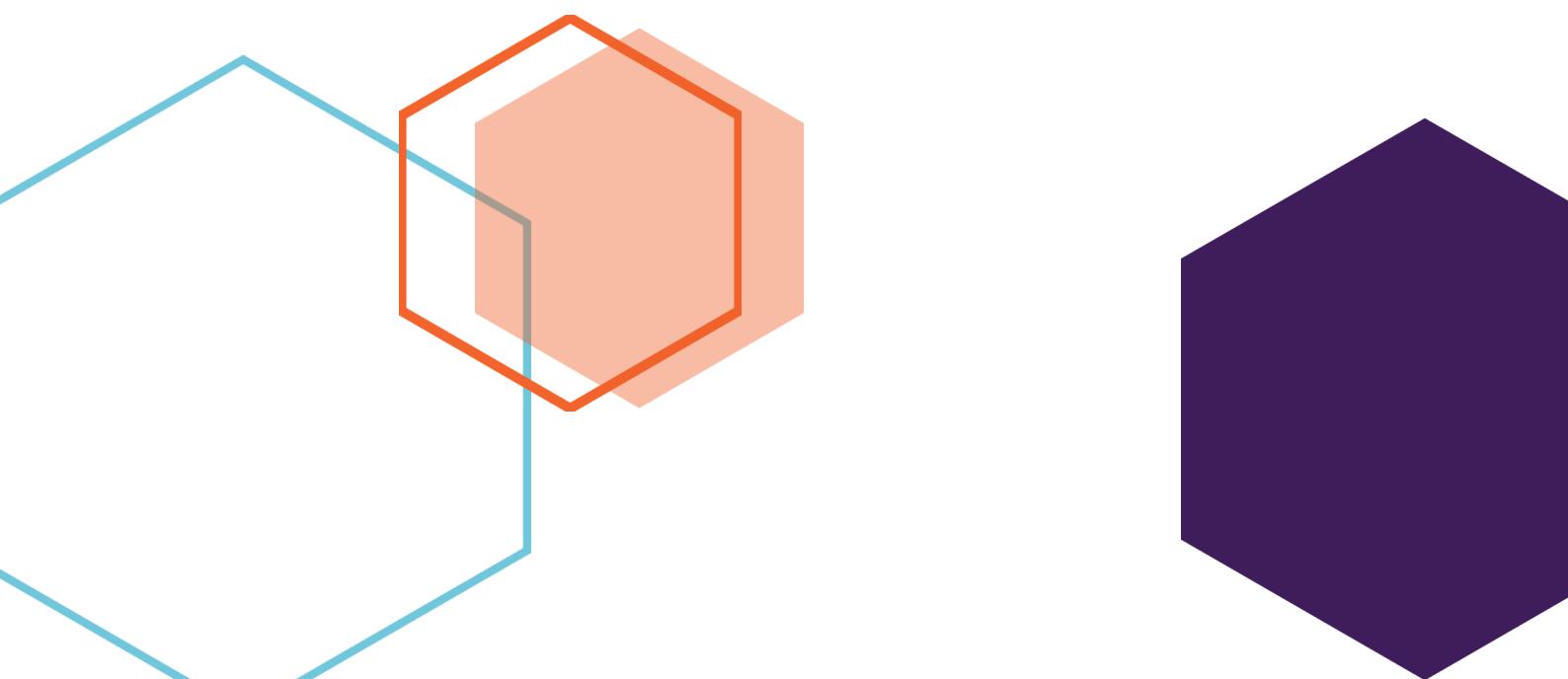


گزارش کار با RapidMiner

پرهام زیلوچیان مقدم

این گزارش فنی مربوط به پروژه درس داده کاوی در نیمسال تحصیلی اول سال ۱۳۹۷ میباشد.



[Type the document title]

• • •



University of Kashan

دانشکده برق و کامپیوتر

گزارش پروژه درس داده کاوی

توسط:

پرهام زیلوچیان مقدم

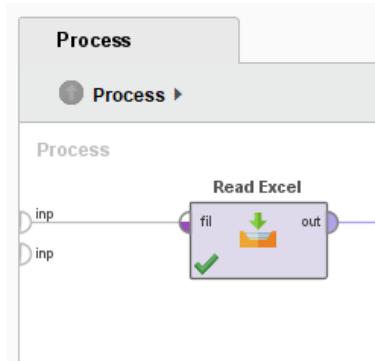
استاد درس:

دکتر سید مهدی وحیدی پور

زمستان 97

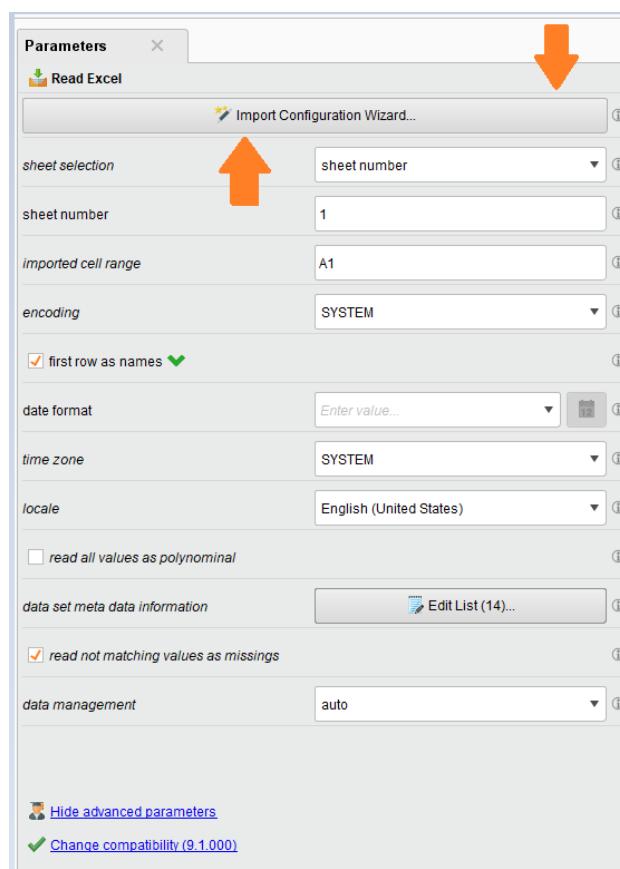
سوال 1:

در قدم اول به این دلیل که ما میخواهیم یک فایل Excel را بخوانیم و داده‌های آن را استخراج کنیم از لیست Operator ها ما Read Excel را انتخاب میکنیم و درون آن مسیر قرارگیری فایل اکسل را وارد میکنیم تا داده‌ها را از آنجا بخواند.



شکل 1: اضافه کردن Read Excel

و سپس از قسمت import Configuration Wizard می‌آییم و قسمت Parameters رو میزنیم.

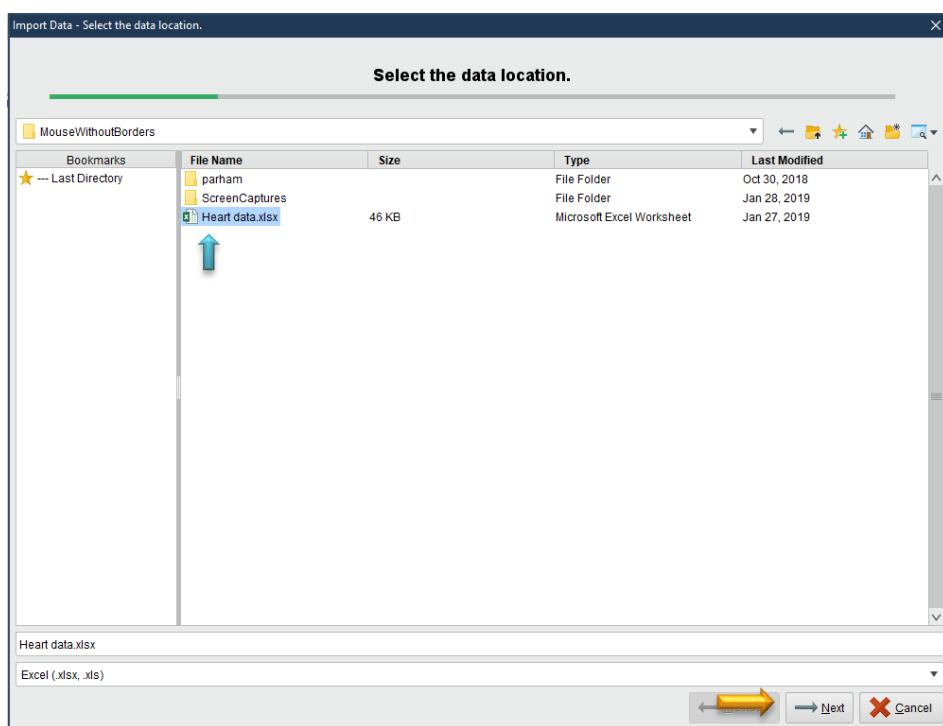


شکل 2: بخش Read Excel در Parameters

[Type the document title]

• • •

در پنجره جدیدی که باز می شود باید فایل Excel مورد نظری را که میخواهیم از آن استفاده کنیم را از مسیر مورد نظر انتخاب می کنیم.



شکل 3: کلیک روی بخش Read Excel import Data

در ادامه برروی Next کلیک کرده و به مرحله بعد میرویم که یک پیش‌نمایشی از داده‌های موجود درون فایل انتخاب شده را به ما نشان می‌دهد.

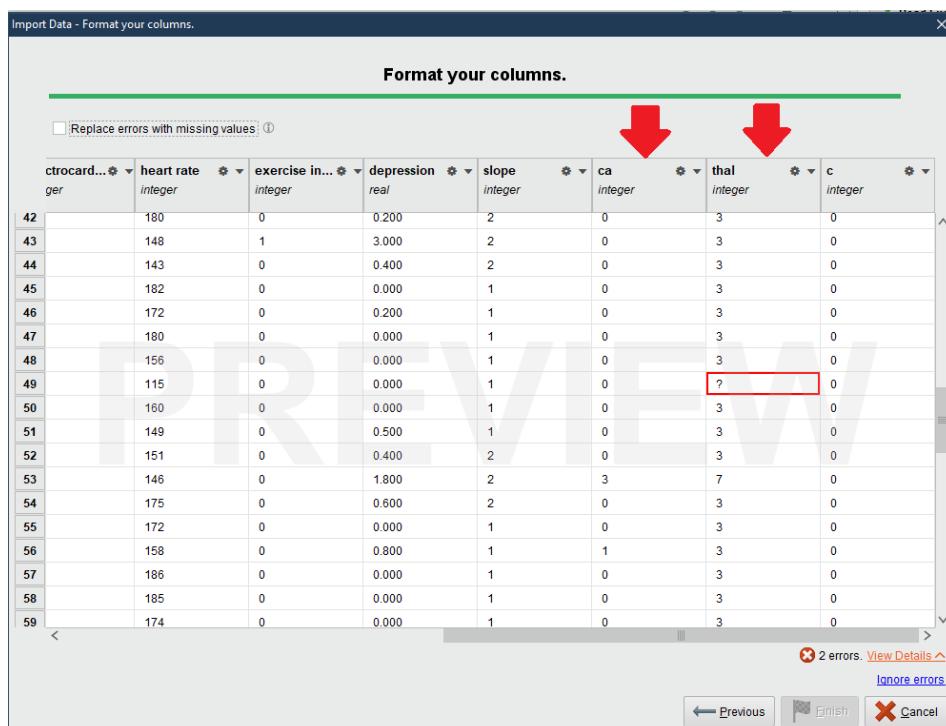
Select the cells to import.																
Sheet:		Cell range:		Select All		Define header row:										
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
3	37.000	1.000	3.000	130.000	250.000	0.000	0.000	187.000	0.000	3.500	3.000	0.000	3.00			
4	41.000	0.000	2.000	130.000	204.000	0.000	2.000	172.000	0.000	1.400	1.000	0.000	3.00			
5	56.000	1.000	2.000	120.000	236.000	0.000	0.000	178.000	0.000	0.800	1.000	0.000	3.00			
6	57.000	0.000	4.000	120.000	354.000	0.000	0.000	163.000	1.000	0.600	1.000	0.000	3.00			
7	57.000	1.000	4.000	140.000	192.000	0.000	0.000	148.000	0.000	0.400	2.000	0.000	6.00			
8	56.000	0.000	2.000	140.000	294.000	0.000	2.000	153.000	0.000	1.300	2.000	0.000	3.00			
9	44.000	1.000	2.000	120.000	263.000	0.000	0.000	173.000	0.000	0.000	1.000	0.000	7.00			
10	52.000	1.000	3.000	172.000	199.000	1.000	0.000	162.000	0.000	0.500	1.000	0.000	7.00			
11	57.000	1.000	3.000	150.000	168.000	0.000	0.000	174.000	0.000	1.600	1.000	0.000	3.00			
12	54.000	1.000	4.000	140.000	239.000	0.000	0.000	160.000	0.000	1.200	1.000	0.000	3.00			
13	48.000	0.000	3.000	130.000	275.000	0.000	0.000	139.000	0.000	0.200	1.000	0.000	3.00			
14	49.000	1.000	2.000	130.000	266.000	0.000	0.000	171.000	0.000	0.600	1.000	0.000	3.00			
15	64.000	1.000	1.000	110.000	211.000	0.000	2.000	144.000	1.000	1.800	2.000	0.000	3.00			
16	58.000	0.000	1.000	150.000	283.000	1.000	2.000	162.000	0.000	1.000	1.000	0.000	3.00			
17	50.000	0.000	3.000	120.000	219.000	0.000	0.000	158.000	0.000	1.600	2.000	0.000	3.00			
18	58.000	0.000	3.000	120.000	340.000	0.000	0.000	172.000	0.000	0.000	1.000	0.000	3.00			
19	66.000	0.000	1.000	150.000	226.000	0.000	0.000	114.000	0.000	2.600	3.000	0.000	3.00			
20	43.000	1.000	4.000	150.000	247.000	0.000	0.000	171.000	0.000	1.500	1.000	0.000	3.00			
21	69.000	0.000	1.000	140.000	239.000	0.000	0.000	151.000	0.000	1.800	1.000	2.000	3.00			
22	59.000	1.000	4.000	135.000	234.000	0.000	0.000	161.000	0.000	0.500	2.000	0.000	7.00			
23	44.000	1.000	3.000	130.000	233.000	0.000	0.000	179.000	1.000	0.400	1.000	0.000	3.00			

شکل 4: بخش انتخاب قسمت‌هایی که میخواهیم import کنیم.

• • •

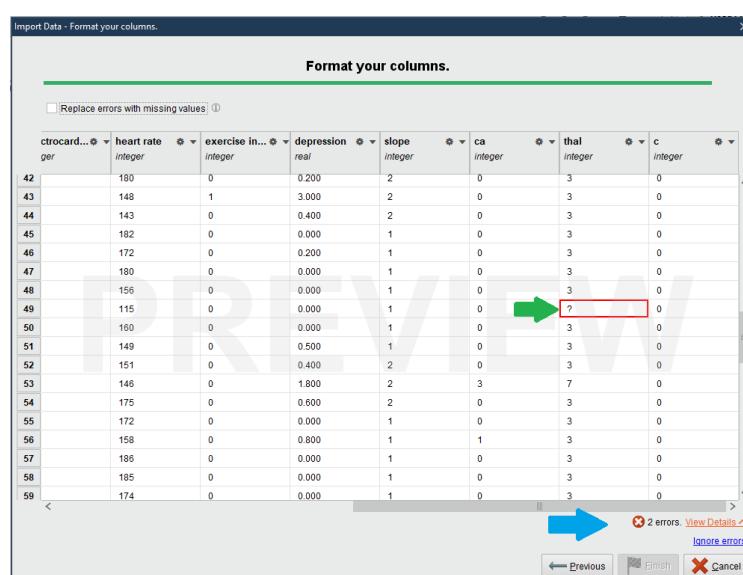
و همین طور از این قسمت میتوانیم Cell هایی رو که میخواهیم از فایل استخراج کند را انتخاب کنیم که خب من در این حالت روی پیشفرض قرار دادم یعنی تمامی داده‌ها را استخراج کند. و سپس مجدداً روی Next کلیک کرده و به مرحله بعدی می‌رویم.

سپس در این قسمت به اجراه Format خونه‌ها رو می‌دهد. فقط نکته‌ای که در اینجا وجود دارد این است که چون ما میخواهیم در قسمت Missing Value از ... استفاده کنیم که این موارد هم با داده‌های عددی کار میکنند باید ما نوع دوتا از ستون‌ها رو عوض کنیم، نوع آن‌ها را از polynomial تغییر می‌دهیم. و در این حالت برای داده‌های با مقدار ""؟"" که در واقع همان Missing Value هاستند مشکل به وجود می‌آید.



شکل ۵: تغییر دادن نوع پارامترهای مشخص شده در شکل

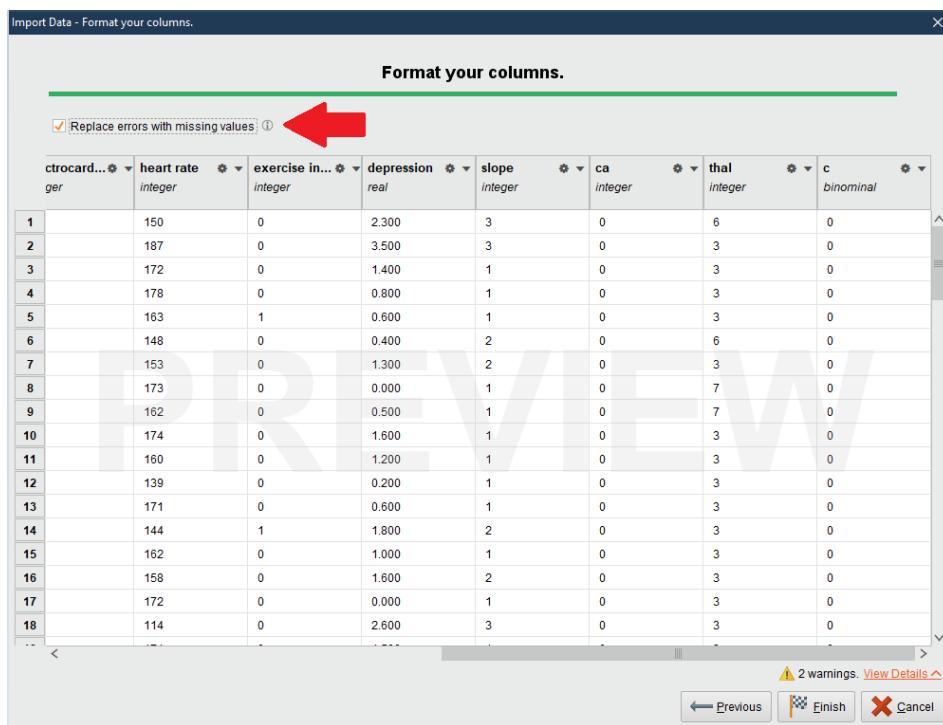
در این حالت پیام خطای زیر ظاهر می‌شود:



شکل ۶: ظاهر شدن پیام‌ها خطای خانه‌های دارای مقدار ""؟""

• • •

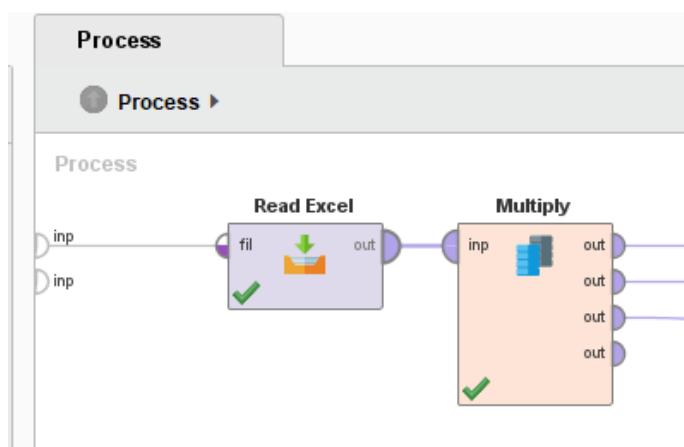
حال برای رفع این خطاهای یک تیک در بالای صفحه وجود دارد و ما باید آن تیک را فعال کنیم تا مقدیر "?" را به نوع Missing Value تبدیل کند:



شکل 7: فعال کردن تیک مشخص شده در شکل

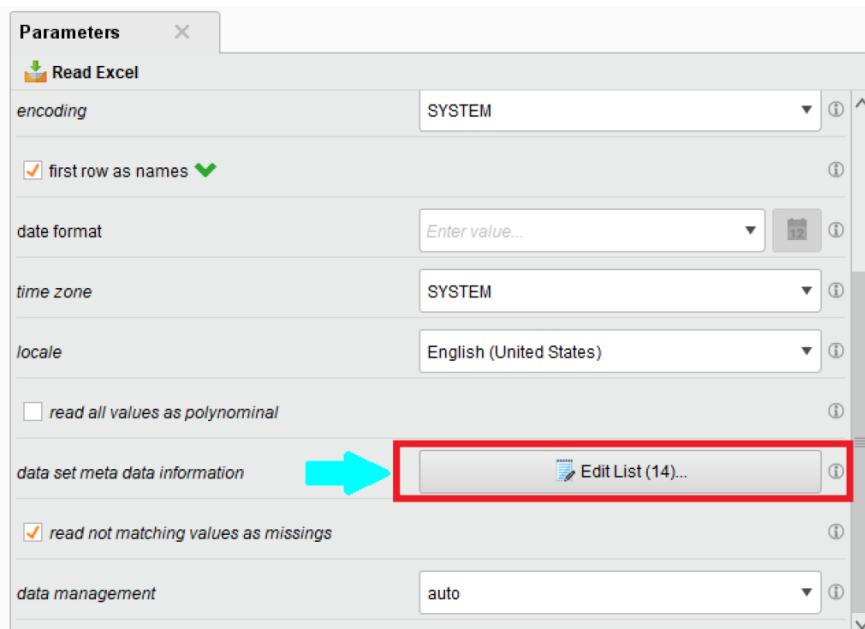
و با این کار خطاهای رفع شده و در انتهای روی Finish کلیک کرده و داده‌ها تماماً import می‌شوند.

در قدم بعدی به این دلیل که در سوال اول از ما خواسته‌اید تا تمامی موارد را در یک Process انجام دهیم به همین دلیل در اینجا ما از Operator ای به نام استفاده می‌کنیم که خروجی ما را تبدیل به چند خروجی می‌کند تا بتوانیم از این خروجی‌ها در قسمت‌های مختلف سوال اول (مانند: ..., a, b, c) استفاده کنیم.



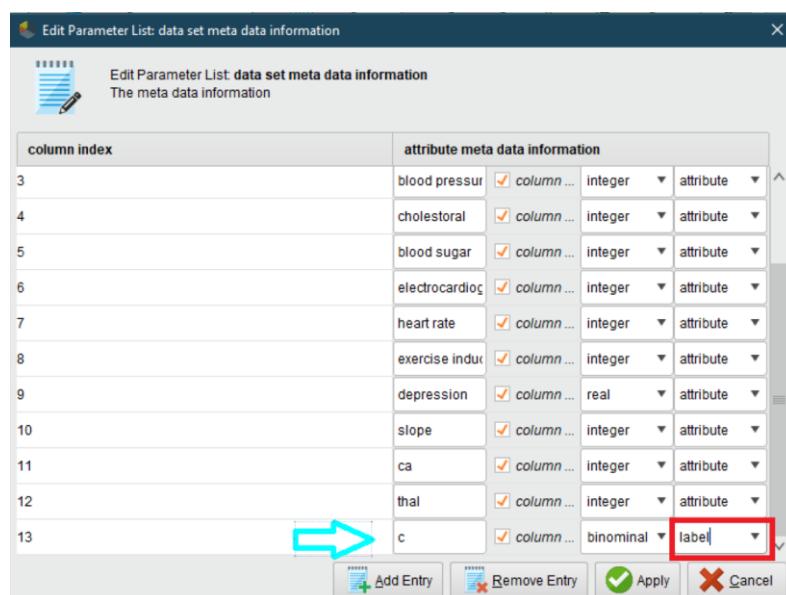
شکل 8: اضافه کردن اپراتور Multiply

نکته پایانی ای که باقی می‌ماند این است که چون در صورت سوال گفته است که متغیر هدف ما C است پس ما باید این متغیر را از نوع Label بگذاریم. برای این کار نیز کافی است روی اپراتور Read Excel برویم و در قسمت اپراتورهای آن این کار را انجام دهیم، و نیز در شکل 9 این مورد به وضوح قابل مشاهده است که باید روی قسمت Edit List کلیک کنیم.



