Machine Learning (IC272)

Assignment 1

Instructor: Dr. Indu Joshi

Deadline: 30th August

The School of Computing and Electrical Engineering (SCEE)
Indian Institute of Technology Mandi
August 21, 2025

Instructions

1. **Submit one Jupyter Notebook or MATLAB script with both tasks, clearly structured and commented.**
2. **Attach a concise 2–3 page report summarizing preprocessing, implementation, and results, with at least two visualizations per task.**
3. **Do not use `scikit-learn`. Allowed libraries: `numpy`, `pandas`, `matplotlib`, and `seaborn`.**
4. **Marks: Task 1 (50 points), Task 2 (50 points), Total = 100 points. Late submissions: -20 points/day.**

Dataset Overview

The **Indian Districts Population** dataset provides comprehensive demographic information across Indian districts. It includes data on population, growth rates, sex ratios, and literacy rates. This dataset enables demographic analyses, regional studies, and decision-making processes.
Dataset Link: **https://www.kaggle.com/datasets/shiivvvaam/indian-districts-population-data**

Task 1: Data Preprocessing (50 points)

- Load and explore the Indian Districts Population dataset.
- Handle missing values and clean anomalies (e.g., invalid percentages).
- Create a target column by categorizing literacy as Low ($< 65\%$), Medium (60–85%), or High ($> 85\%$).
- Encode categorical features, scale numeric variables, and split data into 70% training, 15% validation, and 15% test sets.
- Provide visualizations before and after preprocessing.

Task 2: KNN Classifier (50 points)

- Implement the K-Nearest Neighbors algorithm from scratch.
- Define a distance function, identify k nearest neighbors, and predict the majority class.

- Train and evaluate the model using literacy categories as labels.
- Compare results for $k = 3, 5, 7$ using class distribution plots and confusion matrices.
- Employ k-fold cross-validation to assess model stability, keeping the 15% test set fixed for final evaluation.