مرکز آموزش های کاربردی و حرفه ای دانشگاه تهران

پروژه پیش بینی ابتلا به دیابت

استاد:

دکتر محمدرضا محتاط

دانشجو:

بهنام امان اللهی

دی ۱۴۰۲

فهرست

۱. شناخت کسب و کار:

این مجموعه داده در اصل از موسسه ملی دیابت و بیماری های گوارشی و کلیوی تهیه شده است. این داده‌ها از اعمال محدودیت هایی روی یک پایگاه داده بزرگتر استخراج شده است. به طور خاص، همه بیماران در اینجا زنان حداقل ۲۱ ساله هندی پیما هستند. هدف مجموعه داده این است که بر اساس اندازه‌گیری‌های تشخیصی خاص موجود در مجموعه داده، پیش بینی کند که آیا بیمار مبتلا به دیابت است یا خیر. هدف از این پروژه پیش بینی ابتلای یک فرد به بیماری دیابت بر اساس ویژگی های زیر است:

۱. Pregnancies:

تعداد بارداری فرد

۲. Glucose:

تست تحمل گلوکز خوراکی (OGTT) یک آزمون پزشکی است که در آن ۷۵ گرم گلوکز خوراکی به بیمار داده شده و پس از ۲ ساعت از بیمار نمونه خون گرفته و سرعت پاک سازی آن از خون در آن بررسی می‌شود. از این تست معمولاً برای تشخیص دیابت، مقاومت به انسولین و اختلال سلول های بتا استفاده می‌شود.

میزان گلوکز خون کمتر از ۱۴۰ mg/dL نرمال، ۱۴۰-۱۹۹ mg/dL نشان دهنده اختلال در تحمل گلوکز و ۲۰۰ mg/dL و بیشتر از آن نشان دهنده ابتلا به دیابت است.

۳. BloodPressure:

قلب بین ضربان ها، استراحت می‌کند تا بتواند دوباره با خون پر شود. پزشکان این مکث بین ضربان‌ها را «دیاستول» می‌نامند. فشار خون دیاستولیک، در واقع اندازه گیری در مدت این مکث و پیش از ضربان قلب بعدی است. فشار خون دیاستولیک نشان می‌دهد در حالی که عضله قلب در بین انقباضات در حال استراحت است، خون چقدر فشار روی دیواره شریان وارد می‌کند.

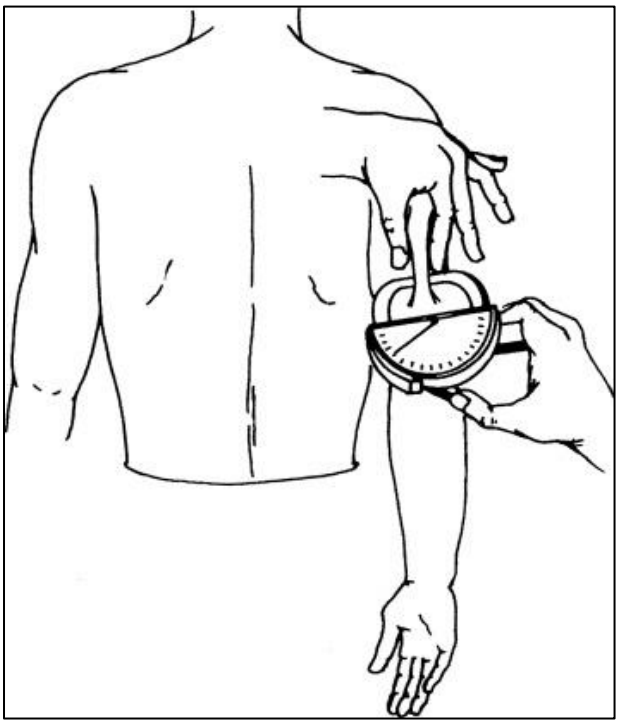
مقادیر مختلف این ویژگی در جدول زیر آورده شده است.

BLOOD PRESSURE CATEGORY	**DIASTOLIC mm Hg (lower number)**
NORMAL	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 120

جدول ۱: درجه بندی فشار خون

۴. SkinThickness:

ضخامت چین های پوستی سه سر بازو که توسط کولیس لانگ بر حسب میلیمتر اندازه گیری می‌شود. (تصویر زیر)



شکل ۱: نحوه اندازه گیری ضخامت چین های پوستی

۵. Insulin:

انسولین سرم ۲ ساعته یک روش اندازه گیری سطح انسولین در خون دو ساعت پس از مصرف یک نوشیدنی شیرین است. انسولین هورمونی است که به بدن کمک می کند از گلوکز یا قند برای انرژی استفاده کند. سطوح بالای انسولین می‌تواند نشانه‌ای از مقاومت به انسولین باشد، وضعیتی که استفاده موثر از انسولین را برای بدن دشوار می کند.

در اینجا جدولی از سطوح طبیعی انسولین در زمان های مختلف پس از آزمایش چالش گلوکز آمده است:

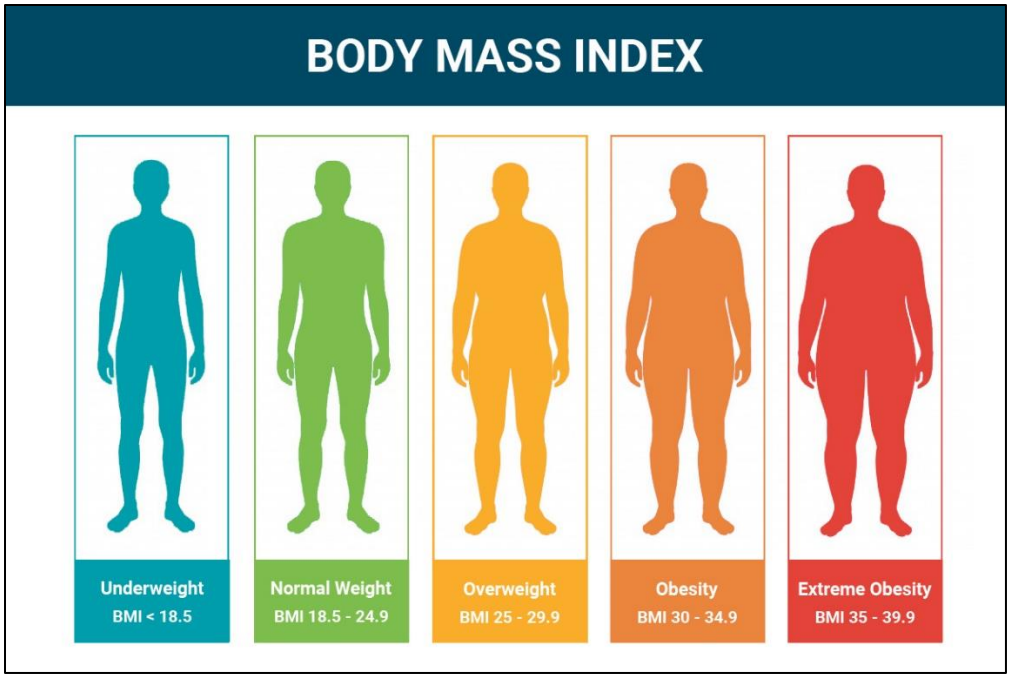
سطح انسولین	توضیح
< 25 µU/mL	ناشتا
30-230 µU/mL	۳۰ دقیقه پس از مصرف گلوکز
18-276 µU/mL	۱ ساعت پس از مصرف گلوکز
16-166 µU/mL	۲ ساعت پس از مصرف گلوکز
< 25 µU/mL	۳ ساعت و بیشتر پس از مصرف گلوکز

جدول ۲: درجه بندی میزان انسولین

۶. BMI:

شاخص توده بدنی یا بی‌ام‌آی سنجشی آماری برای مقایسه وزن و قد یک فرد است. در واقع این سنجش میزان چاقی را اندازه‌گیری نمی‌کند بلکه ابزاری مناسب است تا سلامت وزن فرد با توجه به قدش تخمین زده شود.

$$BMI = \frac{weight(kg)}{height^2(m)}$$



شکل ۲: درجه بندی شاخص توده بدنی

۷. DiabetesPedigreeFunction:

Diabetes pedigree function (DPF) یک اندازه‌گیری کمی است که استعداد ژنتیکی فرد برای ابتلا به دیابت نوع ۲ (T2DM) را بر اساس سابقه خانوادگی بیماری تخمین می‌زند. DPF بالاتر نشان‌دهنده خطر ژنتیکی بالاتر برای T2DM است. افراد با DPF برابر ۰.۲ یا بیشتر در معرض خطر ابتلا به T2DM قرار دارند. DPF یک ابزار تشخیصی برای T2DM نیست، اما می‌توان از آن برای غربالگری زودهنگام T2DM استفاده کرد.

۸. Age:

امکان ابتلا به دیابت نوع ۲ در هر سنی، حتی در دوران کودکی وجود دارد. با این حال، دیابت نوع ۲ اغلب در افراد میانسال و مسن رخ می دهد. هرچند احتمال ابتلا به دیابت نوع ۲ در افراد بالای ۴۵ سال بیشتر است.

۹. Outcome:

نشان دهنده ابتلا (1) یا عدم ابتلای فرد (0) به دیابت است.

۲. درک داده:

تمامی ویژگی ها بجز outcome داده های عددی هستند لذا از نوع continuous در نظر گرفته شده است.

ویژگی outcome چون دو حالت 0 و 1 دارد، از نوع flag و نقش آن Target تعریف شده است.

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

شکل ۳: نوع ویژگی های ورودی

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33....	0.627	50	1
2	1	85	66	29	0	26....	0.351	31	0
3	8	183	64	0	0	23....	0.672	32	1
4	1	89	66	23	94	28....	0.167	21	0
5	0	137	40	35	168	43....	2.288	33	1
6	5	116	74	0	0	25....	0.201	30	0
7	3	78	50	32	88	31....	0.248	26	1
8	10	115	0	0	0	35....	0.134	29	0
9	2	197	70	45	543	30....	0.158	53	1
10	8	125	96	0	0	0.0...	0.232	54	1

شکل ۴: بخشی از داده ورودی

۲-۱. نوع توزیع ویژگی ها:

برای یافتن نوع توزیع ویژگی ها دو روش Anderson-Darling و Kolmogorov-Smirnov وجود دارد که تفاوت آنها در انتخاب نوع توزیع ویژگی BMI است. در روش اول نوع توزیع را گاما و در روش دوم نوع توزیع را نرمال محاسبه کرده است. با توجه به اینکه این ویژگی دارای میانگین ۳۲.۴ میانه ۳۲.۳ و مد ۳۲ است بنابراین میتوان با اندکی اغماض آن را نرمال در نظر گرفت. پس از روش Kolmogorov-Smirnov برای این کار استفاده شده است.

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...	

شکل ۵: نوع توزیع ویژگی ها

۲-۲. چالش هم‌خطی:

با فرض اینکه مقدار همبستگی مجاز برابر ۰.۷ باشد، چالش هم خطی ندارد. اما ویژگی های Pregnancies و Age و همچنین SkinThickness و BMI به ۰.۷ نزدیک هستند. از این رو بعد از پاکسازی داده ها مجددا آنها را بررسی میکنیم.

	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

شکل ۶: همبستگی ویژگی ها

۳. آماده سازی و پیش پردازش داده:

۳-۱. مدیریت داده های نویزی:

پنج ویژگی Glucose, BloodPressure, SkinThickness, Insulin, BMI مقادیر 0 دارند که مجاز نیست. بنابراین با کمک یک نود Filler آنها را به null تبدیل شده است.

	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

شکل ۷: مدیریت داده های نویزی

۳-۲. مدیریت داده های پرت:

با توجه به نرمال بودن نوع توزیع دو ویژگی Blood Pressure و BMI از روش تبدیل Z استفاده شده است.

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	None
Glucose	Continuous	0	0	None
BloodPressu...	Continuous	8	0	Coerce
SkinThickness	Continuous	1	1	None
Insulin	Continuous	7	1	None
BMI	Continuous	3	1	Coerce
DiabetesPed...	Continuous	7	4	None
Age	Continuous	5	0	None
Outcome	Flag	--	--	--

شکل ۸: تبدیل Z برای مدیریت داده های پرت

- به دلیل اینکه تعداد داده های پرت و خیلی پرت در ویژگی BloodPressure و BMI زیاد نیست از Coerce برای جایگزینی آنها استفاده شده است.

برای سایر ویژگی ها از روش IQR استفاده شده است.

Field	Measurement	Outliers	Extremes	Action
Pregnancies	Continuous	4	0	Coerce
Glucose	Continuous	0	0	None
BloodPressu...	Continuous	14	0	None
SkinThickness	Continuous	2	1	Coerce
Insulin	Continuous	16	8	Nullify
BMI	Continuous	8	0	None
DiabetesPed...	Continuous	23	6	Nullify
Age	Continuous	9	0	Coerce
Outcome	Flag	--	--	--

شکل ۹: IQR برای مدیریت داده های پرت

- برای ویژگی های Pregnancies, SkinThickness, Age به دلیل کم بودن تعداد داده های پرت و خیلی پرت از Coerce استفاده شده است.
- برای ویژگی های Insulin, DiabetesPedigreeFunction به دلیل زیاد بودن تعداد داده های پرت و خیلی پرت از Nullify استفاده شده است تا از ایجاد یک Bin بلند در نمودار فراوانی جلوگیری شود. در نتیجه آنها را مانند دادگان مفقوده در نظر گرفته و در مرحله بعد با الگوریتم پر می‌شود.
- روی ویژگی Glucose به دلیل نداشتن داده پرت و خیلی پرت هیچ عملی انجام نشده است.

