Module Assignment

1. write the answer to these questions.

Note: give at least one example for each of the questions.

- What is the difference between static and dynamic variables in Python?

Dynamic variable: Python is a Dynamic Programming Language. Dynamic variables in Python refer to how Python handles variable types and how variables can be reassigned to different types during runtime. In Python variables are dynamically typed, you don't have to declare the type of a variable when you assign a value to it.

Example:

```
x =5
print(type(x)  # <class 'int'>
x= "Hello"
print(type(x) #<class 'str'>
```

Static Variable: Static variables in Python usually refer to variables that retain their value across multiple function calls or class instances. They are often used in the context of class attributes, where a variable is shared across all class cases.

Example :

```
class Counter:
    count = 0
    def __init__(self):
        Counter. count += 1
    def show_count(self):
        print(f"Number of instances created: {Counter.count}")
c1 = Counter()
c2 = Counter()
c3 = Counter()
c1.show_count()
c2.show_count()
c3.show_count()
```

- Explain the purpose of pop, popitem, clear() in a dictionary with suitable examples.

pop : pop method is used to remove the item with the specified key name.

Example:

```
Dict = {student name : "Supriya",
Course : "Python",
Year : 2023}
```

Dict.pop("year")

Print(Dict)

Popitem :

The popitem() method is used to remove the last inserted item.

Example:

Dict = {student name : "Supriya",
Course : "Python",

Year: 2023}

Dict.popitem[]

Print(Dict)

Clear():

The clear() method is used to empty the dictionary.

Example:

Dict = {student name: "Supriya",
Course: "Python",

Year: 2023}

Dict.clear()

Print(Dict)

- What do you mean by frozen set? Explain it with a suitable example.

FrozenSet is an immutable datatype. It can not be changed after the set is created. Like sets, it is an unordered collection of unique elements. This immutability makes frozen sets useful for scenarios where a collection of items should not change, such as using it as a dictionary key.

Example:

Frozen_set=frozenset([1,2,3,4,5])

Print(Frozen_set)

- Difference between mutable and immutable data types in Python and give examples of mutable and immutable data types.
  Mutable: objects whose state or value can be changed after creation.
  Example:
  List=[1,2,3]
  List.append(4)
  Print(list)
  Immutable: objects whose state or value cannot be changed after creation.
  Example:
  tuple = (1, 2, 3, 4, 5)
  print("Original tuple:", tuple)
- What is __init__? Explain with an Example.

The __init__ () is a special method and it refers to a constructor that is used to initialized newly created objects. This method is used automatically when a new instance of a class is created.

Example:

```
Class student:
    def __init__(self,name,age):
        self.name = name
        self.age= age
   def greet(self):
       return f"Hello,my name is (self..name} and I am {self.age} years old"
 person1 =student("noshmitha',2)
print(person1.name)
print(person1.age)
print(person1.greet)
```

- What is docstring in python? explain with an example.

  The first statement in a module, function, class, or method definition is the Python docstring, a string literal. It makes it easier to link the documents together.

  Example:

  ```
  Def Hello:
     Return "Hello"
  ```

- What are unit tests in Python?

  Unit tests in Python are small tests that check the correctness of individual components of your code. They ensure each function or class behaves as expected by comparing actual outcomes to expected results.

- What is break, pass, and continue in Python?

  Pass: Pass specifies an operation-free Python statement. It is in a compound statement as a placeholder.

  Break: the break statement is used to exit or terminate a loop before a loop iteration is over when a specified condition is met.

  Continue: the continue statement is used to skip the rest of the code inside the current iteration of a loop and proceed directly to the next iteration. It is particularly useful when you want to skip certain elements or conditions within a loop without breaking out of the loop entirely.

- What is the use of self in Python?

  The self is a reference to the current instance of a class and is used to access variables and methods associated with that instance. It allows each instance to maintain its own state and behaviors. self is automatically passed as the first parameter to instance methods in a class, distinguishing instance variables from local variables.

- What are global, private, and protected attributes in Python?

  Global: Global scope is defined outside of the function. if you want to insert the outside function within the inner function by using global keyword. Outside variable used anywhere in the program.

  Private: Private attributes are intended to be used only within their own class. They are not accessible directly outside the class.

  Protected: Protected attributes are meant to be accessed within their class and by subclasses (children classes). They are not intended for direct use outside of these scopes.

- What are modules and packages in Python?

  Modules:

  Modules in Python are files with the Python code that create functions, classes and variables. Their main property is allowing code organization, reusability, and maintainability by dividing related functionalities into individual files. Modules can be imported by any Python script or module using the 'import' statement.

  Packages:

a package is a way of organizing related modules into a single directory hierarchy. This structure allows for a more systematic approach to managing large collections of modules and facilitates modular programming and code reuse.

- What are lists and tuples? what are the key differences between the two.

  a.list:
  -list is a collection data type.
  -It can store multiple items in a single variable.
  -It is a mutable data type.
  -It is ordered, indexed, and changeable, and allows duplicate values.
  -It can store any type of data types are allowed.
  -After creating the list we can change, add, and retrieve the data from the list.
  b.Tuple:
  - Tuple is a collection data type.
  -It can store multiple items in a single variable.
  -it is an immutable data type.
  -it is ordered, unchangeable, indexed, and allows duplicate values.
  -After creating the Tuple, you can add, and remove the elements from the tuple.

