

به نام خدا

پروژه نهایی دوره مبانی علم داده
پیش بینی دیابت با IBM Modeler

نام دانشجو :

صفا سامانیان

کد ۳۰

نام استاد :

دکتر محمدرضا محتاط

فهرست:

۳	شناسایی داده های معرفی شده.....
۵	تشخیص خطا یا نویز.....
۵	مراحل کار در نرم افزار.....
۶	تشخیص نویز.....
۸	تشخیص داده پرت.....
۱۱	داده ها مفقودی.....
۱۴	مدل سازی.....
۱۵	1) داده های نامتوازن.....
۱۶	a. SVM.....
۱۸	b. KNN.....
۱۸	c. Neural Net.....
۲۰	d. C&R Tree.....
۲۱	2) داده های دیابت منفی کاهش یافته.....
۲۲	a. SVM.....
۲۳	b. KNN.....
۲۳	c. Neural Net.....
۲۵	d. C&R Tree.....
۲۵	e. CHAID.....
۲۶	f. QUEST.....
۲۶	g. Random Tree.....
۲۷	h. C5.....
۲۷	انتخاب بهترین مدل.....

شناسایی داده های معرفی شده

ما در حال بررسی تاثیر مولفه های زیر بر دیابت در افراد هستیم

1. ستون اول Pregnancies :

دیابت بارداری به شرایطی گفته می شود که افزایش قند خون برای اولین بار، در طی دوران بارداری دیده شود. دیابت بارداری، تقریباً در ۴ درصد از بارداری ها بروز می کند. پس تعداد دفعات بارداری در هر فرد می تواند یکی از عوامل مهم در تشخیص دیابت باشد

2. ستون دوم Glucose :

در دیابت سرعت و توانایی بدن در استفاده و سوخت و ساز کامل گلوکز کاهش می یابد از این رو میزان قند خون افزایش یافته که به آن هیپرگلیسمی می گویند. قند خون در انسان بدون دیابت بین ۷۰ تا ۱۰۰ دسی لیتر می باشد ولی در افراد مبتلا به دیابت از ۱۳۰ دسی لیتر بیشتر است

3. ستون سوم BloodPressure :

فشار خون اندازه گیری نیرویی است که قلب برای پمپاژ کردن خون به سایر نقاط بدن از آن استفاده می کند.

دیابت و فشار خون بالا باعث ایجاد یکدیگر نمی شوند، اما مبتلایان به دیابت به طور معمول مستعد ابتلا به بیماری های دیگری از جمله فشار خون بالا و کلسترول خون بالا هستند. انسان با فشار خون نرمال بین ۸۰ تا ۱۲۰ میلی متر جیوه می باشد ولی این مقدار می تواند بیشتر از این بازه یا کمتر هم باشد

4. ستون چهارم SkinThickness :

حدود یک سوم از بیماران مبتلا به دیابت نوع ۱ از عارضه ی پوستی اسکروز دیجیتال رنج می برند. در این عارضه پوست انگشتان دست و پا ضخیم و سفت و مومی شکل می شوند. در اینجا هم در حال اندازه گیری ضخامت پوست هستیم ضخامت پوست انسان در حالت عادی بین ۱ تا ۴ میلی متر است ولی با دیابت این مقدار زیاد می شود

5. ستون پنجم Insulin :

انسولین یک پیام رسان شیمیایی است که به سلول ها اجازه می دهد گلوکز (قند خون) را جذب کنند. در برخی افراد، سیستم ایمنی بدن به جزایر لانگرهانس حمله می کند و این مسئله باعث می شود که آن ها، انسولین تولید نکنند یا تولید را کاهش دهند. وقتی این اتفاق می افتد، گلوکز در خون می ماند و سلول ها نمی توانند آن را جذب کنند و قندها را به انرژی تبدیل کنند. این امر به معنای شروع دیابت نوع 1 است و فرد مبتلا به این نوع دیابت برای زنده ماندن به تزریق منظم انسولین نیاز دارد. در برخی افراد به ویژه افرادی که دچار اضافه وزن، چاقی یا کم تحرکی هستند انسولین در انتقال گلوکز به سلول ها موثر عمل نمی کند و قادر به انجام وظایف خود نیست. ناتوانی انسولین در ایجاد اثر بر روی بافت ها، مقاومت به انسولین نامیده می شود. هنگامی که جزایر لانگرهانس نتوانند انسولین کافی برای غلبه بر مقاومت به انسولین تولید کنند، دیابت نوع 2 ایجاد می شود. افراد مبتلا به این نوع دیابت نیز در نهایت به تزریق انسولین نیاز پیدا می کنند.

سطح انسولین و گلوکز خون باید متعادل باشد. بعد از غذا، کربوهیدرات ها معمولاً به گلوکز و سایر قندهای ساده تجزیه می شوند. اینها در خون جذب می شوند و باعث افزایش سطح گلوکز خون می شوند که به نوبه خود پانکراس را برای ترشح انسولین در خون تحریک می کند. با حرکت گلوکز به داخل سلول ها، سطح خون کاهش می یابد و ترشح انسولین توسط پانکراس کاهش می یابد.

اگر فردی نتواند انسولین کافی تولید کند یا سلول های بدن در برابر اثرات آن مقاوم شوند (مقاومت به انسولین)، گلوکز نمی تواند به بیشتر سلول های بدن برسد و سلول ها از گرسنگی می میرند. در همین حال، سطح گلوکز در خون به سطوح ناسالم افزایش می یابد.

میزان انسولین نرمال در انسان بزرگسالان : $pmo1/L$ 43_186 یا uU/mL 6_26 و نوزادان : uU/mL 3_20.

مقادیر بحرانی آزمایش Insulin مقدار بیشتر از uU/mL 50 می باشد

6. ستون ششم BMI :

شاخص توده بدنی یا بی ام آی (به انگلیسی body mass index، مخفف BMI) سنجشی آماری برای مقایسه وزن و قد یک فرد است. در واقع این سنجش میزان چاقی را اندازه گیری نمی کند بلکه ابزاری مناسب است تا سلامت وزن فرد با توجه به قدش تخمین زده شود. هرچه BMI بیشتر باشد احتمال ابتلا به دیابت افزایش می یابد

7. ستون هفتم DiabetesPedigreeFunction :

"DiabetesPedigreeFunction" تابعی است که احتمال ابتلا به دیابت را بر اساس سابقه خانوادگی با دامنه واقعی 0.08 تا 2.42 نمره می دهد.

