Machine Learning - Unit-2 Notes
Session: 2023-24

Branch: CSE **Semester: 6th sem**

Syllabus:

Supervised Learning:
Regression: Data pre-processing**,** Dimensionality reduction, feature subset selection,
Types of regression: Multiple linear regression, Polynomial regression model.

1. Machine Learning

Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.

Machine learning methods enable computers to operate autonomously without explicit programming. ML applications are fed with new data, and they can independently learn, grow, develop, and adapt.

Machine learning derives insightful information from large volumes of data by leveraging algorithms to identify patterns and learn in an iterative process. ML algorithms use computation methods to learn directly from data instead of relying on any predetermined equation that may serve as a model.

The performance of ML algorithms adaptively improves with an increase in the number of available samples during the 'learning' processes. For example, deep learning is a sub-domain of machine learning that trains computers to imitate natural human traits like learning from examples. It offers better performance parameters than conventional ML algorithms.

While machine learning is not a new concept – dating back to World War II when the Enigma Machine was used – the ability to apply complex mathematical calculations automatically to growing volumes and varieties of available data is a relatively recent development.

Today, with the rise of big data, IoT, and ubiquitous computing, machine learning has become essential for solving problems across numerous areas, such as

- Computational finance (credit scoring, algorithmic trading)
- Computer vision (facial recognition, motion tracking, object detection)
- Computational biology (DNA sequencing, brain tumor detection, drug discovery)
- Automotive, aerospace, and manufacturing (predictive maintenance)
- Natural language processing (voice recognition)

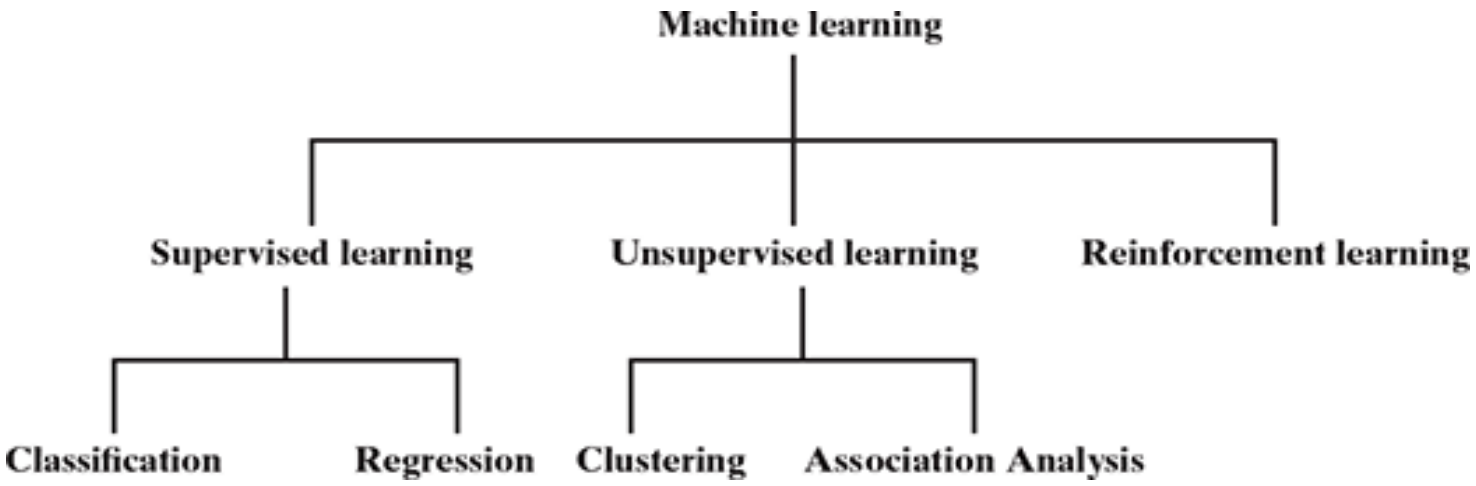


Fig-1: Types of Machine Learning

2. Supervised Learning

Supervised learning: Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

- If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –Apple.
- If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –Banana.

