DATA MINING FOR LIFE SATISFACTION IN AUSTRALIA 2020

Yuhui Pang
u7211790@anu.edu.au
May 2022

Australian National University

CONTENTS

INTRODUCTION

Background

Life satisfaction is the subjective experience of one's quality of life. It is a comprehensive psychological indicator of a person's quality of life. The study of residents' life satisfaction is very important for the development of the country and society. The government can make timely adjustments to relevant policies and related construction according to the residents' satisfaction with their lives. In addition, good resident life satisfaction is conducive to building citizens' confidence in the country and the long-term development of the country.

Research on life satisfaction comes from three main fields namely, research on mental health, research on the quality of life and research on gerontology (Righi & Masserini, 2021). This report analyzes the factors influencing residents' life satisfaction and the prediction of life satisfaction by data mining the public opinion survey on the perception of topical social issues conducted by the Centre for Social Research of the Australian National University in August 2020 (Biddle, 2020).

Problem Introduction

This report will use data mining methods such as association mining, decision tree classification, neural networks, logistic regression, and linear regression to analyze data on life satisfaction and find its related factors to predict life satisfaction by using the data source of *ANU Poll 35 (August 2020)*.

The project primarily employs R and Rattle for model development and evaluation, as well as Excel and Python for data pre-processing.

The goal of this project is to identify the factors that influence residents' life satisfaction and to develop a useful model to classify and forecast resident satisfaction. In addition, by linearizing the satisfaction of the residents, the discrete attributes are continuous and the specific values of satisfaction are predicted by linear regression, which may not be accurate but can get a specific trend of satisfaction instead of a simple classification.

Through the data mining on life satisfaction, a relatively accurate classifier can be constructed, and the project aims to give timely feedback to the government and related organizations to improve the life satisfaction of the country's residents.

DATA DESCRIPTION

Data Source

The original data was obtained from the results of a questionnaire from the Centre for Social Research at the Australian National University in August 2020 on a public opinion survey on topical social issues.

The data contains 335 attributes and 3062 rows of data, each row representing the findings of one surveyed user. The specific data content is shown in Table 1.

Column	Section	Topic	Data Type	Correlation Level
1-4	Routine	ID, Date, Mode, Order	Nominal & Numerical	Low
5-15	Section A	Satisfaction	Ordinal & Numerical	High
16-30	Section B	Covid-19 Experience	Ordinal & Numerical	High
31-116	Module V	Covid-19 Vaccine	Ordinal & Numerical	Low
117-123	Section D	Mental Health	Ordinal & Numerical	High
124-155	Section E	Employment and Income	Ordinal & Numerical	High
156-288	Module G	COVID-19 Policy	Nominal & Numerical	Low
289-316	Module F	Bushfire Policy	Ordinal & Numerical	Low
317-335	Section P	Personal Information	Ordinal & Numerical	Low

Table 1 Data source overview

Data Pre-processing

Data Selection & Data Cleaning

Based on the relevance of data and topics, we assume that life satisfaction is related to satisfaction with the state and related institutions, mental health and employment income levels. Therefore, $Section\ A$, $Section\ D$ and $Section\ E$ were selected as the main data for data mining, where $A3$ was the target value and other attributes were the input values.

Moreover, the data cleaning process removes the data rows with values of -98, -99 and blank. The data selection and data distribution are shown in the following table.

Section	Section A			Section D		Section E	
Attribute	A1	A3	A4a-A4f	D1a-D1f	D3	E11a	E13
Type	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical
Use	Input	Target	Input	Input	Input	Input	Input
Min	1	0	1	1	1	1	1
Median	2	7	2	1,1,2,2,1,1	1	2	4
Max	5	10	4	5	4	4	4
Mean	2.5	6.6	2.4,2.4,2.2, 1.8,2.0,1.6	2.1,1.7,2.1, 2.1,1.6,1.5	1.6	1.9	3.7
Unique	5	11	4	5	4	4	4
Description	Country satisfaction	Life satisfaction	Institutional satisfaction	Negative emotions	Loneliness level	Income level	Housing issues

Table 2 Data selection and data distribution

Data Recoding & Data Normalization

Considering the different value domains and directionality of values for each attribute, for example, the $A1$ attribute from 1 to 5, represents positive to negative attitudes, and the $A3$ attribute from 0 to 10, represents negative to positive attitudes. Therefore, the data need to be recoded and normalized. After the processing, the value domain of all attributes is unified as $[0,1]$, with 0 to 1 representing negative to positive attitudes.

$$Normalization(A) = \frac{A - 1}{range(A)}$$

In addition, some attributes in the attribute set are similar in meaning and multiple-use increases the weight of that aspect, so these attributes can be averaged. For example, $A4a$ to $A4f$ represents residents' confidence in each different institution and can be simply averaged to represent residents' confidence in all institutions.

