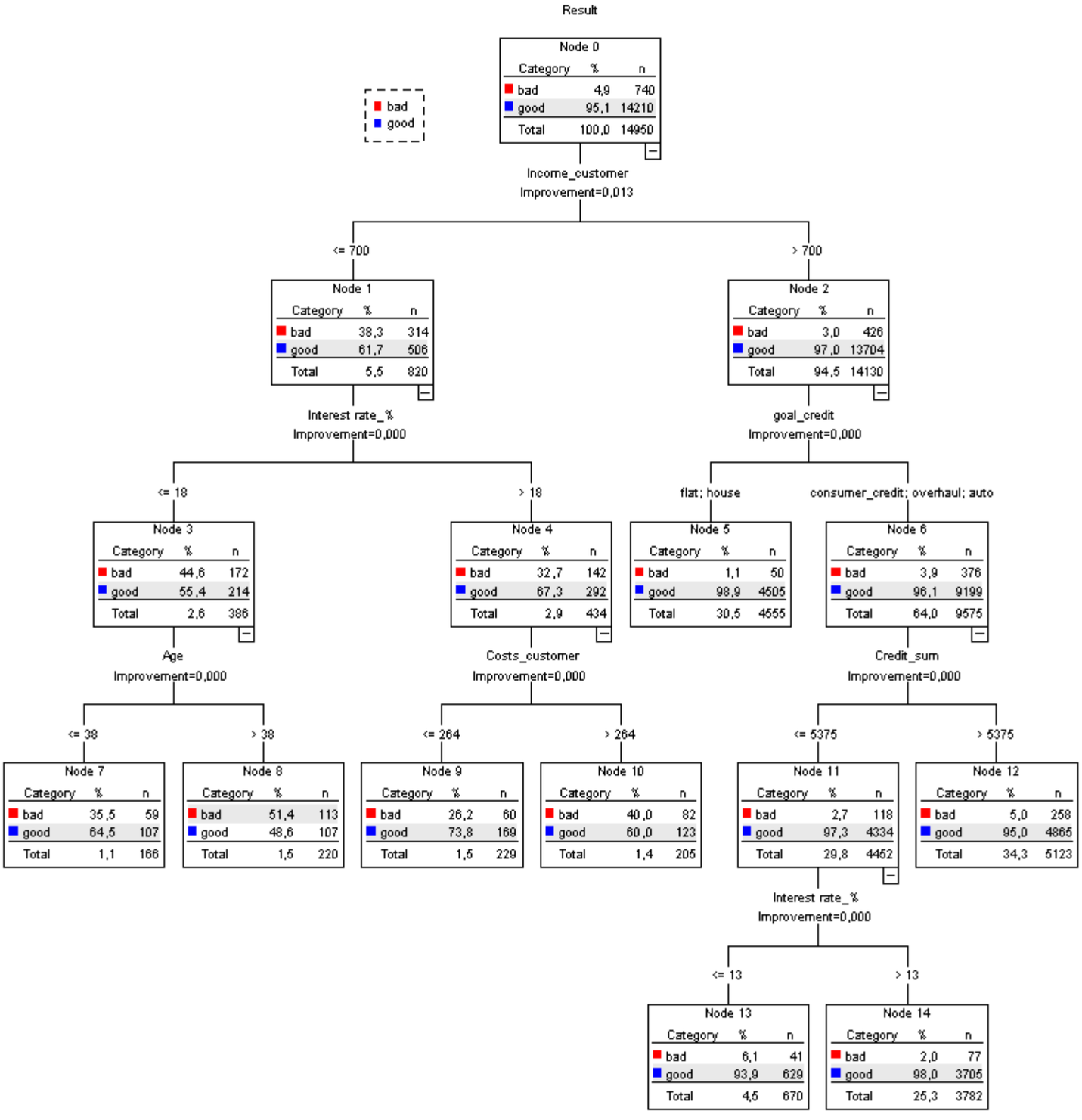
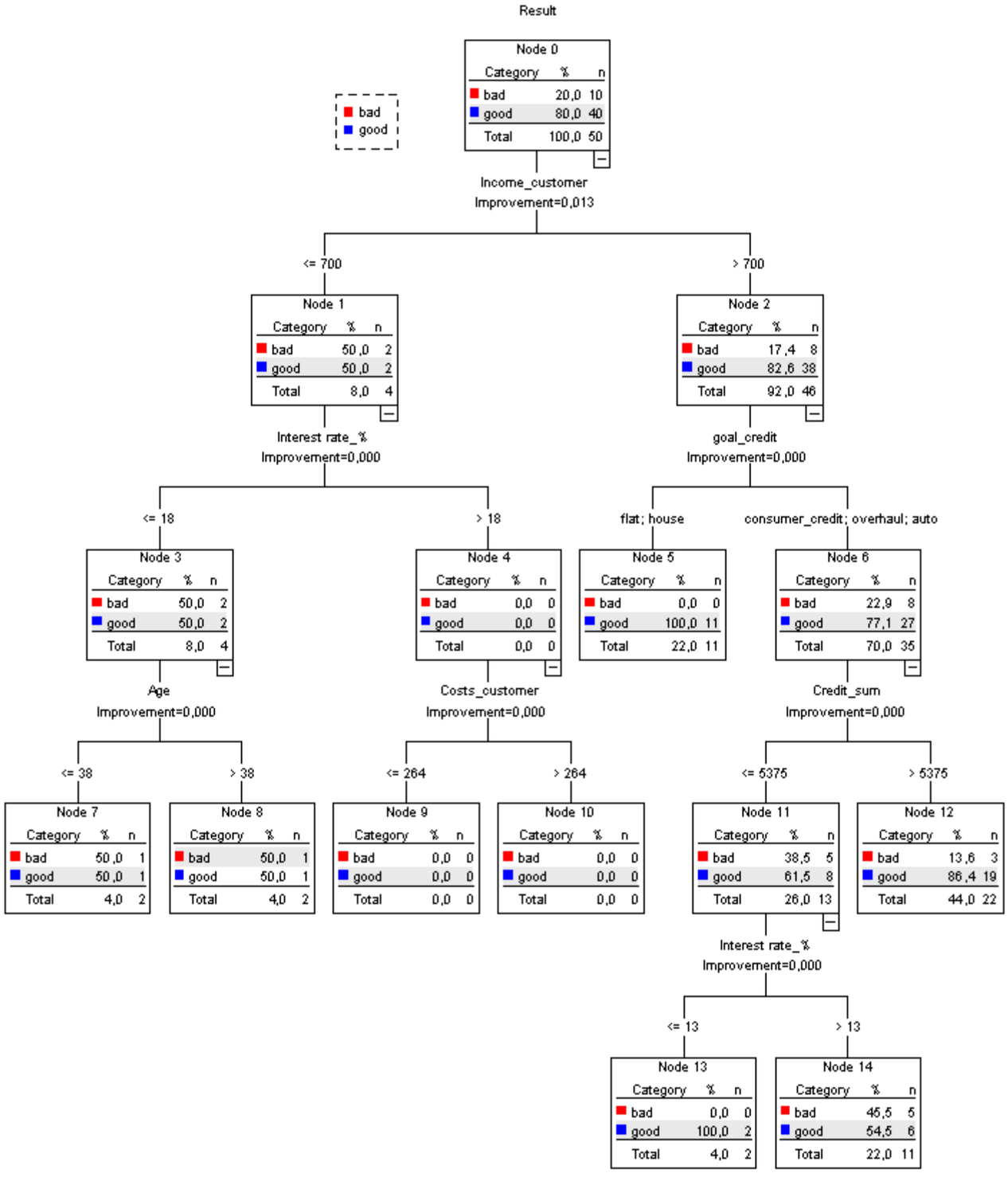
Олійник Богдан КА-83
Аналіз фінансово-економічних даних
Завдання 3
Дерева рішень