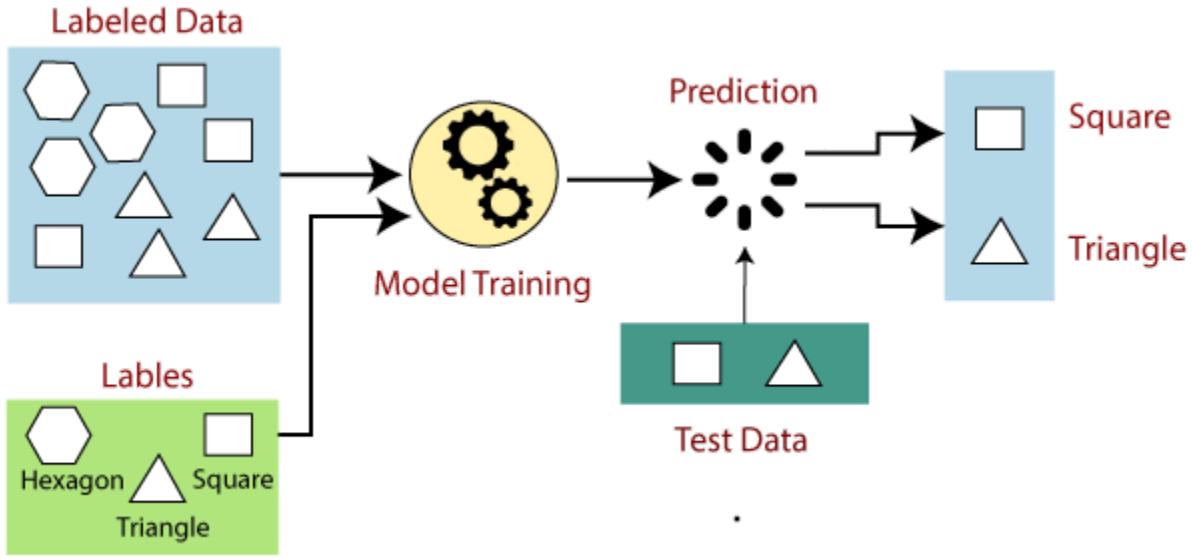
Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.

Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

Supervised learning is classified into two categories of algorithms:

- Classification: A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.
- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.



Types:-

- Regression
- Logistic Regression
- Classification
- Naive Bayes Classifiers
- K-NN (k nearest neighbors)
- Decision Trees
- Support Vector Machine

Advantages:-

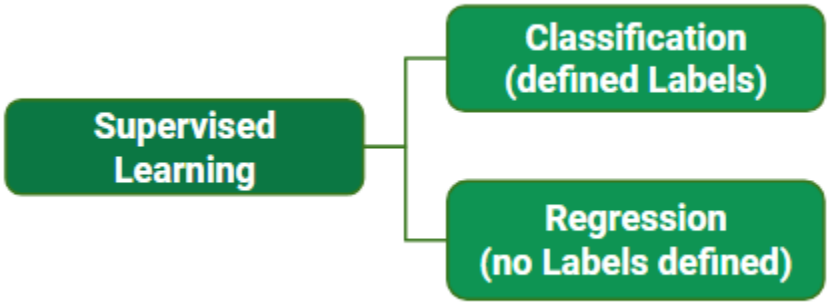
- Supervised learning allows collecting data and produces data output from previous experiences.
- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.
- It performs classification and regression tasks.
- It allows estimating or mapping the result to a new sample.
- We have complete control over choosing the number of classes we want in the training data.

Disadvantages:-

- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
- Supervised learning cannot handle all complex tasks in Machine Learning.
- Computation time is vast for supervised learning.

- It requires a labelled data set.
- It requires a training process.

3. Types of Supervised Learning



Supervised learning can be further classified into two categories:

Regression: In regression, the target variable is a continuous value. The goal of regression is to predict the value of the target variable based on the input variables. Linear regression, polynomial regression, and decision trees are some of the examples of regression algorithms.

Classification: In classification, the target variable is a categorical value. The goal of classification is to predict the class or category of the target variable based on the input variables. Some examples of classification algorithms include logistic regression, decision trees, support vector machines, and neural networks.

