UTAR UNIVERSITI TUNKU ABDUL RAHMAN Wholly owned by UTAR Education Foundation (Co. No. 578227-M) DU012(A)	**UNIVERSITI TUNKU ABDUL RAHMAN (UTAR)** **LKC FACULTY OF ENGINEERING AND SCIENCE (LKC FES)**

ACADEMIC YEAR 2020/2021

UECM3993 Predictive Modeling

BACHELOR OF SCIENCE (HONS) ACTUARIAL SCIENCE

YEAR 2/3

Title : **Divorce Predictors using supervised learning and unsupervised learning**
Data : **Divorce predictor Data Set**
Group name : **PTS**

Name	Course	ID
Tan Eng Sim	AS	1803672
Chao Xin Yi	AS	1800709
Lim Li Ting	AS	1802044
Lee Shu Ying	AS	1800412
Lee Yang	AS	1803551
Rachel Chin Chi Shan	AS	1801619
Sio Wen Kang	AS	1801927

Table of Contents

1.0 Introduction

The report is done based on the divorce predictors data set available on UCI website https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set . The main purpose of the report is to investigate the most suitable algorithms including supervised learning and unsupervised learning in predicting divorce. The data set has 170 instances and 54 attributes which represent each of the following 54 questions of survey:

1.If one of us apologizes when our discussion deteriorates, the discussion ends.
2. I know we can ignore our differences, even if things get hard sometimes.
3. When we need it, we can take our discussions with my spouse from the beginning and correct it.
4. When I discuss with my spouse, contacting him will eventually work.
5. The time I spent with my wife is special for us.
6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
8. I enjoy our holidays with my wife.
9. I enjoy traveling with my wife.
10. Most of our goals are common to my spouse.
11. I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12. My spouse and I have similar values in terms of personal freedom.
13. My spouse and I have a similar sense of entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams with my spouse are similar and harmonious.
16. We're compatible with my spouse about what love should be.
17. We share the same views about being happy in our life with my spouse
18. My spouse and I have similar ideas about how marriage should be
19. My spouse and I have similar ideas about how roles should be in marriage
20. My spouse and I have similar values in trust.
21. I know exactly what my wife likes.
22. I know how my spouse wants to be taken care of when she/he is sick.
23. I know my spouse's favourite food.
24. I can tell you what kind of stress my spouse is facing in her/his life.
25. I have knowledge of my spouse's inner world.
26. I know my spouse's basic anxieties.
27. I know what my spouse's current sources of stress are.
28. I know my spouse's hopes and wishes.
29. I know my spouse very well.
30. I know my spouse's friends and their social relationships.
31. I feel aggressive when I argue with my spouse.
32. When discussing with my spouse, I usually use expressions such as 'you always' or 'you never'.
33. I can use negative statements about my spouse's personality during our discussions.
34. I can use offensive expressions during our discussions.
35. I can insult my spouse during our discussions.
36. I can be humiliating when we discuss.
37. My discussion with my spouse is not calm.
38. I hate my spouse's way of opening a subject.
39. Our discussions often occur suddenly.
40. We're just starting a discussion before I know what's going on.

41. When I talk to my spouse about something, my calm suddenly breaks.
42. When I argue with my spouse, ı only go out and I don't say a word.
43. I mostly stay silent to calm the environment a little bit.
44. Sometimes I think it's good for me to leave home for a while.
45. I'd rather stay silent than discuss with my spouse.
46. Even if I'm right in the discussion, I stay silent to hurt my spouse.
47. When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.
48. I feel right in our discussions.
49. I have nothing to do with what I've been accused of.
50. I'm not actually the one who's guilty about what I'm accused of.
51. I'm not the one who's wrong about problems at home.
52. I wouldn't hesitate to tell my spouse about her/his inadequacy.
53. When I discuss, I remind my spouse of her/his inadequacy.
54. I'm not afraid to tell my spouse about her/his incompetence.

The data sets are obtained by conducting a survey for 170 couples who were already divorced or happily married. The data consists of their corresponding Divorce Predictors Scales variables (DPS),(Yöntem & İlhan 2017, 2018). Of all the participants 84 were divorced and 86 were married. The group consists of 84 males and 86 females. The age of participants was from 20 to 63. The data was collected from 7 regions of Turkey.They were requested to answer the 54 questions provided through face-to-face interview. They will rate each of the questions on a 5 points scale (0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always). All the questions will act as a predictor for Class represented by "0" (divorced) and "1" (happily married). The Class is the variable for our study. (Yöntem & et al., 2019)
Our task is to conduct a classification predictive modelling based on the data.

We perform Exploratory Data Analysis in the first place to have a better understanding in the data set. We inspect the data and explore the characteristics of each column and row. We also test the correlation of each variable in the dataset. Principal component analysis (PAC) is also carried out to reduce dimensionality of data set, to have better visualization on the data we plot out the data.

The splitting of the data set into training and testing data is based on stratified splitting. Next, we train supervised and unsupervised learning algorithms to predict the Class. The result of each algorithm is tested using the training data by the method confusion matrix. The model is then further validated using cross validation. In unsupervised learning we drop the label Class and to the model. The predicted results are then compared with the actual results.

The accuracy of each model is then compared to obtain the best model.

Purpose

We want to select the best model which is able to predict the condition of marriage of the respondents. The selection is based on the best accuracy of the model. We also want to know how important each predictor is in predicting divorce.

2.0 Exploratory Data Analysis

2.1 Data Inspection

```
  Atr1 Atr2 Atr3 Atr4 Atr5 Atr6 Atr7 Atr8 Atr9 Atr10 Atr11 Atr12 Atr13 Atr14 Atr15 Atr16 Atr17 Atr18 Atr19 Atr20 Atr21 Atr22
1    2    2    4    1    0    0    0    0    0     0     1     0     1     1     0     1     0     0     0     1     0     0
2    4    4    4    4    4    0    0    4    4     4     4     3     4     0     4     4     4     4     3     2     1     1
3    2    2    2    2    1    3    2    1    1     2     3     4     2     3     3     3     3     3     3     2     1     0
4    3    2    3    2    3    3    3    3    3     3     4     3     3     4     3     3     3     3     3     4     1     1
5    2    2    1    1    1    1    0    0    0     0     0     1     0     1     1     1     1     1     2     1     1     0
6    0    0    1    0    0    2    0    0    0     1     0     2     1     0     2     0     2     1     0     1     0     0
  Atr23 Atr24 Atr25 Atr26 Atr27 Atr28 Atr29 Atr30 Atr31 Atr32 Atr33 Atr34 Atr35 Atr36 Atr37 Atr38 Atr39 Atr40 Atr41 Atr42
1     0     0     0     0     0     0     0     1     1     2     1     2     0     1     2     1     3     3     2     1
2     0     2     2     1     2     0     1     1     0     4     2     3     0     2     3     4     2     4     2     2
3     1     2     2     2     2     2     3     2     3     3     1     1     1     1     2     1     3     3     3     3
4     1     1     2     1     1     1     1     3     2     3     2     2     1     1     3     3     4     4     2     2
5     0     0     0     2     1     2     1     1     1     1     1     1     0     0     0     0     2     1     0     2
6     0     0     2     2     0     0     0     0     4     1     1     1     1     1     1     2     0     2     2     1
  Atr43 Atr44 Atr45 Atr46 Atr47 Atr48 Atr49 Atr50 Atr51 Atr52 Atr53 Atr54 Class
1     1     2     3     2     1     3     3     3     2     3     2     1     1
2     3     4     2     2     2     3     4     4     4     4     2     2     1
3     2     3     2     3     2     3     1     1     1     2     2     2     1
4     3     2     3     2     2     3     3     3     3     2     2     2     1
5     3     0     2     2     1     2     3     2     2     2     1     0     1
6     2     3     0     2     2     1     2     1     1     1     2     0     1
>
```

Figure 1.0

Figure 1.0 shows the first six rows of the data set. We can see that the data consists of columns "Atr1", "Atr2", …, "Atr54" and "Class". The Atr1 to Atr54 represents the 54 questions we mentioned earlier. The "Class" is our targeted variable which indicates divorce or happily married

```
'data.frame':   170 obs. of  55 variables:
 $ Atr1 : int  2 4 2 3 2 0 3 2 2 1 ...
 $ Atr2 : int  2 4 2 2 2 0 3 1 2 1 ...
 $ Atr3 : int  4 4 2 3 1 1 3 2 1 1 ...
 $ Atr4 : int  1 4 2 2 1 0 2 2 0 1 ...
 $ Atr5 : int  0 4 1 3 1 0 1 2 0 1 ...
 $ Atr6 : int  0 0 3 3 1 2 3 1 4 2 ...
 $ Atr7 : int  0 0 2 3 0 0 4 0 1 0 ...
 $ Atr8 : int  0 4 1 3 0 0 3 3 3 2 ...
 $ Atr9 : int  0 4 1 3 0 0 2 3 3 2 ...
 $ Atr10: int  0 4 2 3 0 1 2 2 3 2 ...
 $ Atr11: int  1 4 3 4 0 0 2 4 3 3 ...
 $ Atr12: int  0 3 4 3 1 2 2 3 3 0 ...
 $ Atr13: int  1 4 2 3 0 1 2 2 3 0 ...
 $ Atr14: int  1 0 3 4 1 0 3 3 3 2 ...
 $ Atr15: int  0 4 3 3 1 2 2 4 3 1 ...
 $ Atr16: int  1 4 3 3 1 0 3 3 3 0 ...
 $ Atr17: int  0 4 3 3 1 2 3 2 3 1 ...
 $ Atr18: int  0 4 3 3 1 1 3 3 3 2 ...
 $ Atr19: int  0 3 3 3 2 0 3 2 3 1 ...
 $ Atr20: int  1 2 2 4 1 1 2 1 3 0 ...
 $ Atr21: int  0 1 1 1 1 0 3 2 2 0 ...
 $ Atr22: int  0 1 0 1 0 0 3 1 2 0 ...
 $ Atr23: int  0 0 1 1 0 0 3 1 2 0 ...
 $ Atr24: int  0 2 2 1 0 0 3 2 3 1 ...
 $ Atr25: int  0 2 2 2 0 2 2 3 2 1 ...
 $ Atr26: int  0 1 2 1 2 2 3 3 3 1 ...
 $ Atr27: int  0 2 2 1 1 0 3 2 2 1 ...
 $ Atr28: int  0 0 2 1 2 0 2 2 3 1 ...
 $ Atr29: int  0 1 3 1 1 0 2 2 2 1 ...
 $ Atr30: int  1 1 2 3 1 0 2 3 3 1 ...
```

```
 $ Atr30: int  1 1 2 3 1 0 2 3 3 1 ...
 $ Atr31: int  1 0 3 2 1 4 1 1 1 1 ...
 $ Atr32: int  2 4 3 3 1 1 2 1 1 1 ...
 $ Atr33: int  1 2 1 2 1 1 2 0 1 0 ...
 $ Atr34: int  2 3 1 2 1 1 1 2 1 1 ...
 $ Atr35: int  0 0 1 1 0 1 1 2 1 0 ...
 $ Atr36: int  1 2 1 1 0 1 2 1 1 0 ...
 $ Atr37: int  2 3 2 3 0 1 3 4 1 1 ...
 $ Atr38: int  1 4 1 3 0 2 2 4 2 1 ...
 $ Atr39: int  3 2 3 4 2 0 2 4 2 2 ...
 $ Atr40: int  3 4 3 4 1 2 3 4 2 2 ...
 $ Atr41: int  2 2 3 2 0 2 3 4 2 1 ...
 $ Atr42: int  1 2 3 2 2 1 3 4 2 2 ...
 $ Atr43: int  1 3 2 3 3 2 3 3 2 3 ...
 $ Atr44: int  2 4 3 2 0 3 4 2 2 2 ...
 $ Atr45: int  3 2 2 3 2 0 3 0 2 2 ...
 $ Atr46: int  2 2 3 2 2 2 3 0 1 2 ...
 $ Atr47: int  1 2 2 2 1 2 2 1 1 0 ...
 $ Atr48: int  3 3 3 3 2 1 3 2 1 2 ...
 $ Atr49: int  3 4 1 3 3 2 2 2 1 2 ...
 $ Atr50: int  3 4 1 3 2 1 3 2 1 2 ...
 $ Atr51: int  2 4 1 3 2 1 3 1 1 2 ...
 $ Atr52: int  3 4 2 2 2 1 2 1 1 4 ...
 $ Atr53: int  2 2 2 2 1 2 2 1 1 3 ...
 $ Atr54: int  1 2 2 2 0 0 2 0 1 3 ...
 $ Class: int  1 1 1 1 1 1 1 1 1 1 ...
```

Figure 1.1

From the structure of the data set all the columns have the class of integers. However, for the column "Class" supposed to be categorised as "factor" because we need to treat it as a categorical variable.

