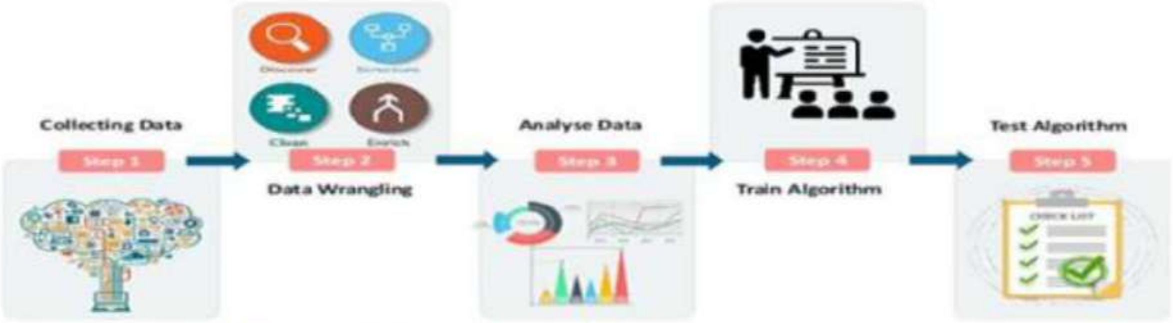
Unit –3:Preparing to Model and Preprocessing

3.1.1 Machine Learning activities



MACHINE LEARNING ACTIVITIES [11].

1. Preparing to Model

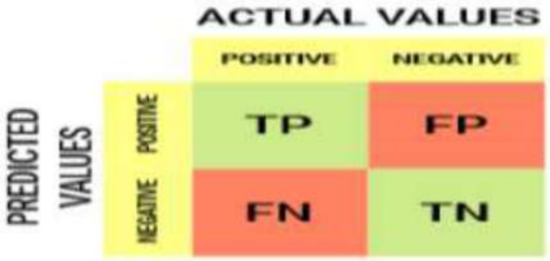
- This stage involves data collection, data preprocessing, and feature engineering. It's essential to gather relevant data, clean it, handle missing values, normalize, and transform it into a suitable format for training. Feature engineering involves selecting or creating relevant features from the data that can improve the model's performance.

2. Learning: Data Partition-k-fold, cross validation, Model Selection

- To avoid overfitting and assess a model's performance, the dataset is split into multiple subsets (folds) using k-fold cross-validation. The model is trained on k-1 folds and validated on the remaining fold, repeating this process k times to obtain a reliable performance estimation. Model selection involves trying out different machine learning algorithms or architectures and choosing the best one based on cross-validation performance.
- Cross-validation is a statistical method used to estimate the skill of machine learning models.
- It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.
- In this tutorial, you will discover a gentle introduction to the k-fold cross-validation procedure for estimating the skill of machine learning models.
- After completing this tutorial, you will know:
- That k-fold cross validation is a procedure used to estimate the skill of the model on new data.
- There are common tactics that you can use to select the value of k for your dataset.
- There are commonly used variations on cross-validation such as stratified and repeated that are available in scikit-learn.

3. Performance Evaluation: confusion matrix

- A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the total number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.
- For a binary classification problem, we would have a 2 x 2 matrix, as shown below, with 4 values:



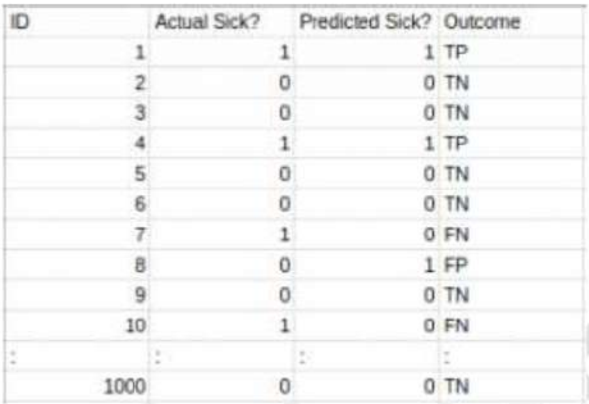
		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