- What is an interpreted language and dynamic typed language? Write 5 differences between them.

  Interpreted Language:
  1. interpreted Language is used to execute the code line by line by an interpreter.
  2. Highly portable across different systems.
  3. Not relevant, focus on how code is executed.
  4. Typically slower line-to-line execution.
  5. No separate compilation step is needed.

  Dynamic Typed Language:
  1. variable types are determined at runtime.
  2. It focuses on type flexibility rather than different systems.
  3. May be slower due to runtime type checks.
  4. Not related to the type determined.
  5. Variables can change types during the execution.

- What are dict and list comprehension?

  List Comprehension: A concise way to create lists by embedding an expression inside a square-bracketed syntax, often with optional filtering.

  Dict Comprehension: A concise way to create dictionaries by embedding key-value pairs within curly braces, typically from an iterable or other dictionary.

- What are decorators in Python? Explain with an example. Write down its use cases.

  Decorators is a useful Python tool that allows programmers to add functionality to existing code. They are very powerful. Because a component of the program attempts to modify another component at compile time, this is also known as metaprogramming. It permits the client to wrap one more capability to expand the way of behaving of the wrapped capability, without forever changing it.

- What is memory managed in python?

  In Python, memory management is handled automatically by the Python runtime, using techniques like reference counting and garbage collection. Each object maintains a reference count to track how many references point to it, and when this count drops to zero, the object's memory is deallocated. Additionally, Python uses garbage collection to reclaim memory from circular references that reference counting alone can't resolve. To optimize performance, Python also employs a memory pool for small objects and dynamically allocates memory as needed during program execution.

- What is lambda in python? Why is it used?

Lambda function is also known as known as anonymous function. A lambda function is a small function that is defined anonymously by the keyword 'lambda'. It can have any number of arguments and it can only have one expression. Lambda functions are frequently used instead of the full 'def' statement when a function definition is very simple and not complicated enough.

- Explain split() and join() functions in python?

  Join(): The join() function in Python is used to concatenate elements of an iterable (like a list or tuple) into a single string, with a specified separator between each element.

  Split(): the split() function is used to divide a string into list of substrings based on specified delimiter.

- What are iterators, iterable and generators in python?

  Iterators: Iterators are objects that represent a flow of data and which can be used to begin a process repeatedly.

  Iterable: an iterable is an object capable of returning its members one at a time, allowing iteration over its elements using a loop or comprehension. Examples include lists, tuples, strings, and dictionaries, among others.

  Generator: the generator is a way that determines how to execute iterators. Except for the fact that it produces expression in the function, it is a normal function. It eliminates the __itr__ and next() methods and reduces additional overheads.

- What is the difference between xrange and range in python?

  Range(): Range() function is used to returns a list of containing the specified range of integers.

  Rangex(): Rangex() function is used to returns an rangex object that evaluates lazily, generating numbers on the flying during iterations.

- Pillars of oops.

  There are six fundamental concepts of object-oriented programming languages.

  They are:-

  a. class: class creates a user-defined data structure, which holds its data members, and member functions, which can be accessed and used by the instances of that class. a class is like a blueprint for an object.

  b.object: an object is an instance of a class.

  c. Polymorphism: it refers to having many forms. student in college, son at home, etc.

  d.inheritances: it allows us to define a class that inherits all the methods and properties from another class.

  e.encapsulations: It describes the idea of wrapping data and the methods that work on data within one unit.

  f.abstraction: it is used to hide unnecessary information and display only necessary information to the user interacting. (or) abstraction is a process of hiding the internal details of an application from the outside world.

- How will you check if a class is a child of another class?

  if a class is a child (subclass) of another class in Python, you can use the built-in issubclass () function.

  Example:

  ```
  Class parent:
     Pass
  Class child(parent):
     Pass
  Print(issubclass(child,parent))
  ```

- How does inheritance work in Python? Explain all types of inheritances with an example.

  Inheritance is a mechanism that allows a class to derive or inherit properties and behaviors (methods) from another class. There are Five types of inheritance.They are:

  1.single inheritance: A class inherits from one single parent class.

  Example:

  ```
  class parent:
     def greet(self):
  ```

```
        print(“Hello from parent.”)
class child(parent):
   pass
child = child()
child. greet()
```
2. Multiple inheritance: A class inherits from more than one parent class.
Example:
```
class parent1:
    def greet(self):
        print(“Hello from parent1.”)
class child(parent2):
    def greet(self):
        print(“Hello from parent2”)
class child(parent1,parent2):
    pass
child = child()
child.greet()
```
3. Multilevel inheriatance:
A class is derived from a class that is also derived from another class.
Example:
```
class Parent(Grandparent):
    pass
class Child(Parent):
    pass
child = Child()
child.greet()
```
4. Hierarichal inheritance:
Multiple classes are inherited from a single-parent class.
Example:
```
class Parent:
    def greet(self):
        print("Hello from Parent")
class Child1(Parent):
    pass
class Child2(Parent):
    pass
child1 = Child1()
child2 = Child2()
child1.greet()
child2.greet()
```
5. Hybrid Inheritance:
A combination of two or more types of inheritance. It involves a complex hierarchy where a class can inherit from more than one class and multiple levels.
Example:
```
class Base:
    def base_method(self):
        print("Base method")
class Parent1(Base):
    def parent1_method(self):
        print("Parent1 method")
```

```
class Parent2(Base):
    def parent2_method(self):
        print("Parent2 method")
class Child(Parent1, Parent2):
    def child_method(self):
        print("Child method")
child = Child()
child.base_method()
child.parent1_method()
child.parent2_method()
child.child_method()
```

- What is encapsulation? Explain it with an example.