شکل 9: رفتن به قسمت تنظیمات نوع متغیرها در اپراتور Read Excel در قسمت آن Parameters

پس از این کار نیز باید در صفحه باز شده مقدار نوع متغیر را به Label مطابق شکل 10 تغییر می‌دهیم.

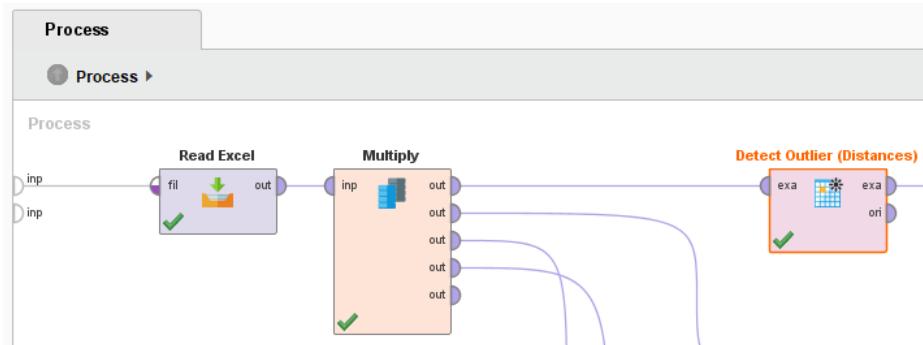


شکل 10: انجام تغییرات روی متغیرهای خوانده نشده به وسیله اپراتور Read Excel

و در پایان نیز روی Apply کلیک کرده تا تغییرات ثبت شوند.

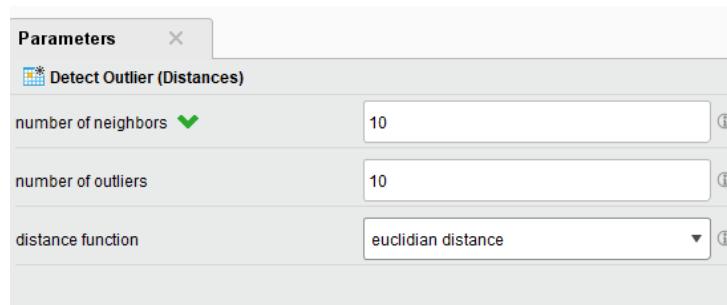
مورد a :

در این قسمت از ما خواسته شده که نویزها را با استفاده از روش Detect Outlier Distances تشخیص داده و آنها را با استفاده از فیلتر حذف کنیم. برای این کار ما از منوی Detect Outlier Distance Operator میدهیم:



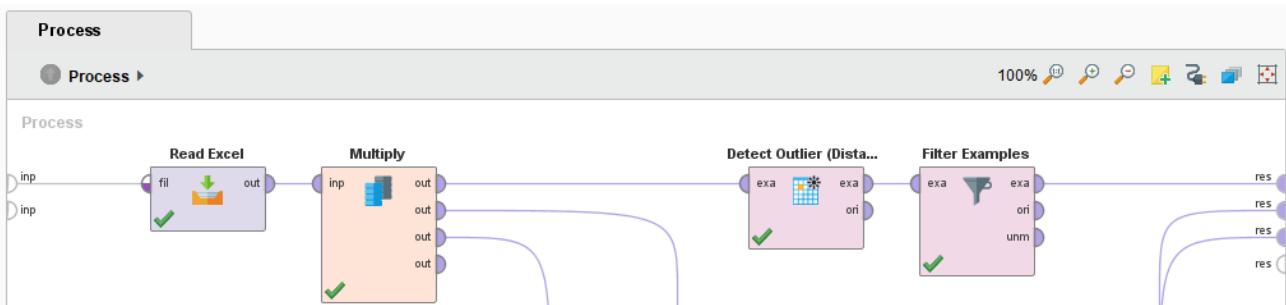
شکل 11: اتصال خروجی فایل Read Excel به Detect Outlier Distance

و سپس پارامترهای مربوط به Detect Outlier Distance را تغییر نمی‌دهیم و به حالت پیش‌فرض می‌گذاریم. تعداد همسایه‌ها برابر با 10 و همین‌طور تعداد Outlier‌ها نیز برابر با 10 می‌باشد.تابع مورد استفاده برای محاسبه فاصله هم فاصله اقلیدسی هست:



شکل 12: مقادیر پارامترهای موجود در Detect Outlier Distance

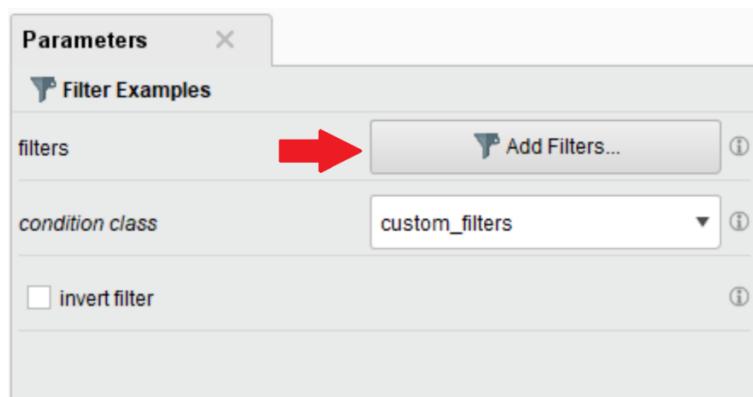
سپس ما از قسمت Process را انتخاب کرده و آن را به خود اضافه می‌کنیم:



شکل 13: اضافه کردن و اتصال Filter Example به موارد رسم شده تاکنون

در ادامه کار باید به مقادیر مورد نیاز برای فیلتر کردن نتایج را بدهیم. در قسمت Parameters باید آن‌ها را وارد کنیم.

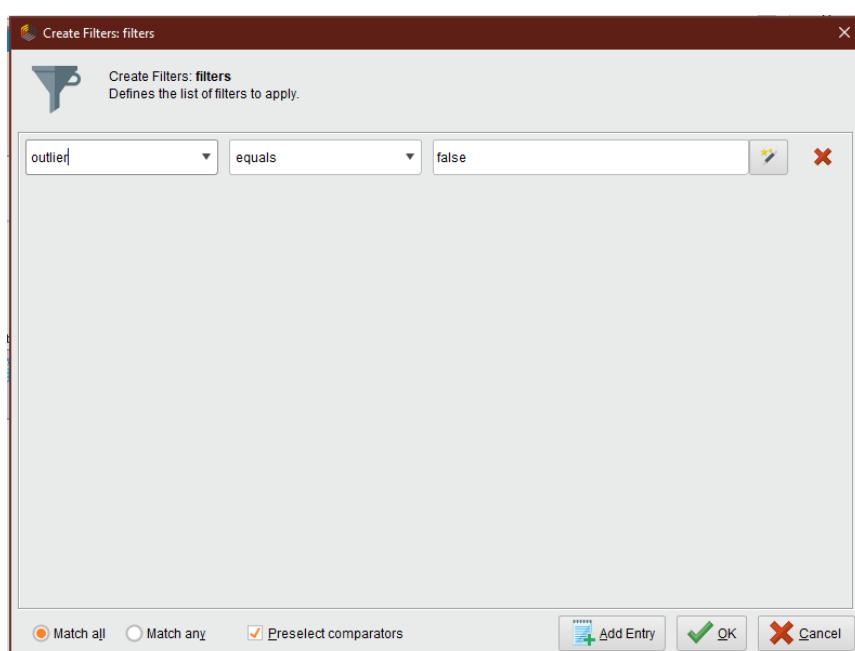
• • •



شکل 14: قسمت Filter Example در Parameters

و ما مقدار Condition Class را نیز برابر با custom_filter قرار می‌دهیم، تا بتوانیم فیلترهای دلخواه خود را اعمال کنیم.

قسمت مشخص شده در شکل 14 را نیز برای اضافه کردن فیلتر می‌زنیم تا فیلتر جدیدی را تعریف کنیم:

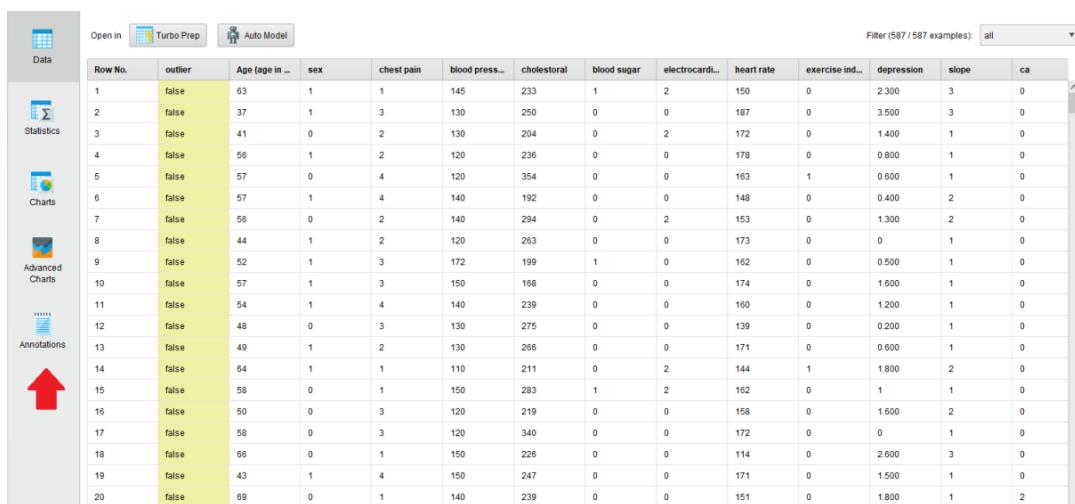


شکل 15: اضافه کردن فیلتر در Filter Example

ما طبق صورت سوال اول که از ما خواسته است Outlier Distance ها را با فیلتر حذف کنیم و برای این کار ما از داخل لیست آن مشخصه‌ای که می‌خواهیم فیلتر کنیم را انتخاب می‌کنیم و سپس نوع بررسی که می‌خواهیم انجام دهیم را انتخاب می‌کنیم و چون ما در اینجا می‌خواهیم آن‌های که مقدار Outlier آن‌ها است را انتخاب کنیم تا بر این اساس آن‌ها را دسته‌بندی کنیم. اگر بخواهیم فیلتر جدیدی هم اضافه کنیم کافی است که Add Entry در پایین شکل 15 را بزنیم که خب در این سوال نیاز به فیلتر بیشتری نیست.

حال برنامه را اجرا می‌کنیم تا خروجی را ببینیم:

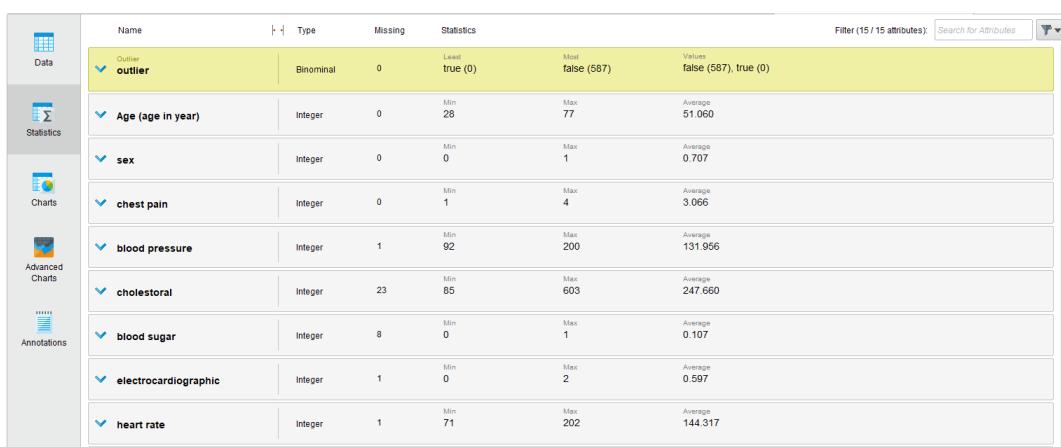
• • •



The screenshot shows the KNIME interface with the 'Data' view selected. The main area displays a table with 20 rows and 13 columns. The columns are: Row No., outlier, Age (age in ...), sex, chest pain, blood pressur..., cholestral, blood sugar, electrocardiograp..., heart rate, exercise ind..., depression, slope, and ca. The first row is highlighted in yellow. A red arrow points upwards from the 'Annotations' section on the left towards the Data table.

شکل 16: خروجی نوع Data در سوال 1 قسمت a

حال همان طور که با فلاش قرمز در شکل 16 مشخص شده است میتوانیم نتایج با چندین نوع تحلیل مشاهده کنیم که شکل 16 نوع دادهای آن را نشان می‌دهد و ما می‌توانیم با نوع نموداری و تحلیلی و ... نیز آن را مشاهده کنیم.



The screenshot shows the KNIME interface with the 'Statistics' view selected. The main area displays a table with 15 attributes and their statistics. The columns are: Name, Type, Missing, Statistics, and Filter (15 / 15 attributes). The attributes listed are: Outlier, Age (age in year), sex, chest pain, blood pressure, cholestral, blood sugar, electrocardiographic, and heart rate. The 'Outlier' attribute is highlighted in yellow.

شکل 17: پاسخ قسمت a سوال 1 با تحلیل Statistics

به عنوان مثال در شکل 17 تحلیل Statistics را از سوال 1 میبینید. همان طور که در شکل مشخص است تعداد Outliers برابر با 0 است و این نشان می‌دهد که فیلتر به درستی انجام شده است.

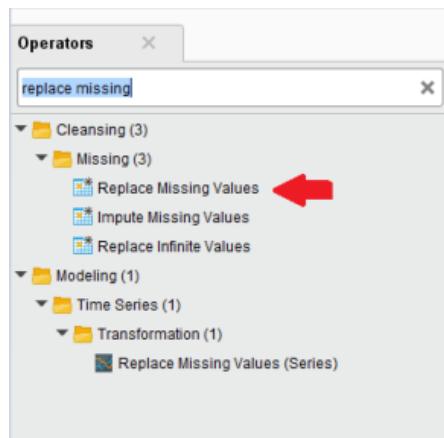
: b مورد

در این قسمت صورت سوال از ما خواسته است که Replace Missing Value را با استفاده از روش Maximum و با مقدار جایگزین کنیم.

همان‌طور که در ابتدای گزارش و در قسمت وارد کردن فایل اکسل مشاهده کردیم ما مقادیری برابر با ”?” داشتیم که به آن‌ها Missing Value می‌گوییم. حال در این سوال میخواهیم آن‌ها را با مقداری طبق فرمول داده شده جایگزین کنیم.

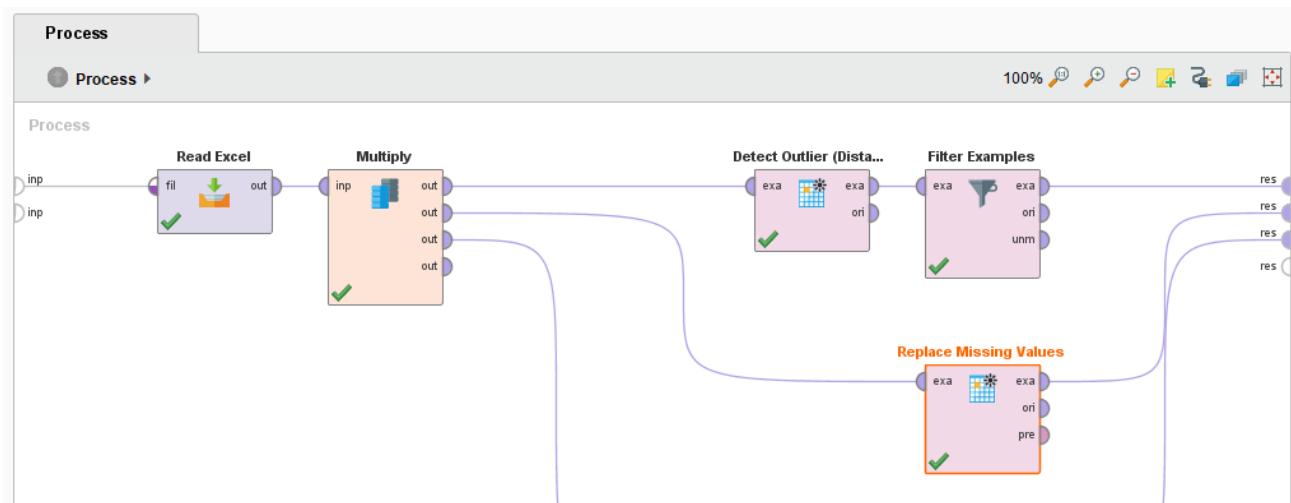
در این قسمت کافی است که در قسمت Replace Missing Operator را جستجو کنیم:

• • •



شکل 18: جستجوی Replace Missing Operator در قسمت ها

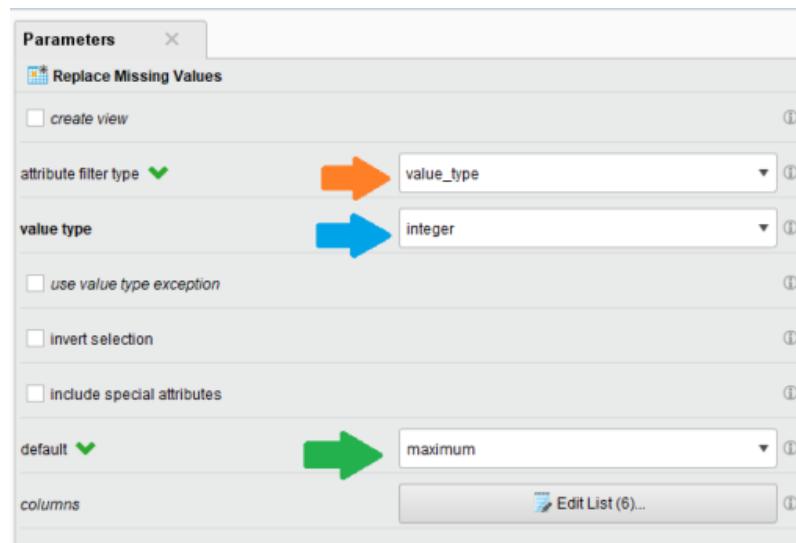
قسمت مشخص شده در شکل 18 را Drag & Drop به داخل محیط طراحی می‌آوریم، و یکی از خروجی‌های قسمت خواند فایل را به ورودی آن وصل کرده و خروجی قسمت exa آن را نیز به خروجی Process وصل می‌کنیم و کار طراحی ما تمام است. و میتوانید نتیجه آن را در شکل 19 مشاهده کنید.



شکل 19 : اضافه کردن اپراتور Replace Missing Value

پس از رسم و اتصال اپراتور اضافه شده به سیستم، حال باید برویم سراغ تنظیمات آن که با کلیک روی Replace Missing Value این کار را انجام میدهیم و به قسمت پارامترهای آن میرویم.

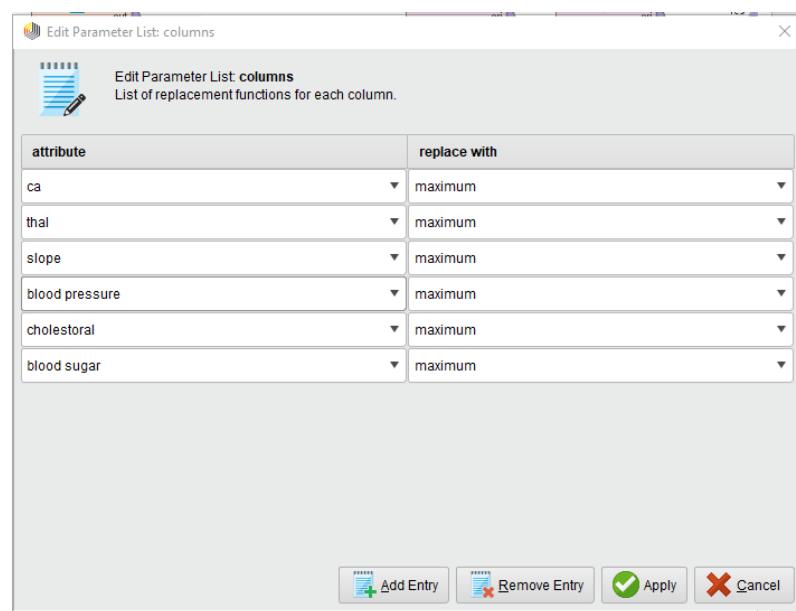
• • •



شکل 20 : قسمت Replace Missing Value در اپراتور Parameters

ابتدا قسمت مشخص شده در شکل 20 با فلش نارنجی رنگ (Attribute filter type) را روی value_type (Attribute filter type) تنظیم میکنیم تا فقط صفاتی را جایگزین کنیم که نوع آنها از نوع integer است و قابلیت پیدا کردن Maximum را دارند و نیز از قسمت قبلی سوال میدانیم که فقط چندتا از صفات که بودند Missing Value داشتند و مابقی نداشتند به همین دلیل مشکلی در کار ما ایجاد نمیکند.

و قسمت مشخص شده با فلش سبز رنگ (default) را نیز برابر با Maximum قرار میدهیم که یعنی به صورت پیشفرض مقادیر از دست رفته را با مقدار Maximum جایگزین کند. اما در این قسمت همچنین ما میتوانیم با کلیک روی گزینه Edit List به صورت دلخواه به هر صفتی یک حالت به خصوصی را برای جایگزینی بدهیم.

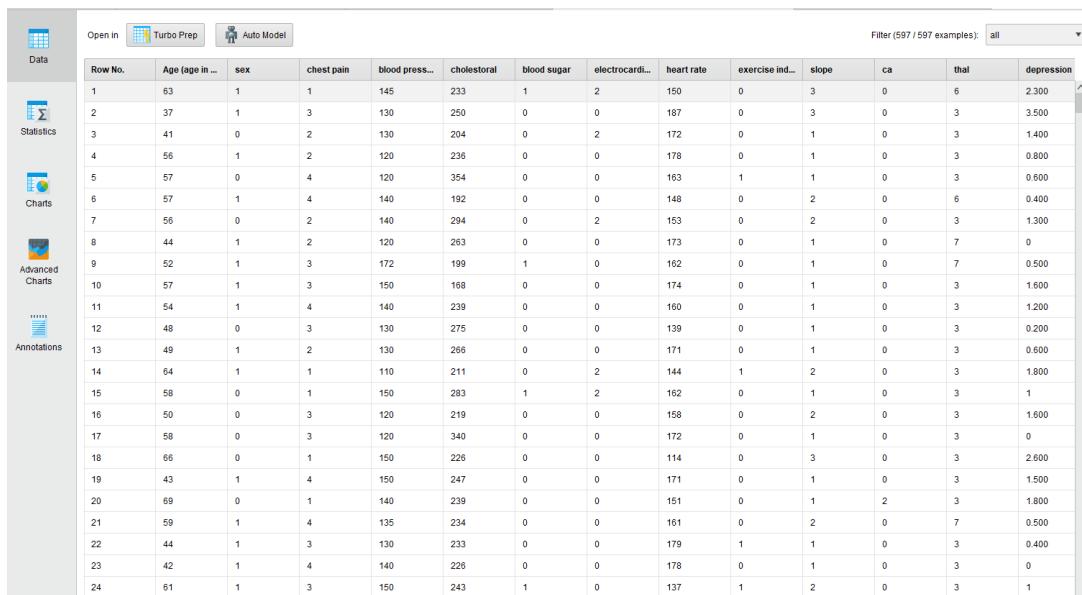


شکل 21 : قسمت Replace Missing Value در اپراتور Edit Parameter List

• • •

همان طور که در شکل بالا مشاهده میکنید میتوانیم نوع Missing Value attribute را که میخواهیم برای چک کردن مشخص کنیم و بگوییم که این صفت را با چه مقداری جایگزین کند مثلا Average Maximum Minimum و ... که چون در صورت سوال از ما خواسته که با جایگزین کنیم ما این مقادیر را برابر با maximum گذاشتیم.

