Department of Artificial Intelligence, SVNIT,SURAT
B.Tech-III ,SEM-V
Subject- Machine Learning(AI301)

LAB ASSIGNMENT-5

Q-1 Decision Tree for Heart Disease Prediction

Part A – Basic Implementation

- Train a Decision Tree on the Heart Disease dataset.
- Report training and testing accuracy.
- Visualize the decision tree and interpret the first two splits.

Part B – Model Evaluation

- Print the **confusion matrix** and **classification report** (precision, recall, F1-score).
- Plot the **ROC curve** and report the **AUC score**.

Part C – Hyperparameter Tuning

- Vary max_depth, min_samples_split, min_samples_leaf.
- Compare training vs testing accuracy → shows **overfitting vs underfitting**.
- Use **GridSearchCV** for automated hyperparameter tuning.

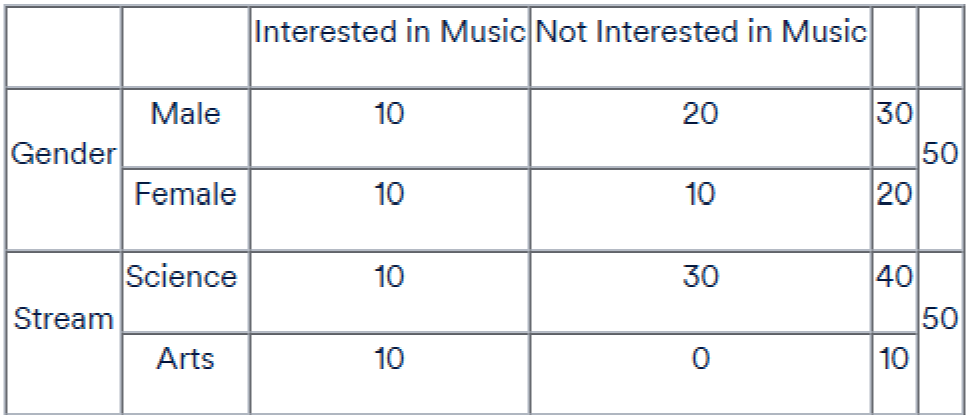
Part D – Error Analysis

- Find the patients (rows) that were misclassified by the decision tree.
- Compare their features with correctly classified patients.
- What patterns do you notice? (e.g., “younger patients with mild symptoms were misclassified”).

Part E – Class Imbalance Check

- Check class distribution (how many patients with disease vs without).
- If imbalanced → use metrics like **balanced accuracy** or apply resampling (SMOTE, undersampling).

Q-2 Answer the following question. (Optional)



		Interested in Music	Not Interested in Music		
Gender	Male	10	20	30	50
	Female	10	10	20	
Stream	Science	10	30	40	50
	Arts	10	0	10	

Consider a sample of 50 students in the age group from 15 to 22 years with some information on their Gender (Boy/ Girl) and Stream(Science/ Arts). 20 out of these 50 are interested in learning music. Now, suppose we are interested in creating a model to predict who will be interested in music?

Now answer the following question based on the above information.

a) What is the Gini index of the original/ parent data set ?

b) If we split the dataset based on gender what is the Gini index for male node ?

c) What is the weighted Gini index for Split on Gender ?

d) If we split the dataset based on gender what is the Gini index for Female node ?

e) If we split the dataset based on stream what is the Gini index for Science node ?

f) If we split the dataset based on stream what is the Gini index for Arts node ?

g) What is the weighted Gini index for Split on Stream?

h) What is the Gini index of the original/parent data set ?

i) Based on the Gini Index which is a better choice for splitting Gender or Stream ?