



دانشگاه قم
دانشکده فنی و مهندسی
گروه مهندسی کامپیوتر و فناوری اطلاعات
گزارش پژوهش درس داده کاوی

عنوان:
بررسی معیار شاخص توده بدنی بر روی فشار خون

دانشجو:
رامین بدرا

استاد:

آقای دکتر امیرخانی

مرداد ماه ۱۳۹۵

چکیده

امروزه در دانش پزشکی جمع آوری داده‌های فراوان در مورد بیماری‌های مختلف از اهمیت فراوانی برخوردار است. مراکز پزشکی با مقاصد گوناگونی به جمع آوری این داده‌ها می‌پردازنند. تحقیق روی این داده‌ها و به دست آوردن نتایج و الگوهای مفید در رابطه با بیماری‌ها، یکی از اهداف استفاده از این داده‌ها است. حجم زیاد این داده‌ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می‌شود.

بنابراین از داده‌کاوی برای غلبه بر این مشکل و به دست آوردن روابط مفید بین عوامل خطرزا در بیماری‌ها استفاده می‌شود.

ما در این تحقیق با استفاده از داده‌های موجود و تکنیک دسته بندی درخت تصمیم و تئوری بیز و الگوریتم KNN، با استفاده از نرم افزار RapidMiner به پیش‌بینی وقوع فشار خون در افراد با تغییر شاخص توده بدنی می‌پردازیم و با کمک تکنیک‌های خوبه بندی و مناسب ترین الگوریتم بدست آمده داده‌های خود را در چندین خوش‌قرار می‌دهیم.

کلمات کلیدی: داده‌کاوی، درخت تصمیم، تئوری بیز، الگوریتم KNN، الگوریتم K-means، الگوریتم K-medoids

فهرست مطالب

۵	بخش اول: مقدمه
۵	بخش دوم: اهداف کلی تحقیق
۶	بخش سوم: شرح داده ها
۶	پایگاه داده
۶	خصیصه های پایگاه داده
۷	بخش چهارم: پیش پردازش داده ها
۸	بخش پنجم: شرح پردازش ها
۸	طبقه بندی
۹	خوش بندی
۱۰	بخش ششم: پیاده سازی
۱۰	طبقه بندی
۱۱	عملگر Retrieve
۱۱	عملگر Filter Example
۱۱	عملگر Generate Attributes
۱۱	عملگر Select Attributes
۱۲	عملگر Reorder Attributes
۱۲	عملگر Set Role
۱۲	عملگر Multiply
۱۲	عملگر Validation
۲۱	نتیجه گیری
۲۲	تخمین برحسب داده های آزمایشی با استفاده از مدل درخت تصمیم
۲۷	میزان همبستگی خصیصه ها
۲۹	خوش بندی
۳۰	عملگر Nominal to Numerical

۳۰	عملگر Multiply
۳۰	عملگر Clustering 1
۳۲	عملگر Clustering 2
۳۲	عملگر Performance
۳۴	نتیجه گیری
۳۵	خوش بندی داده ها با استفاده از الگوریتم برگزیده مرحله قبل الگوریتم K-means
۴۱	نتایج حاصل از پیاده سازی
۴۳	بخش هفتم: نتیجه گیری نهایی

بخش اول: مقدمه

پرفشاری خون یکی از عوامل خطرساز در بروز بیماری‌های قلبی-عروقی است. از طرفی مرگ ناشی از بیماری‌های قلبی-عروقی بیشترین میزان میرایی را در اکثر کشورهای صنعتی به خود اختصاص می‌دهد و میزان آن در کشورهای در حال توسعه نیز رو به افزایش است. طبق طبقه‌بندی سازمان جهانی بهداشت (WHO) ۱۷ علت اصلی برای میرایی وجود دارد که بیماری‌های قلبی-عروقی، رتبه هفتم آن را در جهان به خود اختصاص می‌دهند، اما در ایران مرگ ناشی از بیماری‌های قلبی-عروقی رتبه اول را دارا می‌باشد.

در بسیاری از موارد، علت اصلی ابتلا به پرفشاری خون نامشخص است. اما عوامل متعددی در تشدید این عارضه نقش دارند که از آن جمله می‌توان به چاقی و افزایش نمایه توده بدن (BMI) اشاره نمود.

نمایه توده بدن Body Mass Index (BMI) یک اندازه انتropometریکی از وزن (kg) است که با تقسیم بر مجدور قد (m^2) به دست می‌آید و اغلب به طور گسترده برای ارزیابی اضافه وزن و چاقی در جوامع بزرگ مورد استفاده قرار می‌گیرد.

بخش دوم: اهداف کلی تحقیق

تشخیص بیماری به عهده پزشک است که با مشاهده علائم و انجام آزمایشات صورت می‌گیرد. از جمله بیماری‌هایی که پزشکان را در تشخیص با مشکلاتی مواجه ساخته، بیماری‌های مربوط به فشار خون و قلب است که به دلیل نقش حیاتی قلب در سلامت انسان، از اهمیت بسزایی برخوردار است. به همین دلیل نیاز به سیستم‌هایی برای کمک به تصمیم گیری احساس می‌شود.

سیستم‌های پشتیبان تصمیم با رویکردی متفاوت برای اینگونه از تصمیم گیری‌ها وجود دارد که مبنای اکثر این رویکردها استفاده از تکنیک‌هایی است که بتواند ارتباط بین داده‌ها را کشف کند. سه مورد از این تکنیک‌ها دسته بندی درخت تصمیم، تئوری بیز و الگوریتم KNN است.

ما در این گزارش در ابتدا به توضیح پایگاه داده مورد استفاده می‌پردازیم و توضیح مختصه برای خصیصه‌های داده ارائه خواهیم داد. سپس اشاره‌ای به تکنیک دسته بندی درخت تصمیم، الگوریتم بیز و الگوریتم KNN می‌کنیم و با استفاده از نرم افزار RapidMiner، به پیاده سازی این تکنیک‌ها بر روی داده‌ها می‌پردازیم و سپس با استفاده از الگوریتم K-medoid و K-Means سازی قابل مشاهده است.

بخش سوم: شرح داده ها

پایگاه داده

برای بکارگیری روش‌های داده کاوی قبل از هرچیز نیاز به داده است. پایگاه داده ارتباط شاخص BMI و فشار خون موجود مربوط به کشور چین بوده که بین سالهای ۲۰۰۶ تا ۲۰۱۱ جمع آوری شده است. که شاخص بهینه‌ی چاقی (BMI) را برای پیش‌بینی وقوع فشارخون در افراد با سن‌های بین ۱۸ تا ۶۵ سال مورد بررسی قرار داده است.

خصیصه‌های پایگاه داده

پایگاه داده ارتباط شاخص BMI و فشار خون موجود اطلاعات مربوط به ۳۲۵۳ نفر را که هر یک دارای ۱۵ خصیصه است، در خود ذخیره کرده است.

نوع	توضیحات	مشخصه
عددی	سن	Age
اسمی	جنسیت	Gender
اسمی	استعمال دخانیات	Smoke2006
اسمی	صرف مشروبات الکلی	Drink2006
اسمی	سابقه فشار خون بالا	Hypertension2011
عددی	دور کمر (بر حسب سانتی متر)	Waist Circumference2006
عددی	دور باسن (بر حسب سانتی متر)	Hip Circumference 2006
اسمی	منطقه سکونت	Residence2006
عددی	قد در سال ۲۰۰۶ (سانتی متر)	height 2006
عددی	وزن در سال ۲۰۱۱ (کیلوگرم)	Weight 2006
عددی	قد در سال ۲۰۱۱ (سانتی متر)	Height 2011
عددی	وزن در سال ۲۰۱۱ (کیلوگرم)	Weight 2011
عددی	فشار خون سیستولی (میلیمتر جیوه)	Systol 2006
عددی	فشار خون دیاستولی (میلیمتر جیوه)	diastol2006
عددی	صرف روزانه نمک (۱= کمتر از ۶ گرم، ۲= بین ۶ تا ۱۲ گرم، ۳= بین ۱۲ تا ۱۸ گرم، ۴= بیشتر از ۱۸ گرم)	Salt

خصیصه سابقه فشار خون بالا به عنوان خصیصه تشخیص (برچسب) در نظر گرفته شده است.

بخش چهارم: پیش پردازش داده ها

در مهمترین گام تحقیق (آماده سازی داده ها یا پیش پردازش داده ها) به بررسی داده ها پرداختیم. در جهان واقعی، داده همیشه کامل نیست و در مورد اطلاعات پزشکی، این موضوع همیشه درست است. برای حذف تعدادی از تناقض ها و داده های ناقص در ارتباط با داده ها از پردازش داده استفاده کردیم. بعضی از فیلدها مانند قد و وزن که به تنها ی اهمیتی ندارند، بلکه شاخص BMI آن ها تأثیرگذار است، حذف شدند و به جای آن ها از شاخص های مرتبط استفاده شد. شاخص BMI به این صورت محاسبه می شود:

$$BMI = \text{Weight(kg)} / \text{Height(} m^2 \text{)}$$

در نتیجه پس از پالایش داده ها به رکوردهایی با مشخصات جدول زیر رسیدیم.

نوع	توضیحات	مشخصه
عددی	سن	Age
اسمی	جنسیت	Gender
اسمی	استعمال دخانیات	Smoke2006
اسمی	صرف مشروبات الکلی	Drink2006
اسمی	سابقه فشار خون بالا	Hypertension2011
اسمی	منطقه سکونت	Residence2006
عددی	شاخص توده بدنی در سال ۲۰۰۶	BMI2006
عددی	شاخص توده بدنی در سال ۲۰۱۱	BMI2011
عددی	صرف روزانه نمک	Salt

بخش پنجم: شرح پردازش‌ها

طبقه‌بندی

✓ درخت تصمیم: متد طبقه‌بندی بر اساس تولید درخت تصمیم، یکی از روش‌های یادگیری ماشین به حساب می‌آید که به دلیل استفاده از یک مجموعه آموزشی اولیه، جزء روش‌های یادگیری با ناظر است. برای تولید درخت تصمیم، ابتدا یک مجموعه اولیه در نظر گرفته می‌شود و درخت تصمیم آن ساخته می‌شود. چنانچه این درخت پاسخگوی همه حالات نبود، با انتخاب مجموعه‌ای دیگر، درخت توسعه داده می‌شود. این فرایند تا تکمیل درخت برای پاسخگویی به همه حالات ادامه می‌یابد. درخت تصمیم تولید شده، درختی است که برگ‌های آن کلاس‌های مختلف و گره‌های میانی، ویژگی‌ها و حالات مختلف آنها را نشان می‌دهد.

✓ **Naive Bayse**: تئوری بیز یکی از روش‌های آماری برای طبقه‌بندی به شمار می‌آید. در این روش کلاس‌های مختلف، هر کدام به شکل یک فرضیه دارای احتمال در نظر گرفته می‌شوند. هر رکورد آموزشی جدید، احتمال درست بودن فرضیه‌های پیشین را افزایش و یا کاهش می‌دهد و در نهایت، فرضیاتی که دارای بالاترین احتمال شوند، به عنوان یک کلاس در نظر گرفته شده و برچسبی بر آنها زده می‌شود. این تکنیک با ترکیب تئوری بیز و رابطه سببی بین داده‌ها، به طبقه‌بندی می‌پردازد.

✓ **الگوریتم KNN**: روش K نزدیک‌ترین همسایه یک گروه شامل K رکورد از مجموعه رکوردهای آموزشی که نزدیک‌ترین رکوردها به رکورد آزمایشی باشند را انتخاب کرده و بر اساس برتری رده برچسب مربوط به آن‌ها در مورد دسته رکورد آزمایشی مذبور تصمیم‌گیری می‌نماید. به عبارت ساده‌تر این روش رده‌ای را انتخاب می‌کند که در همسایگی انتخاب شده بیشترین تعداد رکورد مناسب به آن دسته باشند. بنابراین رده‌ای که از همه رده‌ها بیشتر در بین K نزدیک‌ترین همسایه مشاهده شود، به عنوان رده رکورد جدید در نظر گرفته می‌شود.

استفاده از الگوریتم KNN نیازمند تعیین سه موضوع می‌باشد:

- باید یک مجموعه رکورد داشته باشیم.
- یک معیار محاسبه شباهت نیز باید داشته باشیم.
- مقدار K نیز باید مشخص شود تا بتوان بر اساس آن عمل نمود.

خوشه بندی

✓ **الگوریتم K-Means :** روش K-Means یکی از روش های خوشه بندی داده ها در داده کاوی است. این

روش علی رغم سادگی آن یک روش پایه برای بسیاری از روش های خوشه بندی دیگر (مانند خوشه بندی فازی) محسوب می شود. این روش روشنی انحصاری و مسطح محسوب می شود. برای این الگوریتم شکل های مختلفی بیان شده است. ولی همه آنها دارای روالی تکراری هستند. در الگوریتم K-means ابتدا k عضو (که k تعداد خوشه ها است) بصورت تصادفی از میان n عضو به عنوان مراکز خوشه ها انتخاب می شود. سپس n-k عضو باقیمانده به نزدیک ترین خوشه تخصیص می یابند. بعد از تخصیص همه اعضا مراکز خوشه مجدداً محاسبه می شوند و با توجه به مراکز جدید به خوشه ها تخصیص می یابند و این کار تا زمانی که مراکز خوشه ها ثابت بماند ادامه می یابد.

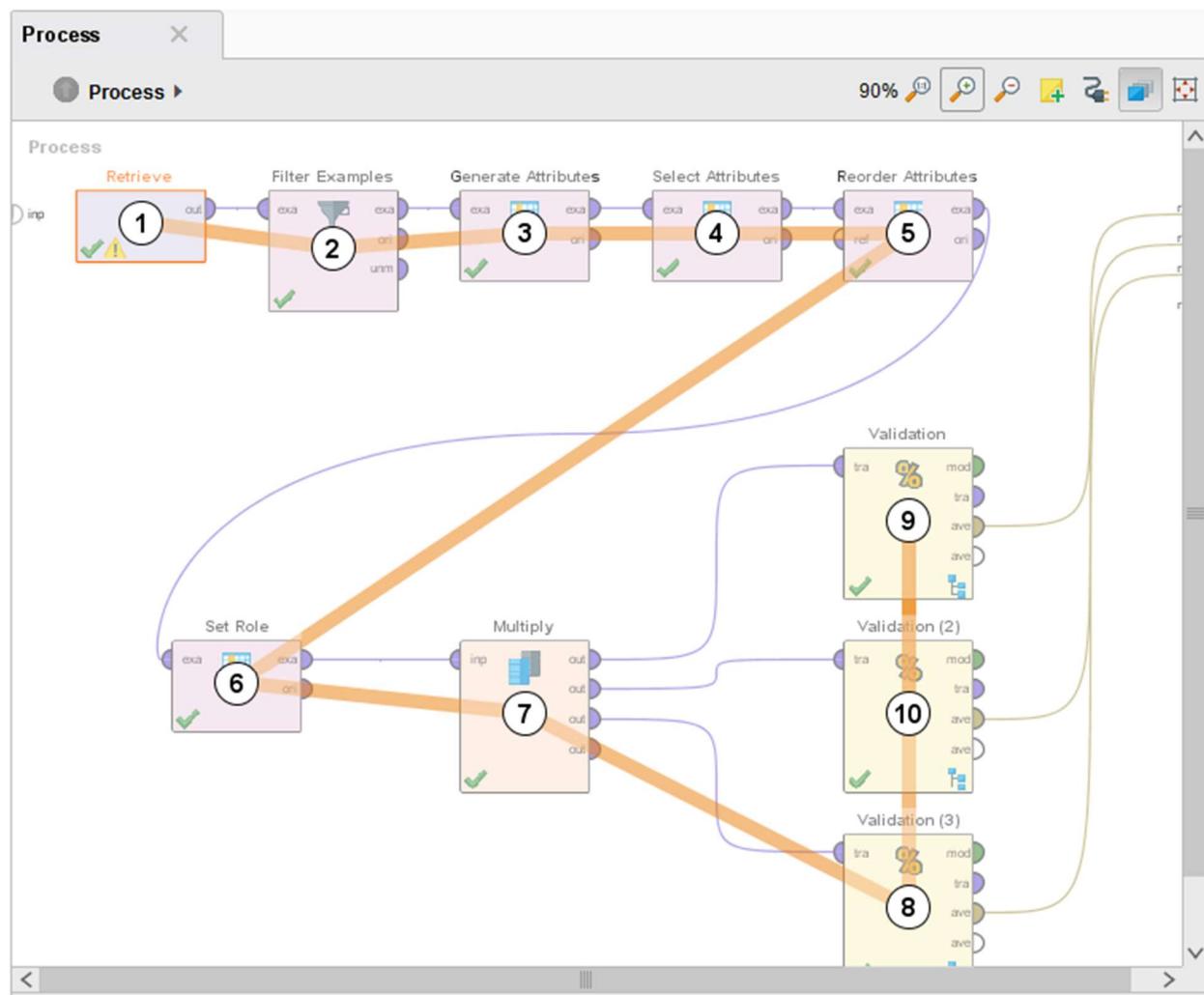
✓ **الگوریتم k-medoids :** الگوریتم k-medoids مبتنی بر شی می باشد و نماینده خوشه ها را

از میان خود داده ها و نه میانگین گیری از آن ها انتخاب می کند. در واقع medoids یک خوشه، مرکزی ترین عنصر یک خوشه است. هدف این روش، کم کردن حساسیت نسبت به مقادیر بزرگ در مجموعه داده هاست. در این الگوریتم هر خوشه با یکی از داده های نزدیک به مرکز معرفی می شود.