  Encapsulation is a fundamental concept in object-oriented programming (OOP) that focuses on bundling data (attributes) and the methods that operate on that data within a single unit called a class. It describes the idea of wrapping data and the methods that work on data within one unit.

  Example:

```
class Bank:
def __init__(self, account_number, balance):
    self.__account_number = account_number
    self.balance = balance
def deposit(self, amount):
     self.balance += amount
def withdraw(self, amount):
    if amount <= self.balance:
        self.balance -= amount
    else:
        print("Insufficient funds")
def get_account_number(self):
    return self.__account_number
```

- What is polymorphism? Explain it with an example.

  Polymorphism is defined as its many forms. For example, students can play many roles. In school, students can play a role as students. Students can as a daughter or son at home, and students can play a role as friend with friends.

Q.2 Which of the following identifier names are invalid and Why?

a) Serial_no.
b) 1st_Room (invalid)
c) Hundred$(invalid)
d) Total_Marks
e) total_Marks
f) Total Marks
g) True
h) _percentage

A Name in a Python Program is called an Identifier.
It can be a Class Name, Function Name, Module Name, OR Variable Name.
The identifier has some rules:
1) Alphabet Symbols (Either Upper case OR Lower case)
2) If the Identifier starts with Underscore (_), it indicates it is private.

3) The Identifier should not start with Digits.
4) Identifiers are case-sensitive.
5) We cannot use reserved words as identifiers Eg: def = 10 ρ
6) There is no length limit for Python identifiers. But it is not recommended to use too lengthy identifiers.
7) Dollar ($) Symbol is not allowed in Python.

Machine learning:

1. what is the difference between Series and Dataframes?

   Series:

   -a series is a one-dimensional, labeled array that can store a single column of any data type.
   -series can only have one.
   -Series can store more simple and homogeneous data.
   -series can only have one data type for the whole array.
   -series can only handle one data type at a time.

   DataFrames:

   -a dataframe is a two-dimensional tabular data structure that can store multiple columns of different data types.
   -data frames can have multiple columns.
   -data frames can store more complex and heterogeneous data.
   -data frames can have different data types and can only have one data type for the whole array.
   -data frames can handle mixed data types such as numbers, strings, Booleans, or dates.

2. Create a database name Travel_Planner in mySql, and create a table name bookings in that which having attributes (user_id INT, flight_id INT, hotel_id INT, activity_id Int, booking_date DATE), fill with some dummy value. Now you have to read the content of this table using pandas as a data frame. Show the output.

```
CREATE DATABASE Travel_Planner;
USE Travel_Planner;
CREATE TABLE bookings (
    user_id INT,
    flight_id INT,
    hotel_id INT,
    activity_id INT,
    booking_date DATE
);
INSERT INTO bookings (user_id, flight_id, hotel_id, activity_id, booking_date)
VALUES
    (1, 101, 201, 301, '2024-07-17'),
    (2, 102, 202, 302, '2024-07-18'),
    (3, 103, 203, 303, '2024-07-19');
```

Output:

	user_id	flight_id	hotel_id	activity_id	booking_date
0	1	101	201	301	2024-07-17
1	2	102	202	302	2024-07-18
2	3	103	203	303	2024-07-19

3. Difference between loc and iloc.

   Loc:

   -Loc function is mainly used when we want to select rows and columns based on their labels.

   Iloc:

   -Iloc function is used when choosing rows and columns based on specific.

4. What is the difference between supervised and unsupervised learning?

   Supervised learning:

   -supervised learning is a learning setup commonly used in machine learning where human supervision is involved as they label the data and provide target output to the algorithm to map the input to it.

   -it handles the input data as the label.

   -the data has x features and y variables, and the model finds y = f(x).

   -supervised learning is classified into two types they are regression and classification.

   -The supervised learning goal is to predict the outcomes for new data based on training data.

   Unsupervised learning:

   -unsupervised learning is a learning setup commonly used in machine learning where human supervision is minimal, as the model finds the patterns in the data without any supervision.

   -unsupervised learning data doesn't have labels.

   -patterns are found in the x features of the data as no y variable is present.

   -unsupervised learning is classified into three types they are clustering, Dimensionality Reduction, and association.

   -unsupervised learning has a handle to get hidden patterns and useful insights from large datasets.

5. Explain the bias and variances tradeoff.

   The bias is known as the difference between the prediction of the values by the machine learning model and the correct value. Being high in biasing gives a large error in training as well as testing data. It recommended that an algorithm should always be low-biased to avoid the problem of underfitting.