Розглянемо метод CRT

Training Tree


Result
bad
good
Node 0
Category % n
bad 4,9 740
good 95,1 14210
Total 100,0 14950
Income_customer
Improvement=0,013
<= 700
> 700
Node 1
Category % n
bad 38,3 314
good 61,7 506
Total 5,5 820
Node 2
Category % n
bad 3,0 426
good 97,0 13704
Total 94,5 14130
Interest rate_%
Improvement=0,000
goal_credit
Improvement=0,000
<= 18
> 18
flat; house
consumer_credit; overhaul; auto
Node 3
Category % n
bad 44,6 172
good 55,4 214
Total 2,6 386
Node 4
Category % n
bad 32,7 142
good 67,3 292
Total 2,9 434
Node 5
Category % n
bad 1,1 50
good 98,9 4505
Total 30,5 4555
Node 6
Category % n
bad 3,9 376
good 96,1 9199
Total 64,0 9575
Age
Improvement=0,000
Costs_customer
Improvement=0,000
Credit_sum
Improvement=0,000
<= 38
> 38
<= 264
> 264
<= 5375
> 5375
Node 7
Category % n
bad 35,5 59
good 64,5 107
Total 1,1 166
Node 8
Category % n
bad 51,4 113
good 48,6 107
Total 1,5 220
Node 9
Category % n
bad 26,2 60
good 73,8 169
Total 1,5 229
Node 10
Category % n
bad 40,0 82
good 60,0 123
Total 1,4 205
Node 11
Category % n
bad 2,7 118
good 97,3 4334
Total 29,8 4452
Node 12
Category % n
bad 5,0 258
good 95,0 4865
Total 34,3 5123
Interest rate_%
Improvement=0,000
<= 13
> 13
Node 13
Category % n
bad 6,1 41
good 93,9 629
Total 4,5 670
Node 14
Category % n
bad 2,0 77
good 98,0 3705
Total 25,3 3782


Result
bad
good
Node 0
Category % n
bad 20,0 10
good 80,0 40
Total 100,0 50
Income_customer
Improvement=0,013
<= 700
> 700
Node 1
Category % n
bad 50,0 2
good 50,0 2
Total 8,0 4
Node 2
Category % n
bad 17,4 8
good 82,6 38
Total 92,0 46
Interest rate_%
Improvement=0,000
goal_credit
Improvement=0,000
<= 18
> 18
flat; house
consumer_credit; overhaul; auto
Node 3
Category % n
bad 50,0 2
good 50,0 2
Total 8,0 4
Node 4
Category % n
bad 0,0 0
good 0,0 0
Total 0,0 0
Node 5
Category % n
bad 0,0 0
good 100,0 11
Total 22,0 11
Node 6
Category % n
bad 22,9 8
good 77,1 27
Total 70,0 35
Age
Improvement=0,000
Costs_customer
Improvement=0,000
Credit_sum
Improvement=0,000
<= 38
> 38
<= 264
> 264
<= 5375
> 5375
Node 7
Category % n
bad 50,0 1
good 50,0 1
Total 4,0 2
Node 8
Category % n
bad 50,0 1
good 50,0 1
Total 4,0 2
Node 9
Category % n
bad 0,0 0
good 0,0 0
Total 0,0 0
Node 10
Category % n
bad 0,0 0
good 0,0 0
Total 0,0 0
Node 11
Category % n
bad 38,5 5
good 61,5 8
Total 26,0 13
Node 12
Category % n
bad 13,6 3
good 86,4 19
Total 44,0 22
Interest rate_%
Improvement=0,000
<= 13
> 13
Node 13
Category % n
bad 0,0 0
good 100,0 2
Total 4,0 2
Node 14
Category % n
bad 45,5 5
good 54,5 6
Total 22,0 11

