

به نام خدا



دوره جامع علم داده دانشگاه تهران (کد ۲۷)

پیش بینی تشخیص دیابت

پروژه دوره مبانی و مفاهیم علم داده

دانش پذیر:

بهزاد دانش

استاد:

محمد رضا محتاط

اسفند ۱۴۰۲ – فروردین ۱۴۰۳

چکیده:

در این پژوهش گام های مربوط به علم داده و یادگیری ماشین در پروژه تشخیص دیابت بیماران آزمایشگاه هند پیاده سازی شده است. این پروژه دارای دیتاست با تعداد ۷۶۸ رکورد و ۸ ستون ورودی و یک ستون خروجی می باشد. با توجه به اطلاعاتی که از این دیتاست در وب سایت ها در اختیار است، میزان دقت این مسئله مابین ۷۵ تا ۸۳ درصد می باشد که در این پروژه با استفاده از مدل CART به دقت ۸۲ درصد رسیدیم. در اینجا طبق مراحل کریسپ هر فصل را جدا کردیم. در فصل یک به درک کسب و کار پرداخته و بیماری دیابت را تعریف و مورد بررسی قرار دادیم. در فصل دو به شناسایی داده و درک آن پرداختیم و هر ویژگی را مورد بررسی قرار دادیم. فصل سه به آماده سازی داده پرداخته و مراحل لازم جهت پاکسازی داده ها را انجام دادیم. در فصل چهار انواع مدل های طبقه بندی را مورد بررسی قرار دادیم و در نهایت بهترین مدل را از بین آن ها انتخاب کردیم و فصل پنجم به ارزیابی این مدل پرداخته و شاخص های ارزیابی را روی این مدل محاسبه کردیم.

فهرست

۶.....	فصل ا: فهم مسئله و کسب و کار
۷.....	تعریف:
۸.....	انواع دیابت
۹.....	دیابت نوع ۱
۱۰.....	دیابت نوع ۲
۱۱.....	دیابت بارداری
۱۲.....	پیش دیابت
۱۳.....	دیابت بی مژه
۱۴.....	علت دیابت
۱۵.....	فصل ۲: درگ داده
۱۶.....	مقدمه
۱۷.....	شناسایی داده ها:
۱۸.....	ستون اول: Pregnancies
۱۹.....	ستون دوم: Glucose
۲۰.....	ستون سوم: BloodPressure
۲۱.....	ستون چهارم: SkinThickness
۲۲.....	ستون پنجم: Insulin
۲۳.....	ستون ششم: BMI
۲۴.....	ستون هفتم: DiabetesPedigreeFunction
۲۵.....	ستون هشتم: Age
۲۶.....	ستون نهم: Outcome
۲۷.....	تشخیص خطا یا نویز
۲۸.....	وارد گردن داده به نرم افزار
۲۹.....	بررسی توزیع داده
۳۰.....	بررسی همبستگی داده ها

۱۶	بررسی کیفیت داده ها	
۱۷	فلاصه	
۱۸	فصل ۲۰: آماده سازی داده	
۱۹	ساختاربندی و مجتمع کردن داده ها	
۲۰	مدیریت داده های نویز	
۲۱	شناسایی دادگان پرت	
۲۲	تشخیص توزیع داده ها	
۲۳	استفاده از (وش ۳ تا ۵ سیگما) (وش Z):	
۲۴	استفاده از (وش ۱.۵ تا ۳):	
۲۵	مدیریت داده های مفقوده	
۲۶	تبديل داده کیفی به کمی	
۲۷	نرم‌ال و استاندارد سازی	
۲۸	مدیریت دادگان نامطوازن	
۲۹	فلاصه	
۳۰	فصل ۲۱: مدلسازی	
۳۱	مقدمه	
۳۲	انتخاب بهترین مدل	
۳۳	Auto Classifier	•
۳۴	مدل SVM (ماشین های پشتیبان بردار)	•
۳۵	مدل KNN	•
۳۶	مدل Neural net (شبکه های عصبی)	•
۳۷	مدل Cart	•
۳۸	مدل Quest	•
۳۹	مدل CHAID	•
۴۰	مدل Random Trees	•
۴۱	مدل C5	•

۱۴۶	خلاصه
۱۴۷	فصل ۵: ارزیابی مدل
۱۴۸	مقدمه
۱۴۸	صمت
۱۴۸	دقت
۱۴۸	حساسیت صمت
۱۴۸	ویرگی
۱۴۸	امتیاز F1
۱۴۸	س्रعت
۱۴۹	پایداری
۱۴۹	تفسیر پذیری
۵۲	خلاصه

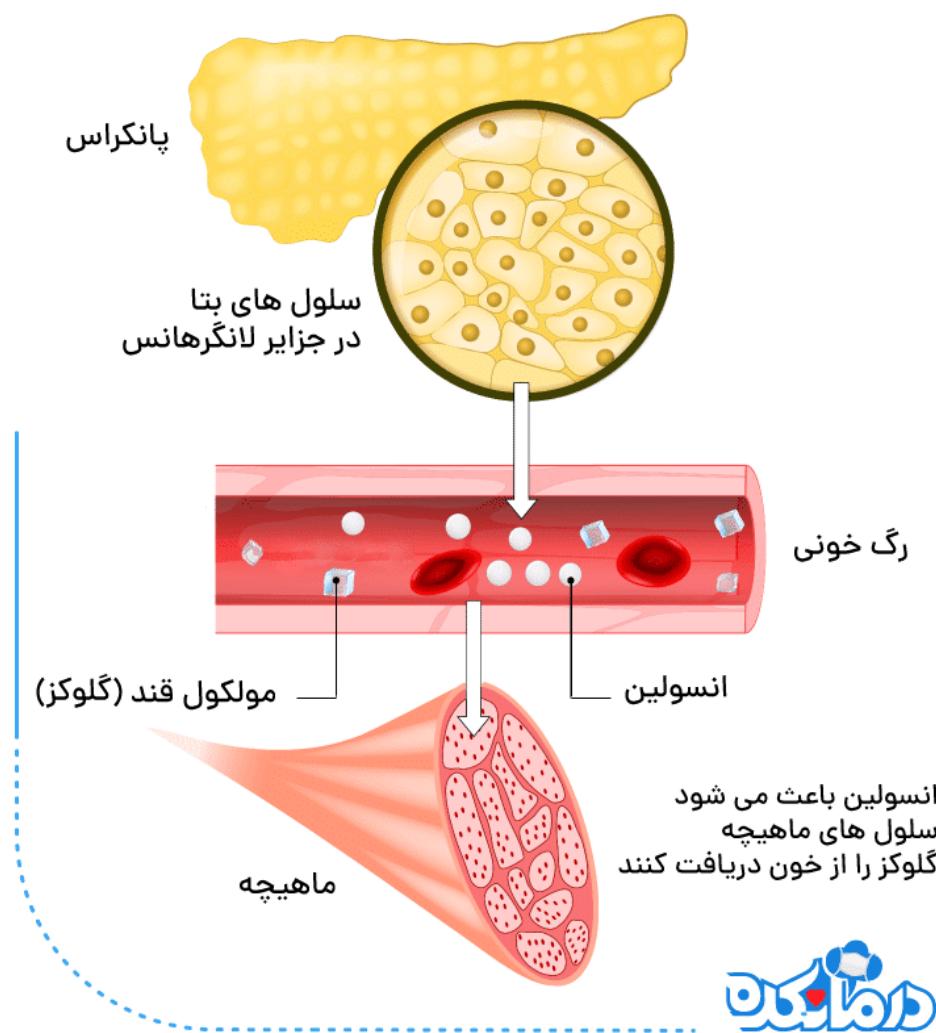
فصل ا: فہم مسئلہ و کسب و کار^۱

^۱ Ref: <https://www.darmankade.com>

تعریف:

دیابت یا مرض قند، نوعی بیماری متابولیک (سوزن و ساز مواد غذایی) مزمن است. ماجرا از آنجا آغاز می‌شود که هورمونی به نام انسولین عملکرد خود را از دست می‌دهد. این هورمون توسط غده پانکراس یا همان لوزالمعده ترشح می‌شود. وظیفه اصلی آن، این است که گلوكز را به سمت سلول‌های بدن هدایت کند. زمانی که انسولین در انجام وظیفه خود ناتوان است، گلوكز به جای ورود به سلول، در خون انشابته می‌شود و به این ترتیب سطح آن در بدن بالا می‌رود؛ دقیقاً همینجاست که می‌گوییم دیابت گرفتم.

عملکرد انسولین



تصویر ۱ - عملکرد انسولین

انواع دیابت

به طور کلی می‌توان مرض قند را به پنج دسته زیر تقسیم کرد:

- دیابت نوع ۱
- دیابت نوع ۲
- دیابت بارداری
- پیش دیابت
- دیابت بی‌مزه

دیابت نوع ۱

دیابت نوع ۱ که به دیابت نوجوانان یا دیابت وابسته به انسولین هم معروف است. این بیماری، نوعی اختلال خودایمنی شناخته می‌شود. در این وضعیت سیستم ایمنی بدن به سلول‌های تولیدکننده انسولین در لوزالمعده حمله می‌کند و توانایی بدن در تولید انسولین را از بین می‌برد. حالا به راحتی می‌توانید حدس بزنید در بدن چه اتفاقی می‌افتد. هورمونی به نام انسولین در بدن وجود ندارد که بخواهد قند را به سمت سلول‌های بدن هدایت کند. این نوع مرض قند مادرزادی بوده و به همین دلیل در کودکان بیشتر دیده می‌شود.

دیابت نوع ۲

در دیابت نوع دوم، انسولین در بدن وجود دارد اما به وظیفه خود به درستی عمل نمی‌کند. این بیماری بسیار شایع‌تر از نوع اول است، می‌توانید حدس بزنید چرا؟ پاسخ آن چندان پیچیده نیست. این بیماری به طور مستقیم با سبک زندگی و رژیم غذایی در ارتباط است. اگر جزء آن دسته از افرادی هستید که رابطه خوبی با ورزش ندارند، شیرینی و نوشیدنی‌های صنعتی پای ثابت سفره غذایی آنهاست، اضافه وزن دارند، چاق هستند و سیگار می‌کشند، متناسفانه شرایط را برای ابتلا به قند خون بالا به خوبی فراهم کرده‌اید.

این بیماری بیشتر بزرگسالان را درگیر می‌کند اما شیوع آن در بین کودکان و نوجوانان نیز در حال افزایش است. بی‌تحرکی و دریافت کالری بیش از حد دو فاکتور اصلی برای ابتلای کودکان به این بیماری است.

دیابت بارداری

دیابت بارداری معمولاً در ماههای آخر بارداری اتفاق می‌افتد. بروز آن بیشتر با تغییرات هورمونی در ارتباط است. هورمون‌هایی که توسط جفت تولید می‌شود به طور خاص بر روی انسولین تاثیر گذاشته و عملکرد آن را مختل می‌کنند.

پیش دیابت

اگر نتیجه آزمایش خون نشان داده است که در مرحله پیش دیابت هستید، باید بگوییم که شانس آورده‌اید. چون این حرف به این معناست سطح گلوکز در بدن بالاست اما شما هنوز به مرض قند مبتلا نشده‌اید. نکته مهم بعدی این است که اگر فکری به حال سبک زندگی و رژیم غذایی خود نکنید، به این بیماری مبتلا خواهید شد.

دیابت بی مزه

دیابت بی مزه، نوعی بیماری مادرزادی است که به دلیل اختلال در غدد درون ریز ایجاد می شود. این بیماری با تشنگی زیاد و تولید بیش از حد ادرار رقیق همراه است و برای آن درمانی وجود ندارد. برای تشخیص آن باید به لیست فوق متخصص غدد مراجعه کنید. این اختلال در اثر فقدان هورمون ادراری (وازوپرسین) یا مسدودسازی عملکرد آن ایجاد می گردد.

علل دیابت

علل دیابت براساس وضعیت عمومی شما، پیشینه خانوادگی، تراز، سلامتی و عوامل محیطی متفاوت است. هیچ علت متدالی برای بروز آن وجود ندارد. از دلایل شناخته شده ایی که سبب بروز این بیماری می شوند، به موارد زیر اشاره شده است:



تصویر ۴- علل دیابت

فصل ۲: درگ داده

مقدمه

این مجموعه داده با دیتاست **Pima Indians Diabetes Database** یک پروژه تشخیصی بوده و هدف این است که تشخیص داده شود با استفاده از مولفه های موجود؛ آیا بیمار مبتلا به دیابت می باشد یا خیر. این دیتاست یک مسئله دسته بندی با تعداد مشاهدات ۷۶۸، ورودی ها ۸ و یک ستون خروجی می باشد. همه بیماران در اینجا زنان با حداقل ۲۱ سال سن می باشند. در ادامه به تشریح اسامی ستون های ویژگی میپردازیم.

شناسایی داده ها:

قبل از شروع کار در ابتدا باید هر یک از ویژگیهای ورودی و همچنین ستون خروجی یا هدف را بشناسیم.

ستون اول: Pregnancies

این ویژگی تعداد دفعات حاملگی را مشخص می کند که با توجه به شناخت پروژه در فصل قبل می تواند یکی از عوامل موثر در دیابت باشد. تعداد بارداری بیمار احتمال مبتلا شدن به دیابت را در او افزایش می دهد. این داده از نوع گسسته می باشد.

ستون دوم: Glucose

این ستون که اهمیت بسیار بالایی در تشخیص دیابت دارد؛ مقدار غلظت گلوکز در آزمایش تحمل گلوکز خوراکی (OGTT)^۲ را نشان می دهد. در این آزمایش ابتدا در حالت ناشتا، نمونه خون از شما گرفته خواهد شد. مدت زمان ناشتایی ۸ تا ۲ ساعت است. پس از آن باید یک محلول گلوکز را بخورید. سپس سه نمونه خون در بازه های زمانی یک ساعته از شما گرفته خواهد شد. در واقع برای این تست، لازم است ۴ بار نمونه خون دهید. نوع این داده نیز گسسته می باشد.

- نتایج کمتر از ۱۴۰ میلی گرم / دسی لیتر، عادی تلقی می شوند.
- نتایج بین ۱۴۰-۱۹۹، شرایط قبل از مرض قند را نشان می دهند.
- نتایج بزرگتر مساوی ۲۰۰، نشان دهنده قند خون بالا هستند.

ستون سوم: BloodPressure

این ستون، مقدار فشار خون دیاستولیک بیمار را نمایش می دهد.

فشار خون^۳، فشاری است که قلب ما برای پمپ کردن خون به دیواره رگها وارد می کند. فشار خون بر حسب میلی متر جیوه (mmHg) اندازه گیری می شود و دو عدد برای توصیف آن استفاده می گردند. این اعداد به شکل $130/80$ mmHg نوشته و «۱۳۰ روی ۸۰» یا «۱۳۰ روی ۸» خوانده می شود. عدد اول، فشار سیستولیک نام دارد. این عدد بیشترین فشاری است که قلب هنگام انقباض و تپیدن برای پمپاژ خون به سراسر بدن از آن استفاده می کند. عدد دوم، فشار دیاستولیک، کمترین فشاری است که قلب در زمانی که بین ضربانها در حال استراحت است، مورد استفاده قرار می دهد. در مثال $130/80$ mmHg، فشار

² Ref: <https://www.darmankade.com>

³ Ref: <https://abidipharma.com>

سیستولیک 130 mmHg و فشار دیاستولیک 80 mmHg است. این ستون رابطه مستقیم با هدف دارد. این داده نسبی می باشد.

SkinThickness:

این ویژگی مقدار ضخامت چین های پوستی سه سر بازو (میلی متر) را نمایش می دهد. همانطور که در فصل قبل گفته شد یکی از عوامل موثر در دیابت نوع 2 ، وزن و کم تحرکی افراد بود و این مولفه یکی از مشخصه های چاقی را بیان می کند. نوع آن گستره می باشد.

Insulin:

این مشخصه، جواب آزمایش انسولین سرم 2 ساعته را نمایش میدهد و هدف از این آزمایش، جهت تایید و اثبات مقاومت به انسولین می باشد. این ستون نیز گستره می باشد.

BMI:

این ستون که مشخص کننده شاخص توده بدنی (نقسیم وزن بر مجذور قد) بیمار میباشد یکی از عوامل موثر در تشخیص بیماری دیابت می باشد. این ویژگی با هدف رابطه مستقیم دارد. این داده پیوسته است.

DiabetesPedigreeFunction:

یکی دیگر از عوامل موثر در دیابت، سابقه خانوادگی بیمار می باشد. این ستون یک تابع جهت امتیاز بندی بیمار با سابقه دیابت در خانواده او می باشد. بدیهی است هرچه مقدار این ستون بیشتر باشد احتمال دیابت در بیمار بیشتر می شود. نوع این داده پیوسته می باشد.

Age:

این ویژگی سن بیمار را نمایش می دهد که یکی دیگر از علل بیماری دیابت می باشد. این ستون با دیابت داشتن بیمار رابطه مستقیم دارد. این داده نسبی می باشد.

Outcome:

این ویژگی همان هدف پروژه می باشد و از نوع داده دودویی بوده و دارای دو مقدار 0 و 1 می باشد که بیانگر عدم دیابت و 1 نشانگر دارا بودن دیابت است.

تشخیص خطأ یا نویز

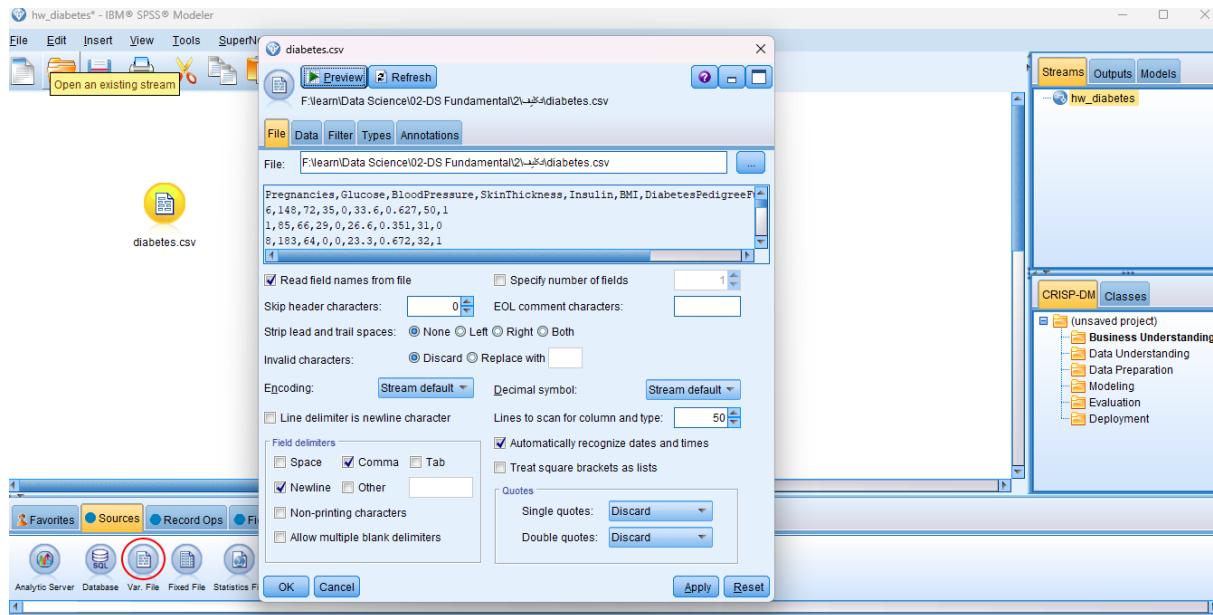
با توجه به تعریف داده ها، بطور کلی مقادیری که داده ها نمی پذیرند در جدول زیر مشخص شده است.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
منفی	صفر یا منفی	کمتر از 60	صفر یا منفی	صفر یا منفی
BMI	DiabetesPedigreeFunction	Age	Outcome	
کمتر از 16	منفی و بیشتر از 20.7	صفر یا منفی	غیر از 0 و 1	

جدول ا- تشخیص خطأ در ویژگی ها

وارد کردن داده به نرم افزار

فایل CSV مربوطه را از سایت [Kaggle](#) دانلود کرده و در نرم افزار IBM var file source ابزار را انتخاب کرده و داخل آن ایمپورت می کنیم.



