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.

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.

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

🡪**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. There are Two Types

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.

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:-

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.

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.

🡪Basic data collection with some sample records, each containing attributes from different machine learning data types:

1. Data Type: Numeric

Attribute: Age (in years)

Sample Records:

- Record 1: Age: 25

- Record 2: Age: 32

- Record 3: Age: 40

2. Data Type: Categorical (Nominal)

Attribute: Gender

Sample Records:

- Record 1: Gender: Male

- Record 2: Gender: Female

- Record 3: Gender: Non-binary

3. Data Type: Categorical (Ordinal)

Attribute: Educational Level

Sample Records:

- Record 1: Educational Level: High School

- Record 2: Educational Level: Bachelor's Degree

- Record 3: Educational Level: Master's Degree

4. Data Type: Boolean

Attribute: Employment Status

Sample Records:

- Record 1: Employment Status: True (employed)

- Record 2: Employment Status: False (unemployed)

- Record 3: Employment Status: True (employed)

5. Data Type: Text/String

Attribute: Occupation

Sample Records:

- Record 1: Occupation: Software Engineer

- Record 2: Occupation: Teacher

- Record 3: Occupation: Sales Representative

6. Data Type: Date/Time

Attribute: Registration Date

Sample Records:

- Record 1: Registration Date: 2022-01-15

- Record 2: Registration Date: 2023-03-10

- Record 3: Registration Date: 2021-11-27

7. Data Type: Floating-Point

Attribute: Weight (in kilograms)

Sample Records:

- Record 1: Weight: 68.5

- Record 2: Weight: 75.2

- Record 3: Weight: 62.8

8. Data Type: Integer

Attribute: Number of Children

Sample Records:

- Record 1: Number of Children: 0

- Record 2: Number of Children: 2

- Record 3: Number of Children: 1

This data collection includes attributes representing various machine learning data types like numeric, categorical (nominal and ordinal), Boolean, text, date/time, floating-point, and integer. It can be used for a variety of machine learning tasks, such as classification, regression, and clustering.

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

🡪 Machine learning data issues can significantly impact the performance and reliability of machine learning models. There are several causes of these issues, and they can have various ramifications on the effectiveness and robustness of the models.

1. Insufficient data: If the dataset used for training is too small, the model may not learn the underlying patterns effectively, leading to poor generalization to new data.

2. Imbalanced data: When certain classes or categories in the dataset are significantly underrepresented, the model may become biased toward the majority class and struggle to correctly predict the minority class.

3. Noisy data: Noise in the data can be caused by errors during data collection, labeling, or processing. This noise can mislead the model and lead to incorrect predictions.

4. Missing data: If there are missing values in the dataset, the model may not have enough information to make accurate predictions, potentially leading to biased results.

5. Biased data: If the data used for training contains inherent biases, the model can learn and perpetuate those biases, leading to discriminatory or unfair predictions.

6. Outliers: Outliers in the data can significantly affect the training process, leading to suboptimal model performance or overfitting.

7. Data distribution mismatch: If the distribution of the training data differs significantly from the distribution of the test data or real-world data, the model may not generalize well to new instances.

8. Feature selection and engineering: Choosing irrelevant or poorly constructed features can lead to poor model performance and ineffective learning.

9. Data leakage: When information from the test set or future data is inadvertently included in the training process, the model may perform well during evaluation but fail in real-world scenarios.

10. Inconsistent data representation: Different data sources might have varying formats or representations, making it difficult to integrate and use the data effectively for training.

**Ramifications of these data issues:**

1. Reduced model performance: Data issues can lead to models that have low accuracy, precision, recall, or other performance metrics.

2. Overfitting or underfitting: Poor data quality can cause the model to overfit (perform well on training data but poorly on new data) or underfit (fail to learn the underlying patterns in the data).

3. Biased predictions: Biases present in the training data can result in biased predictions, leading to unfair or discriminatory outcomes.

4. Increased development time: Addressing data issues often requires additional time and effort, prolonging the model development process.

5. Higher operational costs: Models that perform poorly due to data issues may lead to increased operational costs, as they might require more frequent updates or human intervention.

6. Loss of trust: Models with unreliable or biased outcomes can erode trust in the system, leading to reduced adoption and acceptance.

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

🡪1. Unique value count: One of the first things which can be useful during data exploration is to see how many unique values

are there in categorical columns.

2. Frequency Count: Frequency count is finding how frequent individual values occur in column.

3. Variance: Variance gives a good indication how the values are spread.

4. Pareto Analysis: Pareto analysis is a creative way of focusing on what is important. Pareto 80–20 rule can be effectively used in data exploration.

5. Histogram: Histogram are one of the data scientists favourite data exploration techniques. It gives information on the range of values in which most of the values fall. It also gives information on whether there is any skew in data.

6. Correlation Heat-map between all numeric columns: The term correlation refers to a mutual relationship or association between two things.

