# *Chapter 2: Data Pre-processing in Machine Learning*

## Data processing

Data processing is the processes which help us to convert the raw data into meaningful information with the help this meaningful information, we are able to take an effective decision. We create different models to make the machine do the entire task based on the data which we feed with.

The steps followed in Data processing are:

1. ***Data Selection*****:** In this step, it involves collecting the data from reliable sources and the data which is collected should be of a better-quality data. Here, we don't focus on quantity. Our goal is to collect the quality data, based on the objective of the task from reliable sources.

2. ***Data Preprocessing*****:** This step involves with the part of formatting the data according to which algorithm will be able to understand it. In short, here, we are making our data ready to be processed. This involves three parts which are:
   a. ***Formatting*****:** We get the files in different formats like it can be a CSV format or may be in the form tables. We need format the file according to the requirement so that the machine is able to understand.
   b. ***Cleaning*****:** Here, we remove the unwanted data and fix the instances of missing data. We either fix them by removing the whole row or maybe by replacing it with the central values like means or median.
   c. ***Sampling*****:** Sampling is one method to understand the whole population in which we are working with. It saves the time and space for processing the data. We extract a small group of data which can represent the whole data.

3. ***Data Transformation*****:** We work on the preprocessed data to transform it as per requirement which we need to do the analysis. The data transformation involves with the following steps:
   a. ***Scaling*****:** Scaling is one of the processes with which the data is either made easier for the machine to process. Let's say for example if we make the attributes of the data much closer to each other instead of having outliers then it will be able to contribute in the insights. In the similar manner, if we normalize the data, the storage of the data is reduced.
   b. ***Decomposition*****:** With the help this process, we will be able to convert small, noisy or dirty data into useful and effective data. There are different ML methods which help to augmentation of the data based on the requirement.

c. ***Data Aggregation Process (DAP)*:** This process helps to understand the demographic study of the data, and their behavior patterns. It helps us to identify the useful information about a group and trace the root cause of errors in data.

4. ***Data Output*:** We can create different output in various formats like graphs, reports, video, image, audio, etc. This process involves two steps:
   a. Decoding the data to an understandable form, that earlier was encoded for the ML algorithm.
   b. Then, the decoded data is communicated to various locations that are accessible to any user at any time.

5. ***Data storage*:** This is the step where the data or the metadata is stored in a specific location for the future purpose.

## Data cleaning

It involves with the following:

✓ ***Removal of unwanted observation*:** The observations which are not required should be removed from the data. This might include with deleting the duplicate value or irrelevant values from the dataset. There are redundant observations which change the efficiency to a greater extent as the data repeats and may add towards the correct side or towards the incorrect side, which in turn produces unfaithful results.

✓ ***Fixing Structural errors*:** There are some errors which arise during measurement, transfer of data, typos in the name of features, the same attribute with a different name or mislabeled classes are called structural errors.

✓ ***Managing unwanted outliers*:** Outliers can cause problems to some types of models. Generally, we should remove the outliers. But sometimes removing the outliers will affect the results. Hence, we should have a good reason to remove the outliers from the data.

✓ ***Handling missing data*:** When we get missing values from the data, handling them is one of most important steps. We cannot simply remove those missing values or assign any value to it. It should be handled carefully. By using the technique of flagging the missing values, we are able to allow the algorithm to estimate the optimal constant for missing values, instead of just filling the missing value with mean.

## Data Transformation

1. ***Data Smoothing*****:** It is a process which is used to remove noise from the dataset using some algorithms. By smoothing the data, it highlights the important features which are present in the dataset. With the help of data smoothing, we able to identify the changes to help predict the trends and patterns in the data.

2. ***Data Aggregation*****:** Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results. The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.

3. ***Data Discretization*****:** This is a process of converting continuous data into a set of data intervals. Continuous attribute values are substituted by small interval labels. This makes the data easier to study and analyze. If a data mining task handles a continuous attribute, then its discrete values can be replaced by constant quality attributes. This improves the efficiency of the task. This method is also called a data reduction mechanism as it transforms a large dataset into a set of categorical data. Discretization also uses decision tree-based algorithms to produce short, compact, and accurate results when using discrete values.

4. ***Attribution construction*****:** In the attribute construction method, the new attributes consult the existing attributes to construct a new data set that eases data mining. New attributes are created and applied to assist the mining process from the given attributes. This simplifies the original data and makes the mining more efficient. For example, suppose we have a data set referring to measurements of different plots, i.e., we may have the height and width of each plot. So here, we can construct a new attribute 'area' from attributes 'height' and 'weight'. This also helps understand the relations among the attributes in a data set.

5. ***Generalization*****:** It converts low-level data attributes to high-level data attributes using concept hierarchy. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data. Data generalization can be divided into two approaches:
   - Data cube process (OLAP) approach.
   - Attribute-oriented induction (AOI) approach.

   For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old).

6. ***Data normalization*****:** It is generally considered the development of clean data. Diving deeper, however, the meaning or goal of data normalization is twofold:
   - Data normalization is the organization of data to appear similar across all records and fields.
   - It increases the cohesion of entry types leading to cleansing, lead generation, segmentation, and higher quality data.

