داده کاوی با نرم افزار WEKA

استاد: دکتر جماعت

تهیه و تنظیم: خانم مهندس بمانی

در طول سال های اخیر با توجه به گسترش فناوری اطلاعات و ارتباطات و با افزایش داده های دیجیتال و رشد قدرت محاسباتی کامپیوترها، نرم افزارهای تجاری و آموزشی فراوانی برای داده کاوی در حوزه های مختلف عرضه شده اند. استفاده از ابزارها و الگوریتم های داده کاوی، می تواند ماهیت پیچیده ی داده ها و روابط نامحسوس میان این داده ها را مدل کند. در واقع ابزار داده کاوی، داده را گرفته و یک تصویر از واقعیت رابه شکل مدل ایجاد می کنند.، این مدل روابط موجود در داده ها را شرح می دهد.

امروزه داده کاوی گسترش زیادی یافته است، به طوری که اکثر نرم افزارهای پایگاه داده ای مانند SQL Server و ORACLE نیز شامل ابزارهای داده کاوی شده اند. از جمله نرم افزارهای متداول در زمینه داده کاوی WEKA، Clementine، Rapid Miner، Orange، Matlab، نرم افزار R و غیره می باشند. جدول (۸-۱) نقش متداول ترین نرم افزارهای داده کاوی را نمایش می دهد.

جدول (۸-۱) – متداول ترین نرم افزارهای داده کاوی

Clementine	قوی ترین و پر کاربردترین ابزار در حوزه ی داده کاوی می باشد. طراحی واسط کاربر این نرم افزار باعث شده است تا کاربران به راحتی با نرم افزار ارتباط برقرار کنند. مدل سازی در این نرم افزار نیازی به برنامه نویسی ندارد، از این رو برای افرادی که برنامه نویسی نمی دانند نیز قابل استفاده است.
Matlab	ابزاری بسیار قدرتمند برای انجام پردازش های حجیم و سنگین مهندسی است؛ به طوری که رسم برخی نمودارها یا انجام بعضی محاسبات را جز به کمک Matlab نمی توان تصور کرد. این نرم افزار برای انجام محاسبات مختلفی از جمله پردازش سیگنال، پردازش های آماری، شبکه های عصبی، دریافت تصویر، پردازش تصویر، منطق فازی، الگوریتم های ژنتیک و غیره می باشد. کاربرد موثر Matlab در حل مسایل پردازش سیگنال و سیستم های کنترل است. کاربرد موثر نرم افزار Matlab در حل مسائل داده کاوی می باشد.
R	تحلیل فعل و انفعالات داده و متدهای جدید در روش های آماری، توسط این نرم افزار انجام می گیرد. این نرم افزار دارای ساختاری مشابه دستورات برنامه ی آماری spluse با تغییراتی در نحوه ی برنامه نویسی آماری دارد.
Rapid Miner	نرم افزاری برای متن کاوی در سیستم داده کاوی است، که به عنوان یک برنامه مستقل برای تجزیه و تحلیل داده ها و به عنوان یک موتور داده کاوی در دسترس می باشد.
WEKA	مجموعه ای از الگوریتم های یادگیری ماشینی و ابزارهای پیش پردازش داده ها می باشد. این نرم افزار پشتیبانی های ارزشمندی را برای کل فرآینهای داده کاوی های تجربی فراهم می کنند.
Orange	ای نرم افزاری دانشجویی است و چندان حرفه ای نبوده و بیشتر خصوصیات تجربی دارد.

در این فصل با محیط نرم افزار WEKA، آشنا و روش های داده کاوی در آن مورد بررسی قرار می گیرد. همچنین، امکانات گسترده ی نرم افزار، واسط های گرافیکی، سازگاری با سایر برنامه های ویندوزی مورد بررسی قرار می گیرد.

۸-۱ معرفی نرم افزار WEKA

نرم افزار WEKA، نرم افزاری آزاد و اپن سورس است که توسط دانشگاه وایکاتو[1] در نیوزلند طراحی شده است و اسم آن از عبارت «Waikato Environment For Knowledge» استخراج گشته است. این نرم افزار

[1]- Waikato

برای نخستین بار در سال ۱۹۹۷ به شکل مدرنش نوشته شده و تحت مجوز GPL عرضه شد. چارچوب جاوایی WEKA باعث می شود تا اجرای آن روی سیستم عامل های لینوکس، ویندوز و مکینتاش و حتی روی یک منشی دیجیتال شخصی[1] امکان پذیر باشد.

نرم افزار WEKA، مجموعه ای از الگوریتم های یادگیری ماشین برای انجام وظایف داده کاوی است. این محیط شامل روش هایی برای پردازش اولیه داده، طبقه بندی، رگرسیون، خوشه بندی و قوانین انجمنی می باشد.

۸-۲ معرفی پایگاه داده

جهت آموزش کار از یک مجموعه داده ای بانکی جهت تفهیم بیشتر استفاده خواهد شد که ابتدا این مجموعه داده شرح داده می شود. پیش زمینه داده کاوی، درک درست داده ها و تعریف مسئله است و بدون درک صحیح، نمی توان مسائل را به درستی تعریف کرد و داده ها را جهت کاوش آماده نمود و نتایج را به طور صحیح تفسیر کرد. در واقع هیچ الگوریتمی صرف نظر از خبره بودن آن نمی تواند نتیجه ی مطمئنی حاصل نماید. جهت درک بهتر مطالب این فصل، دو پایگاه داده ۱ Dataset و ۲ Dataset در برنامه Excel با فرمت csv. به عنوان مثال در نظر گرفته می شود،که برای ساخت مدل از قسمت open file نرم افزار قابل استفاده است.

از بانک اطلاعاتی ۱Dataset برای پیاده سازی الگوریتم های رگرسیون، درخت تصمیم و خوشه بندی استفاده شده است. بانک اطلاعاتی شامل اطلاعاتی در مورد ۴۱۱۷ فرد مختلف است که فیلدهای تشکیل دهنده ی آن را در جدول (۸-۲) مشاهده می کنید.

جدول (۸-۲) – نام فیلدهای بانک اطلاعاتی

سن مشتری (numeric)	**سن (Age)**
درآمد مشتری (numeric)	**درآمد (Income)**
مرد یا زن	**جنسیت (Gender)**
مجرد، متاهل، مطلقه	**وضعیت تاهل (Marital)**
تعداد فرزندان (numeric)	**فرزندان (Numkid)**
تعداد وام های دریافتی (numeric)	**وام دریافتی (Loan)**
اعتبارخوب، اعتبار بد	**ریسک (Risk)**

[1]- Personal Digital Assistant

با در نظر گرفتن فیلد هدف (اعتبار خوب، اعتبار بد)، عوامل موثر بر اعتبار متقاضیان وام مورد بررسی قرار می گیرد.

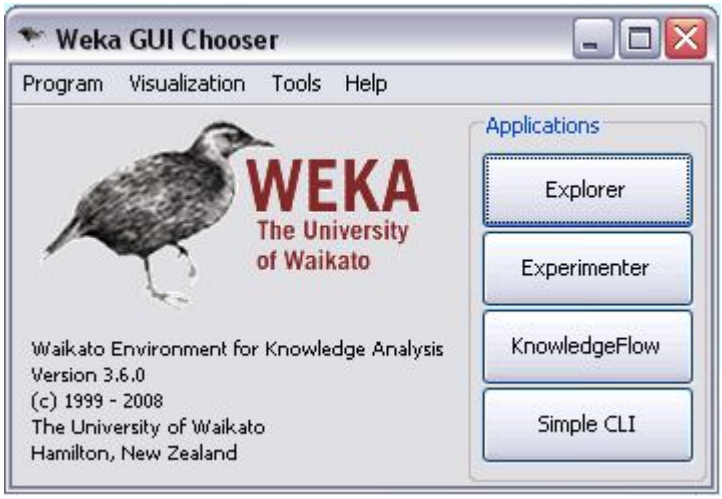
مقادیر برخی فیلدها از نوع عددی می باشند، اگرچه برخی مدل ها ممکن است به الزامات خاص دیگری نیاز داشته باشند.

۸-۳ شروع کار با WEKA

برای اجرای نرم افزار WEKA به دو نکته باید توجه کرد:

۱) برای نصب WEKA می بایست برنامه ی جاوا ماشین(jre – ٦u٢٤ – windows – i٥٨٦- s) نصب گردد. در غیر این صورت با عدم اجرا یا پرش صفحه ی گرافیکی WEKA مواجه می شوید.

۲) داده ها بهتر است در اکسل وارد شده و با فرمت Text delimited ذخیره گردند تا بتوان در نرم افزار WEKA از آن استفاده نمود.



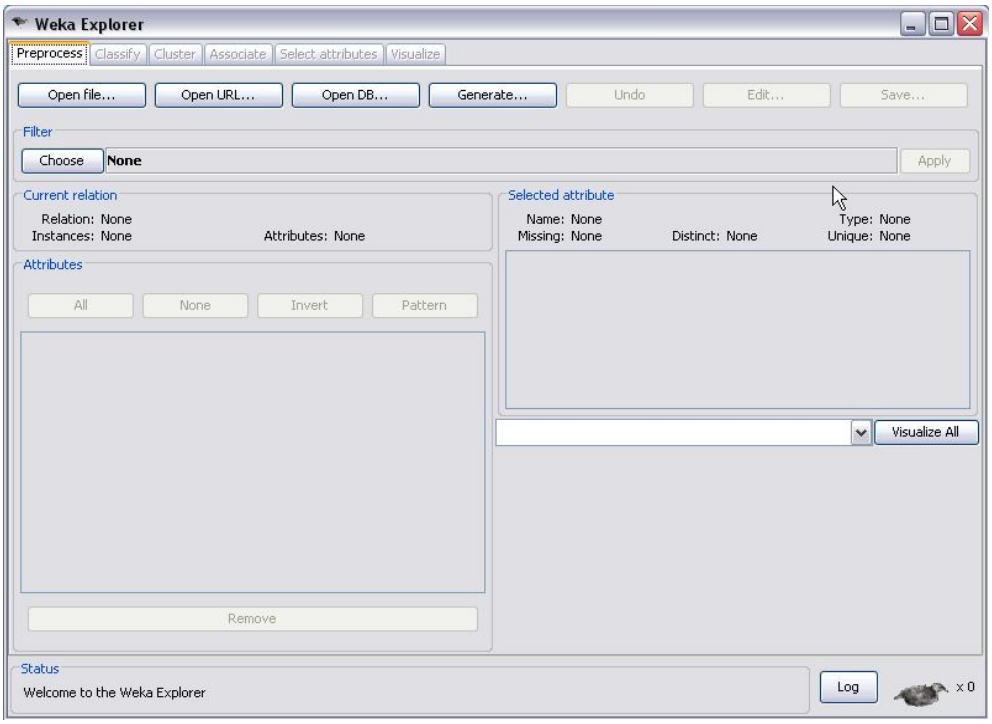
شکل (۸-۲) – صفحه آغازین WEKA

در صورتی که این دو نکته رعایت شود، مشکل خاص دیگری نخواهید داشت. صفحه شروع WEKAبه شکل (۸-۲) خواهد بود.

این صفحه شامل چهار گزینه می باشد:

۱- **Explorer**: این واسط گرافیکی، محیطی برای مکاشفه در داده ها با استفاده از WEKA می باشد، که به وسیله ی انتخاب منوها و پر کردن فرم های مربوطه، دسترسی به همه امکانات را فراهم کرده است. واسط Explorer کمک می کند تا الگوریتم های متعددی نیز آزمایش شوند. شکل (۸-۳) نمای Explorer را نشان می دهد.

در این واسط، شش پانل مختلف وجود دارد. کارکرد تمام گزینه ها به شرح ذیل می باشد:



شکل (۸-۳) – واسط گرافیکی Explorer

- Preprocess : انتخاب مجموعه داده و اصلاح آن از راه های گوناگون.
- Classify : آموزش برنامه های یادگیری که رده بندی یادگیری رگرسیون انجام می دهند و ارزیابی آن ها.
- Cluster : یادگیری خوشه ها برای مجموعه های داده.
- Associate : یادگیری قواعد انجمنی برای داده ها و ارزیابی آن ها.
- Select Attributes : انتخاب مرتبط ترین جنبه ها در مجموعه های داده.
- Visualize : مشاهده نمودارهای مختلف دو بعدی داده ها و تعامل با آن ها.

۲-Experimenter: محیطی برای انجام آزمون و انجام آزمایش های آماری میان روش های مختلف یادگیری می باشد. این واسط کمک می کند تا به هنگام استفاده از تکنیک های رده بندی و رگریون، چه روش ها و پارامترهایی برای مساله داده شده، بهتر عمل می کنند. با این وجود، Experimenter با ساده کردن اجرای رده بندی کننده ها و فیلترها با پارامترهای گوناگون روی تعدادی از مجموعه های داده، جمع آوری کارآیی و انجام آزمایش های معنا، پردازش را خودکار می کند. معمولا کاربرهای پیشرفته، از Experimenter برای توزیع بار محاسباتی بین چندین ماشین، استفاده می کنند.

۳- Knowledge Flow: محیطی که تقریبا تمامی کارایی Explorer را پشتیبانی می کند، اما در عوض از یک رابط «کشیدن و رها کردن[1]» استفاده می کنند که با اتصال الگوریتم های یادگیری و منابع داده ها، ترکیب وچینش دلخواه ساخته می شود.

٤- Simple CLI: رابط ساده خط فرمان که به شما اجازه دسترسی مستقیم به دستورات WEKA را می دهد.

ما Explorer به عنوان واسط کاربری انتخاب می نمائیم. در محیط WEKA Explorer، در پایین هر پانل، جعبه ی Status و دکمه Log قرار دارد. جعبه ی Status، پیغام های مبنی بر انجام عملیات را نشان می دهد و دکمه Log، گزارش عملکرد متنی کارهایی که WEKA تاکنون در این بخش انجام داده را به همراه برچسب زمانی ارائه می کند.
توجه داشته باشید زمانی که WEKA در حال عملیات است، پرنده کوچکی در پایین سمت راست پنجره بالا و پایین می پرد. اگر پرنده بایستد و حرکت نکند، او مریض است؛ یعنی اشتباهی رخ داده است و باید Explorer دوباره اجرا شود.

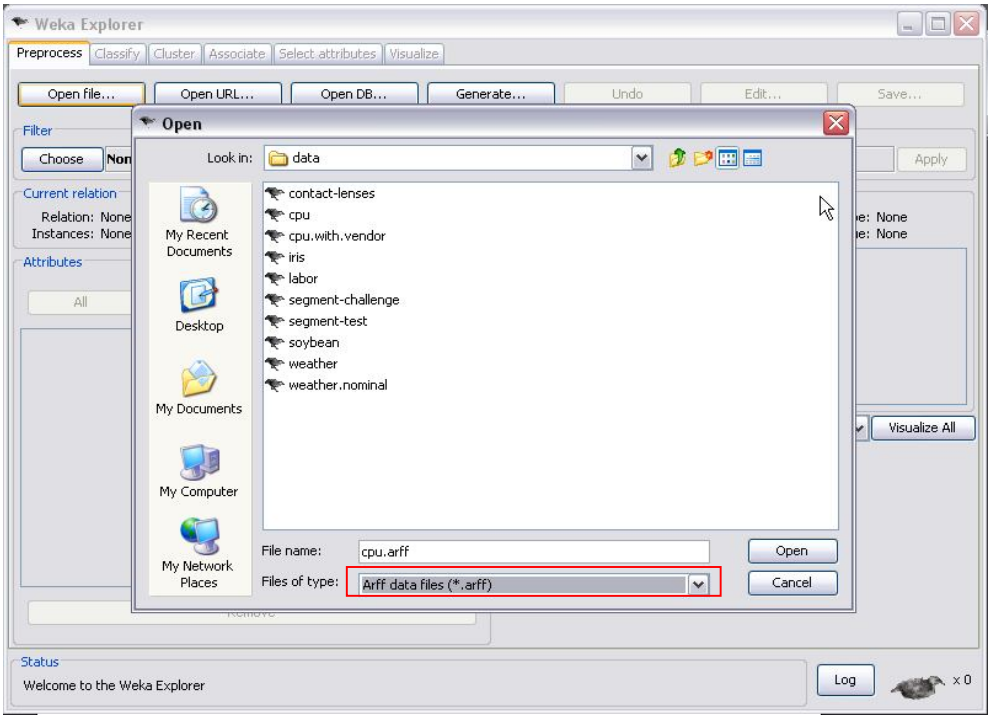
۴-۸ پیش پردازش

ابتدا داده ها به WEKA وارد می شوند که برای این کار باید آن ها را به فرمتی درآورد که برای WEKA فابل فهم باشد. مراحل وارد کردن داده ها در WEKA که به شرح ذیل می باشند:

الف) خواندن و فیلتر کردن فایل‌ها

در بالای پانل Preprocess در شکل (۸-۴)، دکمه های open File، open URL، open database وجود دارد، با کلیک بر روی دکمه open File، ابتدا فایل هایی با پسوند arff در browser فایل نمایش داده می شود. برای دیدن فایل مورد نظر باید گزینه Format در جعبه انتخاب فایل تغییر داده شود.

