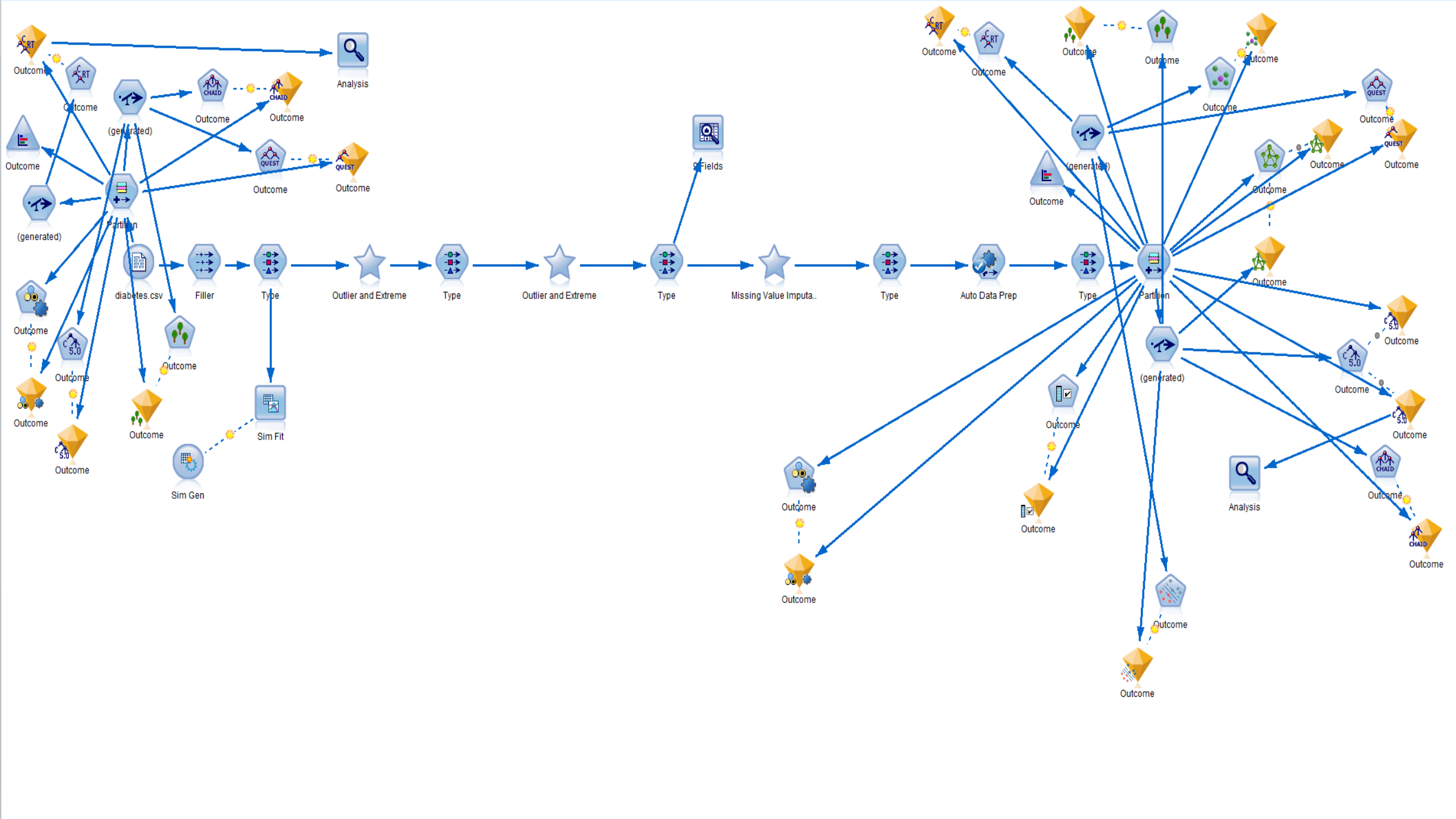
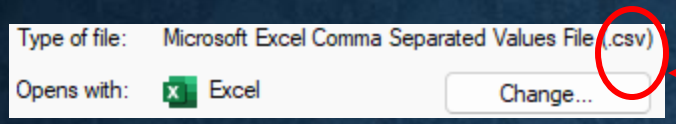
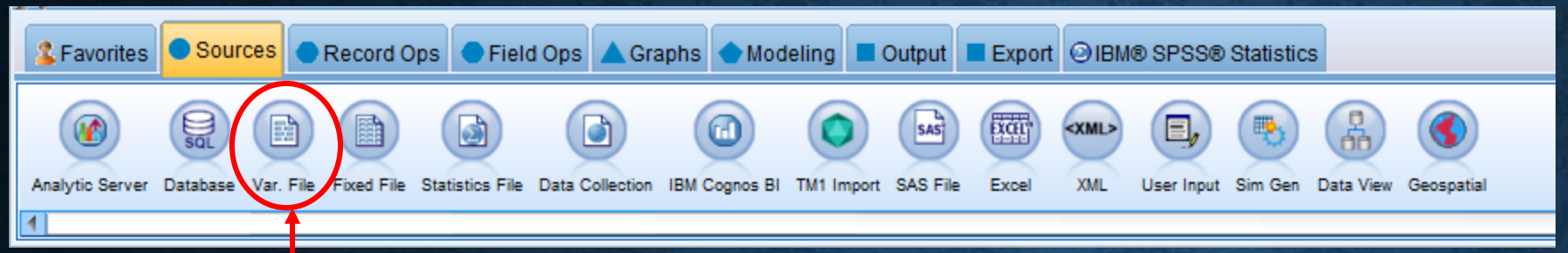
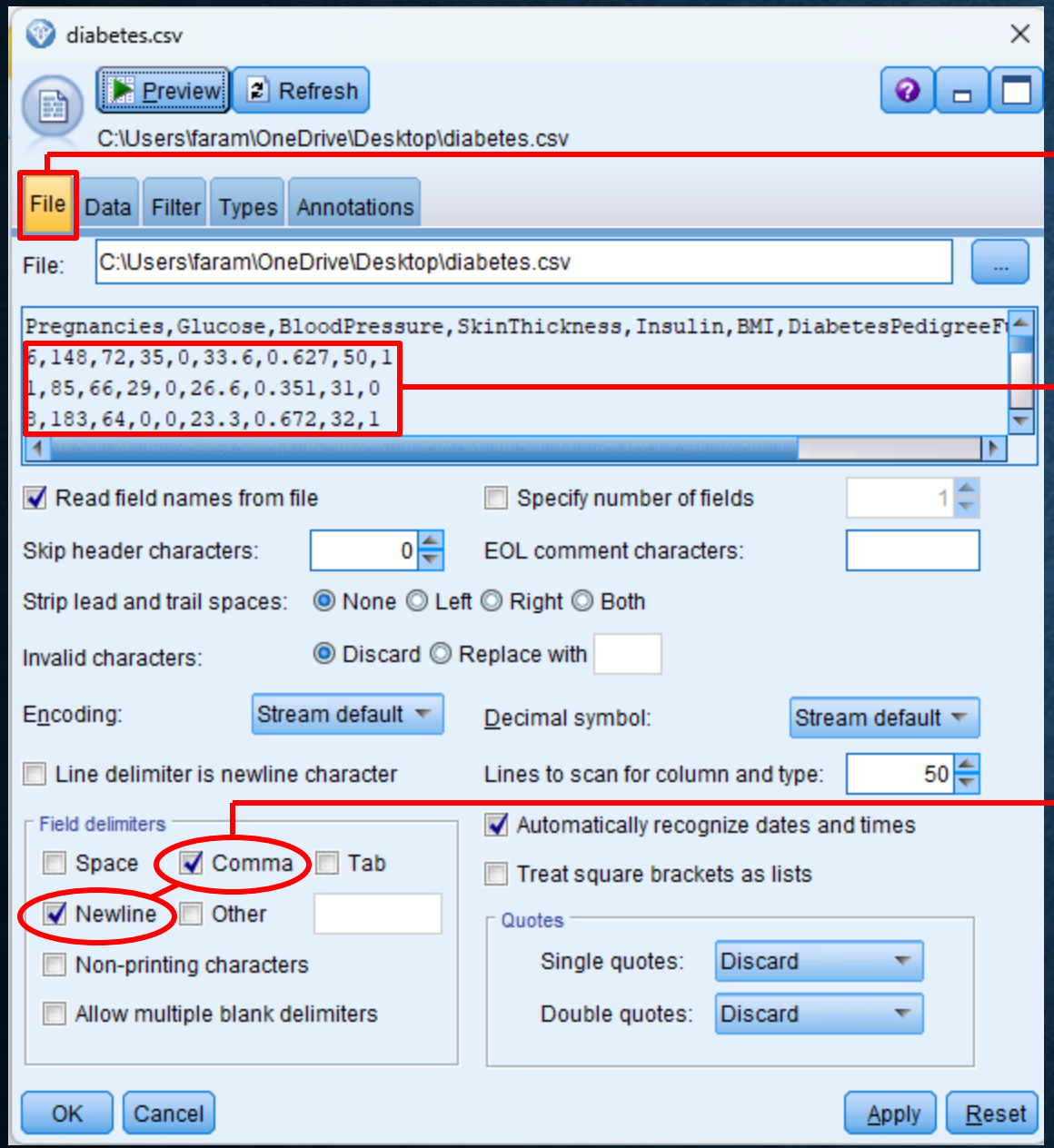

Outcome
Analysis
(generated)
Partition
diabetes.csv
Filler
Type
Outlier and Extreme
Fields
Missing Value Imputa..
Auto Data Prep
Sim Fit
Sim Gen

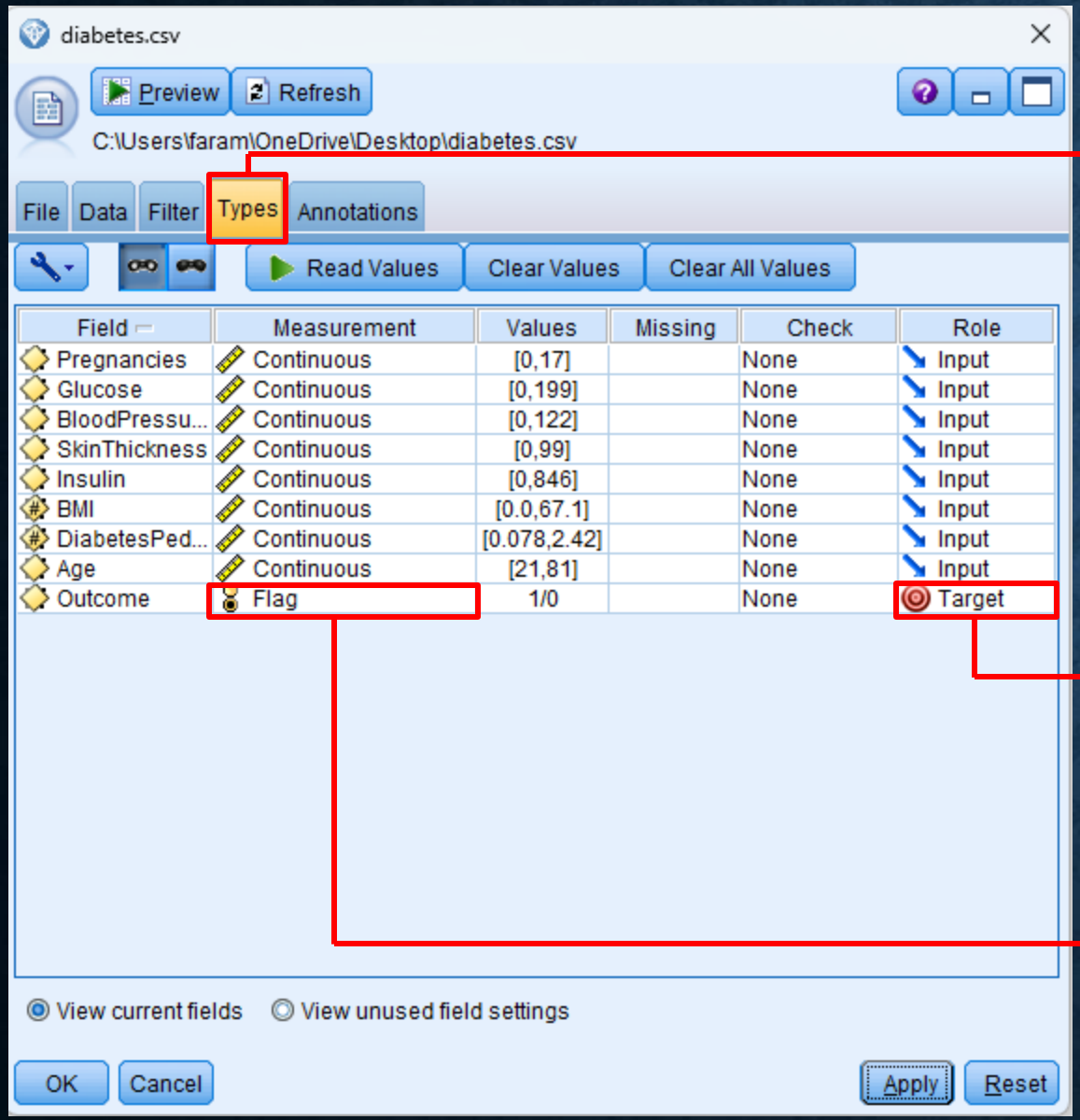




به دلیل اینکه فرمت فایل تمرین **CSV.** بوده ؛ علارغم اینکه به صورت اکسل نمایش داده می شود، از گزینه **Var. File** در تب **Source** استفاده میکنیم.



در تب File به دلیل اینکه اعداد با Comma از هم دیگر جدا شده اند و همچنین هر فرد جدید در سطر جدیدی (NewLine) قرار گرفته این دو را "✓" می زنیم



به منظور تعیین تکلیف(ساختاربندی) توزیع اعدادی که داریم به این تب رفته

به دلیل اینکه متغییر هدف ما ستون Outcome است ؛ نقش این ستون را به Target تغییر می دهیم.

چون اعدادی که در این ستون اعداد داریم 0 و 1 است، از توزیع Flag استفاده می کنیم

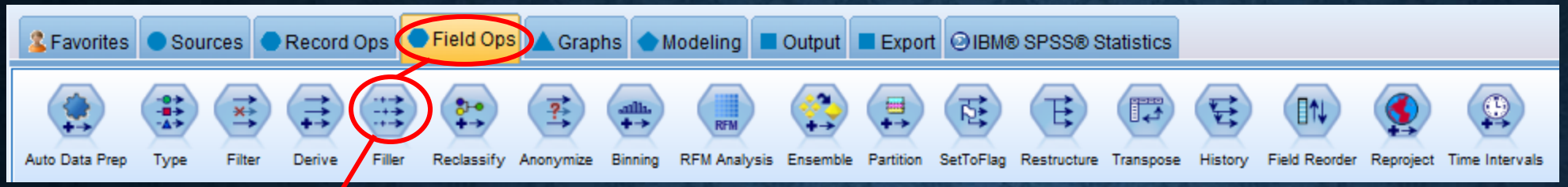


Table (9 fields, 768 records)

File Edit Generate

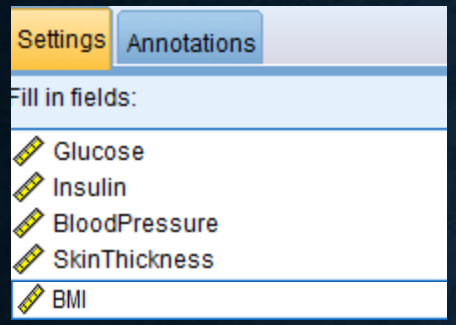
Table Annotations

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33....	0.627	50	1
2	1	85	66	29	0	26....	0.351	31	0
3	8	183	64	0	0	23....	0.672	32	1
4	1	89	66	23	94	28....	0.167	21	0
5	0	137	40	35	168	43....	2.288	33	1
6	5	116	74	0	0	25....	0.201	30	0
7	3	78	50	32	88	31....	0.248	26	1
8	10	115	0	0	0	35....	0.134	29	0
9	2	197	70	45	543	30....	0.158	53	1
10	8	125	96	0	0	0.0...	0.232	54	1
11	4	110	92	0	0	37....	0.191	30	0
12	10	168	74	0	0	38....	0.537	34	1
13	10	139	80	0	0	27....	1.441	57	0
14	1	189	60	23	846	30....	0.398	59	1
15	5	166	72	19	175	25....	0.587	51	1
16	7	100	0	0	0	30....	0.484	32	1
17	0	118	84	47	230	45....	0.551	31	1
18	7	107	74	0	0	29....	0.254	31	1
19	1	103	30	38	83	43....	0.183	33	0
20	1	115	70	30	96	34....	0.529	32	1
21	3	126	88	41	235	39....	0.704	27	0
22	8	99	84	0	0	35....	0.388	50	0
23	7	196	90	0	0	39....	0.451	41	1
24	9	119	80	35	0	29....	0.263	29	1
25	11	143	94	33	146	36....	0.254	51	1
26	10	125	70	26	115	31....	0.205	41	1
27	7	147	76	0	0	39....	0.257	43	1

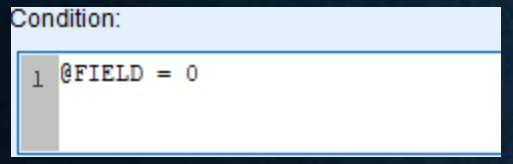
OK

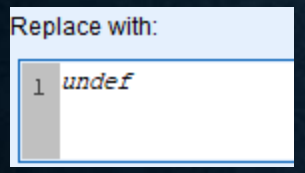
- با توجه به توضیحات داده در زمان تحویل تمرین داده های صفر به معنای نویز است؛ پس ما به **Filler** در تب **Field Ops** نیاز داریم.