   Simply put, this process includes eliminating unstructured data and redundancy (duplicates) in order to ensure logical data storage. When data normalization is done correctly, you will end up with standardized information entry. For example, this process applies to how URLs, contact names, street addresses, phone numbers, and even codes are recorded. These standardized information fields can then be grouped and read swiftly.

***a. First Normal Form (1NF)***

The most basic form of data normalization is 1NFm which ensures there are no repeating entries in a group. To be considered 1NF, each entry must have only one single value for each cell and each record must be unique. For example, you are recording the name, address, gender of a person, and if they bought cookies.

***b. Second Normal Form (2NF)***

Again, working to ensure no repeating entries, to be in the 2NF rule, the data must first apply to all the 1NF requirements. Following that, data must have only one primary key. To separate data to only have one primary key, all subsets of data that can be placed in multiple rows should be placed in separate tables. Then, relationships can be created through new foreign key labels.

For example, you are recording the name, address, gender of a person, if they bought cookies, as well as the cookie types. The cookie types are placed into a different table with a corresponding foreign key to each person’s name.

***c. Third Normal Form (3NF)***

For data to be in this rule, it must first comply with all the 2NF requirements. Following that, data in a table must only be dependent on the primary key. If the primary key is changed, all data that is impacted must be put into a new table.

For example, you are recording the name, address, and gender of a person but go back and change the name of a person. When you do this, the gender may then change as well. To avoid this, in 3NF gender is given a foreign key and a new table to store gender.

As you begin to better understand the normalization forms, the rules will become clearer while separating your data into tables and levels will become effortless. These tables will then make it simple for anyone within an organization to gather information and ensure they collect correct data that is not duplicated.

## Steps in data pre-processing

When it comes to creating a Machine Learning model, data preprocessing is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data preprocessing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models.

1. ***Acquire the dataset***
   Acquiring the dataset is the first step in data preprocessing in machine learning. To build and develop Machine Learning models, you must first acquire the relevant dataset. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. For instance, a business dataset will be entirely different from a medical dataset. While a business dataset will contain relevant industry and business data, a medical dataset will include healthcare-related data.

2. ***Import all the crucial libraries***
   Since Python is the most extensively used and also the most preferred library by Data Scientists around the world, we should know how to import Python libraries for data preprocessing in Machine Learning. The predefined Python libraries can perform specific data preprocessing jobs. Importing all the crucial libraries is the second step in data preprocessing in machine learning. The three core Python libraries used for this data preprocessing in Machine Learning are:

- ***NumPy*****:** NumPy is the fundamental package for scientific calculation in Python. Hence, it is used for inserting any type of mathematical operation in the code. Using NumPy, you can also add large multidimensional arrays and matrices in your code.
- ***Pandas*****:** Pandas is an excellent open-source Python library for data manipulation and analysis. It is extensively used for importing and managing the datasets. It packs in high-performance, easy-to-use data structures and data analysis tools for Python.
- ***Matplotlib*****:** Matplotlib is a Python 2D plotting library that is used to plot any type of charts in Python. It can deliver publication-quality figures in numerous hard copy formats and interactive environments across platforms (IPython shells, Jupyter notebook, web application servers, etc.).

3. ***Import the dataset***
   In this step, you need to import the dataset/s that you have gathered for the ML project at hand. Importing the dataset is one of the important steps in data preprocessing in machine learning. However, before you can import the dataset/s, you must set the current directory as the working directory.

4. ***Identifying and handling the missing values***

   In data preprocessing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data. Needless to say, this will hamper your ML project.

   Basically, there are two ways to handle missing data:

- ***Deleting a particular row*****:** In this method, you remove a specific row that has a null value for a feature or a particular column where more than 75% of the values are missing. However, this method is not 100% efficient, and it is recommended that you use it only when the dataset has adequate samples. You must ensure that after deleting the data, there remains no addition of bias.
- ***Calculating the mean*****:** This method is useful for features having numeric data like age, salary, year, etc. Here, you can calculate the mean, median, or mode of a particular feature or column or row that contains a missing value and replace the result for the missing value. This method can add variance to the dataset, and any loss of data can be efficiently negated. Hence, it yields better results compared to the first method (omission of rows/columns). Another way of approximation is through the deviation of neighboring values. However, this works best for linear data.

5. ***Encoding the categorical data***

   Categorical data refers to the information that has specific categories within the dataset. In the dataset cited above, there are two categorical variables – country and purchased.

   Machine Learning models are primarily based on mathematical equations. Thus, you can intuitively understand that keeping the categorical data in the equation will cause certain issues since you would only need numbers in the equations.

6. ***Splitting the dataset***

   Splitting the dataset is the next step in data preprocessing in machine learning. Every dataset for Machine Learning model must be split into two separate sets – training set and test set. Training set denotes the subset of a dataset that is used for training the machine learning model. Here, you are already aware of the output. A test set, on the other hand, is the subset of the dataset that is used for testing the machine learning model. The ML model uses the test set to predict outcomes.

   Usually, the dataset is split into 70:30 ratio or 80:20 ratio. This means that you either take 70% or 80% of the data for training the model while leaving out the rest 30% or 20%. The splitting process varies according to the shape and size of the dataset in question.