Types of Machine learning algorithms

In machine learning, parametric and non-parametric are two different approaches for modeling the underlyingrelationship between the input and output variables.

Parametric learning involves making assumptions about the functional form of the relationship between the input and output variables, and then estimating the parameters of that function using the training data. In other words, the model structure is fixed and the parameters are learned from the data. Examples of parametric models include linear regression, logistic regression, and neural networks. Once the parameters are learned, the model can be used to make predictions on new data.

Non-parametric learning, on the other hand, makes no assumptions about the functional form of the relationship between the input and output variables. Instead, it uses the training data to build a flexible model thatcan fit any underlying relationship between the variables. Non-parametric models include decision trees, k- nearest neighbors, and support vector machines. Non-parametric models generally require more data than parametric models, but they can be more accurate when the underlying relationship is complex and difficult to model with a fixed functional form.

Here are some key differences between parametric and non-parametric learning:

- Model flexibility: Parametric models have a fixed model structure, which limits their ability to capture complex relationships between variables. Non-parametric models are more flexible and can fit awider range of relationships between variables.
- Parameter estimation: In parametric models, the parameters are estimated using the training data. Non-parametric models do not have fixed parameters, but instead build a flexible model based on thetraining data.
- Data requirements: Parametric models generally require less data than non-parametric models, as theymake assumptions about the functional form of the relationship between variables.
- Non-parametric models require more data, as they need to build a flexible model that can fit any underlying relationshipbetween variables.
- Interpretability: Parametric models are generally more interpretable than non-parametric models, asthey have a fixed model structure that can be easily understood. Non-parametric models can be more difficult to interpret, as they do not have a fixed model structure.

In summary, parametric learning makes assumptions about the functional form of the relationship between the input and output variables, while non-parametric learning makes no such assumptions and builds a flexible modelbased on the training data.
Regenerate response

Types of Learning

In general, machine learning algorithms can be classified into three types.

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning

A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalises to respond correctly to all possible inputs. This is also called learning from exemplars. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

In supervised learning, each example in the training set is a pair consisting of an input object (typically a vector) and an output value. A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples. In the optimal case, the function will correctly determine the class labels for unseen instances. Both classification and regression problems are supervised learning problems. A wide range of supervised learning algorithms are available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems.

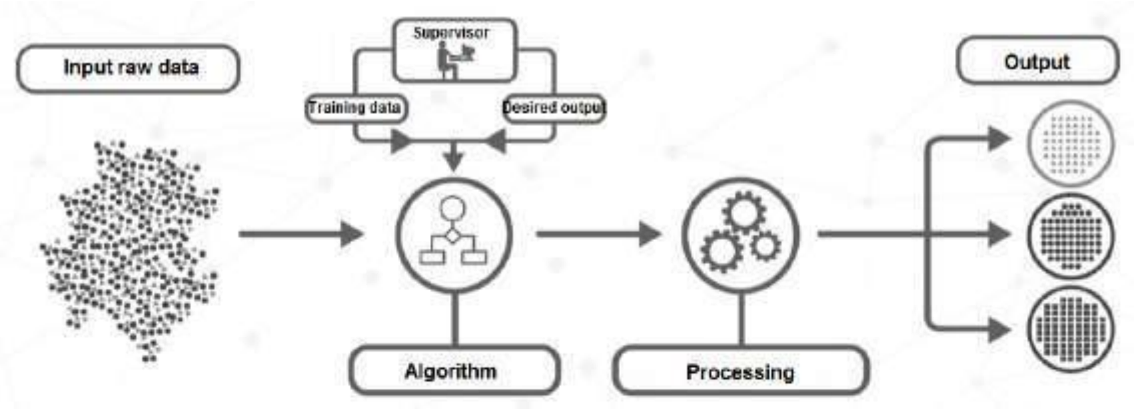


Figure 1.4: Supervised learning

Remarks

A “supervised learning” is so called because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers (that is, the correct outputs), the algorithm iteratively makes predictions on the training data and is

corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Example

Consider the following data regarding patients entering a clinic. The data consists of the genderand age of the patients and each patient is labeled as “healthy” or “sick”.

gender	age	label
M	48	sick
M	67	sick
F	53	healthy
M	49	healthy
F	34	sick
M	21	healthy

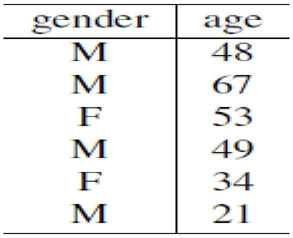
Unsupervised learning

Correct responses are not provided, but instead the algorithm tries to identify similaritiesbetween the inputs so that inputs that have something in common are categorised together. The statistical approach to unsupervised learning is known as density estimation.

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. In unsupervised learning algorithms, a classification or categorization is not included in the observations. There are no output values and so there is no estimation of functions. Since the examples given to the learner are unlabeled, the accuracy of the structure that is output by the algorithm cannot be evaluated. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns Or grouping in data.

Example

Consider the following data regarding patients entering a clinic. The data consists of the genderand ageof the patients.



gender	age
M	48
M	67
F	53
M	49
F	34
M	21

Based on this data, can we infer anything regarding the patients entering the clinic?

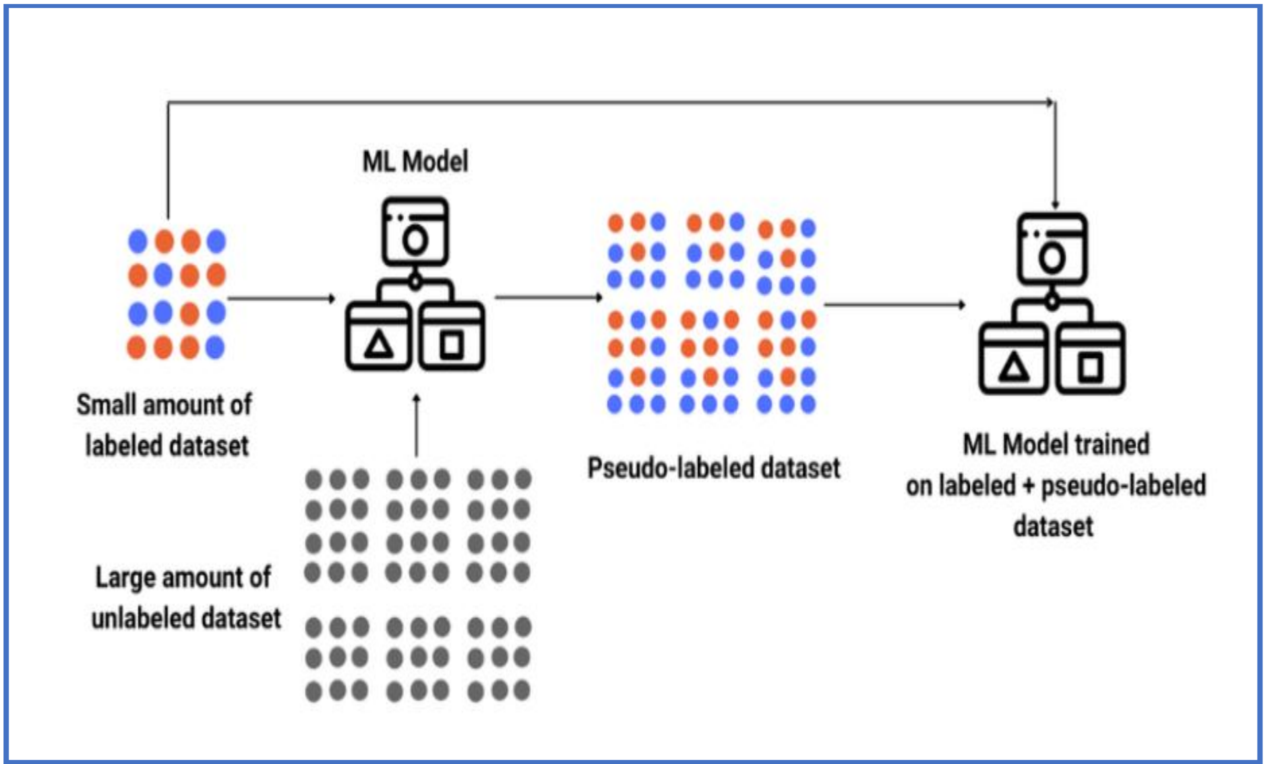
Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predictingcorrect output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with theoutput.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predictthe output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns anduseful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train themodel.

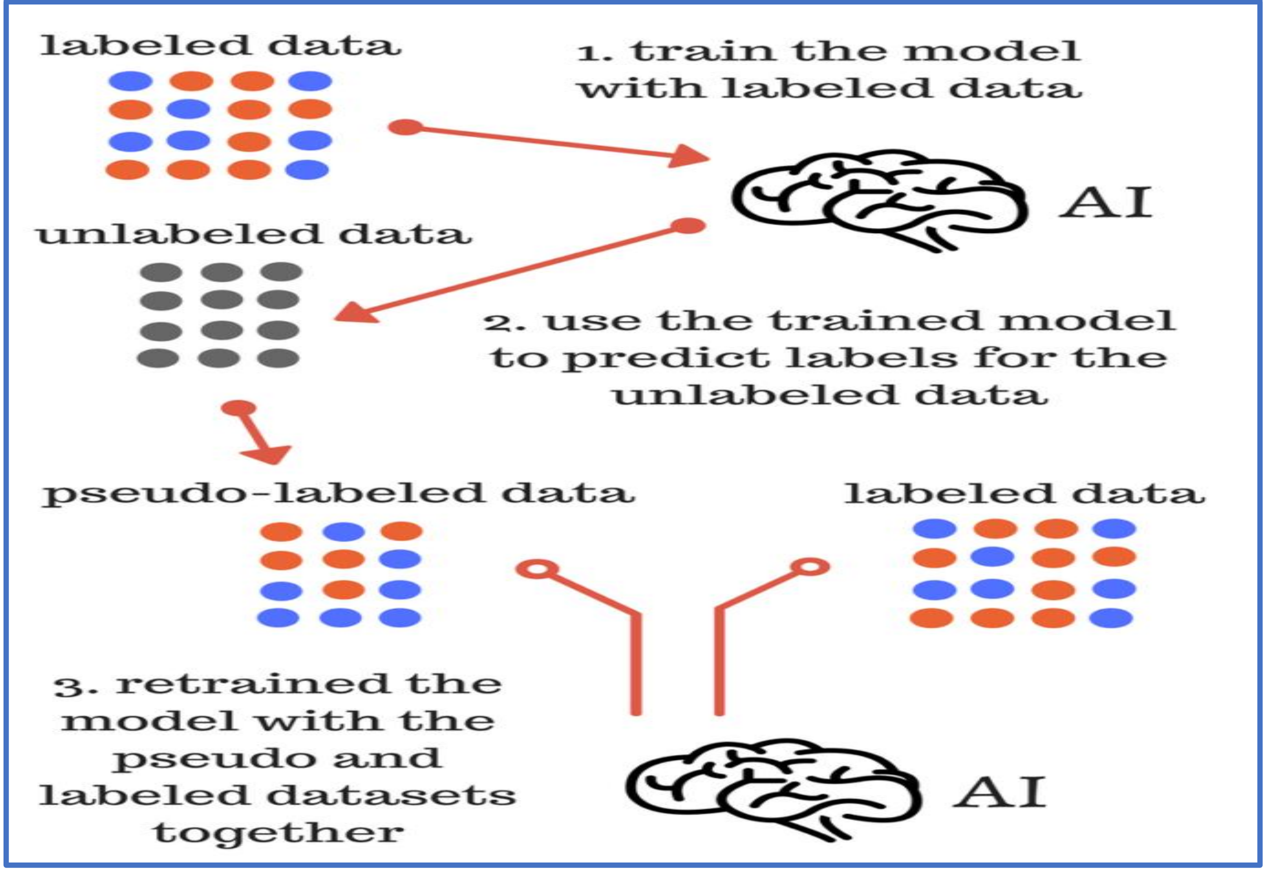
Supervised learning can be categorized in **Classification** and **Regression** problems.	Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
Supervised learning can be used for those cases where we know the inputas well as corresponding outputs.	Unsupervised learning can be used for those cases where we haveonly input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
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 correctoutput.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routinethings by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decisiontree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriorialgorithm.