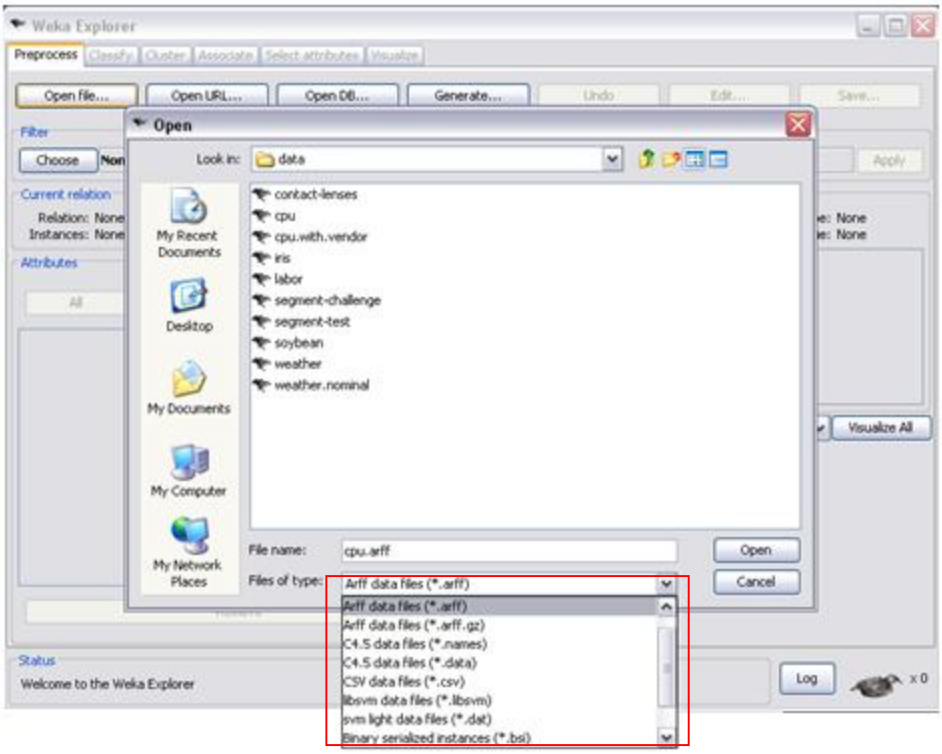
[1]- Drag and Drop



شکل (۸-۴) – باز کردن فایل

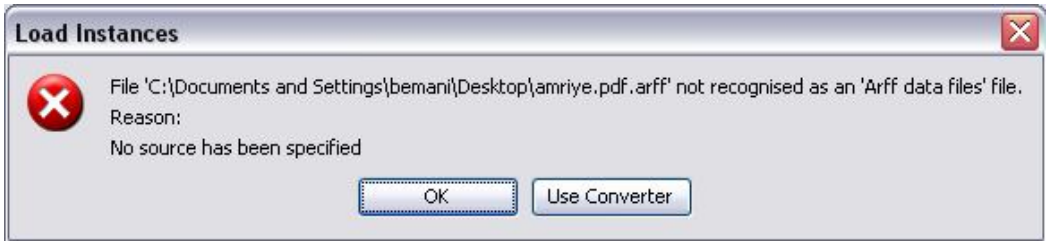
ب) تبدیل فایل‌ها به فرمت ARFF

نرم افزار Weka دارای سه مبدل فرمت فایل می باشد، شکل (۸-۵) برای فایل‌های صفحه گسترده با پسوند CSV، فرمت فایل C٤.٥ با پسوند names و data و برای نمونه های سری با پسوند bsi.

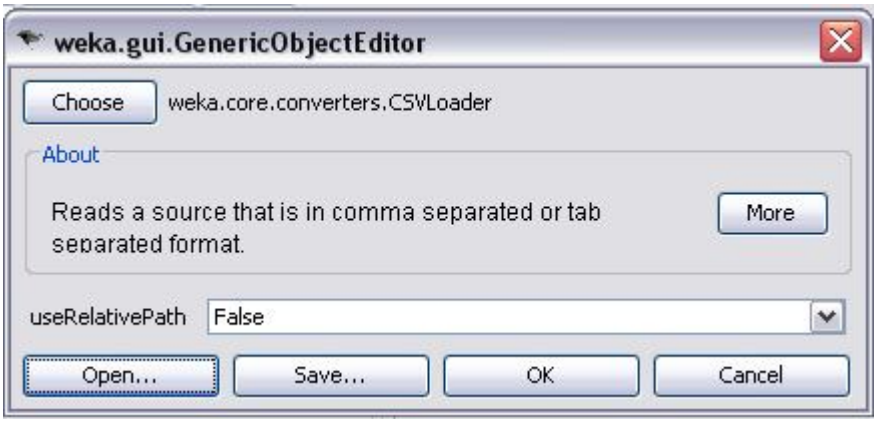


شکل (۸-۵) - تبدیل فرمت های فایل

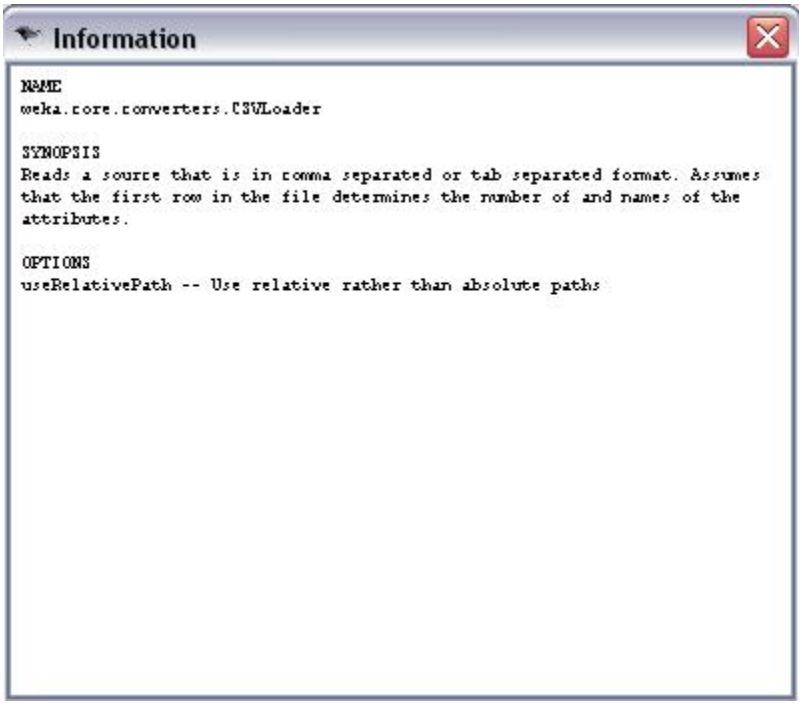
اگر Weka قادر به خواندن داده ها نباشد، آن را به صورت ARFF تفسیر می کند و پیغامی مطابق با شکل (۸-۶- الف) ظاهر می‌شود. با انتخاب گزینه Use Converter، پیغام شکل های (۸-۶- ب) و (۸-۶- ج) ظاهر می شود.



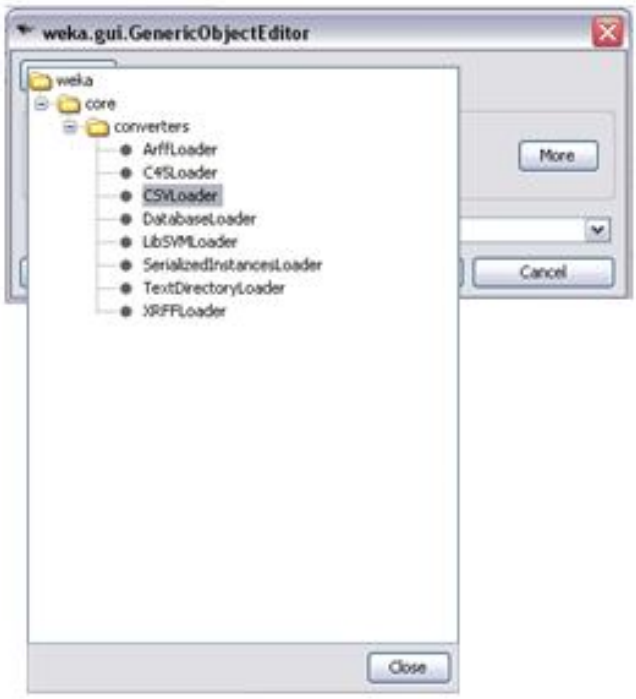
شکل (۸-۶- الف) – پیغام خطا



شکل (۸-۶- ب) - ویرایشگر



شکل (۸-۶- ج) – اطلاعات بیشتر فشردن دکمه more



شکل (۸-۶- د) – انتخاب یک مبدل با فشردن دکمه choose

شکل (۸-۶- ب)، ویرایشگری عمومی در WEKA برای انتخاب و تنظیم اشیا می باشد و برای موارد مختلف لازم است بر روی گزینه choose کلیک شود تا از لیست شکل (۸-۶- د) انتخاب انجام شود. به طور مثال زمانی که پارامتری برای classifier تنظیم می شود، جعبه ای با نوع مشابه بکار برده می شود که در جدول (۸-۳) مشاهده می نمایید.

جدول (۸-۳) – لیست مبدل ها

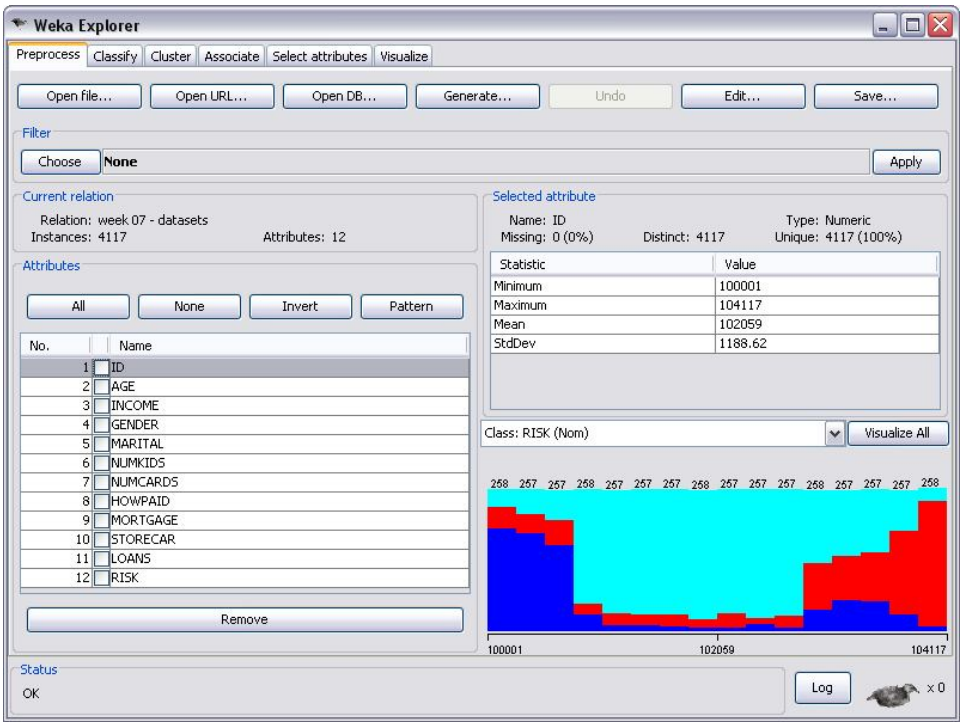
svLoader	فایل هایی با پسوند csv.
ArffLoader	فایل هایی ناموفق
C۴.۵Loader	دو فایل (اسم ها و داده های واقعی)
DatabaseLoader	نمونه های سریالی (برای بازخوانی مجموعه داده ای که به صورت شی سریال شده جاوا ذخیره شده است.)

از ویژگی های دیگر ویرایشگر عمومی در شکل (۸-۶- ب)، save و open می باشد که به ترتیب برای ذخیره اشیای تنظیم شده و باز کردن شی که پیش از این ذخیره شده است، به کار می رود.

در پانل preprocess شکل (۸-۴) open file از فایل های موجود روی کامپیوتر استفاده می کند، علاوه بر آن می توان یک URL باز کرد تا WEKA از پروتکل HTTP برای دانلود کردن یک فایل Arff از شبکه استفاده کند. همچنین می توان با باز کردن open DB، یک پایگاه داده را که درایو اتصال به مجموعه داده به زبان جاوا (JDBC) را دارد، به وسیله دستور select زبان sql، بازیابی کرد. داده ها به کمک save به همه فرمت های ذکر شده، ذخیره می شوند.

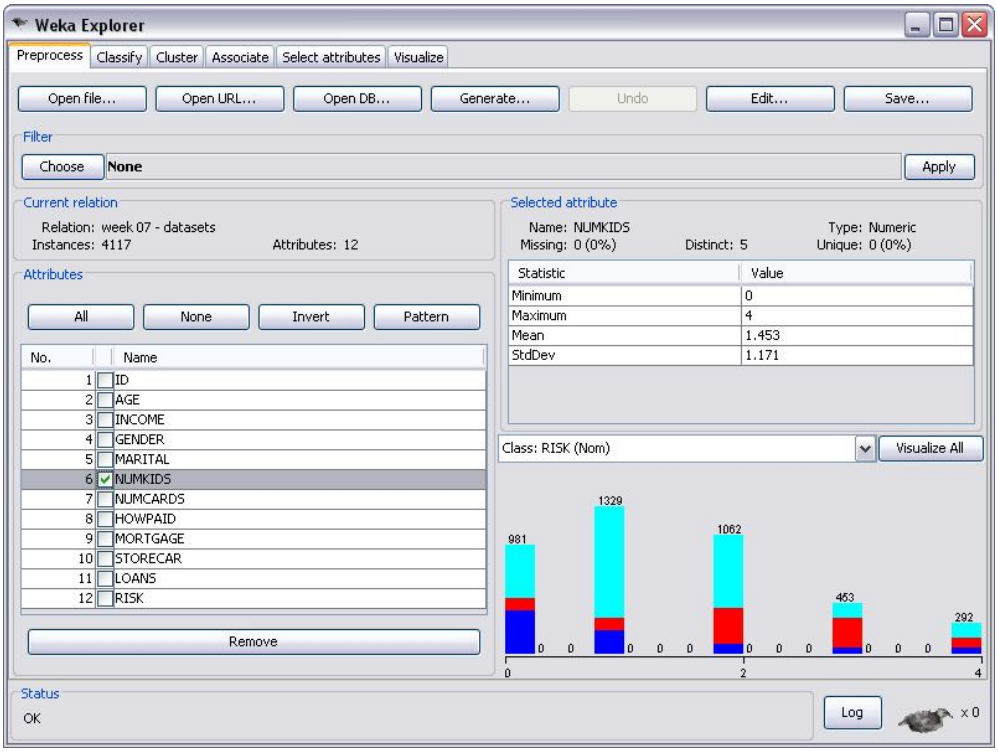
پانل preprocess به کاربر اجازه ی فیلتر کردن داده ها را نیز می دهد. فیلترها اجزای مهم WEKA هستند.

بعد از اینکه فایل بارگذاری شد، WEKA فیلدها را تشخیص می‌دهد و حین بررسی آنها، اطلاعات آماری پایه ای را برای هر کدام از صفات محاسبه می‌کند. همان طور که در شکل (۸-۷) نشان داده شده است، لیست صفات تشخیص داده شده، در سمت چپ، پایین و اطلاعات پایگاه داده مربوطه در بالای آن نشان داده می شود.



