**Technology : I used R for the analysis (all my code is join en the email as a attachment).**

**Question 1**

What is the count of people that earned <=50K salary?

**Answer : 23265 persons**

**Question 2**

What percentage of people with higher education (Masters or Doctorate) earned >50k?

**Answer :** **1181 persons = 3,87%**

**Question 3**

1. Create a model to predict whether someone would earn more than 50k (Income Group) based on factors in the data.

Include an accuracy metric: Number of Correct Predictions/Total Predictions

**Answer :** The following pages are how the methodoloy I used to to it…

**Fives phases of CRISPDM**

**Business Understanding Phase**

In this project, I analyze the Census dataset. The objective is to identify the best algorithm to be employed in order to predict whether a certain adult has annual income more than 50,000 $ based on several attributes such as: age, education, occupation, capital gain, and capital loss as well as other variables as illustrated in details within this document. Such prediction is required to decide if a certain adult could be granted a loan or be targeted by a specific marketing campaign.

**Data Understanding Phase**

The dataset contains 30511 observations of 14 variables: 6 numeric variables and 8 factor level variables. Certain numeric variables contain outliers whereas certain factor variables contain missing data. Here are the **missing data**:

* workclass 1769
* occupation Status 1774
* native country 531

Moreover, the feature Income Group was giving 4 choices because some of the responses were either <=50K or <=50. (with a dot at the end) and >50 or >50. (with a dot at the end also)

**Data Preparation Phase**

Data preparation is done in order to clean and recover missing data and elimination of outliers. Missed data are recovered using the method of *'Replace Missing Values with the values «not mention » for occupation status and native country and the values « Other-service » that was already the logic choice available in the workclass feature*, while outliers are removed using '*Min‐max*' technique.

**Modeling Phase**

Three different models are applied on both training and validation datasets after data cleaning for the purpose of predicting the target variable and thereafter identifying the best algorithm:

1Decision trees (before normalization of data)

1. Neural networks (after normalization)

**Evaluation Phase**

The performance of each algorithm is initially evaluated for each model by getting the percentage between estimated values and actual values. False negative and false positive rates are also evaluated.

.

Data pre-processing and exploratory data analysis

Data preprocessing is taking place through two steps:

* Data cleaning: for recovering missing values.
* Elimination of outliers.

**Data cleaning**

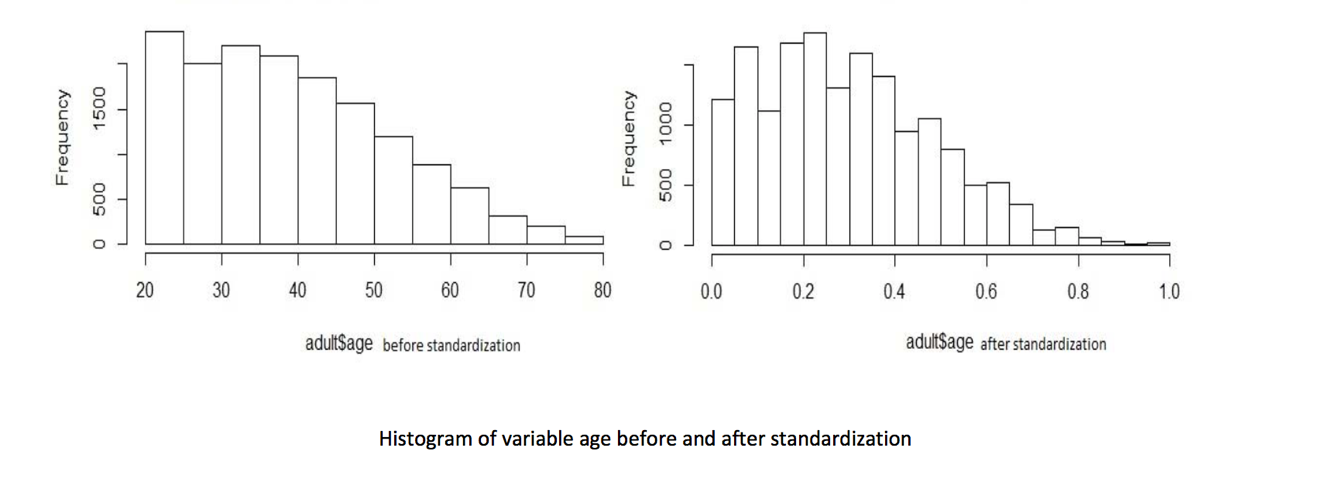
By examining the existing data of the adult dataset, we discover that there missing in three fields: workclass, occupation, and native.country. The missing data are marked as question marks ' ?'. There was a space before the question mark.