8. ستون Age :

مؤلفه ی سن هر فرد در این داده چون کمترین سن در حدود ۲۰ سال است پس ما با دیابت نوع ۲ درگیر هستیم

من برای صحت سنجی نویز داده ها از روش های سرچ در گوگل و استفاده از chatgpt و سوال از خبره استفاده کردم

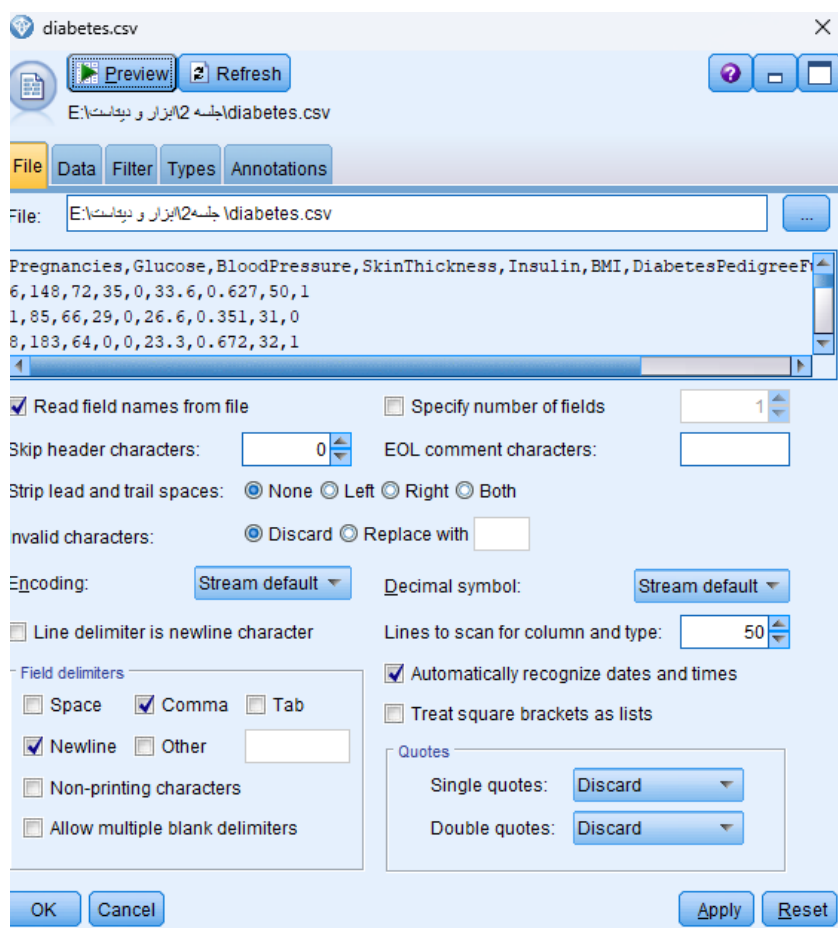
تشخیص خطا یا نویز

با توجه به تعاریف داده ها مقادیری که داده ها نمی پذیرند به صورت زیر است

SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age
صفر و منفی	صفر و منفی	صفر و منفی	غیر از [0.08-2.42]	صفر و منفی
	Pregnancies	Glucose	BloodPressure	
	منفی و اعشاری	صفر و منفی	صفر و منفی	

مراحل کار در نرم افزار

ابتدا باید فایل مورد نظر را وارد برنامه کنیم برای اینکار از قسمت `var.file` → `sources` را وارد کرده سپس فایل مورد نظر را داخل آن بارگذاری میکنیم



سپس باید از قسمت type جنی هر داده را در برنامه معرفی کنیم

File

Data

Filter

Types

Annotations

▶ Read Values

Clear Values

Clear All Values

Field	Measurement	Values	Missing	Check	Role
Pregnancies	Continuous	[0,17]		None	Input
Glucose	Continuous	[0,199]		None	Input
BloodPressu...	Continuous	[0,122]		None	Input
SkinThickness	Continuous	[0,99]		None	Input
Insulin	Continuous	[0,846]		None	Input
BMI	Continuous	[0.0,67.1]		None	Input
DiabetesPed...	Continuous	[0.078,2.42]		None	Input
Age	Continuous	[21,81]		None	Input
Outcome	Flag	1/0		None	Target

تشخیص نویز

در مرحله بعد باید با فیلر مقادیر نویز را مشخص و نول کنیم

Filler

Preview

Settings
Annotations

Fill in fields:

SkinThickness

Replace: Based on condition

Condition:

1 SkinThickness = 0

Replace with:

1 undef

OK
Cancel
Apply
Reset

و خروجی جدول بعد از فیلتر

Table (9 fields, 768 records)

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

برای تمامی ستون هایی که مقدار صفر در آنها تعریف پذیر نیست این عمل را تکرار می کنیم مشابه حالی قبل شرط را در مورد DiabetesPedigreeFunction نیز به صورت زیر تکرار می کنیم چون بازه این داده بین [0.08 – 2.42] است

Filler

Preview

Settings Annotations

Fill in fields:

DiabetesPedigreeFunction

Replace: Based on condition

Condition:

1 DiabetesPedigreeFunction < 0.08 or DiabetesPedigreeFunction

Replace with:

1 undef

OK Cancel Apply Reset

داده های نهایی بعد از انجام کلیه فیلرها

Table (9 fields, 768 records) #1

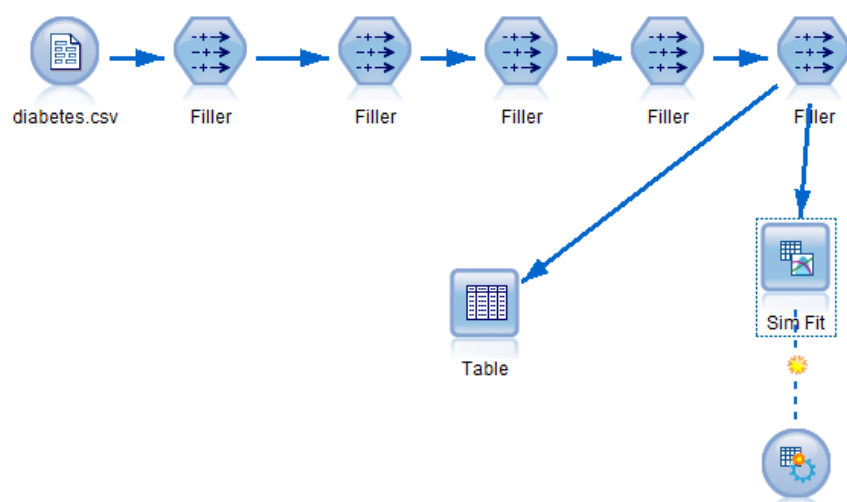
File Edit Generate

Table Annotations

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

تشخیص داده پرت

در این مرحله باید داده های پرت را شناسایی کنیم برای این کار باید ابتدا توزیع هر داده را مشخص کنیم



با دو روش می توان توزیع ها را مشخص کرد

روش Anderson-Darling :

Field	Storage	Status		Distribution	Parameters
Pregnancies	Integer	✓	<input type="checkbox"/>	Exponential	[scale=0.3029366...
Glucose	Integer	✓	<input type="checkbox"/>	Lognormal	[a=118.872658692...
BloodPressure	Integer	✓	<input type="checkbox"/>	Normal	[mean=70.663265...
SkinThickness	Integer	✓	<input type="checkbox"/>	Weibull	[shape1=32.66296...
Insulin	Integer	✓	<input type="checkbox"/>	Lognormal	[a=123.115102908...
BMI	Real	✓	<input type="checkbox"/>	Gamma	[shape=22.699650...
DiabetesPedigree...	Real	✓	<input type="checkbox"/>	Lognormal	[a=0.43208120745...
Age	Integer	✓	<input type="checkbox"/>	Lognormal	[a=29.4789886441...
Outcome	Integer	✓	<input type="checkbox"/>	Categorical	[0=0.66836734693...