- Let's decipher the matrix:
  a) The target variable has two values: Positive or Negative
  b) The columns represent the actual values of the target variable
  c) The rows represent the predicted values of the target variable
- Important Terms in a Confusion Matrix
- **True Positive (TP)**
- The predicted value matches the actual value, or the predicted class matches the actual class.
- The actual value was positive, and the model predicted a positive value.
- **True Negative (TN)**
- The predicted value matches the actual value, or the predicted class matches the actual class.
- The actual value was negative, and the model predicted a negative value.
- **False Positive (FP) – Type I Error**
- The predicted value was falsely predicted.
- The actual value was negative, but the model predicted a positive value.
- Also known as the type I error.
- **False Negative (FN) – Type II Error**
- The predicted value was falsely predicted.
- The actual value was positive, but the model predicted a negative value.
- Also known as the type II error.
- Why Do We Need a Confusion Matrix?
- Before we answer this question, let's think about a hypothetical classification problem.
- Let's say you want to predict how many people are infected with a contagious virus in times before they show the symptoms and isolate them from the healthy population (ringing any bells, yet?). The two values for our target variable would be Sick and Not Sick.
- Now, you must be wondering why we need a confusion matrix when we have our all-weather friend – Accuracy. Well, let's see where classification accuracy falters.

- Our dataset is an example of an **imbalanced dataset**. There are 947 data points for the negative class and 3 data points for the positive class. This is how we'll calculate the accuracy:

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

- Let's see how our model performed:



ID	Actual Sick?	Predicted Sick?	Outcome
1	1	1	TP
2	0	0	TN
3	0	0	TN
4	1	1	TP
5	0	0	TN
6	0	0	TN
7	1	0	FN
8	0	1	FP
9	0	0	TN
10	1	0	FN
:	:	:	:
1000	0	0	TN

- The total outcome values are:
- TP = 30, TN = 930, FP = 30, FN = 10
- So, the accuracy of our model turns out to be:

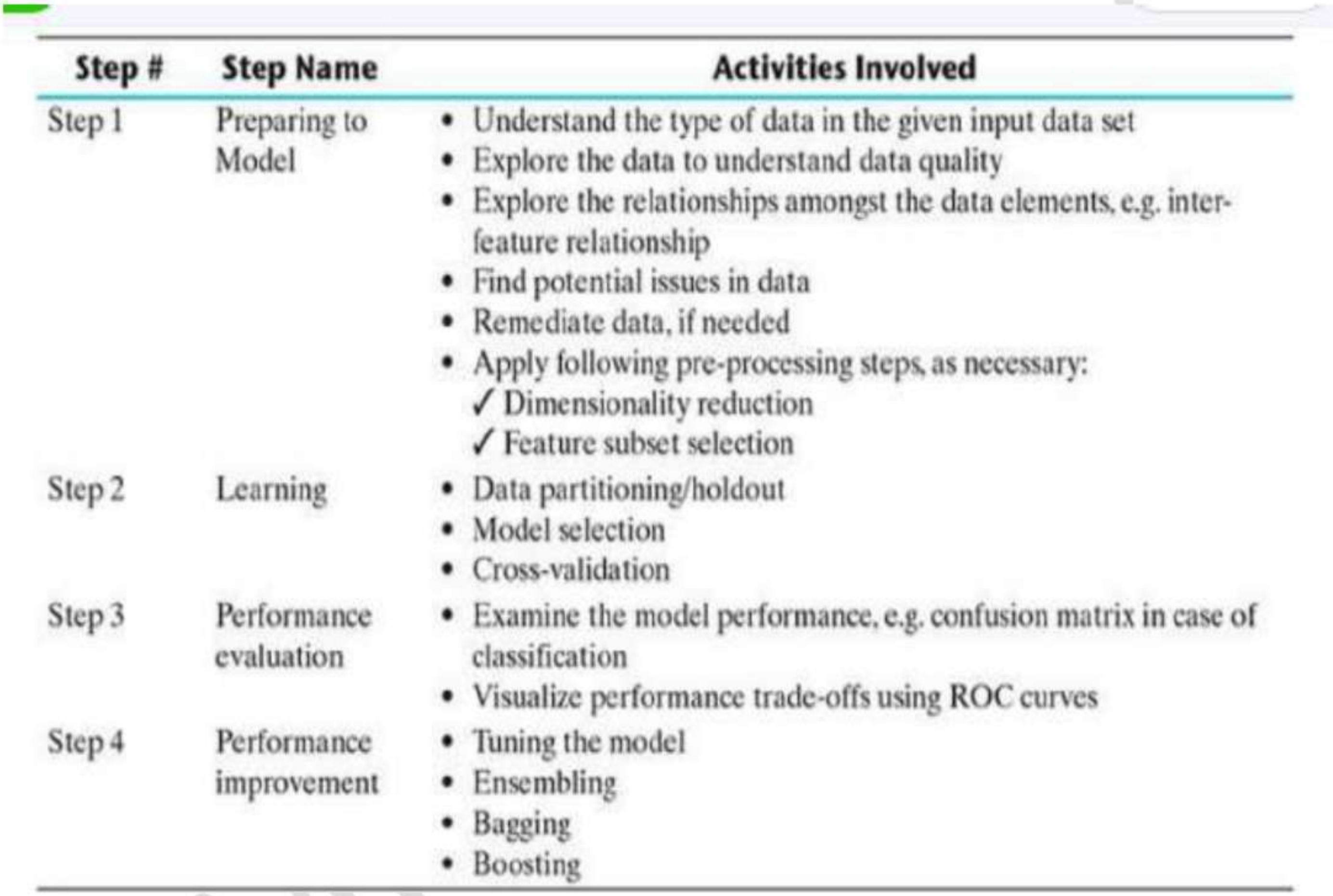
$$Accuracy = \frac{30 + 930}{30 + 30 + 930 + 10} = 0.96$$