بخش ششم: پیاده سازی

طبقه بندی

با استفاده از نرم افزار Rapidminer به طبقه بندی داده ها می پردازیم. قبل از هر کاری ابتدا طبق تصویر زیر و عملگر شماره ۱ داده ها به پردازش اضافه شده اند. سپس قبل از انجام هر کاری باید آماده سازی داده ها یا پیش پردازش صورت بگیرد با استفاده از عملگرهای ۲، ۳، ۴، ۵ و ۶ اینکار صورت میگیرد.

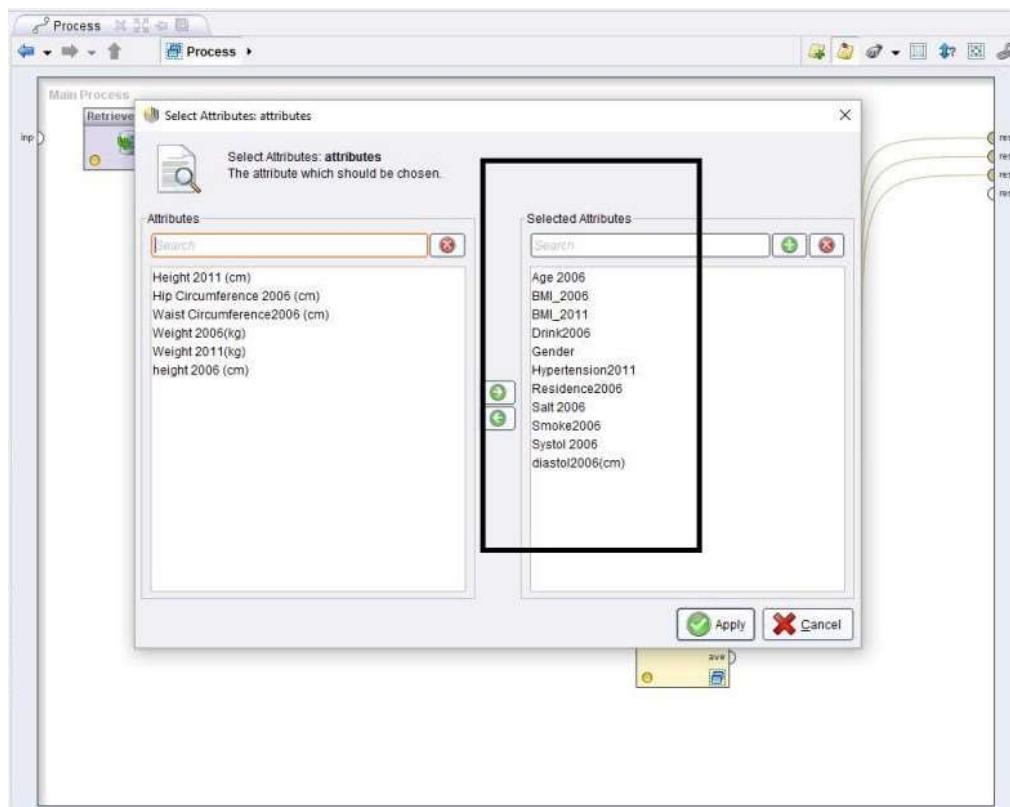


عملگر Retrieve : نخستین موضوع در آماده سازی دادهها مهیا نمودن آن برای عملیاتی است که پس از این بر روی آن انجام میشود با این عملگر دادهها خوانده می شوند برای استفاده از این عملگر باید قبل از این دادهها را در مخزن ذخیره کرده باشید

عملگر Filter Example : به دلایلی ممکن است بعضی از مقادیر مربوط به بعضی از ویژگی ها Null باشد به این گونه از مقادیر دادههای از دست رفته می گوییم. با استفاده از این عملگر دادههای پرت و دادههای از دست رفته (Missing value) حذف می شوند.

عملگر Generate Attributes : در بسیاری از کاربردها شما نیاز به ستونی دارید که با محاسبات خاص و معینی از دیگر ستونها منتج می شود. در این موقع این عملگر انتخاب مناسبی است. در اینجا ما با استفاده از دو خصیصه وزن و قد در سالهای ۲۰۰۶ و ۲۰۱۱ دو خصیصه شاخص توده بدنی در سال ۲۰۰۶ و ۲۰۱۱ را با استفاده از این عملگر ایجاد کردیم.

عملگر Select Attributes : در بسیاری از مواقع شما نیاز به همه صفات خاصه موجود در مجموعه دادهها ندارید و برای تحلیل شما برخی از آنها کافی است. در اینصورت با استفاده از این عملگر دادهها فیلتر می شوند مطابق تصویر زیر صفات خاصه که داخل کادر در سمت راست تصویر مشاهده می کنید برای تحلیل ما انتخاب شدهاند.



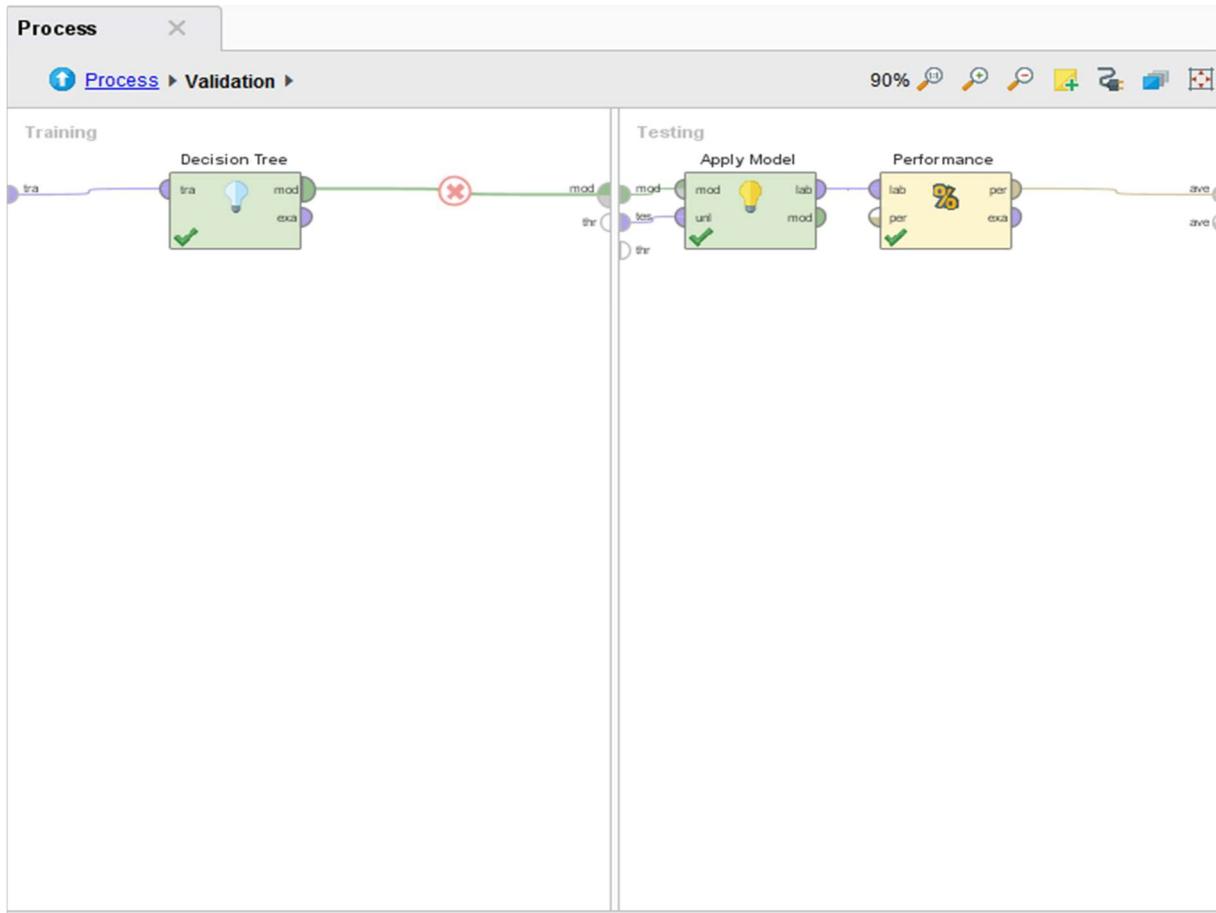
عملگر Reorder Attributes: این عملگر برای جایی ستونها استفاده می‌شود.

عملگر Set Role: یکی از موضوعات مهم برای کار با داده‌ها تعیین نقش برای صفات خاصه است با کمک این عملگر می‌توان برای صفات خاصه نقشی را تعیین کرد. در اینجا برای ستون فشار خون نقش برجسب را انتخاب می‌کنیم.

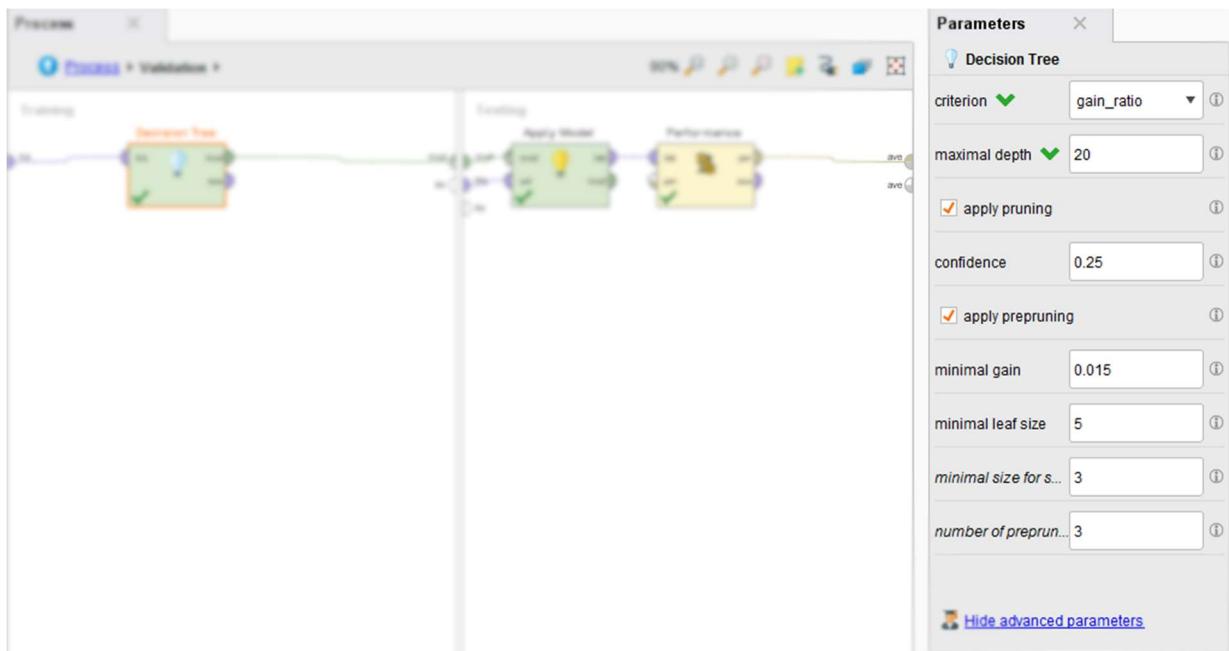
عملگر Multiply: این عملگر هیچ گونه تغییری بر روی داده‌ها اعمال نمی‌کند و برای موقعی مناسب است که مایلید خروجی چند پردازش را برروی داده‌های یکسانی مشاهده کنید.

پس از طی مراحل ۱ تا ۷ داده‌های ما آماده برای انجام پردازش هستند. ما برروی این داده‌ها از سه الگوریتم درخت تصمیم، الگوریتم بیز و الگوریتم KNN استفاده می‌کنیم و نتایج حاصل از هر الگوریتم را ارزیابی کرده و سپس در قسمت پایانی به این نتیجه می‌رسیم که کدام الگوریتم بر روی این مجموعه داده دارای عملکرد مناسب‌تری است.

عملگر Validation: برای اعتبار سنجی مدلی که بدست آورده ایم یا مقایسه چند مدل با هم از این عملگر استفاده می‌شود. یکی از تکنیک‌های رایج و پرکاربرد در برنامه‌های کاربردی جهت ارزیابی مدل به خصوص دسته بندی k-fold cross-validation است. که در اینجا ما از این تکنیک استفاده کردیم و مقدار K را برابر با ۱۰ در نظر گرفتیم. با انتخاب مقدار ۱۰ برای K، مجموعه داده‌ها به ۱۰ قسمت مساوی تقسیم می‌شوند. یک قسمت از آن برای آزمایش کنار گذاشته می‌شود و از ۹ قسمت باقی مانده مدلی طراحی می‌شود این کار برای هر ۱۰ قسمت تکرار می‌شود. دقت و نرخ خطأ از میانگین این ۱۰ تکرار بدست می‌آید. همانطور که در شکل زیر می‌بینیم عملگر X-validation دارای زیر پردازش‌های داخلی است که با دوبار کلیک بر روی این عملگر به قسمت زیر پردازش هدایت می‌شویم و می‌توانیم عملگرهای مناسب را داخل ان قرار دهیم.



- عملگر Validation اول: برای ارزیابی نتایج حاصل از درخت تصمیم بکار رفته است وقتی وارد زیر پردازش این عملگر می شویم مشاهده می کنید طبق شکل بالا پردازش به دو قسمت تقسیم شده یک قسمت مربوط به ساخت مدل (Training) و قسمت بعدی برای تست کردن مدل ساخته شده (Testing) است. در قسمت ساخت مدل از عملگر درخت تصمیم استفاده می کنیم و مقدار پارامترهای آنرا مانند تصویر زیر تنظیم می کنیم.