The column "Class" becomes "factor" with 2 levels which are "0" and "1" after we implement the change of class.

```
 $ Class: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Figure 1.2

```
> dim(data)
[1] 170  55
> class(data)
[1] "data.frame"
>
```

Figure 1.3

The data is a data frame with dimension of 170*55 which means it has 170 rows and 55 columns.

```
> sapply(X = data, FUN = function(x) sum(is.na(x)))
 Atr1  Atr2  Atr3  Atr4  Atr5  Atr6  Atr7  Atr8  Atr9 Atr10 Atr11 Atr12 
    0     0     0     0     0     0     0     0     0     0     0     0 
Atr13 Atr14 Atr15 Atr16 Atr17 Atr18 Atr19 Atr20 Atr21 Atr22 Atr23 Atr24 
    0     0     0     0     0     0     0     0     0     0     0     0 
Atr25 Atr26 Atr27 Atr28 Atr29 Atr30 Atr31 Atr32 Atr33 Atr34 Atr35 Atr36 
    0     0     0     0     0     0     0     0     0     0     0     0 
Atr37 Atr38 Atr39 Atr40 Atr41 Atr42 Atr43 Atr44 Atr45 Atr46 Atr47 Atr48 
    0     0     0     0     0     0     0     0     0     0     0     0 
Atr49 Atr50 Atr51 Atr52 Atr53 Atr54 Class 
    0     0     0     0     0     0     0 
>
```

Figure 1.4

All the columns in the data set do not consist of any missing data.

```
      Atr1            Atr2            Atr3            Atr4            Atr5            Atr6              Atr7       
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
Median :2.000   Median :2.000   Median :2.000   Median :1.000   Median :1.000   Median :0.0000   Median :0.0000  
Mean   :1.776   Mean   :1.653   Mean   :1.765   Mean   :1.482   Mean   :1.541   Mean   :0.7471   Mean   :0.4941  
3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:1.0000  
Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.0000   Max.   :4.0000  
      Atr8            Atr9           Atr10           Atr11           Atr12           Atr13           Atr14      
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
Median :1.000   Median :1.000   Median :2.000   Median :1.000   Median :1.500   Median :2.000   Median :1.000  
Mean   :1.453   Mean   :1.459   Mean   :1.576   Mean   :1.688   Mean   :1.653   Mean   :1.835   Mean   :1.571  
3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000  
Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000  
     Atr15           Atr16           Atr17           Atr18           Atr19           Atr20           Atr21      
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
Median :1.000   Median :1.000   Median :1.000   Median :1.000   Median :1.000   Median :1.000   Median :1.000  
Mean   :1.571   Mean   :1.476   Mean   :1.653   Mean   :1.518   Mean   :1.641   Mean   :1.459   Mean   :1.388  
3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000  
Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000  
     Atr22           Atr23           Atr24           Atr25           Atr26           Atr27         Atr28      
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0   Min.   :0.000  
1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0   1st Qu.:0.000  
Median :0.000   Median :0.000   Median :1.000   Median :1.000   Median :1.000   Median :1.0   Median :0.500  
Mean   :1.247   Mean   :1.412   Mean   :1.512   Mean   :1.629   Mean   :1.488   Mean   :1.4   Mean   :1.306  
3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.0   3rd Qu.:3.000  
Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.0   Max.   :4.000  
```

Figure 1.5.1

```
     Atr29           Atr30           Atr31           Atr32           Atr33           Atr34         Atr35
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0   1st Qu.:0.000
 Median :1.000   Median :1.000   Median :2.000   Median :2.000   Median :1.000   Median :1.0   Median :0.500
 Mean   :1.494   Mean   :1.494   Mean   :2.124   Mean   :2.059   Mean   :1.806   Mean   :1.9   Mean   :1.671
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.0   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.0   Max.   :4.000
     Atr36           Atr37           Atr38           Atr39           Atr40           Atr41           Atr42
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
 Median :0.000   Median :2.000   Median :1.000   Median :2.000   Median :1.500   Median :2.000   Median :2.000
 Mean   :1.606   Mean   :2.088   Mean   :1.859   Mean   :2.088   Mean   :1.871   Mean   :1.994   Mean   :2.159
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
     Atr43           Atr44           Atr45           Atr46           Atr47           Atr48           Atr49
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:2.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000
 Median :3.000   Median :2.000   Median :3.000   Median :3.000   Median :2.000   Median :3.000   Median :3.000
 Mean   :2.706   Mean   :1.941   Mean   :2.459   Mean   :2.553   Mean   :2.271   Mean   :2.741   Mean   :2.382
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
     Atr50           Atr51           Atr52           Atr53           Atr54       Class
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   0:86
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.000   1:84
 Median :2.000   Median :3.000   Median :3.000   Median :2.000   Median :2.000
 Mean   :2.429   Mean   :2.476   Mean   :2.518   Mean   :2.241   Mean   :2.012
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
>
```

Figure 1.5.2

We go through the rough summary of each column in the data set. We found that the columns "Atr1" to "Atr54" each have attributed numbers starting from 0 to 4 (min:0, median:2, 3rd Qu:3 and max:4). As discussed earlier, the score from 0 to 4 reflects the agreeness of the participants toward each question. From the summary we know there is no out of range value. The mean of the column is more than 1 and less than 3 except for "Atr6 " and "Atr7”. The ratio of "Class0" to "Class1" is 86:84.

2.2 Variables study

We drop the column "Class" temporarily for convenience in the study of the "Atr1" to "Atr54".

```
          0     1     2     3     4 miss
Atr1  40.59  5.29  8.24 27.65 18.24    0
Atr2  34.71 13.53 16.47 22.35 12.94    0
Atr3  30.00 14.12 15.29 30.59 10.00    0
Atr4  44.12  7.06 17.65 18.82 12.35    0
Atr5  48.24  5.88  4.71 25.88 15.29    0
Atr6  50.59 28.82 17.06  2.35  1.18    0
Atr7  67.06 24.71  2.94  2.35  2.94    0
Atr8  47.65  5.88 12.35 21.76 12.35    0
Atr9  49.41  4.12  7.65 28.82 10.00    0
Atr10 36.47 10.59 22.35 20.00 10.59    0
Atr11 41.76 11.18  2.35 25.88 18.82    0
Atr12 34.12 15.88 12.94 24.71 12.35    0
Atr13 27.65 19.41 10.59 26.47 15.88    0
Atr14 38.82 13.53 12.35 22.35 12.94    0
Atr15 40.59 11.76  7.65 30.00 10.00    0
Atr16 44.12  8.24 15.29 20.59 11.76    0
Atr17 42.94  8.82  3.53 29.41 15.29    0
Atr18 46.47  5.88  9.41 25.88 12.35    0
Atr19 45.29  6.47  2.94 29.41 15.88    0
Atr20 47.65  5.88 12.35 21.18 12.94    0
Atr21 45.88 10.00 10.00 27.65  6.47    0
Atr22 51.18  7.65 15.29 17.06  8.82    0
Atr23 52.94  4.12  4.71 25.29 12.94    0
Atr24 42.35  9.41 15.29 20.59 12.35    0
Atr25 37.06 16.47  7.06 25.29 14.12    0
Atr26 42.35 11.76 12.35 21.76 11.76    0
Atr27 45.29 10.59 10.00 27.06  7.06    0
Atr28 50.00  5.88 17.65 16.47 10.00    0
Atr29 47.65  7.65  5.88 25.29 13.53    0
Atr30 42.35 11.18 13.53 20.59 12.35    0

Atr30 42.35 11.18 13.53 20.59 12.35    0
Atr31 25.88 16.47 12.35 10.00 35.29    0
Atr32 27.06 17.65  7.06 18.82 29.41    0
Atr33 41.76 11.76  3.53 10.00 32.94    0
Atr34 29.41 21.76  5.88 15.29 27.65    0
Atr35 50.00  7.65  0.59  8.82 32.94    0
Atr36 51.76  5.29  1.18 14.12 27.65    0
Atr37 28.82 18.24  5.88  9.41 37.65    0
Atr38 37.65 14.12  3.53 14.12 30.59    0
Atr39 29.41 16.47  8.24  7.65 38.24    0
Atr40 42.35  7.65  3.53 13.53 32.94    0
Atr41 32.35 15.29  8.24  8.82 35.29    0
Atr42 25.88 10.00 15.29 20.00 28.82    0
Atr43  9.41 11.18 19.41 19.41 40.59    0
Atr44 35.29  9.41 10.00 16.47 28.82    0
Atr45 16.47 14.12 12.35 21.18 35.88    0
Atr46 12.94 10.59 16.47 28.24 31.76    0
Atr47 19.41 20.00 11.76 11.76 37.06    0
Atr48  5.88  5.88 27.65 29.41 31.18    0
Atr49 16.47 16.47 15.29 15.88 35.88    0
Atr50 11.18 18.24 21.18 15.29 34.12    0
Atr51  7.06 17.06 25.29 22.35 28.24    0
Atr52 13.53 16.47 13.53 17.65 38.82    0
Atr53 18.24 18.24 15.29 17.65 30.59    0
Atr54 29.41 15.88 12.35  8.82 33.53    0
```

Figure 1.6

We view the frequency of each attribute in each column using the frequency table. The value in the frequency table is shown in percentage.

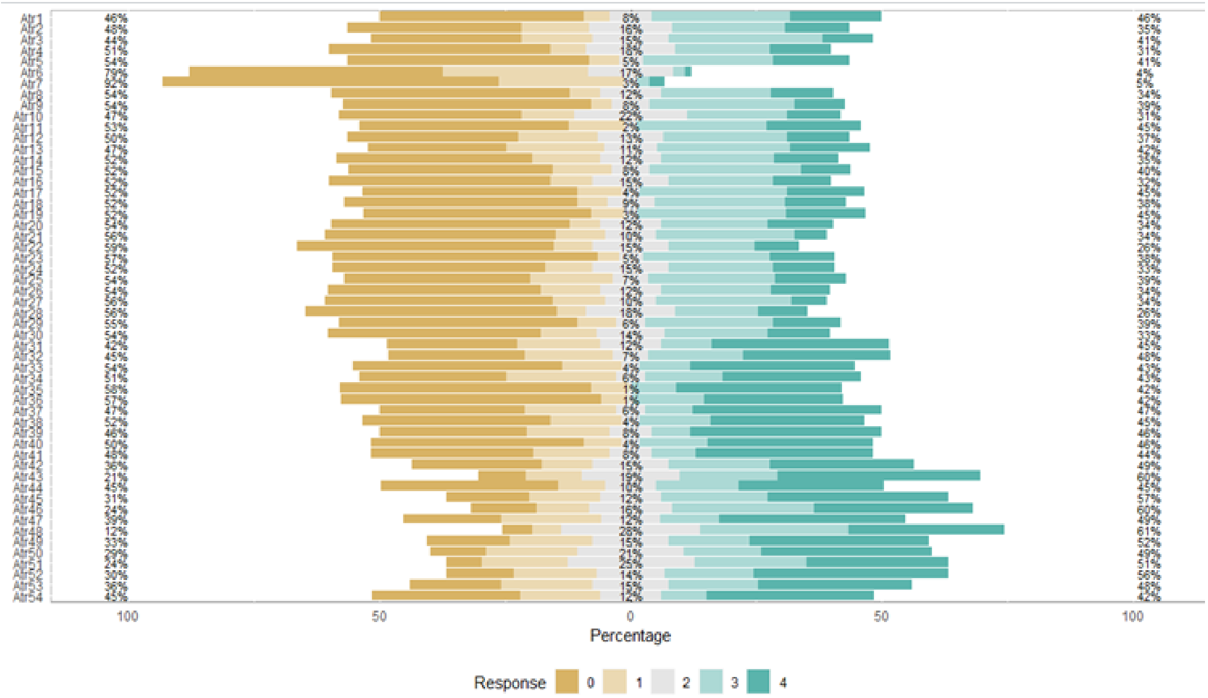


Figure 1.7

We can visualize the percentage of frequency for each column "Atr1" to "Atr54" using the "Likert" plot. Clearly, we can see most of our data is concentrated at 0. "Atr6" and "Atr 7" maybe need our attention as they behave a bit different from the others. Almost all of their data stack at 0 and 1.

Next, we want to study the distribution of each column from "Atr1" to "Atr 54". The distribution of the column can be roughly known by studying the skewness and the kurtosis.

