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

Ans: A machine learning task is the type of prediction or inference being made, based on the problem or question that is being asked, and the available data. For example, the classification task assigns data to categories, and the clustering task groups data according to similarity.

Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process.

There are several different tools and methods used for preprocessing data, including the following:

* sampling, which selects a representative subset from a large population of data;
* transformation, which manipulates raw data to produce a single input;
* denoising, which removes noise from data;
* imputation, which synthesizes statistically relevant data for missing values;
* normalization, which organizes data for more efficient access; and
* feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

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

Ans: **Quantitative data is numbers-based, countable, or measurable.** **Qualitative data is interpretation-based, descriptive, and relating to language**. Quantitative data tells us how many, how much, or how often in calculations. Qualitative data can help us to understand why, how, or what happened behind certain behaviors.

### Advantages of quantitative data

* It’s relatively quick and easy to collect and it’s easier to draw conclusions from.
* When you collect quantitative data, the type of results will tell you which statistical tests are appropriate to use.
* As a result, interpreting your data and presenting those findings is straightforward and less open to error and subjectivity.

Another advantage is that you can replicate it. Replicating a study is possible because your data collection is measurable and tangible for further applications.

### Disadvantages of quantitative data

* Quantitative data doesn’t always tell you the full story (no matter what the perspective).
* With choppy information, it can be inconclusive.
* Quantitative research can be limited, which can lead to overlooking broader themes and relationships.
* By focusing solely on numbers, there is a risk of missing larger focus information that can be beneficial.

### Advantages of qualitative data

* Qualitative data offers rich, in-depth insights and allows you to explore context.
* It’s great for exploratory purposes.
* Qualitative research delivers a predictive element for continuous data.

### Disadvantages of qualitative data

* It’s not a statistically representative form of data collection because it relies upon the experience of the host (who can lose data).
* It can also require multiple data sessions, which can lead to misleading conclusions.

|  |  |
| --- | --- |
| **Quantitative** | **Qualitative** |
| Number-based,countable and measurable | interpretation-based, descriptive and relating to language |
| tells us how much, how many, how often | helps us to understand why, how or what happened |
| fixed and universal | subjective and unique |
| research methods measuring and counting | research methods are interviewing and observing |
| analysed using statistical analysis | analyzed by grouping the data into categories and themes |
| ex Age, Height, Income, Group Size, Test score, Clinical skills performed, Number of Errors | Gender, Religion, Native Language, Social Class, Method of Treatment, Problem-solving strategy used, Type of eaching |

**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:

|  |  |  |
| --- | --- | --- |
| **Quantitative** |  |  |
| Continuous | any numeric value -infinity to + infinity) | 10 ,100 22, 23.1 |
| discrete | only particular number 0 to infinity (whole numbers) | 113,122, |
| Continuous —Interval | extension of ordinal numbers with standardised scale | 0 10 20 30 |
| Continuous– ratio | extension of interval | 1/2 |
| **Qualitative** |  |  |
| Nominal | random named, labeled | city, name, etc |
| Ordinal | categorized in a particular order or on a ranging scale | delhi, calcutta |
| Binary | combination of zeros and one | 0,1 |

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

**Ans Common issues in ML:**

1. inadequate training data
   1. noisy data
   2. incorrect data
   3. generalizing of output data
2. poor quality of data
3. Non-representative training data

Ramification: Overfitting and underfitting, Monitoring and maintenance, Model may not be generalized, may result in bad recommendations and concept drift in the model. data may be biased. An ML system doesn't perform well if the training set is too small or if the data is not generalized, noisy, and corrupted with irrelevant features.

**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

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

Mosaic plot is a visualization technique suitable for contingency tables

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

**Ans:** Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples

Handling missing values:

set a priori targets for the unacceptable level of missing data.

Imputation: mean substitution, regression imputation, Maximum likelihood, Multiple imputation

Sensitivity analysis: Sensitivity analysis is defined as the study which defines how the uncertainty When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for the correctnessy in the output of a model can be allocated to the different sources of uncertainty in its inputs.

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

**Ans:** We can fill the missing data, interpolate the data or drop the missing data.

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

Deleting the Missing value

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values.