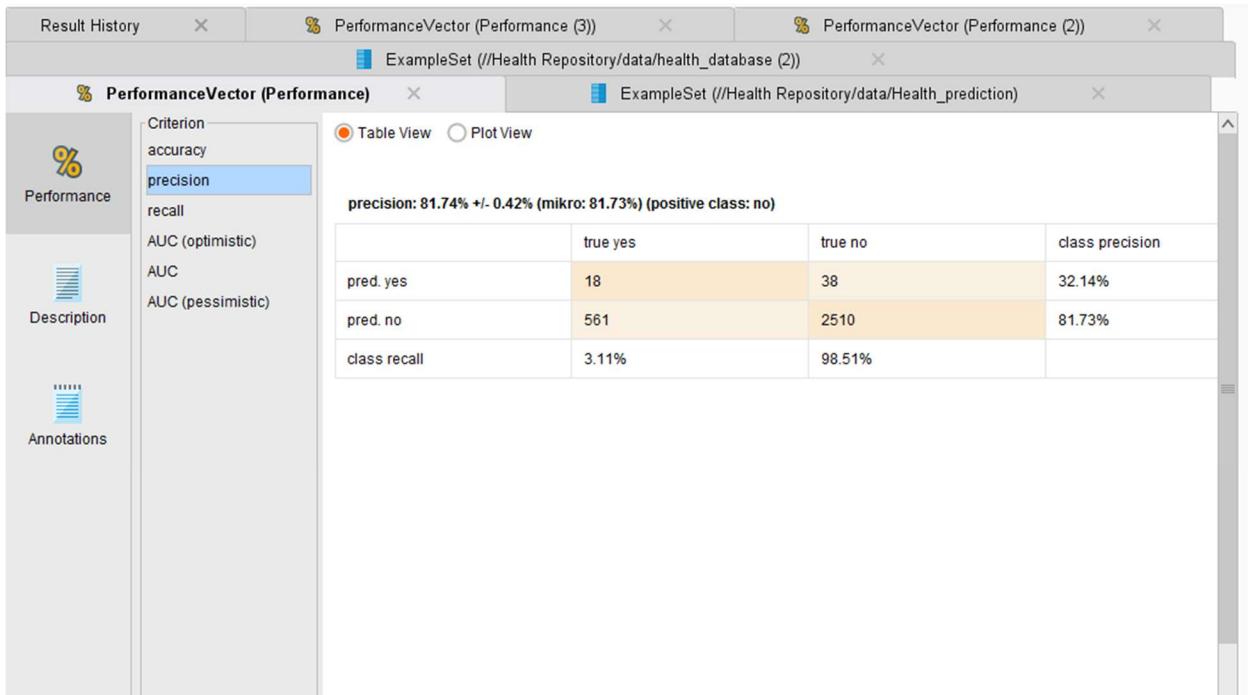
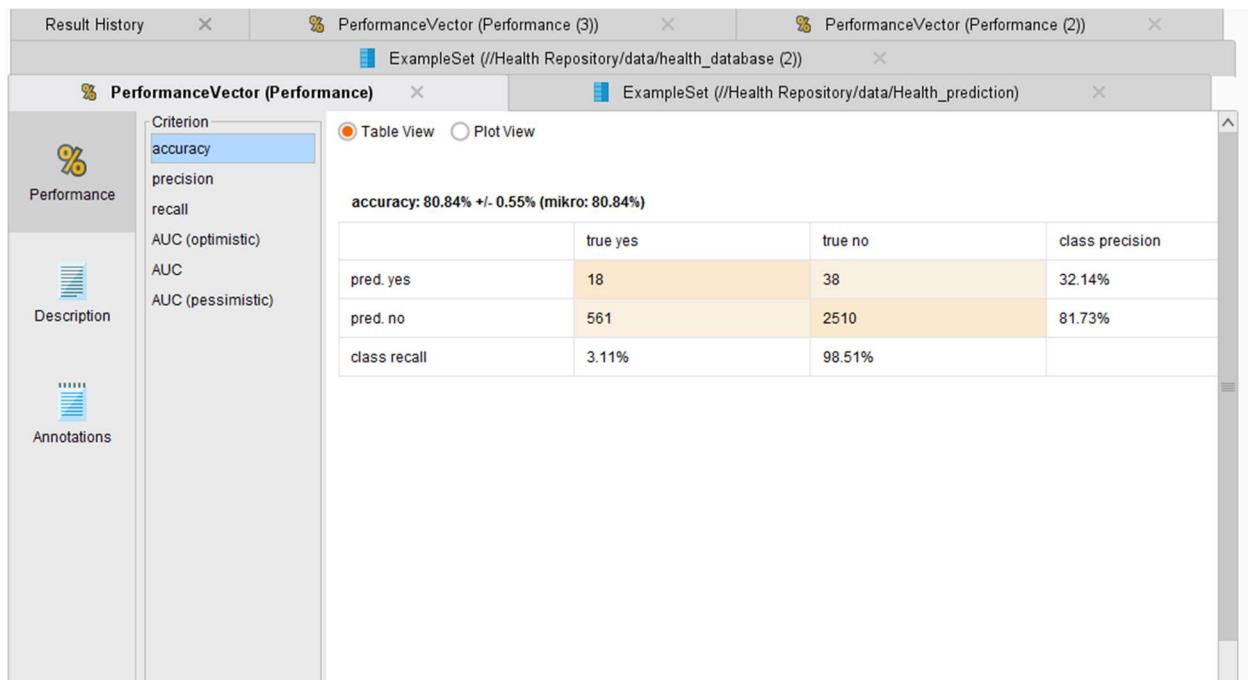
در قسمت تست از دو عملگر Performance و Apply Model استفاده شده است.

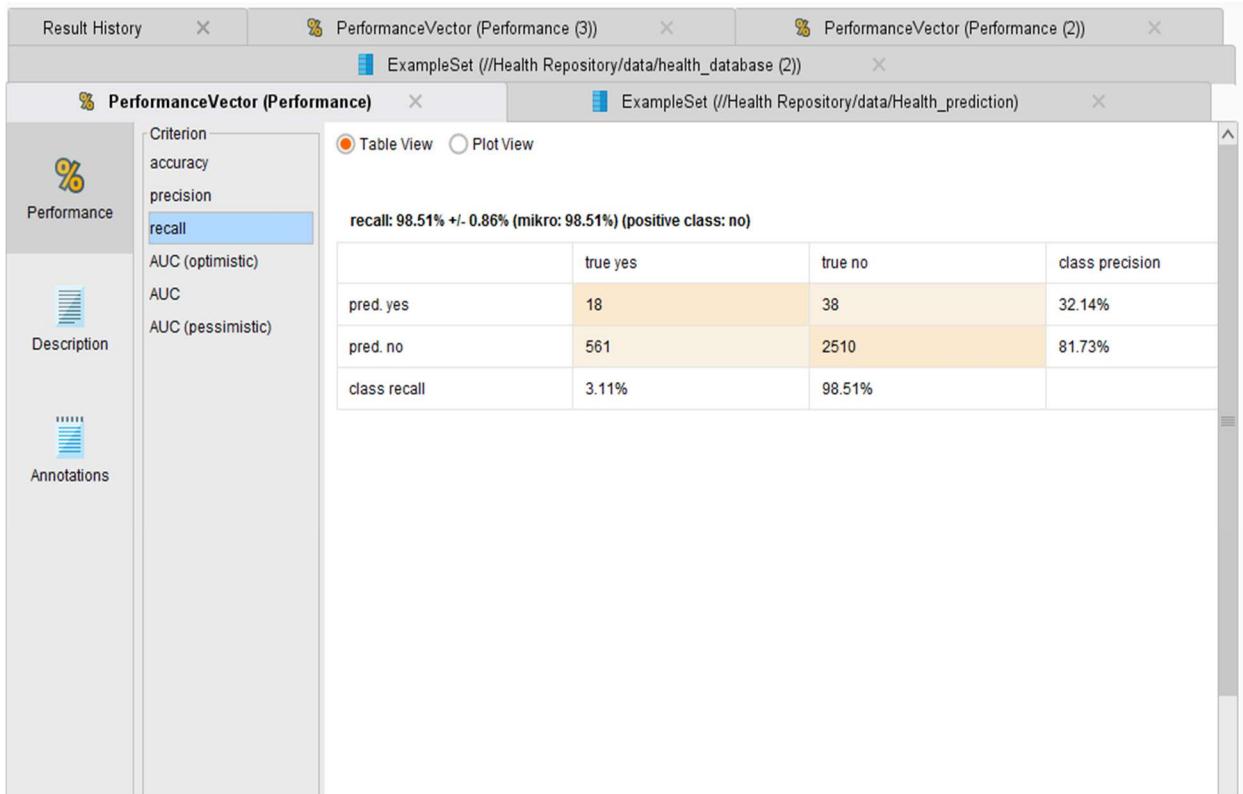
عملگر Apply Model: برای اعمال مدل ساخته شده در مرحله قبل به مجموعه تست استفاده می‌شود پس خروجی عملگر Decision Tree به ورودی این عملگر وصل می‌شود.

عملگر Performance : باستفاده از این عملگر که معیارهای ارزیابی مختلف را می‌توان مشخص کرد که پس از اجرای پردازش می‌توان در خروجی نتایج حاصل از آن را مشاهده کرد. خروجی این عملگر به وسیله پورت Ave به عملگر پدرش در مرحله قبل وصل می‌شود.

خروجی حاصل از عملگر Validation اول (Decision Tree)

در شکل‌های زیر اطلاعات مفیدی از ارزشیابی مدل وجود دارد. دقت مدل حدود ۸۰,۸۴ درصد تخمین زده شده است که با تغییر پارامترها می‌توان به دقت‌های دیگری دست پیدا کرد. در سمت چپ تصویر معیارهایی که درخواست کردید نشان داده می‌شود و با کلیک بر روی هر یک می‌توان مقدار آنرا مشاهده کنید. جدول میان تصویر اطلاعات دقیق تری از تخمین برچسب کلاس نمونه‌ها نشان می‌دهد.

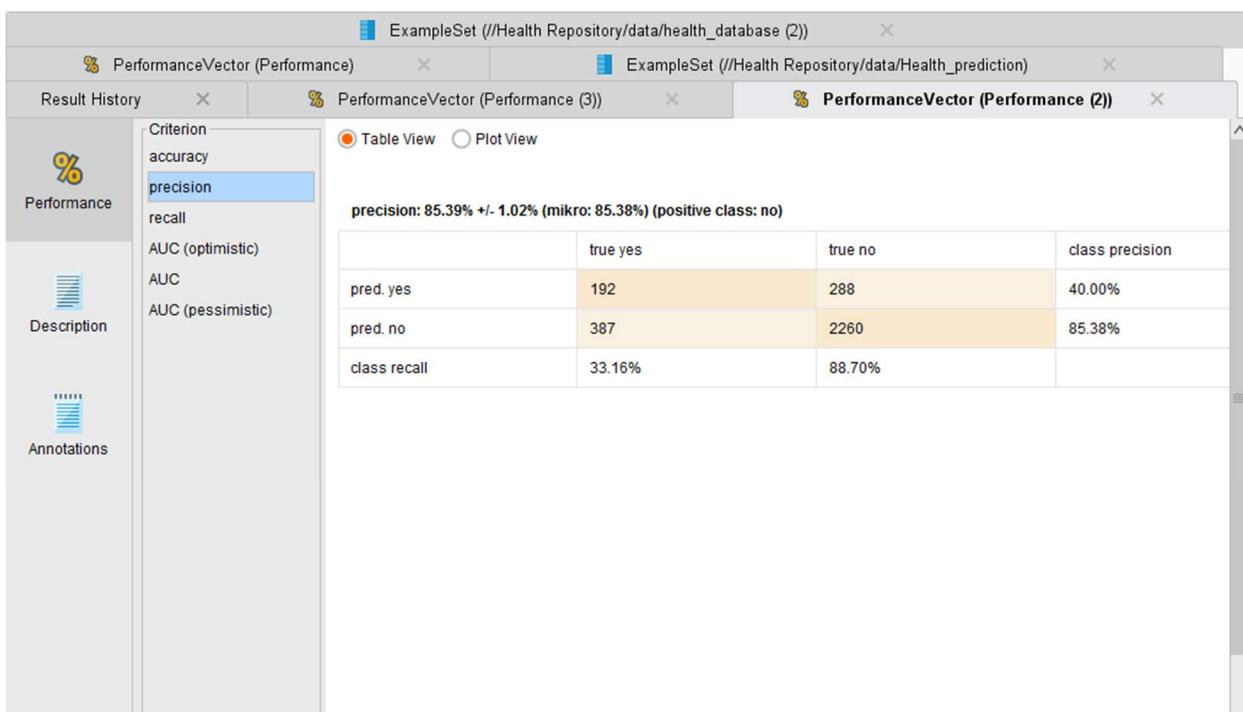
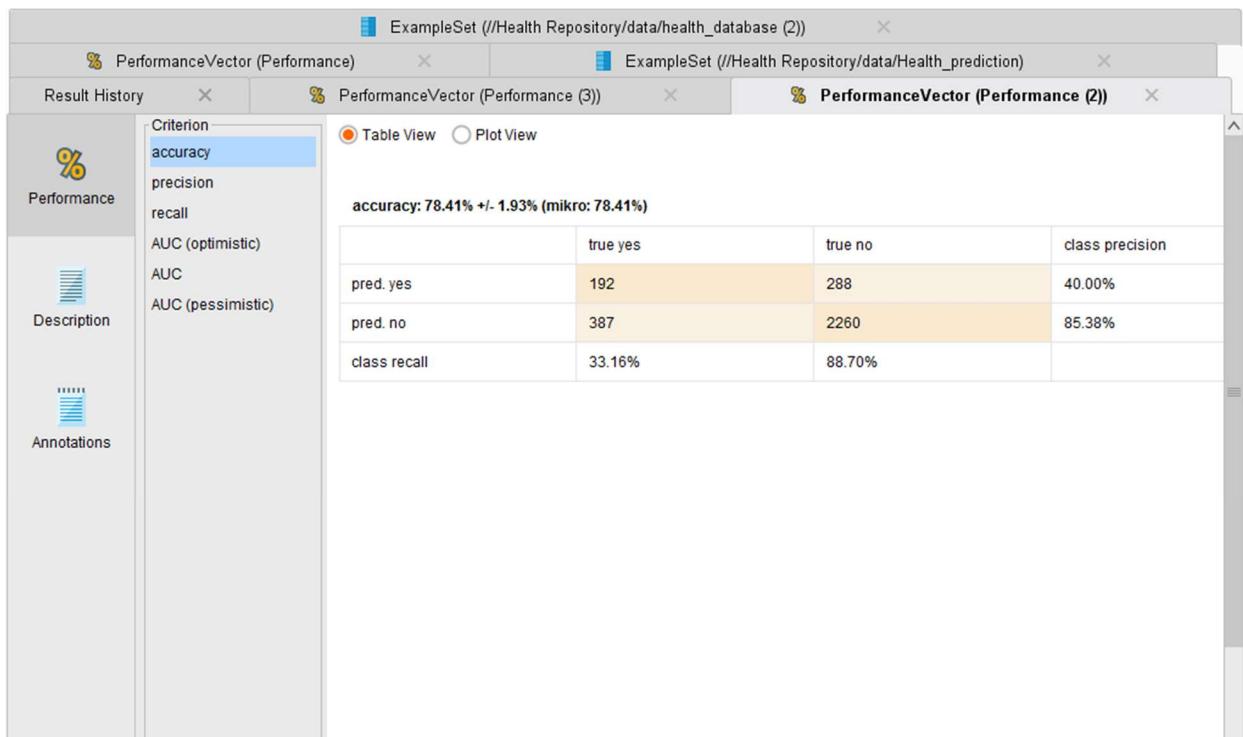


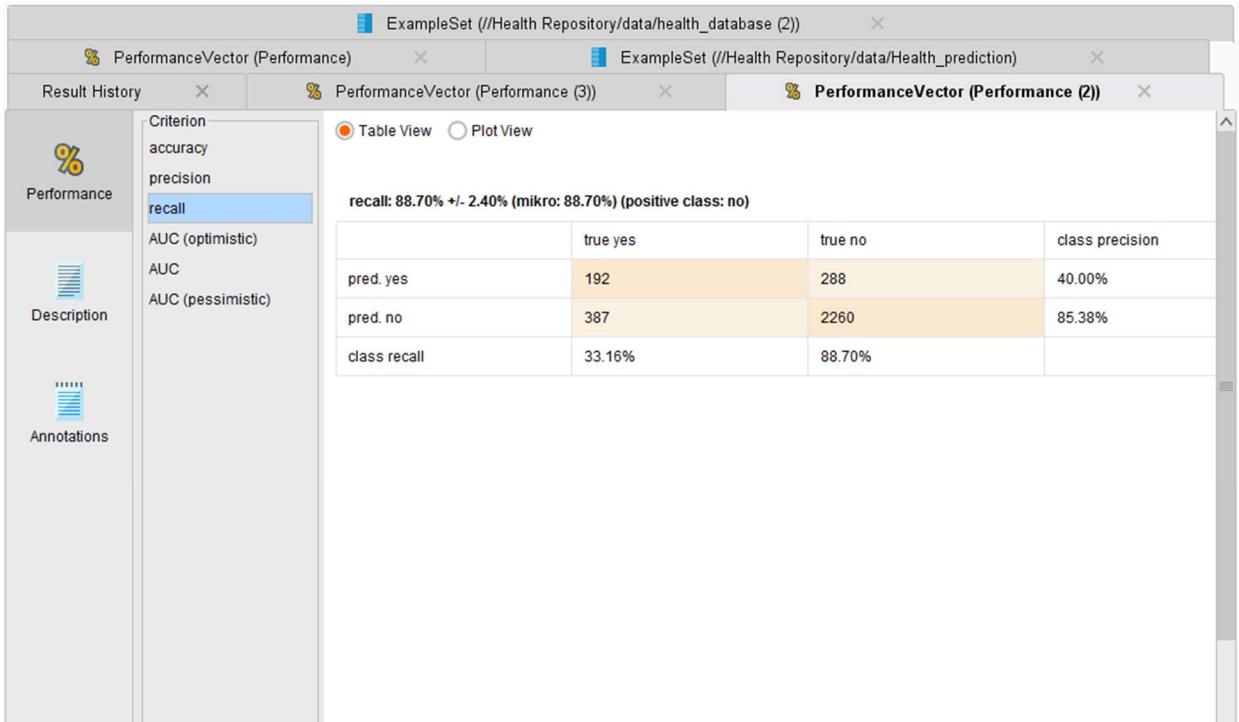


عملگر Validation دوم: برای ارزیابی نتایج حاصل از الگوریتم بیز بکار رفته است وقتی وارد زیر پردازش این عملگر می شویم مشاهده می کنید طبق تصویر بالا پردازش به دو قسمت تقسیم شده یک قسمت مربوط به ساخت مدل (Training) و قسمت بعدی برای تست کردن مدل ساخته شده (Testing) است. در قسمت ساخت مدل از عملگر Naive Bayes استفاده می کنیم و در قسمت تست از دو عملگر Apply Model و Performance دادیم استفاده شده است.