۳-۳. مدیریت داده های مفقوده:

ویژگی هایی که درصد Complete آنها ۱۰۰ نباشد دارای داده مفقوده هستند. برای مدیریت آنها به روش های زیر عمل می‌شود:

Field	Measurement	Impute Missing	Method	% Complete	Valid Records	Null Value
Insulin	Continuous	Never	Fixed	48.177	370	398
SkinThickness	Continuous	Never	Fixed	70.443	541	227
BloodPressu...	Continuous	Never	Fixed	95.443	733	35
DiabetesPed...	Continuous	Never	Fixed	96.224	739	29
BMI	Continuous	Never	Fixed	98.568	757	11
Glucose	Continuous	Never	Fixed	99.349	763	5
Pregnancies	Continuous	Never	Fixed	100	768	0
Age	Continuous	Never	Fixed	100	768	0
Outcome	Flag	Never	Fixed	100	768	0

شکل ۱۰: داده های مفقوده

- ویژگی Insulin به دلیل داشتن بیش از ۵۰ درصد داده مفقوده از مدل حذف می‌شود.
- برای ویژگی SkinThickness به دلیل داشتن تعداد زیاد داده مفقوده، نمی‌توان از روش Fixed استفاده کرد زیرا نوع توزیع را به هم می‌زند. از طرفی چون توزیع نرمال ندارد نمی‌توان آن را با اعداد رندوم از توزیع نرمال پر کرد. از طرفی به دلیل عدم وجود فرد خبره نمیتوان از قوانین خبرگی در Expression استفاده کرد. پس تنها گزینه موجود روش Algorithm است.
- داده های مفقود ویژگی های BloodPressure و BMI به دلیل نرمال بودن، توسط روش Random و با توزیع نرمال پر می‌شوند.

- داده های مفقوده ویژگی DiabetesPedigreeFunction به دلیل کم بودن و همچنین اهمیت سابقه خانوادگی در ابتلا به دیابت، از طریق Algorithm پر می‌شود.
- داده های مفقود ویژگی Glucose به دلیل کم بودن با روش Fixed و با میانگین پر می‌شوند.

Field	Measurement	Impute Missing	Method	% Complete	Valid Records	Null Value
SkinThickness	Continuous	Blank & Null Val...	Algorithm	70.443	541	227
BloodPressu...	Continuous	Blank & Null Val...	Random	95.443	733	35
DiabetesPed...	Continuous	Blank & Null Val...	Algorithm	96.224	739	29
BMI	Continuous	Blank & Null Val...	Random	98.568	757	11
Glucose	Continuous	Blank & Null Val...	Fixed	99.349	763	5
Pregnancies	Continuous	Never	Fixed	100	768	0
Age	Continuous	Never	Fixed	100	768	0
Outcome	Flag	Never	Fixed	100	768	0

شکل ۱۱: مدیریت داده های مفقوده

در نهایت داده های مفقوده تمام ویژگی ها پر شده است.

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	100	768	0	0	0	0
BloodPressu...	Continuous	14	0	None	Never	Fixed	100	768	0	0	0	0
SkinThickness	Continuous	4	0	None	Never	Fixed	100	768	0	0	0	0
BMI	Continuous	8	0	None	Never	Fixed	100	768	0	0	0	0
DiabetesPed...	Continuous	15	0	None	Never	Fixed	100	768	0	0	0	0
Age	Continuous	0	0	None	Never	Fixed	100	768	0	0	0	0
Outcome	Flag	--	--	--	Never	Fixed	100	768	0	0	0	0

شکل ۱۲: مدیریت داده های مفقوده

۴-۳. نرمال سازی:

به منظور نرمال سازی تمامی داده ها را با روش Min/Max Transformation به مقیاس ۰ تا ۱۰۰ تبدیل شده است.

	Outcome	Pregnancies_transformed	Glucose_transformed	BloodPressure_transformed	SkinThickness_transformed	BMI_transformed	DiabetesPedigreeFunction_transformed	Age_transformed
1	1	44.444	67.097	49.455	56.000	43.959	49.326	63.736
2	0	7.407	26.452	41.378	44.000	23.978	24.528	21.978
3	1	59.259	89.677	38.686	16.435	14.558	53.369	24.176
4	0	7.407	29.032	41.378	32.000	28.260	7.996	0.000
5	1	0.000	60.000	6.382	56.000	71.077	42.663	26.374
6	0	37.037	46.452	52.147	28.426	21.123	11.051	19.780
7	1	22.222	21.935	19.842	50.000	36.538	15.274	10.989
8	0	74.074	45.806	59.708	51.266	48.812	5.031	17.582
9	1	14.815	98.710	46.763	76.000	35.110	7.188	70.330
10	1	59.259	52.258	81.759	63.026	75.098	13.836	72.527

شکل ۱۳: نرمال سازی داده ها

۵-۳. چالش هم‌خطی:

نکته قابل توجه این است که بعد از پاکسازی ویژگی SkinThickness و BMI بخاطر همبستگی بالاتر از ۰.۷ دچار چالش هم‌خطی شده اند. از این رو از ویژگی SkinThickness بخاطر داده های مفقود زیاد (۲۲۷ رکورد) در مدلسازی صرف نظر شده است.

	Age	BMI	BloodPressure	DiabetesPedig...	Glucose	Pregnancies	SkinThickness
Age	1.000	0.034	0.320	0.050	0.269	0.550	0.089
BMI	0.034	1.000	0.289	0.143	0.235	0.021	0.736
BloodPressure	0.320	0.289	1.000	0.025	0.217	0.200	0.232
DiabetesPedigr...	0.050	0.143	0.025	1.000	0.092	-0.003	0.102
Glucose	0.269	0.235	0.217	0.092	1.000	0.127	0.213
Pregnancies	0.550	0.021	0.200	-0.003	0.127	1.000	0.087
SkinThickness	0.089	0.736	0.232	0.102	0.213	0.087	1.000

شکل ۱۴: چالش هم‌خطی

در نهایت داده ها به دو بخش Train و Test با نسبت ۸۰٪ و ۲۰ ٪ تقسیم شدند.

۴. انتخاب ویژگی (Feature Selection):

برای انتخاب ویژگی از روش F-score استفاده شده است. F-score یک تکنیک ساده است که تمایز دو مجموعه از اعداد حقیقی را اندازه گیری می کند. با توجه به داده x_k که k=1,...,m و m تعداد داده های با کلاس مثبت یا منفی است، اگر تعداد نمونه های مثبت و منفی به ترتیب +n و -n باشد، F-score برای ویژگی i ام به صورت زیر تعریف می‌شود:

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2}$$

در اینجا $\overline{x}_i$، $\overline{x}_i^{(+)}$ و $\overline{x}_i^{(-)}$ به ترتیب میانگین ویژگی i ام در کل داده ها، داده های کلاس مثبت و داده های کلاس منفی هستند. همینطور $\overline{x}_{k,i}^{(+)}$ ویژگی i ام از داده k ام در داده های با کلاس مثبت و $\overline{x}_{k,i}^{(-)}$ ویژگی i ام از داده k ام در داده های با کلاس منفی هستند.

صورت نشان دهنده تمایز بین مجموعه های مثبت و منفی و مخرج نشان دهنده یکی از هر یک از دو مجموعه است. هرچه F-score بزرگتر باشد، این ویژگی بیشتر متمایز است. بنابراین از این امتیاز به عنوان معیار انتخاب ویژگی استفاده می‌کنیم.

برای این دیتاست مقدار F-score به صورت جدول زیر محاسبه شده است.

	Feature	F-Score
1	Glucose	0.218
2	BMI	0.104
3	Age	0.066
4	Pregnancies	0.057
5	DiabetesPedigreeFunction	0.028
6	BloodPressure	0.025

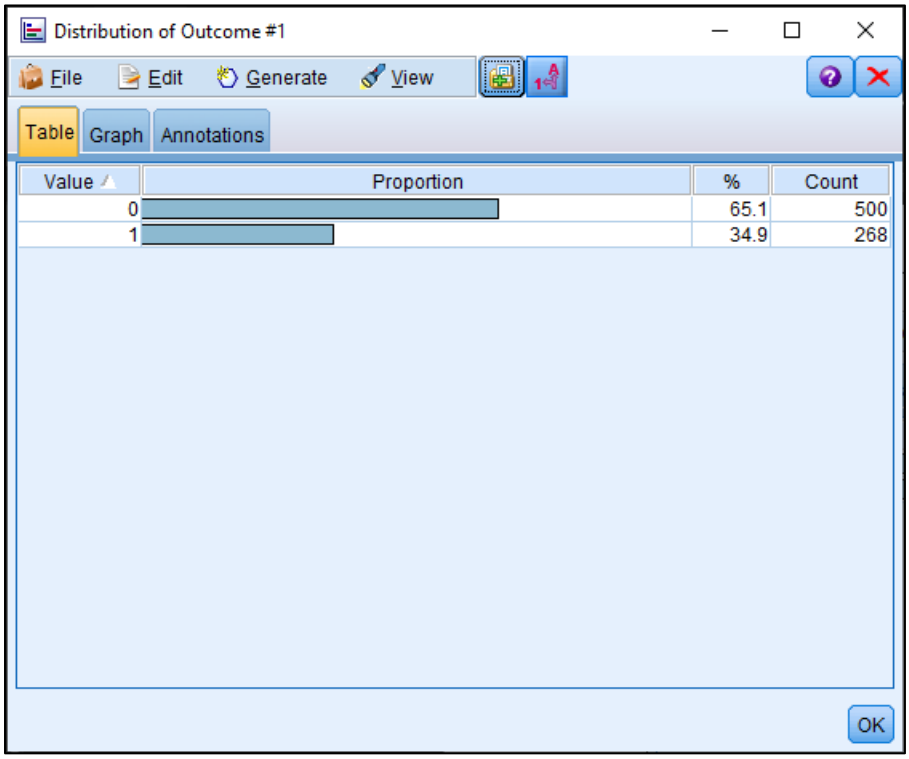
جدول ۳: مقدار F-Score برای ویژگی ها

از نصف میانگین F-score ها به عنوان حد آستانه برای انتخاب ویژگی استفاده شده که در نهایت ۴ ویژگی Glucose، BMI، Age و Pregnancies برای مدلسازی انتخاب شده‌اند.

اطلاعات بالا در فایل f-score-feature-selection.xslx قابل مشاهده است.

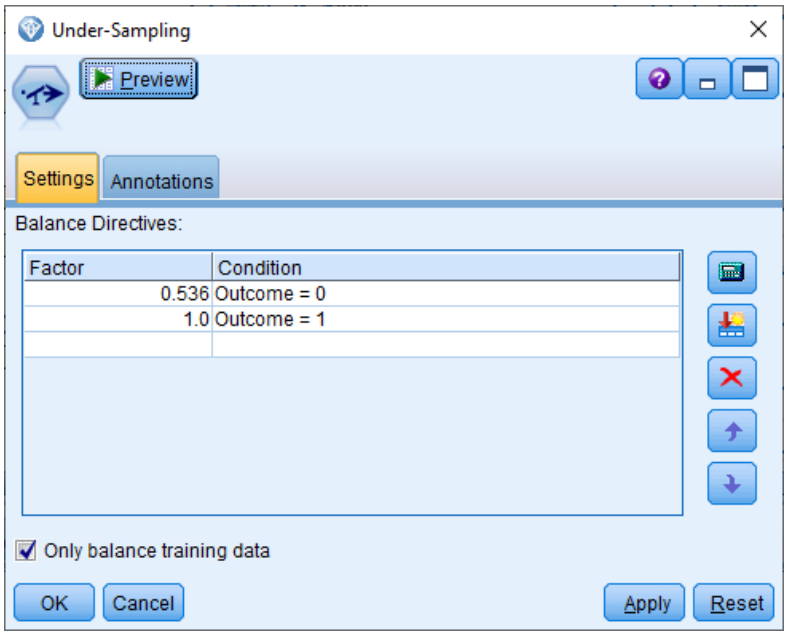
۵. مدیریت دادگان نامتوازن:

برای بررسی توازن دادگان از نمودار توزیع ویژگی Outcome استفاده شده است.

