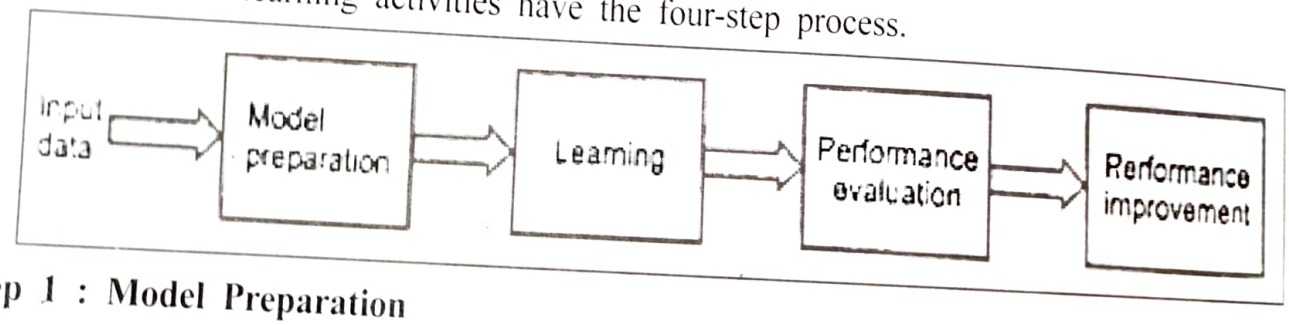
2 Preparing to Model

2.1 Machine Learning Activities

In machine learning, the detailed view of understanding the nature of input data is very important. Using the detailed view of input data, we can select the ML model and apply it on the data.

The machine learning activities have the four-step process.



Step 1 : Model Preparation

In this step, activities like understanding the type, nature and quality of data are involve. Inter-feature relationship between data, data pre-processing and finding any issues in data have been involved through Step 1.

Step 2 : Learning

In this step, activities like data partitioning, model selection and cross-validation are involved.

If it is supervised learning approach then the data is divided into two parts: Training

Data and Test Data in Step-2 and then the appropriate model will be applied.

If it is unsupervised learning approach then the unsupervised learning model will be directly applied to the input data.

Step 3: Performance Evaluation

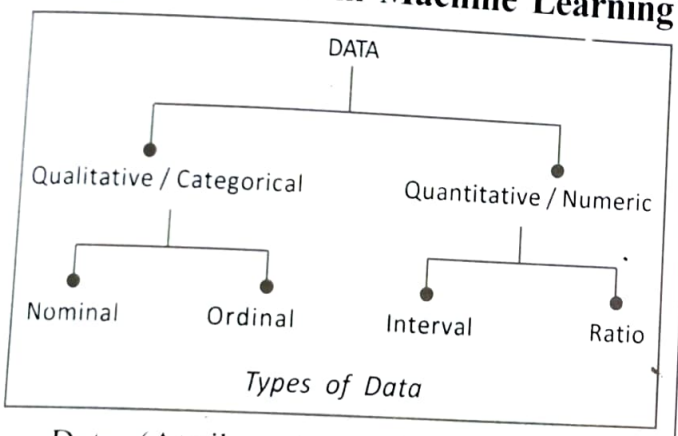
In this step, the performance of the model is evaluated. If it is a classification then the performance will be examined using confusion matrix. Also, performance tade-offs should be visualized using ROC curves.

Evaluation will be done using different criteria like accuracy, F-Score, etc.

Step 4: Performance Improvement

This is the last step of Machine Learning Process. In this step, tuning the model, bagging, boosting activities are involved to improve the performance of the model.

2.2 Types of Data in Machine Learning



Types of Data

Data (Attributes) are broadly classified into Qualitative Data and Quantitative Data.

2.2.1 Qualitative Data :

The data collected on grounds of categorical variables are **qualitative data**. It is also known as categorical data.

Qualitative data are more descriptive and conceptual in nature.

It measures the data on basis of the type of data, collection, or category.

Qualitative data talks about the experience or quality and explains the questions like 'why' and 'how'.

For example, if we consider the performance of employee in terms of Good, Average or Poor then it is called qualitative data.

The qualitative data are subdivided into two parts: Nominal Data and Ordinal Data.

Nominal data is the data which has no numeric value, but a named value.

Example : Nationality: Indian, American, British, etc.

On nominal data, mathematical functions like addition, multiplication or statistical functions like mean, median cannot be applied. Only basic count should be applied. So we can get mode of the nominal data.

Ordinal data is the data which has named value which can be naturally ordered in increasing or decreasing.

Example: Students Satisfaction: Very Happy, Happy, Unhappy

On ordinal data, the ordered data is available so we can perform mode, median and quartiles operations. But, we cannot perform operations like mean.

2.2.2 Quantitative Data :

The data collected on the grounds of the numerical variables are **quantitative data**. It is also known as numeric data.

Quantitative data are more objective and conclusive in nature.

It measures the values and is expressed in numbers. The data collection is based on "how much" is the quantity.

Quantitative data talks about the quantity and explains the questions like 'how much', 'how many'.

For example, if we consider marks of Mid Semester Exam like '30', '25', '12' then it is called quantitative data.

The quantitative data are subdivided into two parts: Interval Data and Ratio Data.

Interval Data is numeric data for which exact difference between values is known.

Such data do not have something called a 'true zero' value.

Example: Celsius Temperature, Date, Time

Ratio Data is numeric data for which exact value can be measured.

Such data have the 'true zero' value, too.

Example: Height, Weight, Age, Marks

2.3 Structure of Data

2.3.1.1 Exploring numerical data Understanding Measure of Central Tendency

The central tendency measure is defined as the number used to represent the center or middle of a set of data values. The three commonly used measures of central tendency are the mean, median, mode.

The purpose of the central tendency is to provide an exact representation of the entire collected data. It is often defined as the single value that is representative of the data.

Mean: In statistics, arithmetic mean is the average of the given set of numbers or observations.

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Where,

A is the arithmetic mean.

n is the number of items or numbers.

x_i is the value of every individual item being averaged

Example: The mean of first 10 natural numbers are

Arithmetic mean = Sum of all values / Total number of values

= (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10) / 10

= 55 / 10

= 5.5

Median: Median is the middle value of the given list of data when arranged in an order. The arrangement of data or observations can be made either in ascending order or descending order.

If the total number of observations given is odd, then the formula to calculate the median is:

$$Median = \left(\frac{n+1}{2}\right)^{th} term$$

where n is the number of observations.

For example, the median of 2, 4, 7, 9, 12 is 7.

If the total number of observation is even, then the median formula is:

$$Median = \frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2}$$

where n is the number of observations

If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values.

So the median of 3, 6, 8, 10 is (6+8)/2 = 7.

Mode: The mode is the value that is repeatedly occurring in a given set.

For example, in the data set 2, 4, 5, 5, 6, 8, the mode of the data set is 5 since it has appeared in the set twice.

When there are two modes in a data set, then the set is called bimodal.

When there are three modes in a data set, then the set is called trimodal.

When there are four or more modes in a data set, then the set is called multimodal

Understanding Measure of Dispersion

Measures of dispersion are statistical measures that describe how spread out the data is around the central tendency measures. They provide information on the variability or diversity of the dataset.

The four commonly used measures of dispersion are the range, variance, standard deviation and Interquartile range (IQR).

Range: The range is the difference between the largest and smallest values in a dataset. It provides a simple measure of how spread out the data is.

For example, suppose we have the following dataset:

5, 6, 8, 10, 12, 14, 20

The largest value is 20 and the smallest value is 5.

Therefore, the range is:

Range = 20 - 5 = 15

So, the range of this dataset is 15.

Variance : Variance is a measure of how data points vary from the mean. It is denoted as 'ó²'.

Properties of Variance:

- It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.
- Variance always has squared units.

The variance formula is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Here,

s^2 = Sample variance

n = Number of observations in sample

x_i = i^{th} observation in the sample

$\bar{x}$ = Sample mean

Standard Deviation:

Standard deviation is the measure of the distribution of statistical data. It is denoted as 'ó'.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Here,

s = Population standard deviation

Properties of Standard Deviation

- It describes the square root of the mean of the squares of all values in a data set and is also called the root-mean-square deviation.
- The smallest value of the standard deviation is 0 since it cannot be negative.
- When the data values of a group are similar, then the standard deviation will be very low or close to zero. But when the data values vary with each other, then the standard variation is high or far from zero.

Example : If a die is rolled, then find the variance and standard deviation of the possibilities.

When a die is rolled, the possible outcome will be 6. So the sample space, n = 6 and the data set = { 1;2;3;4;5;6}.

To find the variance, first, we need to calculate the mean of the data set.

Mean, $\bar{x} = (1+2+3+4+5+6)/6 = 3.5$

We can put the value of data and mean in the formula to get;

$S^2 = \Sigma\ (xi - \bar{x})^2/n$

$S^2 = \frac{1}{6}\ (6.25+2.25+0.25+0.25+2.25+6.25)$

$S^2 = 2.917$

Now, the standard deviation,

$S = \sqrt{2.917} = 1.708$

Interquartile range (IQR) :

The IQR is the range between the first quartile (25th percentile) and third quartile (75th percentile) of a dataset. It provides a measure of the spread of the middle 50% of the data. For example, suppose we have the following dataset:

4, 5, 6, 8, 10, 12, 14, 18, 20

To calculate the IQR, we first need to find the median, which is:

Median = (10 + 12) / 2 = 11

Next, we need to find the first and third quartiles. The first quartile (Q1) is the median of the lower half of the dataset, and the third quartile (Q3) is the median of the upper half of the dataset. In this case, we have:

Lower half :

4, 5, 6, 8 Upper half: 12, 14, 18, 20

Q1 = (5 + 6) / 2 = 5.5 Q3 = (14 + 18) / 2 = 16

The IQR is then :

IQR = Q3 - Q1 = 16 - 5.5 = 10.5

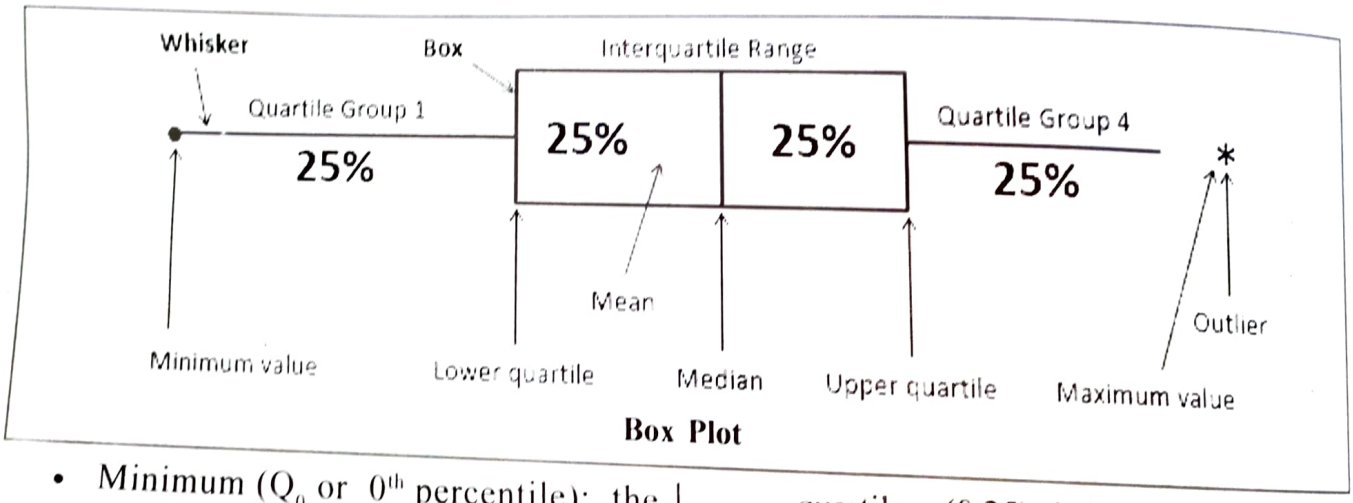
So, the IQR of this dataset is 10.5. This tells us that the middle 50% of the data is spread out over a range of 10.5.

2.3.2 Plotting Numerical data

One of the best ways to analyze any process is to plot the data on a graph or chart. Numerical data can be plotted using box plots or histograms.

Box Plots : A box plot is a graph that shows the frequency of numeric data values. It is mainly used to explore data as well as to present the data in an easy and understandable manner.

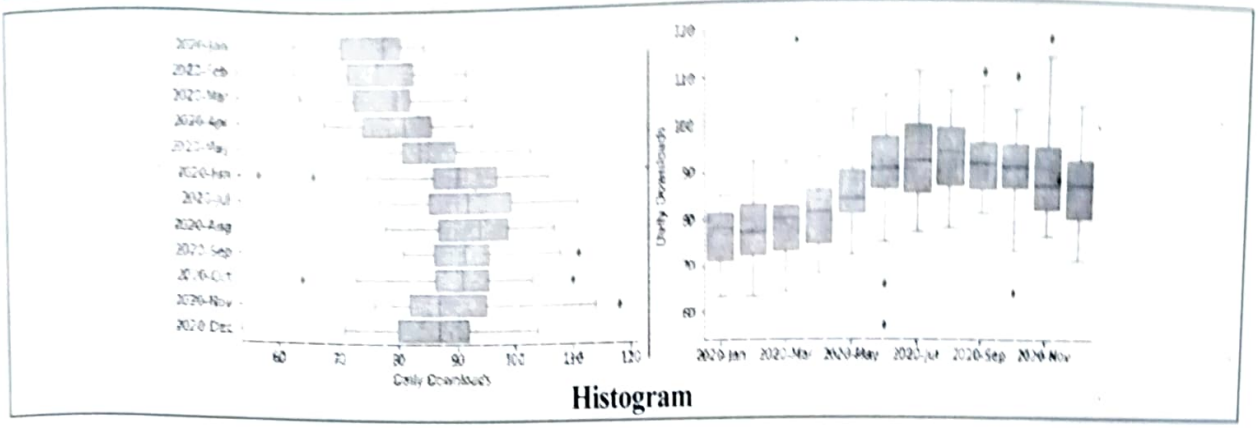
A box plot is a standardized way of displaying the dataset based on the five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles.



Box Plot

- Minimum (Q_0 or 0^{th} percentile): the lowest data point in the data set excluding any outliers
- Maximum (Q_4 or 100^{th} percentile): the highest data point in the data set excluding any outliers
- Median (Q_2 or 50^{th} percentile): the middle value in the data set
- First quartile (Q_1 or 25^{th} percentile): also known as the lower quartile $q_n(0.25)$, it is the median of the lower half of the dataset.
- Third quartile (Q_3 or 75^{th} percentile): also known as the upper quartile $q_n(0.75)$, it is the median of the upper half of the dataset.

We can also find Inter-quartile range (IQR) : the distance between the upper and lower quartiles. IQR = Q_3 - Q_1



Histogram

Histogram :

A **histogram** is a graph which shows the frequency of continuous data values. It is a two-dimensional figure.

Histograms are widely used in statistics, scientific research, economics, in social and human sciences, and it is especially important in process improvement and operational excellence.

Histograms are mainly used to explore data as well as to present the data in an easy and understandable manner. They are often used as the first step to determine the underlying probability distribution of a data set or sample.

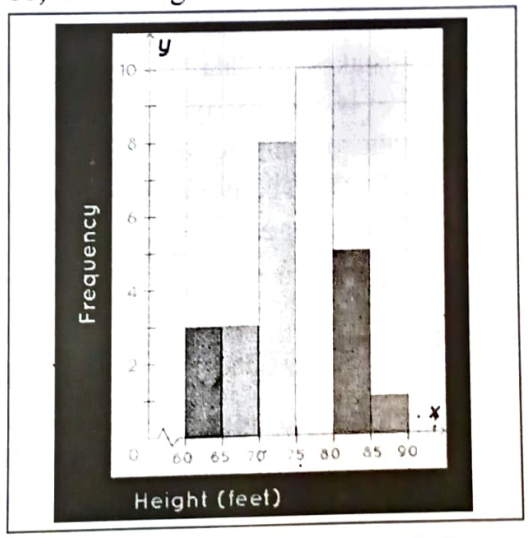
Follow the below mentioned steps to create histogram manually:

1. Collect the data set and prepare it for the analysis..
2. Draw a horizontal line and divide it into equal intervals or bins (12 intervals for example). The total width should be equal to the range of the data.
3. Draw bars above each bin to represent the frequency of the data values within each interval. The bars should be adjacent with no gaps between them.
4. Indicate the mean of the data and other important information such as the standard deviation and the specification limits.

Example : The following table shows the number of trees in a garden with different height range.

Height Range (ft)	Number of Trees (Frequency)
60-75	3
66-70	3
71-75	8
76-80	10
81-85	5
86-90	1

So, the histogram for the stated data is:



2.4 Data Quality and Remediation

2.4.1 Data Quality

Data which have the right quality helps to achieve better performance in Machine Learning Model.

Data quality problems are :

1. Certain data elements which have no value or missing values
2. Data elements which have irrelevant values known as outliers

There are mainly two factors which create data quality problems:

1. Incorrect sample set selection

   The data may not show normal quality due to incorrect selection of sample set. For example, in sales transactions, if we collect the data set of Diwali period then the prediction should be differing from our daily routine.

2. Errors in data collection

   It leads towards outliers and missing values. It may happen that a person is responsible for the wrongly collected data. It may also happen that the data is not recorded at all. So, in record it shows missing values.

2.4.2 Data Remediation

Data remediation is about correcting errors and mistakes in data to eliminate data-quality issues. It is the process of cleansing, organizing, and migrating data so that it is properly protected and serves its intended purpose.

Mainly five steps are involved in data remediation:

Step 1 : Understand the current state of your data.

Step 2 : Understand the nature of your data via classification.

Step 3 : Apply data governance. Identify outliers, if any.

Step 4 : Discuss various data remediation options and select appropriate one.

Step 5 : Implement the selected data remediation strategy.

Outliers and handling outliers :

An outlier is a data point that is remarkably distinct from other data points in a dataset.

There are three main strategies to handle outliers:

1. Remove outliers

   If there are few outliers in record then simply remove it.

2. Imputation (Replace the value)

   Impute the value with mean, median or mode.

3. Capping

   For values that lie outside the $1.5|x|$ IQR limits, we can replace those observations with the following conditions:

   a. If below the lower limit then replace with the value of 5^{th} percentile

   b. If above the upper limit the replace with the value of 95^{th} percentile

Handling missing values :

One or more elements may have missing values in multiple records of data set.

Following strategies have been used to handle missing values:

1. Ignore the tuple
2. Eliminate records having a missing value of data elements
3. Filling missing values using mean, median or mode.
4. Use a global constant to fill in the missing value:
5. Use the most probable value to fill in the missing value.

2.5 Data Pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model.

Data Pre-processing Methods/ Techniques :

- **Data Cleaning** routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- **Data Integration** which combines data from multiple sources into a coherent data store, as in data warehousing.
- **Data Transformation**, the data are transformed or consolidated into forms appropriate for mining
- **Data Reduction** obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

2.5.1 Dimensionality reduction

Dimensionality reduction is the data reduction technique where encoding mechanisms are used to reduce the data set size.

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.

The number of input variables or features for a dataset is referred to as its dimensionality. Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset. Avoiding overfitting is a major motivation for performing dimensionality reduction.

Example :

Dimensional reduction can be discussed through a simple e-mail classification problem, where we need to classify whether the e-mail is spam or not.

This can involve a large number of features, such as whether or not the e-mail has a generic title, the content of the e-mail, whether the e-mail uses a template, etc.

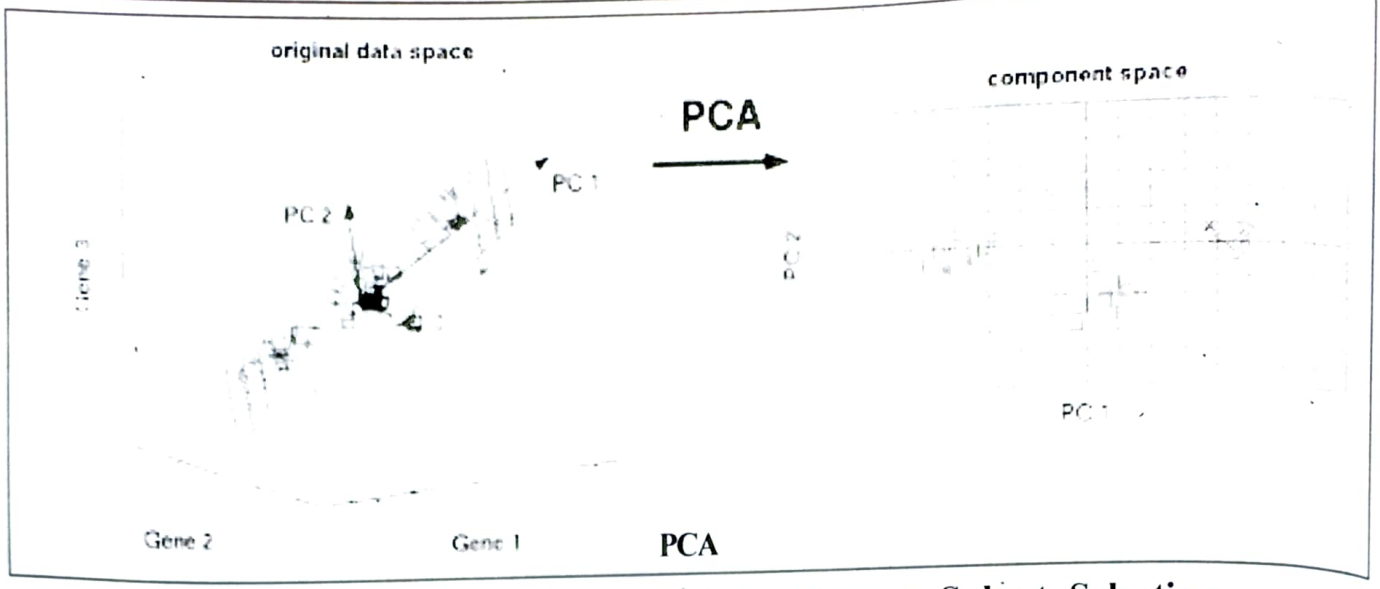
Dimensionality reduction methods :

1. Principal Component Analysis (PCA) :

Frequently used for dimensionality reduction in continuous data, PCA rotates and projects data along the direction of increasing variance. The features with the maximum variance are the principal components.

PCA is a statistical technique used to reduce the dimensionality of a large dataset.

- PCA works by identifying patterns in the data and then creating new variables that capture as much of the variation in the data as possible. These new variables, known as principal components, are linear combinations of the original variables in the dataset.
- The first principal component captures the most variation in the data; the second captures the second most, and so on. The number of principal components created is equal to the number of original variables in the dataset.

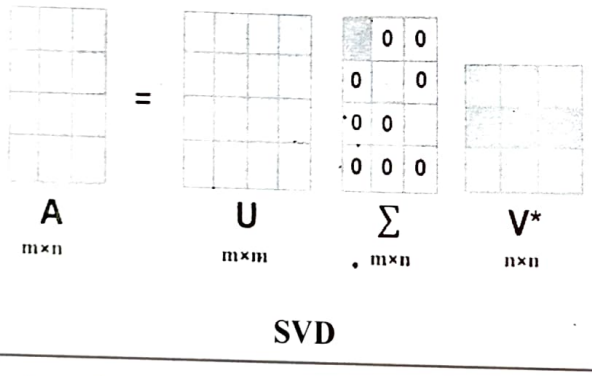


PCA

PCA can be used for a variety of purposes, including data visualization, feature selection, and data compression.

2. Singular Value Decomposition (SVD):

The main intuition behind Singular Value Decomposition (SVD) is, that Matrix *A* transforms a set of orthogonal vectors **(*v*)** to another set of orthogonal vectors **(*u*)** with a scaling factor of **ó**. So, **ó** is called the Singular Value corresponding to the respective singular vectors ***u*** and ***v***.



SVD

It helps reduce datasets containing a large number of values. Furthermore, this method is also helpful to generate significant solutions for fewer values.

2.5.2 Feature Subset Selection

Feature Subset Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them.

It helps in cutting down the noise in our data and reducing the size of our input data.

Feature Subset Selection is the most critical pre-processing activity in any machine learning process.

