



Diabetes Dataset (Cleaning Phase)

Arash Nikzad

Email: Arash1382N@gmail.com

Professor : Mohammadreza Mohtat

فهرست مطالب

3	مقدمه
4	مرحله اول : درک کسب و کار
5	مرحله دوم : درک داده ها
6	تعریف پارامتر های دیتاست
7	تجزیه و تحلیل اولیه داده ها
11	بررسی روابط و correlation فیچرها
12	مرحله سوم : پیش پردازش داده ها
12	ساختار بندی و مجتمع کردن داده ها
14	مدیریت داده های نویزی
15	شناسایی داده های پرت
18	مدیریت داده های مفقوده
20	نرمال و استاندارد سازی
21	مدیریت دادگان نامتوازن
22	مرحله چهارم : خروجی و ذخیره داده های پاکسازی شده

مقدمه

در این پروژه ما به بررسی داده ی دیابت می پردازیم و تلاش در پیاده سازی الگوریتم های یادگیری ماشین برای پیش بینی و وضعیت دیابت افراد متناسب با داده های آنها داریم. منبع داده ها یک دیتاست در kaggle بوده و برای پیاده سازی مدل های classification استفاده خواهد شد. در این پروژه از منطق CRISP-DM استفاده می کنیم و توضیحات و مراحل آن را گام به گام پیش می رویم همچنین در پیاده سازی این مراحل از ابزار IBM Modeler استفاده خواهیم کرد.

مرحله اول : درک کسب و کار

دیابت در رشته های علوم پزشکی به قاتل خاموش معروف است. در دنیای امروزه مشکل دیابت روز به روز در حال افزایش است و در سرتاسر جهان پخش شده است که تاثیر قابل توجهی در کاهش وضعیت سلامت انسان ها به صورت مستقیم یا غیر مستقیم داشته است. به دلیل تنوع دیابت بسیاری از اندام های بدن انسان دچار آسیب شده که ممکن است باعث بروز سکته قلبی - کوری - مشکلات کلیوی و ... شود. طبق اعلامیه ی سازمان جهانی بهداشت (WHO) در حال حاضر بیش از 400 میلیون نفر از دیابت در سرتاسر جهان رنج می برند.

گرفتاری به بیماری دیابت حاصل فاکتور های رژیمی گوناگون افراد است. دیابت نیز همانند بسیاری بیماری ها یک بیماری مزمن است که گام به گام در حال تاثیرگذاری بر روی انسان ها در جهان است که پیش بینی می شود در آینده این روند افزایش نیز پیدا کند

افزایش سطح قند در خون افراد نشان دهنده ی دیابت است. بیماری دیابت درمان قطعی و همیشگی ندارد اما می شود آن را کنترل و یا از آن پیشگیری کرد. مرض قند (Diabetes Mellitus) به دلیل ترشح غیر طبیعی انسولین در بدن شکل می گیرد که دارای 2 نوع دیابت نوع اول و دیابت نوع دوم است. که بیشتر افراد مبتلا به دیابت نوع 2 نسبت به نوع 1 می شوند.

بیشترین عوامل موثر در دیابت نوع دوم عبارتند از : وراثت - عادات غذایی و کمبود تحرک و ... در حالی که

باور بر این است که دیابت نوع اول به دلیل تخریب منطقی خود ایمنی جزایر لانگرهانس میزبان پانکراس است

مرحله دوم : درک داده ها

داده ها به فرمت CSV و کاملاً به زبان انگلیسی اند دارای 10 فیلد که 9 فیلد به عنوان فیچر و یک فیلد به

نام outcome که target مسئله است. داده ها همگی به صورت عددی و به تایپ continuous هستند که

در 768 رکورد قرار دارند

Field	Measurement	Values	Missing	Check	Role
Pregnancies	Continuous	[0,17]		None	Input
Glucose	Continuous	[0,199]		None	Input
BloodPressure...	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

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
727	1	118	78	29	180	36....	0.498	25	0
728	0	141	84	26	0	32....	0.433	22	0
729	2	175	88	0	0	22....	0.326	22	0
730	2	92	52	0	0	30....	0.141	22	0
731	3	130	78	23	79	28....	0.323	34	1
732	8	120	86	0	0	28....	0.259	22	1
733	2	174	88	37	120	44....	0.646	24	1
734	2	106	56	27	165	29....	0.426	22	0
735	2	105	75	0	0	23....	0.560	53	0
736	4	95	60	32	0	35....	0.284	28	0
737	0	126	86	27	120	27....	0.515	21	0
738	8	65	72	23	0	32....	0.600	42	0
739	2	99	60	17	160	36....	0.453	21	0
740	1	102	74	0	0	39....	0.293	42	1
741	11	120	80	37	150	42....	0.785	48	1
742	3	102	44	20	94	30....	0.400	26	0
743	1	109	58	18	116	28....	0.219	22	0

تعریف پارامترهای دیتاست

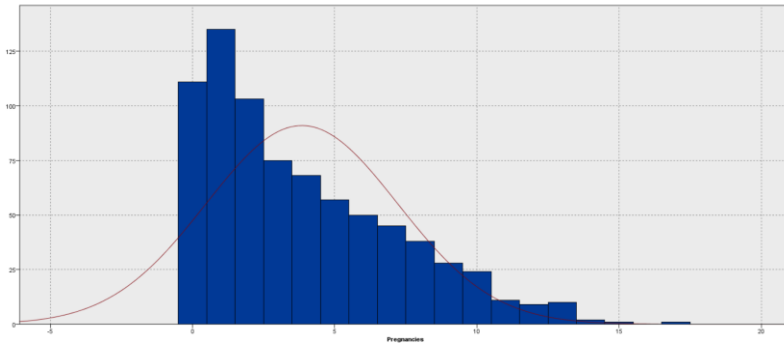
- **Triceps Skinfold Thickness** : ضخامت پوست در درجه اول توسط میزان کالژن تعیین می شود و در دیابت وابسته به انسولین (نوع 1) افزایش می یابد.
- **Diabetes Pedigree Function** : احتمال ابتال به دیابت را بر اساس سابقه خانوادگی امتیاز می دهد
- **Pregnancies** : تعداد حاملگی فرد را نشان می دهد. شیوع ابتال به دیابت در زنان با حاملگی بیش از 4 بار بیشتر است
- **Glucose** : میزان قند خون در تست قند 2 ساعته را نشان می دهد که ارتباط مستقیمی با دیابت دارد. مقدار مجاز گلوکز برای این آزمایش بین 70 تا 140 واحد است.
- **BMI** : احتمال و شدت دیابت نوع 2 ارتباط نزدیکی با شاخص توده بدنی دارد. اما افراد با هر شکل و اندازه و وزن ممکن است به دیابت مبتال شوند.
- **2-Hour Serum Insulin** : تست 2 ساعته انسولین: این تست تحمل گلوکز 2 ساعته با سطوح انسولین برای ارزیابی نحوه پردازش گلوکز توسط فرد و نحوه واکنش انسولین در بدن به این سطوح گلوکز استفاده می شود. این آزمایش معمولاً زمانی تجویز می شود که فردی ممکن است در معرض خطر ابتال به دیابت باشد یا قبلاً سطح گلوکز بالایی داشته است
- **Diastolic Blood Pressure** : فشار خون دیاستولیک: فشار در شریان ها را زمانی که قلب بین ضربان ها استراحت می کند اندازه گیری می کند. درافراد دیابتی، فشار سیستولیک تقریباً چهار واحد بیشتر از فشار دیاستولیک با توجه به فشار نرمال توصیه شده وجود دارد. یعنی اگر فشار نرمال 130/80 است، فشار فرد دیابتی در محدوده 170/80 قرار دارد
- **Age** : سن به عنوان عامل مهمی در دیابت شناخته شده است زیرا غلظت گلوکز خون با بالرفتن سن افزایش می یابد

