Course: Data Mining
Student: Ekaterina Eremina
Topic: HW2

Part_1: written assignment

Exercise_1

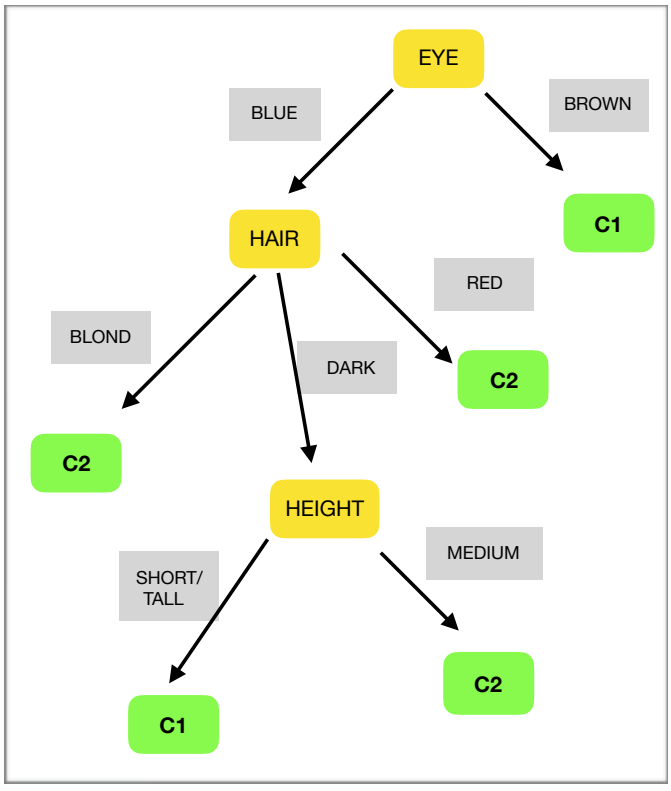
(a) p(C1) = 5, n(C2) = 4.
I(C1, C2) = 0,9911

	p_i	n_i	$I(p_i, n_i)$
Height = Tall	3	2	0,9710
Height = Short	2	1	0,9183
Height = Medium	0	1	0
Hair = Blond	2	2	1
Hair = Dark	3	1	0,8113
Hair =Red	0	1	0
Eye = Brown	3	0	0
Eye = Blue	2	4	0,9183

E(Height) = 5/9*I(3,2) + 3/9*I(2,1)+1/9*I(0,1) = 0,8455
E(Hair) = 4/9*I(2,2) + 4/9*I(3,1) + 1/9*I(0,1) = 0,8050
E (Eye) = 3/9*I(3,0) + 6/9*I(2,4) = 0,6122

Gain(Height) = 0,9911 - E(Height) = 0,1456
Gain(Hair) = 0,9911 - E(Hair) = 0,1861
Gain(Eye) = 0,9911 - E(Eye) = 0,3789

(b)



Exercise_2

Z = {Brown, Blond, Short} -> C1

Let's count probabilities for naive Bayesian classifier.