   The variability of model prediction for a given data point which tells us the spread of our data is called the variance of the model. The model with high variance has a very complex fit to the training data.

   The bias and variance tradeoff implies that as we increase the complexity of a model.ita variance decreases, and its bias increases.

6. What are precision and recall? How are they different from accuracy?

   Recall is used to ability of a model to find all the relevant cases within a dataset.Precisipon is used to the ability of a classification model to identify only the relevant data points.

   Precision shows how often an ML model is correct when predicting the target class. Recall shows whether an ML model can find all objects of the target class.

7. What is overfitting and how can it be prevented?

   Overfitting occurs when a machine learning model learns not only the underlying pattern in the data but also noise and random fluctuations, resulting in a model that performs well on training data but poorly on unseen test data. This happens because the model is too complex relative to the amount and quality of the training data, essentially memorizing the training examples rather than generalizing them. It can prevented by using the techniques such as the cross-validation, train with more data. feature selection, regularization, ensemble methods, early stopping, simplify the model. And data augmentation.

8. Explain the concept of cross-validation.

   Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments. one used to learn or train a model and the other used to validate the model.

9. What is the difference between a classification and a regression problem?

   Classification :

-Classification algorithms are used to forecast or classify distinct values such as real.
- A tree model where the target variable can take a discrete set of values.
- the dependent variables are categorical.

Regression:

-regression algorithms are used to determine continuous values such as price, income, age etc,.
-A tree model where the target variable can take continuous values typically real numbers.
-the dependent variables are numerical.

10. Explain the concept of ensemble learning.

Ensemble learning is refers to a collection of methods that learn a target function by training a number of individual learners and combining their predictions together.

11. What is gradient descent and how does it work?

Gradient descent is an optimization algorithm for finding a local minimum of a differentiable function. Gradient descent in machine learning is simply used to find the values of a functions parameters that minimize a cost function as far as possible.

12. Describe the difference between batch gradient descent and stochastic gradient descent.

Batch gradient descent:

-it computes the gradient using the whole training sample.
-slow and computationally expensive algorithm.
-not suggested for huge training samples.
-it is a deterministic in nature.
-it gives the optimal solution given sufficient time in converge.
-it has no random shuffling of points are required.

Stochastic gradient descent:

-it computes the gradient using a single training sample.
-faster and less computationally expensive than Batch GD.
-it can be used for large training samples.
-It is a stochastic in nature.
-the data sample should be in a random order, and this is why we want to shuffle the training set for every epoch.

13. What is the curse of dimensionality in machine learning?

The curse of dimensionality in machine learning is a phenomenon that states that with a fixed number of training samples, the average predictive power of a classifier or regressor first increases as the number of dimensions or features used is increased but beyond a certain dimensionality it starts deteriorating instead of improving steadily.

14. Explain the difference between L1 and L2 regularization.

The difference between L1 and L2 regularization:
-The L1 regularization penalizes the sum of absolute values of the weights.
-It solution is sparse,.
-L2 regularization penalizes the sum of squares of the weights.
-it solution is a non-sparse.

15. What is a confusion matrix and how is it used?

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

16. Define the AUC-ROC curve.

AUC-ROC stands for “Area under the curve of the receiver operating characteristic curve. The AU C-ROC curve is a way of measuring the performance of an ML model. AUC measures the ability of a binary classifier to distinguish between classes and is used as a summary of the ROC curve.

17. Explain the k-nearest neighbors algorithm.

The k-nearest neighbors algorithm is a supervised machine learning method employed to tackle classification and regression problems. It is widely disposable in real-life scenarios. It assumes the

similarity between the new data and available cases and puts the new case into the category that is most similar to the available categories.

18. Explain the basic concept of a support vector machine.

A support vector machine is defined as a machine learning algorithm that uses supervised learning models to solve complex classification, regression, and outlier detection problems by performing optimal data transformations that determine boundaries between data points on predefined classes and labels.

19. How does the kernel trick work in SVM?

The kernel trick relies on the inner products of vectors. For SVMs, the decision function is based on the dot products of vectors within the input space. Kernel functions replace these dot products with a non-linear function that computes a dot product in a higher dimensional space.

20. What is the hyperplane in SVM and how is it determined?

The hyperplane is a decision boundary that differentiates the two classes in SVM. A data point falling on either side of the hyperplane can be attributed to different classes. The dimension of the hyperplane depends on the number of input features in the dataset.

21. What are the different types of kernels used in SVM and when would you use each?

Support Vector Machines (SVMs) are powerful supervised learning models used for classification and regression tasks. The choice of kernel in SVM determines the decision boundary shape in the feature space.

Types of kernel in SVM:

Linear kernel – the linear kernel is the simplest and is used when the data is linearly separable

Polynomial kernel – the polynomial kernel is effective for non-linear data.

Radial Basis Function kernel – it handles the non-linear decision boundaries well and works efficiently in most scenarios where the relationships between features and targets are complex and non-linear.

22. What are the pros and cons of using a support vector machine?

Pros:

-it performs well at classifying non-linear data.