روش Kolmogorov-Smirnov :

Field	Storage	Status		Distribution	Parameters
Pregnancies	Integer	✓	<input type="checkbox"/>	Exponential	[scale=0.3029366...
Glucose	Integer	✓	<input type="checkbox"/>	Lognormal	[a=118.872658692...
BloodPressure	Integer	✓	<input type="checkbox"/>	Normal	[mean=70.663265...
SkinThickness	Integer	✓	<input type="checkbox"/>	Weibull	[shape1=32.66296...
Insulin	Integer	✓	<input type="checkbox"/>	Lognormal	[a=123.115102908...
BMI	Real	✓	<input type="checkbox"/>	Normal	[mean=33.086224...
DiabetesPedigree...	Real	✓	<input type="checkbox"/>	Lognormal	[a=0.43208120745...
Age	Integer	✓	<input type="checkbox"/>	Lognormal	[a=29.4789886441...
Outcome	Integer	✓	<input type="checkbox"/>	Categorical	[0=0.66836734693...

پس داده های BMI و BloodPressure را با روش Z محاسبه می کنیم و بقیه که نرمال نیستند را با box-plot داده های پرت را مشخص می کنیم

و در اینجا از قسمت همبستگی بررسی می کنیم اگر داده ها همبستگی بیشتر از ۰/۷ داشته باشد میتوان یکی از آنها را حذف کرد

	Age	BMI	BloodPressu...	DiabetesPed...	Glucose	Insulin	Pregnancies	SkinThickness
Age	1.000	0.070	0.300	0.085	0.344	0.217	0.680	0.168
BMI	0.070	1.000	0.304	0.159	0.210	0.226	-0.025	0.664
BloodPressure	0.300	0.304	1.000	-0.016	0.210	0.099	0.213	0.233
DiabetesPedi...	0.085	0.159	-0.016	1.000	0.140	0.136	0.008	0.160
Glucose	0.344	0.210	0.210	0.140	1.000	0.581	0.198	0.199
Insulin	0.217	0.226	0.099	0.136	0.581	1.000	0.079	0.182
Pregnancies	0.680	-0.025	0.213	0.008	0.198	0.079	1.000	0.093
SkinThickness	0.168	0.664	0.233	0.160	0.199	0.182	0.093	1.000

در این اینجا همبستگی بالای ۰/۷ نداریم

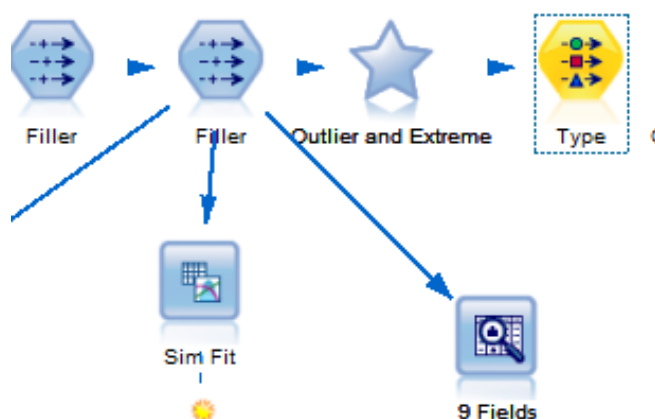
حال با استفاده از data audit به محاسبه داده پرت می پردازیم

برای دو داده نرمال bloodpressure و BMI را از روش Z استفاده می کنیم

چون تعداد داده های bloodpressure و BMI کم است آنها را coerce (یعنی جایگذاری با 3sigma) کردیم

Audit Quality Annotations				
Complete fields (%): 33.33%		Complete records (%): 51.04%		
Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	None
Glucose	Continuous	0	0	None
BloodPressu...	Continuous	14	0	Coerce
SkinThickness	Continuous	2	1	None
Insulin	Continuous	16	8	None
BMI	Continuous	7	1	Coerce
DiabetesPed...	Continuous	23	6	None
Age	Continuous	9	0	None
Outcome	Flag	--	--	--

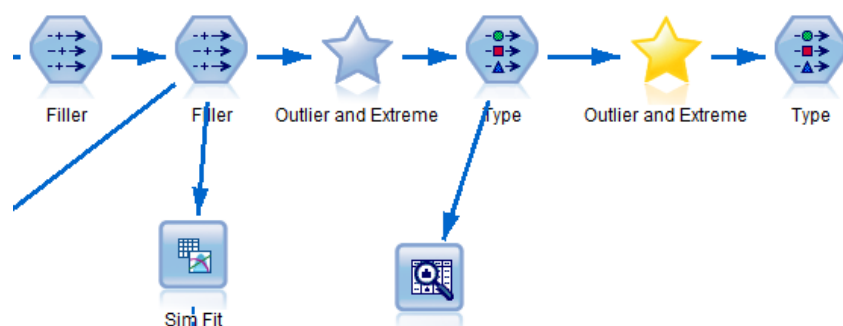
آیکون جدید اضافه شده را متصل کرده و بعد از آن یک type هم اضافه می کنیم تا داده ها شناسایی شود



برای بقیه داده ها از روش باکس پلات استفاده می کنیم و مشابه حالت قبل داده های پرت را مدیریت می کنیم

Audit Quality Annotations				
Complete fields (%): 33.33%		Complete records (%): 51.04%		
Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	Coerce
Glucose	Continuous	0	0	None
BloodPressu...	Continuous	0	0	None
SkinThickness	Continuous	2	1	Coerce
Insulin	Continuous	16	8	Coerce outliers / nullify extremes
BMI	Continuous	0	0	None
DiabetesPed...	Continuous	23	6	Coerce outliers / discard extremes
Age	Continuous	9	0	Coerce
Outcome	Flag	--	--	--

دو داده ی پرت مدیریت شده را به این صورت در امتداد همدیگر قرار می دهیم سپس با یک type داده ها را می خوانیم



داده ها مفقودی

در این مرحله باید داده های مفقودی را مدیریت کنیم برای این کار از type آخر دوباره data audit میگیریم

Field	Measurement	Outliers	Extre...	Action	Impute Missing	Method	% Complete	Valid Reco...	Null Val...
Pregnancies	Continuous	0	0	None	Never	Fixed	100	761	0
Age	Continuous	0	0	None	Never	Fixed	100	761	0
Outcome	Flag	--	--	--	Never	Fixed	100	761	0
Glucose	Continuous	0	0	None	Blank & Null Values	Fixed	99.343	756	5
BloodPressu...	Continuous	0	0	None	Blank & Null Values	Fixed	95.401	726	35
SkinThickness	Continuous	0	0	None	Blank & Null Values	Fixed	70.434	536	225
Insulin	Continuous	16	0	None	Blank & Null Values	Fixed	50.329	383	378
BMI	Continuous	7	0	None	Blank & Null Values	Fixed	98.555	750	11
DiabetesPed...	Continuous	27	0	None	Blank & Null Values	Fixed	100	761	0

که با توجه به میزان اهمیت هر کدام از مولفه ها باید برای روش پر کردن مفقودی ها اقدام کنیم

