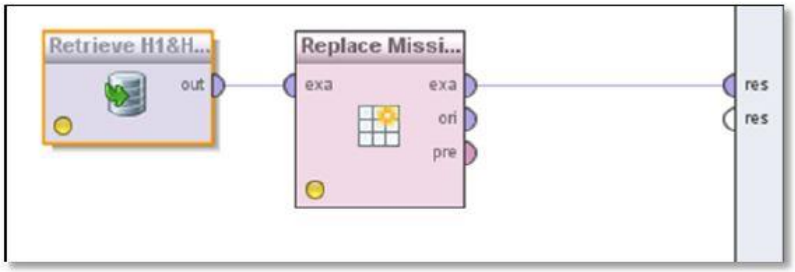
۴-۳-۲- مقادیر مفقود شده:

تقریبا در تمام برنامه‌های کاربردی واقعی، می‌توان نمونه‌هایی را یافت که مقداری برای صفات خاصه آنها وجود ندارد. اجرای الگوریتم‌های داده‌کاوی بر روی داده‌های با مقادیر مفقوده زیاد، نتایج الگوریتم‌ها را می‌تواند نادرست کند. ساده‌ترین راه برای حل این مشکل صرف‌نظر از نمونه‌هایی است که برخی از صفات خاصه آنها دارای مقدار نیست. این درصورتی امکان‌پذیر است که تعداد رکوردهای داده‌ها بسیار زیاد باشد.

در مواردی که تعداد ستون‌های یک متغیر کم باشند می‌توان به صورت دستی نمونه‌های ناقص را کامل کرد. که این راهکار برای پایگاه داده‌های حجیم امکان‌پذیر نیست.

در نرم افزار رپیدماینر ما می‌توانیم به صورت سیستمی و با چندین روش با داده‌های مفقوده رفتار کنیم. برای مثال می‌توان ابتدا از داده‌های ناقص صرف نظر نموده و نتیجه داده‌کاوی را مشاهده و پس از آن با جایگزین نمودن یک مقدار دوباره نتیجه حاصله را مشاهده و تحلیل کنیم. یکی از رایجترین عملگر۱۱هایی که برای داده‌های مفقود شده می‌توان استفاده کرد Replace Missing Values می‌باشد. که ما می‌توانیم مقادیر مفقود شده را با مقادیر جدید به روش‌های مختلفی از جمله میانگین‌گیری، گذاشتن یک مقدار خاصی به جای مقادیر مفقود شده و ... جایگزین کنیم. ما در این تحقیق از این روش استفاده نموده ایم.



۴-۳-۳- تغییر نوع داده:

شکل مناسب داده‌ها به عنوان ورودی الگوریتم‌های داده‌کاوی نقش به سزایی در این فرایند بازی می‌کنند و در مرحله آماده‌سازی داده‌ها این نقش پررنگ است. بسیاری از الگوریتم‌های داده‌کاوی برای نوع خاصی از صفات خاصه مناسب هستند و حتی برخی از آنها تنها بر روی نوع‌های مشخصی از صفات خاصه اجرا می شوند. برای مثال اکثر روش‌های رگرسیون بر روی داده‌های عددی اجرا می‌شوند. حال اگر در میان داده‌های ما داده‌ای از نوع اسمی باشد چگونه می‌توانیم از این روش‌ها برای تحلیل استفاده کنیم؟ عملگرهای متعددی در رپید ماینر وجود دارند که به ما کمک خواهد کرد تا این تبدیلات را انجام دهید(شکل ۴-۲).

عملکرد بعضی از این عملگرها به شرح زیر می باشد.

صفات خاصه اسمی چند مقداره را به نوع اسمی دو مقداره۱۲

صفات خاصه اسمی (چند مقداره) را به نوع عددی[۱۳]

صفات خاصه عددی را به اسمی (دومقداره)[۱۴]

صفات خاصه عددی را به اسمی (چند مقداره)[۱۵]

۴-۳-۴- نرمال سازی داده ها:

یکی دیگر از گونه‌های تغییر شکل داده‌ها نرمال‌سازی است. روش‌های نرمال‌سازی معمولا برای مواردی که محاسبه فواصل بین نمونه‌ها در داده کاوی مطرح است مفید می‌باشد. در این روش مقادیر اندازه گیری شده به محدوده جدیدی نگاشت می‌شوند. این روش برای داده‌هایی که صفات خاصه آن واحد اندازه گیری متفاوتی دارندبسیار مناسب است.

