**ASSIGNMENT**

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

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

**them.**

**3. Distinguish:**

**1. Numeric vs. categorical attributes**

**2. Feature selection vs. dimensionality reduction**

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

**1. The histogram**

**2. Use a scatter plot**

**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?**

**6. What are the various histogram shapes? What exactly are ‘bins&#39;?**

**7. How do we deal with data outliers?**

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

**certain data sets?**

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

**outliers using a scatter plot?**

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

**SOLUTIONS**

1. ***Key tasks involved in getting ready to work with machine learning modeling:***

* ***Data collection and preprocessing***
* ***Data exploration and visualization***
* ***Feature engineering and selection***
* ***Model selection and evaluation***
* ***Hyperparameter tuning and optimization***

1. ***Different forms of data used in machine learning:***

* ***Numerical data (continuous or discrete): e.g. height, age, temperature***
* ***Categorical data (nominal or ordinal): e.g. gender, color, rating***
* ***Text data: e.g. tweets, articles, reviews***
* ***Image data: e.g. pixels in an image***
* ***Audio data: e.g. sound waves in a recording***

1. ***Distinguish:***

* ***Numeric vs. categorical attributes: Numeric attributes are measurable quantities, while categorical attributes are labels or categories. For example, age is a numeric attribute, while gender is a categorical attribute.***
* ***Feature selection vs. dimensionality reduction: Feature selection involves selecting a subset of the original features to use for modeling, while dimensionality reduction involves transforming the original features into a lower-dimensional space. Feature selection is usually done manually, while dimensionality reduction is often automated.***

1. ***Quick notes on:***

* ***The histogram: A histogram is a graphical representation of the distribution of numerical data, showing the frequencies of values in intervals called "bins".***
* ***Scatter plot: A scatter plot is a graphical representation of the relationship between two variables, where each point represents an observation.***
* ***PCA (Principal Component Analysis): PCA is a dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables called principal components, which capture the most important patterns in the data.***

1. ***Why is it necessary to investigate data? Investigating data is necessary to understand its characteristics, identify patterns and relationships, detect anomalies and outliers, and make informed decisions about modeling and analysis. Qualitative and quantitative data can be explored differently, but the overall goal is to gain insights and make meaningful conclusions.***
2. ***Various histogram shapes:***

* ***Symmetric: the data is evenly distributed around the mean***
* ***Skewed right: the tail extends towards the right, indicating more extreme values on the higher end***
* ***Skewed left: the tail extends towards the left, indicating more extreme values on the lower end***
* ***Bimodal: the data has two distinct peaks Bins are intervals of equal width that divide the range of the data into smaller groups for visualization and analysis.***

1. ***Dealing with data outliers: Outliers are extreme values that deviate significantly from the rest of the data. They can be detected and handled in different ways, such as removing them, replacing them with more representative values, or using robust statistical methods that are less sensitive to outliers.***
2. ***Various central inclination measures:***

* ***Mean: the average of all values, sensitive to outliers***
* ***Median: the middle value when the data is sorted, less sensitive to outliers***
* ***Mode: the most frequent value, useful for categorical data The mean can vary too much from the median in certain data sets when there are outliers or the data is skewed.***

1. ***Using a scatter plot to investigate bivariate relationships: A scatter plot can show how two variables are related by plotting their values as points on a 2D graph. It can reveal patterns such as linear or non-linear correlations, clusters, or outliers. Outliers can be identified as points that are far away from the rest of the data.***
2. ***Cross-tabulation, or crosstabs, is a statistical tool used to explore the relationship between two variables. It involves tabulating the frequency or percentage of observations that fall into each combination of categories for the two variables. The resulting table is often called a contingency table.***

***To use crosstabs to figure out how two variables are related, you first need to identify the two variables you are interested in exploring. For example, you may want to know if there is a relationship between gender and smoking status.***

***Once you have identified the variables, you can create a contingency table that displays the frequency or percentage of observations that fall into each combination of categories. For example, you could create a table that shows the number of males and females who smoke and who do not smoke.***

***You can then use various statistical tests to determine whether there is a significant relationship between the two variables. For example, you could use a chi-square test to determine if there is a significant association between gender and smoking status.***

***By analyzing the crosstab results, you can also gain insights into the strength and direction of the relationship between the two variables. For example, if the majority of smokers are male, you might conclude that gender is a significant predictor of smoking status.***

***In summary, cross-tabs can be used to explore the relationship between two variables by creating a contingency table that displays the frequency or percentage of observations that fall into each combination of categories, and then using statistical tests to determine the strength and direction of the relationship between the two variables.***