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

Ans: Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. It involves below steps: Getting the dataset. Importing libraries.

1) Get the Dataset To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset. Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient. So each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML or xlsx file.

2) Importing Libraries In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are: Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:import numpy as nm Here we have used nm, which is a short name for Numpy, and it will be used in the whole program. Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code.

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

Ans:.   Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values.Quantitative research deals with numbers and statistics, while qualitative research deals with words and meanings. Quantitative methods allow you to systematically measure variables and test hypotheses. Qualitative methods allow you to explore concepts and experiences in more detail.research   Qualitative Research

Focuses on testing theories and hypotheses	Focuses on exploring ideas and formulating a theory or hypothesis
Analyzed through math and statistical analysis	Analyzed by summarizing, categorizing and interpreting
Mainly expressed in numbers, graphs and tables	Mainly expressed in words
Requires many respondents	Requires few respondents
Closed (multiple choice) questions	Open-ended questions

Key terms: testing, measurement, objectivity, replicability   Key terms: understanding, context, complexity, subjectivity.

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

Ans:. Primary data is one of the 2 main types of data, with the second one being secondary data. These 2 data types have important uses in research, but in this article, we will be considering the primary data type.

We will introduce you to what primary data is, examples, and the various techniques of collecting primary data.

What is Primary Data?

Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source—where the data originally originates from and are regarded as the best kind of data in research.

The sources of primary data are usually chosen and tailored specifically to meet the demands or requirements of particular research. Also, before choosing a data collection source, things like the aim of the research and target population need to be identified.

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

Ans: ML algorithms will consistently require a lot of data when being trained. Frequently, these ML algorithms will be trained over a specific data index and afterward used to foresee future data, a cycle which you can only expect with a significant amount of effort.Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experienceCommon issues in Machine Learning Inadequate Training Data. ... Poor quality of data.

... Non-representative training data. ... Overfitting and Underfitting. ... Monitoring and maintenance. ... Getting bad recommendations. ... Lack of skilled resources. ... Customer Segmentation.The definition of a ramification is an effect that came from a specific action. An example of a ramification is having trouble getting a job after quitting college. Noun. 1. An arrangement of branches or branching parts.

5. Demonstrate various approaches to categorical data exploration with appropriate Examples.?

Ans:. What is Data Exploration?

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. Using interactive dashboards and point-and-click data exploration, users can better understand the bigger picture and get to insights faster.

Why is Data Exploration Important?

Starting with data exploration helps users to make better decisions on where to dig deeper into the data and to take a broad understanding of the business when asking more detailed questions later. With a user-friendly interface, anyone across an organization can familiarize themselves with the data, discover patterns, and generate thoughtful questions that may spur on deeper, valuable analysis.

Data exploration and visual analytics tools build understanding, empowering users to explore data in any visualization. This approach speeds up time to answers and deepens users' understanding by covering more ground in less time. Data exploration is important for this reason because it democratizes access to data and provides governed self-service analytics. Furthermore, businesses can accelerate data exploration by provisioning and delivering data through visual data marts that are easy to explore and use.

What are the Main Use Cases for Data Exploration?

Data exploration can help businesses explore large amounts of data quickly to better understand next steps in terms of further analysis. This gives the business a more manageable starting point and a way to target areas of interest. In most cases, data exploration involves using data visualizations to examine the data at a high level. By taking this high-level approach, businesses can determine which data is most important and which may distort the analysis and therefore should be removed. Data exploration can also be helpful in decreasing time spent on less valuable analysis by selecting the right path forward from the start.

6. How would the learning activity be affected if certain variables have missing Values? What can be done about it?

Ans:. Introduction

“Data is the fuel for Machine Learning algorithms”.

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. In this blog, you will see how to handle missing values for categorical variables while we are performing data preprocessing. 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.Missing data is an everyday problem that a data professional need to deal with. Though there are many articles, blogs, videos already available, I found it is difficult to find a piece of concise consolidated information in a single place.

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

Ans:

Introduction

The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model. This article describes what is missing data, how it is represented, and the different reasons for the missing data. Along with the different categories of missing data, it also details out different ways of handling missing values with examples.

The following topics are covered in this guide:

What Is Missing Data (Missing Values)?

How Missing Data/Values Are Represented In The Dataset?

Why Is Data Missing From The Dataset?

Types Of Missing Values

Missing Completely At Random (MCAR)

Missing At Random (MAR)

Missing Not At Random (MNAR)

What is a Missing Value?

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.

Below is a sample of the What is a Missing Value?

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.

Below is a sample of the missing data from the Titanic dataset. You can see the columns ‘Age’ and ‘Cabin’ have some missing values.

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

Ans:.

Data Preprocessing in Data Mining

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

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:

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.

(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 :

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: 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:

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)

Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

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”.