# Chapter- Six

# Model Evaluation

Compiled By: Alemisa E

# Outline

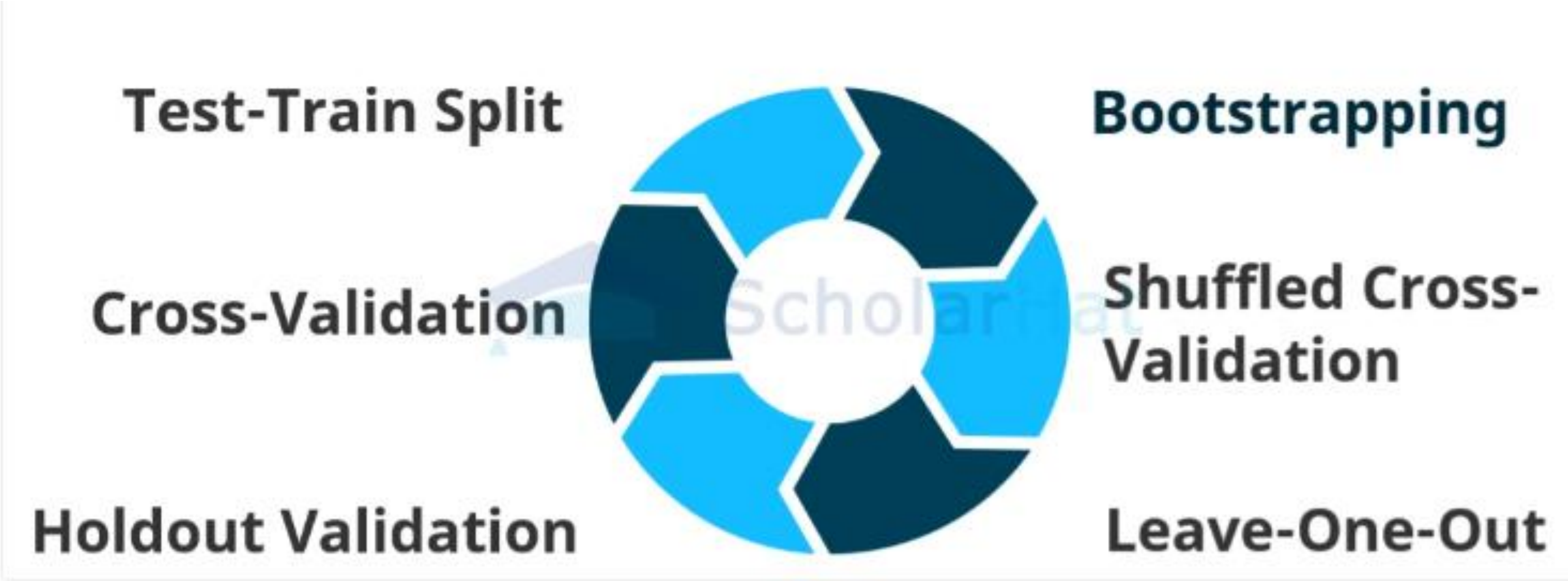
- Model Evaluation
- Data processing
- Data cleaning and transforming
- Feature selection and visualization
- Model selection and tuning
- Methods of dimensional reduction
  - Principal component analysis (PCA)
  - Singular value decomposition (SVD)
  - T-distributed Stochastic Neighbor Embedding (t-SNE)
- Optimize the performance of the model
- Control model complexity
- Over-fitting and Under-fitting
- Cross-Validation and Re-sampling methods
  - K-Fold Cross-Validation
  - 5 ×2 Cross-Validation
  - Bootstrapping
- Gradient descent (batch, stochastic)
  - Bias, variance
- Performance evaluation methods
- Tool kit

# Model Evaluation

- ❑***Model Evaluation*** is a key component of ML, which is the process of determining a trained model's effectiveness and quality using a variety of metrics and approaches.
- ❑Its main goal is to evaluate if the model achieves the ***desired goals*** and how well it generalizes to new data.
- ❑By assessing models, we can choose the best one out of several options and learn more about their limits.
- ❑The usage of data from ***testing*** and ***training*** is often involved in the evaluation process.
- ❑A part of the available data is used during training to instruct the ***model*** in pattern recognition and prediction.
- ❑However, a different set of data that wasn't utilized during training must be used to test the model to assess its performance.

# Model Evaluation Techniques

- ❑*In ML, a variety of strategies are employed for model evaluation.*
- ❑*These methods aid in evaluating the functionality and generalizability of trained models.*
- ❑*Typical model evaluation techniques in machine learning include:*



Test-Train Split

Cross-Validation

Holdout Validation

Bootstrapping

Shuffled Cross-Validation

Leave-One-Out

## i. Test-Train Split:

- The training set and testing set are separated from the available dataset in this method.
- The training set is used to develop the model, and the testing set is used to assess it.
- The performance of the model on the testing set is used to calculate evaluation measures like ***accuracy and mean squared error.***
- This method offers a quick and easy way to assess the performance of the model, but it could be sensitive to how the data are split.

## ii. Cross-Validation:

- Cross-validation addresses the drawback of a single train-test split and is a more reliable technique.
- It entails partitioning the dataset into several folds or subgroups.
- The model is tested on the last fold after being trained on a variety of folds.
- Each fold serves as the assessment set as this process is done numerous times.
- The performance of the model is then estimated more accurately by averaging the evaluation metrics across iterations.

# iii. Holdout Validation

- ❑Similarly, to the train-test split, holdout validation entails reserving a unique validating set-in addition to the training & testing sets.
- ❑ The training set is used to develop the model, the validation set to fine-tune it, and the testing set to assess it.
- ❑This method aids in picking the top-performing model and fine-tuning the model's hyperparameters.

# iv. Bootstrapping

- ❑A ***bootstrapping process*** is a revising approach that entails randomly sampling replacement data from the original dataset in order to produce numerous datasets.
- ❑ A model is trained and evaluated using each bootstrap sample, and the performance of the model is estimated using an aggregate of the evaluation results.
- ❑A measure of the model's reliability & robustness can be obtained by bootstrapping.

# v. Shuffled Cross-Validation

- This method extends k-fold cross-validation by randomly shuffling the data before the cross-validation step.
- It helps to minimize any potential ***bias*** induced by the ordering of the data by ensuring that the data distribution is more evenly represented across the folds.

# vi. Leave-One-Out:

- ***Leave-One-Out (LOO)*** is a specific instance of cross-validation in which the model is trained using the remaining data points while each data point is utilized as the evaluation set.
- The performance of the model is impartially estimated by LOO, but it might be computationally costly for large datasets.

# Performance Evaluation Methods

❑Evaluating the performance of a ML model is one of the important steps in building an effective ML model.

❑***Model evaluation*** is the process that uses some metrics which help us to analyze the performance of the model.

❖ ***To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics.***

❑ These performance metrics help us understand how well our model has performed for the given data.

❑In this way, we can improve the model's performance by tuning the hyper-parameters.

# Performance Evaluation Methods

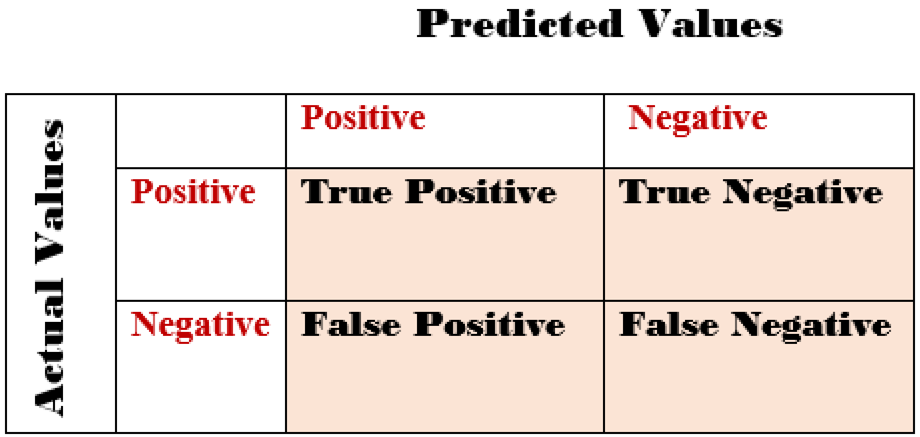
- ❑Each ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.
- ❑In ML, each task or problem is divided into ***classification and Regression.***
- ❑ Not all metrics can be used for all types of problems; hence, it is important to know and understand which metrics should be used.
- ❑Different evaluation metrics are used for both ***Regression and Classification*** tasks.
- ❑In this topic, we will discuss metrics used for classification and regression tasks.

