ML_ASSISMENT_05

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

Sol:
Machine learning involves several key tasks that are crucial for building and deploying effective models. These tasks can be broadly categorized into the following:

- Data Collection and Preparation
- Data Preprocessing
- Data Splitting
- Model Selection and Training
- Model Evaluation: accuracy, precision, recall, F1-score, etc.
- Model Deployment
- Monitoring and Maintenance

Data preprocessing, involves transforming raw data into a format that can be readily used for training machine learning models. It addresses challenges such as noise, missing values, outliers, and inconsistent formats.

The goal is to prepare the data in a way that allows the model to learn patterns effectively.
Data preprocessing steps include data cleaning, transformation, normalization, encoding categorical variables.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.
Sol:
Quantitative and qualitative data are two fundamental types of data used in research and analysis.

- Quantitative Data:
  1. Quantitative data consists of numerical values that can be measured and expressed in terms of quantities.
  2. This type of data focuses on quantities, measurements, and counts.
  3. It is typically used to answer questions that involve **"how much" or "how many."** Quantitative data can be further categorized into **discrete and continuous** data.

- Qualitative Data:
  1. qualitative data, involves non-numeric characteristics and is used to describe qualities, attributes, properties, or characteristics that cannot be measured in terms of quantities.
  2. This type of data is concerned with understanding "why" and "how" things happen, and it delves into the complexities of human behavior, emotions, perceptions, and opinions.
  3. **Examples of Qualitative Data:** Personal narratives , Interviews or transcripts of conversations , Observations of behaviors in natural settings, Textual analysis of literature, art, or media ,Open-ended survey responses.

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

Sol:

Dataset: Student Performance Records

Student ID	Age	Name	Gender	Grade	Height(cm)	Course taken	GPA	Passed
1	18	Alpha	F	12	165	Math, Physics	3.8	Y
2	17	Bravo	M	11	178	Economics,IR	3.2	N
3	19	Charlie	M	12	172	Chemistry ,Economics	3.5	Y
4	18	Delta	F	12	160	History, Math	3.7	Y
5	16	Echo	M	10	175	Biology, hindi	2.9	N
6	17	foxtrot	M	12	174	Hindi, sanskrit	3.5	Y

Explanation of Attributes:

1. **Student** ID: Unique identifier for each student.
2. **Name**: Name of the student (Categorical data type).
3. **Age**: Age of the student (Numerical data type).
4. **Gender**: Gender of the student (Categorical data type).
5. **Grade**: Grade level of the student (Ordinal data type).
6. **Height** (cm): Height of the student in centimeters (Continuous numerical data type).
7. **Courses** Taken: List of courses taken by the student (Categorical data type).
8. **GPA**: Grade Point Average of the student (Continuous numerical data type).
9. **Passed** Exam: Whether the student passed an exam (Binary categorical data type: Yes/No)

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

Sol:

- Incomplete Data:
  Cause: Missing values, incomplete records, or fields not being properly filled.

- Imbalanced Data:
  Cause: Unequal distribution of classes in the dataset.

- Noisy Data:
  Cause: Outliers, errors, or inconsistencies in the data.

- Duplicate Data:
  Cause: Repeated instances in the dataset.

- Biased Data:
  Cause: Systematic errors in data collection, leading to underrepresentation or overrepresentation of certain groups.

- Out-of-Distribution Data:
  Cause: Data that is significantly different from the training data distribution.

- Data Leakage:
  Cause: Inclusion of information from the target variable in the input features, usually due to errors in data preprocessing.

- Small Data:
  Cause: Insufficient amount of data for model training.

- Changing Data Distributions:
  Cause: Drastic shifts in data distributions over time.

- Labeling Errors:
  Cause: Mistakes made during the process of labeling data.

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

- Frequency Distribution: displays the count of each category within a categorical variable.

- Bar Plots: visually represent the frequency distribution of categorical variables.

- Pie Charts:show the proportion of each category within the whole categorical variable.

- Cross-tabulations (Crosstabs):show the distribution of one categorical variable with respect to another categorical variable.

- Heatmaps: represent the frequency or proportion of categories using color intensity.

- Chi-square test :measures the association between two categorical variables to determine if they are independent or related

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

Sol:
Impact of Missing Values:

- Biased Insights: Missing values can introduce bias in the analysis

- Reduced Sample Size: reduce the effective sample size, potentially limiting the statistical power of analyses.

- Incorrect Relationships: affects the relationships between variables, leading to distorted correlations.

- Model Performance: Missing values can affect the performance of predictive models

Dealing with Missing Values:

- Identify Missingness

- Imputation: Imputation involves filling in missing values with estimated or predicted values.
  Common imputation methods include :
  - mean
  - median,
  - mode imputation,
  - regression imputation,
  - k-nearest neighbor imputation

- Dropping Missing Data: In cases where missing values are limited to a small percentage of the dataset, you might consider removing rows or columns with missing values.

- Advanced Techniques: For predictive modeling, techniques like multiple imputation using models that can handle missing data (e.g., decision trees, random forests)

- Collect More Data: collecting more data to reduce the proportion of missing values can improve the quality of analysis.