Test Tree

Risk

Sample	Estimate	Std. Error
Training	,049	,002
Test	,200	,057

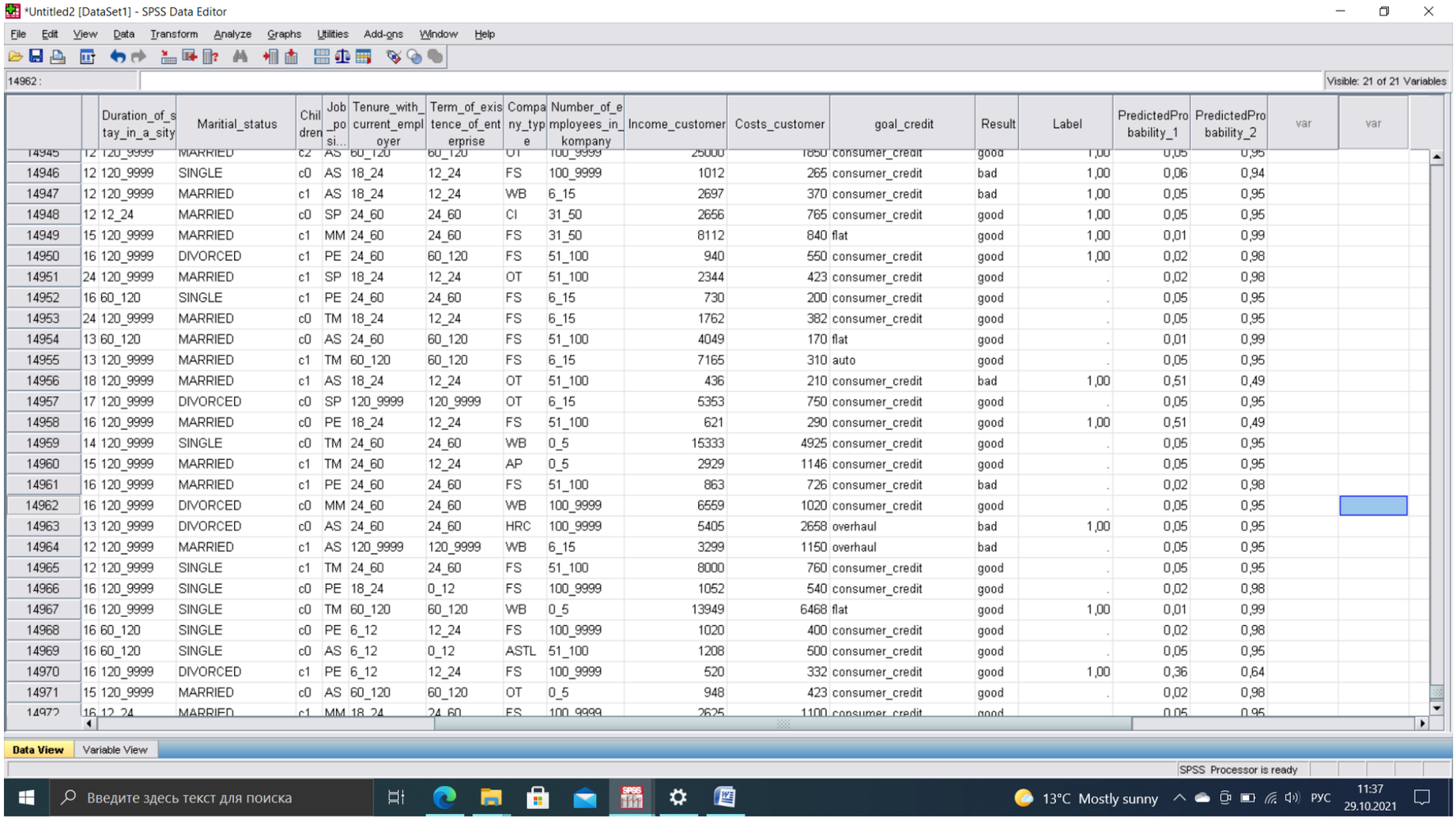
Growing Method: CRT
Dependent Variable: Result

Classification

Sample	Observed	Predicted bad	Predicted good	Percent Correct
Training	bad	113	627	15,3%
	good	107	14103	99,2%
	Overall Percentage	1,5%	98,5%	95,1%
Test	bad	1	9	10,0%
	good	1	39	97,5%
	Overall Percentage	4,0%	96,0%	80,0%

Growing Method: CRT
Dependent Variable: Result

Щоб покращити результати, було вирішено додати до навчальної вибірки деякі кейси, вірогідність яких(за першою спробою) була або близька до 50/50, а також деякі рандомні данні, щоб наростити кількість у навчальній виборці.