I have also combine the fields (<=50K with <=50k.) witch was written 50k with a dot at the end and (>50K and >50k.)

**Elimination of outliers**

Data transformation method was chosen in order to eliminate the outliers of the dataset. For this project, the min-max normalization was applied to certain variables on the dataset in order to achieve the normalization. The concerning variables are the numeric variables such as: age, Demographic ajustment,Years of education, hours.per.week, capital.loss and capital.gain.

Here is for example the feature Age before and after Normalisation.



Evaluation of algorithms :

# Decision tree :

Before I normalized the data, because it doesn’t make sens for instance to split the tree with a node at the age level with a number between 0 and 1, which is the case after normalization.

**The result was :**

The method that was used was a classification method because the target variable is a categorical feature. The tree was split as follow : First by relationship, then with education and finally with capital-gain.

The model was rather precise and the performance (accuracy metric) of the validation set was 84,3% as in figure 1.

**Here’s how the tree was split :**

**The training data : 20000 and the rest, 10511 were the validation data.**

As seen on figure 0,

15% of the training data (20000) had the relationship (husband or wife) and had the eduction (Assoc-acdm,Bachelors,Doctorate,Masters,Prof-school) were predicted to make over 50K as in figure 2.

Figure 0

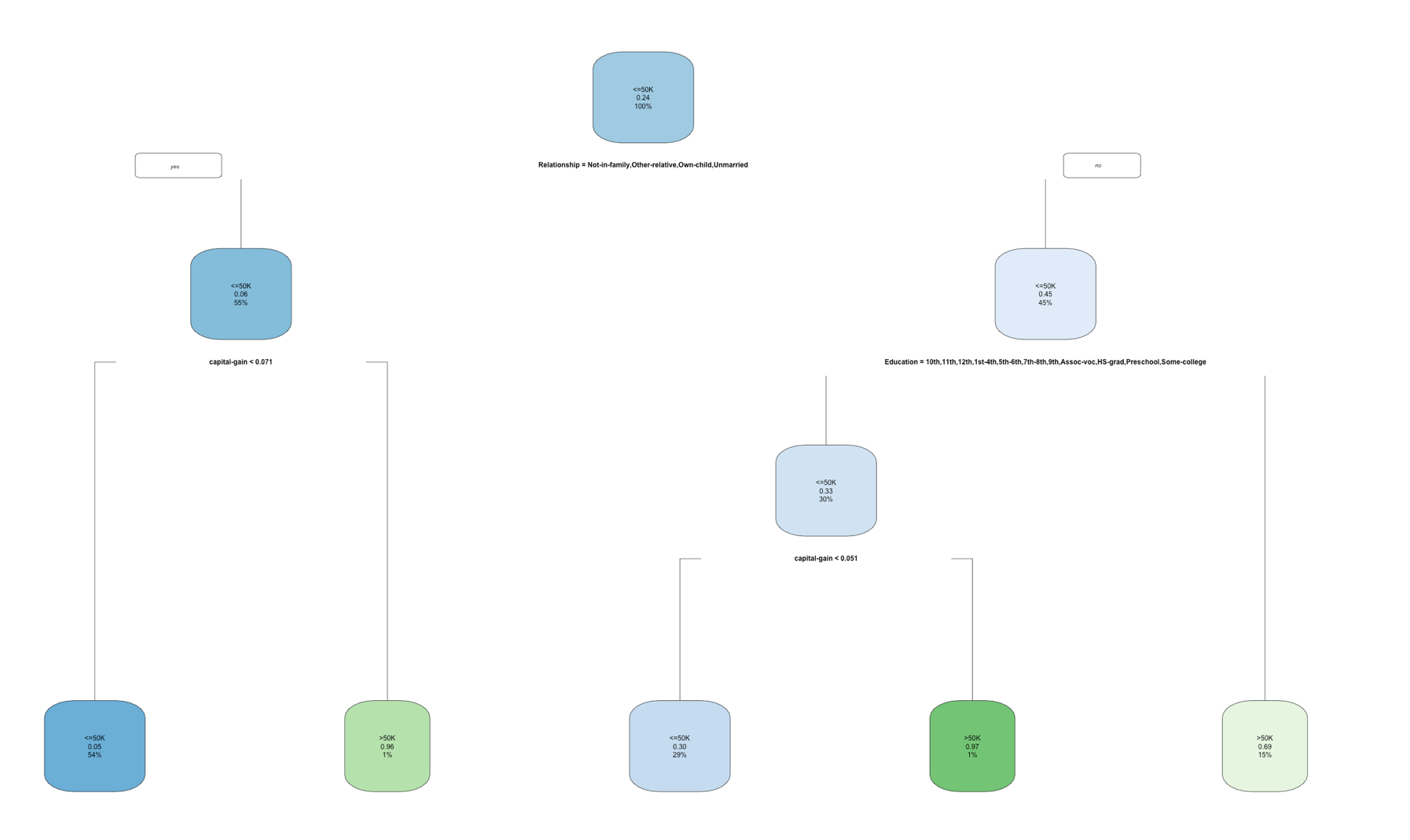


Figure 1

estincome.class