- باتوجه به درک از داده های موجود از هر ستون اگر تعداد حاملگی در ستون اول 0 باشد، غیرطبیعی نیست ولی فشار خون 0 یا ضخامت پوست 0 به معنی است.

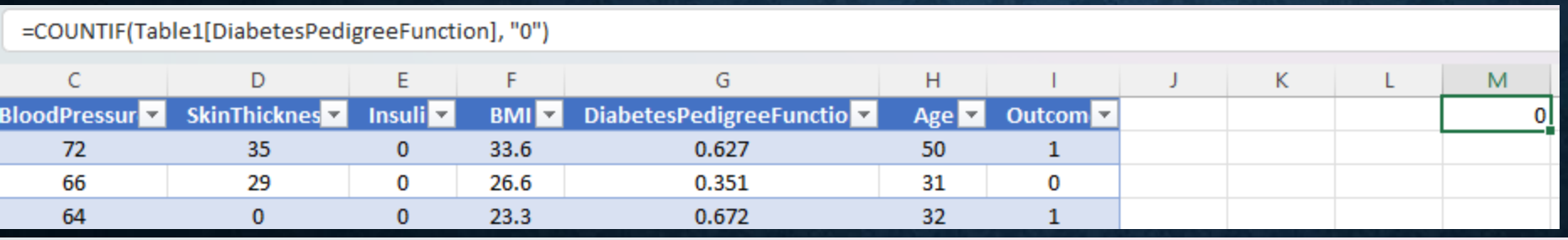


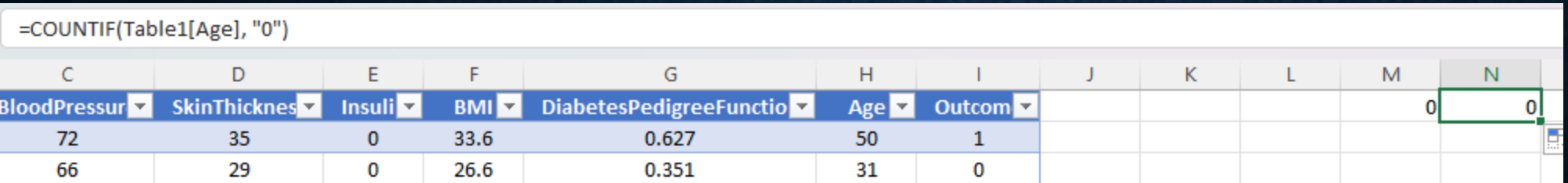
با توجه به نتیجه گیری اسلاید قبلی در ستون داده های مشخص شده ، شرطی میزاریم که اگر اعداد مساوی با "0" بود ، برام تبدیل به **undef** کن. (به معنای نامشخص)

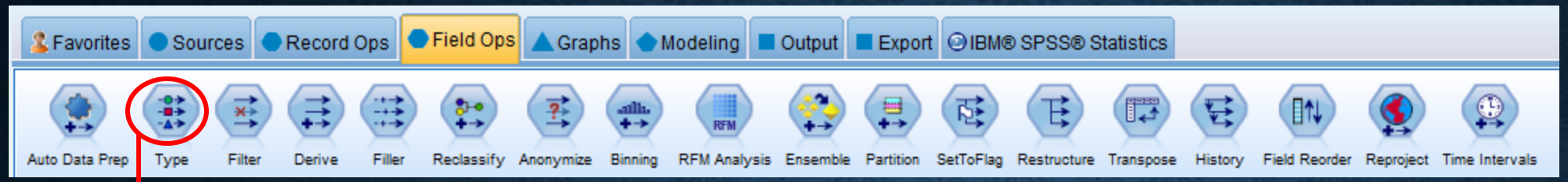




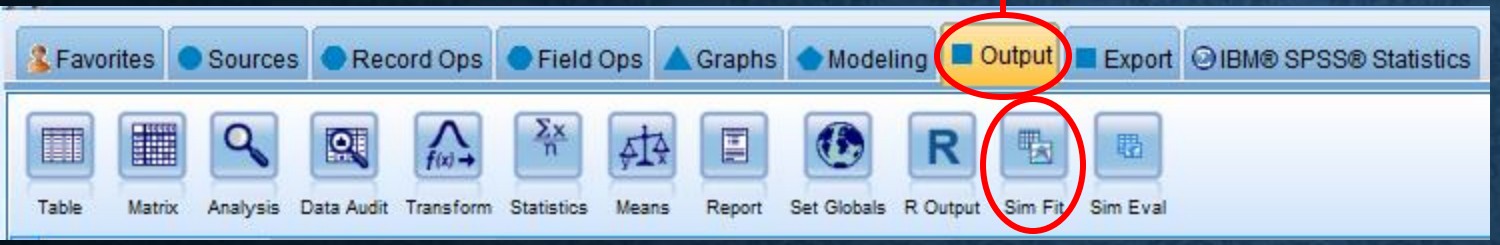
با استفاده از جدول کردن داده ها در اکسل و استفاده از فرمول **"Countif"** تعداد " 0 " موجود در دیگر ستون ها را نیز بررسی میکنیم. (می دانیم که وجود 0 در ستون های حاملگی و نتیجه غیر طبیعی نیست)

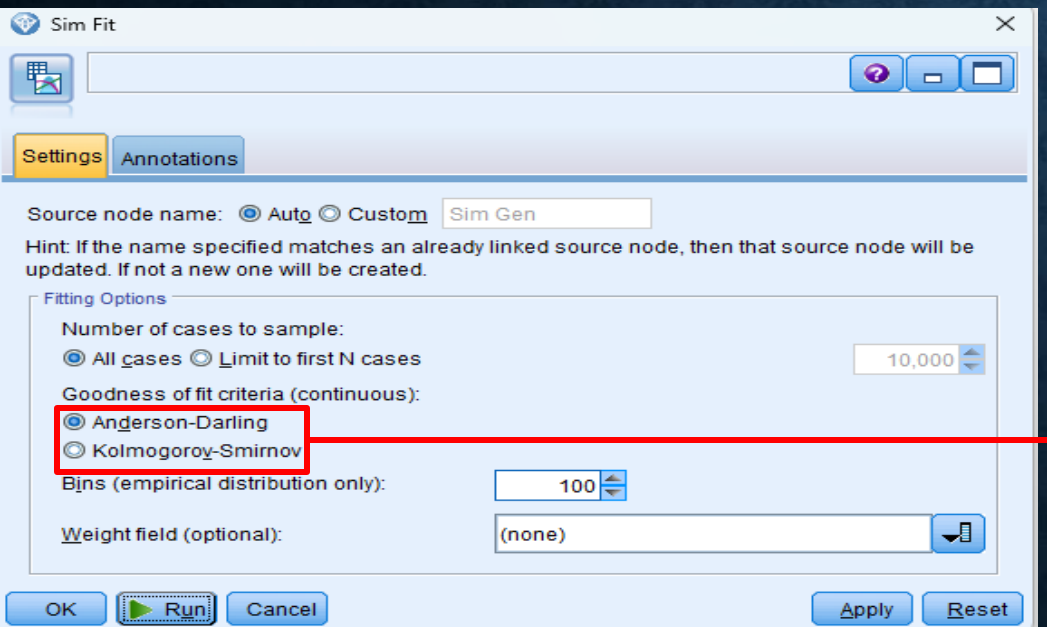






از type بعد از هر تغییری در داده ها استفاده میکنیم؛
سپس به سراغ SimFit برای تست نرمال بودن داده ها می رویم، تا ببینیم داده های پرت را با چه روشی شناسایی کنیم.





از دو روش Anderson-Darling و Kolmogoroy-Smirnov برای تست نرمال بودن می شود استفاده کرد ، که ما در اینجا از هر دو روش پیش می رویم

Field	Storage	Status		Distribution	Parameters	Min,Max
Pregnancies	Integer	✓		Exponential	[scale=0.3029366...	[Max=,Min=]
Glucose	Integer	✓		Lognormal	[a=118.872658692...	[Max=,Min=]
BloodPressure	Integer	✓		Normal	[mean=70.663265...	[Max=,Min=]
SkinThickness	Integer	✓		Weibull	[shape1=32.66296...	[Max=,Min=]
Insulin	Integer	✓		Lognormal	[a=123.115102908...	[Max=,Min=]
BMI	Real	✓		Gamma	[shape=22.699650...	[Max=,Min=]
DiabetesPedigree...	Real	✓		Lognormal	[a=0.43208120745...	[Max=,Min=]
Age	Integer	✓		Lognormal	[a=29.4789886441...	[Max=,Min=]
Outcome	Integer	✓		Categorical	[0=0.66836734693...	

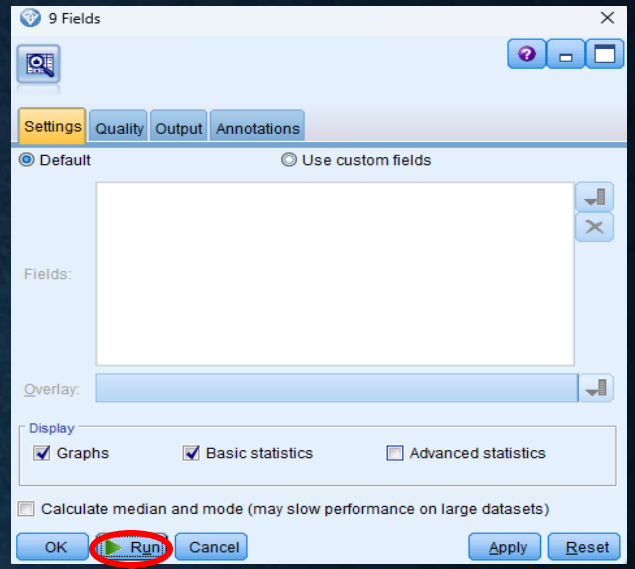
Anderson - Darling

همانطور که مشاهده می کنید هر دو روش ستون داده فشار خون را نرمال تعیین کرده است.

Kolmogoroy - Smirnov

Field	Storage	Status		Distribution	Parameters	Min,Max
Pregnancies	Integer	✓		Exponential	[scale=0.3029366...	[Max=,Min=]
Glucose	Integer	✓		Lognormal	[a=118.872658692...	[Max=,Min=]
BloodPressure	Integer	✓		Normal	[mean=70.663265...	[Max=,Min=]
SkinThickness	Integer	✓		Weibull	[shape1=32.66296...	[Max=,Min=]
Insulin	Integer	✓		Lognormal	[a=123.115102908...	[Max=,Min=]
BMI	Real	✓		Normal	[mean=33.086224...	[Max=,Min=]
DiabetesPedigree...	Real	✓		Lognormal	[a=0.43208120745...	[Max=,Min=]
Age	Integer	✓		Lognormal	[a=29.4789886441...	[Max=,Min=]
Outcome	Integer	✓		Categorical	[0=0.66836734693...	

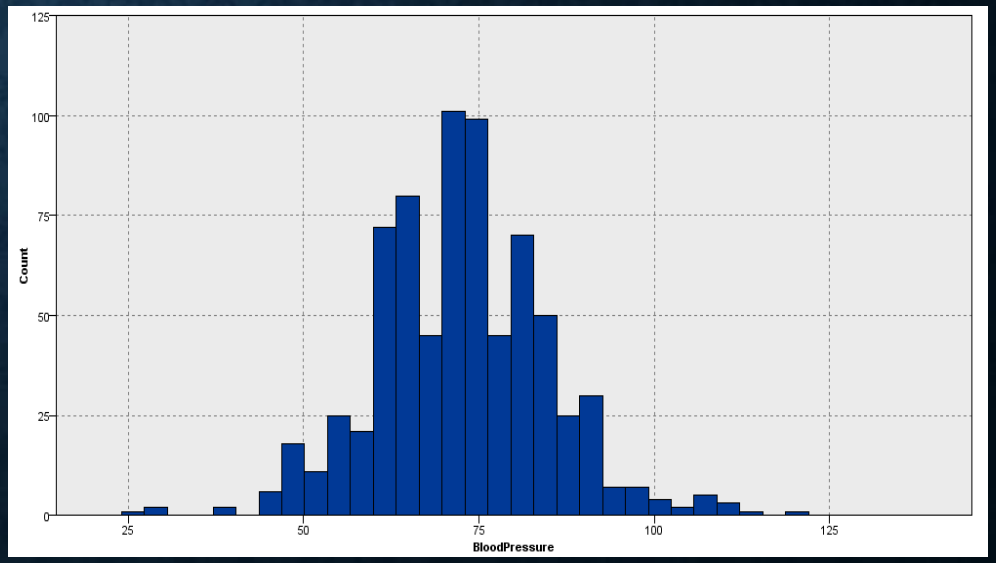
از **Data Audit** در تب **Output** برای تست شکل های توزیع ها نیز می توان استفاده کرد.

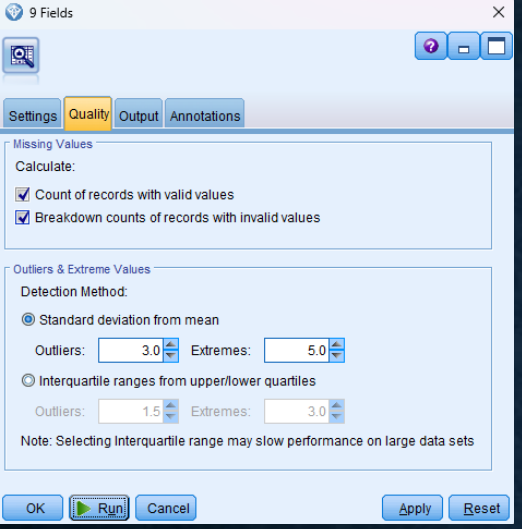


با استفاده از گزینه **Run** شکل ها را می آوریم.

همانطور که در اشکال زیر نیز مشاهده می شود تنها هیستوگرام دارای توزیع نرمال فشار خون می باشد.

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Pregnancies		Continuous	0	17	3.845	3.370	0.902	--	768
Glucose		Continuous	44	199	121.687	30.536	0.531	--	763
BloodPressure		Continuous	24	122	72.405	12.382	0.134	--	733
SkinThickness		Continuous	7	99	29.153	10.477	0.691	--	541
Insulin		Continuous	14	846	155.548	118.776	2.166	--	394
BMI		Continuous	18.200	67.100	32.457	6.925	0.594	--	757
DiabetesPedigreeFunction		Continuous	0.078	2.420	0.472	0.331	1.920	--	768
Age		Continuous	21	81	33.241	11.760	1.130	--	768





در Sim Fit داخل تب Quality رفته و در قسمت Outliers & Extreme Values ، گزینه Standard Devation را انتخاب می کنیم زیرا میخواهیم عملی روی داده های پرتی که داخل ستون فشار خون که دارای توزیع نرمال بود ؛ انجام دهیم.