تجزیه و تحلیل اولیه داده ها (EDA)

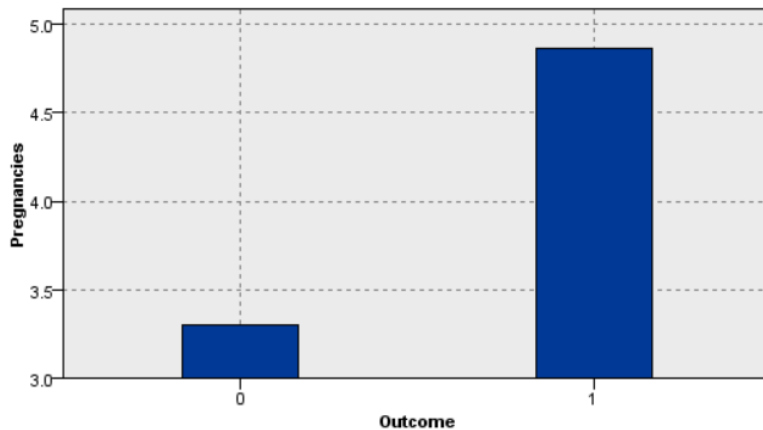
در این بخش به تحلیل اولیه و شهودی نمودار ها برای درک بهتر از داده ها می پردازیم

Pregnancies

- مطابق با نمودار هیستوگرام از داده های بارداری و نمودار توزیع نرمال بدیهی است که این فیچر دارای توضیح نرمال نیست و چولگی به سمت راست دارد

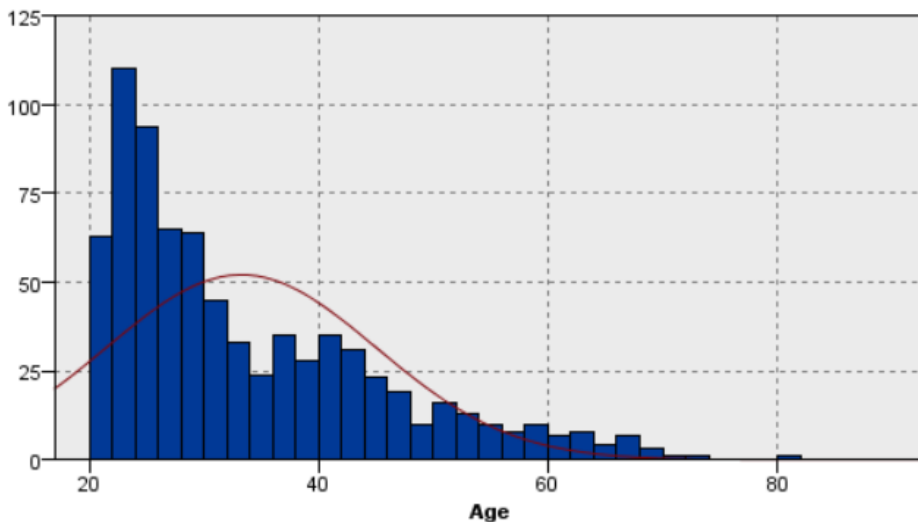


- و مطابق با این نمودار می توان دریافت که افراد باردار بیشتر تحت تاثیر دیابت قرار می گیرند



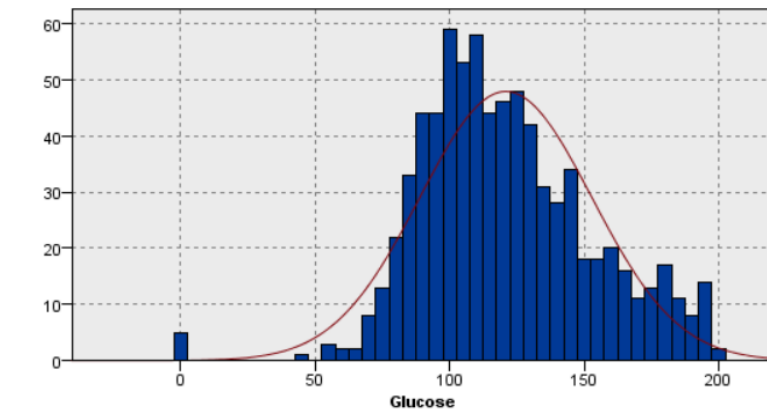
Age

- نمودار سن افراد دارای توزیع نرمال نیست و همچنین دارای تعداد ناچیزی داده پرت است. این داده همان طور که مشاهده می شود برخلاف عموم فیلد ها دارای مقدار 0 نیست که به معنای نداشتن داده های مفقوده است

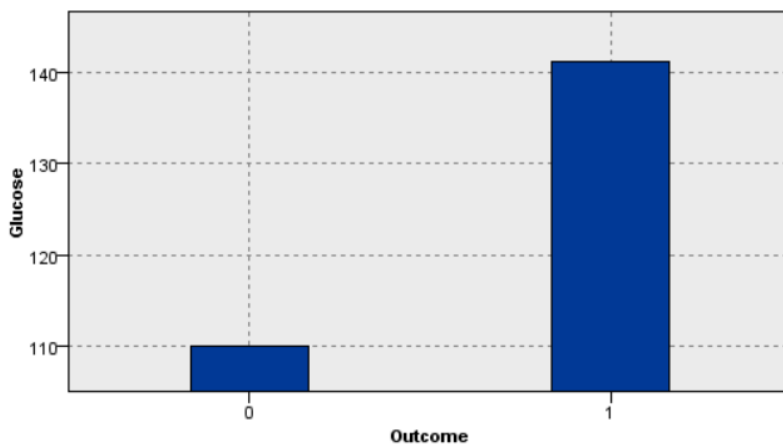


Glucose

- مطابق با هیستوگرام اینطور به نظر می رسد که گلوکز خون از توزیع نسبتاً نرمالی پیروی می کند

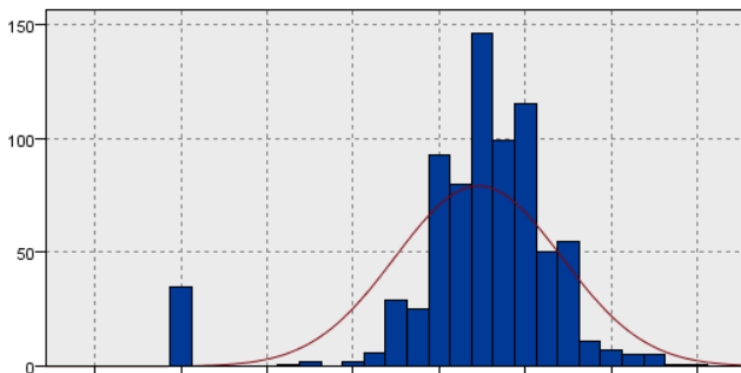


- مطابق با نمودار مقابل می توان دریافت که افزایش گلوکز خون رابطه قابل توجهی با ابتلا به دیابت دارد