- 96%
- The confusion matrix is a performance evaluation tool that summarizes the results of a classification model. It presents the number of true positive, true negative, false positive, and false negative predictions, which helps calculate metrics like accuracy, precision, recall, and F1-score.

4. Performance Improvement: Ensemble

- Ensemble methods combine multiple models to improve predictive performance. Common ensemble techniques include:
- Bagging: Building multiple instances of the same model on different subsets of the data and aggregating their predictions (e.g., Random Forest).
- Boosting: Sequentially training weak learners, where each learner focuses on the mistakes of its predecessor, boosting their collective performance (e.g., AdaBoost, Gradient Boosting Machines - GBM).
- Stacking: Combining multiple models by using their predictions as input for a meta-model, which makes the final prediction (e.g., Stacked Generalization).
- Ensemble methods often lead to more accurate and robust models compared to using a single model.

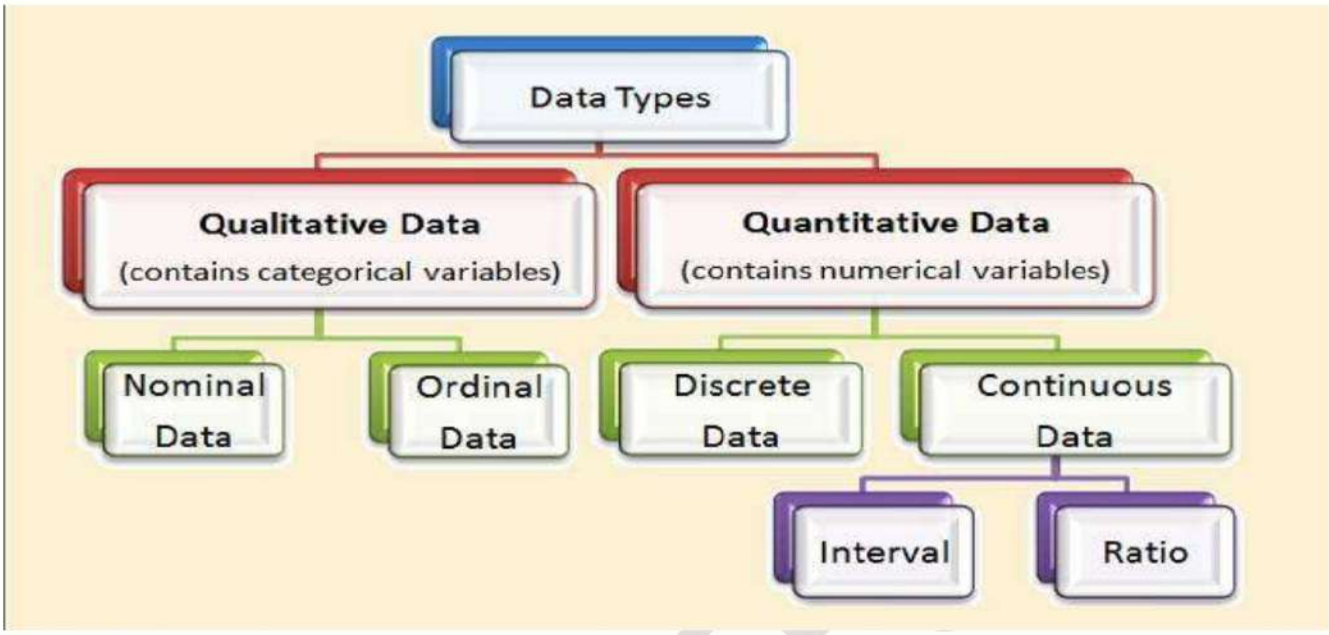
- ➢ Ensemble learning helps improve machine learning results by combining several models. Ensembles can give you a boost in accuracy on your dataset. In this post we will discover how to create some of the most powerful types of ensembles in Python using scikit-learn.
- ➢ These machine learning activities form a structured workflow for developing and improving machine learning models. However, it's important to note that the exact steps and techniques employed can vary depending on the specific problem, dataset characteristics, and available resources.



Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

3.2.1 Types of Data

- ➢ Data is the new oil." Today data is everywhere in every field. Whether you are a data scientist, marketer, businessman, data analyst, researcher, or you are in any other profession, you need to play or experiment with raw or structured data. This data is so important for us that it becomes important to handle and store it properly, without any error. While working on these data, it is important to know the types of data to process them and get the right results.
- ➢ **There are two types of data: Qualitative and Quantitative data, which are further classified into**:



1. Qualitative/Categorical Data: Nominal, Ordinal

- Qualitative data represents qualities or characteristics and is typically non-numeric.
- Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.
- Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.
- It can be divided into two subtypes:

a. Nominal Data: Nominal data represents categories that have no intrinsic order or ranking. Each category is distinct and unrelated to others. Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

- Examples of nominal data include gender (e.g., male, female), colors (e.g., red, blue, green), and countries (e.g., USA, Canada, India).

b. Ordinal Data: Ordinal data also represents categories, but these categories have a specific order or ranking. However, the differences between the categories are not necessarily uniform. Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

- Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered “in-between” qualitative and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

- Examples of ordinal data include survey responses with options like "strongly disagree," "disagree," "neutral," "agree," and "strongly agree."

2. Quantitative/Numeric Data: Interval, Ratio

- Quantitative data represents quantities and can be measured or counted.
- Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like “how much,” “how many,” and “how often.” For example, the price of a phone, the computer’s ram, the height or weight of a person, etc., falls under quantitative data.
- Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.
- It can be further divided into two subtypes:
- a. Interval Data: Interval data has numeric values where the difference between any two values is meaningful, but there is no true zero point. For example, temperature measured in Celsius or Fahrenheit is interval data because differences in temperature are meaningful, but zero degrees does not indicate the absence of temperature.
- b. Ratio Data: Ratio data is similar to interval data but has a true zero point, indicating the absence of the quantity being measured. Ratios between values are meaningful. Examples of ratio data include height, weight, time, and distance. For instance, if someone's weight is 0 kg, it means they have no weight.
- Understanding the types of data is essential for choosing appropriate statistical analysis methods, data visualization techniques, and machine learning algorithms for a given dataset. Different data types require different treatment and consideration during data analysis and modeling processes.