حال این برنامه را اجرا کرده و نتایج آن مطابق زیر است:

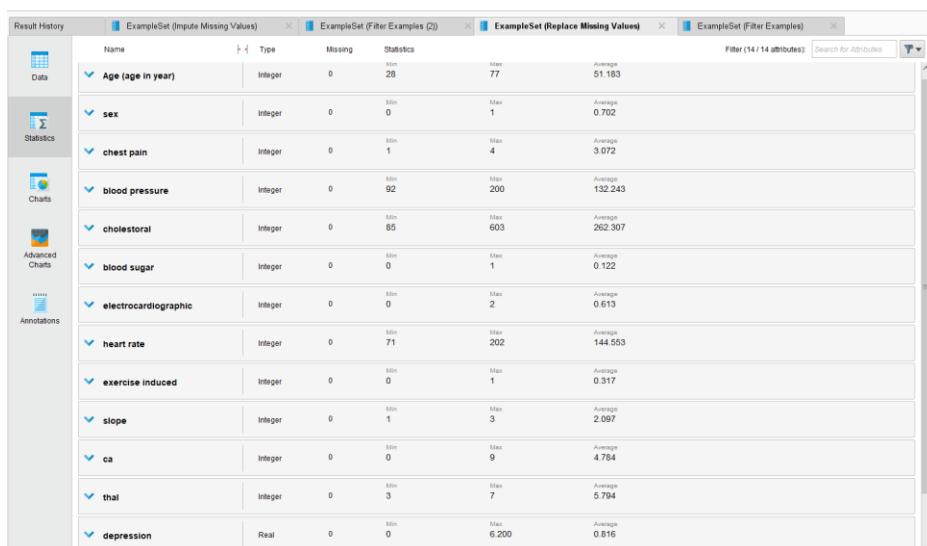


The screenshot shows the Orange data mining interface. On the left, there's a sidebar with icons for Data, Statistics, Charts, Advanced Charts, and Annotations. The main area displays a data table titled "Auto Model". The table has 14 columns: Row No., Age (age in years), sex, chest pain, blood press..., cholesterol, blood sugar, electrocardiograp..., heart rate, exercise indu..., slope, ca, thal, and depression. There are 24 rows of data. At the top of the table, there are buttons for "Open in" (Turbo Prep, Auto Model), a search bar, and a filter dropdown set to "all".

Row No.	Age (age in years)	sex	chest pain	blood press...	cholesterol	blood sugar	electrocardiograp...	heart rate	exercise indu...	slope	ca	thal	depression
1	63	1	1	145	233	1	2	150	0	3	0	6	2.300
2	37	1	3	130	250	0	0	187	0	3	0	3	3.500
3	41	0	2	130	204	0	2	172	0	1	0	3	1.400
4	56	1	2	120	236	0	0	178	0	1	0	3	0.800
5	57	0	4	120	354	0	0	163	1	1	0	3	0.600
6	57	1	4	140	192	0	0	148	0	2	0	6	0.400
7	56	0	2	140	294	0	2	153	0	2	0	3	1.300
8	44	1	2	120	263	0	0	173	0	1	0	7	0
9	52	1	3	172	199	1	0	162	0	1	0	7	0.500
10	57	1	3	150	188	0	0	174	0	1	0	3	1.600
11	54	1	4	140	239	0	0	160	0	1	0	3	1.200
12	48	0	3	130	275	0	0	139	0	1	0	3	0.200
13	49	1	2	130	266	0	0	171	0	1	0	3	0.600
14	64	1	1	110	211	0	2	144	1	2	0	3	1.800
15	58	0	1	150	283	1	2	162	0	1	0	3	1
16	50	0	3	120	219	0	0	158	0	2	0	3	1.600
17	58	0	3	120	340	0	0	172	0	1	0	3	0
18	66	0	1	150	226	0	0	114	0	3	0	3	2.600
19	43	1	4	150	247	0	0	171	0	1	0	3	1.500
20	69	0	1	140	239	0	0	151	0	1	2	3	1.800
21	59	1	4	135	234	0	0	161	0	2	0	7	0.500
22	44	1	3	130	233	0	0	179	1	1	0	3	0.400
23	42	1	4	140	226	0	0	178	0	1	0	3	0
24	61	1	3	150	243	1	0	137	1	2	0	3	1

شکل 22: نتیجه اجرای قسمت b سوال 1

و همین طور برای این که مطمئن شویم که Missing Value نداریم تحلیل Statistics این پاسخ را نیز چک میکنیم.



The screenshot shows the Orange data mining interface. On the left, there's a sidebar with icons for Data, Statistics, Charts, Advanced Charts, and Annotations. The main area displays a statistics table titled "ExampleSet (Replace Missing Values)". The table lists 14 attributes with their data types, minimum, maximum, and average values. For example, "Age (age in years)" is an Integer type with a minimum of 28, maximum of 77, and average of 51.183. Other attributes like "ca" and "thal" are also listed. A search bar at the top right allows filtering by attribute name.

Name	Type	Missing	Min	Max	Average
Age (age in years)	Integer	0	28	77	51.183
sex	Integer	0	0	1	0.702
chest pain	Integer	0	1	4	3.072
blood pressure	Integer	0	92	200	132.243
cholesterol	Integer	0	85	603	262.307
blood sugar	Integer	0	0	1	0.122
electrocardiographic	Integer	0	0	2	0.613
heart rate	Integer	0	71	202	144.553
exercise induced	Integer	0	0	1	0.317
slope	Integer	0	1	3	2.097
ca	Integer	0	0	9	4.784
thal	Integer	0	3	7	5.794
depression	Real	0	0	6.200	0.816

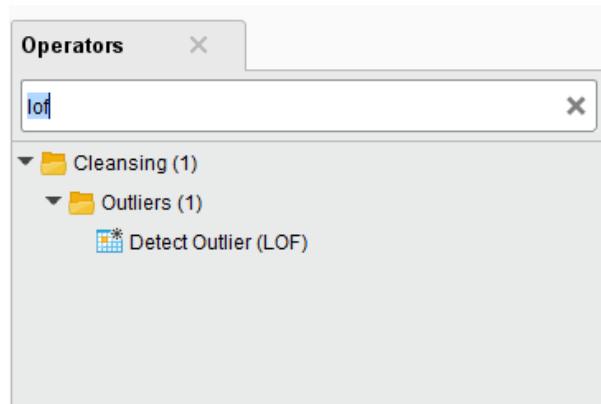
شکل 23: تحلیل Statistics از قسمت b سوال 1

همان طور که میبینیم مقدار Missing برابر با صفر است که یعنی تماماً جایگزین شده‌اند.

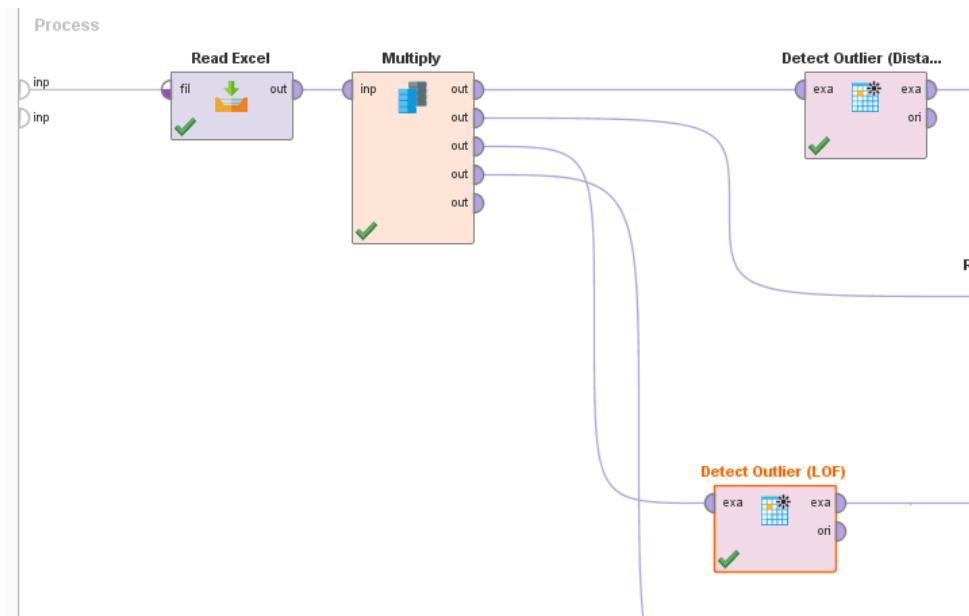
مورد C :

در این قسمت صورت سوال از ما خواسته است که نویزها را با استفاده از روش LOF و با "حد پایین: 3 نزدیکترین همسایگی"، "حد بالا: 7 نزدیکترین همسایگی" مشخص کنیم و با شرط پیشنهادی خودمان آنها را فیلتر کرده و حذف کنیم.

برای این کار ابتدا اپراتور LOF را در لیست اپراتورها جستجو کرده و آن را با استفاده از Drag & Drop آن را به درون Process خود میاوریم:



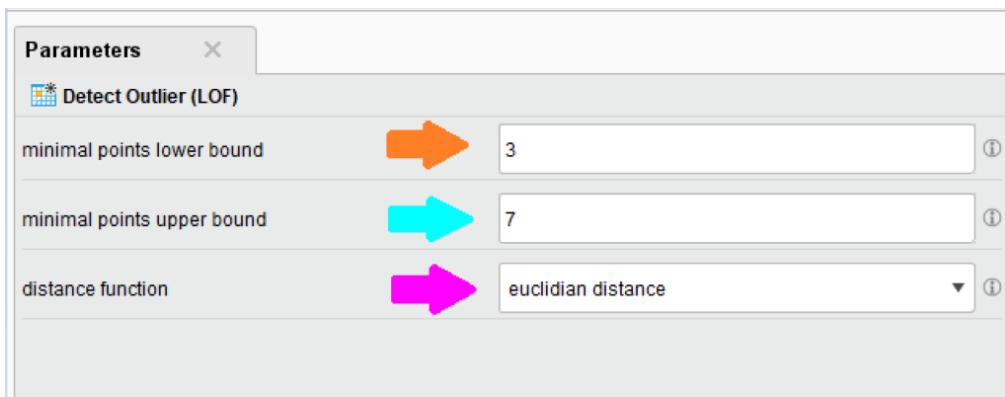
شکل 24: جستجوی اپراتور LOF در لیست اپراتورها



شکل 25: اضافه کردن اپراتور LOF به موجود Process

پس از اضافه کردن LOF حال به سراغ تنظیمات این اپراتور میرویم که مطابق با موارد بیان شده در صورت سوال آنها را تنظیم کنیم. برای این کار روی LOF کلیک کرده و به قسمت Parameters آن می‌رویم تا این مقادیر را تنظیم کنیم.

• • •

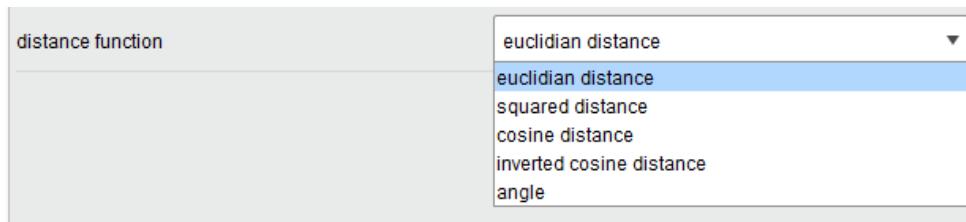


شکل 26: قسمت LOF مربوط به اپراتور Parameters

مقدار minimal points lower bound (فلش نارنجی رنگ در شکل 26) را همان‌طور که در صورت سوال گفته بود برابر با مقدار 3 میگذاریم که همان حد پایین ما هست. (مقدار این متغیر میتواند از 0 تا بی‌نهایت باشد و پیش فرض آن در RapidMiner برابر با 10 هست.)

و مقدار minimal points upper bound (فلش آبی رنگ در شکل 26) را طبق صورت سوال که گفته است حد بالا را برابر با مقدار 7 قرار دهیم ما نیز همین کار را میکنیم. (مقدار این متغیر نیز میتواند از 0 تا بی‌نهایت باشد و مقدار پیش فرض آن در RapidMiner برابر با 20 هست.)

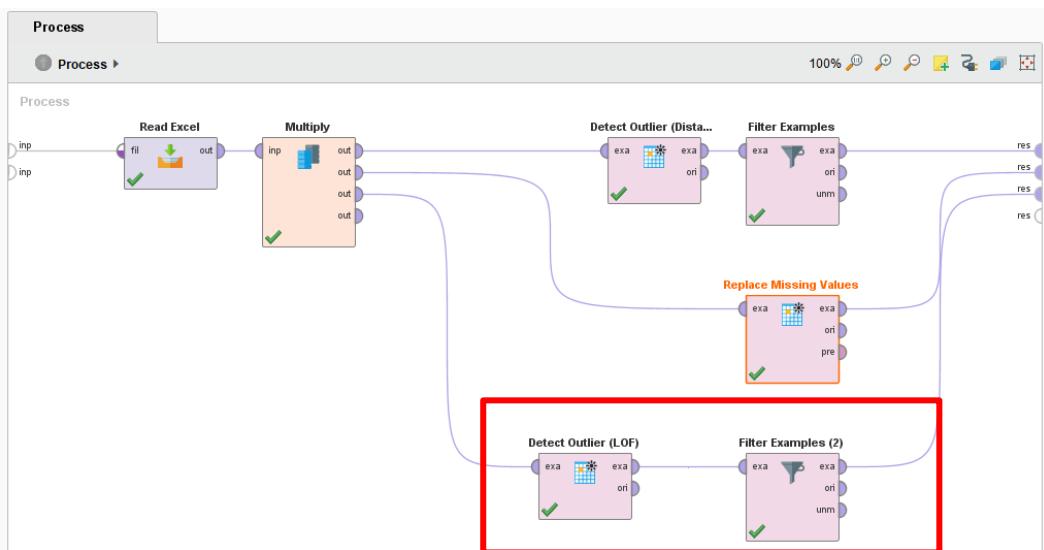
و نیز برای مشخص کردنتابع فاصله‌ای که برای محاسبه فاصله بین دو شی استفاده میشود، به کار می‌رود. و مقدار آن را نیز میتوانیم از بین مقادیر مختلف موجود در شکل 27 انتخاب کنیم.



شکل 27: LOF های متفاوت موجود در اپراتور distance function

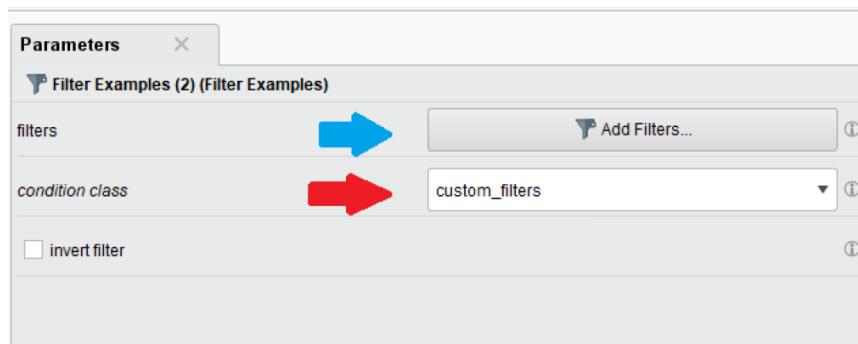
حال در قدم بعدی پس از تنظیم کردن اپراتور LOF میرویم سراغ گذاشتن فیلتر روی خروجی این اپراتور تا بتوانیم نتایج را فیلتر کنیم. برای این کار مانند قسمت A همین سوال یک اپراتور Filter Example را اضافه میکنیم تا نتایج را برای ما فیلتر کند. در ادامه ما اتصالات را نیز مانند شکل 28 متصل کرده و روی filter example کلیک کرده و به قسمت filter example آن می‌رویم.

• • •



شکل 28: نحوه ترسیم قسمت c از سوال 1

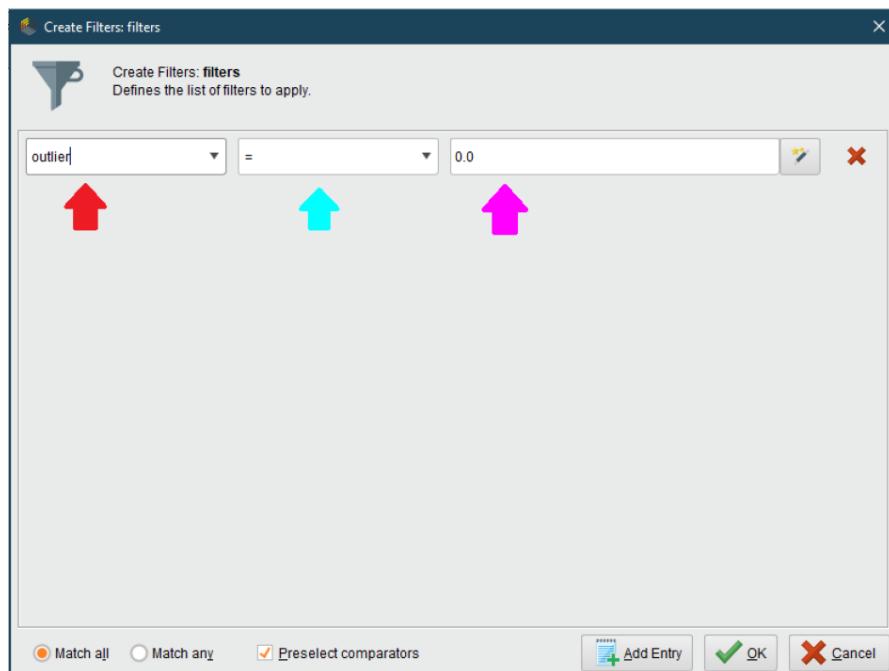
حال مطابق شکل 29 که با فلش قرمز رنگ مشخص شده است را برابر با `custom_filters` قرار می‌دهیم. و روی دکمه Add Filters که با رنگ آبی مشخص شده است کلیک می‌کنیم.



شکل 29: تنظیم کردن مقادیر Parameters در اپراتور فیلتر کردن

پس از کلیک روی اضافه کردن فیلتر صفحه جدیدی باز می‌شود که در شکل 30 آن را مشاهده می‌کنید.

• • •



شکل 30: اضافه کردن فیلتر جدید در اپراتور Filter Example

حال مطابق شکل 30 یک فیلتر روی صفت outlier (که با رنگ قرمز مشخص شده است) قرار می‌دهیم و ما می‌خواهیم آن‌ها را انتخاب کنیم که برابر با مشخص شده با رنگ آبی) صفر (مشخص شده با رنگ صورتی) هستند.

حال با انجام موارد بالا و رسم کردن اتصالات مانند شکل 28 اقدام به اجرای برنامه می‌کنیم و نتایجی مشابه آن‌چه در شکل 31 می‌بینید را مشاهده خواهید کرد.

	Row No.	outlier	Age (age in ...)	sex	chest pain	blood press...	cholesterol	blood sugar	electrocardi...	heart rate	exercise ind...	depression	slope	ca
1	0	53	0	3	128	216	0	2	115	0	0	1	0	?
2	0	52	1	3	138	223	0	0	169	0	0	1	?	?
3	0	58	1	2	125	220	0	0	144	0	0.400	2	?	?
4	0	38	1	3	138	175	0	0	173	0	0	1	?	?
5	0	40	1	2	140	289	0	0	172	0	0	?	?	?
6	0	37	1	2	130	283	0	1	98	0	0	?	?	?
7	0	54	1	3	150	?	0	0	122	0	0	?	?	?
8	0	39	1	3	120	339	0	0	170	0	0	?	?	?
9	0	45	0	2	130	237	0	0	170	0	0	?	?	?
10	0	54	1	2	110	208	0	0	142	0	0	?	?	?
11	0	48	0	2	120	284	0	0	120	0	0	?	?	?
12	0	37	0	3	130	211	0	0	142	0	0	?	?	?
13	0	39	1	2	120	204	0	0	145	0	0	?	?	?
14	0	42	0	3	115	211	0	1	137	0	0	?	?	?
15	0	54	0	2	120	273	0	0	150	0	1.500	2	?	?
16	0	43	0	2	120	201	0	0	165	0	0	?	?	?
17	0	43	0	1	100	223	0	0	142	0	0	?	?	?
18	0	44	1	2	120	184	0	0	142	0	1	2	?	?
19	0	49	0	2	124	201	0	0	164	0	0	?	?	?
20	0	40	1	3	130	215	0	0	138	0	0	?	?	?
21	0	36	1	3	130	209	0	0	178	0	0	?	?	?
22	0	53	1	4	124	260	0	1	112	1	3	2	?	?
23	0	52	1	2	120	284	0	0	118	0	0	?	?	?
24	0	53	0	2	113	468	?	0	127	0	0	?	?	?

شکل 31: نتیجه اجرای قسمت c از سوال 1

و همین طور که در شکل 31 می‌بینید مقدار outlier همه برابر با صفر است. حال برای این که مطمئن شویم نمایش Statistics آن را هم می‌بینیم:

• • •

Name	Type	Missing	Statistics			
			Min	Max	Average	
outlier	Real	0	0	0	0	
Age (age in year)	Integer	0	Min 28	Max 66	Average 47.860	
sex	Integer	0	Min 0	Max 1	Average 0.726	
chest pain	Integer	0	Min 1	Max 4	Average 2.983	
blood pressure	Integer	1	Min 92	Max 200	Average 132.503	
cholesterol	Integer	23	Min 85	Max 603	Average 250.141	
blood sugar	Integer	8	Min 0	Max 1	Average 0.076	
electrocardiographic	Integer	1	Min 0	Max 2	Average 0.228	
heart rate	Integer	1	Min 82	Max 190	Average 139.487	
exercise induced	Integer	1	Min 0	Max 1	Average 0.302	

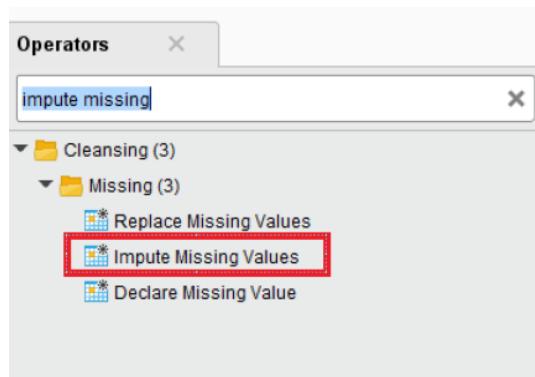
شکل 32: مشاهده نتایج نوع statistics سوال 1 قسمت c

همان طور که در شکل 32 مشاهده می‌کنید و ما نیز مشخص کردیم، همه outlier‌ها برابر با صفر است. حالا اگر بخواهید می‌توانید فیلتر را مقدار دیگری نیز تنظیم کنید که خب ما روی صفر تنظیم کردیم. حتی میتوانید به آن بازه هم بدهید، مثلاً آن‌هایی که outlier آن‌ها از 1 کوچکتر است و ...

: d مورد

در این قسمت از ما خواسته شده است که نویزها را با استفاده از روش K-NN و با دسته‌بند K-NN با معیار سنجش 10 نزدیکترین همسایگی پیش‌بینی کنیم.

