# 1 Introduction

The goal of this project is to apply machine learning (ML) techniques introduced in class to solve real-world problems. Specifically, you will develop a functional advice or action-suggestion system. Your system can operate in various domains, such as healthcare (e.g., suggesting a medical specialist based on symptoms), lifestyle (e.g., providing fashion advice), or general guidance. The project emphasizes not just implementation but a deep understanding of **HOW** and **WHY** different ML methods work and how they can be integrated to create a more robust system.

You will incrementally build this system over **five phases**:

1. **Problem Understanding and Data Exploration**
2. **Supervised Learning**
3. **Unsupervised Learning**
4. **Integrating Generative AI**
5. **Final Submission and Presentation**

Throughout these phases, you will use Python or R to implement appropriate machine learning techniques, evaluate and compare different models, and refine your system to achieve accurate and meaningful results.

Each group will organize their work in a **GitHub repository**, with deliverables from each phase added incrementally. The repository should be structured to allow the instructor to easily review submissions. The final phase will include a short presentation summarizing your work.

# 2 Phase by Phase guide

Below is a detailed description of each phase, including the required deliverables

## 2.1 Phase1: Problem Understanding and Data Exploration

**[Deadline: October 16, 2025]**

The aim of this phase is to define the problem and determine the scope of the project. Specifically, you will explain the type of advice system you intend to build (e.g., medical, fashion, general guidance). In addition, you will collect and explore a dataset relevant to solving this problem. The data may consist of categorical variables, Boolean data, text, or figures. Please note:

- The dataset should contain at least **several hundred rows** or more.
- You can explore publicly available datasets from websites such as:
    - Kaggle
    - GitHub
    - Hugging Face Datasets

After selecting the problem and identifying a suitable dataset, you will preprocess the data. This includes tasks such as handling missing values, extracting key features (if necessary), and visualizing relationships (e.g., symptoms vs. advice, or actions vs. context). Ensure that the dataset you choose can be successfully accessed and read using Python or R programming languages.

**Note**: Before starting work on the data, you have to get your instructor's approval to save time and effort (From the Group Formation Google sheet).

**Deliverables**: Each group is required to submit a GitHub project link that contains the following:

1. **A README.md file with:**:
    - Group member names and IDs.
    - Project title and motivation (why this problem/dataset?)
2. **A /Dataset folder containing the raw data file:**
3. **A Jupyter Notebook (Phase1_Data_Exploration.ipynb) that includes:**

- **The Dataset Goal & Source**: The purpose of the dataset (e.g., requirements classification, defect prediction, etc.). and a URL to its source.
- **General Information**: Number of observations/features, data types, and description of the target variable/classes.
    - **Summary & Visualization:** Statistical summaries (mean, variance, etc.), visualizations of variable distributions, missing value analysis, and class imbalances.
- **Preprocessing Techniques**: Document the preprocessing steps applied to the data with justifications. This may include variable transformation, discretization, value or variable removal, and normalization.

## 2.2 Phase2: Build a Supervised Learning Model

**[Deadline: November 2, 2025]**

In this phase, you will design a supervised model to predict advice based on user input (e.g., predict the type of doctor based on symptoms). To achieve this, you must use at least two different supervised machine learning algorithms (e.g., neural network and SVM) on your dataset. The selection of these models should be clearly justified, as not all algorithms are equally effective for every problem. Avoid choosing algorithms randomly or simply because they are the easiest to implement.

After selecting the algorithms, compare their performance and discuss the results. Highlight which algorithm performed best and explain the key findings observed during your analysis of the results.

**Deliverables**: An updated GitHub repository containing:

1. **Corrections/updates based on feedback from Phase 1.**
2. **A Jupyter Notebook in a /Supervised_Learning folder (Phase2_Supervised_Learning.ipynb) that includes:**
    - **Algorithm Selection & Justification:** Clear reasoning for choosing the two (or more) models (e.g., Neural Network, SVM, Decision Tree, etc.).
    - **Implementation:** Well-commented code for model training, hyperparameter tuning (if attempted), and prediction.

- **valuation & Comparison:** Use of appropriate metrics (e.g., Accuracy, Precision, Recall, F1-Score, ROC-AUC) and techniques (e.g., train-test split, cross-validation) to compare model performance.

- **Results Interpretation:** A discussion on which model performed best, why, and the key findings from the results.