Semi-supervised Learning

- Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning.
- It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model.
- The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning.
- However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

Semi-supervised learning is particularly useful when there is a large amount of unlabeled data available, but it’s too expensive or difficult to label all of it


ML Model
Small amount of labeled dataset
Large amount of unlabeled dataset
Pseudo-labeled dataset
ML Model trained on labeled + pseudo-labeled dataset


labeled data
1. train the model with labeled data
AI
unlabeled data
2. use the trained model to predict labels for the unlabeled data
pseudo-labeled data
labeled data
3. retrained the model with the pseudo and labeled datasets together
AI

Some examples of semi-supervised learning applications include:

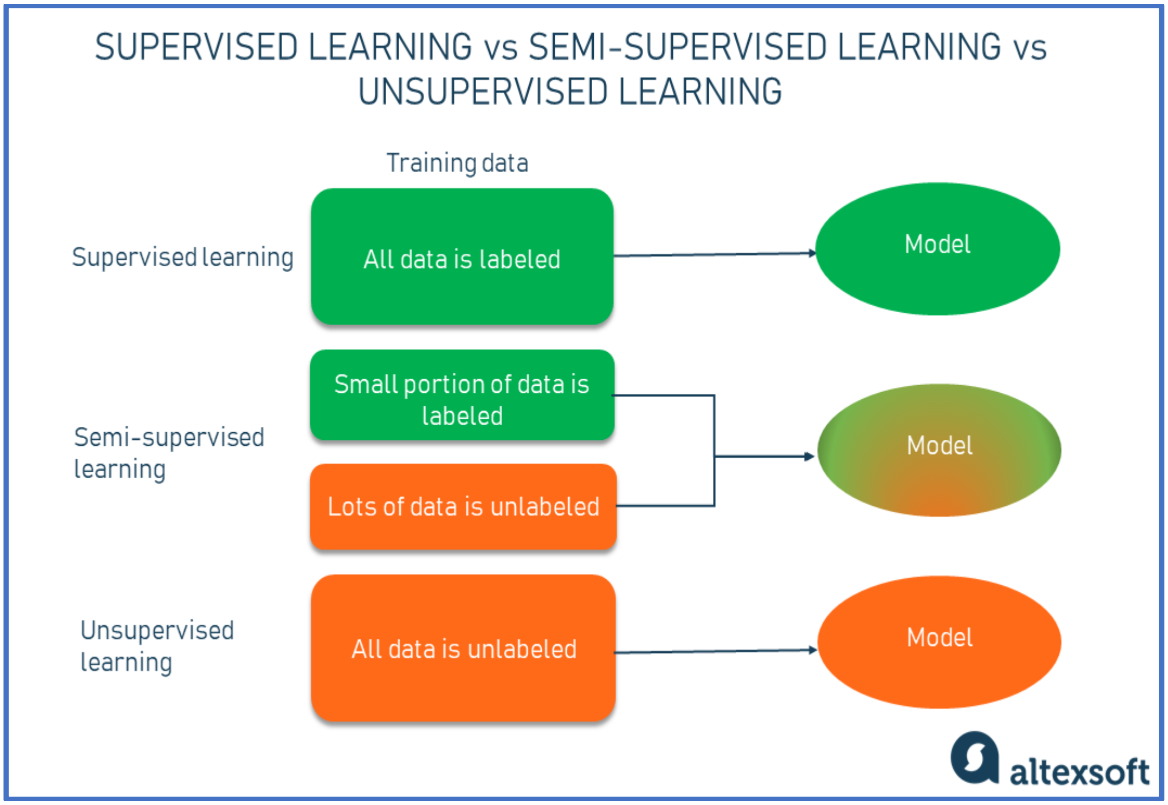
- **Text classification**

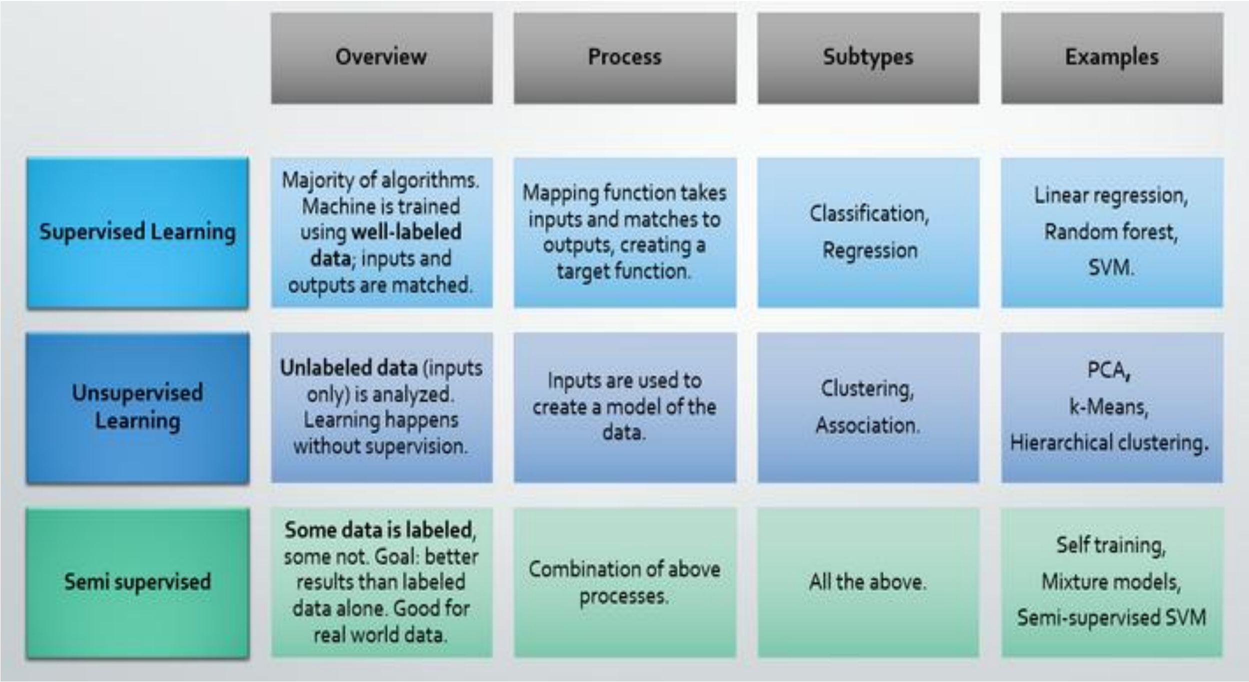
  In text classification, the goal is to classify a given text into one or more predefined categories. Semi-supervised learning can be used to train a text classification model using a small amount of labeled data and a large amount of unlabeled text data.
- **Image classification**

  In image classification, the goal is to classify a given image into one or more predefined categories. Semi-supervised learning can be used to train an image classification model using a small amount of labeled data and a large amount of unlabeled image data.
- **Anomaly detection**

  In anomaly detection, the goal is to detect patterns or observations that are unusual or different from the norm.

Summary of Three learning algorithms





	Overview	Process	Subtypes	Examples
Supervised Learning	Majority of algorithms. Machine is trained using **well-labeled data**; inputs and outputs are matched.	Mapping function takes inputs and matches to outputs, creating a target function.	Classification, Regression	Linear regression, Random forest, SVM.
Unsupervised Learning	**Unlabeled data** (inputs only) is analyzed. Learning happens without supervision.	Inputs are used to create a model of the data.	Clustering, Association.	PCA, k-Means, Hierarchical clustering.
Semi supervised	**Some data is labeled**, some not. Goal: better results than labeled data alone. Good for real world data.	Combination of above processes.	All the above.	Self training, Mixture models, Semi-supervised SVM

Reinforcement learning

This is somewhere between supervised and unsupervised learning. The algorithm gets told when the answer is wrong, but does not get told how to correct it. It has to explore and try out different possibilities until it works out how to get the answer right. Reinforcement learning is sometime called learning with a critic because of this monitor that scores the answer, but does not suggest improvements.

Reinforcement learning is the problem of getting an agent to act in the world so as to maximize its rewards. A learner (the program) is not told what actions to take as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situations and, through that, all subsequent rewards.

Example

Consider teaching a dog a new trick: we cannot tell it what to do, but we can reward/punish it if it does the right/wrong thing. It has to find out what it did that made it get the reward/punishment. We can usea similar method to train computers to do many tasks, such as playing backgammon or chess, scheduling jobs, and controlling robot limbs. Reinforcement learning is different from supervisedlearning. Supervised learning is learning from examples provided by a knowledgeable expert.

Training,Testing and Validation datasets

Training Dataset

***Training Dataset**: The sample of data used to fit the model.*

The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). Themodel *sees* and *learns* from this data.

Validation Dataset

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset

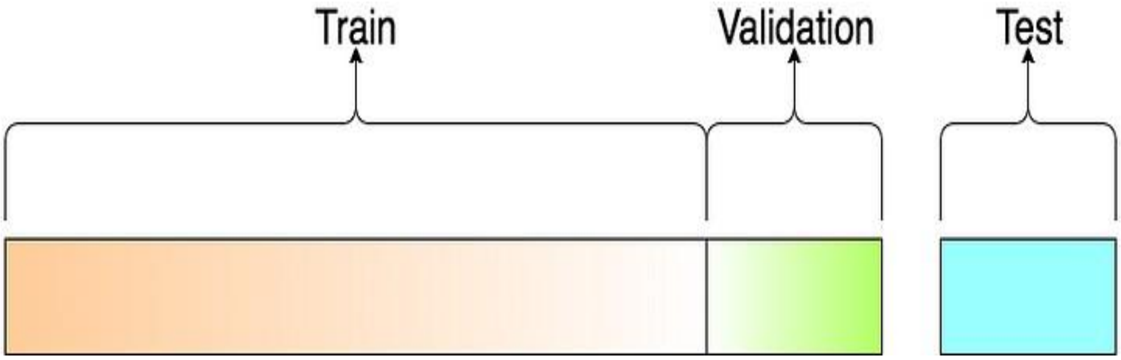
while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

The validation set is used to evaluate a given model, but this is for frequent evaluation. We, as machine learning engineers,use this data to fine-tune the model hyperparameters. Hence the model occasionally *sees* this data, but never does it "*Learn*" from this. We use the validation set results, and update higher level hyperparameters. So the validation set affects a model, but only indirectly. The validation set is also known as the Dev set or the Development set. This makes sense since this dataset helps during the "development" stage of the model.

Test Dataset

***Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.* The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained(using the train and validation sets). The test set is generally what is used to evaluate competing models (For example on many Kaggle competitions, the validation set is released initially along with the training set and the actual

test set is only released when the competition is about to close, and it is the result of the the model on the Test set thatdecides the winner). Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

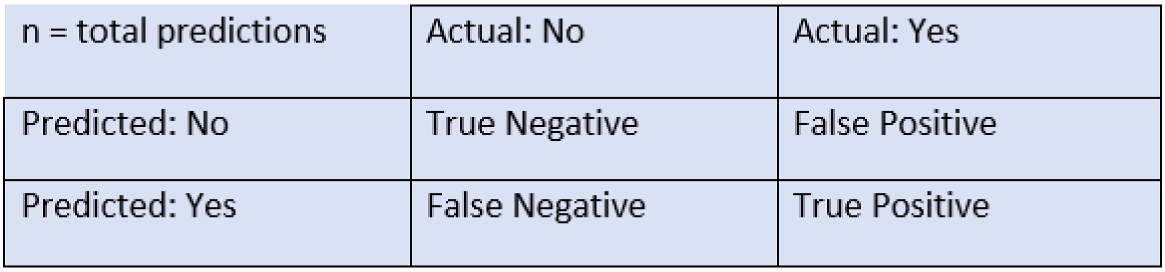


Performance Metrics:

Confusion matrix,Accuracy metrics Confusion Matrix in Machine Learning

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the totalnumber of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for thegiven observations.
- It looks like the below table:



n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

The above table has the following cases

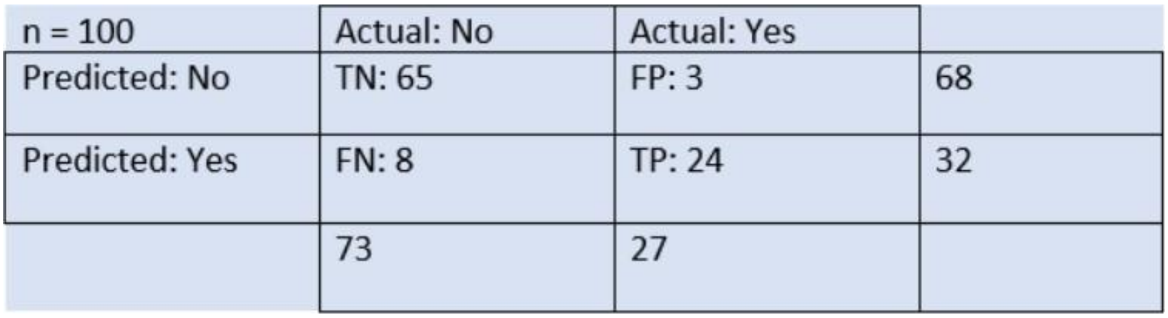
- **True Negative:** Model has given prediction No, and the real or actual value was also No.
- **True Positive:** The model has predicted yes, and the actual value was also true.
- **False Negative:** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
- **False Positive:** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error.**

Need for Confusion Matrix in Machine learning

- It evaluates the performance of the classification models, when they make predictions on test data, and tells howgood our classification model is.
- It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-IIerror.
- With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy,precision, etc.

Example: We can understand the confusion matrix using an example.