P(C1) = 5/9

P(C2) = 4/9

P(Brown|C1) = 3/5

P(Brown|C2) = 0

P(Blond|C1) = 2/5

P(Blond|C2) = 2/4

P(Short|C1) = 2/5

P(Short|C2) = 1/4

P(Z|C1) = 2/5*2/5*3/5 = 12/125

P(Z|C2) = 1/4*2/4*0 = 0

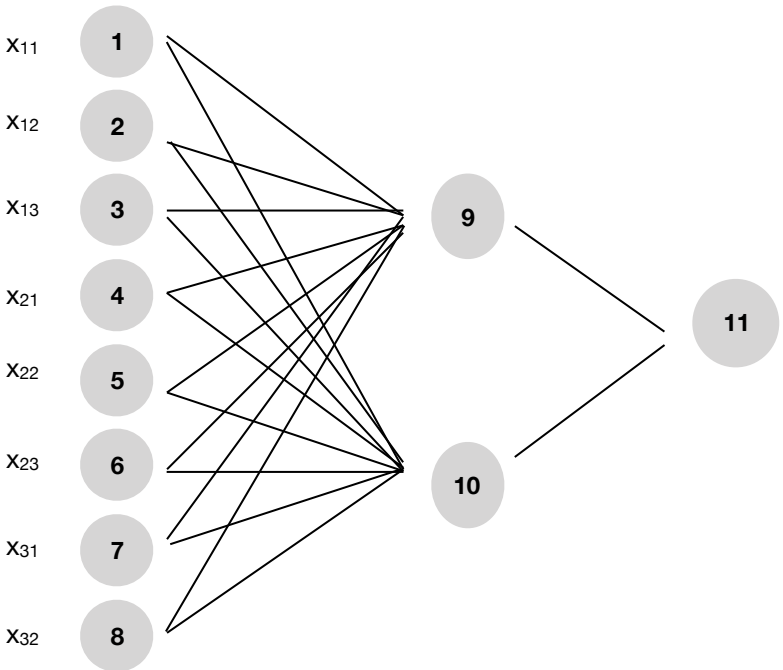
P(Z|C1)* P(C1) = 12/125*5/9 = 0,053

P(Z|C2)* P(C2) = 0

Result will be the same: = {Brown, Blond, Short} -> C1.

Exercise_3

(a)



(b) Training = {Medium, Blond, Blue -> C1}

x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}
tall	short	medium	blond	dark	red	brown	blue
0	0	1	1	0	0	0	1
$w_{1,9}$	$w_{1,10}$	$w_{2,9}$	$w_{2,10}$	$w_{3,9}$	$w_{3,10}$	$w_{4,9}$	$w_{4,10}$
0,2	0,2	-0,1	0,4	0,3	-0,1	-0,2	0,1
$w_{5,9}$	$w_{5,10}$	$w_{6,9}$	$w_{6,10}$	$w_{7,9}$	$w_{7,10}$	$w_{8,9}$	$w_{8,10}$
0,2	0,2	0,1	-0,1	0,3	0,2	0,1	0,2
$w_{9,11}$	$w_{10,11}$	θ_9	θ_{10}	θ_{11}			
-0,3	0,1	-0,2	0,2	0,3			

Node, j	Input, I_j	Output, O_j
9	0+0+0,3-0,2+0+0+0+0,1-0,2 = 0	1/(1+e^0) = 0,5
10	0+0-0,1+0,1+0+0+0+0,1+0,2 = 0,3	1/(1+e^(-0,3)) = 0,5744
11	(-0,3)*0,5+0,1*0,5744+0,1 = 0,00744	1/(1+e^(-0,00744)) = 0,502

Node, j	Error, Err_j
11	0,502*(1-0,502)*(1-0,502) = 0,1245
10	0,5744*(1-0,5744)*0,125*0,1 = 0,0031
9	0,5*(1-0,5)*0,125*(-0,3) = - 0,0094