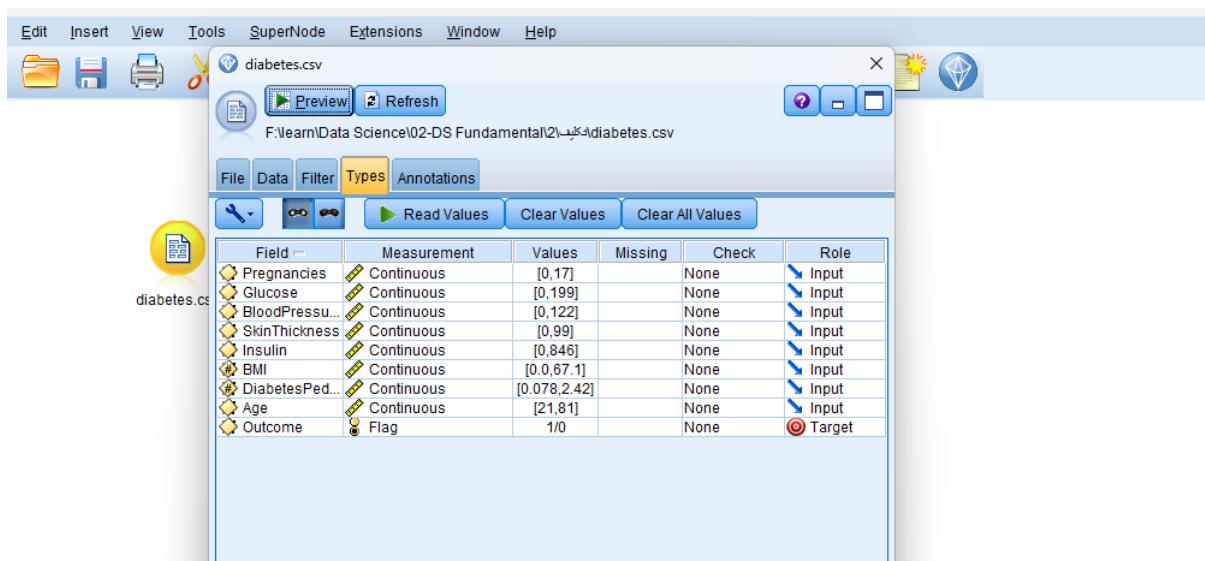
تصویر ۳- وارد کردن داده در IBM

د رکورد اول داده ها بصورت زیر نمایش داده می شود:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33....		0.627	50
2	1	85	66	29	0	26....		0.351	31
3	8	183	64	0	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	0	0	25....		0.201	30
7	3	78	50	32	88	31....		0.248	26
8	10	115	0	0	0	35....		0.134	29
9	2	197	70	45	543	30....		0.158	53
10	8	125	96	0	0	0.0....		0.232	54

تصویر ۴- د رکورد اول داده ها

سپس باید Type داده ها را مشخص کنیم که طبق تصویر زیر بدین صورت می باشد:



تصویر ۵ - داده type

بررسی توزیع داده

با استفاده از ابزار Data Audit در قسمت Output می توان به صورت چشمی توزیع داده ها را مشاهده کرد.

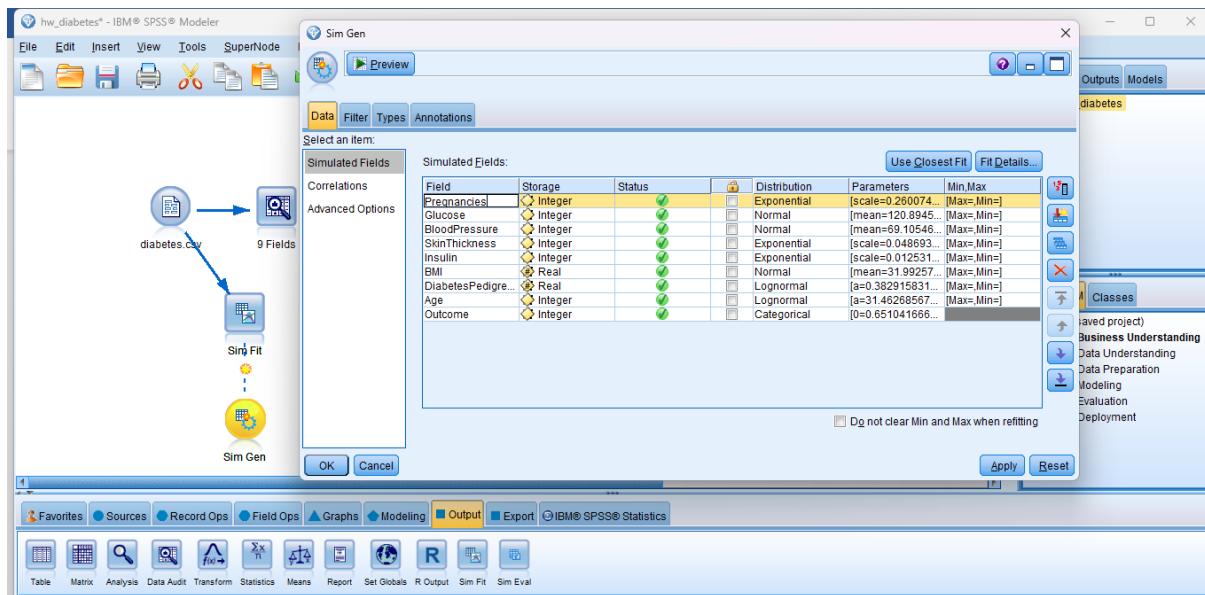


تصویر ۶ - توزیع داده ها با استفاده از ابزار Data Audit

اینطور که از تصویر مشخص است بصورت چشمی می توان گفت که ویژگی های زیر از نوع نرمال می باشند:

Glucose, BloodPressure, SkinThickness, BMI

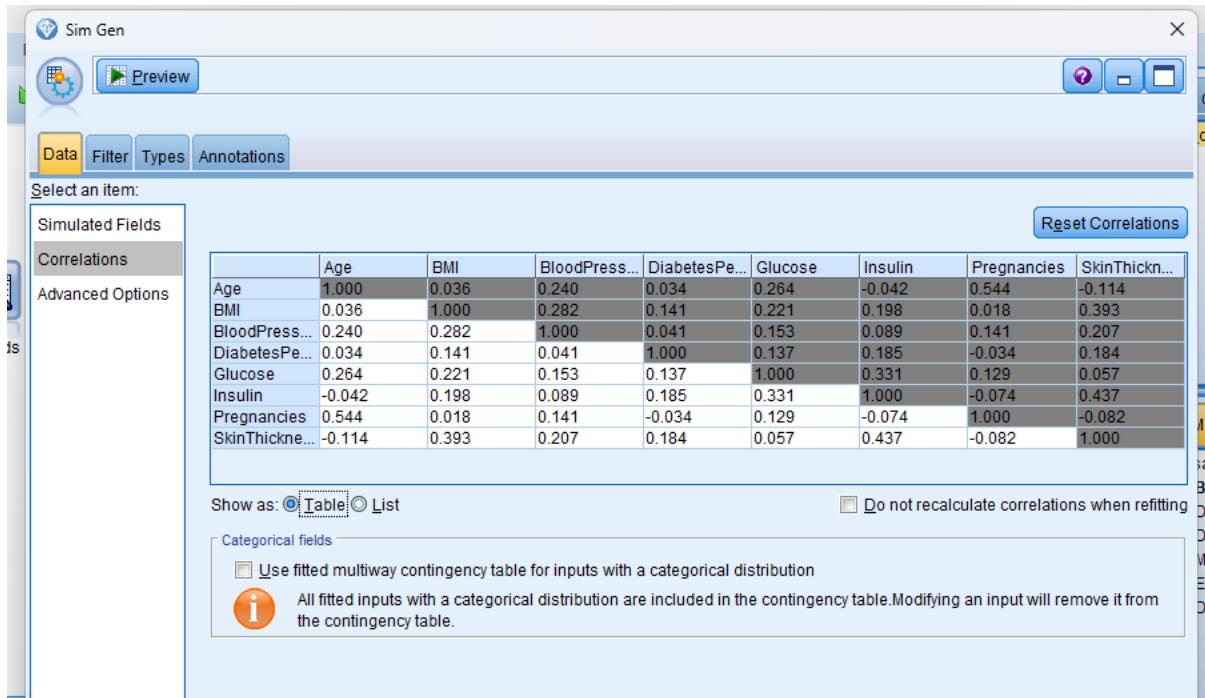
تشخیص توزیع سایر ویژگی ها کمی سخت می باشد که بدین منظور می توان از ابزار Sim Fit از تب Output و انتخاب آزمون اندرسون – دارلینگ استفاده کرد که توزیع هر ویژگی را مشخص می کند.



تصویر ۷- تمثیل داده ها با استفاده از ابزار Sim Fit

بررسی همبستگی داده ها

از تب Correlations در ابزار Sim Fit می توان میزان همبستگی ویژگی ها را نسبت به هم مشاهده کرد.

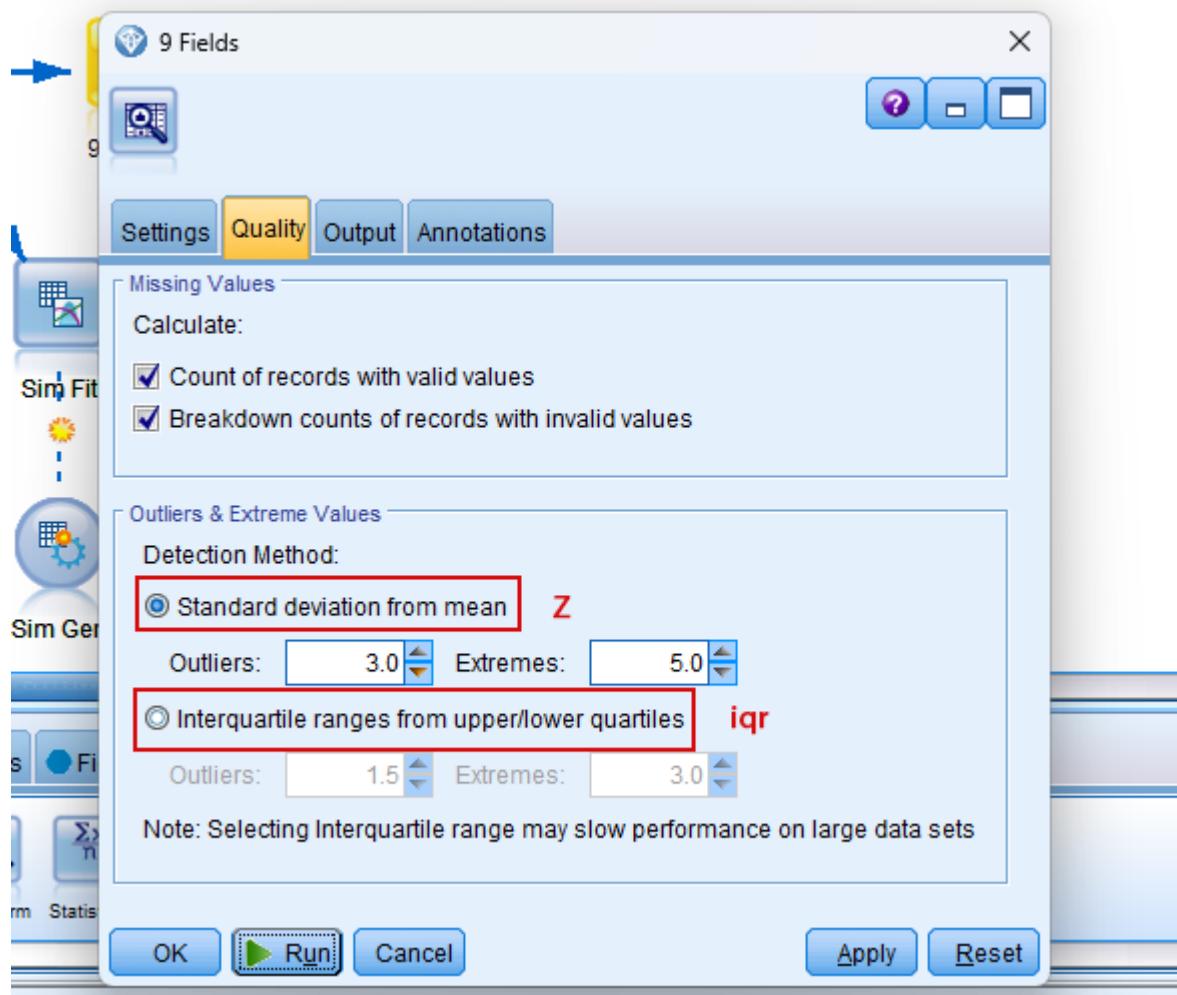


تصویر ۸- میزان همبستگی ویژگی ها نسبت به هم

در اینجا مشاهده می کنید که ویژگی ها همبستگی زیادی به هم ندارند. در صورتی که همبستگی ویژگی ها بـ ۰.۷ به بالا بود میبایست با استفاده از PCA بعد ها را تا حد امکان کاهش داد.

بررسی کیفیت داده ها

جهت بررسی کیفیت داده ها می توان از ابزار Data Audit که برای داده هایی که توزیع نرمال دارند میباشد از متدهای Z و برای مابقی از iqr استفاده کرد.



تصویر ۹ - بررسی کیفیت داده ها با روشن Z

Data Audit of [9 fields] #6															
Audit		Quality		Annotations											
Complete fields (%): 100%		Complete records (%): 100%													
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value			
Pregnancies	Continuous	4	0	None	Never	Fixed	100	768	0	0	0	0			
Glucose	Continuous	5	0	None	Never	Fixed	100	768	0	0	0	0			
BloodPressure	Continuous	35	0	None	Never	Fixed	100	768	0	0	0	0			
SkinThickness	Continuous	1	0	None	Never	Fixed	100	768	0	0	0	0			
Insulin	Continuous	15	3	None	Never	Fixed	100	768	0	0	0	0			
BMI	Continuous	14	0	None	Never	Fixed	100	768	0	0	0	0			
DiabetesPed	Continuous	7	4	None	Never	Fixed	100	768	0	0	0	0			
Age	Continuous	5	0	None	Never	Fixed	100	768	0	0	0	0			
Outcome	Flag	--	--	Never	Fixed		100	768	0	0	0	0			

تصویر ۱۰ - بررسی کیفیت داده های نرمال با روشن Z

Data Audit of [9 fields] #7

The screenshot shows a software interface titled "Data Audit of [9 fields] #7". The top menu bar includes "File", "Edit", "Generate", and various icons. Below the menu is a toolbar with buttons for "Audit", "Quality", and "Annotations". A status bar at the bottom shows "Complete fields (%): 100%" and "Complete records (%): 100%". The main area is a table with the following columns:

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Pregnancies	Continuous	4	0 None	Never	Fixed	100	768	0	0	0	0	0
Glucose	Continuous	5	0 None	Never	Fixed	100	768	0	0	0	0	0
BloodPressure	Continuous	10	35 None	Never	Fixed	100	768	0	0	0	0	0
SkinThickness	Continuous	1	0 None	Never	Fixed	100	768	0	0	0	0	0
Insulin	Continuous	26	8 None	Never	Fixed	100	768	0	0	0	0	0
BMI	Continuous	18	1 None	Never	Fixed	100	768	0	0	0	0	0
DiabetesPed	Continuous	23	6 None	Never	Fixed	100	768	0	0	0	0	0
Age	Continuous	9	0 None	Never	Fixed	100	768	0	0	0	0	0
Outcome	Flag	-	--	Never	Fixed	100	768	0	0	0	0	0

تصویر ۱۱- بررسی کیفیت داده های غیرنرمال با روش *iqr*

خلاصه

در این فصل داده ها را بررسی کردیم و شناخت کافی به هر کدام از ویژگی ها پیدا کردیم؛ این که هر ویژگی چی هست، چه نوعی هست و توزیع آن چیست. همچنین کیفیت و همبستگی بین داده ها را بررسی کردیم. در ادامه داده ها را جهت مدلسازی آماده می کنیم.

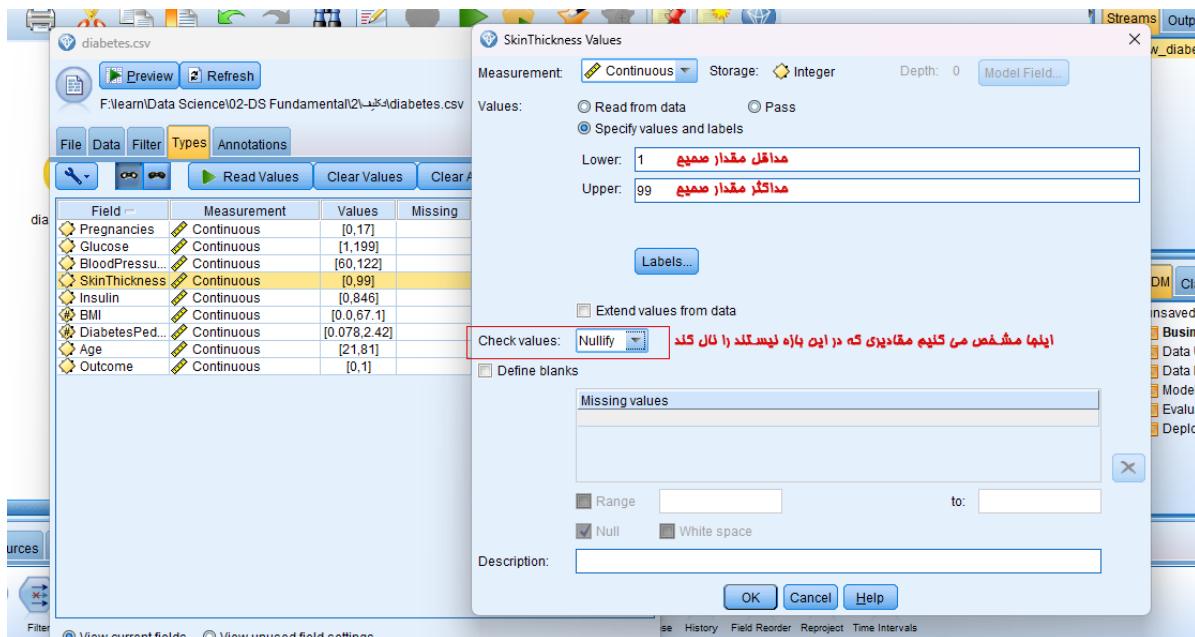
فصل ۳: آماده سازی داده

ساختاربندی و مجتمع کردن داده ها

در این پروژه از آنجا که داده ها پراکنده نیست و تنها یک دیتاست وجود دارد نیاز به انجام این کار نیست.

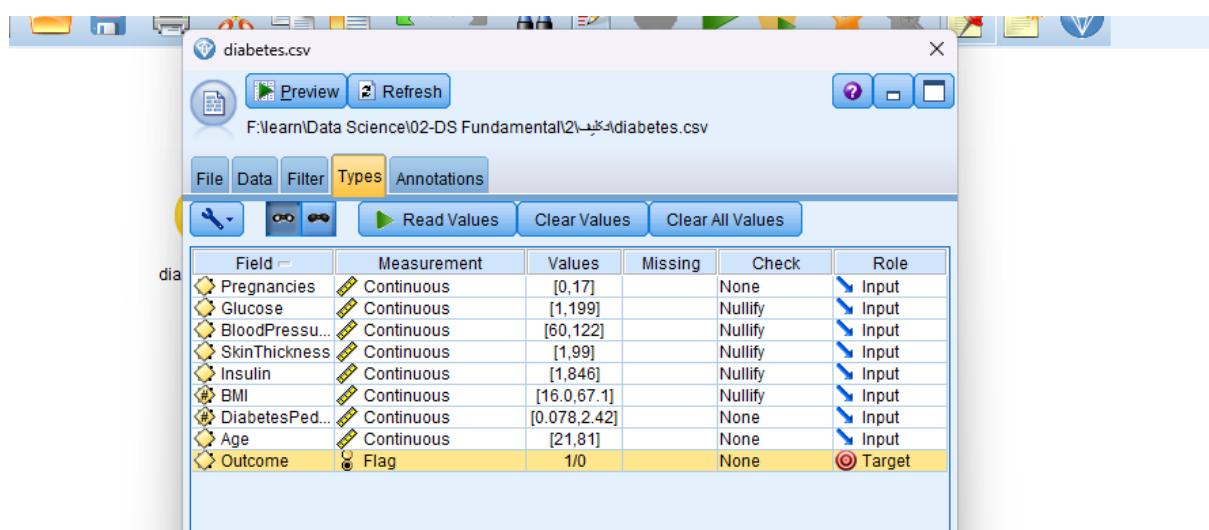
مدیریت داده های نویز

در این قسمت با توجه به جدول ۱ و تصویره مشاهده می کنیم که داده های ما دارای نویز می باشند که برای رفع آن زمانی که داده ها را ایمپورت کردیم روی هر یک از ویژگی ها دابل کلیک کرده تا پنجره زیر نمایان شود.



تصویر ۱/۲- مدیریت داده های نویز

در اینجا حداقل و حدکثر مقدار صحیح هر ویژگی را وارد کرده و مقادیری که در این بازه نیستند را نال می کنیم و تائید می کنیم.



تصویر ۳/۴- مدیریت داده های نویز(مقادیر صمیع هر ویژگی)

شناسایی دادگان پرت

برای مدیریت داده های پرت میبایست از ابزار Data Audit استفاده کرد. همانند فصل قبل که با این ابزار و ابزار Sim Fit جهت شناسایی توزیع داده ها آشنا شدیم یکبار دیگه از این ابزارها جهت توزیع داده ها استفاده می کنیم زیرا که در اینجا داده های نویز را مدیریت کردیم و بهتر است یکبار دیگه توزیع داده ها بررسی شود تا بتوان نسبت به نرمال یا غیرنرمال بودن ویژگی ها روش مربوطه را جهت مدیریت داده های پرت استفاده کرد.