New Attributes

$A4\_avg$ and $D1\_avg$ were generated to represent the average confidence in each institution and the average attitude towards some negative sentiments.

$$A3\_avg = \frac{1}{6}\sum_{i=a}^{e} A3_i \qquad D1\_avg = \frac{1}{6}\sum_{i=a}^{e} D1_i$$

In addition, for predictive classification, the target value $A3$ was recoded using satisfaction score 6 as the cut-off point to obtain a binary type $A3\_binary$ for binary classification of life satisfaction. Where 0 represents dissatisfaction and 1 represents satisfaction.

$$A3\_binary = \begin{cases} 0 & if\ A3 \leq 6 \\ 1 & if\ A3 > 6 \end{cases}$$

At last, for linear regression prediction, a continuous target value is required, while the target value $A3$ in the data source is discrete. A random decimal r is used to simulate a continuous value, which is denoted by $A3\_linear$ in this case.

$$A3\_linear = \begin{cases} 0 & if\ A3 + r < 0 \\ 10 & if\ A3 + r > 10 \\ A3 + r & other \end{cases} \qquad r \in [-0.5, 0.5)$$

METHOD & RESULT

Associate Mining

Correlation analyzes

Using all attributes as input, the correlation analysis result is shown in the right figure, where the blue colour represents positive correlation, and the darker colour represents higher correlation.

From the last row (or the last column) of the figure, we can see that $A3$ is positively correlated with all the other attributes, which is due to the isotropy treatment in the

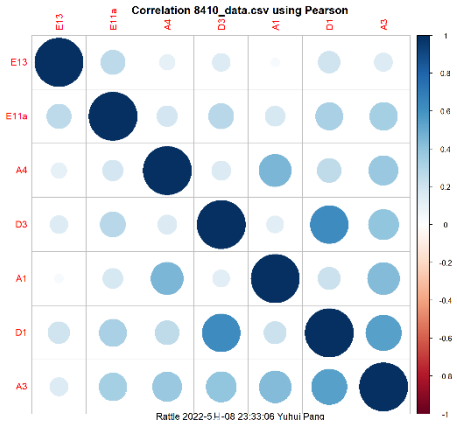
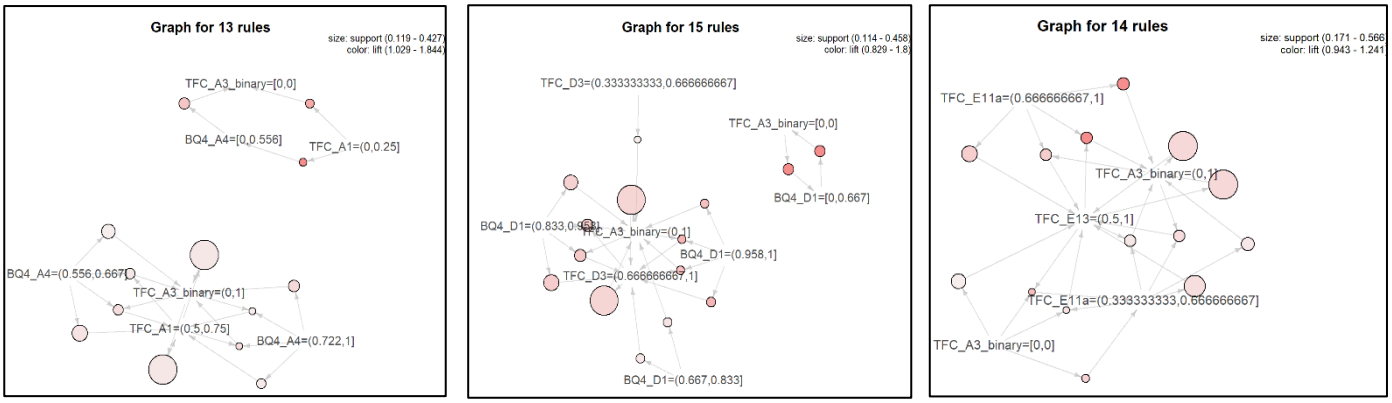


Figure 1 Data correlation between attributes

data pre-processing. We can also find that the degree of correlation among the attributes is relatively high, which supports the previous hypothesis and reflects the value of the study.

