The Four Horsemen of the Divorce

Li PinZhao, Liu YanYu, Dong GengShang, Vincent William Hadiasali

May 2024

SUSTech
Southern University of Science and Technology

Department of Statistics & Data Science

1 Data Introduction

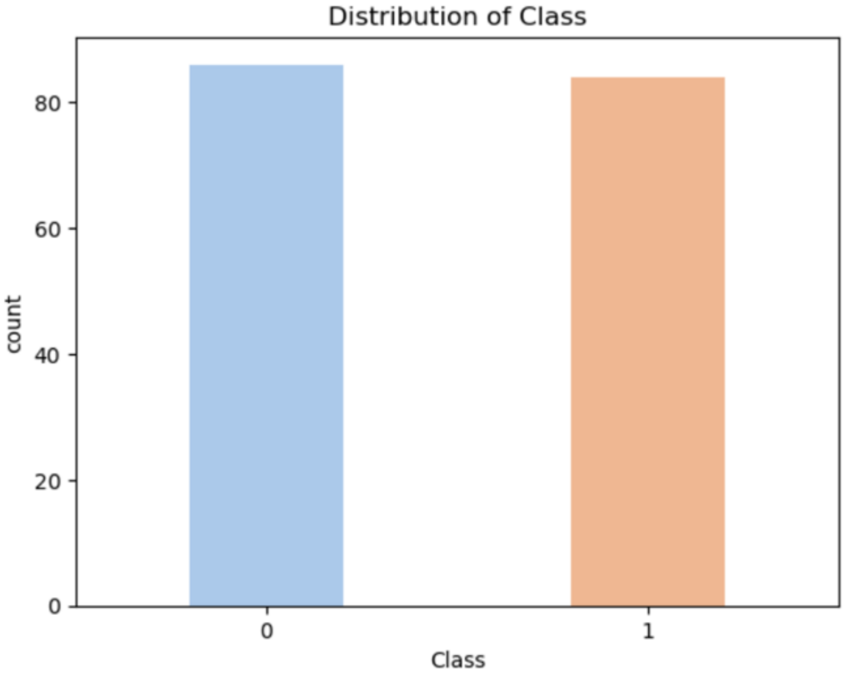
Gottman highlights the combination of predictors such as the ”four horsemen” (Criticism, Contempt, Stonewalling, Defensiveness) along with failed repair attempts as highly indicative of future divorce. We will explore the relationship between the ”four horsemen” and divorce.

The Divorce Predictors Scale (DPS) dataset is derived from a study focused on predicting divorce using the DPS within the framework of Gottman couples therapy. The dataset comprises responses collected from participants, consisting of both divorced and married couples.

The dataset includes two main components:

- Personal Information Form: This section contains demographic and personal details provided by the participants.
- Divorce Predictors Scale: Participants completed the DPS, which consists of items related to various aspects of relationships, including creating a common meaning, failed attempts to repair, love map, and negative conflict behaviors.

From the plot below, we can find out that the class of the dataset is almost balanced, which means there are as many divorced people as there are not.

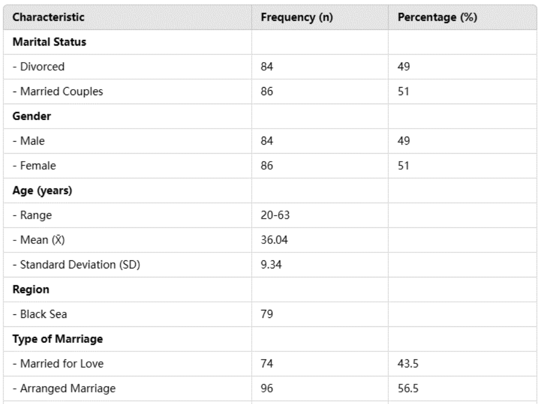


1.1 Questionnaire setting

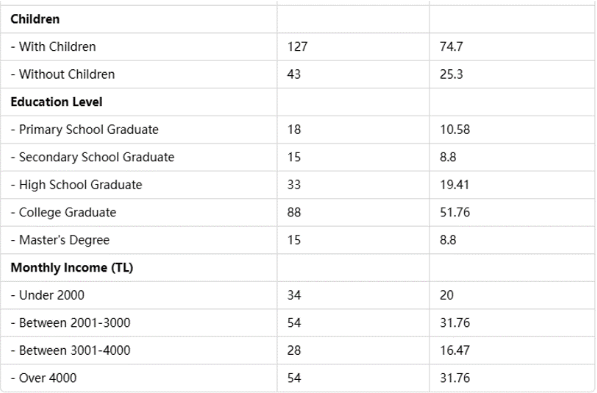
There are a total of 54 questions in the questionnaire, which are named ”Art1” to ”Art54” for convenience. All the questions will be available in .txt file.