ارتباط بین فشار خون بالا و دیابت

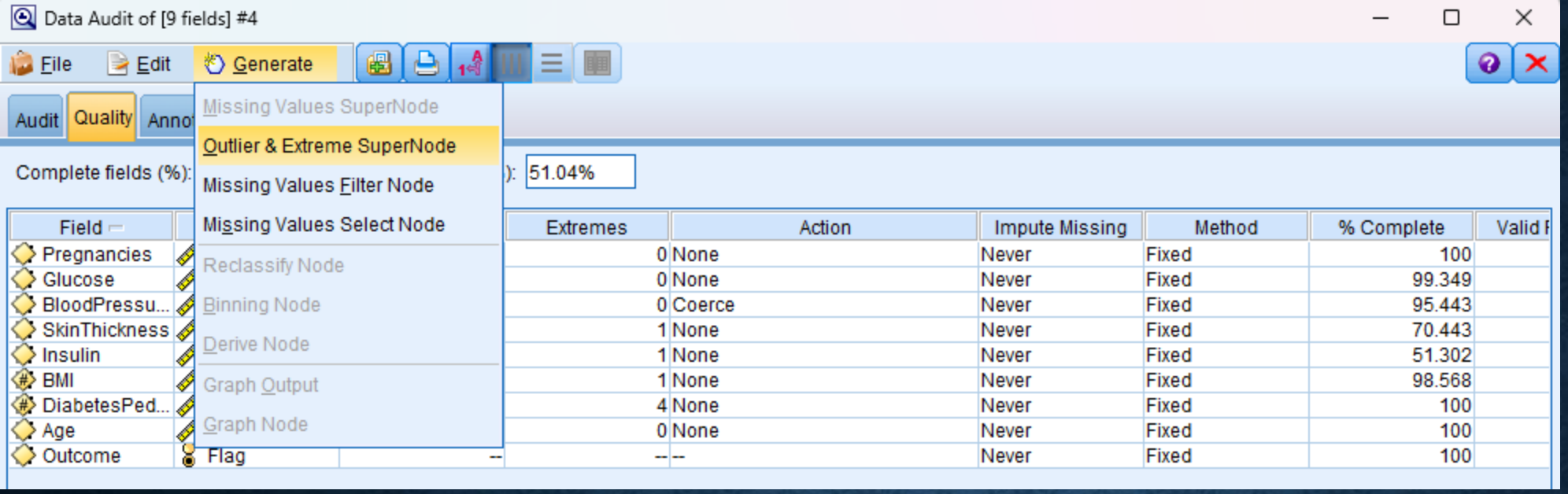
دیابت و فشار خون بالا باعث ایجاد یکدیگر نمی‌شوند، اما مبتلایان به دیابت معمولا مستعد ابتلا به بیماری‌های دیگری از جمله فشار خون بالا و کلسترول خون بالا هستند.

بعد از کلیک روی گزینه RUN در Sim Fit در تب Quality همان طور که مشاهده می کنید ، ما به تعداد 8 تا داده ی پرت در ستون فشار خون داریم و تعداد 8 تا در 768 مقداره مشخصی نمی شود و از آن جایی که تاثیر فشارخون بر دیابت کم است با توجه به مقاله ی پیش رو ؛ پس عمل Coerce (به معنای کم یا زیاد کردن داده ی پرت برای رساندن به حد بالا یا حد پایین داده ها) را انجام می دهیم.

Audit | Quality | Annotations

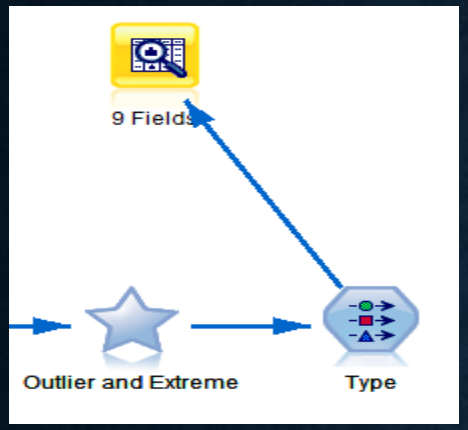
Complete fields (%): 44.44% Complete records (%): 51.04%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid f
Pregnancies	Continuous	4	0	None	Never	Fixed	100	
Glucose	Continuous	0	0	None	Never	Fixed	99.349	
BloodPressu...	Continuous	8	0	Coerce	Never	Fixed	95.443	
SkinThickness	Continuous	1	1	None	Never	Fixed	70.443	
Insulin	Continuous	7	1	None	Never	Fixed	51.302	
BMI	Continuous	3	1	None	Never	Fixed	98.568	
DiabetesPed...	Continuous	7	4	None	Never	Fixed	100	
Age	Continuous	5	0	None	Never	Fixed	100	
Outcome	Flag	--	--	--	Never	Fixed	100	

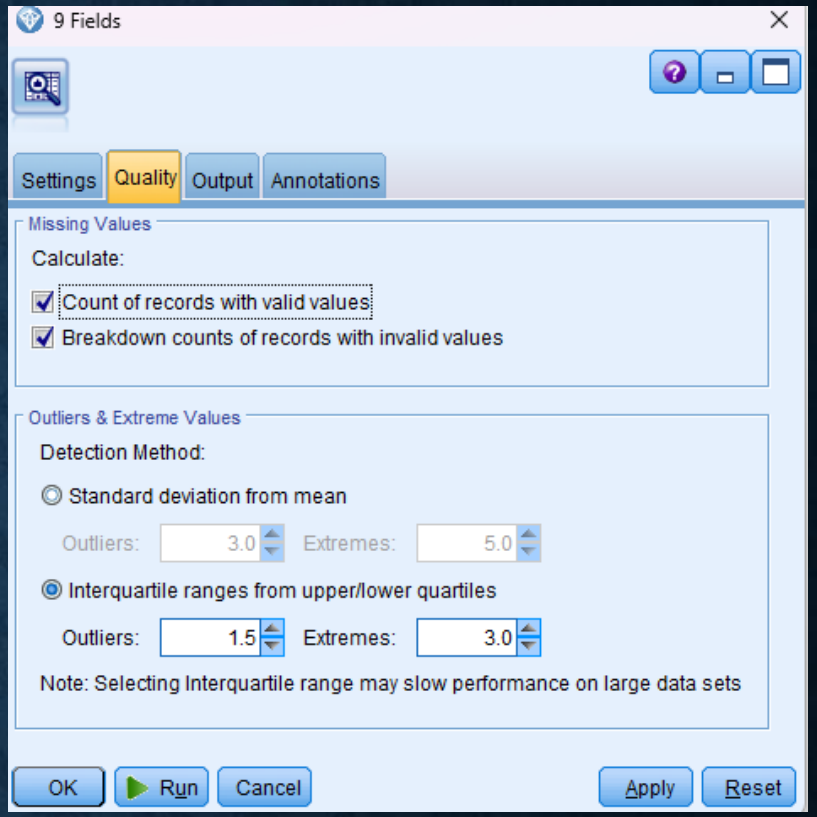


به منظور عمل انجام شده روی داده ها در این قسمت روی گزینه **Generate** کلیک کرده و **Outlier and Extreme SuperNode** را انتخاب می کنیم.

که به شکل ستاره ای در محیط **Steram** ما ظاهر خواهد شد؛ سپس به همانطور که اشاره کردیم **Type** دیگری بعد از اعمال تغییرات به مراحل کاری خود اضافه میکنیم.



بار دیگر در **Sim Fit** داخل تب **Quality** رفته و در قسمت **Outliers & Extreme Values** و این بار گزینه **Interquartile** را انتخاب می کنیم زیرا میخواهیم عملی روی داده های پرتی که داخل ستون های غیر فشار خون که دارای توزیع غیر نرمال بودند ؛ انجام دهیم.
BMI , Skin Thickness , Pregnancies ,)
(Insulin , Age



Data Audit of [9 fields] #5

File Edit Generate

Audit Quality Annotations

Complete fields (%): 44.44% Complete records (%): 51.04%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid F
Pregnancies	Continuous	4	0	Coerce	Never	Fixed	100	
Glucose	Continuous	0	0	None	Never	Fixed	99.349	
BloodPressu...	Continuous	14	0	None	Never	Fixed	95.443	
SkinThickness	Continuous	2	1	Coerce	Never	Fixed	70.443	
Insulin	Continuous	16	8	Nullify	Never	Fixed	51.302	
BMI	Continuous	7	1	Coerce	Never	Fixed	98.568	
DiabetesPed...	Continuous	23	6	Nullify	Never	Fixed	100	
Age	Continuous	9	0	None	Never	Fixed	100	
Outcome	Flag	--	--	--	Never	Fixed	100	

در اینجا تمامی داده هایی که تعداد پرت آن ها پایین بوده و تاثیر ناچیزی در دیابت دارند را **Coerce** کردیم و ستون هایی که ممکن است تاثیر بالایی در دیابت داشته باشد را **Nullify** کردیم همچنین وقتی تعداد مفقودی داده بالاست ممکن است داده ی پرت اشتباه باشد.

Data Audit of [9 fields] #7

File Edit Generate

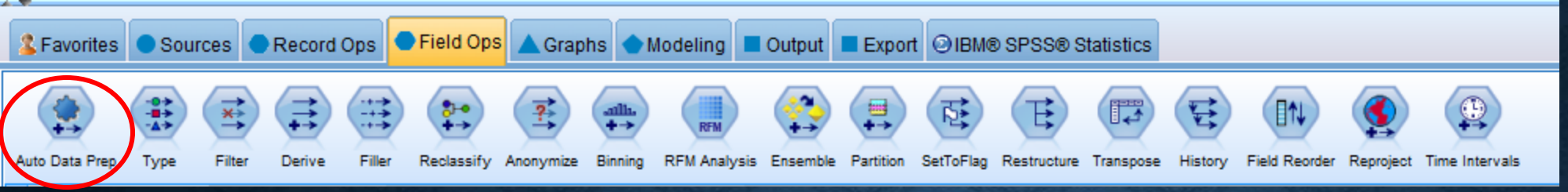
Audit Quality Annotations

Complete fields (%): 33.33% Complete records (%): 45.96%

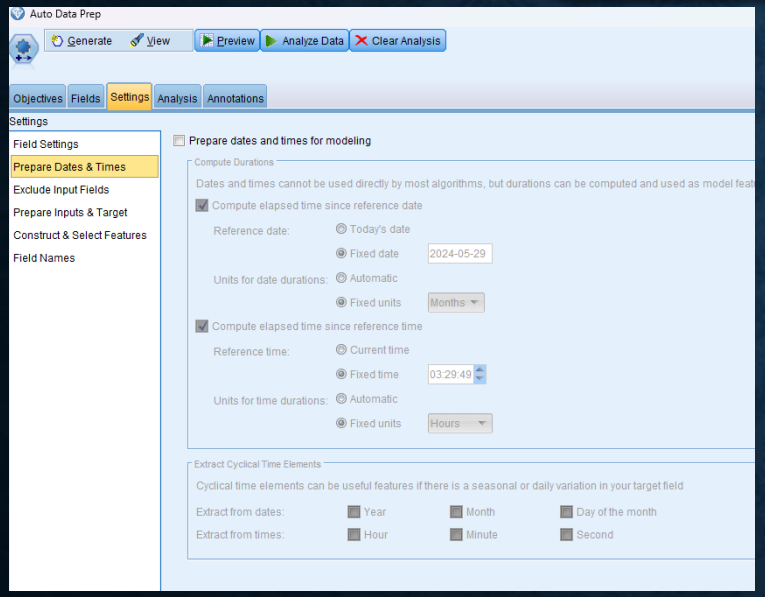
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid F
Insulin	Continuous	3	0	None	Blank & Null Val...	Algorithm	48.177	
SkinThickness	Continuous	0	0	None	Blank & Null Val...	Algorithm	70.443	
BloodPressu...	Continuous	14	0	None	Blank & Null Val...	Random	95.443	
DiabetesPed...	Continuous	12	0	None	Blank & Null Val...	Random	96.224	
BMI	Continuous	0	0	None	Blank & Null Val...	Fixed	98.568	
Glucose	Continuous	0	0	None	Blank & Null Val...	Fixed	99.349	
Pregnancies	Continuous	0	0	None	Never	Fixed	100	
Age	Continuous	9	0	None	Never	Fixed	100	
Outcome	Flag	--	--	--	Never	Fixed	100	

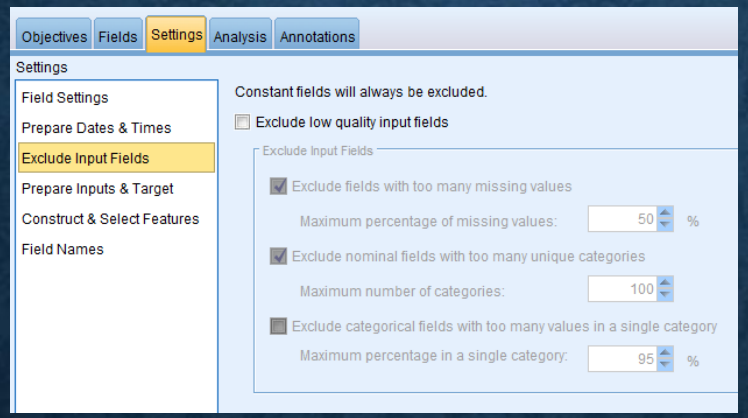
نکته : پس از بررسی دقیق روی داده ها و تاثیرشان بر روی هدف اگر تعداد مفقودی در ستونی بالا بود اما آن ستون بی تاثیر؛ حدالمکان از روش Algorithm استفاده نمی کنیم.

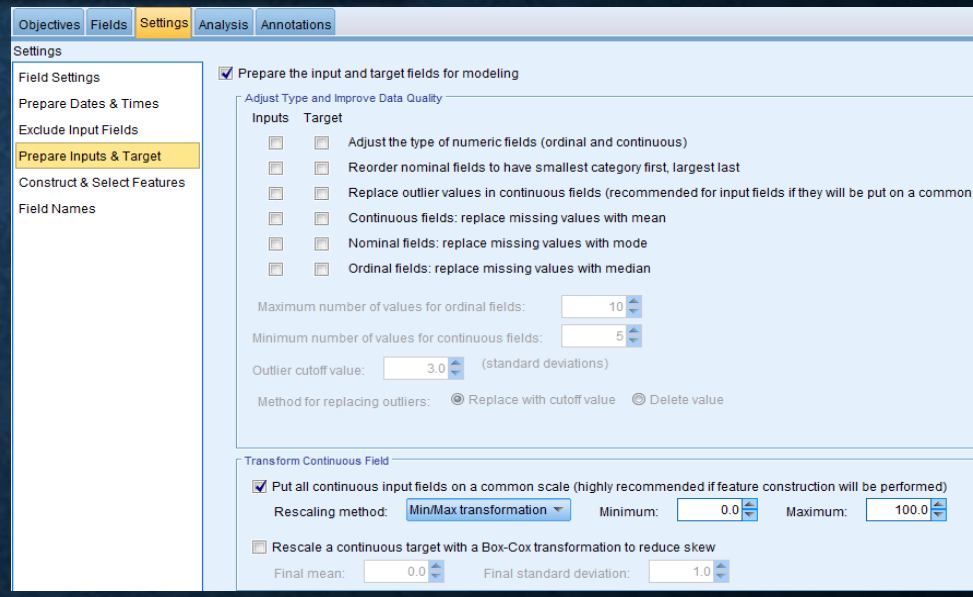
به منظور پیدا کردن داده های مفقودی در ستون های نمایش داده شده از روش Algorithm برای ستون هایی که داده های پرت بالایی داشته و فرض می شود که تاثیر بالایی در مدل داشته باشند استفاده می شود؛ از روش Random نرمال برای پیدا کردن اعداد مفقودی ستون هایی که دصد مفقودیشان متوسط است برای اینکه توزیع به هم نخورد و از میانگین برای ستون هایی که درصد مفقودی آن ها کم است استفاده می کنیم.



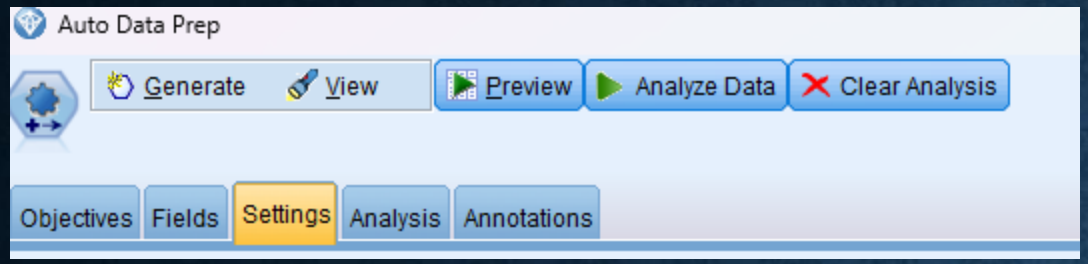
از type بعد از هر تغییری در داده ها استفاده میکنیم؛
سپس به سراغ Auto Data Prep برای اسکیل (Scale) داده ها می رویم، تا ببینیم داده های پرت را با چه روشی شناسایی کنیم.