-optimizing margins can help reduce the overfitting of data and allow for capacity control.

-learning without a local minimum

-there are many kernels that could be used to fit he data unlike any mother algorithm.

-often provides sparse solutions.

-performs well on data sets that have many attributes, even if there are relatively very cases on which to train the model.

Cons:

-the number of possible kernels is infinite and can make it hard to choose the right one.

- most software uses a few kernels that generalize to many situations, but no kernel generalizes to every situation.

-it can be computationally intensive.

-it can overfit the model.

23. Explain the difference between a hard margin and a soft margin SVM.

Hard margin:

-when the data is linearly separable, and we don't want to have any misclassifications, we use SVM with a hard margin.

-it assumes that the hyperplane separating the two classes is defined.

-it does not allow misclassifications.

-it used to maximize the distance between the two hyperplanes. To find the distance of a point from a plane.

Soft margin:

-when a linear boundary is not feasible, we want to allow some misclassifications in the hope of achieving better generality.
-the soft margin follows a somewhat similar optimization procedure with a couple of differences.
-it allows misclassifications to happen. so we need to minimize the misclassification error.

24. Describe the process of constructing a decision tree.

Decision trees are constructed using a top-down, greedy approach, where each split is chosen to maximize information gain or minimize impurity at the current node. This may not always result in the globally optimal tree.

25. Describe the working principle of a decision tree.

A decision tree is a non-parametric supervised learning algorithm for classification and regression tasks. it has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. decision trees are used for the classification and regression tasks, providing easy-to-understand models.
Working principles of a decision tree:
-starting at the root: the algorithm begins at the top called the root node representing the entire dataset.
- Ask the best questions: it looks for the most important feature that splits the data into the most distinct groups.
-branching out: based on the answer to that question, it divides the data into similar subsets, creating new branches. Each branch represents a possible route through the tree.
-Repeating process: the algorithm continues asking questions and splitting the data at each branch until it reaches the final" leaf node" representing the predicted outcomes or classifications.

26. What is information gain and how is it used in decision trees?

Information gain is the basic criterion to decide whether a feature should be used to split a node. The feature with the optimal split that is the highest value of information gain at a node of a decision tree is used as the feature for splitting the node. Information gain is a measure used to determine which feature should be used to split the data at each internal node of the decision tree. It is calculated using entropy. entropy is a metric to measure the impurity in a given attribute.

27. Explain gini impurity and its role in decision trees.

Gini impurity measures how well a node splits the data set between the two outcomes.it aims to reduce the impurity score from the root of the tree to the leaf node. Gini impurity is a measurement uses to build decision trees to determine how the features of a dataset should split nodes to form the tree.

28. What are the advantages and disadvantages of decision trees?

Advantages of decision trees:
-interpretability.
-deals with unbalanced data.
-variable selection.
- handles missing values.
-non-parametric nature.
Disadvantages of decision trees:
-pruning
-ensemble methods.
-addressing class imbalance.

29. How do random forests improve upon decision trees?

Instead of relying on a single decision tree, random forests generate a collection of decision trees, each trained on a random subset of the input data. The idea behind this approach is that by combining multiple models, the overall prediction accuracy will be improved.

30. How does a random forest algorithm work?

Random forest is an algorithm that generates a forest of decision trees. It then takes these many decision trees and combines them to avoid overfitting and produce more accurate predictions.. the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output. the greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

31. What is bootstrapping in the context of random forests?

Bootstrapping is the practice of random sampling with replacement.in data science, the boostrapping is used in many different contexts,but here, bootstrap to makes more trees.

32. Explain the concept of feature importance in random forest.

Feature importance is calculated as the decrease in mode impurity weighted by the probability of reaching that node. The need probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

33. What are the key hyperparameters of a random forest and how do they affect the model?

The key hyperparameters of a random forest and they affect the model.

a)max_depth: among the parameters of a decision tree, max_depth works on the macro level by greatly reducing the growth of the decision tree.

b)min_sample_split: by increasing the value of the min_sample_split, we can reduce the number of spits that happen in the decision tree and therefore prevent the model from overfitting.

c)max_terminal_nodes: this hyperparameter sets a condition on the splitting of the nodes io the tree and hence restricts the growth of the tree.

d)min_samples_leaf: the growth of the tree by setting a minimum sample criterion for terminal nodes.this hyperparameter also helps prevent overfitting as the parameter value increases.

e) n_estimators: this means that choosing a large number of estimators in a random forest model is not the best idea.although it will not degrade the computational complexity and prevent the use of a fire extinguisher on your CPU.

d) max_samples: the max_samples hyperparameter determines what fraction of the original dataset is given to any individual tree.

34. Describe the logistic regression model and its assumptions.

Logistic regression assumes the observations to be independent of each other and independent of repetitive measurement. Any individual should not be measured more than once and neither should it be taken in for the model. A way to check this assumption is by maintaining an order for the observations.

-explanatory variable shave no multicollinearity.

-no extreme outliers.

-the explanatory variables and the logit of the response variable have a linear relationship between them.

-sufficient sample size.

-normally distributed residuals.