# i. Evaluation Metrics for Classification Task

- In a classification problem, the category or classes of data are identified based on training data.
- The model learns from the given dataset and then classifies the new data into classes or groups based on the training.
- It predicts class labels as the output, such as ***Yes or No, 0 or 1, Spam or Not Spam***, etc.
- Different metrics used for classification performance:
  - ***Confusion Matrix***
  - ***Precision***
  - ***F-Score***
  - ***Accuracy***
  - ***Recall***
  - ***AUC(Area Under the Curve)-ROC***

# Confusion Matrix

- **Confusion Matrix:** 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.
- *It helps us to display the performance of a model or how a model has made its prediction in Machine Learning.*
- A confusion matrix is a table of at least size ***M by M.*** It can be understood well through a ***2×2 matrix.***
- Where the row represents the ***actual value***,
- column represents ***the predicted values.***
- Here is a representation of the confusion matrix



**Predicted Values**

| | | Positive | Negative |
|---|---|---|---|
| **Actual Values** | Positive | **True Positive** | **True Negative** |
| | Negative | **False Positive** | **False Negative** |

- This matrix consists of 4 main elements that show different metrics to count a number of correct and incorrect predictions. Each element has two words either as follows: True or False and Positive or Negative

# i. Confusion Matrix…

- If the predicted and truth labels match, then the prediction is said to be **correct, if the prediction is** mismatched, then the prediction is said to be incorrect.
- Further, positive and negative represent the predicted labels in the matrix.
- Some terminologies in the matrix are as follows:
  - ***True Positives:*** It is also known as TP.
    - ✓ It is the output in which the actual and the predicted values are YES.
  - ***True Negatives:*** It is also known as TN.
    - ✓ It is the output in which the actual and the predicted values are NO.
  - **False Positives:** It is also known as FP.
    - ✓ It is the output in which the actual value is NO but the predicted value is YES.
  - **False Negatives:** It is also known as FN.
    - ✓ It is the output in which the actual value is YES but the predicted value is NO.

# ii. Accuracy

- The **accuracy metric** is one of the simplest Classification metrics to implement
- **Accuracy** is defined as the ratio of the number of correct predictions to the total number of predictions.
- It can be formulated as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions}$$

- The formula is given by

$$\mathbf{Accuracy = (TP+TN)/(TP+TN+FP+FN)}$$

- However, Accuracy has a drawback.
- It cannot perform well on an imbalanced dataset.
- Suppose a model classifies that the majority of the data belongs to the major class label.
- It yields higher accuracy. But in general, the model cannot classify on minor class labels and has poor performance.

# Precision

❑***Precision*** is defined as the *ratio of correctly classified positive samples (True Positive) to the total number of classified positive samples* (either correctly or incorrectly).

❑It basically analyses the positive predictions.

**Precision = True Positive/True Positive + False Positive**

**Precision = TP/(TP+FP)**

❑The drawback of Precision is that it does not consider the ***True Negatives and False Negatives.***

# Recall or Sensitivity

- ❑**Recall** is the ratio of true positives to the summation of true positives and false negatives.
- ❑The ***recall measures the model's ability to detect positive samples***.
- ❑The higher the recall, the more positive samples detected.
- ❑It basically analyses the number of correct positive samples.

**Recall = True Positive/True Positive + False Negative**

**Recall = TP/(TP+FN)**

- ❑The drawback of Recall is that often it leads to a higher false positive rate.

# F1 score

- **F-score or F1 Score** is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class.

- ***F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.***

- The formula for calculating the ***F1 score*** is given below:

$$\text{F1-score} = \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}$$

- It is seen that during the precision-recall trade-off if we increase the precision, recall decreases and vice versa.

- The goal of the ***F1 score*** is to combine ***precision*** and ***recall***.

# AUC-ROC Curve

❑Sometimes we need to visualize the performance of the classification model on charts; then, we can use the ***AUC-ROC curve***.

❑It is one of the popular and important metrics for evaluating the performance of the classification model.

❑**AUC (Area Under Curve)**

- ❖is an evaluation metric that is used to analyze the classification model at different threshold values.

❑ **The Receiver Operating Characteristic(ROC) curve**

- ❖ is a probabilistic curve used to highlight the model's performance.

# AUC: Area Under the ROC curve

❑ The curve is plotted between two parameters, which are:

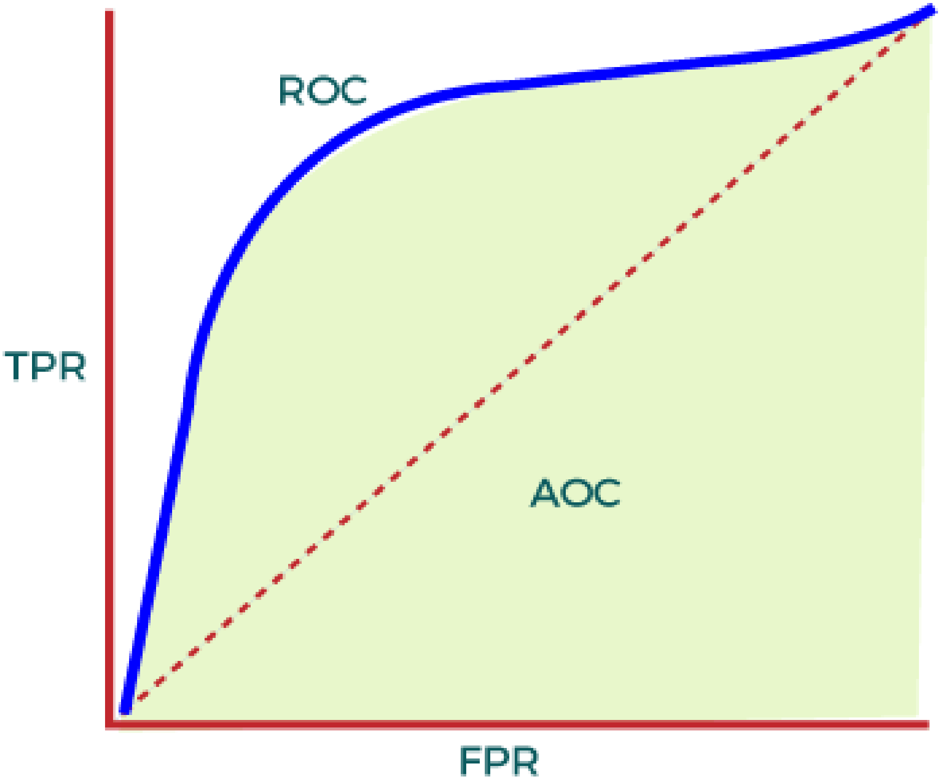
❑ TPR: It stands for True positive rate.

❑ It follows the formula of Recall.

$$TPR = \frac{TP}{TP + FN}$$

❑ **FPR:** It stands for False Positive rate.

$$TPR = \frac{FP}{FP + TN}$$



❑ This curve is useful as it helps us to determine the model’s capacity to distinguish between different classes.

# Evaluation Metrics for Regression Task

❑Regression is used to determine ***continuous values.***

❑ It is mostly used to find a relation between a dependent and an independent variable.

❑For regression analysis, we are predicting a numerical value that may differ from the ***actual output***.

❑So we consider the error calculation as it helps to summarize how close the prediction is to the actual value.

❑The metrics used for regression are different from the classification metrics.

❑ There are many metrics available for evaluating the regression model.

✓ **Mean Absolute Error, Mean Squared Error, R2 Score and Adjusted R2**

# i. Mean Absolute Error (MAE)

- ❑ MAE is one of the simplest metrics, which measures the absolute difference between actual and predicted values,
- ❑where absolute means taking a number as Positive.
- ❑The formula ito calculate MAE:

$$MAE = 1/N \sum |Y - Y'|$$

- ❑Here,
- ❑Y is the Actual outcome,
- ❑Y' is the predicted outcome, and N is the total number of data points.
- ❑MAE is much more robust for the outliers.
- ❑One of the limitations of MAE is that it is not differentiable
- ❑However, to overcome this limitation, another metric can be used, which is Mean Squared Error or MSE.

# ii. Mean Squared Error (MSE)

- MSE measures the average of the Squared difference between predicted values and the actual value given by the model.
- Since in MSE, errors are squared, therefore it only assumes non-negative values, and it is usually positive and non-zero.
- Moreover, due to squared differences, it penalizes small errors also, and hence it leads to over-estimation of how bad the model is.
- MSE is a much-preferred metric compared to other regression metrics as it is differentiable and hence optimized better.
- The formula for calculating MSE is given below:

$$MSE = 1/N \sum (Y - Y')^2$$

- Here,
- Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points.