تیک ها را مطابق صفحات بالا برداشته زیرا در تب **Date & Time** ما داده زمانی نداریم،
در تب **Prep Inputs & Target** در قسمت **Exclude low quality input fields** شامل داده های ما نمی شود یا ما در مراحل قبل درستشون کردیم ؛ در قسمت **Improve Data Quality** تمامی داده ها در قسمت های قبل فیکس شده؛ در قسمت **Transform Continuous Field** برای اسکیل کردن داده ها از **Min/Max** استفاده می کنیم زیرا داده ها رگرسیونی نمی باشند.



روی **Analyze Data** کلیک کرده، سپس بعد از کلیک رو **Preview** مشاهده می کنیم که تمامی اعداد ما بین 0 تا 100 تقسیم شده و داده ها آماده ی مدل سازی می باشند.

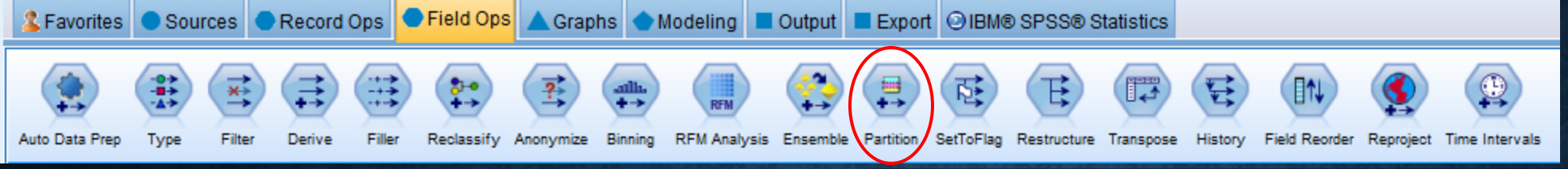
Preview from Auto Data Prep Node (9 fields, 10 records) #1

File Edit Generate

Table Annotations

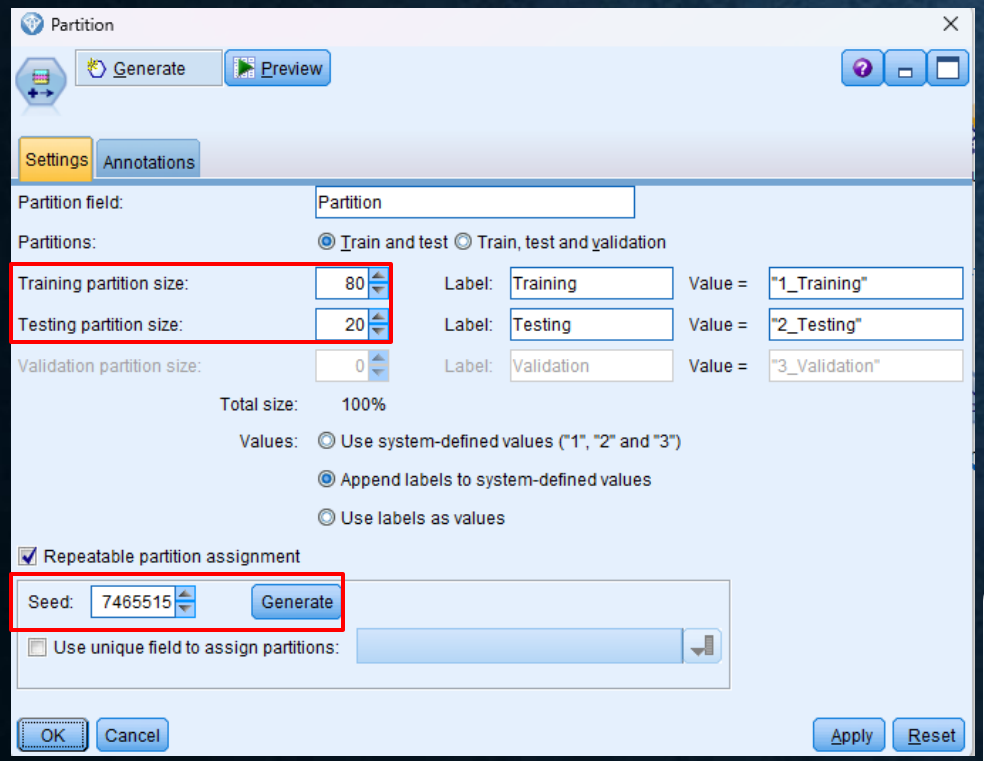
	Outcome	Pregnancies_transformed	Glucose_transformed	BloodPressure_transformed	SkinThickness_transformed	Insulin_transformed	BMI_transformed	DiabetesPedigreeFunction_transformed	Age_transformed
1	1	44.444	74.372	49.455	56.000	48.555	66.866	49.326	48.333
2	0	7.407	42.714	41.378	44.000	12.717	52.935	24.528	16.667
3	1	59.259	91.960	38.686	17.333	76.590	46.368	53.369	18.333
4	0	7.407	44.724	41.378	32.000	23.121	55.920	7.996	0.000
5	1	0.000	68.844	6.382	56.000	44.509	85.771	54.352	20.000
6	0	37.037	58.291	52.147	34.326	35.260	50.945	11.051	15.000
7	1	22.222	39.196	19.842	50.000	21.387	61.692	15.274	8.333
8	0	74.074	57.789	51.267	49.000	35.260	70.249	5.031	13.333
9	1	14.815	98.995	46.763	76.000	76.590	60.697	7.188	53.333
10	1	59.259	62.814	81.759	62.286	35.260	0.000	13.836	55.000

OK



برای مدل سازی می بایست تعدادی از داده ها رو به صورت **Train** و تعدادی را به صورت تست قرار می دهیم.
که اینکار را باید **Partition** انجام داد.

- همانطور که در قسمت پایین مشاهده می کنید ، توزیع داده ها را به صورت **Train 80%** و **Test 20%** قرار می دهیم.
- **Generate Seed** برای تست یکسان بودن عملکرد مدل استفاده می شود که در ادامه به آن می پردازیم.

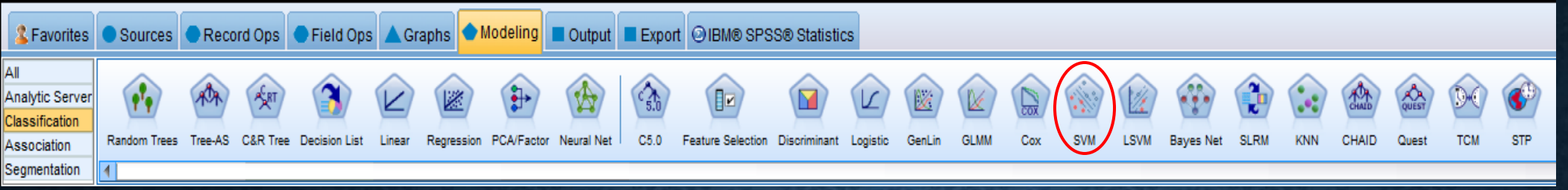


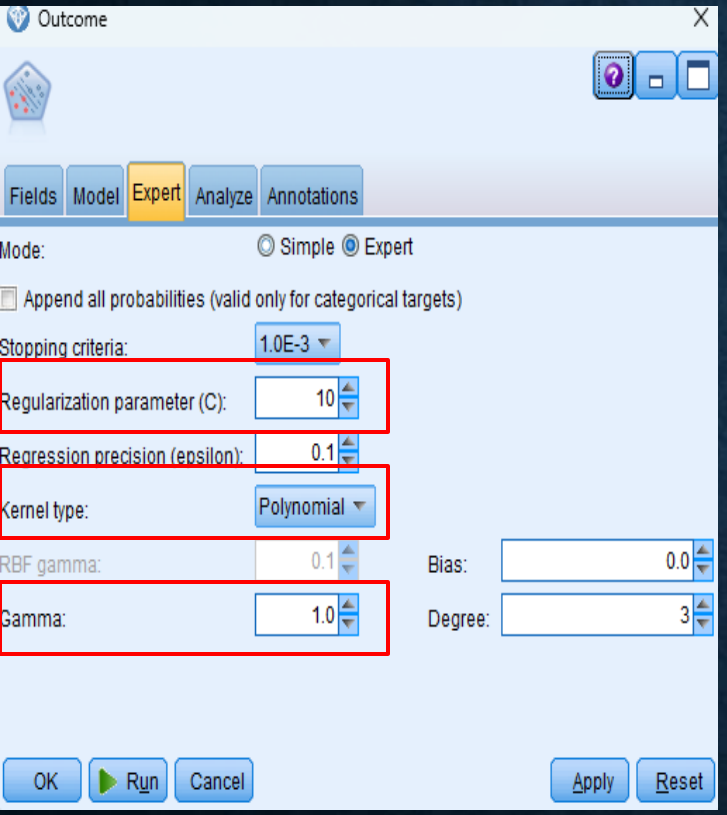
مطابق **Table** مقابل **80%** داده ها به صورت **Train** و **20%** به صورت **Test** پارتیشن بندی شدند.

	Outcome	Pregnancies_transformed	Glucose_transformed	BloodPressure_transformed	SkinThickness_transformed	Insulin_transformed	BMI_transformed	DiabetesPedigreeFunction_transformed	Age_transformed	Partition
1	1	44.444	74.372	49.455	56.000	48.555	66.866	49.326	48.333	1_Training
2	0	7.407	42.714	41.378	44.000	12.717	52.935	24.528	16.667	2_Testing
3	1	59.259	91.960	38.686	17.333	76.590	46.368	53.369	18.333	1_Training
4	0	7.407	44.724	41.378	32.000	23.121	55.920	7.996	0.000	1_Training
5	1	0.000	68.844	6.382	56.000	44.509	85.771	52.847	20.000	1_Training
6	0	37.037	58.291	52.147	34.326	35.260	50.945	11.051	15.000	1_Training
7	1	22.222	39.196	19.842	50.000	21.387	61.692	15.274	8.333	1_Training
8	0	74.074	57.789	87.574	49.000	35.260	70.249	5.031	13.333	1_Training
9	1	14.815	98.995	46.763	76.000	76.590	60.697	7.188	53.333	1_Training
10	1	59.259	62.814	81.759	62.286	35.260	0.000	13.836	55.000	1_Training
11	0	29.630	55.276	76.375	62.286	35.260	74.826	10.153	15.000	1_Training
12	1	74.074	84.422	52.147	62.286	48.555	75.622	41.240	21.667	1_Training
13	0	74.074	69.849	60.223	34.326	48.555	53.930	4.989	60.000	2_Testing
14	1	7.407	94.975	33.302	32.000	52.601	59.900	28.751	63.333	1_Training
15	1	37.037	83.417	49.455	24.000	46.532	51.343	45.732	50.000	1_Training
16	1	51.852	50.251	20.169	34.326	22.543	59.701	36.478	18.333	1_Training
17	1	0.000	59.296	65.607	80.000	62.428	91.144	42.498	16.667	1_Training
18	1	51.852	53.769	52.147	34.326	35.260	58.905	15.813	16.667	1_Training
19	0	7.407	51.759	-0.000	62.000	19.942	86.169	9.434	20.000	1_Training
20	1	7.407	57.789	46.763	46.000	23.699	68.856	40.521	18.333	1_Training
21	0	22.222	63.317	70.991	68.000	63.873	78.209	56.244	10.000	1_Training
22	0	59.259	49.749	65.607	49.000	12.717	70.448	27.853	48.333	2_Testing
23	1	51.852	98.492	73.683	62.286	76.590	79.204	33.513	33.333	2_Testing
24	1	66.667	59.799	60.223	56.000	35.260	57.711	16.622	13.333	1_Training
25	1	81.481	71.859	79.067	52.000	38.150	72.836	15.813	50.000	1_Training
26	1	74.074	62.814	46.763	38.000	29.191	61.891	11.411	33.333	2_Testing
27	1	51.852	73.869	54.839	62.286	48.555	78.408	16.083	36.667	1_Training
28	0	7.407	48.744	41.378	16.000	36.416	46.169	36.748	1.667	1_Training
29	0	96.296	72.864	62.915	24.000	27.746	44.179	15.004	60.000	1_Training
30	0	37.037	58.794	76.375	49.000	35.260	67.861	23.270	28.333	1_Training
31	0	37.037	54.774	53.493	38.000	35.260	71.642	42.049	65.000	1_Training
32	1	22.222	79.397	54.839	58.000	66.763	62.886	69.452	11.667	2_Testing
33	0	22.222	44.221	30.610	8.000	11.561	49.353	16.981	1.667	1_Training
34	0	44.444	46.231	76.375	17.333	12.717	39.602	9.883	11.667	2_Testing
35	0	74.074	61.307	57.531	48.000	35.260	54.925	38.994	40.000	1_Training
36	0	29.630	51.759	33.302	52.000	51.445	47.761	79.784	20.000	1_Training
37	0	81.481	69.347	54.839	49.000	48.555	66.070	30.728	23.333	1_Training
38	1	66.667	51.256	54.839	60.000	35.260	65.473	52.740	41.667	1_Training
39	1	14.815	45.226	44.071	70.000	22.543	76.020	38.185	10.000	1_Training
40	1	29.630	55.779	49.455	80.000	55.780	73.831	25.190	58.333	1_Training
41	0	22.222	90.452	38.686	36.000	16.185	67.662	17.341	8.333	1_Training
42	0	51.852	66.834	65.607	62.286	48.555	80.000	55.526	26.667	1_Training
43	0	51.852	53.266	76.375	22.000	35.260	45.174	14.106	45.000	2_Testing
44	1	66.667	85.930	100.000	34.000	65.318	90.348	57.772	55.000	1_Training
45	0	51.852	79.899	38.686	34.326	48.555	54.527	19.407	31.667	1_Training
46	1	0.000	90.452	41.378	64.000	48.555	83.582	43.150	6.667	1_Training
47	0	7.407	73.367	27.918	34.326	48.555	59.104	43.666	13.333	1_Training
48	0	14.815	35.678	46.763	40.000	14.451	55.721	45.642	1.667	1_Training
49	1	51.852	51.759	41.378	50.000	35.260	77.811	23.899	16.667	1_Training
50	0	51.852	52.764	43.531	34.326	35.260	0.000	20.395	5.000	1_Training
51	0	7.407	51.759	60.223	8.000	19.653	38.607	37.107	1.667	1_Training
52	0	7.407	50.754	19.842	16.000	6.358	48.159	40.252	8.333	2_Testing
53	0	37.037	44.221	41.378	28.000	2.601	48.557	23.720	15.000	1_Training
54	1	59.259	88.442	73.683	54.000	82.659	67.065	34.951	61.667	1_Training
55	0	51.852	75.377	41.378	70.000	94.798	69.055	57.502	35.000	1_Training
56	0	7.407	36.683	19.842	6.000	19.364	45.771	15.274	0.000	1_Training
57	1	51.852	93.970	44.071	64.000	83.815	75.025	15.813	33.333	2_Testing
58	0	0.000	50.251	70.991	100.000	27.746	93.134	79.425	16.667	1_Training
59	0	0.000	73.367	62.915	62.286	48.555	80.597	-1.866	38.333	1_Training
60	0	0.000	52.764	38.686	68.000	36.994	82.587	8.535	1.667	1_Training
61	1	14.815	42.211	44.948	34.326	12.717	0.000	20.305	0.000	1_Training
62	1	59.259	66.834	49.455	49.000	48.555	65.473	17.251	30.000	1_Training
63	0	37.037	22.111	35.994	34.326	14.451	49.751	45.732	25.000	1_Training
64	0	14.815	70.854	30.610	54.000	32.948	50.547	55.795	5.000	1_Training
65	1	51.852	57.286	41.378	49.000	35.260	65.274	16.173	35.000	1_Training
66	0	37.037	49.749	52.147	40.000	12.717	57.711	11.231	18.333	2_Testing
67	1	0.000	54.774	70.991	46.000	35.260	64.677	69.811	28.333	1_Training
68	0	14.815	54.774	76.375	62.286	35.260	84.975	68.913	55.000	1_Training
69	0	7.407	47.739	41.378	12.000	6.936	39.005	23.001	6.667	2_Testing
70	0	29.630	73.367	66.953	40.000	24.855	57.512	9.973	10.000	1_Training
71	1	14.815	50.251	41.378	26.000	21.965	65.473	70.889	11.667	1_Training
72	0	37.037	69.849	38.686	56.000	36.416	56.915	29.919	8.333	1_Training
73	1	96.296	63.317	73.683	62.286	35.260	86.368	45.373	35.000	2_Testing
74	0	29.630	64.824	68.299	26.000	73.988	69.851	13.747	3.333	1_Training
75	0	7.407	39.698	53.493	46.000	12.717	63.682	28.571	1.667	1_Training
76	0	7.407	0.000	17.150	26.000	35.260	49.154	5.571	1.667	2_Testing
77	0	51.852	31.156	57.531	49.000	12.717	64.876	28.122	33.333	1_Training

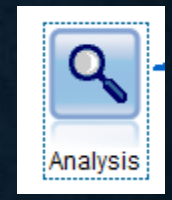
In the real project, we separate the Train and Test data and perform the relevant operations on the Train, then finally we measure our performance with the Test data.