Risk

Sample	Estimate	Std. Error
Training	,050	,002
Test	,081	,045

Growing Method: CRT
Dependent Variable: Result

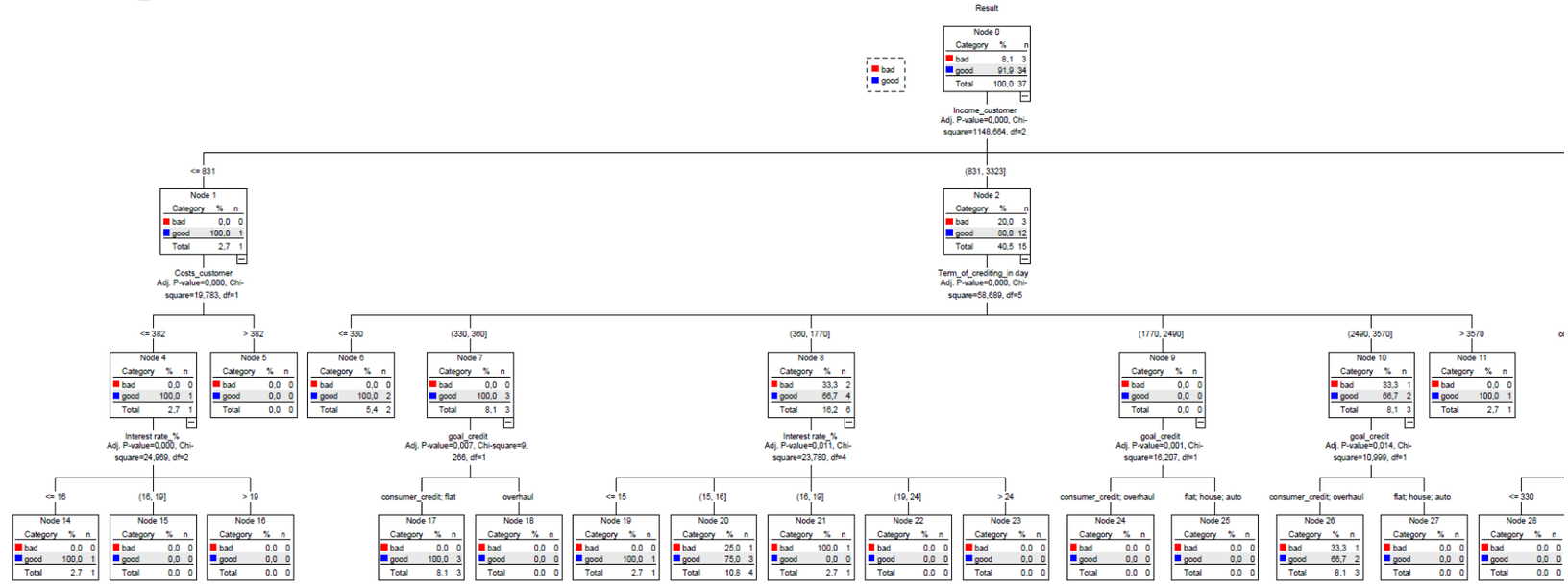
Classification

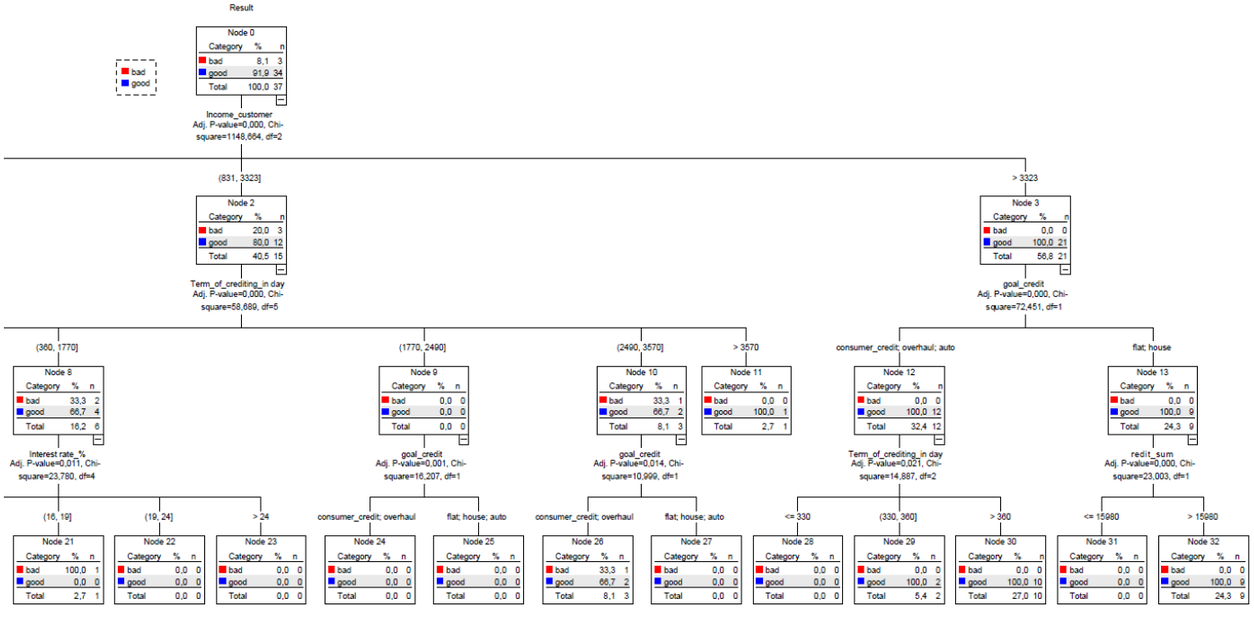
Sample	Observed	Predicted bad	Predicted good	Percent Correct
Training	bad	114	633	15,3%
	good	108	14108	99,2%
	Overall Percentage	1,5%	98,5%	95,0%
Test	bad	0	3	,0%
	good	0	34	100,0%
	Overall Percentage	,0%	100,0%	91,9%