# iii. R Squared Score

- R squared error is also known as Coefficient of Determination, which is another popular metric used for Regression model evaluation.
- The R-squared metric enables us to compare our model with a constant baseline to determine the performance of the model.
- To select the constant baseline, we need to take the mean of the data and draw the line at the mean.
- The R squared score will always be less than or equal to 1 without concerning if the values are too large or small.

$$R^2 = 1 - \frac{MSE(Model)}{MSE(Baseline)}$$

# iv. Adjusted R Squared

- Adjusted R squared, as the name suggests, is the improved version of R squared error.
- R square has a limitation of improvement of a score on increasing the terms
- To overcome the issue of R square, adjusted R squared is used, which will always show a lower value than R².
- It is because it adjusts the values of increasing predictors and only shows improvement if there is a real improvement.
- We can calculate the adjusted R squared as follows:

$$R_a^2 = 1\text{-}\left[\left(\frac{n-1}{n-k-1}\right) \times \left(1 - R^2\right)\right]$$

Here,

- *n is the number of observations*
- *k denotes the number of independent variables and Ra2 denotes the adjusted R2*

# Data Processing

## ❑ What is Data Processing?

❑ ***Is a process of preparing the raw data and making it suitable for a Machine learning model.***

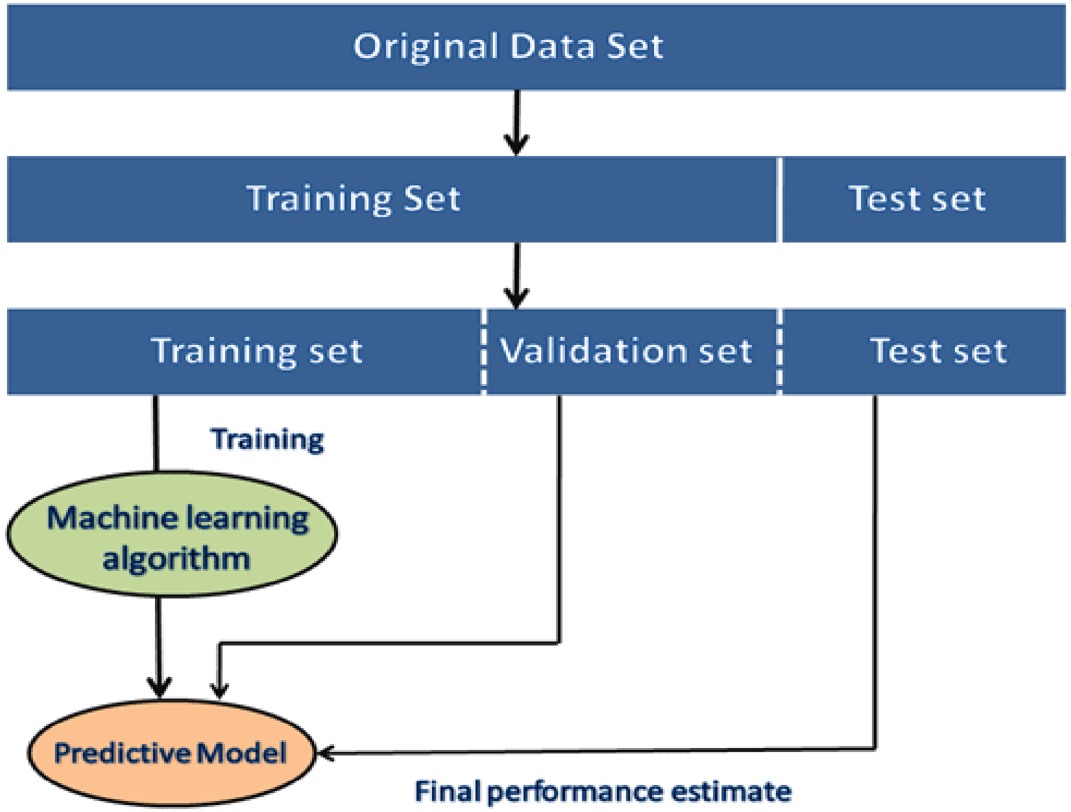
❑ ***Data processing*** is a key component of ML, which is the manipulation or transformation of unprocessed data to make it appropriate for modeling and analysis.

❑When creating a machine learning project, we come across clean and formatted data.

❑While doing any operation with data, it is mandatory to clean it and put it in a formatted way. So for this, we use a ***data preprocessing task.***

# Data Processing

- ❑ In building ML applications,
- ❑ datasets are divided into two parts:
  - **i. Training dataset:**
  - **ii. Test Dataset**



- ❑ The training dataset is utilized to prepare the machine learning model,
- ❑ while the test dataset is utilized to assess the model's exhibition.

# Why do we need to Preprocess Data?

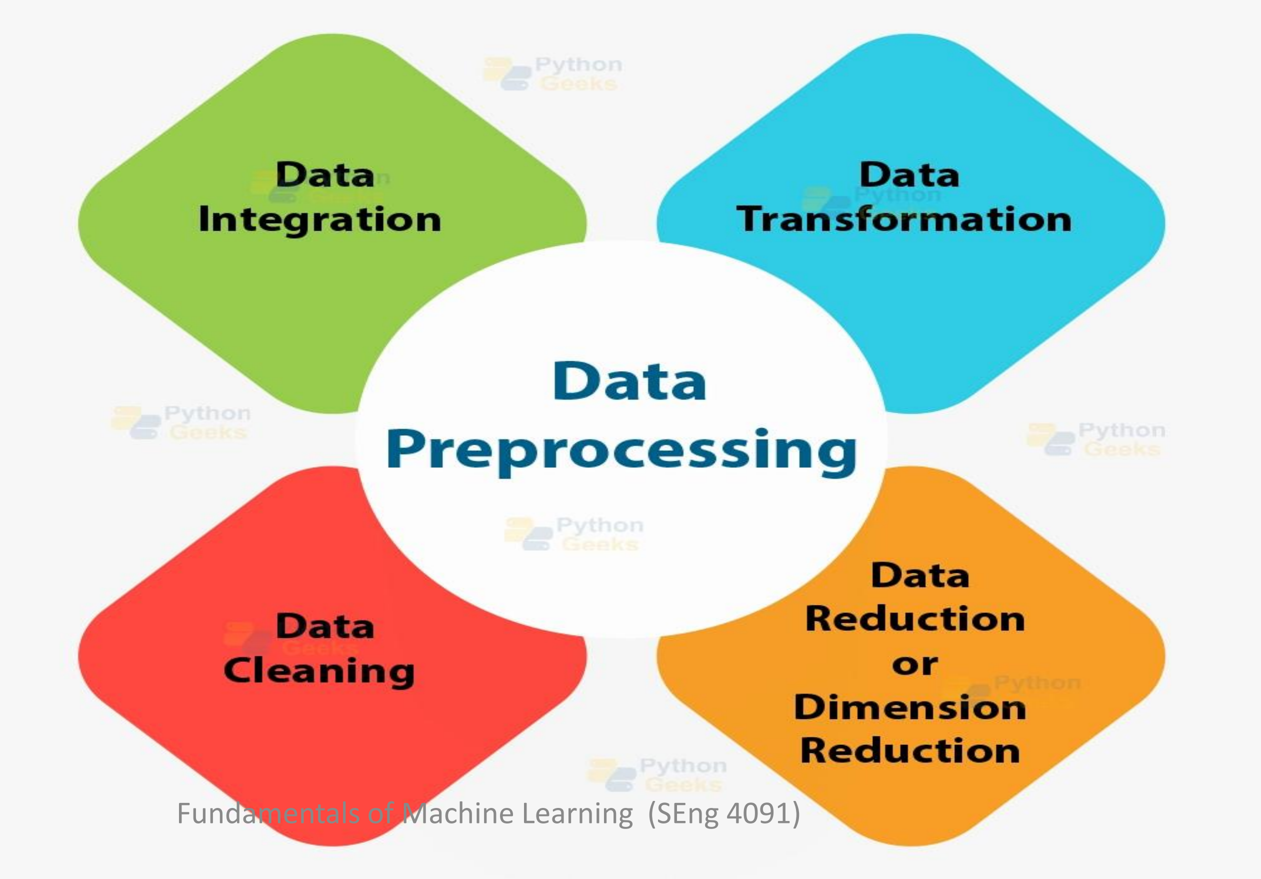
- ❑ Because, real-world data generally contains ***noises, missing values, Inconsistent,*** and maybe in an unusable format which cannot be directly used for ML models.

- ❑ ***Data preprocessing*** is a required task for cleaning the data and making it suitable for a ML model, which also increases the accuracy and efficiency of a ML model.