Suppose we are trying to create a model that can predict the result for the disease that is either a person has that diseaseor not. So, the confusion matrix for this is given as:



n = 100	Actual: No	Actual: Yes	
Predicted: No	TN: 65	FP: 3	68
Predicted: Yes	FN: 8	TP: 24	32
	73	27	

- The table is given for the two-class classifier, which has two predictions "Yes" and "NO." Here, Yes defines that patient has the disease, and No defines that patient does not has that disease.
- The classifier has made a total of **100 predictions**. Out of 100 predictions, **89 are true predictions**, and **11 areincorrect predictions**.
- The model has given prediction "yes" for 32 times, and "No" for 68 times. Whereas the actual "Yes" was 27, andactual "No" was 73 times.

Calculations using Confusion Matrix:

We can perform various calculations for the model, such as the model's accuracy, using this matrix. These calculationsare given below:

- **Classification Accuracy:** It is one of the important parameters to determine the accuracy of the classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula is given

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

- **Misclassification rate:** It is also termed as Error rate, and it defines how often the model gives the wrong predictions. The value of error rate can be calculated as the number of incorrect predictions to all number of the predictions made by the classifier. The formula is given below:

$$\text{Error rate}=\frac{FP+FN}{TP+FP+FN+TN}$$

-

- **Precision:** It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below

$$\text{Precision}= \frac{TP}{TP+FP}$$

- **Recall:** It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F-measure:** If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision. It can be calculated using the below formula:

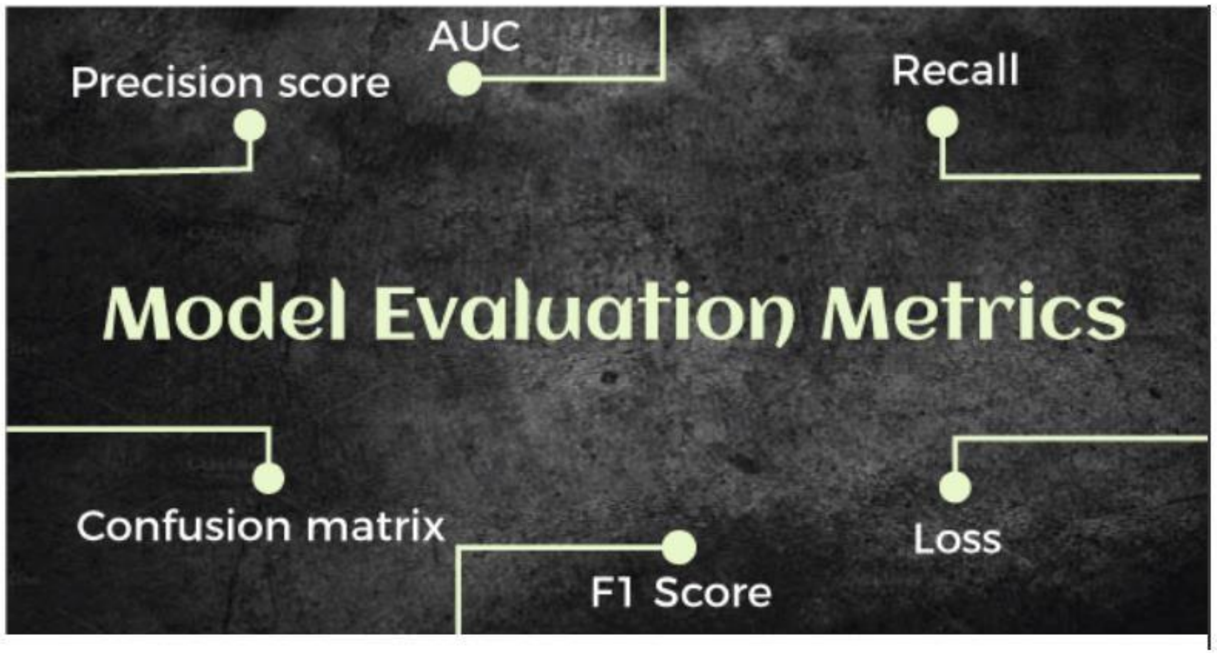
$$\text{F-measure} = \frac{2*Recall*Precision}{Recall+Precision}$$

- **Null Error rate:** It defines how often our model would be incorrect if it always predicted the majority class. Asper the accuracy paradox, it is said that "*the best classifier has a higher error rate than the null error rate.*"
- **ROC Curve:** The ROC is a graph displaying a classifier's performance for all possible thresholds. The graph isplotted between the true positive rate (on the Y-axis) and the false Positive rate (on the x-axis).

Performance Metrics in Machine Learning

Evaluating the performance of a Machine learning model is one of the important steps while building an effective ML model. ***To evaluate the performance or quality of the model, different metrics are used, and these metrics are knownas 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. EachML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.

Other important terms used in Confusion Matrix:



In machine learning, 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.

1. Performance Metrics for Classification

In a classification problem, the category or classes of data is 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. To evaluate the performance of a classification model, differentmetrics are used, and some of them are as follows:

- **Accuracy**
- **Confusion Matrix**

- **Precision**
- **Recall**
- **F-Score**
- **AUC(Area Under the Curve)-ROC**

I. Accuracy

The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number ofcorrect predictions to the total number of predictions.

It can be formulated as:

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

To implement an accuracy metric, we can compare ground truth and predicted values in a loop, or we can also use the scikit-learn module for this.

Firstly, we need to import the *accuracy_score* function of the scikit-learn library as follows:

1. from sklearn.metrics import accuracy_score2.
3. Here, metrics is a class of sklearn.4.
5. Then we need to pass the ground truth and predicted values in the function to calculate the accuracy.6.
7. print(f'Accuracy Score is {accuracy_score(y_test,y_hat)}')

Although it is simple to use and implement, it is suitable only for cases where an equal number of samples belong to eachclass.

When to Use Accuracy?

It is good to use the Accuracy metric when the target variable classes in data are approximately balanced. For example, if60% of classes in a fruit image dataset are of Apple, 40% are Mango. In this case, if the model is asked to predict whether the image is of Apple or Mango, it will give a prediction with 97% of accuracy.

When not to use Accuracy?

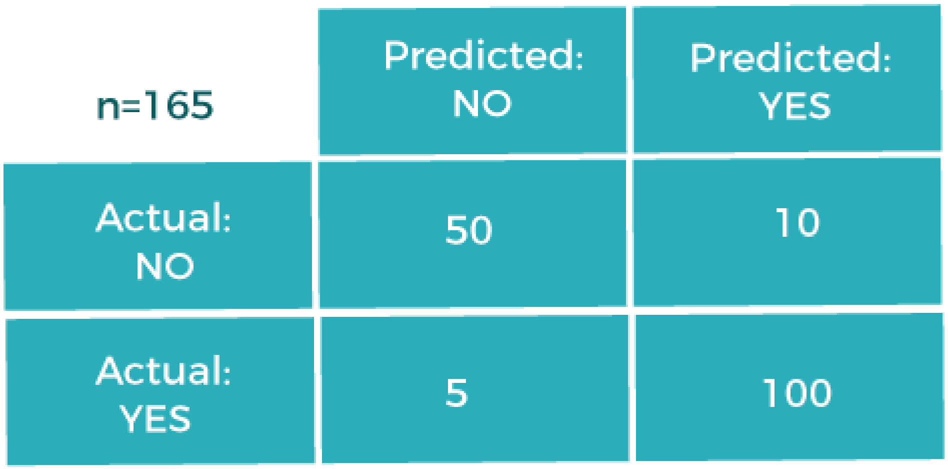
It is recommended not to use the Accuracy measure when the target variable majorly belongs to one class. For example, Suppose there is a model for a disease prediction in which, out of 100 people, only five people have a disease, and 95 people don't have one. In this case, if our model predicts every person with no disease (which means a bad prediction),the Accuracy measure will be 95%, which is not correct.

II. Confusion Matrix

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.

The confusion matrix is simple to implement, but the terminologies used in this matrix might be confusing for beginners.

A typical confusion matrix for a binary classifier looks like the below image(However, it can be extended to use for classifiers with more than two classes).



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

We can determine the following from the above matrix:

- In the matrix, columns are for the prediction values, and rows specify the Actual values. Here Actual and prediction give two possible classes, Yes or No. So, if we are predicting the presence of a disease in a patient, the Prediction column with Yes means, Patient has the disease, and for NO, the Patient doesn't have the disease.
- In this example, the total number of predictions are 165, out of which 110 time

predicted yes, whereas 55 times predicted No.

- However, in reality, 60 cases in which patients don't have the disease, whereas 105 cases in which patients have the disease.

In general, the table is divided into four terminologies, which are as follows:

1. **True Positive(TP):** In this case, the prediction outcome is true, and it is true in reality, also.
2. True Negative(TN): in this case, the prediction outcome is false, and it is false in reality, also.
3. False Positive(FP): In this case, prediction outcomes are true, but they are false in actuality.
4. False Negative(FN): In this case, predictions are false, and they are true in actuality.

III. Precision

The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

$$\boldsymbol{Precision = \frac{TP}{(TP + FP)}}$$

IV. Recall or Sensitivity

It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

The formula for calculating Recall is given below:

$$\text{Recall} = \frac{TP}{TP+FN}$$

When to use Precision and Recall?

From the above definitions of Precision and Recall, we can say that recall determines the performance of a classifier with respect to a false negative, whereas precision gives information about the performance of a classifier with respect to a false positive.

So, if we want to minimize the false negative, then, Recall should be as near to 100%, and if we want to minimize the false positive, then precision should be close to 100% as possible.

In simple words, *if we maximize precision, it will minimize the FP errors, and if we maximize recall, it will minimize the FN error.*

V. F-Scores

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. It is calculated with the help of Precision and Recall. It is a type of single score that represents both Precision and Recall. So, ***the 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:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

When to use F-Score?

As F-score make use of both precision and recall, so it should be used if both of them are important for evaluation, but one (precision or recall) is slightly more important to consider than the other. For example, when False negatives are comparatively more important than false positives, or vice versa.

VI. AUC-ROC

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.

Firstly, let's understand ROC (Receiver Operating Characteristic curve) curve. ***ROC represents a graph to show the performance of a classification model at different threshold levels***. The curve is plotted between two parameters, which are:

- **True Positive Rate**
- **False Positive Rate**

TPR or true Positive rate is a synonym for Recall, hence can be calculated as:

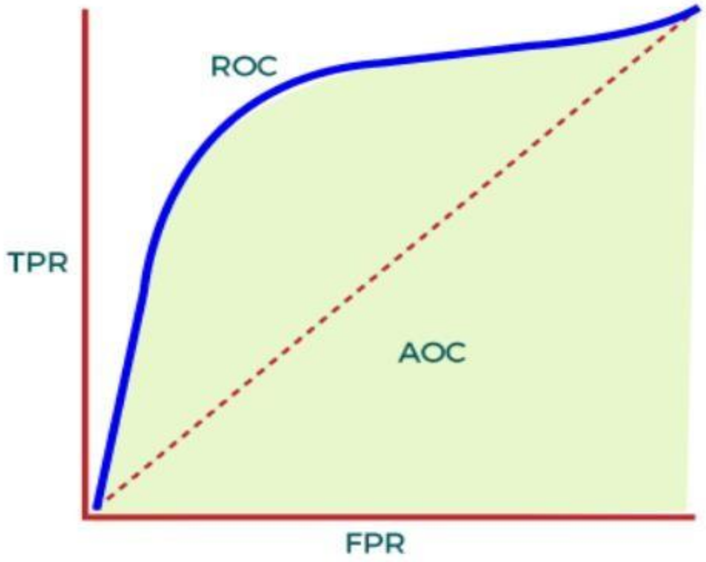
$$TPR = \frac{TP}{TP + FN}$$

FPR or False Positive Rate can be calculated as:

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

To calculate value at any point in a ROC curve, we can evaluate a logistic regression model multiple times with different classification thresholds, but this would not be much efficient. So, for this, one efficient method is used, which is known as AUC.

VII. AUC: Area Under the ROC curve



AUC is known for **Area Under the ROC curve**. As its name suggests, AUC calculates the two-dimensional area under the entire ROC curve, as shown below image:

AUC calculates the performance across all the thresholds and provides an aggregate measure. The value of AUC ranges from 0 to 1. It means a model with 100% wrong prediction will have an AUC of 0.0, whereas models with 100% correct predictions will have an AUC of 1.0.

When to Use AUC

AUC should be used to measure how well the predictions are ranked rather than their absolute values. Moreover, it measures the quality of predictions of the model without considering the classification threshold.

When not to use AUC

As AUC is scale-invariant, which is not always desirable, and we need calibrating probability outputs, then AUC is not preferable.

Further, AUC is not a useful metric when there are wide disparities in the cost of false negatives vs. false positives, and itis difficult to minimize one type of classification error.

Descriptive statistics

1. Central Tendency of Data

1.1 Mean

1.2 Median

1.3 Mode

2. Dispersion of Data

2.1 Inter Quartile Range (IQR)

2.2 Range

2.3 Standard Deviation

2.4 Variance

Once we have collected the data, what will we do with it? Data can be analysed and used in various methods and formats. There are two types of statistical methods widely used for analyzing data.