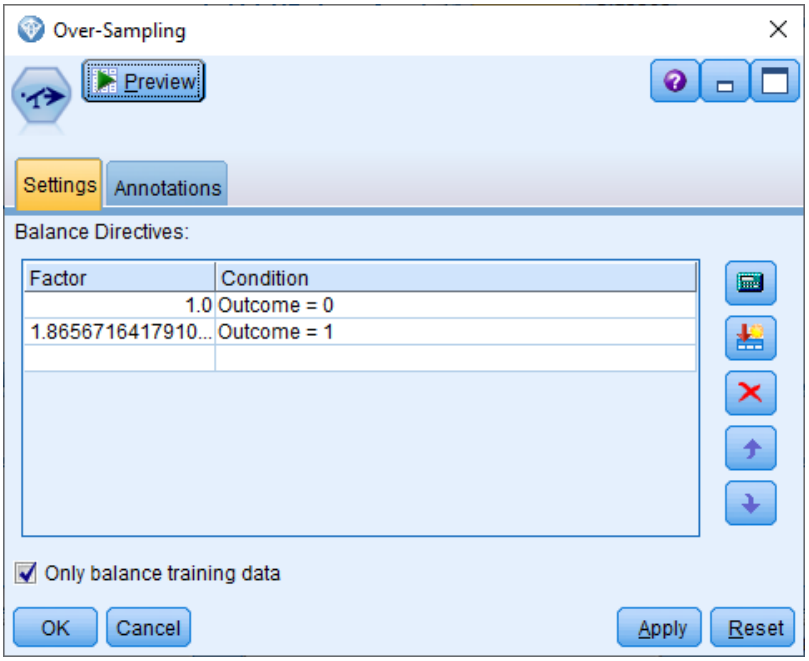


شکل ۱۵: عدم توازن داده ها

همانطور که مشخص است ۶۵.۱ درصد از داده ها به کلاس 0 (فرد سالم) و ۳۴.۹ درصد از داده ها به کلاس 1 (فرد مبتلا به دیابت) تعلق دارد. به دلیل نامتوازن بودن داده ها از تکنیک های Under-Sampling و Over-Sampling برای متوازن کردن داده ها استفاده شده است.



شکل ۱۶: مدیریت دادگان نامتوازن به روش Under-Sampling



شکل ۱۷: مدیریت دادگان نامتوازن به روش Over-Sampling

لازم به ذکر است که مدلسازی هم با داده نامتوازن و هم متوازن انجام شده و نتایج آن در بخش مربوطه آمده است.

۶. مدلسازی:

در مدلسازی از الگوریتم های KNN، SVM، Neural Network، C&RT، QUEST، CHAID، C5.0 و Random Forest استفاده شده است.

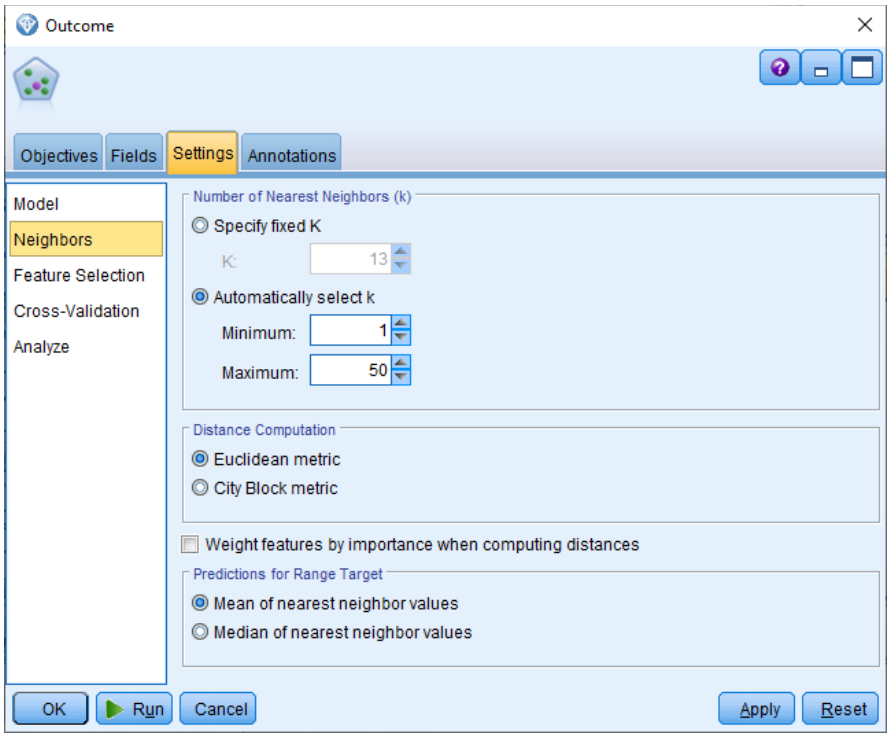
مدلسازی در سه بخش شامل موارد زیر انجام شده است:

۱. مدلسازی بدون Feature Selection: در این بخش از تمام ویژگی های موجود پس از پاکسازی داده ها استفاده شده است.
۲. مدلسازی با Feature Selection: در این بخش از روش f-score برای انتخاب بهترین ویژگی ها استفاده شده است.
۳. مدلسازی بدون پاکسازی داده ها: در این بخش که مختص به الگوریتم های درختی است از تمام ویژگی های موجود بدون پاکسازی داده ها استفاده شده است.

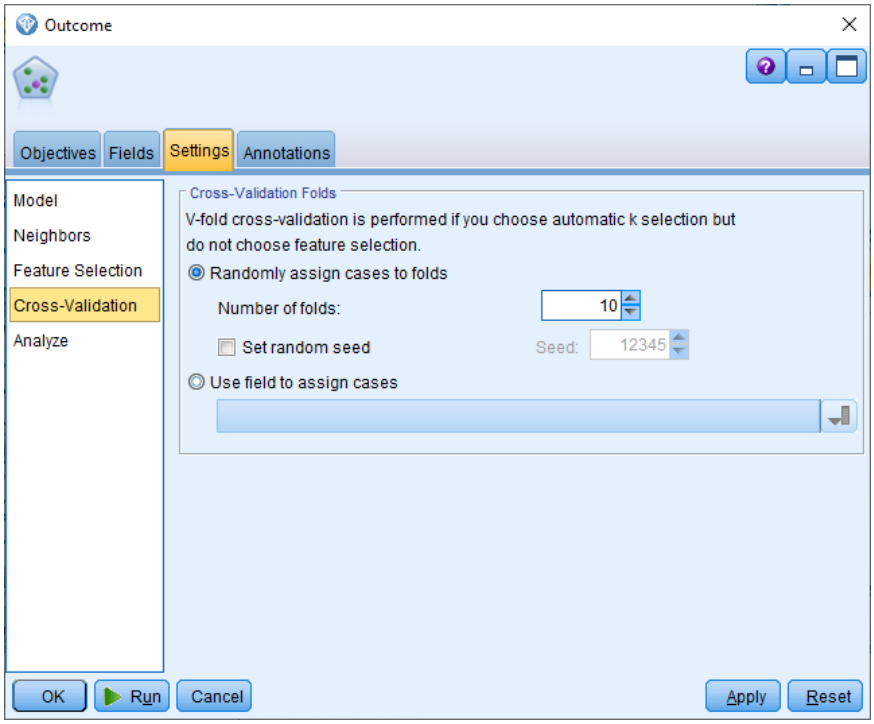
در ادامه، در هر سه بخش بالا، مدلسازی به سه حالت داده نامتوازن، Under-Sampling و Over-Sampling انجام شده است.

۶-۱. الگوریتم KNN:

در این الگوریتم، از K=[1..50] و تعداد ۱۰ فولد برای Cross-Validation استفاده شده است. لازم به ذکر است که این الگوریتم با فولد های بیشتر نیز امتحان شد که نتایج بدست آمده بهبودی نداشت.



شکل ۱۸: تنظیمات الگوریتم KNN



شکل ۱۹: تنظیمات الگوریتم KNN

نتایج مدل به صورت زیر است. باید توجه داشت که K های بهینه با تغییر داده های train ممکن است تغییر کند.

الگوریتمKNN	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Acc	81.26%	78.66%	75.24%	75.45%	93.93%	94.61%
K	7		34		1	

جدول4: دقت الگوریتم KNN بدون Feature Selection

مشاهده می‌شود که روش Over-Sampling دقت بیشتری دارد.

با استفاده از Feature Selection نتیجه به صورت زیر است:

الگوریتم KNN	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Acc	79.58%	79.47%	77.16%	74.17%	94.36%	91.39%
K	9		11		1	

جدول5: دقت الگوریتم KNN با Feature Selection

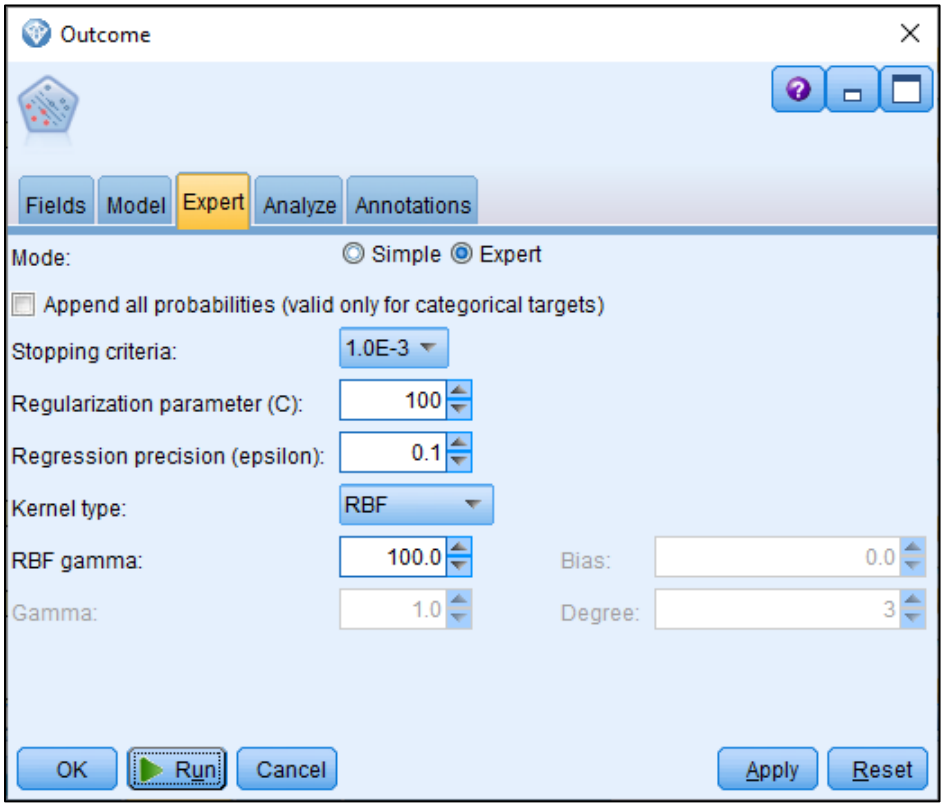
باز هم مشاهده می‌شود که روش Over-Sampling دقت بیشتری دارد.

۲-۶. الگوریتم SVM:

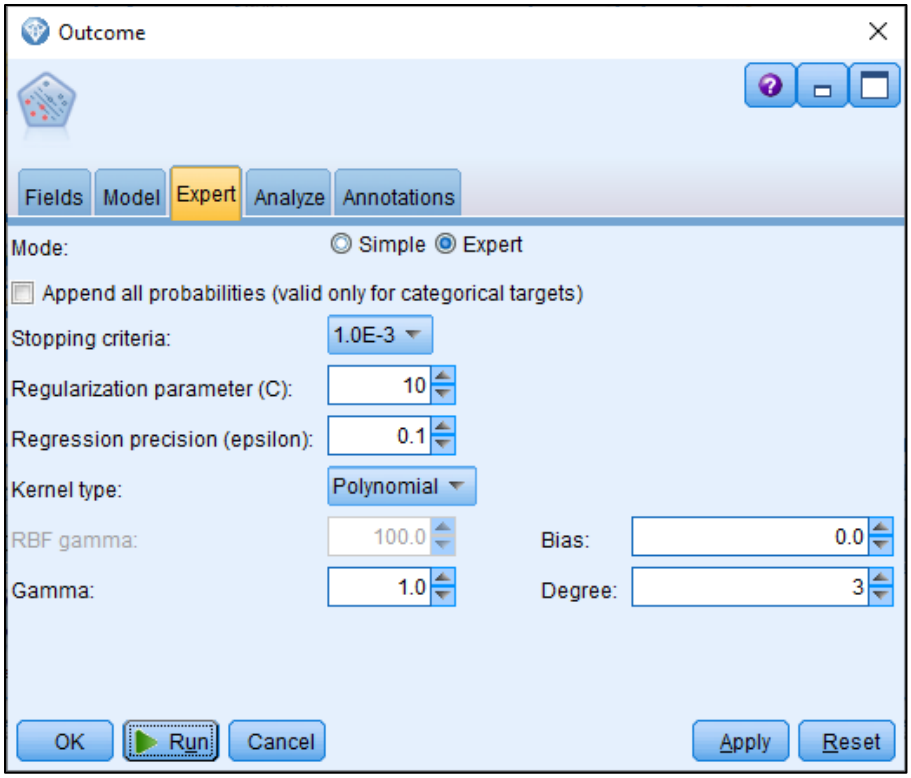
در این روش از C=[1, 10, 100, 1000, 10000] و Gamma=[0.1, 1, 10, 100, 1000, 10000] برای تمام کرنل ها تست شد که در هر کرنل مقادیر بهینه به صورت زیر بدست آمد:

Kernel Type	C	Gamma
RBF	100	100
Polynomial	10	1
Sigmoid	1000	10
Linear	100	-

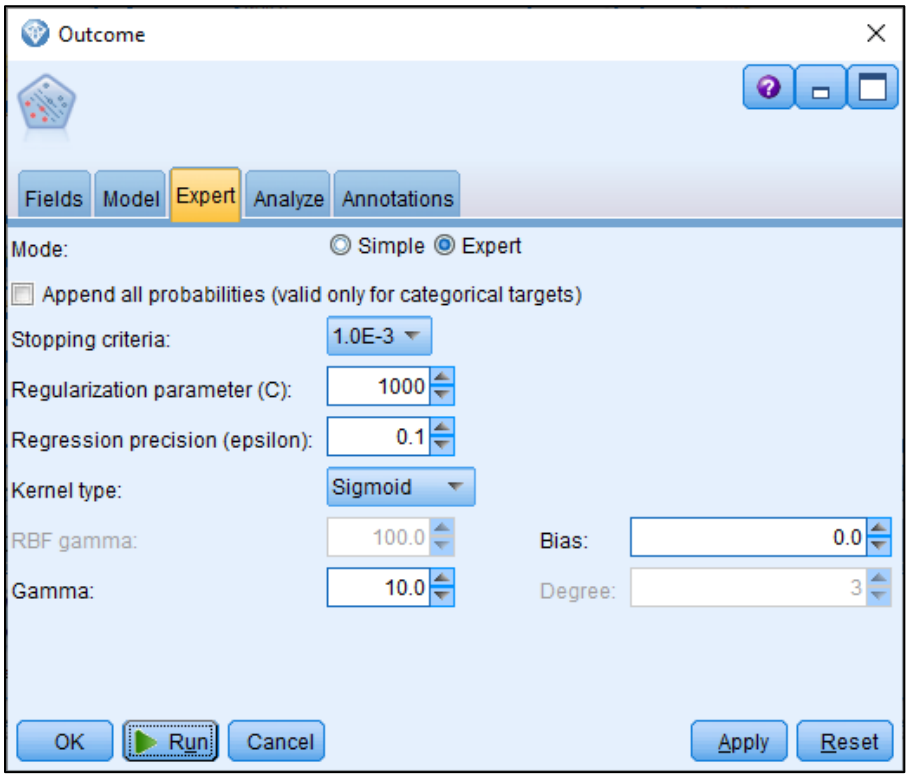
جدول ۶: پارامتر های بهینه برای الگوریتم SVM



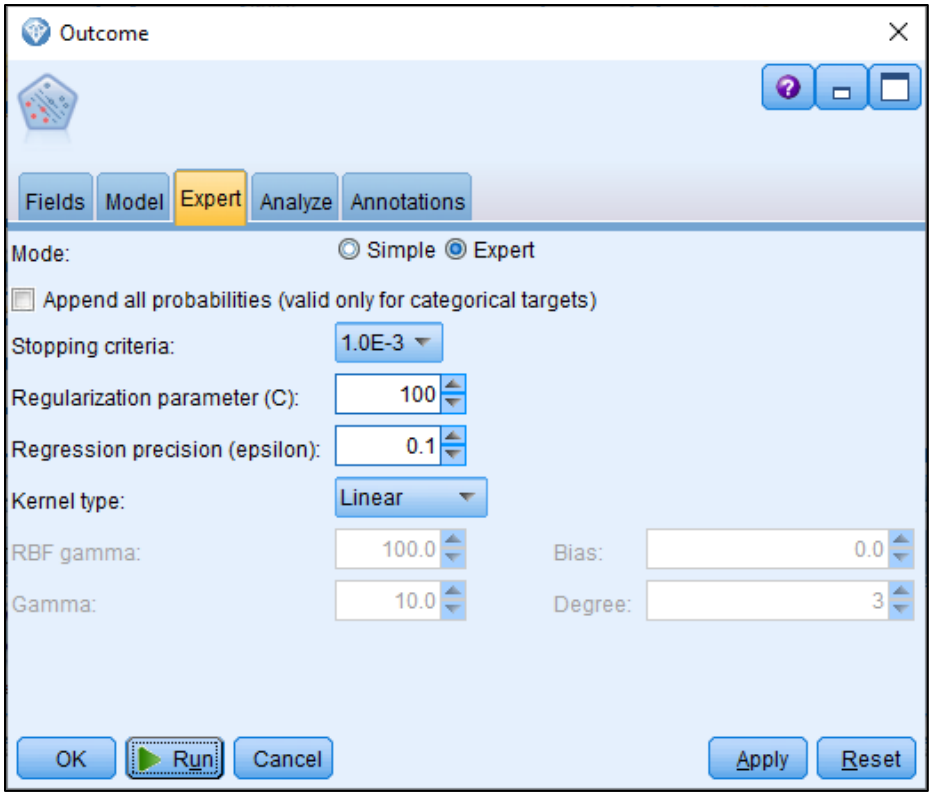
شکل ۲۰: تنظیمات الگوریتم SVM



شکل ۲۱: تنظیمات الگوریتم SVM



شکل ۲۲: تنظیمات الگوریتم SVM



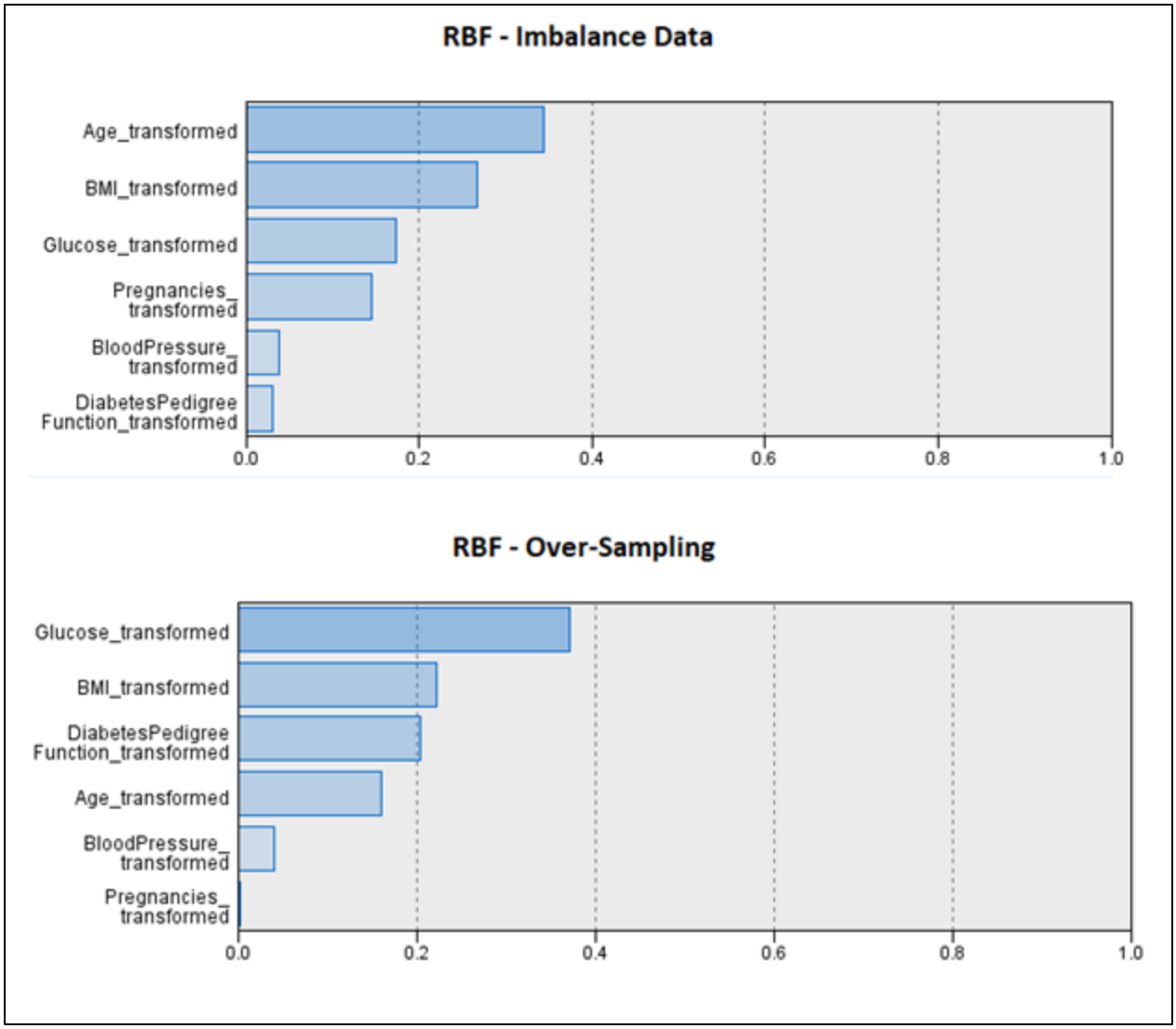
شکل ۲۳: تنظیمات الگوریتم SVM

خروجی مدل ها به صورت زیر است:

الگوریتم SVM	No Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
RBF	91.85%	91.02%	86.47%	86.96%	95.43%	94.12%
Polynomial	79.20%	79.04%	77.68%	75.32%	76.10%	78.48%
Sigmoid	76.18%	72.26%	70.33%	65.52%	69.43%	67.09%
Linear	77.76%	75.16%	76.85%	74.51%	75.47%	77.85%

جدول 7: دقت الگوریتم SVM بدون Feature Selection

مشاهده می‌شود که روش RBF دقت بیشتری به نسبت سایر روش ها دارد. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



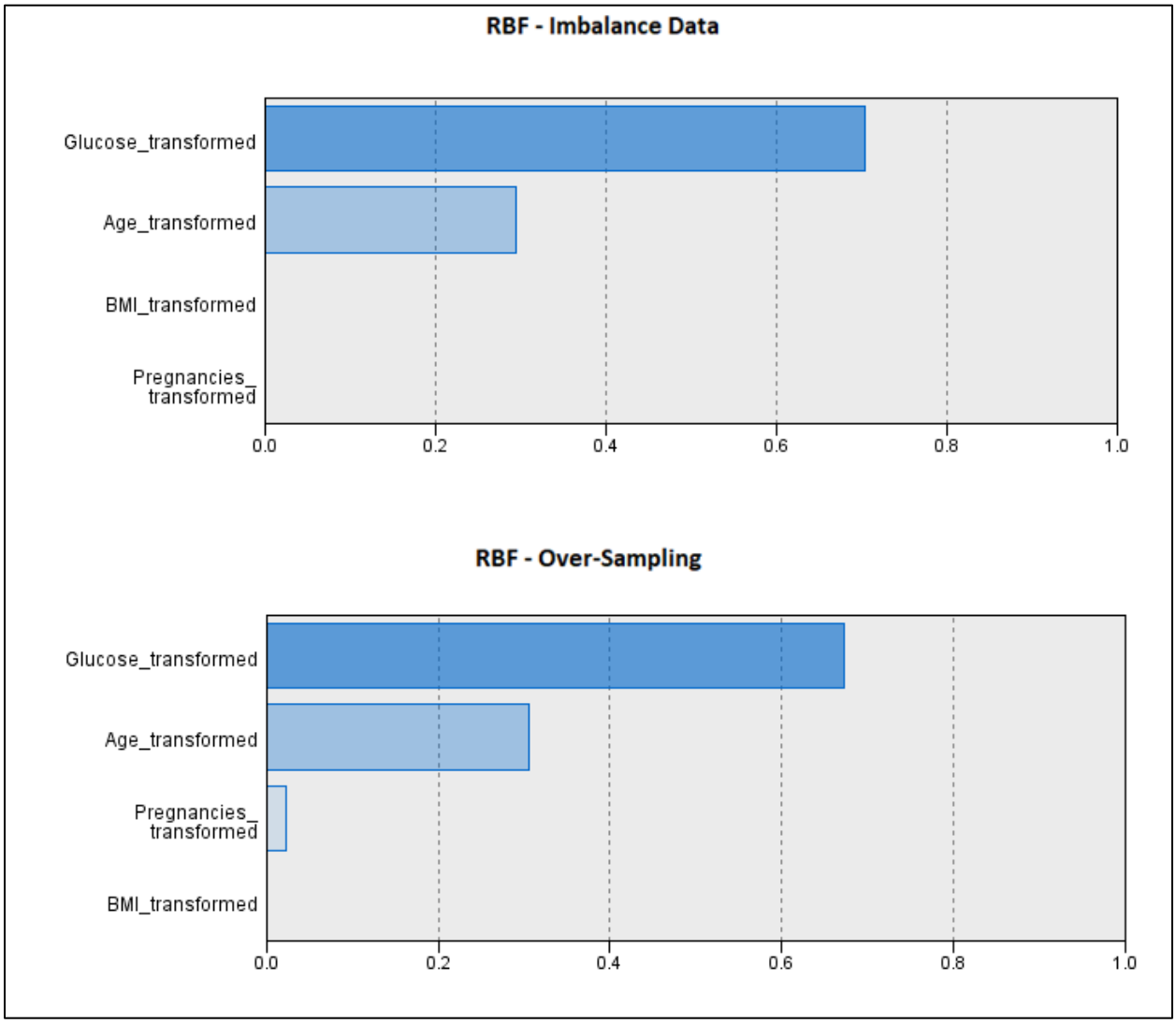
شکل ۲۴: اهمیت ویژگی های الگوریتم SVM بدون Feature Selection