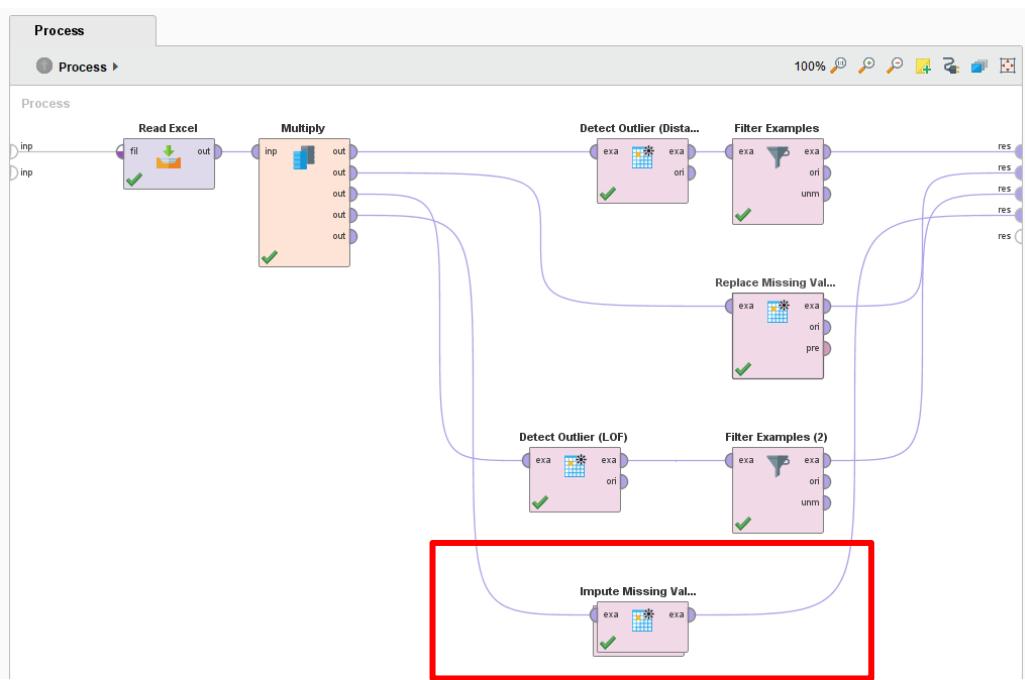
برای این کار از قسمت اپراتورها ما برای impute missing جستجو می‌کنیم و ازین نتایج آن موردنظر ما هست که مورد نظر ما هست را همان‌طور که در شکل 33 مشخص شده است را به داخل Drag & Drop Process موجود می‌کنیم.



شکل 33: جستجو در لیست اپراتورها برای اپراتور Impute Missing Value

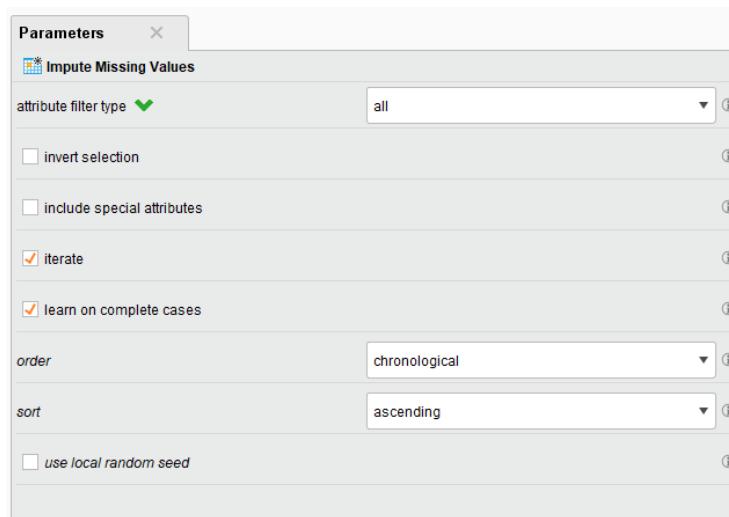
حال که مورد نظر را به Process خود اضافه کردیم اتصالات را مطابق شکل 34 برقرار می‌کنیم.

• • •



شکل 34: رسم اتصالات مورد نظر در قسمت D سوال 1

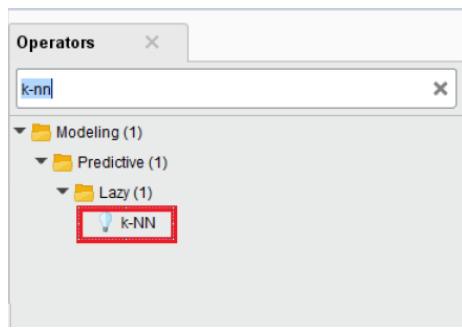
حال که رسم ما نیز تمام شد میرویم سراغ قسمت Parameters ان و تنظیمات رو مطابق با شکل 35 اعمال میکنیم.



شکل 35: قسمت Impute Missing Value مربوط به اپراتور Parameters

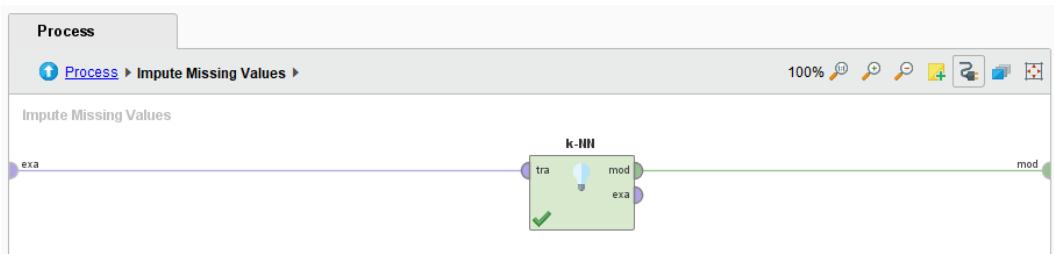
توضیح خاصی در رابطه با شکل 35 نیست که نیاز باشد بدهم، برای همین مستقیم به سراغ قسمت بعدی می‌رویم، برای تنظیم دسته‌بند K-NN کافی است روی اپراتور Impute Missing Value دبل کلیک کرده تا وارد آن شویم، وقتی وارد آن شدیم، حالا در قسمت اپراتورها برای k-NN جستجو میکنیم.

• • •



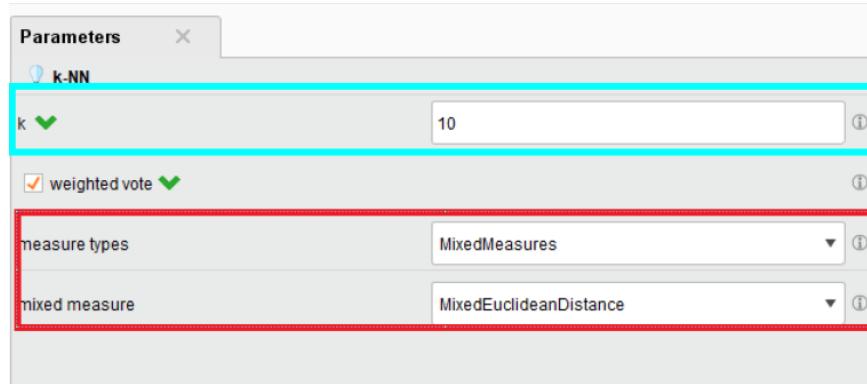
شکل 36: جستجو برای اپراتور k-nn

حال که اپراتور را پیدا کردیم آن را Drag & Drop داخل Process impute missing value می‌آوریم. و مانند شکل 37 اتصالات آن را برقرار می‌کنیم.



شکل 37: اضافه کردن اپراتور k-nn به داخل impute missing value Process

حال می‌رویم سراغ تنظیمات مربوط به k-nn که باشد با کلیک روی اپراتور k-nn و رفتن به قسمت Parameters آن این تنظیمات را انجام دهیم.



شکل 38: تنظیمات قسمت parameters اپراتور k-nn

مقدار k که در واقع همان مقدار سنجش نزدیکترین همسایگی گفته شده در صورت سوال است باید مقدار آن را 10 بدهیم. اما برای توضیح بیشتر k تعداد ما هایی هست که به ما از همه نزدیک‌تر هستند. و مقدار آن میتواند از 1 تا بی‌نهاین باشد و به طور پیش‌فرض برابر با 5 است اما سوال از ما خواسته که برابر با 10 قرار دهیم. در RapidMiner و در قسمت بعدی هم Measure types هست که نوع اندازه‌گیری فاصله را بیان می‌کند، که با چه معیاری این سنجش را انجام دهیم.

• • •



شکل 39: معیارهای مختلف برای سنجش فاصله در k-nn

نوع های مختلف برای اندازه گیری فاصله هست که بر اساس نوع متغیرها که Nominal یا Numerical هستند آن ها را انتخاب می کند. اما ما چون متغیر هامون از هر دو نوع هست، نوع آن را از نوع Mixed قرار می دهم. و متند اندازه گیری آن هم در این حالت فقط نوع mixed را دارد. اما در حالت های دیگر نوع ها مختلف دیگر مانند فاصله اقلیدسی و ... را دارد که در اینجا نیست.

پس از انجام این تنظیمات کار طراحی تمام است و به سراغ اجرا میرویم. و نتایج مطابق زیر هستند:

The screenshot shows the 'ExampleSet (Impute Missing Values)' tab selected in the top navigation bar. The main area displays a table with 597 rows of data, each containing 14 columns corresponding to the attributes: Row No., Age, sex, chest pain, blood press..., cholesterol, blood sugar, electrocardiogram, heart rate, exercise ind., depression, slope, ca, and thal. The data includes various numerical values and categorical labels (e.g., sex: 0 or 1, chest pain: 1 or 3).

شکل 40: نتایج قسمت d از سوال 1

حال اگر تحلیل statistics آن را هم بینیم میبینیم که مقدار missing ای دیگر وجود ندارد:

[Type the document title]

• • •

Name	Type	Missing	Min	Max	Average
Age (age in year)	Integer	0	28	77	51.183
sex	Integer	0	0	1	0.702
chest pain	Integer	0	1	4	3.072
blood pressure	Integer	0	92	200	132.125
cholesterol	Integer	0	85	603	248.520
blood sugar	Integer	0	0	1	0.110
electrocardiographic	Integer	0	0	2	0.612
heart rate	Integer	0	71	202	144.466
exercise induced	Integer	0	0	1	0.315
depression	Real	0	0	6.200	0.816
slope	Integer	0	1	3	1.661
ca	Integer	0	0	9	0.693
thal	Integer	0	3	7	4.932

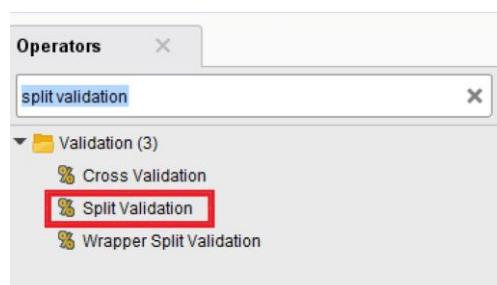
شکل 41: تحلیل statistics قسمت d از سوال 1

: e مورد

در این سوال از ما خواسته است که خروجی قسمتهای d, a, b, c, d را با استفاده از روش split validation و با 80٪ داده آموزشی و روش K-NN به طوری که هر نمونه با 8تا از نزدیکترین همسایگی‌هایش سنجیده شود، مدلی را ایجاد کنیم و معیارهای کارایی accuracy, recall, precision را در خروجی نمایش دهیم.

❖ راه حل اول: در این راه برای هر کدام از قسمتهای d, a, b, c, d به صورت جداگانه برای آنها split validation را رسم می‌کنیم و معیارهای کارایی را اندازه گیری می‌کنیم.

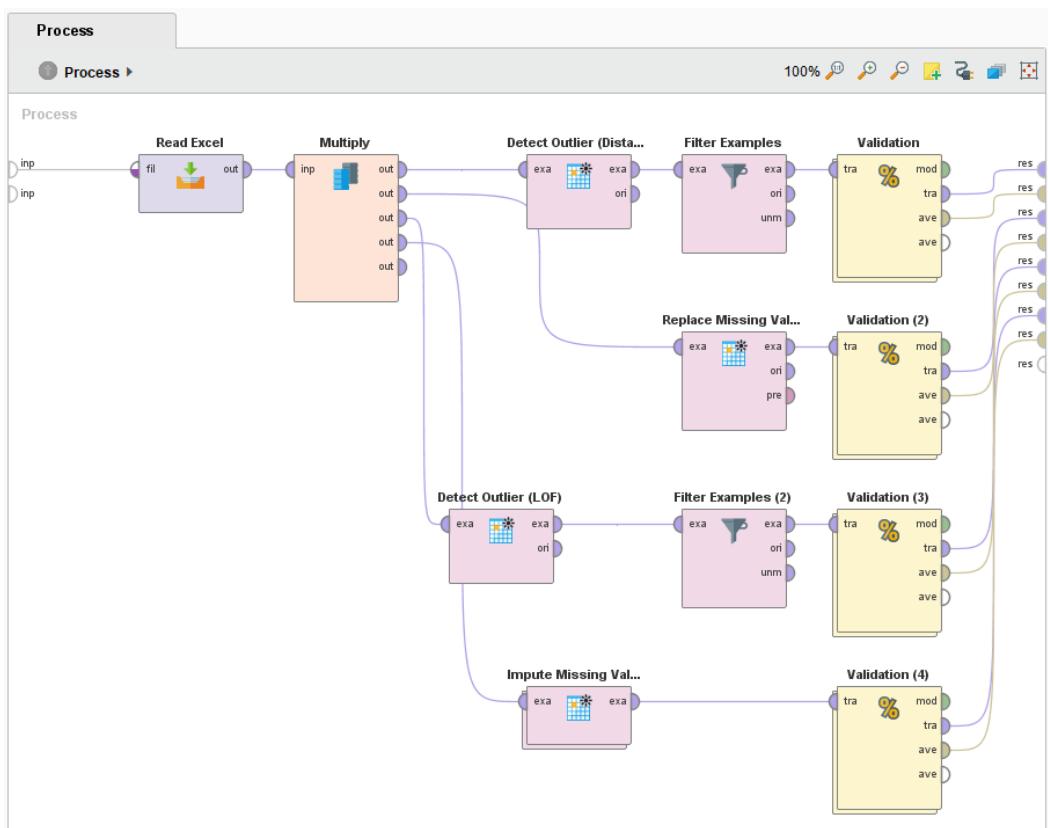
برای این کار اول از همه در قسمت جستجوی اپراتورها برای split validation جستجو می‌کنیم.



شکل 42: جستجوی operator split validation در قسمت operatorها

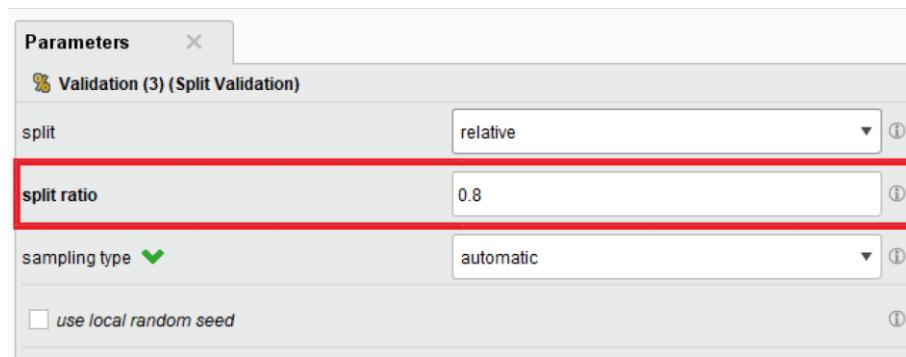
پس از پیدا کردن اپراتور مورد نظر مانند شکل 42 حال با استفاده از Drag & Drop 4 تا آن را داخل محیط Process خود می‌آوریم برای هر یک از قسمتهای d, a, b, c, d تا خروجی هر یک را به این مورد وصل کنیم. و اتصالات را مانند شکل 43 برقرار می‌کنیم.

• • •



شکل 43: طراحی مربوط به سوال 1 قسمت e

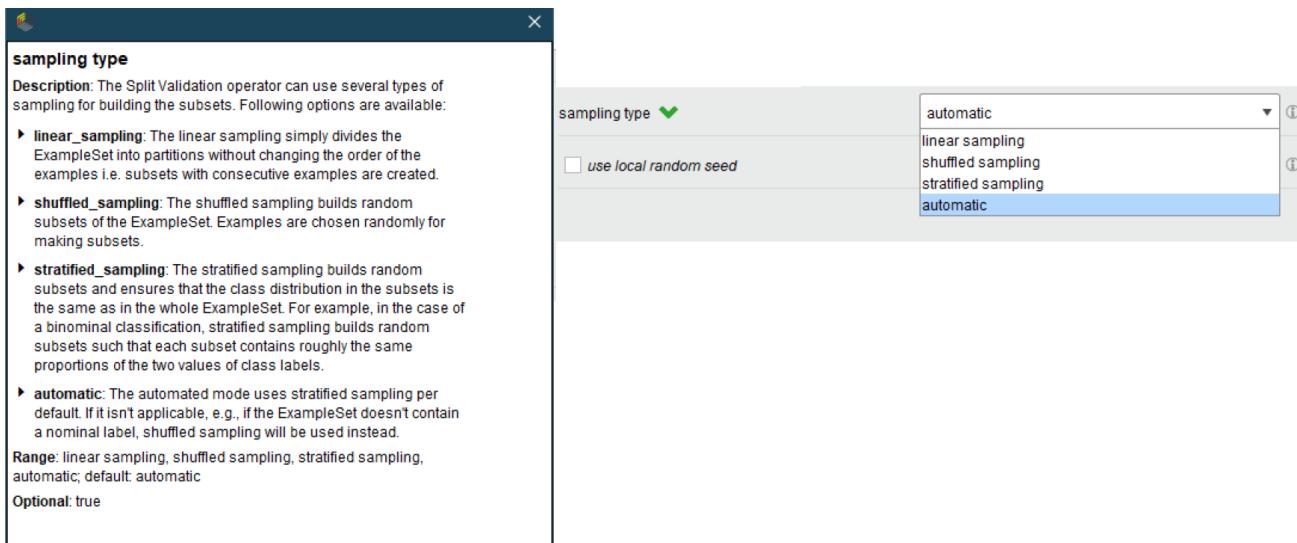
حال که رسم کامل شد ابتدا به سراغ تنظیم پارامترهای مربوط به Split Validation می‌رویم. که آن را مطابق با صورت سوال که گفته بود با ۸۰٪ داده آموزشی باشد تنظیم می‌کنیم.



شکل 44: تنظیم قسمت Split Validation برای اپراتور Parameters

مقدار split ratio که مقداری بین ۰ و ۱ دارد و همان مقدار داده‌های آموزشی است و مقدار ۱ یعنی از همه داده‌ها برای آموزش استفاده کند و ۰ یعنی استفاده نکند. چون در صورت سوال گفته بود ۸۰٪ داده‌ها آموزشی هستند، پس ما هم مقدار ۰.۸ را برای این مقدار قرار می‌دهیم.

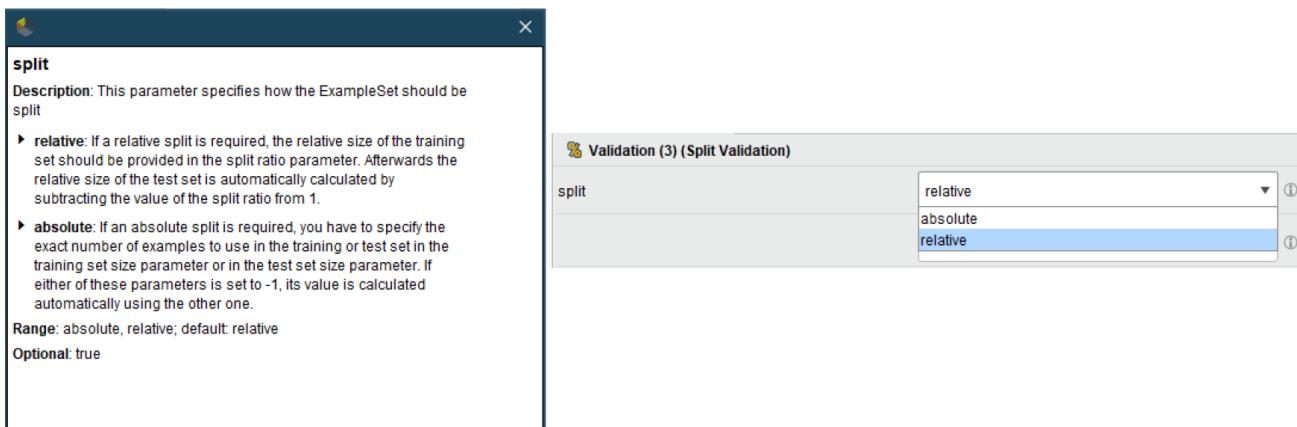
• • •



شکل 45: مقادیر مختلف برای sampling type به همراه توضیحات و تفاوت آن‌ها

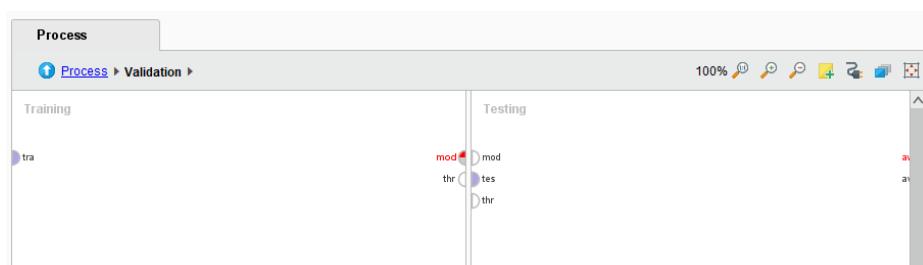
با توجه به موارد شکل 45 و توضیحات آن‌ها من مقدار را روی automatic قرار دادم تا برای همه نوع‌ها داده‌ای مناسب باشد و کار کند. مقایسه کامل بین نوع‌ها مختلف در شکل 45 موجود می‌باشد.

مقدار split در شکل 44 نیز که نحوه تقسیم کردن example set را مشخص می‌کند را برابر با relative قرار دادم تا به صورت اتوماتیک اندازه مجموعه را محاسبه کرده و تقسیم را باتوجه به آن انجام دهد. مقایسه و توضیحات دقیق‌تر را در شکل 46 می‌توانید مشاهده کنید.



شکل 46: مقادیر مختلف برای split به همراه توضیحات و مقایسه آن‌ها

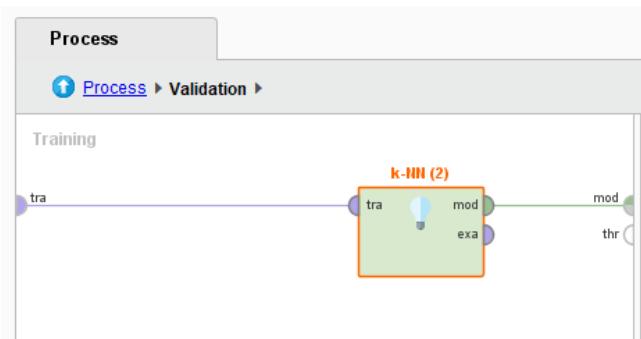
پس از انجام این تنظیمات حال به سراغ تنظیمات داخلی split validation می‌رویم، برای این کار کافی است روی اپراتور split validation دو بار کلیک کرده تا وارد آن شوید. پس از آن که وارد آن شدید با محیطی مطابق شکل 47 مواجه می‌شوید.



شکل 47: محیط داخل split validation

همان طور که در شکل 47 مشاهده می‌کنید split validation از دو قسمت Testing و Training تشکیل شده است. که قسمتی برای آموزش داده‌ها است و قسمتی نیز برای سنجش آن‌ها می‌باشد.

حال در ادامه ابتدا چون در سوال گفته است با روش k-nn آموزش داده‌ها آمورش بینند، پس ما این اپراتور را به درون فرآیند موجود اضافه می‌کنیم.



شکل 48: اضافه کردن اپراتور k-nn به درون split validation

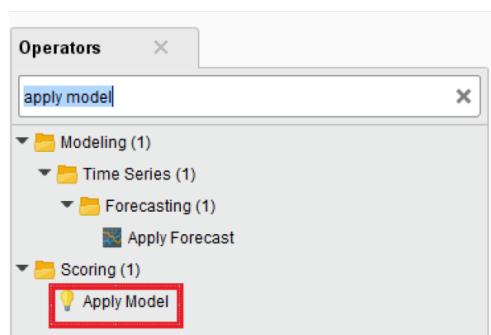
حال که این مورد را اضافه کردیم پارامترهای آن را مطابق صورت سوال تنظیم می‌کنیم، که چون گفته بود داده‌های آموزشی 80% هستند مطابق شکل 49 این کار را انجام می‌دهیم.



