SPC707P Machine and Deep Learning — Week 01 Project

Kavit Tolia
September 30, 2025

1 Pick 5 of your favourite datasets

I have picked the following 5 datasets for this week's project:
1. Adult Income or Census Income from US Census Bureau: Adult Income
2. Air Quality or AirQualityUCI using sensors in Italy: Air Quality
3. Micro Gas Turbine Electrical Energy Prediction: Electrical Energy Prediction
4. Heart Disease (Cleveland): Heart Disease
5. Wine Quality (Red and White Vino Verde): Wine Quality

2 How many instances and features are in each dataset?

Dataset	Instances	Features
Adult Income	32,561 (train only)	14
Air Quality	9,258	15
Electrical Energy	71,225	1
Heart Disease	303	13
Wine Quality	1,599 (red) / 4,898 (white)	11

Table 1: Basic properties of selected datasets.

3 What type of person might have collected this data?

1. The adult income dataset would be collected by a **census bureau**
2. The air quality dataset would have been collected by a **government or environmental agency**
3. The electrical energy prediction dataset would have been collected by an **energy department**
4. The heart disease dataset would have been collected by a **medical institute**
5. The wine quality dataset would have been collected by a **wine producer**

4 Why do I find each of the datasets interesting?

1. Adult Income: It would be interesting to see how well demographics can predict income
2. Air Quality: I'm interested in understanding how certain attributes can determine extent of air pollution
3. Electrical Energy: I find time series analysis quite interesting, and this seemed quite realistic
4. Heart Disease: Understanding the effect of someone's attributes on heart disease can have very positive health impact
5. Wine Quality: **I love wine!**

5 What are some deeper insights the datasets might reveal?

1. Adult Income: If demographics can predict income, governments can guide policy to support their citizens
2. Air Quality: The data can be used to plan cities in a way which reduces air pollution without other changes
3. Electrical Energy: The data can show insights into how the turbine reacts to changes in control voltage
4. Heart Disease: If there is a link between certain factors and heart disease, we could have early prediction and reduce medical costs for individuals and governments
5. Wine Quality: A winemaker can derive an exact recipe to provide the highest quality wine

6 What are some ethical or representative issues with the data?

1. Adult Income: This data is from 1994 and may no longer represent the world today. While the features can help us with predicting income, it can be difficult to attribute rationale if there is a deeper layer of inherent socio-economic bias.
2. Air Quality: This is only from one location and might not be representative of other places across the globe.
3. Electrical Energy: The data might not generalise to other turbine types or environments.
4. Heart Disease: The data might not generalise to the broader demographic and could also lead to privacy issues given medical data.
5. Wine Quality: The ratings for each of the wine is subjective to an individual's preferences - it's difficult to remove human bias from the target variable.

7 Data Cleansing

For this task, I have decided to use the Adult Income dataset. I will attempt to fix any missing values and examine any other issues with the data.
I will explain the data issue, provide the code used to make any changes and highlight the fixed data.
- Once I loaded the data, I used `df.info()` to understand the data types of each of the features.
- For object types, I used `df[col].value_counts` to understand missing data.
- For numerical types, I used `df.describe(include='all')` to understand missing data.
The dataset has '?' as a proxy for missing values. If we left these as is, this could have downstream implications when feeding the data into an ML model.
For example, the model might fail to even work due to these missing value. Or, more worryingly, it could work and find patterns with this distorted data.
We can then use `df.replace` and `df.dropna` to delete rows with missing data. This removed 2,399 instances out of a total of 32,561.

Listing 1: Cleaning Adult Income dataset

```
import pandas as pd
import numpy as np
cols = ['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital-status',
       'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss',
       'hours-per-week', 'native-country', 'income_class']
data_loc = '../data/raw/adult/adult.data'
df = pd.read_csv(data_loc, header=None, names=cols)
df.replace('?', np.nan, inplace=True)
df.dropna(inplace=True)
```

8 Exploratory Data Analysis (EDA)

I have also created a scatter matrix (on the following page) for a handful of numerical features.
The target is highlighted in a different colour (red for greater than 50K and blue for less than 50K).
We can make the following observations based on the scatter matrix:
- Higher levels of education are clearly linked to with a greater probability of earning more than 50K
- Very high capital gains also tend to appear only among higher income individuals (although these large gains are rare and the feature is very skewed with lots of zeros and a few hard-coded 99,999 values)
- Age doesn't seem to have a strong relationship with income
- There seems to be some trend between the number of hours worked and income, although this could be due to other factors