1.Descriptivestatistics

2. Inferential statistics

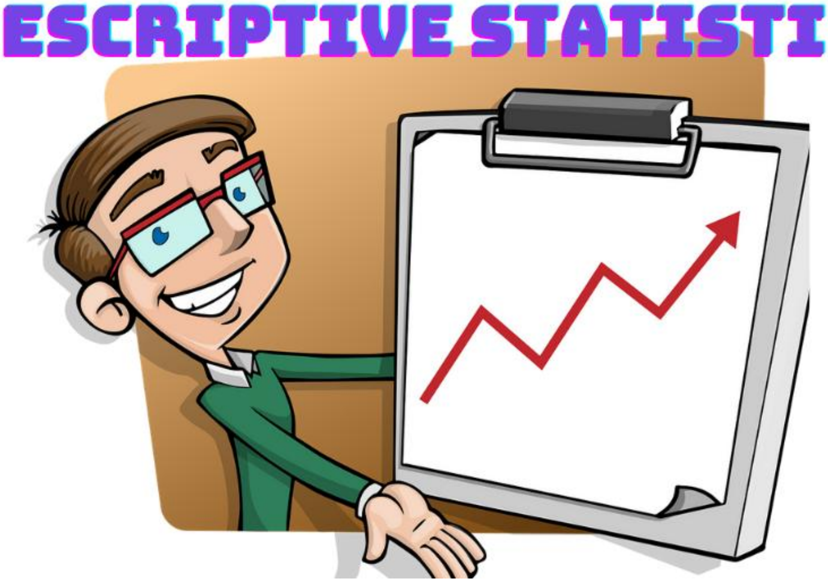
While analyzing a dataset, We use statistical methods to arrive at a conclusion. Data-driven decision-making also depends on how efficiently we use these methods.

Now, let us dive into these methods deeply.

1. Descriptive statistics

The study of numerical and graphical ways to describe and display your data is called descriptive statistics. It describes the data and helps us understand the features of the data by summarizing the given sample set or population of data. In descriptive statistics, we usually take the sample into account.

DESCRIPTIVE STATISTICS



Statisticians use graphical representation of data to get a clear picture of the data. Business trends can be analysed easily with these representations. visual representation is more effective than presenting huge numbers.

We can describe these data in various dimensions. Various dimensions of describing data are

1. Central Tendency of Data

2. Dispersion of Data

1. Central Tendency of Data

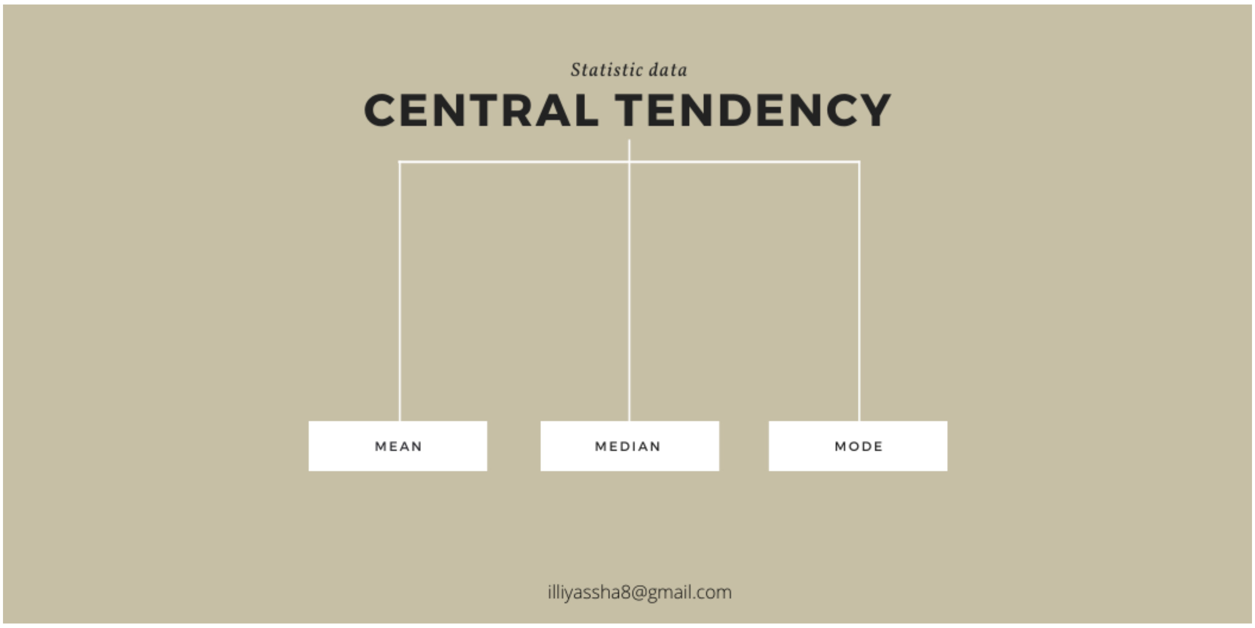
This is the center of the distribution of data. It describes the location of data and concentrates where the data is located.

The three most widely used measures of the “**center**” of the data are

1.1 Mean

1.2 Median

1.3 Mode



Let us see these measures in detail,

1.1 Mean

The “Mean” is the average of the data.

Average can be identified by summing up all the numbers and then dividing them by the number of observations.

Mean = $X_1 + X_2 + X_3 + \ldots + X_n / n$

Example:

Data – 10,20,30,40,50 and Number of observations = 5

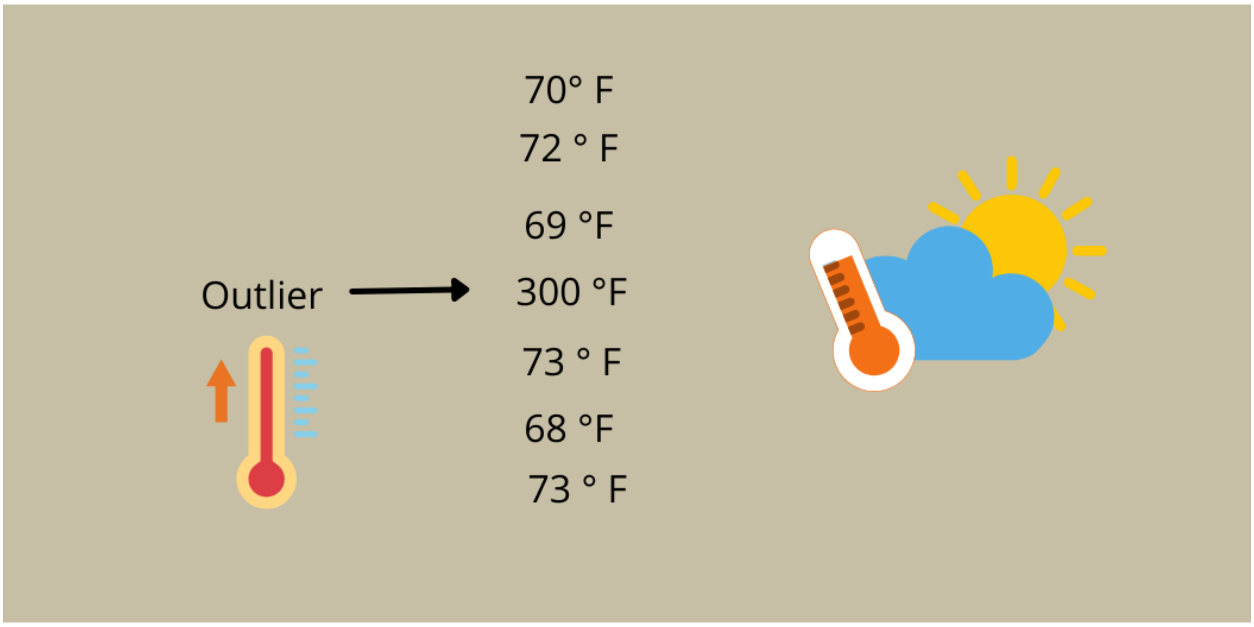
Mean = [10+20+30+40+50] / 5

Mean = 30

Outliers influence the central tendency of the data.

What are Outliers?

Outliers are extreme behaviours. An outlier is a data point that differs significantly from other observations. It can cause serious problems in analysis.



Example:

Data – 10,20,30,40,200

Mean = [10+20+30+40+200] / 5

Mean = 60

Solution for Outliers problem

Removing the outliers while taking average will give us good results.

1.2 Median

Median is the 50%th percentile of the data. It is exactly the center point of the data.

Median can be identified by ordering the data and splits the data into two equal parts and find the number. It is the best way to find the center of the data.

Because the central tendency of the data is not affected by outliers. Outliers don't influence the data.

Example:

Odd number of Data – 10,20,30,40,50

Median is 30.

Even number of data – 10,20,30,40,50,60

Find the middle 2 data and take the mean of that two values.

Here 30 and 40 are middle values.

30+40 / 2 =35

Median is 35

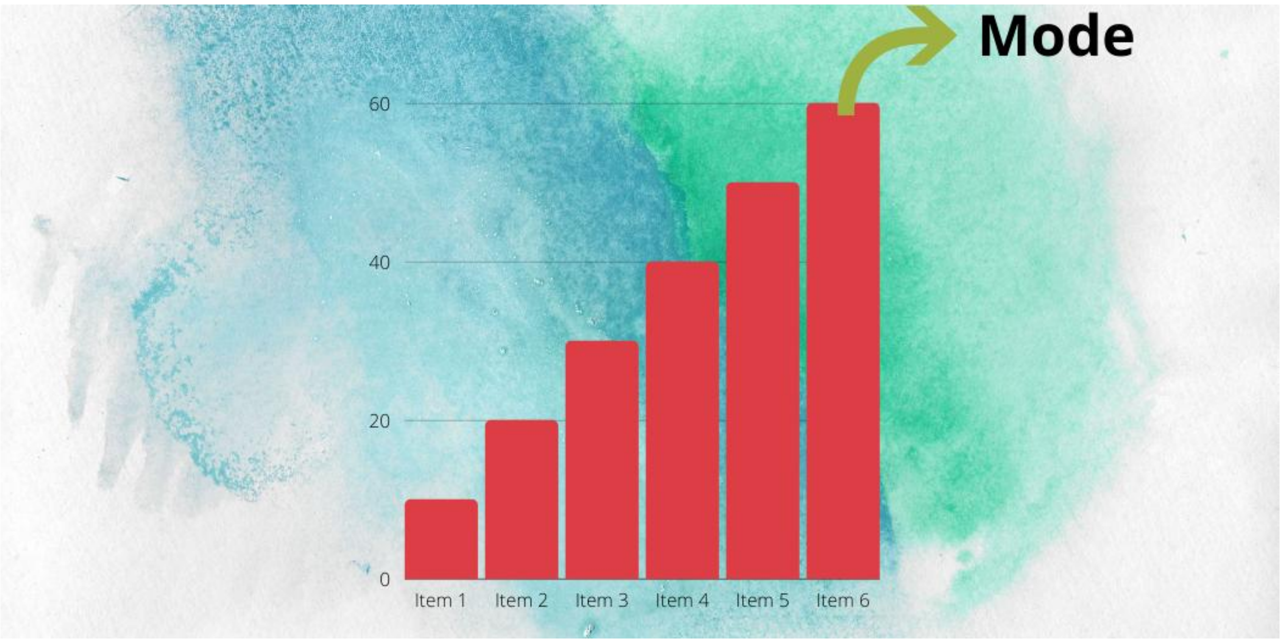
1.3 Mode

Mode is frequently occurring data or elements.

If an element occurs the highest number of times, it is the mode of that data. If no number in the data is repeated, then there is no mode for that data. There can be more than one mode in a dataset if two values have the same frequency and also the highest frequency.

Outliers don’t influence the data.

The mode can be calculated for both quantitative and qualitative data.



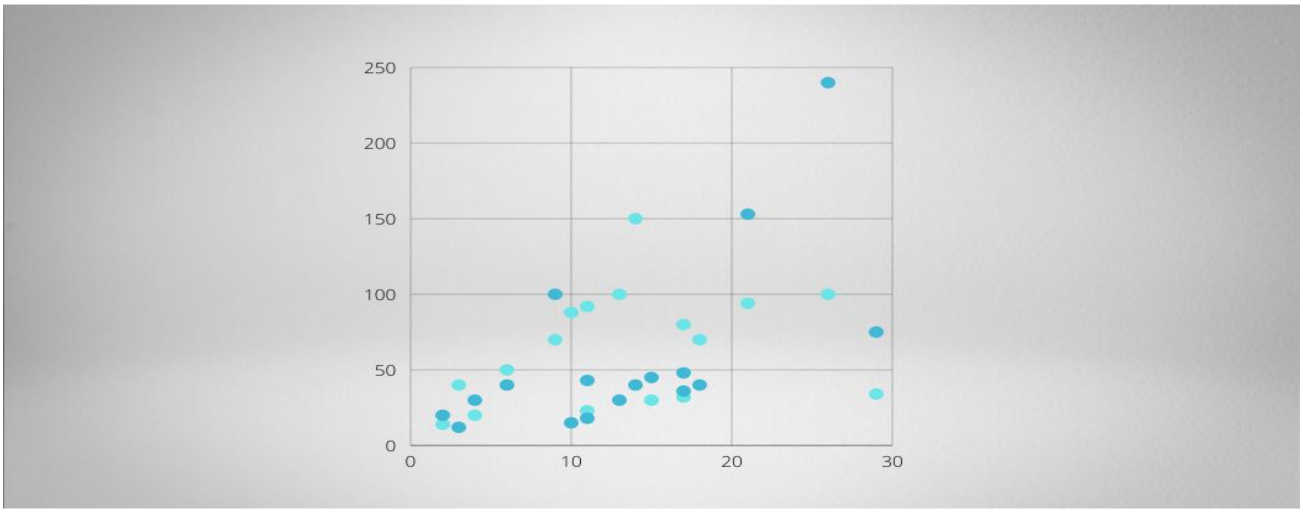
Example

Data – 1,3,4,6,7,3,3,5,10, 3

Mode is 3

because 3 has the highest frequency (4 times)

2. Dispersion of Data



The dispersion is the **"Spread of the data".** It measures how far the data is spread.