4. Data pre-processing

Data Preprocessing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine.

The main agenda for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the data's features.

4.1 Why is Data Preprocessing important?

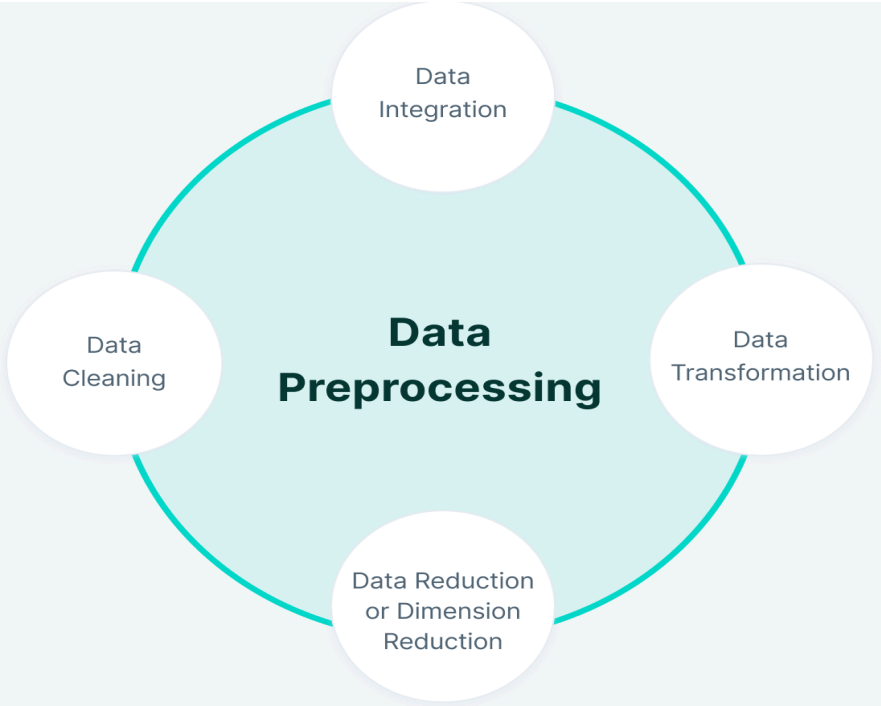
The majority of the real- world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy due to their heterogeneous origin.

Applying data mining algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality.

- Duplicate or missing values may give an incorrect view of the overall statistics of data.
- Outliers and inconsistent data points often tend to disturb the model's overall learning, leading to false predictions.

Quality decisions must be based on quality data. Data Preprocessing is important to get this quality data, without which it would just be a Garbage In, Garbage Out scenario.

4.2 Data Preprocessing Steps



4.2.1 Data Cleaning

Data Cleaning is particularly done as part of data preprocessing to clean the data by filling missing

values, smoothing the noisy data, resolving the inconsistency, and removing outliers.

4.2.1.1. Missing values

Here are a few ways to solve this issue:

- Ignore those tuples

This method should be considered when the dataset is huge and numerous missing values are present within a tuple.

- Fill in the missing values

There are many methods to achieve this, such as filling in the values manually, predicting the missing values using regression method, or numerical methods like attribute mean.

4.2.1.2 Noisy Data

It involves removing a random error or variance in a measured variable. It can be done with the help of the following techniques:

- Binning

It is the technique that works on sorted data values to smoothen any noise present in it. The data is divided into equal-sized bins, and each bin/bucket is dealt with independently. All data in a segment can be replaced by its mean, median or boundary values.

- Regression

This data mining technique is generally used for prediction. It helps to smoothen noise by fitting all the data points in a regression function. The linear regression equation is used if there is only one independent attribute; else Polynomial equations are used.

- Clustering

Creation of groups/clusters from data having similar values. The values that don't lie in the cluster can be treated as noisy data and can be removed.

4.2.1.3 Removing outliers

Clustering techniques group together similar data points. The tuples that lie outside the cluster are outliers/inconsistent data.

