# Data Preprocessing

Group: B
Name: Tykea Ly

1. **Data Cleaning**

   Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, weird, duplicate, or incomplete data within a dataset.

   Example:

   Imagine you have a spreadsheet with info on your classmates—like their names, ages, grades, favorite subjects, and activities. But when you look at it, you notice some problems:

   - Some people didn’t fill in their age or favorite subject.
   - Someone misspelled their name.
   - Someone wrote their age as "200," which doesn’t make sense.
   - Data cleaning is just fixing these issues so that all the info is correct, complete, and easy to understand.

2. **Dimensionality Reduction**

   Dimensionality Reduction is an act of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information. In simple terms when you reduce the dimensions, you’re simplifying the data by keeping only the most important parts. This helps make it easier to analyze, understand, or visualize.

   Example: Dropping columns that do not contribute to exploration of the data.

3. **Feature Engineering**

   Features are the individual columns or pieces of information in your dataset. Feature engineering involves taking the existing data and transforming or combining it to create new features that can provide more insight or improve the performance of a model.

   Example:

   - creating new features: using existing features to create new features
   - transforming features:

- binning or grouping: If you age you can group them into “teen”, “adult”, etc..
- scaling or normalizing: This involves adjusting the range of your data, like turning all your numbers into a scale from 0 to 1.

4. **Imbalanced Data/Data Augmentation**

   Imbalanced data occurs when the classes in your dataset are not represented equally. To handle imbalanced data, you can use techniques like oversampling where you can duplicate the minority class or undersampling the majority class. Data augmentation is a technique used to increase the size of your dataset by creating new data points from existing ones.

   For example:

   - Images: You can flip, rotate, or crop images to create new variations.
   - Texts: You could rephrase sentences or use synonyms to create variations of the text data.

5. **Data Transformation**

   Data transformation is the process of converting data from one format or structure into another to make it more suitable for analysis or modeling.

   - Normalization: Adjusting the data so that it fits within a specific range, usually between 0 and 1.
   - Standardization: This technique transforms data to have a mean of 0 and a standard deviation of 1. It’s particularly useful when the features have different units or scales.
   - Encoding Categorical Variables: Converting categorical data (like "red," "blue," "green") into numerical format.
   - Handling Missing Values: Transforming the dataset by either filling in missing values (imputation) with the mean, median, mode, or another method, or by removing rows/columns that contain missing data.

6. **Sampling Data**

   Sampling data is a technique used to select a subset of data from a larger dataset to analyze or draw conclusions about the whole.

   Types of Sampling:

- **Random Sampling**: Every member of the population has an equal chance of being selected.
- **Systematic Sampling**: Members are selected at regular intervals from a randomly ordered list.
- **Stratified Sampling**: The population is divided into subgroups (strata), and random samples are taken from each subgroup.
- **Cluster Sampling**: the population is divided into clusters, usually geographically, and entire clusters are randomly selected.
- **Convenience Sampling**: Members are selected based on ease of access rather than random selection.