**Assignment\_4**

1.What are the key tasks involved in getting ready to work with machine learning modeling?

**Ans: 1) Collect Data**

**2) Prepare the data**

**3) Choose the model**

**4) Train your machine model**

**5) Evaluation**

**6) Parameter Tuning**

**7) Prediction or Inference**

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

**Ans: Forms of data used in ML could be :**

**Numerical data,**

**Categorical data,**

**Time – series data, and**

**Text.**

3. Distinguish:

1. Numeric vs. categorical attributes:

**Categorical data refers to a data type that can be stored and identified based on the names or labels givien to them. Numerical data refers to the data that is in the form of numbers, and not in any language or descriptice form.**

1. Feature selection vs. dimensionality reduction

**Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.**

4. Make quick notes on any two of the following:

1. The histogram: **A histogram is a variation of a bar chart in which data values are grouped together and put into different classes. This grouping enables you to see how frequently data in eacg class occur in the dataset. The histogram graphically shows the frequency of different data points in the dataset.**

1. Use a scatter plot

**Scatter plots shows how much one variable is affected by another or the relation between them with the help of dots in two dimensions. Scatter plots are very much like graphs in the concept that they use horizontal and vertical axes to plot data points.**

3.PCA (Personal Computer Aid)

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

**Ans: Qualitative researched is exploratory. As opposed to quantitative researched which is conclusive. The reasoning used to synthesise data in qualitative researched is inductive whereas in the case of quantitative research the reasoning is deductive.**

6. What are the various histogram shapes? What exactly are ‘bins'?

**Ans: Various histograms shapes are : Skewed Distribution, double-peaked , plateau or multimodal distribution, edge peak distribution. A histogram displays numerical data by grouping data into “bins” of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in bin**.

7. How do we deal with data outliers?

**Ans: 1. Sort the dataset in ascending order**

**2. calculate the 1st and 3rd quartiles(q1,q3)**

**3. compute IQR = Q3-Q1**

**4. Compute lower bound = (Q1-1.5\*IQR), upper bound = (Q3+1.5\*IQR)**

**5. Loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound mark them as outliers.**

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

**Ans: No matter what value we add to the set, the mean, mediam, and mode will shift by that amount but the range and the IQR will remain the same.**

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

**Ans: In a scatterplot of bivariate data, we use the horizontal or “x-axis” for the explanatory variable and the vertical, or “y-axis” for the response variable. If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line. There is at least one outlier ona scatter plot in most cases.**

10. Describe how cross-tabs can be used to figure out how two variables are related.

**Ans: Cross tabulation is a statistical technique used to display a breakdown of the data by these two variables**.