Parul® University

Data Visualization and Data Analytics

Prof. Khushbu Chauhan, Assistant Professor
Information Technology

CHAPTER 3

Data Preparation

Outlines

- Dealing with missing values
- Data Cleaning using various methods
- Principal Component Analysis
- Feature Selection methods

❖ What is a Missing Value?

- Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.
- Next diagram is a sample of the missing data from the Titanic dataset. You can see the columns ‘Age’ and ‘Cabin’ have some missing values.

DEALING WITH MISSING VALUES

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

❖How is Missing Value Represented In The Dataset?

- In the dataset, blank shows the missing values.
- In Pandas, usually, missing values are represented by NaN.
- It stands for Not a Number.

	StudentID	Gender	DOB	Race	Ethnicity	Class	Weight	Height	Enrollment_Date	State_Residency
1	5	1	08/15/1991	2	1	1	226	70	08/15/2012	In state
2	9	1	11/01/1991	3	1	1	144	71	08/15/2012	
3	35	1	10/29/1990	1	.	1	.	.	08/15/2012	Out of state
4	70	2	04/06/1994	1	2	1	175	63	08/15/2012	In state
5	44	1	01/31/1991	1	2	2	170	77	.	In state
6	51	1	.	1	1	2	177	71	08/15/2011	Out of state
7	85	2	09/26/1991	1	2	2	141	.	08/15/2010	Out of state
8	19	1	05/25/1991	1	2	3	184	.	08/15/2010	In state
9	40	1	10/29/1990	1	2	3	170	67	08/15/2010	In state
10	43	1	02/03/1990	2	2	3	.	.	08/15/2010	Out of state
11	24	1	09/04/1993	1	2	4	167	73	08/15/2007	In state
12	39	1	08/12/1993	3	2	4	150	73	08/15/2006	Out of state
13	45	1	03/09/1994	1	2	4	161	71	08/15/2007	In state
14	79	2	02/16/1992	1	2	4	143	62	08/15/2008	In state
15	89	.	09/11/1993	1	2	4	128	64	08/15/2009	Out of state

Missing numeric values are a period.

Missing character values are blank.

	Height	Weight	Country	Place	Number of days	Some column
0	12.0	35.0	India	Bengaluru	1.0	NaN
1	NaN	36.0	US	New York	2.0	NaN
2	13.0	32.0	UK	London	NaN	NaN
3	15.0	NaN	France	Paris	4.0	NaN
4	16.0	39.0	US	California	5.0	12.0
5	NaN	NaN	NaN	Mumbai	NaN	NaN
6	NaN	NaN	NaN	NaN	6.0	NaN

Why Is Data Missing From The Dataset

- There can be multiple reasons why certain values are missing from the data.
- Reasons for the missing data from the dataset affect the approach to handling missing data. So it's necessary to understand why the data could be missing.

Some of the reasons are listed below

- Past data might get corrupted due to improper maintenance.
- Observations are not recorded for certain fields due to some reasons.
- There might be a failure in recording the values due to human error.

❖ Types Of Missing Value

1. Missing Completely At Random (MCAR)

- In MCAR, the probability of data being missing is the same for all the observations.
- In this case, there is no relationship between the missing data and any other values observed or unobserved (the data which is not recorded) within the given dataset.
- That is, missing values are completely independent of other data. There is no pattern.
- In the case of MCAR, the data could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values.
- For Example, suppose in a library there are some overdue books. Some values of overdue books in the computer system are missing. The reason might be a human error like the librarian forgot to type in the values. So, the missing values of overdue books are not related to any other variable/data in the system.

2. Missing At Random (MAR)

- Missing at random (MAR) means that the reason for missing values can be explained by variables on which you have complete information as there is some relationship between the missing data and other values/data.
- In this case, the data is not missing for all the observations. It is missing only within sub-samples of the data and there is some pattern in the missing values.
- For example, if you check the survey data, you may find that all the people have answered their 'Gender' but 'Age' values are mostly missing for people who have answered their 'Gender' as 'female'. (The reason being most of the females don't want to reveal their age.)