```
> summary(data2_desc[,"skewness"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.6849 -0.0316  0.1948  0.1528  0.3475  2.3272
> summary(data2_desc[,"kurtosis"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.825  -1.666  -1.500  -1.317  -1.419   5.552
```

Figure 1.8

It may appear to be difficult for us to study the skewness and kurtosis directly from the list generated from the function "stat.desc" using library "pastec" so we use the summary method to view the skewness and kurtosis in overall. As stated in McNeese et al., 2020, the rule of thumb the approximate normal distribution will have values from -0.8 to 0.8 for skewness and

-3.0 to 3.0 for kurtosis. From the summary overall mean for kurtosis and skewness look acceptable. We need to choose out those columns which are not in acceptable range.

```
> data2_desc[data2_desc$skewness > 0.8, ]
     median   mean SE.mean CI.mean.0.95    var std.dev coef.var skewness skew.2SE kurtosis
Atr6      0 0.7471  0.0693       0.1369 0.8173  0.9040   1.2101   1.0872   2.9189   0.7466
Atr7      0 0.4941  0.0689       0.1361 0.8077  0.8987   1.8188   2.3272   6.2479   5.5520
     kurt.2SE normtest.W normtest.p
Atr6   1.0079     0.7733          0
Atr7   7.4947     0.5910          0
> data2_desc[data2_desc$skewness < -0.8, ]
 [1] median       mean         SE.mean      CI.mean.0.95 var          std.dev      coef.var
 [8] skewness     skew.2SE     kurtosis     kurt.2SE     normtest.W   normtest.p
<0 rows> (or 0-length row.names)
> data2_desc[data2_desc$kurtosis > 3, ]
     median   mean SE.mean CI.mean.0.95    var std.dev coef.var skewness skew.2SE kurtosis
Atr7      0 0.4941  0.0689       0.1361 0.8077  0.8987   1.8188   2.3272   6.2479    5.552
     kurt.2SE normtest.W normtest.p
Atr7   7.4947      0.591          0
> data2_desc[data2_desc$kurtosis < -3, ]
 [1] median       mean         SE.mean      CI.mean.0.95 var          std.dev      coef.var
 [8] skewness     skew.2SE     kurtosis     kurt.2SE     normtest.W   normtest.p
<0 rows> (or 0-length row.names)
>
```

Figure 1.9

The "Atr6" and "Atr7" are not in the acceptable range. "Atr7" has high kurtosis and skewness and "Atr6" has high skewness. Other columns are approximately normal and no standardization is needed. “Atr6” and “Atr7” are highly positive skewed. More participants rate low scores for Atr6 and Atr7.

2.3 Correlation

We are also interested in the correlation of each and one column. We perform the correlation calculation. It is impossible to show the whole correlation table as it is too large. We only select out the weak correlated pairs. According to the thumb rule the moderate correlated pair must have correlation more than 0.4 to 0.7 and correlation more than 0.7 consider strong, (Schober et al., 2018, p. 1764).

```
 Atr1 Atr10 Atr11 Atr12 Atr13 Atr14 Atr15 Atr16 Atr17 Atr18
    1     2     1     2     2     1     1     1     1     1
Atr19  Atr2 Atr20 Atr21 Atr22 Atr23 Atr24 Atr25 Atr26 Atr27
    1     2     1     1     2     2     1     2     1     1
Atr28 Atr29  Atr3 Atr30 Atr31 Atr32 Atr33 Atr34 Atr35 Atr36
    2     2     3     1     2     2     2     2     2     2
Atr37 Atr38 Atr39  Atr4 Atr40 Atr41 Atr42 Atr43 Atr44 Atr45
    1     1     1     2     1     1     2     4     1     2
Atr46 Atr47 Atr48 Atr49  Atr5 Atr50 Atr51 Atr52 Atr53 Atr54
    8     2     2     2     2     2     2     2     3     2
 Atr6  Atr7  Atr8  Atr9
   52    24     1     1
>
```

Figure 1.10

We observed that most of the uncorrelated items were mostly made up of "Atr6" and "Atr7", "Atr46" and "Atr43". These are columns with correlation less than 0.4.

The strong correlated columns are as below:

```
 Atr1 Atr10 Atr11 Atr12 Atr13 Atr14 Atr15 Atr16 Atr17
   41    40    45    42    42    41    43    43    44
Atr18 Atr19  Atr2 Atr20 Atr21 Atr22 Atr23 Atr24 Atr25
   44    44    39    45    47    45    46    42    44
Atr26 Atr27 Atr28 Atr29  Atr3 Atr30 Atr31 Atr32 Atr33
   44    45    41    47    37    43    37    43    47
Atr34 Atr35 Atr36 Atr37 Atr38 Atr39  Atr4 Atr40 Atr41
   45    48    47    45    46    44    39    46    45
Atr42 Atr43 Atr44 Atr45 Atr46 Atr47 Atr48 Atr49  Atr5
   29     3    42     3     1    10     1    26    47
Atr50 Atr51 Atr52 Atr53 Atr54  Atr6  Atr7  Atr8  Atr9
   37    17     5    19    42     1     1    43    45
> 
```

Figure 1.11

These are pairs with correlation more than 0.7, we can consider these columns as important features.

We also need to test that the correlation we calculate is significant. We calculate the p value of the correlation and set the p-value as 0.05.

