MINISTRY OF
EDUCATION
IMAM ABDULRAHMAN
BIN
FAISAL UNIVERSITY
COLLEGE OF COMPUTER
SCIENCE
INFORMATION &
TECHNOLOGY

وزارة التعليم
جامعة الإمام
عبد الرحمن بن
فيصل
كلية علوم
الحاسب
وتقنية المعلومات

جامعة الإمام عبدالرحمن بن فيصل
IMAM ABDULRAHMAN BIN FAISAL UNIVERSITY

**<u>Lab 2</u>**

**ARTI 308 – Machine Learning**
*Academic Year (2025/2026) – 2nd Semester*

# Lab 2: Identifying ML Problems, Selecting Open Datasets, and Drawing a Methodology Diagram

## Learning Goals:

By the end of this lab, you will be able to:

1. Identify whether a problem is Regression**,** Classification, or Clustering
2. Select a suitable dataset from open sources
3. Define the target variable (if supervised)
4. Write a clear problem statement
5. Draw a methodology diagram

## Part 1: Choosing an Open Dataset

Start by selecting one tabular dataset from an open-source platform. You may use any of the following resources:

- Kaggle: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu
- OpenML: https://www.openml.org
- Government open-data portals (for example: data.gov)
- Etc…

Your dataset should be in CSV or Excel format and should be suitable for a machine learning task such as prediction or pattern discovery.

**Examples of suitable datasets:**

- Predicting house prices based on size and location

- Classifying emails as spam or not spam
- Predicting whether a customer will churn
- Grouping customers based on behavior (clustering)
- Avoid image, audio, or text-heavy datasets for this lab.

## Part 2: Defining the Machine Learning Problem

Once you select a dataset, your next task is to clearly define the machine learning problem**.**

Ask yourself:

- Is this a **regression**, **classification**, or **clustering** problem?
- Is there a target variable?
- What is the model expected to learn or predict?

Write a short problem description in your own words.

## Part 3: Loading and Inspecting the Dataset in Python

Create a Jupyter Notebook and load your dataset using Pandas.

At this stage, you only need to:

- Load the dataset
- Display its shape
- Preview the first few rows
- Check column names and data types

## Part 4: Designing the Methodology Diagram

The final task in this lab is to create a methodology diagram that explains your machine learning workflow. This diagram should visually represent how the project would proceed from start to finish.

A typical methodology includes steps such as:

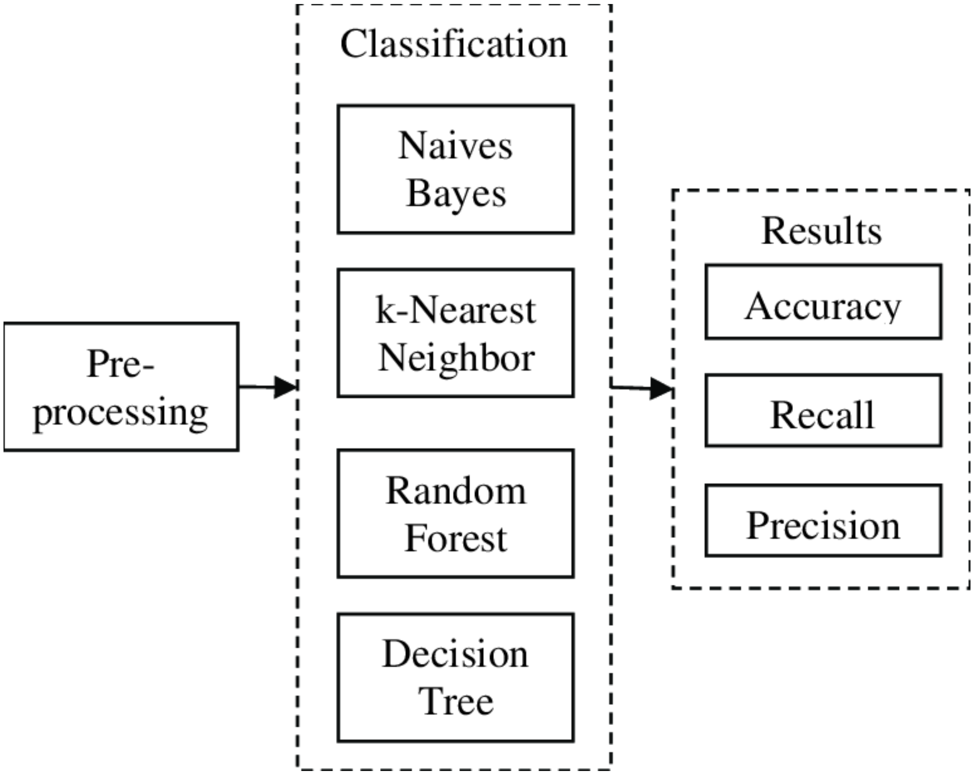
- Dataset selection
- Data loading

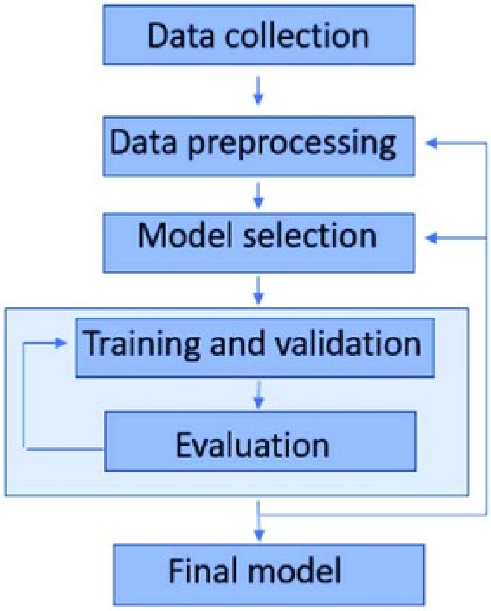
- Data preprocessing
- Train/test split
- Model training
- Model evaluation
- Results

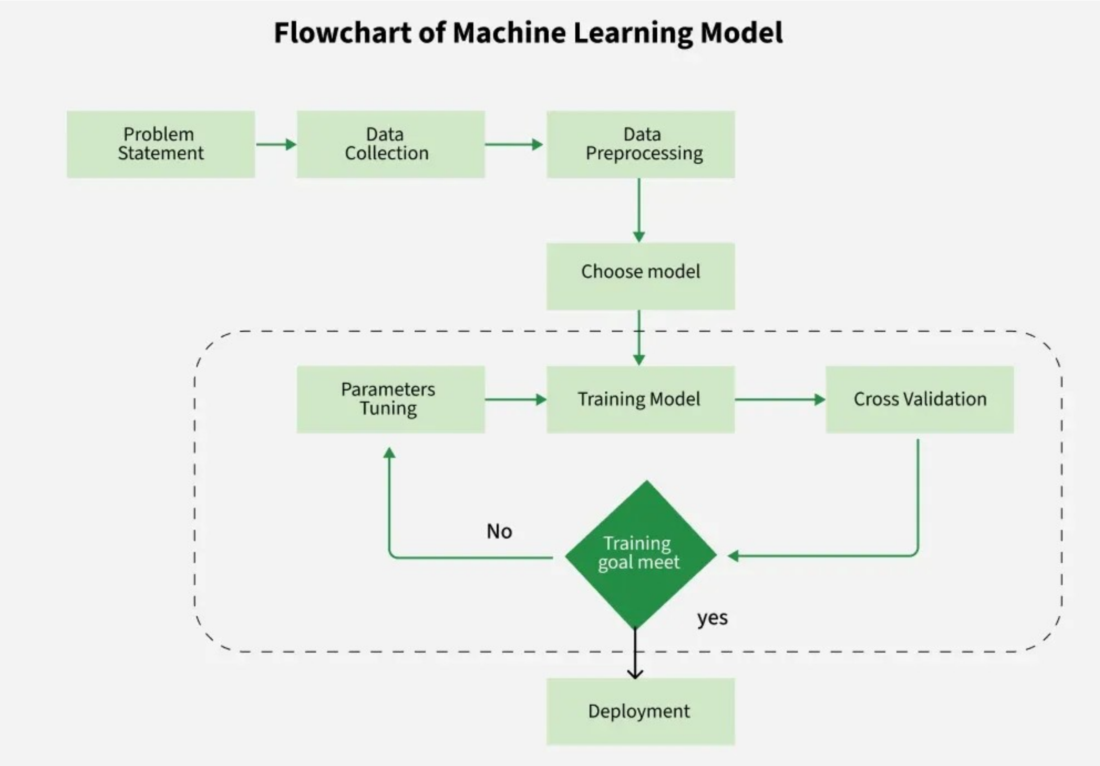
You may create the diagram using any free tool, such as:

