Batch: A3 **Experiment Number:6**

Roll Number: 16010421073 **Name:Keyur Patel**

Aim of the Experiment: Design and implement Decision tree based ID3 algorithm.

Program/ Steps:

```
from sklearn.datasets import load_iris

from sklearn.model_selection import train_test_split
from sklearn.tree  import  DecisionTreeClassifier
from sklearn.metrics import accuracy_score


# Load Iris dataset from scikit-learn
iris = load_iris()

X  =  iris.data
y = iris.target
```

```
# Initialize and train the decision tree classifier (ID3 algorithm)
classifier = DecisionTreeClassifier(criterion='entropy', random_state=42)
classifier.fit(X_train, y_train)


# Make predictions on the test set
predictions = classifier.predict(X_test)


# Calculate accuracy

accuracy = accuracy_score(y_test, predictions)
print('Accuracy:', accuracy)


# Example: Predict a new sample

# Prepare the new sample data (features of an Iris flower)

new_sample = [[5.1, 3.5, 1.4, 0.2]] # Sepal length, sepal width, petal
length, petal width
```

Output/Result:

```
Accuracy: 1.0
Predicted Iris Species: setosa
```

Post Lab Question-Answers:

1. Which one is more prone to overfitting, Decision Trees or ID3? Why?

Ans: Both Decision Trees and ID3 (Iterative Dichotomiser 3) are prone to overfitting, but ID3 is often more susceptible due to its specific nature.

Decision Trees:

- Decision Trees are prone to overfitting, especially when they are deep and complex.
- A deep Decision Tree can memorize the training data, capturing noise in the data instead of the underlying patterns.
- Overfitting occurs when the tree is too specific to the trainingdata and does not generalize well to unseen data.

ID3:

- ID3 is a specific algorithm used to create Decision Trees.
- It constructs trees using a top- down, greedy approach.
- ID3 tends to overfit more because it creates a tree by selecting the attribute that provides the best information gain at each step.
- This means it can create branchesfor noise in the data, especially if some attributes have a large number of possible values (highcardinality).
- ID3 does not have mechanisms like pruning, which are used in more advanced Decision Tree algorithms (like CART) to reduce overfitting.

Why ID3 is More Prone to Overfitting:

1. **Greedy Nature:** ID3 makes locally optimal choices at each step without considering the

global context. This can lead to a tree structure that fits the training data too closely.

1. **Handling Noise:** ID3 does not handle noisy data well. If the dataset has errors or outliers, ID3 might create branches to accommodate them, leading to overfitting.
2. **Lack of Pruning:** ID3 doesn't employ pruning techniques to remove branches that donot provide significant predictive power. Pruning is essential to prevent overfitting bysimplifying the tree structure.
3. **High Cardinality Attributes:** If the dataset contains attributes with many possible values, ID3 might create branches for each value, leading to a complex and overfittedtree.

2. What is the role of Pruning in the Decision Tree Algorithm, and why is itimportant?

Ans: Pruning in the context of Decision Trees refers to the process of cutting off some branches (subtrees) from the tree, i.e., removing parts of the tree that do not provide significant predictive power.Pruning is important in Decision Tree algorithms for several reasons:

1) **Preventing Overfitting:** One of the main reasons for pruning is to prevent overfitting. Decision Trees can become overly complex, capturing noise in the training data. Pruning helps in simplifying the tree, reducing its complexity and making it more generalizable to unseen data. A simpler tree often has better predictive performance on new, unseen data.
2) **Improving Computational Efficiency:** Smaller trees are computationally less intensive. Pruning can significantly reduce the size of the tree, making it faster to evaluate and use for predictions. This is crucial, especially when dealing with large datasets or real-time prediction tasks.
3) **Enhancing Interpretability:** Pruned trees are simpler and easier to interpret. A simpler tree structure is more understandable to humans, allowing easier extraction of insights and decision-making based on the model.
4) **Dealing with Noisy Data:** Decision Trees are sensitive to noisy data, and they can create branches to accommodate outliers or errors in the training data. Pruning helps in removing these branches, making the model more robust to noisy data.

There are two main types of pruning in Decision Trees:

- **Pre-Pruning (Early Stopping):** In pre-pruning, the tree is cut off during the construction phase. The tree-growing process stops early, before it becomes toocomplex. This can be done by limiting the maximum depth of the tree, setting aminimum number of samples required to split a node, or requiring a minimum improvement in impurity to split a node.
- **Post-Pruning (Reduced Error Pruning):** Post-pruning, also known as reduced error pruning, involves first creating the complete tree and then removing branches that do notimprove the model's accuracy on validation data. It works by iteratively collapsing nodesof the tree and replacing them with leaf nodes.

3. What is entropy, and how it is used in the ID3 algorithm?

Ans: Entropy is a measure of disorder or impurity in a set of data. In the context of the ID3 (Iterative Dichotomiser 3) algorithm and other Decision Tree algorithms, entropy is used to quantify the impurity of a dataset. Specifically, it helps in deciding how to split the data based on the values of its attributes. The ID3 algorithm uses entropy to find the best attribute to split the dataset at each node of the decision tree. The attribute with the highest information gain (which corresponds to the attributethat reduces the entropy the most) is chosen as the splitting criterion. Lower entropy indicates that the dataset is more organized and, therefore, more useful for making predictions.

Outcomes:

CO2 : Apply concepts of different types of Learning and Neural Network.

Conclusion (based on the Results and outcomes achieved):
Therefore we learnt to Design and implement Decision tree based ID3 algorithm.

References: Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition