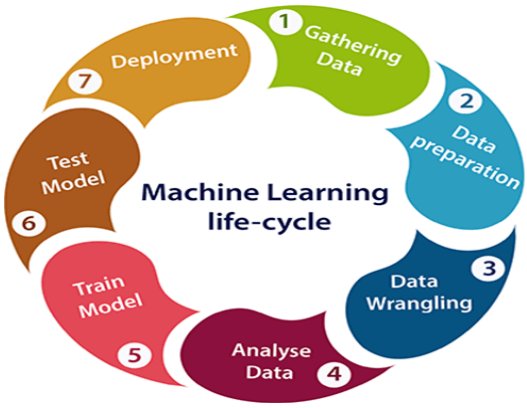
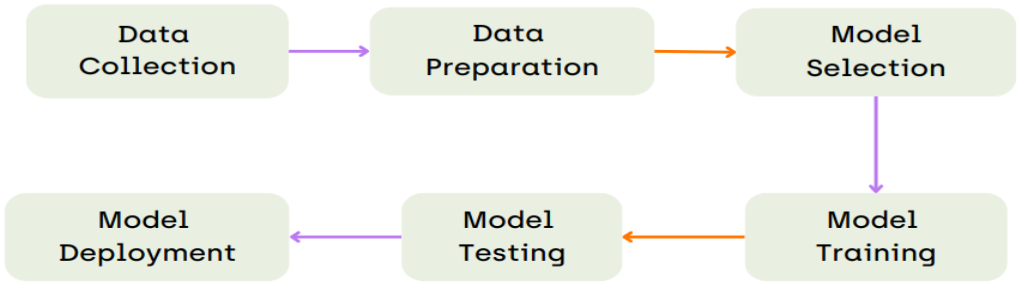
UNIT-3
PREPARING TO MODEL AND PREPROCESSING

Machine Learning (ML) is used to learn from data and make predictions or decisions. The key activities involved in machine learning are:



OR



A. <u>Preparing To Model:</u>

Preparation activities are:

- Understand the type of data.
- Explore the data.
- Explore the relationships.
- Find potential issues in data.
- Apply pre-processing steps.
- Start learning:
  - Input data is divided into training data and test data.
  - Consider different models or the learning algorithms for selection.
  - Train the model based on the training data.

B. <u>Learning activities:</u>

This involves feeding the model with labeled data (data with known outcomes).

1. Data Partition:

Splitting a dataset into training, validation, and testing. Dividing the dataset into three main subsets:

1. **Training set :**

It is used to teach the machine, it has input samples and their matching target values.

2. **Validation set :**

It is used to check how well the model is doing during training , it helps in adjusting the model.

3. Test set :

It is used to give a fair assessment of how well Trained model performs , it checks accuracy, precision, recall or F1 score.

2. Selection of the model

The most important factors to select Model are:

1.Predictive models:

There are models known as "Predictive models" that aim to make guesses or predictions based on information. These models examine data and search for relationships. They focus on learning and determining how to make precise predictions. Ex: Predicting whether a transaction is fraud

2.Descriptive Models:

There are models known as "descriptive models" or "unsupervised learning models" that assist us and learning from a large amount of data. Their primary focus is on uncovering patterns. Ex: Grouping of commodities in an inventory

C. PERFORMANCE EVALUATION

The specific performance metrics used for evaluation.

1. Cross validation:

- It is a technique used to assess the performance and general ability of a model.
- The validation set is used to measure how well the model performs on new, unseen data.
- The data in the validation set might have some specific characteristics that either improve or worsen the model's performance.
- To overcome this, cross-validation is used.
- The dataset is divided into equally sized subsets, "folds ". The model is trained and evaluated multiple times.
- Every fold gets a chance to be the validation set.
- It includes k-fold cross-validation, where k= number of folds.

2. Confusion Matrix:

A matrix containing correct and incorrect predictions in the form of TPs,FPs,FNs and TNs is known as confusion matrix .The win/loss prediction of cricket match has two classes.

- True Positive(TP): The model predicted win and the team won
- False Positive (FP): The model predicted win and the team lost
- False Negative(FN): The model predicted loss and the team won
- True Negative(TN): The model predicted loss and the team lost

The model accuracy is found by:

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

$$\text{accuracy} = \frac{80 + 10}{80 + 10 + 3 + 7}$$

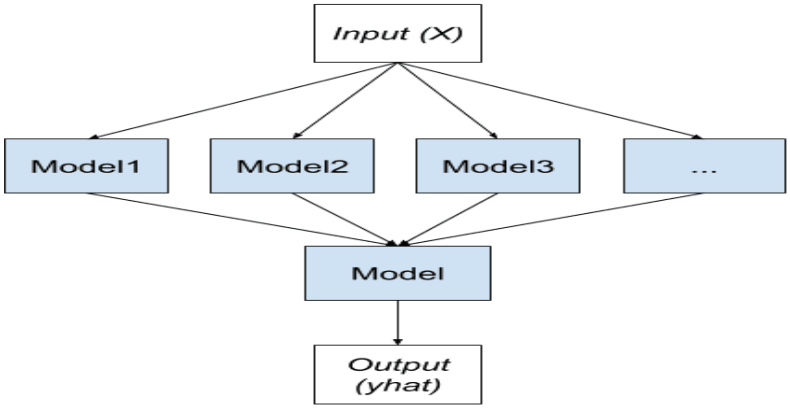
$$= 0.9 \approx 90\%$$

Prediction	Actual win	Actual loss
Win	80 (TP)	3 (FP)
Loss	7 (FN)	10 (TN)

Confusion Matrix

D. PERFORMANCE IMPROVEMENT

1. **Turning model:** Model parameter tuning involves adjusting the options used for fitting the model.
   Ex: Widely used k-Nearest Neighbours(kNN) classification model.
2. **Ensemble model:** To enhance performance, combine multiple models. This approach is known as ensemble modeling. By combining weaker learners, ensemble methods create stronger and more robust models.



Ensemble the model

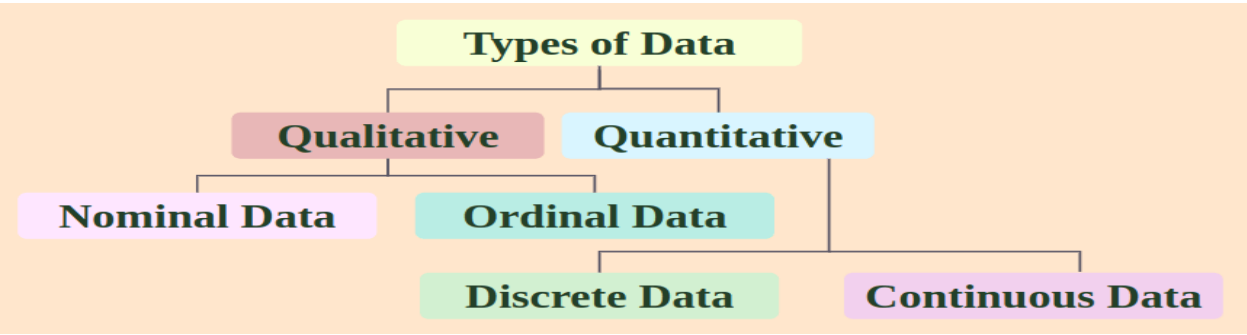
3.2 TYPES OF DATA

A dataset consist of multiple attributes. Each attribute can be referred as features, variables, dimensions or fields. Various types of data in machine learning problems are given in figure.

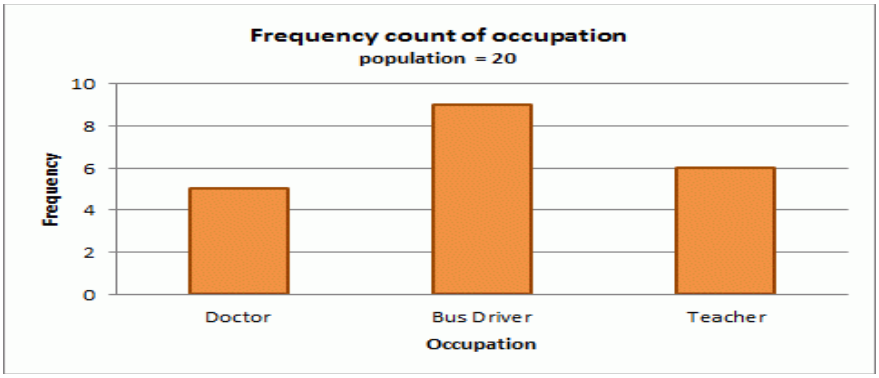
[Introduction:
Quantitative and Qualitative are the two sides of the coin named "Data in Statistics" but as many people are familiar with quantitative data (i.e., numerical data of various sorts), qualitative data is often less understood. Understanding the qualitative data is essential for researchers, analysts, decision-makers, or anyone who wants to gain deep insights into people's behaviors, attitudes, and experiences.]

Types of Data in Statistics
The grouping of data can be based on the quantitative and qualitative aspects of the gathered information, so data can be classified into the following types:

1. **Qualitative Data**
2. **Quantitative Data**



1. **What is Qualitative Data? (Categorical)**
   - **"Qualitative Data uses variables to represent labels or characteristics of entities or objects".**
   - **Ex**: movie genres or travel methods.
   - The labels cannot be represented in **numerical form**, and their numerical values may not hold any significance.
   - It is also known as **categorical** data as it is expressed through indicators and deals with perceptions.
   - It **cannot** be averaged and aggregated.
   - It **can** be grouped based on categories. Ex: the color of hair can be categorized into three main colors like black, brown or blonde.
   - Examples in the real world are:
     - **Interview transcripts:** Data collected from survey forms after the interviews can provide rich qualitative data that describes the opinions, attitudes, and experiences of participants.
     - **Observation notes:** Observing a behavior, recorded data is an example of qualitative data as it can tell us about the characteristics, context.
     - **Open-ended survey responses:** In a survey, there are some open-ended questions sometimes to know about the participant's experiences, perceptions, and opinions on a given topic. This data is also an example of qualitative data.
     - **Ex:**



Types of Qualitative Data

Qualitative data can be further categorized into the following types:

A. Nominal Data

B. Ordinal Data

A. Nominal Data:

a. Nominal data is represented using names, not with numbers.

b. It is used for assigning named values to attributes. It cannot be quantified.

c. Ex1: Blood Group: A, B, O, AB etc. Ex2: Movie or series genres: horror, sci-fi, and comedy etc.

B. Ordinal Data:

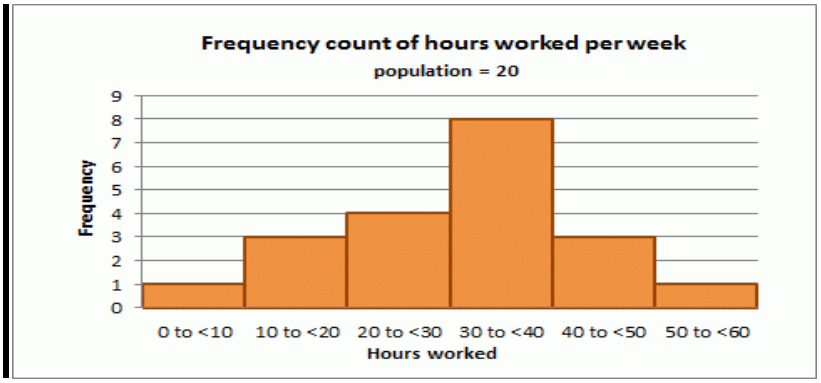
a. Ordinal qualitative data uses a certain scale or measure to group data.

b. The data is ordered or measured, but the scale used to represent the data may not be standard or specific.

c. This type of data includes numerical values so we can arrange values in sequence or compare to analyze.

d. It can be visually represented using bar graphs.

2. **What is Quantitative Data? (Numerical)**
   - **"Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis."**
   - **Ex** the price of a phone, height or weight of a person, time, temperature.
   - These kinds of data are also known as Numerical data. It answers the questions like "how much," "how many," and "how often."
   - It can be used for statistical manipulation.
   - These data can be represented on a wide variety of graphs and charts, like bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.
   - Ex:



Types of Quantitative Data

Quantitative data can be further categorized into the following types:

A. Discrete Data

B. Continuous Data

A. Discrete Data:

a. This will have whole numbers. That means, integer values and not decimals.

b. The most important point to note is that the values will not change with time.

c. **Ex: The total number of students in a class.**

B. Continuous Data

a. It can take any numerical value and has infinite possibilities. Point to note is that the values of a continuous variable can change with time.

b. Ex: Height/ Weight of a person, mileage of a car, age of a person in a 10years dataset, GDP/ population of a country in a 5years dataset.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

Continuous data can be further classified depending on whether it's **interval data** or **ratio data**.

Interval vs. ratio data

Interval Data

- Interval data can be measured, where there is an equal distance between each point on the scale.
- **Interval data has no true or meaningful zero value**.
- Ex: Temperature: a temperature of zero degrees does not mean that there is "no temperature"—it just means that it's extremely cold!

Ratio Data

- Ratio data is the same as interval data in terms of equally spaced points on a scale.
- Ratio data **does have a true zero**.
- Ex: Weight in grams would be classified as ratio data; the difference between 20 grams and 21 grams is equal to the difference between 8 and 9 grams, **and** if something weighs zero grams, it truly weighs nothing.

3.2.2 Data quality and remediation

➢ Data Quality:

- Data quality is the measure of how well suited a data set is to serve its specific purpose.
- Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness, and timeliness.
- Success of predictive models heavily relies on the quality of the data used.
- Two types of problems are: missing values and outliers.
- Some data quality issues are:
    - Incorrect sample set selection: Data used for analysis and prediction does not accurately represent the desired population or time period. Ex: sales data collection during festival seasons are differ from normal day sales.
    - Errors in data collection: Errors occurs during data collection process. When data is collected manually , mistakes can happen. This errors record incorrect values. Ex. Recording 20.67 as 206.7

➢ Data Remediation:

- Data remediation is the process of cleansing, organizing and migrating data so that it's properly protected and best serves its intended purpose.
- There is a misconception that data remediation simply means deleting business data that is no longer needed.
- It's important to remember that the key word "remediation" derives from the word "remedy," which is to correct a mistake.
- The data remediation process typically involves replacing, modifying, cleansing or deleting any "dirty" data.

A. Handling outliers: It is the dataitem/object that deviates(differs) from rest of the objects (normal objects).

- If outliers are detected and it is determined that they need to be addressed. It is important to note if they are valid and occur naturally.
- **Remove outliers**: If they are few then simple approach is to remove.
- **Imputation**: Other way is to impute the value with mean or median or mode.
- **Capping**: We can cap them by replacing those observations.
- If number or outliers are more, then treat them separately.

B. Handling missing values: Within a dataset it is possible that one or more data elements are absent in multiple records. It is possible if data collector collect wrong data, or ask irrelevant questions for data. Many strategies are there to address missing values:

- **Eliminate records having a missing value**: If the missing values remain within an acceptable range then remove them. EX: Auto MPG dataset having 6 records missing from 400 records, so if we have 394 records remaining, which are considerably high so we can work on those 394 records. If records are less in size then it will impact on model performance.
- **Imputing missing values(Alternative values)**: **Imputation** means assign values to data elements that have missing values. Use approach to impute the mean, mode or median as assigned value, when dealing with quantitative attributes. For qualitative attributes, missing values are replaced with mode of all remaining values of attribute.

Ex: Imputed by Mean:
Mean imputation (or mean substitution) replaces missing values of a particular variable with the mean of non-missing cases of that variable.

What is the missing value when you know mean? Dataset Set: {3,5,X,9} Average: 6 #(3+5+9)/3 = 6 Now, Data Set: {3,5,6,9}

3.2.3 Data Pre-Processing

- Data preprocessing is a process of **preparing** the raw data and making it **suitable** for a machine learning model. It is the first and crucial step while creating a machine learning model.
- Always we don't get clean and formatted data so while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

□ Dimensionality reduction:

- *"It is a way of converting the **higher dimensions** dataset into lesser dimensions dataset ensuring that it provides similar information."*
- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.
- Use: For obtaining a better fit predictive model while solving the classification and regression problems
- Ex: high-dimensional data, such as speech recognition, signal processing, bioinformatics, etc. It can also be used for data visualization, noise reduction, cluster analysis, etc.

□ Feature subset selection:

- It is a way of selecting the **optimal** features from the input dataset.
- Three methods are used for the feature selection:
  - **Filters Methods**: only the relevant features is taken. Ex: Correlation
  - **Wrappers Methods**: only the relevant features is taken but with machine learning model, performance decides whether to add those features or remove Ex: Forward, Backward, Bi-directional
  - **Embedded Methods**: check the different training iterations of the machine learning model and evaluate the importance of each feature. Ex: LASSO

Difference between Nominal and Ordinal Data

Feature	Nominal Data	Ordinal Data
Definition	Data that is not ranked or ordered in any way.	Data that is ranked or ordered in a specific way,
Examples	Gender, Color, Marital Status, Nationality	Education Level, Income Range, Satisfaction Level
Arithmetic operations	Cannot perform any arithmetic operations.	Can perform basic arithmetic operations such as addition and subtraction, but not multiplication or division.
Measures of Central Tendency	Mode	Mode, Median
Measures of Dispersion	None	Range, Interquartile Range

Difference between Discrete and Continuous Data

Discrete Data	Continuous Data
They are **countable and finite.**	They are **measurable**.
They are **whole** numbers or integers.	They are in the form of **fractions** or decimal.
They are represented mainly by **Bar** graphs	They are represented in the form of a **Histogram**
The values **cannot** be divided into subdivisions into smaller pieces	The values can be divided into subdivisions into smaller pieces
They have **spaces** between the values.	They are in the **form** of a continuous sequence.
Examples: Total students in a class, number of days in a week, size of a shoe, etc.	**Example:** Temperature of room, Weight of a person, Length of an object, etc.

Difference between **Qualitative Data** *AND* **Quantitative Data**

Qualitative Data	**Quantitative Data**
1. It uses methods like interviews, participant observation, and focus on a grouping to gain collective information.	1. It uses methods as questionnaires, surveys, and structural observations to gain collective information.
2. Data format used in it is textual. Datasheets are contained of audio or video recordings and notes.	2. Data format used in it is numerical. Datasheets are obtained in the form of numerical values.
3. It talks about the experience or quality and explains the questions like 'why' and 'how'.	3. It talks about the quantity and explains the questions like 'how much', 'how many' .
4. The data is analyzed by grouping it into different categories.	4. The data is analyzed by statistical methods.
5. They are subjective and can be further open for interpretation.	5. They are fixed and universal.