```
> inds <- which(data2_cor2$P>0.5, arr.ind = TRUE)
> more_p=data.frame(Var1 = rownames(data2_cor2$P)[inds[, 1]],
+                   Var2 = colnames(data2_cor2$P)[inds[, 2]],
+                   P = data2_cor2$P[inds])
> more_p
[1] Var1 Var2 P
<0 rows> (or 0-length row.names)
> 
```

Figure 1.12

None of the correlations have p value more than 0.05, which means all the correlation is statistically important.

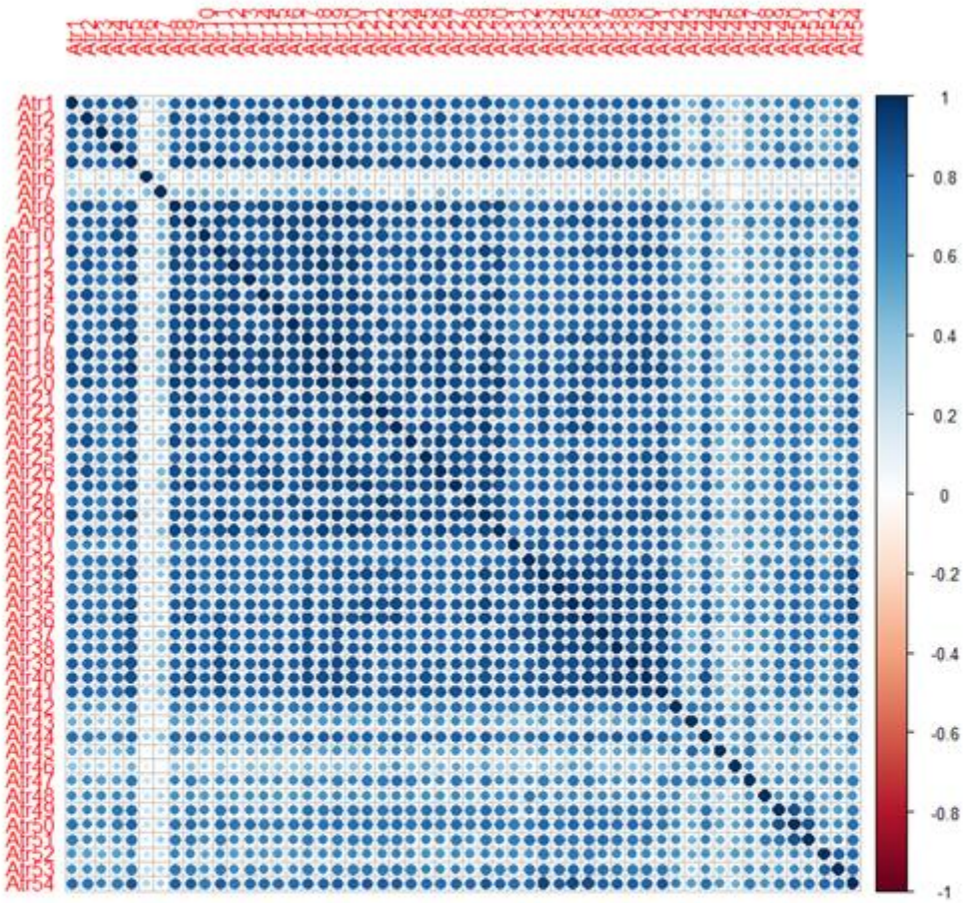


Figure 1.13

We can observe the correlation between each pair clearly from the plot. All the columns have positive correlation with each other. "Atr6", "Atr7" have relatively low correlation.We can say "Atr6" and "Atr7" do not have a direct relationship to others.

2.4 Analysis between two Classes

We plot the frequency of each Class according to score.

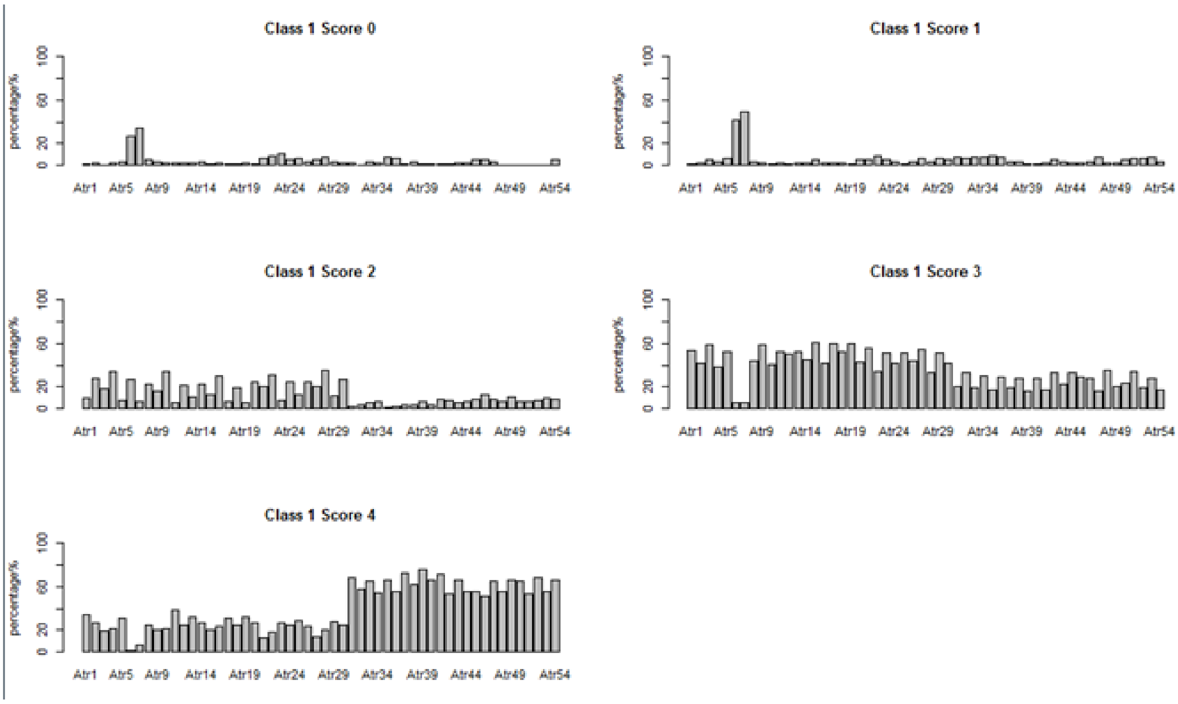


Figure 1.14

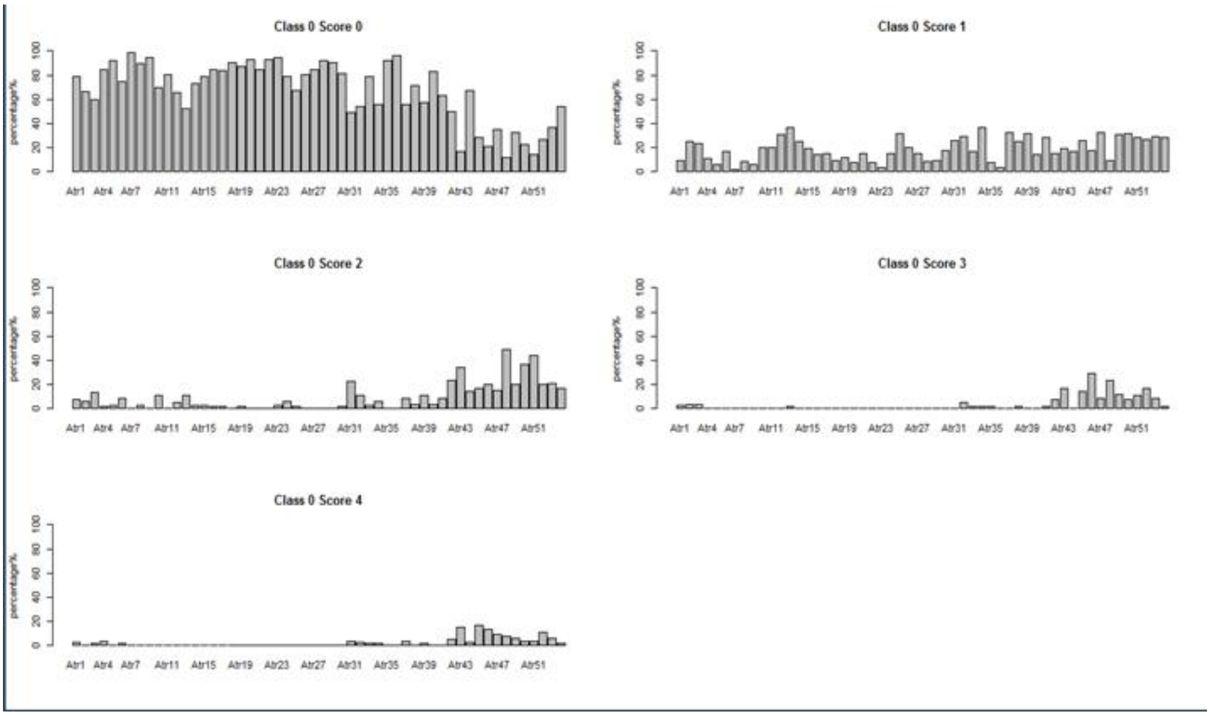


Figure 1.15

0- divorced

1-happily married

From the plot we can view that the happily married couple rate almost all the questions with scores more than 3. The effect is even more obvious for question 31 onward to question 54.

Above 60 to 80% of the respondents with a happily married couple rate these questions with rate 4.

However, most divorced couples rate almost all the questions between 0 to 1.

The graph also shows that Atr6 and Atr7 are not a good indicator for divorce predictor. This is because both happily married, and divorce couples have low rates in Atr6 and Atr7. We cannot really depend on the two questions in predicting the divorce.

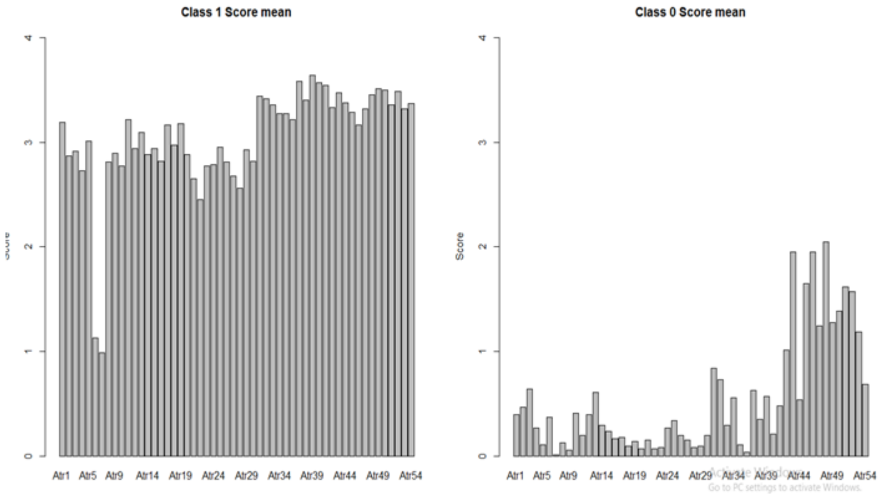


Figure 1.16

The graph shows the mean score for each column of different classes. We can see the mean score for each column in Class 1 is above 2.5 except for "Atr6" and "Atr7". In Class 0 all the columns have scores less than 2.

3.0 Application of algorithms

3.1 Supervised Learning

Data Splitting

We use stratified sampling to split our dataset into training and testing data. We randomly choose elements from each Class (1 or 0) in proportion to the group's size versus the population. We choose this method as it can provide a more accurate representation of the population based on what's used to divide it into different classes.

```
> table(data.train$Class)

 0  1
60 59
> table(data.test$Class)

 0  1
26 25
>
```

Figure 2.0

As a result, we can retain the ratio of class in training and testing data based on the original "Class 1":"Class 0" about 84:86.

3.1.1 kNN

K-nearest neighbors (KNN) is a non-parametric method that is used for classification and regression. In this case, we used the KNN classification method to train the dataset. KNN measures the by the following equation:

$$P(Y = j \mid X = x) = \frac{1}{k} \sum_{i \in N(x)} I(y_i = j)$$

Figure 2.1

The choice of k has a radical effect on the results obtained in the KNN classifier. The optimal k is depending on the data, larger k can decrease the noise on the classification. To find the best k, we have tested with different values of k. In order to get the most accurate accuracy stratification method, and further test with k-fold validation method. After running the predictive model, the results show that the accuracy is the same which is 0.9803922 with different k, k=1,2,3,4,5.

```
-------------------------------------------
Confusion Matrix and performance with K = 1

prediction  0  1
         0 26  1
         1  0 24

 Accuracy: 0.9803922
 Kappa: 0.960739
 Sensitivity: 1
 Specificity: 0.96
 Positive Prediction Value: 0.962963
 Negative Prediction Value: 1
 FPR: 0.04
 FNR: 0
-------------------------------------------
```

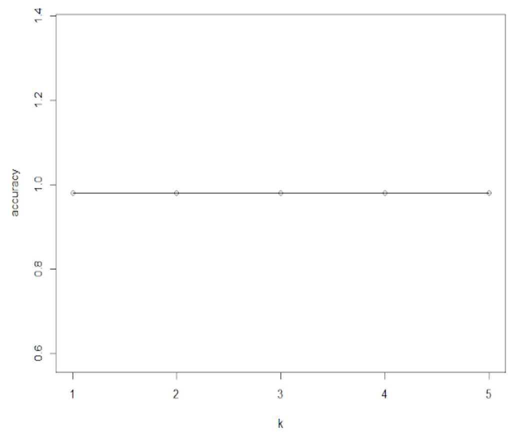


Figure 2.2

Figure 2.2 has shown the confusion Matrix of K=1 and graph of the accuracy plot of K = 1,2,3,4,5. In the k-fold validation method, we have applied a 10-fold validation method. The results show that the overall accuracy of the dataset is 0.9764706.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 12  0
         1  0  5

               Accuracy : 1
                 95% CI : (0.8049, 1)
    No Information Rate : 0.7059
    P-Value [Acc > NIR] : 0.002682

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7059
         Detection Rate : 0.7059
   Detection Prevalence : 0.7059
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

Figure 2.3

Figure 2.3 shown the confusion matrix of one of the folds. The overall accuracy of k-fold validation is calculated as 97.64%. In short, the stratified method is better in using k-fold validation in the KNN predictive model.

3.1.2 Logistic regression

Logistic Regression (LR) algorithm is a predictive analysis and a parametric method used to determine if an independent variable has an effect on a binary dependent variable. This indicates that the logistic regression model has only 2 potential outcomes given an input. Logistics regression model has the ability to handle categorical features. There are types of logistic regression model, here we are performing generalized linear model (GLM) and multinomial logistic regression. The hypothesis function for logistic regression is shown as figure 2.4.

$$P\left(Y = 1 \middle| X_1 = x_1, \dots, X_p = x_p\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}}$$

Figure 2.4

We have distributed 70% of the dataset into training model in order to train the dataset.
With GLM, the accuracy is 0.9803922.

```
data.test.pred  0  1
        pred_0 26  1
        pred_1  0 24

      Accuracy : 0.9803922

         Kappa : 0.960739

   Sensitivity : 1
   Specificity : 0.96
Pos Pred Value : 0.962963
Neg Pred Value : 1
           FPR : 0.04
           FNR : 0
```

Figure 2.5

With Multinomial, the accuracy is 0.9411765.

```
yhat  0  1
   0 25  2
   1  1 23

      Accuracy : 0.9411765

         Kappa : 0.8822171

   Sensitivity : 0.9615385
   Specificity : 0.92
Pos Pred Value : 0.9259259
Neg Pred Value : 0.9583333
           FPR : 0.08
           FNR : 0.03846154
```

Figure 2.6

We also performed 10-fold validation for the GLM model, we got an average accuracy of 0.9529.

As shown above, the generalized linear model showed higher accuracy than multinomial logistic regression. One error is observed for the GLM where one of the couples is happily married but predicted as divorced. The same scenario happened to two couples, and a contradicting result is shown in one couple on multinomial logistic regression.

