Università degli Studi di Milano

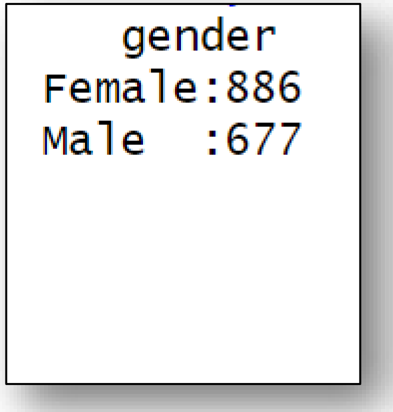
Statistical Learning

Danial Forouzanfar - 12988A

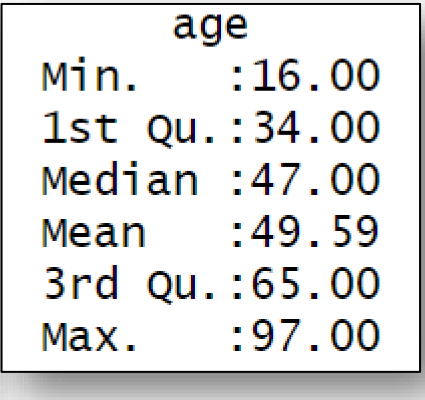
Supervised:

- Decision Tree,
- Random Forest,
- eXtreme Gradient Boosting

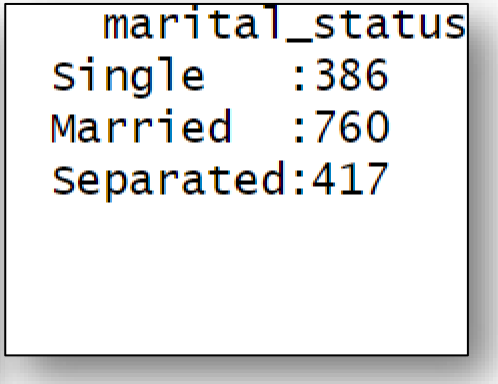
Data on smoking habits



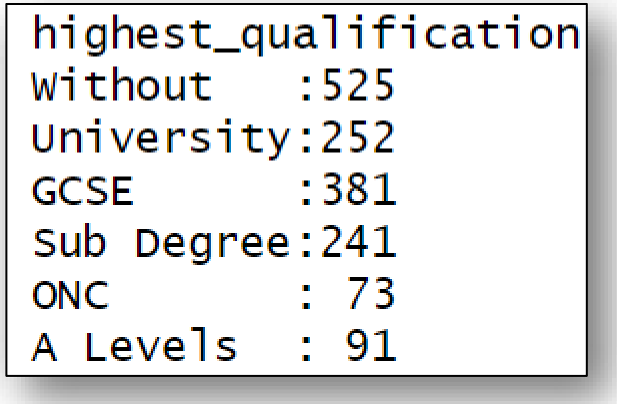
```
    gender
 Female:886
 Male  :677
```



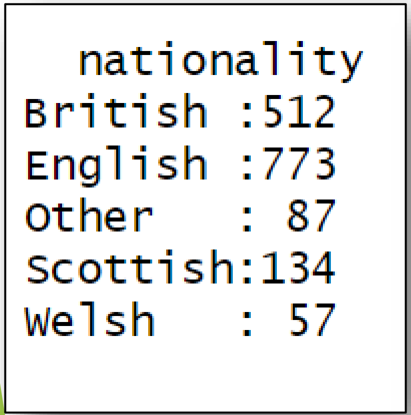
```
      age
 Min.   :16.00
 1st Qu.:34.00
 Median :47.00
 Mean   :49.59
 3rd Qu.:65.00
 Max.   :97.00
```



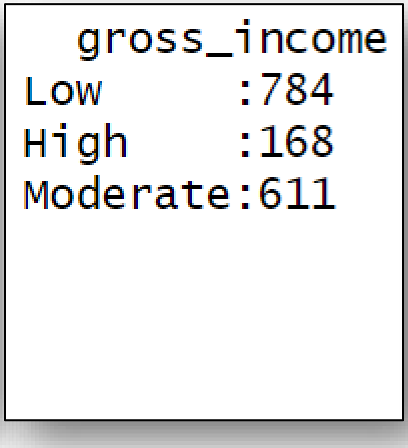
```
   marital_status
 Single    :386
 Married   :760
 Separated:417
```



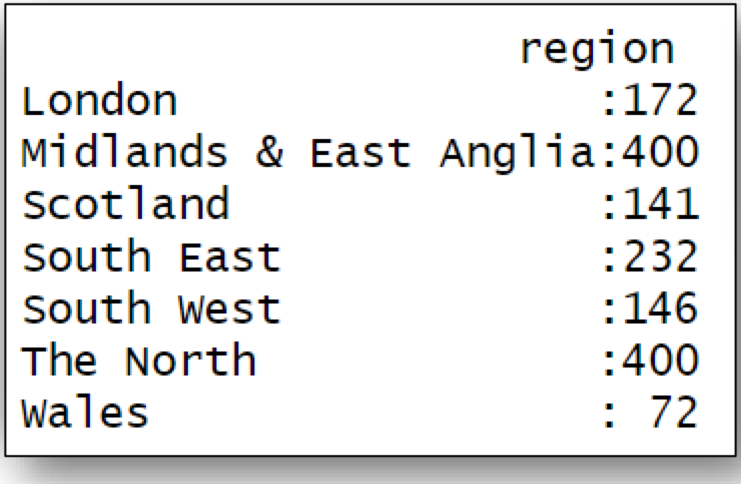
```
highest_qualification
Without     :525
University:252
GCSE        :381
Sub Degree:241
ONC         : 73
A Levels    : 91
```



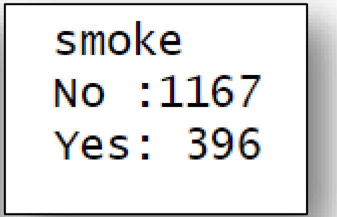
```
   nationality
 British :512
 English :773
 Other   : 87
 Scottish:134
 Welsh   : 57
```



