Exercise 11:

Lets use the ‘census *income*‘ dataset and apply decision tree and naïve Bayesian methods to predict whether a person’s income will exceed $50K/yr. The dataset used is also called ‘*adultsData*‘ data. Some of the attributes available to predict the income are *age, employment type, education, marital status, work hours per week* etc. Please finish the tasks below:

1. Please split the dataset as training (80%) and test data (20%).
2. Build the **decision tree** on the training data and predict the response on the test data. Calculate the accuracy and generate the **Confusion Matrix**.
3. Build the **naïve Bayesian model** on the training data and predict the response on the test data. Calculate the accuracy and generate the **Confusion Matrix**
4. Compare the two models. Which model is better? And why?

- age – The age of the individual
- workclass – The type of employer the individual has. Whether they are government, military, private, and so on.
- fnlwgt – The # of people the census takers believe that observation represents.
- education – The highest level of education achieved for that individual
- education_num – Highest level of education in numerical form
- maritalstatus – Marital status of the individual
- occupation – The occupation of the individual
- relationship – A bit more difficult to explain. Contains family relationship values like husband, father, and so on, but only contains one per observation. I’m not sure what this is supposed to represent
- race – descriptions of the individuals race. Black, White, Eskimo, and so on
- sex – Biological Sex
- capital_gain – Capital gains recorded
- capital_loss – Capital Losses recorded
- hr_per_week – Hours worked per week
- class – Boolean Variable and the label. Whether or not the person makes more than \$50,000 per annum income.

Deliverable

Please write a report (e.g., ½ page) explaining the machine learning models you used on the data. What is the portion of the data that was used to train the models and the size of the testing data that was used to evaluate the accuracy of the models? Which model is better? decision tree or naïve Bayes? Please explain which model achieved better results based on the confusion matrix (e.g., True Positives, True Negatives, False Positives, and False Negatives)