

Machine Learning

- 1) A) GridSearchCV()
- 2) D) All of the above
- 3) A) The regularization will increase
- 4) C) both A & B
- 5) A) It's an ensemble of weak learners.
- 6) C) Both of them
- 7) B) Bias will decrease, Variance increase
- 8) B) model is overfitting

12)

- 1) Scaling can make it easier to interpret the importance of the input features in the optimization problem
- 2) Scaling can help to avoid this problem by making the cost function smoother and easier to optimize

13)

In the case of a highly imbalanced dataset for a classification problem, accuracy may not be a good metric to measure the performance

This is because accuracy can be misleading in such cases, as it only measures the proportion of correctly classified

14)

F-score is a common performance metric for binary classification problems that combines precision and recall into a single score.

$$F1 \text{ score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

15)

Fit() - fit() is used to train the model using the database. It consider the parameters of model using training database and return trained model

Transform() - transform() is used to apply the trained model for new dataset. It take input as argument and return transformed data

Fit_transform() - It first trains the model on the input data using the fit() method, and then applies the trained model to transform the input data using the transform() method

11)

Scaling all numerical features in a dataset is important because many machine learning algorithms are sensitive to the scale of the input features

Standardization : This technique scales the features so that they have zero mean and unit variance. It involves subtracting the mean of each feature from the data and dividing by its standard deviation

Min-max scaling : This technique scales the features so that they have a range between 0 and 1. It involves subtracting the minimum value of each feature from the data and dividing by the difference between the maximum and minimum values

9)

$$\text{Gini index} = 1 - (0.4^2 + 0.6^2) = 0.48$$

$$\text{Entropy} = -0.4\log_2(0.4) - 0.6\log_2(0.6) = 0.971$$

10)

- 1) This reduces the risk of overfitting, which can occur when a single decision tree is too complex
- 2) Decision Trees can become computationally expensive and Random Forests can handle large datasets more efficiently by using a subset of the data
- 3) Random Forests are more robust to outliers and missing data which can balance out the effects of individual outliers or missing values

STATISTICS

- 1) **b) 0.135**
- 2) **d) 0.53**
- 3) **c) 0.745**
- 4) **b) 0.577**
- 5) **c) 0.6**

- 6) a) 0.33
- 7) c) 0.33
- 8) b) 0.22
- 9) a) 0.66
- 10) b) 0.45
- 11) c) 0.5
- 12) a) 0.166
- 13) d) 0.25
- 14) =
- 15) b) 2/3

SQL

- 1) B. Candidate keys
- 2) B. Primary keys cannot contain NULL values...
- 3) C. Insert
- 4) C. ORDERBY
- 5) C. SELECT
- 6) C. 3NF
- 7) C. All of the above can be done by SQL
- 8) B. DML
- 9) B. Table
- 10) B. 2 NF
- 11) Joins in SQL are used to combine data from two or more tables in a relational database
- 12) There are several types of joins
 - 1) Inner join
 - 2) Left join
 - 3) Right join
 - 4) Full outer join
- 13) SQL Server uses a client-server model to manage the interaction between the database and the users
- 14) primary key is a column or a combination of columns that uniquely identifies each row in a table, The primary key constraint is used to create a primary key for a table
- 15) ETL stands for Extract, Transform, Load, and it is a process used to integrate data from multiple sources into a single destination database
 - 1) Extract: Data is extracted from various sources, such as flat files, databases, and web services
 - 2) Transform: The extracted data is transformed into a format that can be used by the destination system

3) Load: The transformed data is loaded into the destination database or data warehouse