Lab 5 Exploratory Analysis - Analysing the Census Income Data Set

When performing tasks in the area of machine learning, it is important to understand your dataset *before* diving straight into the analytic task.

Explore the Census Income Dataset to gain an insight into its characteristics and find inherent biases in the data. (data source: https://archive.ics.uci.edu/ml/datasets/Census+Income)

As you perform the tasks below, keep the following fairness-related questions in mind:

- **What's missing?**
- **What's being overgeneralized?**
- **What's being underrepresented?**
- **How do the variables, and their values, reflect the real world?**
- **What might we be leaving out?**

Task 1 Description of data

Find:

Total number of records, Number of fields, List field names

Task 2 Create visualisations for each of the features in the dataset to explore the data in terms of fairness and completion.

Some important questions to investigate when auditing a dataset for fairness:

- **Are there missing feature values for a large number of observations?**
- **Are there features that are missing that might affect other features?**
- **Are there any unexpected feature values?**
- **What signs of data skew do you see?**

Task 3 Answer the following questions

1. How would you describe the relationship between education level and income bracket?
2. What noteworthy observations can you make about the gender distributions for each marital-status category?

https://colab.research.google.com/github/google/eng-edu/blob/master/ml/cc/exercises/intro_to_fairness.ipynb?utm_source=mlcc&utm_campaign=colab-external&utm_medium=referral&utm_content=fairness-colab&hl=en#scrollTo=qZ-9vJgSEpHj