Q.1. Random forest is supervised or not?

->Random forest is a **supervised** learning algorithm. A random forest is an ensemble of decision trees combined with a technique called bagging

Q.2 DESCR -> **The full description of the dataset**

Q.4 Max_iter

->Maximum number of iterations taken for the solvers to converge.

Q.5 Fit function

-> The 'fit' method trains the algorithm on the training data, after the model is initialized. The fit() method **takes the training data as arguments, which can be one array in the case of unsupervised learning, or two arrays in the case of supervised learning**.

Q.6 What does high mean square error value indicate?

->a large mean squared error indicates the variance or the bias in your estimator is large.

Q.7 Mean Squared Error: The **mean squared error** (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs.

Q.8 Scatter plots are the **graphs that present the relationship between two variables in a data-set**.
It uses dots to represent values for two different numeric variables

Q.9 Scatter plot package - matplotlib.pyplot

Q.10 **Pandas** is an open-source library that allows to you perform data manipulation and analysis in Python. Pandas Python library offers data manipulation and data operations for numerical tables. It is built on top of NumPy.

Q.11 Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Q.12 A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the **data**, **rows**, and **columns**.

Q.13 A Set is an unordered collection data type that is iterable, mutable and has no duplicate elements.

Q.14 The `split()` method in Python splits characters in a string into separate items in a list.

```
str.split(separator, maxsplit)
```

Q.15 Types of classification -

- Linear regression.
  Linear regression analysis is **used to predict the value of a variable based on the value of another variable**
- Logistic Regression.
  Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
- Naive Bayes.
  Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems
- K-Nearest Neighbors.
  The **k-nearest neighbors (KNN) algorithm** is a data classification method for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to.
- Decision Tree.
  is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome**
- Support Vector Machines.
  Supervised Learning algorithms, which is used for Classification as well as Regression problems.
  The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future

Q.16 Datasets used - boston, breast_cancer()

- Numpy - NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays
- The Sklearn train_test_split function splits a dataset into training data and test data.
- Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc
- The make_blobs() function can be used to **generate blobs of points with a Gaussian distribution**. You can control how many blobs to generate and the number of samples to generate, as well as a host of other properties.
- The multilayer perceptron (MLP) is a feedforward artificial neural network model that maps input data sets to a set of appropriate outputs. An MLP consists of multiple layers and each layer is fully connected to the following one.
- A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging.
- The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.
- Prolog is a logic programming language. It has important role in artificial intelligence. Prolog stands for programming in logic. In the logic programming paradigm, prolog language is most widely available. Prolog is a declarative language, which means that a program consists of data based on the facts and rules (Logical relationship) rather than computing how to find a solution. A logical relationship describes the relationships which hold for the given application.

1. predict() : **given a trained model, predict the label of a new set of data**. This method accepts one argument, the new data X_new (e.g. model. predict(X_new)), and returns the learned label for each object in the array.
2. The numpy.*reshape*() *function* shapes an array without changing the data of the array.

3. load_digits() function **helps to load and return the digit dataset**.
4. The **mean squared error** (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them.
5. matplotlib. pyplot is **a collection of functions that make matplotlib work like MATLAB**. Each pyplot function makes some change to a figure e.g. plots some lines in a plotting area, decorates the plot with labels, etc. In matplotlib.
6. A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.
7. matplotlib inline to **enable the inline plotting, where the plots/graphs will be displayed just below the cell where your plotting commands are written**.
8. Lasso regression is **a machine learning algorithm that can be used to perform linear regression while also reducing the number of features used in the model**. Lasso stands for least absolute shrinkage and selection operator
9. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models
10. Ridge regression is **the regularization technique that performs L2 regularization**. It modifies the loss function by adding the penalty (shrinkage quantity) equivalent to the square of the magnitude of coefficients.
11. corr() is **used to find the pairwise correlation of all columns in the dataframe.** The correlation of a variable with itself is 1.
12. The arange() function is **used to get evenly spaced values within a given interval**. Values are generated within the half-open interval [start, stop]