DEALING WITH MISSING VALUES

- So, the probability of data being missing depends only on the observed data.
- In this case, the variables ‘Gender’ and ‘Age’ are related and the reason for missing values of the ‘Age’ variable can be explained by the ‘Gender’ variable but you can not predict the missing value itself.
- Suppose a poll is taken for overdue books of a library. Gender and the number of overdue books are asked in the poll. Assume that most of the females answer the poll and men are less likely to answer. So why the data is missing can be explained by another factor that is gender.
- In this case, the statistical analysis might result in bias.

3.Missing Not At Random (MNAR)

- Missing values depend on the unobserved data.
- If there is some structure/pattern in missing data and other observed data **can not explain** it, then it is Missing Not At Random (MNAR).
- If the missing data does not fall under the MCAR or MAR then it can be categorized as MNAR.
- It can happen due to the reluctance of people in providing the required information. A specific group of people may not answer some questions in a survey.
- For example, suppose the name and the number of overdue books are asked in the poll for a library. So most of the people having no overdue books are likely to answer the poll. People having more overdue books are less likely to answer

the poll.

- So in this case, the missing value of the number of overdue books depends on the people who have more books overdue.
- Another example, people having less income may refuse to share that information in a survey.
- In the case of MNAR as well the statistical analysis might result in bias.

❖Why Do We Need To Care About Handling Missing Value?

- It is important to handle the missing values appropriately.
- Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
- You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.
- Missing data can lead to a lack of precision in the statistical analysis.

❖How To Handle Missing Value?

→7 ways to handle missing values in the dataset:

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Algorithms that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library

What is Data Cleaning

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Methods of Data Cleaning

- There are many data cleaning methods through which the data should be run. The methods are described below:

Methods of Data Cleaning

1. **Ignore the tuples:** This method is not very feasible, as it only comes to use when the tuple has several attributes is has missing values.

2. **Fill the missing value:** This approach is also not very effective or feasible. Moreover, it can be a time-consuming method. In the approach, one has to fill in the missing value. This is usually done manually, but it can also be done by attribute mean or using the most probable value.

3. **Binning method:** This approach is very simple to understand. The smoothing of sorted data is done using the values around it. The data is then divided into several segments of equal size. After that, the different methods are executed to complete the task.

Methods of Data Cleaning

4. Regression: The data is made smooth with the help of using the regression function. The regression can be linear or multiple. Linear regression has only one independent variable, and multiple regressions have more than one independent variable.

5. Clustering: This method mainly operates on the group. Clustering groups the data in a cluster. Then, the outliers are detected with the help of clustering. Next, the similar values are then arranged into a "group" or a "cluster".

Process of Data Cleaning

The following steps show the process of data cleaning in data mining.

1. Monitoring the errors: Keep a note of suitability where the most mistakes arise. It will make it easier to determine and stabilize false or corrupt information. Information is especially necessary while integrating another possible alternative with established management software.

2. Standardize the mining process: Standardize the point of insertion to assist and reduce the chances of duplicity.

3. Validate data accuracy: Analyse and invest in data tools to clean the record in real-time. Tools used Artificial Intelligence to better examine for correctness.

Process of Data Cleaning

4. Scrub for duplicate data: Determine duplicates to save time when analyzing data. Frequently attempted the same data can be avoided by analyzing and investing in separate data erasing tools that can analyze rough data in quantity and automate the operation.

5. Research on data: Before this activity, our data must be standardized, validated, and scrubbed for duplicates. There are many third-party sources, and these Approved & authorized party sources can capture information directly from our databases. They help us to clean and compile the data to ensure completeness, accuracy, and reliability for business decision-making.

6. Communicate with the team: Keeping the group in the loop will assist in developing and strengthening the client and sending more targeted data to prospective customers.

Reference from :- https://www.javatpoint.com/data-cleaning-in-data-mining

❖What is data cleaning?

• Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

DATA CLEANING USING VARIOUS METHODS

❖Benefits of Data Cleaning

- Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

. Removal of errors when multiple sources of data are at play.

. Fewer errors make for happier clients and less-frustrated employees.

. Ability to map the different functions and what your data is intended to do.