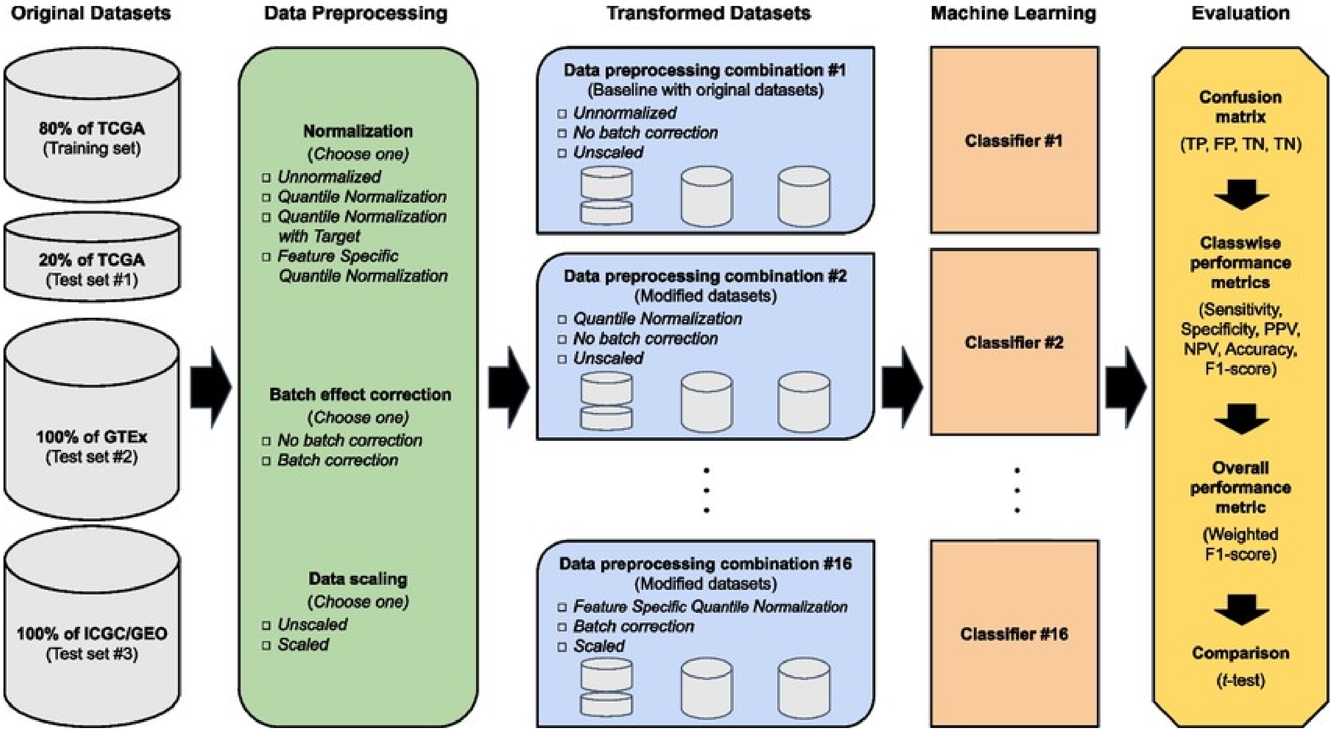
- draw.io: https://www.drawio.com
- Canva: https://www.canva.com
- Figma: https://www.figma.com
- Napkin: https://www.napkin.ai
- PowerPoint
- Etc…