شکل (۸-۷) – بانک اطلاعاتی

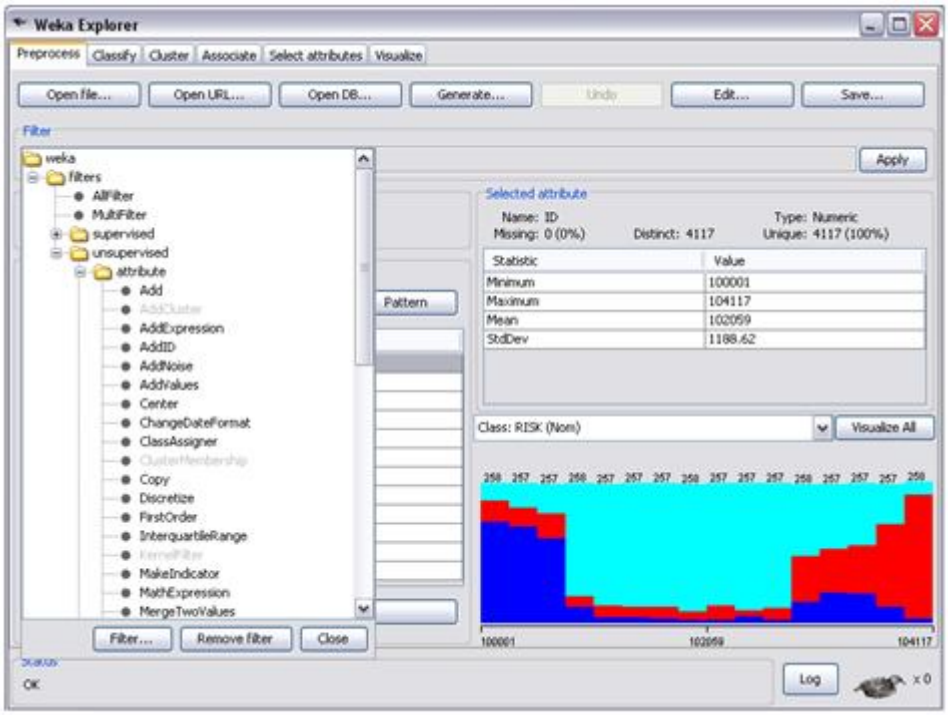
با کلیک کردن بر روی هر کدام از صفات، اطلاعات آماری اصلی آن را در سمت راست قابل مشاهده خواهد بود. نمودار ترسیم شده در سمت راست، بر اساس فیلدهای موجود در پایگاه داده می باشد که می توان آن ها را به دلخواه تغییر داد. به عنوان مثال شکل (۸-۸) از انتخاب فیلد NUMKID حاصل شده است.



شکل (۸-۸) – اطلاعات آماری فیلد NUMKID

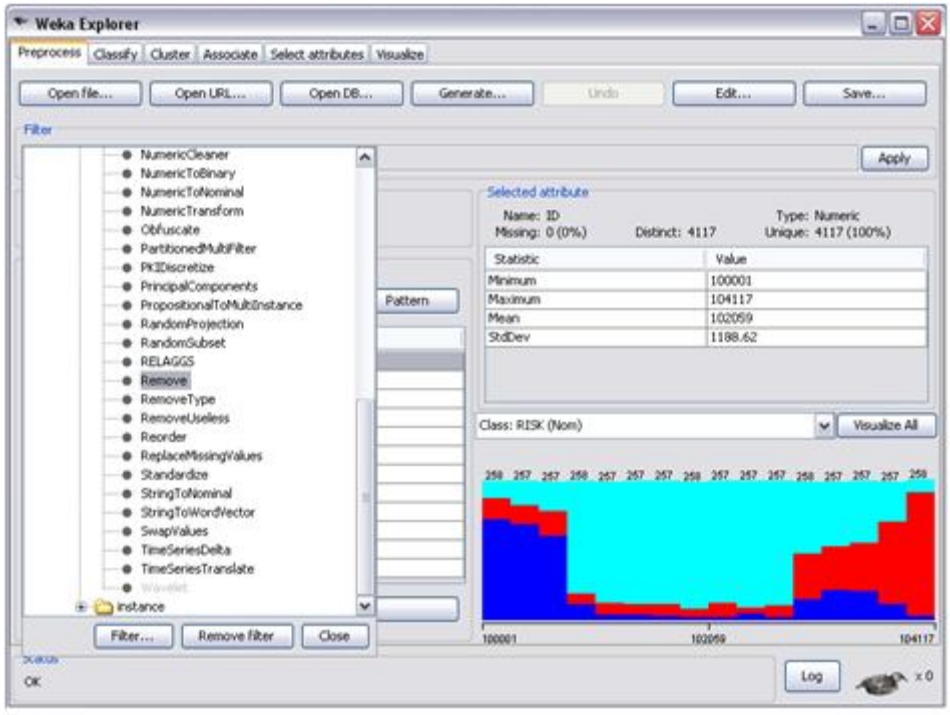
ج) به کارگیری فیلترها

با کلیک کردن choose، در شکل (۸-۹) لیستی از فیلترها مشاهده می شود. از فیلترها برای حذف ویژگی های مورد نظر از یک مجموعه داده استفاده می شود. البته با انتخاب دستی ویژگی ها، با انتخاب آنها(تیک زدن) و انتخاب کلید Remove، امکان حذف وجود دارد.

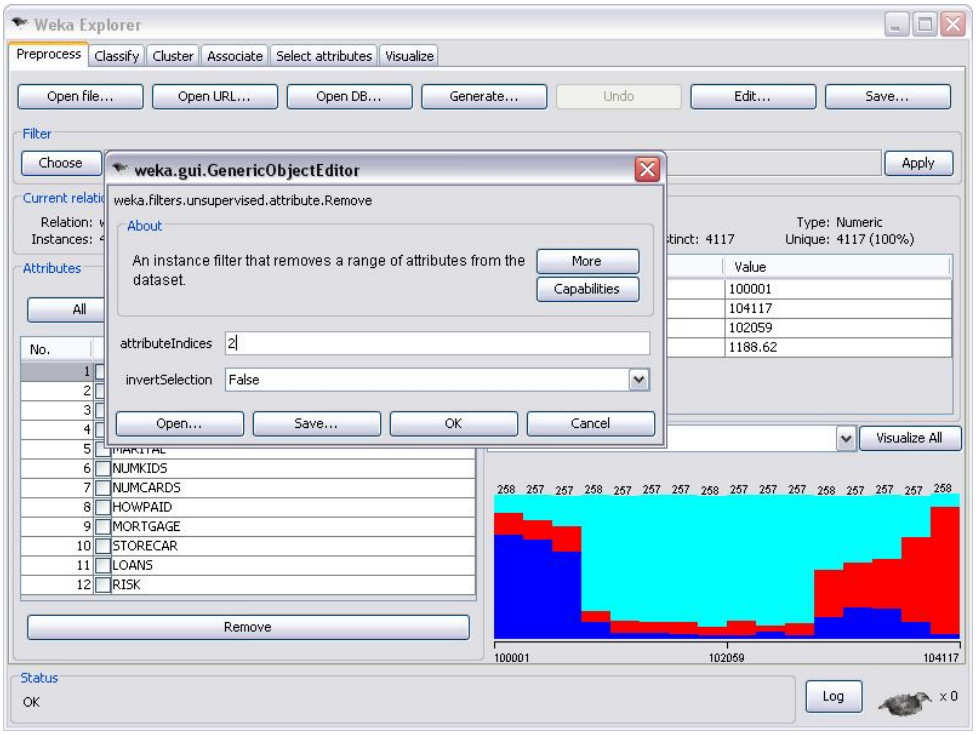


شکل (۸-۹) – انواع فیلترها

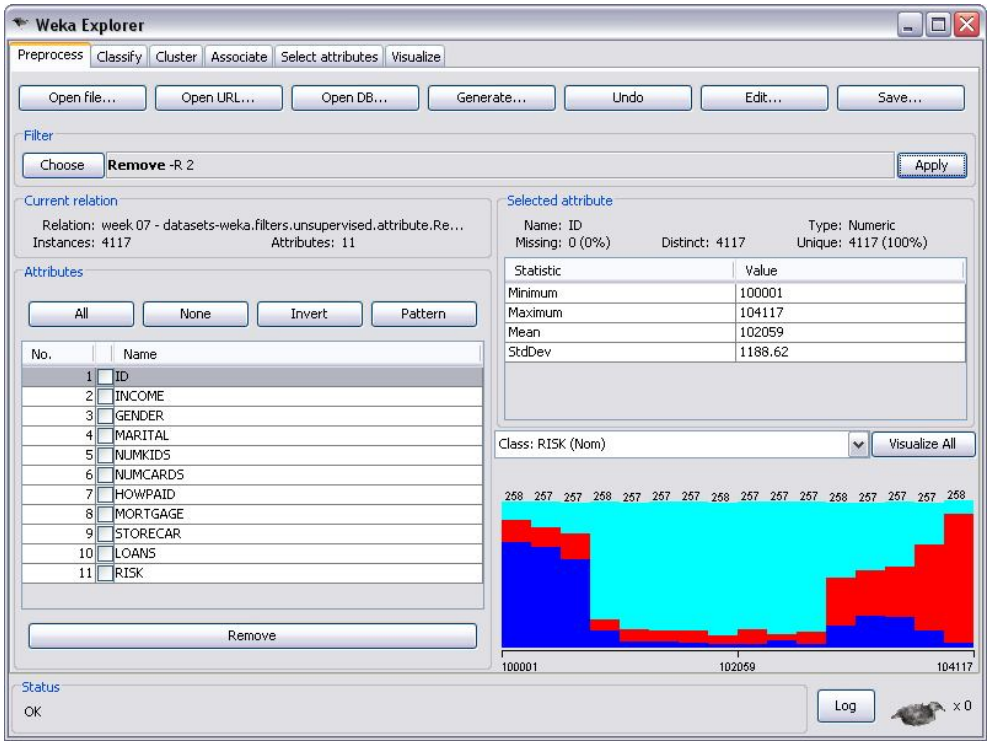
به عنوان مثال مراحل لازم برای حذف فیلد age از بانک اطلاعاتی با استفاده از روش اول، در شکل (۸-۱۰) مشاهده می شود.



شکل (۸-۱۰- الف) – انتخاب فیلتر Remove

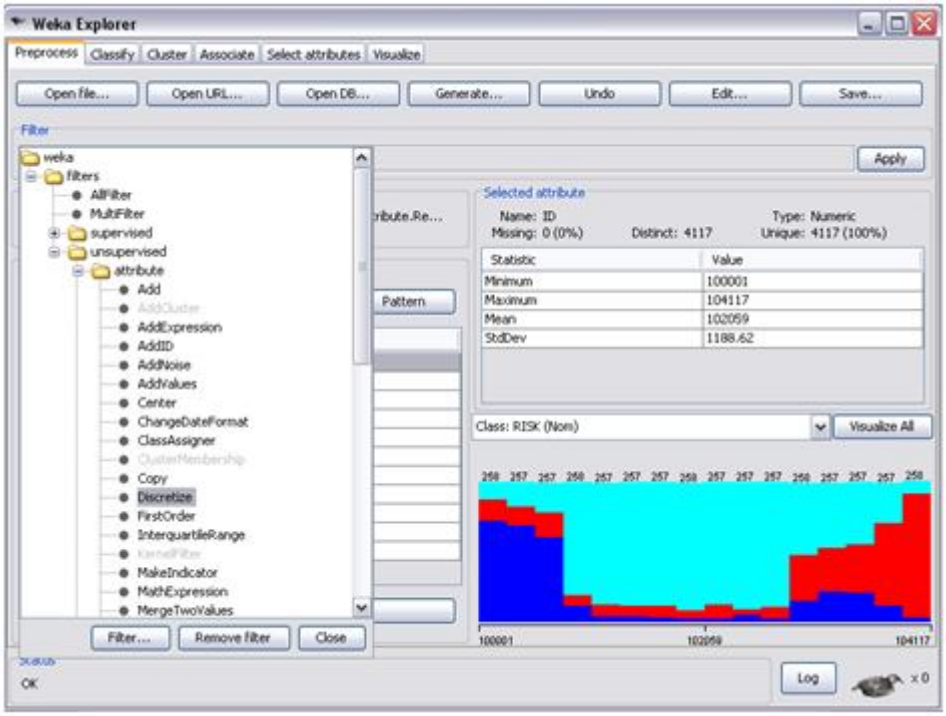


شکل (۸-۱۰- ب) – وارد کردن شماره فیلد مورد نظر

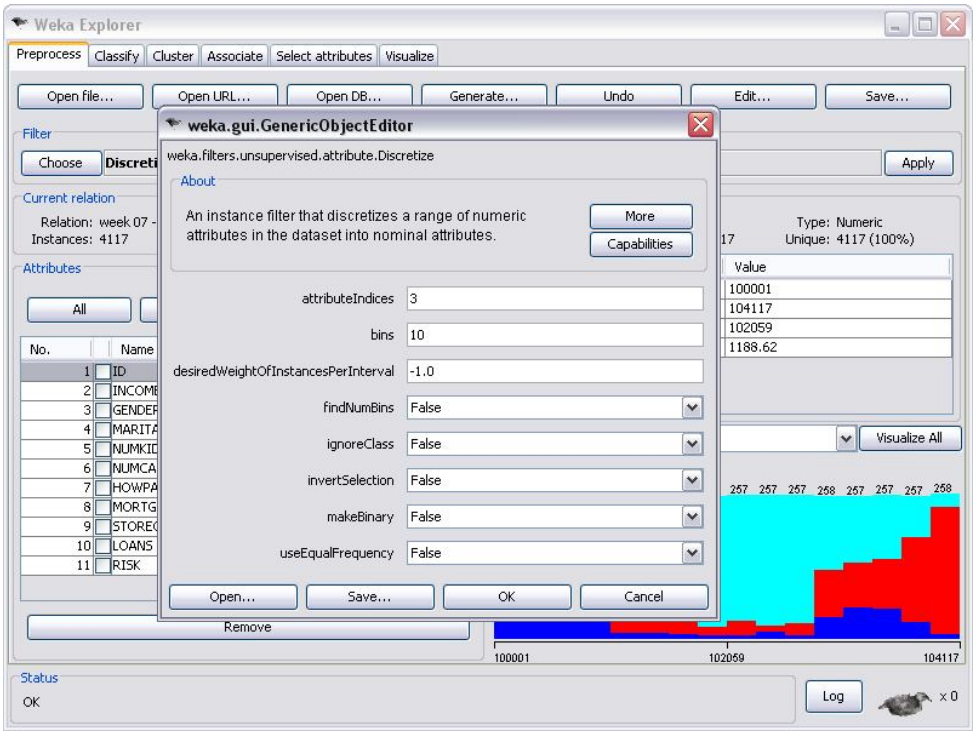


شکل (۸-۱۰- ج) – انتخاب گزینه Apply و حذف فیلد Age

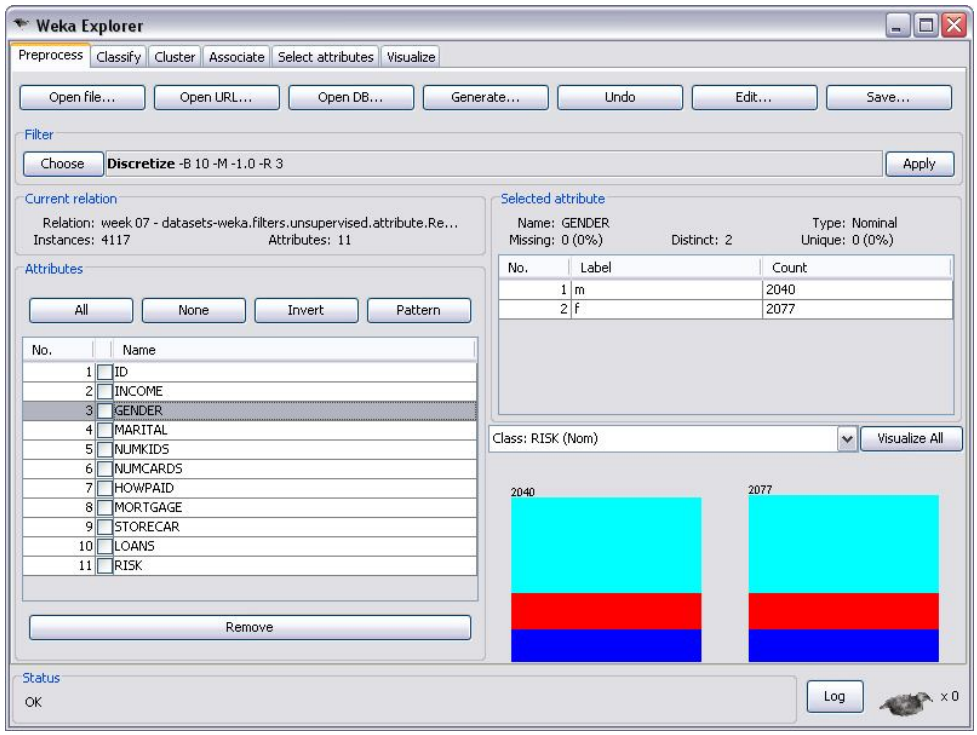
Discretize از دیگر فیلترهای موجود است که با استفاده از آن مقادیر یک صفت پیوسته به تعداد دلخواه بازه گسسته تبدیل می شود. شکل (۸-۱۱) مراحل لازم برای شکستن مقادیر صفت جنسیت به ۲ بازه را نشان می دهد.