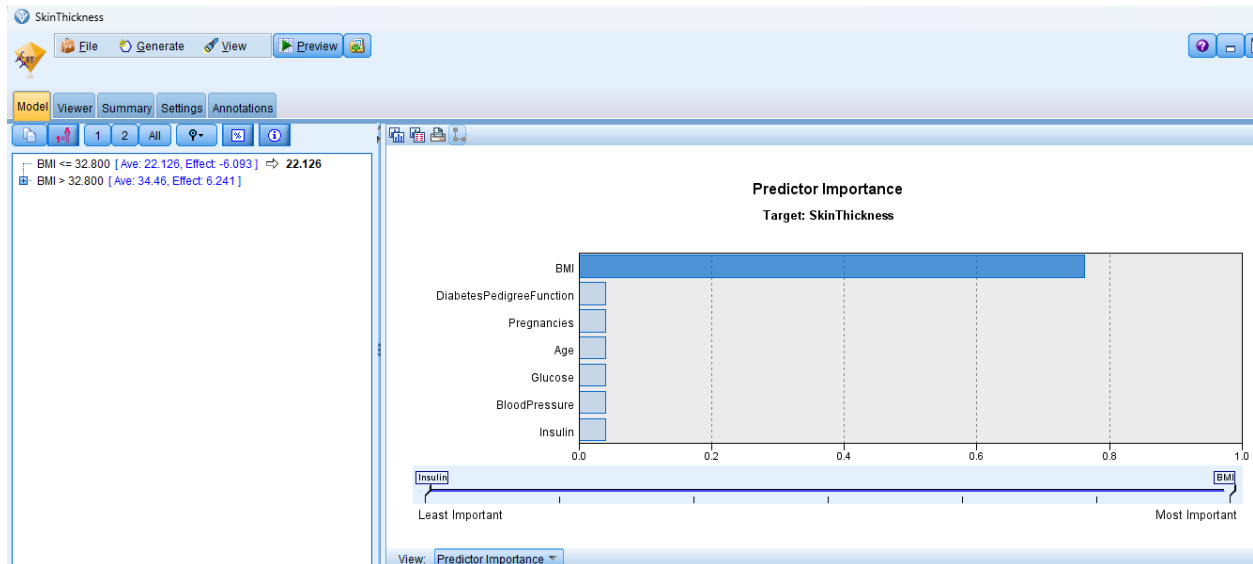
Complete fields (%): 44.44% Complete records (%): 50.07%

Field	Measurement	Outli...	Extr...	Action	Impute Missing	Method	% Complete	Valid Reco...	Null Val...
Pregnancies	Continuous	0	0	None	Never	Fixed	100	761	0
Age	Continuous	0	0	None	Never	Fixed	100	761	0
Outcome	Flag	--	--	--	Never	Fixed	100	761	0
Glucose	Continuous	0	0	None	Blank & Null Values	Fixed	99.343	756	5
BloodPressu...	Continuous	0	0	None	Blank & Null Values	Random	95.401	726	35
SkinThickness	Continuous	0	0	None	Blank & Null Values	Algorithm	70.434	536	225
Insulin	Continuous	16	0	None	Blank & Null Values	Random	50.329	383	378
BMI	Continuous	7	0	None	Blank & Null Values	Random	98.555	750	11
DiabetesPed...	Continuous	27	0	None	Blank & Null Values	Fixed	100	761	0

من برای مقدارهای glucose و DiabetesPedigreeFunction که مقدارهای کمتری است از روش مقدار ثابت مد استفاده کردم برای bloodpressure و BMI از روش رندم نرمال و برای Insulin از روش رندم غیر نرمال و برای میزان skinthickness از روش الگوریتم استفاده کردم

پس از انجام عملیات مدل های انجام شده توسط برنامه را بررسی می کنیم

در مدل **skinthickness** پیشنهادی توسط برنامه مشخص شده که بیشترین تاثیر بر این متغیر از میزان **BMI** و بعد از آن تاثیر بقیه مولفه ها با میزان کمتر را بر **skinthickness** داریم

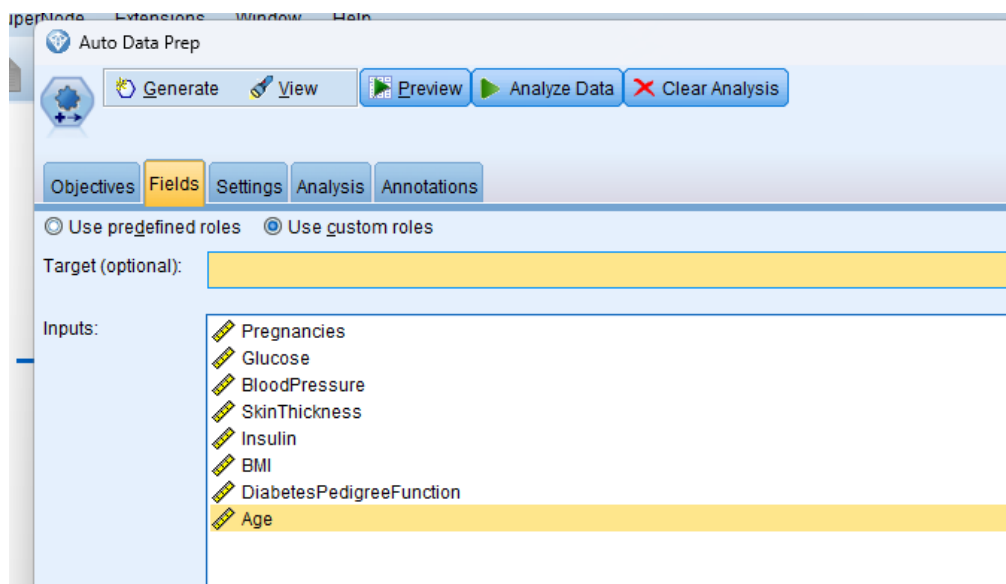


خروجی بعد از انجام عملیات پر کردن مفقودی

Audit Quality Annotations								
Complete fields (%): 100%			Complete records (%): 100%					
Field	Measurement	Outliers	Extr...	Action	Impute Mi...	Method	% Co...	Valid ...
# Pregnancies	Continuous	0	0	None	Never	Fixed	100	761
# Glucose	Continuous	0	0	None	Never	Fixed	100	761
# BloodPressu...	Continuous	1	0	None	Never	Fixed	100	761
# SkinThickness	Continuous	3	0	None	Never	Fixed	100	761
# Insulin	Continuous	0	0	None	Never	Fixed	100	761
# BMI	Continuous	7	0	None	Never	Fixed	100	761
# DiabetesPed...	Continuous	27	0	None	Never	Fixed	100	761
# Age	Continuous	0	0	None	Never	Fixed	100	761
# Outcome	Flag	--	--	--	Never	Fixed	100	761

در مرحله آخر باید داده ها را استاندارد سازی کنیم

برای استانداردسازی داده ها باید از یک **Auto Data Prep** استفاده کنیم



نتیجه preview براساس min/max transformation (بین ۰ تا ۱۰۰)

Table (9 fields, 761 records) #1

File Edit Generate

Table Annotations

	Outcome	Pregnanci...	Glucose...	BloodPress...	SkinThickne...	Insulin_tr...	BMI_transf...	DiabetesPedi...	Age_transfor...
1	1	44.444	67.097	50.000	56.000	91.050	48.050	48.591	63.736
2	0	7.407	26.452	40.625	44.000	30.207	26.209	23.893	21.978
3	1	59.259	89.677	37.500	30.252	94.161	15.913	52.617	24.176
4	0	7.407	29.032	40.625	32.000	22.832	30.889	7.427	0.000
5	0	37.037	46.452	53.125	30.252	8.867	23.089	10.470	19.780
6	1	22.222	21.935	15.625	50.000	21.098	39.938	14.676	10.989
7	0	74.074	45.806	55.365	49.212	8.974	53.354	4.474	17.582
8	1	14.815	98.710	46.875	76.000	26.087	38.378	6.622	70.330
9	1	59.259	52.258	87.500	61.414	55.132	45.250	13.244	72.527
10	0	29.630	42.581	81.250	61.414	3.113	60.530	9.575	19.780
11	1	74.074	80.000	53.125	61.414	21.669	61.778	40.537	28.571
12	0	74.074	61.290	62.500	30.252	69.745	27.769	100.000	79.121
13	1	7.407	93.548	31.250	32.000	5.178	37.129	28.098	83.516
14	1	37.037	78.710	50.000	24.000	46.243	23.713	45.011	65.934
15	1	51.852	36.129	52.335	30.252	81.990	36.817	35.794	24.176
16	1	0.000	47.742	68.750	80.000	62.139	86.115	41.790	21.978
17	1	51.852	40.645	53.125	30.252	51.476	35.569	15.213	21.978
18	0	7.407	38.065	0.000	62.000	19.653	78.315	8.859	26.374
19	1	7.407	45.806	46.875	46.000	23.410	51.170	39.821	24.176
20	0	22.222	52.903	75.000	68.000	63.584	65.835	55.481	13.187
21	0	59.259	35.484	68.750	49.212	72.743	53.666	27.204	63.736