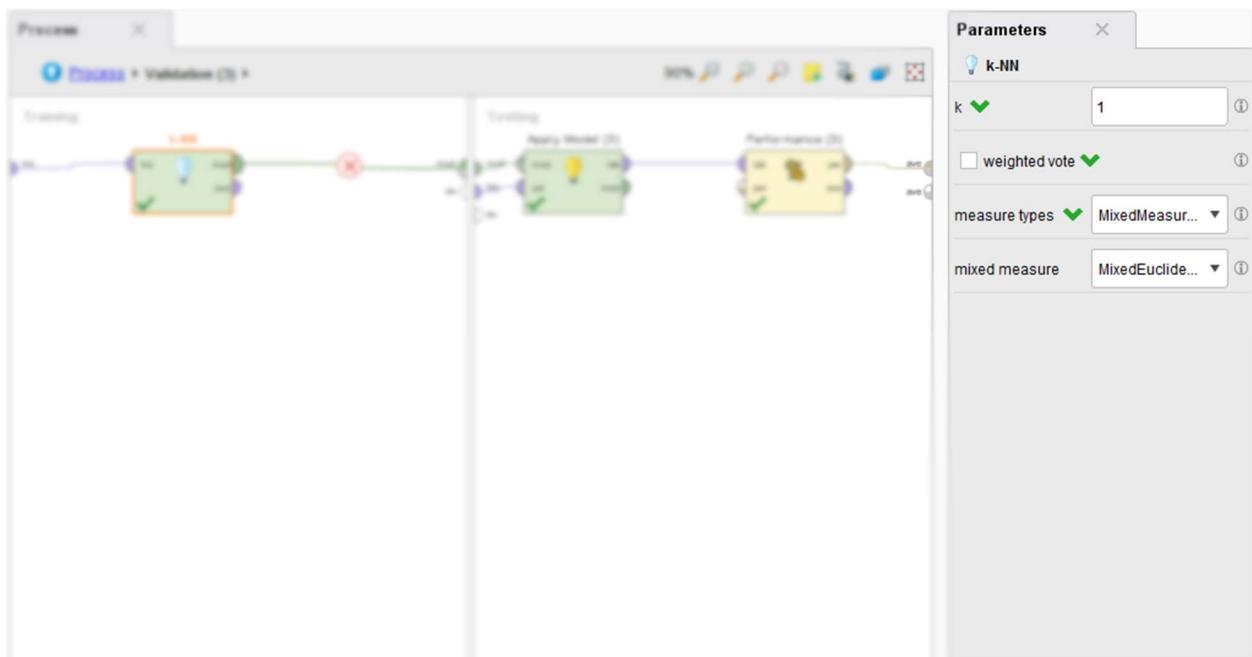
خروجی حاصل از عملگر Validation دوم (Naive Bayes)

طبق تصویر زیر دقیق در این روش ۷۸,۴۱ درصد تخمین زده شده است و همینطور معیارهای دیگر بر اساس تصاویر زیر مشخص است.





- عملگر Validation سوم: برای ارزیابی نتایج حاصل از الگوریتم K-NN بکار رفته است وقتی وارد زیر پردازش این عملگر می‌شویم مشاهده می‌کنید طبق تصویر بالا پردازش به دو قسمت تقسیم شده یک قسمت مربوط به ساخت مدل (Training) و قسمت بعدی برای تست کردن مدل ساخته شده (Testing) است. در قسمت ساخت مدل از عملگر K-NN استفاده می‌کنیم مقدار پارامترهای آنرا مانند تصویر زیر تنظیم می‌کنیم. و در قسمت تست از دو عملگر Apply Model و Performance که قبلاً توضیح دادیم استفاده شده است.



خروجی حاصل از عملگر Validation سوم (K-NN)

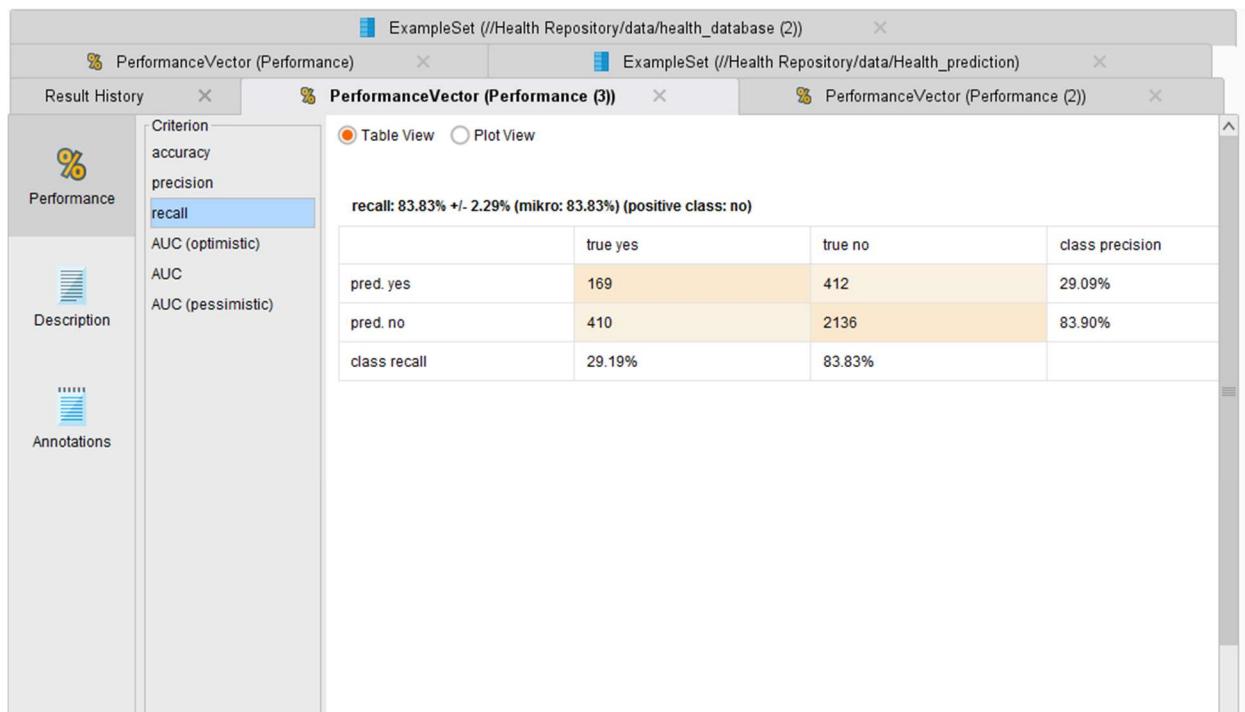
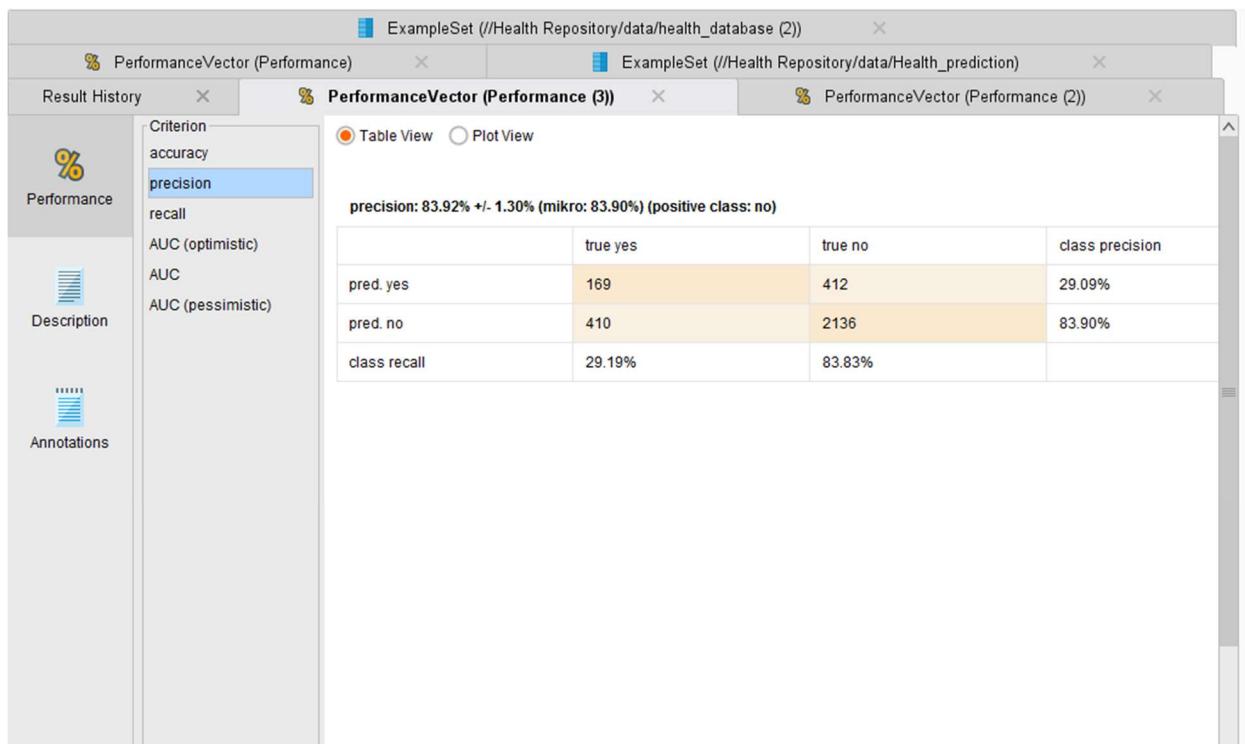
نتایج حاصل از این عملگر در تصاویر زیر مشخص است.

The screenshot shows the KNIME interface with three open windows:

- PerformanceVector (Performance)**: This window is active and displays performance metrics for a single dataset. The sidebar on the left shows categories like Criterion, Performance, Description, and Annotations. Under Criterion, 'accuracy' is selected, highlighted in blue. The main content area shows the accuracy as 73.71% +/- 1.92% (mikro: 73.71%). Below this is a confusion matrix table:

	true yes	true no	class precision
pred. yes	169	412	29.09%
pred. no	410	2136	83.90%
class recall	29.19%	83.83%	

- PerformanceVector (Performance (3))**: Shows results for three datasets.
- PerformanceVector (Performance (2))**: Shows results for two datasets.



نتیجه گیری

مقایسه سه الگوریتم دسته بندی بکار رفته:

	Decision Tree ✓	Naive Bayes	KNN
Accuracy	80.84%	78.41%	73.71%
Precision	81.74%	85.39%	83.92%
Recall	98.51%	88.70%	83.83%
AUC	0.638	0.731	0.500

مهمترین معیار برای تعیین کارایی یک الگوریتم دسته بندی دقت یا نرخ دسته بندی Accuracy است که این معیار دقت کل یک دسته بند را محاسبه می کند در واقع این معیار مشهورترین و عمومی ترین معیار محاسبه کارایی الگوریتم های دسته بندی است که نشان می دهد، دسته بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را بدسترسی دسته بندی کرده است.

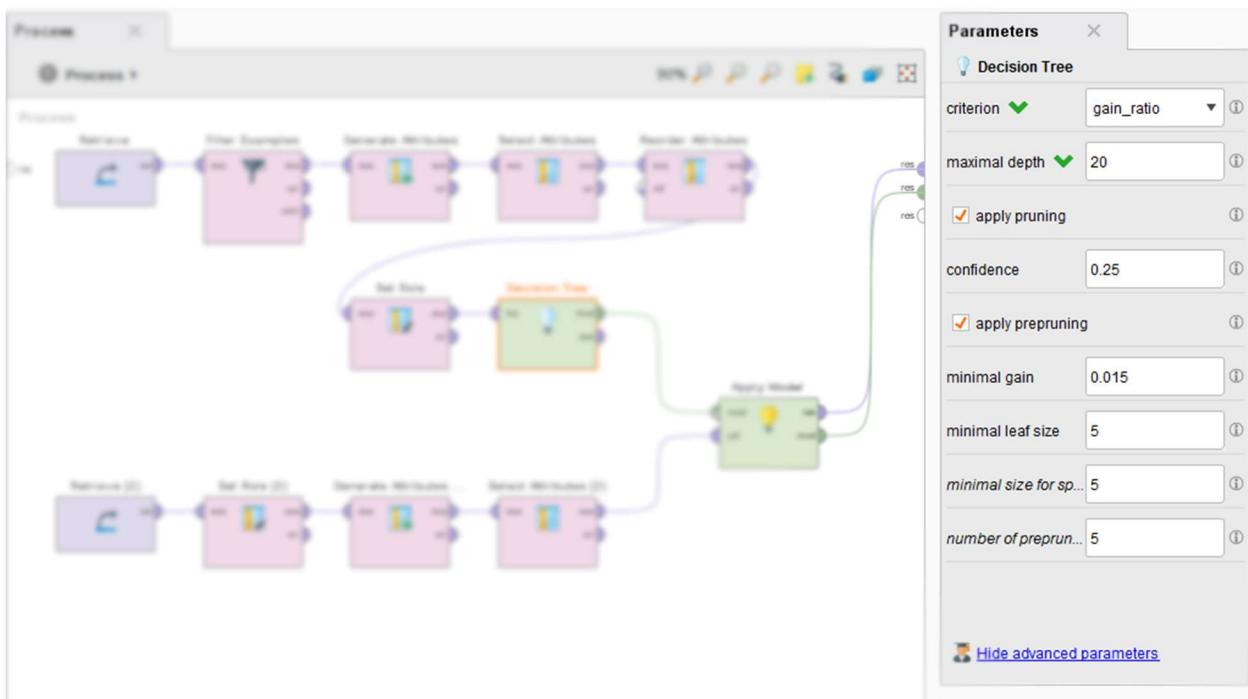
معیار مهم دیگری که برای تعیین میزان کارایی یک دسته بند استفاده می شود معیار AUC یا (Area Under Curve) است.

AUC نشان دهنده سطح زیر نمودار می باشد که هر چه مقدار این عدد مربوط به یک دسته بند بزرگتر باشد کارایی نهایی دسته بند مطلوب تر ارزیابی می شود.

در جدول بالا نتایج حاصل از اعمال سه الگوریتم دسته بندی به یک مجموعه از داده وجود دارد. همانطور که مشاهده می کنید درخت تصمیم دارای Recall و Accuracy بیشتری نسبت به دو الگوریتم دیگر است پس اعمال الگوریتم درخت تصمیم بر روی این مجموعه داده دارای نتایج مطلوب تری نسبت به دو الگوریتم دیگر است.

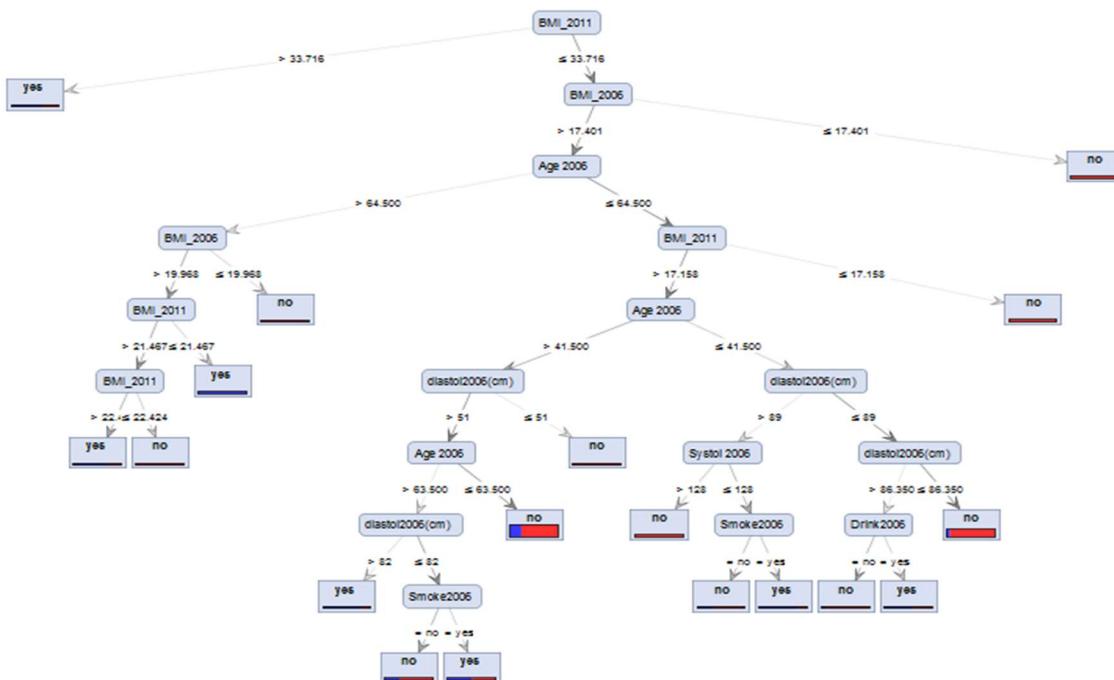
تخمین برچسب داده های آزمایشی با استفاده از مدل درخت تصمیم

در اینجا ما یک مجموعه داده ۱۰۰ تایی داریم که مقدار خصیصه فشار خون آنها نامشخص است می خواهیم با استفاده از مدل درخت تصمیم ساخته شده مقدار این خصیصه را برای این مجموعه از داده ها پیش بینی کنیم. مقادیر موجود برای صفات خاصه داده های آزمایشی باید در محدوده مقادیر صفات خاصه در داده های آموزشی باشد اگر اینچنان نبود با استفاده از عملگر Filter Example داده های آزمایشی خارج از محدوده را حذف می کنیم. عملگرهای Set Role و Reorder Attributes ، Select Attributes ، Generate Attributes طبق توضیحاتی که در بخش قبل دادم برای پیش پردازش داده ها استفاده شده اند. سپس عملگر درخت تصمیم به مجموعه داده آموزشی اعمال می شود و پارامترهای آن مانند تصویر زیر تعیین می شوند.