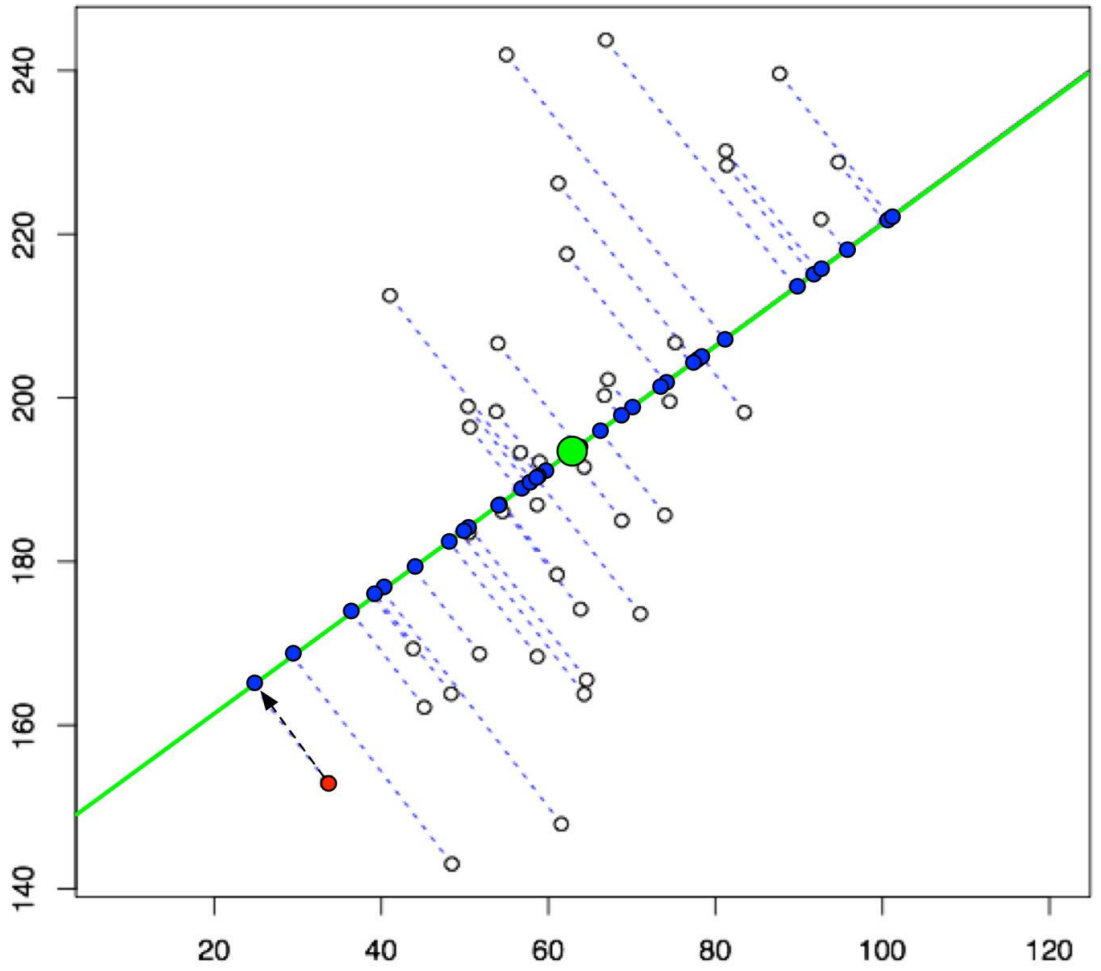
. Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.

. Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

❖What Is Principal Component Analysis?

- Principal Components Analysis, also known as PCA, is a technique commonly used for reducing the dimensionality of data while preserving as much as possible of the information contained in the original data.
- PCA achieves this goal by projecting data onto a lower-dimensional subspace that retains most of the variance among the data points.
- What is dimensionality reduction, and what is a subspace? Let's illustrate this with an example.
- If you have data in a 2-dimensional space, you could project all the data points onto a line using PCA.