If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.

If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted.

The disadvantage of this method is one might end up deleting some useful data from the dataset.

There are 2 ways one can delete the missing values:

1. Deleting the entire row

If a row has many missing values then you can choose to drop the entire row.

If every row has some (column) value missing then you might end up deleting the whole data

1. Deleting the entire column

If a certain column has many missing values then you can choose to drop the entire column.

Imputing the Missing Value

1. Replacing With Arbitrary Value

If you can make an educated guess about the missing value then you can replace it with some arbitrary value

1. Replacing With Mean

This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.

1. Replacing With Mode

Mode is the most frequently occurring value. It is used in the case of categorical features.

1. Replacing With Median

Median is the middlemost value. It’s better to use the median value for imputation in the case of outliers.

1. Replacing with previous value – Forward fill

In some cases, imputing the values with the previous value instead of mean, mode or median is more appropriate. This is called forward fill. It is mostly used in time series data.

1. Replacing with next value – Backward fill

In backward fill, the missing value is imputed using the next value.

1. Interpolation

Missing values can also be imputed using interpolation.

Imputing Missing Values For Categorical Features

There are two ways to impute missing values for categorical features as follows:

1. Impute the Most Frequent Value
2. Impute the Value “missing”, which treats it as a Separate Category

Nearest Neighbors Imputations (KNNImputer)

Adding missing indicator to encode “missingness” as a feature

In some cases, while imputing missing values, you can preserve information about which values were missing and use that as a feature.

Because sometimes there may be a relationship between the reason for missing values (also called the “missingness”) and the target variable you are trying to predict.

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

**Ans:** Data preprocessing is a Data Mining method that entails converting raw data into a format that can be understood. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific activities or trends, as well as including numerous inaccuracies. This might result in low-quality data collection and, as a result, low-quality models based on that data. Preprocessing data is a method of resolving such problems.

Data Preprocessing can be done in four different ways. Data cleaning/cleaning, data integration, data transformation, and data reduction

1. Data Cleaning :
   1. Missing Data :

It’s fairly common for your dataset to contain missing values. It could have happened during data collection or as a result of a data validation rule, but missing values must be considered anyway.

* + 1. Dropping rows/columns: If the complete row is having NaN values then it doesn't make any value out of it. So such rows/columns are to be dropped immediately. Or if the % of row/column is mostly missing say about more than 65% then also one can choose to drop.
    2. Checking for duplicates: If the same row or column is repeated then also you can drop it by keeping the first instance. So that while running machine learning algorithms, so as not to offer that particular data object an advantage or bias.
    3. Estimate missing values: If only a small percentage of the values are missing, basic interpolation methods can be used to fill in the gaps. However, the most typical approach of dealing with missing data is to fill them in with the feature’s mean, median, or mode value.
  1. Noisy Data:

Noisy data is meaningless data that machines cannot interpret. It can be caused by poor data collecting, data input problems, and so on. It can be dealt with in the following ways:

* + 1. Binning Method: This method smooths data that has been sorted. The data is divided into equal-sized parts, and the process is completed using a variety of approaches. Each segment is dealt with independently. All data in a segment can be replaced by its mean, or boundary values can be used to complete the task.
    2. Clustering: In this method, related data is grouped in a cluster. Outliers may go unnoticed, or they may fall outside of clusters.
    3. Regression: By fitting data to a regression function, data can be smoothed out. The regression model employed may be linear (with only one independent variable) or multiple (with numerous independent variables) (having multiple independent variables).

1. Data Integration

It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases. Do you think how data can be matched up?? For a data analyst in one database, he finds Customer\_ID and in another he finds cust\_id, How can he sure about them and say these two belong to the same entity. Databases and Data warehouses have Metadata (It is the data about data) it helps in avoiding errors.

1. Data transformation:
   1. Normalization:

It is done to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

* 1. Concept Hierarchy Generation:

Using concept hierarchies, low-level or primitive/raw data is substituted with higher-level concepts in data generalization. Categorical qualities, for example, are generalized to higher-level notions such as street, city, and nation. Similarly, numeric attribute values can be translated to higher-level concepts like age, such as youthful, middle-ag3.

