MACHINE LEARNING

1)d

2)a

3)b

4)a

5)c

6)a

7)b

8)b

9)

```
# Define the probabilities of each class

p_A = 0.4

p_B = 0.6


# Calculate the Gini index

gini_index = 1 - (p_A**2 + p_B**2)

print("Gini index:", gini_index)


# Calculate the entropy

entropy = -p_A*math.log2(p_A) - p_B*math.log2(p_B)

print("Entropy:", entropy)
```

10) Below are the advantages of Random forest over decision tree:

- Reduce over fitting: It reduces overfitting of data by building multiple decision trees and thus reducing variance in the model.
- Accuracy: By building multiple trees and aggregating their prediction. It can capture more complex patterns.

- Noise robustness: Random Forest can filter out noise and outliers and make more accurate predictions
- Feature Importance: Random Forest helps in measuring of features and identifying of important features in the dataset.
- Parallel processing: By building multiple trees it reduces tie and cost for large datasets.

11) Scaling of all numerical into dataset is done so that they contribute equally an large value no does not dominate.

Standardization and Min Max scaling

12) Using Gradient descent algorithm following are the advantages

- Faster convergence
- Avoid numerical instability
- Helps reach Global minimum
- Improves accuracy

13) In case of highly imbalanced datasets, accuracy is not a good metric to measure the performance off a model.

Accuracy does not take the consideration of imbalanced datasets.

Lest say there are two Datasets A and B. A has a sample 99% of datasets while B has a sample of 1% datasets. Then the accuracy of the model will be 99%, but it will not identify any of sets belonging to B.

I such cases F1, precision, AUC – ROC curve are better metrics to evaluate the performance of model.

14) F score metrics is known as F1 score. It is a metric combination of precision and recall into a single score. It is harmonic mean of precision and recall.

Formula of F1 score:

F1 score= 2*(precision 8 recall)/(precision + recall)

Where precision is no. of true positive divided by true positive and false positive and recall is no. of true positive divided by sum of true positive and false negatives.

1 meaning perfect performance and 0 mean poor performance

15

15) In scikit learn , fit(),transform() and fit_transform() are common method of prerocesssing.

Fit() method is used to fit the model too the training data. It calculates and saves the necessary parameters of the model on training data.

Transform() method applies the calculated parameters of the fit() method on the new data. It transformed the data according to learned parameters.

Fit_transform() is a combination of fit and transform. It fits the model on training data and then transforms the same data in one step.

SQL

1)b

2)b and c

3)c

4)a

5)c

6)c

7)c

8)b

9)b

10)b

11) Joins in SQL is used to combine data from 2 or more tables in a relational database . Joins are performed by specifying common columns between tables.

12) There are 5 types of Joins.

- Inner join : It returns the matching rows between the tables based on join condition.
- Right join: It returns the matching rows between the tables and all the rows from right table based on condition.
- Left join : it returns the matching and left rows between the tables based on condition.

- Full Outer join: It returns all the rows from both tables and matching rows based on the condition.
- Cross join : It returns all the possible combination of rows between two tables.

Statistics

1)b

2)d

3)c

4)b

5)c

6)a

7)c

8)b

9)