**Assignment\_5**

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. Ex – the classification task assigns data to categories, and the clustering task groups data groups data according to similarity.**

**Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data.**

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

**Ans: Quantitative data is anything that can be counted or measured; it refers to numerical data. Qualitative data is descriptive, referring to things can be observed but not measured – such as colors or emotions.**

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:

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

**Ans: Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results. Lack of training data, Slow implementation, Overfitting of Training data are the various causes of machine learning.**

**Ramification is a complex or unwelcome consequence of an action or event.**

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

Ans:

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 can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. We can use two primary methods to solve the error: imputation or removal of data. The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low.**

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

**Ans: Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the row 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.**

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

**Ans: Various data pre-processing techniques are :**

* **Data cleaning/Cleansing – cleaning “dirty” data.**
* **Data Integration – combining data from multiple sources.**
* **Data Transformation – Constructing data cube.**
* **Data Reduction – Reducing representation of data set.**

***Dimensionality reduction :***

**It is a way of converting the higher dimensions dataset into lesser dimension dataset ensuring that it provides similar information.**

**Feature selection :**

**It is the process of reducing the number of input variable when developing a predictive model. It Is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases to improve the performance of the model.**

9.

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

**Ans: The difference between Q3 and Q1 is called the Inter – Quartile Range or IQR. IQR = Q3 – Q1. To detect the outliers using this method, we define a new range, let’s call it decision range, and any data point lying outside this range is considered as outliers and is accordingly dealt with.**

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?

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

1. Data collected at regular intervals

**Data collected at regular intervals help the model more accurate. There will be more data available for training the model.**

2. The gap between the quartiles : **Quartiles are values that divide data into quarters. The data is divided into four segment according to where the number falls on the number line**

3. Use a cross-tab

1. Make a comparison between:

1. Data with nominal and ordinal values : **Nominal data assigns names to each data point without placing it in some sort of order. For example, the results of a test could be classified nominally as a “pass” or “fail”. Ordinal data groups data according to some sort of ranking system. It orders the data.**

2. Histogram and box plot : **Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis**.

3.The average and median**: The mean of the data is found by adding all numbers in the data set and then divide by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest.**