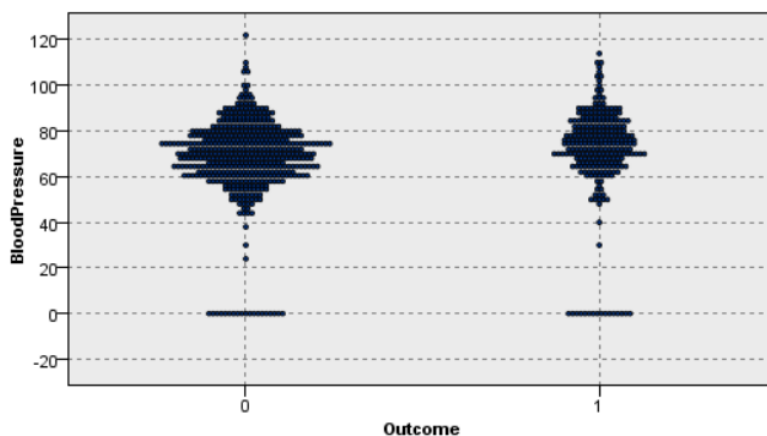


Blood Pressure

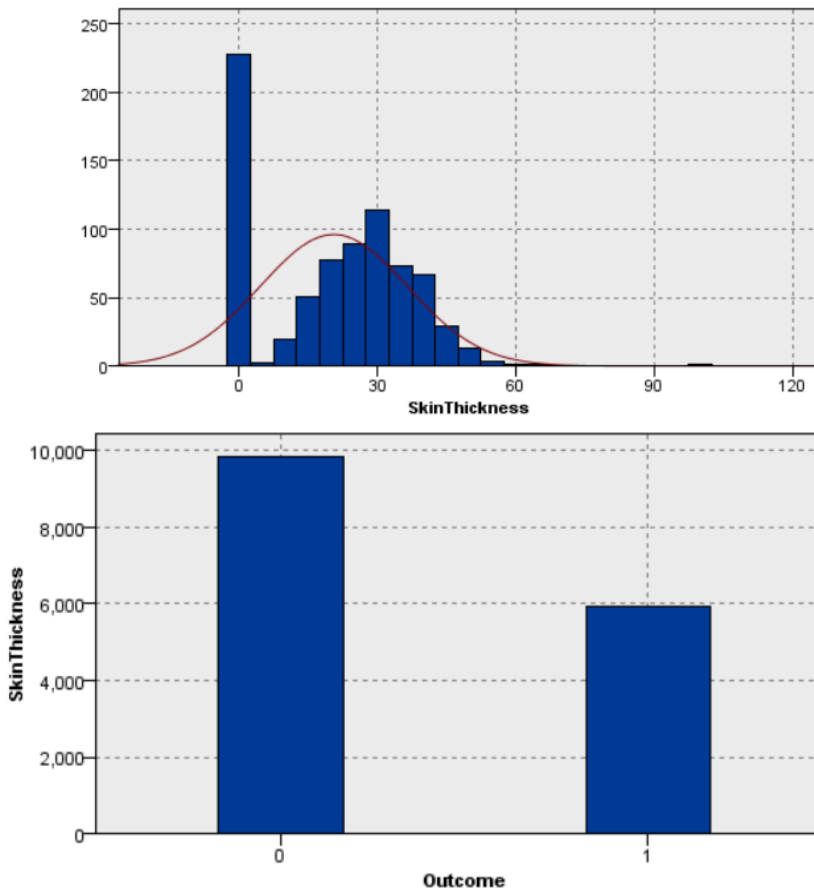
- داده فشار خون نیز از توزیع نرمال پیروی می کند اما دارای تعدادی زیاد مقدار 0 است که داده های نویز و یا مفقوده است



- از این نمودار نیست می توان دریافت که فشار خون در هر دو گروه افراد سالم و مبتلا به دیابت دارای توزیع نرمال است



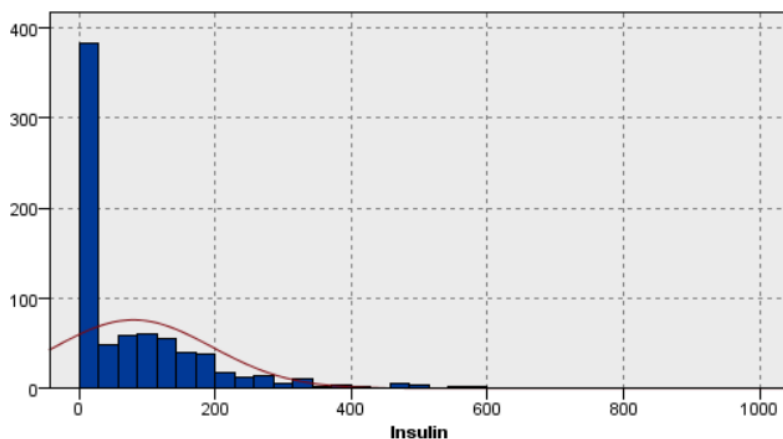
Skin Thickness



- همانند داده های قبلی ضخامت پوست نیز از توزیعی نسبتاً نرمال پیروی می کند اما دارای تعداد زیادی داده 0 است که داده های مفقوده ما در این دیتاست هستند

- همان طور که مشاهده می شود این داده ها به دلیل داشتن مقادیر مفقوده ی زیاد بر روی نمودار تاثیر گذاشته و داده هارا نامتوازن کرده است

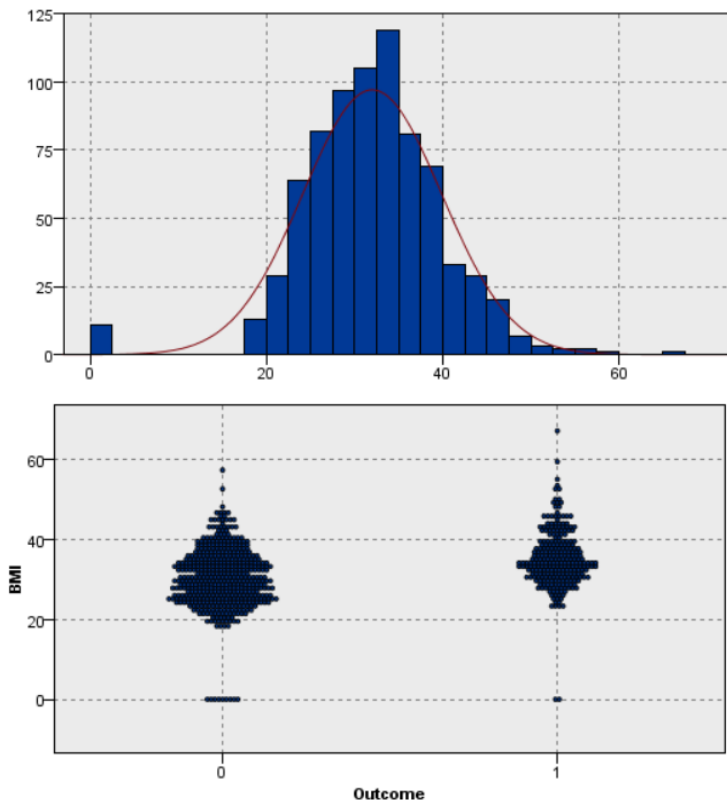
Insulin



- در داده های انسولین مشاهده می شود به دلیل داشتن تعداد زیاد مقادیر مفقوده تحلیل نمودار آن اطلاعات درستی به ما نخواهد داد و در ادامه باید متناسب با correlation آن با فیلد هدف درباره آن تصمیم گیری کرد

