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

* Get Data: The first step in the Machine Learning process is getting data.
* Cleaning, Preparing & Manipulating Data: Real-world data often has unorganized, missing, or noisy elements.
* Train Model: This step is where the magic happens!
* Testing Model.
* Improving model.
* Prediction
* Data preprocessing involves transforming raw data to well-formed data sets so that data mining analytics can be applied.
* Preprocessing involves both data validation and data imputation
* The Goal of Data Validation is to assess whether the data in question is both complete and accurate.
* The Goal of Data Imputation is to correct errors and input missing values, Either Manually or Automatically through business process automation (BPA) programming.

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

The Data Type Is Broadly Classified Into:

1. Quantitative
2. Qualitative

* **Quantitative Data Type:** This Type Of Data Type Consists Of Numerical Values. Anything Which Is Measured By Numbers. E.G., Profit, Quantity Sold, Height, Weight, Temperature, Etc. This Is Again Of Two Types
  1. **Discrete Data Type**: – The Numeric Data Which Have Discrete Values Or Whole Numbers. This Type Of Variable Value If Expressed In Decimal Format Will Have No Proper Meaning. Their Values Can Be Counted. E.G.: – No. Of Cars You Have, No. Of Marbles In Containers, Students In A Class, Etc.
  2. **Continuous Data Type**: – The Numerical Measures Which Can Take The Value Within A Certain Range. This Type Of Variable Value If Expressed In Decimal Format Has True Meaning. Their Values Can Not Be Counted But Measured. The Value Can Be Infinite E.G.: Height, Weight, Time, Area, Distance, Measurement Of Rainfall, Etc.
* **Qualitative Data Type:** These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type. This Can Be Divided Into:-
  1. **Structured Data:** This Type Of Data Is Either Number Or Words. This Can Take Numerical Values But Mathematical Operations Cannot Be Performed On It. This Type Of Data Is Expressed In Tabular Format. E.G.) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.
  2. **Unstructured Data:** This Type Of Data Does Not Have The Proper Format And Therefore Known As Unstructured Data.This Comprises Textual Data, Sounds, Images, Videos, Etc.

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

1. **Determine What Information You Want to Collect:** The first thing you need to do is choose what details you want to collect. You’ll need to decide what topics the information will cover, who you
2. want to collect it from and how much data you need. Your goals — what you hope to accomplish using your data — will determine your answers to these questions. As an example, you may decide to collect data about which type of articles are most popular on your website among visitors who are between the ages of 18 and 34. You might also choose to gather information about the average age of all of the customers who bought a product from your company within the last month.
3. **Set a Timeframe for Data Collection:** Next, you can start formulating your plan for how you’ll collect your data. In the early stages of your planning process, you should establish a timeframe for your data collection. You may want to gather some types of data continuously. When it comes to transactional data and website visitor data, for example, you may want to set up a method for tracking that data over the long term. If you’re tracking data for a specific campaign, however, you’ll track it over a defined period. In these instances, you’ll have a schedule for when you’ll start and end your data collection.
4. **Determine Your Data Collection Method:** At this step, you will choose the data collection method that will make up the core of your data-gathering strategy. To select the right collection method, you’ll need to consider the type of information you want to collect, the timeframe over which you’ll obtain it and the other aspects you determined.
5. **Collect the Data:** Once you have finalized your plan, you can implement your data collection strategy and start collecting data. You can store and organize your data in your DMP. Be sure to stick to your plan and check on its progress regularly. It may be useful to create a schedule for when you will check in with how your data collection is proceeding, especially if you are collecting data continuously. You may want to make updates to your plan as conditions change and you get new information.
6. **Analyze the Data and Implement Your Findings:** Once you’ve collected all of your data, it’s time to analyze it and organize your findings. The analysis phase is crucial because it turns raw data into valuable insights that you can use to enhance your marketing strategies, products and business decisions. You can use the analytics tools built into our DMP to help with this step. Once you’ve uncovered the patterns and insights in your data, you can implement the findings to improve your business.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| car\_name | brand | model | vehicle\_age | km\_driven | seller\_type | fuel\_type | transmission\_type | mileage | engine | max\_power | seats |
| Maruti Alto | Maruti | Alto | 9 | 120000 | Individual | Petrol | Manual | 19.7 | 796 | 46.3 | 5 |
| Hyundai Grand | Hyundai | Grand | 5 | 20000 | Individual | Petrol | Manual | 18.9 | 1197 | 82 | 5 |
| Hyundai i20 | Hyundai | i20 | 11 | 60000 | Individual | Petrol | Manual | 17 | 1197 | 80 | 5 |
| Maruti Alto | Maruti | Alto | 9 | 37000 | Individual | Petrol | Manual | 20.92 | 998 | 67.1 | 5 |
| Ford Ecosport | Ford | Ecosport | 6 | 30000 | Dealer | Diesel | Manual | 22.77 | 1498 | 98.59 | 5 |
| Maruti Wagon R | Maruti | Wagon R | 8 | 35000 | Individual | Petrol | Manual | 18.9 | 998 | 67.1 | 5 |
| Hyundai i10 | Hyundai | i10 | 8 | 40000 | Dealer | Petrol | Manual | 20.36 | 1197 | 78.9 | 5 |
| Maruti Wagon R | Maruti | Wagon R | 3 | 17512 | Dealer | Petrol | Manual | 20.51 | 998 | 67.04 | 5 |
| Hyundai Venue | Hyundai | Venue | 2 | 20000 | Individual | Petrol | Automatic | 18.15 | 998 | 118.35 | 5 |

