Aghasalim Agharahimov CS-22
Report: Titanic Survival Prediction Using Decision Trees

1. Introduction This lab focuses on manually building a Decision Tree by applying entropy and information gain to classify Titanic passengers based on survival.

2. Dataset Overview The dataset used in this analysis differs from the standard Titanic dataset. It consists of the following columns: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. Since the dataset contains null values, rows with missing values were dropped to ensure accurate calculations.

3. Data Preprocessing

- Selected relevant features: Survived, Pclass, Sex, and Age.
- Encoded the 'Sex' column to numerical values (Male = 0, Female = 1).
- Dropped rows containing null values to maintain data consistency.

4. Entropy Calculation Entropy measures impurity in a dataset. The entropy of the 'Survived' column was calculated using:

5. Information Gain Calculation Information gain measures the effectiveness of an attribute in classifying the data. It was calculated for:

- **Sex**
- **Pclass**
- **Age**

The attribute with the highest information gain is selected as the first split in the Decision Tree.

6. Results

- The entropy of the survival column was computed.
- Information gain was calculated for each attribute.
- The attribute with the highest information gain was identified as the best for the first split.

7. Conclusion This experiment demonstrated how entropy and information gain help in decision tree construction. The most informative attribute was chosen as the root node, and further splits would be based on entropy values to improve classification accuracy.

This process is a crucial step in understanding Decision Trees and their application in machine learning for classification problems.