BMI

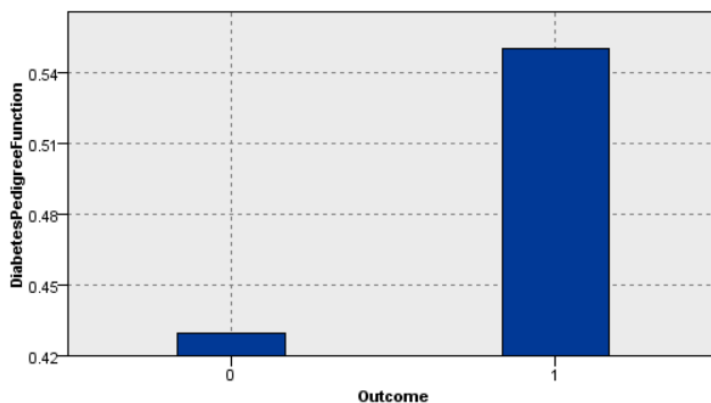
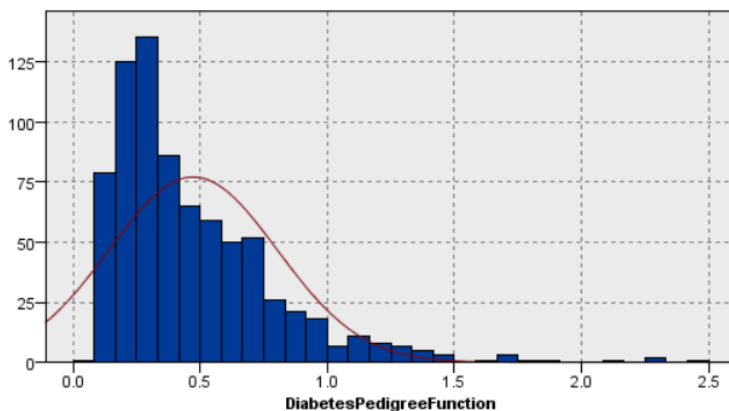
- نمودار bmi که تا حدی از توزیع نرمال پیروی می کند و در کنار داده های مفقوده 0 مشاهده می شود که داده های دور افتاده ای نیز در سمت راست دارد که باید برطرف شوند



- همان طور که از قبل اطلاع داشتیم مشاهده می شود که افرادی که به دیابت مبتلا شده اند به طور کلی دارای bmi بیشتری نیز هستند و عموم افراد با bmi نرمال در بخش افراد سالم قرار دارند

Diabetes Pedigree Function

- تابع شجرنامه دیابت از توضیح نرمال پیروی نمی کند و چولگی به سمت راست دارد که می تواند به دلیل وجود داده های پرت زیاد آن در سمت باشد



- همچنین می توان مشاهده کرد که به طور میانگین افرادی که تابع شجرنامه آنها بیشتر بوده یعنی به صورت ارثی این بیماری را داشته اند بیشتر به دیابت مبتلا شده اند

بررسی روابط و correlation فیچرها

	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 روابط از داده های پرت استفاده نشده است

خوشبختانه مطابق با جدول روابط فیچرها عموماً ارتباط بالایی با یکدیگر ندارد و ما را درگیر مشکل multicollinearity نخواهد کرد. اما در بین این روابط مشاهده می شود که انسولین با گلوکز رابطه ی نسبتاً بالایی دارد همچنین در مراحل قبل مشاهده شد که بسیاری از داده های انسولین مفقوده هستند پس یکی از روش هایی که میتوان پیش گرفت حذف این فیلد در مراحل بعدی است. ضخامت پوست و bmi نیز دارای رابطه ای نسبتاً بالا هستند اما در صورت اهمیت بالای هردو می توان از آن چشم پوشی کرد

"البته شایان ذکر است که این مقادیر به دلیل داشتن داده های پرت و نویز ممکن است دقیق نباشد و به صورت خیلی

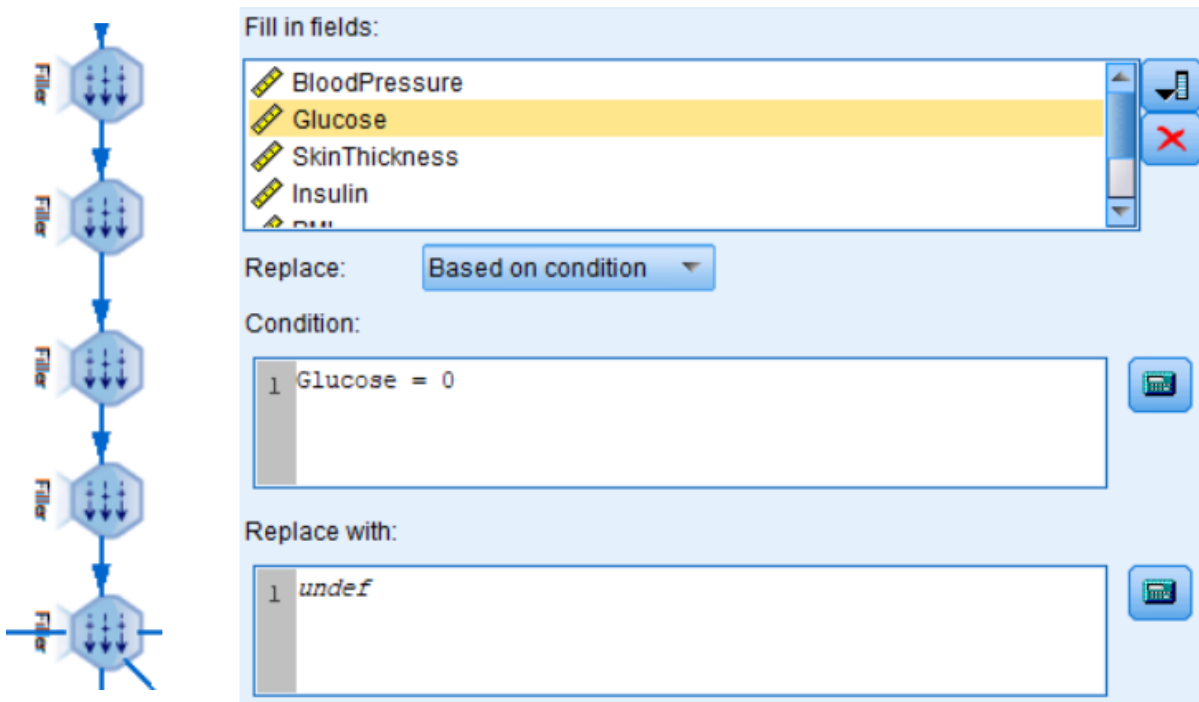
"دقیق نمی توان از آن استناد کرد.

