**1. What are the key tasks that machine learning entails? What does data pre-processing imply?**

**Ans:** Data preparation or its pre-processing is the procedure for organizing original data for usage in a Machine Learning system. It’s perhaps the most important stage in and the first one for building a ML model.

This is not always the situation that we come upon cleaning and preparing data when engaging on a ML project. And, before performing any data-related function, it is great for cleaning the information and reformat it. As a response, we utilize a data-preprocessing activity for all this.

1] Collect data The initial requirement for constructing a ML model is a database, as a model is entirely backed up by evidence. ...

* 2] Import the necessary libraries ...
* 3] Import the desired data ...
* 4] Find and handle the missing or null values ...
* 5] Encode the categorical quantities ...
* 6] Split the whole data into train-test sets ...
* 7] Scaling the features ...

**2. Describe quantitative and qualitative data in depth. Make a distinction between the two.**

**Ans:** The analysis in any research project involves summarizing the mass of information that has been collected and presenting the end results in such a way that it communicates the foremost necessary findings or options. For example, if a vesture complete is making an attempt to spot the most recent trends among young girls, the complete can initially reach young girls and raise their queries relevant to the analysis objective. Once collecting this information, the vesture can analyze the data to spot patterns – for example, it should discover that almost all young girls would really like to examine additional sort of jeans. There are many alternative data analysis ways, but the two most commonly and majorly used are Qualitative and Quantitative Analysis.

One variety of data is objective, up-to-the-point, and conclusive. The other variety of data is subjective, interpretive, and explained easily. Quantitative data can be counted, measured, and expressed using numbers. Qualitative data is descriptive and abstract and may be classified on traits and characteristics. The key variations between Qualitative and Quantitative data are as prescribed below:

* The data type, in which the classification of objects is based on attributes (quality) is called qualitative data. The type of information that might be counted and expressed in numbers and values is called quantitative data.
* Quantitative data relies on numbers. Simple arithmetic or additional advanced applied mathematics analysis is employed to get commonalities or patterns within the information. The results are usually seen in graphs and tables. Applications like Excel, SPSS, or R can be accustomed to calculate things like Average scores, range of times a specific answer was given, the correlation between two or additional variables, dependability, and validity of the results.
* The approach to the inquiry within the case of qualitative data is subjective and holistic, whereas quantitative information has an associative objective and targeted approach.

**3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.**

**Ans**: Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all

fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that then translates to rich data analysis and

allows the building of a convincing and credible answer to questions that have been posed. Regardless of the field of study or preference for defining data (quantitative, qualitative), accurate data collection is essential to maintaining the integrity of research. Both the selection of appropriatedata collection instruments (existing, modified, or newly developed) and clearly delineated instructions for their correct use reduce the likelihood of errors occurring. Data collection is one of the most important stages in conducting a research. You can have the best research design in the world but if you cannot collect the required data you will be not be able to complete your project. Data collection is a very demanding job which needs thorough planning, hard work, patience, perseverance and more to be able to complete the task successfully. Data collection starts with determining what kind of data required followed by the selection of a sample from a certain population. After that, you need to use a certain instrument to collect the data from the selected sample.

**4. What are the various causes of machine learning data issues? What are the ramifications?**

**Ans:** Machine Learning" is one of the most popular technology among all data scientists and machine learning enthusiasts. It is the most effective Artificial Intelligence technology that helps create automated learning systems to take future decisions without being constantly programmed. It can be considered an algorithm that automatically constructs various computer software using past experience and training data. It can be seen in every industry, such as healthcare, education, finance, automobile, marketing, shipping, infrastructure, automation, etc. Almost all big companies like Amazon, Facebook, Google, Adobe, etc., are using various machine learning techniques to grow their businesses. But everything in this world has bright as well as dark sides. Similarly, Machine Learning offers great opportunities, but some issues need to be solved.

Common issues in Machine Learning

Although machine learning is being used in every industry and helps organizations make more informed and data-driven choices that are more effective than classical methodologies, it still has so many problems that cannot be ignored. Here are some common issues in Machine Learning that professionals face to inculcate ML skills and create an application from scratch.

**1. Inadequate Training Data**

**2. Poor quality of data**

**3. Non-representative training data**

**4. Overfitting and Underfitting**

**5. Monitoring and maintenance**

**6. Getting bad recommendations**

**7. Lack of skilled resources**

**8. Customer Segmentation**

**9. Process Complexity of Machine Learning**

**10. Data Bias**

**Methods to ramifications :**

* Increase training data in a dataset.
* Reduce model complexity by simplifying the model by selecting one with fewer parameters
* Ridge Regularization and Lasso Regularization
* Early stopping during the training phase
* Reduce the noise
* Reduce the number of attributes in training data.
* Constraining the model.

