1-Ans—

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs**that can access data and use it to learn for themselves.

2-Ans--

Supervised Learning—

1. Supervised learning algorithms are trained using labelled data.

2. Supervised learning model takes direct feedback to check if it is predicting correct output or not.

3. Supervised learning model predicts the output.

4. In supervised learning, input data is provided to the model along with the output.

5. The goal of supervised learning is to train the model so that it can predict the output when it is given new data.

6. Supervised learning needs supervision to train the model.

7. Supervised learning can be categorized in **Classification** and **Regression** problems

8. Supervised learning can be used for those cases where we know the input as well as corresponding outputs.

9. Supervised learning model produces an accurate result.

10. Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.

11. It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.

Unsupervised Learning---

1. Unsupervised learning algorithms are trained using unlabelled data.
2. Unsupervised learning model does not take any feedback.
3. Unsupervised learning model finds the hidden patterns in data.
4. In unsupervised learning, only input data is provided to the model.
5. The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
6. Unsupervised learning does not need any supervision to train the model.
7. Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
8. Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
9. Unsupervised learning model may give less accurate result as compared to supervised learning.
10. Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
11. It includes various algorithms such as Clustering, KNN, and Apriorism algorithm.

3-Ans—

VALIDATION SET--A validation dataset is a [dataset](https://en.wikipedia.org/wiki/Dataset) of examples used to tune the [hyperparameters](https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)) of a classifier. It is sometimes also called the development set or the dev set. An example of a hyperparameter for [artificial neural networks](https://en.wikipedia.org/wiki/Artificial_neural_networks) includes the number of hidden units in each layer. It, as well as the testing set should follow the same probability distribution as the training dataset. In order to avoid overfitting, when any [classification](https://en.wikipedia.org/wiki/Statistical_classification) parameter needs to be adjusted, it is necessary to have a validation dataset in addition to the training and test datasets. For example, if the most suitable classifier for the problem is sought, the training dataset is used to train the different candidate classifiers, the validation dataset is used to compare their performances and decide which one to take and, finally, the test dataset is used to obtain the performance characteristics such as [accuracy](https://en.wikipedia.org/wiki/Accuracy), [sensitivity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity), [specificity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity), [F-measure](https://en.wikipedia.org/wiki/Precision_and_recall#F-measure), and so on. The validation dataset functions as a hybrid: it is training data used for testing, but neither as part of the low-level training nor as part of the final testing. The basic process of using a validation dataset for [model selection](https://en.wikipedia.org/wiki/Model_selection) is: Since our goal is to find the network having the best performance on new data, the simplest approach to the comparison of different networks is to evaluate the error function using data which is independent of that used for training. Various networks are trained by minimization of an appropriate error function defined with respect to a training data set. The performance of the networks is then compared by evaluating the error function using an independent validation set, and the network having the smallest error with respect to the validation set is selected. This approach is called the *hold out* method. Since this procedure can itself lead to some overfitting to the validation set, the performance of the selected network should be confirmed by measuring its performance on a third independent set of data called a test set. An application of this process is in [early stopping](https://en.wikipedia.org/wiki/Early_stopping), where the candidate models are successive iterations of the same network, and training stops when the error on the validation set grows, choosing the previous model.

Use of validation dataset-- The validation dataset is different from the test dataset that is also held back from the training of the model, but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

TEST SET--A test dataset is a [dataset](https://en.wikipedia.org/wiki/Dataset) that is [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of the training dataset, but that follows the same [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal [overfitting](https://en.wikipedia.org/wiki/Overfitting) has taken place. A better fitting of the training dataset as opposed to the test dataset usually points to overfitting. A test set is therefore a set of examples used only to assess the performance of a fully specified classifier. To do this, the final model is used to predict classifications of examples in the test set. Those predictions are compared to the examples' true classifications to assess the model's accuracy. In a scenario where both validation and test datasets are used, the test dataset is typically used to assess the final model that is selected during the validation process. In the case where the original dataset is partitioned into two subsets, the test dataset might assess the model only once. Note that some sources advise against such a method. However, when using a method such as [cross-validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)), two partitions can be sufficient and effective since results are averaged after repeated rounds of model training and testing to help reduce bias and variability.

Use of test dataset--the model should be evaluated on samples that were not used to build or fine-tune the model, so that they provide an unbiased sense of model effectiveness.

--We want to use them because ML needs data to train itself for a skill This type of data set models help us to evaluate the data effectually and accurately.

4-Ans---

7 main processing steps

* Acquire the dataset
* Import all the crucial libraries
* Import the dataset
* Identifying and handling the missing values
* Encoding the categorical data
* Splitting the dataset
* Feature scaling

1. Acquire the dataset- This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. For instance, a business dataset will be entirely different from a medical dataset. While a business dataset will contain relevant industry and business data, a medical dataset will include healthcare-related data.
2. Import all the crucial libraries-- Python is the most extensively used and also the most preferred library by Data Scientists around the world, we’ll show you how to import Python libraries for data pre-processing in Machine Learning. Read more about [Python libraries for Data Science here.](https://www.upgrad.com/blog/python-libraries-for-data-science/) The predefined Python libraries can perform specific data pre-processing jobs.

Like-NumPy, pandas etc..

1. Import the dataset-- you need to import the dataset/s that you have gathered for the ML project at hand. However, before you can import the dataset/s, you must set the current directory as the working directory.

4. Identifying and handling the missing values

In data pre-processing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data. Needless to say, this will hamper your ML project.

there are two ways to handle missing data:

* 1. Deleting a particular row
  2. Calculating the mean

### 5. Encoding the categorical data

Categorical data refers to the information that has specific categories within the dataset. In the dataset cited above, there are two categorical variables – country and purchased.

Machine Learning models are primarily based on mathematical equations. Thus, you can intuitively understand that keeping the categorical data in the equation will cause certain issues since you would only need numbers in the equations.

5-Ans—

A discrete variable is a variable whose value is obtained by counting.

*Examples*:     number of students present

                                    number of red marbles in a jar

                                    number of heads when flipping three coins

                                    students’ grade level

A continuous variable is a variable whose value is obtained by measuring.

*Examples*:     height of students in class

                        weight of students in class

                        time it takes to get to school

                        distance travelled between classes

or

A continuous variable is one which can take on an [uncountable set](https://en.wikipedia.org/wiki/Uncountable_set) of values.

For example, a variable over a non-empty range of the [real numbers](https://en.wikipedia.org/wiki/Real_number) is continuous, if it can take on any value in that range. The reason is that any range of real numbers between {\displaystyle a}and {\displaystyle b}with {\displaystyle a,b\in \mathbb {R} ;a\neq b}is infinite and uncountable.

6Ans—The given diagram represents a histogram and it’s variable type is discrete variable type.