1.2 Data Statistics

The statistical data are shown in the table below:



Characteristic	Frequency (n)	Percentage (%)
Marital Status		
- Divorced	84	49
- Married Couples	86	51
Gender		
- Male	84	49
- Female	86	51
Age (years)		
- Range	20-63	
- Mean (X̄)	36.04	
- Standard Deviation (SD)	9.34	
Region		
- Black Sea	79	
Type of Marriage		
- Married for Love	74	43.5
- Arranged Marriage	96	56.5

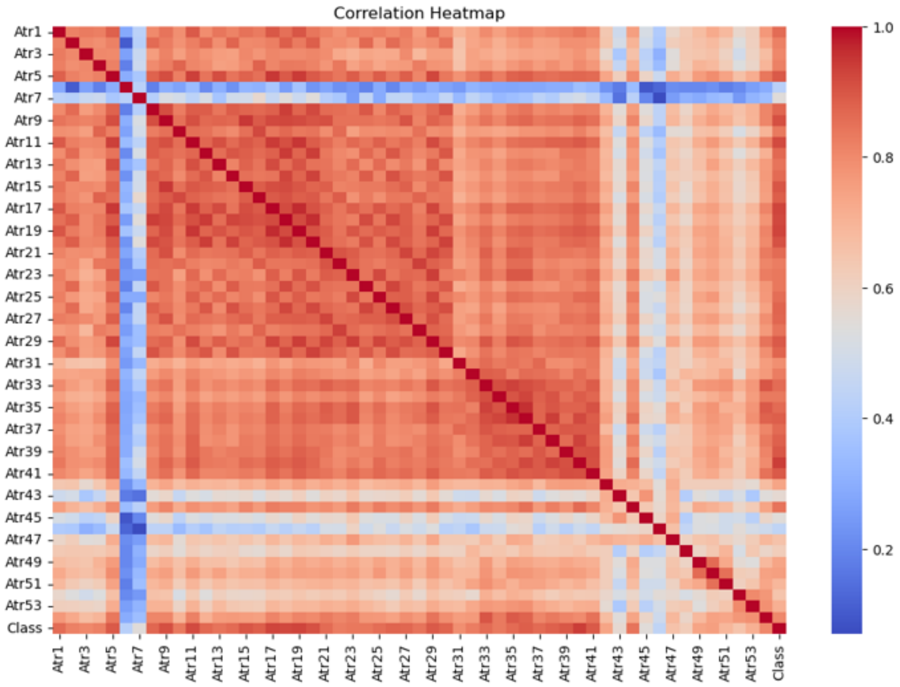


Children		
- With Children	127	74.7
- Without Children	43	25.3
Education Level		
- Primary School Graduate	18	10.58
- Secondary School Graduate	15	8.8
- High School Graduate	33	19.41
- College Graduate	88	51.76
- Master's Degree	15	8.8
Monthly Income (TL)		
- Under 2000	34	20
- Between 2001-3000	54	31.76
- Between 3001-4000	28	16.47
- Over 4000	54	31.76

2 Data Analysis

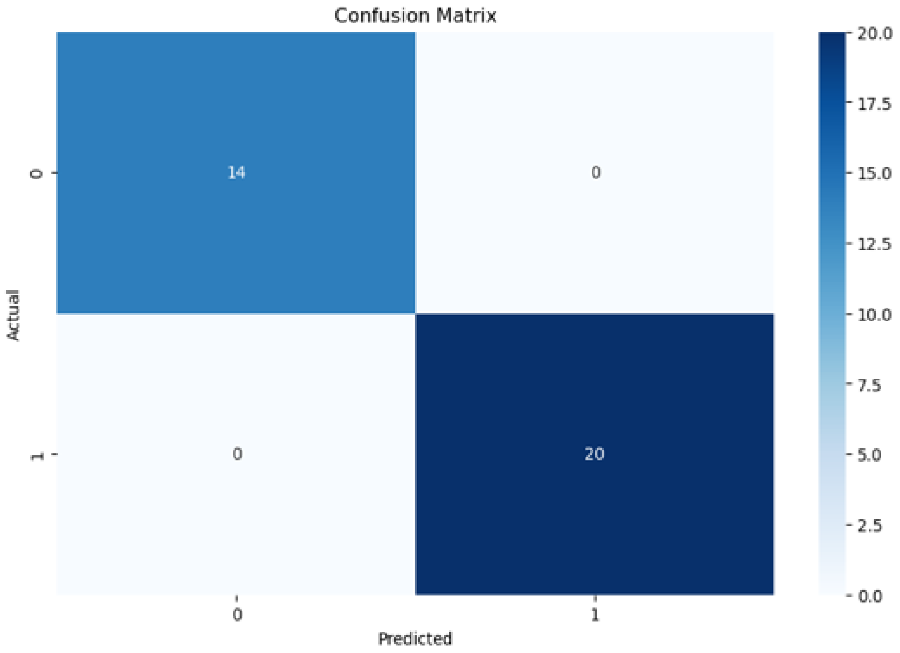
2.1 Correlation Analysis

In order to study the correlation between class(whether divorce or not) and the 54 problems, we drew a heat map as shown below. We can discovered that there exists high correlation between class and many questions.