Principal Component Analysis



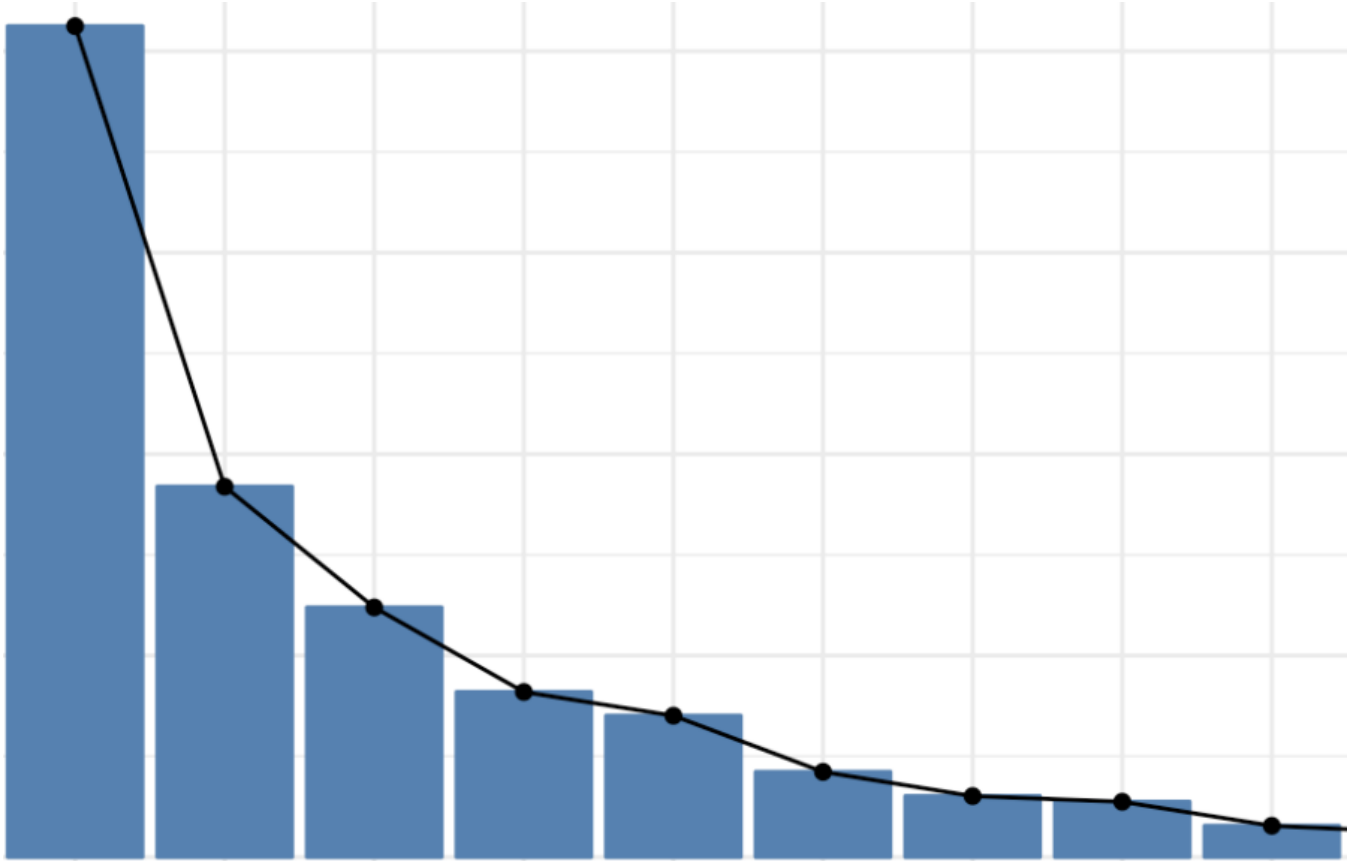
Principal Component Analysis

- As shown in the figure, You have essentially reduced the dimensionality of your data from 2D to 1D. The 1D space (your line) is a subspace of the 2D coordinate system.
- The green line has been constructed using mathematical optimization to maximize the variance between the data points as much as possible along that line.
- We call this line our first principal component. Naturally, the points on the line are still closer to each other than in the original 2D space because you are losing a dimension to distinguish them.
- But in many cases, the simplification achieved by dimensionality reduction outweighs the loss in information, and the loss can be partly or fully reconstructed.

Principal Component Analysis

- You are reducing the dimensionality from 3D (the real world) to a 2D representation. You are losing some explicit 3D information, such as the distance between a person in the front and another person further in the back.
- However, you will have a pretty decent idea of how far these two persons are apart in reality because the person in the back will appear smaller than the person in the front. So the 3D information is not completely lost but sort of encoded in the 2D image

Principal Component Analysis



As you can see, the first principal component explains vastly more than the following ones. This allows us to project even highly dimensional data down to relatively low-dimensional subspaces

- **How does Principal Components Analysis Work?**

- Principal Components Analysis achieves dimensionality reduction through the following steps.

