1_Introduction

January 14, 2025

[ ]:
```
'''
   Data Preprocessing -->

   Data preprocessing in machine learning (ML) is the process of transforming␣
↪raw data into a clean
   and structured format that can be effectively used by machine learning␣
↪models. It is a crucial step
   in the ML pipeline, as raw data often contains inconsistencies, missing␣
↪values, noise, or irrelevant
   features that can negatively impact model performance
'''
```

[ ]:
```
'''
   Key Steps To Clean Data -->

   1.  Data Cleaning ->

   +   Handling Missing Values ->
       Filling in missing data with mean/median values, or removing rows/
↪columns with missing data

   +   Removing Noise ->
       Filtering out outliers or errors that can distort the analysis

   +   Correcting Inconsistencies ->
       Fixing data entry errors or formatting issues

   2.  Data Integration ->

       Combining data from multiple sources or tables into a cohesive dataset,␣
↪ensuring that relationships
       between data points are preserved

   3.  Data Transformation ->

   +   Normalization/Standardization ->
```

```
       Scaling data to a standard range (e.g., 0 to 1) or to have a mean of␣
↪zero and a standard deviation
       of one, making it easier for the model to learn patterns

   +   Encoding Categorical Variables ->
       Converting categorical data into numerical format using techniques like␣
↪one-hot encoding or label
       encoding

   +   Feature Engineering ->
       Creating new features that might help the model perform better by␣
↪combining or transforming
       existing features

   4.  Data Reduction ->

   +   Dimensionality Reduction ->
       Reducing the number of features using techniques like Principal␣
↪Component Analysis (PCA) to
       focus on the most important features and reduce the complexity of the␣
↪model

   +   Feature Selection ->
       Choosing only the most relevant features for model training to avoid␣
↪overfitting and improve
       model interpretability

   5.  Data Splitting ->

       Dividing the dataset into training, validation, and test sets to␣
↪evaluate model performance and
       generalization ability
'''
```