SVM





- در این مدل با رفتن در تب **Expert** و انتخاب **Expert** مقدار **Gamma** و **C** و **Kernel type** را تغییر دهیم برای رسیدن به دقت بالاتر در مدل سازی داده ها
- پس از تعیین آن ها مدل را **RUN** کرده تا دقت را بسنجیم. اینکار را با استفاده از **Analysis** در تب **Output** می توانیم انجام دهیم.



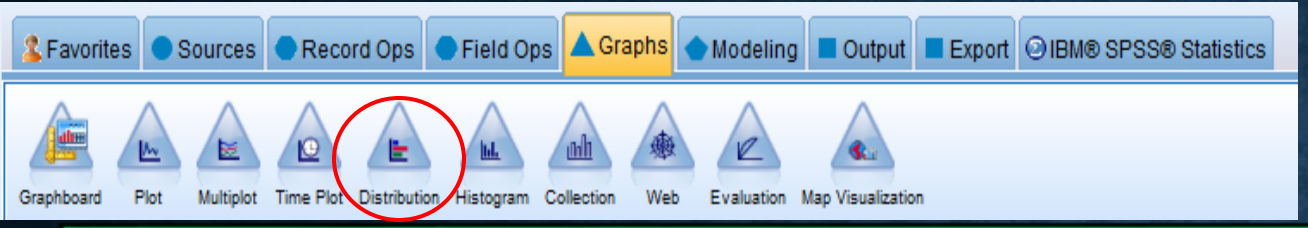
'Partition'	1_Training		2_Testing	
Correct	333	76.91%	122	86.52%
Wrong	100	23.09%	19	13.48%
Total	433		141	

Coincidence Matrix for $S-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	185	28
1	72	148

'Partition' = 2_Testing	0	1
0	82	11
1	8	40

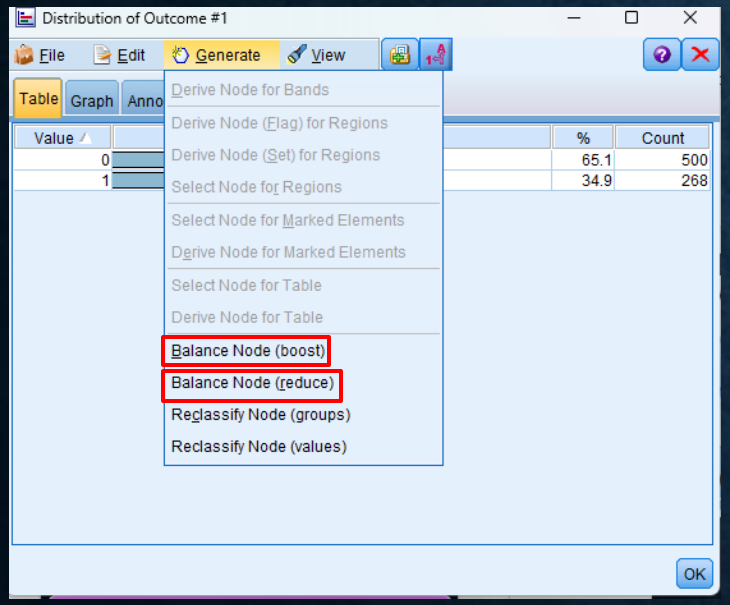
- همانطور که مشاهده میکنید؛ دقت در مدل **SVM** در پارتیشن **Train 76%** و در پارتیشن **Test 86%** است. که به علت بالاتر بودن اختلاف آن ها بیشتر از **5%** مدل **Overfit** می باشد.



Value	Proportion	%	Count
0		65.1	500
1		34.9	268

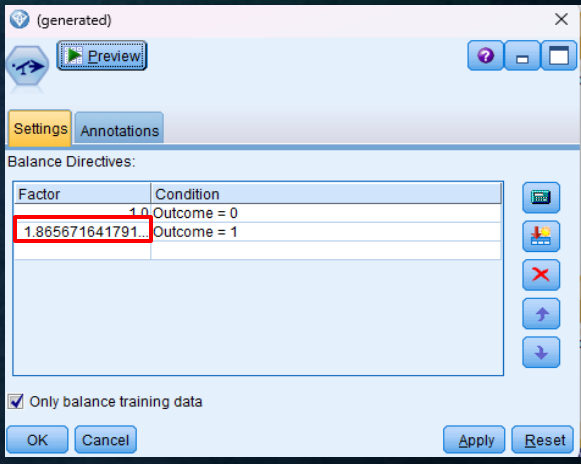
❁ برای اینکه بفهمیم آیا داده های ما **Unbalance** هستند یا خیر؛ از تب **Graph** نمودار **Distribution** را انتخاب کرده و به **Partition** وصلش میکنیم و اجراش میکنیم.

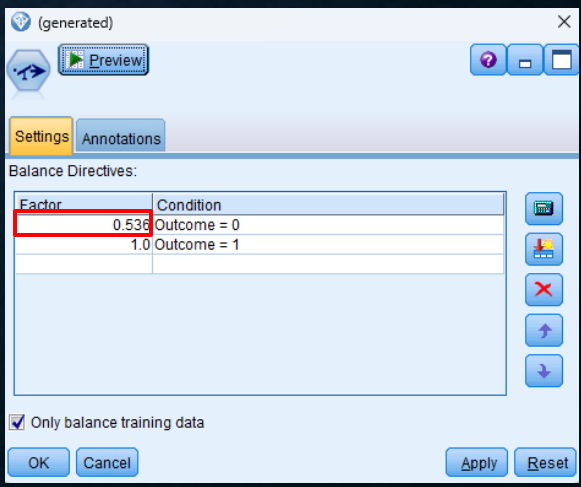
❁ همانطور که مشاهده میکنیم داده ها براساس خروجی مربوطه **Unbalance**هستند.



برای مدیریت **Imbalance Data** در همان نمودار **Distribution** تب **Generate** را انخاب کرده و گزینه های **boost** و **reduce** را انتخاب میکنیم. (**Oversampling** , **Undersampling**)

دوتا از آیکون های پشتی را به ما می دهد که روی هرکدام کلیک کنیم ، جداول رو به رو را به ما نشان خواهد داد.

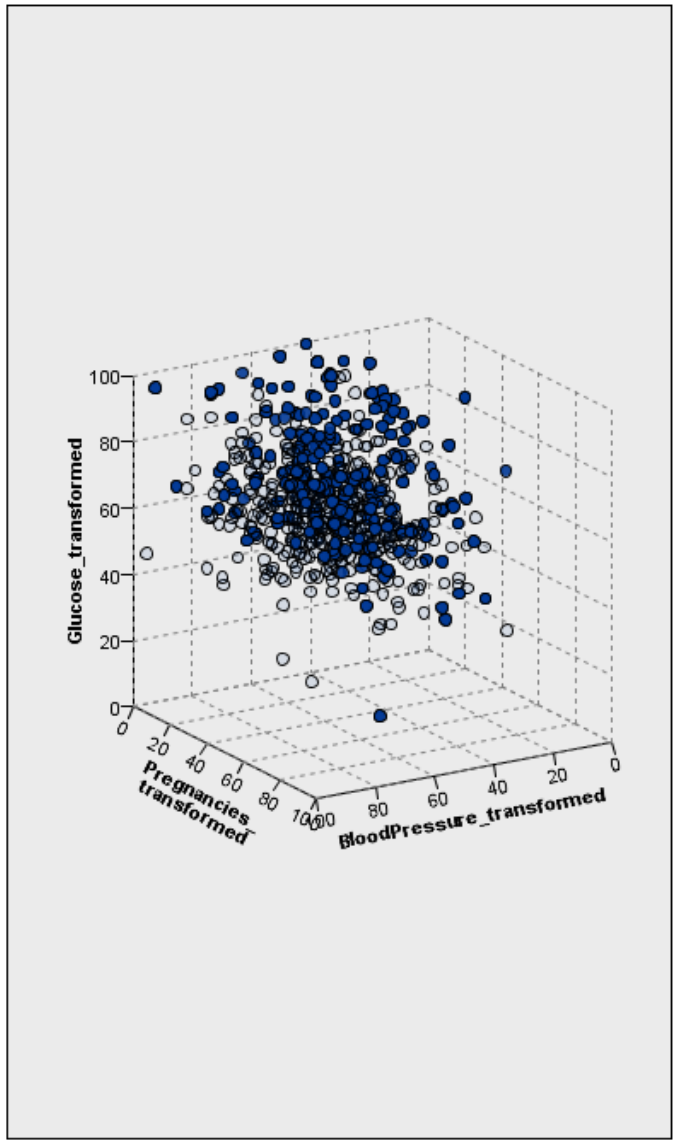




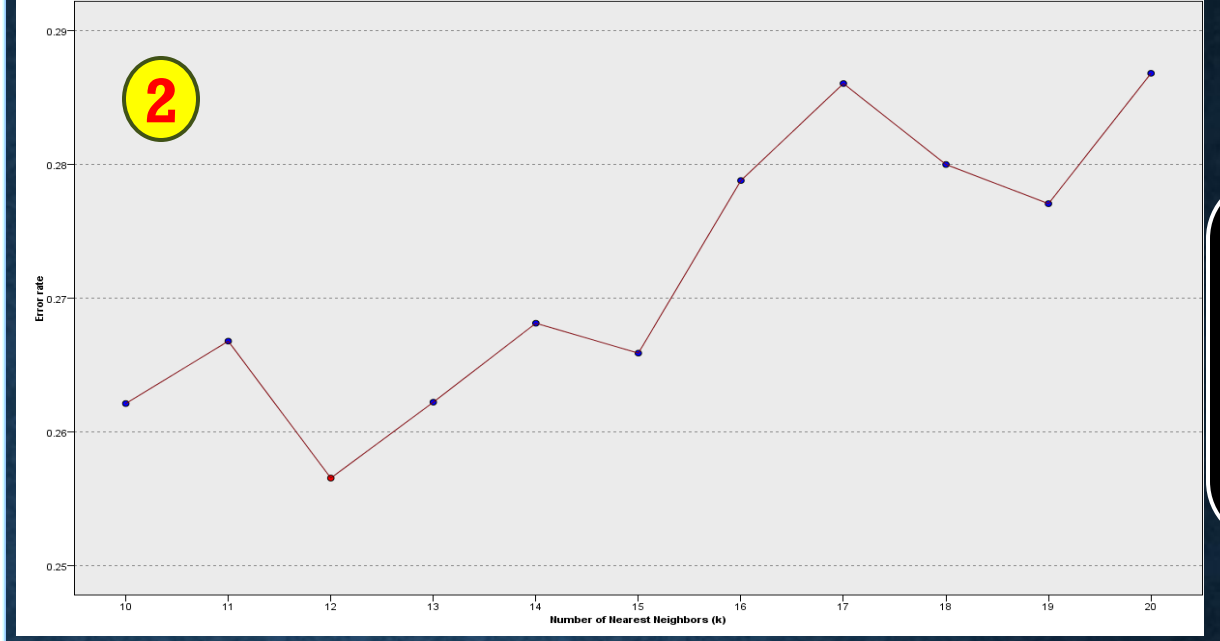
❁ در **Oversampling** به داده ای که فراوانی بالاتر دارد نسبت **1** داده می شود و داده ای که فراوانی کمتر دارد در نسبت اشاره شده در قسمت **Factor** ضرب می شود.

❁ در **Undersampling** به داده ای که فراوانی کمتری دارد نسبت **1** داده می شود و داده ای که فراوانی بیشتری را دارا می باشد در نسبت اشاره شده ضرب می شود.

1



2



KNN

لازم به ذکر است زین پس برای جلوگیری از داده های نابالانس از Reduce و Boost هم به صورت آزمون خطا استفاده می شود.

3

1- مدل سه بعدی KNN برای پیش بینی داده های جدید

2 – انتخاب K12 به عنوانه بهترین K به علت اینکه کمترین Error را دارد.

3 – میزان دقت داده ها با استفاده از Classification Table در روش KNN

4 – دقت داده های Train و Test با استفاده از Analysis Data

دقت نسبتا مناسب ولی تقریبا Overfit است

Classification Table

Partition	Observed	Predicted		
		1.000	0.000	Percent Correct
Training	1.000	284	94	75.1%
	0.000	105	301	74.1%
	Overall Percent	49.6%	50.4%	74.6%

Comparing $KNN-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	478	76.24%	116	82.27%
Wrong	149	23.76%	25	17.73%
Total	627		141	

Coincidence Matrix for $KNN-Outcome (rows show actuals)

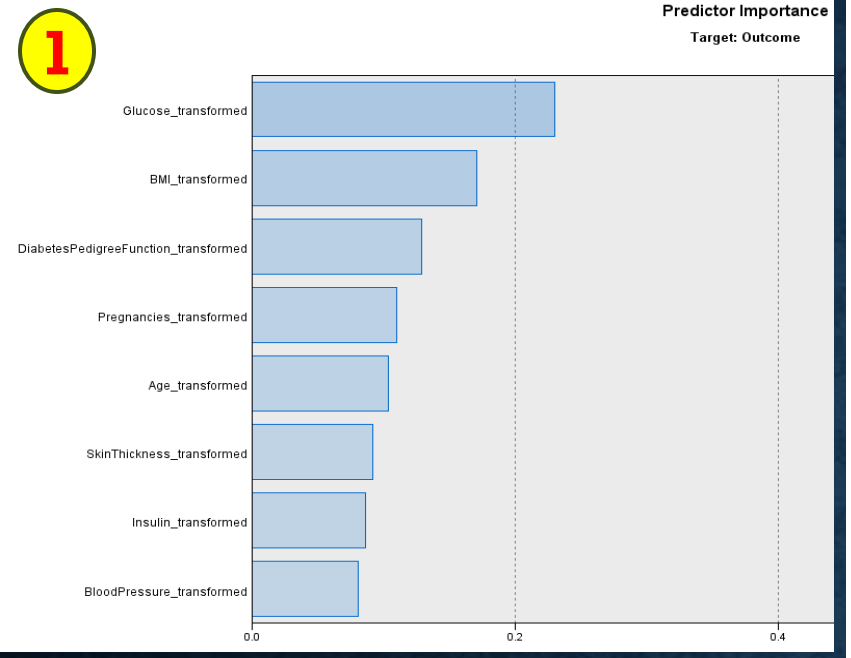
'Partition' = 1_Training	0	1
0	320	87
1	62	158