$w_{1,10}$	0,2 + 0,9*(0,0031)*0 = 0,2	$w_{6,10}$	0,1 + 0,9*(0,0031)*0 = 0,1
$w_{1,9}$	0,2 + 0,9*(-0,0094)*0 = 0,2	$w_{7,9}$	0,3 + 0,9*(-0,0094)*0 = 0,3
$w_{2,10}$	-0,1 + 0,9*(0,0031)*0 = -0,1	$w_{7,10}$	0,3 + 0,9*(0,0031)*0 = 0,3
$w_{2,9}$	-0,1 + 0,9*(-0,0094)*0 = -0,1	$w_{8,9}$	0,1 + 0,9*(-0,0094)*1 = 0,092
$w_{3,10}$	0,3 + 0,9*(0,0031)*1 = 0,303	$w_{8,10}$	0,1 + 0,9*(0,0031)*1 = 0,103
$w_{3,9}$	0,3 + 0,9*(-0,0094)*1 = 0,292	$w_{9,11}$	-0,3+0,9*(0,1245)*0,5 = -0,244
$w_{4,9}$	-0,2 + 0,9*(-0,0094)*1 = -0,209	$w_{10,11}$	0,1+0,9*(0,1245)*0,5744 = 0,164
$w_{4,10}$	-0,2 + 0,9*(0,0031)*1 = -0,197	θ_9	-0,2+0,9*(-0,0094) = -0,209
$w_{5,9}$	0,2 + 0,9*(-0,0094)*0 = 0,2	θ_{10}	0,2+0,9*(0,0031) = 0,203
$w_{5,10}$	0,2 + 0,9*(0,0031)*0 = 0,2	θ_{11}	0,3+0,9*(0,1245) = 0,412
$w_{6,9}$	0,1 + 0,9*(-0,0094)*0 = 0,1		

Exercise_4

(a) We have clusters centroids A_1, B_1 and C_1. Let's calculate Euclidian distances.

	A_1	B_1	C_1
A_2	7,348	9,899	4,123
A_3	6,403	10,05	7,874
B_2	3,742	2,449	10,817
B_3	5,745	9,644	11,045
C_2	3,742	5,831	11,874
C_3	5,477	10	9,644
C_4	4,583	8,775	8,367

After the first round execution we will have three clusters $\{A_1A_3B_3C_2C_3C_4\}$, $\{B_1B_2\}$ and $\{A_2C_1\}$. Cluster centers are (4,5; 4,5; 6,83), (1,5; 2; 1,5) and (10,5; 2; 2) respectively.

(b)

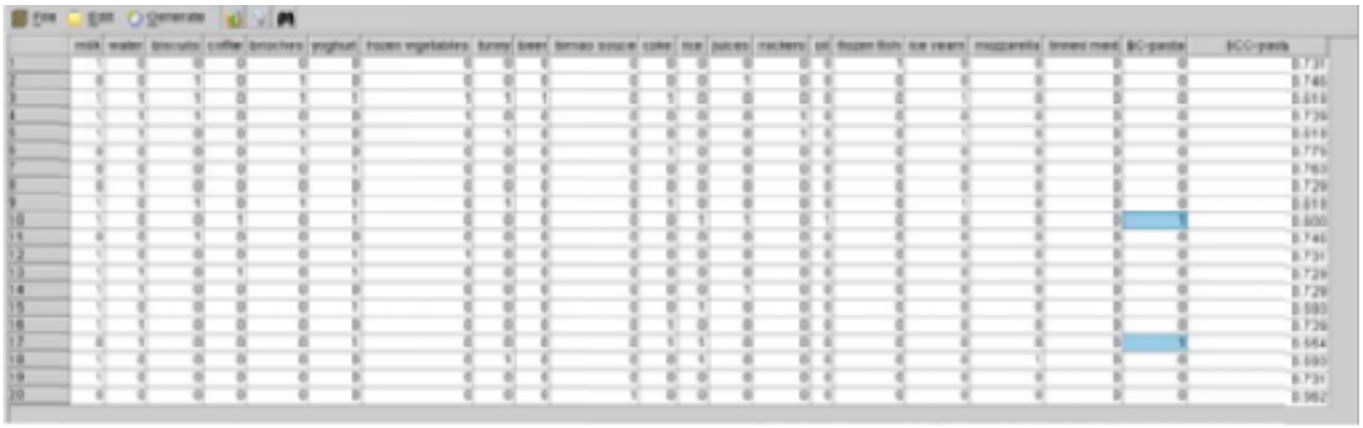
	(4,5; 4,5; 6,83)	(1,5; 2; 1,5)	(10,5; 2; 2)
A_1	3,138	4,301	7,159
A_2	7,337	9,028	3,041
A_3	3,54	8,86	9,552
B_1	7,684	1,225	9,605
B_2	5,642	1,225	8,559
B_3	3,035	8,631	11,011
C_1	9,264	11,811	7,018
C_2	3,632	4,95	10,5
C_3	5,703	9,354	5,315
C_4	1,59	7,649	8,441

Finally, we have three clusters: $\{A_1A_3B_3C_2C_3C_4\}$, $\{B_1B_2\}$ and $\{A_2C_1\}$.