Numeric-continous

Numeric-Discrete

Categorical

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

Noisy data, dirty data, and incomplete data are the quintessential enemies of ideal Machine Learning. The solution to this conundrum is to take the time to evaluate and scope data with meticulous data governance, data integration, and data exploration until you get clear data. Ramifications or major issues in machine learning are:

1. Five practical issues in machine learning and the business implications Data quality.
2. Machine learning systems rely on data.
3. The complexity and quality trade-off.
4. Sampling bias in data.
5. Changing expectations and concept drift.
6. Monitoring and maintenance.

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

* Label Encoding or Ordinal Encoding
* One hot Encoding
* Dummy Encoding
* Effect Encoding
* Binary Encoding
* BaseN Encoding
* Hash Encoding
* Target Encoding

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

Even in a Well-Designed & Controlled study, Missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions.

* Missing value correction is required to reduce bias and to produce powerful suitable models.
* Methods for dealing with missing values. The popular methods which are used by the machine learning community to handle the missing value for categorical variables in the dataset are as follows: Delete the observations: If there is a large number of observations in the dataset, where all the classes to be predicted are sufficiently represented in the training data, then try deleting the missing value observations, which would not bring significant change in your feed to your model

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

1. **Delete the observations:** If there is a large number of observations in the dataset, where all the classes to be predicted are sufficiently represented in the training data, then try deleting the missing value observations, which would not bring significant change in your feed to your model. For Example Implement this method in a given dataset, we can delete the entire row which contains missing values.
2. **Replace missing values with the most frequent value:** You can always impute them based on Mode in the case of categorical variables, just make sure you don’t have highly skewed class distributions.
3. **Develop a model to predict missing values:** One smart way of doing this could be training a classifier over your columns with missing values as a dependent variable against other features of your data set and trying to impute based on the newly trained classifie

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

Data cleaning

Missing value handling,

Outlier handling

Duplicates removal

Catergorical data handling

Scaling

Dimension reduction

Feature selection

**Dimension 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).

**Function selection**

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

9.

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

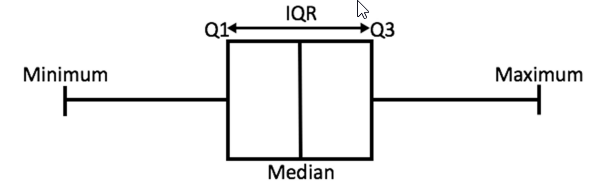
* Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
* Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.
* The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

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?

* minimum is the minimum value in the dataset
* maximum is the maximum value in the dataset.
* So the difference between the two tells us about the range of dataset.
* The median is the median (or centre point), also called second quartile, of the data (resulting from the fact that the data is ordered).
* Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
* Q3 is the third quartile of the data, i.e., to say 75%
* When the data is left skewed, lower whisker will be longer than upper whisker.
* To detect the outliers this method is used, we define a new range, let’s call it decision range, and any data point lying outside this range is considered as outlier and is accordingly dealt with. The range is as given below:

Lower Bound: (Q1 - 1.5 \* IQR)Upper Bound: (Q3 + 1.5 \* IQR)

* The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR



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

1. Data collected at regular intervals

* Interval data is one of the two types of discrete data.
* An example of interval data is the data collected on a thermometer—its gradation or markings are equidistant.
* Unlike ordinal data, interval data always take numerical values where the distance between two points on the scale is standardised and equal.

2. The gap between the quartiles

* Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1.
* Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3.
* The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR.

3. Use a cross-tab

Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyse categorical (nominal measurement scale) data.

**contingency table** (also known as a **cross tabulation** or **crosstab**) is a type of [table](https://en.wikipedia.org/wiki/Table_(information)) in a [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) format that displays the (multivariate) [frequency distribution](https://en.wikipedia.org/wiki/Frequency_distribution) of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them

1. Make a comparison between:

1. Data with nominal and ordinal values

Nominal data:

Nominal data is defined as data that is used for naming or labelling variables, without any quantitative value.

There is usually no intrinsic ordering to nominal data. For example, Race is a nominal variable having a number of categories, but there is no specific way to order from highest to lowest and vice versa.

Ordinal data:

Ordinal data is a type of categorical data with an order. The variables in ordinal data are listed in an ordered manner. The ordinal variables are usually numbered, so as to indicate the order of the list. However, the numbers are not mathematically measured or determined but are merely assigned as labels for opinions.

2. Histogram and box plot

* Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed 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 media are nearly the same, and the two whiskers has almost the same length.

3. The average and median

The mean (informally, the “average“) is found by adding all of the numbers together and dividing by the number of items in the set The median is found by ordering the set from lowest to highest and finding the exact middle.

Eg: 10,20,25,40,50

Mean=29

Median=25