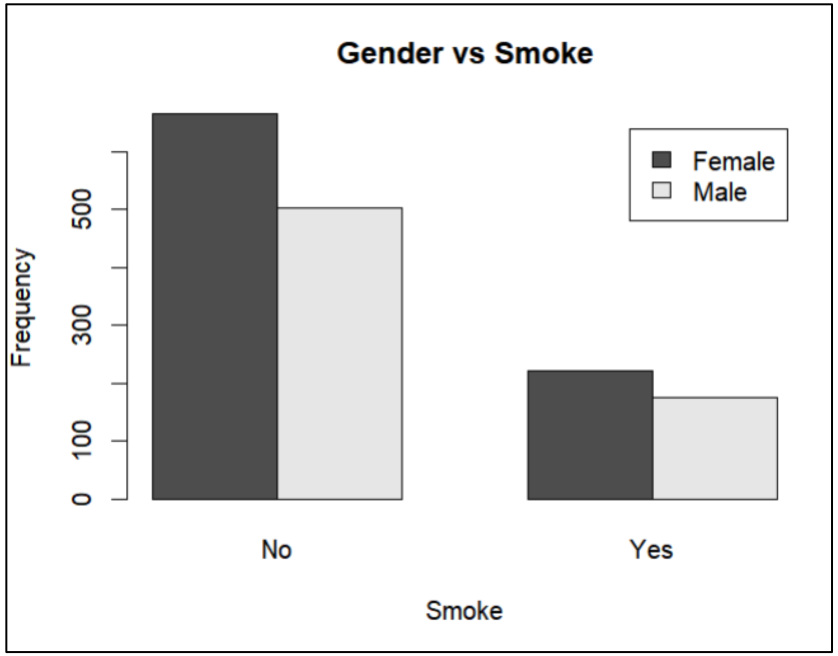
```
    gross_income
 Low      :784
 High     :168
 Moderate:611
```

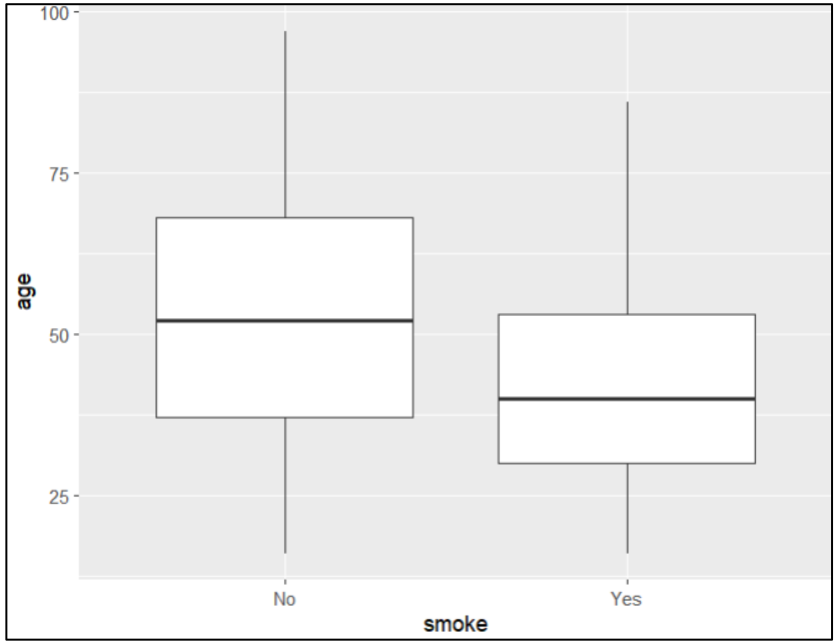


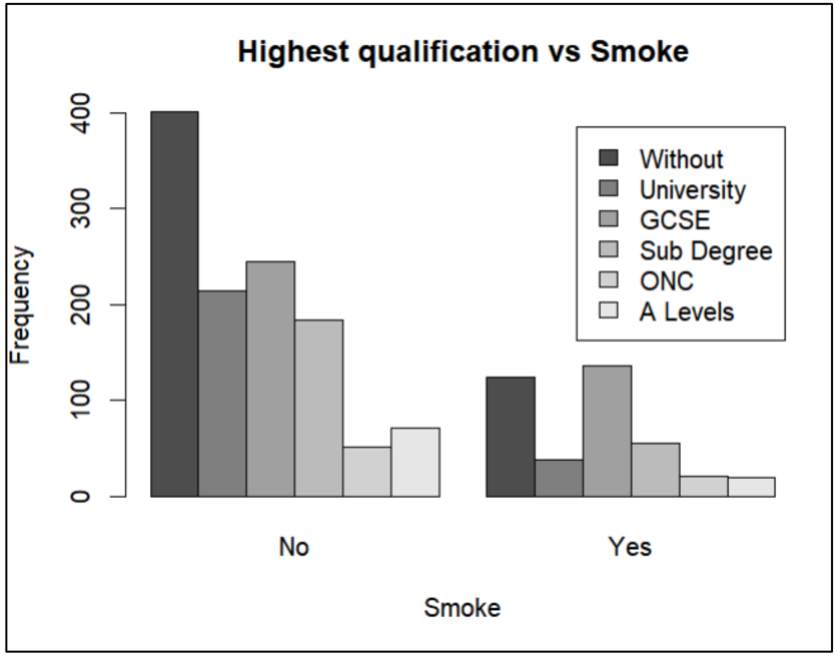
```
                          region
London                   :172
Midlands & East Anglia:400
Scotland                 :141
South East               :232
South West               :146
The North                :400
Wales                    : 72
```

