**Classification :**

Classification is a process of categorizing data or objects into predefined classes or categories based on their features or attributes.

**Classification Types**

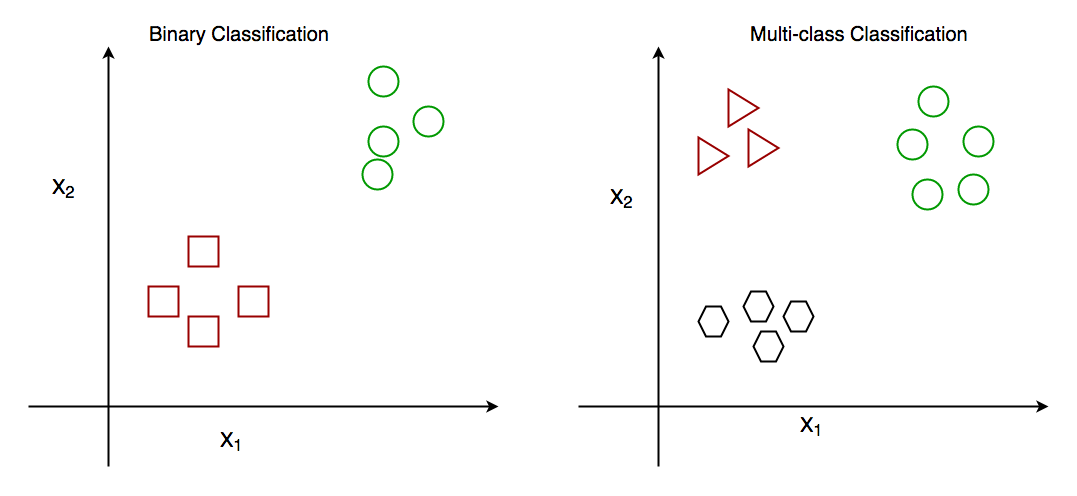
There are two main classification types in machine learning:

**Binary Classification**

In binary classification, the goal is to classify the input into one of two classes or categories. Example – On the basis of the given health conditions of a person, we have to determine whether the person has a certain disease or not.

**Multiclass Classification**

In multi-class classification, the goal is to classify the input into one of several classes or categories. For Example – On the basis of data about different species of flowers, we have to determine which specie our observation belongs to.



*Binary vs Multi class classification*

Other categories of classification involves:

**Multi-Label Classification**

In, [**Multi-label Classification**](https://www.geeksforgeeks.org/an-introduction-to-multilabel-classification/) the goal is to predict which of several labels a new data point belongs to. This is different from multiclass classification, where each data point can only belong to one class. For example, a multi-label classification algorithm could be used to classify images of animals as belonging to one or more of the categories cat, dog, bird, or fish.

**Imbalanced Classification**

In, [**Imbalanced Classification**](https://www.geeksforgeeks.org/handling-imbalanced-data-for-classification/)the goal is to predict whether a new data point belongs to a minority class, even though there are many more examples of the majority class. For example, a medical diagnosis algorithm could be used to predict whether a patient has a rare disease, even though there are many more patients with common diseases.

## **Comparison between Classification and Regression**

| **Classification** | **Regression** |
| --- | --- |
| In this problem statement, the target variables are discrete. | In this problem statement, the target variables are continuous. |
| Problems like [Spam Email Classification](https://www.geeksforgeeks.org/detecting-spam-emails-using-tensorflow-in-python/), [Disease prediction](https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/) like problems are solved using Classification Algorithms. | Problems like [House Price Prediction](https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/), [Rainfall Prediction](https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/) like problems are solved using regression Algorithms. |
| In this algorithm, we try to find the best possible decision boundary which can separate the two classes with the maximum possible separation. | In this algorithm, we try to find the best-fit line which can represent the overall trend in the data. |
| [Evaluation metrics](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/) like Precision, Recall, and F1-Score are used here to evaluate the performance of the classification algorithms. | Evaluation metrics like [Mean Squared Error,](https://www.geeksforgeeks.org/python-mean-squared-error/) [R2-Score](https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/), and  [MAPE](https://www.geeksforgeeks.org/how-to-calculate-mape-in-python/) are used here to evaluate the performance of the regression algorithms. |
| Here we face the problems like [binary Classification](https://www.geeksforgeeks.org/getting-started-with-classification/) or [Multi-Class Classification](https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/) problems. | Here we face the problems like [Linear Regression](https://www.geeksforgeeks.org/ml-linear-regression/) models as well as non-linear models. |
| Input Data are Independent variables and categorical dependent variable. | Input Data are Independent variables and continuous dependent variable. |
| The classification algorithm’s task mapping the input value of x with the discrete output variable of y. | The regression algorithm’s task is mapping input value (x) with continuous output variable (y). |
| Output is Categorical labels. | Output is Continuous numerical values. |
| Objective is to  Predict categorical/class labels. | Objective is to Predicting continuous numerical values. |
| Example use cases are Spam detection, image recognition, sentiment analysis | Example use cases are Stock price prediction, house price prediction, demand forecasting. |
| **Examples of classification algorithms are:**  Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Naive Bayes, Neural Networks, K-Means Clustering, Multi-layer Perceptron (MLP), etc. | **Examples of regression algorithms are:**  Linear Regression, Polynomial Regression, Ridge R egression, Lasso Regression, Support Vector Regression (SVR), Decision Trees for Regression, Random Forest Regression, K-Nearest Neighbors (K-NN) Regression, Neural Networks for Regression, etc. |