مرحله سوم: پیش پردازش داده ها

ساختار بندی و مجتمع کردن داده ها

داده های ما همگی در یک فایل CSV قرار دارند که در 10 ستون و 768 سطر است به همین خاطر نیازی به مجتمع کردن داده ها در این مرحله نیست. همچنین تمام داده های ما عددی بوده و از نوع continuous هستند اما همان طور که در بخش قبل مشاهده کردیم در این دیتا ست داده های مفقود به عنوان مقدار 0 ست شده است پس برای ادامه ی کار نیاز است در ابتدا این داده ها را با مقدار null که ابزار IBM با عنوان (undef) قرار دارد جایگزین

کنیم



با این تغییر مقادیر 0 به مقادیر null در داده های ما تغییر پیدا کرد. (این عمل بر روی فیلد های سن و بارداری انجام نشد زیرا داده ی سن ما دارای مقدار 0 نبود ولی داده های بارداری چون مقدار 0 مقداری معنادار بود و متناسب با توزیع آن وجود مقادیر 0 به عنوان یک داده واقعی منطقی تر بود آن را تغییر ندادیم)

داده های جدید به صورت زیر خواهند بود

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	\$null\$	33....	0.627	50	1
2	1	85	66	29	\$null\$	26....	0.351	31	0
3	8	183	64	\$null\$	\$null\$	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	\$null\$	\$null\$	25....	0.201	30	0
7	3	78	50	32	88	31....	0.248	26	1
8	10	115	\$null\$	\$null\$	\$null\$	35....	0.134	29	0
9	2	197	70	45	543	30....	0.158	53	1
10	8	125	96	\$null\$	\$null\$	\$n....	0.232	54	1
11	4	110	92	\$null\$	\$null\$	37....	0.191	30	0
12	10	168	74	\$null\$	\$null\$	38....	0.537	34	1
13	10	139	80	\$null\$	\$null\$	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	\$null\$	\$null\$	\$null\$	30....	0.484	32	1
17	0	118	84	47	230	45....	0.551	31	1
18	7	107	74	\$null\$	\$null\$	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	\$null\$	\$null\$	35....	0.388	50	0
23	7	196	90	\$null\$	\$null\$	39....	0.451	41	1
24	9	119	80	35	\$null\$	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	\$null\$	\$null\$	39....	0.257	43	1
28	1	97	66	15	140	23....	0.487	22	0
29	13	145	82	19	110	22....	0.245	57	0
30	5	117	92	\$null\$	\$null\$	34....	0.337	38	0
31	5	109	75	26	\$null\$	36....	0.546	60	0
32	3	158	76	36	245	31....	0.851	28	1
33	3	88	58	11	54	24....	0.267	22	0
34	6	92	92	\$null\$	\$null\$	19....	0.188	28	0
35	10	122	78	31	\$null\$	27....	0.512	45	0
36	4	103	60	33	192	24....	0.966	33	0
37	11	138	76	\$null\$	\$null\$	33....	0.420	35	0
38	9	102	76	37	\$null\$	32....	0.665	46	1
39	2	90	68	42	\$null\$	38....	0.503	27	1
40	4	111	72	47	207	37....	1.390	56	1
41	3	180	64	25	70	34....	0.271	26	0

مدیریت داده های نویزی

همان طور که مشاهده شد داده های اصلی که به عنوان نویز و یا داده های مفقوده بودند در مرحله قبلی به

null تبدیل شد حال برای بررسی وجود noise ها می توان از مفاهیم هر یک از پارامتر ها استفاده کرد. اما مطابق با

تعاریف آنها این داده ها از نظر علمی مقادیر ماکسیمم دقیقی ندارند ولی بازه های منطقی و میانگین دارند که برای

بررسی داده های غیر طبیعی و دور افتاده در مرحله ی بعد باید اقدام کرد.

Field	Measurement	Values	Missing	Check	Role
Pregnancies	Continuous	[0,17]		None	Input
Glucose	Continuous	[44,199]		None	Input
BloodPressu...	Continuous	[24,122]		None	Input
SkinThickness	Continuous	[7,99]		None	Input
Insulin	Continuous	[14,846]		None	Input
BMI	Continuous	[18.2,67.1]		None	Input
DiabetesPed...	Continuous	[0.078,2.42]		None	Input
Age	Continuous	[21,81]		None	Input
Outcome	Flag	1/0		None	Target

همان طور که می توان مشاهده کرد داده ها همگی در رنج واقعی خود هستند که در این صورت داده ای الزاما

نویز نخواهیم داشت و بهتر است به این داده ها را در بخش مدیریت داده های پرت بپردازیم

شناسایی داده های پرت

برای شناسایی داده های پرت آنها را مطابق با توزیعشان به دو دسته نرمال و غیر نرمال تقسیم می کنیم. سپس

داده ها با توزیع نرمال را با کمک Z-SCORE و بقیه توزیع هارا با کمک IQR مدیریت می کنیم.

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...	

همان طور که مشاهده می شود دو فیچر BloodPressure و BMI از توزیع نرمال پیروی می کنند و

سایر موارد غیر نرمال هستند.

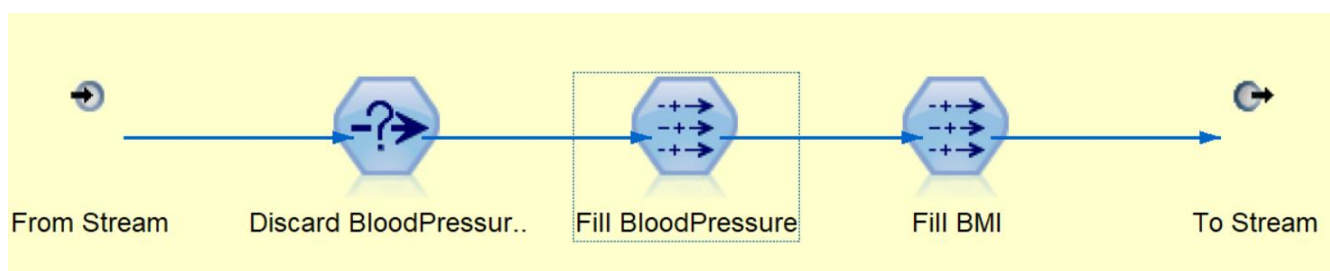
مدیریت داده های پرت با Z-score

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 None		Never	Fixed	94.792	
SkinThickness	Continuous	1	1 None		Never	Fixed	69.792	
Insulin	Continuous	7	1 None		Never	Fixed	51.172	
BMI	Continuous	3	1 None		Never	Fixed	97.917	
DiabetesPed...	Continuous	7	4 None		Never	Fixed	99.349	
Age	Continuous	5	0 None		Never	Fixed	99.349	
Outcome	Flag	--	--		Never	Fixed	100	

با کمک data audit اطلاع بالا را به دست آوردیم و داده های پرت را با حد 3 و داده های خیلی پرت را

با حد 5 مشخص کردیم. شایان ذکر است که اطلاعاتی که درباره تعداد outlier ها و extreme ها هست فقط

برای داده هایی قابل استناد هستند که از توزیع نرمال پیروی می کنند.