```
Call:
glm(formula = Class ~ Atr1 + Atr2 + Atr3 + Atr4 + Atr5 + Atr6 +
    Atr7 + Atr8 + Atr9 + Atr10 + Atr11 + Atr12 + Atr13 + Atr14 +
    Atr15 + Atr16 + Atr17 + Atr18 + Atr19 + Atr20 + Atr21 + Atr22 +
    Atr23 + Atr24 + Atr25 + Atr26 + Atr27 + Atr28 + Atr29 + Atr30 +
    Atr31 + Atr32 + Atr33 + Atr34 + Atr35 + Atr36 + Atr37 + Atr38 +
    Atr39 + Atr40 + Atr41 + Atr42 + Atr43 + Atr44 + Atr45 + Atr46 +
    Atr47 + Atr48 + Atr49 + Atr50 + Atr51 + Atr52 + Atr53 + Atr54,
    family = binomial, data = data.train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-4.951e-06  -2.241e-06  -2.110e-08   2.249e-06   5.736e-06

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.084e+01  5.639e+05       0        1
Atr1        -4.354e+00  1.833e+05       0        1
Atr2         3.154e+00  2.032e+05       0        1
Atr3        -2.269e+00  2.367e+05       0        1
Atr4        -1.792e+00  3.985e+05       0        1
Atr5         7.158e+00  4.661e+05       0        1
Atr6         2.841e+00  1.800e+05       0        1
Atr7         7.405e+00  2.955e+05       0        1
Atr8         4.760e+00  2.456e+05       0        1
Atr9        -3.477e+00  3.130e+05       0        1
Atr10        5.654e+00  2.789e+05       0        1
Atr11       -2.417e+00  3.632e+05       0        1
```

Figure 2.7

However, we must be concerned that the GLM has high P-values mainly due to the glm.fit algorithm does not converge, thus resulting in confusion in the output. This suggests a classic case of overfitting where observations are insufficient to support the model, leading to perfect separation (Allison, 2008). To further illustrate, there are one or more variables,which in this case are the questions, that predict the outcome perfectly and subsequently push the conditional likelihood to infinite. In short, we can conclude that the GLM is the more effective model to predict divorce as compared to multinomial logistic regression.

3.1.3 Tree Decision

Decision trees are used to represent choices and their results in the form of a tree. The graph's nodes denote events or choices, while the graph's edges represent decision rules or circumstances. It's mostly used in R-based Machine Learning and Data Mining applications. In most cases, a model is built using observed data, also known as training data (Tutorials Point, 2020). The model is then checked and improved using a set of validation data. R has packages for creating and visualising decision trees. We use this model to decide on the data's category (yes/no, spam/not spam) for a new set of predictor variables (Tutorials Point, 2020).

We train our model using two different libraries, namely the tree library and rpart library. Using two similar algorithms, the results produced from tree library and rpart library are similar as the accuracy obtained from these algorithms are the same.

```
data with tree::tree

tree.pred  0  1
        0 25  1
        1  1 24

      Accuracy : 0.9607843

         Kappa : 0.9215385

   Sensitivity : 0.9615385
   Specificity : 0.96
Pos Pred Value : 0.9615385
Neg Pred Value : 0.96
           FPR : 0.04
           FNR : 0.03846154
```

```
data with tree::Rpart

tree.pred.rpart  0  1
              0 25  1
              1  1 24

      Accuracy : 0.9607843

         Kappa : 0.9215385

   Sensitivity : 0.9615385
   Specificity : 0.96
Pos Pred Value : 0.9615385
Neg Pred Value : 0.96
           FPR : 0.04
           FNR : 0.03846154
```

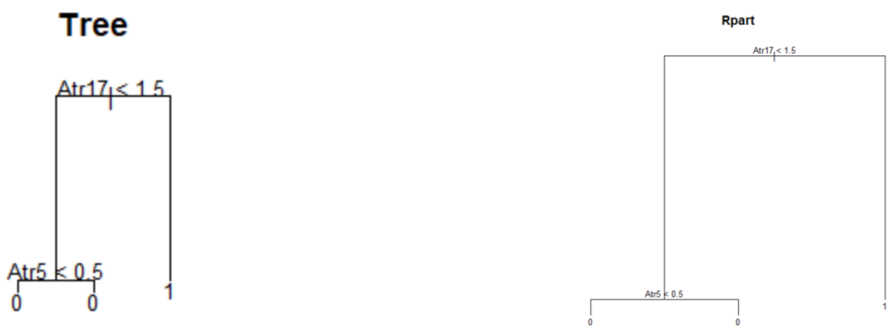


Figure 2.8

As we can infer from the decision tree graph, if the respondent answer on the question Atr 17 (We share the same views about being happy in our life with my spouse) are either seldom or never, no matter what their answer on the Atr 5 (The time I spent with my wife is special for us) are, they are destined to be divorced. On the other hand, if their answers on question Atr17 are average, frequently or always, they are happily married. By applying this decision tree into our testing data, the accuracy we obtained are as high as 96.07843%. We performed 10 folds validation to obtain a further test on the accuracy of this model. Therefore, we get an average mean of 97.06%.

3.1.4 Random Forest

There are some disadvantages of using a decision tree. The decision tree will have some inaccuracy. Hastie et al., (2008) stated that often the decision tree appears to be inflexible as it fixed the training data too well. However, the random forest does well as it combines a huge variety of trees resulting in flexibility. It uses bootstraps which randomly select samples with the same size from the original dataset. Same sample is allowed to pick more than once, (Yiu, 2019).

```
> rf.data=randomForest(Class~.,data=data.train,importance=TRUE)
> rf.data

Call:
 randomForest(formula = Class ~ ., data = data.train, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7

        OOB estimate of  error rate: 2.52%
Confusion matrix:
   0  1 class.error
0 60  0  0.00000000
1  3 56  0.05084746
>
```

Figure 2.9

We train our model using the default setting. It results for us that 500 trees are built, and the number of variables tried at each split is 7. That means that the model will use a random subset of variables of 7 at each step of the tree.

Next, we want to check whether the random forest produced enough trees for modelling or not. We plot Out of Bag error (OOB) with respect to the number of trees until 500 trees. The OOB is obtained by testing accuracy of the Out of Bag data set which is not included in the boosting dataset (Yiu, 2019).

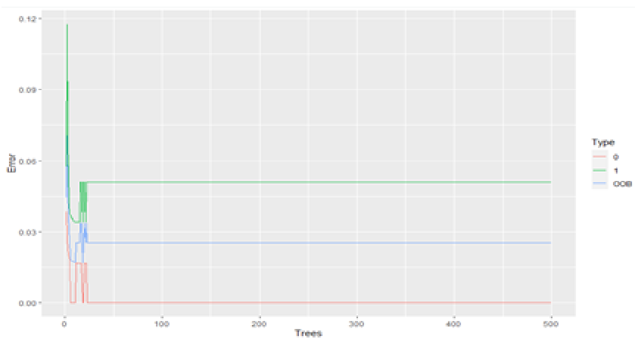
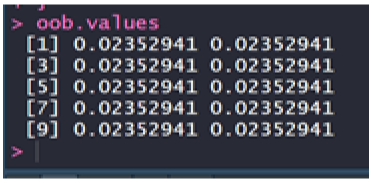


Figure 2.10

We can observe that the error is consistent from 100 to 500 trees. And it would not show any reduction if we increase the number of trees further.

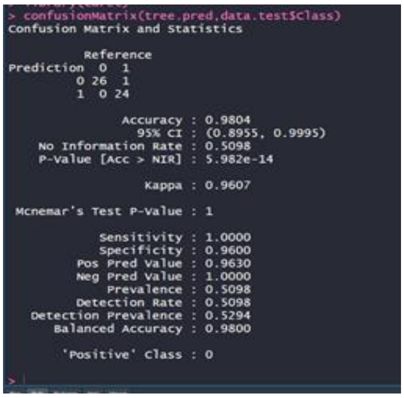
We also try the number of variables tried each step to obtain the most optimal number. We try one to ten.



```
> oob.values
 [1] 0.02352941 0.02352941
 [3] 0.02352941 0.02352941
 [5] 0.02352941 0.02352941
 [7] 0.02352941 0.02352941
 [9] 0.02352941 0.02352941
>
```

Figure 2.11

The error rate does not change from one to ten. It is fine for us to use the default setting which is 7.



```
> confusionMatrix(tree.pred,data.test$Class)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 26  1
         1  0 24

               Accuracy : 0.9804
                 95% CI : (0.8955, 0.9995)
    No Information Rate : 0.5098
    P-Value [Acc > NIR] : 5.982e-14

                  Kappa : 0.9607

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9600
         Pos Pred Value : 0.9630
         Neg Pred Value : 1.0000
             Prevalence : 0.5098
         Detection Rate : 0.5098
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9800

       'Positive' Class : 0

>
```

Figure 2.12

The model gives us 98.04% if we test it with the testing data.

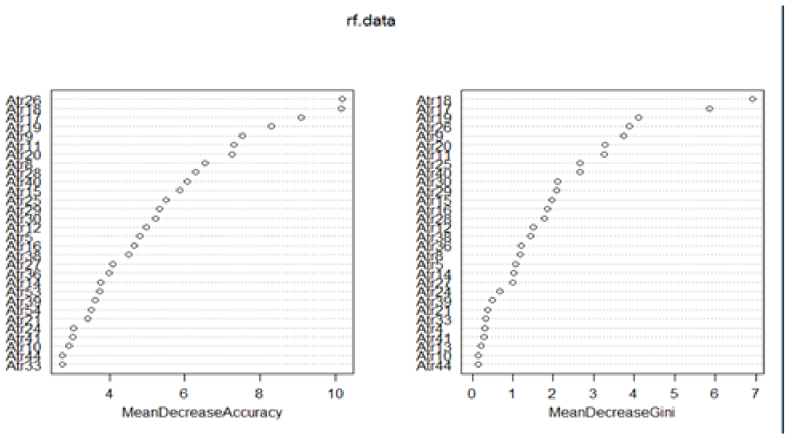


Figure 2.13

According to Martinez-Taboada & Redondo, (2018),the Mean Decrease Accuracy plot expresses how much accuracy the model losses by excluding each variable and. The Mean Decrease Gini measures how each variable contributes to the production of trees. The higher the MDA and MDG the more important is the variables.

To further test the accuracy of the model, we also perform 10 folds validation using the random forest the average mean is 97.64706%.

3.1.5 Linear Discriminant Analysis

LDA uses the linear combinations of predictors to predict the class of a given observation. Using the LDA we assume that Atr=(AtrX) where X=1,2, …,54 follow approximately multivariate Gaussian distribution. From the EDA part we know that each of the columns of our data is followed approximately normal therefore this statement is acceptable. The distribution of P(Atr=Atr X|Class=k) is

$$\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \boldsymbol{\Sigma}^{-1}(x-\mu)\right)$$

Figure 2.14

It also assumes that each Attr=AttrX has same covariance matrices ∑.

Using Bayesian classifier the P(Class=k|Attr=Attr X) can be represent in form:

$$\delta_k(x) = x^T \boldsymbol{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1}\mu_k + \log \pi_k$$

Figure 2.15

```
Prior probabilities of groups:
        0         1
0.5042017 0.4957983
```

Figure 2.16

When we train the model the prior probabilities π_(Class 1)and π_(Class 0) is calculated.