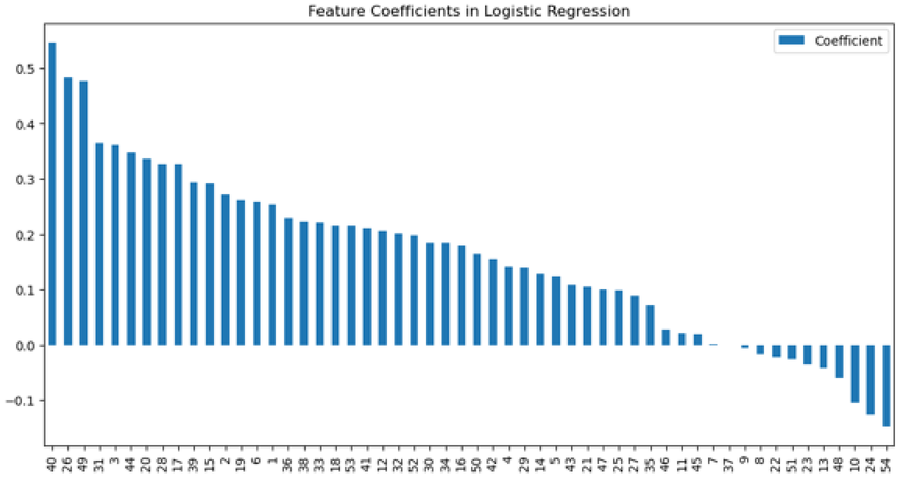


2.2 Logistic Regression

We divide the data set into test sets and training sets, and use logistic regression to classify, and use confusion matrices to evaluate the model.



As can be seen from the confusion matrix, the classification results are completely correct. Then, in order to study how important or influential each feature is to predict the target variable, we obtain the coefficients of the features from the logistic regression model and visualize them as a bar chart in descending order.

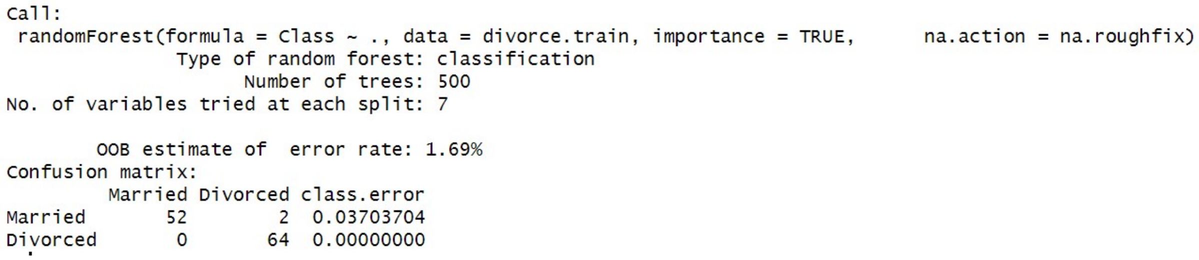


Features with negative coefficients are the following: ”I enjoy traveling with my wife”, ”I enjoy our holidays with my wife”, ”I know how my wife wants to be taken care of when she’s sick”, ”I’m not the one who’s wrong about problems at home”, ”I know my wife’s favorite food”, ”My husband and I have similar entertainment”, ”I feel right in our discussions”, ”My wife and most of our goals are common”, ”I can tell you what kind of stress my wife is facing in her life” and ”I’m not afraid to tell her about my wife’s incompetence”.

This suggests that those couples who share the same interests, values, and goals are more likely to have a harmonious relationship, and this last Art 54 shows that honesty is also very important in marriage.

2.3 Random Forest

We also use a random forest which is an ensemble learning approach to supervised learning. Multiple predictive models are developed, and the results are aggregated to improve classification rates. The algorithm for a random forest involves sampling cases and variables to create a large number of decision trees. Each case is classified by each decision tree. The most common classification for that case is then used as the outcome. However, we must prepare the data first. We randomly divided into a training sample (70%) and a validation sample (30%).