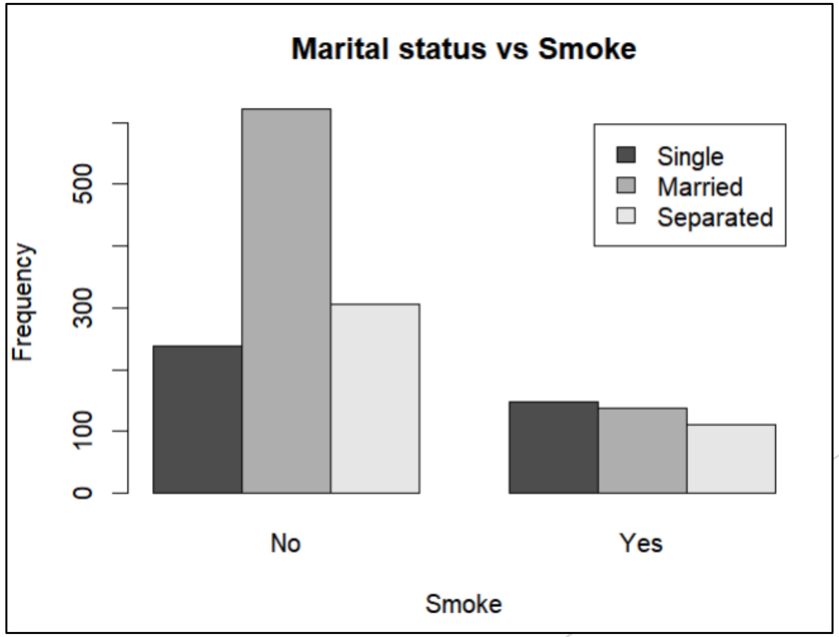


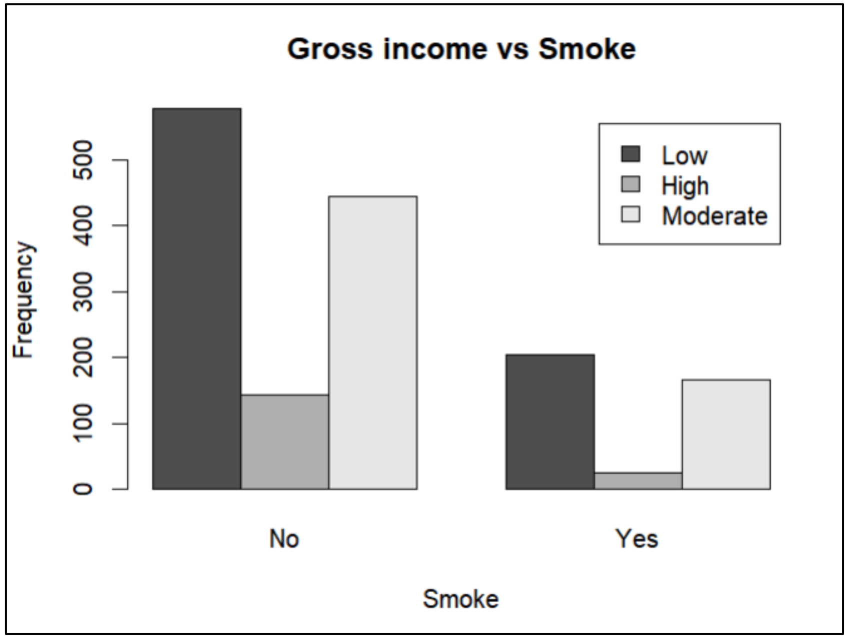
```
  smoke
 No  :1167
 Yes:  396
```

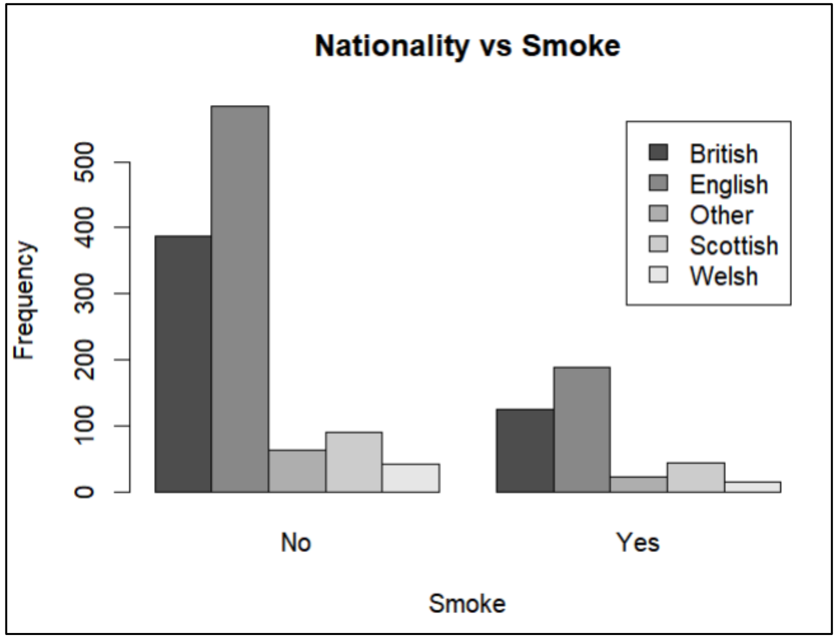

Gender vs Smoke
Female
Male
Frequency
0
100
300
500
No
Yes
Smoke

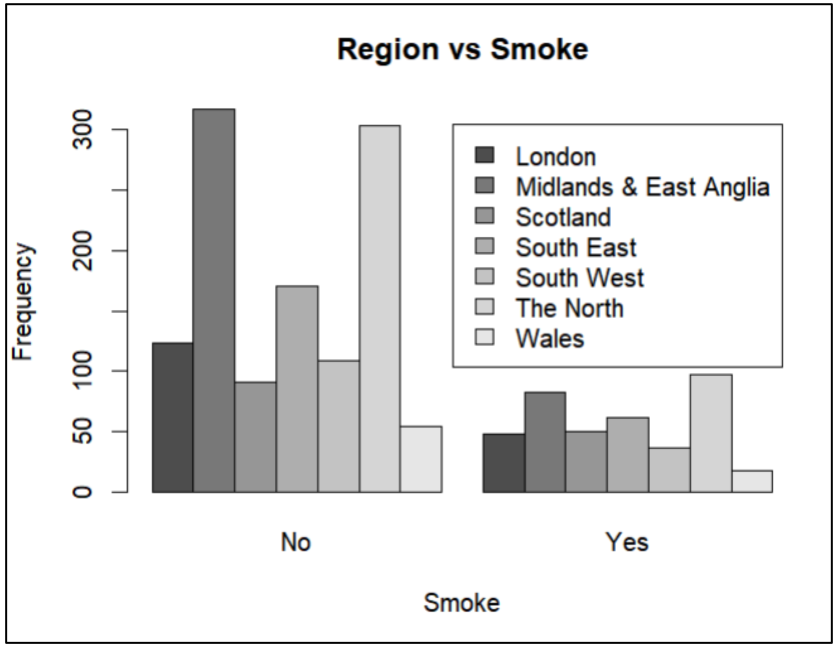

100
75
50
25
age
No
Yes
smoke


Highest qualification vs Smoke
Without
University
GCSE
Sub Degree
ONC
A Levels
Frequency
0
100
200
300
400
No
Yes
Smoke


Marital status vs Smoke
Single
Married
Separated
Frequency
0
100
300
500
No
Yes
Smoke