تشخیص توزیع داده ها

ابتدا از ابزار Data Audit استفاده می کنیم که بصورت چشمی می توان گفت ویژگی Glucose، BMI و SkinThickness نرمال هستند.

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	60	122	75.096	10.316	0.841	-	647
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
Outcome		Flag	0	1	--	--	--	2	768

تصویر ۱۴- بررسی توزیع داده ها بعد از مدیریت نویز (Data Audit)

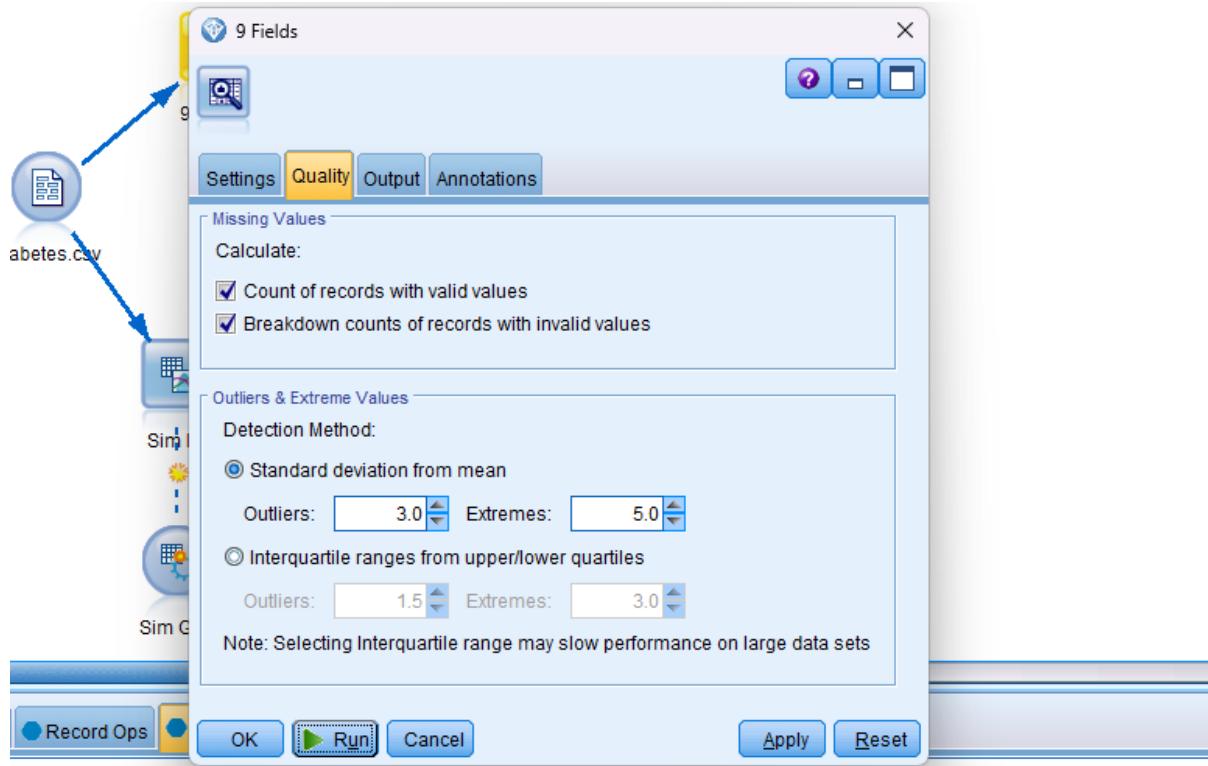
حال با استفاده از ابزار Sim Fit و آزمون کولموگروف-اسمیرنوف بررسی می کنیم.

Simulated Fields:					
Field	Storage	Status		Distribution	F
Pregnancies	Integer			Exponential	[S]
Glucose	Integer			Lognormal	[S]
BloodPressure	Integer			Lognormal	[S]
SkinThickness	Integer			Weibull	[S]
Insulin	Integer			Lognormal	[S]
BMI	Real			Normal	[r]
DiabetesPedigree...	Real			Lognormal	[S]
Age	Integer			Lognormal	[S]
Outcome	Integer			Categorical	[C]

تصویر ۱۵- توزیع داده ها بعد از مدیریت نویز (Sim Fit)

داده هایی که توزیع نرمال و لگ نرمال دارند را با روش z و مابقی را با روش iqr مدیریت می کنیم.

استفاده از روش ۳ تا ۵ سیگما (روش Z):
از ابزار Data Audit تنظیمات زیر را انجام می دهیم.



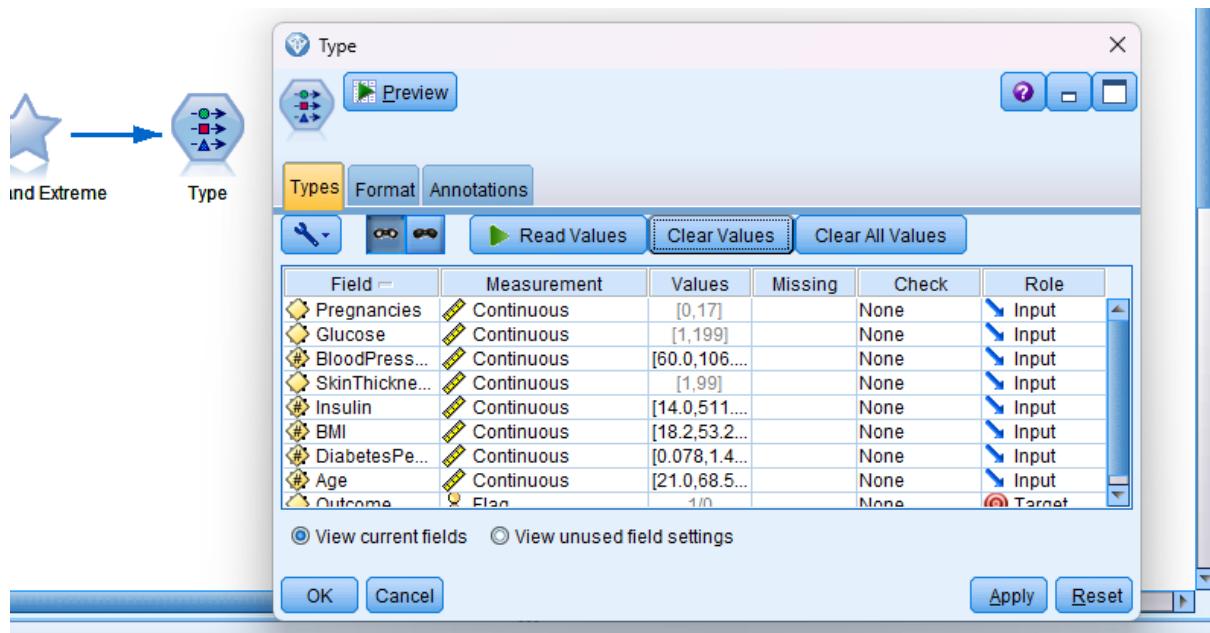
تصویر ۶- تنظیم ۳ تا ۵ سیگما در Data Audit

مشخصه هایی که در تصویر ۱۵ دارای توزیع نرمال و لاغ نرمال بودند را در اینجا مدیریت کردیم و مواردی که تعداد بالایی نداشتند را Coarse و یکی که تعداد قابل قبولی داده پرت داشت را برای داده های پرت و داده های خیلی پرت را null کردیم.

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	None
Glucose	Continuous	0	0	None
BloodPressure	Continuous	7	0	Coerce
SkinThickness	Continuous	1	1	None
Insulin	Continuous	7	1	Coerce
BMI	Continuous	3	1	Coerce
DiabetesPed...	Continuous	7	4	Coerce outliers / nullify extremes
Age	Continuous	5	0	Coerce
Outcome	Flag	--	--	N

تصویر ۷- مدیریت داده های پرت و خیلی پرت مولفه های نرمال و لاغ نرمال

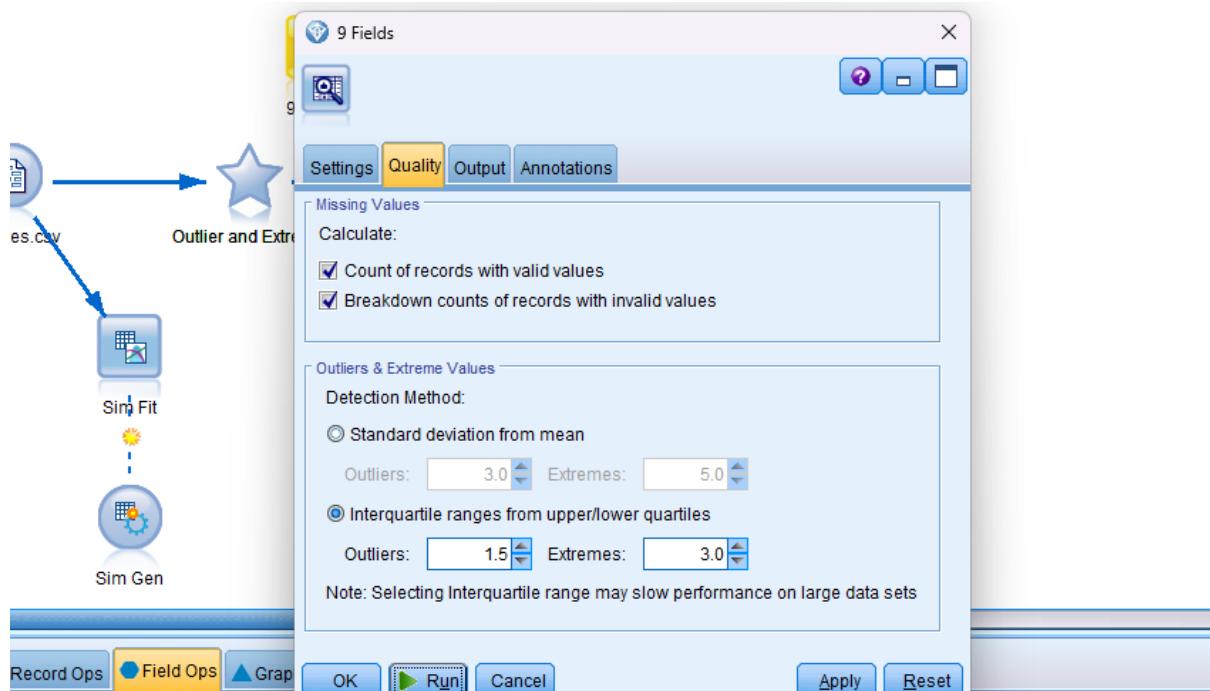
بعد از اینکار میبایست از ابزار Type Field Ops در قسمت استفاده کرد و مقادیر داده هایی که مدیریت شده را دوباره خواند زیرا که رنج داده ها عوض شده است.



تصویر ۱۸- بروزرسانی رنچ داده های مدیریت شده با ابزار Type

استفاده از روش Iqr ۳ تا ۱.۵

ابزار Data Audit را به ابزار Type وصل می کنیم و اینبار تنظیمات زیر را انجام می دهیم.



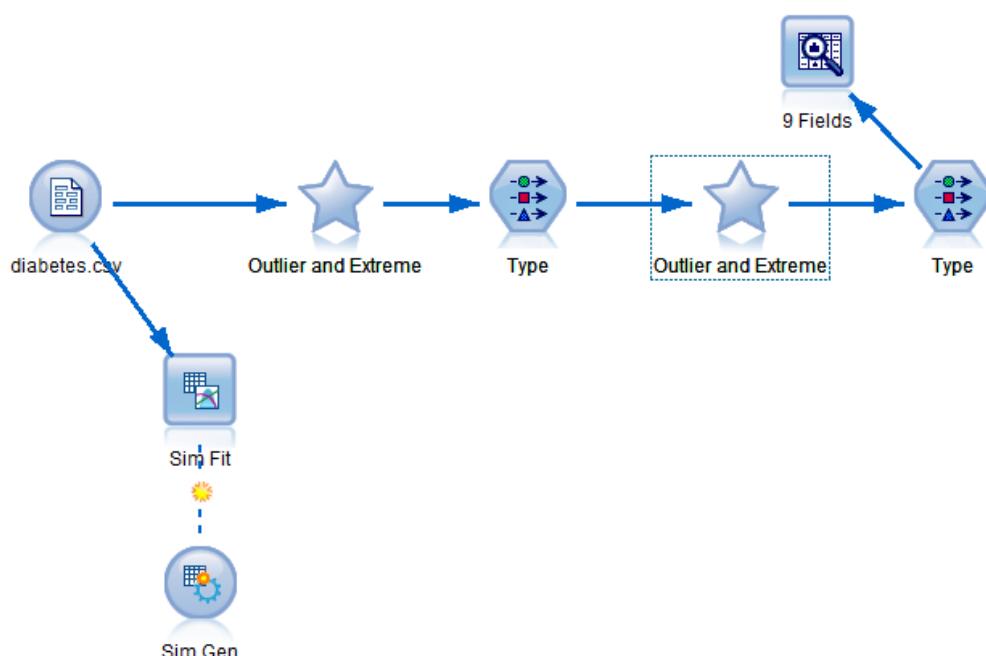
تصویر ۱۹- تنظیمه Iqr ۳ تا ۱.۵ در Data Audit

در اینجا هم مشخصه هایی که توزیع غیرنرمال داشتند را چون تعداد کمی بودند coarse کردیم.

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	Coerce
Glucose	Continuous	0	0	None
BloodPressure	Continuous	12	0	None
SkinThickness	Continuous	2	1	Coerce
Insulin	Continuous	24	0	None
BMI	Continuous	8	0	None
DiabetesPed...	Continuous	28	0	None
Age	Continuous	9	0	None
Outcome	Flag	--	--	None

تصویر ۲۰- مدیریت داده های پرت و فیلی پرت مولفه های غیرنرمال

اینجا هم باید بعد از مدیریت داده های غیرنرمال از ابزار Type جهت بروزرسانی رنچ مقادیر استفاده کرد.



تصویر ۲۱- مسیر پژوهش تا به الان

مدیریت داده های مفقوده

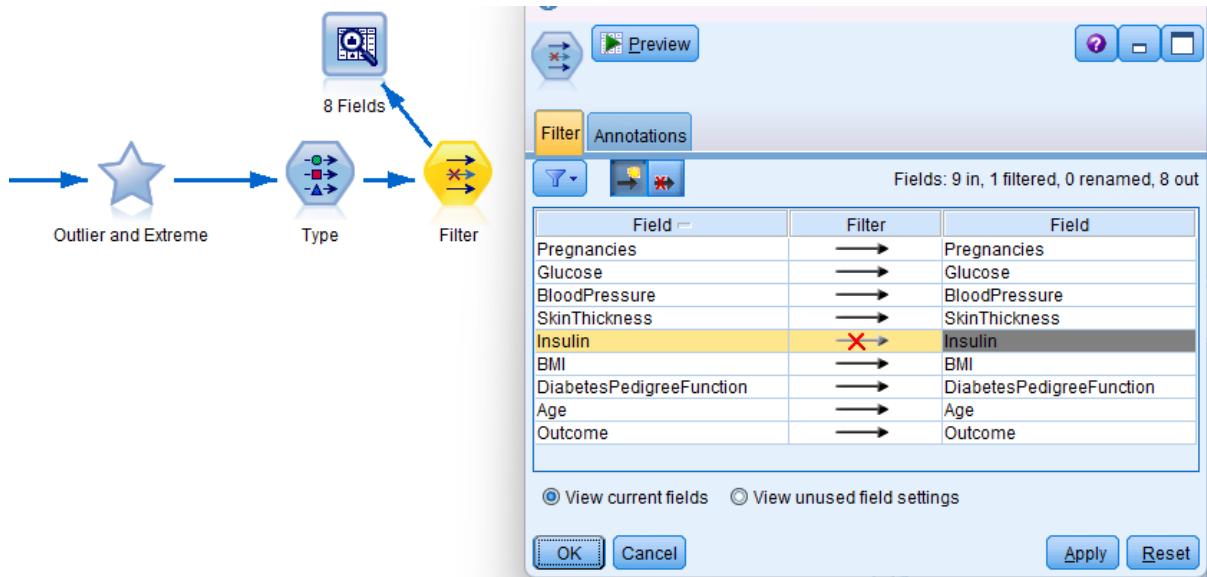
برای مدیریت داده های مفقوده می توان با استفاده از Data Audit درصد داده ها را مشاهده کرد.

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Pregnancies	Continuous	0	0	None	Never	Fixed	100	768	0	0	0	0
Glucose	Continuous	0	0	None	Never	Fixed	99.349	763	5	0	0	0
BloodPressure	Continuous	12	0	None	Never	Fixed	64.245	647	121	0	0	0
SkinThickness	Continuous	0	0	None	Never	Fixed	70.443	541	237	0	0	0
Insulin	Continuous	24	0	None	Never	Fixed	51.302	394	374	0	0	0
BMI	Continuous	8	0	None	Never	Fixed	98.568	757	11	0	0	0
DiabetesPed...	Continuous	28	0	None	Never	Fixed	99.479	764	4	0	0	0
Age	Continuous	9	0	None	Never	Fixed	100	768	0	0	0	0
Outcome	Flag	--	--	Never	Fixed		100	768	0	0	0	0

تصویر ۲۲- درصد داده های هر ویژگی

در اینجا چون نزدیک به ۵۰ درصد از داده های انسولین مفقوده هستند تصمیم به حذف این ویژگی میگیریم. البته که حذف راه حل آخر است ولی چون نیمی از داده ها مفقوده هستند و هم اهمیت زیادی در

این پروژه ندارد آن را حذف می کنیم. برای اینکار از ابزار Filter در تب Field Ops استفاده می کنیم و طبق تصویر این ویژگی را حذف می کنیم.



تصویر ۳۴- حذف مولفه Insulin

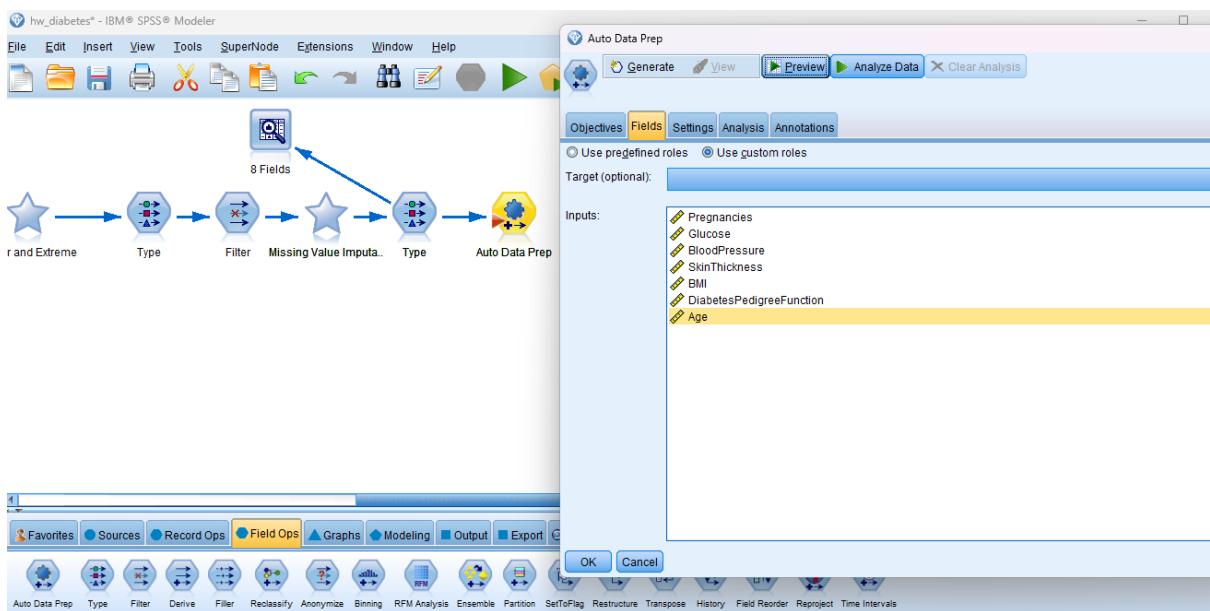
در ادامه برای ویژگی Glucose از متاد میانگین، ویژگی BloodPressure و SkinThickness به علت تعداد بالای داده مفقوده از متاد الگوریتم، ویژگی BMI از متاد میانه و ویژگی DiabetesPedigreeFunction از متاد رندم استفاده کردم. اینجا هم همانند داده های پرت میباشد بعد از انجام ابزار Type قرار داد.

تبديل داده کيفي به كمي

در اين دياتاست چون داده کيفي نداريم نيازي به انجام اين مرحله نيسست.

