IN6227 Data Mining – Assignment 1

Objective

This assignment is an individual task designed to give you hands-on experience with classification models. You will train, test, and compare two classifiers using the provided dataset.

Dataset

You will use the Census Income dataset from the UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Census+Income

Task

1. Choose any two classification models from existing implementations (e.g., Decision Tree, Logistic Regression, Naïve Bayes, k-NN, Random Forest, SVM, Neural Networks, etc.).
2. Train and test both models on the Census Income dataset.
3. Compare the models using appropriate classification evaluation techniques.

Report Requirements

1. Format: PDF file, maximum 2 pages.
2. At the top of the first page, include:
   a) Your matric number
   b) Your full name
   c) The line: "IN6227-2023-Assignment-1"
3. Content:
   a) Dataset preparation
      i. How you handle missing values
      ii. Any preprocessing or feature engineering (with justification)
   b) Model setup

i. The two classifiers chosen (state which libraries/implementations used)
ii. Any hyperparameter tuning (explain your decisions)
iii. Stopping criteria (if applicable)

c) Model evaluation
i. Performance metrics (accuracy, precision, recall, F1-score, ROC/PR curves, etc.)
ii. Training and prediction time for each model
iii. Clear comparison and discussion of results

d) Conclusion
i. Which model performed better and why

4. Include a link to your GitHub repository containing the full source code.

Submission

Upload your PDF report to NTULearn → Assignments
Deadline: Wednesday, 2025-Oct-08, 23:59:59.
Late submissions may not be accepted.