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

        OOB estimate of  error rate: 1.69%
Confusion matrix:
         Married Divorced class.error
Married       52        2  0.03703704
Divorced       0       64  0.00000000
```

Attribute	MeanDecreaseGini	Attribute	MeanDecreaseGini	Attribute	MeanDecreaseGini
Atr1	0.127040190	Atr19	3.729249011	Atr37	0.266028965
Atr2	0.146133302	Atr20	2.671412044	Atr38	0.564290193
Atr3	0.050394239	Atr21	0.932763217	Atr39	0.945559518
Atr4	0.874325869	Atr22	0.298579167	Atr40	4.160531614
Atr5	0.982997654	Atr23	0.011518418	Atr41	0.359349767
Atr6	0.070689322	Atr24	0.134139524	Atr42	0.046470896
Atr7	0.030287352	Atr25	2.264458709	Atr43	0.008229220
Atr8	1.472582662	Atr26	2.281173853	Atr44	0.331835600
Atr9	4.825427521	Atr27	1.235028739	Atr45	0.012666667
Atr10	0.084210849	Atr28	0.265553310	Atr46	0.011000000
Atr11	5.422090713	Atr29	1.358955467	Atr47	0.033139387
Atr12	1.765801822	Atr30	1.814237194	Atr48	0.036708131
Atr13	0.154723816	Atr31	0.118786895	Atr49	0.043704936
Atr14	1.633689605	Atr32	0.154061449	Atr50	0.025586592
Atr15	1.590513636	Atr33	0.021719791	Atr51	0.009327273
Atr16	3.437695757	Atr34	0.017442495	Atr52	0.030143747
Atr17	4.518763895	Atr35	0.027301341	Atr53	0.056813279
Atr18	5.820740479	Atr36	0.708404338	Atr54	0.077652775

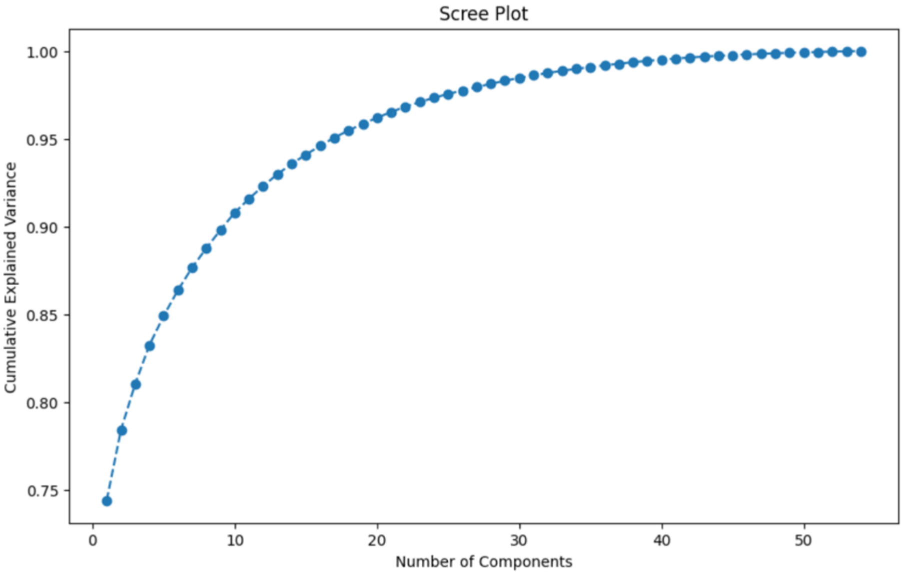
表 1: Mean Decrease Gini for Attributes