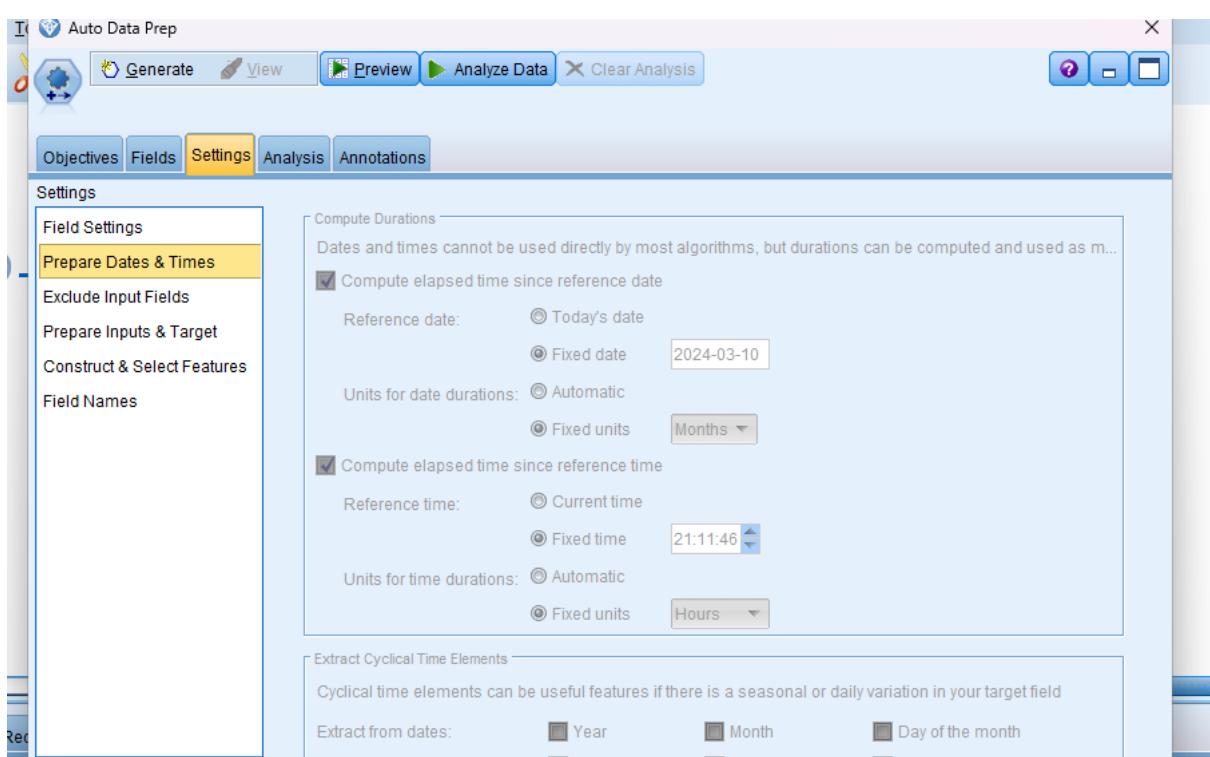
نرمال و استانداردسازی

برای استانداردسازی از ابزار Auto Data Prep که در تب Field Ops میباشد استفاده می کنیم و از روش Min/Max داده ها را به مقیاس ۰ تا ۱۰۰ تبدیل می کنیم. در نظر داشته باشید که نوع داده ها قبل از این عمل باید continuous باشد. در نهايیت طبق تصویر تنظيمات را انجام می دهيم.



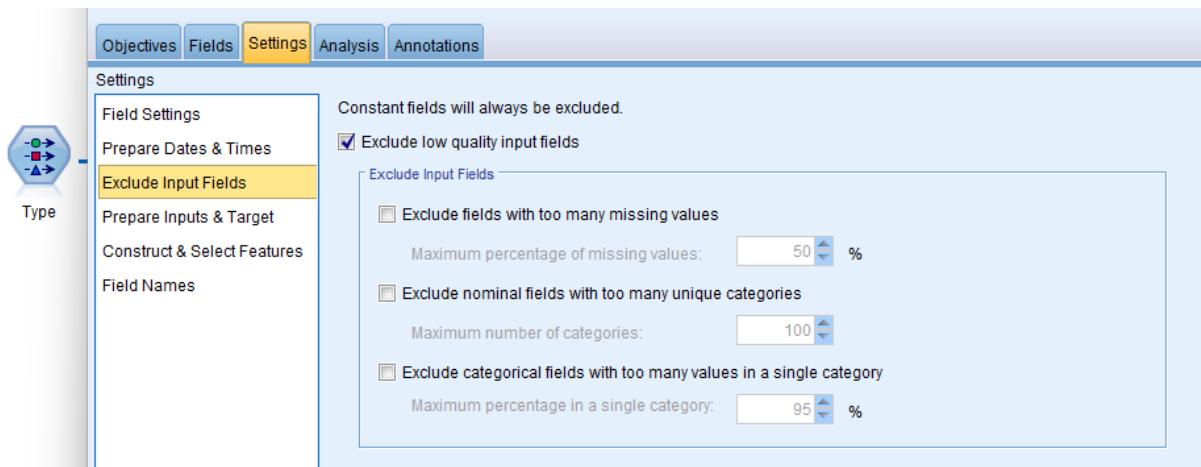
تصویر ۲۴- اسکریپت‌سازی با ابزار Auto Data Prep (انتخاب مولفه ها)

در انتخاب مولفه ها تارگت را انتخاب نمی کنیم.



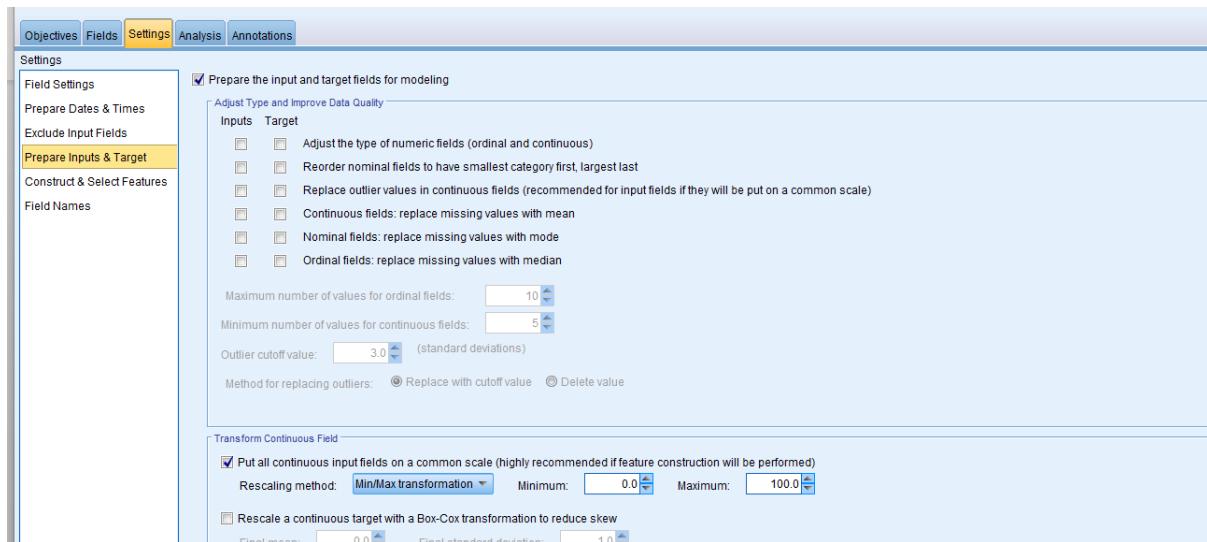
تصویر ۲۵- تنظیمات Prepare Dates & Times

در قسمت Prepare Dates & Times تنظیمات مربوطه به پروژه های سری زمانی است و در این پروژه نیازی به تنظیم نیست.



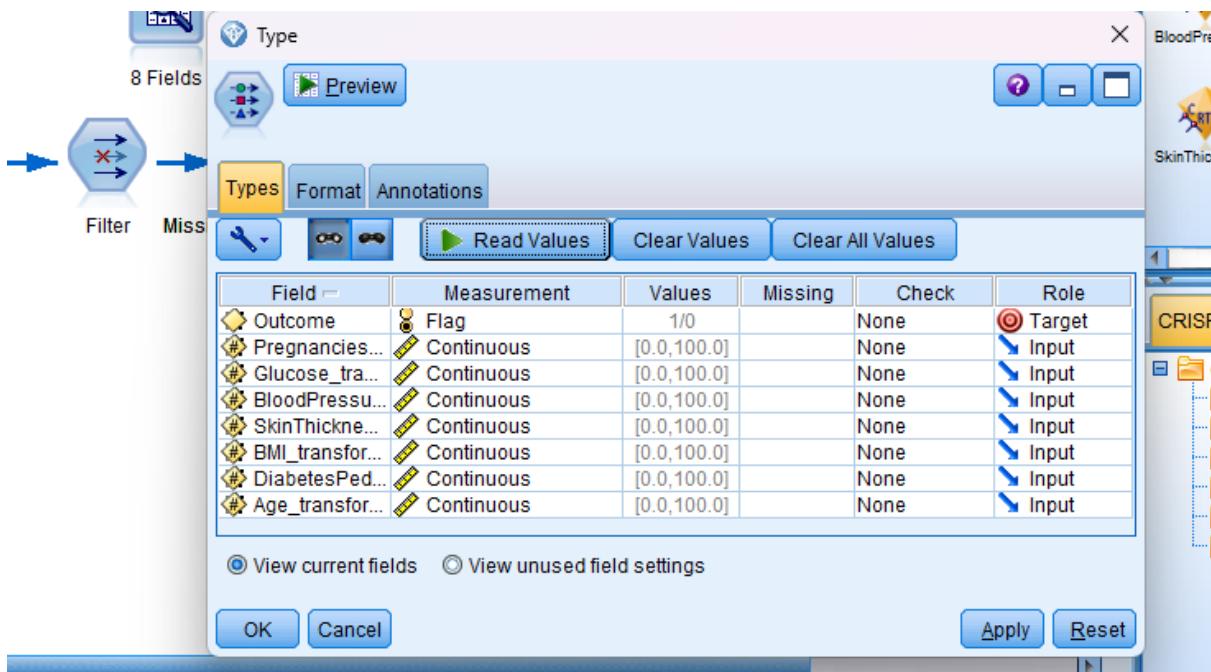
تصویر ۲۶- تنظیمات Exclude Input Fields در Auto Data Prep

در تنظیمات Exclude Input Fields در صورتی که داده های مفقوده و ... را مدیریت نکرده باشیم می توانیم تنظیم کنیم که در پروژه ما نیازی نیست زیرا که در مراحل قبلی این موارد مدیریت شده است.



تصویر ۲۷- تنظیمات Prepare Inputs & Target در Auto Data Prep

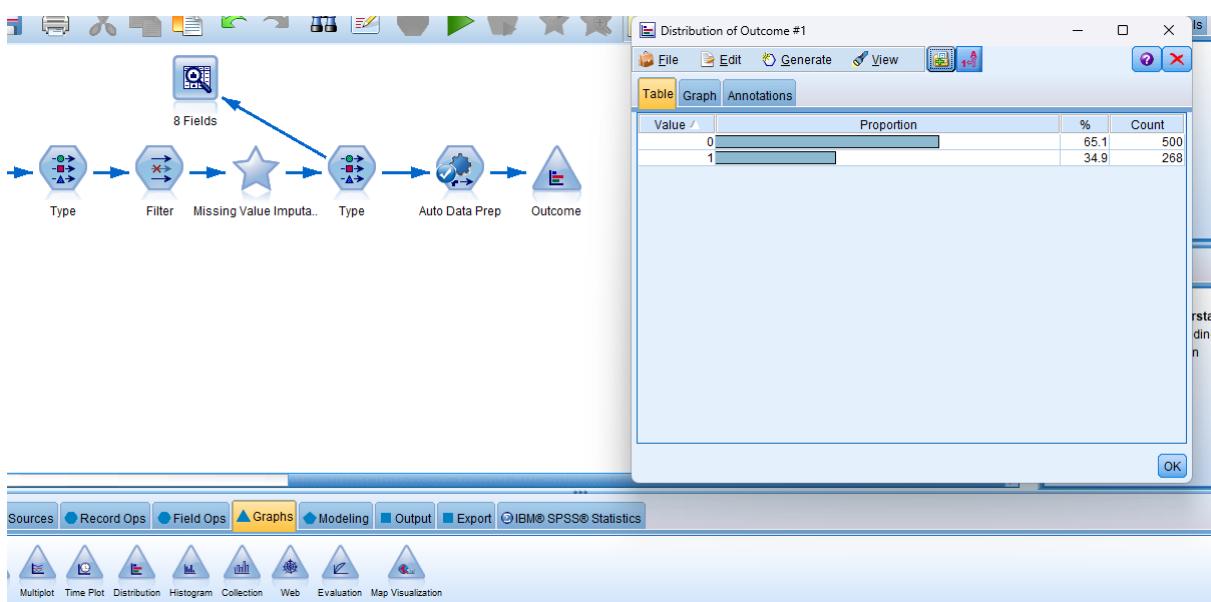
در تنظیمات Prepare Inputs & Target روش مدنظر را جهت هم مقیاس سازی انتخاب می کنیم که در اینجا ما روش min/max را انتخاب کردیم. جهت اطمینان از استانداردسازی بعد از آن ابزار Type قرار داده و رنج داده ها را مشاهده می کنیم.



تصویر ۲۸- مشاهده رنج داده های هم مقیاس شده از طریق ابزار Type

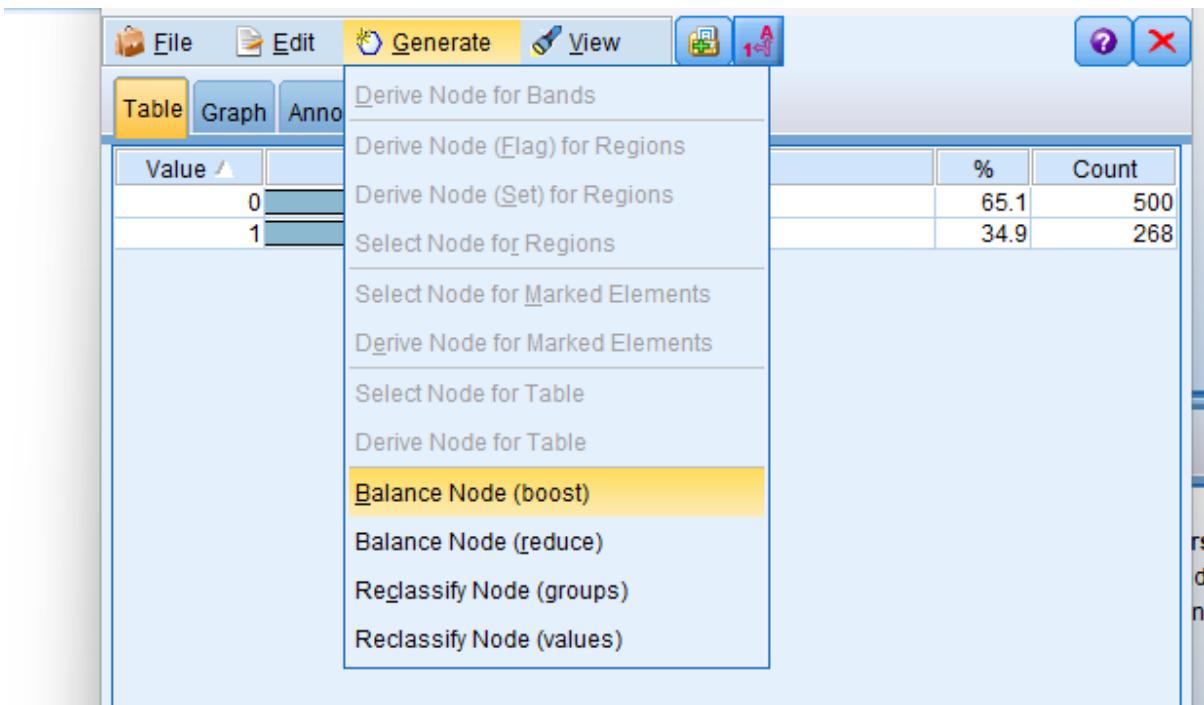
مدیریت دادگان نامتوافق

برای مشاهده این که فیلد هدف بالанс هست یا نه از ابزار Graphs از قسمت Distribution استفاده می کنیم و فیلد تارگت را به آن معرفی می کنیم و نتیجه را نمایش می دهد.



تصویر ۲۹- مشاهده میزان بالанс فیلد هدف

برای مدیریت آن از روش over sampling استفاده می کنم و تنها داده های آزمایش را انجام می دهم. نودی که تولید می شود را در مدلسازی استفاده می کنم به صورتی که یکبار مدل را بدون over sampling و یکبار با over sampling انجام داده و یکبار با آن را ارزیابی می کنیم.



تصویر ۳۰ - مدیریت داده نامتعارف با استفاده از روش over sampling

خلاصه

در این فصل داده ها را پاکسازی کردیم و برای عمل مدلسازی آماده کردیم. مشخص شد که دیتابست ما یکسری داده های نویز داشتند که با صفر مشخص شده بود و برای تشخیص آن لازم بود که کسب و کار و داده های موجود به خوبی درک شوند تا بهتر بتوان مدیریت کرد و آماده سازی را انجام داد. همچنین از روش های مختلف جهت مدیریت داده های مفقوده استفاده کردیم که بهترین نتیجه را برای هر داده بدهد. با توجه به اینکه این پروژه نیازی به مهندسی ویژگی ها اعم از کاهش بعد یا ... ندارد مستقیم عمل مدلسازی را انجام خواهیم داد که در فصل بعد با آن آشنا خواهیم شد.

فصل ۱۴: مدلسازی

مقدمه

در این فصل به منظور مدلسازی با استفاده از ابزار Field Ops از تاب partition داده ها را به دو بخش Test و Train به نسبت ۸۰ به ۲۰ تقسیم می کنیم.

انتخاب بهترین مدل

در این فصل تمام مدل های مربوط به مسئله طبقه بندی را در تمام جهات بررسی می کنیم تا در نهایت بهترین مدل برای پروژه تشخیص دیابت را شناسایی کنیم.

Auto Classifier •

در ابتدا با استفاده از این ابزار تمام مدل های مسئله طبقه بندی را با حالات مختلف بررسی کرده تا ۱۰ تا از بهترین مدل ها را به ما معرفی کند که تیجه آن با استفاده از over sampling و بدون استفاده از آن بصورت زیر شد.

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
<input checked="" type="checkbox"/>		C&R Tree 10	2	145.000	36	2.072	82.143	7	0.840
<input checked="" type="checkbox"/>		C&R Tree 22	2	135.000	37	1.984	80.714	7	0.820
<input checked="" type="checkbox"/>		C&R Tree 1	2	134.545	41	2.019	80.0	7	0.862
<input checked="" type="checkbox"/>		C&R Tree 4	2	134.545	41	2.019	80.0	7	0.862
<input checked="" type="checkbox"/>		C&R Tree 7	2	134.545	41	2.019	80.0	7	0.862
<input checked="" type="checkbox"/>		C&R Tree 13	2	129.000	35	2.019	80.0	7	0.856
<input checked="" type="checkbox"/>		C&R Tree 16	2	129.000	35	2.019	80.0	7	0.856
<input checked="" type="checkbox"/>		C&R Tree 19	2	129.000	35	2.019	80.0	7	0.856

تصویر ۱۳- ده مدل برتر با استفاده از Auto Classifier و بدون Over sampling

تصویر ۱۰ - نمایش ایجاد مدل های SVM در Auto Classifier

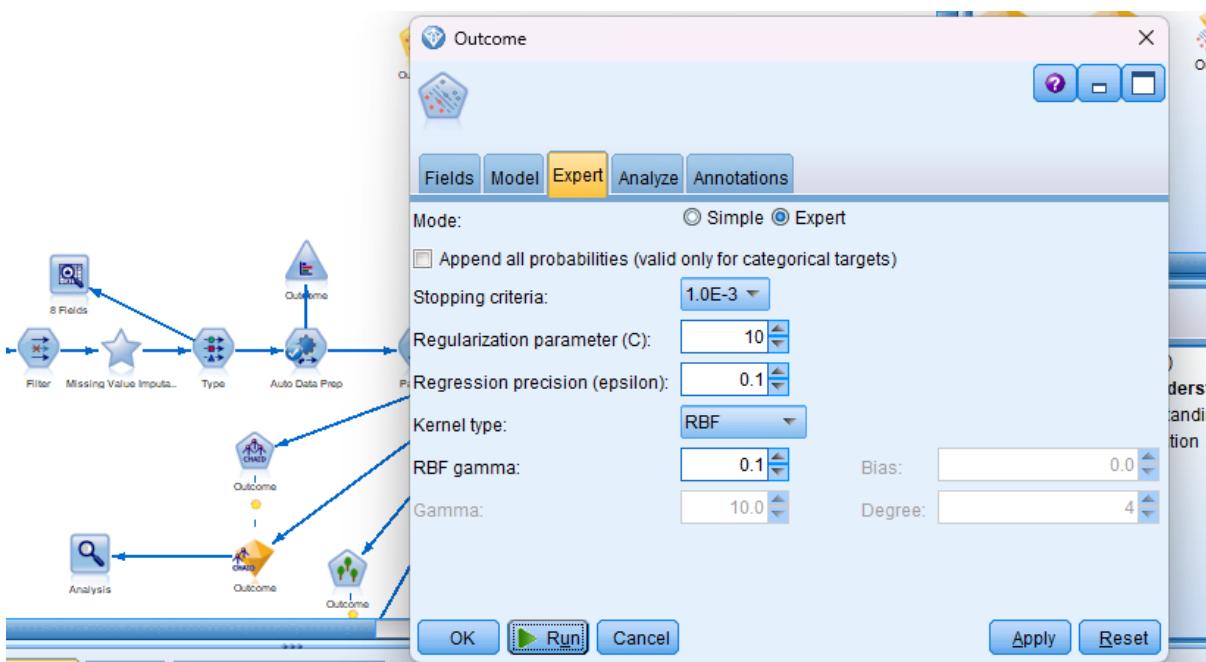
Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
<input checked="" type="checkbox"/>		SVM 1	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 2	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 3	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 5	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 6	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 7	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 11	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 12	3	130.0	45	1.852	80.0	7	0.855
<input checked="" type="checkbox"/>		SVM 13	3	130.0	45	1.852	80.0	7	0.855

تصویر ۱۱ - نمایش تنظیمات مدل over sampling

همانطور که مشاهده می کنید زمانی که از over sampling استفاده شد بهترین مدل را SVM و بدون آن مدل CART را بهترین انتخاب کرد که در ادامه تمام مدل ها را به صورت جداگانه بررسی کرده و در نهایت بهترین را انتخاب می کنیم.