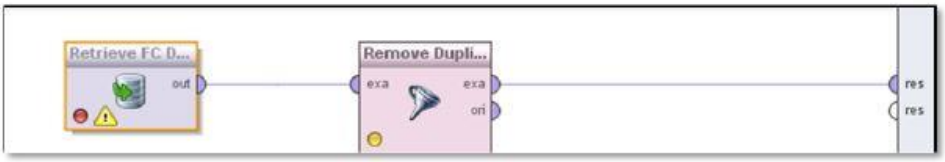
۴-۳-۵- گسسته سازی داده ها:

در واقع هدف گسسته‌سازی تبدیل داده‌های عددی به نوع اسمی است. با این عمل مقادیر یک صفت خاصه به چندین بازه تقسیم می‌شود. برای مثال فرض کنید ستونی از داده‌های شما نشان دهنده سن افراد است و نوع این صفت خاصه نیز عددی تعریف شده است. با عمل گسسته‌سازی شما می‌توانید این اعداد را به چندین بازه تقسیم کنید و برای هر بازه نامی را در نظر بگیرید برای مثال کلیه افراد زیر ۲۰ سال را در یک بازه و افراد بزرگتراز ۲۰ سال را در بازه دیگری قرار دهید بدین طریق شما محدوده‌ای از اعداد را در ۲ بازه قرار داده‌اید که می‌توانید با نام‌های جوان و میانسال آنرا جایگزین کنید. با این کار داده‌های عددی را به نوع اسمی تبدیل کرده‌اید. برخی از الگوریتم‌های داده کاوی با گسسته‌سازی کاراتر عمل می‌کنند. در ضمن این عمل برای نمایش قابل فهم داده‌ها نیز مفید است.

در نرم افزار رپیدماینر چندین روش برای انجام عمل گسسته‌سازی وجود دارد. این عملگرها در شکل زیر مشخص شده اند.

شکل(۴-۴) عملگر گسسته سازی داده ها در رپیدماینر

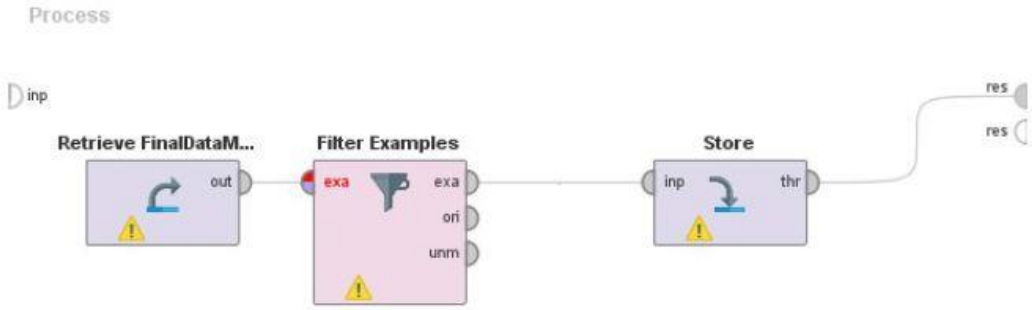
۴-۳-۷- حذف نمونه های تکراری:



شکل(۴-۸) کاهش ابعاد داده ها(حذف سطرهای تکراری)

۴-۳-۸- حذف برخی از نمونه ها[۱۷]:

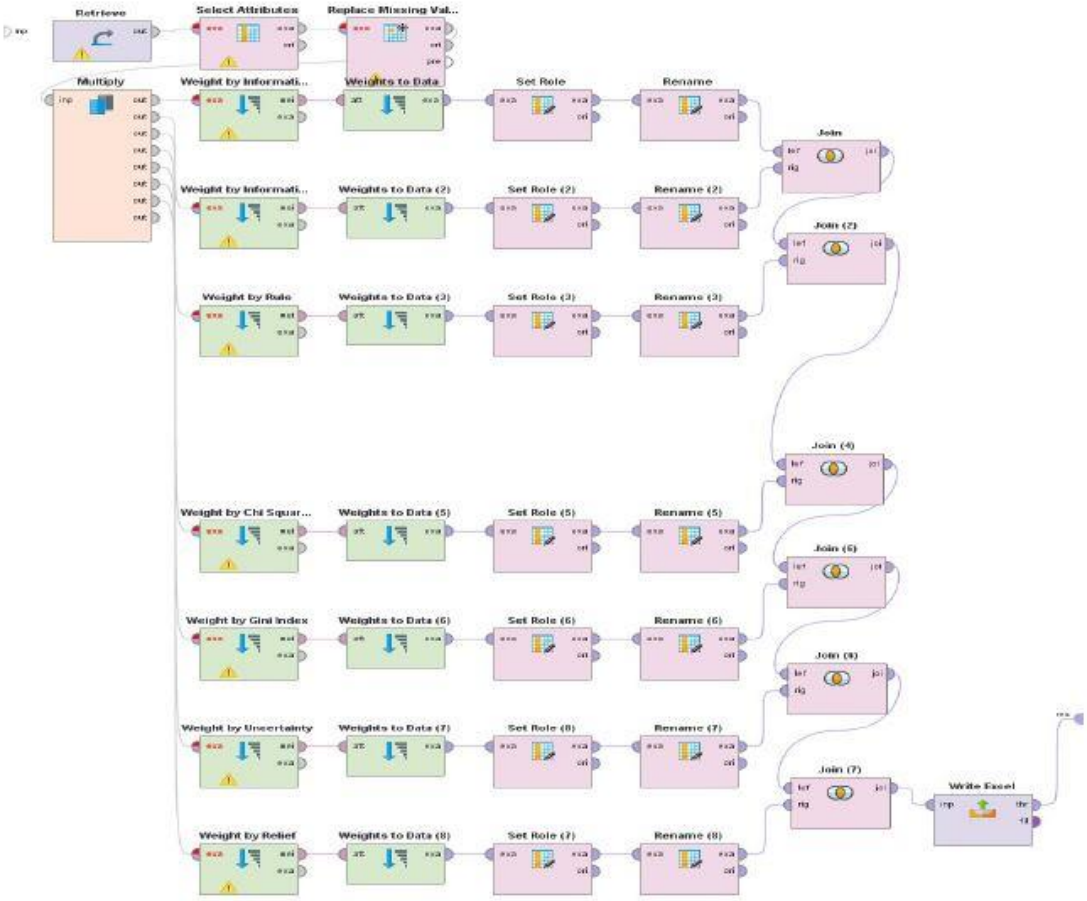
این عملگر در حقیقت وظیفه حذف برخی از نمونه ها را به عهده دارند.



شکل (۴-۹) حذف برخی از نمونه ها

۴-۴- مدل وزن دهی[۱۸]:

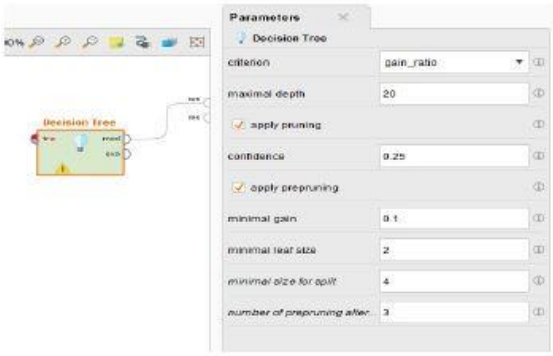
از طریق اعمال روش های وزن دهی جهت شناسایی مهمترین عوامل موثر بر پیشرفت تحصیلی استفاده شده است. در این تحقیق به دلیل ماهیت داده ها از ۷ عملگر وزن دهی در مدل وزن دهی استفاده شده است. این مدل در شکل زیر آمده است، همچنین نتایج اعمال روش های وزن دهی دربخش ۸-۴-۴ جدول ۴-۲ آمده است.

