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 specific dataset that is used for building, training, and evaluating a machine learning model. It is a collection of input features and corresponding target variables or labels that serve as the basis for training and evaluating the model's performance. |

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

|  |
| --- |
| To transform a raw dataset into a modeling dataset, the following steps can be taken:   * Data Cleaning: Handle missing values, remove duplicate records, and correct any inconsistencies or errors in the data. * Data Integration: Combine multiple data sources if necessary, ensuring the compatibility and consistency of the data. * Feature Selection: Identify relevant features that are most likely to contribute to the model's predictive power and remove irrelevant or redundant features. * Feature Engineering: Create new features or transform existing features to better represent the underlying patterns and relationships in the data. * Data Transformation: Normalize or scale numerical features and encode categorical features into numerical representations suitable for modeling. * Splitting into Training and Testing Sets: Divide the dataset into separate subsets for training the model and evaluating its performance. |

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

|  |
| --- |
| Nominal data and ordinal data are both types of categorical data, but they differ in terms of the level of measurement and the relationships between categories:   * Nominal Data: Nominal data represents categories or labels without any inherent order or ranking. Each category is considered equal and distinct. For example, hair color (e.g., red, brown, blonde) or country of origin (e.g., USA, UK, Germany). * Ordinal Data: Ordinal data represents categories with a specific order or ranking. The categories have a meaningful relationship in terms of their relative position. For example, educational levels (e.g., high school, bachelor's degree, master's degree) or satisfaction ratings (e.g., very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, very satisfied). |

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

|  |
| --- |
| Data visualization is an important part of the data preparation process for several reasons:   * It helps in understanding the data: Visualization provides insights into the distribution, patterns, and relationships within the dataset, allowing for a better understanding of the data's characteristics. * It aids in identifying data issues: Visualizations can reveal outliers, missing values, skewed distributions, or inconsistent patterns that require further investigation or data cleaning. * It facilitates feature selection: Visualizations can assist in selecting relevant features by identifying correlations or dependencies between variables and the target variable. * It supports decision-making: Visualizations provide a clear and intuitive way to communicate findings and insights to stakeholders, facilitating informed decision-making. |

1. What is an outlier?

|  |
| --- |
| An outlier is an observation that significantly deviates from other observations in a dataset. It is an extreme value that lies outside the expected range of values. Outliers can occur due to various reasons such as measurement errors, data entry errors, or genuinely unusual instances in the data. Outliers can have a significant impact on data analysis and modeling results, as they can skew statistical measures and affect the performance of machine learning algorithms. |

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

|  |
| --- |
| Some libraries used for data analysis and visualization are:   * Pandas: Pandas provides data structures and functions for data manipulation and analysis. It is widely used for data preprocessing, cleaning, filtering, and transformation tasks. * NumPy: NumPy is a library for numerical computing in Python. It provides efficient data structures and mathematical functions for working with arrays and matrices, which are fundamental for data analysis and modeling. * Matplotlib: Matplotlib is a plotting library that provides a wide range of customizable visualizations, including line plots, scatter plots, histograms, and bar plots. It is suitable for creating static, publication-quality visualizations. * Seaborn: Seaborn is a higher-level data visualization library built on top of Matplotlib. It offers additional statistical |

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