- ❑ 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.***

# Methods of data processing

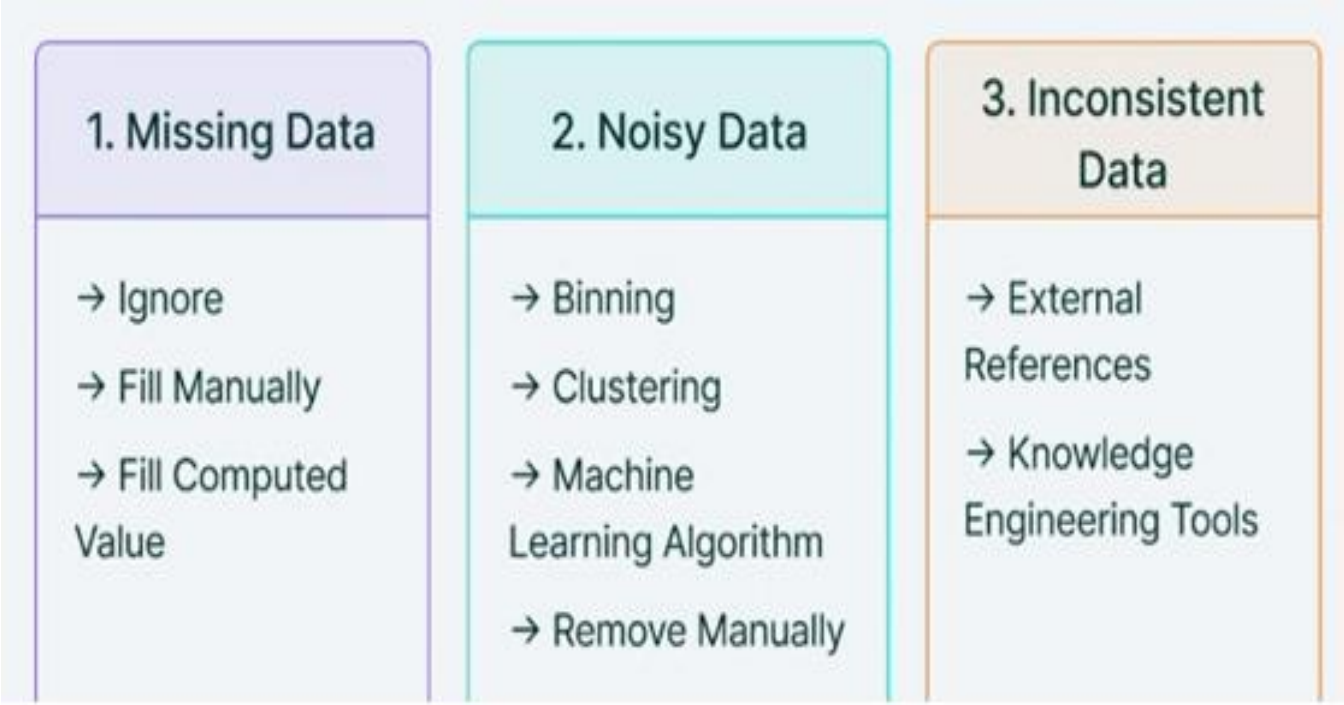
- ❑ To use the data for ***training a model,*** it must first be transformed and prepared, which is known as ***data processing.***
- ❑Here are some typical ways that methods of data processing:



# Data Cleaning

- ❑ Data cleaning is the first phase, which identifies and fixes mistakes, discrepancies, dealing with outliers & and missing values in the dataset to improve data quality.

## How to do that??

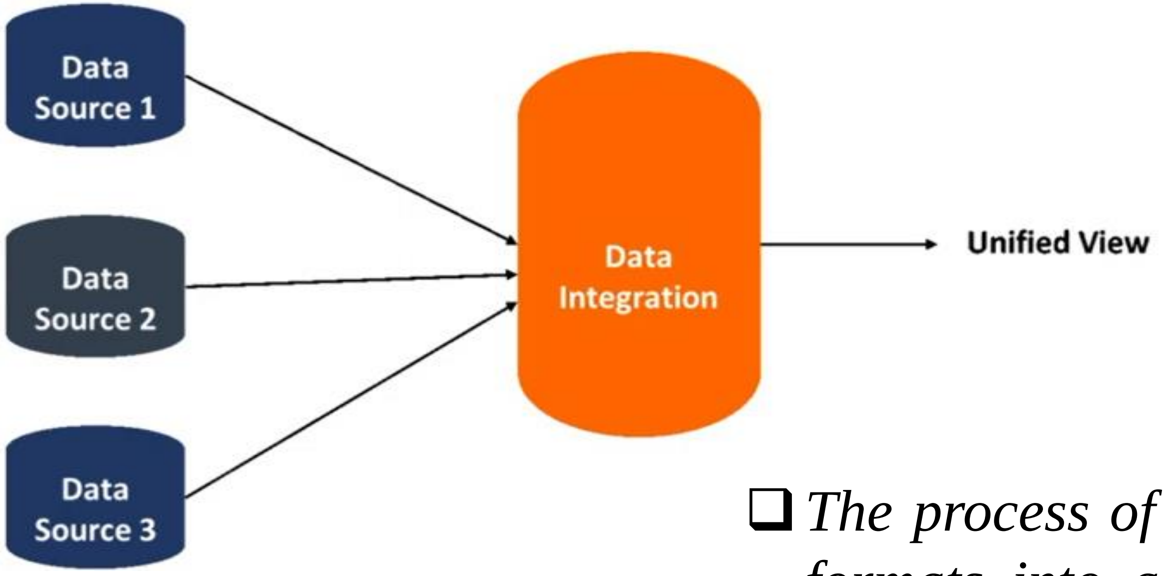


| 1. Missing Data | 2. Noisy Data | 3. Inconsistent Data |
|---|---|---|
| → Ignore<br>→ Fill Manually<br>→ Fill Computed Value | → Binning<br>→ Clustering<br>→ Machine Learning Algorithm<br>→ Remove Manually | → External References<br>→ Knowledge Engineering Tools |

- ❑ ***Note***: *The mean, median, or interpolation approaches can be used to impute missing data.*

# Data Integration

❑When data is gathered from several sources, integration is necessary to combine and consolidate the ***datasets*** into a single, comprehensive dataset.



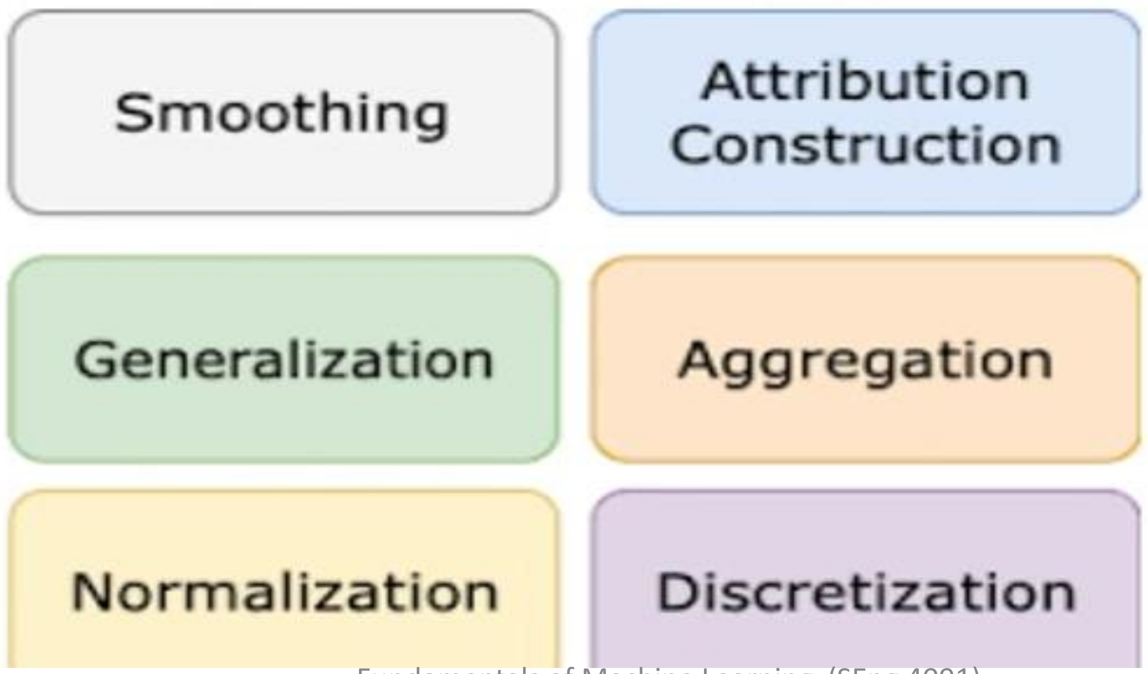
❑ *The process of merging data from several sources or formats into a single cohesive dataset is known as **data integration.***

❑ It entails addressing discrepancies, dealing with duplicate records, and merging data based on shared keys or identifiers.

# Data Transformation

❑Data transformation entails changing or scaling the data to satisfy the presumptions or specifications of the selected machine learning algorithm.

❑Common methods include standardization (scaling to zero mean and unit variance), logarithmic transformation, and normalization (scaling features to a given range).