شکل (۸-۱۱- الف) – انتخاب فیلتر Discretize



شکل (۸-۱۱- ب) –وارد کردن شماره فیلد مورد نظر و انجام تنظیمات



شکل (۸-۱۱- ج) –انتخاب گزینه Apply

Classify ۵-۸

WEKA الگوریتم های طبقه بندی و پیش بینی بسیار متنوعی را پیاده سازی می کند. طبقه بندی روشی است برای مشخص کردن گروهی که یک نمونه داده در آن قرار می گیرد. الگوریتم های طبقه بندی به Bayesian، Functions، lazy، Meta، Misc، trees و Rules تقسیم می شوند. در این بخش به معرفی برخی از اسامی طبقه بندی های WEKA پرداخته می شود.

- **Trees**

✓ Decision stump، برای مجموعه داده های عددی یا رده‌ای، درخت تصمیم گیری یک سطحی می سازد، که با مقادیر از دست رفته، به صورت مقادیر مجزا برخورد کرده و شاخه سومی از درخت را توسعه می دهد.

- **Rules**

✓ Decision Table، طبقه بندی بر اساس اکثریت جدول تصمیم‌گیری می‌سازد. این الگوریتم، با استفاده از بهترین جستجوی اول، زیر دسته های ویژگی‌ها را ارزیابی نموده و می‌تواند از اعتبارسنجی تقاطعی برای ارزیابی استفاده کند.

✓ N-nge، به جای استفاده از اکثریت جدول تصمیم گیری که بر اساس دسته ویژگی‌های مشابه عمل می‌کند، از روش نزدیک‌ترین همسایه برای تعیین رده هر یک از نمونه ها که توسط مدخل (Entry) جدول تصمیم گیری پوشش داده نشده اند، استفاده می شود.

✓ Conjunctive Rule، قاعده‌ای را یاد می‌گیرد که مقادیر رده های عددی را پیش‌بینی می کند. نمونه های آزمایشی به مقادیر پیش فرض رده نمونه های آموزشی، منسوب می‌شوند. سپس تقویت اطلاعات (برای رده های رسمی)، یا کاهش واریانس (برای رده های عددی) مربوط به هر والد محاسبه شده و به روش هرس کردن با خطای کاهش یافته (Reduced-error pruning)، قواعد هرس می‌شوند.

✓ ZeroR، برای رده های اسمی، اکثریت داده های مورد آزمایش و برای رده های عددی، میانگین آن‌ها را پیش‌بینی می کند. این الگوریتم بسیار ساده است.

✓ M۵ Rules، به کمک M۵ از روی درخت‌های مدل، قواعد رگرسیون را استخراج می کند.

- Functions

✓ Simple Linear Regression، مدل رگرسیون خطی یک ویژگی مشخص را یاد می گیرد، آنگاه مدل با کمترین خطای مربعات، را انتخاب می کند. در این الگوریتم، مقادیر از دست رفته و مقادیر غیرعددی مجاز نیستند.

✓ Linear Regression، رگرسیون خطی استاندارد کمترین خطای مربعات را انجام می دهد و می تواند به طور اختیاری به انتخاب ویژگی بپردازد، این کار می تواند به صورت حریصانه، با حذف عقب رونده (Backwarde limination) انجام شود، یا با ساختن یک مدل کامل از همه ویژگی ها و حذف یکی یکی جمله ها با ترتیب نزولی ضرایب استاندارد شده آن ها، تا رسیدن به شرط توقف مطلوب انجام گیرد.

✓ Least Med sq، یک روش رگرسیون خطی مقاوم است که به جای میانگین مربعات انحراف از خط رگرسیون، میانه را کمینه می کند. این روش به طور مکرر رگرسیون خطی استاندارد را به زیرمجموعه هایی از نمونه ها اعمال می کند و نتایجی را بیرون می دهد که کمترین خطای مربع میانه را دارند.

✓ SMOreg، الگوریتم بهینه سازی حداقل ترتیبی را روی مسایل رگرسیون اعمال می کند.

✓ Pace Regression، با استفاده از تکنیک رگرسیون pace، مدل‌های رگرسیون خطی تولید می‌کند. رگرسیونpace، زمانی که تعداد ویژگی ها خیلی زیاد است، به طور ویژه ای در تعیین ویژگی هایی که باید صرف‌نظر شوند، خوب عمل می‌کند. در واقع در صورت وجود نظم و ترتیب خاصی، ثابت می شود که با بینهایت شدن تعداد ویژگی ها، این الگوریتم بهـینه عمل می کند.

✓ RBF Network، یک شبکه با تابع پایه ای گوسی شعاعی را پیاده سازی می کند. مراکز و عرضه ای واحدهای مخفی به وسیله روش میانگین K (K-means) تعیین می شود. سپس خروجی های فراهم شده از لایه های مخفی (Hidden layer)، با استفاده از رگرسیون منطقی در مورد رده های اسمی و رگرسیون خطی در مورد رده های عددی، با یکدیگر ترکیب می‌شوند. فعال سازی های توابع پایه پیش از ورود به مدل های خطی، با جمع شدن با عدد یک، نرمالیزه می شوند. در این الگوریتم می‌توان K، تعداد خوشه ها، بیشترین تعداد تکرارهای رگرسیون های منطقی برای مساله های رده های رسمی، حداقل انحراف معیار خوشه ها، و مقدار بیشینه رگرسیون را تعیین نمود. اگر رده‌ها رسمی باشد، میانگین K به طور جداگانه به هر رده اعمال می شود تا K خوشه مورد نظر برای هر رده استخراج گردد.

- رده بندهای Lazy

یادگیرنده های lazy نمونه های آموزشی را ذخیره می کنند و تا زمان رده بندی هیچ کار واقعی انجام نمی دهند.

- IB۱

یک یادگیرنده ابتدایی بر پایه نمونه است که نزدیکترین نمونه های آموزشی به نمونه های آزمایشی داده شده را از نظر فاصله اقلیدسی پیدا کرده و نزدیکترین رده‌ای مشابه رده همان نمونه های آموزشی را تخمین می زند.

- IBK

یک رده بند با K همسایه نزدیک است که معیار فاصله ذکر شده را استفاده می‌کند. تعداد نزدیکترین فاصله ها (پیش فرض k=۱)، می‌تواند به طور صریح در ویرایشگر شیء تعریف شود. پیش‌بینی های متعلق به پیش از یک همسایه می تواند بر اساس فاصله آنها تا نمونه های آزمایشی، وزن دار گردد.

دو فرمول متفاوت برای تبدیل فاصله به وزن، پیاده سازی شده اند. تعداد نمونه های آموزشی که به وسیله رده بندی نگهداری می شود، می تواند با تنظیم گزینه اندازه پنجره محدود گردد. زمانی که نمونه های جدید اضافه می شوند، نمونه های قدیمی حذف شده تا تعداد کل نمونه های آموزشی در اندازه تعیین شده باقی بماند.

- Kstar

یک روش نزدیکترین همسایه است که از تابع فاصله ای عمومی شده بر اساس تبدیلات استفاده می کند.

- LWL

یک الگوریتم کلی برای یادگیری وزن دار شده به صورت محلی است. این الگوریتم با استفاده از یک روش بر پایه نمونه، وزن‌ها را نسبت می دهد و از روی نمونه های وزندار شده، رده‌بند را می سازد. رده‌بند Nave Bayes، در ویرایشگر شیء LWL انتخاب می‌شود. برای مسایل رده بندی و رگرسیون خطی برای مسایل رگرسیون، انتخابهای خوبی هستند. می‌توان در این الگوریتم، تعداد همسایه های مورد استفاده را که پهنای باند هسته و شکل هسته مورد استفاده برای وزن دار کردن را (خطی، معکوس، یا گوسی) مشخص می کند، تعیین نمود. نرمال سازی ویژگی‌ها به طور پیش فرض فعال است.

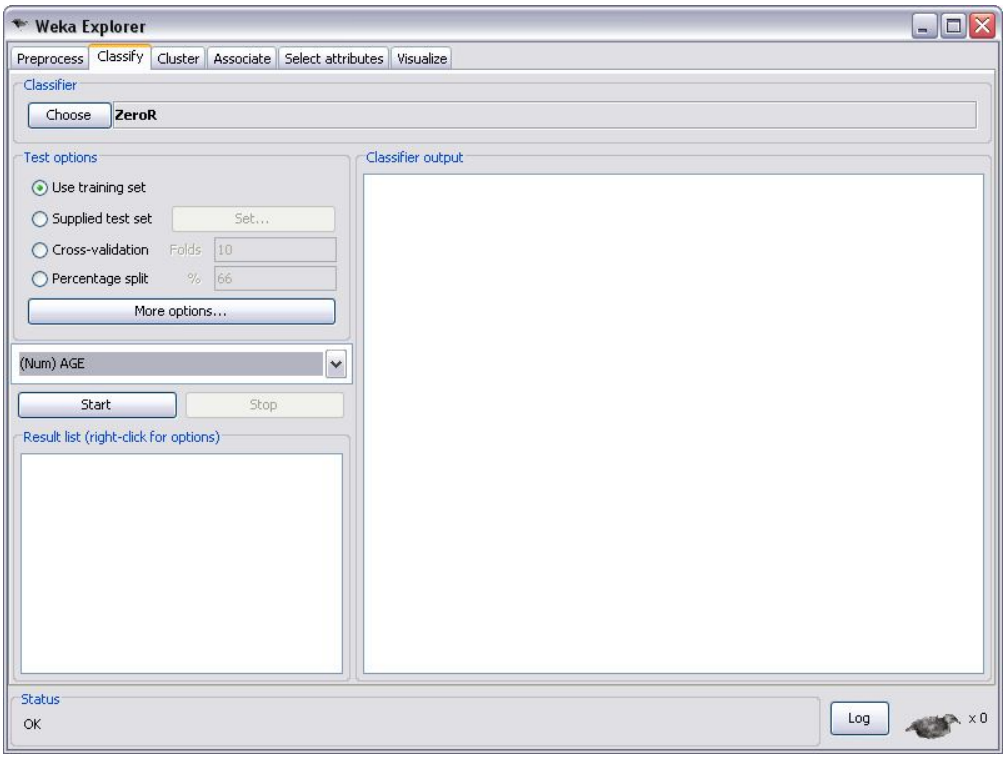
الگوریتم های طبقه بندی به صورت تئوری معرفی شدند. با مثال های عملی نحوه کار با classifier ها را نشان خواهیم داد.

۸-۵-۱ رگرسیون

مدل رگرسیون در عین سادگی می تواند اطلاعات مفیدی را در اختیار پیش بینی کننده قرار دهد. شیوه ی کار به این صورت است که رابطه میان یک سری متغیر مستقل و یک متغیر وابسته کشف می شود. بدین معنی که با داشتن مقادیر متغیرهای مستقل، متغیر وابسته پیش بینی می شود.

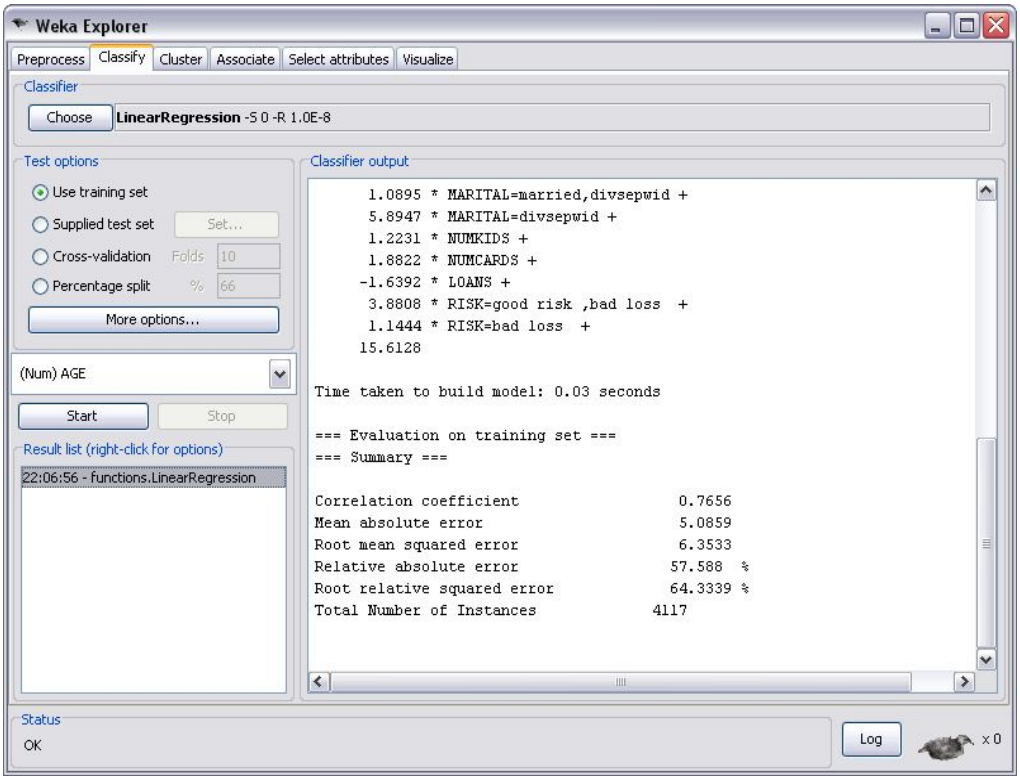
به عنوان مثال از بانک اطلاعاتی Dataset [۱] استفاده می شود. متغیرهای مستقل این پیش بینی، ریسک، درآمد، جنسیت، وضعیت تاهل، تعداد فرزندان و تعداد وام دریافتی می باشند. حال با استفاده از این شش متغیر، متغیر وابسته «سن» پیش بینی می شود.

بعد از باز کردن فایل مورد نظر و با کلیک کردن بر روی پانل Classify، پنجره های مطابق شکل (۸-۱۲) باز می شود. در این قسمت باید مدل انتخاب شود که مدل، رگرسیون خطی می باشد. برای این کار روی choose کلیک کره، سپس به بخش Function رفته و LinearRegression را انتخاب کنید (با کلیک کردن بر روی remove filter تمامی آیتم های function فعال می شود). همان طور که مشاهده می کنید، برای مدلسازی گزینه های متعدد دیگری نیز وجود دارد که نشان از قدرت بالای نرم افزار WEKA دارد. حال باید برای نرم افزار مشخص کرد تا از چه داده هایی برای ساخت مدل استفاده کرد. چهار گزینه وجود دارد، اولی استفاده از Training Set است که در اینجا از آن استفاده می شود. این گزینه به این معناست که WEKA باید داده ها را از فایلی که در بخش قبل وارد شده، بگیرد. Supplied test set برای زمانی است که شما می خواهید از داده های متفاوتی برای مدل سازی استفاده کنید. Cross-validation در واقع از روشی با همین نام استفاده می کند که در آن ابتدا داده ها به یک سری زیر مجموعه تقسیم شده و آنالیز روی یک زیر مجموعه آن که مجموعه تمرینی نام دارد، انجام می شود. پس این آنالیز روی زیر مجموعه های دیگر که مجموعه آزمون نام دارد، تایید می شود، شکل (۸-۱۷).



شکل (۸-۱۲) – انتخاب پانل classify

در نهایت، percentage split نیز تعداد اندکی از داده ها را انتخاب کرده و از آن برای ساخت مدل استفاده می کند. مرحله نهایی انتخاب متغیر وابسته می باشد که به پیش بینی آن پرداخته می شود. این متغیر AGE است. پس در کامبو باکس، AGE را انتخاب و روی START کلیک کنید. در شکل (۸-۱۳) مدل ایجاد شده، مشاهده می شود.