1. Standardize the data

- The variables that make up your dataset will often have different units and different means.
- This can cause issues such as producing extremely large numbers during the calculation.
- To make the process more efficient, it is good practice to center the data at mean zero and make it unit-free.

- You achieve this by subtracting the current mean from the data and dividing by the standard deviation

$$z = \frac{value - mean}{standard\ deviation}$$

2. Calculate the Covariance Matrix:-

- Principal components analysis attempts to capture most of the information in a dataset by identifying the principal components that maximize the variance between observations

Principal Component Analysis

- The covariance matrix is a symmetric matrix with rows and columns equal to the number of dimensions in the data. It tells us how the features or variables diverge from each other by calculating the covariance between the pairwise means.

$$\begin{bmatrix} cov(x_1, x_1) & \dots & cov(x_1, x_n) \\ \dots & & \\ cov(x_n, x_1) & \dots & cov(x_n, x_n) \end{bmatrix}$$

3. **Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix**

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the *principal components* of the data.
- Eigenvectors are linearly independent vectors that do not change direction when a matrix transformation is applied. Eigenvalues are scalars that indicate the magnitude of the Eigenvector.
- The Eigenvectors of the covariance matrix point in the direction of the largest variance. The larger the Eigenvalue, the more of the variance is explained.

- In other words, the Eigenvector with the largest Eigenvalue corresponds to the first principal component, which explains most of the variance.
- The Eigenvector with the second-largest Eigenvalue corresponds to the second principal component, etc.

4. Reduce Dimensionality:-

- As stated previously, the principal components are efficient feature combinations that ensure that the information explained does not overlap between features.
- Eliminating information redundancy already helps in reducing dimensionality. But since the percentage of the overall variance explained declines with every new principal component.

Principal Component Analysis

- we can reduce dimensionality further by eliminating the least important principal components
- At this stage, we have to decide how many principal components are sufficient and how much information loss we can tolerate.
- Lastly, we need to project the data from our original feature space down to the reduced space spanned by our principal components.

- **FEATURE SELECTION METHOD**
- A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.
- Feature selection is the process of reducing the number of input variables when developing a predictive model.
- Feature selection is a way of reducing the input variable for the model by using only relevant data in order to reduce overfitting in the model.

FEATURE SELECTION METHOD

- Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics
- and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective,
- although the choice of statistical measures depends on the data type of both the input and output variables

FEATURE SELECTION METHOD

- Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

FEATURE SELECTION METHOD

- Need for Feature Selection
- We collect a huge amount of data to train our model and help it to learn better. Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data. Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well. So, it is very necessary to remove such noises and less-important data from the dataset to do this, and Feature selection techniques are used.