Smoothing

Attribution Construction

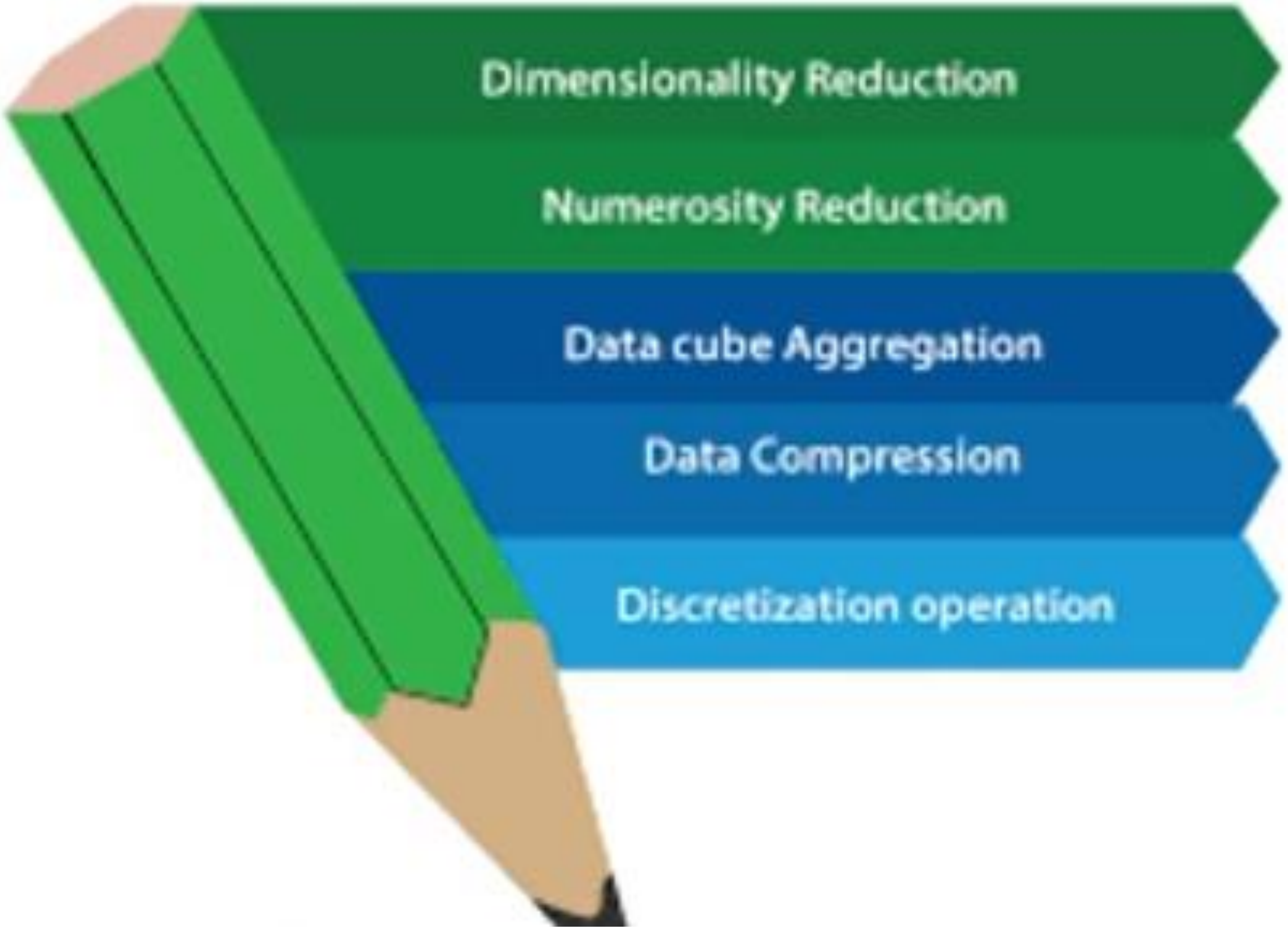
Generalization

Aggregation

Normalization

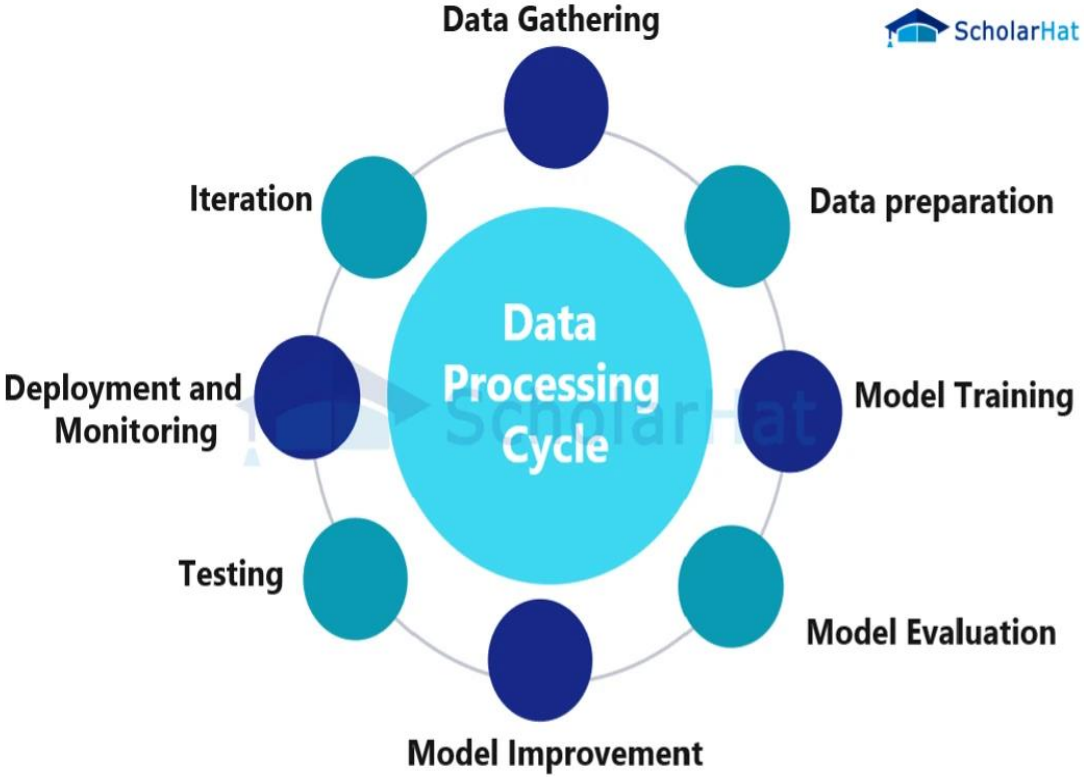
Discretization

# Data Reduction or Dimension Reduction



# Data Processing Cycle

- ❑ The continual process of transforming & analyzing data to train and enhance machine learning models is known as the "***data processing cycle***" in machine learning.
- ❑ It entails a set of actions that are frequently repeated until successful results are obtained.



# Data Processing Cycle

## 1.Data Gathering:

- ❑Gathering pertinent data from numerous sources, including databases, files, and online platforms, is the initial stage.
- ❑ Depending on the particular issue, the data may include *characteristics*, *labels*, *or target variables.*

## 2. Data preparation:

- ❑ In this process, the data is cleaned to get rid of mistakes, discrepancies, and missing values.
- ❑Additionally, it entails formatting the data in a way that is appropriate for analysis, such as processing text or picture data, scaling numerical features, or encoding categorical variables.

# Data Processing Cycle

## 3. Model Training:

- ❑During this phase, a model based on ML is trained using the preprocessed data.
- ❑The model attempts to identify patterns and connections between the features as well as the target variable by learning from the input data.

## 4. Model Evaluation:

- ❑ Following training, the effectiveness of the model is assessed using cross-validation methods or validation data.
- ❑The model's ability to forecast the target variable and complete the desired task is measured using a variety of criteria.

## 5. Model Improvement:

- ❑ Changes are made to the model's operation based on the findings of the evaluation.
- ❑This could entail modifying the model's architecture, adjusting the hyperparameters, or adding new features.

# Data Processing Cycle

## 6. Testing:

- ❑Using test data that hasn't been seen before, the improved model is put to the test.
- ❑ This step offers an unbiased evaluation of the model's capacity for generalization as well as how well it operates in previously untested situations.

## 7. Deployment and Monitoring:

- ❑The model can be deployed for use in real-world applications if it operates well on the test data.
- ❑ It is possible to spot any deterioration or required improvements by tracking the model's performance over time as well as gathering feedback.

## 8. Iteration:

- ❑Since data processing cycles are frequently iterative, the aforementioned stages are repeated as necessary.
- ❑To continuously develop the model and react to shifting circumstances, new data can be gathered, & the preprocessing, training, evaluation, & improvement phases can be repeated.

# How to get datasets for Machine Learning

- ❑ ML depends vigorously on datasets for preparing models and making precise predictions.
- ❑Datasets assume a vital part in the progress of AIML projects and are fundamental for turning into a gifted information researcher.

## ❑ What is a Dataset?

- ❖ **A dataset** is a collection of data in which data is arranged in some order.
- ❖ A ***dataset*** can contain any data from a series of an array to a database table.
- ❖ The table shows an example of the dataset:

| Country | Age | Salary | Purchased |
|---|---|---|---|
| India | 38 | 48000 | No |
| France | 43 | 45000 | Yes |
| Germany | 30 | 54000 | No |
| France | 48 | 65000 | No |
| Germany | 40 |  | Yes |
| India | 35 | 58000 | Yes |

- ❖ *Where each column corresponds to a* ***particular variable,*** *and each row corresponds to the* ***fields of the dataset.***
- ❖ *The most supported file type for a tabular dataset is* ***"Comma Separated File,"*** *or* ***CSV.***
- ❖ *But to store "tree-like data," we can use the JSON file more efficiently.*

# Dataset vs Database

## ❑Dataset

- ❖A *dataset* is a collection of data that is usually stored in a **flat-file format, such as a CSV or Excel file.**
- ❖A dataset may contain data from a variety of *sources, including databases, websites, and manual entry.*

## ❑*Database*

- ❖Database is a collection of data that is organized and stored in a way that allows easy access and retrieval.
- ❖There are many different types of databases, including relational databases, object-oriented databases, and NoSQL databases.