7. Pearson Correlation and Trend between two numeric columns: Once you have visualised correlation heat-map , the next step is to see the correlation trend between two specific numeric columns.

8. Outlier overview: Finding something unusual in data is called Outlier detection (also known as anomaly detection). These outliers represent something unusual, rare , anomaly or something exceptional.

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.

--Real-world data collection has its own set of problems, It is often very messy which includes missing data, presence of outliers, unstructured manner, etc.

--Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model.

--Missing value in a dataset is a very common phenomenon in the reality.

--Missing value correction is required to reduce bias and to produce powerful suitable models.

--Most of the algorithms can’t handle missing data, thus you need to act in some way to simply not let your code crash. So, let’s begin with the methods to solve the problem.

--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.

🡪 The Various Methods for dealing with missing data values are:

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.

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.

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 classifier.

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

🡪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.

Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

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.

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:

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.

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).

Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

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 use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:

Data Cube Aggregation: Aggregation operation is applied to data for the construction of the data cube.

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.

Numerosity Reduction: This enables to store the model of data instead of whole data, for example: Regression Models.

Dimensionality Reduction: This reduces 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.

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)

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

1. Data collected at regular intervals

2. The gap between the quartiles

3. Use a cross-tab

**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.

**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.

11. Make a comparison between:

1. Data with nominal and ordinal values

🡪

**Data with Nominal Values:**

Nominal data consists of categories or labels that represent different groups or characteristics. These categories have no inherent order or ranking. In other words, they are merely distinct labels without any quantitative meaning. Examples of nominal data include:

- Colors: Red, Blue, Green, etc.

- Gender: Male, Female, Non-binary, etc.

- Countries: USA, Canada, Australia, etc.

In nominal data, you can't perform mathematical operations like addition or subtraction, as there is no numerical significance associated with the categories. You can only use frequencies or percentages to describe the distribution of categories within the dataset.

**Data with Ordinal Values:**

Ordinal data, on the other hand, also consists of categories like nominal data, but there is a clear order or ranking among these categories. The relative differences between the categories are meaningful, but the magnitude of differences is not precisely defined. Examples of ordinal data include:

- Education Levels: High School, Bachelor's Degree, Master's Degree, Ph.D.

- Economic Status: Low Income, Middle Income, High Income

- Rating Scales: 1 (Poor), 2 (Fair), 3 (Good), 4 (Excellent)

In ordinal data, you can determine the order of categories, but you can't perform arithmetic operations like multiplication or division. However, you can use median and other non-parametric statistical tests to analyse and compare such data.

2. Histogram and box plot

🡪 **Histogram**:

A histogram is a graphical representation of the distribution of continuous or discrete data. It consists of a series of bars, where each bar represents a range (or bin) of data values, and the height of the bar represents the frequency or count of data points falling within that range. Histograms are particularly useful for visualizing the overall shape of the data distribution, identifying peaks and clusters, and detecting any potential outliers.

Key features of a histogram:

- Continuous data is grouped into intervals or bins.

- The bars are connected without any gaps since there are no distinct categories.

- Histograms provide insights into the data's central tendency, spread, skewness, and modality.

**Box Plot (Box-and-Whisker Plot)**:

A box plot is a graphical representation that summarizes the distribution of continuous data through five key summary statistics: minimum, first quartile (Q1), median (second quartile, Q2), third quartile (Q3), and maximum. It also includes "whiskers" that extend from the box to show the range of the data, excluding outliers. Box plots are valuable for providing a quick overview of the data's spread, skewness, and presence of outliers.

Key features of a box plot:

- Box plots display the median as a horizontal line inside a rectangular box, with the box's top and bottom edges representing Q3 and Q1, respectively.

- The whiskers extend from the box's edges to the minimum and maximum values within a certain range (usually 1.5 times the interquartile range, IQR). Data points beyond the whiskers are considered outliers and plotted individually.

3. The average and median

🡪 **Average (Mean):**

- Imputation: In machine learning, missing data is a common issue. The average can be used as a method of imputation, where missing values in a feature are replaced with the mean of the available data for that feature. This helps in maintaining the overall distribution of the data and prevents bias caused by missing data.

- Evaluation Metrics: The average is often used as part of evaluation metrics in machine learning models. For example, in regression tasks, the Mean Squared Error (MSE) is calculated, which involves taking the average of squared differences between predicted and actual values. Similarly, the Mean Absolute Error (MAE) uses the average of absolute differences.

**Median:**

- Handling Outliers: In machine learning, datasets may contain outliers that can disproportionately influence the average. Using the median as a measure of central tendency is more robust against the effect of outliers, making it useful for summary statistics in scenarios where the data is skewed or contains extreme values.

- Data Preprocessing: In certain data preprocessing steps, such as feature scaling, the median can be used as a reference point for centering data, especially when using techniques like median scaling.