Figure shown on next page →

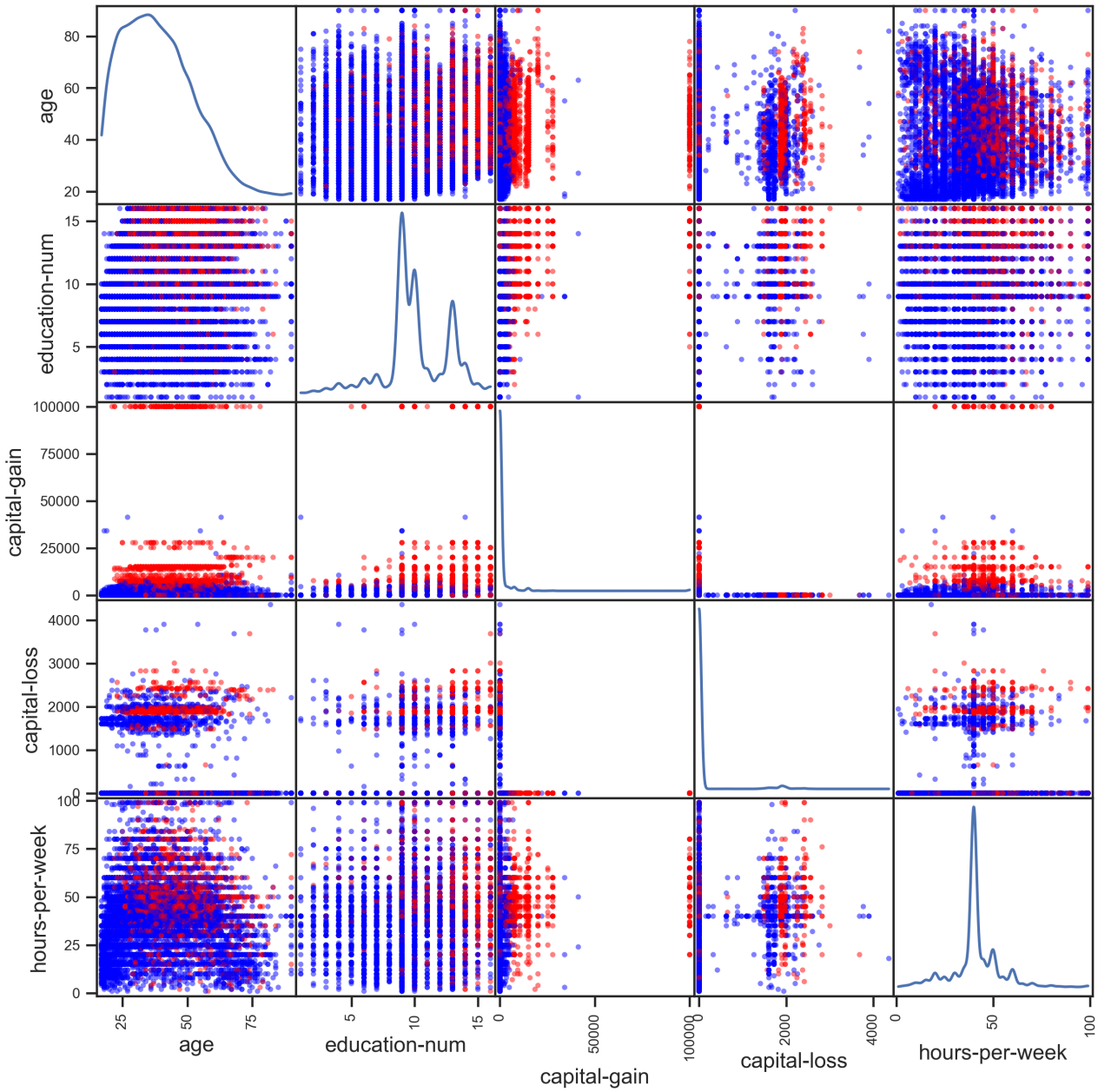

Figure 1: Scatter matrix of selected continuous variables in the Adult Income dataset.
The red dots highlight individuals with income higher than 50K and the blue dots for lower than 50K.