مدل سازی:

برای مدل سازی ما ابتدا از قسمت Field Ops گزینه ی Partition را انتخاب می کنیم و مطابق شکل زیر مقدار ۸۰ درصد داده ها را برای train و ۲۰ درصد را برای test انتخاب می کنیم

Settings Annotations

Partition field: Partition

Partitions: ☒ Train and test ☐ Train, test and validation

Training partition size: 80 Label: Training Value = "1_Training"

Testing partition size: 20 Label: Testing Value = "2_Testing"

Validation partition size: 0 Label: Validation Value = "3_Validation"

Total size: 100%

Values: ☐ Use system-defined values ("1", "2" and "3")
☒ Append labels to system-defined values
☐ Use labels as values

☒ Repeatable partition assignment

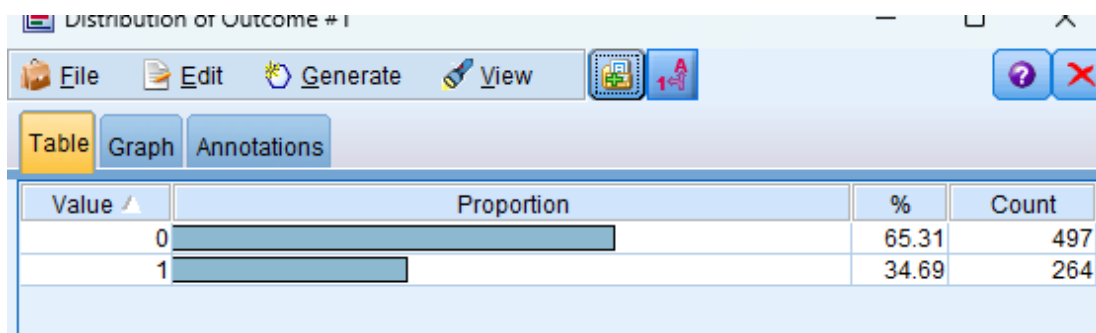
Seed: 1234567 Generate

☐ Use unique field to assign partitions:

سپس باید در این مرحله هر یک از مدل های موجود را بر روی داده ها بررسی کرده و با توجه به خروجی ها تصمیم بگیریم داده ما با کدام یکی از این مدل ها بهترین دقت را دارد برای داده های نا متوازن معمولاً از سه روش می توان مدل سازی را پیش برد

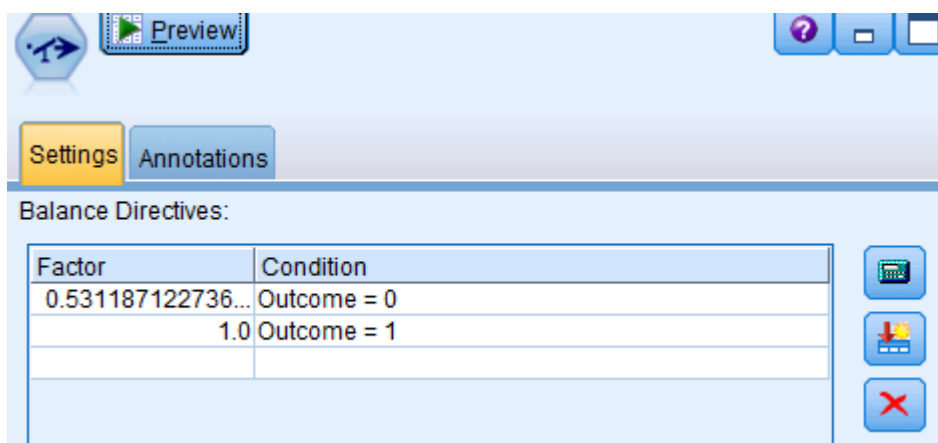
- استفاده از داده ها در همان صورت موجود
- روش نمونه گیری از داده های بیشتر به اندازه ی داده های کمتر
- روش چند برابر کردن داده های کمتر به اندازه ی داده های بیشتر

برای مشاهده ی توازن داده ها در قسمت **Graphs → Distribution** بعد از وارد کردن و اجرا داریم:

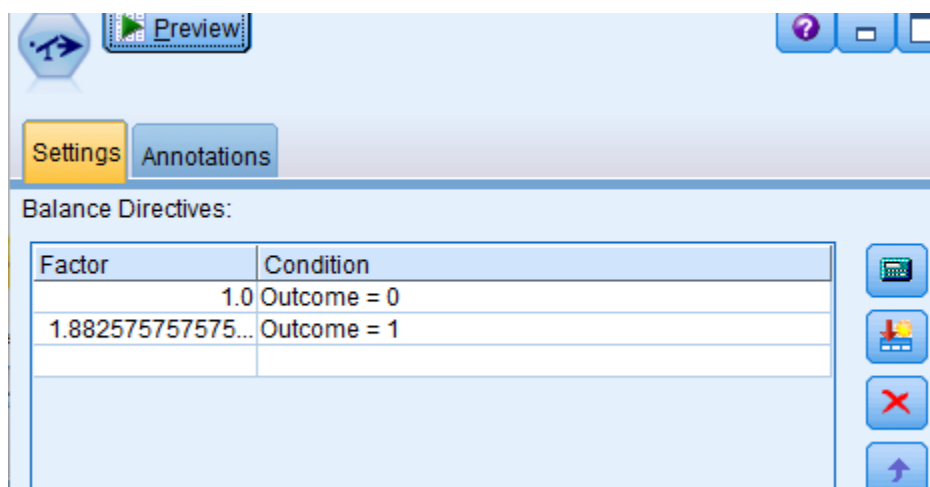


که مشاهده می شود در اینجا داده های دیابت مثبت تقریباً نصف داده های دیابت منفی است

برای متوازن کردن داده ها از قسمت generate دوگزینه ی Balance Node(boost) برای افزایش داده های کمتر به اندازه ی داده های بیشتر و گزینه ی Balance Node(reduce) برای کاهش داده های بیشتر به اندازه ی داده های کمتر را داریم.



در این حالت داده های دیابت مثبت را نگه داشته ولی دیابت منفی را نصف کرده



در این حالت داده های دیابت مثبت را ۱.۸۸ برابر کرده تا با داده های دیابت منفی برابر شود

حال به بررسی مدل ها در تک تک این حالت ها می پردازیم

۱. داده های نامتوازن

در ابتدا ویژگی ها را تست می کنیم تا میزان اهمیت هر کدام را بدست بیاوریم تا در صورت لزوم اقدامات لازم را انجام دهیم برای این کار از قسمت Feature Selection → Modeling را اجرا می کنیم

همانطور که می بینیم تمامی داده ها از نظر مدل مهم هستند و درصد اهمیت بالایی دارند و داده ای حذف نمی شود.