<=50K >50K

<=50K 21866 1399

>50K 3389 3857

> (21866+3857)/30511

[1] 0.843073

Figure 2

\* denotes terminal node

> mytreeadult

n= 20000

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 20000 4760 <=50K (0.76200000 0.23800000)

2) Relationship=Not-in-family,Other-relative,Own-child,Unmarried 11003 707 <=50K (0.93574480 0.06425520)

4) capital-gain< 0.07055571 10821 532 <=50K (0.95083634 0.04916366) \*

5) capital-gain>=0.07055571 182 7 >50K (0.03846154 0.96153846) \*

3) Relationship=Husband,Wife 8997 4053 <=50K (0.54951651 0.45048349)

6) Education=10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-voc,HS-grad,Preschool,Some-college 6029 1996 <=50K (0.66893349 0.33106651)

12) capital-gain< 0.05095551 5735 1712 <=50K (0.70148213 0.29851787) \*

13) capital-gain>=0.05095551 294 10 >50K (0.03401361 0.96598639) \*

7) Education=Assoc-acdm,Bachelors,Doctorate,Masters,Prof-school 2968 911 >50K (0.30694070 0.69305930) \*

Figure 3 (algorithm used = c.4.5)

Classification tree:

rpart(formula = `Income Group` ~ ., data = pwcClean, method = "class",

control = rpart.control(minsplit = 1))

Variables actually used in tree construction:

[1] capital-gain Education Relationship

Neural network

This model had a accuracy metric of 85,3% with 10 hidden layers as in figure 4.

Figure 4.

> pwc.net = nnet(`Income Group`~.,data=pwc.test,size=10)

pwc.valid$est.income = predict(pwc.net,pwc.valid,type="class")

> Tpwc=table(pwc.valid$`Income Group`,pwc.valid$est.income)

> Tpwc

<=50K >50K

<=50K 7457 577

>50K 958 1519

> (7457+1519)/(30511-20000)

[1] 0.8539625