```
Group means:
       Atr1      Atr2      Atr3    Atr4     Atr5      Atr6       Atr7
0 0.3833333 0.5166667 0.6833333 0.30000 0.100000 0.3333333 0.01666667
1 3.1355932 2.8813559 2.8813559 2.79661 2.966102 1.1355932 0.98305085
        Atr8      Atr9     Atr10     Atr11 Atr12     Atr13     Atr14     Atr15
0 0.08333333 0.050000 0.4333333 0.2166667  0.45 0.6666667 0.3333333 0.2333333
1 2.83050847 2.864407 2.8813559 3.2203390  3.00 3.0338983 2.9322034 2.9322034
      Atr16     Atr17     Atr18     Atr19    Atr20    Atr21    Atr22
0 0.1333333 0.1333333 0.1166667 0.1833333 0.050000 0.150000 0.050000
1 2.8983051 3.1355932 3.0338983 3.1355932 2.915254 2.644068 2.542373
       Atr23     Atr24     Atr25    Atr26     Atr27      Atr28    Atr29
0 0.06666667 0.2333333 0.3333333 0.200000 0.1333333 0.08333333 0.100000
1 2.66101695 2.7966102 2.9322034 2.881356 2.6610169 2.69491525 2.915254
      Atr30    Atr31     Atr32    Atr33     Atr34     Atr35      Atr36
0 0.1333333 1.016667 0.7666667 0.250000 0.5333333 0.1333333 0.03333333
1 2.8813559 3.406780 3.4237288 3.338983 3.2542373 3.2372881 3.16949153
      Atr37    Atr38    Atr39    Atr40     Atr41    Atr42    Atr43     Atr44
0 0.6333333 0.300000 0.600000 0.250000 0.4833333 1.100000 2.016667 0.6333333
1 3.5593220 3.440678 3.661017 3.559322 3.5084746 3.338983 3.457627 3.3389831
     Atr45    Atr46    Atr47    Atr48    Atr49    Atr50    Atr51    Atr52
0 1.750000 2.050000 1.266667 2.150000 1.383333 1.466667 1.700000 1.600000
1 3.186441 3.118644 3.254237 3.423729 3.508475 3.440678 3.322034 3.457627
     Atr53     Atr54
0 1.150000 0.6166667
1 3.305085 3.3559322
```

Figure 2.17

The vector of $\mu_{Class\ 1}[Attr1, \ldots, Attr54]$ and $\mu_{Class0}[Att1, \ldots, Attr54]$ are shown in table form.

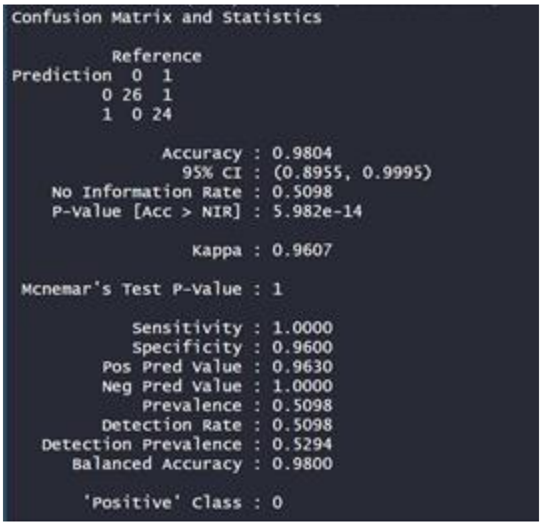
```
Coefficients of linear discriminants:
               LD1
Atr1  -0.341592026
Atr2   0.949494550
Atr3  -0.123941382
Atr4  -0.334413677
Atr5   1.104534024
Atr6   1.067015789
Atr7   1.106925393
Atr8   0.988641506
Atr9  -0.110200648
Atr10  0.007089952
Atr11 -0.138812473
Atr12  0.667723381
Atr13 -1.200445560
Atr14 -0.052574837
Atr15  1.004032347
Atr16 -1.146749756
Atr17  1.601370104
Atr18  0.463839881
Atr19  0.230041763
Atr20 -0.586156923
Atr21 -0.008766651
Atr22 -0.569214257
Atr23 -0.539907357
Atr24 -0.277095754
Atr25  0.011946353
Atr26  0.527344900
Atr27 -0.210467536
Atr28  2.052813793
Atr29  0.785024481
```


```
Atr28  2.052813793
Atr29  0.785024481
```


```
Atr30 -1.283886999
Atr31  0.457221745
Atr32  0.004115706
Atr33 -0.121260288
Atr34 -0.054459807
Atr35  0.010404720
Atr36 -0.597765606
Atr37 -0.387202775
Atr38  0.818671559
Atr39 -0.081226010
Atr40  0.518759830
Atr41  0.070089683
Atr42  0.055790165
Atr43  0.334128326
Atr44 -0.138982423
Atr45 -0.162511222
Atr46 -0.152809931
Atr47 -0.138954318
Atr48 -0.753722537
Atr49  0.463394429
Atr50 -0.284623979
Atr51  0.778368700
Atr52  0.494088765
Atr53  0.487190263
Atr54 -0.668159348
```

Figure 2.18

The linear combination of variables shown by the coefficients of linear discriminant is used to form LDA decision rules.


```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 26  1
         1  0 24

               Accuracy : 0.9804
                 95% CI : (0.8955, 0.9995)
    No Information Rate : 0.5098
    P-Value [Acc > NIR] : 5.982e-14

                  Kappa : 0.9607

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9600
         Pos Pred Value : 0.9630
         Neg Pred Value : 1.0000
             Prevalence : 0.5098
         Detection Rate : 0.5098
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9800

       'Positive' Class : 0
```

Figure 2.19

In testing we separate the group using the default cut-off value of posterior probability 0.5. From the testing we can obtain accuracy of 0.9804 using the LDA model. By doing 10-folds validation, we obtain mean accuracy of 0.9823529.

2 Unsupervised Learning

Unsupervised learning, we train the data without depending on the label or target variable. In our case we drop our target variable which is the column "Class". We have the aim to discover the group based on the similar characteristic of data.

3.2.1 PCA

Principal Component Analysis, aka, PCA is one of the commonly used approaches to do unsupervised learning, feature extraction, and dimensionality reduction (cmdline, 2019).

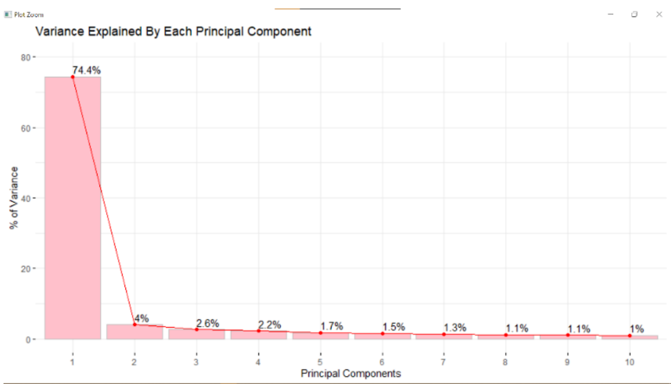


Figure 2.20

In our case, we see that the first principal component explains most of the variation in our data. Actually, it explains 74% of the variance (variance of 40) and the remaining 53 PCs explains the rest of the variation. It suggests that the first principal component is driving almost all of the variation in our data. In essence it helps us reduce the dimension of the data. In our example, with just one dominant principal component, we have reduced the dimension 54 to only 1

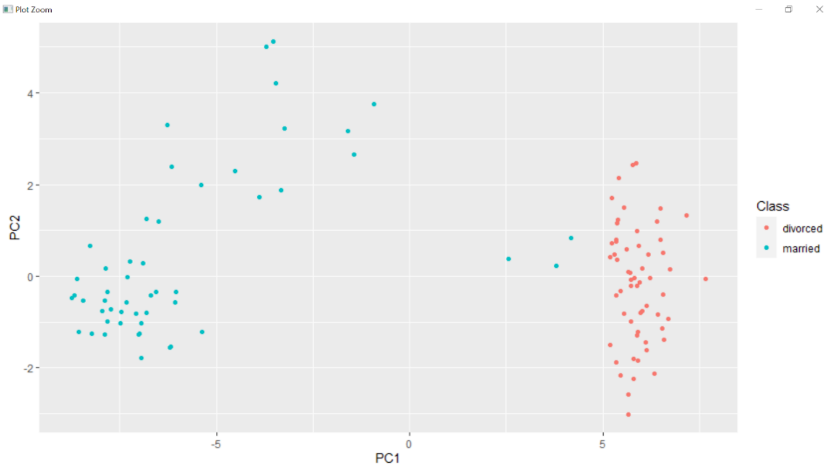


Figure 2.21

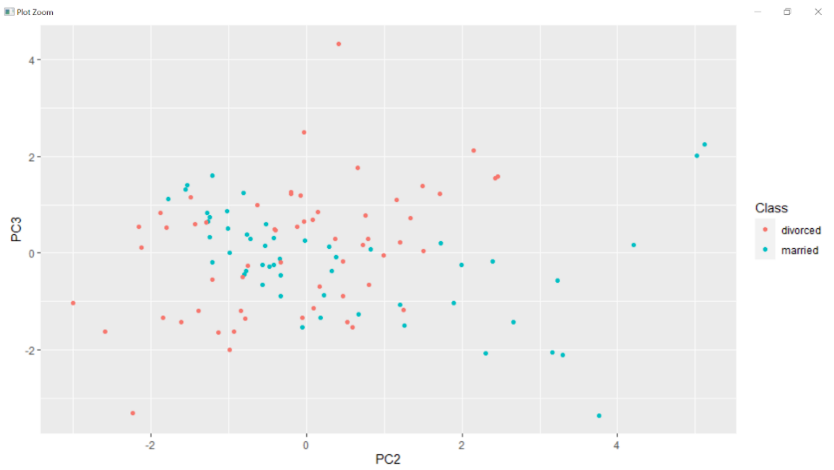


Figure 2.22

By comparing the the difference between the figures of comparison between PC1 and PC2 , and PC2 and PC3, we found that the first principal components eventually is more effective the data points

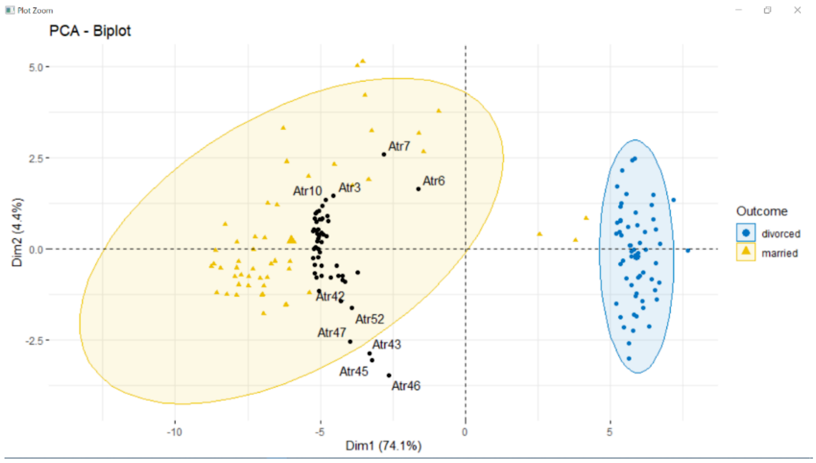


Figure 2.23

In essence, there are a lot of things that we can derive from the biplot above. For instance, it is worth noting that Atr46, Atr6, and Atr7 have no significant correlation relationship with other Atr as the angles between them are greater compared to others'. On the other hand, It is noted that Atr46, Atr45, and Atr7 do have strong influence on PC2 as their value on PC2 are greater than others', while there are many attributes as we can observe from the figure above their value are around -5 on PC1.

From the information granted above, It suffices to say that PCA might be the best way to visualize our data by reducing the dimension.

3.2.2 K-mean clustering

K-mean clustering is a distance-based clustering; it clusters the point based on the distance. The internal distances should be small while the external distances should be large. The algorithm will separate the data into k clusters based on the distance to the centroid of the cluster, (Fonseca, 2019).

K-mean clustering is suitable for our data set because from the EDA we found that there is an obvious score trend in each question for the data separate in each class. Class 1 is above 2.5 except for "Atr6" and "Atr7". In Class 0 all the columns have scores less than 2. This provides us a confident, the algorithm will work well as there is a clear distance between each class.

We use the fviz_nbclust() function to estimate the optimal number of each cluster.

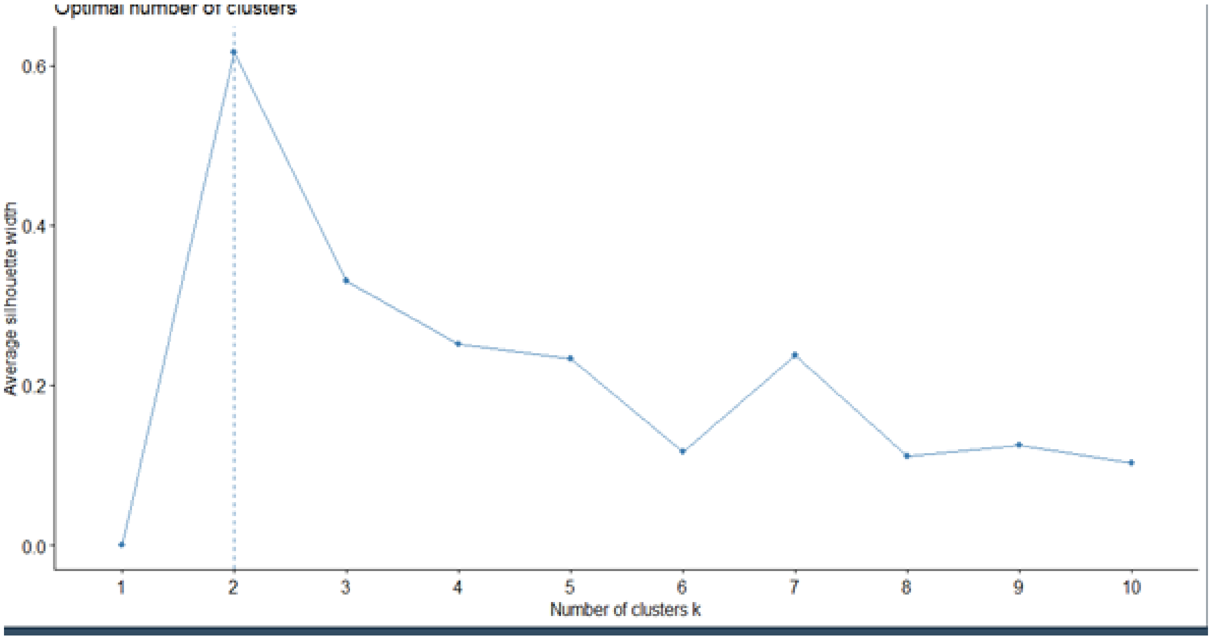


Figure 2.24

We use the silhouette method to measure the quality of clustering. The silhouette method measures how well each object lies within its cluster. Highest average silhouette indicates the optimal number of cluster k. In our case the most optimal k is k=2. This result coincides with our 2 classes of target variables which are Class1 and Class2.

