ~~the~~ are represented by attribute-value pairs.

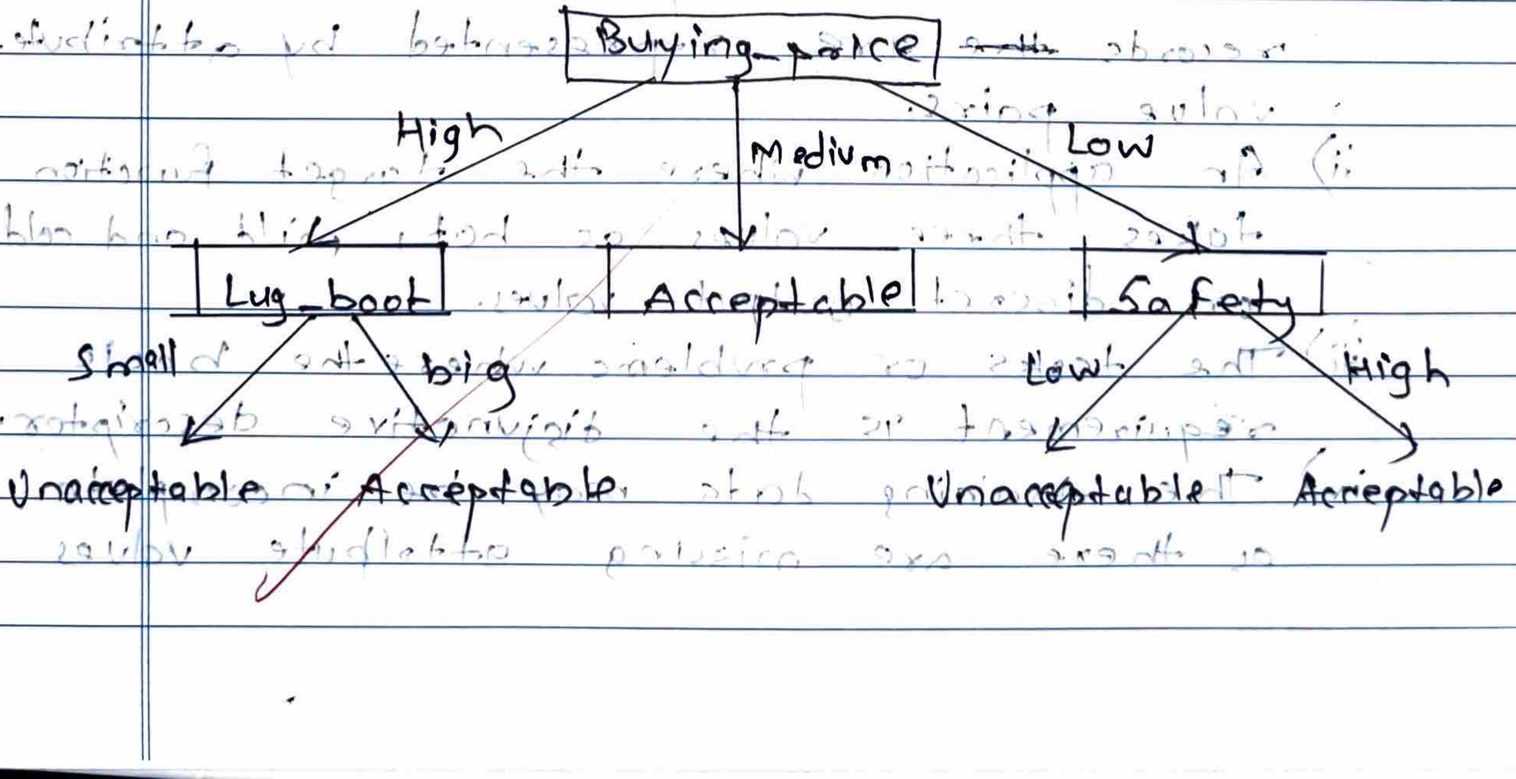
ii) An application where the target function takes three values as hot, mild and cold or discret output values.

iii) The tasks or problems where the basic requirement is the disjunctive descriptor.

iv) The training data may be incomplete as there are missing attribute values

- Decision tree representation:

- Decision tree is a classifier which is represented in the form of a tree structure where each node is either a leaf node or decision node.
- Leaf node represents the value of target or responce attribute of examples.
- Decision node represents some test to be carried out on a single attribute value, with one branch and sub tree for each possible outcome of the test.
- Decision tree generates regression or classification models in the form of a tree structure.
- Decision tree divides a dataset into smaller subsets with increase in depth of tree.
- The final decision tree is a tree with decision nodes and leaf node.



- A decision node (e.g., buying price) has two or more branches (e.g., High, Medium & Low)
- Leaf node shows a classification or decision.
- The topmost decision node in a tree which represents the best predictor is called root node.
- Decision trees can be used to represent categorical as well as numerical data.

- Gini Index:

- All attributes are assumed to be continuous valued.
- It is assumed that there exist several possible split values for each attribute.
- Gini index method can be modified for categorical attributes.
- Gini Index is used in CART.
- If a dataset T contains example from n classes, gini index gini(T) is defined as-

$$Gini(T) = 1 - \sum_{j=1}^{n} (P_j)^2$$

- In the above equation Pj represents the relative frequency of class j in T.

A+
27/9/24