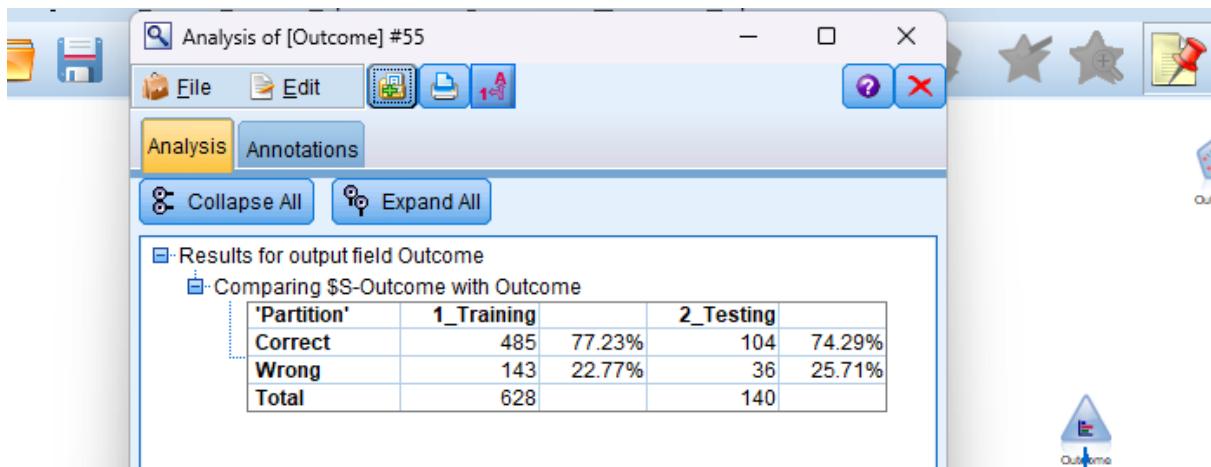
• مدل SVM (ماشین های پشتیبان بردار)

بهترین نتیجه ای که از این مدل توانستم بگیرم با استفاده از تنظیمات زیر بود:

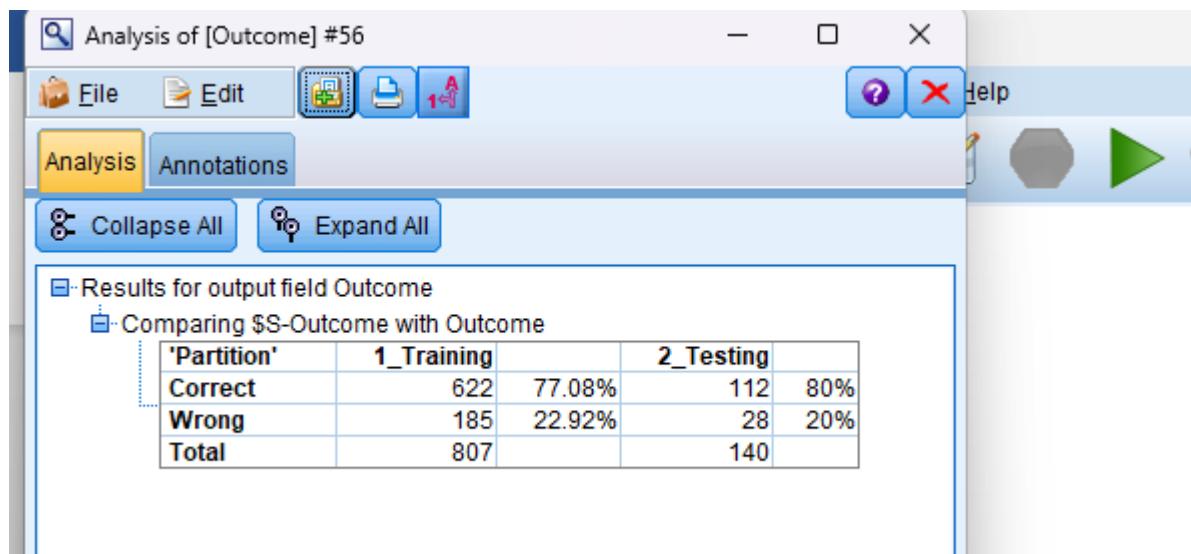


تصویر ۱۳ - تنظیمات مدل SVM

که نتیجه آن بدون استفاده و با استفاده از over sampling به صورت زیر شد:



تصویر ۱۴ - آنالیز مدل over sampling svm بدون

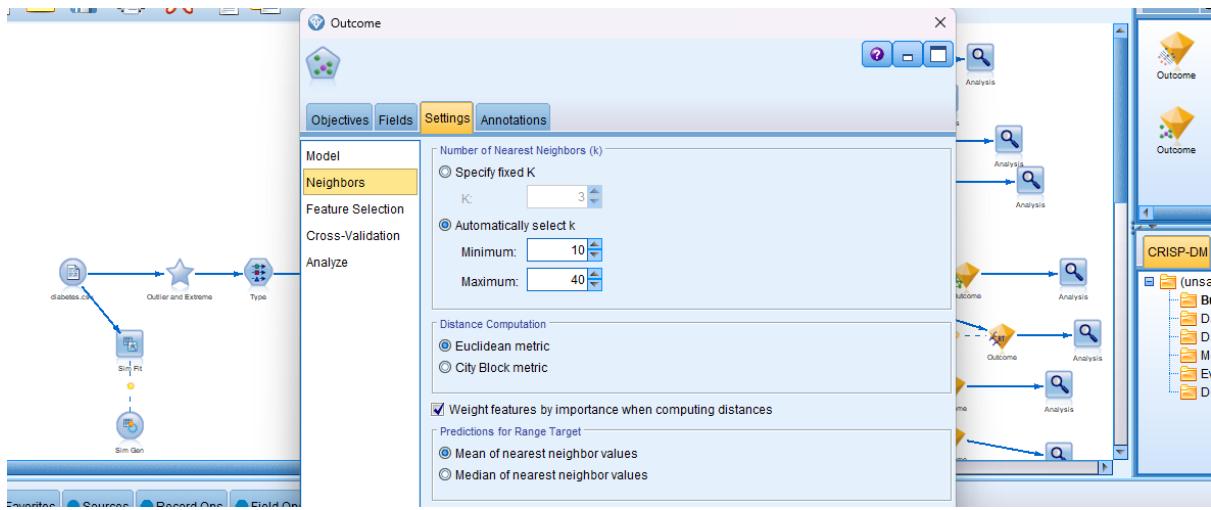


تصویر ۱۵ - آنالیز مدل over sampling ۴ svm

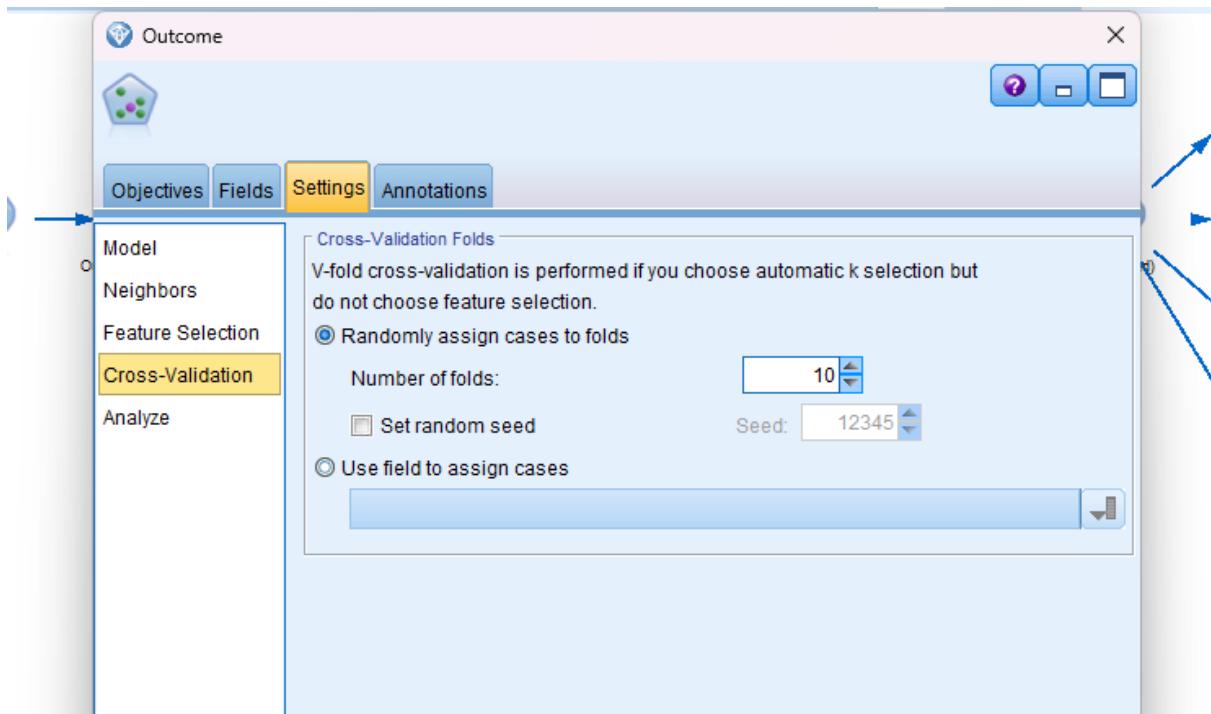
این مدل با استفاده از over sampling نتیجه بهتری داد.

KNN • مدل

در این مدل هم بهترین نتیجه ای که گرفتم با استفاده از تنظیمات زیر بود:

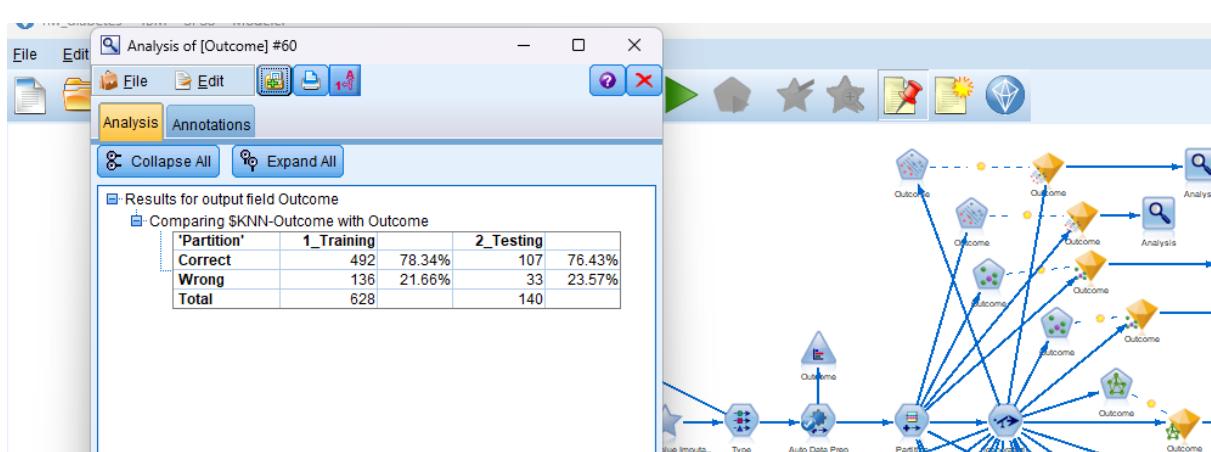
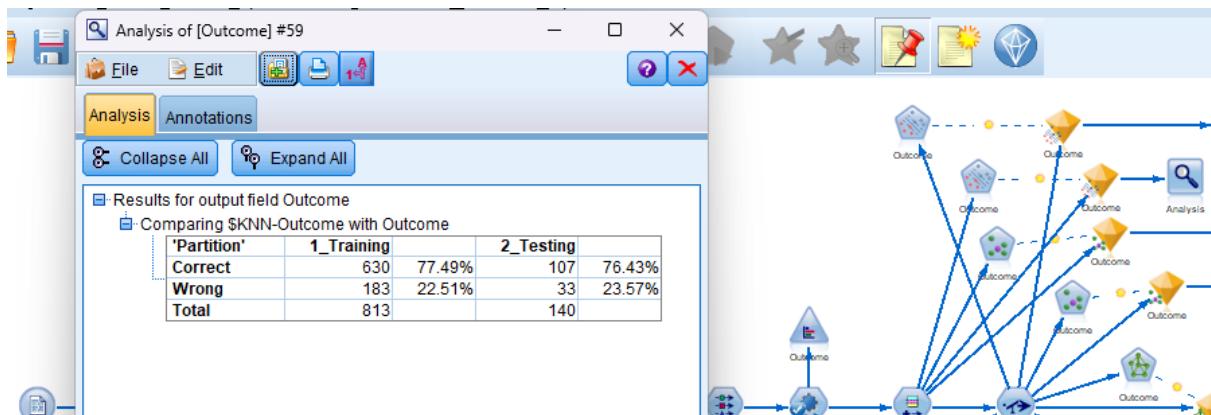


تصویر ۱۶- تنظیمات مدل (Neighbors) KNN



تصویر ۱۷- تنظیمات (Cross-Validation) KNN

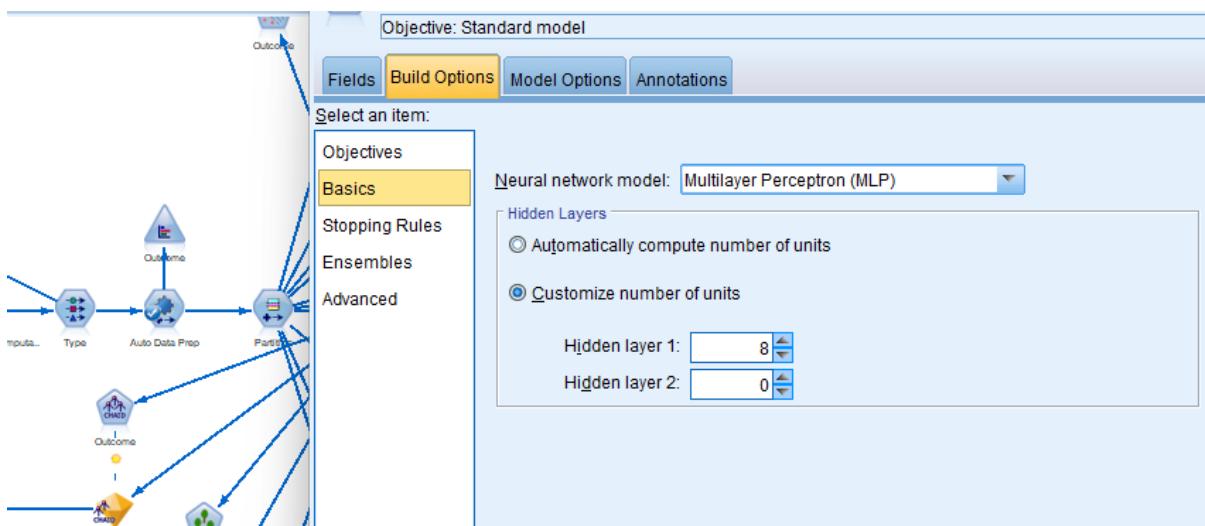
آنالیز این مدل نیز بصورت زیر شد.



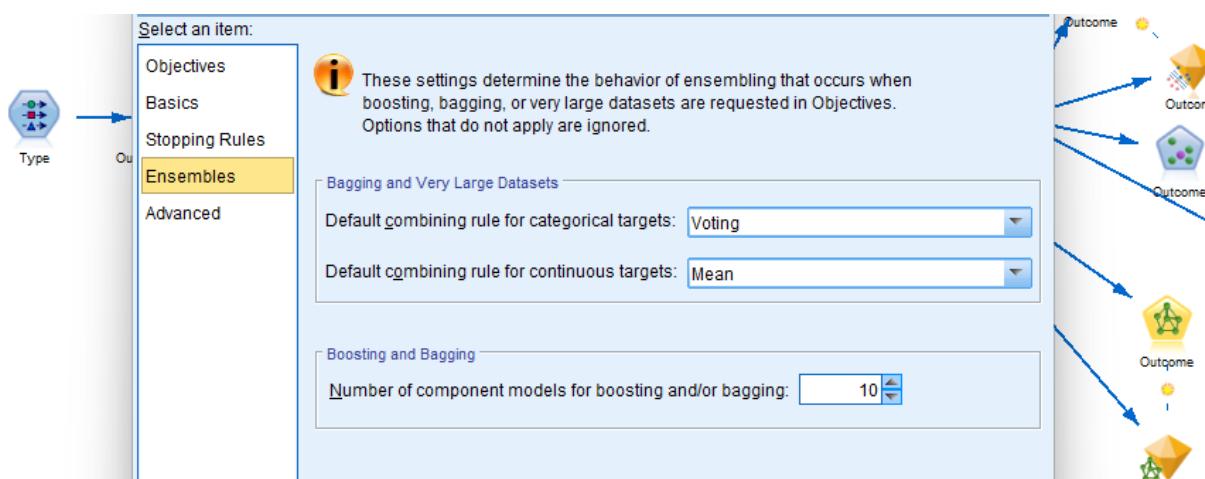
در این مدل تفاوت خاصی بین Over sampling و بدون آن نشد و نتیجه بهتری نسبت به مدل SVM نداد.

• مدل Neural net (شبکه های عصبی)

در این مدل بهترین نتیجه ای که توانستم بگیرم استفاده از روش استاندارد و تنظیمات زیر بود که نتیجه بهتری به من نداد.

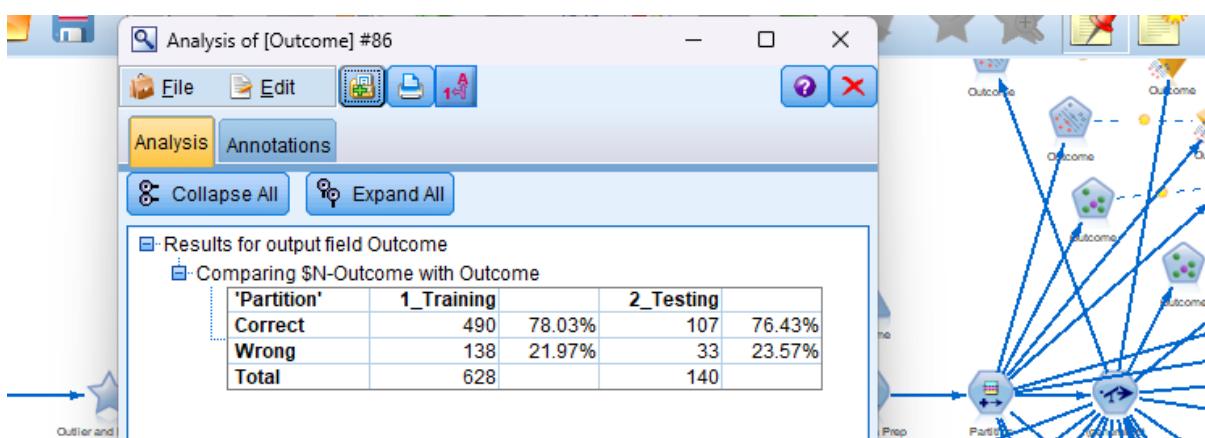


تصویر ۱۳- تنظیمات شبکه عصبی (Basics)

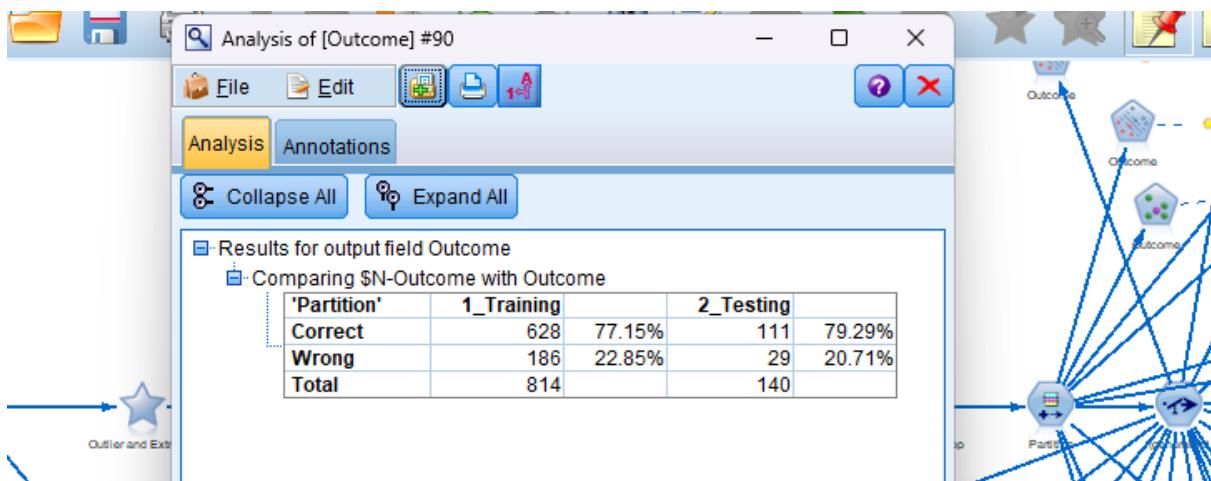


تصویر ۱۴- تنظیمات شبکه عصبی (Ensembles)

آنالیز این مدل با over sampling و بدون over sampling بصورت زیر می باشد.