First, we grow 500 traditional decision trees by sampling 118 observations with replacement from the training sample and sampling 7 variables at each node of each tree. Random forests can provide a natural measure of variable importance. The relative importance measure is the total decrease in node impurities (heterogeneity) from splitting on that variable, averaged over all trees. Node impurity is measured with the Gini coefficient. Atr18 is the most important variable and

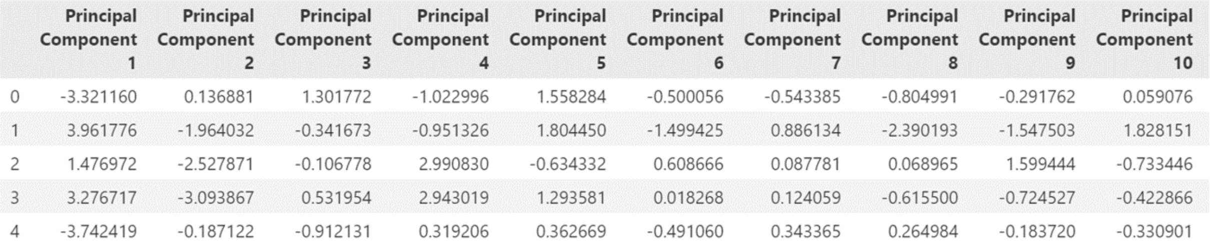
Atr43 is the least important. the validation sample is classified using the random forest and the predictive accuracy is calculated. The confusion matrix shows an accuracy of 98.31% which is good.

2.4 Principal Component Analysis

In this section, PCA was used to reduce the dimensionality of the dataset while preserving its key characteristics. By computing the eigenvectors and eigenvalues of the dataset's covariance matrix, PCA transformed the original variables into a smaller set of principal components that captured the most significant sources of variation. The analysis revealed that the first ten principal components explained over 90% of the total variance, indicating their effectiveness in summarizing the dataset. This reduction in dimensionality not only simplified subsequent analyses but also provided deeper insights into the underlying structure of the data.



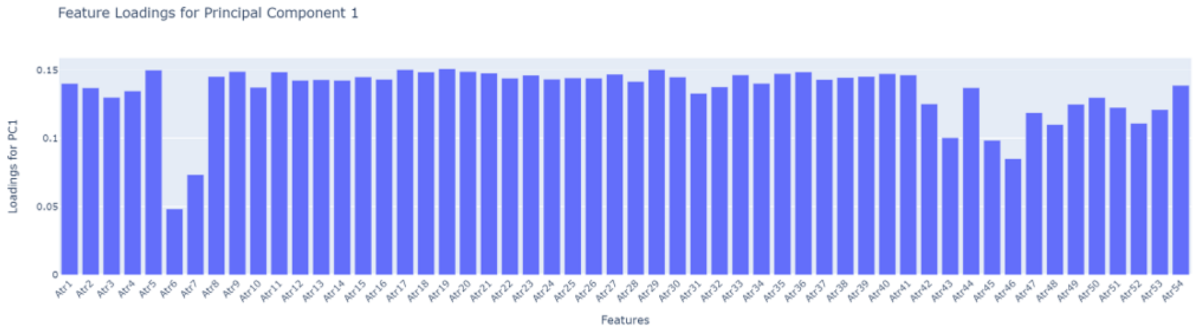
The first ten principal components after dimensionality reduction to five are shown in the table below：