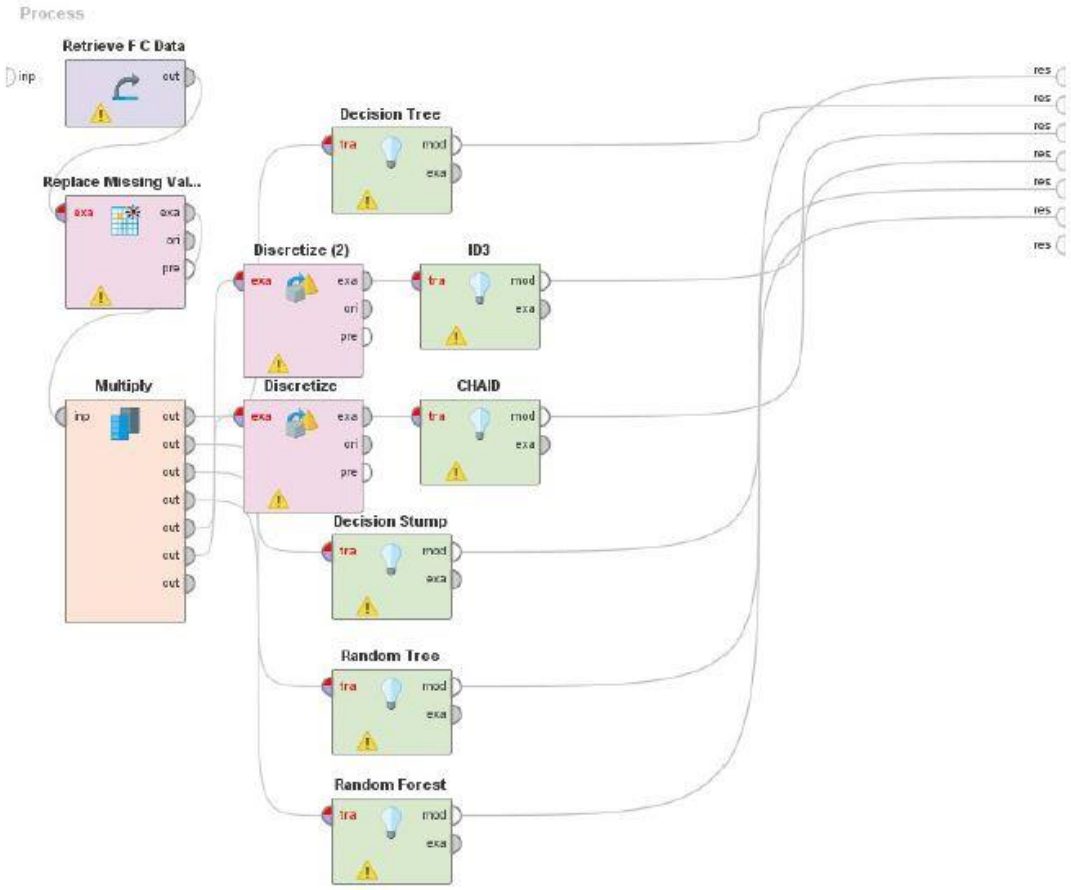


شکل(۴-۱۰) مدل وزن دهی

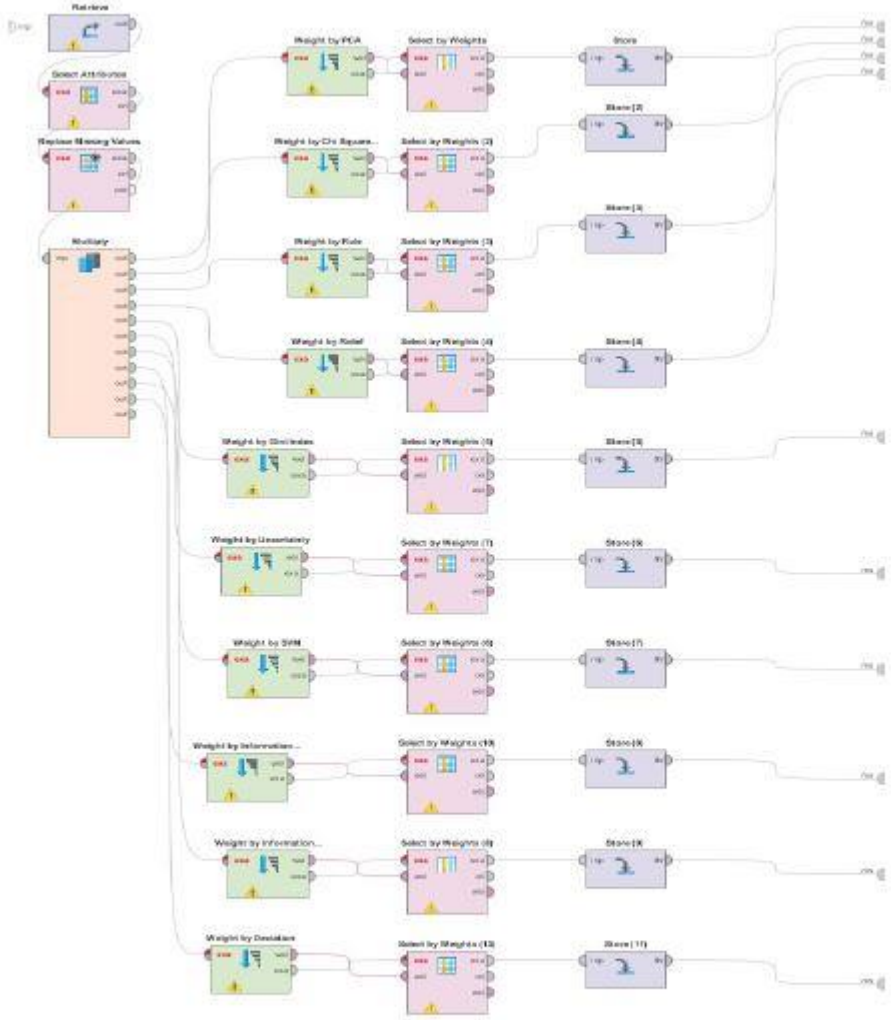
۴-۵- استنتاج های درختی

روش یادگیری دراین بخش روش Decision Tree است. و بقیه روش ها در واقع انشعابی از همین روش و یا انواع مقدماتی تر همین روش با قابلیت دست کاری کمتر هستند.