شکل 49: تنظیم پارامترهای اپراتور k-nn داخل split validation

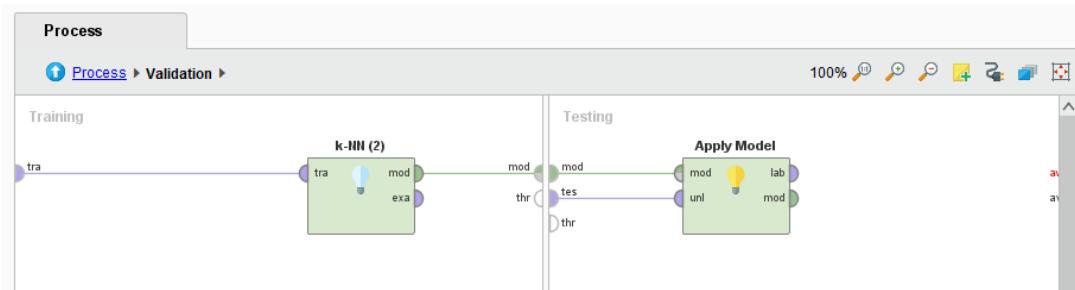
حال که این تنظیمات را هم انجام دادیم باید دو اپراتور دیگر را نیز برای قسمت Testing اضافه کنیم. که این دو اپراتور عبارت‌اند از : Performance

ابتدا به اضافه کردن اپراتور apply model می‌پردازیم که کافی است اسم آن را در قسمت جستجوی اپراتورها جستجو کنید (مطابق شکل 50) و نتیجه مشخص شده را در داخل فرآیند خود Drag & Drop کنید.



شکل 50 : جستجوی برای اپراتور Apply Model

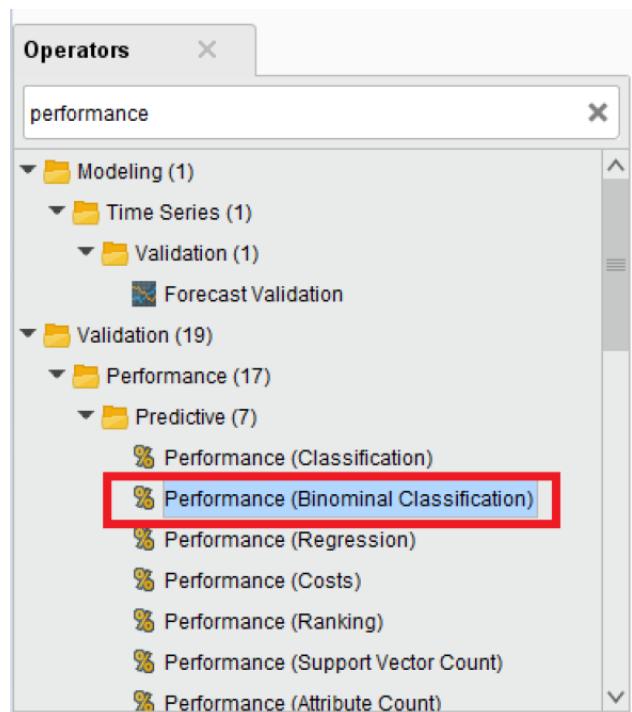
حال که این مورد را نیز مطابق شکل 51 اضافه کردیم به ادامه کار میپردازیم، و بعد از اضافه کردن این اپراتور این نکته را هم بگوییم که نیاز به تغییری در پارامترهای آن نیست و آن‌ها را به حالت پیش‌فرش قرار می‌دهیم.



شکل 51: اضافه کردن اپراتور به داخل Apply Model

یک توضیحی هم در مورد Apply Model بدhem که این اپراتور می‌آید و یک مدل را روی Example set ما اعمال می‌کند تا بتوانیم آن را برای قسمت Performance استفاده کنیم.

بعد از انجام این تغییرات حال به سراغ اضافه کردن اپراتور Performance می‌رویم، برای این کار مانند قسمت‌های قبلی در قسمت جستجوی اپراتورها عبارت Performance را جستجو می‌کنیم، و مورد مشخص شده در شکل 52 را به فرآیند خود اضافه می‌کنیم.



شکل 52: جستجوی Performance در قسمت جستجوی اپراتورها

بعد از جستجو کردن فقط ما چندین نوع مختلف از این اپراتور را می‌بینیم که ما نوع Binomial Classification آن را انتخاب می‌کنیم، چون می‌خواهیم براساس متغیر C کارایی را بسنجیم که نیز از نوع Binomial هست.

در قسمت تنظیم پارامترها نیز باید سه مقدار گفته شده در صورت سوال را مشخص کیم، که در شکل 53 این موارد را مشخص کرده‌ام.

• • •



شکل 53: تنظیمات پارامترهای مربوط به اپراتور Performance

کار ما در قسمت طراحی تمام است فقط نکته‌ای که باقی می‌ماند این است که باید متغیر C را نیز از نوع Label تعییف می‌کردیم امکان دارد به صورت پیش‌فرض این طور نباشد برای این کار را انجام دهیم چون اپراتور Performance حول C دارد اندازه‌گیری می‌کند برای همین باید این کار را انجام دهیم برای این کار هم در ابتدای توضیحات این گزارش آمده است.

حال در ادامه کافی است که تمامی مراحل توضیح داده شده در بالا را برای تمامی Split validation های موجود در شکل 43 انجام دهید. پس از انجام این کار برنامه آماده است و میتوانید آن را اجرا کنید.

Criterion	accuracy	precision	recall
	accuracy: 64.10%		
		true 0	true 1
pred. 0	59	32	64.84%
pred. 1	10	16	61.54%
class recall	85.51%	33.33%	

شکل 54: نتیجه به دست آمده برای a سوال 1 قسمت accuracy, recall, precision

Criterion	accuracy	precision	recall
	accuracy: 65.55%		
		true 0	true 1
pred. 0	52	23	69.33%
pred. 1	18	26	59.09%
class recall	74.29%	53.06%	

شکل 55: نتیجه به دست آمده برای b سوال 1 قسمت accuracy, recall, precision

[Type the document title]

• • •

Criterion
accuracy
precision
recall

accuracy: 64.41%

	true 0	true 1	class precision
pred. 0	36	19	65.45%
pred. 1	2	2	50.00%
class recall	94.74%	9.52%	

شکل 56: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت c

Criterion
accuracy
precision
recall

accuracy: 61.34%

	true 0	true 1	class precision
pred. 0	49	25	66.22%
pred. 1	21	24	53.33%
class recall	70.00%	48.98%	

شکل 57: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d

PerformanceVector

```
PerformanceVector:  
accuracy: 64.10%  
ConfusionMatrix:  
True: 0 1  
0: 59 32  
1: 10 16  
precision: 61.54% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 59 32  
1: 10 16  
recall: 33.33% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 59 32  
1: 10 16
```

شکل 58: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت a با استفاده از Split Validation

PerformanceVector

```
PerformanceVector:  
accuracy: 65.55%  
ConfusionMatrix:  
True: 0 1  
0: 52 23  
1: 18 26  
precision: 59.09% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 52 23  
1: 18 26  
recall: 53.06% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 52 23  
1: 18 26
```

شکل 59: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت b با استفاده از Split Validation

● ● ●

PerformanceVector

```

PerformanceVector:
accuracy: 64.41%
ConfusionMatrix:
True: 0 1
0: 36 19
1: 2 2
precision: 50.00% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 36 19
1: 2 2
recall: 9.52% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 36 19
1: 2 2

```

شکل 60 : نتیجه به دست آمده برای سوال 1 قسمت c با استفاده از accuracy, recall, precision

PerformanceVector

```

PerformanceVector:
accuracy: 61.34%
ConfusionMatrix:
True: 0 1
0: 49 25
1: 21 24
precision: 53.33% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 49 25
1: 21 24
recall: 48.98% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 49 25
1: 21 24

```

شکل 61 : نتیجه به دست آمده برای سوال 1 قسمت d با استفاده از accuracy, recall, precision

		Open in Turbo Prep Auto Model												Filter (299 / 299 examples): all	
		Row No.	c	outlier	Age (age in ...)	sex	chest pain	blood press...	cholesterol	blood sugar	electrocardi...	heart rate	exercise ind...	depression	slope
	Data	1	0	0	53	0	3	128	216	0	2	115	0	0	1
	Statistics	2	0	0	52	1	3	138	223	0	0	169	0	0	1
	Charts	3	0	0	58	1	2	125	220	0	0	144	0	0.400	2
	Advanced Charts	4	0	0	38	1	3	138	175	0	0	173	0	0	1
	Annotations	5	0	0	40	1	2	140	289	0	0	172	0	0	?
		6	0	0	37	1	2	130	283	0	1	98	0	0	?
		7	0	0	54	1	3	150	?	0	0	122	0	0	?
		8	0	0	39	1	3	120	339	0	0	170	0	0	?
		9	0	0	45	0	2	130	237	0	0	170	0	0	?
		10	0	0	54	1	2	110	208	0	0	142	0	0	?
		11	0	0	48	0	2	120	284	0	0	120	0	0	?
		12	0	0	37	0	3	130	211	0	0	142	0	0	?
		13	0	0	39	1	2	120	204	0	0	145	0	0	?
		14	0	0	42	0	3	115	211	0	1	137	0	0	?
		15	0	0	54	0	2	120	273	0	0	150	0	1.500	2
		16	0	0	43	0	2	120	201	0	0	165	0	0	?
		17	0	0	43	0	1	100	223	0	0	142	0	0	?
		18	0	0	44	1	2	120	184	0	0	142	0	1	2
		19	0	0	49	0	2	124	201	0	0	164	0	0	?
		20	0	0	40	1	3	130	215	0	0	138	0	0	?
		21	0	0	36	1	3	130	209	0	0	178	0	0	?
		22	0	0	53	1	4	124	260	0	1	112	1	3	2
		23	0	0	52	1	2	120	284	0	0	118	0	0	?

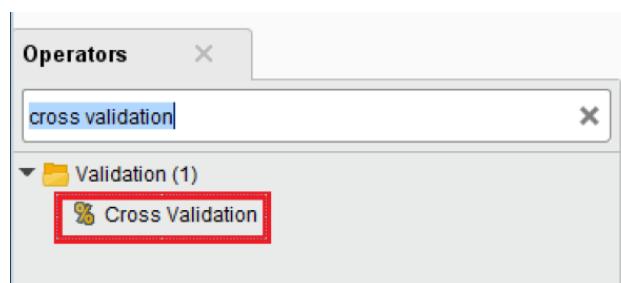
• • •

	Name	Type	Missing	Statistics		Filter (15 / 15 attributes)	Search for Attributes	Print
	c	Binomial	0	Least 1 (240) Most 0 (347) Values 0 (347), 1 (240)				
	outlier	Binomial	0	Least true (0) Most false (587) Values false (587), true (0)				
	Age (age in year)	Integer	0	Min 28 Max 77 Average 51.060				
	sex	Integer	0	Min 0 Max 1 Average 0.707				
	chest pain	Integer	0	Min 1 Max 4 Average 3.066				
	blood pressure	Integer	1	Min 92 Max 200 Average 131.956				
	cholesterol	Integer	23	Min 85 Max 603 Average 247.660				
	blood sugar	Integer	8	Min 0 Max 1 Average 0.107				
	electrocardiographic	Integer	1	Min 0 Max 2 Average 0.597				
	heart rate	Integer	1	Min 71 Max 202 Average 144.317				
	exercise induced	Integer	1	Min 0 Max 1 Average 0.317				
	depression	Real	0	Min 0 Max 6.200 Average 0.809				

قسمت f:

در این قسمت همان موارد سوال قبل را از ما خواسته است فقط با این تفاوت که این دفعه به جای Split Validation خواسته است که از Validation استفاده کنیم.

این مورد را در قسمت جستجوی اپراتورها مطابق شکل زیر جستجو میکنیم:

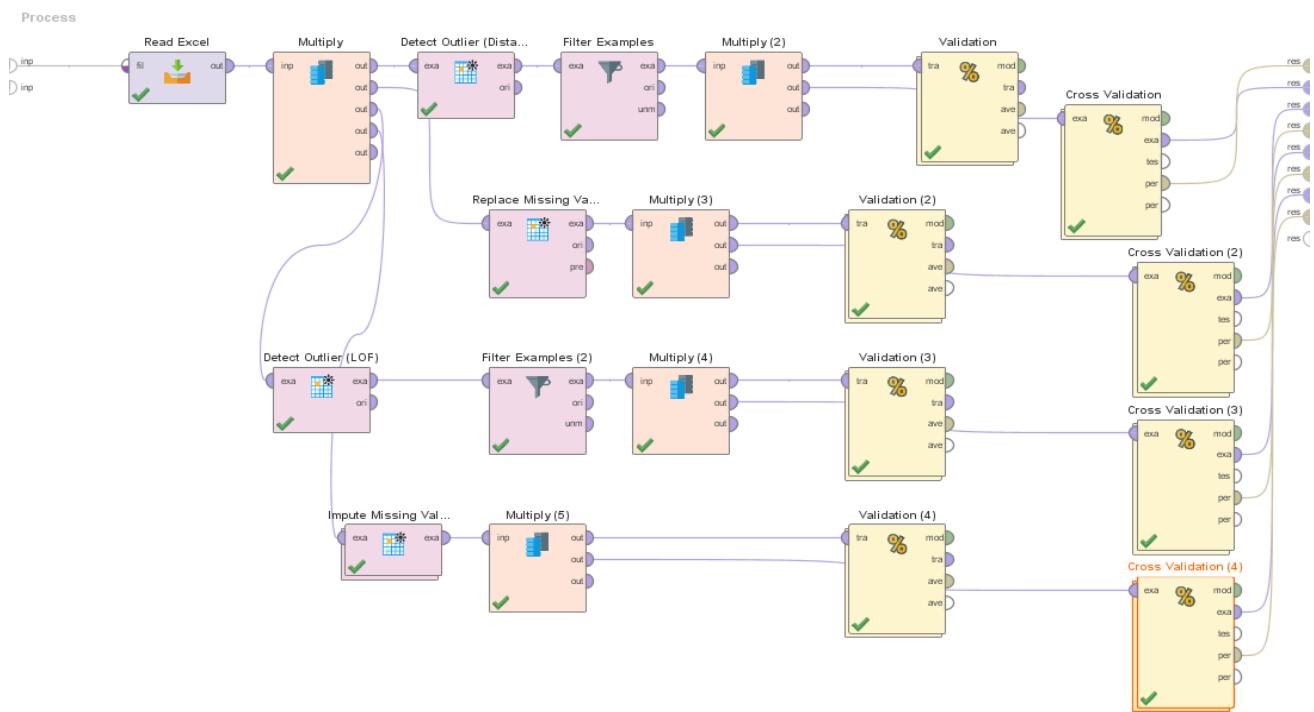


شکل 62 : جستجوی اپراتور Cross Validation

بعد از جستجو آن را Drag & Drop چهارتای آن را به فرآیندی که در آن هستیم می‌وریم. و به شکل زیر اتصالات را برقرار میکنیم: (برای این که فضای کاری شلوغ نشود من اتصالات قسمت قبل رو به خروجی حذف کدم.)

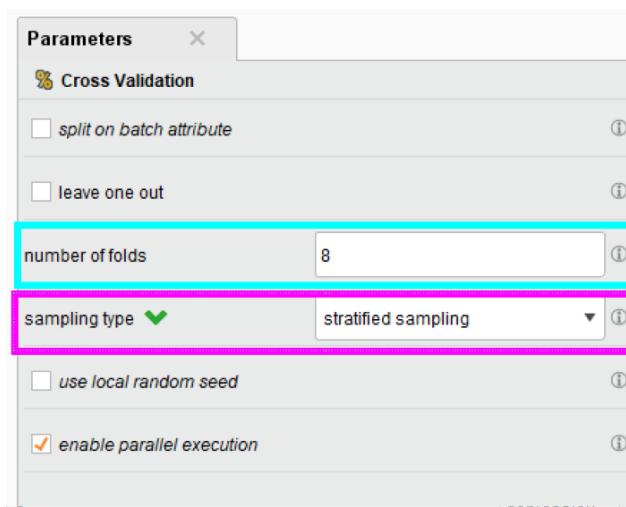
نتیجه مطابق شکل زیر می‌باشد:

• • •



شکل 63: پس از رسم تمامی اتصالات مربوط به سوال 1 قسمت f

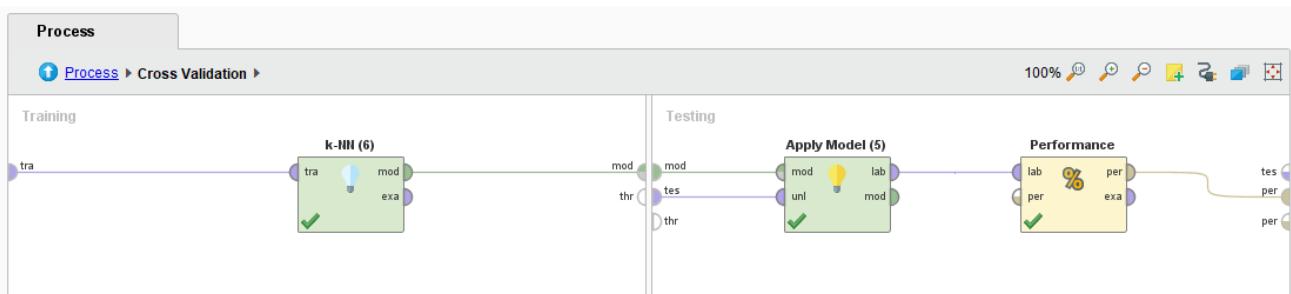
حال باید پارامترهای مربوط به تمامی این Cross Validation های اضافه شده را تنظیم کنیم. که برای این کار روی آنها کلیک کرده و به قسمت پارامتر آنها رفته و مطابق شکل زیر به ان مقدار می دهیم. چون در صورت سوال گفته است که Number of folds : 8 پس ما نیز این مقدار را می دهیم، و نیز چون صورت سوال گفته است که sampling type: stratified sampling پس ما نیز این کار را انجام می دهیم.



شکل 64: تنظیم پارامترهای مربوط به Cross Validation

پس از تنظیم کردن موارد فوق حال به سراغ داخل Cross Validation ها می رویم و مقدار داخلی آنها را دقیقا مشابه با قسمت قبلی قرار می دهیم.

• • •



شکل 65: اتصالات داخل Cross Validation قسمت f سوال 1

تنظیمات هر یک از این اپراتورها دقیقا مشابه با قسمت قبلی یعنی e است، برای همین در اینجا دیگر بیان نمیکنم.

حال به سراغ اجرای آن می رویم که مطابق زیر نتایج آن به دست می آیند.

Criterion	accuracy	precision	recall
	Table View	Plot View	
accuracy: 65.74% +/- 5.91% (micro average: 65.76%)			
pred. 0	300	154	66.08%
pred. 1	47	86	64.66%
class recall	86.46%	35.83%	

شکل 66 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت a با استفاده از Cross Validation

Criterion	accuracy	precision	recall
	Table View	Plot View	
accuracy: 63.98% +/- 4.18% (micro average: 63.99%)			
pred. 0	264	127	67.52%
pred. 1	88	118	57.28%
class recall	75.00%	48.16%	

شکل 67 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت b با استفاده از Cross Validation

Criterion	accuracy	precision	recall
	Table View	Plot View	
accuracy: 69.61% +/- 6.16% (micro average: 69.57%)			
pred. 0	182	81	69.20%
pred. 1	10	26	72.22%
class recall	94.79%	24.30%	

شکل 68 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت c با استفاده از Cross Validation

Criterion	accuracy	precision	recall
	Table View	Plot View	
accuracy: 65.15% +/- 5.44% (micro average: 65.16%)			
pred. 0	269	125	68.27%
pred. 1	83	120	59.11%
class recall	76.42%	48.98%	

شکل 69 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d با استفاده از Cross Validation

● ● ●

PerformanceVector

```

PerformanceVector:
accuracy: 65.74% +/- 5.91% (micro average: 65.76%)
ConfusionMatrix:
True: 0 1
0: 300 154
1: 47 86
precision: 65.94% +/- 16.51% (micro average: 64.66%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 300 154
1: 47 86
recall: 35.82% +/- 9.22% (micro average: 35.83%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 300 154
1: 47 86

```

شکل 70 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d با استفاده از Cross Validation

PerformanceVector

```

PerformanceVector:
accuracy: 63.98% +/- 4.18% (micro average: 63.99%)
ConfusionMatrix:
True: 0 1
0: 264 127
1: 88 118
precision: 58.31% +/- 7.83% (micro average: 57.28%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 264 127
1: 88 118
recall: 48.23% +/- 8.11% (micro average: 48.16%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 264 127
1: 88 118

```

شکل 71 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d با استفاده از Cross Validation

PerformanceVector

```

PerformanceVector:
accuracy: 69.61% +/- 6.16% (micro average: 69.57%)
ConfusionMatrix:
True: 0 1
0: 182 81
1: 10 26
precision: 72.23% +/- 20.03% (micro average: 72.22%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 182 81
1: 10 26
recall: 24.52% +/- 11.57% (micro average: 24.30%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 182 81
1: 10 26

```

شکل 72 : نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d با استفاده از Cross Validation

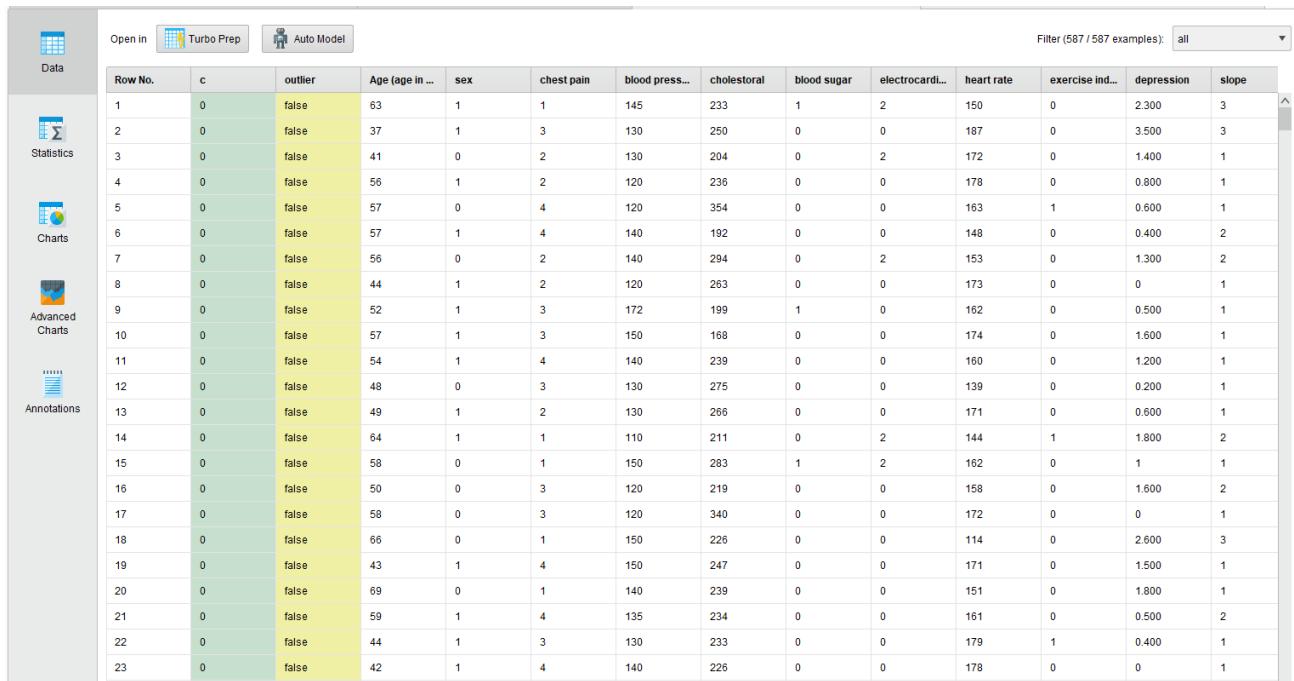
PerformanceVector

```

PerformanceVector:
accuracy: 65.15% +/- 5.44% (micro average: 65.16%)
ConfusionMatrix:
True: 0 1
0: 269 125
1: 83 120
precision: 58.28% +/- 8.12% (micro average: 59.11%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 269 125
1: 83 120
recall: 48.90% +/- 11.55% (micro average: 48.98%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 269 125
1: 83 120