	Principal Component 1	Principal Component 2	Principal Component 3	Principal Component 4	Principal Component 5	Principal Component 6	Principal Component 7	Principal Component 8	Principal Component 9	Principal Component 10
0	-3.321160	0.136881	1.301772	-1.022996	1.558284	-0.500056	-0.543385	-0.804991	-0.291762	0.059076
1	3.961776	-1.964032	-0.341673	-0.951326	1.804450	-1.499425	0.886134	-2.390193	-1.547503	1.828151
2	1.476972	-2.527871	-0.106778	2.990830	-0.634332	0.608666	0.087781	0.068965	1.599444	-0.733446
3	3.276717	-3.093867	0.531954	2.943019	1.293581	0.018268	0.124059	-0.615500	-0.724527	-0.422866
4	-3.742419	-0.187122	-0.912131	0.319206	0.362669	-0.491060	0.343365	0.264984	-0.183720	-0.330901

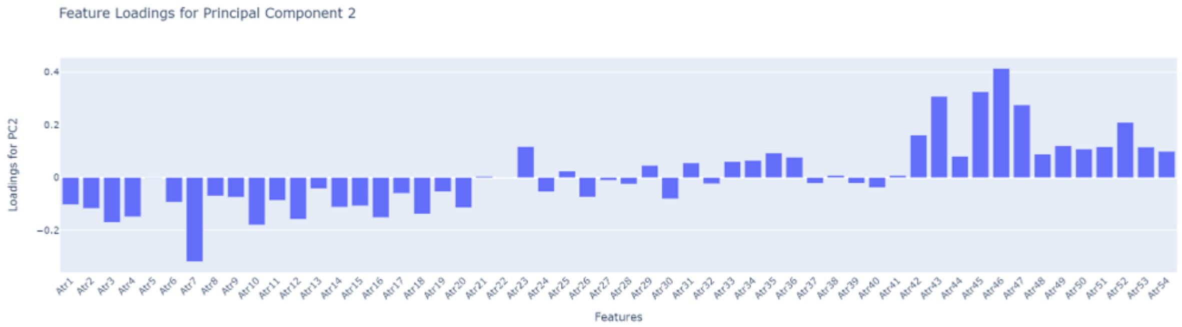
We visualize the feature loadings of the first three principal components to understand the contributions of original variables to these components. The visualization effectively illustrates how each variable influences the principal components, with higher absolute values indicating stronger associations.

- Principal Component 1:



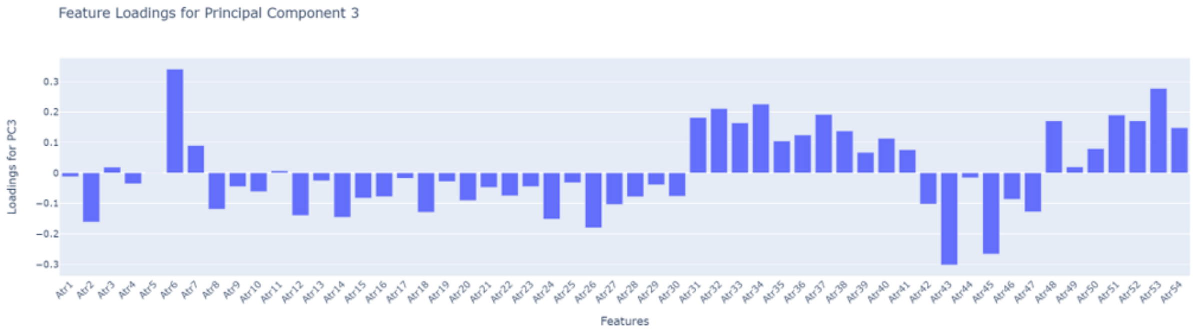
The first principal component shows positive loadings across all features. All features positively contribute to the first principal component without distinctive negative contributions. This suggests a uniform pattern in the influence of features on the first principal component, which may imply a lack of contrasting relationships among the variables.