35. How does logistic regression handle binary classification problems?

Logistic regression is one of the most popular algorithms for binary classification. the goal of logistic regression is to output values between 0 and 1, which can be interpreted as the probabilities of each example belonging to a particular class.

36. What is the sigmoid function and how is it used in logistic regression?

Logistic regression is used for binary classification where the goal is to predict one of two possible outcomes, typically represented as 0 and 1. The sigmoid function is at the core of logistic regression, serving as the link function that maps the linear combination of input features to probabilities.

37. Explain the concept of the cost function in logistic regression.

The logistic regression cost function also known as the cross-entropy loss is a measure of the error between the predicted probabilities and true class labels. The logistic regression cost function is also

known as the cross-entropy loss function.it is often referred to as the cross -entropy cost function and is designed to optimize the parameters to minimize the prediction error for binary classification tasks.

38. How can logistic regression be extended to handle multiclass classification?

It can be extended to handle multiclass problems by using techniques such as softmax regression , these extensions allow logistic regression to classify instances into multiple classes by estimating the probabilities of each class.

39. What is the difference between L1 and L2 regularization in logistic regression?

L1 regularization in logistic regression:
-L1 regularization penalizes the sum of absolute values of the weights
-the L1 regularization solution is sparse.
-The L1 regularization is robust to outliers.
- the Li regularization has built-in feature selection.
L2 regularization in logistic regression:
-L2 regularization penalizes the sum of squares of the weights.
-The L2 regularization solution is non-sparse.
-The L2 regularization doesn't perform feature selection.
-The L2 regularization is not contains the Robust.

40. What is XGBoost and how does it differ from other boosting algorithms?

Extreme Gradient boosting is also known as XGBoost. XGBoost improves the gradient boosting for computational speed and scale in several ways. XGBoost uses multiple cores on the CPU so thayt learning can occur in parallel during training. It is a boosting algorithm that can handle extensive datasets, making it attractive for big data applications.

41. Explain the concept of boosting in the context of ensemble learning.

Boosting creates an ensemble model by combining weak decision trees sequentially. It assigns weights to the output of individual trees. Then it gives incorrect classifications from the first decision tree a higher weight and input to the next tree.

42. How does XGBoost handle missing values?

XGBoost supports missing values by default. tn tree algorithms branch directions for missing values are learned during training. The XGBoost linear treats missing values as zeros. when the missing parameter is specified values in the input predictor that is equal to missing will be treated as missing and removed by default.

43. What are the key hyperparameters in XGBoost and how do they affect model performances?

The key hyperparameters in XGBoost and they affect model performances.

a) n_estimators: Number of boosting rounds. More rounds can improve model performance but increase the risk of overfitting.
b) learning_rate (eta): Shrinks the contribution of each tree. Lower values require more rounds but can lead to better generalization.
c) max_depth: Maximum depth of each tree. Higher values increase model complexity, capturing more patterns but risking overfitting.
d) subsample: Fraction of samples used per tree. Reduces overfitting by introducing randomness.
e) colsample_bytree: Fraction of features used per tree. Helps to prevent overfitting and improve model robustness.

44. Describe the process of gradient boosting in XGBoost.

XGBoost defines Extreme Gradient boosting is a machine learning algorithm under ensemble learning.it is trendy for supervised learning tasks, such as regression and classification, the XGBoost builds a prediction model by combining the predictions of multiple individual models, often decision trees in an iterative manner.

45. What are the advantages and disadvantages of using XGBoost?

Advantages of using XGBoost:
-high accuracy

-fast learning
-easy tuning
-variable ranking.
-it is a new and efficient ensemble learning algorithm with multitudinous advantages.
Disadvantages of using XGBoost:
-it may not be suitable for complex data.
-its classification effect in the case of data imbalance is often not ideal.
-overfitting in high-dimensional data.
-reduced accuracy in small datasets.

Q20. What do you mean by measure of central tendency and measures of dispersion? how it can be calculated?

Measures that indicate the approximate center of distribution are called measures of central tendency. Measures that describe the spending of the data are measures of dispersion. These measures include the mean, median, mode, range, upper, and lower quartiles, variance, and standard deviation.

Q21.what do you mean by skewness? Explain its types. Use the graph to show.

Skewness is a measure of the deviation of a random variable given a distribution from the normal distribution. a normal distribution is without any skewness, as it is symmetric on both sides.

Types of skewness:

a)Positive skewness: if the given distribution is shifted to the left and with its tail on the right side.

b)Negative skewness: if the given distribution is shifted to the right and with its tail on the left side.

c) measuring skewness: measuring newness is measured using the formula that contains mean, medina, and mode.

Q22. Explain probability mass function(PMF) and probability density function(PDF) and what is the differences between them?

Probability mass function:

-The probability mass function denotes the probability that a discrete random variable will take on a particular value.
-it characterizes the distribution of discrete random variable X.
-it takes the number x as input.
-it returns the probability that X = x

Probability density function:

Probability density function gives the probability that a continuous random variable will be between a certain specified interval.it is used for the discrete random variables.

- It characterizes the distribution of discrete random vector X.
- It takes the vector x as input.
- It returns the probability that X = x.