Model Summary Annotations					
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Rank		
	Rank	Field	Measurement	Importance	Value
<input checked="" type="checkbox"/>	1	Glucose_transformed	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	2	BMI_transformed	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	3	Age_transformed	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	4	Pregnancies_transform...	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	5	SkinThickness_transfor...	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	6	BloodPressure_transfor...	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	7	DiabetesPedigreeFunct...	Continuous	Import...	1.0
<input checked="" type="checkbox"/>	8	Insulin_transformed	Continuous	Import...	0.994

بررس مدل ها در این حالت

• SVM

در حالت RBF با $C=10$ و گاما 0.1

Analysis of [Outcome]					
File Edit					
Analysis Annotations					
Collapse All Expand All					
Results for output field Outcome					
Comparing \$\$-Outcome with Outcome					
'Partition'	1_Training		2_Testing		
Correct	459	75.87%	132	84.62%	
Wrong	146	24.13%	24	15.38%	
Total	605		156		
Coincidence Matrix for \$\$-Outcome (rows show actuals)					
'Partition' = 1_Training		0	1		
0		341	53		
1		93	118		
'Partition' = 2_Testing		0	1		
0		96	7		
1		17	36		

در این مدل مشاهده می شود که داده ها به طور واضحی روی حالت صفر بایاس شده و نتیجه خوبی در حالت ۱ ندارد

در حالت Polynomial با $C=10$ و گاما 1

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	479	79.17%	127	81.41%
Wrong	126	20.83%	29	18.59%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		359	35
1		91	120

'Partition' = 2_Testing		0	1
0		94	9
1		20	33

می بینیم که با تغییر گاما باز هم بایاس روی صفر را داریم
درحالت Sigmoid با $C=10$ و گاما 1

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	442	73.06%	127	81.41%
Wrong	163	26.94%	29	18.59%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		337	57
1		106	105

'Partition' = 2_Testing		0	1
0		98	5
1		24	29

باز هم مشاهده می شود که مدل قابلیت پیشبینی خوبی در حالت 1 ندارد
درحالت Sigmoid با $C=1$ و گاما 1

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	444	73.39%	120	76.92%
Wrong	161	26.61%	36	23.08%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		344	50
1		111	100

'Partition' = 2_Testing		0	1
0		96	7
1		29	24

در این جا نیز مشابه قبل مدل خوبی نداریم به جهت بایاس زیاد روی صفر
درحالت Linear با $C=10$

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	460	76.03%	126	80.77%
Wrong	145	23.97%	30	19.23%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		349	45
1		100	111

'Partition' = 2_Testing		0	1
0		93	10
1		20	33

در این حالت نیز بایاس زیادی بر روی صفر داریم و مدل نتیجه خوبی ندارد
پس مدل SVM در با این توازن وزنی نامتعادل بین صفر و یک مدل خوبی نیست

• KNN

در این مدل نیز نامتوازن بودن داده ها و بایاس زیاد بر روی صفر را داریم و این مدل نیز مناسب نیست

Results for output field Outcome

Comparing \$KNN-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	466	77.02%	124	79.49%
Wrong	139	22.98%	32	20.51%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		340	54
1		85	126

'Partition' = 2_Testing		0	1
0		92	11
1		21	32

• Neural Net

Results for output field Outcome

Comparing \$N-Outcome with Outcome

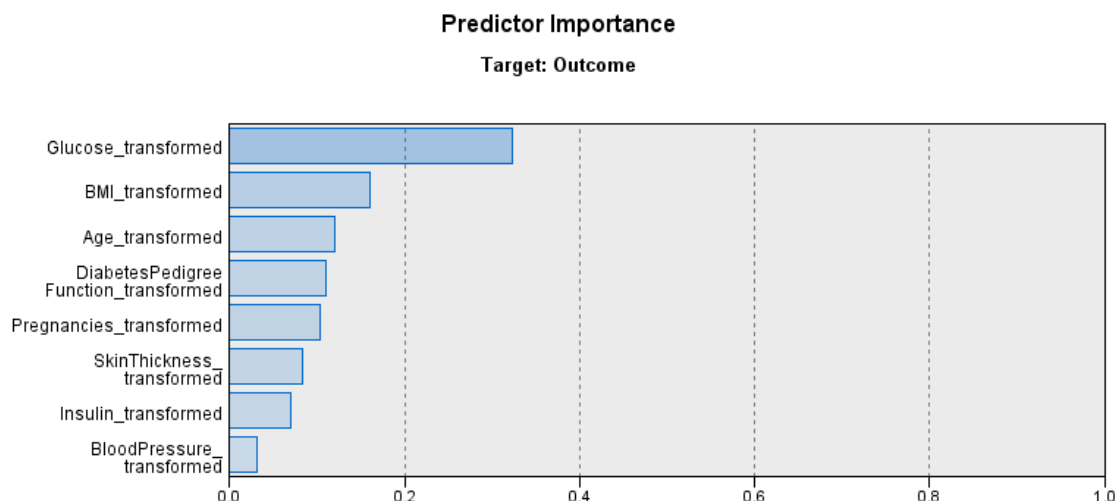
'Partition'	1_Training		2_Testing	
Correct	479	79.17%	125	80.13%
Wrong	126	20.83%	31	19.87%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		345	49
1		77	134

'Partition' = 2_Testing		0	1
0		95	8
1		23	30

مدل شبکه عصبی نیز مانند حالات قبلی مشکل بایاس نبودن روی یک را دارد و توصیه نمی شود



طبق این نمودار مشاهده می شود که بیشترین تاثیر را از گلوکز و کمترین را از فشار خون داشته پس در اینجا ما فشارخون را حذف می کنیم تا ببینیم تاثیر مثبتی بر مدل دارد یا خیر همان طور که در زیر نیز قابل مشاهده است باز هم نتایج پیش بینی خوبی در قسمت دیابت مثبت نداریم

Results for output field Outcome

Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	462	76.36%	127	81.41%
Wrong	143	23.64%	29	18.59%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		339	55
1		88	123
'Partition' = 2_Testing		0	1
0		96	7
1		22	31

برای مدل شبکه عصبی در حالت **boosting** ما جدول زیر را داریم که مدل **overfit** شده مدل خوبی نیست

Results for output field Outcome

Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	542	89.59%	122	78.21%
Wrong	63	10.41%	34	21.79%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		361	33
1		30	181
'Partition' = 2_Testing		0	1
0		85	18
1		16	37

در حالت bagging اگرچه ما دیگر overfit نداریم ولی پر قسمت تست مدل عملکرد خوبی در پیش بینی داده های دیابت مثبت نداشته است

Results for output field Outcome

Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	504	83.31%	125	80.13%
Wrong	101	16.69%	31	19.87%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		355	39
1		62	149

'Partition' = 2_Testing		0	1
0		93	10
1		21	32

پس به طور کلی این مدل نیز مدل خوبی نیست

• C&R Tree

در حالت تک درخت جدول زیر را داریم که همانطور که مشاهده می شود در این حالت نیز باز هم نتایج داده های تست بایاس شده روی صفر است و پیش بینی خوبی در حالت دیابت مثبت ندارد