In most of the dataset, the data values are closely located near the mean. On some other dataset, the values are widely spread out of the mean. These dispersions of data can be measured by

2.1 Inter Quartile Range (IQR)

2.2 Range

2.3 Standard Deviation

2.4 Variance

Let us see these measures in detail,

1. Inter Quartile Range (IQR)

Quartiles are special percentiles.

1st **Quartile Q1** is the same as the 25th percentile.

2nd **Quartile Q2** is the same as 50th percentile.

3rd **Quartile Q3** is same as 75th percentile

Steps to find quartile and percentile

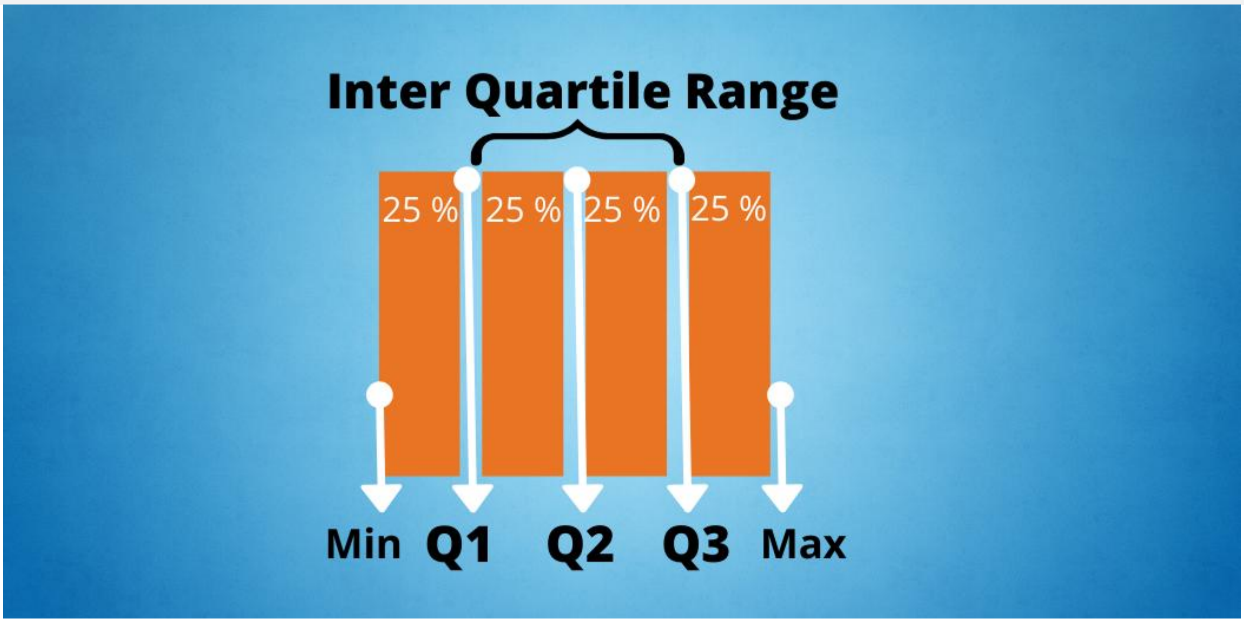
–The data should sort and ordered from the smallest to the largest.

–For Quartiles, ordered data is divided into 4 equal parts.

–For Percentiles, ordered data is divided into 100 equal parts.

Inter Quartile Range is the difference between the third quartile(Q3*)* and the first Quartile (Q1)

IQR = Q3- Q1



Inter Quartile range

It is the spread of the middle half (50%) of the data

2.2 Range

The range is the difference between the largest and the smallest value in the data.

Max – Min = Range

2.3 Standard Deviation

The most common measure of spread is the standard deviation.

The Standard deviation is the measure of **how far the data deviates** from the **mean value**.

The standard deviation formula varies for population and sample. Both formulas are similar, but not the same.

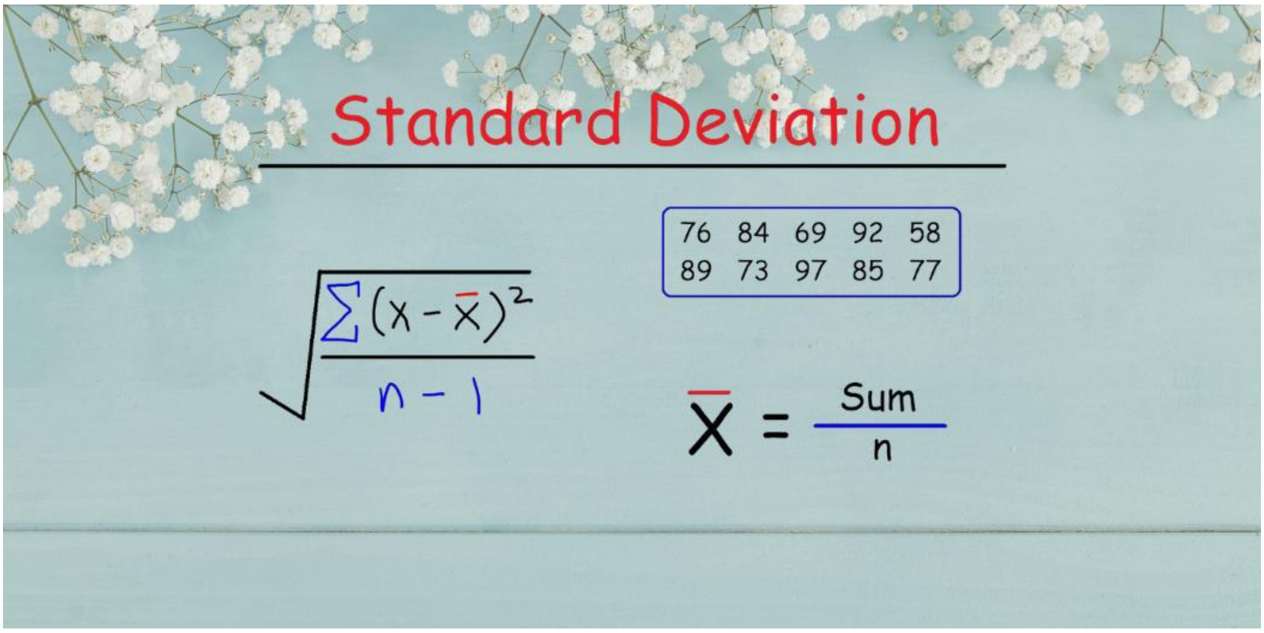
- Symbol used for **Sample Standard Deviation – "s"** (lowercase)
- Symbol used for **Population Standard Deviation – "σ"** (sigma, lower case)

Steps to find **Standard deviation**

If x is a number, then the difference "x – mean" is its deviation. The deviations are used to calculate the standard deviation.

Sample Standard Deviation, s = Square root of sample variance

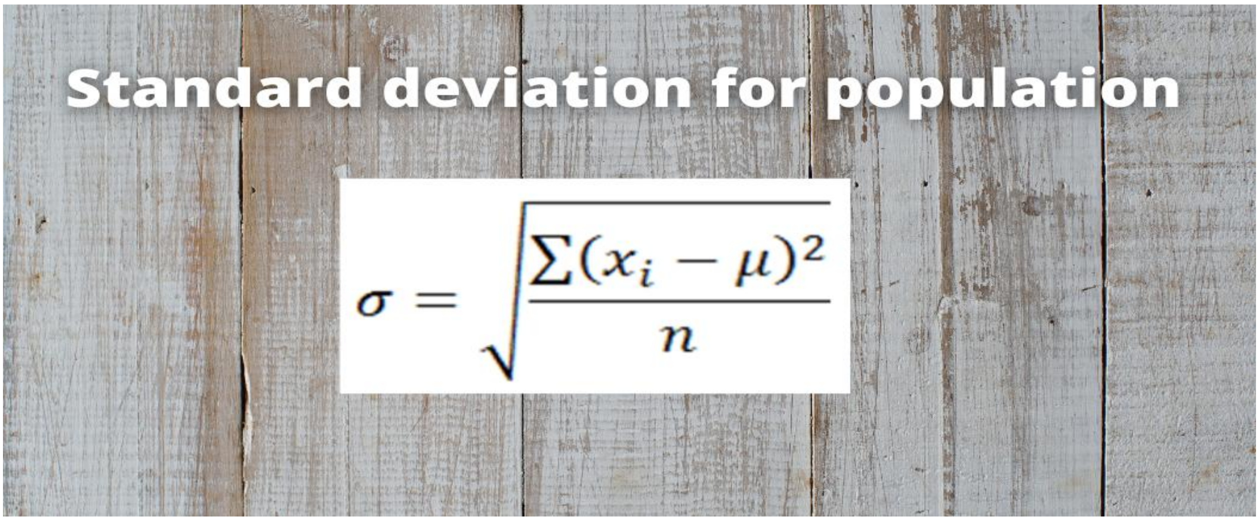
Sample Standard Deviation, s = Square root of $[\Sigma(x - \bar{x})^2/ n-1]$ where $\bar{x}$ is average and n is no. of samples


Standard Deviation
76 84 69 92 58
89 73 97 85 77
$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
$\bar{X} = \frac{\text{Sum}}{n}$

Standard Deviation for sample

Population Standard Deviation, σ = Square root of population variance

Population Standard Deviation, σ = Square root of [$\Sigma(x - \mu)^2 / N$] where μ is Mean and N is no.of population.


Standard deviation for population
$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$

The standard deviation for population

The standard deviation is always positive or zero. It will be large when the data values are spread out from the mean.

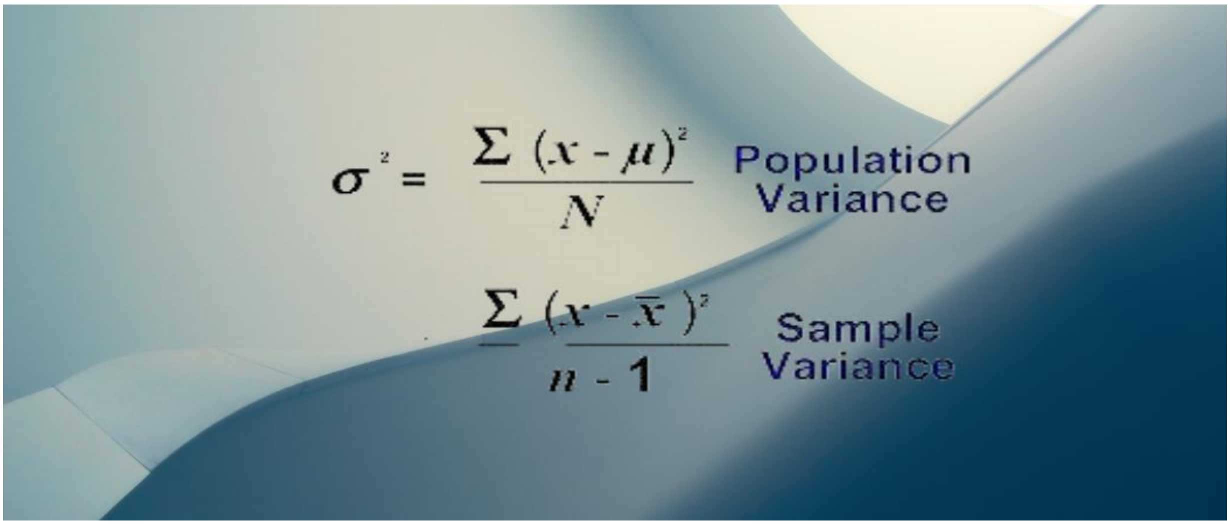
2.4 Variance

The variance is a measure of variability. It is the **average squared deviation from the mean**.

The symbol σ^2 represents the population variance and the symbol for s^2 represents sample variance.