**5. Demonstrate various approaches to categorical data exploration with appropriate examples.**

**Ans: Categorical Variable/Data (or Nominal variable):**Such variables take on a fixed and limited number of possible values. For example – grades, gender, blood group type, etc. Also, in the case of categorical variables, the logical order is not the same as categorical data e.g. “one”, “two”, “three”. But the sorting of these variables uses logical order. For example, gender is a categorical variable and has categories – male and female and there is no intrinsic ordering to the categories. A purely categorical variable is one that simply allows you to assign categories, but you cannot clearly order the variables. **Terms related to Variability Metrics :**

* **Mode :**Most frequently occurring value in the given data

**Example-**

Data = ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"]

Mode = "Bat"

* **Expected Value :**When working in machine learning, categories have to be associated with a numeric value, so as to give understanding to the machine. This gives an average value based on a category’s probability of occurrence i.e. Expected Value. It is calculated by –

-> Multiply each outcome by its probability of occurring.

-> Sum these values

* So, it is the sum of values times their probability of occurrence often used to sum up factor variable levels.
* **Bar Charts :**Frequency of each category plotted as bars. Loading Libraries –

**6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?**

**Ans:**  Delete Rows with Missing Values: Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped

This article covers 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 — Datawig

Every dataset has missing values that need to be handled intelligently to create a robust model. In this article, I have discussed 7 ways to handle missing values that can handle missing values in every type of column. There is no thump rule to handle missing values in a particular manner, the method which gets a robust model with the best performance. One can use various methods on different features depending on how and what the data is about. Having domain knowledge about the dataset is important, which can give an insight into how to preprocess the data and handle missing values..

**7. Describe the various methods for dealing with missing data values in depth.**

**Ans: Data is a powerful weapon to handle the world on the finger. As you know Customer is the king of business but customer’s data is the god of the customer and if you have customer’s god in your database then you can handle customers according to you.**

Now let’s talk about data cleaning, In this tutorial, we are working on **How to handle missing values or data** by different methods to make Machine Learning model powerful to provide an accurate prediction.

**What are the Methods to dealing Missing Values/Data?**

In general, there are the best **6 methods to handle missing data or values**. It is the part of Data Pre-processing and this is the most important step to build Machine Learning/Data Science project. The following are the most popular methods to handle missing data.

**•Ignore missing values row / Delete row**

**•Fill missing value manually**

**•Use global constant**

**•Measure of central tendency (Mean, Median & Mode)**

**•Measure of central tendency for each class**

**•Most probable value ( ML Algorithms)**

1. **What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.**

**Ans:** **Dimensionality** **Reduction**: This reduce the size of **data** by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed **data**, original **data** can be retrieved, such **reduction** are called lossless **reduction** else it is called lossy **reduction**.

**Pre-processing in Data Mining:**   
Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

**1. Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**   
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**   
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**   
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**   
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**   
   This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

**9 i. What is the IQR? What criteria are used to assess it?.**

**Ans:** The interquartile range is a widely accepted method to find outliers in data. When using the interquartile range, or IQR, the full dataset is split into four equal segments, or quartiles.**The distances between the quartiles** is what is used to determine the IQR.

The interquartile range (IQR) measures the spread of the middle half of your data. It is the range for the middle 50% of your sample. Use the IQR to assess the**variability where most of your values lie**. he **interquartile range** (**IQR**) is a **measure** of variability, based on dividing a data set into quartiles. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

**ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?**

**Ans:** Definition

The method to summarize a set of data that is measured using an interval scale is called a box and whisker plot. These are maximum used for data analysis. We use these types of graphs or graphical representation to know:

* Distribution Shape
* Central Value of it
* Variability of it

A box plot is a chart that shows data from a five-number summary including one of the measures of [central tendency](https://byjus.com/maths/central-tendency/). It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. Boxplots are also very beneficial when large numbers of data sets are involved or compared.

In simple words, we can define the box plot in terms of descriptive statistics related concepts. That means box or whiskers plot is a method used for depicting groups of numerical data through their quartiles graphically. These may also have some lines extending from the boxes or whiskers which indicates the variability outside the lower and upper quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers can be indicated as individual points.

It helps to find out how much the data values vary or spread out with the help of graphs. As we need more information than just knowing the measures of central tendency, this is where the box plot helps. This also takes less space. It is also a type of [pictorial representation of data](https://byjus.com/maths/pictorial-representation-of-data/)

**Parts of Box Plots**

Check the image below which shows the minimum, maximum, first quartile, third quartile, median and outliers.

**Minimum**: The minimum value in the given dataset

**First Quartile (Q1)**: The first quartile is the median of the lower half of the data set.

**Median:**The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.

**Third Quartile (Q3):**The third quartile is the median of the upper half of the data.

**Maximum:**The maximum value in the given dataset.

Apart from these five terms, the other terms used in the box plot are:

**Interquartile Range (IQR):**The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

**Outlier:**The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.

(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).

**10. Make brief notes on any two of the following:**

**1. Data collected at regular intervals**

**Ans: Interval Data: Definition, Examples, and Analys**

There are a total of four types of data in statistics primarily. They are nominal data, ordinal data, ratio data, and interval data.

[**Ordinal Data**](https://www.researchprospect.com/ordinal-data-definition-examples-and-interpretation/)**:** This type of data is classified into categories. However, the distance between these categories is unknown.