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم SVM	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
RBF	95.29%	93.64%	85.78%	82.66%	93.67%	91.25%
Polynomial	78.13%	76.72%	74.37%	72.99%	74.91%	72.99%
Sigmoid	74.51%	73.72%	75.70%	73.72%	75.38%	73.72%
Linear	78.10%	74.36%	73.83%	72.44%	74.37%	73.08%

جدول 8: دقت الگوریتم SVM با Feature Selection

باز هم مشاهده می‌شود که روش RBF دقت بیشتری به نسبت سایر روش ها دارد. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



شکل ۲۵: اهمیت ویژگی های الگوریتم SVM با Feature Selection

۳-۶. الگوریتم شبکه عصبی (Neural Network):

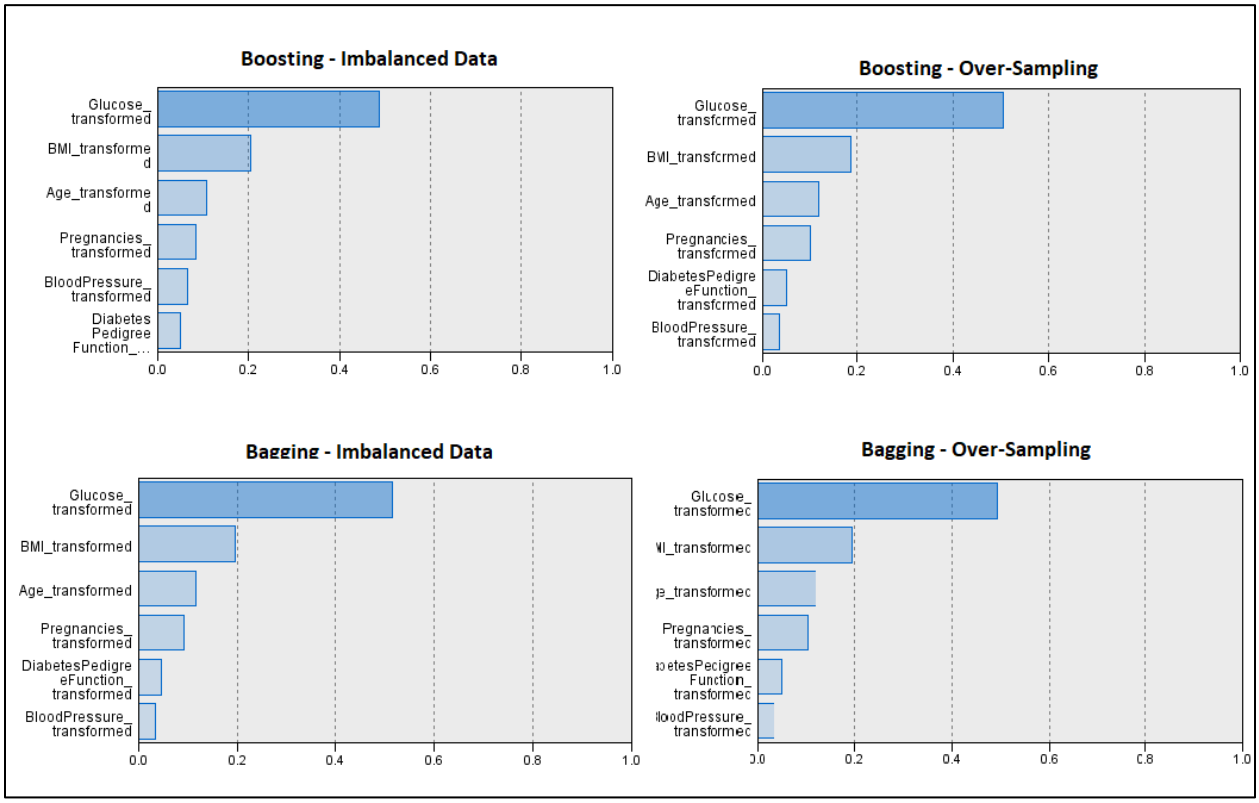
در این روش دو روش MLP و RBF با تعداد نرون های پنهان (5,10,15,20,25,30) با تعداد مدل (10,20,30,40) برای Boosting و Bagging آزمایش شد. در نهایت مدل MLP با دو لایه پنهان که هرکدام ۲۰ نرون دارند و برای Boosting و Bagging استفاده از ۱۰ مدل، نتایج بهتری را نشان داد.

خروجی مدل ها به صورت زیر است:

الگوریتم شبکه عصبی	No Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	78.45%	78.12%	74.77%	71.52%	78.29%	75.15%
Boosting	92.32%	91.67%	86.54%	83.23%	90.39%	92.66%
Bagging	92.58%	90.54%	86.60%	86.36%	90.10%	91.93%

جدول 9: دقت الگوریتم شبکه عصبی بدون Feature Selection

مشاهده می‌شود که روش های Boosting و Bagging در داده های نامتوازن و Over-Sampling دقت بیشتری دارند. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



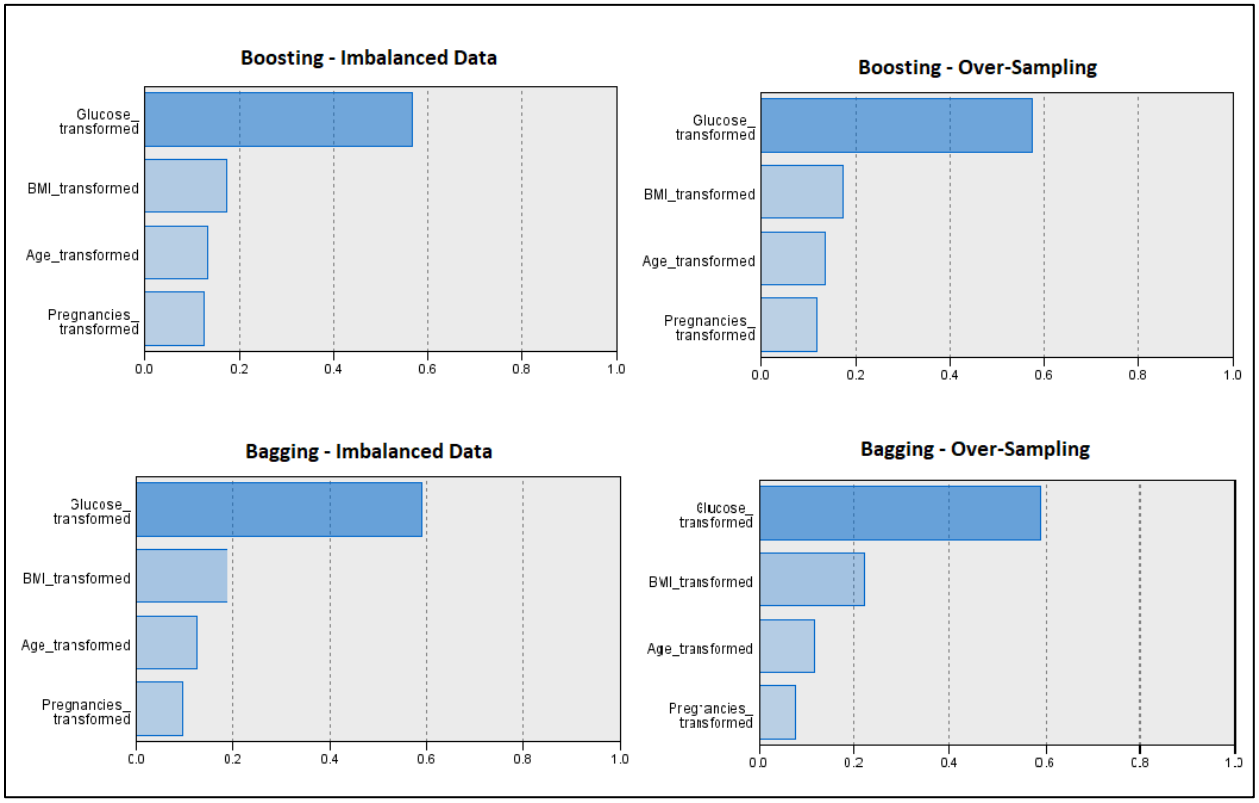
شکل 26: اهمیت ویژگی های الگوریتم شبکه عصبی بدون Feature Selection

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم شبکه عصبی	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	78.43%	74.36%	74.53%	73.72%	80.18%	76.28%
Boosting	94.61%	94.23%	78.97%	78.21%	93.69%	94.87%
Bagging	92.32%	90.38%	81.92%	78.21%	93.22%	90.38%

جدول10 : دقت الگوریتم شبکه عصبی با Feature Selection

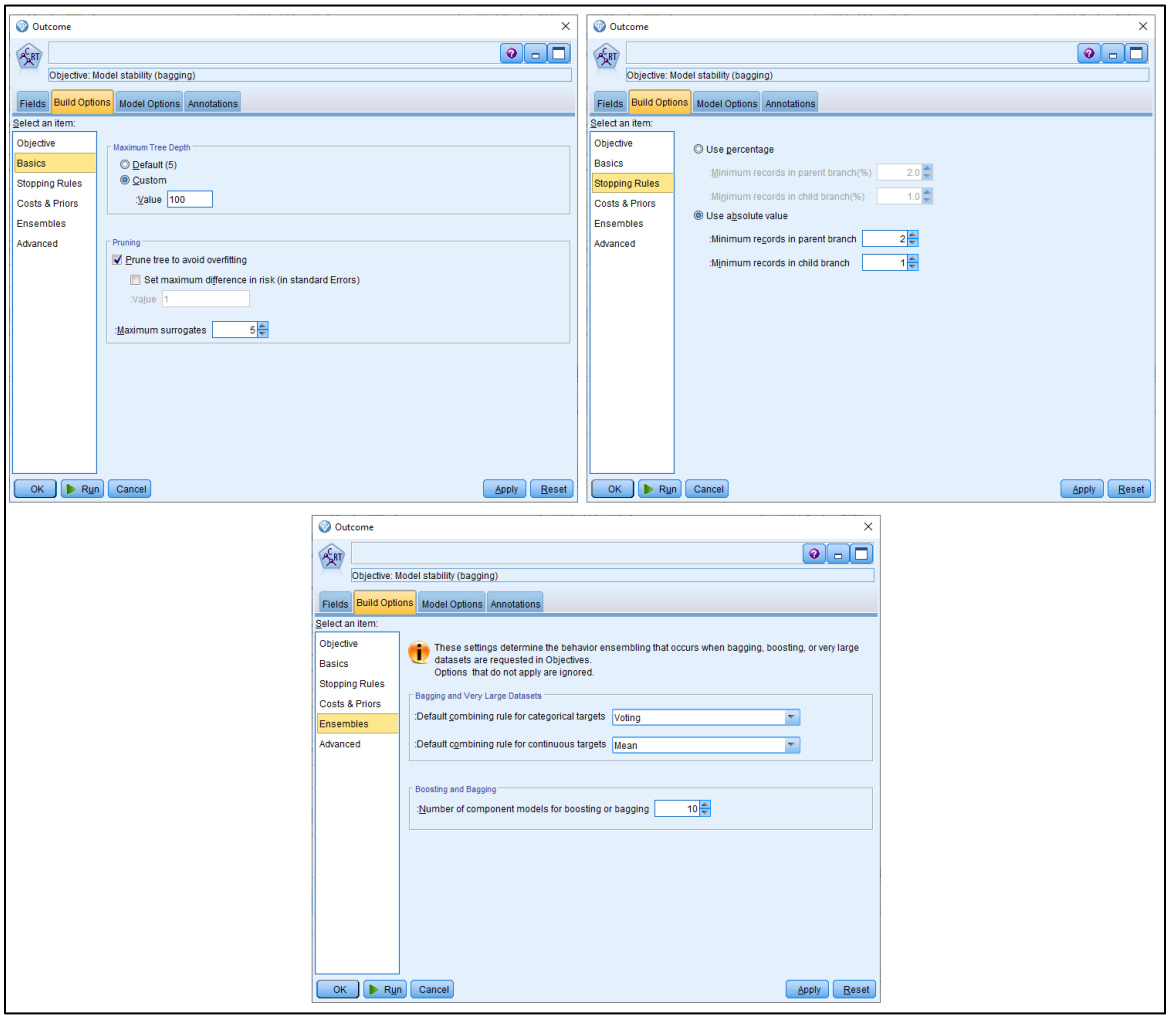
باز هم مشاهده می‌شود که روش های Boosting و Bagging در داده های نامتوازن و Over-Sampling دقت بیشتری دارند. تنها تفاوت این است که دقت به نسبت مدلسازی بدون Feature Selection افزایش داشته است. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



شکل ۲۷: اهمیت ویژگی های الگوریتم شبکه عصبی با Feature Selection

۴-۶. الگوریتم C&RT:

یکی از انواع الگوریتم بر پایه درخت تصمیم است. پس از آزمایش مقادیر مختلف، حداکثر عمق ۱۰۰ نود، شرط توقف ۲ رکورد در والد و ۱ رکورد در فرزند و استفاده از ۱۰ مدل برای Boosting و Bagging ، نتایج بهتری را نشان داد.



