CSSM 501 Assignment 6
Machine Learning on UCI Adult Income dataset

The goal is to predict individuals' income levels as a binary value (>50k, <=50k) using several variables such as age, marital status, etc.

I am more familiar with the PyTorch library. This is why I chose to implement these models from scratch using torch. I started with Logistic Regression, Simple Feed-Forward, and Deep Neural Network models. Below there is architecture of this models.

```
Logistic Regression Model:
 LogisticRegressionModel(
  (linear): Linear(in_features=104, out_features=1, bias=True)
  (sigmoid): Sigmoid()
)

Simple Feedforward Neural Network:
 SimpleFeedforwardNN(
  (fc1): Linear(in_features=104, out_features=64, bias=True)
  (relu): ReLU()
  (fc2): Linear(in_features=64, out_features=1, bias=True)
  (sigmoid): Sigmoid()
)

Deep Neural Network:
 DeepNeuralNetwork(
  (network): Sequential(
    (0): Linear(in_features=104, out_features=128, bias=True)
    (1): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU()
    (3): Dropout(p=0.5, inplace=False)
    (4): Linear(in_features=128, out_features=64, bias=True)
    (5): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (6): ReLU()
    (7): Dropout(p=0.5, inplace=False)
    (8): Linear(in_features=64, out_features=32, bias=True)
    (9): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (10): ReLU()
    (11): Dropout(p=0.5, inplace=False)
    (12): Linear(in_features=32, out_features=1, bias=True)
    (13): Sigmoid()
  )
)
```

I created a hyperparameter space for each model and did grid search in this space. The problem is even with these advanced DNN model, F1-scores were not that good.

```
Model Evaluation Results on Test Set:
LogisticRegressionModel:
  Accuracy: 0.8458
  F1-Score: 0.6643

SimpleFeedforwardNN:
  Accuracy: 0.8529
  F1-Score: 0.6727

DeepNeuralNetwork:
  Accuracy: 0.8541
  F1-Score: 0.6700
```

I realized this is due to class imbalance. When I solve this issue with sampling, F1-scores increased for all of the models. The best one among them seems to be Decision Trees with the following scores

```
Confusion Matrix:
[[5748 1020]
 [ 928 5900]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.86      6768
           1       0.85      0.86      0.86      6828

    accuracy                           0.86     13596
   macro avg       0.86      0.86      0.86     13596
weighted avg       0.86      0.86      0.86     13596


Accuracy Score: 0.8567225654604296
```

Decision trees outperformed logistic regression and multilayer perceptrons (MLPs) likely because they can naturally handle both categorical and numerical data without extensive preprocessing. Additionally, decision trees effectively capture complex, non-linear relationships and feature interactions inherent in the UCI Adult Income dataset.