```
> str(data2Cluster)
List of 9
 $ cluster     : int [1:170] 2 1 1 1 2 2 1 1 1 2 ...
 $ centers     : num [1:2, 1:54] 3.288 0.433 2.95 0.5 2.975 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr [1:54] "Atr1" "Atr2" "Atr3" "Atr4" ...
 $ totss       : num 21635
 $ withinss    : num [1:2] 2898 2886
 $ tot.withinss: num 5784
 $ betweenss   : num 15852
 $ size        : int [1:2] 80 90
 $ iter        : int 1
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

Figure 2.25

Here we set the k=2 as it is the most optimal k value, we also set the initial configuration as 25. We can see from the cluster it separates our data into two clusters with label 1 and 2. The size of group 1:2 is 80:90.

```
> print(data2Cluster)
K-means clustering with 2 clusters of sizes 80, 90

Cluster means:
       Atr1 Atr2      Atr3      Atr4      Atr5      Atr6       Atr7      Atr8       Atr9
1 3.2875000 2.95 2.9750000 2.8250000 3.1375000 1.1250000 1.03750000 2.9250000 3.01250000
2 0.4333333 0.50 0.6888889 0.2888889 0.1222222 0.4111111 0.01111111 0.1444444 0.07777778
      Atr10     Atr11     Atr12 Atr13     Atr14     Atr15     Atr16     Atr17     Atr18
1 2.8750000 3.3250000 3.0500000 3.225 2.9750000 3.0375000 2.9375000 3.2750000 3.0750000
2 0.4222222 0.2333333 0.4111111 0.600 0.3222222 0.2666667 0.1777778 0.2111111 0.1333333
      Atr19  Atr20     Atr21      Atr22      Atr23     Atr24     Atr25     Atr26
1 3.3000000 2.9875 2.7750000 2.57500000 2.91250000 2.9125000 3.0625000 2.8875000
2 0.1666667 0.1000 0.1555556 0.06666667 0.07777778 0.2666667 0.3555556 0.2444444
      Atr27     Atr28     Atr29     Atr30     Atr31     Atr32     Atr33     Atr34
1 2.7875000 2.6500000 3.0500000 2.9250000 3.5250000 3.5250000 3.4875000 3.3750000
2 0.1666667 0.1111111 0.1111111 0.2222222 0.8777778 0.7555556 0.3111111 0.5888889
      Atr35      Atr36     Atr37     Atr38     Atr39     Atr40     Atr41    Atr42
1 3.4250000 3.35000000 3.7125000 3.5250000 3.7375000 3.6500000 3.6625000 3.425000
2 0.1111111 0.05555556 0.6444444 0.3777778 0.6222222 0.2888889 0.5111111 1.033333
     Atr43     Atr44    Atr45    Atr46    Atr47    Atr48    Atr49    Atr50    Atr51
1 3.537500 3.4625000 3.362500 3.225000 3.437500 3.525000 3.562500 3.575000 3.437500
2 1.966667 0.5888889 1.655556 1.955556 1.233333 2.044444 1.333333 1.411111 1.622222
     Atr52    Atr53  Atr54
1 3.537500 3.387500 3.4875
2 1.611111 1.222222 0.7000

Clustering vector:
  [1] 2 1 1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [43] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [85] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[127] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[169] 2 2

Within cluster sum of squares by cluster:
[1] 2897.588 2886.111
 (between_SS / total_SS =  73.3 %)
```

Figure 2.26

We can find the mean of each column in each group. The within cluster sum of squares by cluster is 73.3%. It shows the compactness of the clustering and determines how similar are the members within the same group. 73.3% is in our acceptable range.

Figure 2.27

We visualize the clustering using fviz_cluster(). As our variable is more than 2 the fviz_cluster() will perform Principal Component Analysis(PCA) to reduce the dimension into 2. The data is plotted using the first two principal components coordinates.

Next, we perform clustering validation using silhouette coefficient to evaluate the goodness of our clustering.

```
> fviz_silhouette(sil)
  cluster size ave.sil.width
1       1   80          0.61
2       2   90          0.62
```

Figure 2.28

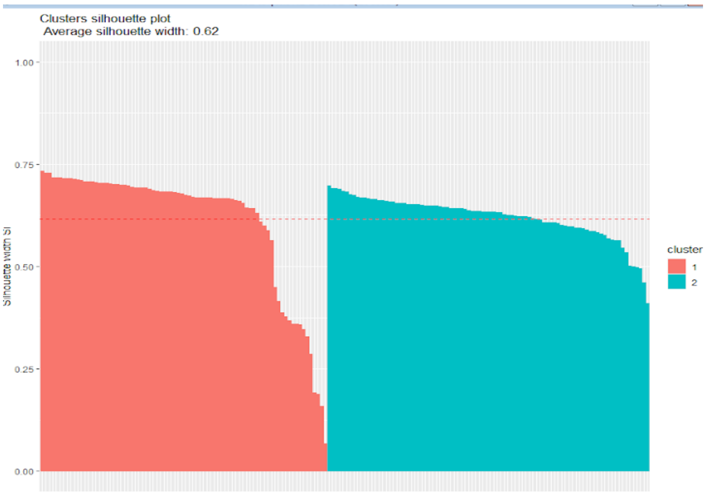


Figure 2.29

The graph shows the average silhouette width of the clustering.Fonseca, (2019) states that avg.sil.width more than 0 represents a well clustered; avg.sil.width less than 0 represents a wrong clustering. Avg.sil.width equal to 0 shows the observation is between two clusters. Our model is considered good as the ave.sil.width is more than 0.5. The average silhouette of our model is 0.62.

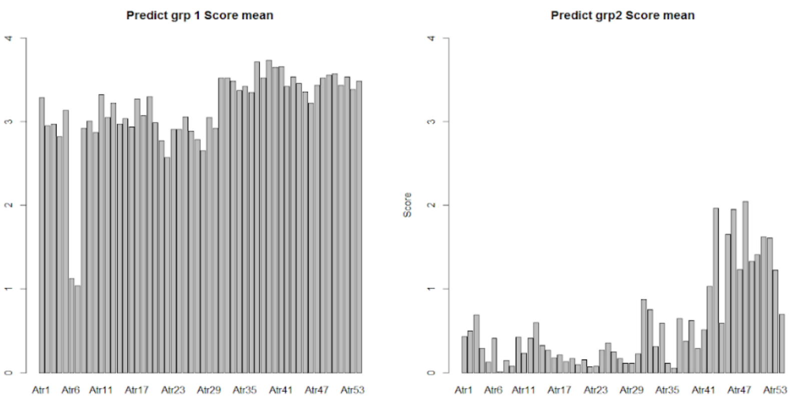


Figure 2.30

As we compare to the original label "Class1" and "Class0", the predicted group1 has mean similar to "Class 1" and predicted group2 has mean similar to "Class 0". Therefore, we can replace the group 1 as "Class 1" and group2 as "Class 0" to test the accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 86  4
         1  0 80

               Accuracy : 0.9765
                 95% CI : (0.9409, 0.9936)
    No Information Rate : 0.5059
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9529

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 1.0000
            Specificity : 0.9524
         Pos Pred Value : 0.9556
         Neg Pred Value : 1.0000
             Prevalence : 0.5059
         Detection Rate : 0.5059
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9762