Growing Method: CRT
Dependent Variable: Result

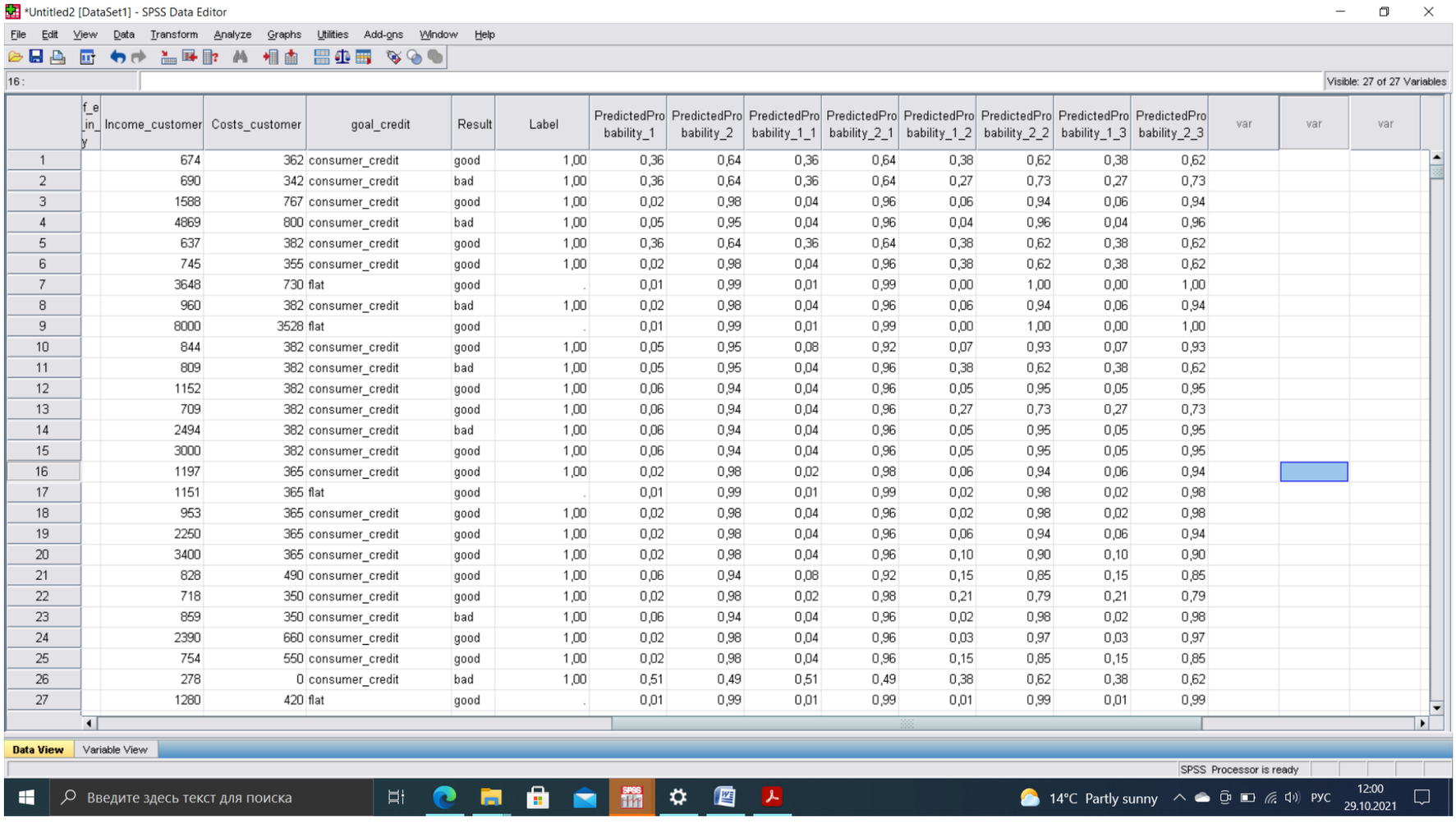
Як ми можемо бачити-це суттєво допомогло покращити якість нашого передбачення.

Розглянемо метод CHAID
Training Tree





Було прийнято рішення спробувати ще покращити результати: видаляємо з навчальної вибірки ті кейси, в яких вірогідність гарного предикту близька до 1