Gross income vs Smoke


Nationality vs Smoke


Region vs Smoke

Libraries

General

- tidymodels
- tidyverse
- caret

Decision Tree

- visNetwork
- rpart

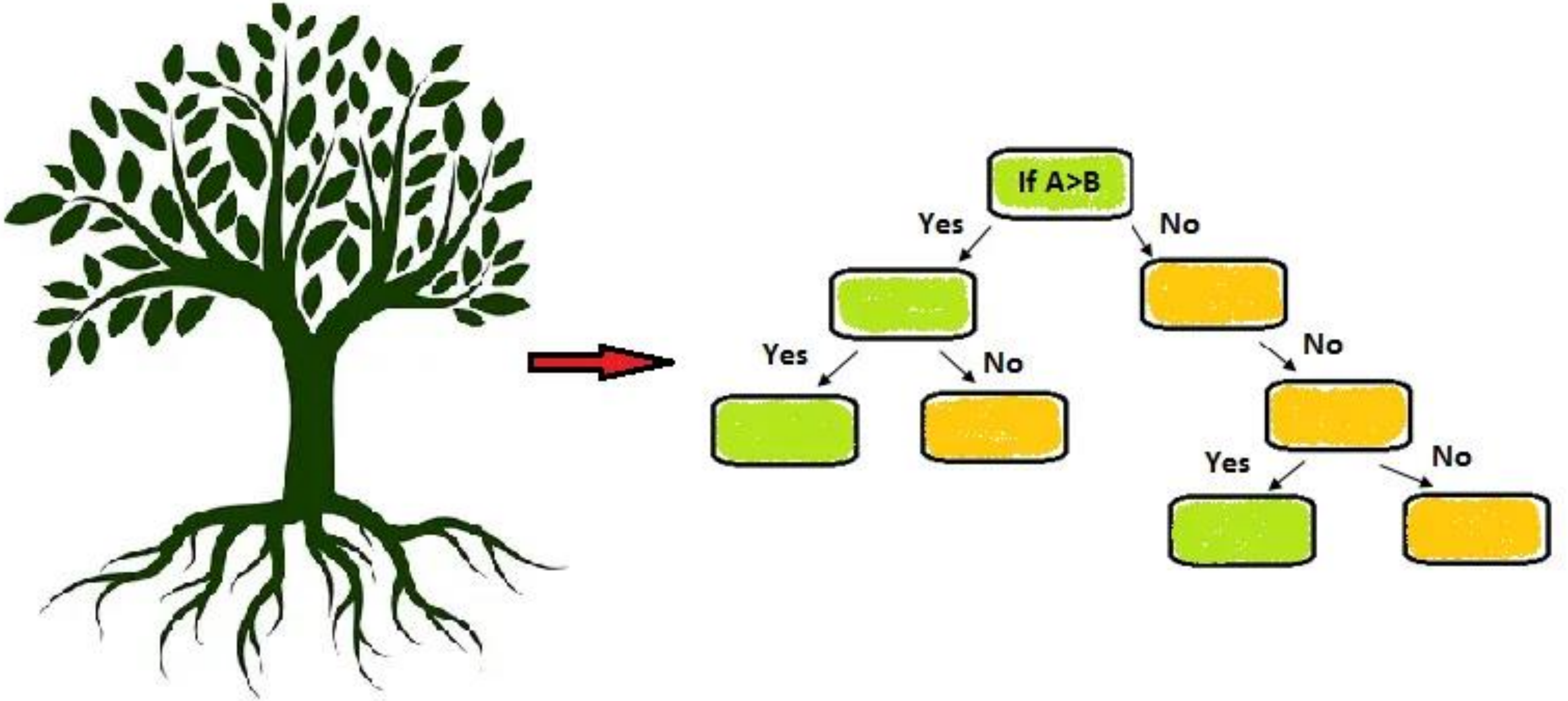
Random Forest

- randomForest
- e1071

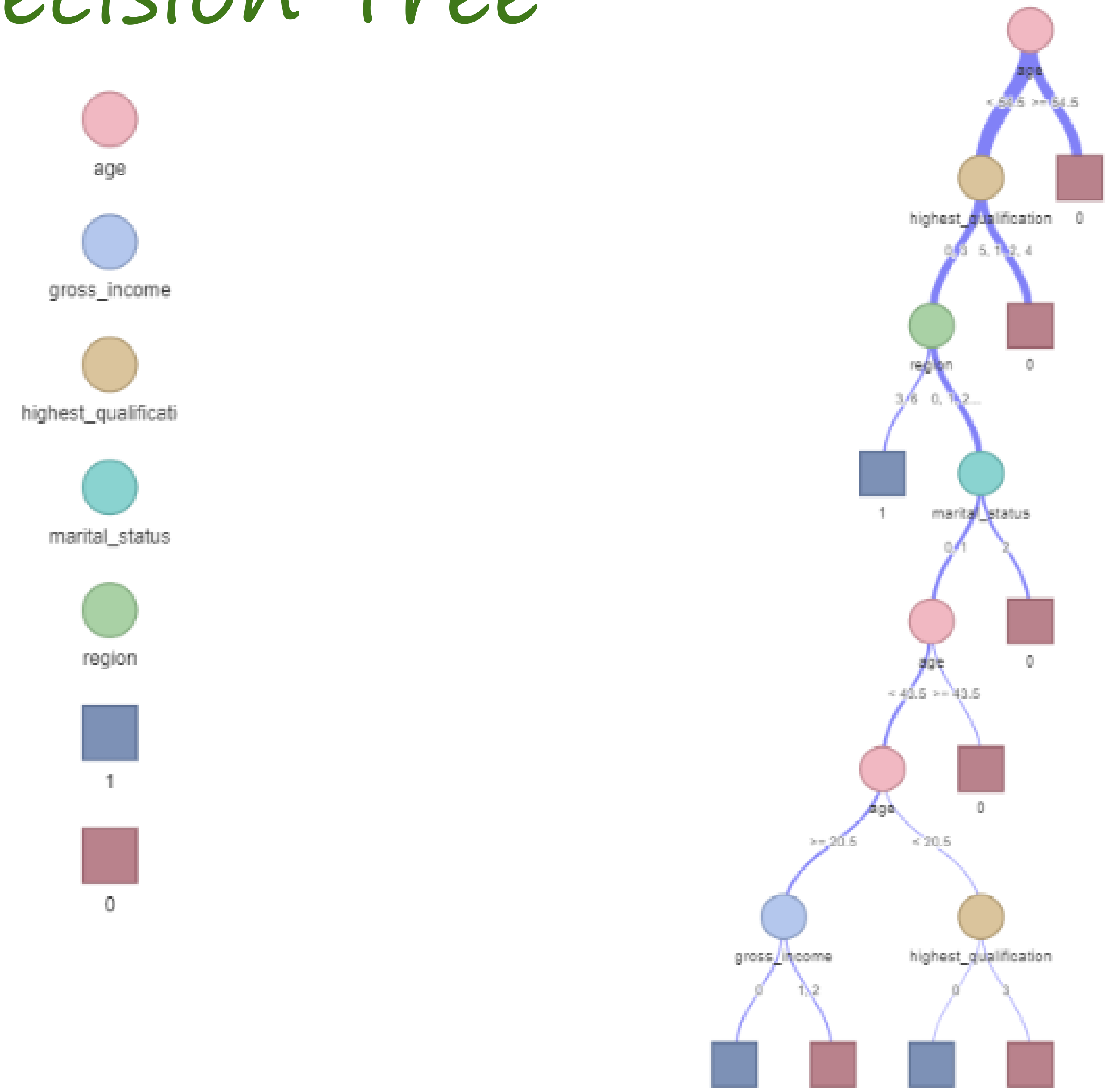
XG Boosting

- xgboost
- rattle

Decision Tree