Population variance **σ^2 = [**$\Sigma (x - \mu)^2 / N$]

Sample Variance $s^2 = [\ \Sigma(x - \bar{x})^2 / n\text{-}1]$



$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$
Population Variance
$\frac{\Sigma (x - \bar{x})^2}{n - 1}$
Sample Variance

2.5 Co-Variance

Covariance formula is a statistical formula which is used to assess the relationship between two variables. In simple words, **covariance** is one of the statistical measurement to know the relationship of the variance between the two variables.

The covariance indicates how two variables are related and also helps to know whether the two variables vary together or change together. The covariance is denoted as Cov(X,Y) and the formulas for covariance are given below.

Comparison of Mean, Median, and Mode

Mean	**Median**	**Mode**
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	It does not require measurement on all observations	It does not require measurement on all observations.
Uniquely and comprehensively defined.	Cannot be determined under all conditions.	Not uniquely defined for multi-modal situations.
Affected by extreme values.	Not affected by extreme values.	Not affected by extreme values.
Can be treated algebraically. In other words, means of several groups can be combined.	Cannot be treated algebraically, meaning, Medians of several groups cannot be combined.	Cannot be treated algebraically, since Modes of several groups cannot be combined.

Descriptive statistics

1. Central Tendency of Data

1.1 Mean

1.2 Median

1.3 Mode

2. Dispersion of Data

2.1 Inter Quartile Range (IQR)

2.2 Range

2.3 Standard Deviation

2.4 Variance

Once we have collected the data, what will we do with it? Data can be analysed and used in various methods and formats. There are two types of statistical methods widely used for analyzing data.

1.Descriptivestatistics
2. Inferential statistics

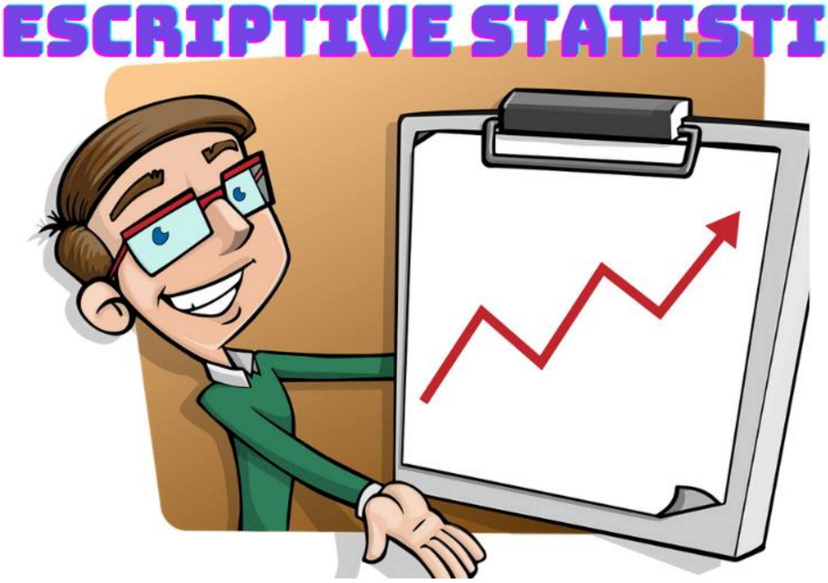
While analyzing a dataset, We use statistical methods to arrive at a conclusion. Data-driven decision-making also depends on how efficiently we use these methods.

Now, let us dive into these methods deeply.

1. Descriptive statistics

The study of numerical and graphical ways to describe and display your data is called descriptive statistics. It describes the data and helps us understand the features of the data by summarizing the given sample set or population of data. In descriptive statistics, we usually take the sample into account.

DESCRIPTIVE STATISTICS



Statisticians use graphical representation of data to get a clear picture of the data. Business trends can be analysed easily with these representations. visual representation is more effective than presenting huge numbers.

We can describe these data in various dimensions. Various dimensions of describing data are

1. Central Tendency of Data

2. Dispersion of Data

1. Central Tendency of Data

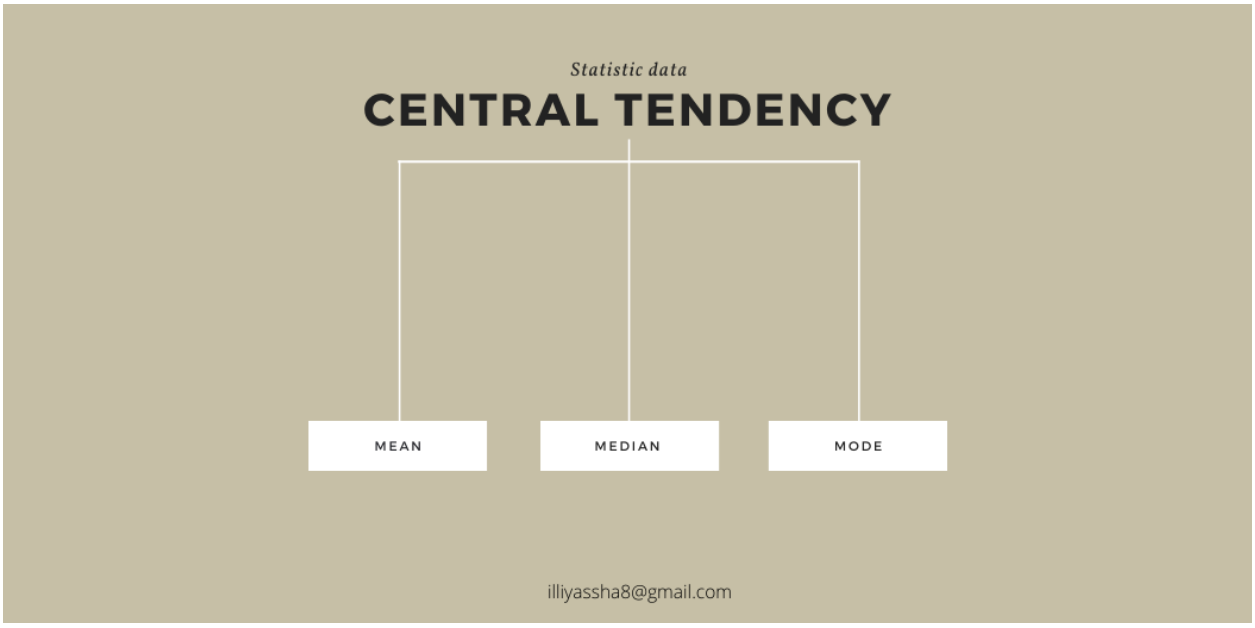
This is the center of the distribution of data. It describes the location of data and concentrates where the data is located.

The three most widely used measures of the "**center**" of the data are

1.1 Mean

1.2 Median

1.3 Mode



Let us see these measures in detail,

1.1 Mean

The “Mean” is the average of the data.

Average can be identified by summing up all the numbers and then dividing them by the number of observations.

Mean = $X_1 + X_2 + X_3 + \ldots + X_n / n$

Example:

Data – 10,20,30,40,50 and Number of observations = 5

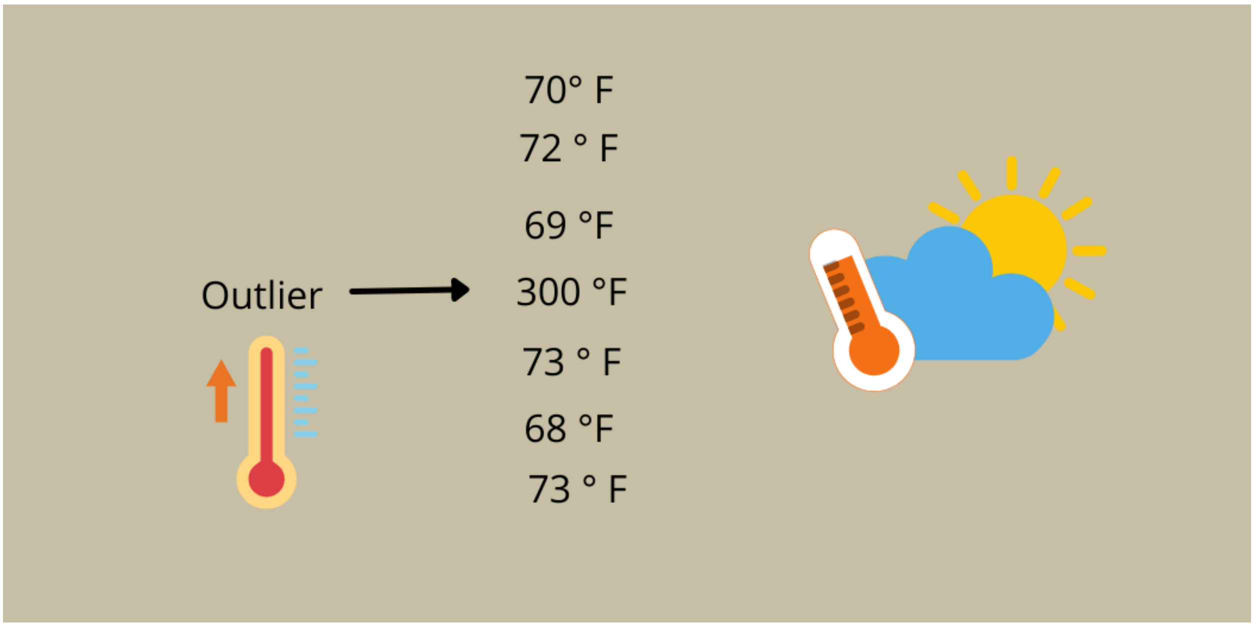
Mean = [10+20+30+40+50] / 5

Mean = 30

Outliers influence the central tendency of the data.

What are Outliers?

Outliers are extreme behaviours. An outlier is a data point that differs significantly from other observations. It can cause serious problems in analysis.



Example:

Data – 10,20,30,40,200

Mean = [10+20+30+40+200] / 5

Mean = 60

Solution for Outliers problem

Removing the outliers while taking average will give us good results.

1.2 Median

Median is the 50%th percentile of the data. It is exactly the center point of the data.

Median can be identified by ordering the data and splits the data into two equal parts and find the number. It is the best way to find the center of the data.

Because the central tendency of the data is not affected by outliers. Outliers don’t influence the data.

Example:

Odd number of Data – 10,20,30,40,50

Median is 30.

Even number of data – 10,20,30,40,50,60

Find the middle 2 data and take the mean of that two values.

Here 30 and 40 are middle values.

30+40 / 2 =35

Median is 35

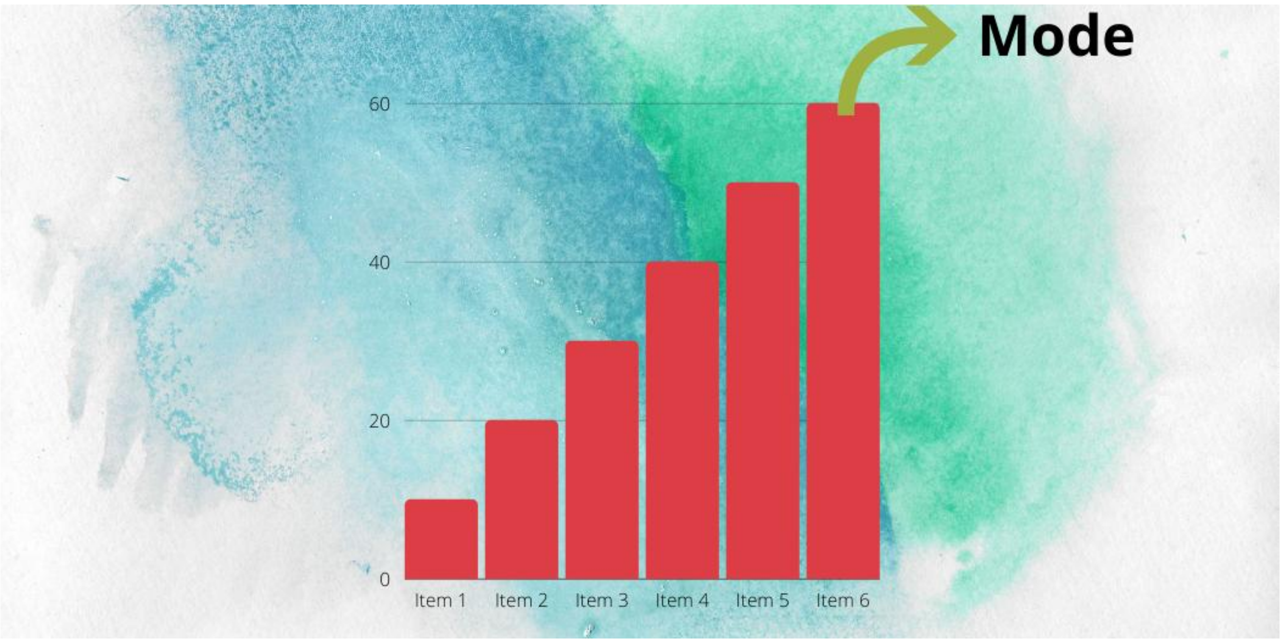
1.3 Mode

Mode is frequently occurring data or elements.

If an element occurs the highest number of times, it is the mode of that data. If no number in the data is repeated, then there is no mode for that data. There can be more than one mode in a dataset if two values have the same frequency and also the highest frequency.

