Parallel and Distributed Computing Semester Project

Objective:

Your task is to design and implement an optimized machine learning pipeline for binary classification using the provided dataset. You are free to choose any approach **parallel processing**, **distributed computing**, **GPU acceleration**, or a **hybrid strategy** to minimize **processing time** and maximize **model accuracy**.

Requirements:

- Preprocess the data (handle missing values, encode categorical variables, normalize features).
- Train a machine learning or deep learning model for binary classification.
- Optimize the pipeline using one or more of the following:
  - **Parallel computing** (e.g., multithreading, multiprocessing)
  - **Distributed systems** (e.g., MPI, Dask, Spark)
  - **GPU acceleration** (e.g., TensorFlow, PyTorch, CUDA)

Evaluate and report:

- Final **accuracy, confusion matrix, f1 score**
- **Total processing time (must be reduced by 70% atleast)**
- Comparative analysis of different setups (e.g., CPU vs GPU, parallel vs serial)

Deliverables:

- Source code (modular and well-commented)
- Performance report (accuracy, processing time, resource usage)
- A brief presentation/demo explaining your architecture, approach, and key findings