**Logistic regression :**

Logistic regression is used for binary [classification](https://www.geeksforgeeks.org/getting-started-with-classification/) where we use [sigmoid function](https://www.geeksforgeeks.org/derivative-of-the-sigmoid-function/), that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It’s referred to as regression because it is the extension of[linear regression](https://www.geeksforgeeks.org/ml-linear-regression/) but is mainly used for classification problems.

### Key Points:

* Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
* It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
* In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).

## **Types of Logistic Regression**

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. **Ordinal:**In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

**Cross validation :**

The main purpose of cross validation is to prevent [overfitting](https://www.geeksforgeeks.org/overfitting-and-regularization-in-ml/), which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. By evaluating the model on multiple validation sets, cross validation provides a more realistic estimate of the model’s generalization performance, i.e., its ability to perform well on new, unseen data.

## Types of Cross-Validation

There are several types of cross validation techniques, including **k-fold cross validation, leave-one-out cross validation, and Holdout validation, Stratified Cross-Validation.**The choice of technique depends on the size and nature of the data, as well as the specific requirements of the modeling problem.

### ****1. Holdout Validation****

In[Holdout Validation](https://www.geeksforgeeks.org/introduction-of-holdout-method/), we perform training on the 50% of the given dataset and rest 50% is used for the testing purpose. It’s a simple and quick way to evaluate a model. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e. higher bias.

### ****2. LOOCV (Leave One Out Cross Validation)****

In this method, we perform training on the whole dataset but leaves only one data-point of the available dataset and then iterates for each data-point. In [LOOCV](https://www.geeksforgeeks.org/loocvleave-one-out-cross-validation-in-r-programming/), the model is trained on  samples and tested on the one omitted sample, repeating this process for each data point in the dataset. It has some advantages as well as disadvantages also.

**An advantage** of using this method is that we make use of all data points and hence it is low bias.

The major**drawback**of this method is that it leads to **higher variation**in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation. Another drawback is it **takes a lot of execution time** as it iterates over ‘the number of data points’ times.

### ****3. Stratified Cross-Validation****

It is a technique used in machine learning to ensure that each fold of the cross-validation process maintains the same class distribution as the entire dataset. This is particularly important when dealing with imbalanced datasets, where certain classes may be underrepresented. In this method,

1. The dataset is divided into k folds while maintaining the proportion of classes in each fold.
2. During each iteration, one-fold is used for testing, and the remaining folds are used for training.
3. The process is repeated k times, with each fold serving as the test set exactly once.

[Stratified Cross-Validation](https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/)is essential when dealing with classification problems where maintaining the balance of class distribution is crucial for the model to generalize well to unseen data.

### ****4. K-Fold Cross Validation****

In [K-Fold Cross Validation](https://www.geeksforgeeks.org/k-fold-cross-validation-in-r-programming/), we split the dataset into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

**Methods for Checking perforfance**

<https://www.geeksforgeeks.org/machine-learning-model-evaluation/>

**Hyper parameter tuning:**

<https://www.geeksforgeeks.org/hyperparameter-tuning/>

**Bagging v/s Boosting:**

<https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/>

**Random forest: :**

<https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

**Gradiend boosting:**

<https://www.geeksforgeeks.org/ml-gradient-boosting/>