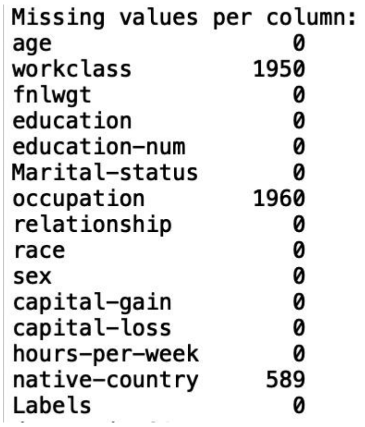
This A2_qhuangak_20548333_prediction.py file is used to predict labels of testing data(TestingFeatures.csv).

It includes four process:

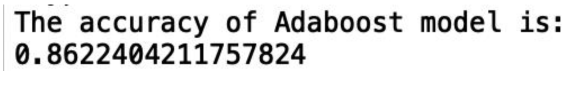
1. The feature engineering process, it use sklearn to complete the process
   1) Import some basic lib
   2) Read the train data set
   3) Combine train features and labels
   4) Count the missing data in the train data(missing error in this dataset is ‘?’)



```
Missing values per column:
age                0
workclass       1950
fnlwgt             0
education          0
education-num      0
Marital-status     0
occupation      1960
relationship       0
race               0
sex                0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country   589
Labels             0
```

   5) Transfer the data into int format, and the missing values in that three column become ‘0’
   6) Fill the missing values by mode
   7) Delete the noise (There is a special value which ‘native country’ is ‘Holand Netherlands’, although it is not important, but if I use dummy, it will influence the result)
   8) Delete the duplicate data
   9) Split data to train and test, the ratio is 0.2
   10) Extract the label and get the x_train, y_train(include labels), x_test, y_test(include labels)

2. The model training process
   Use the AdaBoost to train the model, get the accuracy of model:



```
The accuracy of Adaboost model is:
0.8622404211757824
```

3. The testing data preprocessing process
   1) Print the missing value in the testing data
   2) Transfer the data into int.
   3) Fill the missing value by mode

4. Predict labels of testing data process, and generate the prediction .csv file