```

شکل 73: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت d با استفاده از Cross Validation



Row No.	c	outlier	Age (age in ...)	sex	chest pain	blood press...	cholesterol	blood sugar	electrocardi...	heart rate	exercise ind...	depression	slope
1	0	false	63	1	1	145	233	1	2	150	0	2.300	3
2	0	false	37	1	3	130	250	0	0	187	0	3.500	3
3	0	false	41	0	2	130	204	0	2	172	0	1.400	1
4	0	false	56	1	2	120	236	0	0	178	0	0.800	1
5	0	false	57	0	4	120	354	0	0	163	1	0.600	1
6	0	false	57	1	4	140	192	0	0	148	0	0.400	2
7	0	false	56	0	2	140	294	0	2	153	0	1.300	2
8	0	false	44	1	2	120	263	0	0	173	0	0	1
9	0	false	52	1	3	172	199	1	0	162	0	0.500	1
10	0	false	57	1	3	150	168	0	0	174	0	1.600	1
11	0	false	54	1	4	140	239	0	0	160	0	1.200	1
12	0	false	48	0	3	130	275	0	0	139	0	0.200	1
13	0	false	49	1	2	130	266	0	0	171	0	0.600	1
14	0	false	64	1	1	110	211	0	2	144	1	1.800	2
15	0	false	58	0	1	150	283	1	2	162	0	1	1
16	0	false	50	0	3	120	219	0	0	158	0	1.600	2
17	0	false	58	0	3	120	340	0	0	172	0	0	1
18	0	false	66	0	1	150	226	0	0	114	0	2.600	3
19	0	false	43	1	4	150	247	0	0	171	0	1.500	1
20	0	false	69	0	1	140	239	0	0	151	0	1.800	1
21	0	false	59	1	4	135	234	0	0	161	0	0.500	2
22	0	false	44	1	3	130	233	0	0	179	1	0.400	1
23	0	false	42	1	4	140	226	0	0	178	0	0	1

شکل 74: خروجی داده‌ای مربوط به سوال 1 قسمت f

• • •

Name	Type	Missing	Statistics			Filter (15 / 15 attributes):	Search for Attributes	Print
Label c	Binominal	0	Least 1 (240)	Most 0 (347)	Values 0 (347), 1 (240)			
Outlier outlier	Binominal	0	Least true (0)	Most false (587)	Values false (587), true (0)			
Age (age in year)	Integer	0	Min 28	Max 77	Average 51.060			
sex	Integer	0	Min 0	Max 1	Average 0.707			
chest pain	Integer	0	Min 1	Max 4	Average 3.066			
blood pressure	Integer	1	Min 92	Max 200	Average 131.956			
cholesterol	Integer	23	Min 85	Max 603	Average 247.660			
blood sugar	Integer	8	Min 0	Max 1	Average 0.107			
electrocardiographic	Integer	1	Min 0	Max 2	Average 0.597			
heart rate	Integer	1	Min 71	Max 202	Average 144.317			
exercise induced	Integer	1	Min 0	Max 1	Average 0.317			
depression	Real	0	Min 0	Max 6.200	Average 0.809			

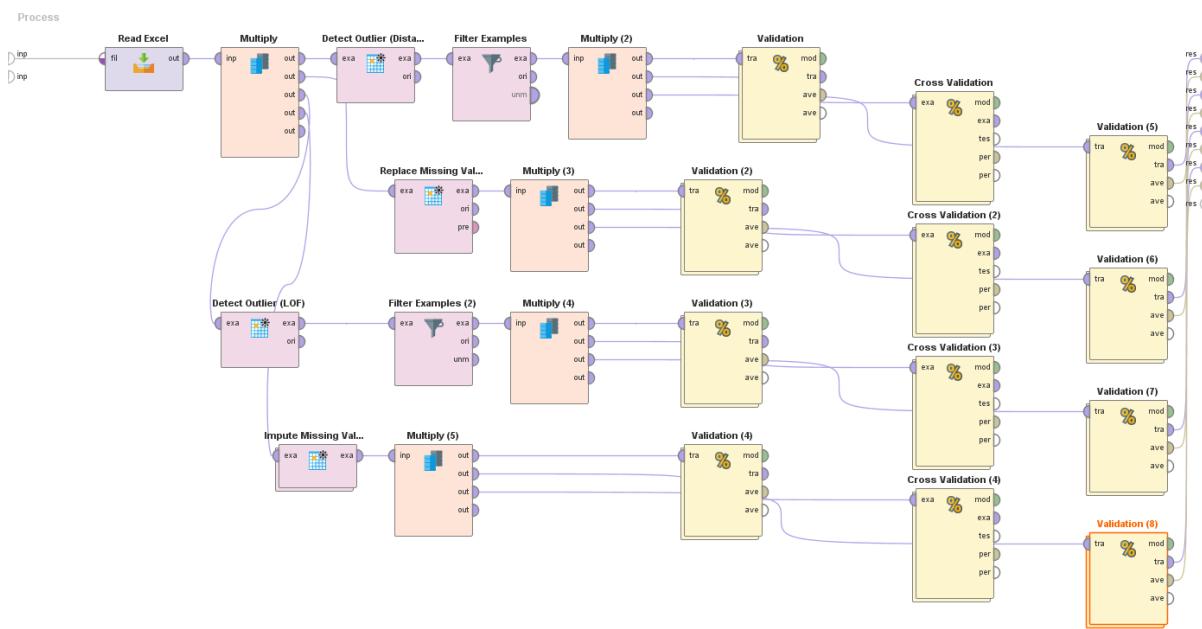
شکل 75 : خروجی Statistics مربوط به سوال 1

قسمت g:

در این قسمت از ما سوال خواسته‌هایش از ما دقیقاً مشابه قسمت e می‌باشد با این تفاوت که در این قسمت به جای استفاده از دسته‌بند K-NN از ما خواسته است که از دسته‌بند Decision tree استفاده کنیم.

برای این کار مانند همان قسمت 4 عدد split validation را به فرآیند خود اضافه می‌کنیم. سپس داخل آن‌ها را به جای K-nn از K-nn استفاده می‌کنیم. شکل نهایی رسم اتصالات مطابق شکل زیر می‌شود:

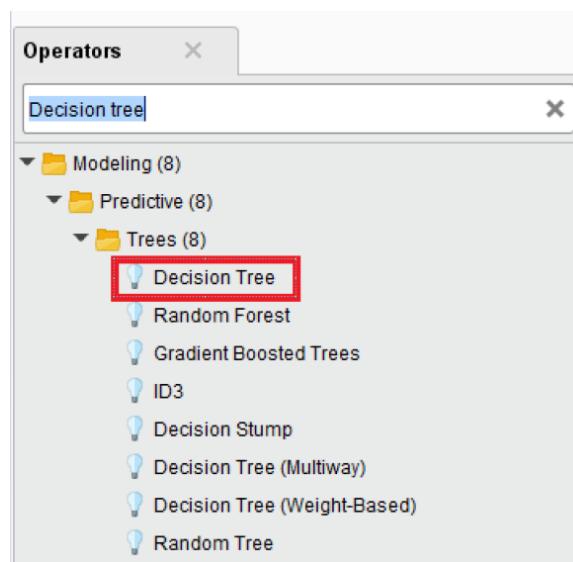
• • •



شکل 76: شکل نهایی طراحی و اتصالات قسمت g سوال 1

لازم به ذکر است که من برای جلوگیری از شلوغ شدن محیط کاری من اتصالات مراحل قبلی را قطع کردم.

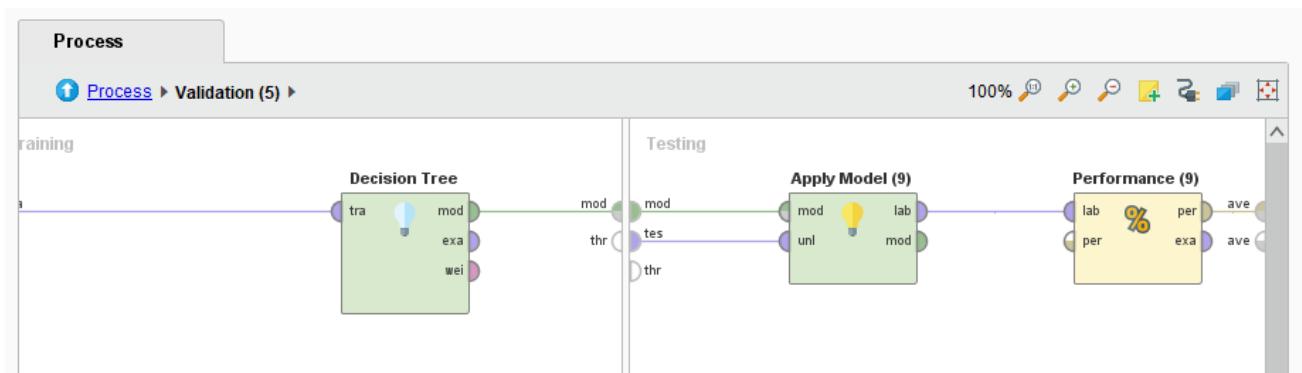
برای این کار در قسمت جستجوی اپراتورها عبارت Decision tree را تایپ کرده و آن را به فرآیند موجود اضافه میکنیم.



شکل 77: جستجو برای اپراتور Decision tree در قسمت اپراتورها

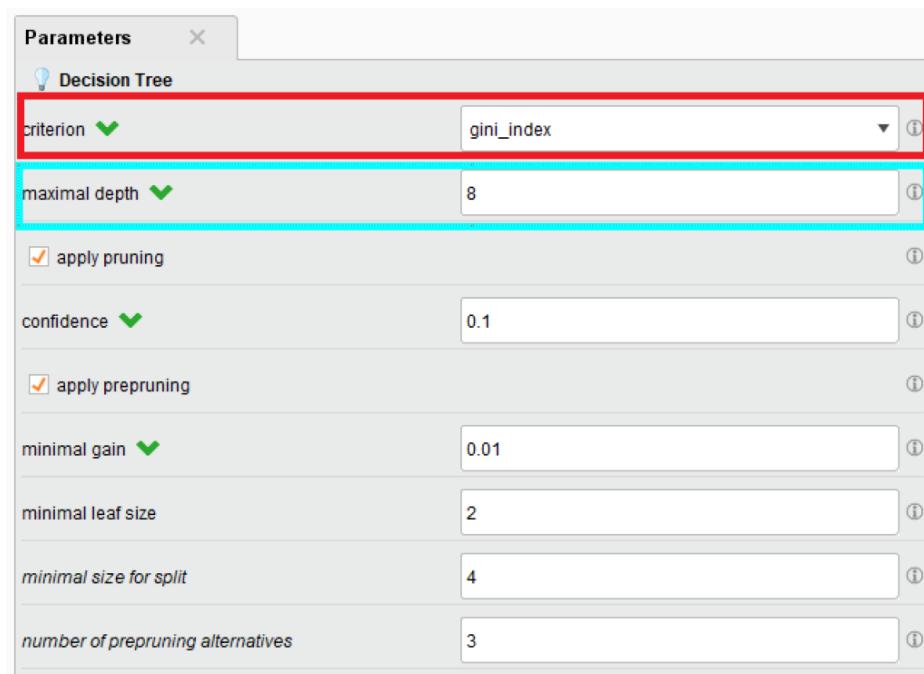
پس از این کار طراحی قسمت داخلی Split Validation مطابق شکل زیر می شود.

• • •



شکل 78: طراحی قسمت داخلی Split Validation

و تنظیمات پارامترهای Decision Tree را نیز طبق اطلاعات داده شده در صورت سوال انجام می‌دهیم.



شکل 79: تنظیمات قسمت Decision Tree در اپراتور Parameters

ما بقی قسمت‌ها نیز که در صورت سوال به آن‌ها اشاره‌ای نشده است را به حالت پیش‌فرض رها می‌کنیم.

حال برنامه طراحی شده را اجرا می‌کنیم:

Criterion	accuracy	precision	recall
	Table View	Plot View	
accuracy: 74.36%			
	true 0	true 1	class precision
pred. 0	54	15	78.26%
pred. 1	15	33	68.75%
class recall	78.26%	68.75%	

شکل 80: نتیجه به دست آمده برای accuracy, recall, precision قسمت 1 Split Validation با استفاده از a سوال Decision Tree

[Type the document title]

• • •

Criterion
accuracy
precision
recall

accuracy: 76.47%

	true 0	true 1	class precision
pred. 0	60	18	76.92%
pred. 1	10	31	75.61%
class recall	85.71%	63.27%	

شکل 81 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت b با استفاده از accuracy, recall, precision

Criterion
accuracy
precision
recall

accuracy: 71.19%

	true 0	true 1	class precision
pred. 0	32	11	74.42%
pred. 1	6	10	62.50%
class recall	84.21%	47.62%	

شکل 82 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت c با استفاده از accuracy, recall, precision

Criterion
accuracy
precision
recall

accuracy: 70.59%

	true 0	true 1	class precision
pred. 0	56	21	72.73%
pred. 1	14	28	66.67%
class recall	80.00%	57.14%	

شکل 83 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت d با استفاده از accuracy, recall, precision

PerformanceVector

```

PerformanceVector:
accuracy: 74.36%
ConfusionMatrix:
True: 0 1
0: 54 15
1: 15 33
precision: 68.75% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 54 15
1: 15 33
recall: 68.75% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 54 15
1: 15 33

```

شکل 84 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت a با استفاده از accuracy, recall, precision

PerformanceVector

```
PerformanceVector:  
accuracy: 76.47%  
ConfusionMatrix:  
True: 0 1  
0: 60 18  
1: 10 31  
precision: 75.61% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 60 18  
1: 10 31  
recall: 63.27% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 60 18  
1: 10 31
```

شکل 85 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت b با استفاده از accuracy, recall, precision

PerformanceVector

```
PerformanceVector:  
accuracy: 71.19%  
ConfusionMatrix:  
True: 0 1  
0: 32 11  
1: 6 10  
precision: 62.50% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 32 11  
1: 6 10  
recall: 47.62% (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 32 11  
1: 6 10
```

شکل 86 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت c با استفاده از accuracy, recall, precision

PerformanceVector

```

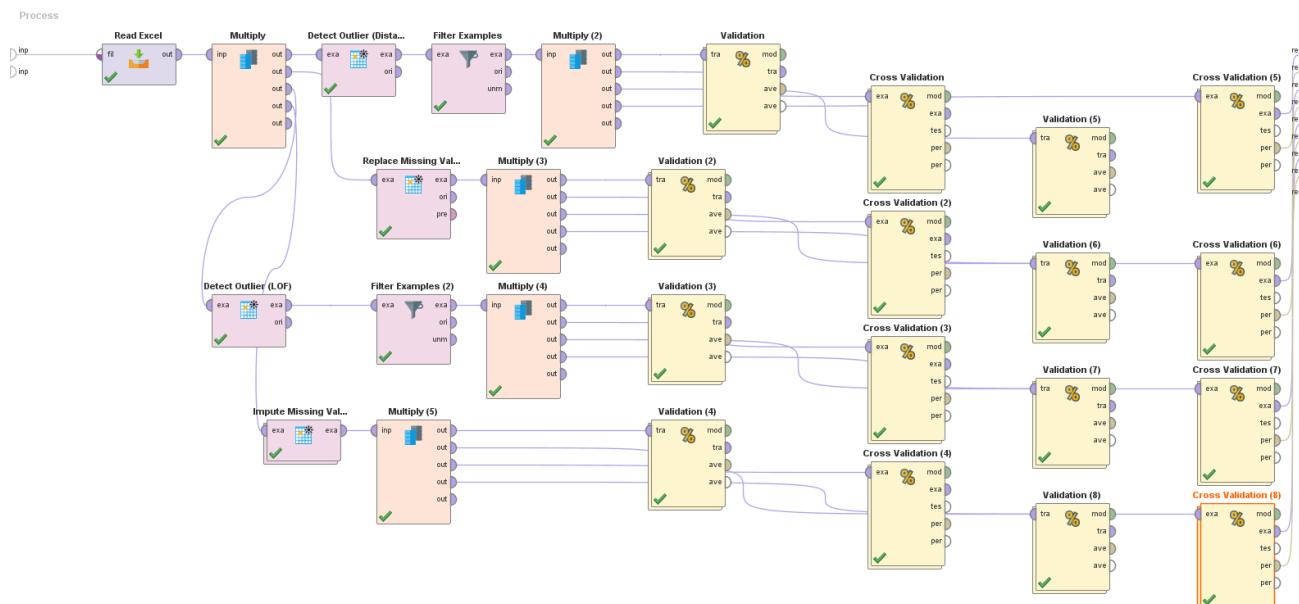
PerformanceVector:
accuracy: 70.59%
ConfusionMatrix:
True: 0 1
0: 56 21
1: 14 28
precision: 66.67% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 56 21
1: 14 28
recall: 57.14% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 56 21
1: 14 28

```

شکل 87 : نتیجه به دست آمده برای Decision Tree و Split Validation سوال 1 قسمت d با استفاده از accuracy, recall, precision

: h قسمت

در این قسمت نیز دقیقا مشابه قسمت f است با این تفاوت که به جای Decision Tree از K-nn استفاده می کنیم. و تنظیمات و اضافه کردن آن دقیقا مانند قسمت g می باشد برای همین من توضیحی نمیدهم و فقط شکل طراحی نهایی و نیز نتایج آن را قرار می دهم.



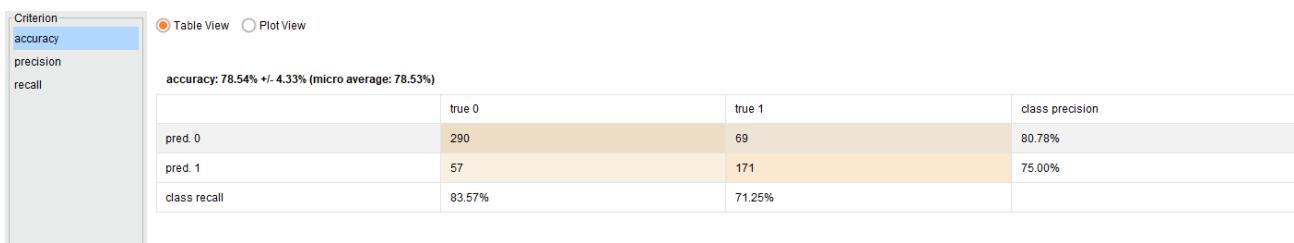
شکل 88 : طراحی نهایی قسمت h از سوال 1

همانند قسمتهای قبلی برای جلوگیری از شلوغی من اتصالات مراحل قبل را پاک کردم.

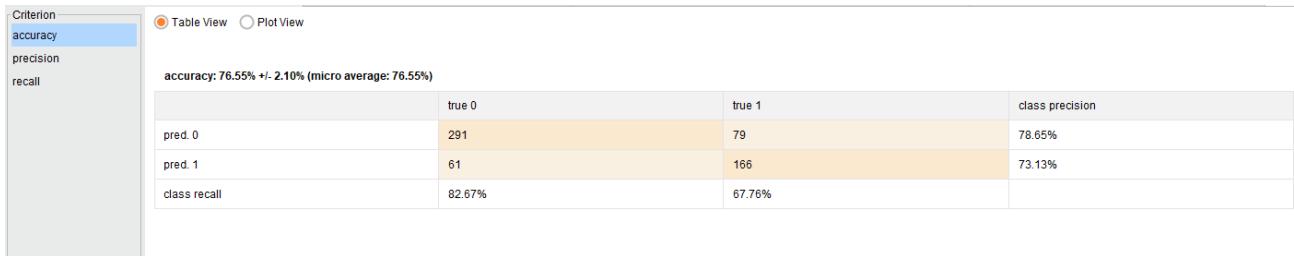
حال نتایج حاصل از اجرای این برنامه را قرار می دهم.

[Type the document title]

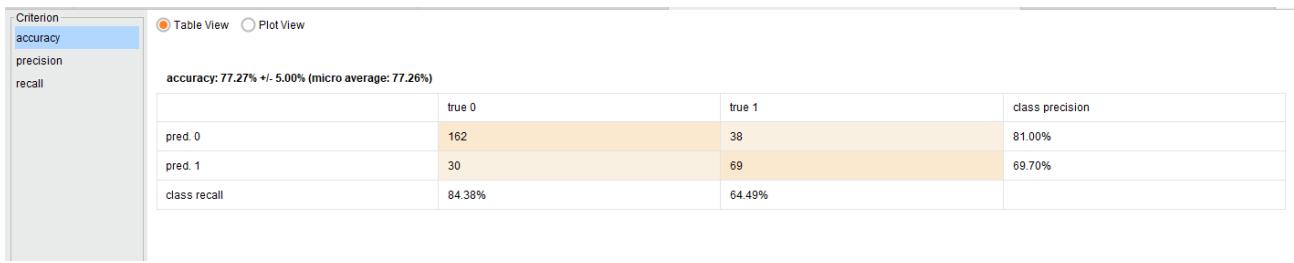
• • •



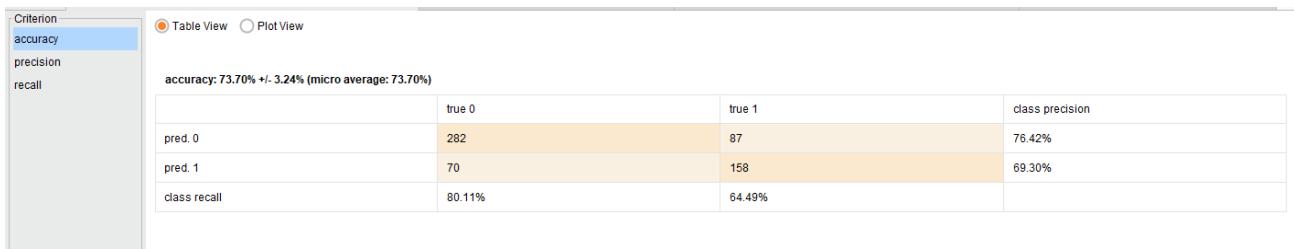
شکل 89: نتیجه به دست آمده برای accuracy, recall, precision قسمت a سوال 1 با استفاده از Decision Tree و Cross Validation



شکل 90: نتیجه به دست آمده برای accuracy, recall, precision قسمت b سوال 1 با استفاده از Decision Tree و Cross Validation



شکل 91: نتیجه به دست آمده برای accuracy, recall, precision قسمت c سوال 1 با استفاده از Decision Tree و Cross Validation



شکل 92: نتیجه به دست آمده برای accuracy, recall, precision قسمت d سوال 1 با استفاده از Decision Tree و Cross Validation

PerformanceVector

```

PerformanceVector:
accuracy: 78.54% +/- 4.33% (micro average: 78.53%)
ConfusionMatrix:
True: 0 1
0: 290 69
1: 57 171
precision: 76.06% +/- 8.59% (micro average: 75.00%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 290 69
1: 57 171
recall: 71.35% +/- 10.69% (micro average: 71.25%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 290 69
1: 57 171

```

شکل 93: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت a با استفاده از Decision Tree و Cross Validation

PerformanceVector

```

PerformanceVector:
accuracy: 76.55% +/- 2.10% (micro average: 76.55%)
ConfusionMatrix:
True: 0 1
0: 291 79
1: 61 166
precision: 73.25% +/- 3.35% (micro average: 73.13%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 291 79
1: 61 166
recall: 67.76% +/- 3.78% (micro average: 67.76%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 291 79
1: 61 166

```

شکل 94: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت b با استفاده از Decision Tree و Cross Validation

PerformanceVector

```

PerformanceVector:
accuracy: 77.27% +/- 5.00% (micro average: 77.26%)
ConfusionMatrix:
True: 0 1
0: 162 38
1: 30 69
precision: 71.78% +/- 12.39% (micro average: 69.70%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 162 38
1: 30 69
recall: 64.49% +/- 8.25% (micro average: 64.49%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 162 38
1: 30 69

```

شکل 95: نتیجه به دست آمده برای accuracy, recall, precision سوال 1 قسمت c با استفاده از Decision Tree و Cross Validation

• • •

PerformanceVector

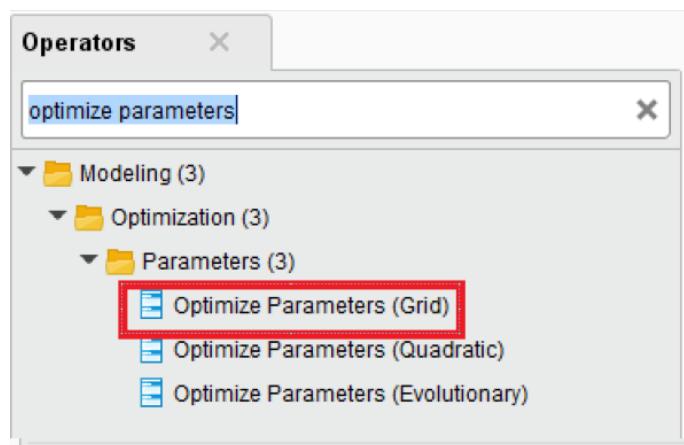
```
PerformanceVector:  
accuracy: 73.70% +/- 3.24% (micro average: 73.70%)  
ConfusionMatrix:  
True: 0 1  
0: 282 87  
1: 70 158  
precision: 69.04% +/- 2.65% (micro average: 69.30%) (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 282 87  
1: 70 158  
recall: 64.48% +/- 9.40% (micro average: 64.49%) (positive class: 1)  
ConfusionMatrix:  
True: 0 1  
0: 282 87  
1: 70 158
```

شکل 96: نتیجه به دست آمده برای accuracy, recall, precision قسمت d سوال 1 با استفاده از Decision Tree و Cross Validation

سوال 3:

در این سوال نیز از ما خواسته است که برای هر کدام از حالت‌های e , f ، بهترین پارامترها را به صورت مستقل به دست بیاوریم. معیارهایی مانند: بهترین تعداد فولدها، بهترین میزان تقسیم داده‌های آموزشی و آزمایشی، بهترین تعداد نزیک ترین همسایگی و بهترین نوع نمونه برداری.