سپس خروجی عملگر درخت تصمیم و خروجی حاصل از پیش پردازش داده های آزمایشی به عملگر Apply Model برای تخمین برچسب داده های آزمایشی اعمال می شود. در ادامه خروجی حاصل از این پردازش را مشاهده می کنیم.

Result History		Tree (Decision Tree)		ExampleSet (Select Attributes (2))					
Data	ExampleSet (100 examples, 4 special attributes, 10 regular attributes)							Filter (100 / 100 examples): all	
	Row No.	Hypertensio...	prediction(H...)	confidence(yes)	confidence(no)	Smoke2006	Drink2006	diastol2006(...	
1	?	no	0.087	0.913	yes	no	73.300		
2	?	no	0.170	0.830	yes	no	82.700		
3	?	no	0.167	0.833	yes	no	80		
4	?	no	0.246	0.754	yes	no	79.300		
5	?	no	0.246	0.754	yes	no	80		
6	?	no	0.389	0.611	yes	no	89.300		
7	?	no	0.371	0.629	yes	no	79.300		
8	?	no	0.014	0.986	yes	no	77.300		
9	?	yes	0.538	0.462	yes	no	80		
10	?	no	0.444	0.556	yes	no	88.700		
11	?	no	0.167	0.833	yes	no	76.700		
12	?	no	0.246	0.754	yes	no	78.700		
13	?	no	0.246	0.754	yes	no	78.700		
14	?	no	0.022	0.978	yes	no	64		
15	?	no	0	1	yes	no	71.300		
16	?	no	0.022	0.978	yes	no	60.700		
17	?	yes	0.500	0.500	yes	no	70.300		



با نگاه به تصویر بالا می بینیم که شاخص توده بدنی در سطح اول قرار دارد. بیشترین اطلاعات از طریق این شاخص بدست می آید و بوسیله این شاخص سریعتر به برچسب میرسیم. بنابر این خصیصه شاخص توده بدنی بیشترین تاثیر را بر روی برچسب کلاس دارد و بیشترین تقسیم بندی بر روی درخت بوسیله این شاخص صورت گرفته است. از ۱۷ انشعاب که بر روی درخت وجود دارد ۷ مورد از آن به شاخص توده بدنی مربوط می باشد که این باز هم نشان از اهمیت این شاخص دارد.

پیش بینی نمونه شماره ۱۳:

Row ...	Hyperten...	prediction...	confidence(yes)	confidence(no)	Smok...	Dri...	diast...	Systol...	Resi...	Gen...	Age ...	Sal...	BMI_2006	BMI_2011
1	?	no	0.076	0.924	yes	no	73.300	111.300	urban	female	32	0	25.520	25.559
2	?	no	0.076	0.924	yes	no	82.700	116.700	urban	female	40	2	25.872	24.550
3	?	no	0.238	0.762	yes	no	80	120.700	urban	female	60	2	25.390	25.921
4	?	no	0.238	0.762	yes	no	79.300	129.300	urban	female	47	1	26.535	27.042
5	?	no	0.238	0.762	yes	no	80	130.700	urban	female	51	1	24.802	20.063
6	?	no	0.238	0.762	yes	no	89.300	126	urban	female	54	2	27.682	25.545
7	?	no	0.238	0.762	yes	no	79.300	110	urban	female	43	4	29.689	26.101
8	?	no	0.076	0.924	yes	no	77.300	111.300	urban	male	28	2	21.096	17.706
9	?	no	0.238	0.762	yes	no	80	129.300	urban	male	59	1	23.951	22.761
10	?	no	0.238	0.762	yes	no	88.700	130.700	urban	male	57	3	23.033	25.991
11	?	no	0.238	0.762	yes	no	76.700	120.700	urban	male	61	2	24.788	23.914
12	?	no	0.238	0.762	yes	no	78.700	127.300	urban	male	53	2	28.217	29.511
13	?	no	0.238	0.762	yes	no	78.700	100	urban	male	54	2	26.873	29.016
14	?	no	0.076	0.924	yes	no	64	98	rural	female	41	2	25.712	25.712
15	?	no	0.238	0.762	yes	no	71.300	100	rural	female	44	2	24.965	27.774
16	?	no	0.076	0.924	yes	no	60.700	92	rural	female	36	2	25.974	25.888
17	?	yes	0.500	0.500	yes	no	79.300	120	rural	female	64	2	26.067	25.606

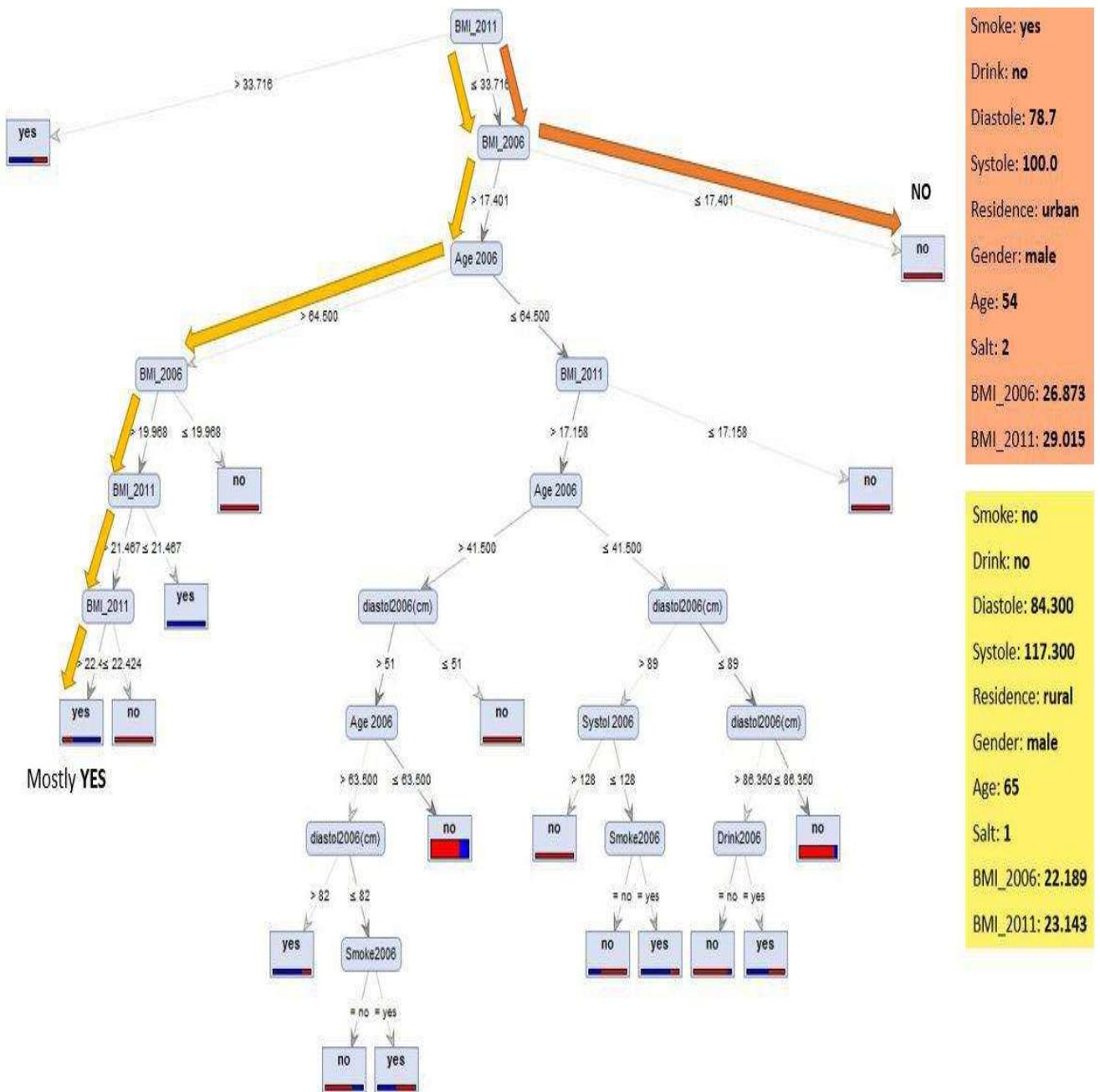
همانطور که در تصویر بالا مشاهده می کنید آقایی ۵۴ ساله که در منطقه شهری زندگی می کند و دارای فشار خون ۱۰ بر روی ۷ می باشد و از دخانیات استفاده می کند. همچنین دارای تغییر شاخص توده بدنی از مقدار ۲۶,۸۷ در سال ۲۰۰۶ به میزان ۲۹,۰۱ در سال ۲۰۱۱ است به احتمال ۰,۷۲ دارای بیماری فشار خون نیست و به احتمال ۰,۲۳ دارای بیماری فشار خون است که در این نمونه پیش بینی شده است که مبتلا به فشار خون نیست.

پیش بینی نمونه شماره ۴۱:

Row ...	Hyperten...	prediction...	confidence(yes)	confidence(no)	Smok...	Dri...	diast...	Systol...	Resi...	Gen...	Age ...	Sal...	BMI_2006	BMI_2011
31	?	no	0.238	0.762	yes	no	80	130	rural	female	59	2	26.317	25.862
32	?	no	0.076	0.924	no	no	64.300	91.700	rural	male	39	2	22.477	21.926
33	?	no	0.238	0.762	no	no	70	120	rural	male	56	3	21.359	22.980
34	?	no	0.076	0.924	no	no	60	100	rural	male	39	3	22.189	22.773
35	?	no	0.238	0.762	no	no	77.700	118	rural	male	50	2	21.333	21.621
36	?	no	0.238	0.762	no	no	80	124	rural	male	51	4	20.957	21.565
37	?	no	0.238	0.762	no	no	70	110	rural	male	49	1	22.575	22.051
38	?	no	0	1	no	no	79.300	112.700	rural	male	28	0	21.083	16.437
39	?	no	0.238	0.762	no	no	68.700	126	rural	male	61	2	21.641	19.411
40	?	no	0.238	0.762	no	no	72	118.700	rural	male	57	1	21.504	21.930
41	?	yes	0.727	0.273	no	no	84.300	117.300	rural	male	65	1	22.189	23.143
42	?	no	0.076	0.924	no	no	62	98	rural	male	34	1	22.600	26.187
43	?	no	0.076	0.924	no	no	78	98	rural	male	28	2	20.812	22.546
44	?	no	0.076	0.924	no	no	74	106	rural	male	34	3	21.641	21.395
45	?	no	0.076	0.924	no	no	60.700	92.700	rural	male	38	2	21.708	21.110
46	?	no	0.238	0.762	no	no	61.300	92.700	rural	male	44	2	22.107	21.852
47	?	no	0.238	0.762	no	no	86	110	rural	male	50	1	21.929	22.171
..

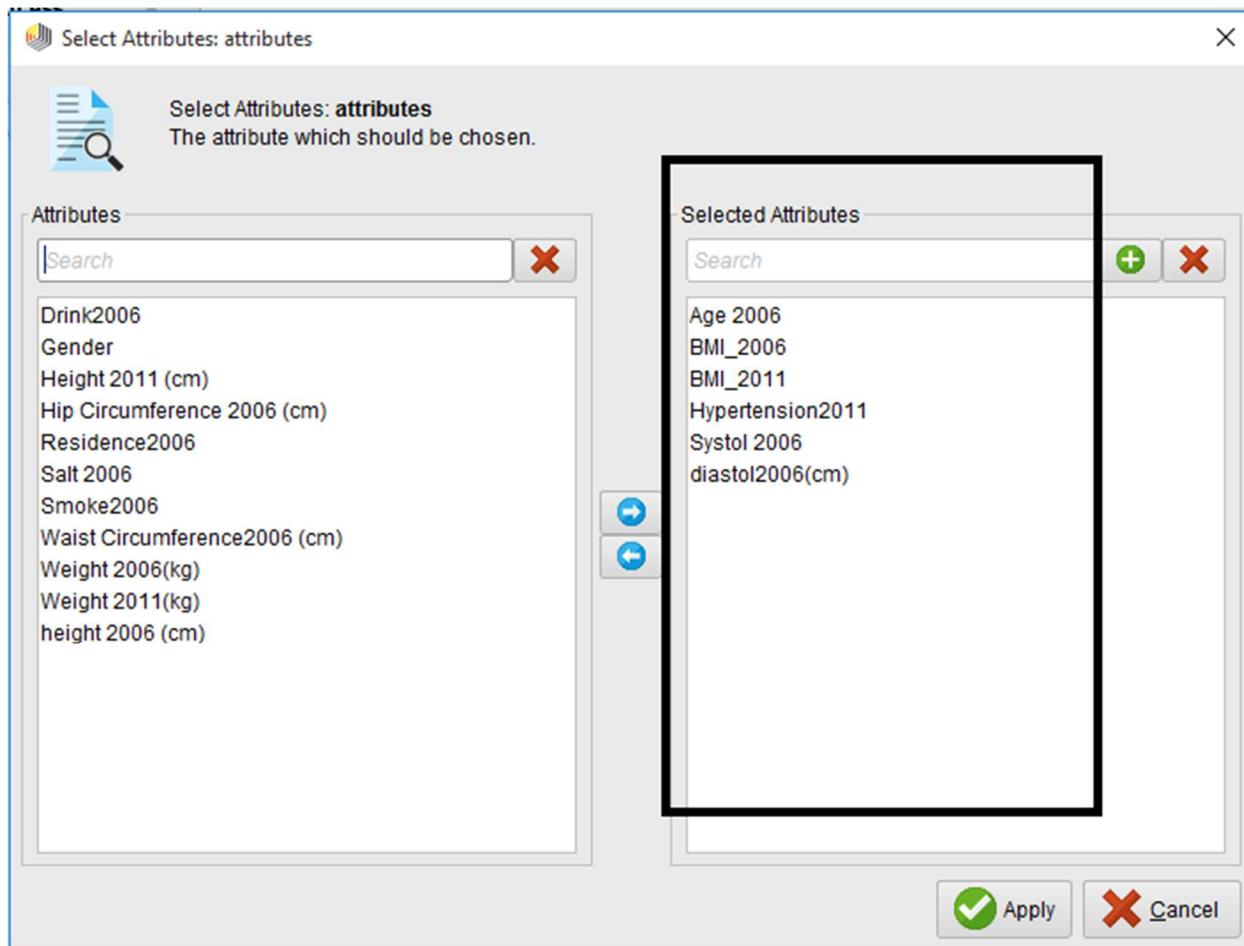
در این نمونه مشاهده می کنید که آقایی ۶۵ ساله که در منطقه روستایی زندگی می کند و دارای فشار خون ۱۱ بر روی ۸ است و میزان تغییر شاخص توده بدنی از ۲۲,۱۸ به مقدار ۲۳,۱۴ است به احتمال ۰,۲۷ فشار خون ندارد و به احتمال ۰,۷۲ دارای فشار خون است که اینجا پیش بینی شده که فشار خون دارد.

مسیر دو نمونه بالا بر روی درخت را در تصویر زیر مشاهده می کنید.

