UNIT 2 ASSIGNMENT

Managing Your Data in ML

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Week 2 Written Portion

Building a Modeling Dataset

Answer the questions below about building a model dataset and understanding your data through analysis and visualization.

## Questions:

1. What does it mean to have a “modeling dataset”?

|  |
| --- |
| A modeling dataset refers to a dataset that has been prepared and transformed into a suitable format for use in machine learning models. This dataset should be clean, consistent, and free of errors, with all necessary features and variables included. The data should be in a format that can be easily processed by the chosen machine learning algorithm, with any necessary transformations, encoding, or scaling applied. |

1. What steps would you take with a raw dataset to end up with a modeling dataset?

|  |
| --- |
| Data Cleaning: Remove any missing or duplicate values, handle outliers, and correct errors in the data.  Data Transformation: Convert data types, perform feature scaling, and apply necessary transformations to make the data suitable for modeling.  Feature Engineering: Extract relevant features from the data, create new features if necessary, and select the most informative features for the model.  Data Encoding: Encode categorical variables into numerical formats using techniques like one-hot encoding or label encoding.  Data Split: Split the dataset into training, validation, and testing sets to evaluate the model's performance.  Data Visualization: Visualize the data to understand relationships, distributions, and patterns, which can inform feature selection and model choice. |

1. What is the difference between nominal data and ordinal data? Explain with an example.

|  |
| --- |
| Nominal Data: Nominal data is categorical data where each category has no inherent order or ranking. Examples include gender (male/female), colors (red, blue, green), or occupations (teacher, engineer, doctor).Ordinal Data: Ordinal data is categorical data where each category has a natural order or ranking. Examples include education levels (high school, bachelor's, master's), income levels (low, medium, high), or satisfaction ratings (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied). |

1. Why is data visualization an important part of the data preparation process?

|  |
| --- |
| Data visualization is crucial in the data preparation process because it helps understand data distributions by revealing patterns, outliers, and relationships that inform feature selection and model choice, identify correlations between features to detect important relationships and potential multicollinearity issues, detect outliers and anomalies that may impact model performance, and effectively communicate insights and findings to stakeholders, facilitating better decision-making. |

1. What is an outlier?

|  |
| --- |
| An outlier is a data point that significantly deviates from the majority of the data, often indicating an error, anomaly, or unusual behavior. Outliers can impact model performance and should be handled appropriately, either by removing them or using robust algorithms that can handle such deviations. |

1. Name a few libraries used for data analysis and visualization and explain when you would use each library.

|  |
| --- |
| Matplotlib: A versatile library for creating static plots, suitable for publication-quality charts and graphs.  Seaborn: Built on top of Matplotlib, Seaborn provides high-level, visually appealing plots for statistical analysis and data exploration.  Plotly: Ideal for interactive visualizations, Plotly creates web-based, interactive plots that can be shared easily. |

*To submit this assignment, please refer to the instructions in the course*.