'Partition' = 2_Testing	0	1
0	77	16
1	9	39

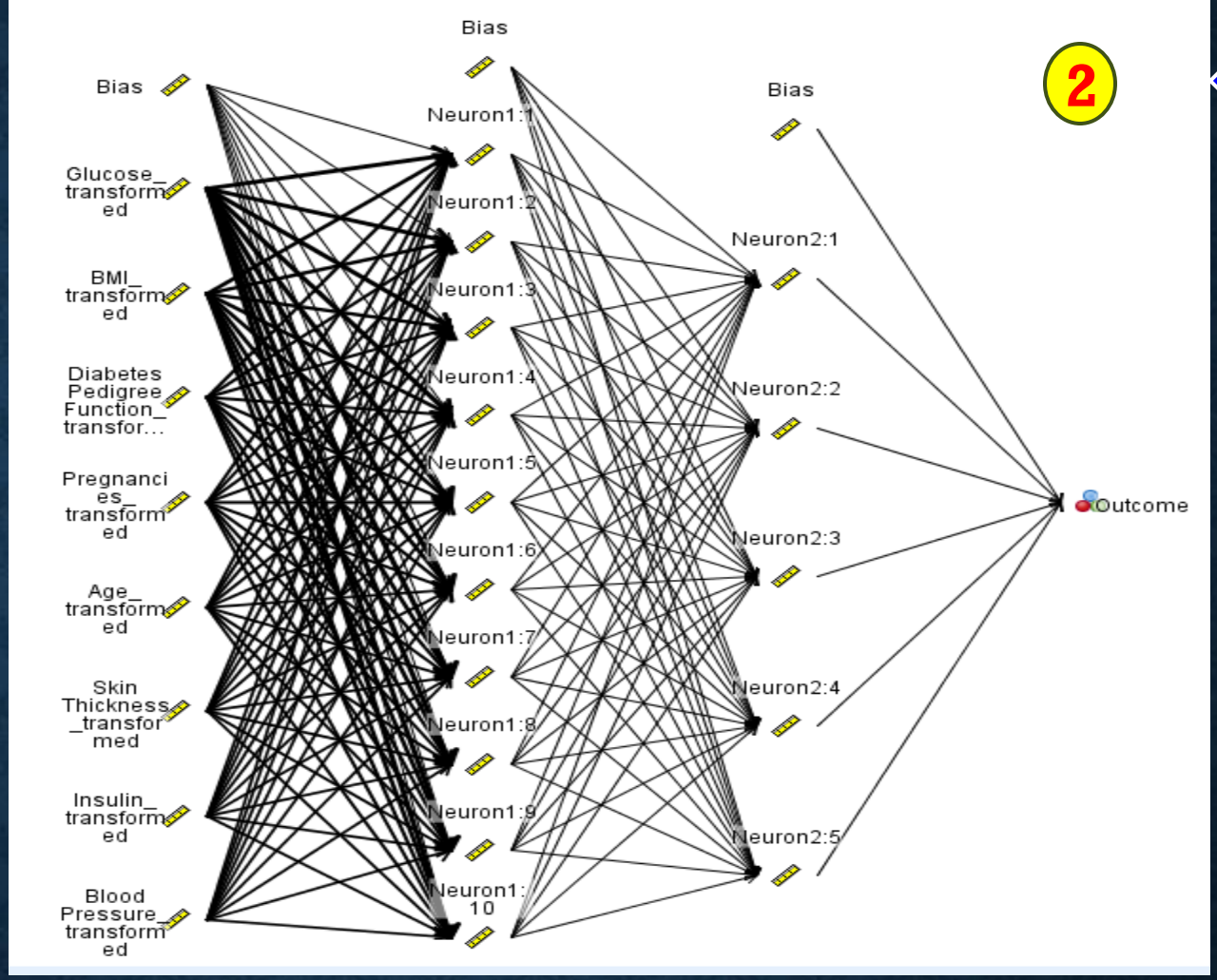
4

Neural Network

1



2



3

Results for output field Outcome

Comparing $N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	332	78.12%	112	79.43%
Wrong	93	21.88%	29	20.57%
Total	425		141	

Coincidence Matrix for $N-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	161	44
1	49	171

'Partition' = 2_Testing	0	1
0	70	23
1	6	42

1 – میزان تاثیر پارامترهای مختلف روی **Target** که می بایست از دید بیزینسی هم همخوانی لازمه را داشته باشد.

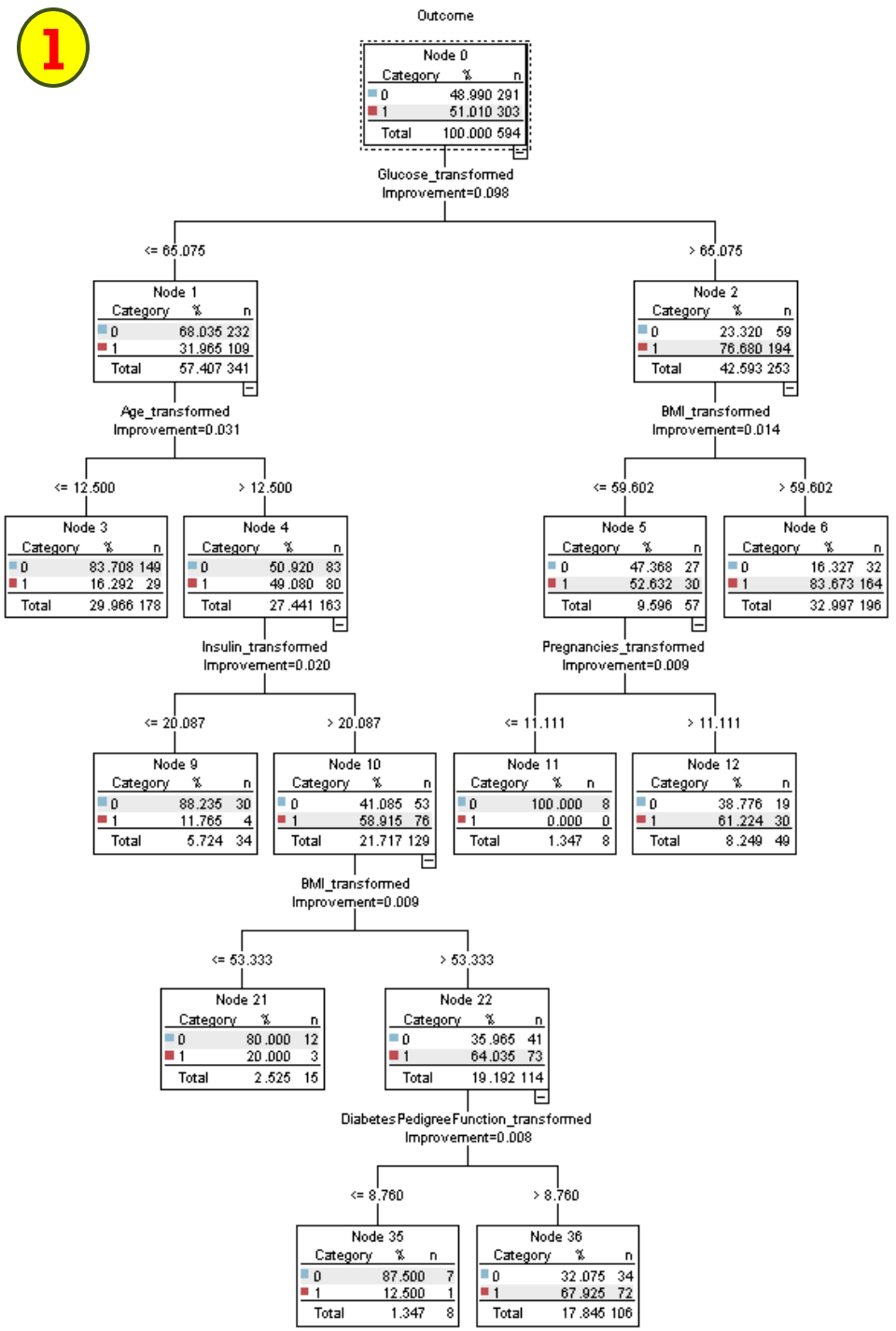
2 – مدل شبکه عصبی که 10 لایه پنهان در مرحله اول و 5 لایه پنهان در مرحله دوم را دارا می باشد.

3 – میزان دقت روش شبکه عصبی که میزان دقت مدل را با استفاده از **Analysis Data** بررسی می کند.

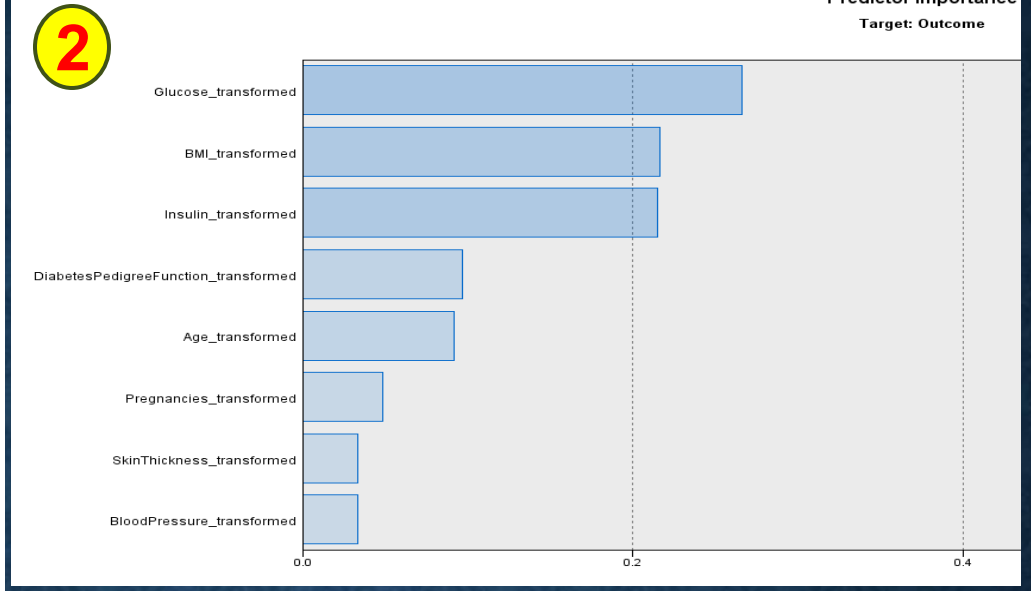
مدل نسبتا مناسب و Overfit یا Underfit نمی باشد.

CRT

1



2



2 – میزان تاثیر پارامتر های مختلف روی Outcome در روش CRT

Collapse All Expand All

Results for output field Outcome

Comparing $R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	470	77.81%	123	75%
Wrong	134	22.19%	41	25%
Total	604		164	

Coincidence Matrix for $R-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	294	107
1	27	176
'Partition' = 2_Testing	**0**	**1**
0	71	28
1	13	52

3

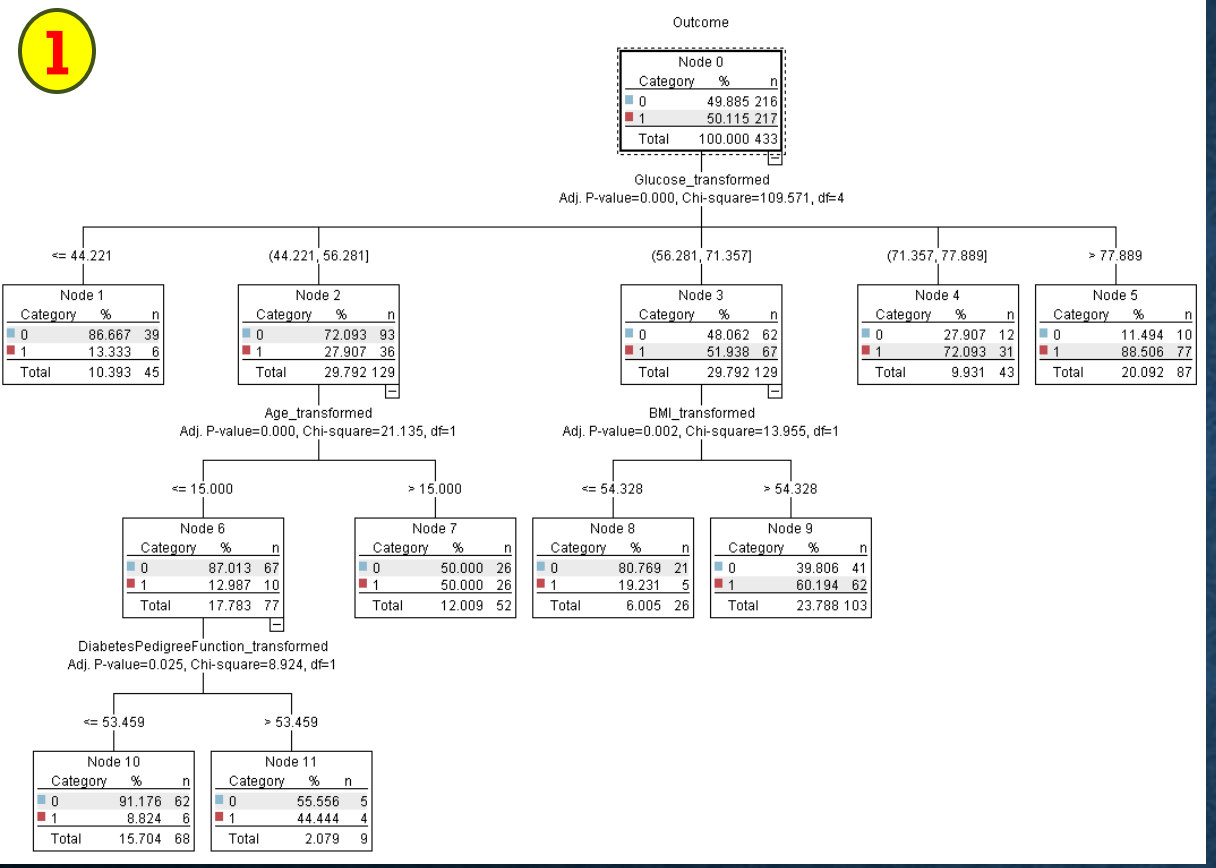
3 – میزان دقت داده های Train و Test در مدل CRT

1 – مدل درخت CRT با عمق 10 با 2% شاخه پدر

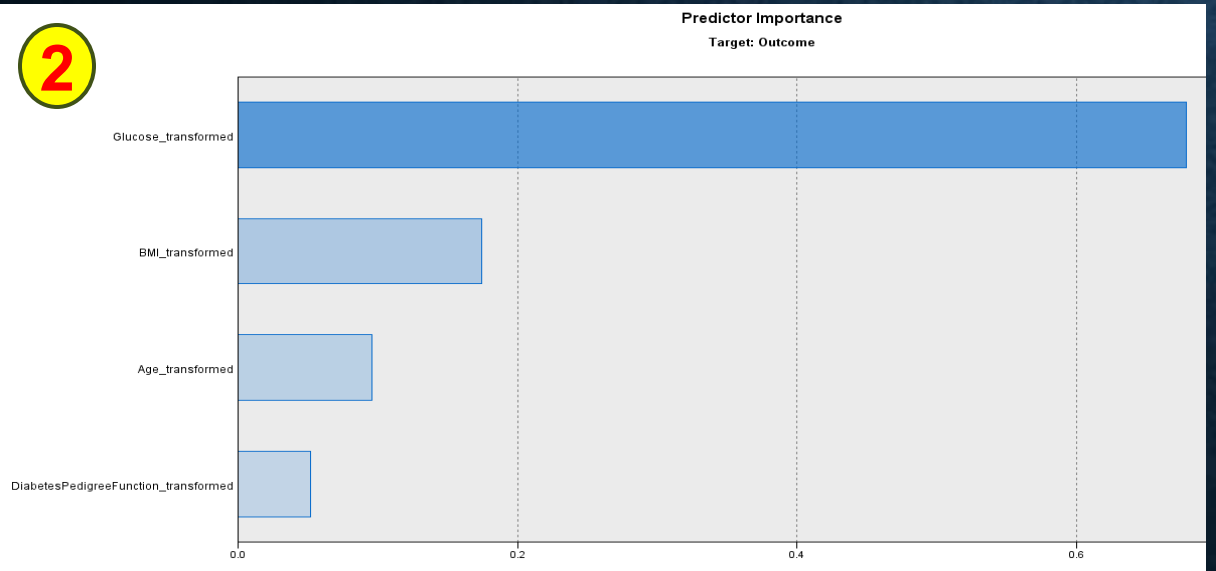
مدل با درصد پایینی مناسب و **Overfit** یا **Underfit** نمی باشد.