Life Satisfaction with Other Factors

The relationship between $A3$ and other attributes was found by association mining of $A3\_binary$ with different groups of data. The result is shown in the following figure.



(a) A3_binary & A1 & A4 (b) A3_binary & D1 & D3 (c) A3_binary & E11a & E13

Figure 2 Association mining

Figure 2a shows the association mining of $A3$ and $A1$, $A4$, where $min\_sup = 0.10$, $min\_conf = 0.51$. From the figure, it is clear that the rules are divided into two piles. Where the larger values of $A1$ and $A4$ are associated with $A3 = 1$ and the smaller values are associated with $A3 = 0$. This indicates that people who are satisfied with the state and related institutions tend to be satisfied with their lives and vice versa.

Figure 2b shows the association mining for $A3$ and $D1$, $D3$, where $min\_sup = 0.11$, $min\_conf = 0.45$. In this figure, the rules are also divided into two piles. Where the larger values of $D1$ and $D3$ are associated with $A3 = 1$ and the smaller values are associated with $A3 = 0$. This indicates that people who have more negative emotions and feel lonely tend to be dissatisfied with life and vice versa.

Figure 2c shows the association mining for $A3$ and $E11a$, $E13$, where $min\_sup = 0.15$, $min\_conf = 0.50$. This case is not as good as the first two cases, however, it can be seen that the larger values of $E11a$ and $E13$ are associated with $A3 = 1$, indicating that people with a high level of employment income tend to be satisfied with their lives.

Decision Tree

In this section, a decision tree algorithm will be used to classify the data for prediction, where the target value is $A3\_binary$, and the input values are $A1$,$A4$,$D1$,$D3$,$E11a$,$E13$. All input data are cleaned and normalized.

Parameter Setting

Partition	Min Split	Min Bucket	Max Depth	Complexity
75/15/15	13	5	30	0.0050

Result

The decision tree is shown in the following figure. We can see that the decision tree generates some rules to divide the data into two classes, where 0 represents dissatisfaction with life and 1 represents satisfaction with life. For example, if $D1 \geq 0.73$ and $A1 \geq 0.63$, It will be classified into the $class$ 1, which means satisfaction with life.