       'Positive' Class : 0
```

Figure 2.31

The accuracy is calculated when we compare our predicted group with the original group. We found that the accuracy of this model is 97.67%. The model is considered good.

3.2.3 Hierarchical Clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is a method of grouping related objects into clusters (Tim, 2018). The endpoint is a series of clusters, each of which is different from the others while the artefacts within each cluster are broadly identical (Tim, 2018).

We performed the agglomerative hierarchical clustering criterion: distance matrix using the euclidean and the complete linkage (that measures the distance between the cluster). Going on, we further add a border around the two largest clusters, k=2 as we have identified that k=2 is the desirable number as indicated by the highest average silhouette.

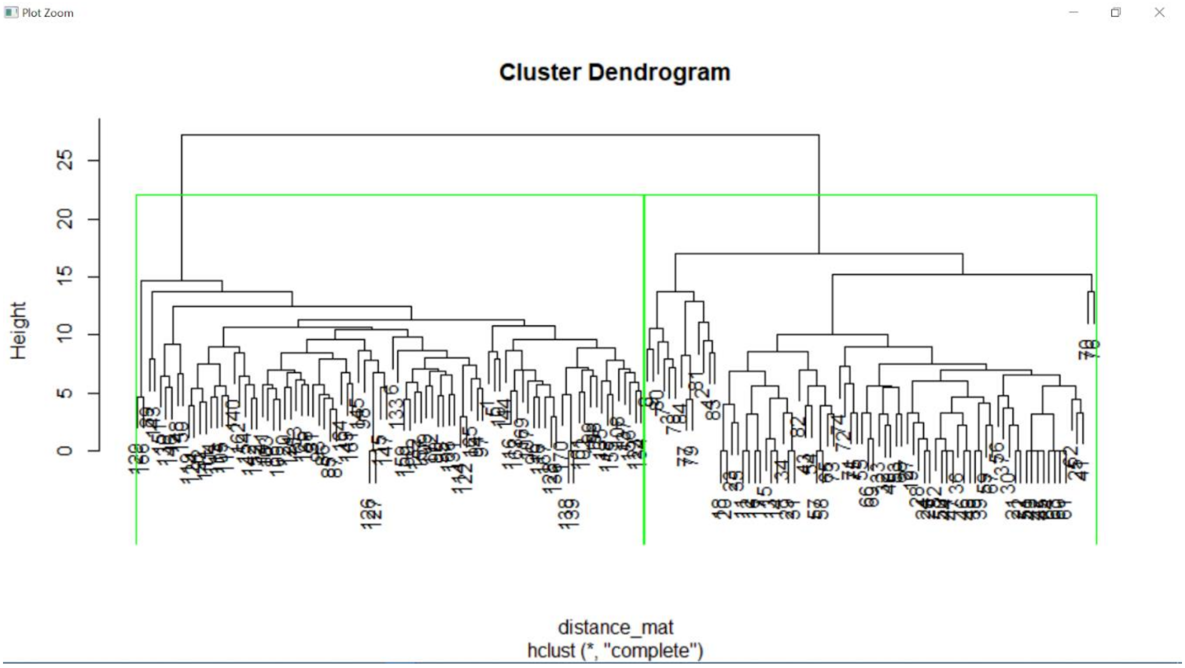


Figure 2.32

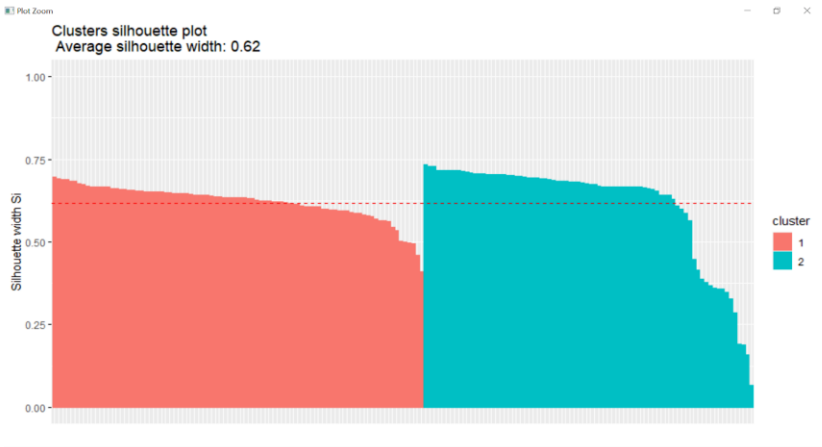


Figure 2.33

As we can infer from the average silhouette width of the clustering graph above, both Cluster 1 and 2 have average widths of 0.61 and 0.62, which could be considered as a good clustering. However, it can be seen that there are relatively low silhouette coefficients in several samples in cluster 2.

On the other hand, the dunn index calculated using the cluster.stats showing a figure of 0.6056253, suggesting a higher-than-average compactness and well separation of the clusters.

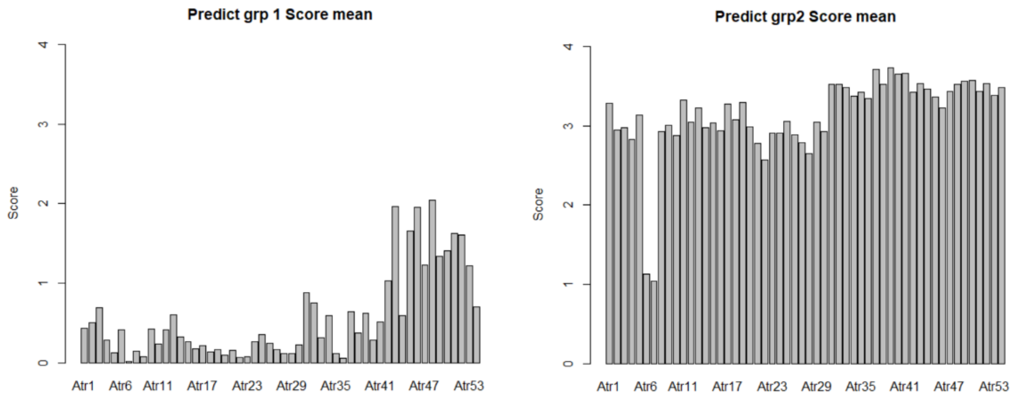


Figure 2.34

As we compare to the original label "Class1" and "Class0", the predicted group1 has mean similar to "Class 0" and predicted group2 has mean similar to "Class 1". Therefore, we can replace the group 1 as "Class 0" and group2 as "Class 1" to test the accuracy.

```
          Reference
Prediction  0  1
         0 86  4
         1  0 80

               Accuracy : 0.9765
                 95% CI : (0.9409, 0.9936)
    No Information Rate : 0.5059
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9529

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 1.0000
            Specificity : 0.9524
         Pos Pred Value : 0.9556
         Neg Pred Value : 1.0000
             Prevalence : 0.5059
         Detection Rate : 0.5059
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9762

       'Positive' Class : 0
```

Figure 2.35

By applying the hierarchical clustering method, we obtained an accuracy of 97.65%.

4.0 Comparison between models and analysis

	Model	Average Accuracy
Supervised	kNN	97.65%
	Logistic regression	95.29%
	Tree Decision	97.06%
	Random Forest	97.65%
	LDA	98.24%
Unsupervised	K-mean clustering	97.65%
	Hierarchical clustering	97.65%

Table 1.0 Model and average accuracy

We can see from the table both supervised models and unsupervised models have high accuracy (>95%) in predicting the outcome. We can say that the predictors “Atr1” to “Atr54” are good predictors to predict target variable “divorce” or “happily marriage”. In our report we calculate the accuracy by splitting the data into test set and train set. The model is further validating using the 10 folds validation. In the comparison part, we determine the best model by comparing the average accuracy of the 10 folds validation.

LDA has resulted in the highest accuracy among all the models which is 98.24%. The LDA can result in high accuracy because our variables follow approximately normal which meet the assumption of LDA. That is why the model can represent the data very well. The logistic regression results in the lowest accuracy. The model is also not statistically reliable due to its perfect separation issue. Other models which are based on the distance have almost the same accuracy (ie kNN, K-mean clustering). The error of K-mean clustering may due to the misclassification as we can see from the PCA analysis there are three points
Of “divorce” is quite close to the cluster of “happily married”. The tree decision has also some disadvantages due to the over-fitting; this may cause it to have lower accuracy compared to the random forest.LDA might be the most suitable model for the divorce dataset.

From the study of correlation, we found that Atr 6 and Atr 7 do not have good correlation as compared to the other variable.

6. We don't have time at home as partners.

7. We are like two strangers who share the same environment at home rather than family

From the mean of both questions, we can see both “Class 1” and “Class 0” rate both the question with very low scores: “Class 1” is about 1 and “Class 0” is less than 1. We can know that don’t having time at home as a partner is not a very good contribution to poor marriage. This is because the advancement of technology has connected people regardless of distance. A lot of young couples are busy working. They depend on the communication technology to redefine their relationship. The technology let them feel connected and seen. Some of the couples like to hang out together instead of staying home. If we refer to the Atr 7, it is impossible for both people to be a couple if they treat each other as strangers who share the same environment. Therefore, they probably would not end up in a marriage.

29. I know my spouse very well.

18. My spouse and I have similar ideas about how marriage should be

17. We share the same views about being happy in our life with my spouse

19. My spouse and I have similar ideas about how roles should be in marriage

9. I enjoy traveling with my wife.

These five questions are the top 5 important variables in predicting divorce. This is because these five questions have the highest Mean Decrease Accuracy and Mean Decrease Gini calculated using the random forest method. From the five questions we can see one thing in common which is about understanding each other. One must understand the roles of one another and provides help and empathy. One also must agree with the roles of one another. Knowing each other well is important to solve all the disputes in marriage. From the question we know understanding can be built through spending time together like travelling.

5.0 Conclusion

After carrying out some supervised and unsupervised learnings, we found out about Linear Discriminant Analysis(LDA) has the best and highest accuracy among the other models. The accuracy of LDA is 98.24%. Therefore, the best model for Divorce predictor data set is LDA. It has met our objective which is to predict whether the husband and wife are happy married or going to divorce based on the 54 questions.

However, we are also going to recommend some unsupervised learnings for future study purposes. We study lesser unsupervised models compared to supervised learning models. For improvement, we can add on some unsupervised learning such as Density-based clustering, grid-based clustering and kernel spectra based clustering. The more models we compared, the more accurate that we can determine for the best model to use in Divorce predators data set.

References

Allison, P. (2008). *Convergence Failures in Logistic Regression*. [online] ResearchGate. https://www.researchgate.net/publication/228813245_Convergence_Failures_in_Logistic_Regression

Fonseca, L. (2019, August 16). *Clustering Analysis in R using K-means - Towards Data Science*. Medium. https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning* (2nd ed.) [E-book]. Springer Publishing. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Publishing.

Martinez-Taboada, F., & Redondo, J. I. (2018). Variable importance plot (mean decrease accuracy and mean decrease Gini). *History Studies International Journal of History*, *10*(6), 215–224. https://doi.org/10.9737/hist.2018.644

McNeese, B., J., Ven, A., P., McNeese, B., S., McNeese, B., A., McNeese, B., Best, K., McNeese, B., Stevenson, T., McNeese, B., Stevenson, T., McNeese, B., R., McNeese, B., Y., McNeese, B., … McNeese, B. (2020, April 25). *Are the Skewness and Kurtosis Useful Statistics?* BPI Consulting. https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics#:%7E:text=The%20rule%20of%20thumb%20seems,the%20data%20are%20highly%20skewed

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Tim, B. (2018, March 28). What is Hierarchical Clustering? | Displayr.com. Retrieved March 30, 2021, from Displayr website: https://www.displayr.com/what-is-hierarchical-clustering/#:~:text=Hierarchical%20clustering%2C%20also%20known%20as

Tutorials Point. (2020). R Tutorial - Tutorialspoint. Retrieved March 23, 2021, from Tutorialspoint.com website: https://www.tutorialspoint.com/r/index.htm

Yiu, T. (2019, August 14). *Understanding Random Forest - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Yöntem, M.K. and İlhan, T. (2018). Boşanma Göstergeleri Ölçeğinin Geliştirilmesi. [Development of the Divorce Predictors Scale]. Sosyal Polika Çalışmaları Dergisi. 41, 339-358.

Yöntem, M , Adem, K , İlhan, T , Kılıçarslan, S. (2019). DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS. Nevşehir Hacı Bektaş Veli University SBE Dergisi, 9 (1), 259-273.

Evaluation Form

Members	**Contribution**
Tan Eng Sim	14%
Chao Xin Yi	14%
Lim Li Ting	14%
Lee Shu Ying	14%
Lee Yang	14%
Rachel Chin Chi Shan	14%
Sio Wen Kang	16%
Total	100%