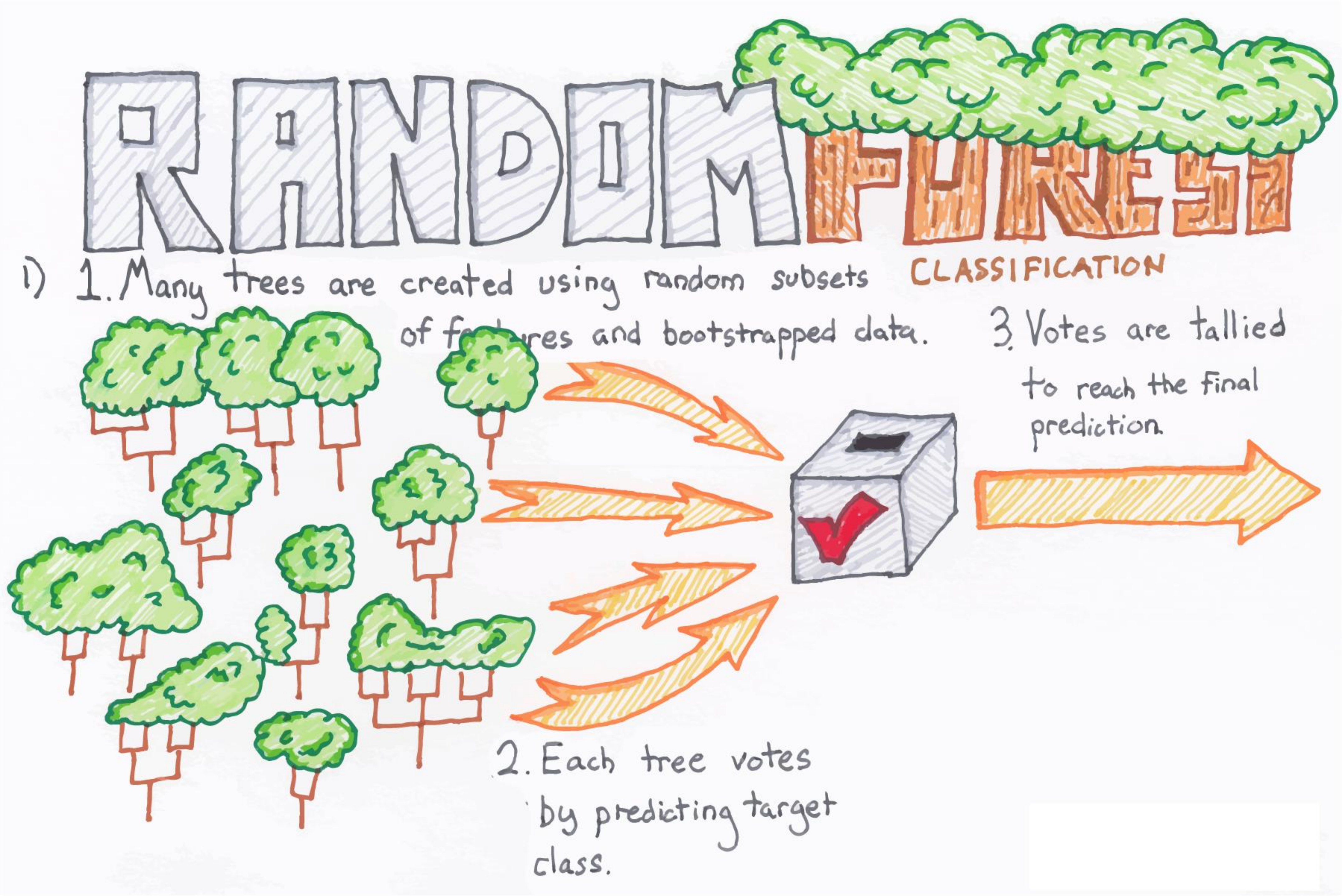


Decision Tree



Decision Tree

	Metric	Value
1	Accuracy	0.7468085
2	Precision	0.9259259
3	Recall	0.7775120
4	F1 score	0.8452536



1) 1. Many trees are created using random subsets of features and bootstrapped data.

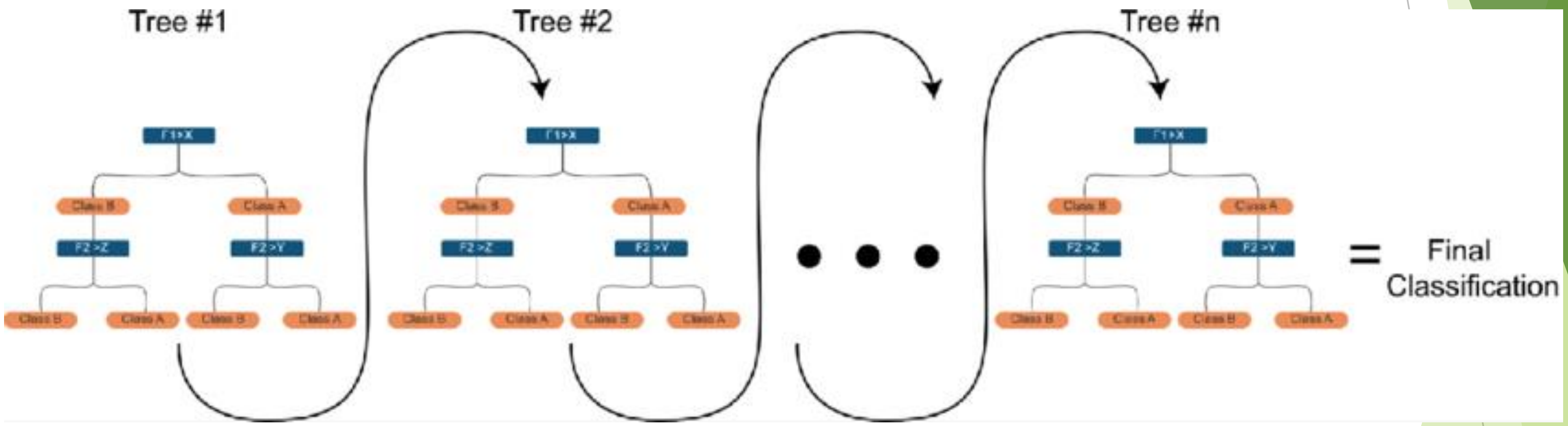
2. Each tree votes by predicting target class.