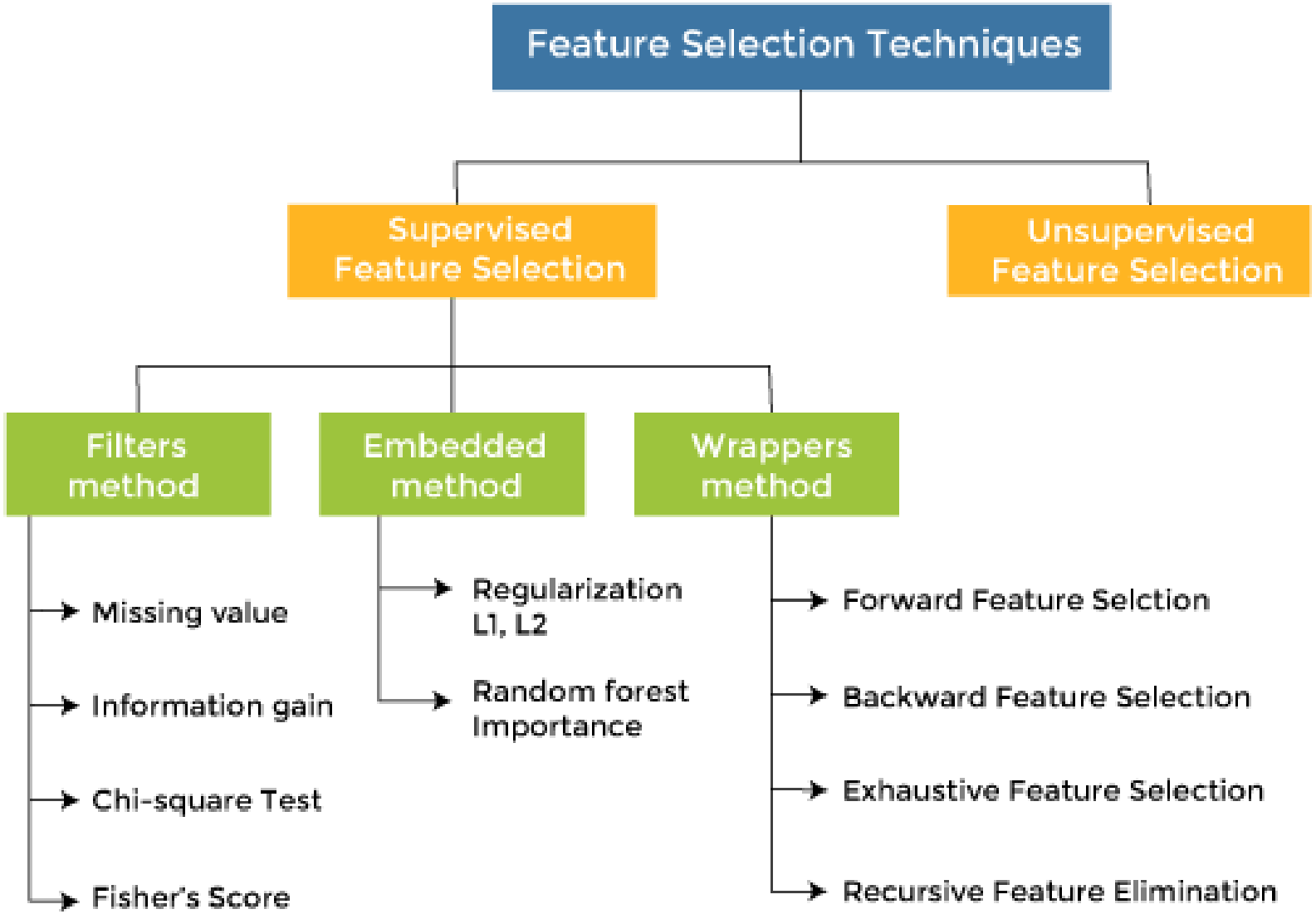
- some benefits of using feature selection
- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhance the generalization.

Feature Selection Techniques

- 1.Supervised Feature Selection technique
- Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.

- 2.Unsupervised Feature Selection technique
- Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset.

FEATURE SELECTION METHOD



- **1. Wrapper Methods**
- In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.
- Wrapper methods techniques

Forward selection - Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.

FEATURE SELECTION METHOD

- Backward elimination - Backward elimination is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering
- all the features and removes the least significant feature. This elimination process continues until removing the features does not improve the performance of the model.

FEATURE SELECTION METHOD

- Exhaustive Feature Selection- Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute force. It means this method tries & makes each possible combination of features and returns the best performing feature set.
- Recursive Feature Elimination-Recursive feature elimination is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features. Now, an estimator is trained with each set of features, and the importance of each feature is determined using coef_attribute or through a
- feature_importances_attribute.

- 2. **Filter Methods**
- In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.

- The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.

- The advantage of using filter methods is that it needs low computational time and does not overfit the data.

FEATURE SELECTION METHOD

- **Information Gain**: Information gain determines the reduction in entropy while transforming the dataset. It can be used as a feature selection technique by calculating the information gain
- of each variable with respect to the target variable.

- **Chi-square Test**: Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and
- the target variable, and the desired number of features with the best chi-square value is selected.

- **Fisher's Score**:
- Fisher's score is one of the popular supervised technique of feature selection. It returns the rank of the variable on the fisher's criteria in descending order. Then we can select the variables with a large fisher's score.
- **Missing Value Ratio**:The value of the missing value ratio can be used for evaluating the feature set against the threshold value.
- The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable is having more than the threshold value can be dropped.

- Some techniques of embedded methods are:

$$\text{Missing Value Ratio} = \frac{Number\ of\ Missing\ values * 100}{Total\ number\ of\ observations}$$

- 3**. Embedded Methods**
- Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.
- These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in