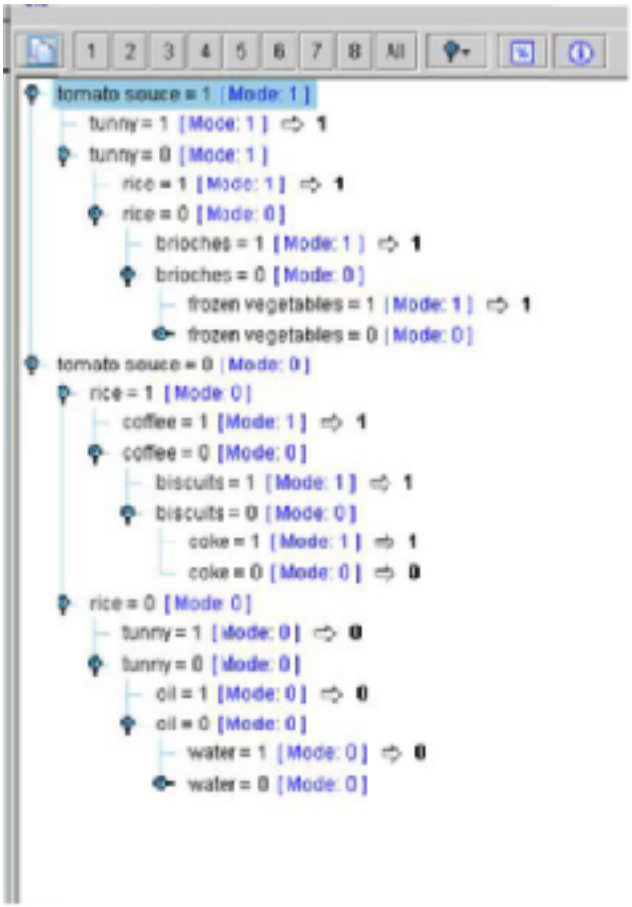
Part_2: Lab
Exercise_1

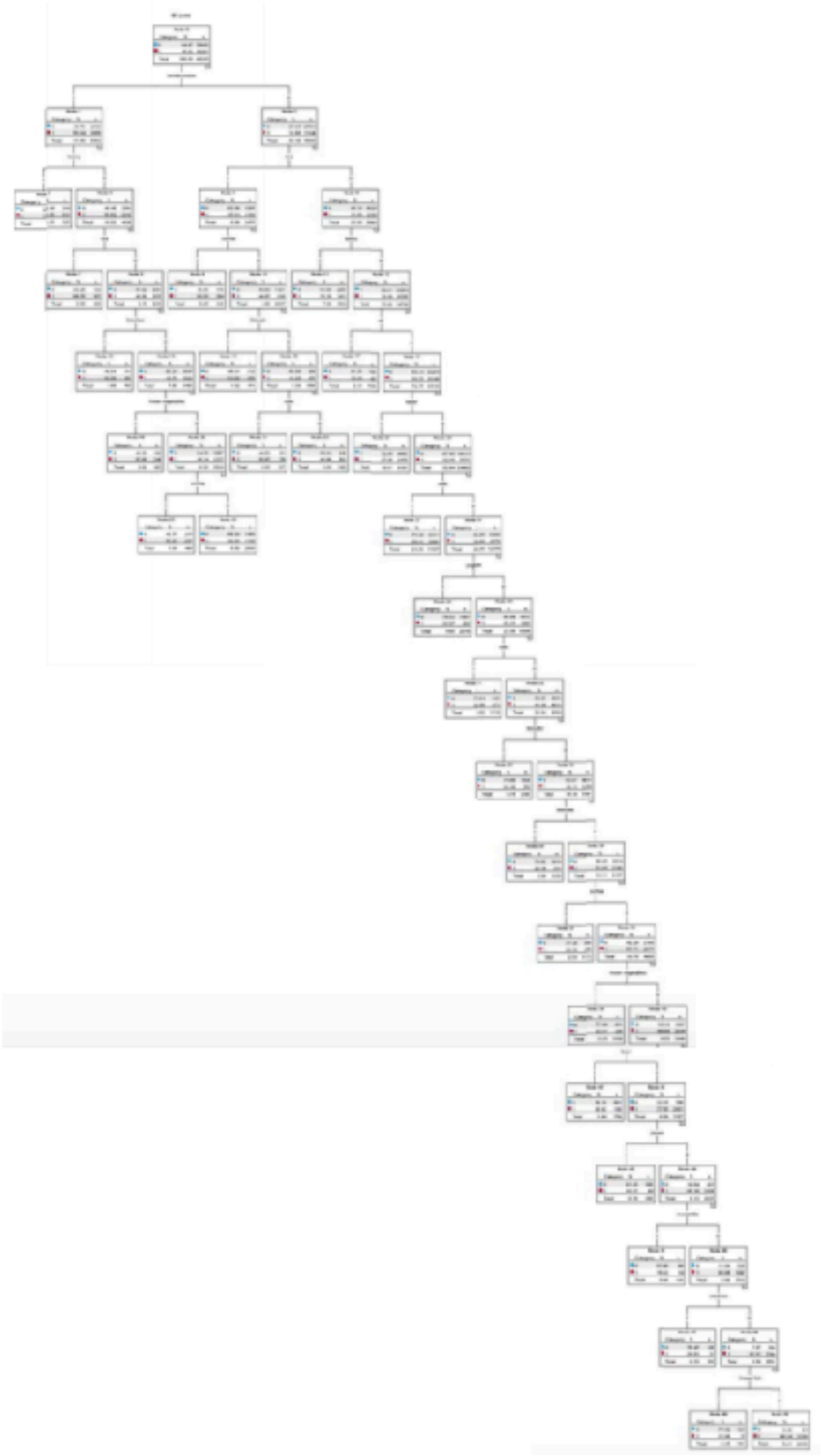
