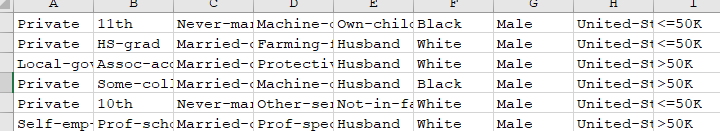
**Part 1:**

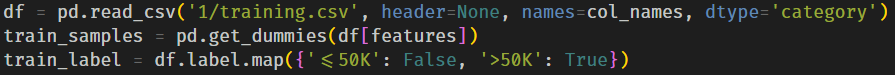
Removal of the unknown values and continuous attributes was done in excel. The first five entries of testing.csv is shown here:



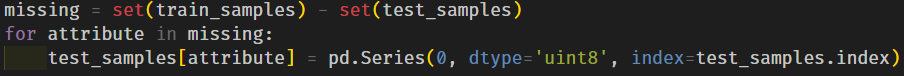
**Part 1 a:**

The decision tree was built in Python using Pandas and Scikit-Learn, and this guide was followed: https://www.ritchieng.com/machine-learning-evaluate-classification-model/  
  
There were two tricky parts:

* Scikit-Learn does not accept categorical attributes, so pandas’ get\_dummies() function was used:



* The test labels were missing an attribute value that existed on the training labels. The missing values were found with the set difference, and then inserted with 0 values. A for loop was used to keep the code more general.



**Part 1 b:**

Completing this part was simple after completing part 1a. it uses all the same data frames and labels, and so a Gaussian Naïve-Bayes classifier was constructed, and the metrics printed in just 2 lines of code.

**Part 1 Results:**

|  |  |  |
| --- | --- | --- |
|  | **Tree** | **Gaussian Naïve-Bayes** |
| Accuracy | 0.807 | 0.748 |
| Recall / TP Rate | 0.528 | 0.035 |
| FP Rate | 0.102 | 0.980 |
| Precision | 0.627 | 0.365 |
| F1 Score | 0.573 | 0.064 |

Notable that the Naïve-Bayes classifier apparently has a very low true positive rate, and a very high false positive rate. Running it multiple times, the same pattern continued.