[**Nominal Data**](https://www.researchprospect.com/what-is-numerical-data-and-what-are-its-types/)**:**Nominal data is used to label variables without assigning any quantitative value to them.

[**Ratio Data**](https://www.researchprospect.com/ratio-data-definition-examples-and-analysis/)**:**This is a kind of qualitative data that measures variables on a continuous scale.

**Interval Data:**This data type is measured along a scale and has an equal distance between its values.

For now, let’s take a deeper look at Interval Data and its analysis.

Key Characteristics of Interval Data

Here are a few characteristics of Interval data:

* **Measurement:**Interval data is measured using an interval scale, which not only shows the order and direction but also shows the exact difference in the value. For example, the markings on a thermometer or a ruler are equidistant, in simpler words they measure the same distance between the two markings.
* **Interval Difference:**The distances between each value on interval data is equal. For example, the difference between 10 cm and 20 cms is the same as 20 cms and 30 cms.
* **Calculation:** In interval data, one can add or subtract values but cannot divide or multiply. Almost all statistical analysis are applicable when calculating interval data, mean, mode, median etc.
* **Point Zero:** Absolute zero point is arbitrary, which means a variable can be measured even if it has a negative value like temperature can be -10 below zero but height cannot be below zero.

**2. The gap between the quartiles**

**Quartiles** . There are three **quartiles** denoted by Q 1, Q 2 and Q 3 divides the frequency distribution in to four equal parts. That is 25 percent of data will lie below Q 1, 50 percent of data below Q 2 and 75 percent below Q 3. Here Q 2 is called the Median. **Quartiles** are obtained in almost the same way as median . **Quartiles** for Raw or Ungrouped ...

Here is a histogram of the age of all 934 Nobel Prize winners up to the year 2020, showing **the quartiles**: **The quartiles** (Q 0 ,Q 1 ,Q 2 ,Q 3 ,Q 4) are the values that separate each quarter. **Between** Q 0 and Q 1 are the 25% lowest values in the data. **Between** Q 1 and Q 2 are the next 25%. And so on. Q 0 is the smallest value in the data.

**3. Use a cross-tab**

Researchers use**cross-tabulation** to examine the relationship within the data that is not readily evident. It is quite useful in market research studies and surveys. A cross-tab report shows the connection between two or more questions asked in the study. Cross-tab is a popular choice for statistical data analysis.

Have you ever looked at the nutrition chart behind a snack pack? This little table gives you a comprehensive breakdown of how a particular snack will contribute to your overall energy levels. The analysis helps you make informed decisions regarding your diet and calorie consumption.

Cross-tabulation is a mainframe statistical model that follows similar lines. It helps you make informed decisions regarding your research by identifying patterns, trends, and the correlation between your study parameters. When conducting a study, the raw data can usually be daunting. They will always point to several chaotic possible outcomes. In such a situation, cross-tab helps you zero in on a single theory beyond doubt by drawing trends, comparisons, and correlations between mutually inclusive factors within your study.

**11. Make a comparison between:**

**1. Data with nominal and ordinal values**



**2. Histogram and box plot**

[Histograms](https://citoolkit.com/articles/histogram/) and [box plots](https://citoolkit.com/articles/box-plot/) are graphical representations for the frequency of numeric data values. They aim to [describe](https://citoolkit.com/articles/descriptive-statistics/) the data and explore the central tendency and variability before using advanced statistical analysis techniques. In this article, we will further discuss the similarities and differences between these two tools. Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points. Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying [probability distribution](https://citoolkit.com/articles/probability-distributions/) of a data.

Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space. Although histograms are better in displaying the [distribution](https://citoolkit.com/articles/probability-distributions/) of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length. You can use histograms and box plots to verify whether an improvement has been achieved by exploring the data before and after the improvement initiative.

Both tools can be helpful to identify whether variability is within specification limits, whether the process is capable, and whether there is a shift in the process over time. Both histograms and box plots are ideal to represent moderate to large amount of data. They may not accurately display the distribution shape if the data size is too small. In practice, a sample size of at least 30 data values would be sufficient for both tools.

Many statistical applications allow the option of summarizing your data [graphically](https://citoolkit.com/articles/graphical-analysis/) (including plotting the data on histograms and box plots as shown below). This can reveal unusual observations in your data that should be investigated before performing detailed statistical analysis.

**3. The average and median**

|  |  |  |
| --- | --- | --- |
|  | **Median** | **Average** |
| Definition | Middle number or average of middle numbers in the sorted list of numbers | Also known as mean, obtained by dividing the sum of quantities by the number of quantities |
| Formula | n= total members in the list  If n =odd  Median = ((n+1)/2)th term  If n=even  Median = ((n/2)th term + (n/2+1)th term)/2 | Sum of all data values/ number of data values |
| |  |  | | --- | --- | | Extreme values in the  data set |  | | Preferred | Not preferred |
| Example in usage | Usually used in income level research | Usually used when the graph falls on a normal distribution |