	f_ein_y	Income_customer	Costs_customer	goal_credit	Result	Label	PredictedProbability_1	PredictedProbability_2	PredictedProbability_1_1	PredictedProbability_2_1	PredictedProbability_1_2	PredictedProbability_2_2	PredictedProbability_1_3	PredictedProbability_2_3	var	var	var
1		674	362	consumer_credit	good	1,00	0,36	0,64	0,36	0,64	0,38	0,62	0,38	0,62			
2		690	342	consumer_credit	bad	1,00	0,36	0,64	0,36	0,64	0,27	0,73	0,27	0,73			
3		1588	767	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,06	0,94	0,06	0,94			
4		4869	800	consumer_credit	bad	1,00	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96			
5		637	382	consumer_credit	good	1,00	0,36	0,64	0,36	0,64	0,38	0,62	0,38	0,62			
6		745	355	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,38	0,62	0,38	0,62			
7		3648	730	flat	good	.	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00			
8		960	382	consumer_credit	bad	1,00	0,02	0,98	0,04	0,96	0,06	0,94	0,06	0,94			
9		8000	3528	flat	good	.	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00			
10		844	382	consumer_credit	good	1,00	0,05	0,95	0,08	0,92	0,07	0,93	0,07	0,93			
11		809	382	consumer_credit	bad	1,00	0,05	0,95	0,04	0,96	0,38	0,62	0,38	0,62			
12		1152	382	consumer_credit	good	1,00	0,06	0,94	0,04	0,96	0,05	0,95	0,05	0,95			
13		709	382	consumer_credit	good	1,00	0,06	0,94	0,04	0,96	0,27	0,73	0,27	0,73			
14		2494	382	consumer_credit	bad	1,00	0,06	0,94	0,04	0,96	0,05	0,95	0,05	0,95			
15		3000	382	consumer_credit	good	1,00	0,06	0,94	0,04	0,96	0,05	0,95	0,05	0,95			
16		1197	365	consumer_credit	good	1,00	0,02	0,98	0,02	0,98	0,06	0,94	0,06	0,94			
17		1151	365	flat	good	.	0,01	0,99	0,01	0,99	0,02	0,98	0,02	0,98			
18		953	365	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,02	0,98	0,02	0,98			
19		2250	365	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,06	0,94	0,06	0,94			
20		3400	365	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,10	0,90	0,10	0,90			
21		828	490	consumer_credit	good	1,00	0,06	0,94	0,08	0,92	0,15	0,85	0,15	0,85			
22		718	350	consumer_credit	good	1,00	0,02	0,98	0,02	0,98	0,21	0,79	0,21	0,79			
23		859	350	consumer_credit	bad	1,00	0,06	0,94	0,04	0,96	0,02	0,98	0,02	0,98			
24		2390	660	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,03	0,97	0,03	0,97			
25		754	550	consumer_credit	good	1,00	0,02	0,98	0,04	0,96	0,15	0,85	0,15	0,85			
26		278	0	consumer_credit	bad	1,00	0,51	0,49	0,51	0,49	0,38	0,62	0,38	0,62			
27		1280	420	flat	good	.	0,01	0,99	0,01	0,99	0,01	0,99	0,01	0,99			

Risk

Sample	Estimate	Std. Error
Training	,050	,002
Test	,070	,039

Growing Method: CHAID
Dependent Variable: Result

Classification

Sample	Observed	Predicted bad	Predicted good	Percent Correct
Training	bad	0	747	,0%
	good	0	14210	100,0%
	Overall Percentage	,0%	100,0%	95,0%
Test	bad	0	3	,0%
	good	0	40	100,0%
	Overall Percentage	,0%	100,0%	93,0%

Growing Method: CHAID
Dependent Variable: Result

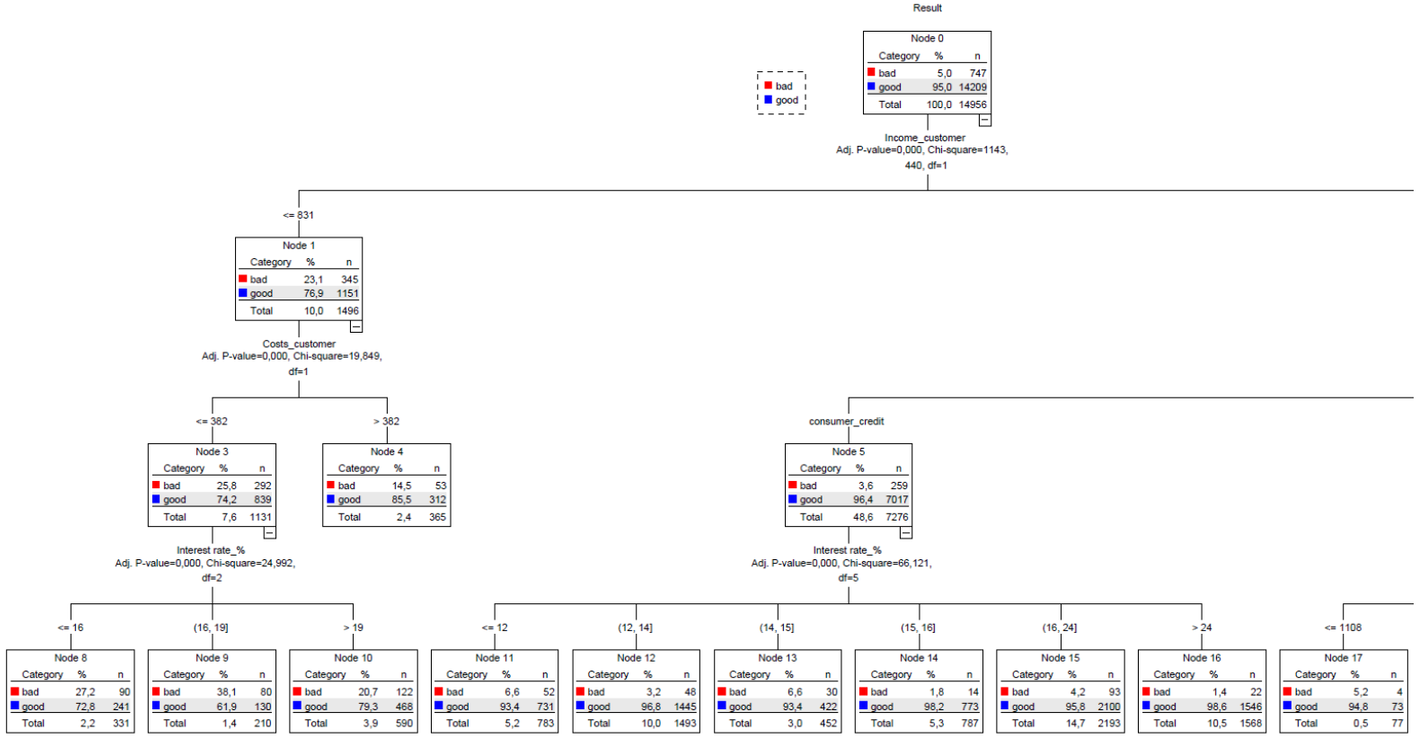
Якщо порівняти ці два методи, то можно сказати, що CHAID більш підходить до нашої моделі