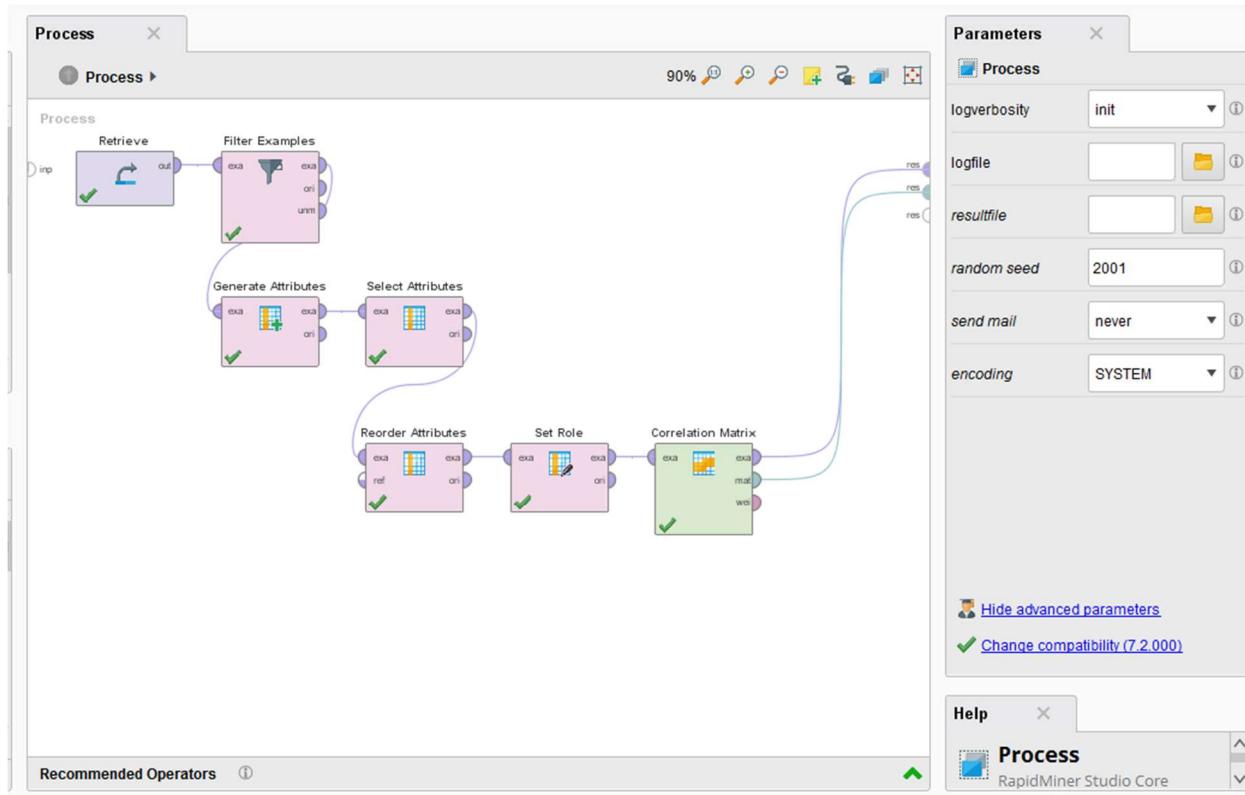


میزان همبستگی خصیصه ها

در این پردازش قصد داریم میزان همبستگی خصیصه ها نسبت به هم را بدست آوریم. عملگرهای ابتدایی برای انجام پیش پردازش و اماده سازی داده ها بر روی مجموعه داده اعمال می شوند. عملگر select Attribute مطابق تصویر زیر برای مشخص کردن خصیصه هایی که میخواهیم میزان همبستگی آنها را نسبت بهم بدست آوریم بکار می رود.



- سپس خروجی آنها مشابه تصویر زیر به عملگر correlation matrix برای محاسبه میزان همبستگی وصل می شود.



پس از اجرای پردازش بالا ماتریس همبستگی زیر بدست می آید.

Attributes	BMI_2006	BMI_2011	Systol 2006	diastol2006(cm)	Age 2006
BMI_2006	1	0.820	0.255	0.240	0.089
BMI_2011	0.820	1	0.208	0.205	0.011
Systol 2006	0.255	0.208	1	0.650	0.201
diastol2006(cm)	0.240	0.205	0.650	1	0.105
Age 2006	0.089	0.011	0.201	0.105	1

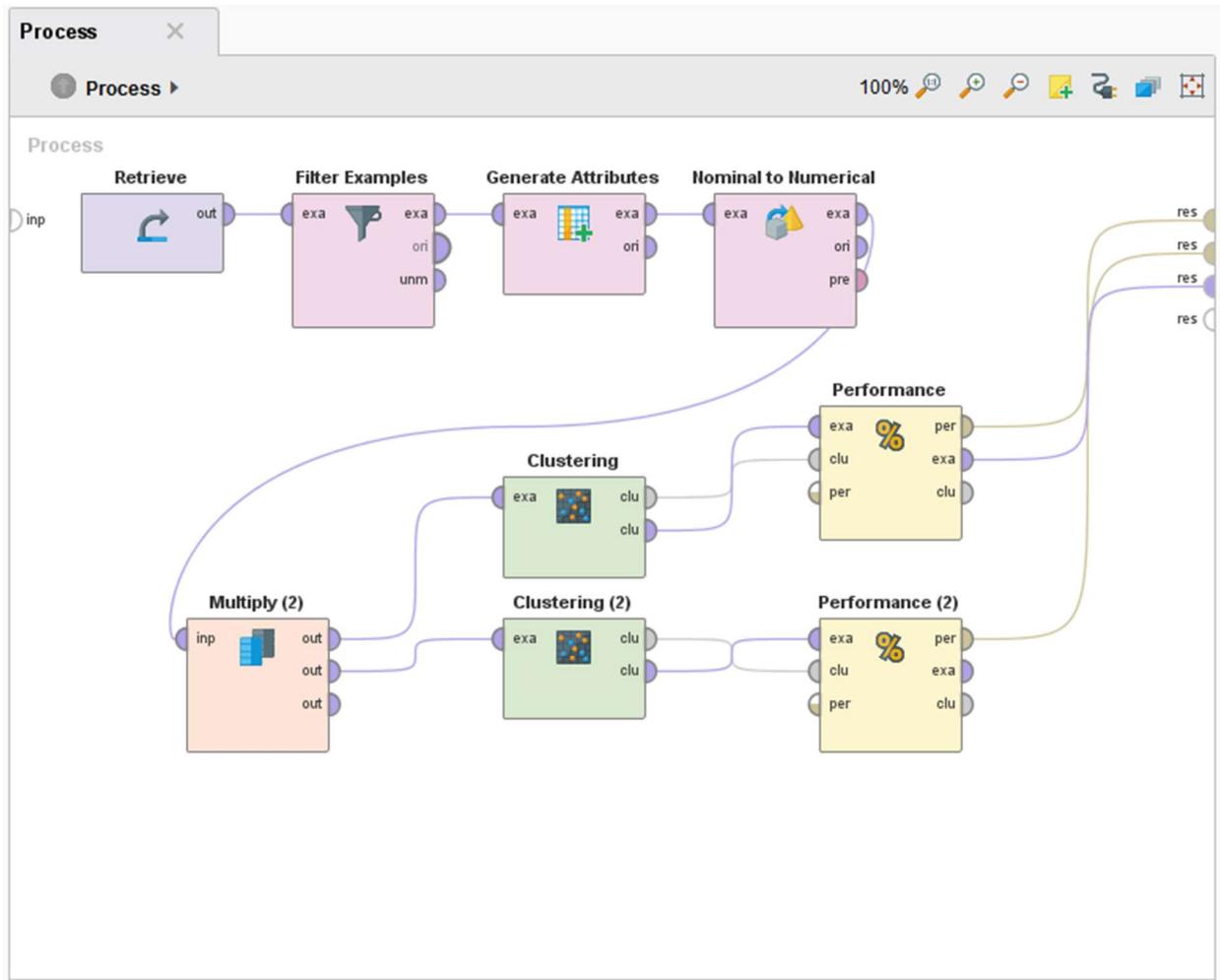
در تصویر بالا مشاهده می کنید که شاخص توده بدنی با فشار خون سیستولی و دیاستولی دارای همبستگی حدود ۰،۲ هستند و این همبستگی نسبتاً خوبی است که با افزایش یا کاهش شاخص توده بدنی فشار خون سیستولی و دیاستولی نیز افزایش یا کاهش پیدا می کند. خصیصه سن با فشار خون سیستولی و دیاستولی نیز دارای همبستگی نسبتاً خوبی هستند و با افزایش سن فشار خون سیستولی و دیاستولی نیز افزایش پیدا می کنند. در مورد هر یک از خصیصه ها تحلیل مشابهی با استفاده از این ماتریس می توان انجام داد.

در تصویر زیر همبستگی دو به دو بین خصیصه‌ها نمایش داده شده است.

First Attribute	Second Attribute	Correlation
BMI_2006	BMI_2011	0.820
BMI_2006	Systol 2006	0.255
BMI_2006	diastol2006(cm)	0.240
BMI_2006	Age 2006	0.089
BMI_2011	Systol 2006	0.208
BMI_2011	diastol2006(cm)	0.205
BMI_2011	Age 2006	0.011
Systol 2006	diastol2006(cm)	0.650
Systol 2006	Age 2006	0.201
diastol2006(cm)	Age 2006	0.105

خوشه بندی

در این روش رکوردها بر اساس شباهتی که به یکدیگر دارند و عدم شباهتشان با رکوردهای دیگر در خوشه‌های متعددی قرار می‌گیرند در این تکنیک مراحلی تحت عنوان آموزشی و آزمایشی وجود نداشته و در پایان یک مدلی ساخته می‌شود که عملاً همان تعیین خوشه‌ها بوده و به همراه کارایی آن به عنوان خروجی ارائه می‌شود در این مرحله ما از الگوریتم‌های K-medoid و K-means استفاده کرده‌ایم و نتایج حاصل از این الگوریتم‌ها را ارزیابی کرده و الگوریتم مناسب‌تر انتخاب می‌شود.



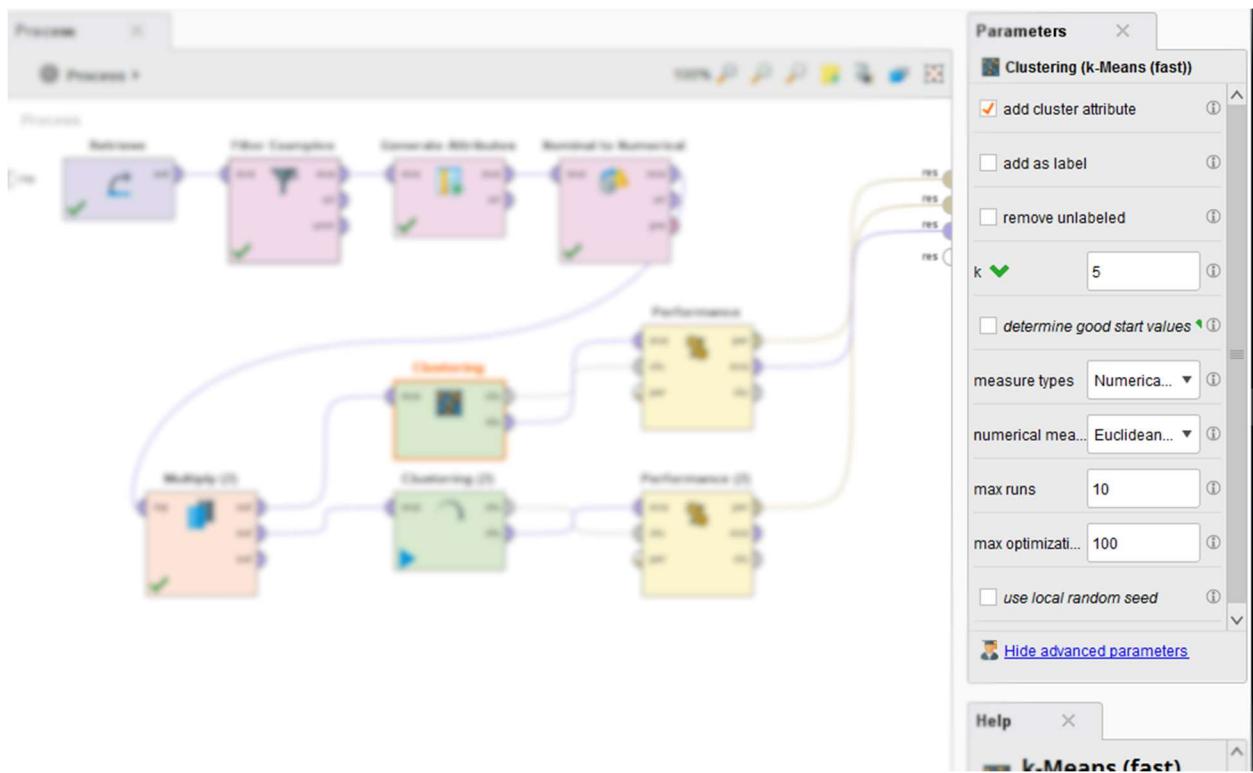
عملگرهای ۱، ۲ و ۳ را در قسمت قبل توضیح دادیم و برای اماده سازی داده‌ها از آن استفاده می‌شود.

عملگر Nominal to Numerical: برای تبدیل داده‌های اسمی به داده‌های عددی استفاده می‌شود و قبل از خوشبندی K-medoids، K-means حتماً باید این عملگر وجود داشته باشد.

عملگر Multiply: این عملگر هیچ گونه تغییری بر روی داده‌ها اعمال نمی‌کند و برای موقعی مناسب است که مایلید خروجی چند پردازش را بر روی داده‌های یکسانی مشاهده کنید.

خروجی حاصل از عملگر Multiply به عملگرهای الگوریتم‌های خوشبندی اعمال می‌شود.

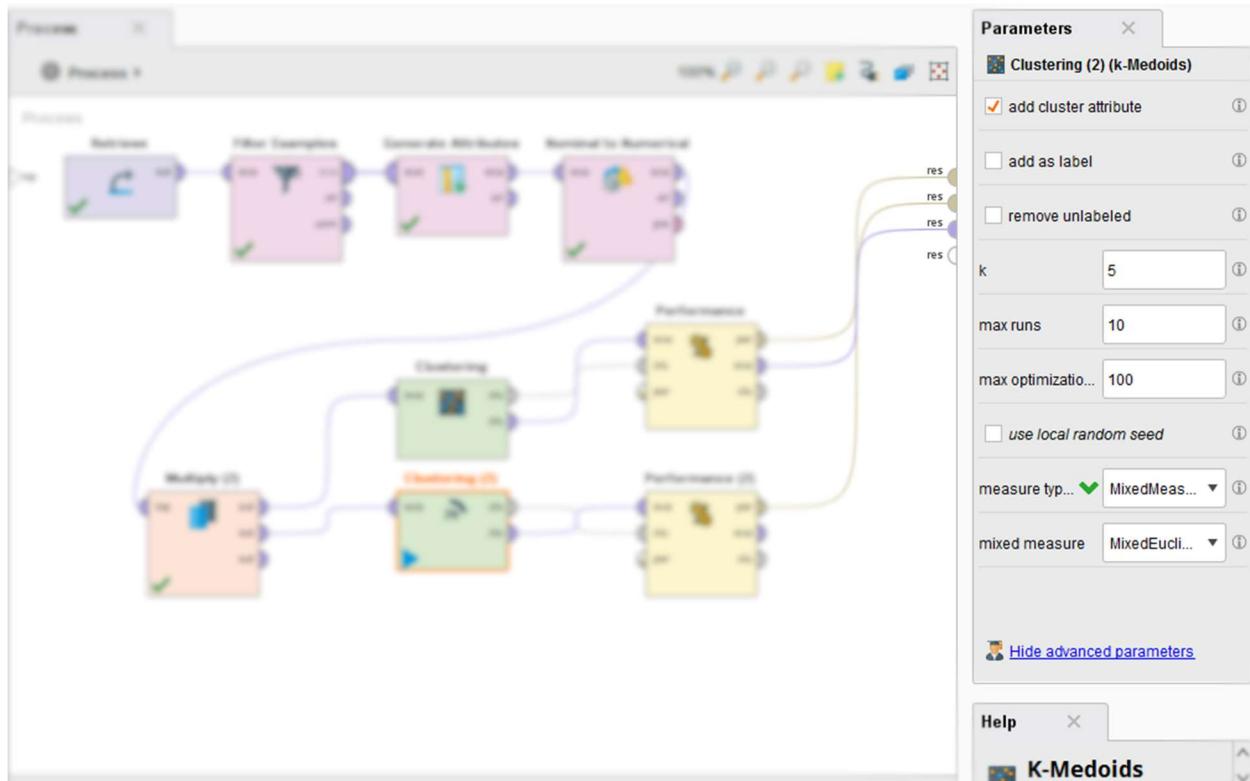
عملگر Clustering 1: در این عملگر از الگوریتم K-means استفاده کردیم. مهم‌ترین پارامتر این الگوریتم مقدار K است. با کمک این پارامتر تعداد خوشبندی‌های خروجی را می‌توان مشخص کرد. مقدار K را برابر با ۵ قراردادیم.



همانطور که در تصویر زیر مشخص است نمونه‌های ۲، ۱۰، ۱۴، ۱۵ و ... در خوشه ۱ قرار گرفته اند و نمونه‌های دیگر در خوشه‌های دیگر این موضوع را از روی ستون Cluster میتوان تشخیص داد.

ExampleSet (3127 examples, 2 special attributes, 22 regular attributes)												Filter (3,127 / 3,127 examples): all
Row No.	id	cluster	Smoke2006 = ...	Smoke2006 = ...	Drink2006 = ...	Drink2006 = ...	Hypertensio...	Hypertensio...	Residence2...	Residence2...		
1	1	cluster_4	1	0	1	0	1	0	1	0		
2	2	cluster_1	1	0	1	0	1	0	1	0		
3	3	cluster_0	1	0	1	0	1	0	1	0		
4	4	cluster_0	1	0	1	0	1	0	1	0		
5	5	cluster_0	1	0	1	0	1	0	1	0		
6	6	cluster_2	1	0	1	0	1	0	1	0		
7	7	cluster_2	1	0	1	0	1	0	1	0		
8	8	cluster_4	1	0	1	0	1	0	1	0		
9	9	cluster_4	1	0	1	0	1	0	1	0		
10	10	cluster_1	1	0	1	0	1	0	1	0		
11	11	cluster_4	1	0	1	0	1	0	1	0		
12	12	cluster_2	1	0	1	0	1	0	1	0		
13	13	cluster_2	1	0	1	0	1	0	1	0		
14	14	cluster_1	1	0	1	0	1	0	1	0		
15	15	cluster_1	1	0	1	0	1	0	1	0		
16	16	cluster_2	1	0	1	0	1	0	1	0		
17	17	cluster_4	1	0	1	0	1	0	1	0		
< 100												>

عملگر 2 Clustering 2 : در این عملگر از الگوریتم K-medoid استفاده کردیم. با کمک پارامتر K تعداد خوشه‌های خروجی را می‌توان مشخص کرد. مقدار K را برابر با ۵ قراردادیم.



