The concept of skewness of data emerges from the fact when we have data that has absolute extremes and when we plot it shows a rising or lowering plot of the datapoints.

When we observe this , it is better to find the median to find the central point of the data rather than the average of the data points

This is called **skewness** and data is called skewed when they have high extremes.

Data can be distributed into different types of distributions like

Normal Distribution, Binomial Distribution, Poisson Distribution

Scaling is the process through which we change our data’s magnitudes in order to make it comparable to each other, we cannot compare apples to watermelons therefore scaling is required

There are two types of scaling mainly to bring equal importance(weights) to the features which have larger magnitudes and smaller magnitudes alike-

1. Nomalization-

Normalization is a process to scale all the values in a range to the range of [0,1] , the formula-

(Xmax-Xmin)=Range ; **(X-Xmin)/Range**

Normalization is mainly required when the data has any other type of distribution other than Normal/Bell Curve/Gaussian distribution.

1. Standardization

In this type of scaling, we want data which follows Gaussian Distribution/Normal Distribution

And it scales it over a range of 1 with mean being zero and std deviation obviously being 1.

Outlier- a data point that is dissimilar to the other data points. Process to identify outliers are

***Mean Deviation = 1/n∑in (xi – x̄)2***

Correlation is important to consider as it can help us determine if the columns are loosely or closely related to each other.

Correlation is of three types-

Perfect Positive correlation

High Positive correlation

Low correlation

**Linear regression -** There are many algorithms to predict but Linear Regression is one of the most basic and robust of them all.

It basically uses the principle wherein a line is drawn(predicted values) and the distance is calculated based upon the distance between actual value and predicted value.

The difference between the two can tell us the accuracy and error margin of the model and ultimately tell us how good or bad is the mathematical model at delivering results based on the training dataset.

Overfitting- when the model works well on the training set but not on the testing set

Underfitting- when the model works well neither on training nor on testing dataset.

The measures to calculate error for Linear Regression are many-

1. MAE

Mean Absolute Error is the difference between actual value and predicted value and dividing that by the total number of values/observations. Performs better with outliers.

1. MSE

Mean Squared Error is the square of the difference by the number of values/observations and is very sensitive to outliers thus not great if dataset has outliers. The reason behind this is MSE squares the difference of the values and thus even if one value is an outlier the square makes it very high.

Classification- a technique through which we classify a dataset into different classes and sub classes.

There are multiple algorithms to do it for ex-

1. KNN
2. Decision Trees
3. Naïve Bayes
4. Random Forest

While Logisitic Regression works on a dataset that is linearly separable , Decision Trees not only requires lesser data but also works with complex data with multiple classes for classification.

Logistic needs to learn the pattern via a huge amount of data to classify in terms of complex decisions while Decision Tree does it effortlessly even in small datasets just because it is very decision based and works with strings too.

This all was about Supervised Learning where the data is structured and we know what to find out how to find out

Unsupervised Learning is a different beast altogether, the data is neither classified nor labelled so we have to find the structure ourselves.

Clustering

Hierarchical Clustering-It is an approach that which mainly has two types Top-down(Divisive) and Bottom Up(Agglomerative).

We can cluster together sub categories based on similarity(cosine) or distance(Euclidean)

Neural Networks

NNs work well if the data has been normalized ,Therefore always ensure it is either normalized or standardized

Softmax is an activation function that is used on the output nodes majorly used in classification problems which produces normalized values(between 0 and 1) that are in a probability distribution(ex-0.7 for label1 then 0.3 will be for label2)