Outliers don't influence the data.

The mode can be calculated for both quantitative and qualitative data.



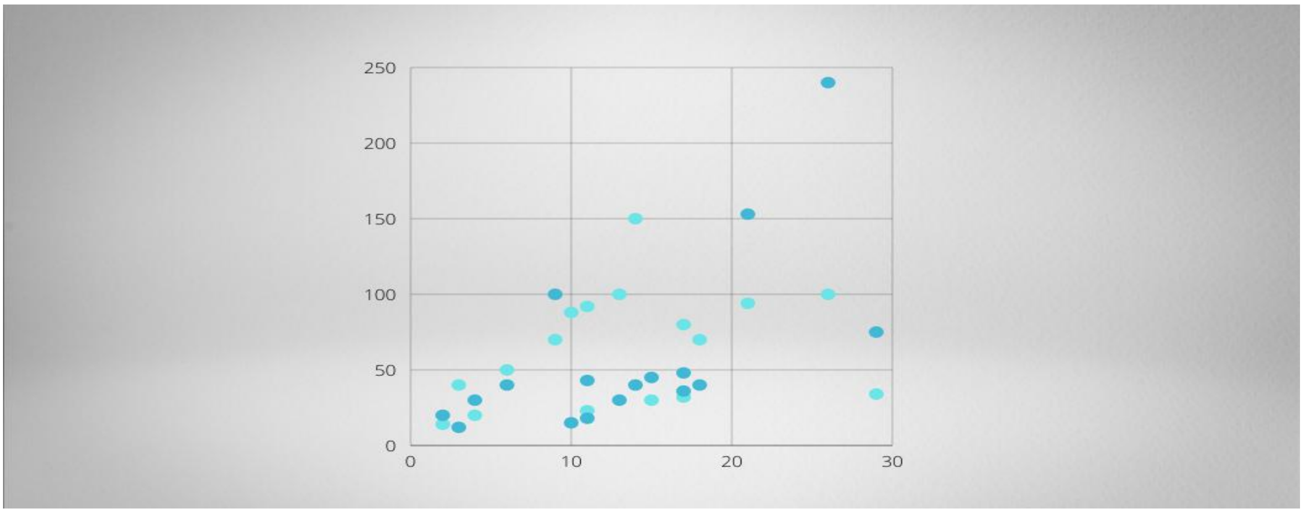
Example

Data – 1,3,4,6,7,3,3,5,10, 3

Mode is 3

because 3 has the highest frequency (4 times)

2. Dispersion of Data



The dispersion is the **“Spread of the data”.** It measures how far the data is spread.

In most of the dataset, the data values are closely located near the mean. On some other dataset, the values are widely spread out of the mean. These dispersions of data can be measured by

2.1 Inter Quartile Range (IQR)

2.2 Range

2.3 Standard Deviation

2.4 Variance

Let us see these measures in detail,

1. Inter Quartile Range (IQR)

Quartiles are special percentiles.

1st **Quartile Q1** is the same as the 25th percentile.

2nd **Quartile Q2** is the same as 50th percentile.

3rd **Quartile Q3** is same as 75th percentile

Steps to find quartile and percentile

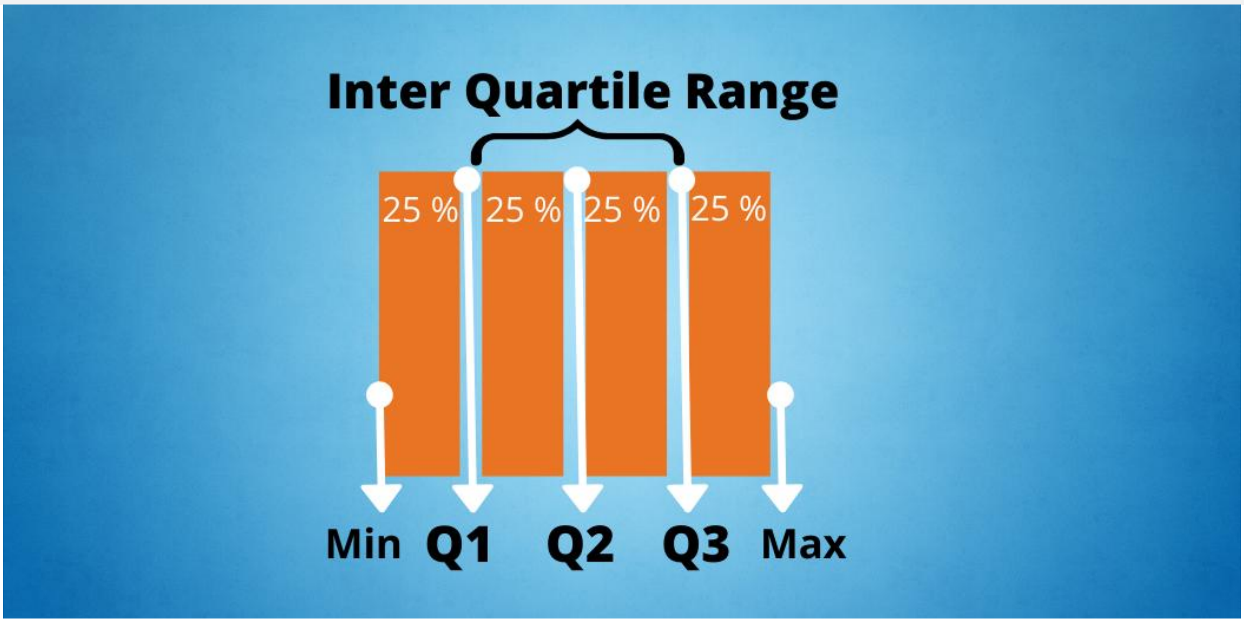
–The data should sort and ordered from the smallest to the largest.

–For Quartiles, ordered data is divided into 4 equal parts.

–For Percentiles, ordered data is divided into 100 equal parts.

Inter Quartile Range is the difference between the third quartile(Q3*)* and the first Quartile (Q1)

IQR = Q3- Q1



Inter Quartile range

It is the spread of the middle half (50%) of the data

2.2 Range

The range is the difference between the largest and the smallest value in the data.

Max – Min = Range

2.3 Standard Deviation

The most common measure of spread is the standard deviation.

The Standard deviation is the measure of **how far the data deviates** from the **mean value**.

The standard deviation formula varies for population and sample. Both formulas are similar, but not the same.

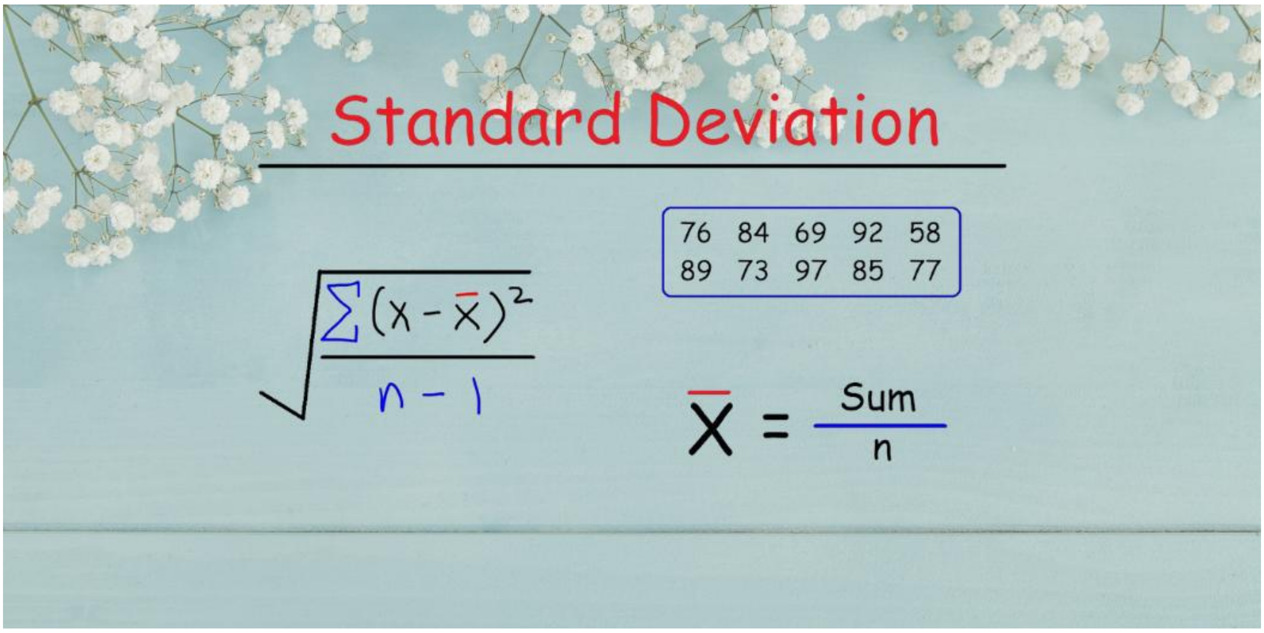
- Symbol used for **Sample Standard Deviation – "s"** (lowercase)
- Symbol used for **Population Standard Deviation – "σ"** (sigma, lower case)

<u>Steps to find **Standard deviation**</u>

If x is a number, then the difference "x – mean" is its deviation. The deviations are used to calculate the standard deviation.

Sample Standard Deviation, s = Square root of sample variance

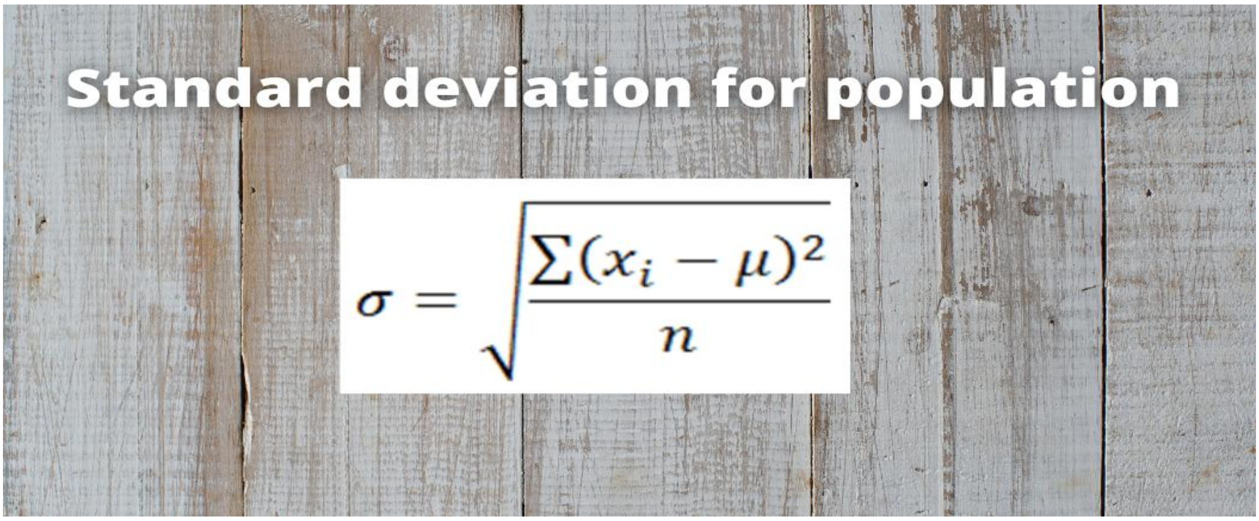
Sample Standard Deviation, s = Square root of $[\Sigma(x - \bar{x})^{2}/ n\text{-}1]$ where $\bar{x}$ is average and n is no. of samples


Standard Deviation
76 84 69 92 58
89 73 97 85 77
$\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$
$\bar{X} = \frac{\text{Sum}}{n}$

Standard Deviation for sample

Population Standard Deviation, σ = Square root of population variance

Population Standard Deviation, σ = Square root of [$\Sigma(x - \mu)^2 / N$] where μ is Mean and N is no.of population.


Standard deviation for population
$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$

The standard deviation for population

The standard deviation is always positive or zero. It will be large when the data values are spread out from the mean.