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	461	76.2%	126	80.77%
Wrong	144	23.8%	30	19.23%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		348	46
1		98	113

'Partition' = 2_Testing		0	1
0		93	10
1		20	33

مشابه حالت قبل boosting هم بررسی شد که حالت overfit داشت ولی در حالت bagging

نتایج بهتری داشتیم به صورت زیر

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	498	82.31%	125	80.13%
Wrong	107	17.69%	31	19.87%
Total	605		156	

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

'Partition' = 1_Training		0	1
0		349	45
1		62	149

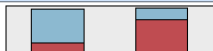
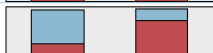

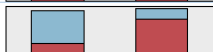
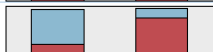
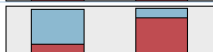



'Partition' = 2_Testing		0	1
0		89	14
1		17	36

پس به دلیل این که داده ها به طور واضحی حالت بایاس شده بر روی صفر را دارند بقیه ی روند را پس از متوازن کردن داده ها انجام می دهیم

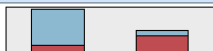


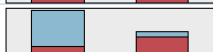
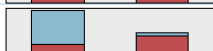
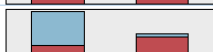
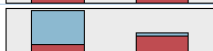

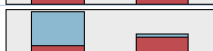
2. داده های دیابت منفی کاهش یافته

ابتدا در این قسمت اول با Auto Classifier ۱۰ داده برتر را بر اساس پیش بینی خود برنامه انتخاب میکنیم تنظیمات مورد نظرم را در آن اعمال میکنیم

این مدل به طور خودکار از این مدل ها مدلی با درصد های بالاتر را پیدا میکند

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	SVM 1	49	600.0	56	1.728	76.386	8	0.861
	SVM 37	49	600.0	56	1.728	76.386	8	0.861
	Neural Net 9	49	550.0	55	1.682	75.422	8	0.835
	Neural Net 2	49	585.000	48	1.696	77.349	8	0.847
	Neural Net 1	49	680.0	55	1.746	80.241	8	0.883
	Neural Net 3	49	680.0	55	1.746	80.241	8	0.883
	Neural Net 4	49	680.0	55	1.746	80.241	8	0.883
	Neural Net 7	49	680.0	55	1.746	80.241	8	0.883
	Neural Net 8	49	680.0	55	1.746	80.241	8	0.883

جدول بالا درصد های train و جدول پایین درصد های test را نشان می دهد

Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
	SVM 1	49	140.0	20	2.352	80.769	8	0.915
	SVM 37	49	140.0	Double click to view model details		80.769	8	0.915
	Neural Net 9	49	140.0	29	2.327	76.923	8	0.889
	Neural Net 2	49	145.0	22	2.289	79.487	8	0.918
	Neural Net 1	49	130.0	29	2.289	79.487	8	0.886
	Neural Net 3	49	130.0	29	2.289	79.487	8	0.886
	Neural Net 4	49	130.0	29	2.289	79.487	8	0.886
	Neural Net 7	49	130.0	29	2.289	79.487	8	0.886
	Neural Net 8	49	130.0	29	2.289	79.487	8	0.886

همان طور که می بینیم بهترین مدل را مدل SVM با درصد train ۷۶.۳۸ درصد و درصد test ۸۰.۷۶ درصد معرفی کرده حال ما با بررسی داده ها در تک تک مدل ها تلاش برای پیدا کردن مدل بهتر و یا مطمئن شدن از این نتیجه هستیم

مدل ها در این حالت

• SVM

در حالت RBF با $C=1$ و گاما 1

در این حالت نسبت به حالت قبل مشاهده می شود که هم مدل بسیار بهتر داده های دیابت مثبت را بهتر تشخیص داده و هم **overfit** نشده است

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	317	77.13%	129	82.69%
Wrong	94	22.87%	27	17.31%
Total	411		156	

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

'Partition' = 1_Training		0	1
0		152	48
1		46	165
'Partition' = 2_Testing		0	1
0		83	20
1		7	46

در حالت Polynomial با $C=5$ و گاما 1

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	315	75.9%	128	82.05%
Wrong	100	24.1%	28	17.95%
Total	415		156	

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

'Partition' = 1_Training		0	1
0		153	51
1		49	162
'Partition' = 2_Testing		0	1
0		86	17
1		11	42

در اینجا نیز نسبت به قبل میزان بیشتری تشخیص درست دیابت داشتیم

در حالت Sigmoid با $C=10$ و گاما 1

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	291	71.67%	109	69.87%
Wrong	115	28.33%	47	30.13%
Total	406		156	

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

'Partition' = 1_Training		0	1
0		100	95
1		20	191
'Partition' = 2_Testing		0	1
0		62	41
1		6	47

می بینیم که در این حالت میزان تشخیص درست دیابت بسیار بیشتر شده و مدل نیز **overfit** نشده حال با مقایسه با بقیه حالات باید ببینیم درصد کدام یک از مدل ها بهتر است

در حالت Linear با $C=1$

Results for output field Outcome

Comparing \$S-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	321	74.13%	126	80.77%
Wrong	112	25.87%	30	19.23%
Total	433		156	

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

'Partition' = 1_Training		0	1
0		168	54
1		58	153
'Partition' = 2_Testing		0	1
0		81	22
1		8	45

• KNN

همان طور که در پایین میبینید مدل نسبت با حالت بایاس نشده خیلی نتایج بهتری را برگردانده است

Results for output field Outcome

Comparing \$KNN-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	316	76.33%	127	81.41%
Wrong	98	23.67%	29	18.59%
Total	414		156	

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

'Partition' = 1_Training		0	1
0		160	43
1		55	156
'Partition' = 2_Testing		0	1
0		88	15
1		14	39

• Neural Net

Results for output field Outcome

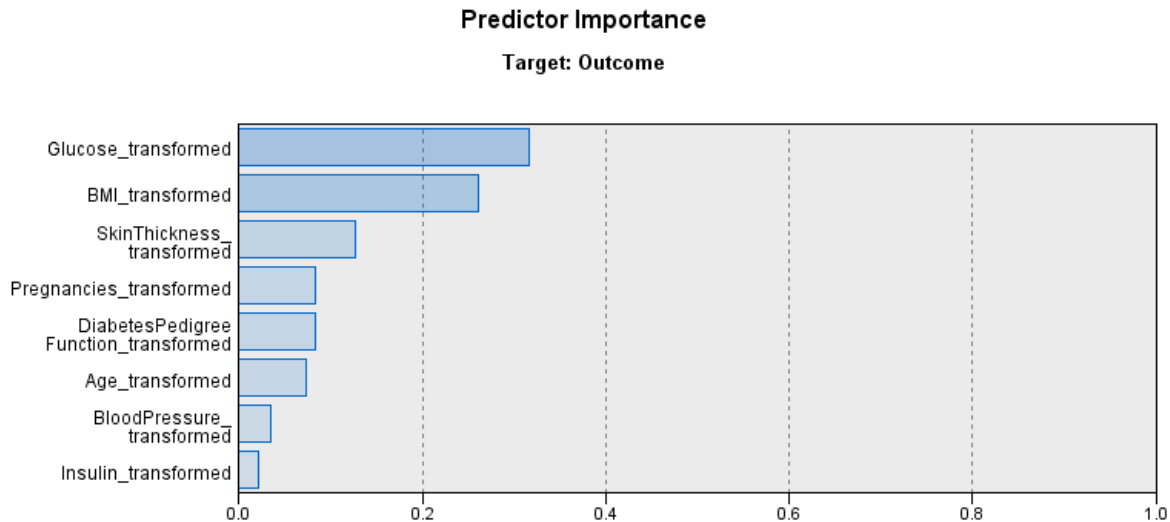
Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	316	74%	122	78.21%
Wrong	111	26%	34	21.79%
Total	427		156	