شکل (۸-۱۳) – اجرای مدل رگرسیون روی داده ها

با توجه به مدل بدست آمده در بالا می توان تفسیرهای زیر را انجام داد:

WEKA تنها ستون هایی را در تحلیل دخالت می دهد که به دقت مدل کمک کنند. بدین ترتیب، ستون هایی که در ساخت یک مدل خوب بی تاثیر هستند، کنار گذاشته می شوند. در شکل (۸-۱۴) عوامل موثر طبق مدل رگرسیون مشاهده می شود.

```
AGE =

      0.0003 * INCOME +
      0.3862 * GENDER=f +
      1.0895 * MARITAL=married,divsepwid +
      5.8947 * MARITAL=divsepwid +
      1.2231 * NUMKIDS +
      1.8822 * NUMCARDS +
     -1.6392 * LOANS +
      3.8808 * RISK=good risk ,bad loss  +
      1.1444 * RISK=bad loss  +
     15.6128
```

شکل (۸-۱۴) – عوامل موثر در مدل

۸-۵-۲ درخت تصمیم

طبقه بندی ماهیتی متفاوت با رگرسیون دارد. در ساده ترین حالت، می توان طبقه بندی را به صورت یک درخت در نظر گرفت. در اینجا خروجی دریافت شده به یک عدد محدود نیست و می توان از نتیجه حاصل، برداشت های عمیق تری کرد.

قبل از شروع کار با WEKA باید چندین نکته مهم را درباره طبقه بندی در نظر گرفت:

۱- در رگرسیون تنها از یک مجموعه داده استفاده می شود و یا آن مدل مشخص می شود. اما در طبقه بندی بین شصت تا هشتاد درصد داده های موجود به عنوان «مجموعه آموزشی[۱]» و بقیه آنان در قالب «مجموعه آزمایشی[۲]» به کار می روند، یعنی در ابتدا مدل با استفاده از مجموعه آموزشی مشخص شده، سپس با استفاده از مجموعه آزمایشی دقت آن امتحان می شود. دلیل استفاده از دو مجموعه داده در زمینه طبقه بندی، جلوگیری از تطابق بیش از اندازه یا Overfitting است. این پدیده زمانی رخ می دهد که شما یک مدل را بیش از اندازه به یک مجموعه داده خاص وابسته سازید. در این حالت با وجود این که مدل درباره ی آن مجموعه از داده ها عملکرد بسیار دقیقی از خود ارائه می دهد، برای دیگر داده ها دقت خوبی را نشان نمی دهد. در واقع در این حالت مدل به جای یادگیری یک مجموعه داده، ان را به خاطر می سپارد. با توجه به این که هدف اصلی از ساخت چنین مدلی ارائه پیش بینی مناسب برای داده های جدید بوده، باید سعی کرد تا جای ممکن از Overfitting جلوگیری شود. به همین دلیل، از مجموعه داده های آزمایشی استفاده می شود تا پس از ساخت مدل، دقت آن در مواجهه با داده های جدید آزمایش شود.

۲- اندازه درخت های طبقه بندی یکی از مسائل مهمی است که در کارکرد آن ها اثر مستقیم دارد. اگر یک درخت زیادی بزرگ شود، ریسک Overfitting روی مجموعه داده های آموزشی پیش می آید. از طرف دیگر یک درخت کوچک ممکن است برخی فاکتورهای مهم را نادیده بگیرد. برای این که اندازه مناسب درخت به دست آورده شود، از روشی با عنوان «هرس کردن» استفاده می شود. در این روش با اعمال هرس روی یک درخت بزرگ سعی می شود تا گره هایی که اطلاعات اضافه ای در بر ندارند، حذف شوند. هرس کردن باعث می شود تا توازنی میان دقت درخت و سادگی آن برقرار شود.

۳- اشاره ای به مثبت کاذب[۳] و منفی کاذب[۴] می شود. مثبت کاذب زمانی رخ می دهد که آزمایشی آماری، فرضی صحیح را رد کند.

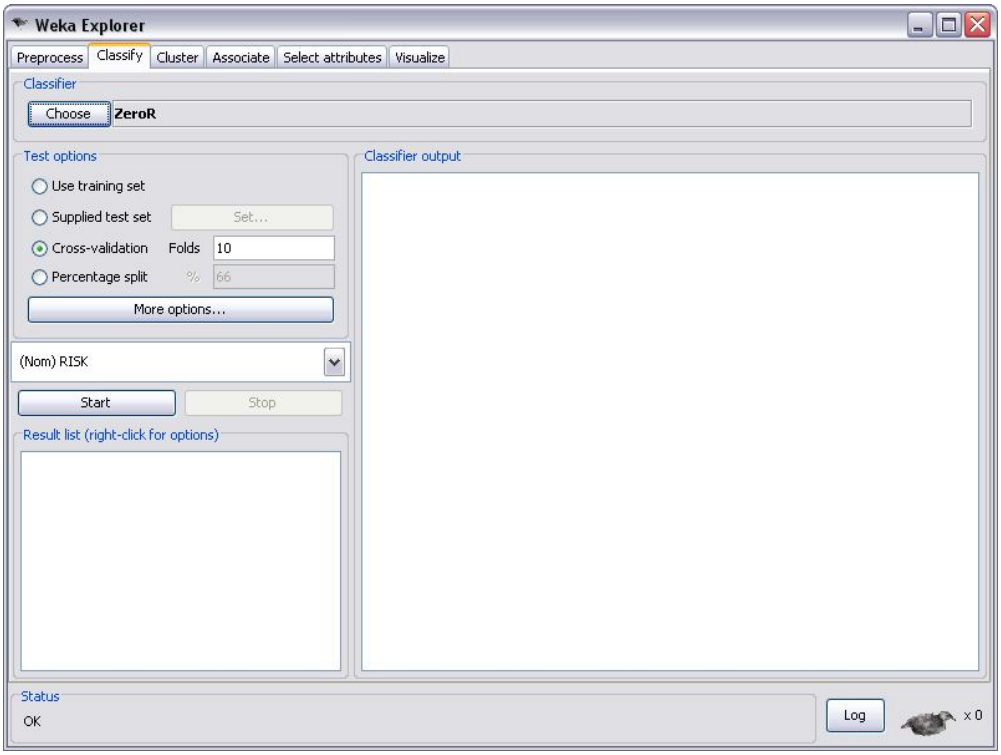
---

[1]- Training set
[2]- Test set
[3]- False Positive
[4]- False Negative

منفی کاذب نیز به صورت مشابه هنگامی رخ می دهد که آزمایش فرضی نادرست را بپذیرد. سازنده یک مدل باید مشخص کند که مفاهیم تا چه اندازه ای قابل قبول است.

در ادامه با یک مثال عملی نحوه ی کار با الگوریتم نشان داده شده است. همان طور که گفته شد برای اجرای مناسب طبقه بندی باید داده ها را به دو دسته ی آموزشی و آزمایشی تقسیم کنید. مهم ترین نکته در این تقسیم بندی تصادفی بودن آن است. باید سعی کنید این تقسیم بندی به صورت کاملا تصادفی انجام شود، در غیر این صورت مجموعه آموزشی و آزمایشی شما نشان دهنده یک جامعه نخواهد بود. بنابر این برای هر مجموعه یک فایل جداگانه بسازید.

بانک اطلاعاتی ۱ Dataset را در نظر گرفته و با باز کردن فایل مورد نظر، بر روی پانل classify کلیک کنید. پنجره ای مطابق شکل (۸-۱۵) باز می شود.

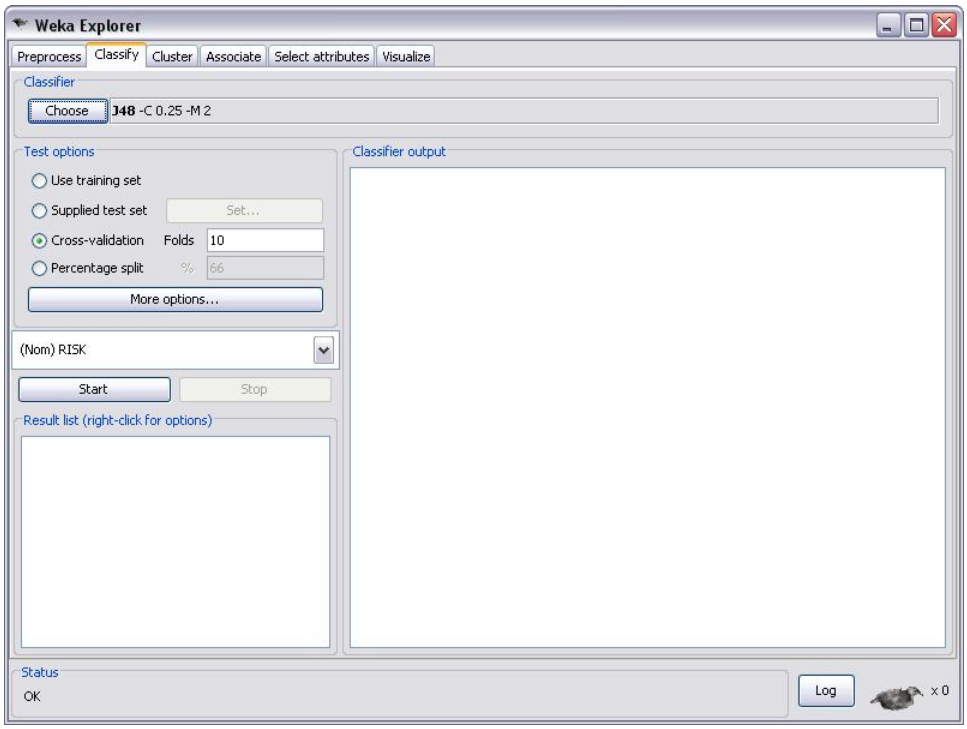


شکل (۸-۱۵) – انتخاب پانل classify

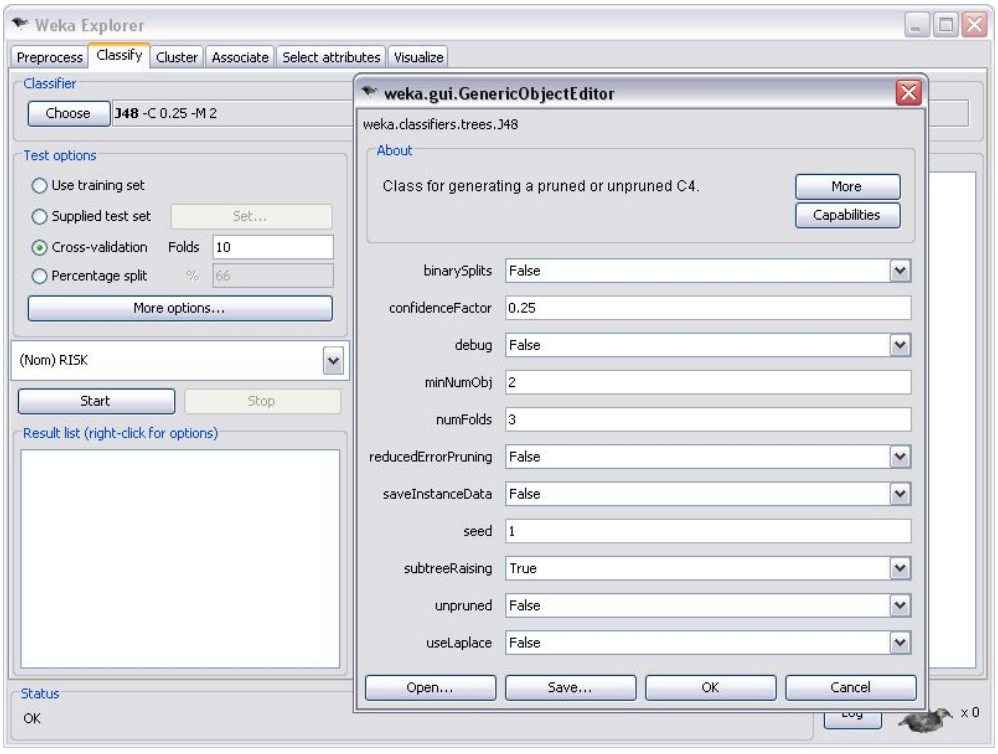
با کلیک کردن بر روی choose در پانل classify می توان الگوریتم رده بندی مورد نظر را انتخاب نمود، شکل (۸-۱۶). در این مثال، الگوریتم j۴۸ انتخاب می شود. زمانی که یک الگوریتم رده بندی انتخاب می شود، نسخه خط فرمانی[1] رده بند در سطری نزدیک به دکمه ظاهر می گردد. این خط فرمان شامل پارامترهای الگوریتم است که با خط تیره مشخص می شوند. برای تغییر آن ها می توان روی آن خط کلیک

[1]- Command line

نمود تا ویرایشگر مناسب شی باز شود، شکل (۸-۱۷). در اینجا همان مقادیر پیش فرض در نظر گرفته می شود.

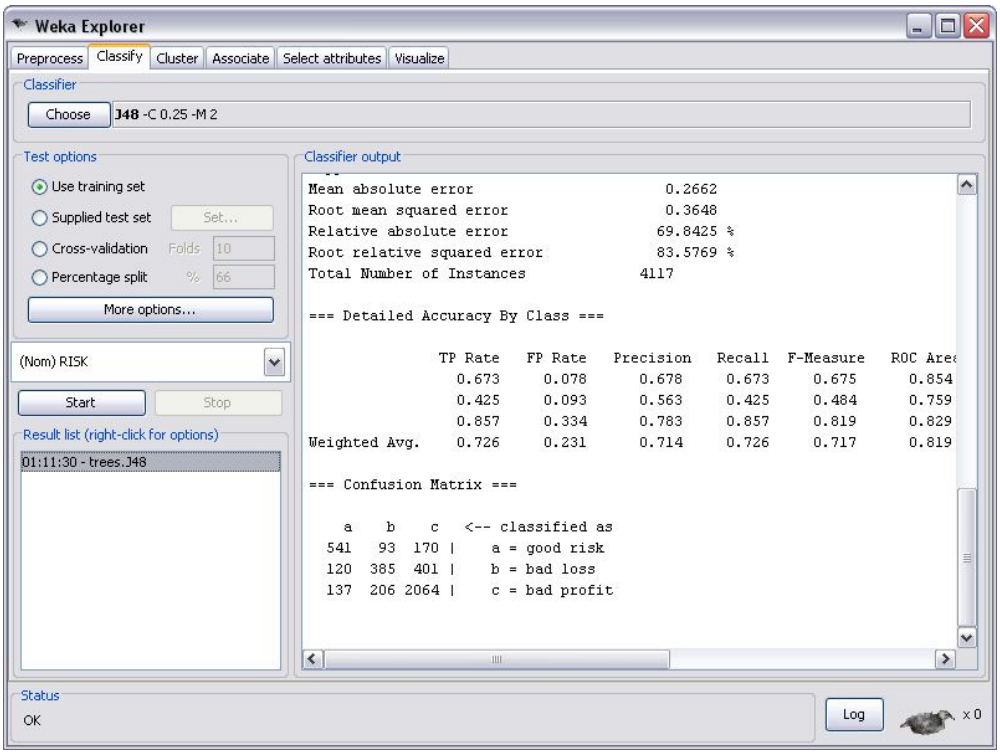


شکل (۸-۱۶) – انتخاب الگوریتم رده بندی



شکل (۸-۱۷) – تنظیم پارامترهای الگوریتم رده بندی

با کلیک کردن بر روی دکمه start مدل مورد نظر تولید می شود، شکل (۸-۱۸).