برخی از درخت ها در این روش با استفاده از یک مقدار آستانه اقدام به هرس می کنند تا درخت از میزان خاصی که مورد نظر ما است وسیع تر نشود.این امر باعث می شود تا قدرت تفسیر نیز بالاتر برود.





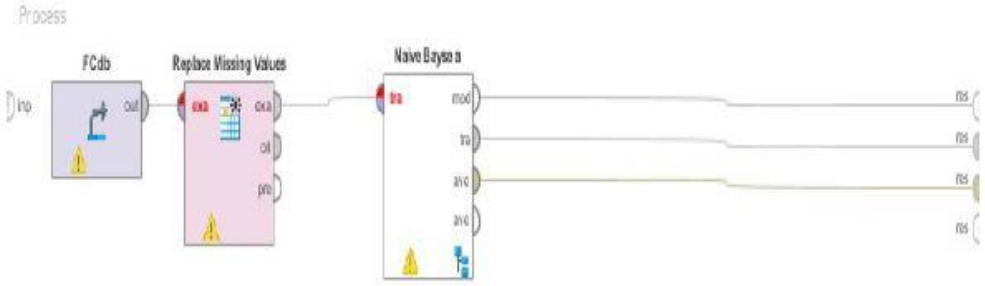
شکل(۴-۱۲) مدل رسم درخت تصمیم برای دسته بندی داده ها