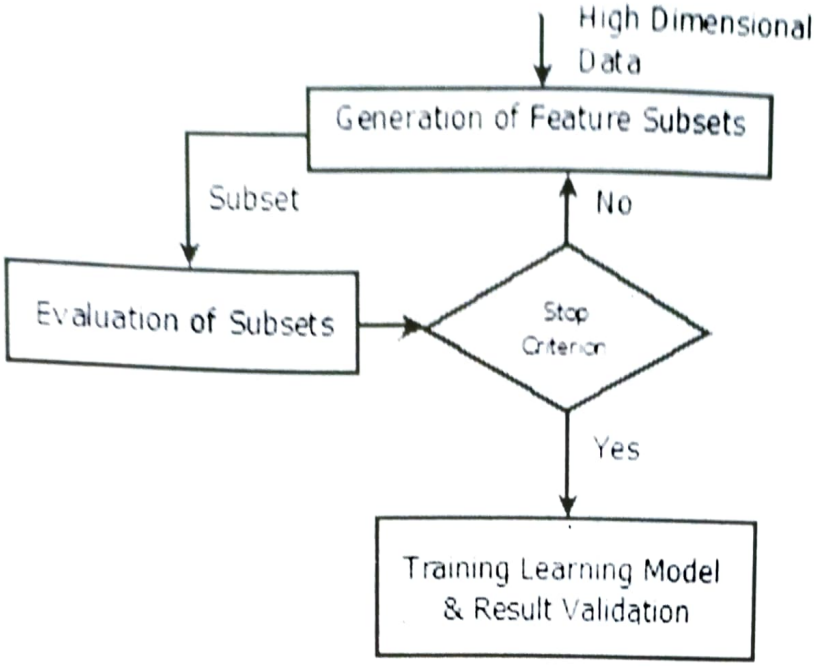
Four basic steps for feature subset selection are as follow:

Step 1: Generation of all possible subset

Step 2: Evaluation of generated subset,

Step 3: Stopping criterion

Step 4: Validation of result



Flow chart for feature subset selection

Generation of all possible subset is a brute force method that generates candidate feature subsets for estimation based on a particular search procedure.

Each generated candidate feature subset is estimated and compared with the preceding most excellent one according to a certain estimating criterion.

If the criteria score of new subset come out to be better, it replaces the previous best subset. The process of generation of subset and estimation is recurring until a given stopping criterion is not satisfied.

Then, the most excellent subset usually needs to be tested by prior knowledge or many separate tests on synthetic datasets and/or real-world datasets.

Feature subset selection models are of two types:

1. Supervised Models: Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model. Filter method, wrapper method and intrinsic method are three methods used in supervised model.
2. Unsupervised Models: Unsupervised feature selection refers to the methods which not need the output labels class for feature selection. We use them for unlabelled data.

Question Bank

■ Short-Answer Questions (3 or 4 marks) :

1. State the activities involved in model preparation stage.
2. Give the difference between qualitative data and quantitative data.
3. Give the difference between histogram and box plot with example.
4. State different measures of central tendency.

5. Find mean, median, mode and standard deviation for the following data:
   1, 1, 2, 4, 5, 5, 6, 7, 7, 7, 8, 9, 10
6. What is IQR? How it is meseaured?
7. Define outliers. How can we take care of outliers in data?
8. State various strategies to handle missing values.
9. Explain PCA in brief.
10. What are the factors which lead to the data quality issues?

■ Long-Answer Questions (7 marks):

1. Describe machine learning activities in detail.
2. Explain data types in machine learning with example.
3. Define data pre-processing. Explain various methods used in data pre-processing.
4. Write a short note on dimensionality reduction.
5. Write a short note on feature subset selection.
6. Define following terms: Data pre-processing, Data remediation, Outliers, Imputation, Standard Deviation, Ratio Data, Ordinal Data
7. Explain steps to create box plot with suitable example.
8. Explain steps to create histogram with suitable example. Also, state the difference between bar-chart and histogram.

■ Check your knowledge :

1. Which of the following types of data is non-numerical and can be classified into distinct categories?

   a. Categorical data b. Ordinal data
   c. Interval data d. All of the above

2. A technique which is used to clean, transform and prepare raw data for analysis is called ____.

   a. Data preparing **b. Data pre-processing**
   c. Data Processing d. Data Mining

3. Which of the following is a data reduction technique ?

   a. Dimensionality reduction b. Data encoding
   c. Dimensionality Imputation d. Data Visualization

4. What is the purpose of feature selection ?

   a. To increase the size of the dataset
   b. To remove irrelevant or redundant features from the dataset
   c. To add new features to the dataset
   d. To convert numerical data into categorical data

5. Which measure of central tendency is most appropriate for nominal or categorical data?

   a) Mean b) Median **c) Mode** d) Range

■