(1) Decision tree can be seen on the next page.

(2)



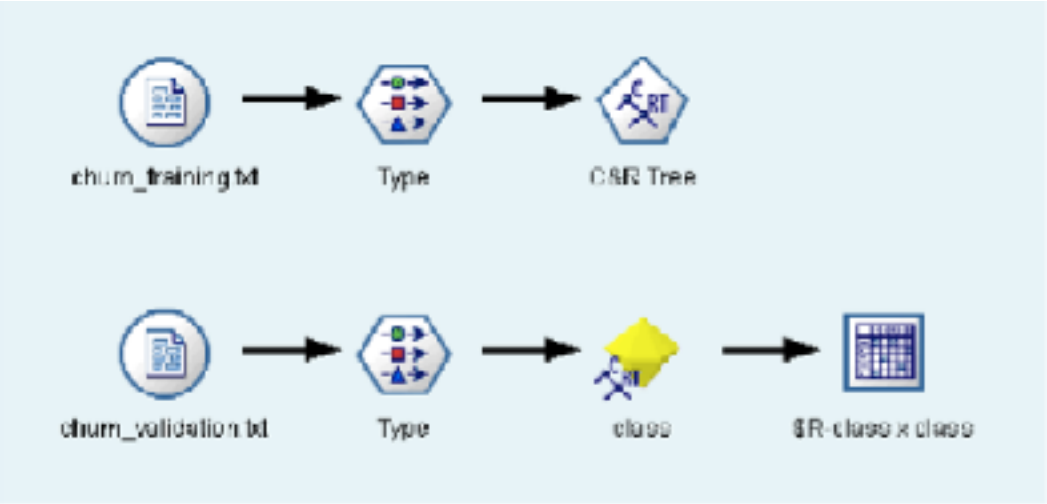
(3)

