SPC707P Machine and Deep Learning — Week 01 Project

Kavit Tolia

September 25, 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 data points or instances in each dataset?

1. The adult income dataset has **48,842** instances, split across train and test data
2. The air quality dataset has **9,358** instances
3. The electrical energy prediction dataset has **71,225** instances
4. The heart disease dataset has **303** instances
5. The wine quality dataset is **1,599 red wine** and **4,898 white wine** instances

3 How many features in each dataset?

1. The adult income dataset has **14** features (attributes per person)
2. The air quality dataset has **15** features (readings from sensors)
3. The electrical energy prediction dataset has **1** feature (time series)
4. The heart disease dataset has **13** features (patient attributes)
5. The wine quality dataset has **11** features (chemical test information)

4 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**

5 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!**

6 What are some deeper insights the datasets might reveal?