4.2.2 Data Integration

Data Integration is one of the data preprocessing steps that are used to merge the data present in multiple sources into a single larger data store like a data warehouse.

Data Integration is needed especially when we are aiming to solve a real-world scenario like detecting the presence of nodules from CT Scan images. The only option is to integrate the images from multiple

medical nodes to form a larger database.

4.2.3 Data Transformation

Once data clearing has been done, we need to consolidate the quality data into alternate forms by changing the value, structure, or format of data using the below-mentioned Data Transformation strategies.

4.2.3.1 Generalization

The low-level or granular data that we have converted to high-level information by using concept hierarchies. We can transform the primitive data in the address like the city to higher-level information like the country.

4.2.3.2 Normalization

It is the most important Data Transformation technique widely used. The numerical attributes are scaled up or down to fit within a specified range. In this approach, we are constraining our data attribute to a particular container to develop a correlation among different data points. Normalization can be done in multiple ways, which are highlighted here:

- Min-max normalization
- Z-Score normalization
- Decimal scaling normalization

4.2.3.3 Attribute Selection

New properties of data are created from existing attributes to help in the data mining process. For example, date of birth, data attribute can be transformed to another property like is_senior_citizen for each tuple, which will directly influence predicting diseases or chances of survival, etc.

4.2.3.4 Aggregation

It is a method of storing and presenting data in a summary format. For example sales, data can be aggregated and transformed to show as per month and year format.

4.2.4 Data Reduction

The size of the dataset in a data warehouse can be too large to be handled by data analysis and data mining algorithms.

One possible solution is to obtain a reduced representation of the dataset that is much smaller in volume but produces the same quality of analytical results.

Here is a walkthrough of various Data Reduction strategies.

4.2.4.1 Data cube aggregation

It is a way of data reduction, in which the gathered data is expressed in a summary form.

4.2.4.2 Dimensionality reduction

Dimensionality reduction techniques are used to perform feature extraction. The dimensionality of a dataset refers to the attributes or individual features of the data. This technique aims to reduce the number of redundant features we consider in machine learning algorithms. Dimensionality reduction can be done using techniques like Principal Component Analysis etc.

4.2.4.3 Data compression

By using encoding technologies, the size of the data can significantly reduce. But compressing data can be either lossy or non-lossy. If original data can be obtained after reconstruction from compressed data, this is referred to as lossless reduction; otherwise, it is referred to as lossy reduction.

4.2.4.4 Discretization

Data discretization is used to divide the attributes of the continuous nature into data with intervals. This is done because continuous features tend to have a smaller chance of correlation with the target variable. Thus, it may be harder to interpret the results. After discretizing a variable, groups corresponding to the target can be interpreted. For example, attribute age can be discretized into bins like below 18, 18-44, 44-60, above 60.

4.2.4.5 Numerosity reduction

The data can be represented as a model or equation like a regression model. This would save the burden of storing huge datasets instead of a model.

4.2.4.6 Attribute subset selection

It is very important to be specific in the selection of attributes. Otherwise, it might lead to high dimensional data, which are difficult to train due to underfitting/overfitting problems. Only attributes that add more value towards model training should be considered, and the rest all can be discarded.

5. Feature Subset Selection

Feature Selection is the most critical pre-processing activity in any machine learning process. It intends to select a subset of attributes or features that makes the most meaningful contribution to a machine learning activity. In order to understand it, let us consider a small example i.e. Predict the weight of students based on the past information about similar students, which is captured inside a 'Student Weight' data set. The data set has 04 features like Roll Number, Age, Height & Weight. Roll Number has no effect on the weight of the students, so we eliminate this feature. So now the new data set will be having only 03 features. This subset of the data set is expected to give better results than the full set.

Age	Height	Weight
12	1.1	23
11	1.05	21.6
13	1.2	24.7
11	1.07	21.3
14	1.24	25.2

12	**1.12**	**23.4**

The above data set is a reduced dataset. Before proceeding further, we should look at the fact why we have reduced the dimensionality of the above dataset OR what are the issues in High Dimensional Data?

