# Introduction

For the income prediction project, I’ve completed data cleaning, data pre-processing and feature exploration. Besides, I’ve applied two basic models: logistic regression and decision tree without fine-tuning parameters or cross validation.

# Data Cleaning

Table

Description automatically generated

Noticing that there is a duplicated column ‘education’ which is in nonnumerical format, we can drop it.

Now we have the matching education level as below:

The matching education level of the education number:

1: Preschool, 2: 1st-4th, 3: 5th-6th, 4: 7th-8th, 5: 9th, 6: 10th, 7: 11th, 8: 12th, 9: HS-grad, 10: Some-college, 11: Assoc-voc, 12: Assoc-acdm, 13: Bachelors, 14: Masters, 15: Prof-school, 16: Doctorate.

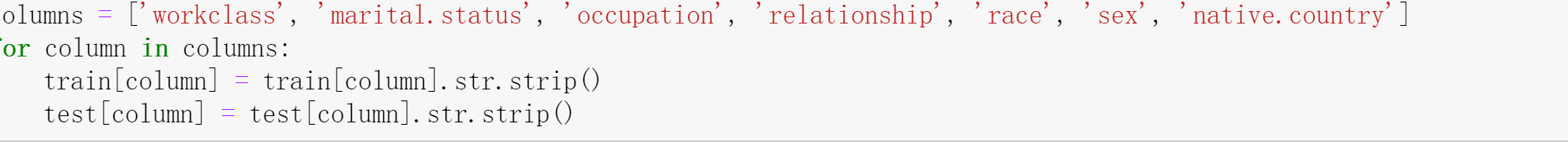
Next, I checked for null values, and it appears that there are no nulls in the whole dataset.

Then I checked duplicates and remove them:

A picture containing text

Description automatically generated

The next step is to remove the spaces in all data entries for easier access.



I noticed that there are some ‘?’ symbols in data entries so I decided to change ‘?’ to category ‘Unknown’.

Table

Description automatically generated

# Feature Exploration

Now the datasets have been prepared and I started feature exploration. In mid-term stage, I only explored a few features: workclass, age, sex and occupation.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Looks like people with >50K has a higher average age than the ones with <=50K. And in both cases of income, we see that the male category has a little bit greater age average than the female category. People working in prof-specialty and exec-managerial industries have higher income.

# Basic models

At this stage, I did not do scaling for numerical features. After one-hot encoding, I noticed that there are some unseen categorical values in test data, so we need more pre-processing work. I have submitted the prediction results using Logistic regression and Decision Tree model.

Graphical user interface, text, application, email

Description automatically generated

# Future Schedule

My following plan is to do more feature engineering work, to explore the feature importance level. I will also perform transformations on features that are highly skewed and perform some scaling on numerical features. Normalization ensures that each feature is treated equally when applying supervised learner. I will test more models like Adaboost, Random Forest, Gradient Boosting, SVM and neural network.