مدیریت داده های پرت با IQR

حال داده های پرت بقیه فیچر ها را با توزیع غیر نرمال با کمک IQR شنا سایی و سپس هندل می کنیم. اما قبل از آن به دلیل داشتن correlation بالای insulin با glucose و فراوانی داده های مفقوده و پرت در این فیلد و همچنین ارتباط نه چندان بالای آن با فیلد هدف آن را دراپ کردیم.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Val
Pregnancies	Continuous	4	0 None		Never	Fixed	100	
Glucose	Continuous	0	0 None		Never	Fixed	100	
BloodPressu...	Continuous	14	0 None		Never	Fixed	100	
SkinThickness	Continuous	2	1 None		Never	Fixed	73.352	
BMI	Continuous	6	0 None		Never	Fixed	99.313	
DiabetesPed...	Continuous	24	5 None		Never	Fixed	100	
Age	Continuous	8	0 None		Never	Fixed	100	
Outcome	Flag	--	--		Never	Fixed	100	

همان طور که مشاهده می شود داده ها دارای outlier هستند و باید هر فیلد را متناسب با ویژگی های آن مدیریت کرد.

- **Pregnancies**: این فیچر فقط دارای 4 داده پرت است که مطابق دید اولیه ما نسبت به داده ها مقادیر

بسیار تعداد بارداری بود با توجه به تعداد کم داده های پرت می توان با **coerce** کردن داده های پرت آن بدون ایجاد تغییر در توزیعش آن را مدیریت کرد.

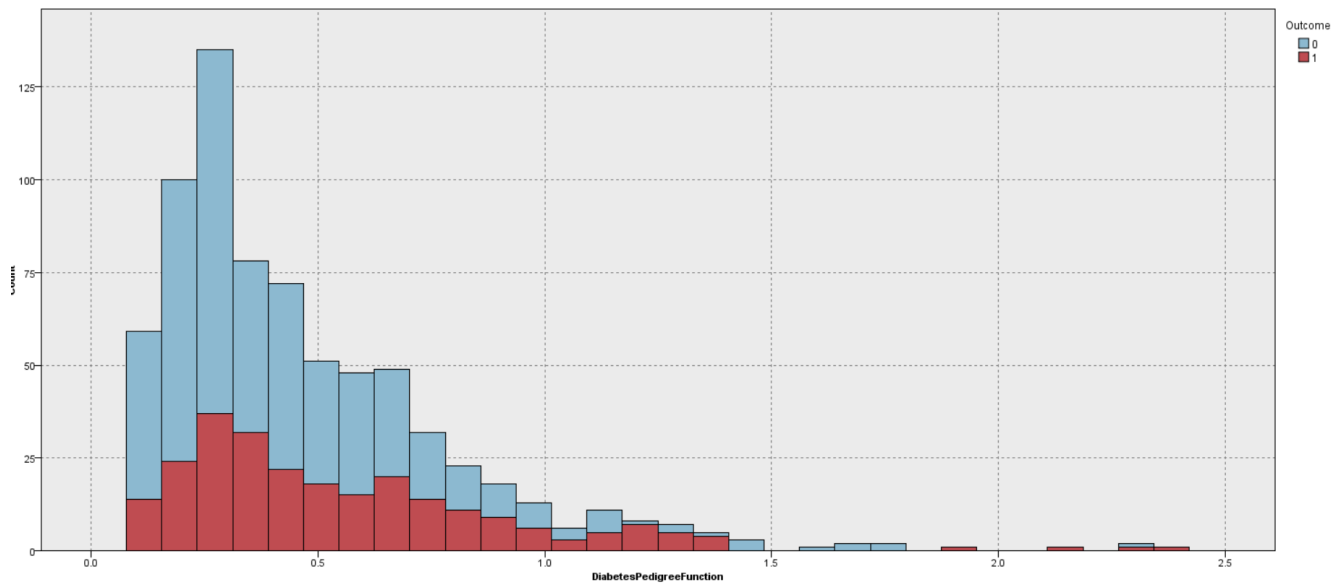
- **Glucose**: همان طور که مشاهده می شود این فیچر دارای هیچ داده پرتی نیست و نیازمند به مدیریت در این بخش نیست

- **SkinThickness**: این فیلد دارای 2 داده پرت و یک داده بسیار پرت است. با توجه به تعداد زیاد داده

های گمشده ما در این فیچر باید آنها را با کمک الگوریتم مدیریت کنیم. از این رو با **coerce** کردن داده های پرت و **nullify** کردن داده های بسیار پرت در این مرحله آن را هندل و در مراحل بعدی به آن میپردازیم.

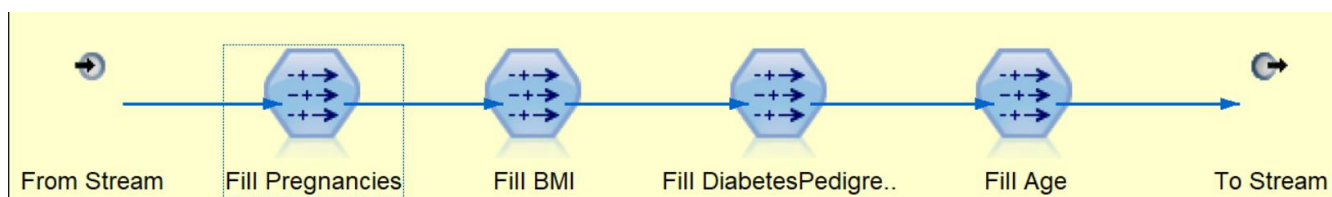
- **Diabetes Pedigree Function**: بیشترین تعداد داده های پرت ما در این فیچر قرار دارد که باید با

توجه بیشتری داده های پرت آن را مدیریت کنیم.



همان طور که مشاهده می شود داده های پرت و بسیار پرت در هر دو کلاس نهایی وجود دارند اما داده های خیلی پرت در کلاس 1 بیشتر اند. در این داده ها اگر هر دو مجموعه داده های پرت و خیلی پرت را `coerce` کنیم به دلیل تعداد رکورد زیاد از تغییر توزیع آن جلوگیری کند همچنین به دلیل اهمیت این فیلد بهتر است داده های آن را تا حد ممکن `discard` نکنیم و از تمام رکورد ها در ادامه فرایند استفاده کنیم.

- Age: همچنین این فیچر با داشتن تنها 8 داده ی پرت و فاقد داده خیلی پرت به راحتی با `coerce` کردن آن می توان به داده های تمیز تری رسید و همان طور که در دید اولیه به داده ها دریافتیم فاصله ی داده های پرت و حتی خیلی پرت با داده های نرمال ما بسیار زیاد نیست که در این صورت با `coerce` کردن آنها تغییر خاصی ایجاد نخواهد شد



مدیریت داده های مفقوده

حال با شناسایی و مدیریت داده های پرت می توان به بررسی و مدیریت داده های مفقوده پرداخت که در

مراحل قبلی از تبدیل 0 های نویزی به null و تعدادی از داده های فوق پرت در مرحله قبلی به دست آمده اند.