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

Sol: Imputation: Imputation involves filling in missing values with estimated or predicted values.
Common imputation methods include :

- mean
- median,
- mode imputation,
- regression imputation,
- k-nearest neighbor imputation

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

Sol:
pre-processing techniques are:

- Data Cleaning
- Data Transformation
- Feature Selection
- Encoding Categorical Data
- Handling Imbalanced Data:
- Handling Outliers
- Data Integration: Combining data from multiple sources into a single dataset for analysis.
- Feature Engineering: Creating new features by combining or transforming existing features to provide more meaningful information to the model.
- Data Reduction: Reducing the size of the dataset while preserving its important characteristics

Dimensionality Reduction:
Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while preserving as much relevant information as possible. This is particularly useful when dealing with high-dimensional data, as it can help to improve the efficiency and interpretability of machine learning models.

One common technique for dimensionality reduction is Principal Component Analysis (PCA).

Function Selection:
Function selection, also known as feature selection, is the process of choosing a subset of relevant features from the original set of features. It aims to improve model performance by reducing overfitting, reducing computational complexity, and enhancing interpretability.

9.
i. What is the IQR? What criteria are used to assess it?
ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can boxplots be used to identify outliers?

Sol:
i. The Interquartile Range (IQR) is a statistical measure that quantifies the spread or dispersion of a dataset.
It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the dataset.

Quartiles are values that divide a dataset into four equal parts, with Q1 representing the value below which 25% of the data falls, Q3 representing the value below which 75% of the data falls, and the median (Q2) representing the value that separates the lower 50% from the upper 50% of the data.

Mathematically, the formula to calculate the IQR is:**IQR = Q3 - Q1**

ii. A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that provides a summary of its distribution, central tendency, and spread. It is particularly useful for visualizing the distribution of numerical data and identifying potential outliers.

A box plot consists of several key components, one of them is (whiskers: The whiskers extend from the edges of the box to the minimum and maximum values within a certain range. The length of the whiskers can vary depending on the data and any potential outliers)

The length of the lower whisker will surpass the upper whisker in length when the data distribution is highly skewed to the left (negatively skewed).
In such cases, the lower whisker will extend further towards the minimum value, indicating that the majority of the data is concentrated towards the higher end of the dataset, with a few low values pulling the median and box downwards.

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

Sol:

1. **Data collected at regular intervals:**

- Data collected at regular intervals refers to information gathered at consistent and predetermined time or space intervals. This approach is often used in various fields, including scientific research, finance, and environmental monitoring.
- Regular interval data collection enables the tracking of trends, patterns, and changes over time or space, allowing for better understanding and analysis of underlying processes.
- **Examples** include daily temperature measurements, hourly stock prices, monthly rainfall records.
- **Advantages**: Provides a structured and systematic approach to data collection, facilitates trend analysis.
- **Challenges**: May overlook irregular events or occurrences that fall outside the regular intervals, might require interpolation or extrapolation for missing data points, and could lead to oversimplification of complex dynamics.

2. **The gap between the quartiles:**

- The gap between the quartiles refers to the difference between the third quartile (Q3) and the first quartile (Q1) in a dataset's distribution. It is also known as the interquartile range (IQR).
- The IQR is a measure of the spread or dispersion of the middle 50% of the data points. It quantifies the spread of data around the median and is less affected by extreme values (outliers) than the full range.
- Calculating IQR: IQR = Q3 - Q1
- The IQR is used in box plots to define the length of the box, extending from Q1 to Q3. It also helps determine the length of the whiskers and identify potential outliers.
- A larger IQR indicates a wider spread of data points within the middle range of the dataset, while a smaller IQR indicates a more concentrated distribution.

11. Make a comparison between:

1. Data with nominal and ordinal values

nominal	ordinal values
● Nominal values are categorical data that represent different categories or groups without any inherent order or ranking.	● Ordinal values are also categorical data, but they come with a specific order or ranking. While the differences between categories are not necessarily uniform, there is a clear sequence of preference or significance.
● Examples:colors, genders, types of animals, etc.	● Examples : educational levels (e.g., high school, bachelor's, master's, PhD), rating scales (e.g., poor, fair, good, excellent), etc.

2. Histogram and box plot

Histogram	box plot
● A histogram is a graphical representation of the distribution of a dataset.	● A box plot is a graphical summary of a dataset's distribution. ● It displays the median, quartiles (25th and 75th percentiles), and potential outliers in the data.
● Histograms are often used for visualizing continuous or discrete numerical data, but they can also be used for ordinal data where the order matters.	● Box plots are particularly useful for comparing the distributions of multiple datasets or visualizing the spread and skewness of a single dataset.
● It displays the frequency or count of data points within certain ranges or "bins" on the x-axis. The height of each bar in the histogram corresponds to the frequency of data points falling within that bin.	● The "box" represents the interquartile range (IQR), which contains the middle 50% of the data. ● The "whiskers" extend from the box to the minimum and maximum data values within a certain range (often defined by 1.5 times the IQR). Data points outside this range are typically shown as individual points and considered potential outliers

3. The average and median

average	median
● The average, also known as the mean, is calculated by summing up all the values in a data set and then dividing by the number of values. Mathematically, if you have a data set with values x1,x2,…,xn , the average ● (μ) is calculated as μ= nx1+x2+…+xn	● The median is the middle value in a data set when it is arranged in numerical order. ● If the data set has an **odd number** of values, the median is simply the middle value. ● If the data set has an **even number** of values, the median is the average of the two middle values.