3. Votes are tallied to reach the final prediction.

Random Forest

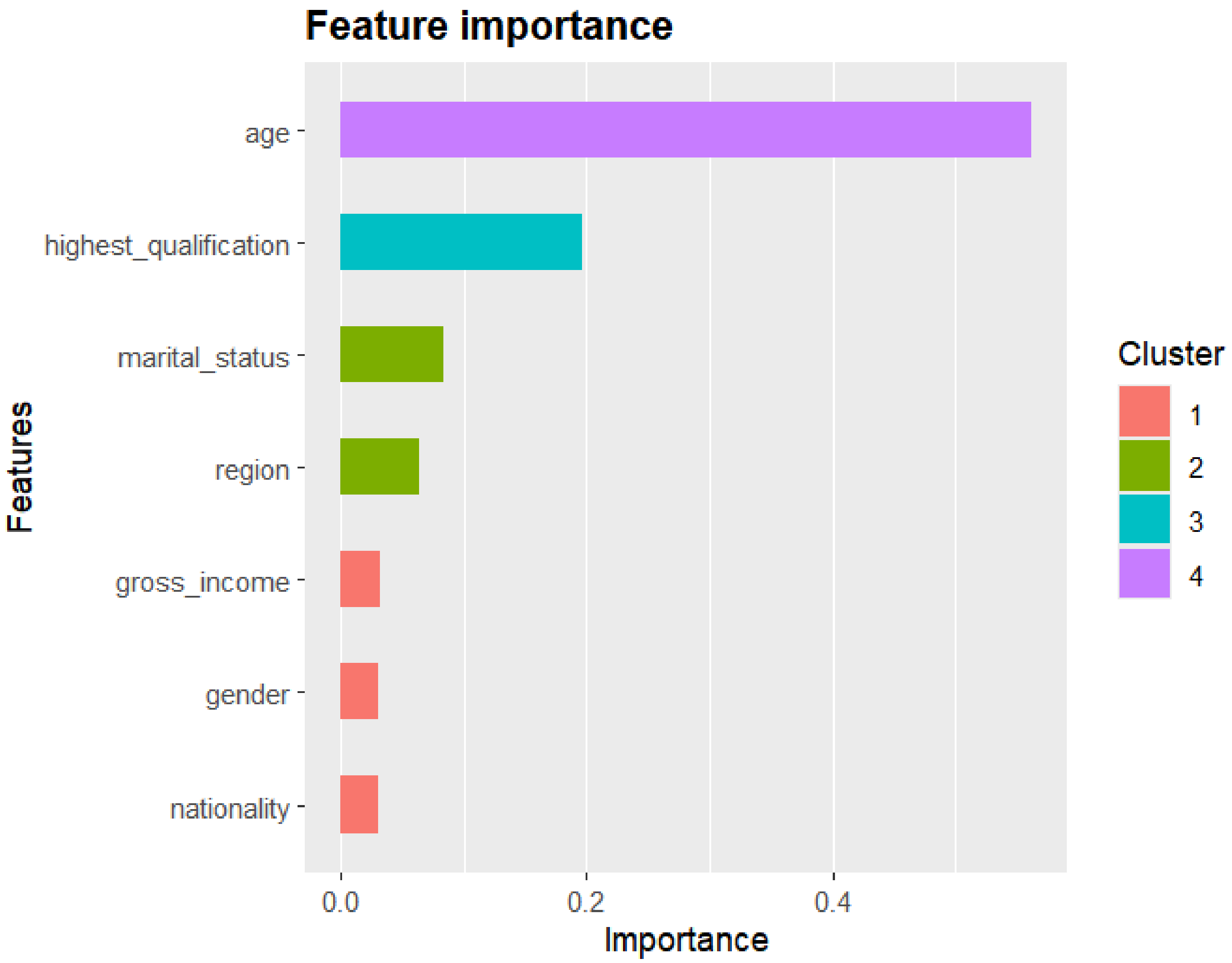
	Metric	Value
1	Accuracy	0.7510638
2	Precision	0.9572650
3	Recall	0.7671233
4	F1 score	0.8817110

eXtreme Gradient Boosting



XG Boosting

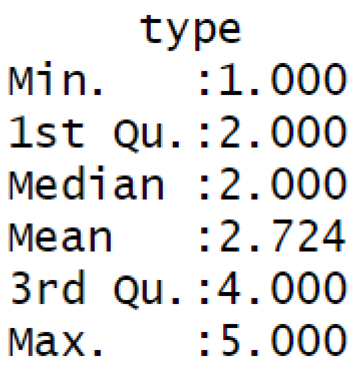
	Metric	Value
1	Accuracy	0.7595745
2	Precision	0.9743590
3	Recall	0.7668161
4	F1 score	0.8582183


Feature importance
Features
age
highest_qualification
marital_status
region
gross_income
gender
nationality
0.0
0.2
0.4
Importance
Cluster
1
2
3
4

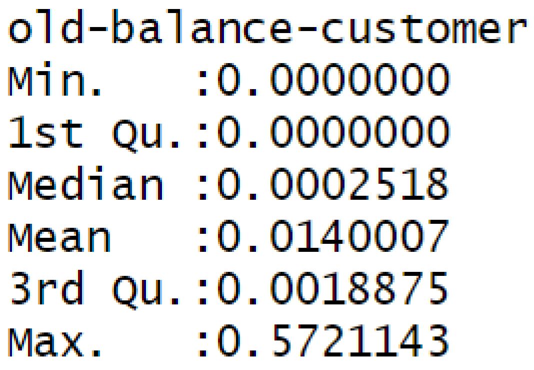
Unsupervised:

- PCA

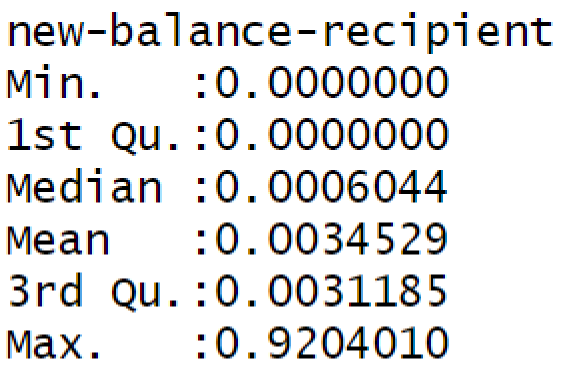
Fraud Detection



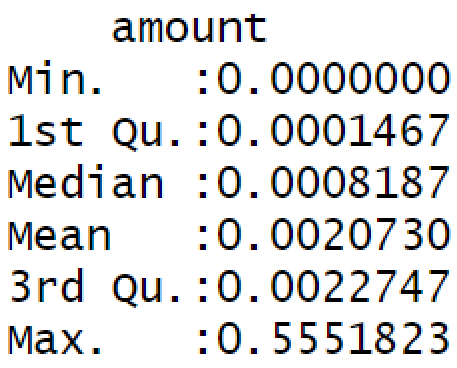
```
      type      
 Min.   :1.000  
 1st Qu.:2.000  
 Median :2.000  
 Mean   :2.724  
 3rd Qu.:4.000  
 Max.   :5.000  
```



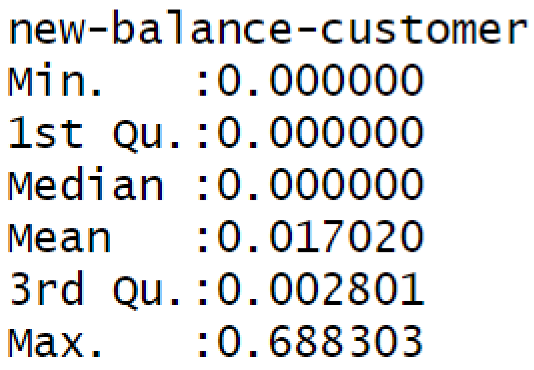
```
old-balance-customer
Min.   :0.0000000   
1st Qu.:0.0000000   
Median :0.0002518   
Mean   :0.0140007   
3rd Qu.:0.0018875   
Max.   :0.5721143   
```