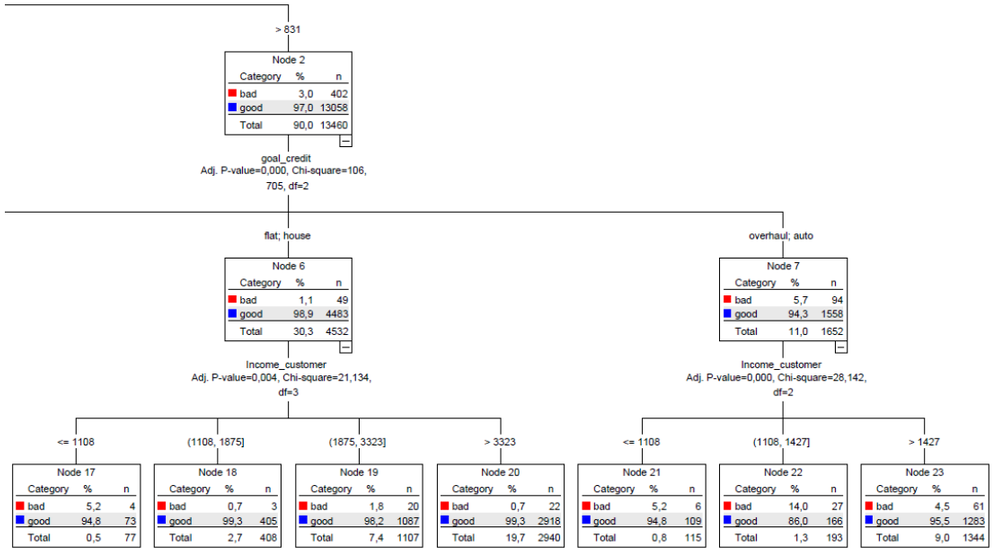
Як ми бачимо, виділений нижче кейс у цьому методі дає результат набагато ближчий до істини.

	Result	Label	PredictedProbability_1	PredictedProbability_2	PredictedProbability_1_1	PredictedProbability_2_1	PredictedProbability_1_2	PredictedProbability_2_2	PredictedProbability_1_3	PredictedProbability_2_3	PredictedProbability_1_4	PredictedProbability_2_4
	bad	1,00	0,01	0,99	0,01	0,99	0,02	0,98	0,02	0,98	0,02	0,98
	good	1,00	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00	0,00	1,00
	good	1,00	0,06	0,94	0,04	0,96	0,05	0,95	0,05	0,95	0,05	0,95
	good	1,00	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96
	good	1,00	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96
	good	1,00	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00	0,00	1,00
	good	1,00	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96
	bad	1,00	0,06	0,94	0,04	0,96	0,05	0,95	0,05	0,95	0,05	0,95
	bad	1,00	0,05	0,95	0,04	0,96	0,10	0,90	0,10	0,90	0,10	0,90
	good	1,00	0,05	0,95	0,08	0,92	0,10	0,90	0,10	0,90	0,10	0,90
	good	1,00	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00	0,00	1,00
	good	1,00	0,02	0,98	0,04	0,96	0,01	0,99	0,01	0,99	0,01	0,99
	good	.	0,02	0,98	0,02	0,98	0,02	0,98	0,02	0,98	0,02	0,98
	good	.	0,05	0,95	0,04	0,96	0,27	0,73	0,27	0,73	0,27	0,73
	good	.	0,05	0,95	0,02	0,98	0,02	0,98	0,02	0,98	0,02	0,98
	good	.	0,01	0,99	0,01	0,99	0,00	1,00	0,00	1,00	0,00	1,00
	good	.	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96
	bad	1,00	0,51	0,49	0,51	0,49	0,38	0,62	0,38	0,62	0,38	0,62
	good	.	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96
	good	.	0,51	0,49	0,51	0,49	0,27	0,73	0,27	0,73	0,27	0,73
	good	.	0,05	0,95	0,04	0,96	0,04	0,96	0,04	0,96	0,04	0,96