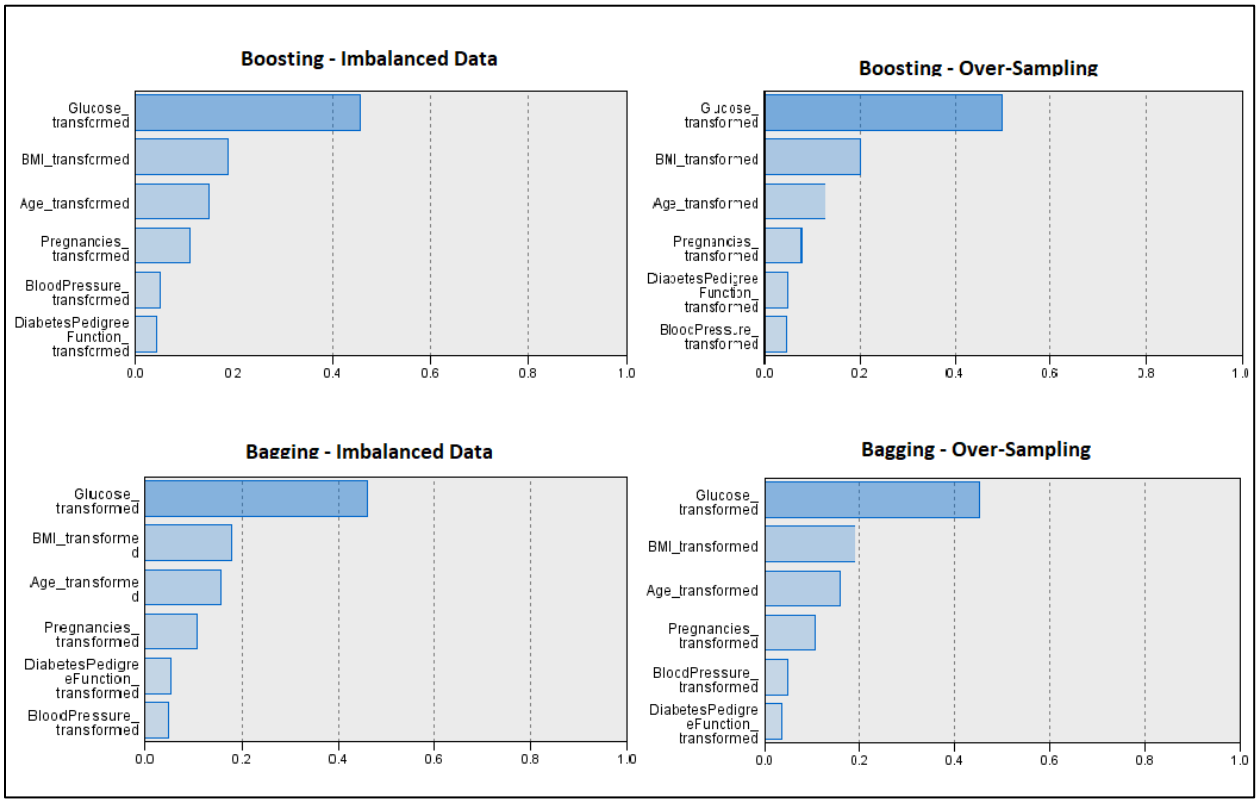
شکل ۲۸: تنظیمات الگوریتم C&RT

خروجی مدل ها به صورت زیر است:

الگوریتم C&RT	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	81.17%	82.89%	81.51%	79.26%	85.93%	84.57%
Boosting	92.77%	93.05%	87.59%	86.21%	94.35%	93.96%
Bagging	92.88%	92.68%	84.82%	84.09%	93.06%	92.26%

جدول 11 : دقت الگوریتم C&RT بدون Feature Selection

مشاهده می‌شود که روش های Boosting و Bagging در داده های نامتوازن و Over-Sampling دقت بیشتری دارند. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



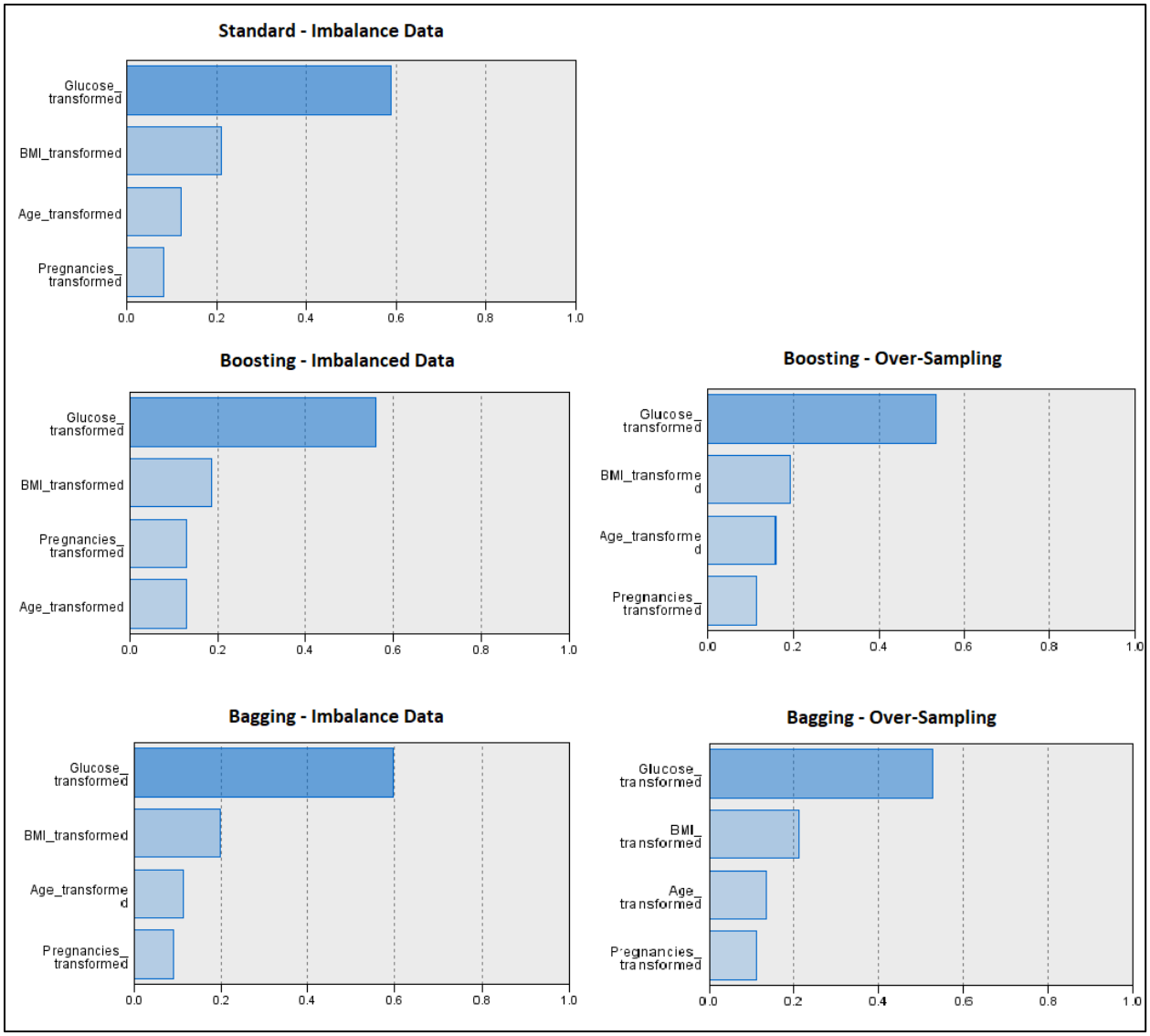
شکل 29: اهمیت ویژگی های الگوریتم C&RT بدون Feature Selection

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم C&RT	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	92.95%	92.65%	81.41%	80.28%	85.68%	84.62%
Boosting	94.27%	93.59%	87.32%	86.54%	93.27%	94.23%
Bagging	94.77%	94.87%	89.44%	85.26%	92.71%	91.03%

جدول 12 : دقت الگوریتم C&RT با Feature Selection

مشاهده می‌شود که تمام روش ها در داده های نامتوازن و Boosting و Bagging در Over-Sampling دقت بیشتری دارند. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



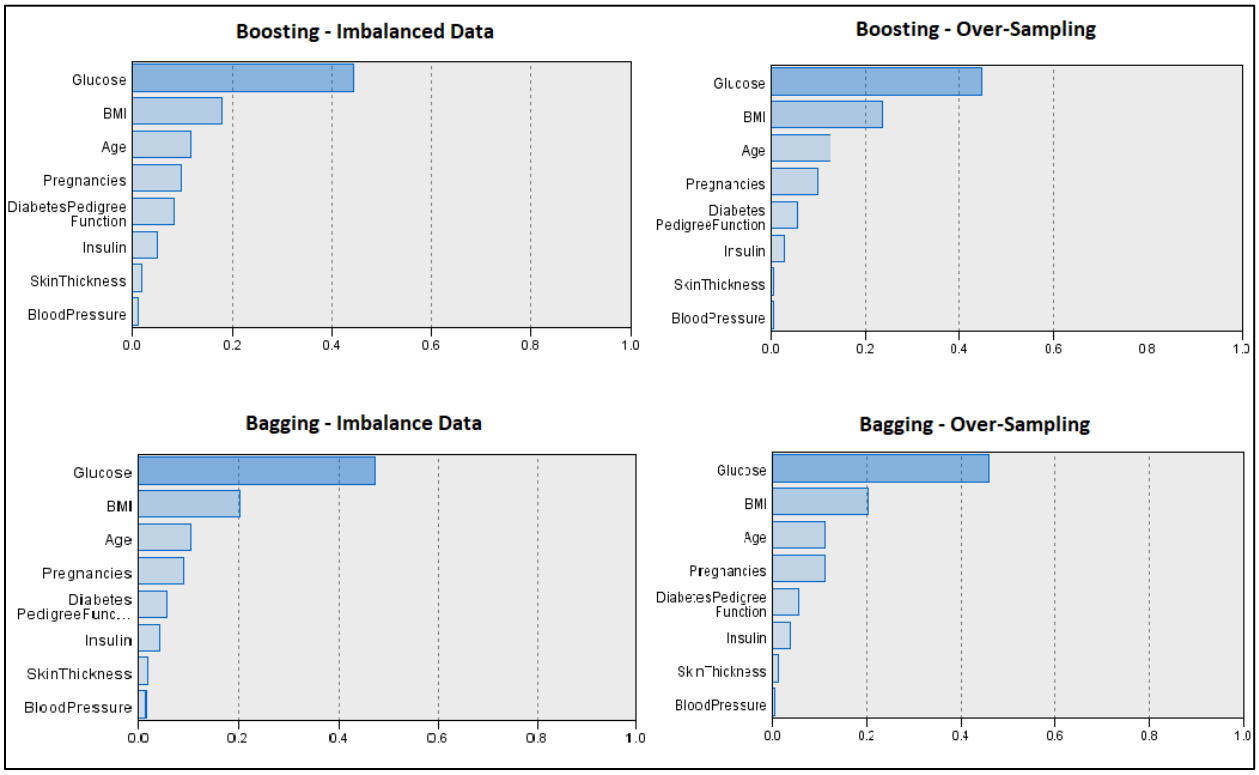
شکل ۳۰: اهمیت ویژگی های الگوریتم C&RT با Feature Selection

برای مدلسازی بدون پاکسازی داده ها نیز نتایج به صورت زیر است:

الگوریتم C&RT	Without Data Cleaning					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	79.32%	77.92%	89.19%	79.19%	86.03%	83.67%
Boosting	96.06%	95.60%	87.32%	83.14%	92.85%	91.84%
Bagging	95.44%	94.16%	88.18%	85.03%	93.35%	90.97%

جدول 13 : دقت های الگوریتم C&RT بدون پاکسازی داده ها

مشاهده می‌شود که روش های Boosting و Bagging در داده های نامتوازن و Over-Sampling دقت بیشتری دارند. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



شکل ۳۱: اهمیت ویژگی های الگوریتم C&RT بدون پاکسازی داده ها

۵-۶. الگوریتم QUEST:

یکی از انواع الگوریتم بر پایه درخت تصمیم است. پس از آزمایش مقادیر مختلف، حداکثر عمق ۱۰۰ نود، شرط توقف ۲ رکورد در والد و ۱ رکورد در فرزند و استفاده از ۱۰ مدل برای Boosting و Bagging ، نتایج بهتری را نشان داد. ولی در نهایت مدلی با دقت بالاتر از ۹۰٪ بدست نیامد.

خروجی مدل ها به صورت زیر است:

الگوریتم QUEST	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	76.43%	72.61%	70.18%	73.12%	74.62%	72.50%
Boosting	85.16%	84.46%	81.38%	80.14%	90.81%	88.31%
Bagging	80.33%	79.11%	77.29%	77.22%	82.57%	79.10%

جدول 14: دقت الگوریتم QUEST بدون Feature Selection

مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling دقت بیشتری به نسبت بقیه دارند.