* 1. Smoothing

Smoothing works to remove the noise from the data. Such techniques include binning, clustering, and regression.

* 1. Aggregation

Aggregation is the process of applying summary or aggregation operations on data. Daily sales data, for example, might be combined to calculate monthly and annual totals. Feature Aggregation — If the features are highly correlated or if the features can be aggregated into another single feature then it is worth doing it. For example in the dataset you have the height and width of an object then they can be featured into a single feature area. This decreases dimensionality. These types of features are highly correlated in nature as a result it also decreases multicollinearity.ed, or elderly.

1. Data Reduction

Because data mining is a methodology for dealing with large amounts of data. When dealing with large amounts of data, analysis becomes more difficult. We employ a data reduction technique to get rid of this. Its goal is to improve storage efficiency while lowering data storage and analysis expenses.

* 1. Dimensionality reduction

A huge number of features may be found in most real-world datasets. Consider an image processing problem: there could be hundreds of features, also known as dimensions, to deal with. As the name suggests, dimensionality reduction seeks to minimize the number of features — but not just by selecting a sample of features from the feature set, which is something else entirely — Feature Subset Selection or feature selection

* 1. Numerosity Reduction:

Data is replaced or estimated using alternative and smaller data representations such as parametric models (which store only the model parameters rather than the actual data, such as Regression and Log-Linear Models) or non-parametric approaches (e.g. Clustering, Sampling, and the use of histograms).

1. Preprocessing of Text data:

Preprocessing the text data is a very important step while dealing with text data because the text at the end is to be converted into features to feed into the model. The objective of preprocessing text data is that we won't get rid of characters, words, others that don’t give value to us. We want to get rid of punctuations, stop words, URLs, HTML codes, spelling corrections, etc. We would also like to do Stemming and Lemmatization so that in features duplication of words is not there which convey almost the same meaning

* 1. Steps to perform for text pre-processing
     1. Read the text— Read the text data and store it in a variable
     2. Store in the list — Using df.tolist() store the sentences in a list.
     3. Initialize the Preprocess object and pass techniques\*
     4. Iterate through the list to get the processed text.
     5. For the after reading text data we will apply Preprocess object present in preprocessing

1. Preprocessing of Image data:

The term “image pre-processing” refers to actions on images at the most basic level. If entropy(degree of randomness) is an information metric, these methods do not improve image information content, but rather decrease it. Pre-processing aims to improve image data by suppressing unwanted distortions or enhancing particular visual properties that are important for subsequent processing and analysis.

Steps to perform for image pre-processing

* 1. Read image — Read the images
  2. Resize image — Resize the images because the image size captured and fed to the model is different. So it is good to establish a base size and resize the images
  3. Remove noise(Denoise) — Using Gaussian blur inside the function processing() we can smooth the image to remove unwanted noise.
  4. Segmentation Morphology (smoothing edges)— We will segment the image in this stage, separating the background from foreground objects, and then we will refine our segmentation with more noise removal.
  5. There are 4 different types of Image Pre-Processing techniques and they are listed below.
     1. Pixel brightness transformations/ Brightness corrections

The most common Pixel brightness transforms operations are

g(x)=αf(x)+β, alpha and beta control contrast and brightness of the image.

* + - 1. Gamma correction — It is a non-linear adjustment to individual pixel values.
      2. Histogram equalization — It is a contrast enhancement technique.
      3. Sigmoid Stretching —It has a contrast factor ‘C’ and a threshold value where we may manage the overall contrast enhancement by lighting and darkening the image.
    1. Geometric Transformations
    2. Image Filtering and Segmentation
    3. Fourier transform and Image restoration

**9.**

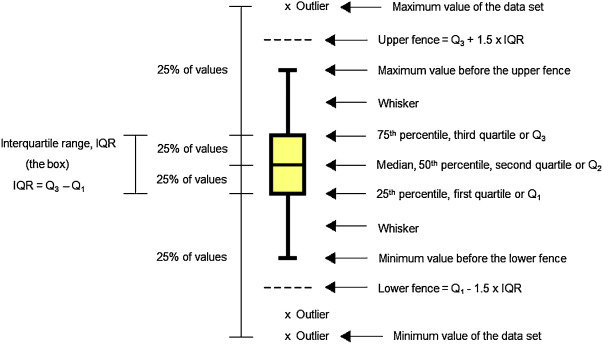
**i. What is the IQR? What criteria are used to assess it?**