CHAID

1



2



Results for output field Outcome

Comparing $R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	439	72.68%	118	71.95%
Wrong	165	27.32%	46	28.05%
Total	604		164	

Coincidence Matrix for $R-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	280	121
1	44	159

'Partition' = 2_Testing	0	1
0	66	33
1	13	52

3

1 – مدل درخت **Chaid** با عمق 8 و 2% شاخه پدری

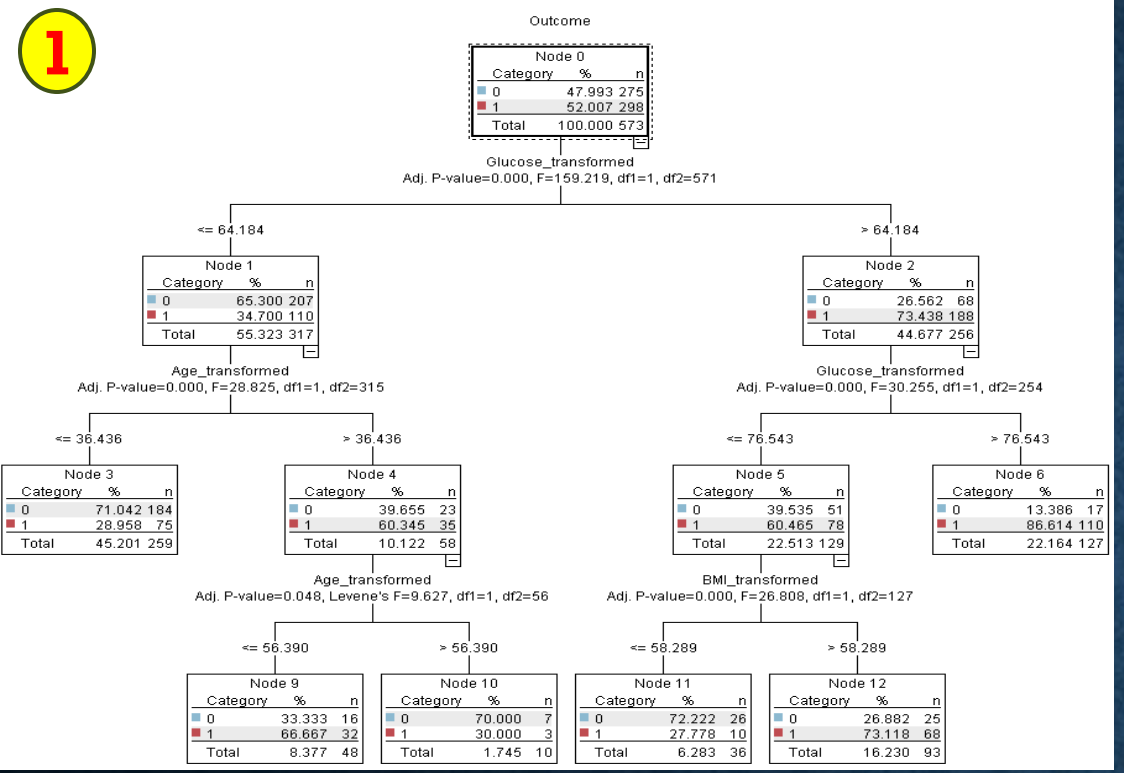
2 – میزان تاثیر پارامترهای مختلف روی **Outcome** در روش **Chaid**

3 – میزان دقت داده های **Train** و **Test** در روش **Chaid** با **Analysis Data**

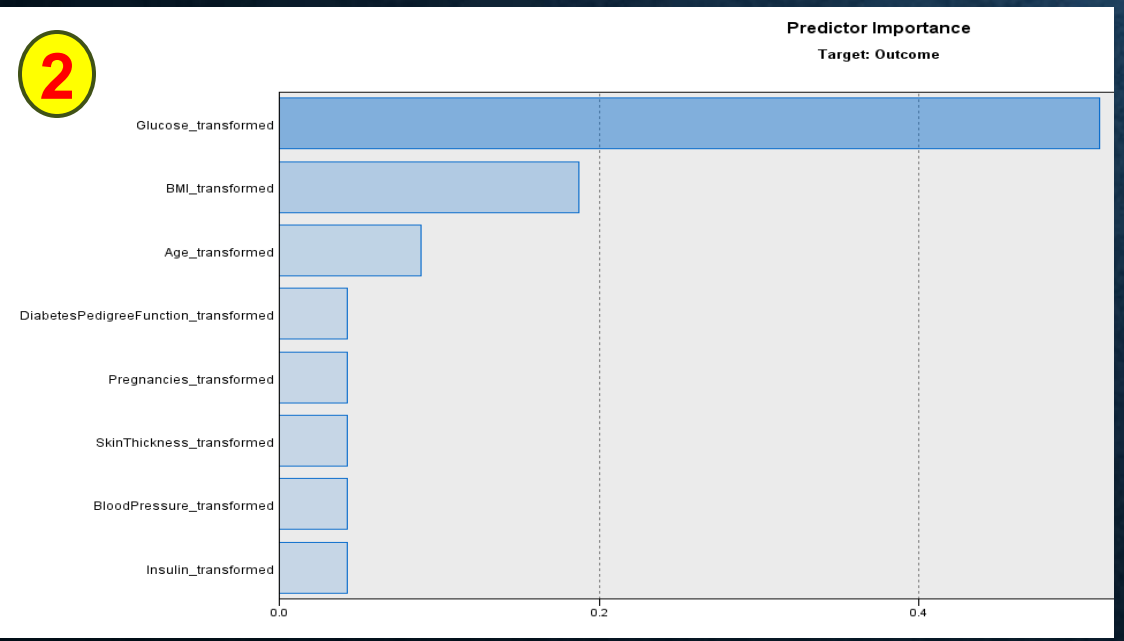
با توجه به اینکه تاثیر گلوکز در این روش بالا در نظر گرفته شده است و تاثیر بقیه پارامترها تقریبا ناچیز است ؛ همچنین میزان دقت داده های **Train** و **Test** تقریبا **Underfit** است، مدل قابل اتکا نمی باشد.

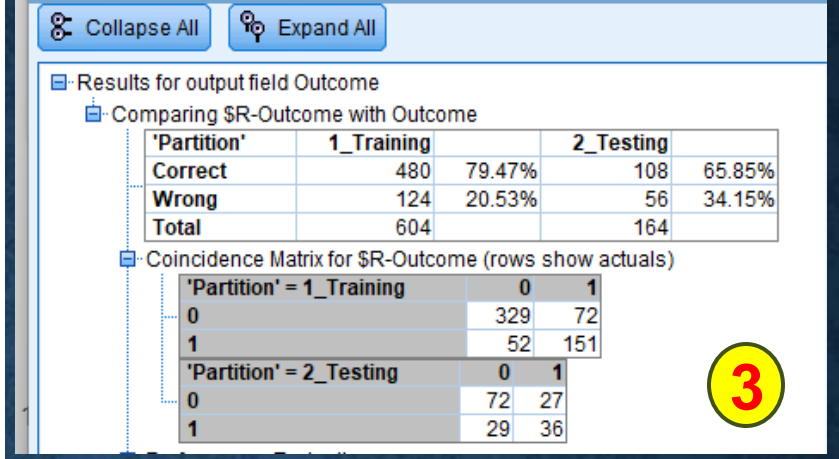
QUEST

1



2





3

1 – مدل درخت Quest با عمق 7 و 2% شاخه پدری

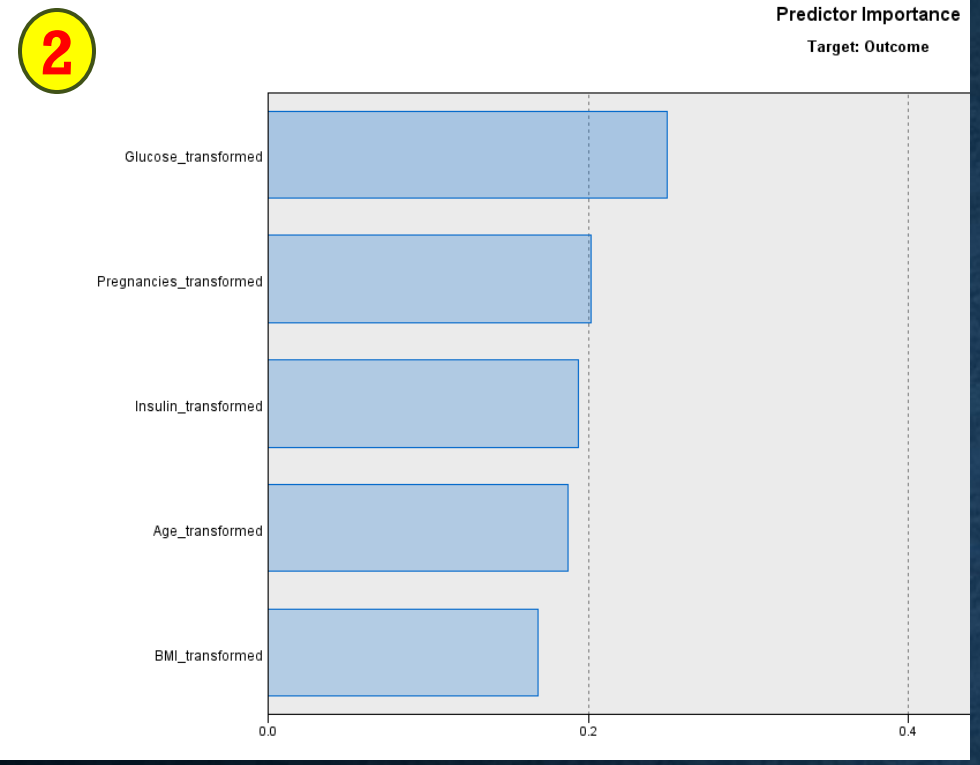
2 – میزان تاثیر پارامترهای مختلف روی Outcome در روش Quest

3 – میزان دقت داده های Train و Test در روش Quest با Analysis Data

با توجه به اینکه تاثیر گلوکز در این روش بالا در نظر گرفته شده است و تاثیر بقیه پارامترها تقریبا ناچیز است ؛ همچنین میزان دقت داده های Train و Test تقریبا Underfit است، مدل قابل اتکا نمی باشد.

C5

2



1

- Rules for 1 - contains 3 rule(s)
  - Rule 1 for **1**
    - if Pregnancies_transformed > 44.444
    - and Insulin_transformed > 20.809
    - then **1**
  - Rule 2 for **1**
    - if Glucose_transformed > 58.794
    - and BMI_transformed > 59.502
    - then **1**
  - Rule 3 for **1**
    - if Pregnancies_transformed > 22.222
    - and Glucose_transformed > 58.794
    - and Age_transformed <= 63.333
    - then **1**
- Rules for 0 - contains 4 rule(s)
  - Rule 1 for **0**
    - if Glucose_transformed > 58.794
    - and BMI_transformed <= 59.502
    - and Age_transformed > 63.333
    - then **0**
  - Rule 2 for **0**
    - if Pregnancies_transformed <= 22.222
    - and BMI_transformed <= 59.502
    - and Age_transformed <= 63.333
    - then **0**
  - Rule 3 for **0**
    - if Glucose_transformed <= 58.794
    - and Insulin_transformed <= 20.809
    - then **0**
  - Rule 4 for **0**
    - if Pregnancies_transformed <= 44.444
    - and Glucose_transformed <= 58.794
    - then **0**
- Default: 1

Collapse All | Expand All

- Results for output field Outcome
  - Comparing $C-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	447	74.01%	116	70.73%
Wrong	157	25.99%	48	29.27%
Total	604		164	

  - Coincidence Matrix for $C-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	276	125
1	32	171

'Partition' = 2_Testing	0	1
0	64	35
1	13	52

3

1 – مدل درخت C5 ، Rule set & Group symbolics

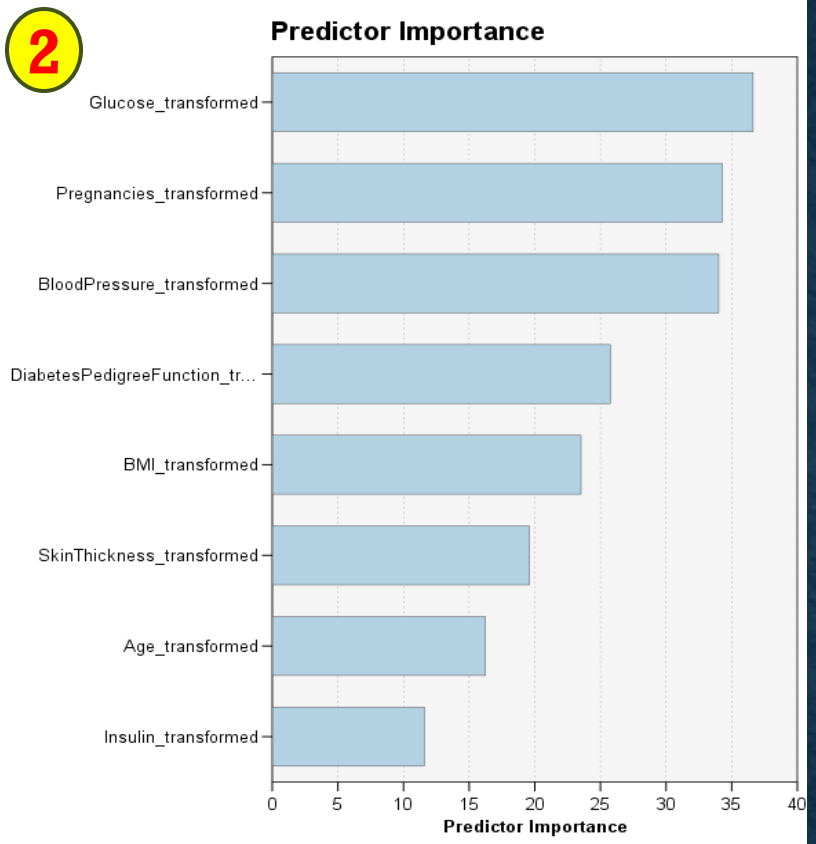
2 – میزان تاثیر پارامترهای مختلف روی Outcome در روش C5

3 – میزان دقت داده های Train و Test در روش C5 با Analysis Data

با توجه به اینکه تاثیرتمامی پارامترها در یک حد است مدل قابل اتکاست، ولی از آنجاییکه میزان دقت داده های Train و Test تقریبا Underfit است، راندمان لازمه را به ما نمی دهد.

Random Trees

2



Results for output field Outcome

Comparing $R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	544	87.18%	107	74.31%
Wrong	80	12.82%	37	25.69%
Total	624		144	

Coincidence Matrix for $R-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	337	73
1	7	207

'Partition' = 2_Testing	0	1
0	66	24
1	13	41

3

1 – مدل درخت Random Trees با عمق 20 و تعداد 100

2 – میزان تاثیر پارامترهای مختلف روی Outcome در روش Random Trees

3 – میزان دقت داده های Train و Test در روش Random Trees با Analysis Data

با توجه به اینکه تاثیرتمامی پارامترها در یک حد است مدل قابل اتکاست، ولی از آنجاییکه میزان دقت داده های Train و Overfit Test , است، مدل قابل اتکا نیست.

1

Random Trees

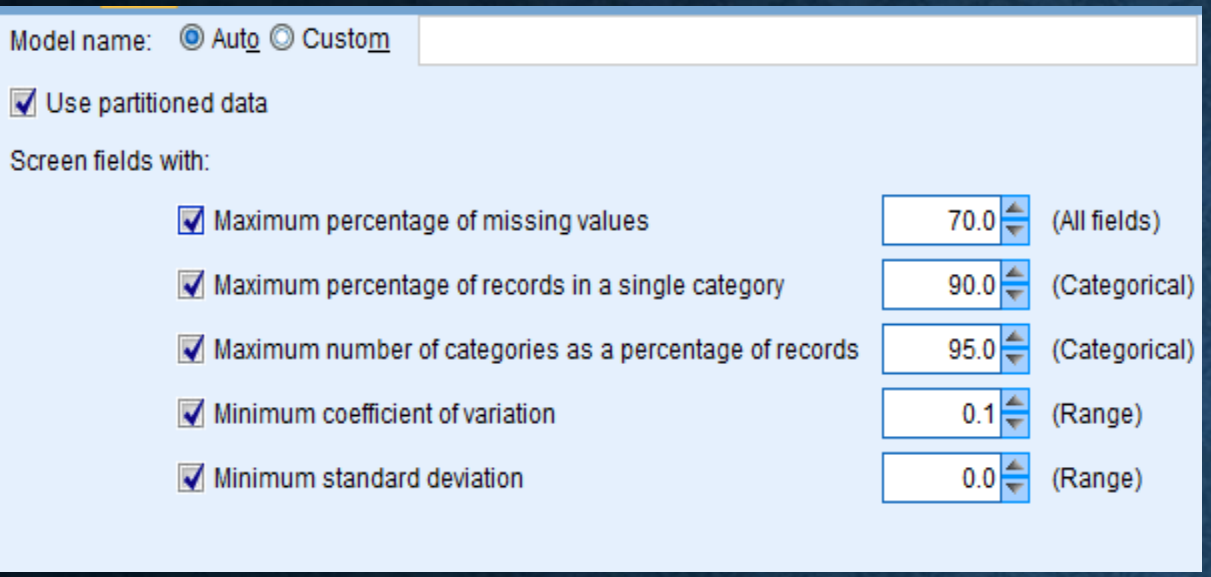
Model Information

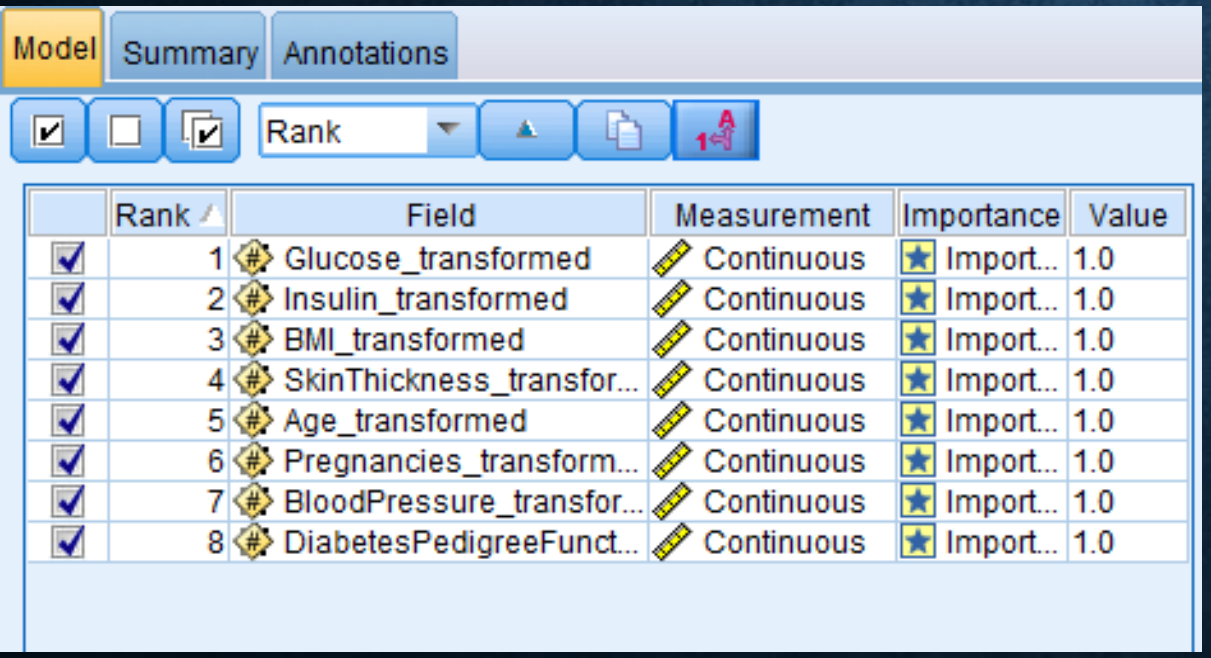
Target Field	Outcome
Model Building Method	Random Trees Classification
Number of Predictors Input	8
Model Accuracy	0.808
Misclassification Rate	0.192