## 2.3 Phase3: Apply Unsupervised Learning

**[Deadline: November 16, 2026]**

In this phase, you will use clustering to group data and enhance recommendations. Clustering is sometimes applied independently to solve specific problems, but it is also often used to enhance and improve other algorithms, such as GAI or supervised learning. From the instructor's point of view, the goal is for you to learn how to apply clustering algorithms and understand their application. However, as a learner, your objective is to find the best way to integrate this approach to enhance the model you previously developed.

You must apply at least one unsupervised learning algorithm to your dataset, providing a clear justification for your choice. Then, compare and discuss the results using different evaluation methods and metrics, such as the **Silhouette Coefficient**, **Total Within-Cluster Sum of Squares**, and **BCubed Precision and Recall**. Use visual representations wherever possible and strive to interpret the results and understand the algorithm's performance.

**Hint**: Ensure you remove the class label from your data before applying clustering.

**Deliverables**: Each group is required to submit a GitHub project link that contains the following:

1. **Corrections/updates based on feedback from Phase 2**.

2. **A Jupyter Notebook in a /Unsupervised_Learning folder (Phase3_Unsupervised_Learning.ipynb) that includes:**

    - **Algorithm Application:** Implementation of at least one clustering after removing the class label.

    - **Evaluation & Visualization:** Use of metrics like Silhouette Score, Within-Cluster-Sum-of-Squares, and visualizations to assess cluster quality.

    - **Integration & Insight:** A clear explanation of how the discovered clusters could be used to improve the supervised model or provide new insights (e.g., creating user profiles, refining advice). If integration is not feasible, a detailed and justified explanation why.

## 2.4 Phase4: Integrate Generative AI

**[Deadline: November 30, 2026]**

In this phase, you will integrate Generative AI into your system. The integration aims to enhance the system by providing detailed advice or explanations based on user input. To achieve this, you will use one of the Generative AI models, such as **LAMA** or **GPT**. You must apply at least two templates and demonstrate the differences between their outcomes.

**Deliverables**: An updated GitHub repository containing:

1. **Corrections/updates based on feedback from Phase 2**.
2. **A Jupyter Notebook in a /Generative_AI folder (Phase4_Generative_AI.ipynb) that includes:**
    - **Implementation:** Code that integrates a Generative AI model (e.g., GPT, LLaMA) via an API. At least two different prompt templates must be implemented and tested
    - **Template Comparison & Analysis:** Demonstration of the outputs from the different prompts, with a clear analysis of their differences in quality, detail, and relevance.
    - **Justification:** A reasoned argument for selecting the final prompt template to be used in the system.

## 2.5 Phase 5: Final submission and presentation

**[Deadline: December 7, 2026]**

This phase aims to consolidate all work into a final, polished deliverable and effectively communicate your project's journey and findings.

**Deliverables**:

1. **Final GitHub Repository:** A complete, well-organized repository containing all deliverables from Phases 1-4, incorporating all feedback.
2. **Team Presentation:** A 5-10 minute presentation summarizing:
    - Problem definition and dataset.
    - Key steps and findings from each phase.

- Overall results, challenges, and lessons learned.
- **All team members must participate.**

# 3 General information

## 3.1 Group Work:

You will work in groups of 5 students. For each phase, you must clearly document how the tasks are distributed among your team members. During the final presentation, questions will be asked randomly to any team member, so every member must be fully aware of the entire project.

For weekly submissions, only one team member is responsible for submitting the GitHub link via LMS.

## 3.2 Submission Instructions

As mentioned above, each group will submit their work as GitHub links, except for the final phase, which will include both the presentation and the GitHub link. Please name the GitHub project as: **"SW485-Project-Group#"**.

Ensure that all work is your own. **Plagiarism is strictly prohibited**, and a score of zero will be given to any submissions with high levels of similarity. **If you use any assistive tools like ChatGPT**, clearly explain how you utilized them. It is acceptable to use such tools for enhancing your writing for example, but not for creating work entirely from scratch. Please be cautious!

## 3.3 Marks Distribution

This project contributes **25%** of your course grade, distributed across the 5 phases as follows:

| **Phase#** | Mark |
|---|---|
| Phase 1: Dataset selection and preprocessing | 5 |
| Phase 2: Supervised learning | 5 |
| Phase 3: Unsupervised learning | 5 |
| Phase 4: Integrate Generative AI | 5 |
| Phase 5: Final GitHub project submission + Presentation | 5 |
| **Project total** | **25** |