3.2.2 Data quality and remediation

- Data quality is a critical aspect of any data-driven project, as the accuracy and reliability of the data directly impact the performance and validity of the analyses and models. Two common challenges in data quality are handling outliers and missing values.
- Benefits of Data Remediation
- Although it may seem like a daunting task, the lasting benefits of data remediation to the organization far outweigh the effort. Because of its many benefits, organizations should include ongoing data remediation as part of their business activities. The main benefits include:
- **Reduced Costs** – Our organization will minimize the risk of financial loss through fines, lawsuits, and reputational damage that come with a data breach. Also, data remediation will reduce your overall volume of data, and therefore data storage costs.

- **Reduced Risk –** Sensitive data that was once exposed and leaving your organization vulnerable to risk of breaches and leaks will now be stored securely or safely deleted.
- **Compliant With Privacy Laws and Regulations –**Your organization can rest assured that it's fully compliant with the ever changing privacy laws and regulations, including GDPR, CCPA, CPPA, PCI DSS, HIPAA and more.
- **Streamlined Operations and Efficiency** – Your team can work more efficiently and effectively by having easy access to reliable data. This will also help your team make faster data-based decisions.
- **Data Minimization –** Your organization will safely delete data that has no legal or business justification of storing. You'll only be left with data that's legally required or provides value to your organization.

- Let's discuss each of them and their potential remediation strategies:

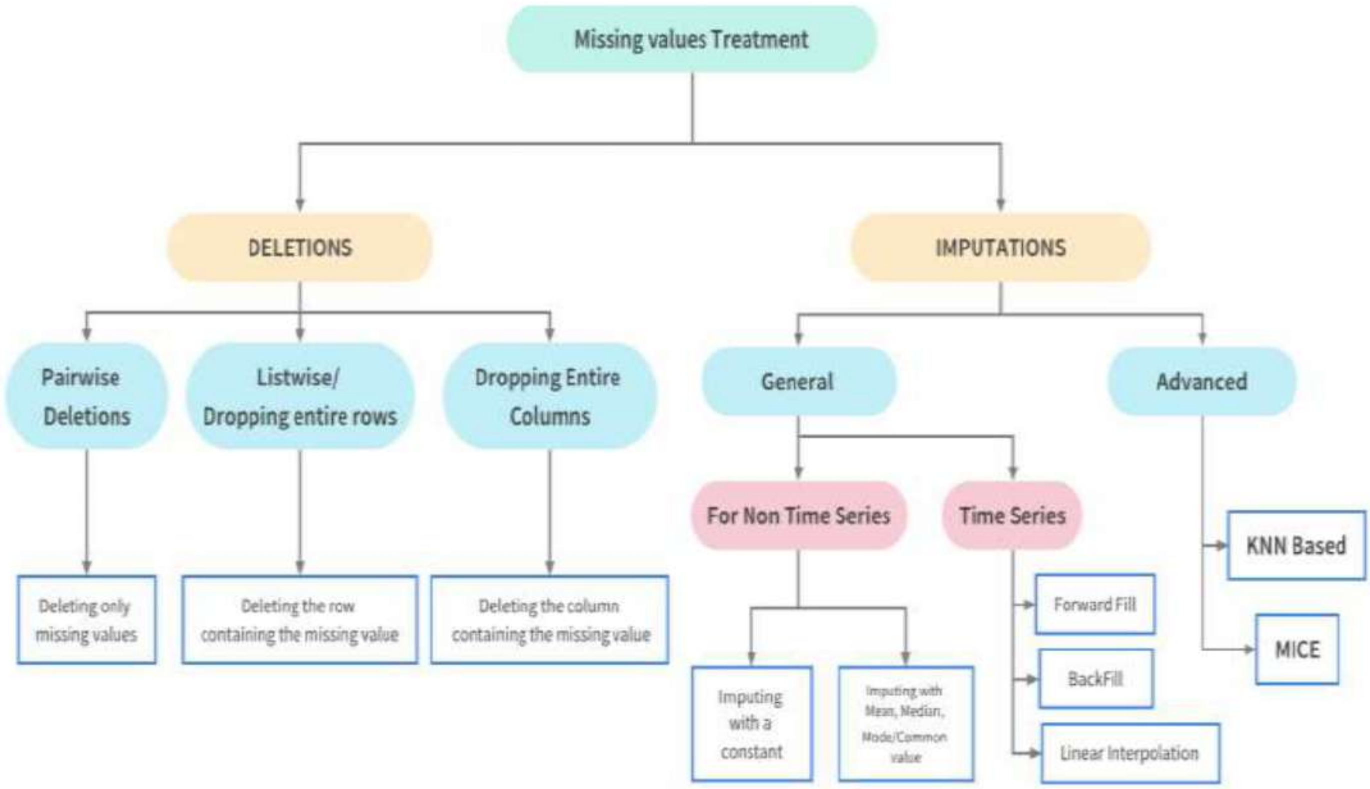
1. Handling outliers

- Outliers are data points that significantly deviate from the rest of the data and may negatively impact statistical analyses and machine learning models. Outliers can occur due to various reasons, such as measurement errors, data entry mistakes, or genuine extreme values.
- Remediation strategies for handling outliers include:
- a. Removing Outliers: In some cases, outliers can be safely removed from the dataset if they are due to data entry errors or do not reflect the underlying patterns in the data. However, care should be taken not to remove outliers without a valid reason, as they might carry valuable information.
- b. Transforming Data: Applying data transformations (e.g., log transformation) can reduce the impact of extreme values while preserving the overall data distribution.
- c. Capping or Flooring: Limiting extreme values by capping the maximum and minimum values within a reasonable range can be a suitable approach.
- d. Robust Statistics: Using robust statistical measures like median and interquartile range (IQR) instead of mean and standard deviation can be less sensitive to outliers.

2. Handling missing values

- Missing values are gaps or null entries in a dataset, which can arise due to various reasons such as data collection errors, non-responses, or data corruption.
- Remediation strategies for handling missing values include:
- a. Removal: If the percentage of missing values in a column or row is substantial and doesn't significantly affect the overall dataset, the rows or columns with missing values can be removed. However, this should be done with caution to avoid losing valuable information.
- b. Imputation: Imputation involves filling in missing values with estimated or predicted values. Common imputation methods include mean, median, mode imputation, as well as using machine learning techniques for more advanced imputation.
- c. Creating a Missing Indicator: In some cases, creating a binary indicator variable that denotes whether a value is missing or not can provide additional information to the model.

- d. Time-Series Interpolation: For time-series data, interpolation methods can be used to estimate missing values based on neighboring time points.
- e. Advanced Imputation: Utilizing advanced imputation techniques such as k-nearest neighbors (KNN) imputation or multiple imputation can handle more complex missing value patterns.
- Handling outliers and missing values should be done thoughtfully, considering the specific context of the data and the impact on the analysis or model. Proper data quality assessment and remediation can significantly improve the reliability and effectiveness of data-driven projects.