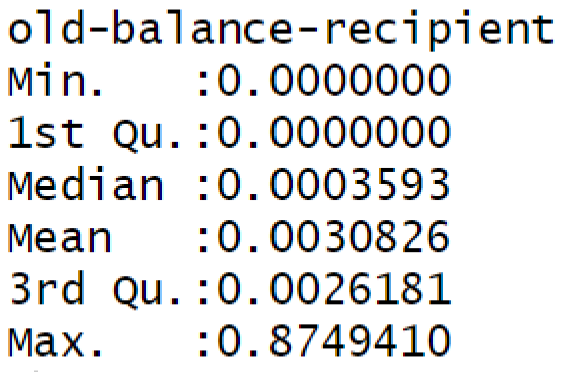
```
new-balance-recipient
Min.   :0.0000000    
1st Qu.:0.0000000    
Median :0.0006044    
Mean   :0.0034529    
3rd Qu.:0.0031185    
Max.   :0.9204010    
```



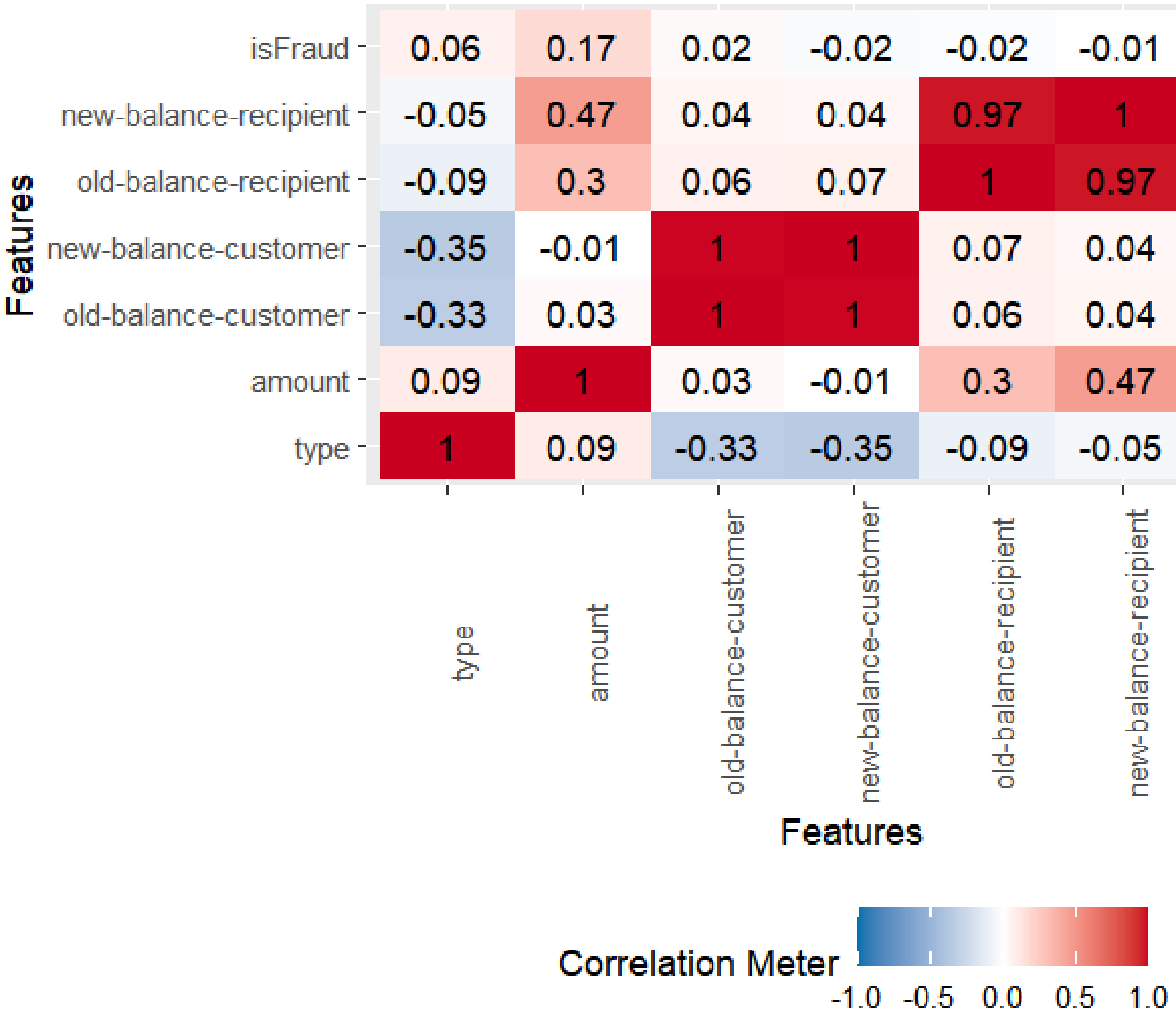
```
     amount         
 Min.   :0.0000000  
 1st Qu.:0.0001467  
 Median :0.0008187  
 Mean   :0.0020730  
 3rd Qu.:0.0022747  
 Max.   :0.5551823  
```



```
new-balance-customer
Min.   :0.000000    
1st Qu.:0.000000    
Median :0.000000    
Mean   :0.017020    
3rd Qu.:0.002801    
Max.   :0.688303    
```

