Python/scikit-learn Mini-Project

Objective: Predict people's incomes based on data from the 1994 U.S. census. You'll find the data for this mini-project in the data folder, and you can get context about the data from https://archive.ics.uci.edu/ml/datasets/adult.

Architecture: Your code will consist of two modules. This document contains the specification for the first module.

Module 1: data_preprocessing.py

This module will handle the preprocessing you need to do to prepare your data to be fed to your learning algorithms. You have complete freedom to design this module any way that makes sense to you. Just a few pointers:

1) You'll notice that some fields are marked with a question mark ('?'), indicating that their values are unknown. You can deal with these values any way you like, but the most obvious strategy would be to remove any rows with these unknown values.

2) For data preprocessing you'll almost certainly want to use the pandas and scikit-learn Python packages, both of which you've encountered earlier in the curriculum.

3) Some of the data you'll find in the dataset are categorical (e.g. gender), and others are numerical (e.g. age). There are a few different ways that you might deal with categorical features, but two popular ones are one-hot encoding and ordinal encoding. We'd recommend one-hot here.

4) You'll want to normalize numerical features to zero mean and unit variance during preprocessing (see what happens when this isn't done!).

5) You'll need to have all your inputs represented together as a single vector. A simple way to do that is to concatenate your numerical and categorical inputs.

6) Rather than manually exploring the hyperparameters that make your algorithm work best, be sure to use scikit-learn's GridSearchCV function!

7) When you finish, send your code to @yazabi, and we'll review it for you!