# Types of data in datasets

- ❑**Numerical data:**
  - ❖Such as house price, temperature, etc.
- ❑**Categorical data:**
  - ❑Such as Yes/No, True/False, Blue/green, etc.
- ❑***Ordinal data:***
  - ❖These data are similar to categorical data but can be measured on the basis of comparison.
- ❑*Note: A real-world dataset is of huge size, which is difficult to manage and process at the initial level. Therefore, to practice machine learning algorithms, we can use any dummy dataset.*

# Types of datasets

- ❑ML incorporates different domains, each requiring explicit sorts of datasets.
- ❑ A few normal sorts of datasets utilized in machine learning include:

## ❑Image Datasets:

- ❖*Image datasets* contain an assortment of images and are normally utilized in computer vision tasks such as ***image classification, object detection, and image segmentation.***
- ❖**Examples:** ImageNet, CIFAR-10, MNIST

## ❑Text Datasets:

- ❖Text datasets comprise textual information, like articles, books, or virtual entertainment posts.
- ❖These datasets are utilized in NLP techniques like ***sentiment analysis, text classification, and machine translation.***
- ❖**Examples:** Gutenberg Task dataset, IMDb film reviews dataset

# Types of datasets

## ❑Time Series Datasets:

❖This type of dataset includes information focuses gathered after some time.

❖They are generally utilized in determining, abnormality location, and pattern examination.

❑ **Examples:** ***Securities exchange information, Climate information, Sensor readings.***

## ❑Tabular Datasets:

❑Tabular datasets are organized information coordinated in tables or calculation sheets.

❑They contain lines addressing examples or tests and segments addressing highlights or qualities.

❑Tabular datasets are utilized for undertakings like relapse and arrangement.

❑ The dataset given before in the article is an illustration of a tabular dataset.

# Popular sources for ML datasets

## 1. Kaggle Datasets

- ❑***Kaggle*** is one of the best sources for providing datasets for Data Scientists and Machine Learners.
- ❑*It also provides the opportunity to work with other ML engineers and solve difficult Data Science related tasks.*
- ❑Kaggle provides a high-quality dataset in different formats that we can easily find and download.  The link is https://www.kaggle.com/datasets.

## 2. UCI Machine Learning Repository

- ❑UCI stands for the University of California Irvine ML repository, and it is a very useful resource for getting open-source and free datasets for machine learning.
- ❑is an important asset that has been broadly utilized by scientists and specialists beginning around 1987.
- ❑It contains a huge collection of datasets sorted by ML tasks such as ***regression, classification, and clustering.***
- ❑The link for the UCI ML repository is https://archive.ics.uci.edu/ml/index.php.

# Popular sources for ML datasets

## 3. Datasets via AWS (Amazon Web Services)

- We can search, download, access, and share the datasets that are publicly available via AWS resources.
- These datasets can be accessed through AWS resources but provided and maintained by different *government organizations, researchers, businesses, or individuals.*
- Anyone can analyze and build various services using shared data via AWS resources.
- Anyone can add any dataset or example to the **Registry of Open Data on AWS.**
- The link for the resource is https://registry.opendata.aws/.

## 4. Google's Dataset Search Engine

- Google's Dataset Web index helps scientists find and access important datasets from different sources across the web.
- The link is https://toolbox.google.com/datasetsearch.

# Popular sources for ML datasets

## 5. Microsoft Datasets

❑Microsoft has launched the **"Microsoft Research Open Data"** repository with the collection of free datasets in various areas such as **NLP, computer vision, and domain-specific sciences.** It gives admittance to assorted and arranged datasets that can be significant for machine learning projects.

❑The link is https://msropendata.com/.

## 6. Awesome Public Dataset Collection

❑Awesome public dataset collection provides high-quality datasets that are arranged in a well-organized manner within a list according to topics such as ***Agriculture, Biology, Climate, Complex networks,*** etc.

❑Most of the datasets are available free, but some may not, so it is better to check the license before downloading the dataset.

❑The link is https://github.com/awesomedata/awesome-public-datasets.

# Popular sources for ML datasets

## 7. Government Datasets

- ❑There are different sources to get government-related data.
- ❑Various countries publish government data for public use collected by them from different departments.
- ❑The goal of providing these datasets is to increase the transparency of government work among the people and to use the data in an innovative approach.
- ❑Below are some links to government datasets:
  - ❖ Indian Government dataset
  - ❖ US Government Dataset
  - ❖ Northern Ireland Public Sector Datasets
  - ❖ European Union Open Data Portal

# Popular sources for ML datasets

## 8. Computer Vision Datasets

- ❑Visual data provides multiple numbers of great datasets that are specific to computer visions such as ***Image Classification, Video classification, Image Segmentation,*** etc.
- ❑Therefore, if you want to build a project on deep learning or image processing, then you can refer to this source. The link is https://www.visualdata.io/.

## 9. Scikit-learn dataset

- ❑Scikit-learn, a well-known machine learning library in Python, gives a few underlying datasets to practice and trial and error.
- ❑These datasets are open through the sci-kit-learn Programming interface and can be utilized for learning different machine-learning calculations.
- ❑ Scikit-learn offers both toy datasets, which are little and improved, and genuine world datasets with greater intricacy.
- ❑Instances of sci-kit-learn datasets incorporate the Iris dataset, the Boston Lodging dataset, and the Wine dataset.
- ❑The link is https://scikit-learn.org/stable/datasets/index.html.

# Feature selection and visualization

❑ ***Feature Selection Techniques in Machine Learning***

- ❑ While developing the ML model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant.
- ❑ If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model.
- ❑ ***Feature selection*** is one of the important concepts of ML, which highly impacts the performance of the model.
- ❑ As ML works on the concept of "**Garbage In Garbage Out",** so we always need to input the most appropriate and relevant dataset to the model in order to get a better result.

# What is Feature Selection?

❑**Feature Selection** is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

❑***Need for Feature Selection***

- ❖It helps in avoiding the curse of dimensionality.
- ❖It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- ❖It reduces the training time.
- ❖It reduces overfitting hence enhance the generalization.

# Data Visualization

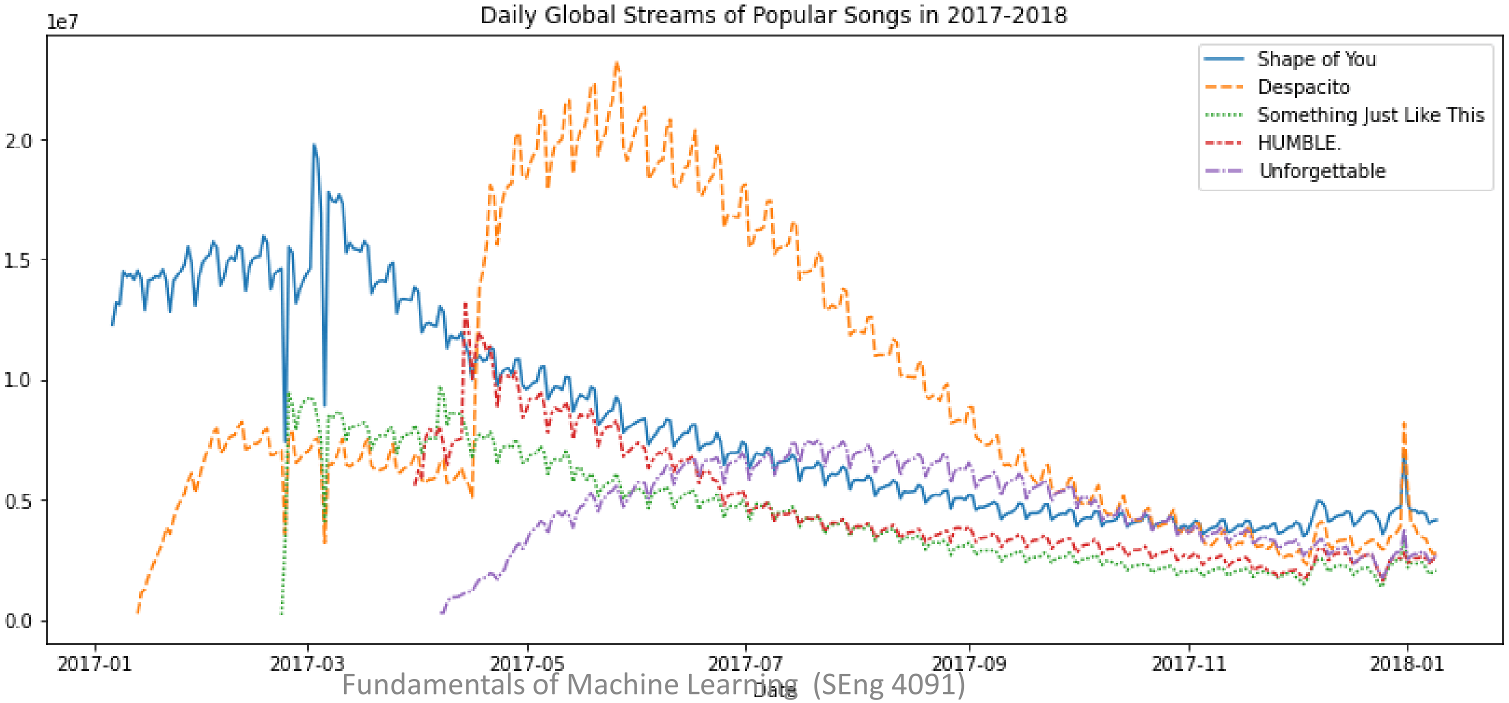
- Data visualization is a crucial aspect of ML that enables analysts to understand and make sense of data patterns, relationships, and trends.
- It also helps to analyze complex data sets by presenting them in an easily understandable format.
- Data visualization is an essential step in data preparation and analysis as it helps to identify outliers, trends, and patterns in the data that may be missed by other forms of analysis.
- With the increasing availability of big data, it has become more important than ever to use data visualization techniques to explore and understand the data.
- Machine learning algorithms work best when they have high-quality and clean data, and data visualization can help to identify and remove any inconsistencies or anomalies in the data.