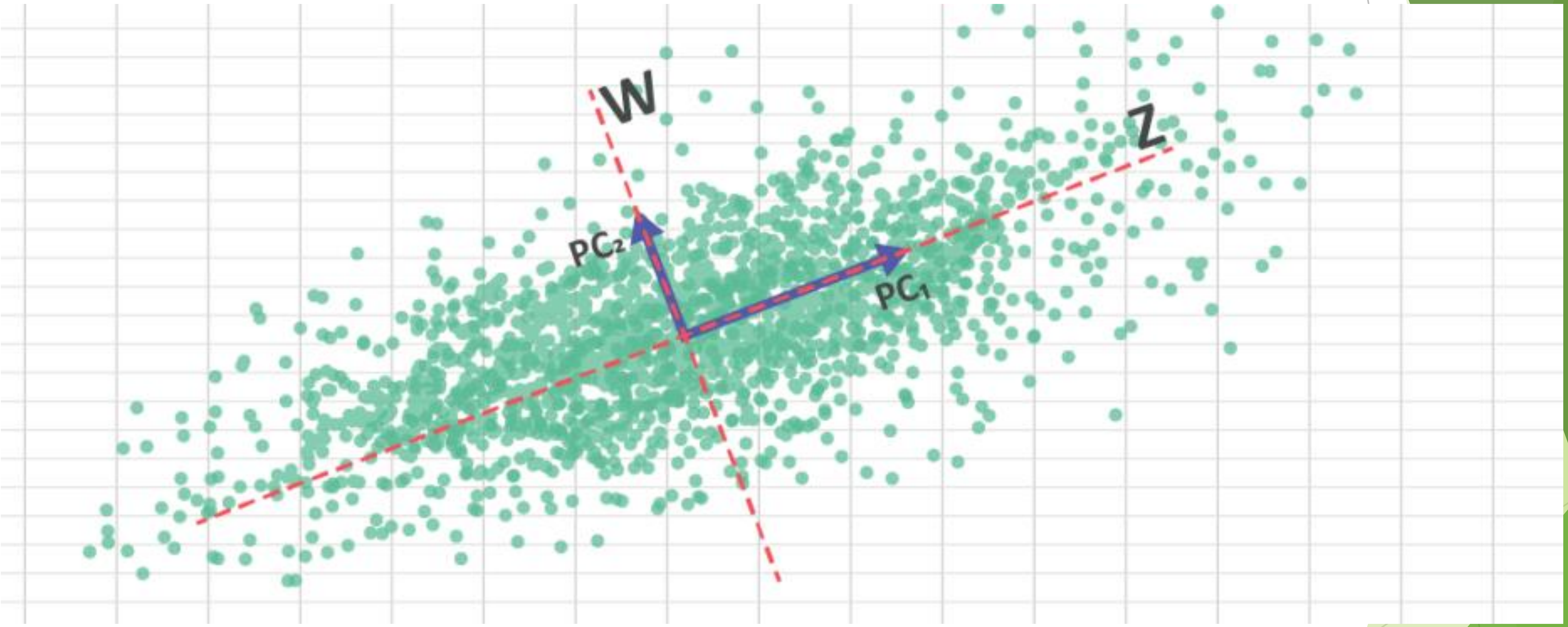


```
old-balance-recipient
Min.   :0.0000000    
1st Qu.:0.0000000    
Median :0.0003593    
Mean   :0.0030826    
3rd Qu.:0.0026181    
Max.   :0.8749410    
```



	type	amount	old-balance-customer	new-balance-customer	old-balance-recipient	new-balance-recipient	isFraud
isFraud	0.06	0.17	0.02	-0.02	-0.02	-0.01	1
new-balance-recipient	-0.05	0.47	0.04	0.04	0.97	1	-0.01
old-balance-recipient	-0.09	0.3	0.06	0.07	1	0.97	-0.02
new-balance-customer	-0.35	-0.01	1	1	0.07	0.04	-0.02
old-balance-customer	-0.33	0.03	1	1	0.06	0.04	0.02
amount	0.09	1	0.03	-0.01	0.3	0.47	0.17
type	1	0.09	-0.33	-0.35	-0.09	-0.05	0.06

PCA



PCA

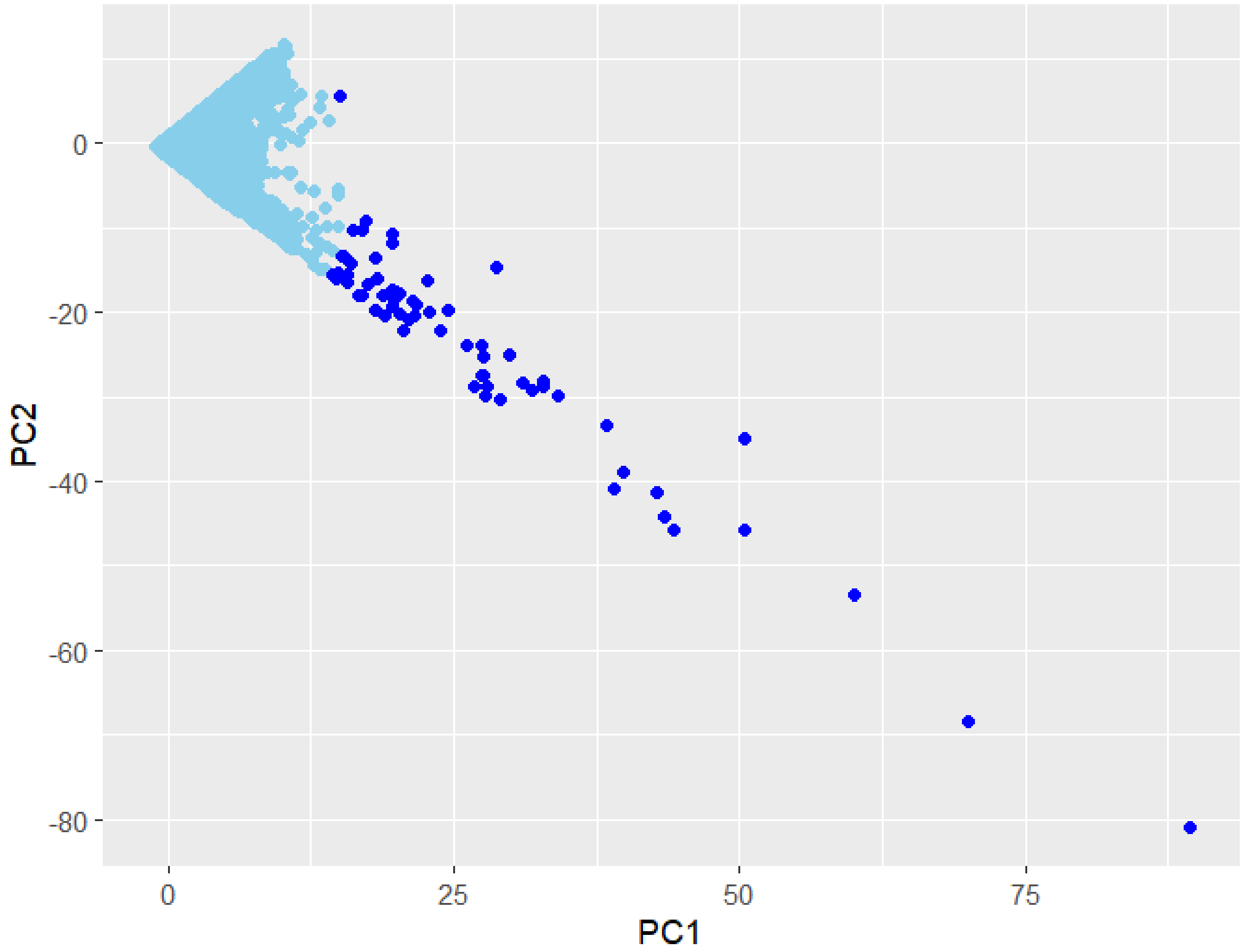
PCA	
Method	Randomized
Transform	Standardize
K	2
Threshold Level	15

PCA

Eigen Values

Var	PC1	PC2
Type	-0.261	-0.274
Amount	0.270	-0.315
Old-Balance-Customer	0.440	0.480
New-Balance-Customer	0.438	0.487
Old-Balance-Recipient	0.484	-0.402
New-Balance-Recipient	0.489	-0.444

PCA


PC2
PC1