تصویر ۱۵- آنالیز مدل شبکه عصبی بدون over sampling

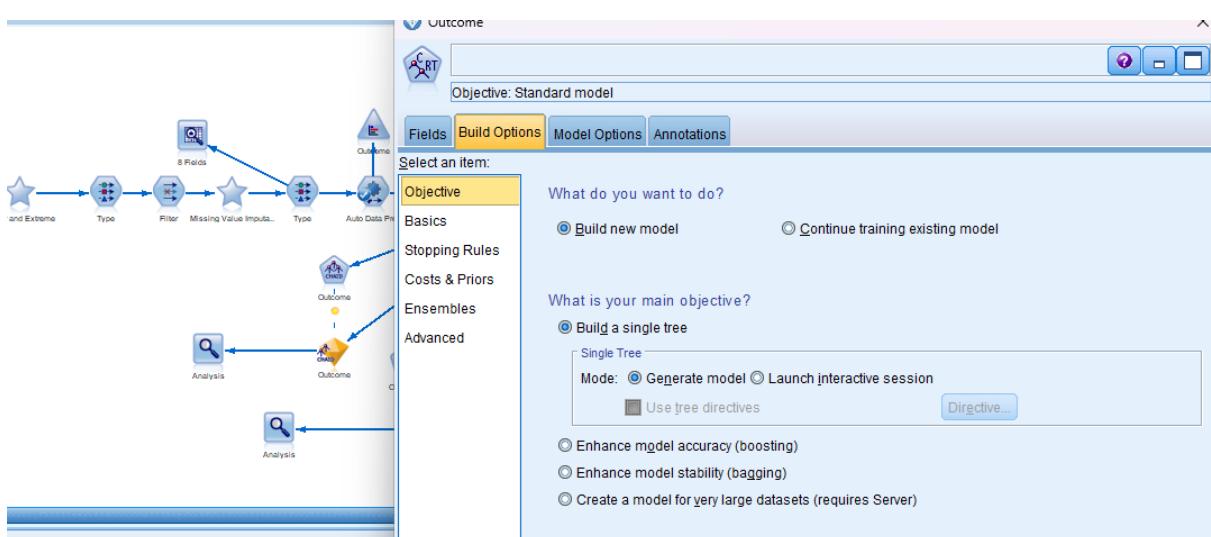


تصویر سعی- آنالیز شبکه عصبی با over sampling

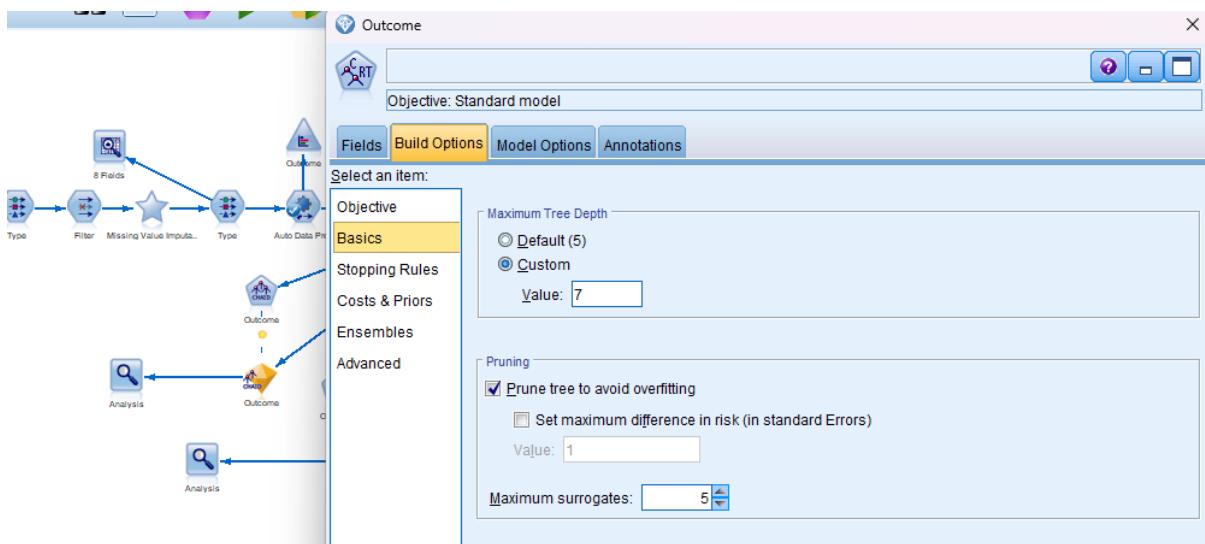
در این مدل هم با استفاده از over sampling نتیجه بهتری داد و نزدیک به مدل SVM شد.

• مدل Cart

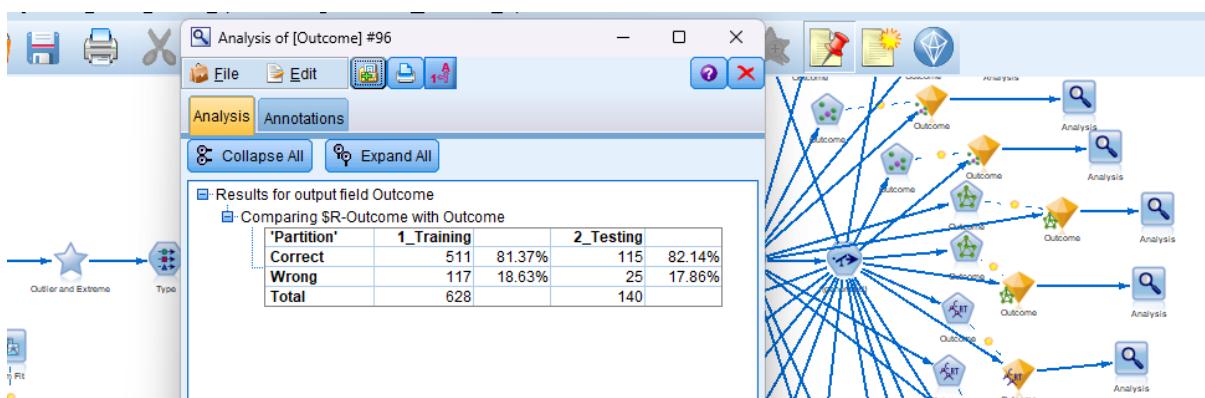
در این مدل با استفاده از تنظیمات زیر و بدون استفاده از over sampling به نتیجه بینه تری نسبت به مدل SVM رسیدم.



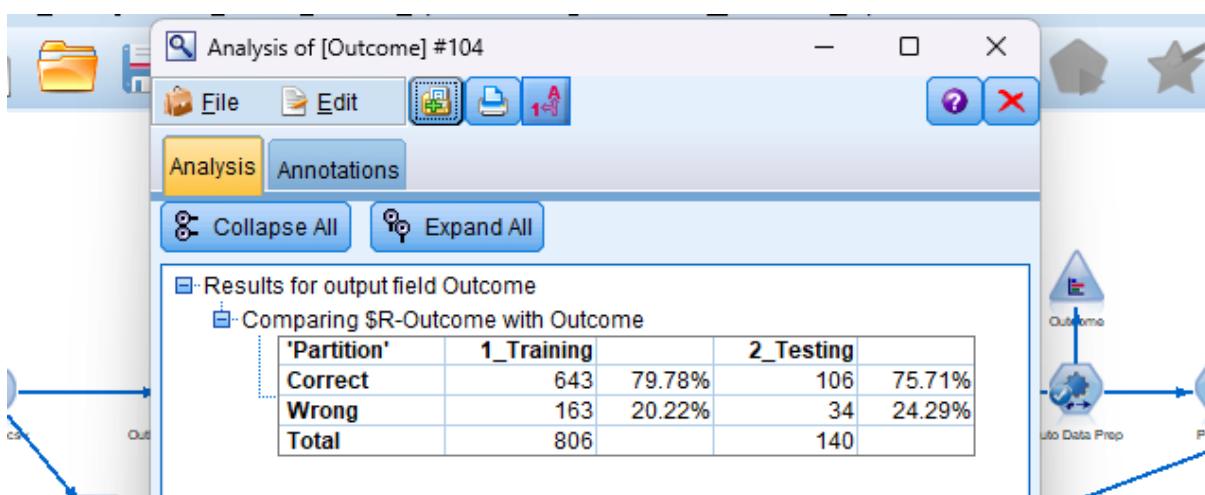
(Objective) Cart - تنظیمات تصویر



تصویر ۱۴-۱۵) ترتیبات مدل CART



تصویر ۱۶-۱۷) Over sampling با استفاده از CART

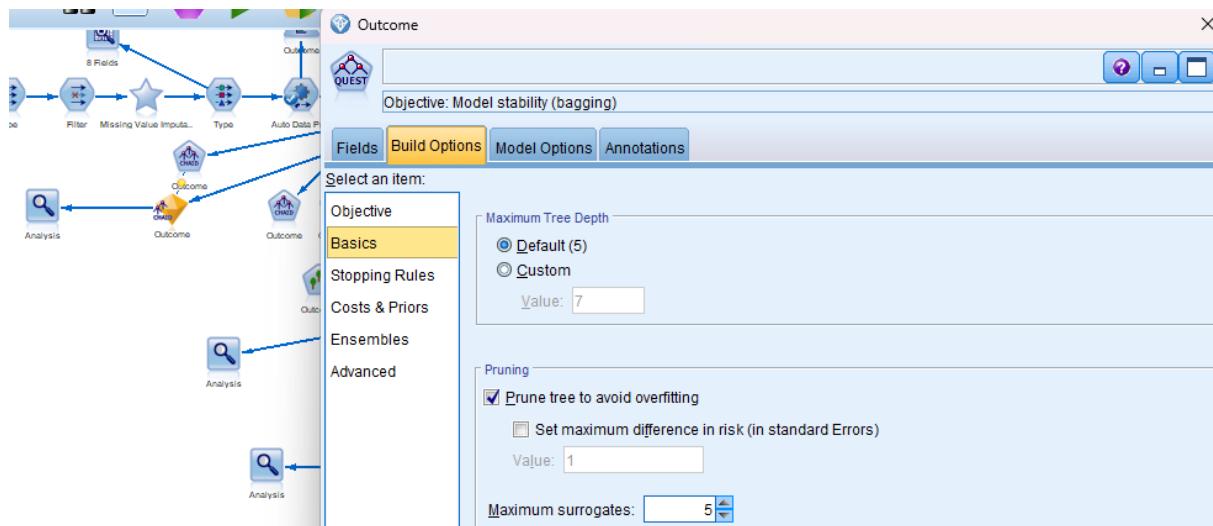


تصویر ۱۸-۱۹) Over sampling با استفاده از CART

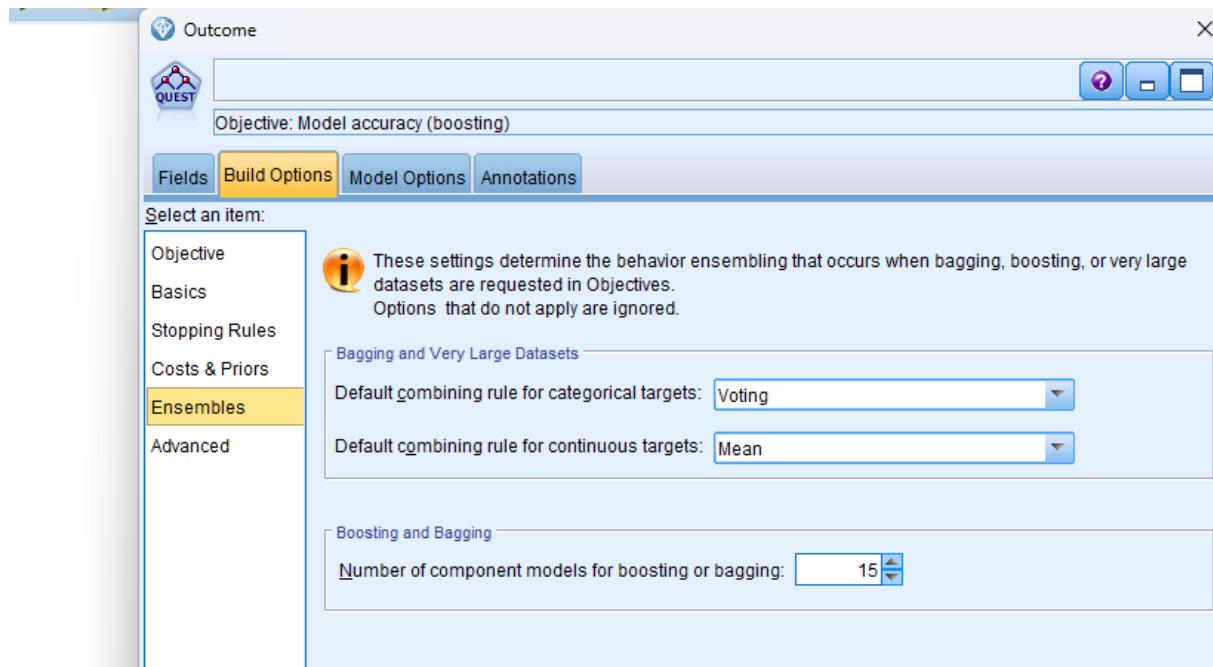
همانطور که مشاهده می کنیم این مدل نتیجه بهتری در زمانی که از over sampling استفاده نکردم داد.

• مدل Quest

این مدل، بهترین نتیجه با استفاده از روش Over sampling و تنظیمات زیر بدست آمد.



(Basics) Quest نتیجه‌ها - معرفی



(Ensembles) Quest نتیجه‌ها - معرفی

Analysis of [Outcome] #199

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field Outcome

Comparing \$R-Outcome with Outcome

'Partition'	1_Training	2_Testing
Correct	637 78.64%	107 76.43%
Wrong	173 21.36%	33 23.57%
Total	810	140

Over sampling jI داده‌ها را با Quest می‌بینیم

و همچنین بهترین نتیجه بدون استفاده از Over sampling و boosting و تنظیمات زیر بدست آمد.

Outcome

QUEST

Objective: Model accuracy (boosting)

Fields Build Options Model Options Annotations

Select an item:

Objective

Basics

Stopping Rules

Costs & Priors

Ensembles

Advanced

Maximum Tree Depth

Default (5)

Custom

Value: 7

Pruning

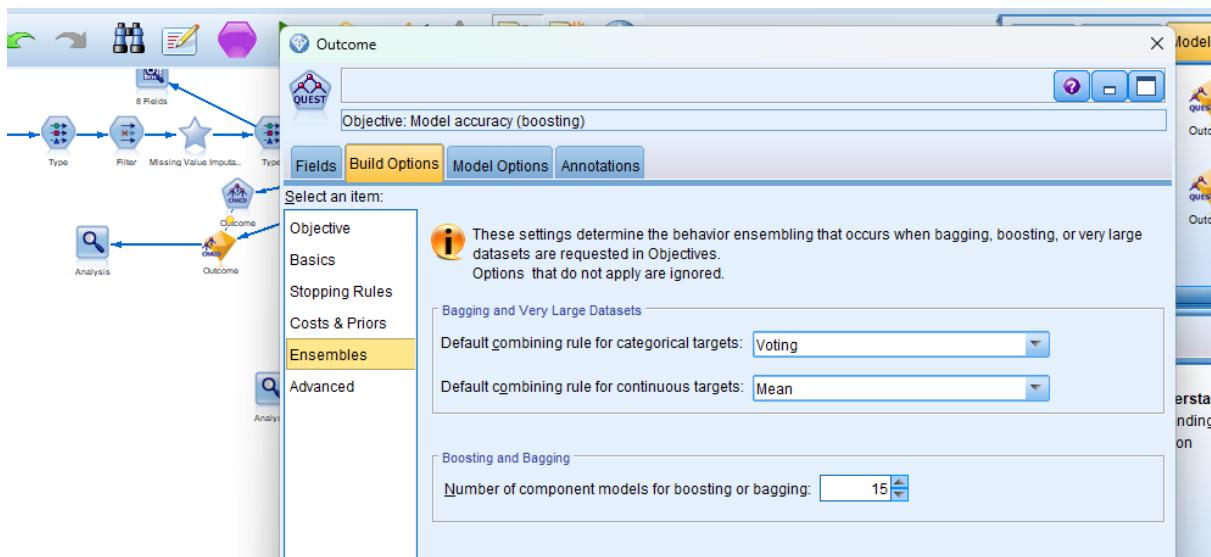
Prune tree to avoid overfitting

Set maximum difference in risk (in standard Errors)