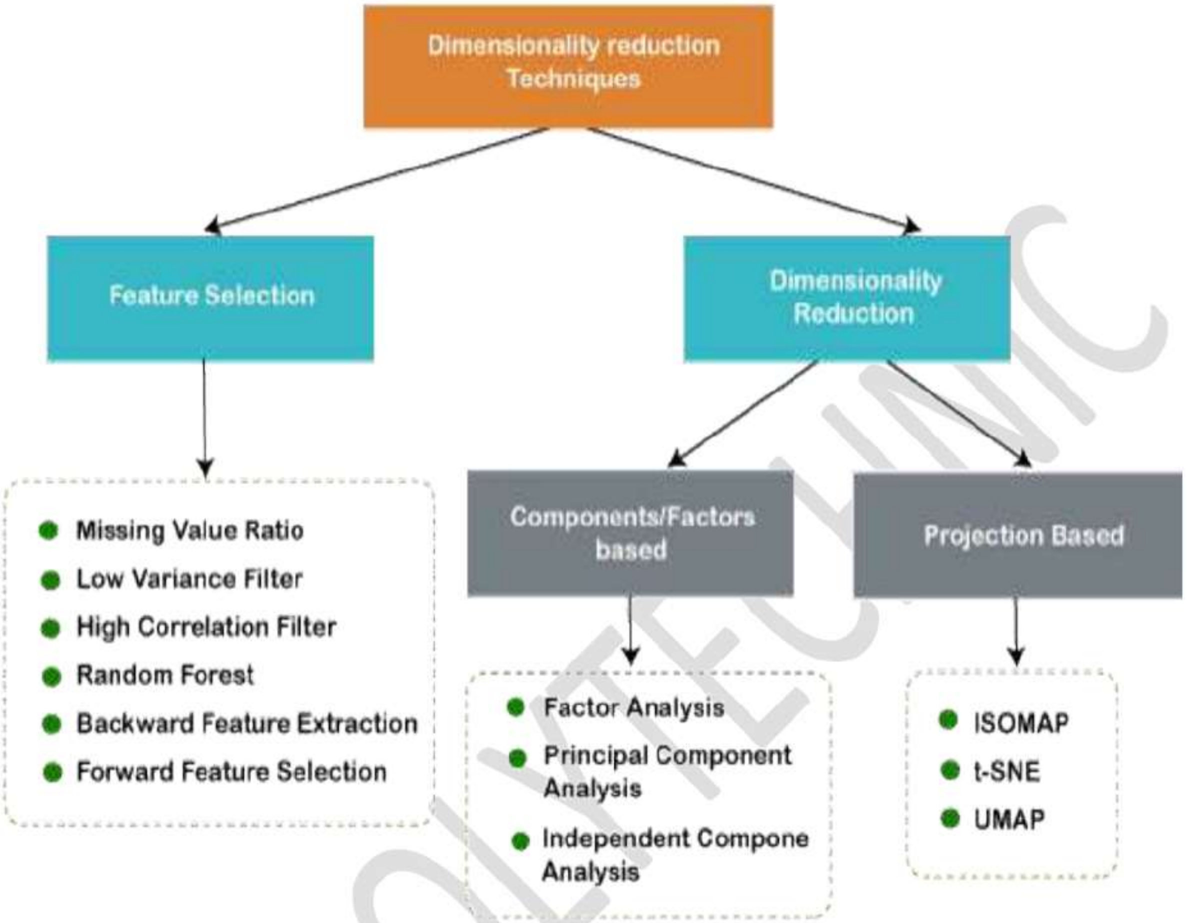


3.2.3 Data Pre-Processing

- Data pre-processing is a crucial step in machine learning that involves transforming raw data into a format suitable for training models. It includes various techniques to improve data quality and reduce the computational complexity of the models. Two important techniques in data pre-processing are dimensionality reduction and feature subset selection:

1. Dimensionality Reduction

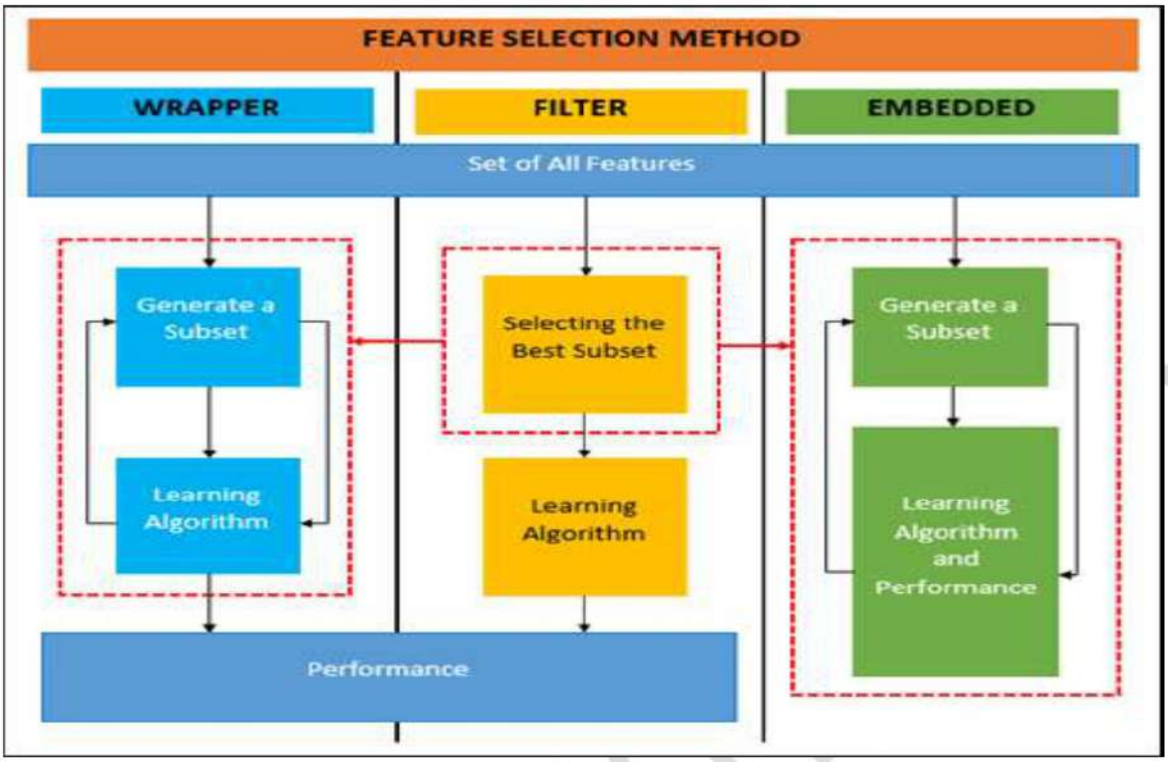
- Dimensionality reduction is the process of reducing the number of features or variables in a dataset while preserving its essential information. High-dimensional datasets can lead to increased computational complexity and the risk of overfitting. Dimensionality reduction techniques aim to represent the data in a lower-dimensional space while minimizing the loss of relevant information.



- Common dimensionality reduction techniques include:
- a. Principal Component Analysis (PCA): PCA transforms the original features into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they explain, allowing for a reduced number of principal components to represent the data.
- b. t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is a technique used primarily for visualization. It reduces high-dimensional data into a two- or three-dimensional space, making it easier to visualize clusters and patterns.
- c. Singular Value Decomposition (SVD): SVD is a linear algebra technique used for matrix factorization. It can be applied for dimensionality reduction in certain cases.

2. Feature subset selection: Filter,Wrapper, Hybrid, Embedded

- Feature Subset Selection: Feature subset selection involves selecting a subset of the most relevant features from the original feature set. The goal is to retain the most informative features while eliminating irrelevant or redundant ones, thus simplifying the model and improving its performance.



- There are various methods for feature subset selection, categorized into four main types:
- a. Filter Methods: These methods rank features based on statistical measures or information-theoretic metrics. Features are selected or removed based on their individual properties, regardless of the target variable or the model. Common filter methods include Information Gain, Chi-Square, and Correlation.
- b. Wrapper Methods: Wrapper methods evaluate feature subsets by training and testing a model on different combinations of features. This approach involves more computational cost since it searches for the best feature subset through an iterative process. Examples include Recursive Feature Elimination (RFE) and Forward Selection.
- c. Embedded Methods: Embedded methods combine feature selection with the model training process. Model-specific techniques, such as Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge Regression, incorporate feature selection directly into the model fitting.
- d. Hybrid Methods: Hybrid methods combine multiple feature selection techniques to achieve better performance. For example, Recursive Feature Elimination with Cross-Validation (RFECV) is a hybrid approach that combines wrapper and filter methods.
- Choosing the appropriate dimensionality reduction and feature subset selection techniques depends on the nature of the data, the model's requirements, and the computational resources available. These techniques help create more efficient and accurate machine learning models while avoiding issues like the curse of dimensionality