شکل (۴-۱۶) مدل Attribute Selection

۴-۶- مدل پیش بینی وضعیت تحصیلی

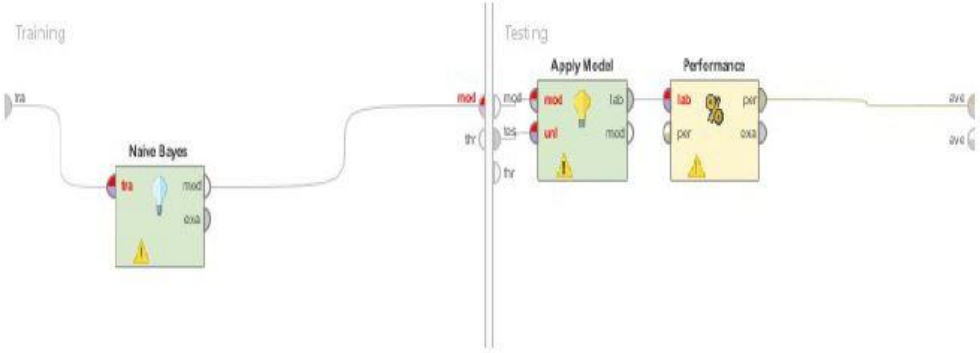
برای پیش بینی وضعیت های مختلف تحصیلی مدل زیر طراحی شد. این مدل که می توان انـواع الگـوریتم های مختلف برای پیش بینی وضعیت تحصیلی از آن استفاده کرد به راحتی وضعیت تحصیلی دانشجویان را پیشبینی می کند. در ادامه نتایج هریک از الگوریتم های استفاده شده در این مدل عبارتند از الگـوریتم هـای Decision Stump، Decision Tree، Naïve Bayes (Kernel) و یادگیری عمیق آمده است.



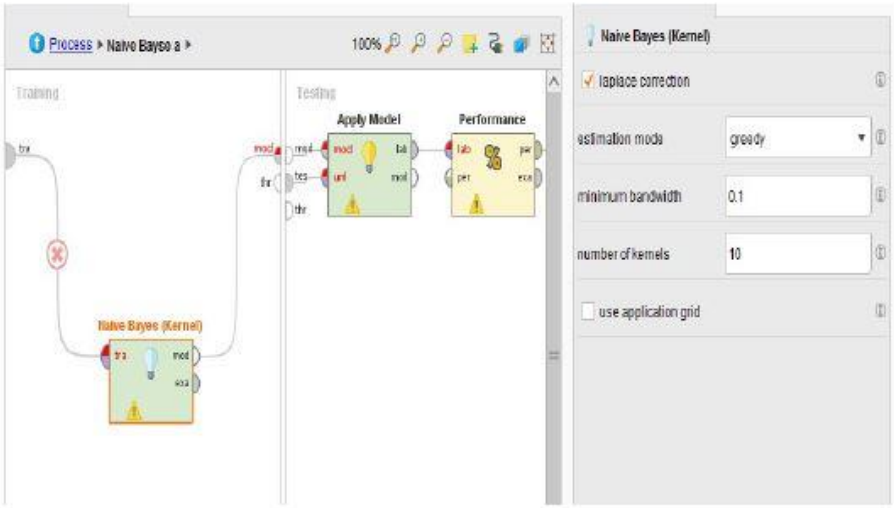
شکل(۴-۱۸) مدل پیش بینی نتایج

نتایج حاصل از اجرای الگوریتم Naïve Bayes:

این عملگر یک مدل طبقه‌بندی بیز ساده ایجاد می‌کند. بیز ساده براساس قـانون احتمـال شـرطی بیـز، بـرای طبقه‌بندی داده‌ها به کار برده می‌شود. از مزایای بیز ساده اجـرای راحـت و نتـایج خـوب بـرای بسـیاری از کاربردها می باشد.این عملگر با استفاده از روش توزیع نرمال تخمینی و بر اساس نظریـه Bayes ، مـدلی را برای دسته بندی رکوردهای آتی، در اختیار ما قرار می دهد.پارامتر Laplace تصـحیح هـای لاپـلاس را بـه منظور جلوگیری از تاثیر زیاد احتمالات صفر انجام می دهد.