High Dimensional refers to the high number of variables or attributes or features present in certain data sets, more so in the domains like DNA analysis, geographic information system (GIS), etc. It may have sometimes hundreds or thousands of dimensions which is not good from the machine learning aspect because it may be a big challenge for any ML algorithm to handle that. On the other hand, a high quantity of computational and a high amount of time will be required. Also, a model built on an extremely high number of features may be very difficult to understand. For these reasons, it is necessary to take a subset of the features instead of the full set. So we can deduce that the objectives of feature selection are:

1. Having a faster and more cost-effective (less need for computational resources) learning model
2. Having a better understanding of the underlying model that generates the data.
3. Improving the efficacy of the learning model.

Main Factors Affecting Feature Selection

a. Feature Relevance: In the case of supervised learning, the input data set (which is the training data set), has a class label attached. A model is inducted based on the training data set — so that the inducted model can assign class labels to new, unlabeled data. Each of the predictor variables, ie expected to contribute information to decide the value of the class label. In case of a variable is not contributing any information, it is said to be irrelevant. In case the information contribution for prediction is very little, the variable is said to be weakly relevant. The remaining variables, which make a significant contribution to the prediction task are said to be strongly relevant variables.

In the case of unsupervised learning, there is no training data set or labelled data. Grouping of similar data instances are done and the similarity of data instances are evaluated based on the value of different variables. Certain variables do not contribute any useful information for deciding the similarity of dissimilar data instances. Hence, those variable makes no significant contribution to the grouping process. These variables are marked as irrelevant variables in the context of the unsupervised machine learning task.

We can understand the concept by taking a real-world example: At the start of the article, we took a random dataset of the student. In that, Roll Number doesn't contribute any significant information in predicting what the Weight of a student would be. Similarly, if we are trying to group together students

with similar academic capabilities, *Roll No* can really not contribute any information. So, in the context of grouping students with similar academic merit, the variable *Roll No* is quite irrelevant. Any feature which is irrelevant in the context of a machine learning task is a candidate for rejection when we are selecting a subset of features.

b. Feature Redundancy: A feature may contribute to information that is similar to the information contributed by one or more features. For example, in the Student Data-set, both the features Age & Height contribute similar information. This is because, with an increase in age, weight is expected to increase. Similarly, with the increase in Height also weight is expected to increase. So, in context to that problem, Age and Height contribute similar information. In other words, irrespective of whether the feature Height is present or not, the learning model will give the same results. In this kind of situation where one feature is similar to another feature, the feature is said to be potentially redundant in the context of a machine learning problem.

All features having potential redundancy are candidates for rejection in the final feature subset. Only a few representative features out of a set of potentially redundant features are considered for being a part of the final feature subset. So in short, the main objective of feature selection is to remove all features which are irrelevant and take a representative subset of the features which are potentially redundant. This leads to a meaningful feature subset in the context of a specific learning task.

6. Linear Regression

The "Supervised Machine Learning" algorithm of regression is used to forecast continuous features.

The simplest regression procedure, linear regression, fits a linear equation or "best fit line" to the observed data in an effort to explain the connection between the dependent variable one and or more independent variables.

Simple Linear Regression Model

A form of regression method called simple linear regression simulates the relationship between a given independent variable and a dependent variable. A Simple Linear Regression model displays a linear or sloping straight-line relationship.

A straight line can be used in simple linear regression to establish the link between two variables. Finding the slope and intercept, which both define the line and reduce regression errors, is the first step in drawing the line.

One x variable and one y variable make up simple linear regression's most basic version. Because it cannot be predicted by the dependent variable, the x variable is the independent variable. The fact that the y variable is contingent on the prediction we make makes it the dependent variable.

The dependent variable for simple linear regression must have a continuous or real value. However, the independent variable can indeed be assessed on continuous or categorical values.

The two major goals of the simple linear regression algorithm are as follows

Model how the two variables are related. For instance, the connection between income and spending, experience and pay, etc.

Anticipating fresh observations Examples include predicting the weather based on temperature, calculating a company's revenue based on its annual investments, etc.