عملگر 3 Performance : هر سه عملگر Performance بکار رفته در این پردازش الگوریتم‌های خوشه‌بندی توسط معیارهایی ارزیابی می‌کند.

پس از اجرای پردازش بالا خروجی‌های زیر بدست می‌آید

نتایج حاصل از Performance اول:

The screenshot shows the KNIME interface with the following tabs in the top bar: Result History, ExampleSet (Multiply (2)), PerformanceVector (Performance (2)), and PerformanceVector (Performance). The main panel displays the 'PerformanceVector' node configuration. The 'Performance' tab is selected, showing the following output:

```
PerformanceVector:  
Avg. within centroid distance: 18.185  
Avg. within centroid distance_cluster_0: 18.086  
Avg. within centroid distance_cluster_1: 21.020  
Avg. within centroid distance_cluster_2: 16.670  
Avg. within centroid distance_cluster_3: 19.278  
Avg. within centroid distance_cluster_4: 16.980  
Davies Bouldin: 0.074
```

نتایج حاصل از Performance دوم

The screenshot shows the KNIME interface with the following tabs in the top bar: Result History, ExampleSet (Multiply (2)), PerformanceVector (Performance (2)), and PerformanceVector (Performance). The main panel displays the 'PerformanceVector' node configuration. The 'Performance' tab is selected, showing the following output:

```
PerformanceVector:  
Avg. within centroid distance: 31.307  
Avg. within centroid distance_cluster_0: 32.326  
Avg. within centroid distance_cluster_1: 32.515  
Avg. within centroid distance_cluster_2: 34.882  
Avg. within centroid distance_cluster_3: 25.037  
Avg. within centroid distance_cluster_4: 27.278  
Davies Bouldin: 0.094
```

نتیجه گیری

مقایسه سه الگوریتم خوشه بندی بکار رفته

	K-means ✓	K-medoid
Avg. within centroid distance	18.185	31.307
Avg. within centroid distance_cluster_0	18.086	32.326
Avg. within centroid distance_cluster_1	21.020	32.515
Avg. within centroid distance_cluster_2	16.670	34.882
Avg. within centroid distance_cluster_3	19.278	25.037
Avg. within centroid distance_cluster_4	16.980	27.278
Davies Bouldin	0.074	0.094

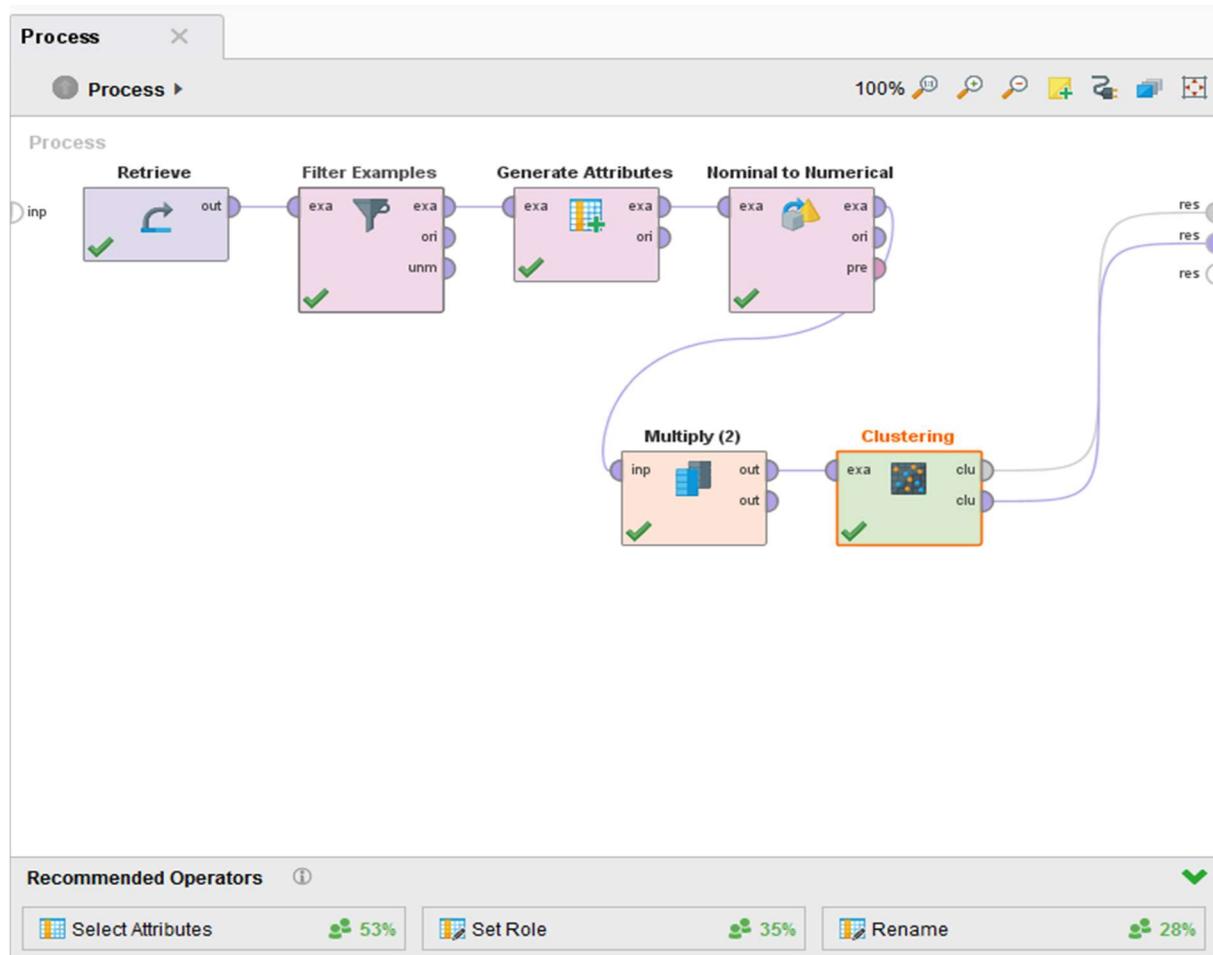
معیارهای ارزیابی الگوریتم‌های خوشه بندی:

- **شباخت درون خوشه‌ای(Avg. within centroid distance):** در این معیار فاصله هر نمونه تا عنصر مرکزی یا همون centroid مشخص شده است. میانگین این مقادیر به عنوان شباخت درون خوشه‌ای می‌باشد که هر چه این مقدار کمتر باشد شباخت درون خوشه‌ای بیشتر و مناسب‌تر است.
- **شباخت بین خوشه‌ای:** در این معیار فاصله مراکز خوشه‌ها از هم بدست می‌آید و هرچقدر این مقادیر بیشتر باشد مناسب‌تر است.
- **دیویس بولدین(Davies Bouldin):** با استفاده از ترکیب دو معیار قبل بر اساس یک فرمول خاص معیاری بدست می‌آید به نام دیویس بولدین که هر چقدر این معیار کمتر باشد مناسب‌تر است.

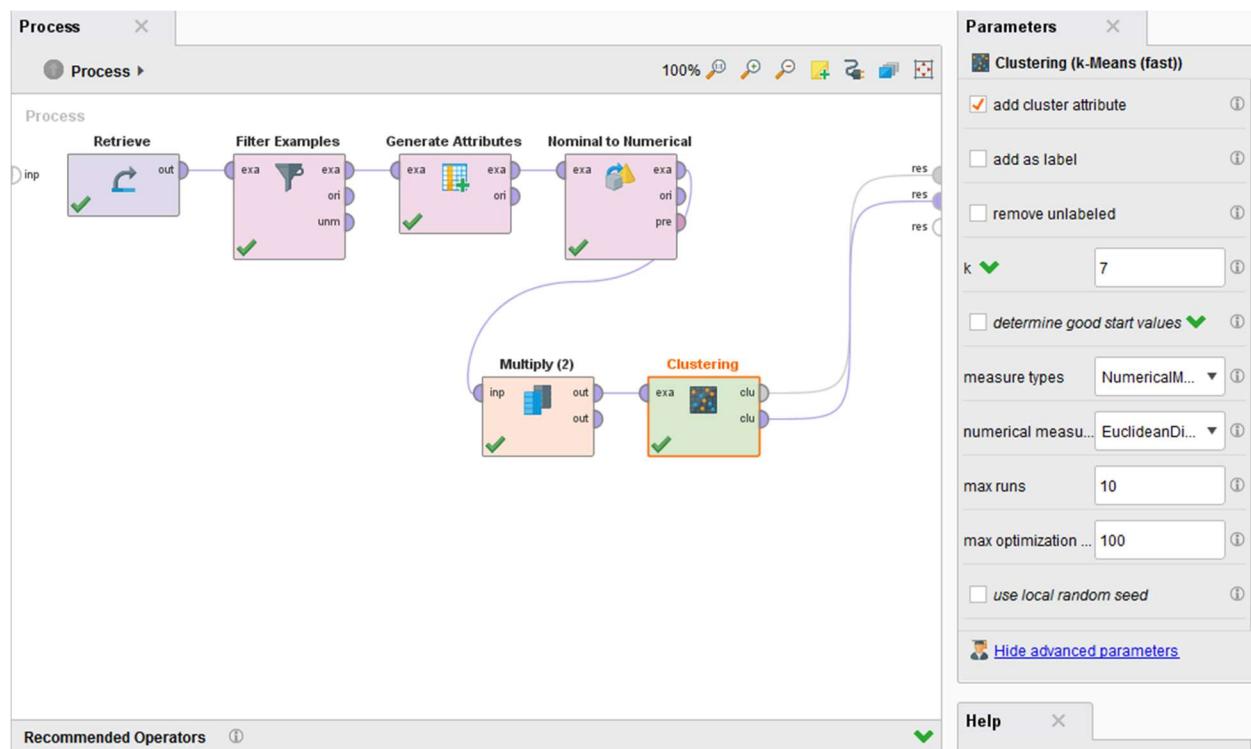
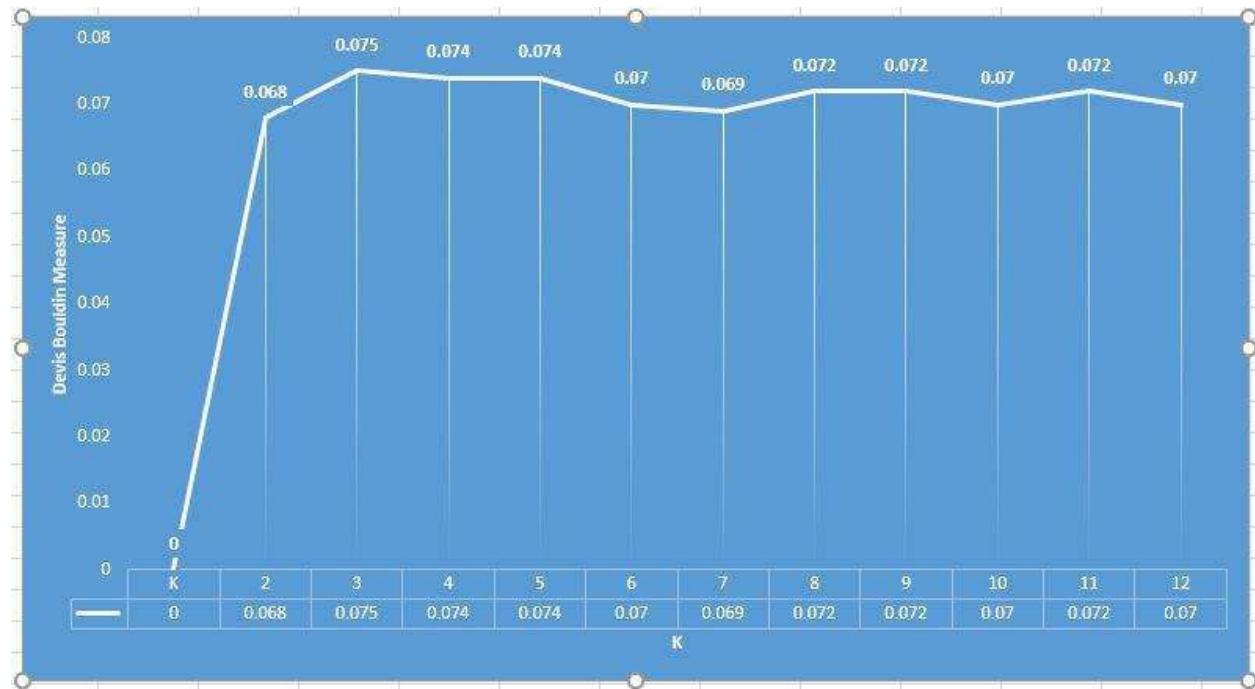
الگوریتم‌های خوش‌بندی بکار رفته در این پردازش را بر اساس معیارهای بالا در جدول مقایسه کردیم که به این نتیجه رسیدیم : در الگوریتم K-medoids معیار شباهت درون خوش‌بندی نسبت به الگوریتم K-means دارای مقادیر بیشتری است پس شباهت درون خوش‌بندی در الگوریتم K-means نسبت به الگوریتم K-medoids بهتر است. بر اساس معیار دیویس بولدین نیز الگوریتم K-means دارای عملکرد بهتری نسبت به الگوریتم دیگر است و همچنین الگوریتم K-means سرعت اجرای بالاتری دارد بنابر این عملکرد الگوریتم خوش‌بندی K-means نسبت به دو الگوریتم دیگر بر روی این مجموعه از داده‌ها مناسب‌تر است.

خوش‌بندی داده‌ها با استفاده از الگوریتم برگزیده مرحله قبل الگوریتم K-means

بر اساس عملگرهای مشخص شده در تصویر مراحل پیش پردازش داده‌ها را انجام می‌دهیم تا به عملگر خوش‌بندی K-means برسیم.



در عملگر Clustering باید اندازه k که همان تعداد خوشه‌ها هست مشخص شود. در نمودار زیر تغییرات معیار دیویس بولدین (DB) را با افزایش مقدار K مشاهده می‌کنیم. پس به این نتیجه می‌رسیم که مقدار ۶ یا ۷ برای تعداد خوشه‌ها مقدار مناسبی است در اینجا ما مقدار ۷ را برای K در نظر می‌گیریم.



پس از اجرای پردازش خروجی‌های زیر نمایش داده می‌شود.