Records Summary

Records	Number	Percent
Included	796	100.00
Excluded	0	0.00
Total	796	100.00

Feature Selection





	Rank	Field	Measurement	Importance	Value
☑	1	Glucose_transformed	Continuous	Import...	1.0
☑	2	Insulin_transformed	Continuous	Import...	1.0
☑	3	BMI_transformed	Continuous	Import...	1.0
☑	4	SkinThickness_transfor...	Continuous	Import...	1.0
☑	5	Age_transformed	Continuous	Import...	1.0
☑	6	Pregnancies_transform...	Continuous	Import...	1.0
☑	7	BloodPressure_transfor...	Continuous	Import...	1.0
☑	8	DiabetesPedigreeFunct...	Continuous	Import...	1.0

با توجه به برآورد مدل ریاضی Feature Selection ، تمامی پارامتر ها دارای اهمیتی یکسانی می باشند.

Auto Classifier

☑	C5	Specify...	64
☐	Logistic regression	Default	1
☑	Decision List	Default	1
☐	Bayesian Network	Default	1
☐	Discriminant	Default	1
☑	KNN Algorithm	Specify...	2
☐	LSVM	Default	1
☑	Random Trees	Specify...	4
☑	SVM	Specify...	108
☐	Tree-AS	Default	1
☑	CHAID	Specify...	12
☑	Quest	Specify...	6
☑	C&R Tree	Specify...	6
☑	Neural Net	Specify...	72

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Testing set

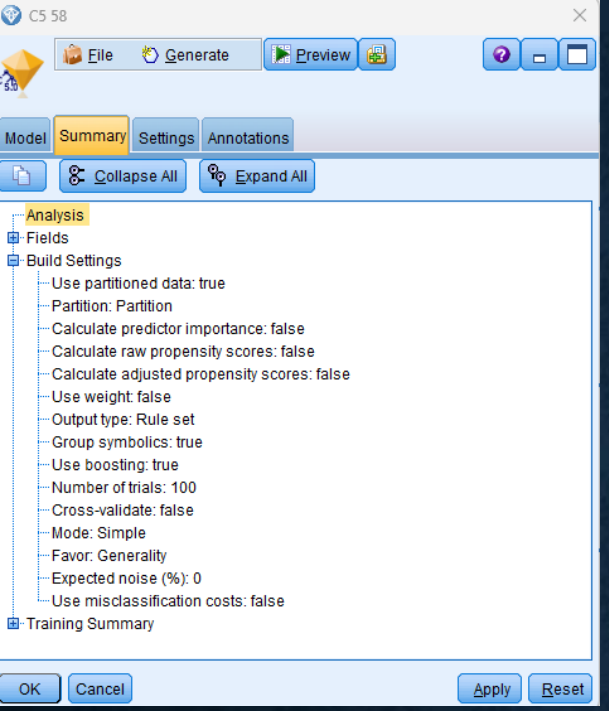
Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		C5 57	< 1	70.0	17	2.097	77.358	8	0.797
☑		C5 59	< 1	70.0	17	2.097	77.358	8	0.797
☑		C5 58	< 1	55.0	9	2.069	77.358	8	0.792
☑		C5 60	< 1	55.0	9	2.069	77.358	8	0.792
☑		C5 3	< 1	50.0	18	2.022	76.101	8	0.781
☑		C5 9	< 1	50.0	18	2.022	76.101	8	0.781
☑		C5 11	< 1	50.0	18	2.022	76.101	8	0.781
☑		C5 17	< 1	50.0	18	2.022	76.101	8	0.781
☑		C5 25	< 1	50.0	18	2.022	76.101	8	0.781
☑		C5 27	< 1	50.0	18	2.022	76.101	8	0.781

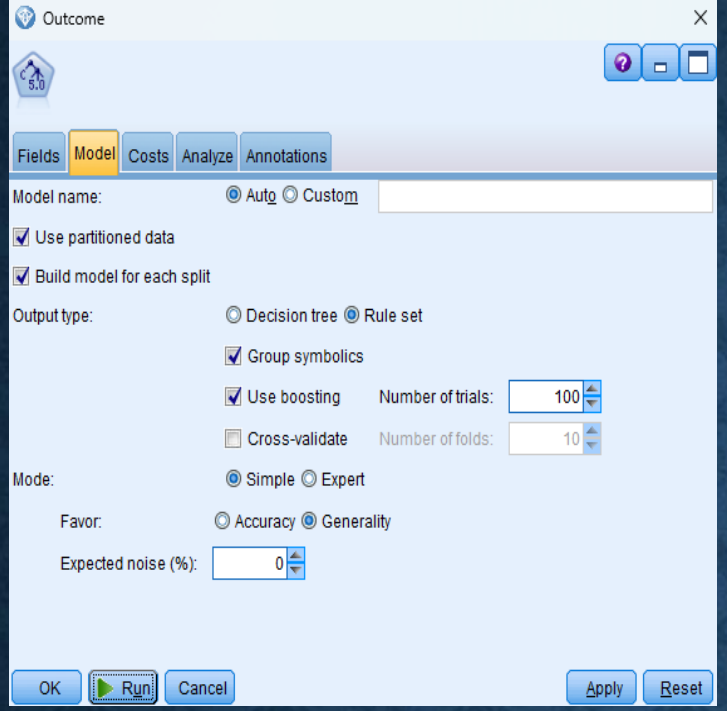
Sort by: Use | Ascending | Descending | Delete Unused Models | View: Training set

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		C5 58	< 1	600.0	36	2.226	82.594	8	0.909
☑		C5 60	< 1	600.0	36	2.226	82.594	8	0.909
☑		C5 57	< 1	605.0	34	2.224	83.580	8	0.908
☑		C5 59	< 1	605.0	34	2.224	83.580	8	0.908
☑		C5 3	< 1	585.0	40	2.173	82.594	8	0.9
☑		C5 9	< 1	585.0	40	2.173	82.594	8	0.9
☑		C5 11	< 1	585.0	40	2.173	82.594	8	0.9
☑		C5 17	< 1	585.0	40	2.173	82.594	8	0.9
☑		C5 25	< 1	585.0	40	2.173	82.594	8	0.9
☑		C5 27	< 1	585.0	40	2.173	82.594	8	0.9

در این مدل ما با آوردن تمامی مدل هایی که تا به الان داده ها را تست کردین و **Specify** کردن و آوردن تمامی **option** ها در آن **Top 10** بهترین مدل برای نتیجه گیری این داده ها می آوریم و همانطور که مشاهده می کنیم ، مدل **C5 58** با توجه به تنظیمات داده شده را انتخاب می نماییم.

Result





Results for output field Outcome

Comparing $C-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	472	78.02%	126	77.3%
Wrong	133	21.98%	37	22.7%
Total	605		163	

Coincidence Matrix for $C-Outcome (rows show actuals)

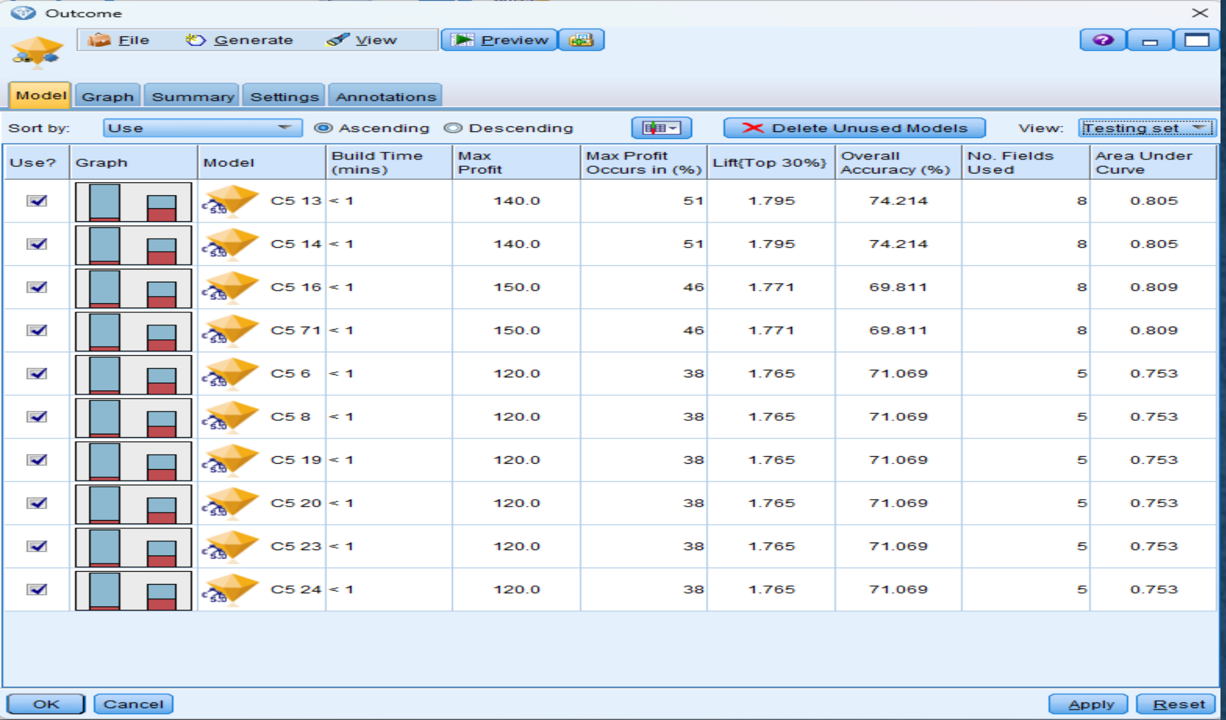
'Partition' = 1_Training	0	1
0	296	97
1	36	176

'Partition' = 2_Testing	0	1
0	81	26
1	11	45

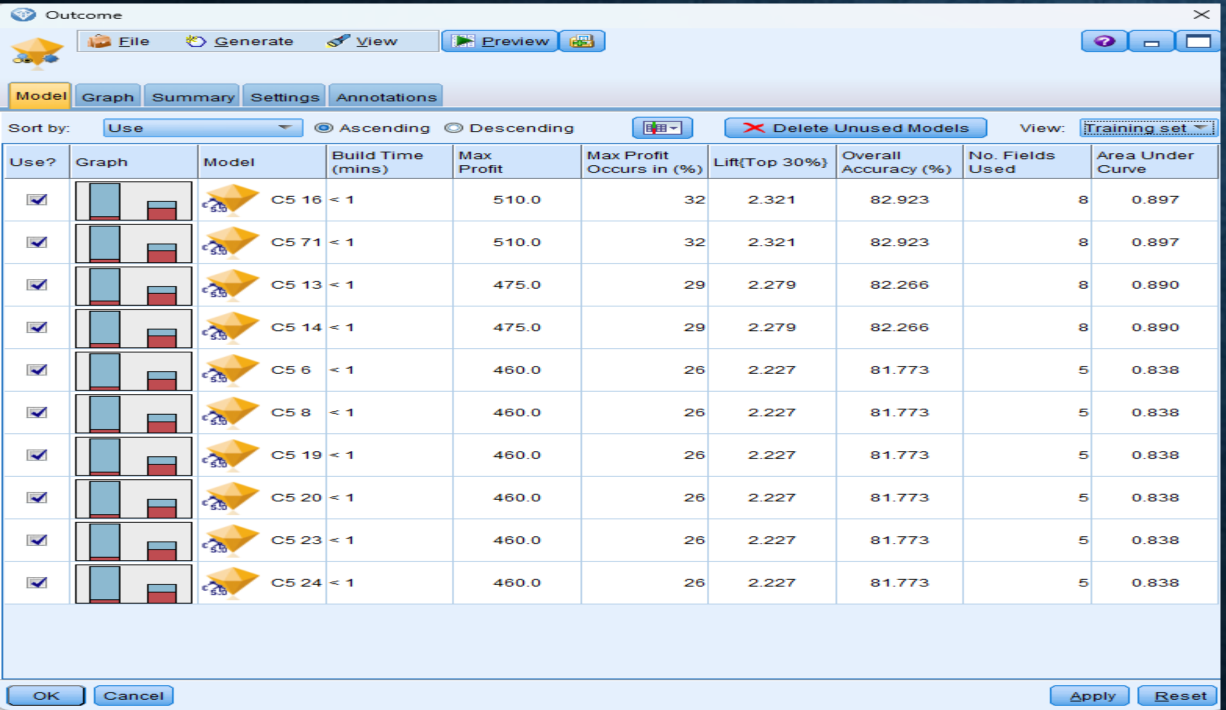
تنظیمات را مطابق C5 28 برای درخت C5 انجام داده و همانطور که انتظار می رفت بهترین دقت در داده های Train و Test به ما داد.

RE-DOING THE TREE MODELS ON RAW DATA

همانطور که مشاهده می کنید ، تمامی مدل های درخت را میتوان قبل از تبدیل اعداد نیز حساب کرد ، پس انجام و تست دوباره تمامی مدل های درخت روی داده اولیه همچنین اجرای **Auto Classifier** و **Boost** و **Reduce** به این نتیجه رسیدیم که تمامی مدل ها **Overfit** می باشند و همان نتیجه قبلی بهترین نتیجه ما تا به اینجا بوده است.



Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
✓		C5 13	< 1	140.0	51	1.795	74.214	8	0.805
✓		C5 14	< 1	140.0	51	1.795	74.214	8	0.805
✓		C5 16	< 1	150.0	46	1.771	69.811	8	0.809
✓		C5 71	< 1	150.0	46	1.771	69.811	8	0.809
✓		C5 6	< 1	120.0	38	1.765	71.069	5	0.753
✓		C5 8	< 1	120.0	38	1.765	71.069	5	0.753
✓		C5 19	< 1	120.0	38	1.765	71.069	5	0.753
✓		C5 20	< 1	120.0	38	1.765	71.069	5	0.753
✓		C5 23	< 1	120.0	38	1.765	71.069	5	0.753
✓		C5 24	< 1	120.0	38	1.765	71.069	5	0.753



Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
✓		C5 16	< 1	510.0	32	2.321	82.923	8	0.897
✓		C5 71	< 1	510.0	32	2.321	82.923	8	0.897
✓		C5 13	< 1	475.0	29	2.279	82.266	8	0.890
✓		C5 14	< 1	475.0	29	2.279	82.266	8	0.890
✓		C5 6	< 1	460.0	26	2.227	81.773	5	0.838
✓		C5 8	< 1	460.0	26	2.227	81.773	5	0.838
✓		C5 19	< 1	460.0	26	2.227	81.773	5	0.838
✓		C5 20	< 1	460.0	26	2.227	81.773	5	0.838
✓		C5 23	< 1	460.0	26	2.227	81.773	5	0.838
✓		C5 24	< 1	460.0	26	2.227	81.773	5	0.838