Q23. What is correlation? Explain its type in detail. What are the methods of determining correlation?

Correlation: correlation is a statistics measure that expresses the extent to which two variables are linearly related. It's a common tool for describing simple relationships without making a statement about cause and effect.

Types of correlation:

a) Positive correlation: when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
b) Negative correlation: when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by a decrease/increase in the value of the other variables.

c) No correlation: when there is no linear dependence or no relation between the two variables.
There are three methods of measuring the correlation between two variables: scatter diagram,karl Pearson's coefficient of correlation, and Spearman's rank correlation coefficient.

Q24. Calculate the coefficient of correlation between the marks obtained by 10 students in accountancy and statistics:

Student	1	2	3	4	5	6	7	8	9	10
Accountancy	45	70	65	30	90	40	50	75	85	60
Statistics	35	90	70	40	95	40	60	80	80	50

Use Karl Pearson coefficient of correlation method to find it.

Answer it is in jupyter

Q25.discuss the 4 differences between correlation and regression.

Correlation:

-correction determines the interconnection or a co-relationship between the variables.

-in correction, both the independent and dependent values have no differences.

-correlation stipulates the degree to which both variables can move together.

-correlation helps to constitute the connection between the two variables.

Regression:

-regression determines the independent variables are numerically associated with the dependent variables.

-in regression, both independent and dependent variables are different.

-regression specifies the effect of the change in the unit in the known variable on the evaluated variable.

-regression helps in estimating a variable value based on another given value.

Q26. Find the most likely price at Delhi corresponding to the price of Rs.70 at Agra from the following data: coefficient of correlation between the prices of the two places +0.8.

Answer is in Jupyter

Q27. In a partially destroyed record of an analysis of correlation data, the following results are only legible: variable of x = 9, regression equations are (i) 8x-10y = -66; (ii) 40x – 18y = 214. What are (a) the mean values of x and y.(b) the coefficient of correlation between x and y.(c) the sigma of y?

Answer is in Jupyter notebook

Q28. What is normal distribution? What are the four assumptions of normal distribution? Explain in detail.

The normal distribution is also known as the Gaussian distribution.it is a symmetric about the mean and indicates that values near the mean occur more frequently then the values that are farther away from the eman.

Assumptions of the normal distribution:

a. Symmetric: a symmetric distribution occurs when the values of variables appear at regular frequencies and often the mean, median, and mode all occur at the same point.
b. Unimodal: unimodal distribution is a probability distribution which has a single peak.
c. Asymptotic: asymptotic distribution is a hypothetical distribution that is in a sense the liming distribution of a sequence of distributions.
d. Mean: adding all the numbers in the data and then dividing the number of values in the set.
e. Median: median is the middle number in a sorted ascending or descending list of numbers.
f. Mode: the number which occurs more times that is called a mode.

Q29. Write all the characteristics or properties of the normal distribution curve.

The characteristics of the normal distribution curve include mean, median, and mode.

Mean: adding all the numbers in the data and then dividing the number of values in the set.

Median: the median is the middle number in a sorted ascending or descending list of numbers.

Mode: the number which occurs more times that is called a mode.

Q30. Which of the following options is correct about the normal distribution curve?

(a) Within a range of 0.6745 of a sigma(variable) on both sides the middle 50% of the observations occur i.e., mean +=0.6745sigma covers 50% area 25% on each side.
(b) Mean +=1S.D. covers 68.268% area, and 34.14% of the area lies on either side of the mean.
(c) Mean +=2S.D. covers 95.45% area, and 47.725% area lies on either side of the mean.
(d) Mean +=3S.D. covers 99.73% area, and 49.856% area lies on either side of the mean.
(e) Only 0.27% area is outside the range simma+=3sigma.

All the options are correct about the normal distribution curve.

Q31. The mean of a distribution is 60 with a standard deviation of 10. Assuming that the distribution is normal, what percentage of items be (i) between 60 and 72, (ii) between 50 snd 60, (iii) beyond 72 snd (iv) between 70 and 80?

Answer is in Jupyter notebook

Q32. 15000 students sat for an examination. The mean mark was 49 and the distribution of marks had a standard deviation of 6. Assuming that the marks were normally distributed what proportion of students scored (a) more than 55 marks, (b) more than 70 marks

Answer is in Jupyter notebook

Q33.if the height of 500 students is normally distributed with mean 65 inches and a standard deviation 5 inches. How many students have height: a) greater than 70 inches, b) between 60 and 70 inches.

Answer is in Jupyter notebook

Q34. What is the statistical hypothesis? Explain the errors in hypothesis testing.

Statistical hypothesis is a statement about the nature of a population. It is often stated in terms of a population parameter.

Types of errors in hypothesis testing:

a) Type 1 error: type 1 error is also known as a false positive, takes the place when a true null hypothesis is rejected.in other words, when you believe something is true when it is really not,you must made a type 1 error.the probability of a type 1 error is denoted by the Greek letter alpha.
b) Type 2 error: Type 2 error is also known as the false negative, it occurs when a flse null hypothesis is accepted. The probability of making this error is denoted by the Greek letter beta.

Q35.explain the sample. What are large samples & small samples.