# Types of Data Visualization Approaches

## i. Line Charts:

❑In a line chart, each data point is represented by a point on the graph, and these points are connected by a line.

❑Time-series data is frequently displayed using line charts. For example. Figure below



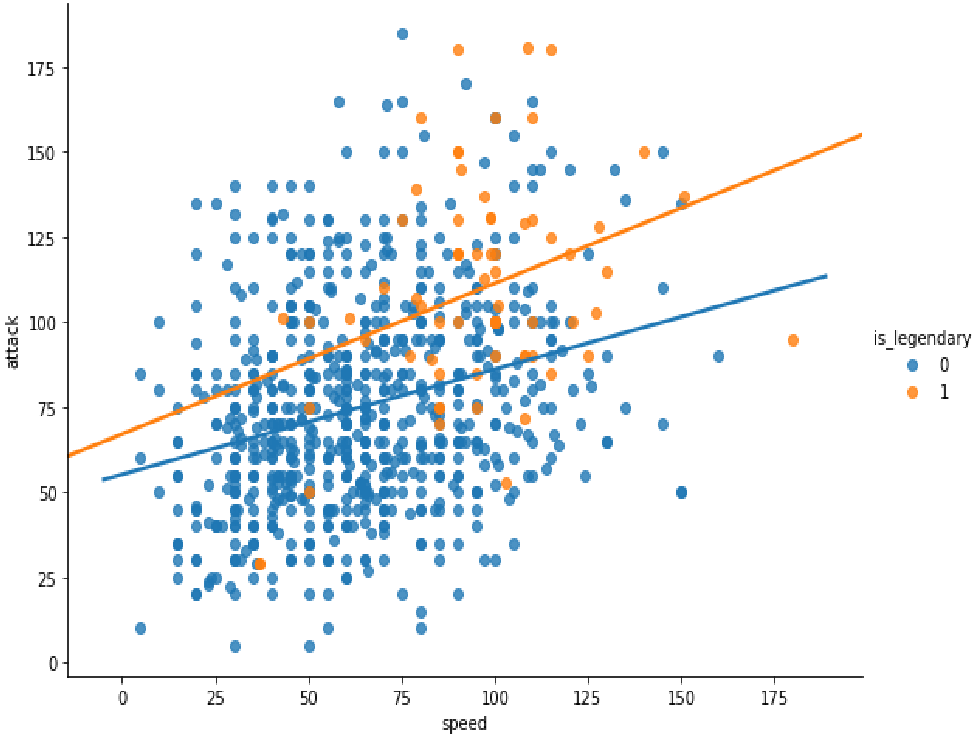
# Types of Data Visualization Approaches

## ii. Scatter Plots:

❑A quick and efficient method of displaying the relationship between two variables is to use scatter plots.

❑With one variable plotted on the x-axis and the other variable drawn on the y-axis,

❑ each data point in a scatter plot is represented by a point on the graph.

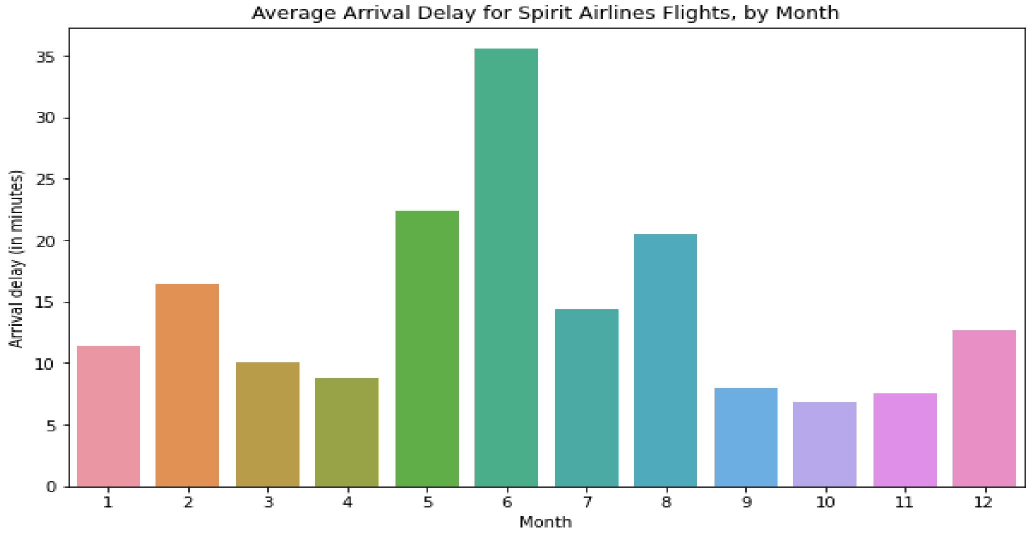


❑We may use scatter plots to visualize data to find patterns, clusters, and outliers.

# Types of Data Visualization Approaches

## iii. Bar Charts:

- ❑ Bar charts are a common way of displaying categorical data.
- ❑In a bar chart, each category is represented by a bar, with the height of the bar indicating the frequency or proportion of that category in the data.
- ❑Bar graphs are useful for comparing several categories and seeing patterns over time.



**Note: Other Data visualization approaches: Heat Maps, Tree Maps, Box Plots**

# Uses of Data Visualization in ML

- **Identify trends and patterns in data:**
  - It may be challenging to spot trends and patterns in data using conventional approaches, but data visualization tools may be utilized to do so.
- **Communicate insights to stakeholders:**
  - Data visualization can be used to communicate insights to stakeholders in a format that is easily understandable and can help to support decision-making processes.
- **Monitor machine learning models:**
  - Data visualization can be used to monitor machine learning models in real time and to identify any issues or anomalies in the data.
- **Improve data quality:**
  - Data visualization can be used to identify outliers and inconsistencies in the data and to improve data quality by removing them.

# Over-fitting and Under-fitting

- ❑Overfitting and underfitting occur while training our machine learning or deep learning models, both are the main problems that occur in ML or DL.
- ❑The main goal of each ML model is **to generalize well**.
- ❑ G**eneralization** defines the ability of an ***ML model*** to provide a suitable output by adapting the given set of unknown inputs.
- ❑It means after providing training on the dataset, it can produce reliable and accurate output.
- ❑Hence, underfitting and overfitting are the two terms that need to be checked for the performance of the model and whether the model is generalizing well or not.

# Over-fitting and Under-fitting

- *Some basic terms that will help to understand this topic well:*
  - **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.
  - **Noise:-** is unnecessary and irrelevant data that reduces the performance of the model.
  - **Bias:-** is a prediction error that is introduced in the model due to oversimplifying the ML algorithms.
  - Or it is the difference between the predicted values and the actual values.
  - **Variance:** If the ML model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

# Over-fitting

❑Overfitting occurs when our ML model tries to cover all the data points or more than the required data points present in the given dataset.

❑i.e. our training has focused on the particular training set so much that it has missed the point entirely.

❑Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.