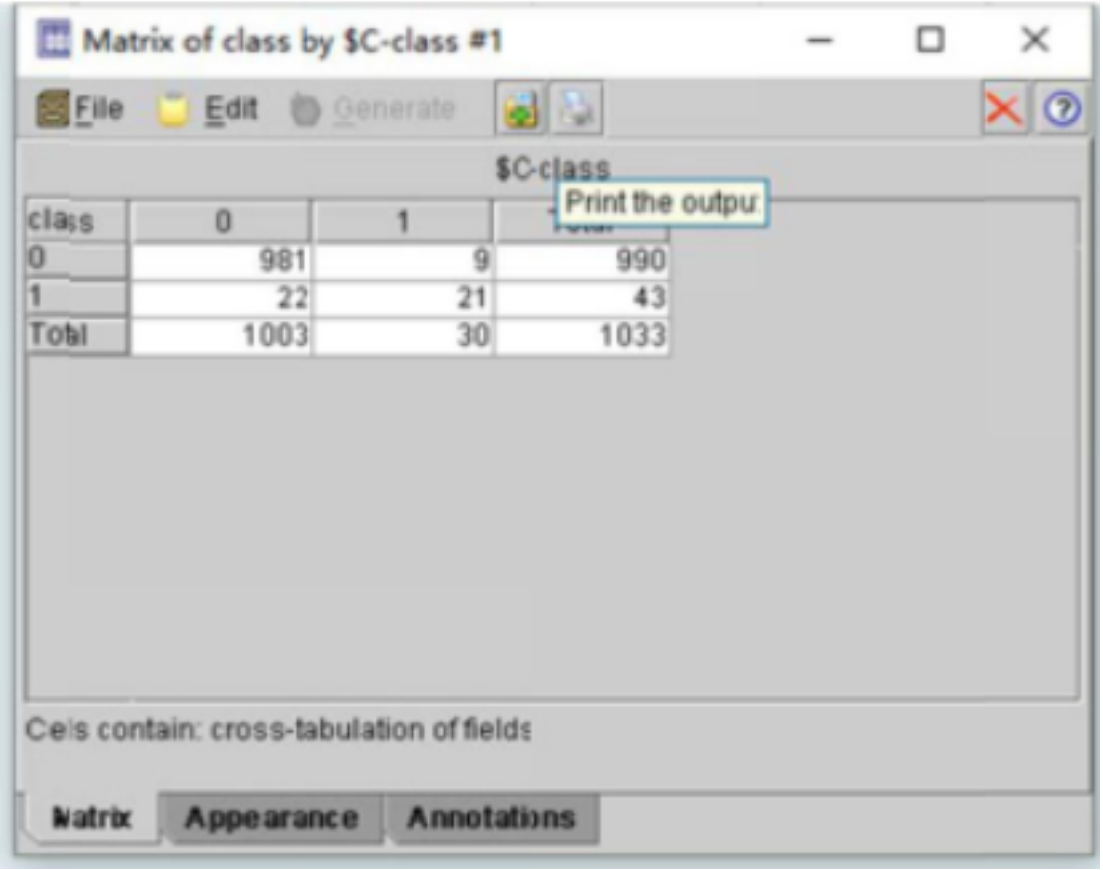




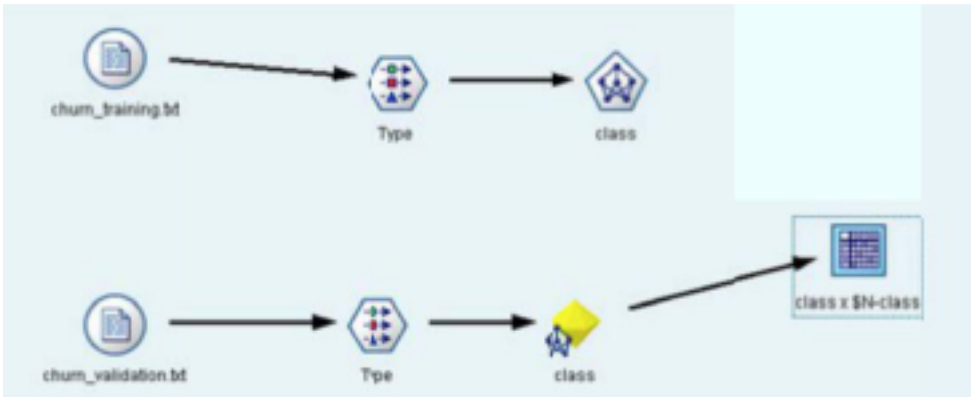
Exercise_2