شکل (۸-۱۸) – مدل حاصل از اجرای الگوریتم طبقه بندی

از جمله اعداد مهمی که در خروجی مشاهده می شود:

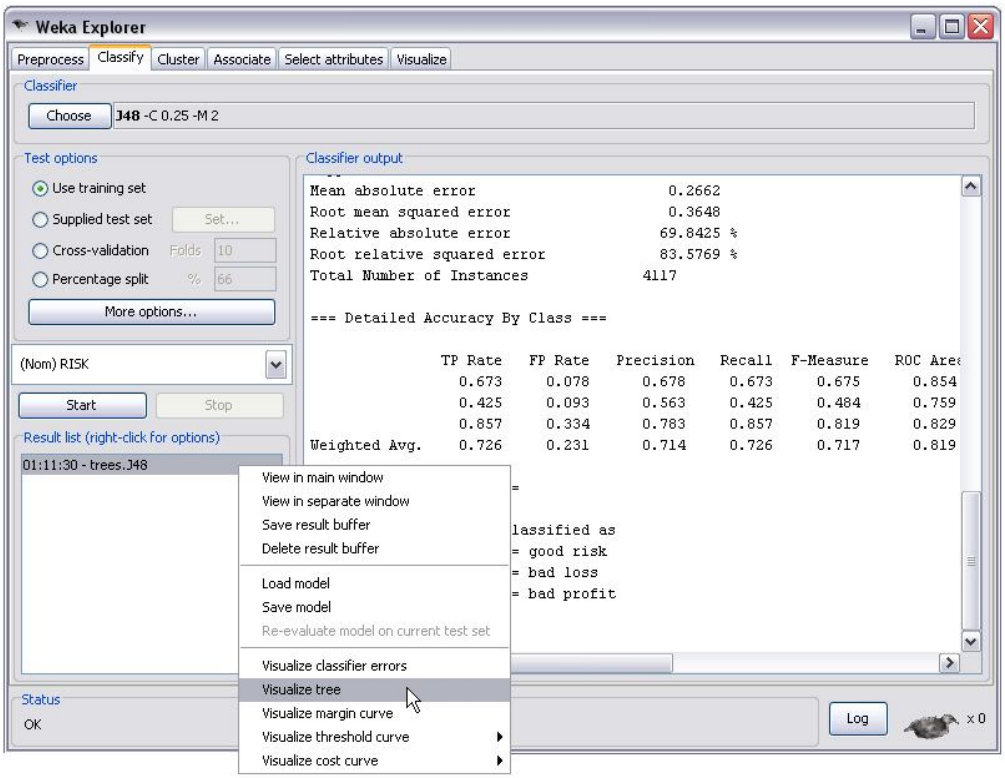
- correctly classified instances: ۷۲/۶۲ %
- incorrectly classified instances: ۲۷/۳۷ %

۷۲/۶۲ درصد نشان می دهد بیش از ۷۲ درصد از نمونه ها به درستی و ۲۷ درصد نادرست دسته بندی شده اند.

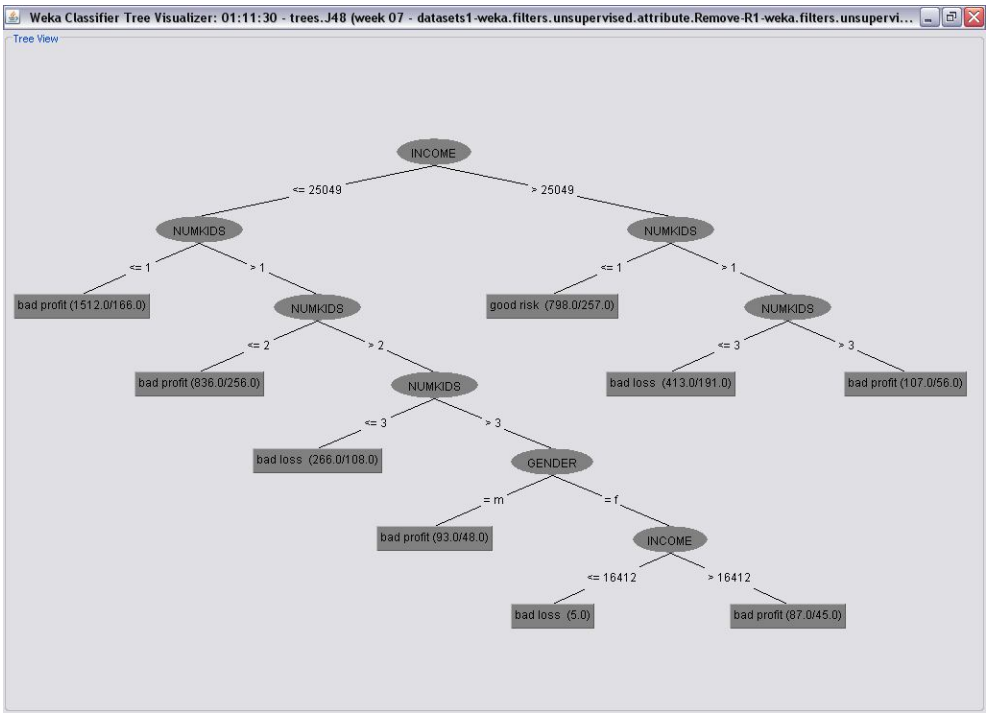
در قسمت Confusion Matrix میزان مثبت کاذب و منفی کاذب را مشخص می کند.

با راست کلیک بر روی مجموعه جواب در پانل Result list در سمت چپ، و با انتخاب visualize tree می توان نتیجه را در پنجره های جداگانه، و یا شکل گرافیکی درخت حاصل از طبقه بندی را مشاهده نمود، شکل (۸-۱۹).

توجه کنید که در شکل (۸-۲۰) با راست کلیک بر روی یک قسمت خالی از صفحه می توان نحوه نمایش درخت را به دلخواه تنظیم کرد.



شکل (۸-۱۹)



شکل (۸-۲۰) – درخت حاصل از طبقه بندی

توضیح درخت در شکل (۸-۲۰) را در نظر بگیرید، نحوه ی تفسیر برای گره ها به صورت زیر می باشد :

۱) گره های هم عمق:

گره یا گره هایی را انتخاب می کنیم که دارای بیشترین مقدار عددی باشند این عدد در قسمت پایین شکل گرافیکی درخت تصمیم نوشته شده است. اگر سه گره هم عمق با مقادیر ۳۰ و ۴۰ و ۱۰ وجود داشته باشد، گره با مقدار ۱۰ حذف می شود.

۲) پایش عمقی گره ها:

گره یا گره هایی در نظر گرفته می شوند که دارای عمق بیشتری نسبت به بقیه باشند، اگر دو گره با عمق ۶ و ۲ وجد داشته باشد گره با عمق ۶ تفسیر می شود. اما اگر اختلاف زیاد نباشد، مثل ۵ و ۶ یا ۷ و ۹ ، هر دو تفسیر می شوند.

به عنوان مثال راست ترین گره NUMKIDS یعنی (۵۶.۰ / ۱۰۷.۰) به این معنا است که INCOME بیشتر از ۲۵۰۴۹ بوده و تعداد فرزندان بیشتر از ۳ نفر می باشد. در واقع به صورت دقیق تر ۵۶ نفر از مجموع ۱۰۷ نفر دارای سود بدی هستند. توجه داشته باشید که تمام این اعداد به مجموعه آموزشی مربوط است، یعنی ۷۵ درصد از کل داده ها. حال به آزمایش مدل با استفاده از مجموعه آزمایشی پرداخته می شود. برای این کار در قسمت Test options، Supplied test set را انتخاب کنید، در ادامه روی start کلیک کنید تا آزمایش انجام شود. اعداد بدست آمده از روی مجموعه آموزشی و آزمایشی را مقایسه نمایید تا کارایی مدل ارزیابی شود.

توجه داشته باشید که موضوع همیشه به همین سادگی نمی باشد. ممکن است برای مجموعه داده های دیگر ، اعداد بسیار پایین تری بدست آورید. این نکته صلاحیت کلی طبقه بندی را زیر سوال نمی برد، بلکه نشان می دهد، طبقه بندی روش مناسبی برای کاوش این مجموعه داده نیست.

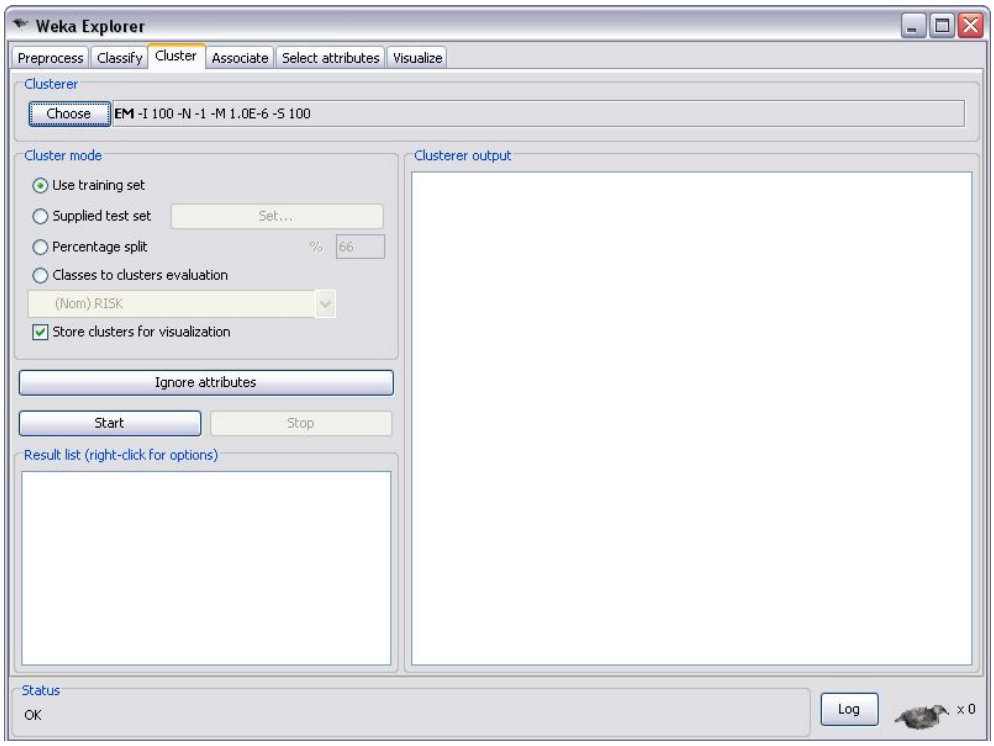
روش های مختلف داده کاوی ممکن است تفاوت های ویژه ای در عملکرد خود روی مجموعه داده مختلف ارائه دهند.

۸-۵-۳ خوشه بندی

خوشه بندی، دسته بندی مجموعه ای از داده ها به گروه ها یا خوشه های مختلف است که در آن خوشه ها به نحوی با یکدیگر شباهت دارند. با دسته بندی یک مجموعه داده به گروه ها، می توان به سرعت نتیجه گیری های لازم را بدست آورد. در الگوریتم های خوشه بندی، دانستن تعداد خوشه ها نیز با اهمیت است، که به دانش پیش نیازی از مدل و داده ها نیاز دارد. در صورتی که ماهیت مسئله به روشنی معلوم نباشد، تعیین این که چند خوشه لازم است می تواند کاری بسیار مشکل باشد. این موضوع یکی از مشکلات روش های خوشه بندی است، در صورتی که این روش ها سعی می کنند تا از تمام ویژگی های داده ها استفاده کنند (بر خلاف روش طبقه بندی که بخشی از ویژگی ها را در نظر می گرفت).

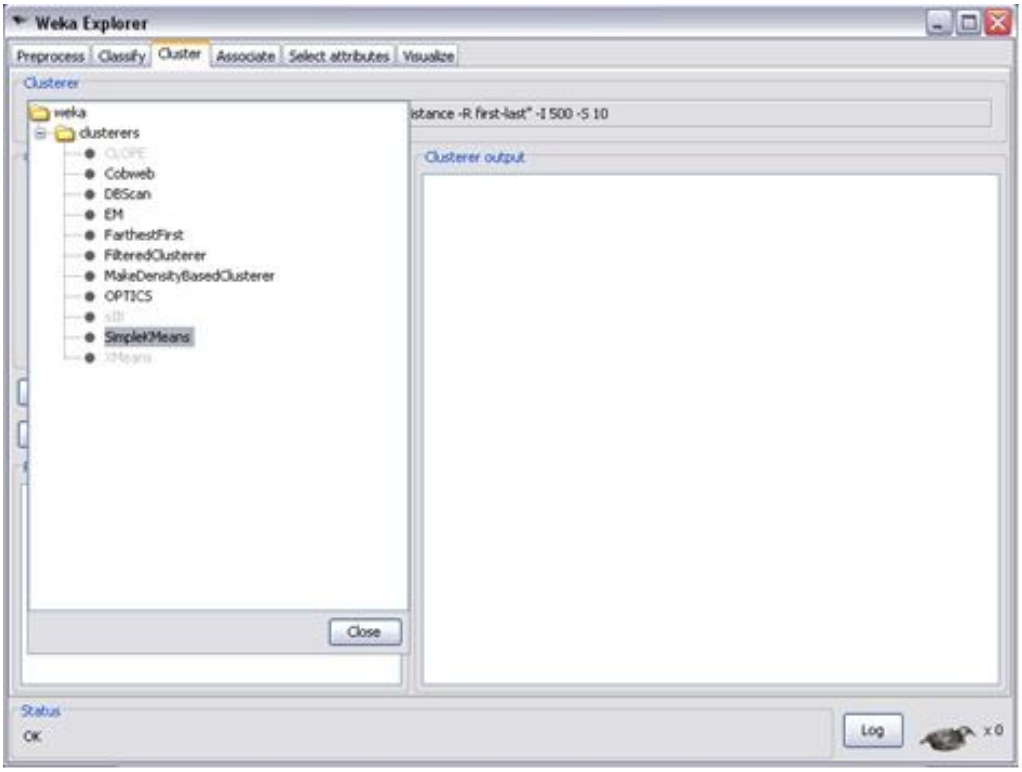
روش های خوشه بندی به دو دسته سلسله مراتبی (Hierarchical) و قسمت بندی (Partitioning) دسته بندی می شوند.در روش های سلسله مراتبی خوشه ها به مرور ساخته می شوند و خوشه بندی نهایی پس از طی مراحلی شکل می گیرد. روش های قسمت بندی در مقابل سعی می کنند تمام خوشه ها را به یکباره مشخص کنند. WEKA الگوریتم های خوشه بندی متنوعی را پیاده سازی می کند، در اینجا الگوریتم K-Means انتخاب می شود.

فایل نمونه مورد استفاده در این قسمت Dataset ۱ است که در مرحله preprocess فیلد id را از آن حذف کنید. بعد از باز کردن فایل مورد نظر و با کلیک بر روی پانل cluster پنجره‌ای مطابق شکل(۸-۲۱) باز می شود.

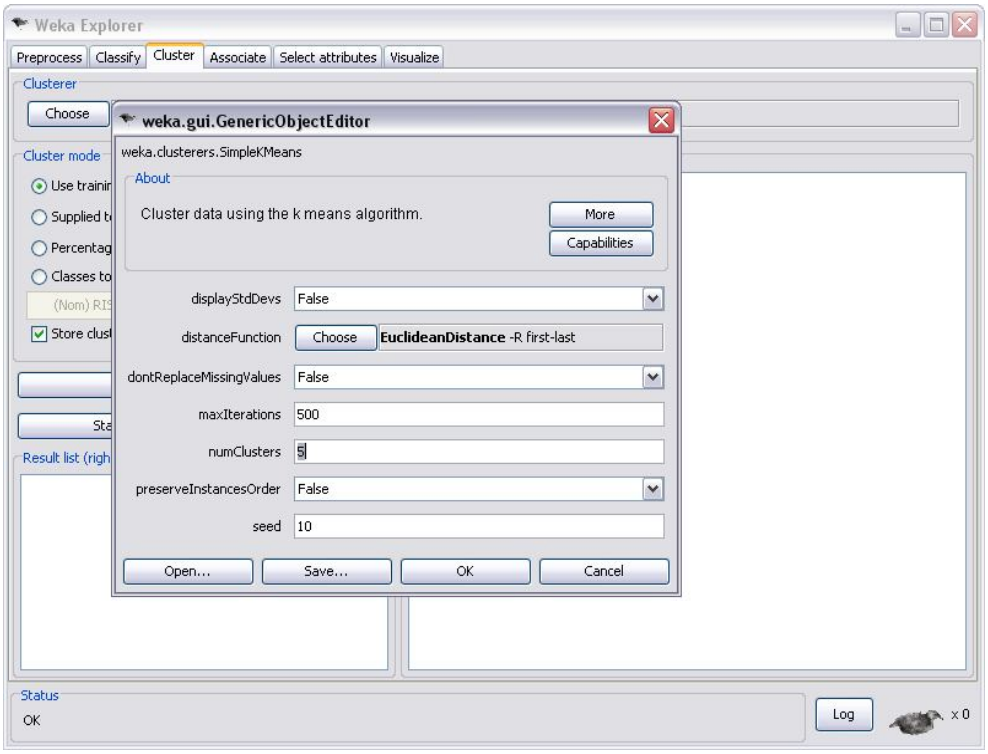


شکل (۸-۲۱) – انتخاب پانل cluster

با کلیک بر روی دکمه choose در پانل cluster می‌توان الگوریتم خوشه بندی مورد نظر را انتخاب نمود، شکل (۸-۲۲). الگوریتم Simple K-Means را بر می گزینیم. زمانی که یک الگوریتم خوشه بندی انتخاب می شود، نسخه خط فرمانی الگوریتم خوشه بندی در سطری نزدیک به دکمه ظاهر می‌گردد. این خط فرمان شامل پارامترهای الگوریتم است که با خط تیره مشخص می شوند. برای تغییر آن‌ها می توان روی آن خط کلیک نمود تا ویرایشگر مناسب شیء باز شود، شکل (۸-۲۳). در این مثال تعداد خوشه ۵ تنظیم می شود.