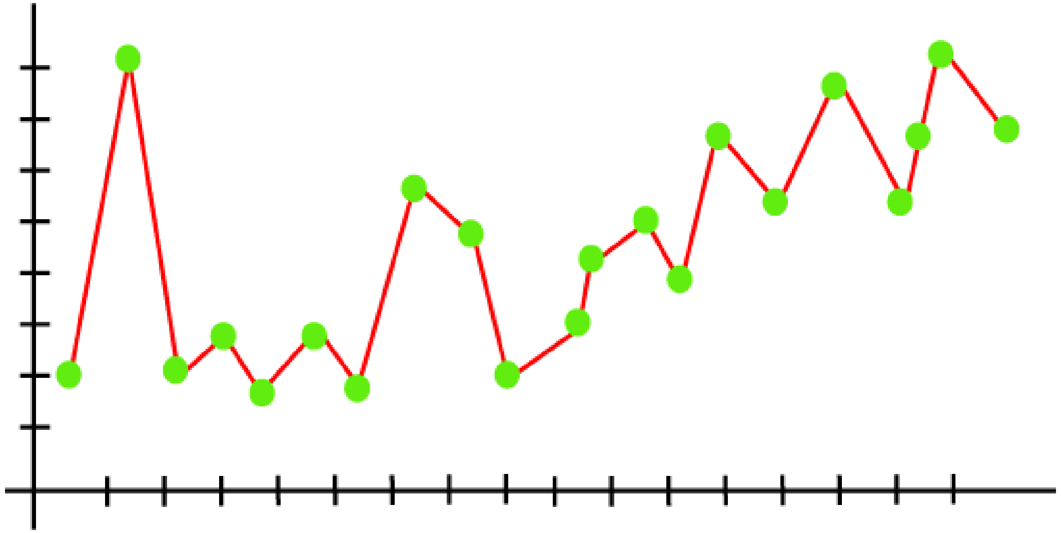
❑The overfitted model has **low bias** and **high variance.**

❑The chances of occurrence of overfitting increase as much as we provide training to our model.

❑ It means the more we train our model, the more chances of occurring the overfitted model.

# Over-fitting : Example

❑Consider the below graph of the linear regression output:



❑ As we can see from the above graph, the model tries to cover all the data points present in the scatter plot.

❑ It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

# How to Avoid the overfitting in Model

- ❑Both overfitting and underfitting cause the degraded performance of the ML model.
- ❑But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.
  - ❖***Cross-Validation***
  - ❖***Training with more data***
  - ❖***Removing features***
  - ❖***Early stopping the training***
  - ❖***Regularization***
  - ❖***Ensembling***

# Under-fitting

- ❑Underfitting occurs when our ML model is not able to capture the underlying trend of the data.

- ❑Underfitting, on the other hand, means the model has not captured the underlying logic of the data.

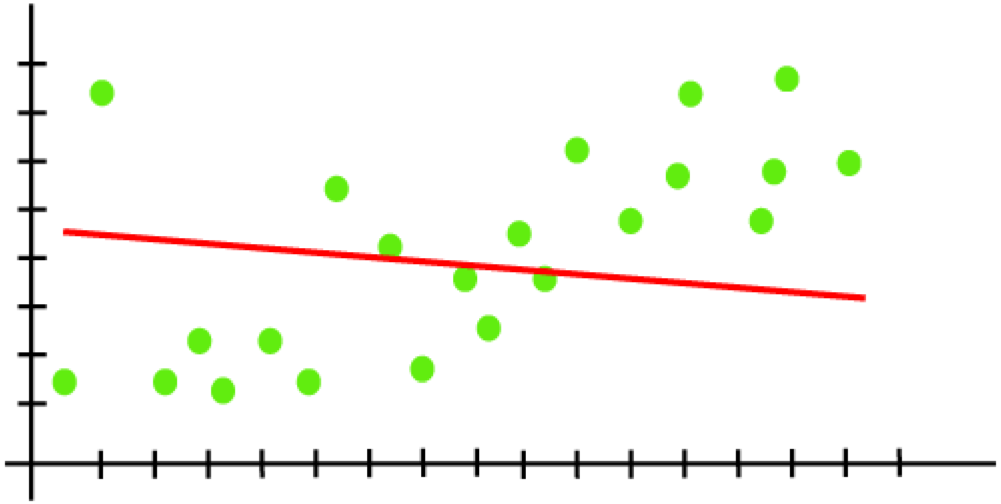
- ❑It doesn’t know what to do with the task we’ve given it and, therefore, provides an answer that is far from correct.

- ❑**In the case of underfitting,** the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

- ❑An underfitted model has high bias and low variance.

# Under-fitting : Example

- ❑ As we can see from the diagram,
- ❑ the model is unable to capture the data points present in the plot.



- ❑ How to avoid underfitting:
  - ❖ By increasing the training time of the model.
  - ❖ By increasing the number of features.

# Goodness of Fit

- ❑The "Goodness of fit" term is taken from the statistics, and the goal of the ML models to achieve the goodness of fit.
- ❑In statistics modeling, *it defines how closely the result or predicted values match the true values of the dataset.*
- ❑The model with a good fit is between the underfitted and overfitted model, and ideally, it makes predictions with 0 errors, but in practice, it is difficult to achieve it.
- ❑As when we train our model for a time, the errors in the training data go down, and the same happens with test data.
- ❑But if we train the model for a long duration, then the performance of the model may decrease due to the overfitting, as the model also learn the noise present in the dataset.
- ❑The errors in the test dataset start increasing, *so the point, just before the raising of errors, is the good point, and we can stop here for achieving a good model.*

# Bias and Variance

- ❑If the ML model is not accurate, it can make prediction errors, and these prediction errors are usually known as **Bias and Variance.**
- ❑ In ML , these errors will always be present as there is a difference between the ***model predictions and actual predictions.***
- ❑ The main aim of ML/data science analysts is to reduce these errors to get more accurate results.
- ❑Let's first understand what Errors in Machine learning.

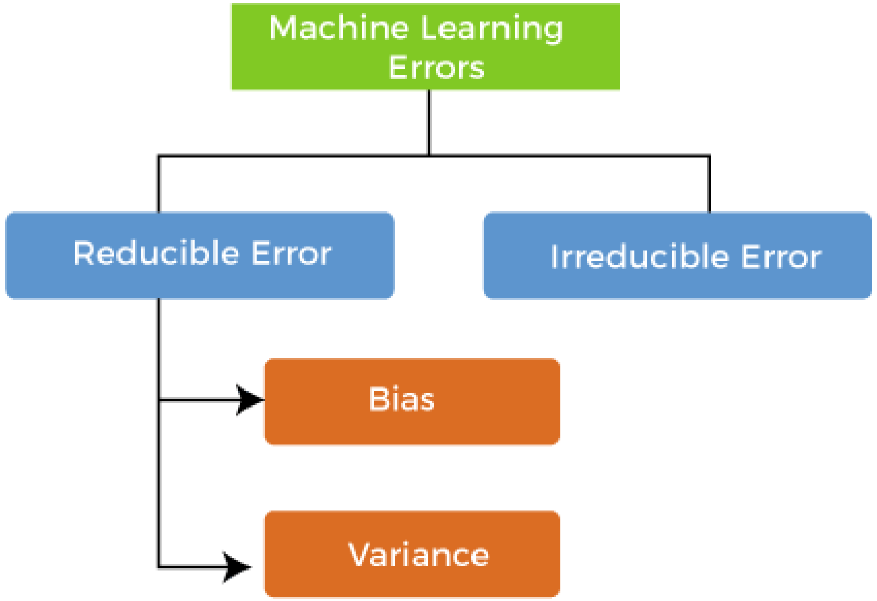
## ❑Errors In ML

- ❖In ML, an error is a measure of how accurately an algorithm can make predictions for the previously unknown dataset.
- ❖On the basis of these errors, the ML model is selected that can perform best on the particular dataset.
- ❖ There are mainly two types of errors in machine learning, which are:

# Types of errors in ML

**i. Reducible errors:** These errors can be reduced to improve the model accuracy.

❑Such errors can further be classified into bias and Variance.



**ii. Irreducible errors:** These errors will always be present in the model regardless of which algorithm has been used.

❑The cause of these errors is unknown variables whose value can't be reduced.

# What is Bias?

- ❑ In general, the ML model analyses the data, finds patterns in it, and makes predictions.
- ❑ While training, the model learns these patterns in the dataset and applies them to test data for prediction.
  - ❖ ***While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias.***
- ❑ **A model has either:**
  - ❖ **Low Bias:** A low-bias model will make fewer assumptions about the form of the target function.
  - ❖ **High Bias:** A model with a high bias makes more assumptions, and the model becomes unable to capture the important features of our dataset.
  - ❖ **A high bias model also cannot perform well on new data.**

# What is a Variance Error?

- ***variance tells that how much a random variable is different from its expected value.***
- Ideally, a model should not vary too much from one training dataset to another, which means the algorithm should be good in understanding the hidden mapping between inputs and output variables.
- Variance errors are either of **low variance or high variance.**
- **Low variance** means there is a small variation in the prediction of the target function with changes in the training data set.
- **High variance** shows a large variation in the prediction of the target function with changes in the training dataset.
- A model that shows high variance learns a lot and perform well with the training dataset, and does not generalize well with the unseen dataset.
- As a result, such a model gives good results with the training dataset but shows high error rates on the test dataset.

# What is a Variance Error?

- ❑ A model with high variance has the following problems:
  - ❖A high variance model leads to overfitting.
  - ❖Increase model complexities.
  - ❖Usually, nonlinear algorithms have a lot of flexibility to fit the model, and have high variance.
- ❑Ways to Reduce High Variance:
  - ❖Reduce the input features or number of parameters as a model is overfitted.
  - ❖Do not use a much complex model.
  - ❖Increase the training data.
  - ❖Increase the Regularization term.