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم QUEST	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	75.37%	72.96%	74.42%	73.87%	76.07%	73.58%
Boosting	83.63%	80.25%	79.18%	78.98%	85.96%	84.08%
Bagging	77.67%	77.99%	76.17%	75.47%	82.40%	81.76%

جدول 15 : دقت الگوریتم QUEST با Feature Selection

باز هم مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling دقت بیشتری به نسبت بقیه دارند.

برای مدلسازی بدون پاکسازی داده ها نیز نتایج به صورت زیر است:

الگوریتم QUEST	Without Data Cleaning					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	75.37%	72.96%	74.42%	73.87%	76.07%	73.58%
Boosting	83.63%	80.25%	79.18%	78.98%	85.96%	84.08%
Bagging	77.67%	77.99%	76.17%	75.47%	82.40%	81.76%

جدول :16 دقت الگوریتم QUEST بدون پاکسازی داده ها

باز هم مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling دقت بیشتری به نسبت بقیه دارند.

۶-۶. الگوریتم CHAID:

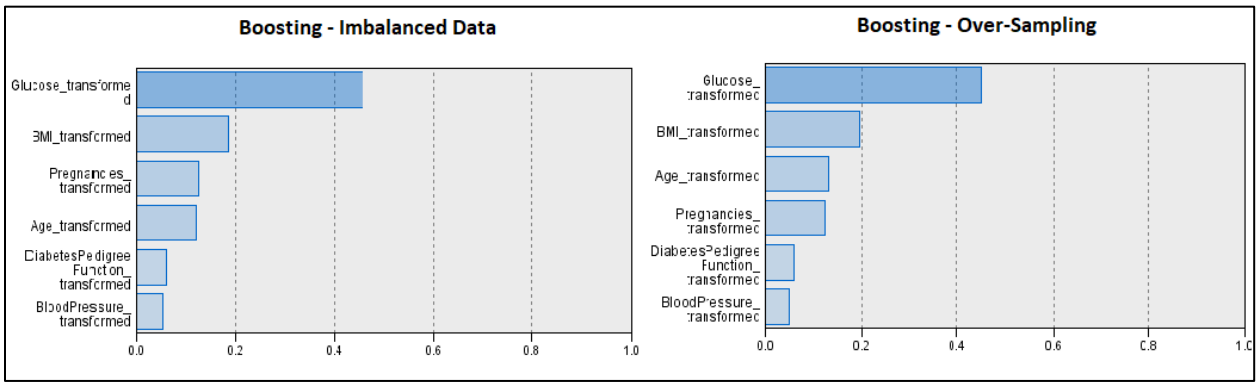
یکی از انواع الگوریتم بر پایه درخت تصمیم است. پس از آزمایش مقادیر مختلف، حداکثر عمق ۱۰۰ نود، شرط توقف ۲ رکورد در والد و ۱ رکورد در فرزند و استفاده از ۱۰ مدل برای Boosting و Bagging ، نتایج بهتری را نشان داد.

خروجی مدل ها به صورت زیر است:

الگوریتم CHAID	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	74.28%	78.08%	76.99%	73.97%	82.18%	78.77%
Boosting	93.58%	91.72%	84.86%	83.45%	94.29%	95.17%
Bagging	87.54%	84.94%	83.02%	83.73%	89.07%	87.95%

جدول 17: دقت الگوریتم CHAID بدون Feature Selection

مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling دقت بیشتری دارد. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



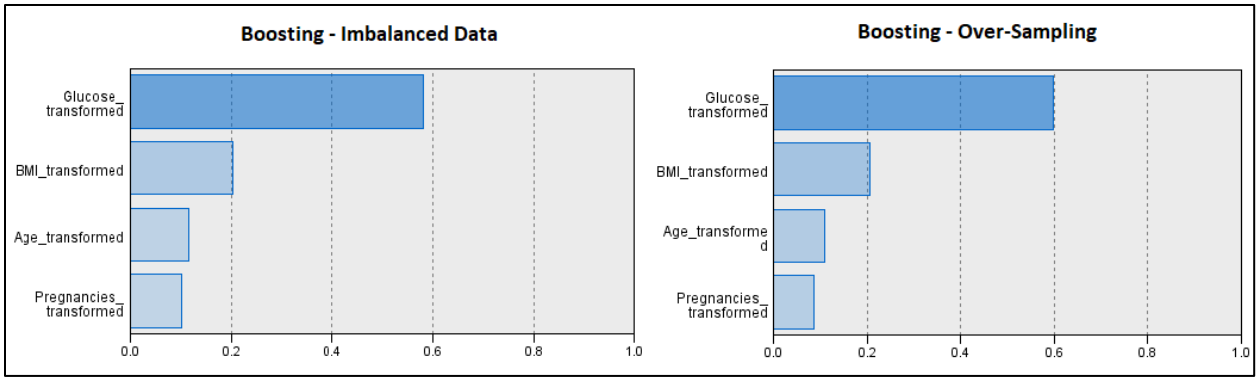
شکل ۳۲:اهمیت ویژگی های الگوریتم CHAID بدون Feature Selection

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم CHAID	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	79.15%	79.25%	77.57%	77.99%	78.95%	76.73%
Boosting	94.25%	90.57%	87.85%	86.79%	92.73%	91.19%
Bagging	83.02%	79.89%	80.34%	78.21%	85.08%	84.36%

جدول 18 : دقت الگوریتم CHAID با Feature Selection

باز هم مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling دقت بیشتری دارد. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



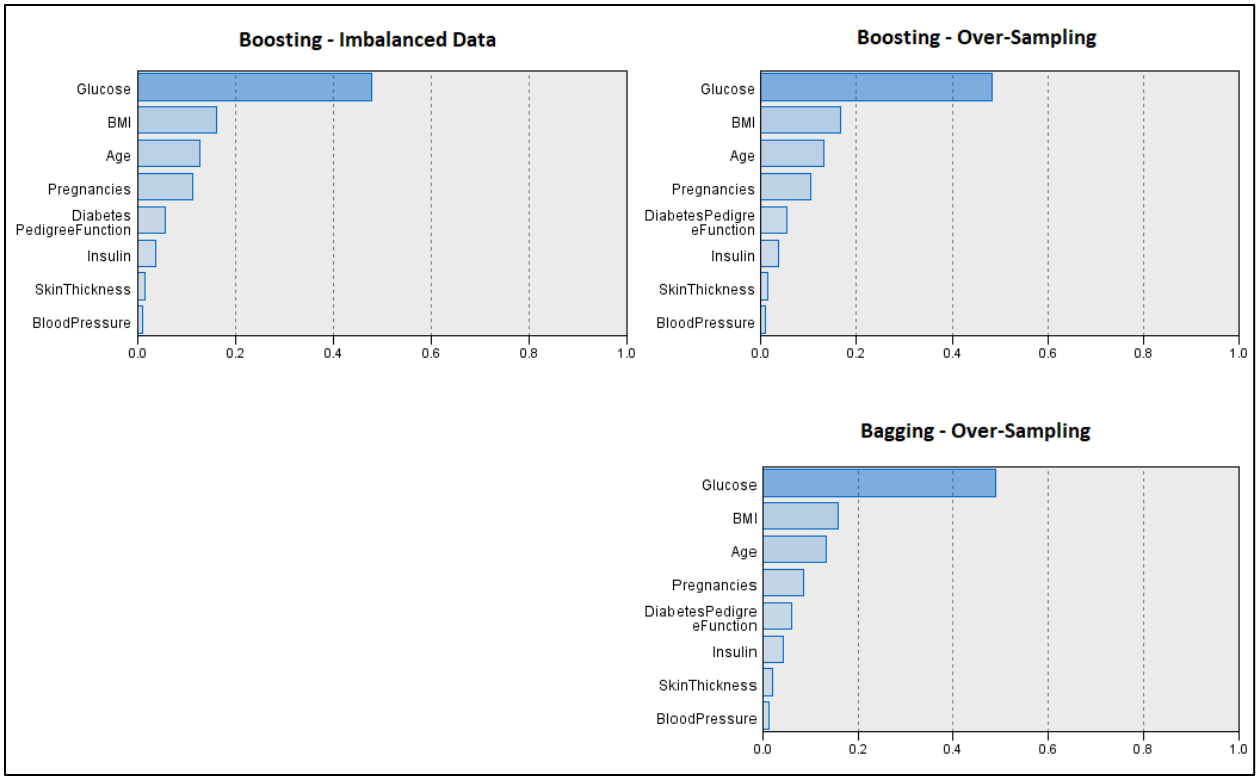
شکل ۳۳: اهمیت ویژگی های الگوریتم CHAID با Feature Selection

برای مدلسازی بدون پاکسازی داده ها نیز نتایج به صورت زیر است:

الگوریتم CHAID	Without Data Cleaning					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	80.00%	77.12%	76.80%	73.20%	81.16%	79.08%
Boosting	95.00%	93.92%	86.56%	87.16%	93.88%	94.59%
Bagging	86.44%	83.97%	85.75%	80.13%	94.14%	91.03%

جدول 19 : دقت الگوریتم CHAID بدون پاکسازی داده ها

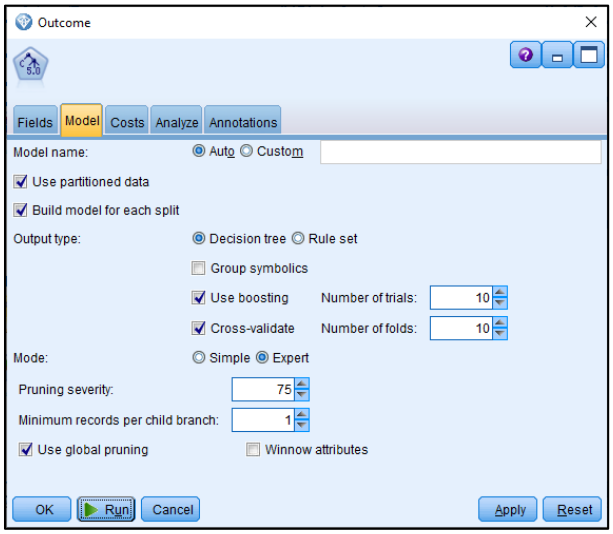
باز هم مشاهده می‌شود که روش Boosting در داده های نامتوازن و Over-Sampling و روش Bagging در Over-Sampling دقت بیشتری دارند. اهمیت ویژگی ها برای مدل های برتر به صورت زیر است:



شکل ۳۴: اهمیت ویژگی های الگوریتم CHAID بدون پاکسازی داده ها

۷-۶. الگوریتم C5.0:

یکی از انواع الگوریتم بر پایه درخت تصمیم است. پس از آزمایش مقادیر مختلف، شرط وجود حداقل ۱ رکورد در فرزند، استفاده از ۱۰ مدل برای Boosting و ۱۰ فولد برای Cross-Validation ، نتایج بهتری را نشان داد.



شکل ۳۵: تنظیمات الگوریتم C5.0

خروجی مدل ها به صورت زیر است:

الگوریتم C5.0	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	81.63%	77.22%	78.07%	75.56%	88.10%	86.67%

جدول20 : دقت الگوریتم C5.0 بدون Feature Selection

مشاهده می‌شود که هیچ مدلی دقت بالای ۹۰٪ را ندارد.

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم C5.0	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	80.14%	77.65%	79.30%	74.61%	80.24%	78.77%

جدول21 : دقت الگوریتم C5.0 با Feature Selection

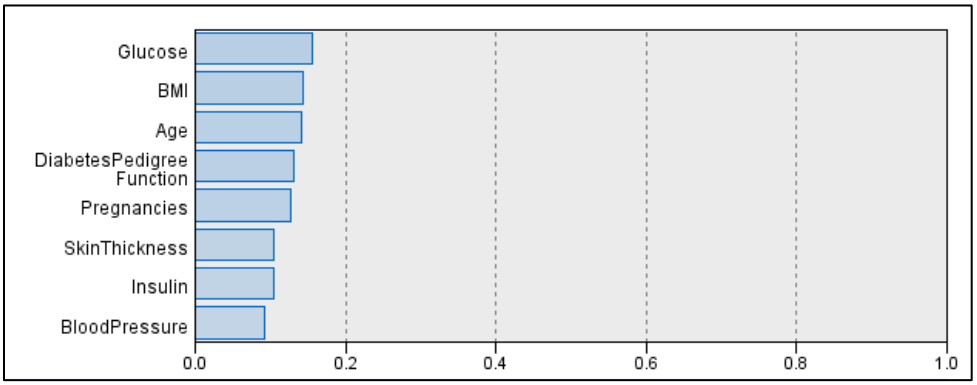
باز هم مشاهده می‌شود که هیچ مدلی دقت بالای ۹۰٪ را ندارد.