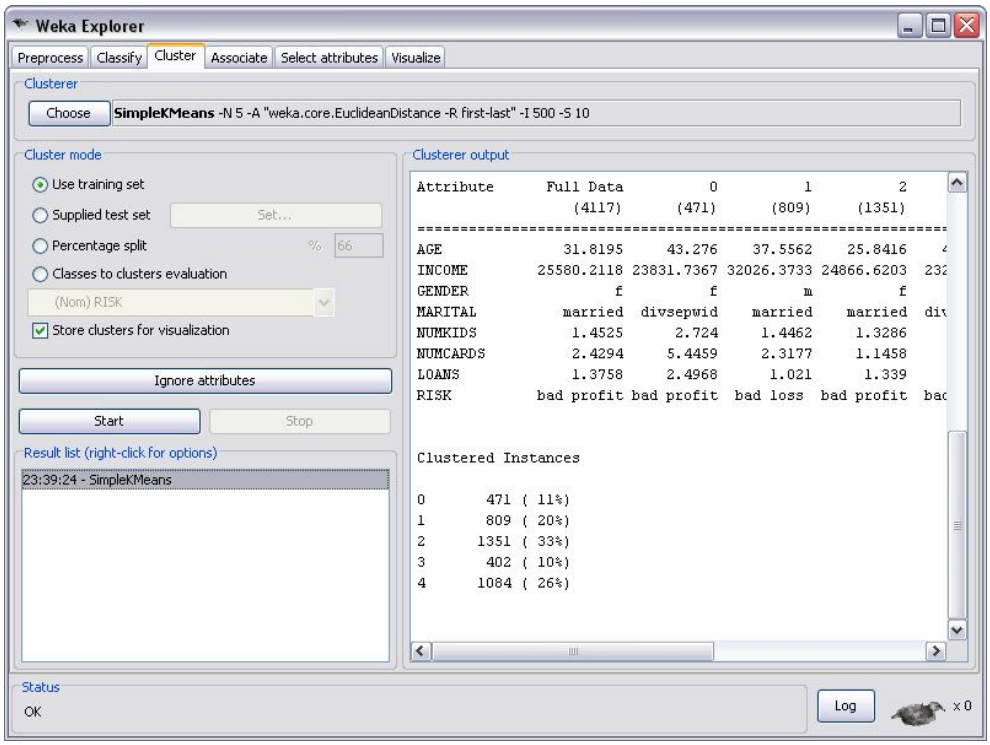


شکل (۸-۲۲) - انتخاب الگوریتم خوشه بندی



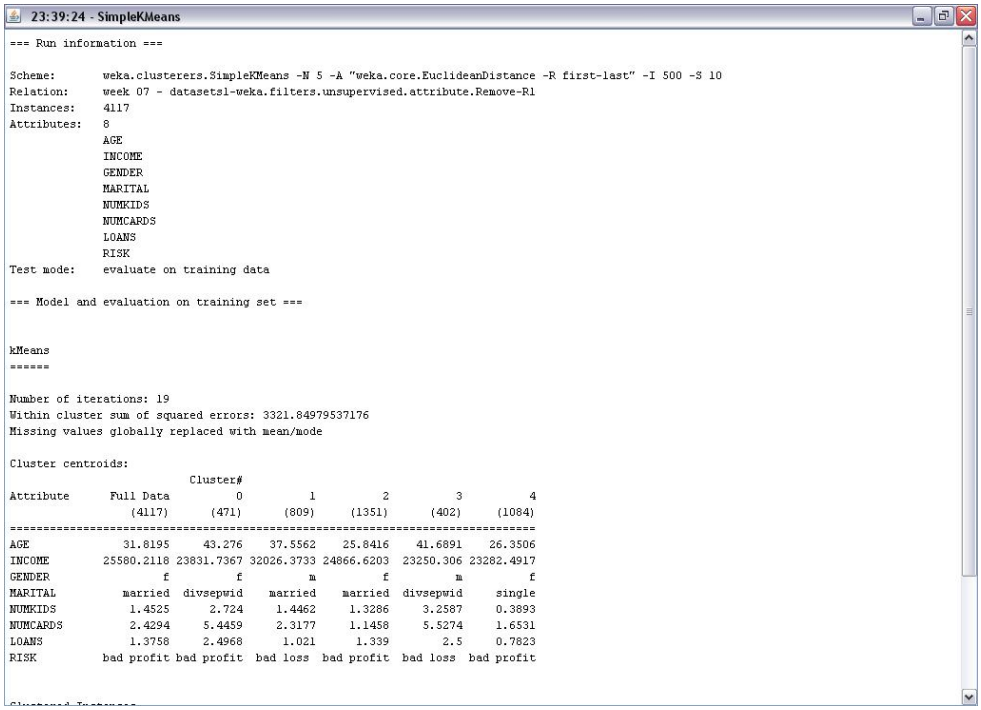
شکل (۸-۲۳) – تنظیم پارامترهای الگوریتم خوشه بندی

باکلیک بر روی دکمه start مدل مورد نظر تولید می شود، شکل (۸-۲۴).



شکل (۸-۲۴) - مدل حاصل از اجرای الگوریتم خوشه بندی

با راست کلیک بر روی مجموعه جواب در صفحه Result list در سمت چپ-پایین و انتخاب گزینه‌ی «View in separate window» می توان نتیجه را در پنجره‌ای جداگانه مشاهده نمود، شکل (۸-۲۵). همان طور که مشاهده می کنید، اطلاعات آماری مربوط به هرکلاستر، از جمله مرکز ثقل هر خوشه، تعداد و درصد اعضای هر خوشه در این پنجره قابل مشاهده است.