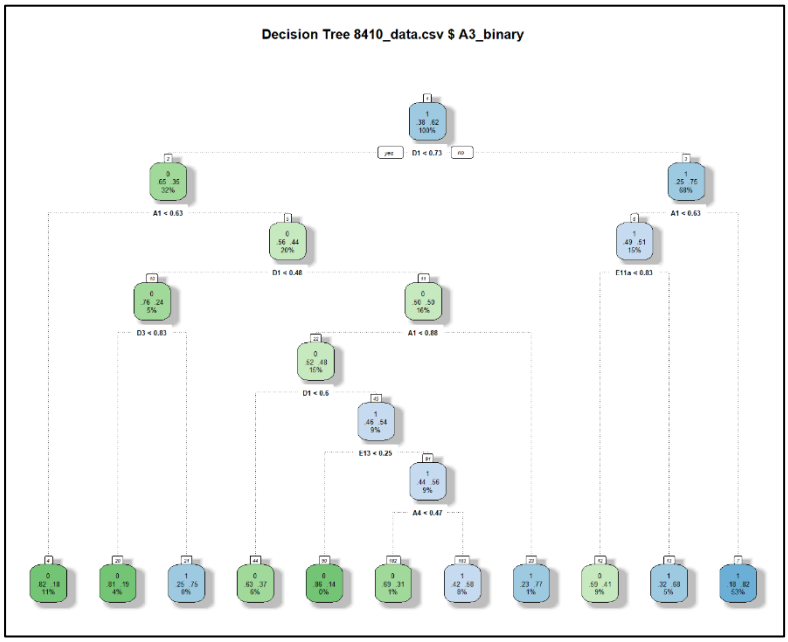


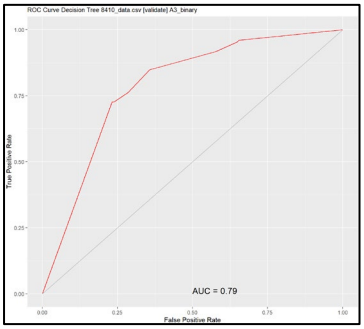
Figure 3 Decision tree

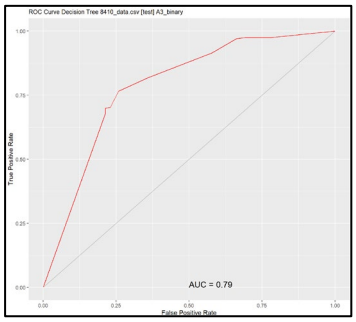
Evaluation

The following figure shows the confusion matrix and roc curve of the model and some evaluation measures of both validation data and testing data.

```
Error matrix for the Decision Tree model on 8410_data.csv [validate] (counts):

      Predicted
Actual   0   1 Error
     0 108  60  35.7
     1  42 235  15.2

Error matrix for the Decision Tree model on 8410_data.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 24.3 13.5  35.7
     1  9.4 52.8  15.2

Overall error: 22.9%, Averaged class error: 25.45%
```

```
Error matrix for the Decision Tree model on 8410_data.csv [test] (counts):

      Predicted
Actual   0   1 Error
     0 114  64  36.0
     1  49 220  18.2

Error matrix for the Decision Tree model on 8410_data.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 25.5 14.3  36.0
     1 11.0 49.2  18.2

Overall error: 25.3%, Averaged class error: 27.1%
```





Accuracy	Precision	Recall	F-Score
0.76	0.80	0.85	0.82

(a) Evaluation for validation data

Accuracy	Precision	Recall	F-Score
0.75	0.77	0.82	0.79

(b) Evaluation for testing data

Figure 4 Confusion matrix and Roc curve of validation data and testing data

Logistic Regression

Parameter Setting

Set the ratio of the training set and test set to $0.75: 0.25$, and set the logistic regression $family = binomial$.

Result

The results of the logistic regression are shown in the figure below.

```
> summary(pre)

Call:
glm(formula = f, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3294  -0.8526   0.4881   0.7611   2.5538

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.67625    0.41272 -16.176  < 2e-16 ***
A1           1.93453    0.20996   9.214  < 2e-16 ***
A4           2.40548    0.39885   6.031 1.63e-09 ***
D1           3.87266    0.35612  10.874  < 2e-16 ***
D3           0.64470    0.22564   2.857  0.00427 **
E11a         1.14357    0.20926   5.465 4.63e-08 ***
E13          0.04338    0.28777   0.151  0.88018
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2981.8  on 2227  degrees of freedom
Residual deviance: 2272.0  on 2221  degrees of freedom
AIC: 2286

Number of Fisher Scoring iterations: 4
```

Figure 4 Logistic regression result

The result shows the weights of each input variable. It can be seen from the figure that the weight of E13 is relatively low, indicating that its influence on the model building is small.

Evaluation

The following figure shows the confusion matrix and roc curve of the model and some evaluation measures. It can be seen that the model works much better in predicting class 1 than class 0.

```
> table(real,predict)
    predict
real   0   1
   0 162  97
   1  57 427
```

Accuracy	Precision	Recall	F-Score
0.79	0.81	0.88	0.84

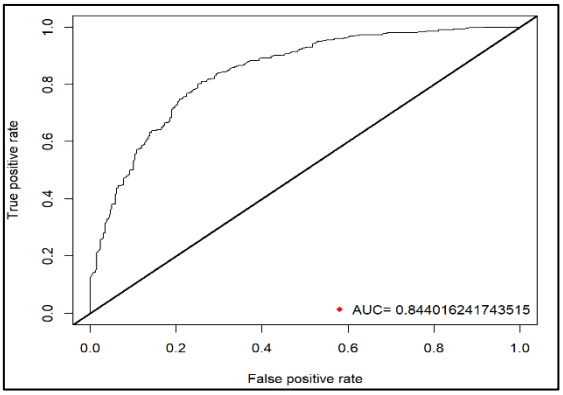


Figure 5 Logistic Regression Model Evaluation

Neural Network

Parameter Setting

Set the ratio of the training set test set to $0.75: 0.25$, set the parameter of the neural network to $hidden = c(3,2), act.fct = "logistic", linear.output = F$.

Result

The following figure shows the result of the neural network structure, which has two hidden layers with 3 and 2 hidden units, and the values on the connecting lines represent the weights of each input. Similarly, we can see that attribute $E13$ has a much smaller weight relative to the other attributes, indicating that its contribution to the classification model is small.