A sample is a subset of individuals or items selected from a larger population, used to make inferences about the entire population. Sampling allows researchers to gather data and make conclusions without needing to study every member of a population.

Large samples: a sample size that is sufficiently large to accurately reflect the populations characteristics and provide reliable statistical analysis.

Small samples: A sample size that is relatively small, which might not fully capture the populations variability and characteristics.

Q36.a random sample of size 25 from a population gives the sample standard derivation to be 9.0. test the hypothesis that the population standard deviation is 10.5

Q37. 100 students of a PW IOI obtained the following grades in data science paper:

Grade :[A,B,C,D,E]

Total Frequency:[15,17,30,22,16,100]

Using the x^2 test, examine the hypothesis that the distribution of grades in uniform.

Q38.Anova Test

To study the performance of three detergents and three different water temperatures the following whiteness readings were obtained with specially designed equipment.

Water temp	Detergents A	Detergents B	Detergents C
Cold Water	57	57	67
Worm Water	49	52	68
Hot water	54	46	58

Q39. How would you create a basic flask route that displays “hello, world!” on the homepage?

The answer is in the Jupyter notebook for the question numbers such as (36,37,38,39)

Q40.Explain how to set up a flask application to handle form submissions using POST requests.

a) Create a html page that will contain our form.
b) Create a flask application that will act as a backend.
c) Run the flask application and fill out the form.
d) Submit the form and view the results.

Q41. Write a flask route that accepts a parameter in the URL and displays it on the page.

This Flask route setup demonstrates how to handle dynamic URL parameters in a web application. By defining a route with a variable placeholder and creating a corresponding route handler function, you can create interactive and personalized web responses based on user input directly in the URL.

Q42. How can you implement user authentication in a flask application?

a) Use the flask-login library for session management.
b) Use the built-in flask utility for hashing passwords.
c) Add protected pages to the app for logged-in users only.
d) Use flask-SQLAIchemy to create a user model.
e) Create sign-up and login forms for the users to create accounts and log in.
f) Flash error messages back to users when something goes wrong.
g) Use information from the users account to display on the profile page.

Q43.describe the process of connecting a flask app to an SQLite database using SQLAIchemy.

Flask is a lightweight Python web framework that provides useful tools and features for creating web applications in the Python programming language. SQLAIchemy is an SQL toolkit that provides efficient and high-performing database access for relational databases.it provides ways to interact with several database engines such as SQLite, MYSQL, and Postgresql. flask SQLAIchemy is a flask extension that makes using SQLAIchemy with flask easier, providing you tools and methods to interact with your database in your flask applications through SQL AIchemy.

Q44. How would you create a RESTFul API endpoint in Flask that returns JSON data?

To create a RESTful API endpoint in Flask that returns JSON data, you define a route using the @app.route decorator and return a Python dictionary from the associated view function. Flask’s jsonify function is used to convert the dictionary into a JSON response. This allows the endpoint to serve structured data in JSON format, which is a common format for APIs.

Q45. Explain how to use Flask-WTF to create and validate forms in a Flask application.

To use flask WTF in the application, first we need to install it using pip and then import it into the application. We then create forms using the WTForms library and use flaks-WTF functions to handle form submissions and validation. In flask WTF forms are defined as classes that extend the FlaskForm class

Q46. How can you implement file uploads in a Flask application?

To implement file uploads in a Flask application, you need to follow these steps:

Create an HTML Form: Design an HTML form with enctype="multipart/form-data" to allow file uploads. This form should include an <input> element of type file for users to select and submit their files.

Define Upload Routes: In Flask, create a route to render the HTML form and another route to handle the file upload.

Configure File Handling: Use Flask's request. files to access the uploaded file. Ensure you check for allowed file extensions and save the file securely to a designated folder. Implement necessary error handling for cases where no file is uploaded or the file type is invalid.

Q47.Describe the steps to create a flask blueprint and why you might use one.

The steps to create a flask blueprint include the first argument "example_blueprint" is the blueprints name which is used by flasks routing mechanism. The second argument "_name_" is the blueprints import name which is flask uses to locate the blueprints resources.

Q48. How would you deploy a Flask by application to a production server using Gunicorn and Nginx?

Gunicorn is a crucial component in hosting the Flask applications. Firstly flasks built in the development server cannot handle multiple requests simultaneously, making it unsuitable for production environments.

Its prefork worker model enables scalability by running numerous worker processes or threads, accommodating increased traffic efficiently. Nginx is a powerful and high-performance web server known for its stability, simple configuration, and low resource consumption. In our deployment nginx is configured as a reverse proxy sitting in front of Gunicorn. It receives client requests and forwards them to Gunicorn which in turn communicates with the flask application. Nginx is also responsible for serving statics files, and potentially load balancing if your application scales to multiple servers,\.

Q49. Make a fully functional web application using flask, MongoDB, signup, sign in page and after successfully login. say hello Greeks message at the webpage.

Flask: A micro web framework used to create web applications easily.

MongoDB: A NoSQL database used to store user data.

Signup & Login: Users can create an account and log in to the application. Credentials are hashed for security.

Personalized Greeting: Once logged in, users see a greeting message, showing that they are authenticated.