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

'Partition' = 1_Training		0	1
0		153	63
1		48	163
'Partition' = 2_Testing		0	1
0		79	24
1		10	43

در اینجا با کلیک بر روی مدل داریم



که در این حالت برای بهتر شدن تست مقدار insulin را حذف میکنیم اما با بررسی داده ها پس از این کار می بینیم با این که درصد accuracy بیشتر شد و داده های کمتری دچار تشخیص اشتباه شدند

Results for output field Outcome

Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	313	75.24%	128	82.05%
Wrong	103	24.76%	28	17.95%
Total	416		156	

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

'Partition' = 1_Training		0	1
0		147	58
1		45	166
'Partition' = 2_Testing		0	1
0		82	21
1		7	46

مدل شبکه عصبی در حالت boosting نیز تست شد که overfit داشت اما در حالت bagging نتیجه بهتر و به صورت زیر است

Results for output field Outcome

Comparing \$N-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	324	80%	121	77.56%
Wrong	81	20%	35	22.44%
Total	405		156	

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

'Partition' = 1_Training		0	1
0		150	44
1		37	174
'Partition' = 2_Testing		0	1
0		81	22
1		13	40

C&R Tree •

در حالت تک درخت جدول زیر را داریم که همانطور که مشاهده می شود در این حالت نیز نتایج خیلی مورد قبول تر است

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	313	74.17%	122	78.21%
Wrong	109	25.83%	34	21.79%
Total	422		156	

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

'Partition' = 1_Training		0	1
0		150	61
1		48	163
'Partition' = 2_Testing		0	1
0		80	23
1		11	42

برای این مدل در حالت boosting نتایج دچار overfit شد ولی در حالت bagging نتایج بهتر و به صورت زیر است

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	332	81.17%	126	80.77%
Wrong	77	18.83%	30	19.23%
Total	409		156	

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

'Partition' = 1_Training		0	1
0		159	39
1		38	173
'Partition' = 2_Testing		0	1
0		85	18
1		12	41

CHAID •

این مدل در حالت تک درخت نتیجه زیر را برگردانده که در مقایسه با حالت بدون توازن بسیار بهتر است

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	335	75.62%	130	83.33%
Wrong	108	24.38%	26	16.67%
Total	443		156	

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

'Partition' = 1_Training		0	1
0		171	61
1		47	164
'Partition' = 2_Testing		0	1
0		87	16
1		10	43

CHAID درحالت boosting دچار overfit شد ولی در حالت bagging نتیجه به صورت زیر بود که مدل accuracy بالاتری دارد

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	343	83.05%	121	77.56%
Wrong	70	16.95%	35	22.44%
Total	413		156	

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

'Partition' = 1_Training		0	1
0		153	49
1		21	190
'Partition' = 2_Testing		0	1
0		75	28
1		7	46

QUEST •

در مدل شبیه به دومدل درخت قبلی است و نتایج آن را به این صورت داریم که نسبت به دو مدل قبلی عملکرد ضعیف تری داشته است

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	306	71%	121	77.56%
Wrong	125	29%	35	22.44%
Total	431		156	

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

'Partition' = 1_Training		0	1
0		162	58
1		67	144
'Partition' = 2_Testing		0	1
0		84	19
1		16	37

در حالت bagging عملکرد این مدل به صورت زیر بهبود میابد

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	326	76.53%	118	75.64%
Wrong	100	23.47%	38	24.36%
Total	426		156	

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

'Partition' = 1_Training		0	1
0		146	69
1		31	180
'Partition' = 2_Testing		0	1
0		77	26
1		12	41

Random Tree •

این مدل نیز مانندهای قبلی است و عملکرد آن به صورت زیر است

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	339	80.91%	124	79.49%
Wrong	80	19.09%	32	20.51%
Total	419		156	

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

'Partition' = 1_Training		0	1
0		166	42
1		38	173
'Partition' = 2_Testing		0	1
0		80	23
1		9	44

C5 •

و آخرین مدل

Results for output field Outcome

Comparing \$C-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	309	76.11%	112	71.79%
Wrong	97	23.89%	44	28.21%
Total	406		156	

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

'Partition' = 1_Training		0	1
0		108	87
1		10	201
'Partition' = 2_Testing		0	1
0		64	39
1		5	48

نتیجه بعد از هرس کردن و در نهایت حالت boost شده مدل به صورت زیر است

Results for output field Outcome

Comparing \$C-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	343	82.25%	122	78.21%
Wrong	74	17.75%	34	21.79%
Total	417		156	

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

'Partition' = 1_Training		0	1
0		151	55
1		19	192
'Partition' = 2_Testing		0	1
0		79	24
1		10	43

انتخاب بهترین مدل

از بین مدل های بالا بخواهیم بهترین مدل ها را انتخاب کنیم به این ۴ مدل زیر می رسم

C&R Tree (1 در حالت bagging

bagging در حالت CHAID (2
Random Tree (3
C5 در حالت boosting (4

C5	Random Tree	CHAID	C&R Tree	مدل
82.25%	80.91%	83.05%	81.17%	درصد train
78.21%	79.49%	77.56%	80.77%	درصد test

از بین مدل های بالا دو مدل اول و سوم بیشترین روابستگی را دارد و اختلاف درصد بین train و test در این دو مدل کمتر است

برای تصمیم گیری بین این دو مدل به جدول اغتشاش این دو نگاه میکنیم و درصد accuracy که در این جا accuracy در C&R Tree بهتر است ولی در جدول اغتشاش Random Tree بهتر عمل کرده و تعداد بیشتری را در دیابت مثبت پیش بینی کرده

C&R Tree

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training	2_Testing
Correct	332 81.17%	126 80.77%
Wrong	77 18.83%	30 19.23%
Total	409	156

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

'Partition' = 1_Training	0	1
0	159	39
1	38	173
'Partition' = 2_Testing	0	1
0	85	18
1	12	41

Random Tree

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training	2_Testing
Correct	339 80.91%	124 79.49%
Wrong	80 19.09%	32 20.51%
Total	419	156

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

'Partition' = 1_Training	0	1
0	166	42
1	38	173
'Partition' = 2_Testing	0	1
0	80	23
1	9	44

برای انتخاب بهترین مدل باید داده های تست را تغییر بدهیم تا ببینیم باز هم مدل نتیجه مشابه ارایه میدهد یا خیر

نتیجه تغییر دادن seed برای C&R Tree

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training	2_Testing
Correct	342 81.82%	119 77.27%
Wrong	76 18.18%	35 22.73%
Total	418	154

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

'Partition' = 1_Training	0	1
0	161	45
1	31	181
'Partition' = 2_Testing	0	1
0	77	25
1	10	42

نتیجه تغییر دادن seed برای Random Tree

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training		2_Testing	
Correct	334	80.87%	115	74.68%
Wrong	79	19.13%	39	25.32%
Total	413		154	

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

'Partition' = 1_Training		0	1
0		166	35
1		44	168
'Partition' = 2_Testing		0	1
0		79	23
1		16	36

پس مدل برتر ما در این پروژه مدل C&R Tree است