برای مدلسازی بدون پاکسازی داده ها نیز نتایج به صورت زیر است:

الگوریتم C5.0	Without Data Cleaning					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	85.92%	84.19%	80.79%	86.62%	94.77%	93.66%

جدول 22: دقت الگوریتم C5.0 بدون پاکسازی داده ها

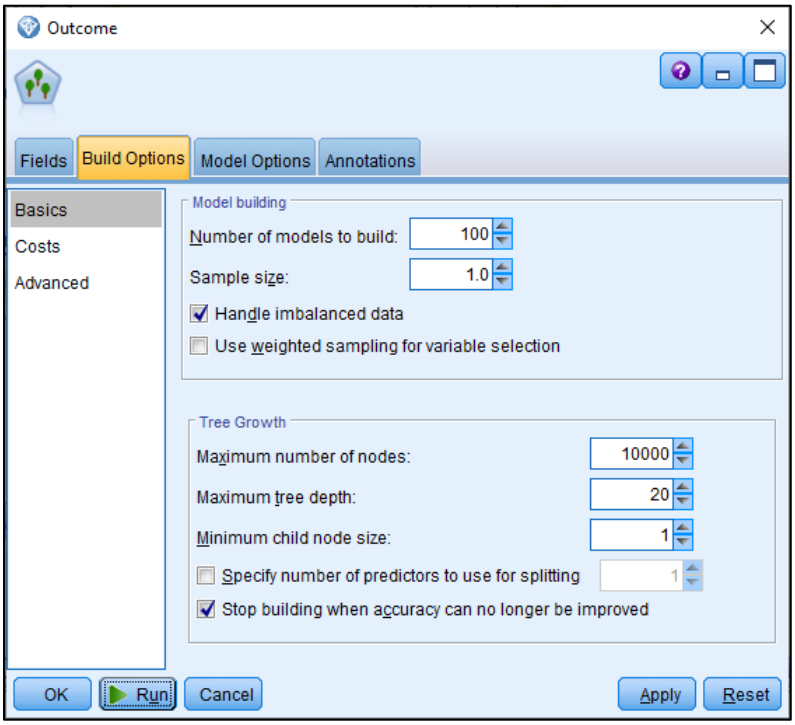
برای داده های پاکسازی نشده Over-Sampling دقت بیشتری دارد. اهمیت ویژگی ها برای مدل برتر به صورت زیر است:



شکل36 : اهمیت ویژگی های الگوریتم C5.0 بدون پاکسازی داده ها

۸-۶. الگوریتم جنگل تصادفی (Random Forest):

این الگوریتم با ساخت تعداد زیادی درخت تصمیم اقدام به پیش بینی هدف میکند. در اینجا با ساخت ۱۰۰ مدل با حداکثر عمق ۲۰ برای هر درخت و تعداد حداقل ۱ رکورد در گره فرزند بهترین نتیجه بدست آمد. لازم به ذکر است با فعال کردن گزینه Handle imbalance data، در تمام سناریوها دقت مدل افزایش یافت.



شکل ۳۷: تنظیمات الگوریتم Random forest

خروجی مدل ها به صورت زیر است:

الگوریتم Random Forest	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	90.35%	86.22%	87.90%	85.71%	94.33%	93.41%

جدول23 : دقت الگوریتم Random forest بدون Feature Selection

مشاهده می‌شود که در داده های تمیز شده و بدون Feature Selection دقت مدل با Over-Sampling بیشینه شده است. اطلاعات مدل برتر به صورت زیر است:

Model Information		
Target Field		Outcome
Model Building Method		Random Trees Classification
Number of Predictors Input		6
Gmean		0.814
True Positive Rate	0	0.726
	1	0.913

جدول24 : اطلاعات مدل Random forest بدون Feature Selection

با استفاده از Feature Selection نتایج به صورت زیر است:

الگوریتم Random Forest	Clean Data With Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	89.08%	85.55%	87.44%	83.82%	93.09%	90.75%

جدول25 : دقت الگوریتم Random forest با Feature Selection

مشاهده می‌شود که در داده های تمیز شده و با Feature Selection نیز دقت مدل با Over-Sampling بیشینه شده است. اطلاعات مدل برتر به صورت زیر است:

Model Information		
Target Field		Outcome
Model Building Method		Random Trees Classification
Number of Predictors Input		4
Gmean		0.819
True Positive Rate	0	0.761
	1	0.881

جدول26 : اطلاعات مدل Random forest با Feature Selection

برای مدلسازی بدون پاکسازی داده ها نیز نتایج به صورت زیر است:

الگوریتم Random Forest	Without Data Cleaning					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
Standard	90.71%	87.97%	87.89%	87.97%	94.06%	93.98%

جدول27 : دقت الگوریتم Random forest بدون پاکسازی داده ها

مشاهده می‌شود که در داده های تمیز نشده نیز دقت مدل با Over-Sampling بیشینه شده است. اطلاعات مدل برتر به صورت زیر است:

Model Information		
Target Field		Outcome
Model Building Method		Random Trees Classification
Number of Predictors Input		8
Gmean		0.792
True Positive Rate	0	0.715
	1	0.878

جدول28 : اطلاعات مدل Random forest بدون پاکسازی داده ها

مدل های ساخته شده در این بخش در فایل diabetes-all-models.str قابل مشاهده است.

۷. انتخاب مدل:

پس از بررسی تمامی مدل ها و با توجه به دقت آنها مدل های زیر انتخاب شدند:

الگوریتم	Clean Data Without Feature Selection					
	Imbalance		Under-Sampling		Over-Sampling	
	Train	Test	Train	Test	Train	Test
KNN	81.26%	78.66%	75.24%	75.45%	93.93%	94.61%
SVM[1]	91.85%	91.02%	86.47%	86.96%	95.43%	94.12%
Neural Network[2]	92.58%	90.54%	86.60%	86.36%	90.10%	91.93%
C&RT[3]	92.77%	93.05%	87.59%	86.21%	94.35%	93.96%
QUEST[4]	85.16%	84.46%	81.38%	80.14%	90.81%	88.31%
CHAID[5]	93.58%	91.72%	84.86%	83.45%	94.29%	95.17%
C5.0	81.63%	77.22%	78.07%	75.56%	88.10%	86.67%
Random Forest	90.35%	86.22%	87.90%	85.71%	94.33%	93.41%
الگوریتم	Clean Data With Feature Selection					
KNN	79.58%	79.47%	77.16%	74.17%	94.36%	91.39%
SVM[1]	95.29%	93.64%	85.78%	82.66%	93.67%	91.25%
Neural Network[2]	94.61%	94.23%	78.97%	78.21%	93.69%	94.87%
C&RT[3]	94.27%	93.59%	87.32%	86.54%	93.27%	94.23%
QUEST[4]	83.63%	80.25%	79.18%	78.98%	85.96%	84.08%
CHAID[5]	94.25%	90.57%	87.85%	86.79%	92.73%	91.19%
C5.0	80.14%	77.65%	79.30%	74.61%	80.24%	78.77%
Random Forest	89.08%	85.55%	87.44%	83.82%	93.09%	90.75%
الگوریتم	Not Cleaned Data					
C&RT[3]	96.06%	95.60%	87.32%	83.14%	92.85%	91.84%
QUEST[4]	83.63%	80.25%	79.18%	78.98%	85.96%	84.08%
CHAID[5]	95.00%	93.92%	86.56%	87.16%	93.88%	94.59%
C5.0	85.92%	84.19%	80.79%	86.62%	94.77%	93.66%
Random Forest	90.71%	87.97%	87.89%	87.97%	94.06%	93.98%

جدول 29 : مدل های برتر

۱. در SVM از روش RBF استفاده شده است.
۲. در شبکه عصبی در حالت Clean Data Without Feature Selection از Bagging و در حالت Clean Data With Feature Selection از Boosting استفاده شده است.
۳. در C&RT از Boosting استفاده شده است.
۴. در QUEST از Boosting استفاده شده است.
۵. در CHAID از Boosting استفاده شده است.

همانگونه که مشاهده می‌شود در هر سه حالت فوق، روش Over-Sampling از دقت بالاتری برخوردار است. پس در هر سه حالت مدل های با دقت بالای ۹۰٪ را انتخاب و به صورت موازی استفاده شده و از خروجی آنها به روش Voting برای پیشبینی استفاده می‌شود. سپس دقت هر سه حالت با هم مقایسه شده و حالت برتر بدست می آید.

اطلاعات زیر در فایل final-evaluation.xlsx قابل مشاهده است.

برای حالت Clean Data Without Feature Selection داریم:

Train Set:

		predicted	
		1	0
actual	1	391	0
	0	4	395

	class	
	1	0
accuracy	0.995	
precision	0.990	1.000
recall	1.000	0.990
F1	0.995	0.995

Test Set:

		predicted	
		1	0
actual	1	60	2
	0	4	95

	class	
	1	0
accuracy	0.963	
precision	0.938	0.979
recall	0.968	0.960
F1	0.952	0.969

جدول ۳۰: ارزیابی مدل های برتر موازی بدون Feature Selection

برای حالت Clean Data With Feature Selection داریم:

Train Set:

		predicted	
		1	0
actual	1	422	2
	0	2	392

	class	
	1	0
accuracy	0.995	
precision	0.995	0.995
recall	0.995	0.995
F1	0.995	0.995

Test Set:

		predicted	
		1	0
actual	1	36	0
	0	1	99

	class	
	1	0
accuracy	0.993	
precision	0.973	1.000
recall	1.000	0.990
F1	0.986	0.995

جدول ۳۱: ارزیابی مدل های برتر موازی با Feature Selection

برای حالت بدون پاکسازی داریم:

Train Set:

		predicted	
		1	0
actual	1	396	2
	0	4	399

	class	
	1	0
accuracy	0.993	
precision	0.990	0.995
recall	0.995	0.990
F1	0.992	0.993

Test Set:

		predicted	
		1	0
actual	1	53	0
	0	2	95

	class	
	1	0
accuracy	0.987	
precision	0.964	1.000
recall	1.000	0.979
F1	0.981	0.990

جدول ۳۲: ارزیابی مدل های برتر موازی بدون پاکسازی داده ها

همانطور که مشاهده می‌شود حالت Clean Data With Feature Selection بهترین عملکرد را داشته است.

مدل های ساخته شده در این بخش در فایل diabetes-best-models.str قابل مشاهده است.

۷. نتیجه گیری:

پس از آزمایش مدل ها و هایپرپارامتر های مختلف، در نهایت استفاده از شرایط و مدل های زیر به صورت موازی بهترین نتیجه را برای پیش بینی ابتلا به دیابت ارائه داد:

۱. حذف ویژگی Insulin به دلیل داده مفقوده زیاد
۲. حذف ویژگی SkinThickness به دلیل وجود چالش هم‌خطی و همچنین داده مفقوده زیاد
۳. استفاده از روش F-Score برای انتخاب ویژگی که ۴ ویژگی Glucose، BMI، Age و Pregnancies انتخاب شدند.
۴. تقسیم دادگان به دو گروه ۸۰ و ۲۰ درصدی برای Train و Test
۵. استفاده از روش Over-Sampling برای مدیریت دادگان نامتوازن
۶. استفاده از مدل های KNN، SVM، Neural Network، C&RT، CHAID و Random Forest برای مدلسازی
۷. پیش بینی با استفاده از روش Voting روی خروجی مدل ها

در نهایت دقت پیش بینی ۹۹.۵٪ برای داده های Train و ۹۹.۳٪ برای داده های Test به صورت زیر بدست آمد:

Train Set:

		predicted	
		1	0
actual	1	422	2
	0	2	392

	class	
	1	0
accuracy	0.995	
precision	0.995	0.995
recall	0.995	0.995
F1	0.995	0.995

Test Set:

		predicted	
		1	0
actual	1	36	0
	0	1	99

	class	
	1	0
accuracy	0.993	
precision	0.973	1.000
recall	1.000	0.990
F1	0.986	0.995

مدل نهایی در فایل diabetes-final-model.str قابل مشاهده است.

۸. فهرست شکل ها:

۹. فهرست جدول ها:

۱۰. منابع:

1. Amatul Zehra: A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset, https://core.ac.uk/download/pdf/159180503.pdf
2. Yi-Wei Chen and Chih-Jen Lin, Combining SVMs with Various Feature Selection Strategies, https://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf
3. Ilango, B.S., Ramaraj, N.: Hybrid Prediction Model with F-score Feature Selection for Type II Diabetes Databases, https://doi.org/10.1145/1858378.1858391
4. Hussan, B.M., Data Mining based Prediction of Medical data Using K-means algorithm, http://dx.doi.org/10.1109/ICCRD.2011.5764179
5. Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.: Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients, https://www.ijeat.org/wp-content/uploads/papers/v1i3/C0211021312.pdf
6. Giveki, D., Salimi, H., Bahmanyar, G.R., Khademian, Y.: Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search, https://arxiv.org/abs/1201.2173
7. Karegowda, A.G., Punya, V., Jayaram, M.A., Manjunath, A.S.: Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5, https://www.ijcaonline.org/archives/volume45/number12/6836-9460
8. Breault, J.L.: Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?, https://www.researchgate.net/publication/215899115_Data_mining_diabetic_databases_Are_rough_sets_a_useful_addition
9. Han, J., Rodriguze, J.C., Beheshti, M.: Diabetes data analysis and prediction model discovery using RapidMiner, https://ieeexplore.ieee.org/document/4734287
10. Pradhan, M., Sahu, R.K.: Predict the onset of diabetes disease using Artificial Neural Network, https://www.researchgate.net/publication/228850003_Predict_the_onset_of_diabetes_disease_using_Artificial_Neural_Network_ANN