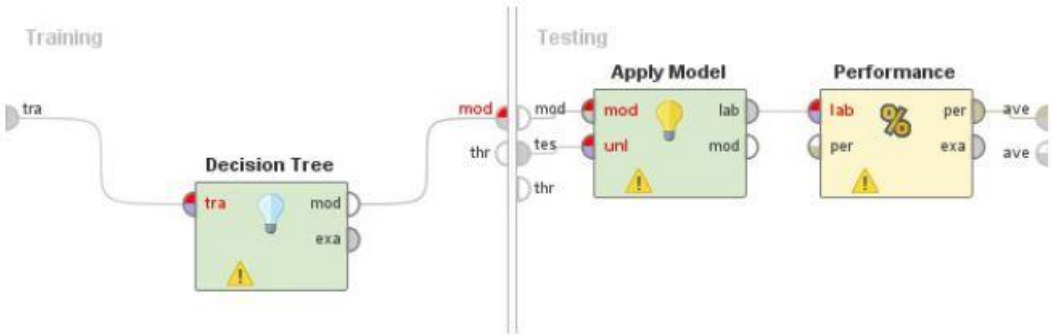
شکل(۴-۱۹) مدل پیش بینی نتایج توسط الگوریتم Naïve Bayes



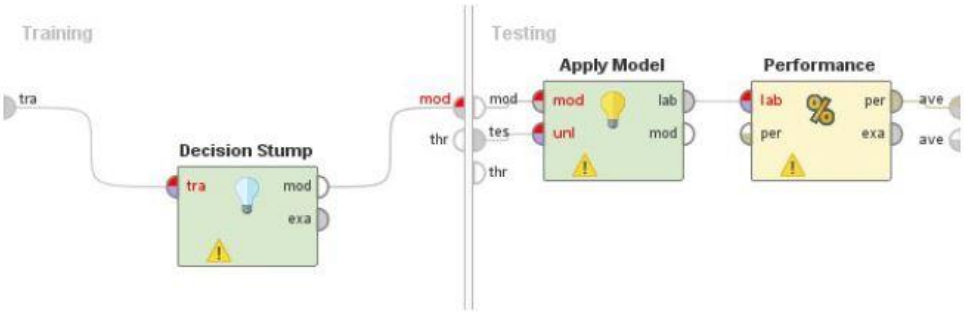
شکل(۴-۲۱) مدل پیش بینی نتایج توسط الگوریتمNaïve Bayes (kernel)

۴-۶-۳- نتایج حاصل از اجرای الگوریتمDecision Tree:

درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته بندی وپیش بینی است. در ساختار درخت تصمیم، پیش‌بینی به دست آمده از درخت، در قالب یکسری قواعد توضیح داده می‌شود. در درخت تصمیم ضرورتی ندارد داده ها لزوما از نوع عددی باشند داده‌ها می‌توانند متنی و اسمی هم باشند. در درخت تصمیم بالاترین گره در درخت گره ریشه است وگره های برگ، دسته‌ها یا توزیع دسته‌ها رانشان می‌دهند. ویژگی-های مهم درخت تصمیم این است که استفاده از آن آسان است، درک مدل ایجاد شده توسط درخت تصمیم آسان است، در تقسیم‌بندی هیچ داده‌ای حذف نمی‌شود.توضیحات در مورد درخت تصمیم در قسمت استنتاج های درختی آمده است.