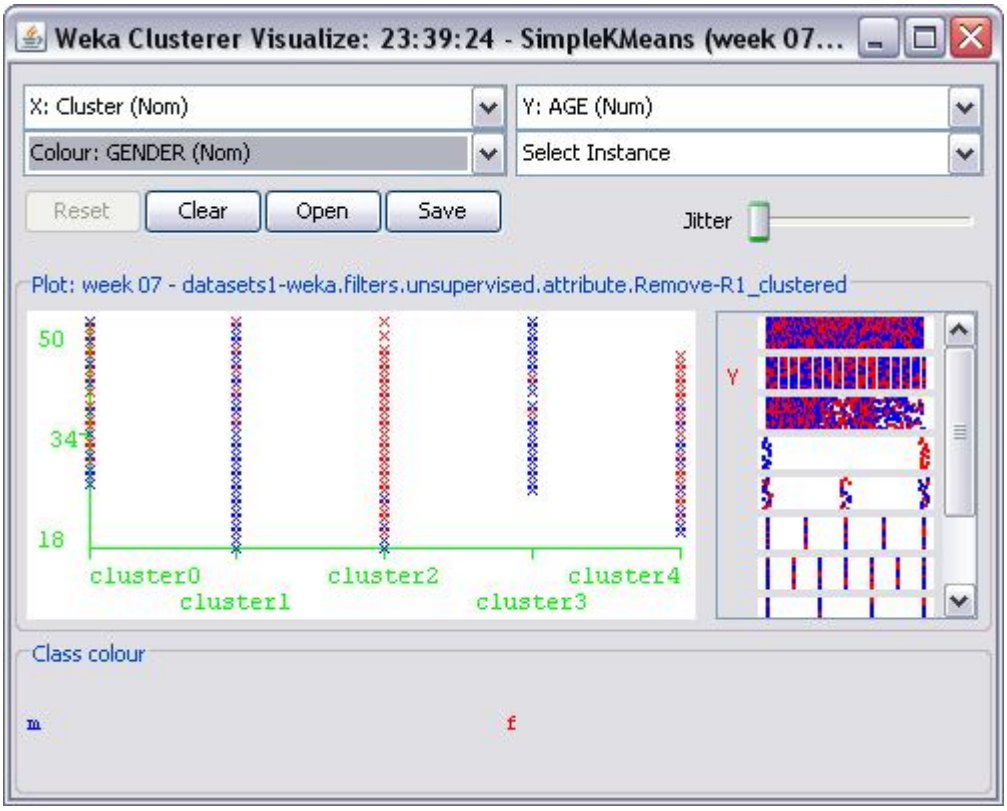


23:39:24 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     week 07 - datasets1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    4117
Attributes:   8
              AGE
              INCOME
              GENDER
              MARITAL
              NUMKIDS
              NUMCARDS
              LOANS
              RISK
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 19
Within cluster sum of squared errors: 3321.84979537176
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute      Full Data          0          1          2          3          4
                  (4117)      (471)      (809)     (1351)      (402)     (1084)
================================================================================
AGE               31.8195     43.276    37.5562    25.8416    41.6891    26.3506
INCOME         25580.2118 23831.7367 32026.3733 24866.6203  23250.306 23282.4917
GENDER                  f          f          m          f          m          f
MARITAL           married  divsepwid    married    married  divsepwid     single
NUMKIDS            1.4525      2.724     1.4462     1.3286     3.2587     0.3893
NUMCARDS           2.4294     5.4459     2.3177     1.1458     5.5274     1.6531
LOANS              1.3758     2.4968      1.021      1.339        2.5     0.7823
RISK          bad profit bad profit  bad loss  bad profit  bad loss  bad profit
```

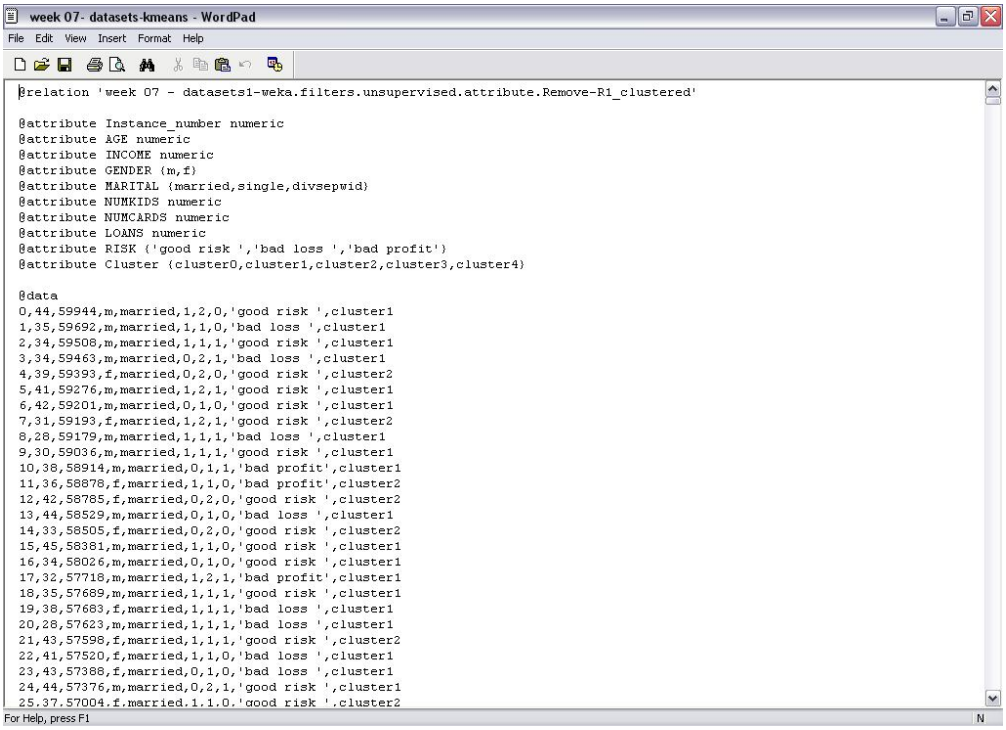
شکل (۸-۲۵) – نتیجه حاصل از اجرای الگوریتم خوشه بندی

روش دیگر برای کسب اطلاعات در مورد هرکلاستر، مصورسازی است. با راست کلیک بر روی مجموعه جواب در پانل Result list در سمت چپ و انتخاب گزینه‌ی « Visualize cluster assignments» پنجره‌ای باز می‌شود، که انتخاب های مختلف برای هرکدام از سه بعد نمودارحاصل (محور X، محور Y، رنگ) نمودارهای مختلفی را نتیجه می‌دهد که می‌توان از آن‌ها اطلاعات مورد نظر را بدست آورد. به عنوان مثال در شکل (۸-۲۶)، محور Xها نماینده شماره خوشه ، محور Yها نماینده شماره نمونه در بانک اطلاعاتی، و رنگ ها نماینده جنسیت هستند (قرمز: زن، آبی: مرد). همانطورکه مشاهده می شود خوشه ۲ بیشتر توسط زنان احاطه شده است و خوشه ۱ توسط مردان. به این صورت که میزان اعتبار متقاضیان وام برای زن ها بدتر از مردان بوده است.



شکل (۸-۲۶) – مصور سازی نتیجه حاصل از خوشه بندی

علاوه بر این ممکن است علاقه مند باشید که بدانید هر نمونه در بانک اطلاعاتی، به کدام خوشه اختصاص داده شده است. برای این منظور در پنجره شکل (۸-۲۶)، گزینه save را انتخاب کرده و فایل مورد نظر را با نام « Datasets ۱-Kmeans» ذخیره کنید. فایل حاصل را می توان از طریق یک نرم افزار پردازش متن مثل Notepad یا word باز کرد. بخش ابتدایی این فایل در شکل (۸-۲۷) نشان داده شده است. همان طورکه مشـاهده می کنید، WEKA ویژگی جدیدی به نام Cluster را به مجموعه ویژگی های موجود اضافه کرده است.



شکل (۸-۲۷) – نتیجه اختصاص نمونه ها به خوشه

۸-۵-۴ قوانین وابستگی

نرم افزار WEKA می تواند الگوریتم های قوانین وابستگی متفاوت حاکم بر بانک اطلاعاتی را پیاده سازی کند.این قوانین در فصل ۴ به طور کامل مورد بررسی قرار گرفته است ولی جهت یادآوری قوانین وابستگی، الگوهای پنهان میان اقلام موجود در پایگاه داده های بزرگ را شناسایی می کند. یک قانون وابستگی، یک قانون به فرم $X \rightarrow Y$ است که X و Y مجموعه ای از آیتم ها هستند که نباید با یکدیگر اشتراک داشته باشند. معنی این قانون این است که حضور X در یک تراکنش دلالت بر حضور Y در همان تراکنش را دارد. در قانون وابستگی اطمینان یک قانون میزان همبستگی بین مجموعه اقلام را اندازه گیری می کند، در حالی که پشتیبانی یک قانون، اهمیت بستگی بین مجموعه اقلام را اندازه گیری می کند. در واقع باید به دنبال قوانینی بود که مینیمم پشتیبانی و اطمینان تعریف شده به وسیله کاربر را برآورده کنند.

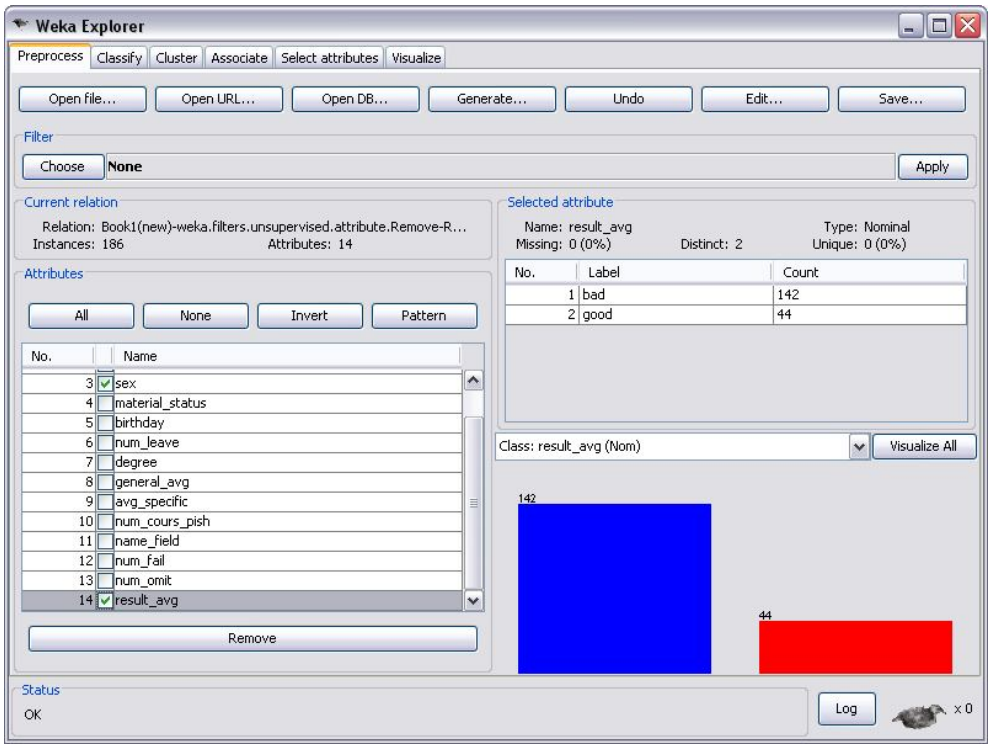
برای اینکار الگوریتم های مختلفی وجو دارد که از معروفتری آنها الگوریتم Apriori می باشد. جهت درک بهتر، از بانک اطلاعاتی ۲ Dataset به عنوان مثال استفاده می شود، که شامل اطلاعاتی در مورد ۲۰۰ دانشجو در مقاطع مختلف تحصیلی است. جدول (۸-۴)، مجموعه داده ها را نشان می دهد.

جدول (۸-۴) – فیلدهای بانک اطلاعاتی

(Sex) جنسیت	مونث، مذکر
(Marital) وضعیت تاهل	مجرد، متاهل
(Birthday) تولد	سال تولد دانشجویان (numeric)
(NumConditional) مشروطی	تعداد مشروطی ها (numeric)
(Degree) مقطع تحصیلی	مقاطع (MA, BA,PHD)
(General-Avg) معدل دروس عمومی	میانگین معدل دروس عمومی (numeric)
(Avg-Specific) معدل دروس اختصاصی	میانگین معدل دروس اختصاصی (numeric)
(Num-cours-pish) دروس پیش نیاز	تعداد دروس پیش نیاز (numeric)
(Field) رشته تحصیلی	رشته های تحصیلی (فنی، علوم انسانی و تجربی)
(Num-omit) حذف	تعداد دروس حذفی دانشجویان در طول کل ترم ها (numeric)
(Num-fail) درس افتاده	تعداد دروسی که نگذراندن (numeric)
(Result-Avg) نتیجه کل معدل	بد، خوب

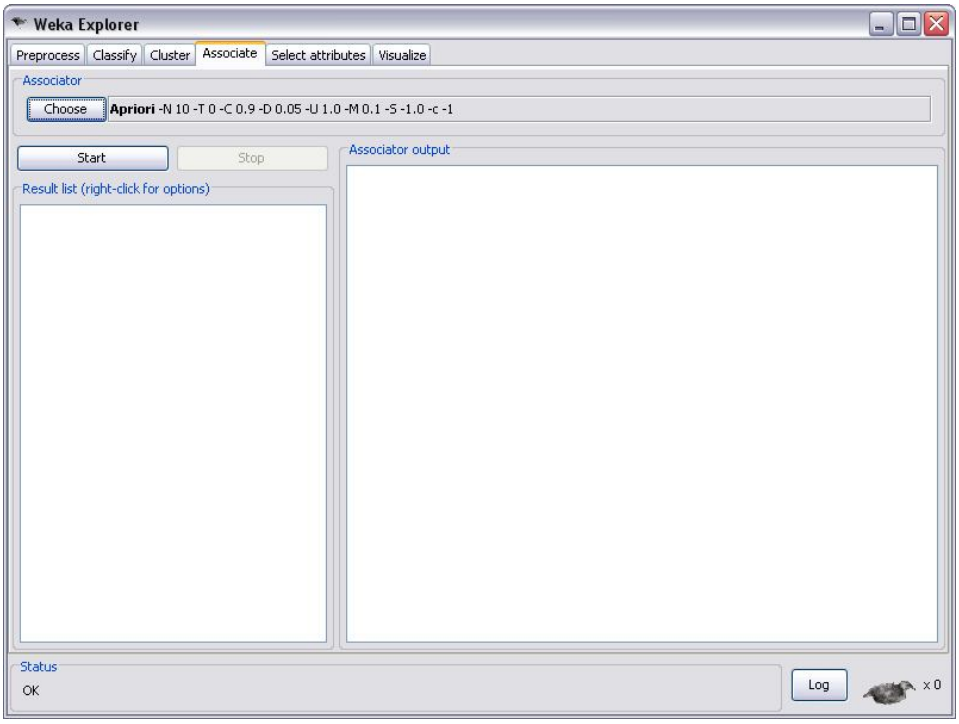
با توجه به فیلد هدف (خوب، بد)، افت تحصیلی دانشجویان مورد بررسی قرار گرفته است. با پیدا کردن الگوهای نهفته در این اطلاعات می توان برنامه ریزی مناسبی جهت بهبود وضعیت تحصیلی دانشجویان ارائه داد.

در قسمت preprocess فیلد ID از آن حذف می شود و دو صفت SEX و RESULT-AVG انتخاب و مقادیر آن ها به صورت گسسته در آمده است، شکل (۸-۲۸).



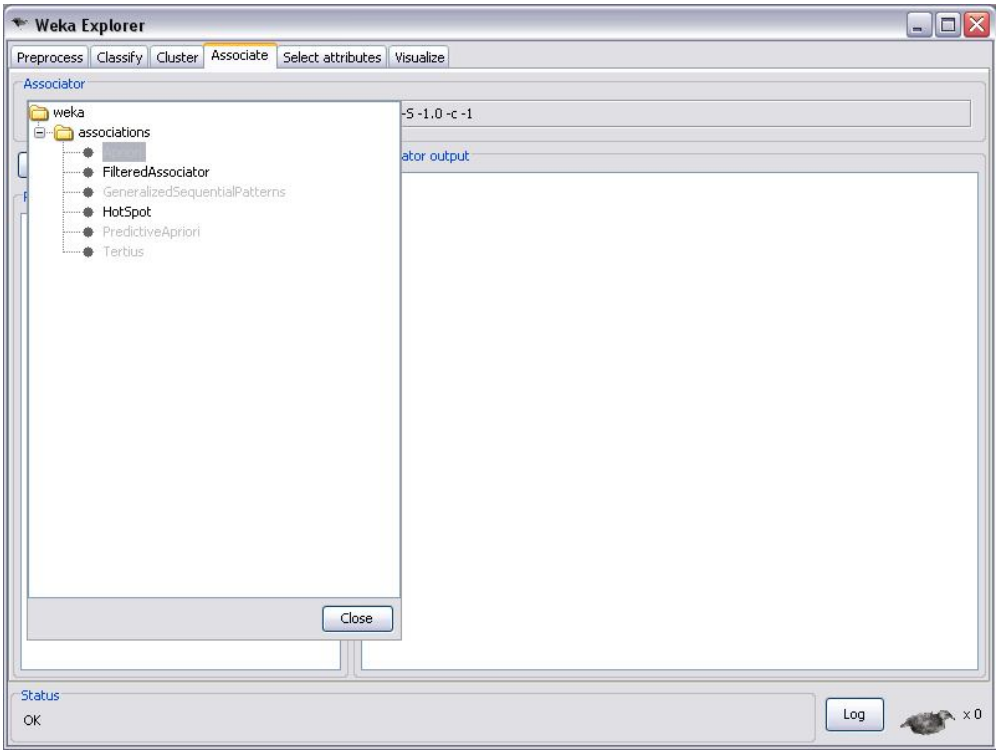
شکل (۸-۲۸) – بارگذاری فایل

با باز کردن فایل مورد نظر، بر روی پانل Associate کلیک کرده، پنجره ای مطابق با شکل (۸-۲۹) باز می شود.

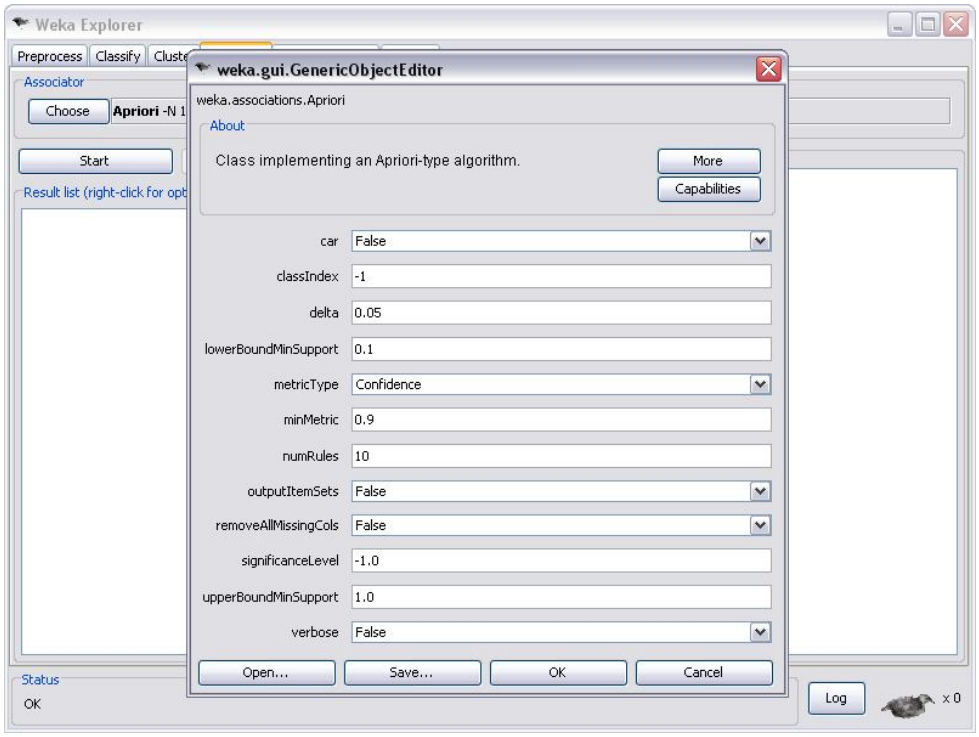


شکل (۸-۲۹) – انتخاب پانل Associate

با کلیک کردن بر گزینه ی choose در پانل Associate، می توان الگوریتم مورد نظر را انتخاب نمود، شکل (۸-۳۰). در اینجا الگوریتم Apriori انتخاب می شود. برای تغییر پارامترهای الگوریتم بر روی خط فرمان نزدیک گزینه ی choose کلیک کرده تا ویرایشگر مناسب شیء باز شود، شکل (۸-۳۱). در این بخش همان پارامترهای پیش فرض در نظر گرفته می شود.

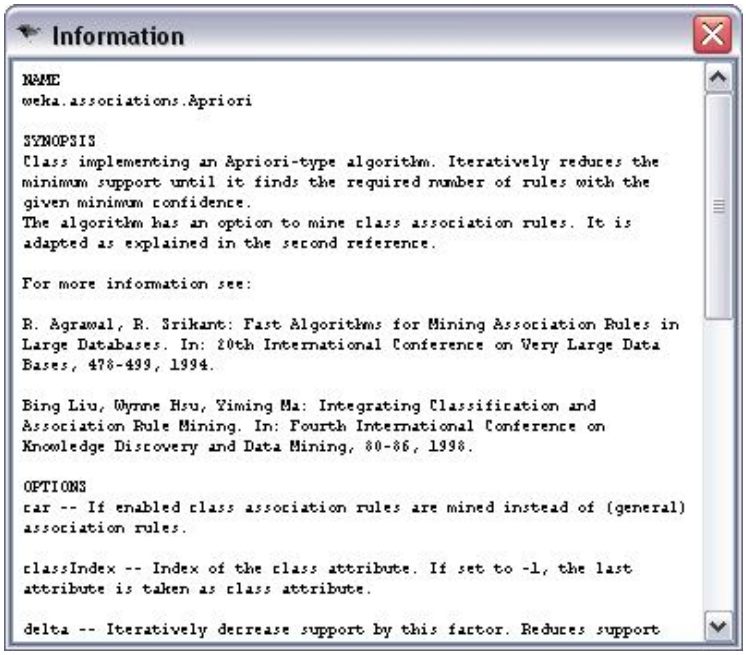


شکل (۸-۳۰) – انتخاب الگوریتم Apriori



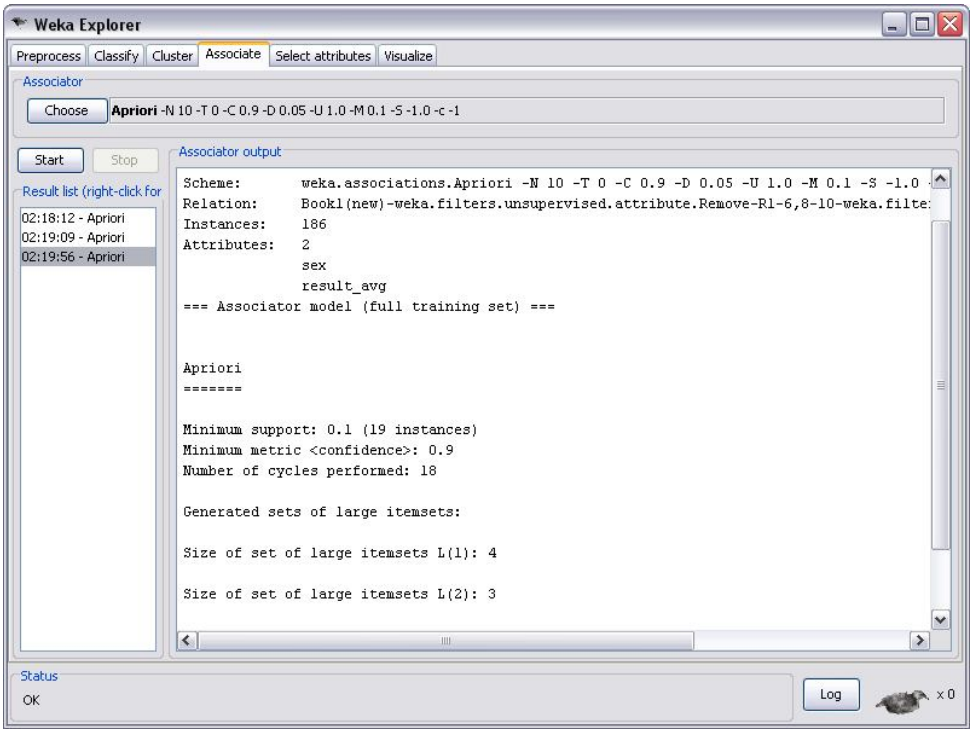
شکل (۸-۳۱) – تنظیم پارامترهای الگوریتم Association

با کلیک کردن بر روی گزینه More در شکل (۸-۳۱)، می توان توضیحات لازم را در مورد هر کدام از پارامترها بدست آورد، شکل (۸-۳۲).



شکل (۸-۳۱) – توضیحات بیشتر در مورد پارامترهای الگوریتم Apriori

در نهایت با انتخاب الگوریتم Apriori بر روی گزینه ی start کلیک کرده و نتایج حاصل از این الگوریتم در قالب ضریب همبستگی (conf) و L_1 و L_2 نمایش داده می شود، شکل (۸-۳۲). نتایج نشان می دهد که احتمال اینکه جنسیت مذکر وضعیت تحصیلی بدی را داشته باشد ۰.۹ می باشد.



شکل (۸-۳۲) - نتیجه اجرای الگوریتم Apriori

البته باید توجه نمود که قواعد انجمنی ممکن است در هر نوع مجموعه داده ای قابل اعمال نبوده و یا نتایج مطلوبی ارائه ننماید انتخاب نوع الگوریتمی که نتایج مطلوبی داشته باشد، نیاز به تجربه بیشتری دارد.