

# ML Module-1 Assignment

1. What is a parameter?

Ans. A parameter is a variable listed inside the parentheses of a function definition. It defines what kind of input the function expects to receive. When the function is called, the actual values passed in are called arguments.

2. What is correlation? What does negative correlation mean?

Ans. Correlation is a statistical measure that shows the relationship between two features — how the change in one feature affects the other.

If increasing one feature causes the other to decrease, it's a **negative correlation**.

3. Define Machine Learning. What are the main components in Machine Learning?

ANS. Machine Learning is a field of computer science where machines learn from data and make predictions or decisions without being explicitly programmed.

The main components of a Machine Learning system are:

1. **Data** – Raw input that the model learns from
2. **Features** – Important variables extracted from the data
3. **Model** – The algorithm that learns the relationship in the data
4. **Loss Function** – Measures the error or difference between predicted and actual values
5. **Optimiser** – Adjusts the model to reduce the error
6. **Evaluation Metrics** – Used to assess the model's performance

4. How does loss value help in determining whether the model is good or not?

ANS. The loss value shows how far the model's predictions are from the actual results. A low and consistently decreasing loss means the model

is learning well, while a high or stagnant loss means it's not performing effectively

5. What are continuous and categorical variables?

ANS. Continuous variables are numeric values that can take any value within a range, like height, weight, or temperature. Categorical variables are non-numeric and represent categories or labels, like gender, yes/no, or present/absent

6. How do we handle categorical variables in Machine Learning? What are the common techniques?

ANS. We handle categorical variables by replacing missing values, often with the mode, and then converting them into numbers so models can understand them. Common techniques include Label Encoding, One-Hot Encoding, and Target Encoding. This ensures the model can use categorical data effectively in training

7. What do you mean by training and testing a dataset?

ANS. Training a dataset means using part of the data to teach the model patterns and relationships, after preprocessing steps like handling null values and converting to numeric form. Testing a dataset means using new, unseen data to check how well the trained model performs and predicts

8. What is sklearn.preprocessing?

ANS. sklearn.preprocessing is a module in the Scikit-learn library that provides functions to transform and scale data, such as normalization, standardization, encoding categorical variables, and handling missing values, to prepare it for machine learning models

9. What is a Test set?

ANS. A test set is unseen data used to evaluate how well a trained model performs on new data and to measure its accuracy or generalization ability.

10. How do we split data for model fitting (training and testing) in Python? How do you approach a Machine Learning problem?

ANS. We usually split data into training and testing sets using `train_test_split` from Scikit-learn, for example 80% training and 20% testing. When approaching a machine learning problem, I first understand the problem, collect and clean the data, explore it with EDA, preprocess it, choose and train a model, evaluate it on the test set, and then improve it if needed.

11. Why do we have to perform EDA before fitting a model to the data?

ANS. EDA helps us understand the data's structure, detect patterns, spot outliers, and find missing values before modeling. This ensures data quality, guides preprocessing, and helps choose the right model for better results.

12. What is correlation?

ANS. Correlation is a statistical measure that shows the relationship between two features — how the change in one feature affects the other.

13. What does negative correlation mean?

ANS. If increasing one feature causes the other to decrease, it's a **negative correlation**.

14. How can you find correlation between variables in Python?

ANS. We can find correlation between variables in Python using the `corr()` method in pandas, which returns a correlation matrix, and visualize it with a heatmap using Seaborn for better understanding.

15. What is causation? Explain difference between correlation and causation with an example.

ANS. Causation means that one event directly affects another. Correlation means two variables move together, but one doesn't necessarily cause the other. For example, ice cream sales and drowning cases may be correlated because both increase in summer, but eating ice cream doesn't cause drowning — the real cause is hotter weather.

16. What is an Optimizer? What are different types of optimizers? Explain each with an example.

ANS. An optimizer is an algorithm that adjusts a model's parameters to minimize the loss function during training. Common types include:

1. **Gradient Descent** – Updates parameters in the direction of the steepest decrease in loss. Example: Linear Regression.
2. **Stochastic Gradient Descent (SGD)** – Updates parameters using one sample at a time for faster but noisier convergence. Example: Logistic Regression.
3. **Adam** – Combines momentum and adaptive learning rates for efficient training. Example: Deep Neural Networks.
4. **RMSprop** – Adjusts learning rate based on recent gradients, good for recurrent networks. Example: RNN training.

17. What is `sklearn.linear_model`?

ANS. `sklearn.linear_model` is a module in Scikit-learn that provides linear models for regression and classification, such as Linear Regression, Logistic Regression, Ridge, and Lasso, to model relationships between variables

18. What does `model.fit()` do? What arguments must be given?

ANS. `model.fit()` trains the model by learning patterns from the training data. It takes the input features (x) and target values (y) as mandatory arguments, and may also include parameters like epochs, batch size, or validation data depending on the model.

19. What does `model.predict()` do? What arguments must be given?

ANS. `model.predict()` uses a trained model to generate predictions for new input data. It takes the feature data (x) as the mandatory argument and returns the predicted output values.

20. What are continuous and categorical variables?

ANS. Continuous variables are numeric values that can take any value within a range, like height, weight, or temperature. Categorical variables

represent discrete categories or labels, like gender, yes/no, or present/absent.

21. What is feature scaling? How does it help in Machine Learning?

ANS. Feature scaling is the process of bringing all numeric features to a similar range, such as 0–1 or -1–1. It helps machine learning models converge faster and improves performance, especially for algorithms sensitive to feature magnitude like KNN, SVM, and gradient descent-based models.

22. How do we perform scaling in Python?

ANS. We perform scaling in Python using Scikit-learn's `sklearn.preprocessing` module. Common methods are `StandardScaler()` for standardization (mean = 0, std = 1) and `MinMaxScaler()` for normalization (range = 0 to 1). We fit the scaler on the training data and transform both training and test sets.

23. What is `sklearn.preprocessing`?

ANS. `sklearn.preprocessing` is a module in Scikit-learn that provides tools for preparing data before modeling, such as scaling, normalization, encoding categorical variables, and handling missing values

24. How do we split data for model fitting (training and testing) in Python?

ANS. We split data into training and testing sets using the `train_test_split` function from Scikit-learn, often with an 80:20 or 70:30 ratio. The training set is used to fit the model, and the test set is used to evaluate its performance on unseen data.

25. Explain data encoding?

ANS. Data encoding is the process of converting categorical data into numerical form so machine learning models can understand and process it. Common methods include Label Encoding, One-Hot Encoding, and Target Encoding.