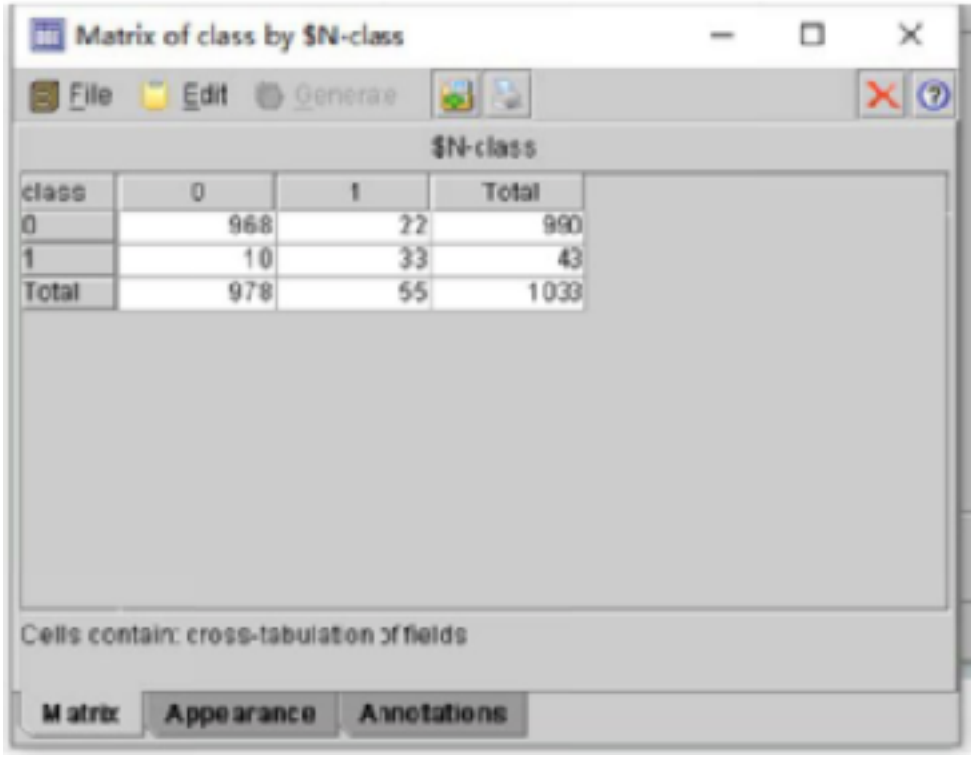
(1)


churn_training.txt
Type
C&R Tree
churn_validation.txt
Type
class
$R-class x class