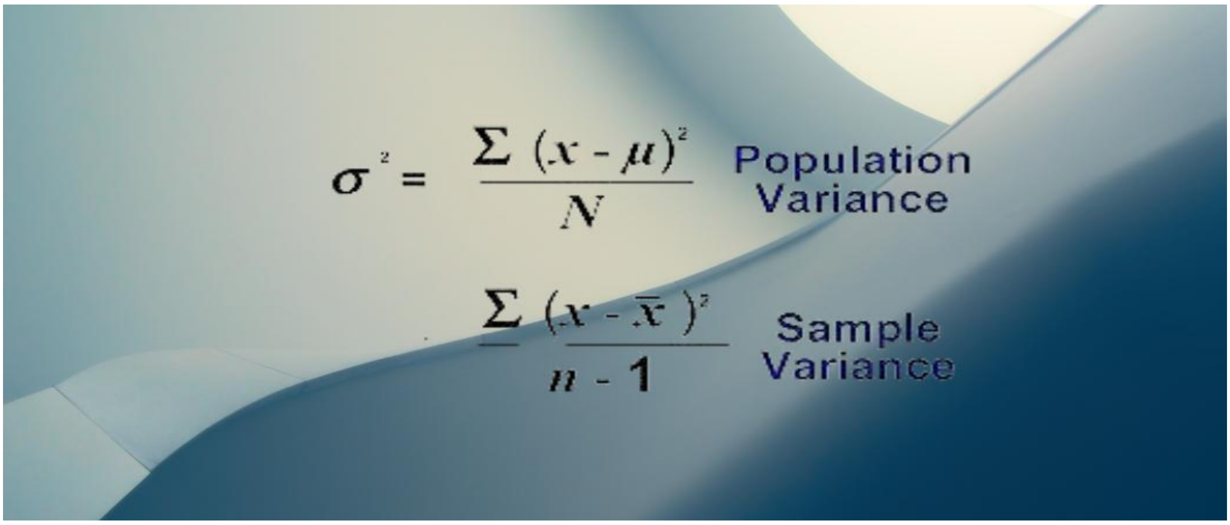
2.4 Variance

The variance is a measure of variability. It is the **average squared deviation from the mean**.

The symbol σ^2 represents the population variance and the symbol for s^2 represents sample variance.

Population variance **σ^2 = [**$\Sigma (x - \mu)^2 / N$]

Sample Variance s^2 = [$\Sigma(x - \bar{x})^2/ n-1$]



$$\sigma^2 = \frac{\Sigma (x-\mu)^2}{N}$$

$$\frac{\Sigma (x-\bar{x})^2}{n-1}$$

2.5 Co-Variance

Covariance formula is a statistical formula which is used to assess the relationship between two variables. In simple words, **covariance** is one of the statistical measurement to know the relationship of the variance between the two variables.

The covariance indicates how two variables are related and also helps to know whether the two variables vary together or change together. The covariance is denoted as Cov(X,Y) and the formulas for covariance are given below.

Comparison of Mean, Median, and Mode

Mean	**Median**	**Mode**
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	It does not require measurement on all observations	It does not require measurement on all observations.
Uniquely and comprehensively defined.	Cannot be determined under all conditions.	Not uniquely defined for multi-modal situations.
Affected by extreme values.	Not affected by extreme values.	Not affected by extreme values.
Can be treated algebraically. In other words, means of several groups can be combined.	Cannot be treated algebraically, meaning, Medians of several groups cannot be combined.	Cannot be treated algebraically, since Modes of several groups cannot be combined.

Feature, Feature vector, feature space, feature extraction and feature selection

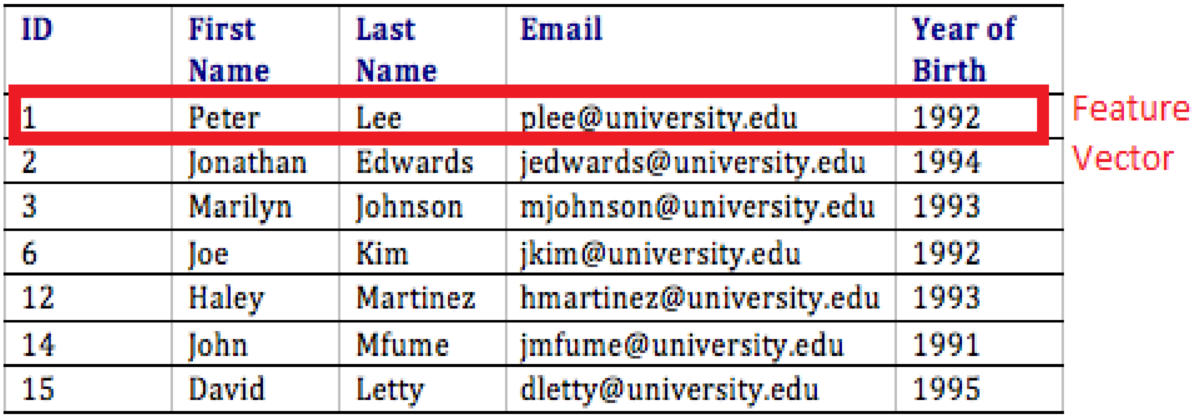
Feature: is a list of numbers eg: age, name, height, weight etc., that means every column is afeature in relational table.

Feature Vector is representation of particular row in relational table. Each row is a feature vector,row 'n' is a feature vector for the 'n'th sample.

Feature Set: Help to predict the output variable.

Example: To predict the age of particular person we need to know the year of birth. Here FeatureSet = Year of Birth.

Normally good feature set can be identified using expert domain knowledge or mathematicalapproach.

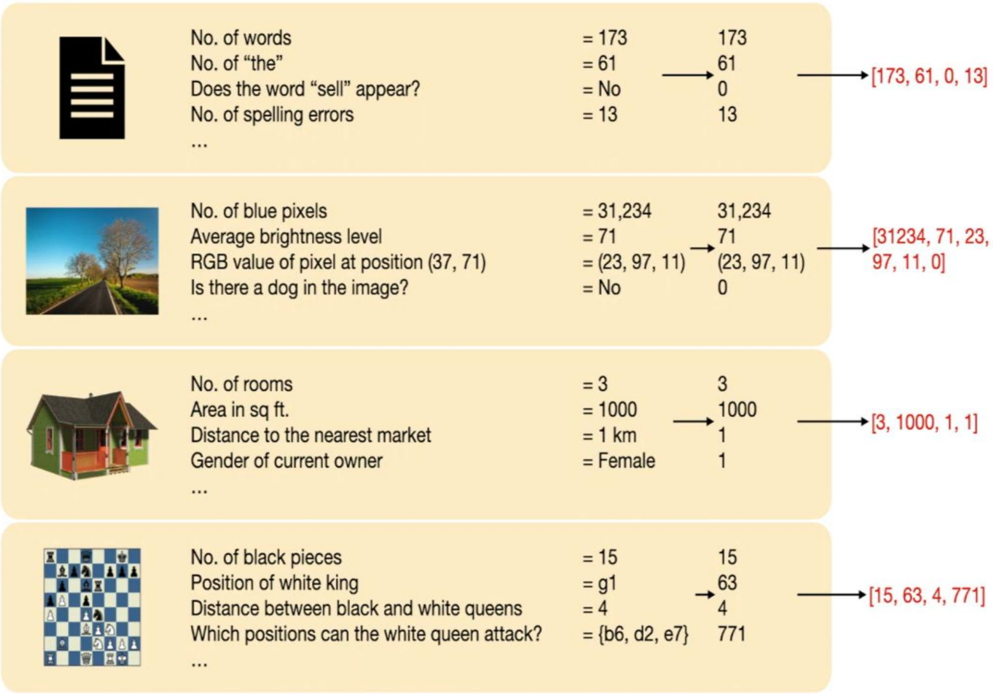


ID	First Name	Last Name	Email	Year of Birth
1	Peter	Lee	plee@university.edu	1992
2	Jonathan	Edwards	jedwards@university.edu	1994
3	Marilyn	Johnson	mjohnson@university.edu	1993
6	Joe	Kim	jkim@university.edu	1992
12	Haley	Martinez	hmartinez@university.edu	1993
14	John	Mfume	jmfume@university.edu	1991
15	David	Letty	dletty@university.edu	1995

Feature Vector

Table: Students

Feature vector: Collection of **numerical** features


No. of words = 173 173
No. of "the" = 61 61
Does the word "sell" appear? = No 0
No. of spelling errors = 13 13
...
[173, 61, 0, 13]
No. of blue pixels = 31,234 31,234
Average brightness level = 71 71
RGB value of pixel at position (37, 71) = (23, 97, 11) (23, 97, 11)
Is there a dog in the image? = No 0
...
[31234, 71, 23, 97, 11, 0]
No. of rooms = 3 3
Area in sq ft. = 1000 1000
Distance to the nearest market = 1 km 1
Gender of current owner = Female 1
...
[3, 1000, 1, 1]
No. of black pieces = 15 15
Position of white king = g1 63
Distance between black and white queens = 4 4
Which positions can the white queen attack? = {b6, d2, e7} 771
...
[15, 63, 4, 771]

In machine learning, **a feature vector** is a numerical representation of an object or entity in a dataset, which is used as input to a machine learning algorithm. A feature vector is a one-dimensional array of numbers, where each number corresponds to a specific feature or attribute of the object being represented. The feature vector is derived from the raw data of the object, through a process of featureextraction or feature engineering.

The feature space is the set of all possible feature vectors that can be created from the raw data. Each feature vector represents a point in the feature space. The dimensionality of the feature space is determined by the number of features in the feature vector. For example, if the feature vector has three features, the feature spacewould be three-dimensional.

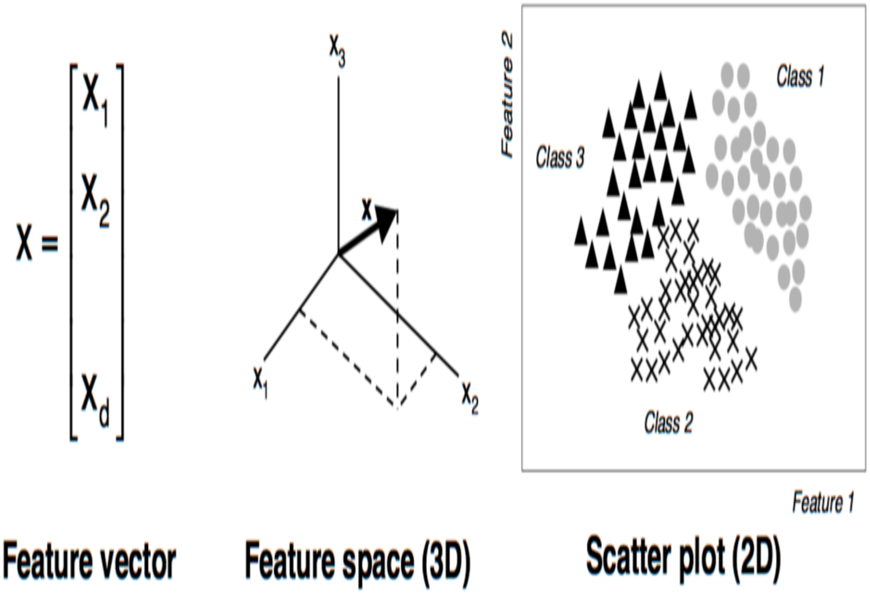
let's consider an example to illustrate this concept. Suppose we have a dataset of houses for sale, and we want to predict their prices based on a set of features. The features we are interested in are the square footage, the numberof bedrooms, the number of bathrooms, and the location of the house.

We can represent each house as a feature vector, where the first element is the square footage, the second elementis the number of bedrooms, the third element is the number of bathrooms, and the fourth element is a one-hot encoding of the location. For example, if there are three possible locations (A, B, C), the fourth element of the feature vector would be a vector of length three, where the element corresponding to the location of the house is 1and the other elements are 0.

Suppose we have a dataset of 100 houses, with the following feature vectors:

Square Footage	**Bedrooms**	**Bathrooms**	**Location**
1500	2	1	A
2000	3	2	B
1200	2	1	C
...	...	...	...

The feature space for this dataset would be four-dimensional, with each house represented as a point in this four- dimensional space. The machine learning algorithm would use this feature space to learn a function that maps the feature vectors to the corresponding prices of the houses. The goal of the algorithm is to find a function that accurately predicts the price of a house based on its features.



Feature vector

Feature space (3D)

Scatter plot (2D)

Feature selection and feature extraction are two different techniques used in machine learning to reduce the dimensionality of a dataset by selecting or creating a subset of the most relevant features that are likely to contribute the most to the predictive performance of a model. Although they have a similar goal, they differ in how they achieve it.

Feature selection involves selecting a subset of the most important features from the original dataset, while discarding the remaining features. This can be done by using statistical methods to measure the correlation or relevance of each feature to the target variable, or by using domain knowledge to manually select the most informative features. The selected features can then be used as input to a machine learning algorithm to build a model.

On the other hand, **feature extraction** involves creating new features by transforming or combining the original features in some way, such that the new features capture the most important information from the original features. This can be done using techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), or non-negative matrix factorization (NMF). The new features can then be used as input to a machine learning algorithm to build a model.

In both cases, the goal is to reduce the dimensionality of the dataset, while preserving the most relevant information for building an accurate machine learning model. However, feature selection tends to be more straightforward and interpretable, as it simply involves selecting a subset of the original features. Feature extraction can be more complex, as it involves creating new features that may not have a direct interpretation in terms of the original dataset.

Difference between Feature extraction and Feature Selection

Sr. No.	Feature Selection	Feature Extraction
1	Selects a subset of relevant features from the original set of features.	Extracts a new set of features that are more informative and compact.
2	Reduces the dimensionality of the feature space and simplifies the model.	Captures the essential information from the original
3	Can be categorized into filter, wrapper, and embedded methods.	Can be categorized into linear and nonlinear methods.
4	Requires domain knowledge and feature engineering.	Can be applied to raw data without feature engineering