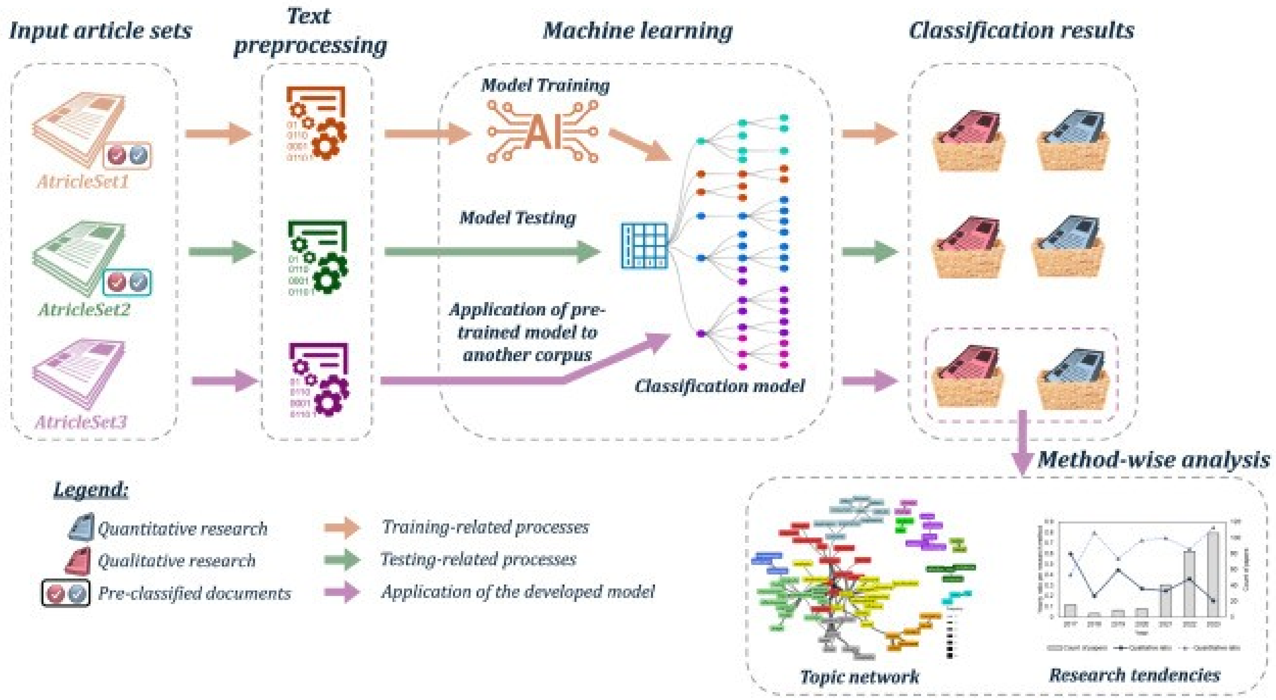
Export your diagram as a PNG or PDF and save it in your GitHub repository.




Data collection
Data preprocessing
Model selection
Training and validation
Evaluation
Final model


Flowchart of Machine Learning Model
Problem Statement
Data Collection
Data Preprocessing
Choose model
Parameters Tuning
Training Model
Cross Validation
No
Training goal meet
yes
Deployment


Original Datasets
Data Preprocessing
Transformed Datasets
Machine Learning
Evaluation
80% of TCGA
(Training set)
20% of TCGA
(Test set #1)
100% of GTEx
(Test set #2)
100% of ICGC/GEO
(Test set #3)
Normalization
(Choose one)
□ Unnormalized
□ Quantile Normalization
□ Quantile Normalization with Target
□ Feature Specific Quantile Normalization
Batch effect correction
(Choose one)
□ No batch correction
□ Batch correction
Data scaling
(Choose one)
□ Unscaled
□ Scaled
Data preprocessing combination #1
(Baseline with original datasets)
□ Unnormalized
□ No batch correction
□ Unscaled
Data preprocessing combination #2
(Modified datasets)
□ Quantile Normalization
□ No batch correction
□ Unscaled
Data preprocessing combination #16
(Modified datasets)
□ Feature Specific Quantile Normalization
□ Batch correction
□ Scaled
Classifier #1
Classifier #2
Classifier #16
Confusion matrix
(TP, FP, TN, TN)
Classwise performance metrics
(Sensitivity, Specificity, PPV, NPV, Accuracy, F1-score)
Overall performance metric
(Weighted F1-score)
Comparison
(t-test)


Input article sets
Text preprocessing
Machine learning
Classification results
Model Training
AI
Model Testing
Application of pre-trained model to another corpus
Classification model
AtricleSet1
AtricleSet2
AtricleSet3
Method-wise analysis
Topic network
Research tendencies
Legend:
Quantitative research
Qualitative research
Pre-classified documents
Training-related processes
Testing-related processes
Application of the developed model

## Assignment 2 - Instructions

Your work for this lab must be submitted through GitHub only**.**

Your GitHub repository must contain:

- A short written summary describing the dataset and the machine learning problem
- A Jupyter Notebook that loads the dataset and displays basic information
- The methodology diagram saved as an image or PDF

Submit only the GitHub repository link through the Assignment 2 submission link on Blackboard.

Do not upload files or screenshots directly to Blackboard. Make sure your repository is accessible.