- Principal Component 2:



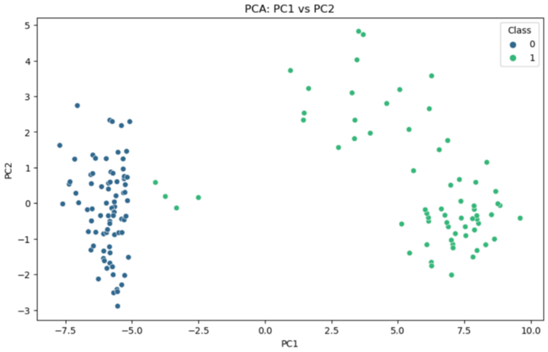
The second principal component may reflect ”Contempt” factor, suggesting that couples are more likely to divorce when one partner ignores the other.

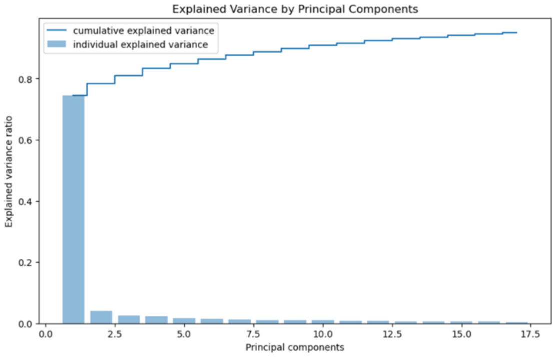
- Principal Component 3:



The third principal component may reflect ”Defensiveness” factor, suggesting that couples are more likely to divorce when one partner tend to use bad words and behavior.

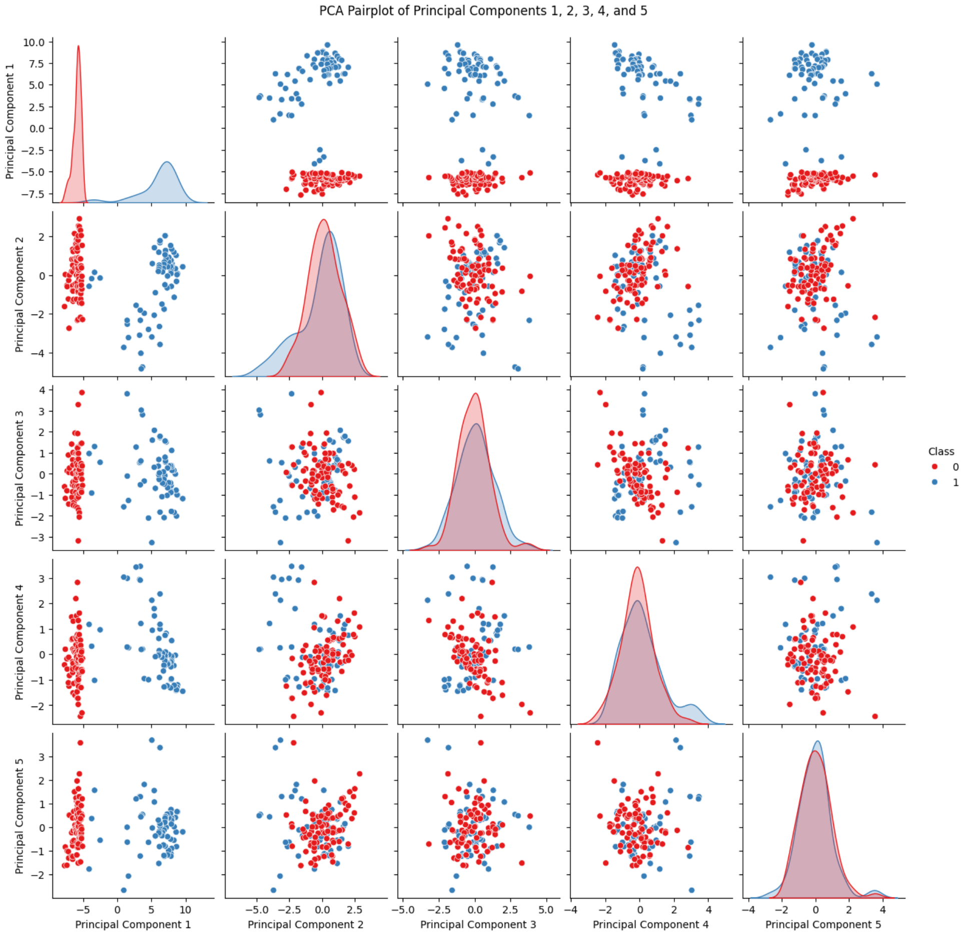
Then we visualize the first two principal components:





From the figure above, we can see that the samples of different classes have been well separated except for a few points, indicating that the first two principal components may capture the most significant directions of change in the data, which may be related to the differentiation between classes. What's more, the cumulative variance contribution rate of the first two principal components has reached about 80%.

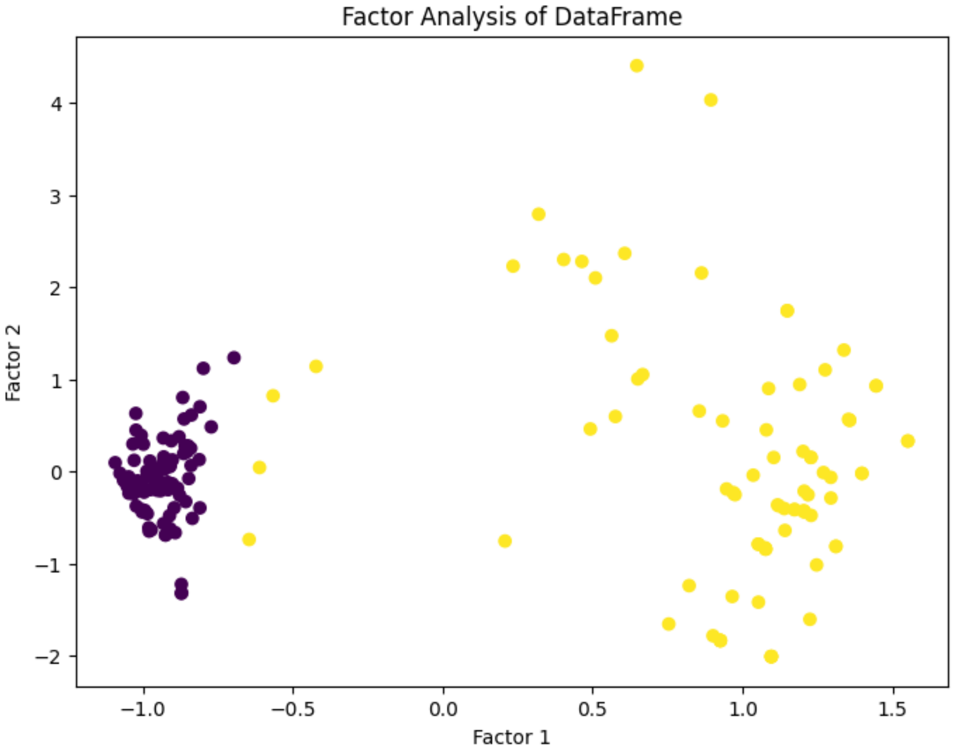
Then we visualize the relationship between the first five principal components:



From the figure above, we can see that the first principal component has the greatest impact on the classification results, while the second, third, fourth and fifth principal components are of similar importance and not much, because the first principal component explained the most variance.

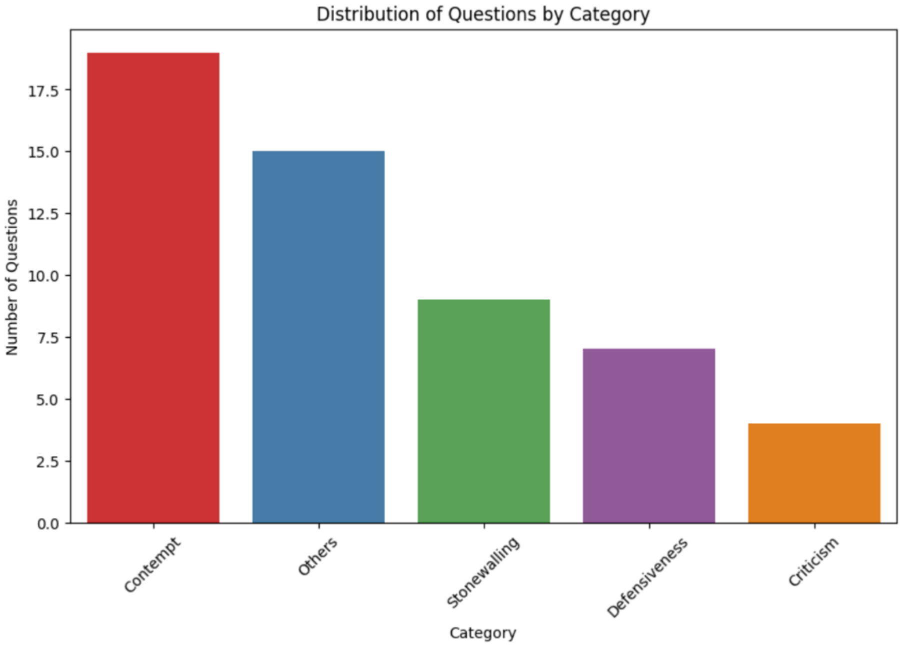