برای این کار ما از یک اپراتور به نام Optimizer استفاده می‌کنیم تا بتوانیم این کار را انجام دهیم. برای این کار در قسمت جستجو برای این اپراتور جستجو کرده و آن را به تعداد 8 عدد برای دو قسمت e , f ، که از ما خواسته قرار می‌دهیم.

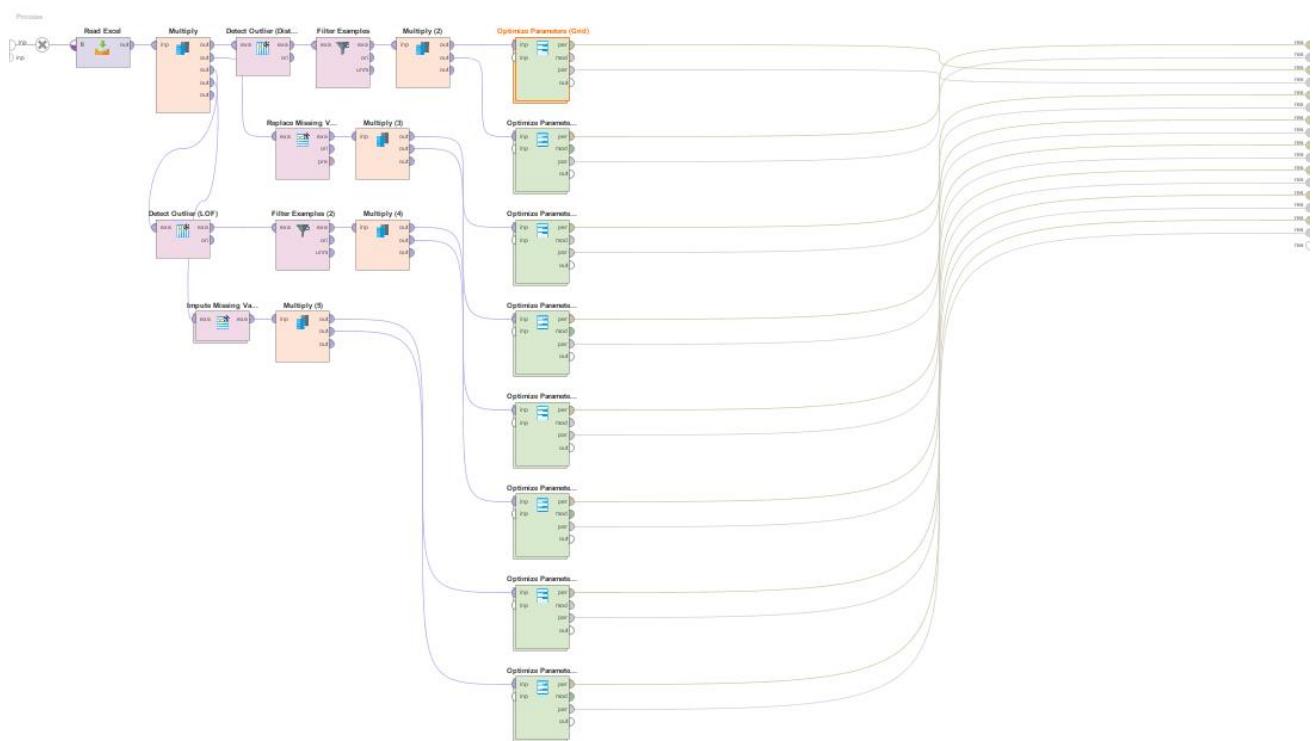


شکل 97: جستجو برای اپراتور Optimizer

پس از اضافه کردن چون به ازای هر کدام از قسمت‌های e و f نیاز به یک Optimizer داریم پس باید به هر کدام از قسمت‌های d سوال 1 ما دو عدد از این اپراتور را وصل کنیم. پس برای این کار ما نیاز به اپراتور Multiply نیز داریم.

پس شکل نهایی به صورت زیر در می‌آیند:

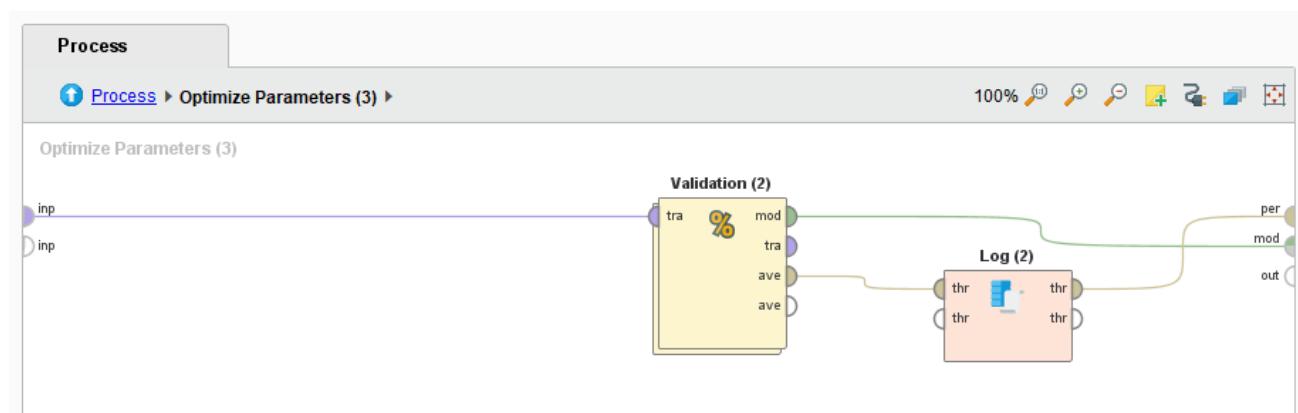
• • •



شکل 98: طراحی سوال 3

پس از این قسمت به سراغ طراحی داخل هر کدام از این Optimizer ها میرویم. درون یکی از آنها باید اپراتورهای مربوط به قسمت e باشد و درون دیگری اپراتورهای مربوط به قسمت f باشد.

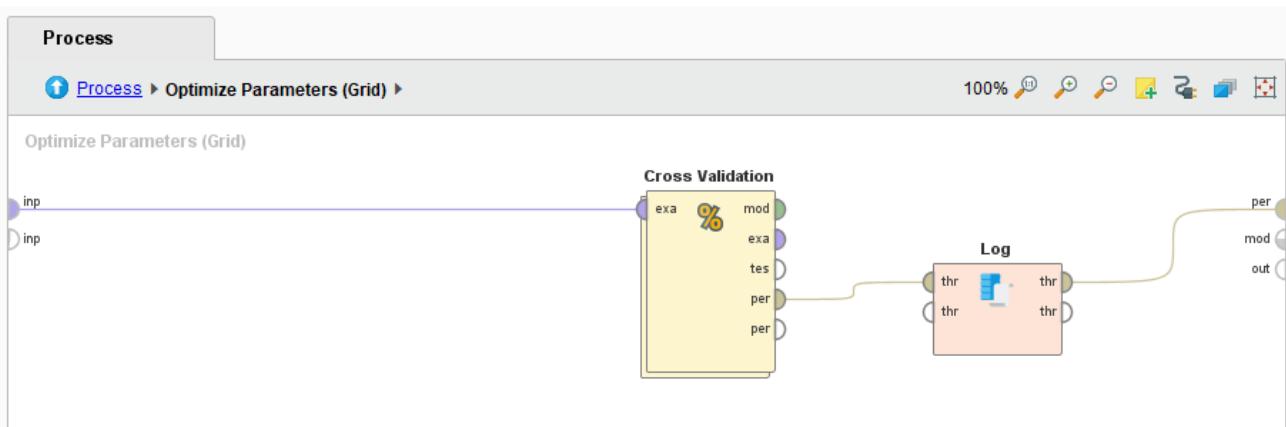
در طراحی داخلی آن موردی که مربوط به قسمت e میشود را به شکل زیر طراحی میکنیم:



شکل 99: طراحی داخلی Optimizer مربوط به قسمت e

• • •

و نیز در طراحی داخلی Optimizer ای که مربوط به قسمت f می شود نیز به صورت زیر عمل می کنیم :



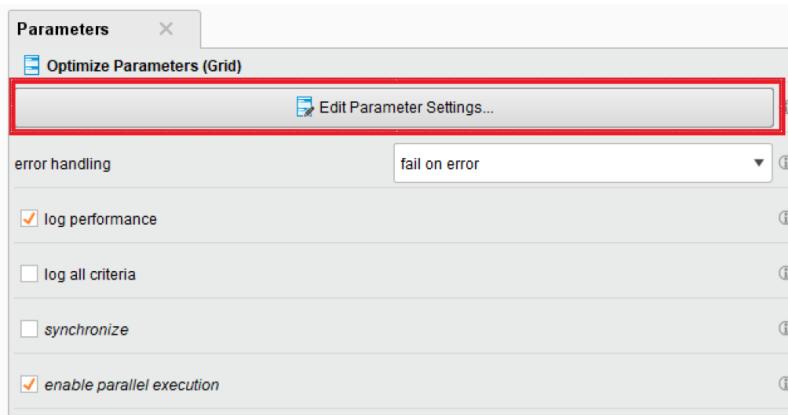
شکل 100 : طراحی داخلی Optimizer مربوط به قسمت f

طراحی داخلی Cross Validation و نیز Split Validation دقیقا مشابه قسمتهای e , f سوال 1 است، پس در اینجا مجددا بیان نمی کنم.

و همان طور که در شکل های بالا مشاهده می کنید یک اپراتور Log نیز قرار داده ام که برای لگ گیری در مراحل مختلف امتحان استفاده می شود.

حال به سراغ تنظیم پارامترها برای سنجش کارایی و پیدا کردن بهترین مقدار هر یک از موارد خواسته شده می رویم:

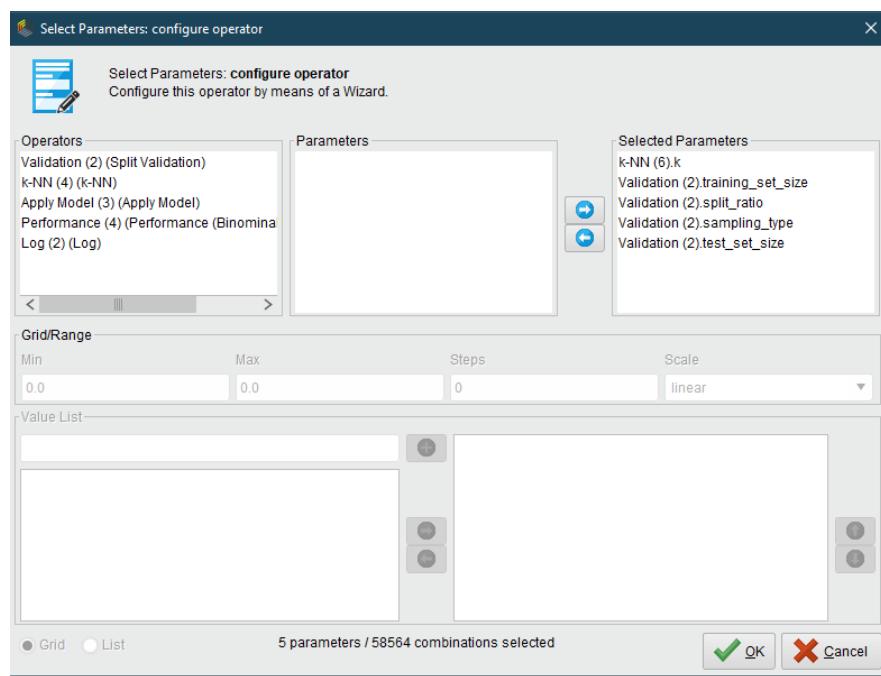
برای این کار روی هر کدام از Optimizer ها که می خواهید این موارد را برای آن تنظیم کنید کلیک کنید و سپس در قسمت Parameters آن اقدام به تنظیم کردن موارد خواسته شده بکنید.



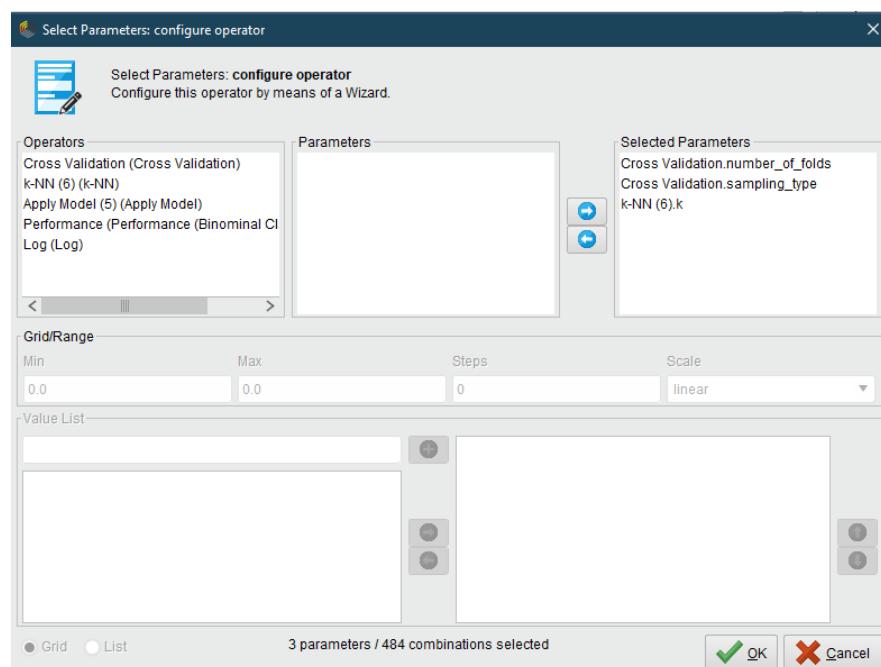
شکل 101 : قسمت پارامترهای اپراتور Optimizer

پس از کلیک روی قسمت مشخص شده در شکل بالا، حال صفحه جدیدی باز می شود و ما اقدام به تنظیم کردن مواردی که می خواهیم در آن بررسی شوند می کنیم.

• • •



شکل 102 : قسمت مربوط به Optimizer در اپراتور Edit Parameter Settings برای قسمت e



شکل 103 : قسمت مربوط به Optimizer در اپراتور Edit Parameter Settings برای قسمت f

حال پس از انجام این تنظیمات برای تمامی Optimizer ها اقدام به اجرای برنامه میکنیم که مدت زیادی اجرای آن به دلیل زیادی محاسبات به طول میانجامد.

● ● ●

iteration	Cross V...	Cross V...	k-NN (6).k	acc... ↓
58	22	shuffled ...	11	0.657
62	61	shuffled ...	11	0.657
82	41	automatic	11	0.656
73	61	stratified ...	11	0.656
60	41	shuffled ...	11	0.653
86	80	automatic	11	0.652
65	90	shuffled ...	11	0.651
84	61	automatic	11	0.650
76	90	stratified ...	11	0.650
87	90	automatic	11	0.649
80	22	automatic	11	0.649
210	2	automatic	41	0.649
81	31	automatic	11	0.649
77	100	stratified ...	11	0.649
83	51	automatic	11	0.649
61	51	shuffled ...	11	0.649
68	12	stratified ...	11	0.647
79	12	automatic	11	0.647
71	41	stratified ...	11	0.647
75	80	stratified ...	11	0.647
72	51	stratified ...	11	0.647

شکل 104: بهترین مقدار به دست آمده برای Optimizer 1

ParameterSet

```

Parameter set:

Performance:
PerformanceVector [
----accuracy: 65.74% +/- 8.58% (micro average: 65.76%)
ConfusionMatrix:
True: 0      1
0:   306    160
1:   41     80
----precision: 65.10% +/- 20.96% (micro average: 66.12%) (positive class: 1)
ConfusionMatrix:
True: 0      1
0:   306    160
1:   41     80
----recall: 33.06% +/- 13.97% (micro average: 33.33%) (positive class: 1)
ConfusionMatrix:
True: 0      1
0:   306    160
1:   41     80
]
Cross Validation.number_of_folds      = 22
Cross Validation.sampling_type      = shuffled sampling
k-NN (6).k                          = 11

```

شکل 105: بهترین مقدار به دست آمده برای Optimizer 1

• • •

Optimize Parameters (2) (484 rows, 5 columns)

iteration	Cross V...	Cross V...	k-NN (6).k	acc... ↓
202	31	stratified ...	41	0.652
320	2	shuffled ...	70	0.652
54	90	linear sa...	11	0.651
94	51	linear sa...	21	0.651
477	31	automatic	100	0.651
314	51	linear sa...	70	0.650
37	31	automatic	1	0.650
131	90	automatic	21	0.650
349	71	automatic	70	0.650
444	31	linear sa...	100	0.650
14	22	shuffled ...	1	0.650
256	22	automatic	51	0.650
307	90	automatic	60	0.650
379	41	stratified ...	80	0.650
224	31	linear sa...	51	0.650
119	80	stratified ...	21	0.649
166	2	automatic	31	0.649
404	71	linear sa...	90	0.649
405	80	linear sa...	90	0.649
410	22	shuffled ...	90	0.649
36	22	automatic	1	0.649
222	12	linear sa...	51	0.649
352	100	automatic	70	0.649
377	22	stratified ...	80	0.649

شکل 106 : بهترین مقدار به دست آمده برای Optimizer 2

ParameterSet

```
Parameter set:
Performance:
PerformanceVector [
-----accuracy: 65.23% +/- 17.01% (micro average: 65.16%)
ConfusionMatrix:
True: 0 1
0: 309 165
1: 43 80
-----precision: 65.04% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 309 165
1: 43 80
-----recall: 33.33% +/- 31.62% (micro average: 32.65%) (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 309 165
1: 43 80
]
Cross Validation.number_of_folds      = 31
Cross Validation.sampling_type   = stratified sampling
k-NN (6).k        = 41
```

شکل 107 : بهترین مقدار به دست آمده برای Optimizer 2

● ● ●

Optimize Parameters (3) (58564 rows, 7 columns)

iteration	k-NN (6).k	Validati...	Validati...	Validati...	Validati...	acc... ↓
23941	41	91	0.900	shuffled ...	46	0.847
13259	31	64	0.900	shuffled ...	28	0.831
13306	60	100	0.900	shuffled ...	28	0.831
41148	70	10	0.900	stratified ...	73	0.831
50490	100	28	0.900	shuffled ...	91	0.831
2623	41	73	0.900	shuffled ...	10	0.814
13266	100	64	0.900	shuffled ...	28	0.814
19945	11	91	0.900	stratified ...	37	0.814
34534	41	46	0.900	shuffled ...	64	0.814
58497	90	46	0.900	automatic	100	0.814
39720	90	28	0.820	shuffled ...	73	0.802
7903	41	37	0.900	shuffled ...	19	0.797
21288	21	100	0.900	automatic	37	0.797
23843	51	10	0.900	shuffled ...	46	0.797
25185	51	19	0.900	stratified ...	46	0.797
26522	1	28	0.900	automatic	46	0.797
35846	70	28	0.900	stratified ...	64	0.797
57134	100	19	0.900	stratified ...	100	0.797
57202	11	82	0.900	stratified ...	100	0.797
39741	80	46	0.820	shuffled ...	73	0.783
7878	11	19	0.900	shuffled ...	19	0.780
7900	11	37	0.900	shuffled ...	19	0.780
7935	31	64	0.900	shuffled ...	19	0.780
9207	100	10	0.900	stratified ...	19	0.780

شکل 108 : بهترین مقدار به دست آمده برای Optimizer 3

ParameterSet

```

Parameter set:

Performance:
PerformanceVector [
-----accuracy: 84.75%
ConfusionMatrix:
True: 0      1
0:    35     5
1:    4      15
-----precision: 78.95% (positive class: 1)
ConfusionMatrix:
True: 0      1
0:    35     5
1:    4      15
-----recall: 75.00% (positive class: 1)
ConfusionMatrix:
True: 0      1
0:    35     5
1:    4      15
]
k-NN (6).k      = 41
Validation (2).training_set_size      = 91
Validation (2).split_ratio          = 0.9
Validation (2).sampling_type        = shuffled sampling
Validation (2).test_set_size        = 46

```

شکل 109 : بهترین مقدار به دست آمده برای Optimizer 3

● ● ●

Optimize Parameters (4) (58564 rows, 7 columns)

iteration	k-NN (6).k	Validati...	Validati...	Validati...	Validati...	Validati...	acc... ↓
40709	80	46	0.580	stratified ...	73	0.881	
38824	41	91	0.180	shuffled ...	73	0.864	
28599	90	37	0.500	shuffled ...	55	0.847	
38254	60	19	0.740	linear sa...	73	0.847	
6709	90	46	0.100	shuffled ...	19	0.831	
14713	51	64	0.100	automatic	28	0.831	
19610	70	10	0.740	stratified ...	37	0.831	
21508	21	82	0.180	linear sa...	46	0.831	
21731	51	64	0.340	linear sa...	46	0.831	
21758	100	82	0.340	linear sa...	46	0.831	
23804	100	73	0.820	shuffled ...	46	0.831	
34939	21	82	0.260	stratified ...	64	0.831	
37394	41	10	0.180	linear sa...	73	0.831	
42488	51	19	0.900	automatic	73	0.831	
46251	60	28	0.740	stratified ...	82	0.831	
47815	80	19	0.900	automatic	82	0.831	
54640	21	64	0.100	shuffled ...	100	0.831	
2950	11	46	0.260	stratified ...	10	0.814	
2956	70	46	0.260	stratified ...	10	0.814	
3905	100	28	0.900	stratified ...	10	0.814	
4580	31	91	0.420	automatic	10	0.814	
5670	41	91	0.260	linear sa...	19	0.814	
7332	51	64	0.500	shuffled ...	19	0.814	