Field	Measurement	% Complete ▲	Valid Records	Null Value
# SkinThickness	Continuous	73.352	534	194
♦ BMI	Continuous	99.313	723	5
# DiabetesPed...	Continuous	99.313	723	5
# Pregnancies	Continuous	100	728	0
# Glucose	Continuous	100	728	0
# BloodPressu...	Continuous	100	728	0
# Age	Continuous	100	728	0
♦ Outcome	Flag	100	728	0

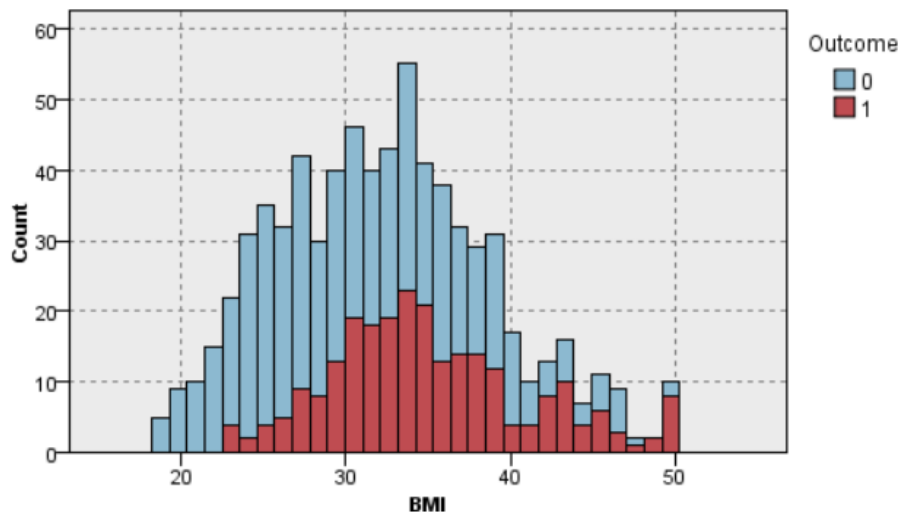
داده های ما از نظر داشتن داده های مفقوده در عموم فیچر ها از شرایط خوبی به سر می برند و فقط دو داده

فیچر دارای 5 داده پرت اما فیچر skin thickness دارای تعداد قابل توجهی از داده های مفقوده هستند که بهتر

است به جای پر کردن آن با مقادیر ثابت و یا اماری از روش های الگوریتمی و imputation استفاده کنیم.

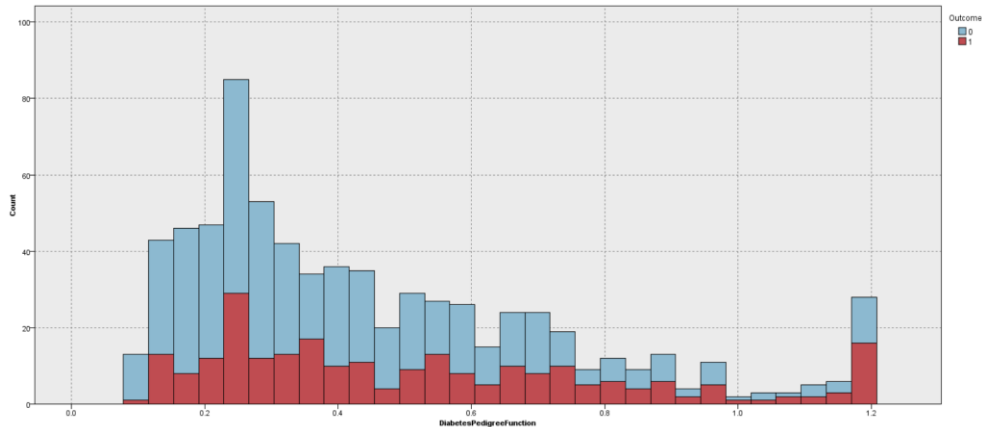
● **BMI** : این فیلد دارای 5 داده مفقوده است و متناسب با توزیع نرمالی که داشت می توان 5 داده را با 5

مقدار رندوم که از توزیع نرمال پیروی می کنند پر کرد.



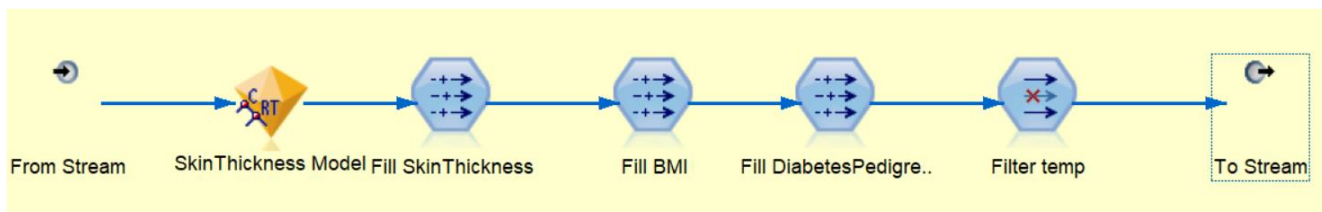
• DiabetesPedigreeFunction: این فیلد نیز فقط دارای 5 داده مفقوده است و با توجه

به توزیع آن میتوان دید که عموم داده ها حول میانه آن قرار دارند ازین رو استفاده از میانگین برای مدیریت داده های مفقوده گزینه ی خوبی است.



• SkinThickness: این فیلد برخلاف بقیه فیلد ها به دلیل حجم بسیار زیاد داده های مفقوده و

نداشتن توزیع نرمال استفاده از مقادیر ثابت و یا مقادیر رندوم می تواند به توزیع پایه ی آن آسیب بزند و بهترین راه مدیریت آن استفاده از الگوریتم است تا بتواند مطابق با بقیه داده ها مقادیر مناسبی را برای آن جایگزین کند.



نرمال و استاندارد سازی


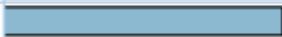
با پاکسازی داده ها نوبت به نرمال سازی آنها می رسد. البته در صورت وجود داده های کیفی باید ابتدا آن ها را به مقادیر کمی تبدیل کنیم و سپس نرمال سازی و scale کردن داده ها را انجام دهیم که این عمل در ادامه در الگوریتم هایی با مبنای فاصله هستند (distance base) مفید و ضروری خواهد بود که در این مرحله ما با min-max transformation داده های همه ی فیلد ها را در بازه 0 تا 100 قرار می دهیم.