Розглянемо метод CHAID-exhaustive

Training Tree





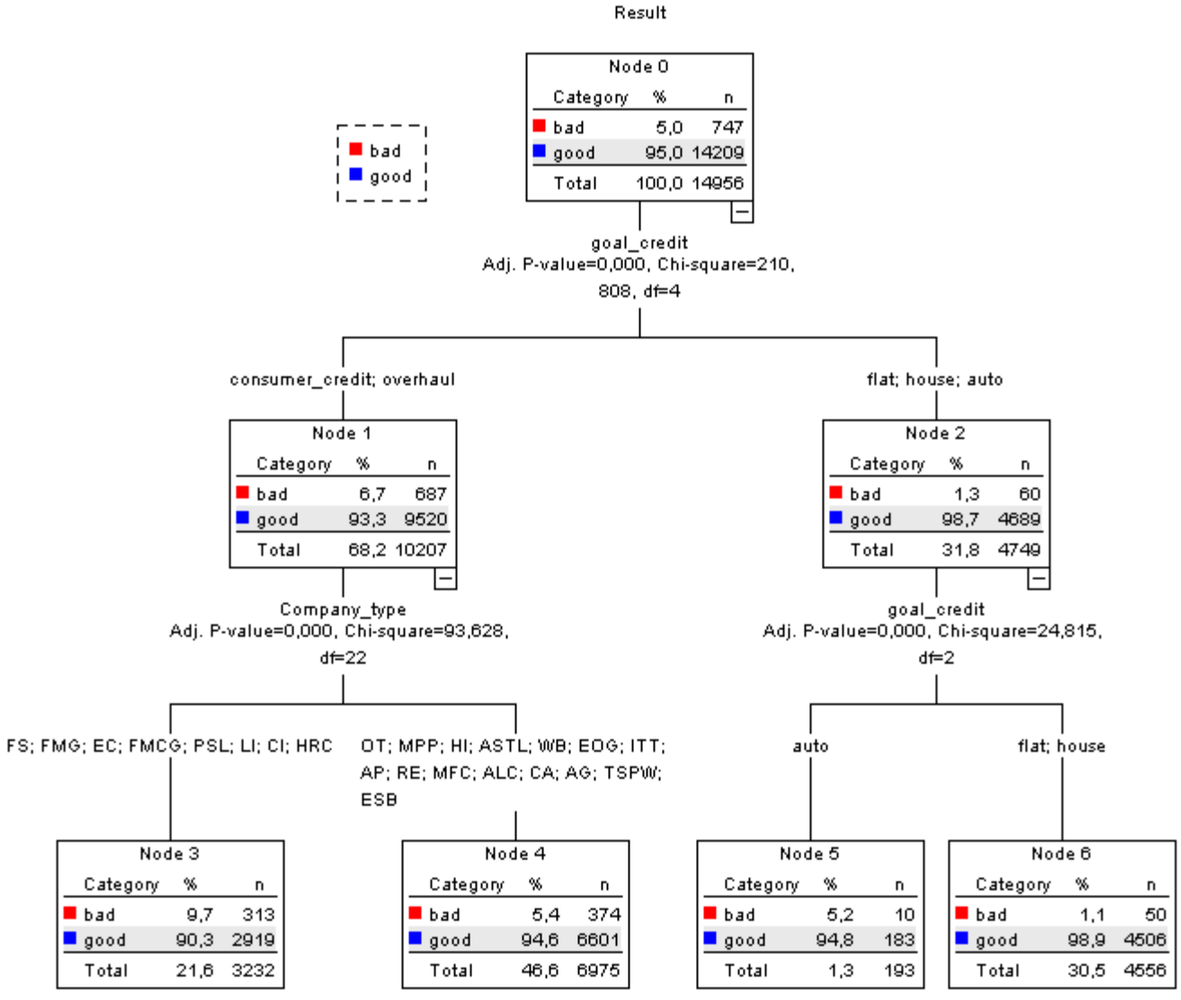
Classification

Sample	Observed	Predicted bad	Predicted good	Percent Correct
Training	bad	0	747	,0%
	good	0	14209	100,0%
	Overall Percentage	,0%	100,0%	95,0%
Test	bad	0	3	,0%
	good	0	41	100,0%
	Overall Percentage	,0%	100,0%	93,2%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Result

Як ми можемо бачити цей метод дає ще більшу точність навіть без корегування вибірки.

Розглянемо метод QUEST
Training Tree



Risk

Sample	Estimate	Std. Error
Training	,050	,002
Test	,068	,038

Growing Method: QUEST
Dependent Variable: Result

Classification

Sample	Observed	Predicted bad	Predicted good	Percent Correct
Training	bad	0	747	,0%
	good	0	14209	100,0%
	Overall Percentage	,0%	100,0%	95,0%
Test	bad	0	3	,0%
	good	0	41	100,0%
	Overall Percentage	,0%	100,0%	93,2%

Growing Method: QUEST
Dependent Variable: Result

good	.	0,51	0,49	0,51	0,49	0,27	0,73	0,27	0,73	0,27	0,73	0,27	0,73	0,10	0,90	
bad	.	0,02	0,98	0,04	0,96	0,01	0,99	0,01	0,99	0,01	0,99	0,02	0,98	0,10	0,90	

Як ми можемо бачити наприкладі цих двох кейсів, метод QUEST дав нам найбільш ближчий результат до істини

ВИСНОВОК

Мною було реалізовано чотири методи дерев рішень на одній й тій самій вибірці. Спочатку я намагався розділити тестові та навчальні вибірки таким чином, щоб процент похибки був найменшим.

Реалізувавши всі ці методи, можна сказати, що всі вони не дають 100% результату, більш реалістичним методом був QUEST, але я вважаю, що якщо ми хочемо покращити наші результати предикту, то потрібно реалізовувати одразу декілька методів і на основі цих результатів аналізувати що нам підходить краще(або використовувати одразу декілька методів у комплексі)