شکل 110 : بهترین مقدار به دست آمده برای Optimizer 4

ParameterSet

```

Parameter set:
Performance:
PerformanceVector [
----accuracy: 88.14%
ConfusionMatrix:
True: 0 1
0: 33 5
1: 2 19
----precision: 90.48% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 33 5
1: 2 19
----recall: 79.17% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 33 5
1: 2 19
]
k-NN (6).k      = 80
Validation (2).training_set_size      = 46
Validation (2).split_ratio      = 0.58
Validation (2).sampling_type      = stratified sampling
Validation (2).test_set_size      = 73

```

شکل 111 : بهترین مقدار به دست آمده برای Optimizer 4

● ● ●

Optimize Parameters (5) (484 rows, 5 columns)

iteration	Cross V...	Cross V...	k-NN (6).k	acc... ↓
27	41	stratified ...	1	0.643
37	31	automatic	1	0.643
43	90	automatic	1	0.643
95	61	linear sa...	21	0.643
101	12	shuffled ...	21	0.643
104	41	shuffled ...	21	0.643
114	31	stratified ...	21	0.643
118	71	stratified ...	21	0.643
205	61	stratified ...	41	0.643
348	61	automatic	70	0.643
382	71	stratified ...	80	0.643
1	2	linear sa...	1	0.643
2	12	linear sa...	1	0.643
3	22	linear sa...	1	0.643
5	41	linear sa...	1	0.643
6	51	linear sa...	1	0.643
7	61	linear sa...	1	0.643
9	80	linear sa...	1	0.643
11	100	linear sa...	1	0.643
12	2	shuffled ...	1	0.643
13	12	shuffled ...	1	0.643
14	22	shuffled ...	1	0.643
15	31	shuffled ...	1	0.643
16	41	shuffled ...	1	0.643

شکل 112 : بهترین مقدار به دست آمده برای Optimizer 5

● ● ●

Optimize Parameters (6) (58564 rows, 7 columns)

iteration	k-NN (6).k	Validati...	Validati...	Validati...	Validati...	acc... ↓
2893	100	91	0.180	stratified ...	10	0.900
6363	41	64	0.740	linear sa...	19	0.900
9523	70	73	0.180	automatic	19	0.900
21401	51	91	0.100	linear sa...	46	0.900
23968	90	10	0.100	stratified ...	46	0.900
27050	1	64	0.340	linear sa...	55	0.900
38817	80	82	0.180	shuffled ...	73	0.900
39848	51	37	0.900	shuffled ...	73	0.900
45276	100	19	0.100	stratified ...	82	0.900
50012	51	37	0.580	shuffled ...	91	0.900
949	21	91	0.660	linear sa...	10	0.867
2204	31	28	0.660	shuffled ...	10	0.867
6203	90	28	0.660	linear sa...	19	0.867
6409	60	100	0.740	linear sa...	19	0.867
6828	70	46	0.180	shuffled ...	19	0.867
7263	21	10	0.500	shuffled ...	19	0.867
7906	70	37	0.900	shuffled ...	19	0.867
10071	51	28	0.580	automatic	19	0.867
11454	21	73	0.580	linear sa...	28	0.867
12951	31	10	0.740	shuffled ...	28	0.867
13687	21	19	0.340	stratified ...	28	0.867
15201	90	64	0.420	automatic	28	0.867
15520	90	28	0.660	automatic	28	0.867
16035	70	55	0.100	linear sa...	37	0.867

شکل 113 : بهترین مقدار به دست آمده برای Optimizer 6

ParameterSet

```

Parameter set:
Performance:
PerformanceVector [
-----accuracy: 90.00%
ConfusionMatrix:
True: 0 1
0: 18 2
1: 1 9
-----precision: 90.00% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 18 2
1: 1 9
-----recall: 81.82% (positive class: 1)
ConfusionMatrix:
True: 0 1
0: 18 2
1: 1 9
]
k-NN (6).k      = 100
Validation (2).training_set_size      = 91
Validation (2).split_ratio           = 0.18000000000000002
Validation (2).sampling_type         = stratified sampling
Validation (2).test_set_size         = 10

```

شکل 114 : بهترین مقدار به دست آمده برای Optimizer 6

● ● ●

Optimize Parameters (7) (484 rows, 5 columns)

iteration	Cross V...	Cross V...	k-NN (6).k	acc... ↓
8	71	linear sa...	1	0.650
335	41	stratified ...	70	0.647
391	51	automatic	80	0.646
21	90	shuffled ...	1	0.646
371	71	shuffled ...	80	0.646
340	90	stratified ...	70	0.646
209	100	stratified ...	41	0.646
243	2	stratified ...	51	0.646
353	2	linear sa...	80	0.646
478	41	automatic	100	0.645
84	61	automatic	11	0.645
38	41	automatic	1	0.645
109	90	shuffled ...	21	0.645
215	51	automatic	41	0.645
405	80	linear sa...	90	0.645
282	61	shuffled ...	60	0.645
111	2	stratified ...	21	0.645
7	61	linear sa...	1	0.645
61	51	shuffled ...	11	0.645
133	2	linear sa...	31	0.645
75	80	stratified ...	11	0.644
102	22	shuffled ...	21	0.644
131	90	automatic	21	0.644
193	51	shuffled ...	41	0.644

شکل 115 : بهترین مقدار به دست آمده برای Optimizer 7

ParameterSet

```

Parameter set:
Performance:
PerformanceVector [
----accuracy: 65.00% +/- 17.39% (micro average: 64.99%)
ConfusionMatrix:
True: 0      1
0:    313    170
1:    39     75
----precision: 65.79% (positive class: 1)
ConfusionMatrix:
True: 0      1
0:    313    170
1:    39     75
----recall: 32.17% +/- 32.17% (micro average: 30.61%) (positive class: 1)
ConfusionMatrix:
True: 0      1
0:    313    170
1:    39     75
]
Cross Validation.number_of_folds      = 71
Cross Validation.sampling_type   = linear sampling
k-NN (6).k      = 1

```

شکل 116 : بهترین مقدار به دست آمده برای Optimizer 7

● ● ●

Optimize Parameters (8) (58564 rows, 7 columns)

iteration	k-NN (6).k	Validati...	Validati...	Validati...	Validati...	acc... ↓
47701	41	28	0.820	automatic	82	0.881
8188	31	73	0.180	stratified ...	19	0.864
35795	1	91	0.820	stratified ...	64	0.864
51164	21	91	0.420	stratified ...	91	0.864
2371	51	64	0.740	shuffled ...	10	0.847
15765	11	37	0.820	automatic	28	0.847
25740	100	73	0.340	automatic	46	0.847
27837	60	10	0.900	linear sa...	55	0.847
32108	90	37	0.180	linear sa...	64	0.847
40479	90	55	0.420	stratified ...	73	0.847
43979	1	55	0.100	shuffled ...	82	0.847
45160	41	28	0.900	shuffled ...	82	0.847
47409	90	82	0.580	automatic	82	0.847
47416	51	91	0.580	automatic	82	0.847
57049	21	55	0.820	stratified ...	100	0.847
57575	1	91	0.260	automatic	100	0.847
926	11	73	0.660	linear sa...	10	0.831
1697	21	10	0.340	shuffled ...	10	0.831
8464	41	100	0.340	stratified ...	19	0.831
11931	60	64	0.900	linear sa...	28	0.831
14439	60	37	0.820	stratified ...	28	0.831
15133	70	10	0.420	automatic	28	0.831
17614	21	64	0.260	shuffled ...	37	0.831
19642	60	37	0.740	stratified ...	37	0.831

شکل 117: بهترین مقدار به دست آمده برای Optimizer 8

ParameterSet

```

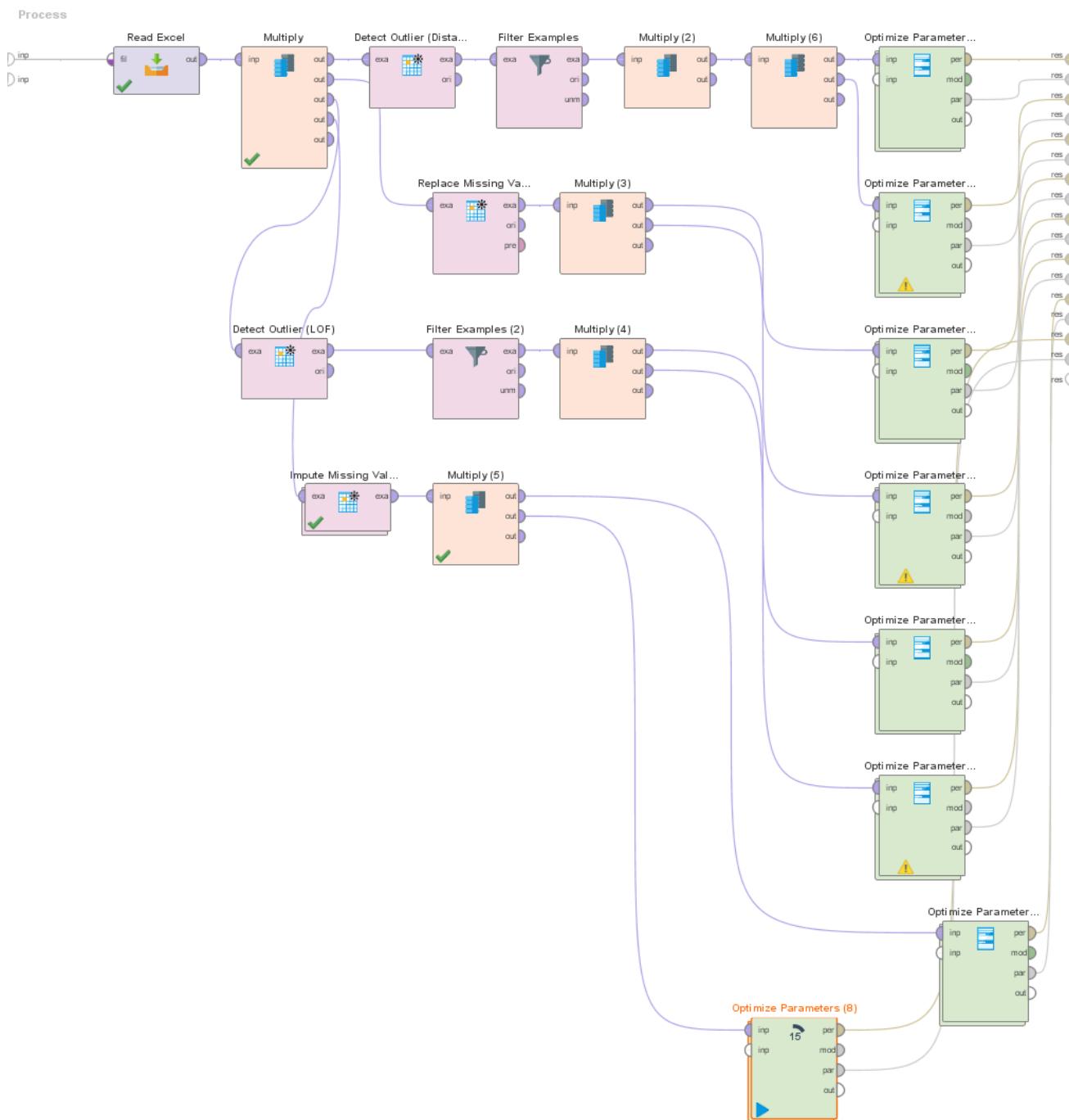
Parameter set:
Performance:
PerformanceVector [
-----accuracy: 88.14%
ConfusionMatrix:
True:   0      1
0:     33      5
1:     2      19
-----precision: 90.48% (positive class: 1)
ConfusionMatrix:
True:   0      1
0:     33      5
1:     2      19
-----recall: 79.17% (positive class: 1)
ConfusionMatrix:
True:   0      1
0:     33      5
1:     2      19
]
k-NN (6).k      = 41
Validation (2).training_set_size      = 28
Validation (2).split_ratio          = 0.8200000000000001
Validation (2).sampling_type        = automatic
Validation (2).test_set_size        = 82

```

شکل 118: بهترین مقدار به دست آمده برای Optimizer 8

سوال 4:

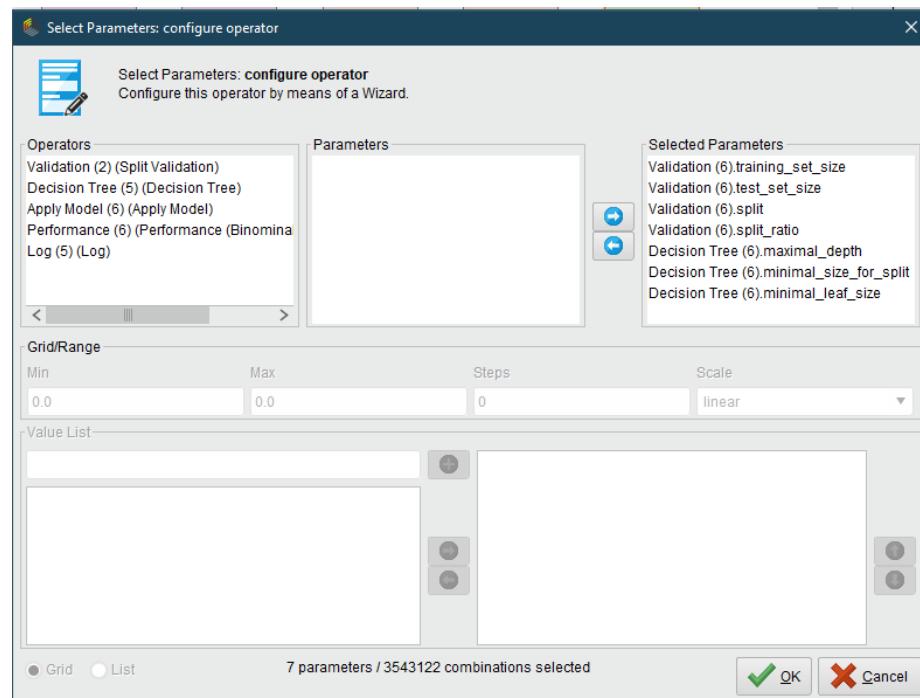
در این سوال نیز دقیقاً مانند سوال 3 اتصالات را برقرار کرده و شکلی مشابه شکل زیر را طراحی می‌کنیم:



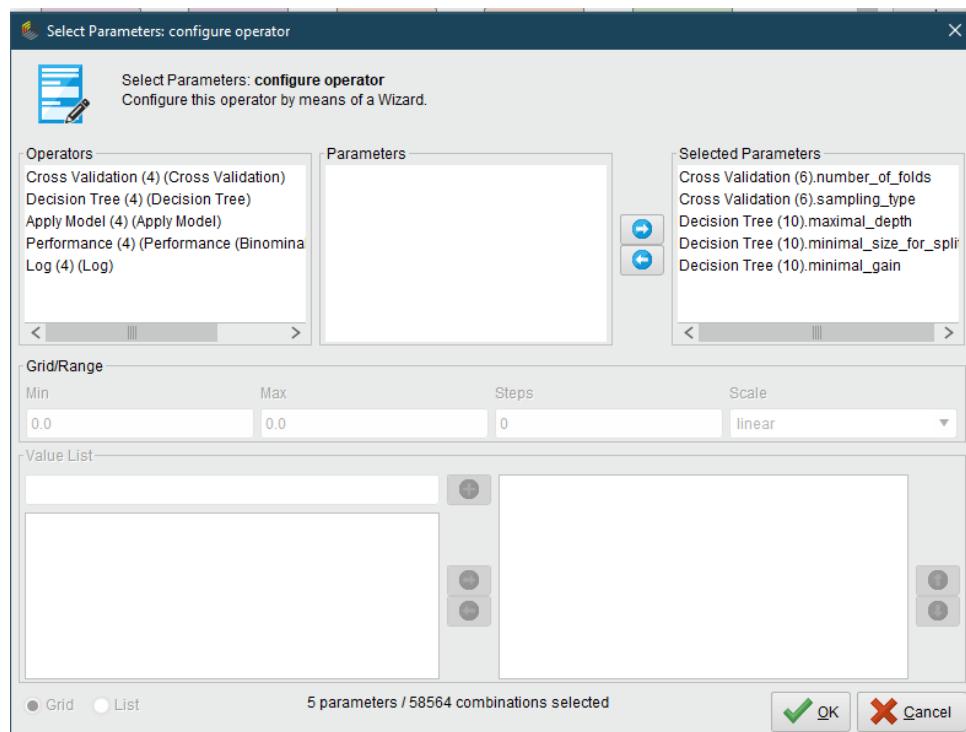
شکل 119: طراحی مربوط به سوال 4

پس از انجام طرحی و قرار دادن مقادیر مربوط به قسمتهای g, h در سوال 1 درون Optimizer ها حال به تنظیم مقادیر درون آنها می پردازیم.

• • •



شکل 120 : تنظیمات مربوط به قسمت g سوال 1



شکل 121 : تنظیمات مربوط به قسمت h سوال 1

حال در زیر نتایج آن را بیان می‌کنم:

Optimize Parameters (2) (119 rows, 11 columns)										
iteration	Detect ...	Detect ...	Replace...	Detect ...	Detect ...	Cross V...	Cross V...	Decisio...	Decisio...	acc... ↓
93	52	84	average	10	15	8	shuffled ...	gain_ratio	8	0.780
12	20	28	average	10	15	8	shuffled ...	gain_ratio	8	0.772
108	84	92	average	10	15	8	shuffled ...	gain_ratio	8	0.770
34	20	44	average	10	15	8	shuffled ...	gain_ratio	8	0.766
40	68	44	average	10	15	8	shuffled ...	gain_ratio	8	0.764
13	28	28	average	10	15	8	shuffled ...	gain_ratio	8	0.761
8	76	20	average	10	15	8	shuffled ...	gain_ratio	8	0.760
21	92	28	average	10	15	8	shuffled ...	gain_ratio	8	0.760
74	76	68	average	10	15	8	shuffled ...	gain_ratio	8	0.759
119	84	100	average	10	15	8	shuffled ...	gain_ratio	8	0.758
32	92	36	average	10	15	8	shuffled ...	gain_ratio	8	0.758
19	76	28	average	10	15	8	shuffled ...	gain_ratio	8	0.758
11	100	20	average	10	15	8	shuffled ...	gain_ratio	8	0.758
101	28	92	average	10	15	8	shuffled ...	gain_ratio	8	0.756
106	68	92	average	10	15	8	shuffled ...	gain_ratio	8	0.756
109	92	92	average	10	15	8	shuffled ...	gain_ratio	8	0.756
116	60	100	average	10	15	8	shuffled ...	gain_ratio	8	0.756
97	84	84	average	10	15	8	shuffled ...	gain_ratio	8	0.754
47	36	52	average	10	15	8	shuffled ...	gain_ratio	8	0.754
70	22	76	average	10	15	8	shuffled ...	gain_ratio	8	0.754

شکل 122 : نتیجه مربوط به Optimizer 1

Optimize Parameters (Grid) (11 rows, 3 column)

iteration	k-NN (2).k	accuracy ↓
7	60	0.830
4	31	0.810
10	90	0.810
1	1	0.800
11	100	0.790
2	11	0.780
6	51	0.780
3	21	0.770
5	41	0.770
8	70	0.770
9	80	0.760

شکل 123 : نتیجه مربوط به Optimizer 2

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Decision Tree.maximal_depth	accuracy ↓
4	29	0.830
2	9	0.810
9	80	0.810
7	60	0.800
11	100	0.800
8	70	0.770
10	90	0.770
3	19	0.750
6	50	0.740
5	39	0.720
1	-1	0.690

شکل 124 : نتیجه مربوط به Optimizer 3

Optimize Parameters (Grid) (4 rows, 3 columns)

iteration	Decision Tree.criterion	accuracy ↓
4	accuracy	0.810
2	information_gain	0.780
1	gain_ratio	0.760
3	gini_index	0.750

شکل 125 : بهترین نتیجه مربوط به Optimizer 4

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Validation.training_set_size	accuracy ↓
7	60	0.830
4	29	0.810
10	90	0.810
1	-1	0.800
11	100	0.790
2	9	0.780
6	50	0.780
3	19	0.770
5	39	0.770
8	70	0.770
9	80	0.760

شکل 126 : بهترین نتیجه مربوط به Optimizer 5

• • •

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Validation.test_set_size	accuracy ↓
7	60	0.830
4	29	0.810
10	90	0.810
1	-1	0.800
11	100	0.790
2	9	0.780
6	50	0.780
3	19	0.770
5	39	0.770
8	70	0.770
9	80	0.760

شکل 127 : بهترین نتیجه مربوط به Optimizer 6

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Detect Outlier (Distances).number_of_neighbors	accura... ↓
5	41	0.828
8	70	0.828
10	90	0.808
4	31	0.798
11	100	0.790
3	21	0.788
6	51	0.788
1	1	0.770
2	11	0.768
7	60	0.758
9	80	0.758

شکل 128 : بهترین نتیجه مربوط به Optimizer 7

• • •

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Detect Outlier (Distances).number_of_outliers	accuracy ↓
4	31	0.825
8	70	0.811
1	1	0.807
11	100	0.790
3	21	0.783
10	90	0.775
2	11	0.769
7	60	0.769
6	51	0.752
5	41	0.739
9	80	0.721

شکل 129 : بهترین نتیجه مربوط به Optimizer 8

Optimize Parameters (Grid) (5 rows, 3 columns)

iteration	Replace Missing Values.default	accuracy ↓
4	average	0.810
1	none	0.800
2	minimum	0.780
3	maximum	0.770
5	zero	0.770

شکل 130 : بهترین نتیجه مربوط به Optimizer 1

• • •

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	k-NN (2).k	accuracy ↓
6	51	0.785
8	70	0.785
9	80	0.777
4	31	0.775
1	1	0.775
2	11	0.769
7	60	0.768
3	21	0.764
5	41	0.751
10	90	0.750
11	100	0.728

شکل 131 : بهترین نتیجه مربوط به Optimizer 2

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Decision Tree.maximal_depth	accuracy ↓
8	70	0.789
5	39	0.787
4	29	0.783
3	19	0.782
11	100	0.779
1	-1	0.777
6	50	0.777
10	90	0.775
9	80	0.774
2	9	0.770
7	60	0.769

شکل 132 : بهترین نتیجه مربوط به Optimizer 3

• • •

Optimize Parameters (Grid) (4 rows, 3 columns)

iteration	Decision Tree.criterion	accuracy ↓
4	accuracy	0.783
2	information_gain	0.778
3	gini_index	0.777
1	gain_ratio	0.751

شکل 133 : بهترین نتیجه مربوط به Optimizer 4

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Cross Validation.number_of_folds	accuracy ↓
8	71	0.787
10	90	0.786
3	22	0.785
11	100	0.779
9	80	0.777
5	41	0.776
4	31	0.775
7	61	0.774
6	51	0.771
2	12	0.763
1	2	0.742

شکل 134 : بهترین نتیجه مربوط به Optimizer 5

Optimize Parameters (Grid) (4 rows, 3 columns)

iteration	Cross Validation.sampling_type	accuracy ↓
3	stratified sampling	0.777
2	shuffled sampling	0.775
1	linear sampling	0.773
4	automatic	0.768

شکل 135 : بهترین نتیجه مربوط به Optimizer 6

● ● ●

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Detect Outlier (Distances).number_of_neighbors	accuracy ↓
8	70	0.793
7	60	0.789
5	41	0.787
3	21	0.787
10	90	0.781
9	80	0.769
1	1	0.763
4	31	0.756
6	51	0.755
2	11	0.740
11	100	0.728

شکل 136 : بهترین نتیجه مربوط به Optimizer 7

Optimize Parameters (Grid) (11 rows, 3 columns)

iteration	Detect Outlier (Distances).number_of_outliers	accuracy ↓
4	31	0.800
5	41	0.782
8	70	0.778
7	60	0.777
10	90	0.773
9	80	0.772
6	51	0.766
2	11	0.763
3	21	0.762
1	1	0.757
11	100	0.728

شکل 137 : بهترین نتیجه مربوط به Optimizer 8

[Type the document title]

• • •