	Pregnancies_transformed	Glucose_transformed	BloodPressure_transformed	SkinThickness_transformed	BMI_transformed	DiabetesPedigreeFunction_transformed	Age_transformed	Outcome
1	44.444	67.097	49.410	30.435	48.050	48.606	63.736	1
2	7.407	26.452	41.337	23.913	26.209	24.170	21.978	0
3	59.259	89.677	38.646	13.373	15.913	52.590	24.176	1
4	7.407	29.032	41.337	17.391	30.889	7.880	0.000	0
5	0.000	60.000	6.351	30.435	77.691	88.508	26.374	1
6	37.037	46.452	52.102	13.373	23.089	10.890	19.780	0
7	22.222	21.935	19.807	27.174	39.938	15.051	10.989	1
8	14.815	98.710	46.719	41.304	38.378	7.083	70.330	1
9	59.259	52.258	81.705	32.352	44.243	13.634	72.527	1
10	29.630	42.581	76.322	32.352	60.530	10.004	19.780	0
11	74.074	80.000	52.102	32.352	61.778	40.637	28.571	1
12	74.074	61.290	60.175	13.373	27.769	100.000	79.121	0
13	7.407	93.548	33.263	17.391	37.129	28.331	83.516	1
14	37.037	78.710	49.410	13.043	23.713	45.064	65.934	1
15	0.000	47.742	65.557	43.478	86.115	41.877	21.978	1
16	51.852	40.645	52.102	19.150	35.569	15.582	21.978	1
17	7.407	38.065	-0.000	33.696	78.315	9.296	26.374	0
18	7.407	45.806	46.719	25.000	51.170	39.929	24.176	1
19	22.222	52.903	70.940	36.957	65.835	55.423	13.187	0
20	59.259	35.484	65.557	32.352	53.666	27.446	63.736	0
21	51.852	98.065	73.631	32.352	67.395	33.023	43.956	1
22	66.667	48.387	60.175	30.435	33.697	16.379	17.582	1

همانطور که مشاهده می شود حال همه ی داده ها در بازه 0 تا 100 قرار گرفته اند و آماده ی استفاده در الگوریتم های distance base هستند اما قبل از بهتر است نگاهی به وضعیت تقسیم داده ها در کلاس های هدف بکنیم و در صورت نامتوازن بودن آنها آن را برطرف کنیم.

مدیریت دادگان نامتوازن

در این مرحله که به عنوان مرحله آخر پیش پردازش داده های ما یاد می شود به بررسی تقسیم record ها در بین کلاس های جواب توجه می کنیم. در صورت نامتوازن بودن آنها با کمک متد های over-sampling و یا under-sampling که متد های ساده اما کاربردی و مفید هستند در برطرف کردن این مشکل استفاده می کنیم.

Table	Graph	Annotations	
Value	Proportion	%	Count ▾
0		65.66	478
1		34.34	250

مطابق داده ی بالا می توان دید که داده ها کاملاً نامتوازن بوده و در این صورت اهمیت کلاس 1 که در واقع کلاس افرادی است که به دیابت مبتلا شده اند کمتر می شود که می تواند precision و recall مدل نهایی ما را کم بکند. با توجه به داشتن فقط 760 رکورد در صورت استفاده از متد under-sampling تعداد داده های ما بسیار کم می شود که در مراحل بعد در modeling پروژه می توانیم به مشکل بربخوریم ازین رو از متد over-sampling استفاده می کنیم و به داده های کلاس 1 وزن بیشتری می دهیم تا داده ها متوازن شوند.

Factor	Condition
1.0	Outcome = 0
1.912	Outcome = 1

حال با وزندهی جدید تعداد داده های بیشتری از کلاس 1 متناسب با وزن آن تولید می شود و این عمل باعث متعادل کردن داده های ما می شود.

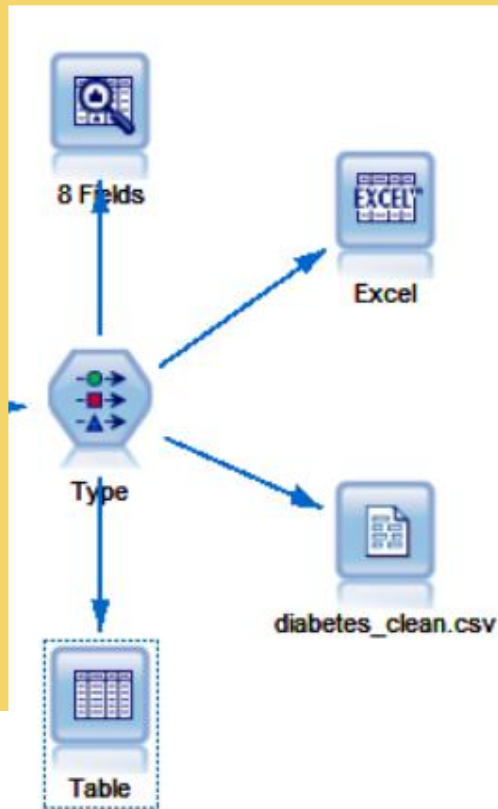
مرحله چهارم : خروجی و ذخیره داده های پاکسازی شده

با اتمام مراحل پیش پردازش داده ها حال نیاز است آنها را ذخیره سازی کنیم تا در مرحله modeling با

ابزار های گوناگون بتوانیم بر روی داده های پاکسازی شده کار کنیم. در این پروژه ما داده ها را به دو صورت فایل

excel و csv خروجی میگیریم. همچنین مشاهده می شود که بعد از مرحله توازن داده ها تعداد record ها به

955 افزایش یافته است.



Outcome	Pregnancie	Glucose_tr	BloodPressure	SkinThickn	BMI_trans	DiabetesP	Age_transf
1	44.44444	67.09677	49.41037873	30.43478	48.04992	48.60558	63.73626
1	44.44444	67.09677	49.41037873	30.43478	48.04992	48.60558	63.73626
0	7.407407	26.45161	41.33681852	23.91304	26.20905	24.16999	21.97802
1	59.25926	89.67742	38.64563179	13.37345	15.91264	52.58964	24.17582
1	59.25926	89.67742	38.64563179	13.37345	15.91264	52.58964	24.17582
0	7.407407	29.03226	41.33681852	17.3913	30.88924	7.879593	0
1	0	60	6.351390966	30.43478	77.69111	37.29091	26.37363
1	0	60	6.351390966	30.43478	77.69111	37.29091	26.37363
0	37.03704	46.45161	52.10156547	13.37345	23.08892	10.88977	19.78022
1	22.22222	21.93548	19.80732464	27.17391	39.9376	15.05091	10.98901
1	22.22222	21.93548	19.80732464	27.17391	39.9376	15.05091	10.98901
1	14.81481	98.70968	46.719192	41.30435	38.37754	7.08278	70.32967
1	14.81481	98.70968	46.719192	41.30435	38.37754	7.08278	70.32967
1	59.25926	52.25806	81.70461955	32.35193	44.24337	13.63435	72.52747
1	59.25926	52.25806	81.70461955	32.35193	44.24337	13.63435	72.52747
0	29.62963	42.58065	76.32224608	32.35193	60.53042	10.00443	19.78022
1	74.07407	80	52.10156547	32.35193	61.77847	40.63745	28.57143
1	74.07407	80	52.10156547	32.35193	61.77847	40.63745	28.57143
0	74.07407	61.29032	60.17512567	13.37345	27.76911	100	79.12088
1	7.407407	93.54839	33.26325832	17.3913	37.12949	28.33112	83.51648
1	7.407407	93.54839	33.26325832	17.3913	37.12949	28.33112	83.51648
1	37.03704	78.70968	49.41037873	13.04348	23.71295	45.06419	65.93407
1	37.03704	78.70968	49.41037873	13.04348	23.71295	45.06419	65.93407
1	0	47.74194	65.55749914	43.47826	86.11544	41.87694	21.97802
1	0	47.74194	65.55749914	43.47826	86.11544	41.87694	21.97802
1	51.85185	40.64516	52.10156547	19.14962	35.56942	15.58212	21.97802
1	51.85185	40.64516	52.10156547	19.14962	35.56942	15.58212	21.97802