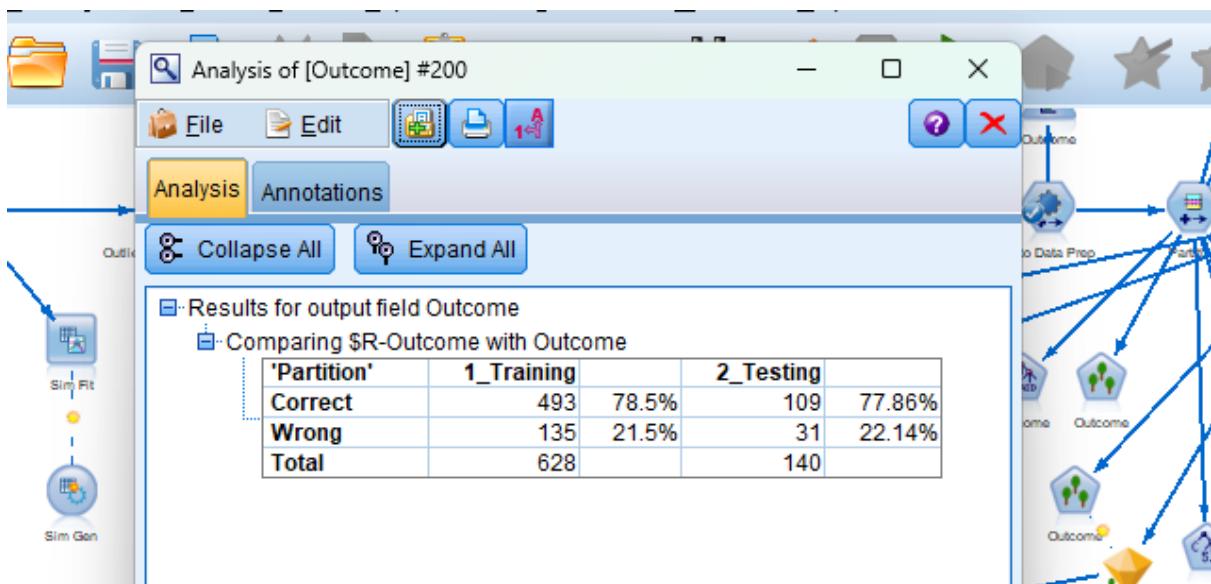
Value: 1

Maximum surrogates: 5

تصویر اول - تنظیمات (Basics) Quest



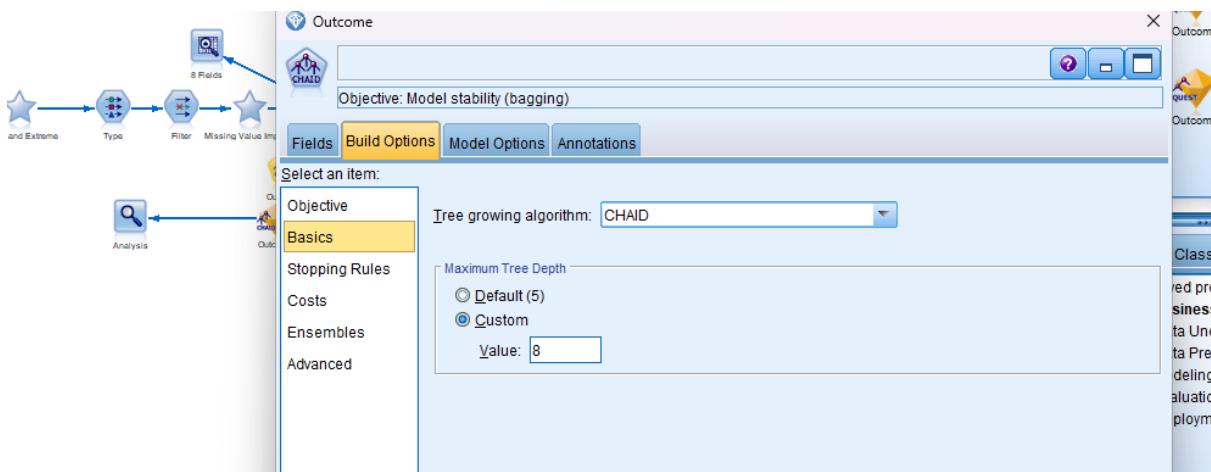
تصویر ۵-۲۷



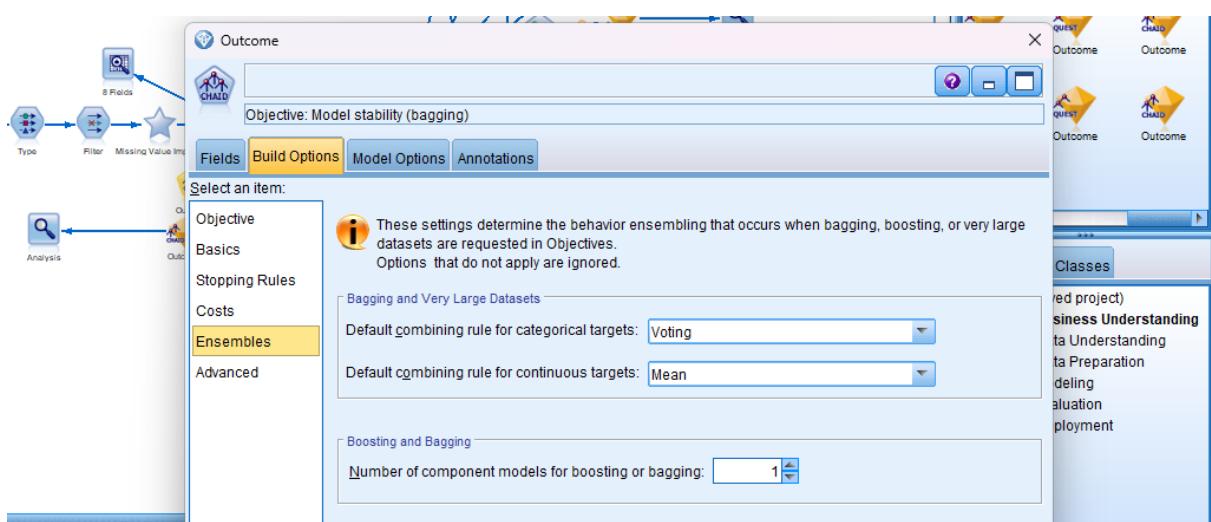
این مدل تیجه بهتری نسبت به مدل CART نداد.

• مدل CHAID

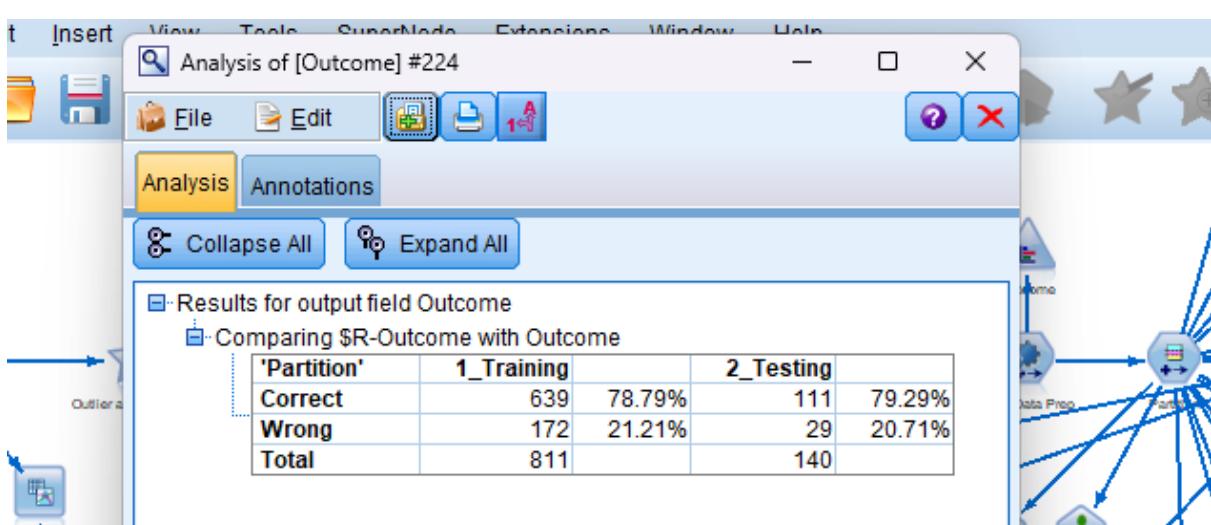
در این مدل هم زمانی که از over sampling استفاده می کنیم بهترین تیجه با استفاده از bagging و CHAID مدل زیر می باشد.



(Basics) CHAID تایپیتیں - مصوبہ

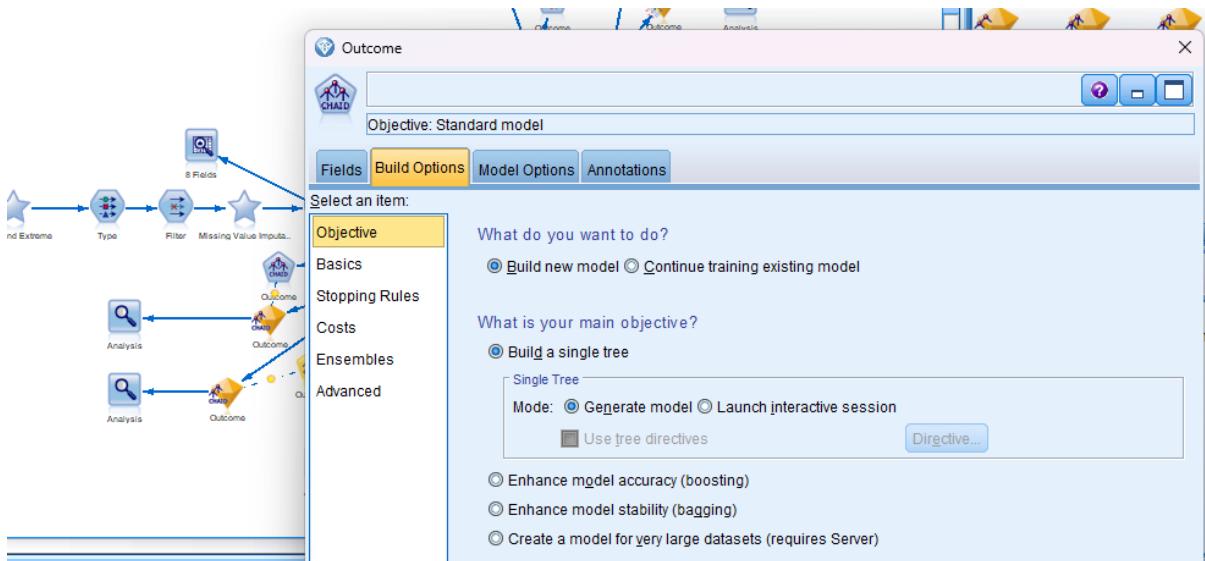


(Ensembles) CHAID تایپیتیں - مصوبہ

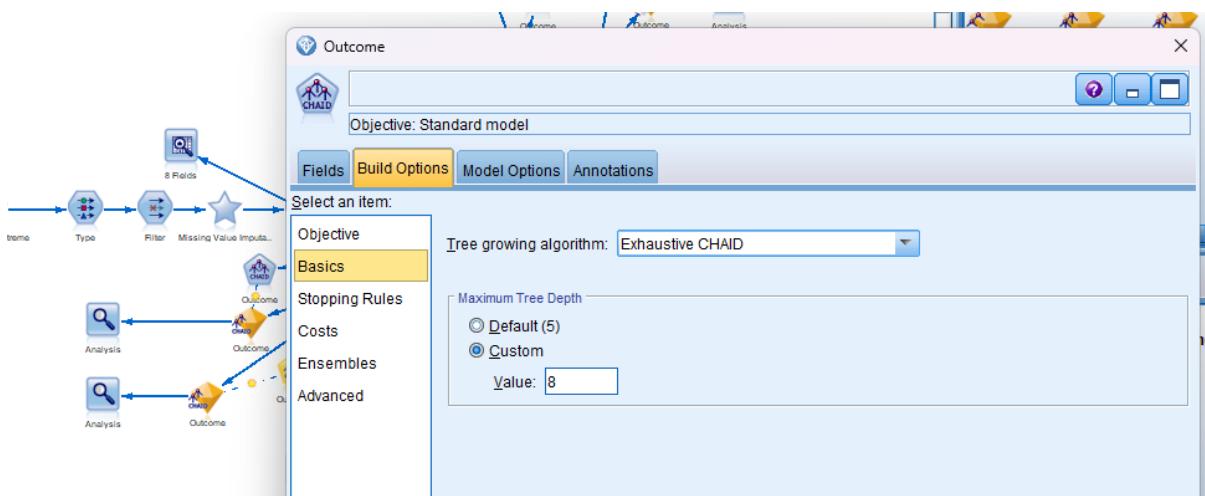


over sampling چاید تایپیتیں - مصوبہ

و بهترین نتیجه که بدون استفاده از over sampling بدست آمد بصورت زیر بود.



over sampling نجاح_ (objective) CHAID تطبيقات - DV مفهوم



over sampling نجاح_ (Basics) CHAID تطبيقات - DV مفهوم

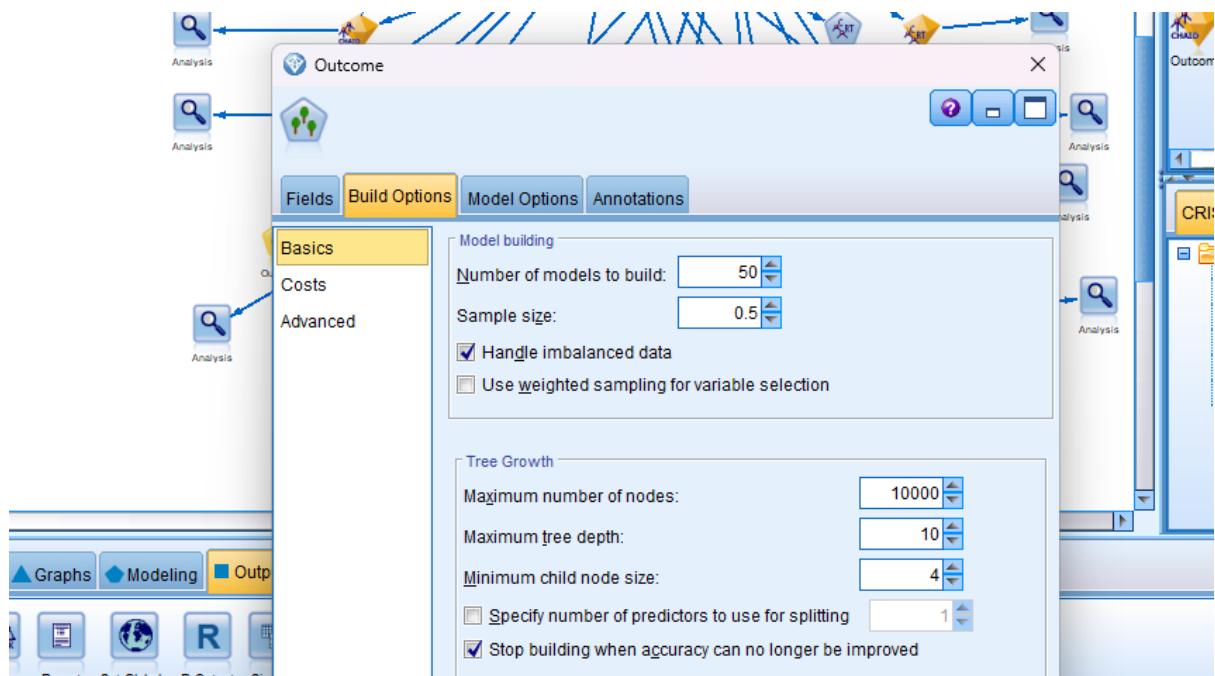
'Partition'	1_Training	2_Testing
Correct	479	76.27%
Wrong	149	23.73%
Total	628	
		74.29%
		25.71%

over sampling نجاح CHAID جعلی - DV مفهوم

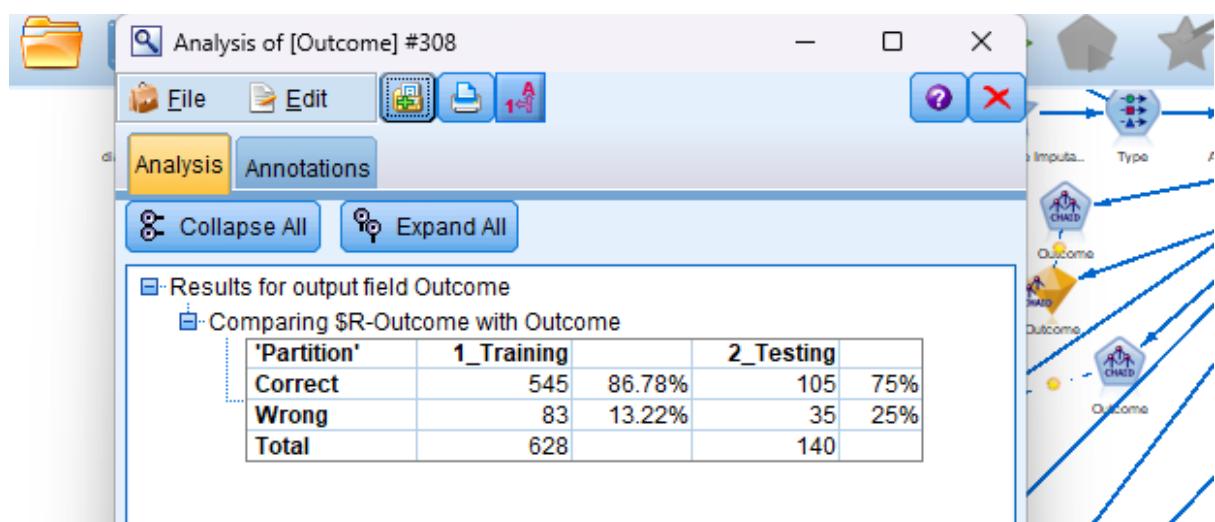
همانطور که از آنالیز مدل پیدا است تیجه بھینه تری نسبت به مدل CART نداد.

Random Trees •

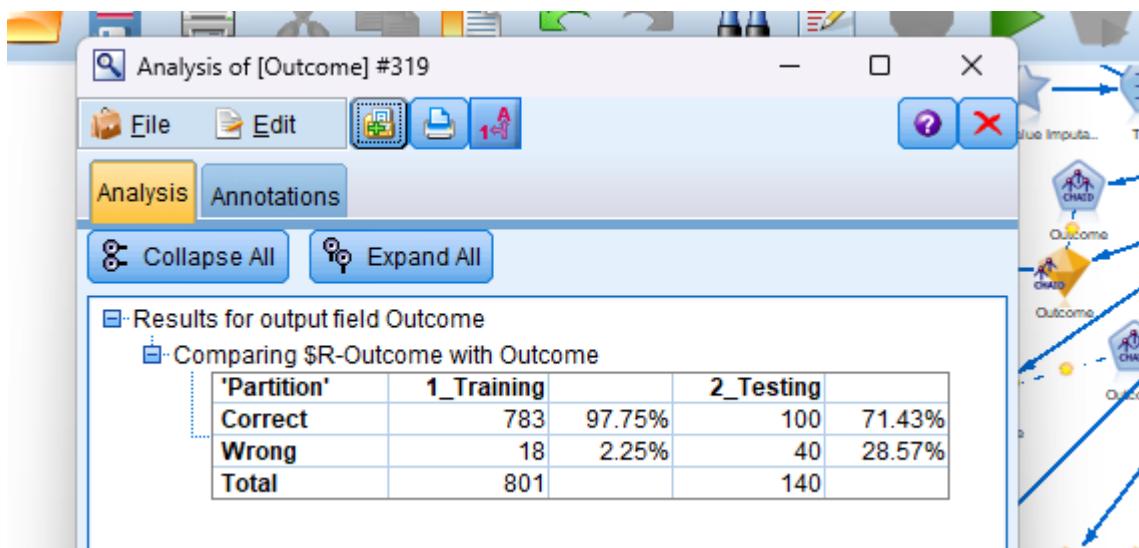
این مدل تیجه خوبی نداد و اکثر موقع اورفیت می شد.



تصویر ۴۰- تنظیمات Random Trees



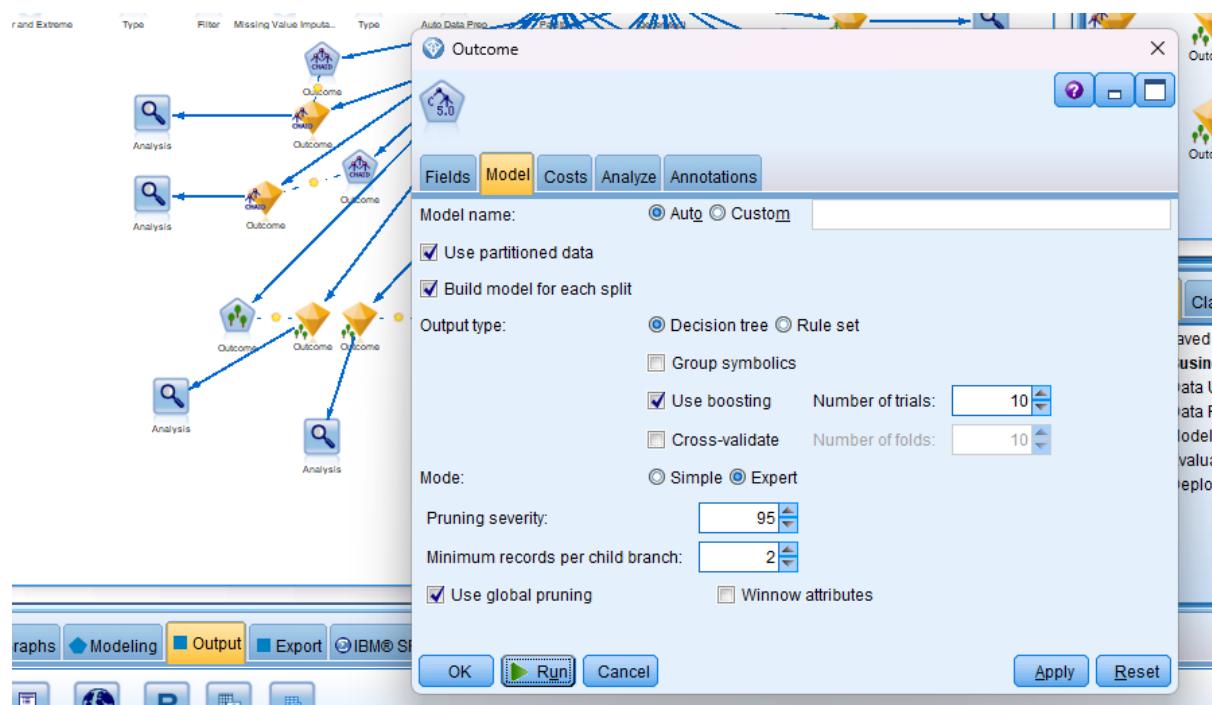
تصویر ۴۱- آنالیز over sampling بذوق Random Trees



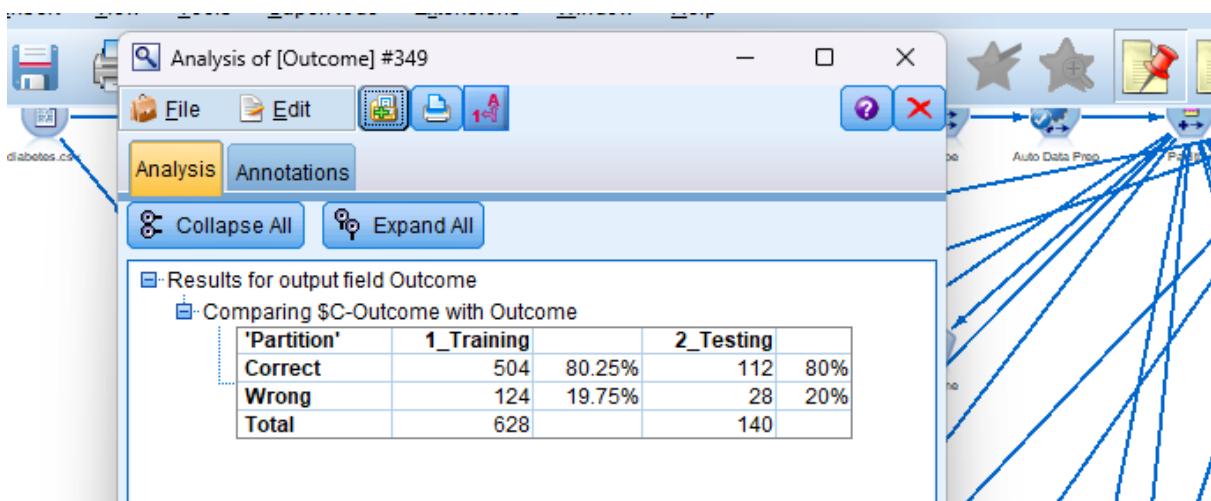
تصویر ۶-۷ over sampling یا Random Trees تأثیر می‌کند.

