# Where to Collect data



# Import Datasets from Kaggle.

* Install Kaggle package
* Import the data
* Unzip the file
* Use it

# Handling Missing Values

1. Imputation (Mean, Median, and Mode)
2. Dropping Data

# Data Standardization

<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

Graphical user interface, text, application

Description automatically generated

# Label Encoding

* Converting the labels to into numerical form using Label Encoder

Graphical user interface, text, application, email

Description automatically generated

# Train Test Split in Machine Learning

Diagram

Description automatically generatedDiagram

Description automatically generated with medium confidence

Text

Description automatically generated

# Graphical user interface, text, application Description automatically generatedHandling Imbalance Datasets

* Compare the total of two data
* Cut the imbalance data
* Create new data from the cut data by stacked both data

# Feature Extraction of Text Data: Tf-idf Vectorizer

Graphical user interface, text, application, email

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

* Fill null with empty string (“ ”)
* Merging the author’s name and news title
* Separating the data & label
* Stemming:

Stemming is the process of reducing a word to its Root word

example: actor, actress, acting --> act

* Separating the data & label (X and Y)
* Convert the textual data to Feature Vectors

# Numerical Datasets Pre-Processing – Use case

Section B -> G

# Text Datasets Pre-Processing – Use case

Section C -> H