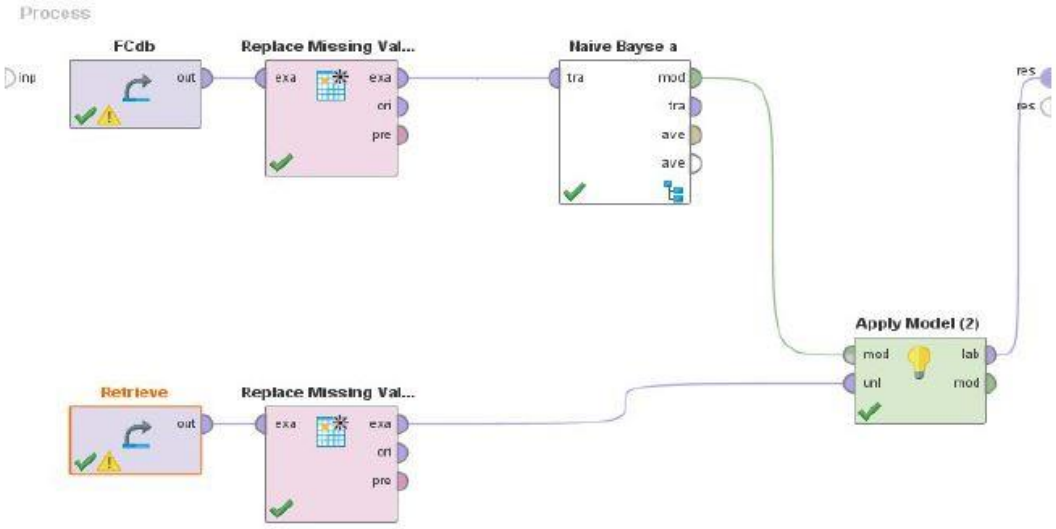
شکل(۴-۲۳) مدل پیش بینی نتایج توسط الگوریتم Decision Tree



شکل(۴-۲۵) مدل پیش بینی نتایج توسط الگوریتم Decision Stump

جدول(۴-۳) مقایسه نتایج مدل های پیش بینی وضعیت تحصیلی

الگوریتم	فارغ التحصیلی		محروم از تحصیل		انتقالی		انصرافی		اخذ کاردانی		دقت کلی
	فراخوانی کلاس	صحت کلاس	فراخوانی کلاس	صحت کلاس	فراخوانی کلاس	صحت کلاس	فراخوانی کلاس	صحت کلاس	فراخوانی کلاس	صحت کلاس	
Naive Bayes	91.45	۹۶.۸۹	51.79	۳۵.۴۶	67.98	۵۳.۵۷	40.48	۵۰.۵۰	20.93	۷.۷۳	84.57
Naive Bayes(Kernel)	85.55	۹۶.۶۶	44.39	۴۱.۲۸	69.01	۵۹.۵۲	40.81	۴۷.۸۳	53.94	۳.۹۱	79.54
Decision Tree	99.16	۹۷.۱۵	45.92	۹۷.۱۶	68.86	۶۴.۳۴	54.03	۴۹.۷۰	8.14	۳۱.۸۲	91.18
Decision Stump	99.6	۸۸.۰۰	57.55	۵۴.۰۰	0	۰	0		0		85.69
Deep Learning	99.34	۹۶.۹۰	55.61	۶۰.۳۰	71.64	۶۵.۵۱	40	۵۵.۶۱	2.23	۱۰۰	91.35



۴-۷- پیش بینی وضعیت تحصیلی دانشجویان جاری بر اساس مدل طراحی شده

accuracy: 87.71% +/- 6.69% (mikro: 87.72%)

	true Fareghottahsil	true AkhzKardani	class precision
pred. Fareghottahsil	76	12	86.36%
pred. AkhzKardani	9	74	89.16%
class recall	89.41%	86.05%	

شکل(۴-۳۵) مدل پیش بینی پیشرفت وضعیت تحصیلی فارغ التحصیلی و اخذ کاردانی

accuracy: 97.02% +/- 1.61% (mikro: 97.02%)

	true Fareghottahsil	true Enseraf	class precision
pred. Fareghottahsil	608	25	96.05%
pred. Enseraf	12	595	98.02%
class recall	98.06%	95.97%	

شکل(۴-۳۶) مدل پیش بینی پیشرفت وضعیت تحصیلی فارغ التحصیلی و انصرافی

accuracy: 96.34% +/- 1.47% (mikro: 96.34%)

	true Fareghottahsil	true Enteghali	class precision
pred. Fareghottahsil	657	24	96.48%
pred. Enteghali	26	860	96.21%
class recall	96.19%	96.49%	

شکل(۴-۳۷) مدل پیش بینی پیشرفت وضعیت تحصیلی فارغ التحصیلی و انتقالی