C5 مدل •

این مدل نیز نتیجه مشابهی نسبت به مدل SVM دارد.

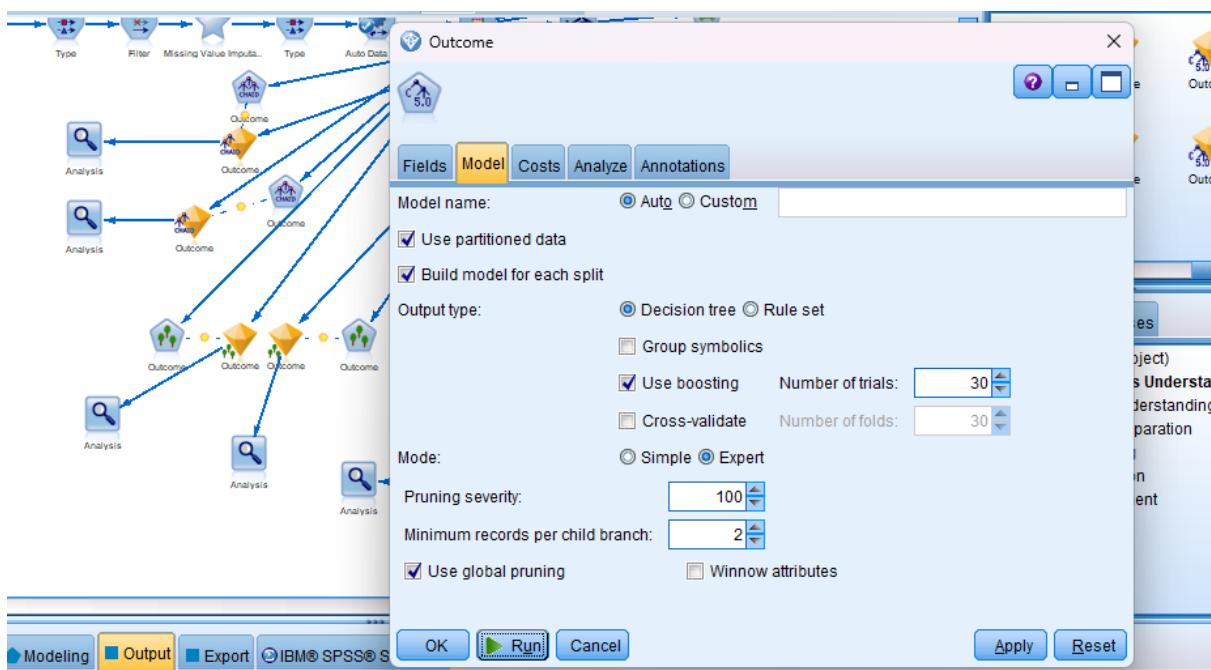


تصویر ۶-۸ تطبیقات C5

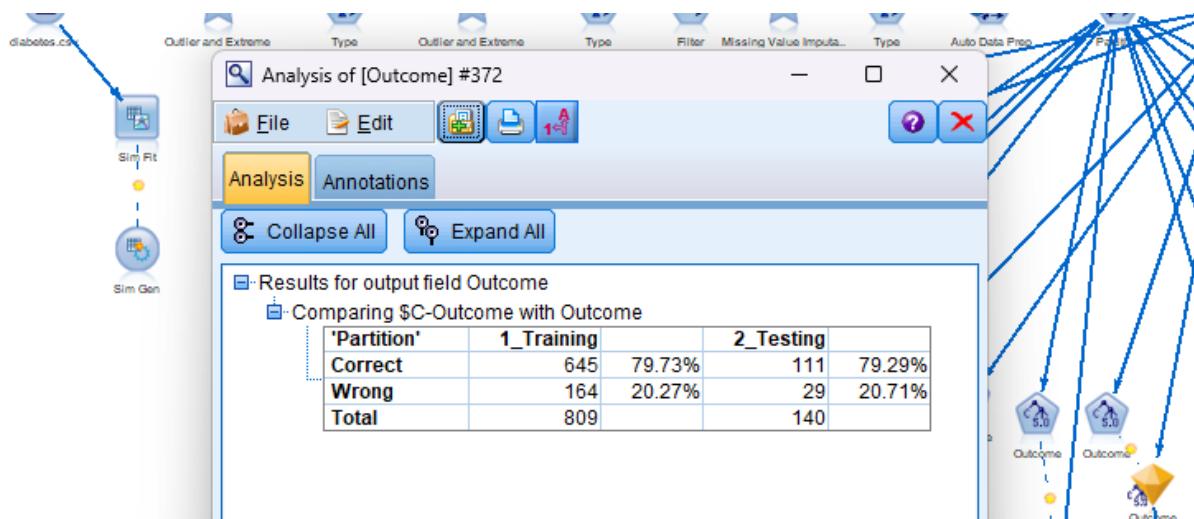


تصویر ۴۵- آنالیز C5 بدون over sampling

و بهترین تنظیمات آن زمانی که از over sampling استفاده کردم بدین صورت بود.



تصویر ۴۶- تنظیمات over sampling



تصویر ۶۶- آنالیز C5 over sampling

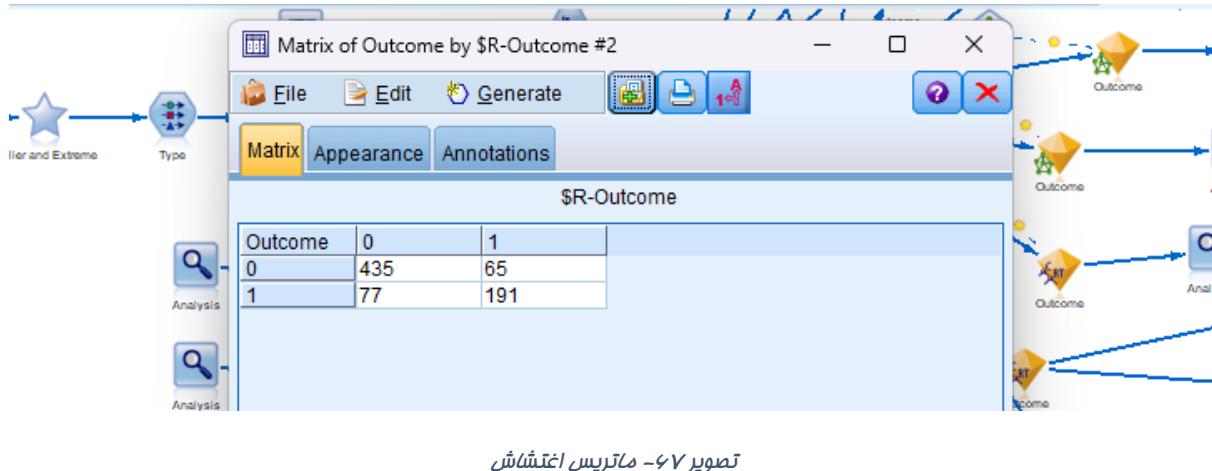
خلاصه

در این فصل انواع مختلف مدل های مربوط به مسئله طبقه بندی را بررسی کردیم و به دقت ۸۲ درصد با استفاده از مدل CART رسیدم که نسبت به سایر مدل ها بهترین نتیجه را داد. در فصل بعد با توجه به مشخصه های ارزیابی مدل این مدل را بررسی می کنیم.

فصل ۵: ارزیابی مدل

مقدمه

در این فصل با استفاده از ابزار Matrix از بخش output ماتریس اغتشاش را رسم کرده و طبق فرمول های هر شاخص ارزیابی آن را برای بهترین مدلی که داشتیم (CART) محاسبه می کنیم.



صحت

در اینجا طبق فرمول $\frac{TP+TN}{TP+TN+FP+FN}$ صحت این مدل با توجه به ماتریس به صورت زیر می شود.

$$\frac{435 + 191}{435 + 191 + 65 + 77} = \frac{626}{768} = 0.81$$

صحت این مدل ۸۱ درصد می باشد.

دقت

طبق فرمول $\frac{TP}{TP+FP} = \frac{191}{191+65} = 0.74$ دقت این مدل ۷۴ درصد می باشد.

حساسیت صحت

طبق فرمول $\frac{TP}{TP+FN} = \frac{191}{191+77} = 0.71$ حساسیت این مدل ۷۱ درصد می باشد.

ویژگی

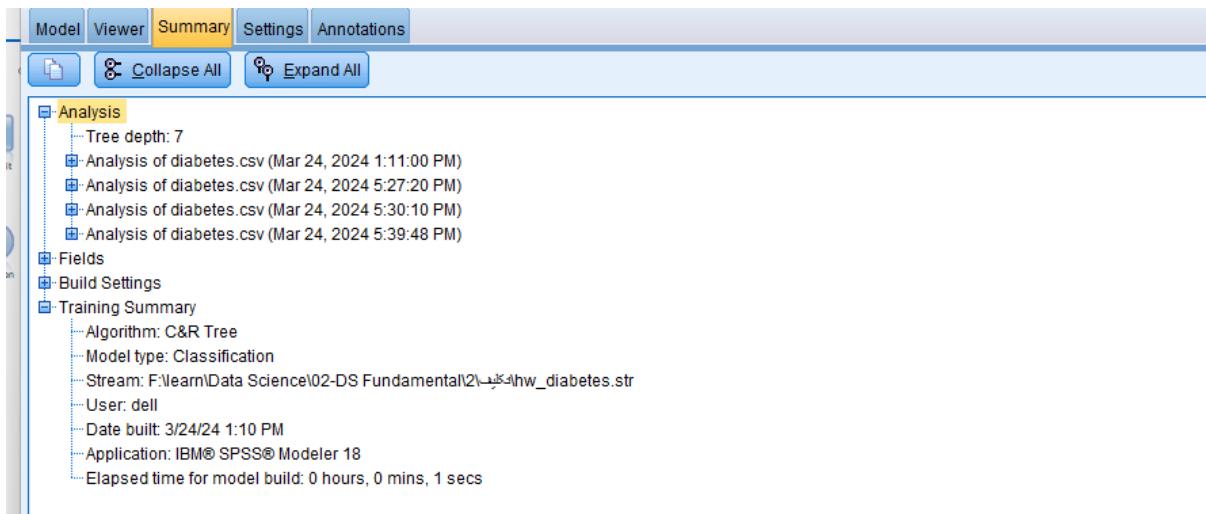
طبق فرمول $\frac{TN}{TN+FP} = \frac{435}{435+65} = 0.87$ ویژگی این مدل ۸۷ درصد می باشد.

F1 امتیاز

طبق فرمول $\frac{2TP}{2TP+FP+FN} = \frac{382}{382+77+65} = 0.72$ امتیاز F1 این مدل ۷۲ درصد می باشد.

سرعت

طبق تصویر این مدل در عرض ۱ ثانیه اجرا شد.



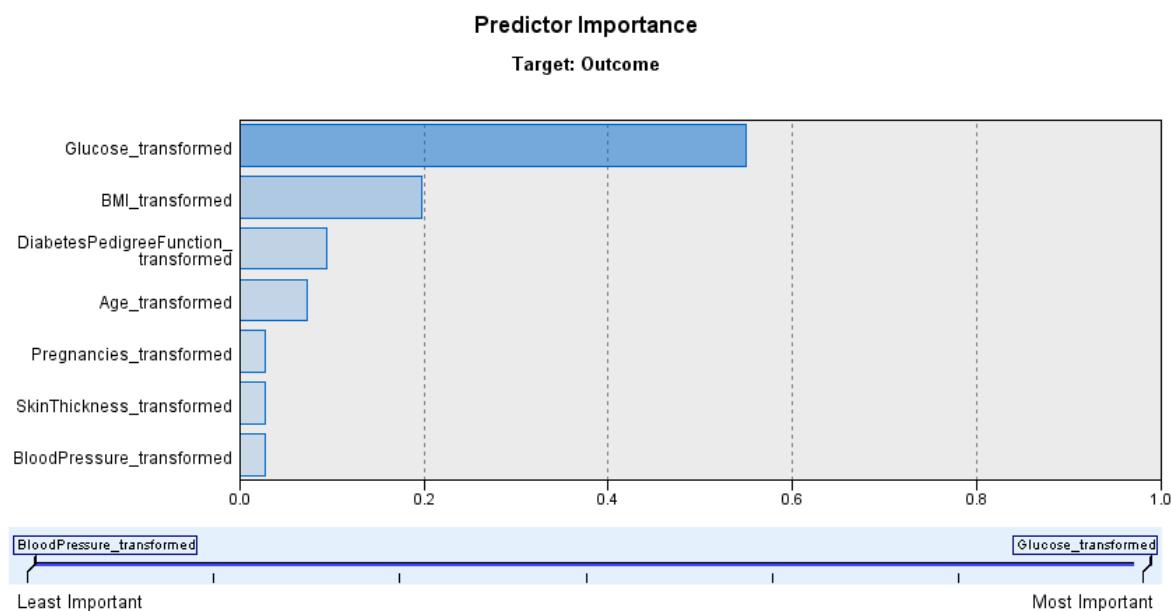
تصویر ۶۸- سرعت اجرای مدل برتر

پایداری

با توجه به تصویر ۲ مشاهده می کنید که تعداد زیادی از داده ها مفقوده بودند و مدل توانست با دقت ۸۱ درصد به جواب برسد.

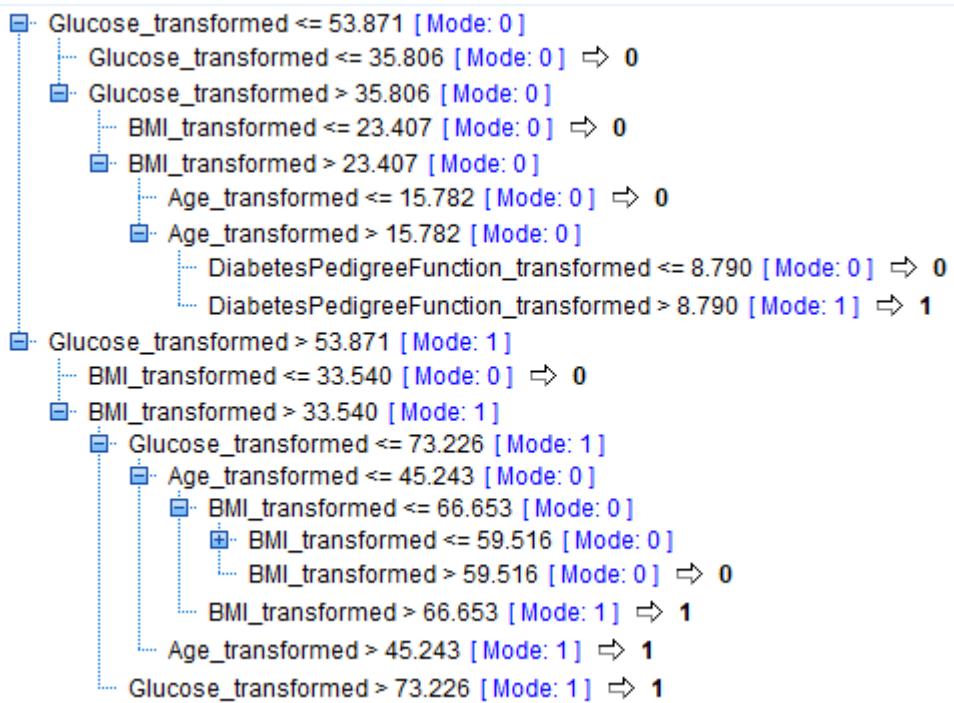
تفسیر پذیری

طبق تصویر اهمیت هر ویژگی را مشاهده می کنید.

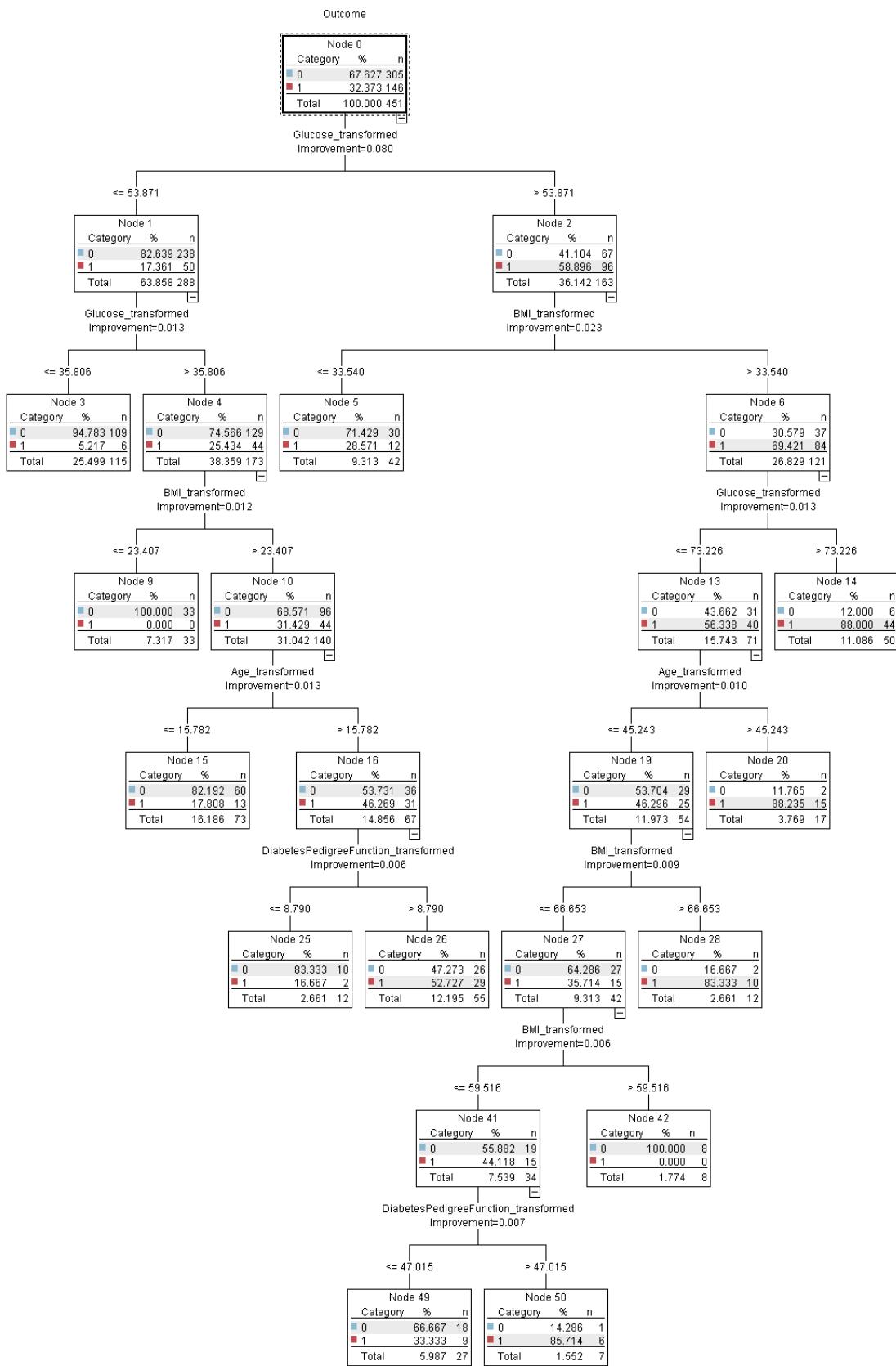


تصویر ۶۹- اهمیت هر ویژگی

همچنین در تصویر زیر مراحل مدل شرح داده شده است که نشان دهنده قابل فهم بودن مدل می باشد.



تصویر ۷ - مرحله انجام مدل برتر



تصویر ۶-۱- مراحل گرافیکی مدل برآورد

خلاصه

شاخص های ارزیابی مدل را بررسی کردیم و طبق آن به دقت ۸۱ درصد رسیدیم که دقت بالایی را نشان نمی دهد ولی نسبت به سایر مدل ها بهترین نتیجه را داد. طبق گوگلی که کرده بودم برای این دیتابست میزان دقت را بین ۷۵ تا ۸۳ به دست آورده بودند که به نظرم این میزان دقت برای حل این مسئله قابل قبول باشد.