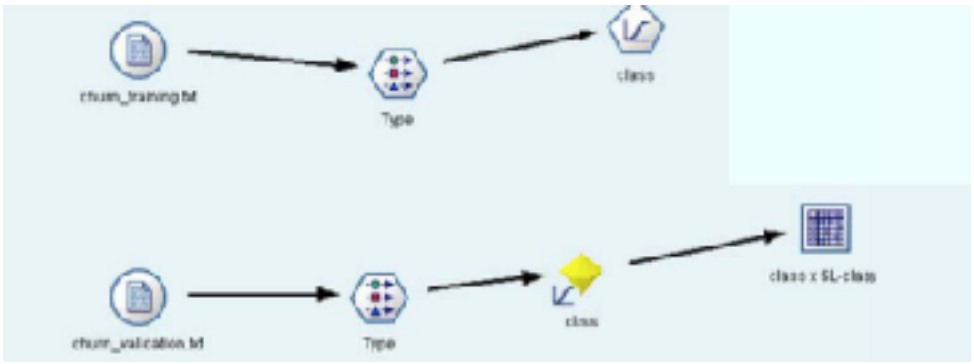

Matrix of class by $C-class #1
File
Edit
Generate
$C-class
Print the outpu
class
0
1
0
981
9
990
1
22
21
43
Total
1003
30
1033
Cells contain: cross-tabulation of fields
Matrix
Appearance
Annotations

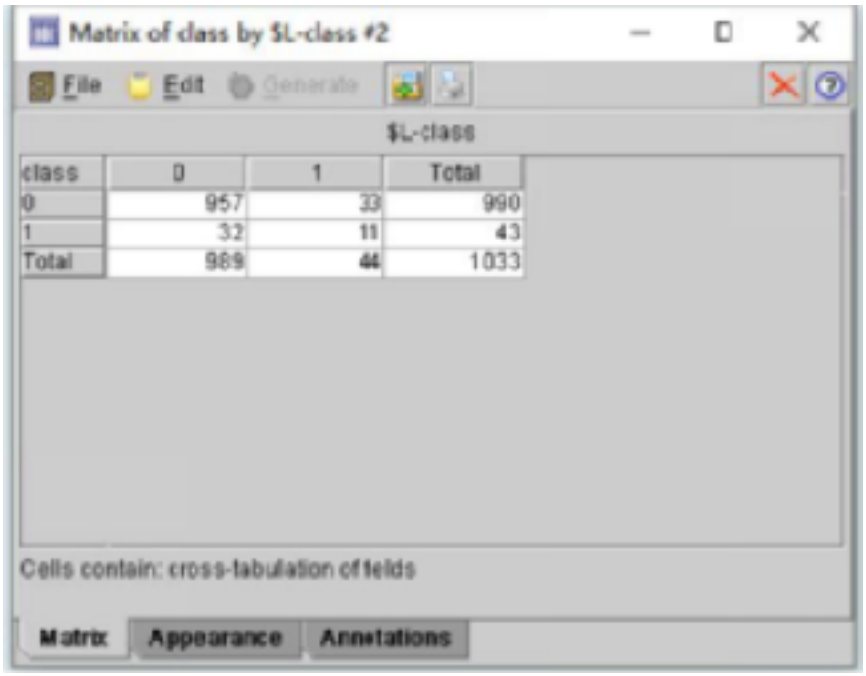
(2)


churn_training.txt
Type
class
class x $N-class
churn_validation.txt
Type
class


Matrix of class by $N-class
File
Edit
Generate
$N-class
class
0
1
Total
0
968
22
990
1
10
33
43
Total
978
55
1033
Cells contain: cross-tabulation of fields
Matrix
Appearance
Annotations

(3)


churn_training.txt
Type
class
class x $L-class
churn_validation.txt
Type
class



(4)

	sensitivity	specifity	precision	accuracy
decision_tree	0,4884	0,9909	0,7000	0,9700
neural_network	0,7674	0,9778	0,6000	0,9690
logistic_regression	0,2558	0,9667	0,2500	0,9371

After we calculated all measures for evaluation, let's choose the best method. Accuracy of decision_tree and neural_network are almost the same. Next we check sensitivity and specificity - neural_network predict positive values much better, than decision_tree, while difference in negative values prediction is not that dramatic. It allows to conclude, that neural_network is better.