**Ans:** The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile range = Upper Quartile – Lower Quartile = Q­3 – Q­1

**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:** The main components of a boxplot: median, quartiles, whiskers, fences and outliers

****

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

When the Box plot is negatively skewed. ie. If the distance from median to minimum is greater than the distance from median to maximum. The lower whisker surpass the upper whisker in length

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

**1. Data collected at regular intervals**

**2. The gap between the quartiles**

Quartiles are divided in to three Q1: median of the lower half of the dataset, Q3 is the median of upper half of the data, IQR: The diff. between the Q3 and Q1.In effect, it is the range of the middle half of the data that shows how spread out the data is.

Statisticians sometimes also use the terms semi-interquartile range and mid-quartile range .

The semi-interquartile range is one-half the difference between the first and third quartiles. It is half the distance needed to cover half the scores. The semi-interquartile range is affected very little by extreme scores. This makes it a good measure of spread for skewed distributions. It is obtained by evaluating (Q3−Q1)/2 .

The mid-quartile range is the numerical value midway between the first and third quartile. It is one-half the sum of the first and third quartiles. It is obtained by evaluating (Q3+Q1)/2 .

(The median, midrange and mid-quartile are not always the same value, although they may be.)

**3. Use a cross-tab**

Cross tabulation is a method to quantitatively analyze the relationship between multiple variables.

Also known as contingency tables or cross tabs, cross tabulation groups variables to understand the correlation between different variables. It also shows how correlations change from one variable grouping to another. It is usually used in statistical analysis to find patterns, trends, and probabilities within raw data.

Cross tabulation is usually performed on categorical data — data that can be divided into mutually exclusive groups.

**11. Make a comparison between:**

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

**Nominal data is classified without a natural order or rank, whereas ordinal data has a predetermined or natural order**

|  |  |
| --- | --- |
| **Nominal data** | **Ordinal Data** |
| group of Non-Parametric variables | group of Non-Parametric ordered vaiables |
| evaluated using nonparametric statistics | can be evaluated using parametric statistics in some cases. |
| Ex country gender race etc of a group of people  data are nouns, with no order | includes data having a position in class as ‘first’ or ‘second’, comes with level of order |
| Test: MCNemar, Cochran Q’s, Fisher’ Exact, Chi-square | Test: Wilconxon signed, Friedman 2-way Anova, Wilconxon rank-sum, Kruskal-wallis 1-way |
| Data analysis done by grouping input variables into categories and cal. percentage or mode of the distribution | analyzed by computing the mode, median and other positional measures like quartiles, percentile etc. |
| Collection techniques, Open ended questions, multiple response choice questions, close ended questions | Collected using Likert scale, interval scale, rating scale etc. |
| Categorical in nature | categorical and quantitative. sometimes assign quantitative values to ordinal data. |
| used for research that involve getting personal data, place or thing | uses to carry investigations that involve getting people's views or opinion on some matter |
| Freedom to express freely and give adequate information.  Researchers have to deal with a lot of irrelevant data.  Inclusiveness of responses | does not give freedom to respondents to express themselves freely. restricted to particular choice of option  Restriction give researchers access to concise data, by eliminating any possibility of having irrelevant data |
| collection techniques are not user friendly. for open and close ended questions, respondents may have to type their inputs, something many respondents find tiring and time-consuming | user friendly as features can be integrated into ordinal data collection forms, making it user-friendly. |

**2. Histogram and box plot**

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.. Histograms and box plots are very similar in that they both help to visualize and describe numeric data

Histograms are preferred to determine the underlying probability distribution 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 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 have almost the same length.

**3. The average and median**

Average and median are both measures of “central tendency,” in that they are intended to provide some indication of a typical or middle value of a set of data. The average is calculated by adding up all of the individual values and dividing this total by the number of observations. The median is calculated by taking the “middle” value, the value for which half of the observations are larger and half are smaller.