The screenshot shows the KNIME interface with three tabs at the top: "Result History", "ExampleSet (Multiply (2))", and "Cluster Model (Clustering)". The "Cluster Model (Clustering)" tab is active. On the left, there is a sidebar with icons for "Description", "Folder View", "Graph", "Centroid Table", and "Plot". The main content area is titled "Cluster Model" and displays the following text:

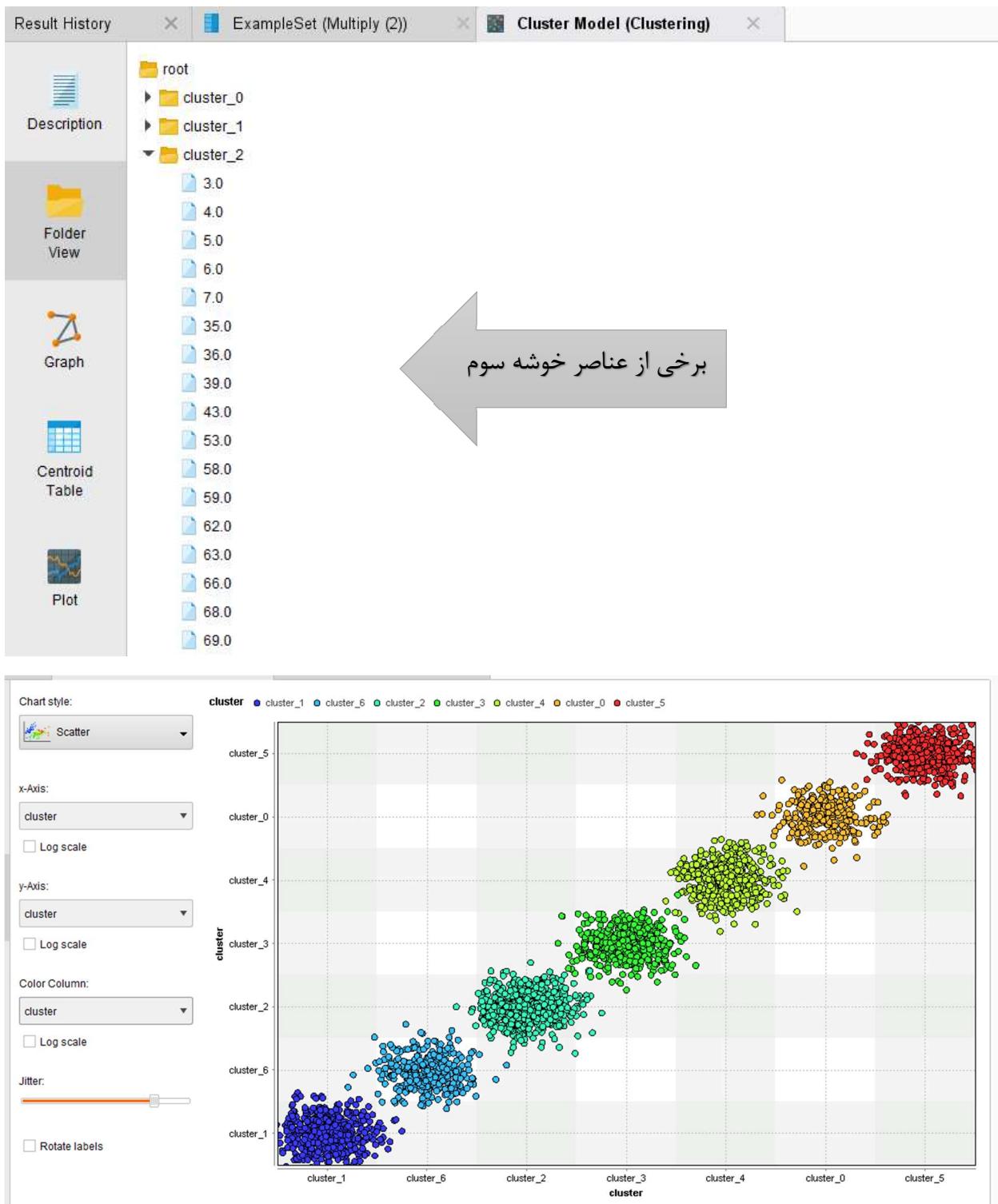
```
Cluster 0: 285 items
Cluster 1: 507 items
Cluster 2: 589 items
Cluster 3: 502 items
Cluster 4: 468 items
Cluster 5: 438 items
Cluster 6: 338 items
Total number of items: 3127
```

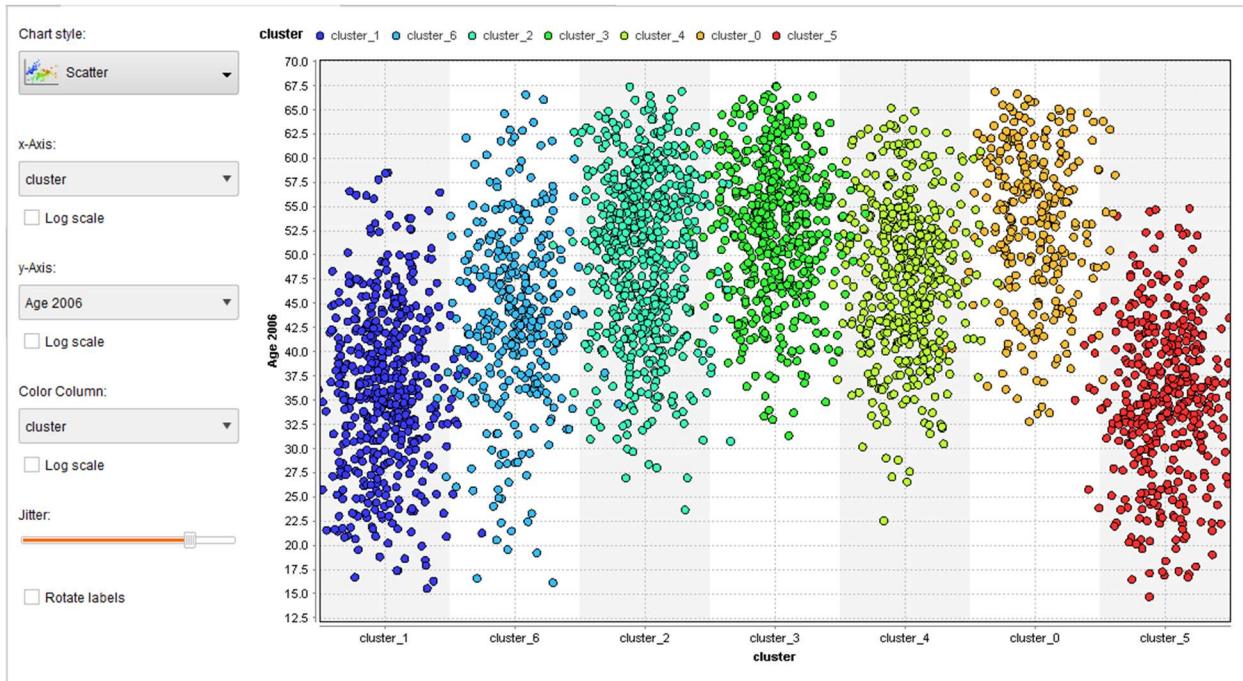
Result History ExampleSet (Multiply (2)) Cluster Model (Clustering)

The screenshot shows the KNIME interface with three tabs at the top: 'Result History', 'ExampleSet (Multiply (2))', and 'Cluster Model (Clustering)'. The 'Cluster Model (Clustering)' tab is active. On the left, there is a sidebar with icons for 'Description', 'Folder View' (selected), 'Graph', 'Centroid Table', and 'Plot'. The main area displays a tree view under 'root': 'cluster_0' is expanded, showing its contents. A large gray arrow points from the Persian text 'برخی از عناصر خوشة اول' (Some good elements of the first cluster) to the 'cluster_0' node. The data listed under 'cluster_0' are: 34.0, 37.0, 220.0, 236.0, 244.0, 245.0, 249.0, 250.0, 251.0, 272.0, 275.0, 277.0, 284.0, 292.0, 300.0, 321.0.

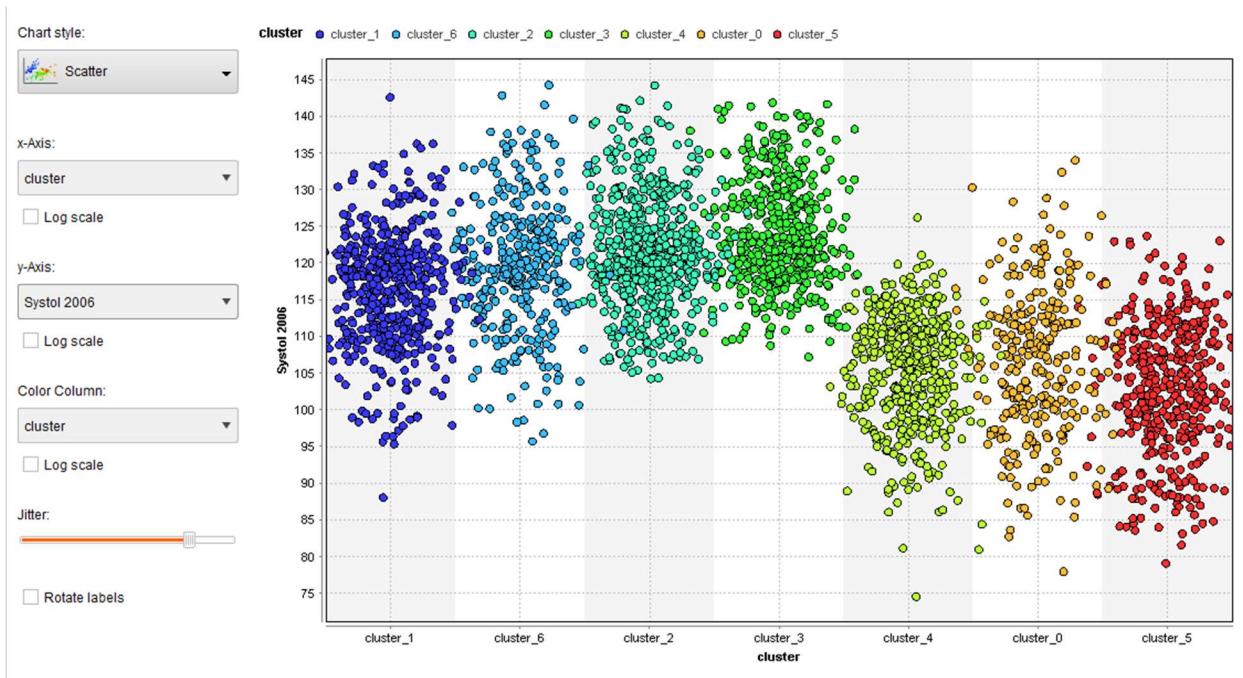
Result History ExampleSet (Multiply (2)) Cluster Model (Clustering)

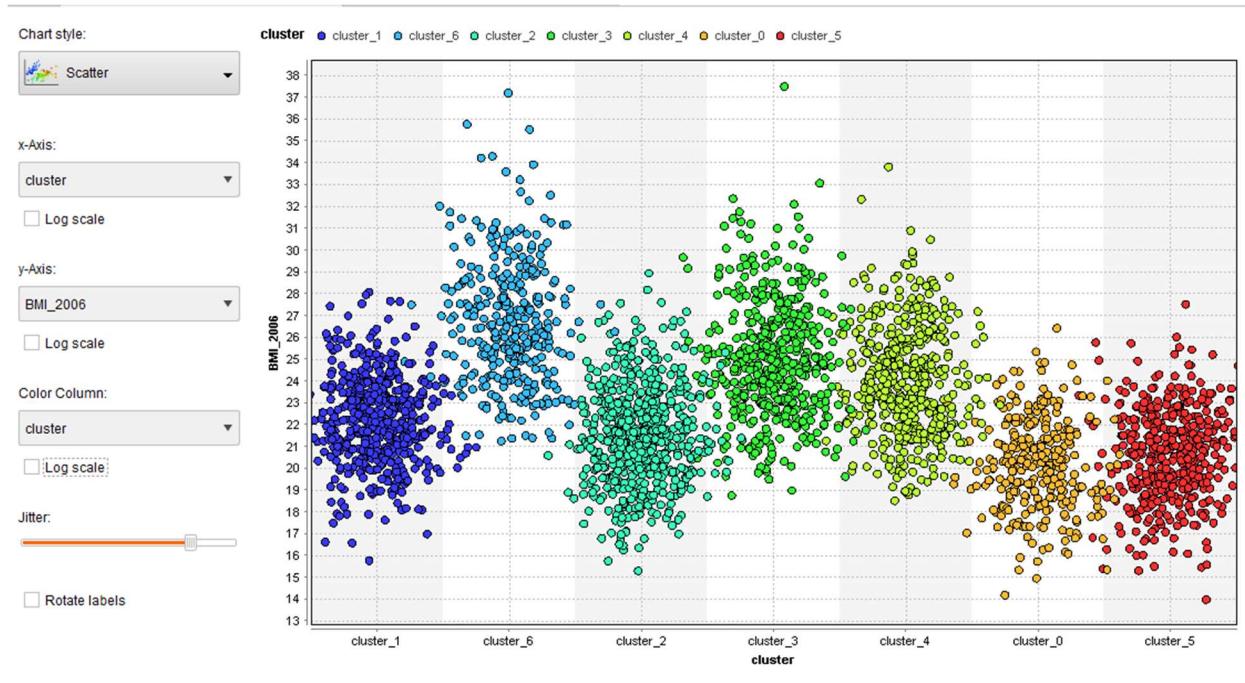
The screenshot shows the KNIME interface with the same three tabs at the top. The 'Cluster Model (Clustering)' tab is active. The sidebar on the left is identical to the previous screenshot. A large gray arrow points from the Persian text 'برخی از عناصر خوشة دوم' (Some good elements of the second cluster) to the 'cluster_1' node. The data listed under 'cluster_1' are: 1.0, 9.0, 11.0, 14.0, 17.0, 45.0, 48.0, 56.0, 64.0, 67.0, 74.0, 79.0, 82.0, 83.0, 84.0, 87.0, 88.0, 92.0.





در تصویر بالا مشاهده می‌کنیم که سن در خوشه سوم از ۳۲ سال به بالا است و در خوشه پنجم سن نمونه‌ها کمتر است حال با توجه به تصاویر زیر می‌بینیم که در خوشه سوم که متوسط سن بالاتر از خوشه پنجم بود دارای فشار خون و شاخص توده بدنی بالاتری است. برای هر کدام از خوشه‌ها با توجه به معیارهای متفاوت تحلیل‌های متفاوتی وجود دارد.





نتایج حاصل از پیاده سازی

طبقه بندی

پس از اجرای الگوریتم های دسته بندی بر روی مجموعه دادهها نتایج در بردارهای ذخیره می شود که این بردار با بردار برچسب داده های آزمایشی هم بعد است و قابل مقایسه هستند.

در تصویر زیر این مقایسه برای مدل درخت تصمیم نشان داده شده است.

accuracy: 80.84% +/- 0.55% (mikro: 80.84%)

	true yes	true no	class precision
pred. yes	18	38	32.14%
pred. no	561	2510	81.73%
class recall	3.11%	98.51%	

همانگونه که مشاهده می شود تعداد ۱۸ نفر دارای فشار خون بودند که درخت تصمیم درست تشخیص داده است. همچنین تعداد ۲۵۱۰ نفر فاقد فشار خون بوده اند که درست تشخیص داده شده است. ولی ۳۸ نفر فاقد فشار خون بودند که به اشتباه دارای فشار خون و تعداد ۵۶۱ نفر دارای فشار خون بودند که به اشتباه فاقد فشار خون تشخیص داده شده است. بنابراین برای درخت تصمیم به دقت ۸۰٪ درصد رسیدیم. توجه داشته باشید با خاطر اعمال

هرس بر روی درخت تصمیم درصدی از Precision کلاس yes کم شده است. که با توجه به کاهش زمان اجرا و حذف شاخه های اضافی درخت در اثر انجام هرس این کاهش قابل چشم پوشی می باشد.

accuracy: 78.41% +/- 1.93% (mikro: 78.41%)

	true yes	true no	class precision
pred. yes	192	288	40.00%
pred. no	387	2260	85.38%
class recall	33.16%	88.70%	

با توجه به تصویر بالا با استفاده از الگوریتم بیز به این نتایج رسیدیم. تعداد ۱۹۲ نفر دارای فشار خون بودند و بدرستی تشخیص داده شده است. ۲۲۵۰ نفر فقد فشار خون بودند که درست تشخیص داده شده است. اما تعداد ۲۸۸ نفر فقد فشار خون بودند که به اشتباه دارای فشار خون تشخیص داده شده و تعداد ۳۸۷ نفر دارای فشار خون بودند که به اشتباه فقد فشار خون تشخیص داده شده است. بنابر این به دقت ۷۸,۴۱ درصد رسیدیم.

accuracy: 73.71% +/- 1.92% (mikro: 73.71%)

	true yes	true no	class precision
pred. yes	169	412	29.09%
pred. no	410	2136	83.90%
class recall	29.19%	83.83%	

در الگوریتم KNN تعداد ۱۶۲ نفر دارای فشار خون بودند و درست تشخیص داده شده است. تعداد ۲۱۳۶ نفر فقد فشار خون بودند که بدرستی تشخیص داده شده است. اما ۴۱۲ نفر فقد فشار خون بودند که به اشتباه دارای فشار خون تشخیص داده شدند و تعداد ۴۱۰ نفر دارای فشار خون بودند که به اشتباه فقد فشار خون تشخیص داده شدند. بنابر این به دقت ۷۳,۷۴ درصد برای این الگوریتم رسیدیم.

خوشه بندی

با توجه به الگوریتم های خوشه بندی و معیارهای ارزیابی الگوریتم خوشه بندی K-means با مقدار $K=7$ به عنوان الگوریتم مناسبتر برای این مجموعه از داده ها انتخاب شد.

بخش هفتم: نتیجه گیری نهایی

تشخیص بیماری فشار خون یک اقدام مهم در حیطه پزشکی به حساب می آید. ما در این گزارش با در دست داشتن پایگاه داده‌ای از میزان شاخص توده بدنی افراد و سابقه فشار خون، در ابتدا به توضیح این اطلاعات پرداختیم و سپس با بکارگیری تکنیک‌های دسته بندی و خوش بندی به نتایجی رسیدیم. با کمک درخت تصمیم به یک سیستم پشتیبان تصمیم گیری دست یافتیم. این سیستم به پزشکان کمک می کند که علاوه بر دانش و تخصص خود از تشخیص این سیستم نیز استفاده کنند و در نهایت تشخیص دقیقتری در مورد بیماری بدهنند. با استفاده از الگوریتم‌های خوبه بندی نمونه‌هایی که دارای شباهت بیشتری نسبت بهم بودن در چندین خوش جمع آوری شدند. با استفاده از شباهت‌های نمونه‌های هر خوش بهم میتوان اقدامات درمانی جداگانه انجام داد.