The equation below can be used to illustrate the Simple Linear Regression model

$$\mathbf{y= a0+a1x+ \varepsilon}$$

Where

The regression line's intercept, denoted by the symbol a0, can be obtained by putting x=0.

The slope of the regression line, or a1, indicates whether the line is rising or falling.

$$\mathbf{\varepsilon = The\ incorrect\ term.}$$

7. Multiple Linear Regression

Multiple linear regression is a style of predictive analysis that is frequently used. You can comprehend the relationship between such a continuous dependent variable and two or more independent variables using this kind of analysis.

The independent variables may be categorical or continuous, such as age and height (like gender and occupation). It's crucial to remember that before performing the analysis, if given dependent variable is categorical, one should pseudo code it.

Formula and Calculation

Multiple regression analysis allows for the simultaneous control of several factors that affect the dependent variable. The link between independent variables and dependent variables can be examined using regression analysis.

Let k stand for the quantity of variables denoted by the letters x1, x2, x3… xk.

To use this strategy, we must suppose that we have k independent variables that we may set. These variables will then probabilistically decide the result Y.

Additionally, we presume that Y is directly dependent on the variables as

$$\mathbf{Y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta k x k + \varepsilon}$$

It depends on or is projected that the variable yi

The y-intercept determines the slope of y, therefore when xi and x2 are both zero, y will be 0.

- The one-unit changes in xi1 and xi2 that cause changes in y are represented by the regression coefficients 1 and 2.
- The slope coefficient of all independent variables is denoted by the symbol p.
- The random error (residual) in the model is described by the phrase.
- Except for the requirement that k not equal 1, this is identical to simple linear regression where is a standard error.

We have more than k observations, with n often being substantially higher.

We measure a value yi for the random variable Yi and assign the independent variables to the values xi1, xi2..., xik, for the ith observation.

As a result, the equations can be used to describe the model.

$$\mathbf{Yi = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + \cdots + \beta k xik + i \text{ for } i = 1, 2, \ldots, n}$$

where the mistakes i are separate standard variables with the same unknown variance of 2 and a mean of 0.

8. Polynomial Regression Model

The link between the dependent and independent variables, Y and X, is modelled as the nth degree of the polynomial in polynomial regression, a type of linear regression. In order to draw the best line using data points, this is done.

One of the rare instances of multiple linear regression models is polynomial regression. In other words, it is a sort of linear regression when the dependent and independent variables have a curvilinear connection to one another. In the data, a polynomial connection is fitted.

Additionally, by incorporating several polynomial parts, a number of linear regression equations are transformed into polynomial regression equations.

The relationship between both the independent variable x and the dependent variable y is modelled as an nth degree polynomial in polynomial regression. A nonlinear relationship between both the value of x and the associated conditional mean of y, given by the symbol E(y |x), is fit via polynomial regression.

Need of Polynomial Regression

A few criteria that specify the requirement for polynomial regression are listed below.

- If a linear model is used to a linear database, as is the case with simple linear regression, a good result is produced. However, a significant output is calculated if this model is applied to a non-linear dataset with no adjustments. These result in increased mistake rates, a drop in accuracy, and an increase inside the loss function.
- Polynomial regression is required in situations when the data points are organized non-linearly.
- A linear model won't cover any data points if a non-linear model is available and we attempt to cover it. In order to guarantee that all of the data points are covered, a polynomial model is employed. Nevertheless, a curve rather than a straight line will work well for most data points when employing polynomial models.
- A scatter diagram of residuals (Y-axis) here on predictor (X-axis) will show regions of many positive residuals inside the middle if we attempt to fit a linear model to curved data. As a result, it is inappropriate in this circumstance.

Polynomial Regression Applications

Basically, these are employed to define or enumerate non-linear phenomena.

- The rate of tissue growth.
- Progression of pandemic disease.
- Carbon isotope distribution in lake sediments.

Modeling the estimated return of a dependent variable y in relation to the value of an independent variable x is the fundamental aim of regression analysis. We used the equation below in simple

regression

$$\mathbf{y = a + bx + e}$$

Here, the dependent variable is y, along with the independent variables a, b, and e.