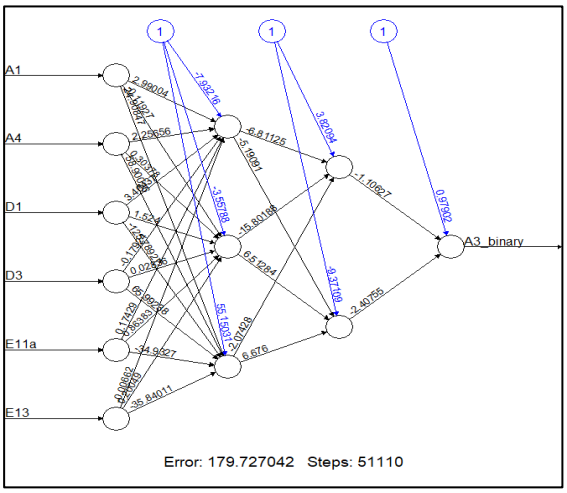


Figure 6 Neural network

Evaluation

The following figure shows the confusion matrix and roc curve of the model and some evaluation measures.

```
> table(real,predict)
    predict
real   0   1
   0 197  97
   1  80 369
```

Accuracy	Precision	Recall	F-Score
0.76	0.79	0.86	0.80

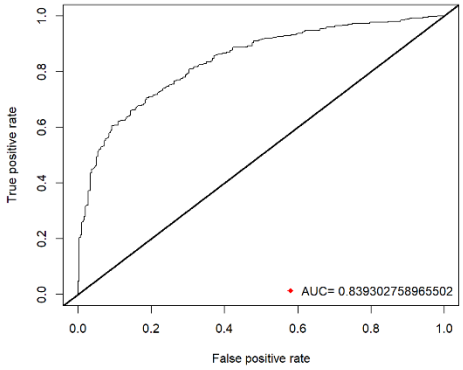


Figure 7 Neural Network Model Evaluation

Linear Regression

The linear regression model is built using Continuous values $A3\_linear$ as the target value, and the results are shown below. We can also see that the weight of $E13$ is relatively small.

In the linear regression model, the $average\ error = 1.09$. The results show a way to predict specific values by using continuous-valued linear regression.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3492  -0.8025   0.1239   0.9183   4.7182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.62510    0.20491   3.051  0.00231 **
A1           1.83483    0.12829  14.303  < 2e-16 ***
A4           1.62790    0.22804   7.139 1.27e-12 ***
D1           3.27546    0.20161  16.247  < 2e-16 ***
D3           0.69587    0.13847   5.026 5.42e-07 ***
E11a         0.80371    0.12486   6.437 1.49e-10 ***
E13          0.09708    0.17368   0.559  0.57625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.060607)

    Null deviance: 8035.7  on 2227  degrees of freedom
Residual deviance: 4576.6  on 2221  degrees of freedom
AIC: 7942.6
```

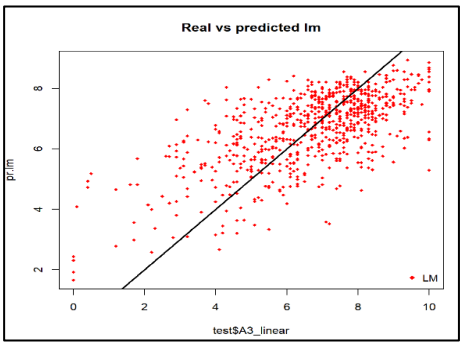


Figure 8 Linear regression

CONCLUSION & FURTHER WORK

Conclusion

From the results of the above models, the models are good at classification and prediction of the residents' life satisfaction with high accuracy. This verifies the assumption presented earlier in the report, which is people who are more satisfied with the state and related institutions as well as those with more positive psychological situations and high levels of employment income. Therefore, the models constructed in this report can be used to make predictions about whether residents are satisfied with their lives.

In addition, this report gives a way to serialize the ordinal attribute $A3$ and thus use linear regression for predicting the specific value of $A3$. The overall average error is acceptable.

However, we can see that the model works much better for positive outcomes than for negative ones. It is because the value of $A3$ in the original data is not very balanced, and most people (62%) are satisfied with life, which affects the construction and meaning of the model. Because even simply predicting that all people are satisfied with their life can get a better accuracy rate.

Further work

Considering the unbalanced nature of the data, it is possible to add some data that are not satisfied with life or remove some positive data before the model is constructed. In addition, since the results of linear regression are not ideal simply because the data are derived from the results after continuousizing the discrete variables, a continuous score containing decimal points can be considered for respondents to rate their life.

REFERENCE

Biddle, Nicholas; Edwards, Ben; Gray, Matthew; Sollis, Kate, 2020, "ANU Poll 35 (August 2020): COVID-19 attitudes and behaviours (Wave 3)", doi:10.26193/ZFGFNE, ADA Dataverse, V2

Haar, J. M., Russo, M., Suñe, A., & Ollier-Malaterre, A. (2014). Outcomes of work-life balance on job satisfaction, life satisfaction and mental health: A study across seven cultures. *Journal of Vocational Behavior, 85*(3), 361-373. https://doi.org/10.1016/j.jvb.2014.08.010

Righi, A., & Masserini, L. (2021). Measuring relational factors underlying subjective happiness. *Current Psychology (New Brunswick, N.J.),* https://doi.org/10.1007/s12144-021-02208-2

APPENDIX

Code of logistic regression:

```
setwd("D:/ANU/2022S1/COMP8410/Assignment2")
dataset <- read.csv("8410_data.csv")
data <- subset(dataset,select=c("A3_binary","A1","A4","D1","D3","E11a","E13"))
index <- sample(1:nrow(data),round(0.75*nrow(data)))
train <- data[index,]
test <- data[-index,]

f <- "A3_binary ~ A1 + A4 + D1 + D3 + E11a + E13"
pre = glm(f,data=train,family=binomial)
real=test$A3_binary
predict_=predict.glm(pre,type="response",newdata=test)
predict=ifelse(predict_>0.5,1,0)
table(real,predict)

error=predict-real
accuracy=(nrow(test)-sum(abs(error)))/nrow(test)
precision=sum(real & predict)/sum(predict)
recall=sum(predict & real)/sum(real)
F_measure=2*precision*recall/(precision+recall)

print(accuracy)
print(precision)
print(recall)
print(F_measure)

library(ROCR)
pred <- prediction(predict_,real)
auc <- performance(pred,'auc')@y.values
auc <- as.character(auc)
perf <- performance(pred,'tpr','fpr')
plot(perf)
abline(0,1,lwd=2)
legend('bottomright',legend=paste("AUC=",auc),pch=18,col='red', bty='n')
```

Code of neural network:

```
library(neuralnet)
setwd("D:/ANU/2022S1/COMP8410/Assignment2")
```

```
dataset <- read.csv("8410_data.csv")
data <- subset(dataset ,select=c("A3_binary","A1","A4","D1","D3","E11a","E13"))

index <- sample(1:nrow(data),round(0.75*nrow(data)))
train <- data[index,]
test <- data[-index,]

f <- "A3_binary ~ A1 + A4 + D1 + D3 + E11a + E13"
nn <- neuralnet(f,data=train,hidden=c(3,2), act.fct = "logistic", linear.output=F)
pr.nn <- compute(nn,test[,-1])
predict_ <- pr.nn$net.result
predict <- ifelse(pr.nn$net.result>0.5,1,0)
real = test$A3_binary
table(predict,real)

accuracy=(nrow(test)-sum(abs(error)))/nrow(test)
precision=sum(real & predict)/sum(predict)
recall=sum(predict & real)/sum(real)
F_measure=2*precision*recall/(precision+recall)

print(accuracy)
print(precision)
print(recall)
print(F_measure)

library(ROCR)
pred <- prediction(predict_,real)
auc <- performance(pred,'auc')@y.values
auc <- as.character(auc)
perf <- performance(pred,'tpr','fpr')
plot(perf)
abline(0,1,lwd=2)
legend('bottomright',legend=paste("AUC=",auc),pch=18,col='red', bty='n')
```

Code of linear regression:

```
setwd("D:/ANU/2022S1/COMP8410/Assignment2")
dataset<- read.csv("8410_data.csv")
data <- subset(dataset,select=c("A3_linear","A1","A4","D1","D3","E11a","E13"))

index <- sample(1:nrow(data),round(0.75*nrow(data)))
```

```
  train <- data[index,]
  test <- data[-index,]

  f <- "A3_linear~ A1 + A4 + D1 + D3 + E11a + E13"
  lm.fit <- glm(f, data=train)
  pr.lm <- predict(lm.fit,test)
  summary(lm.fit)

  MAE.lm <- sum(abs(pr.lm - test$A3_linear))/nrow(test)
  print(MAE.lm)

  plot(test$A3_linear,pr.lm,col='red',main='Real vs predicted lm',pch=18, c
  ex=0.7)
  abline(0,1,lwd=2)
  legend('bottomright',legend='LM',pch=18,col='red', bty='n', cex=.95)
```