# In-Class Exercise: Binary Classification Project Guide

# Project Overview

In this project, you will design and implement a binary classification model using a dataset of your choice.   
Your task is to:  
- Identify a binary outcome (e.g., disease vs. no disease, churn vs. stay, fraud vs. not fraud),  
- Choose appropriate classification algorithms (e.g., logistic regression, decision trees, random forests, SVM, neural networks),  
- Justify your choices using at least 5 peer-reviewed academic sources, and  
- Ensure your dataset is relevant to the context studied in your selected papers.

# Project Goals

- Conduct a brief literature review on classification models used for a similar task.  
- Select a dataset that aligns with the application area found in the literature.  
- Preprocess and prepare the data for modeling.  
- Implement and compare at least two classification algorithms.  
- Evaluate model performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score, AUC).  
- Interpret and visualize results in a real-world context.

# Dataset Requirements

Your dataset must:  
- Have a binary target variable.  
- Be relevant to the domain or topic discussed in your literature review.  
- Contain enough samples (minimum: 500 rows) and meaningful predictors.  
  
Suggested Sources:  
- Kaggle (https://www.kaggle.com/)  
- UCI Machine Learning Repository (https://archive.ics.uci.edu/)  
- OpenML (https://www.openml.org/)  
- Government sources (Data.gov, CDC, CMS)  
- APIs (Twitter, Reddit, etc., if cleaned and converted)

# Your Task (Step-by-Step)

1. Literature Review (Due: [Insert Date])  
- Find and summarize at least 5 academic papers that have used binary classification for similar tasks.  
- Focus on methods, performance metrics, and datasets used.  
- Submit a 1–2 page summary with full citations in APA or MLA.  
  
2. Dataset Selection  
- Choose a dataset relevant to the context of your literature review.  
- Provide a short justification for why this dataset is appropriate.  
  
3. Data Preprocessing  
- Clean and transform data (e.g., handling missing values, encoding categorical variables, feature selection).  
- Split into training and testing datasets.  
  
4. Model Implementation  
- Implement at least two classification models (e.g., logistic regression and random forest).  
- Tune hyperparameters if possible (e.g., using grid search or cross-validation).  
  
5. Model Evaluation  
- Use metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.  
- Present results in tables and visualizations (e.g., confusion matrix, ROC curve).  
  
6. Analysis and Interpretation  
- Compare model performance and justify which one is better for your task.  
- Discuss implications of your findings in a real-world context.

# Deliverables

- Literature Review Summary

- Project Proposal

- Final Report (3–5 pages)  
- Python or R Notebooks  
- Visualizations of Results  
- References in APA/MLA Format

# Proposal Template

Team Members:  
Project Title:  
Binary Outcome:  
Dataset Chosen (source + link):  
Literature Review Summary (brief overview of papers):  
Key Variables:  
Classification Algorithms to Compare:  
Evaluation Metrics:  
Why This Matters (real-world relevance):