2.5 Factor Analysis

In this section, we apply factor analysis on the dataset, and visualized the first two factors, and found that the results of factor analysis were very similar to PCA.

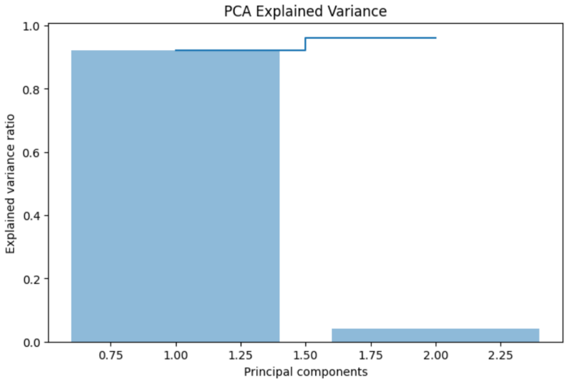


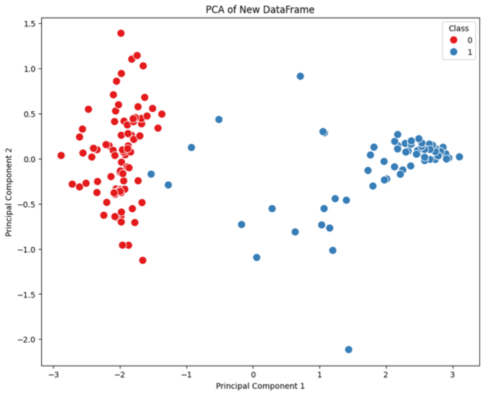
2.6 Feature Extraction

Based on Gottman's ”four horsemen” theory, we manually divided the 54 questions into five categories(Criticism, Contempt, Stonewalling, Defensiveness and others).



Then we apply PCA to dataset again. The variance contribution rate and classification results of the first two principal components are shown in the figure below.

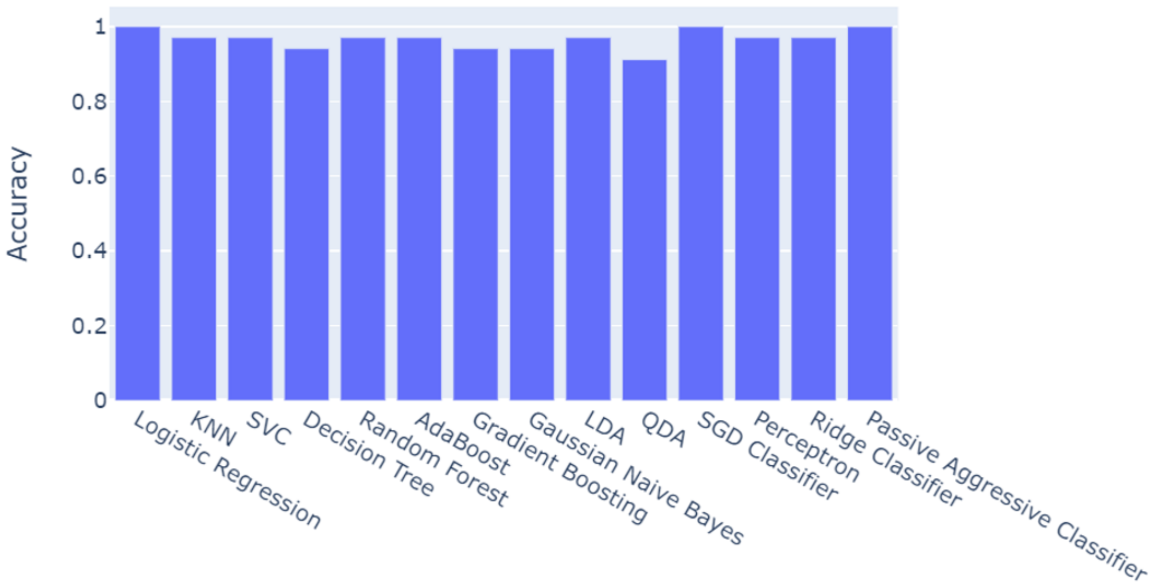




The variance contribution rate and classification results of the first two principal components are shown in the figure below. We can find that the variance contribution rate of the first two principal components has reached more than 90%, and the classification effect is also good.

3 Conclusion

We also use different models to do the classification, all of them work well.



The DPS dataset provides valuable insights into the predictive power of the scale in identifying marital discord and potential divorce risk factors. The results of PCA and factor analysis have verified Goldman's Four Horsemen theory to a certain extent, so in order to have a perfect and stable relationship, we should try to avoid the ”four horsemen” and potential risks in our future marriage.