

به نام خدا

راهنمای استفاده از نرم افزار WEKA

ارائه دهندگان:

میلاذ قهاری

وحید رحمانی فرد

دانشگاه آزاد اسلامی واحد تهران جنوب

پاییز 1389

3	مقدمه
3	استفاده‌ها: خطر فرمانسیست معامل
3	استفاده‌ها: خطر فرمان WEKA
4	استفاده‌ها: واسط کاربری WEKA
4	استفاده‌ها: WEKA در برنامه‌های دیگر
4	پنجره اصلی WEKA
6	قسمت‌های اصلی WEKA
7	فرمت اطلاعات ورودی در WEKA
10	Explorer
14	برگه Classify
15	مثال‌ها: Classifier
18	نمایش درخت تصمیم
22	برگه Cluster
23	برگه Associate
24	برگه Visualize

مقدمه

نرم افزار WEKA مجموعه‌ای از به روزترین الگوریتم‌های یادگیری ماشینی و ابزارهایی برای پیش‌پردازش داده‌ها می‌باشد. با توجه به اینکه کلیه امکانات WEKA در قالب واسط‌های کاربری مناسب در اختیار کاربران قرار می‌گیرد بنابراین کاربران می‌توانند متدهای مختلف را بر روی داده‌های خود پیاده‌سازی کرده و بهترین الگوریتم را برای کار انتخاب نمایند.

این نرم‌افزار در دانشگاه Waikato نیوزلند ایجاد شده است و نام آن از حروف اول کلمات **Waikato Environment for Knowledge Analysis** می‌باشد این نرم‌افزار به زبان برنامه‌نویسی Java نوشته شده است و می‌توان آن را بر روی پلت‌فرم‌های متفاوتی که ماشین مجازی Java بر روی آنها نصب شده است اجرا نمود. همچنین این نرم‌افزار تحت مجوز GNU GPL انتشار یافته است و این بدان معناست که استفاده از آن رایگان بوده و کاربران به راحتی می‌توانند به کدمنبع‌های آن دسترسی داشته و حتی آنها را بر حسب نیاز تغییر داده و روش‌های دیگری را نیز به آنها اضافه کنند.

برای استفاده از WEKA از روش‌های متفاوتی می‌توان استفاده کرد. در ادامه این روش‌ها را مورد بررسی قرار می‌دهیم:



استفاده از خط فرمان سیستم عامل

پس از نصب WEKA می‌توانید با استفاده از ماشین مجازی جاوا و کنسولی که توسط اکثر سیستم‌عامل‌ها ارائه می‌شود از امکانات WEKA استفاده کنید. اگر قبلاً با Java برنامه نویسی کرده باشید قطعاً می‌دانید که فراخوانی کلاس‌های Java از خط فرمان به سادگی امکان‌پذیر است.

تنها مشکل اساسی این روش، تایپ زیاد در آن است و همچنین با توجه به اینکه Java به حروف حساس است بنابراین باید در به کار بردن نام‌ها حداکثر دقت را داشته باشید. البته به علت اینکه در برنامه‌نویسی WEKA از روش‌های نام‌گذاری استاندارد استفاده شده است در این مورد با مشکل کمتری مواجه خواهید بود. همچنین باید با کلیه روش‌هایی که قصد دارید از آنها استفاده کنید آشنا باشید و همچنین نحوه ست کردن پارامترهای آنها را نیز به صورت دستی انجام دهید. این روش کمی خسته کننده و وقت‌گیر است و جز در موارد اضطرار کاربران از آن استفاده نمی‌کنند.

استفاده از خط فرمان WEKA

با توجه به اینکه WEKA با استفاده از نرم‌افزار برنامه‌نویسی Java نوشته شده است و این زبان برنامه‌نویسی یک زبان Cross-platform است، بنابراین می‌توان WEKA را بر روی کلیه ماشین‌هایی که JRE بر روی آنها نصب است اجرا نمود. حال ممکن است در این سیستم‌ها خبری از کنسول و یا همان خط فرمان نباشد. از این رو تیم توسعه WEKA برای حل این مشکل یک خط فرمان داخلی برای WEKA قرار داده است. این خط فرمان قادر به اجرای دستورات خاصی است که در جای مناسبی مورد بررسی قرار می‌گیرد.



استفاده از واسط کاربری WEKA

یکی از مهمترین ویژگی‌های WEKA واسط کاربری آن است. مانند سایر نرم‌افزارها، استفاده از یک واسط کاربری می‌تواند کاربر را در انجام فعالیت‌هایش بسیار کمک کند. WEKA دارای 2 نوع واسط کاربری است. این واسط‌ها نیز در جای مناسب مورد بررسی قرار می‌گیرند.

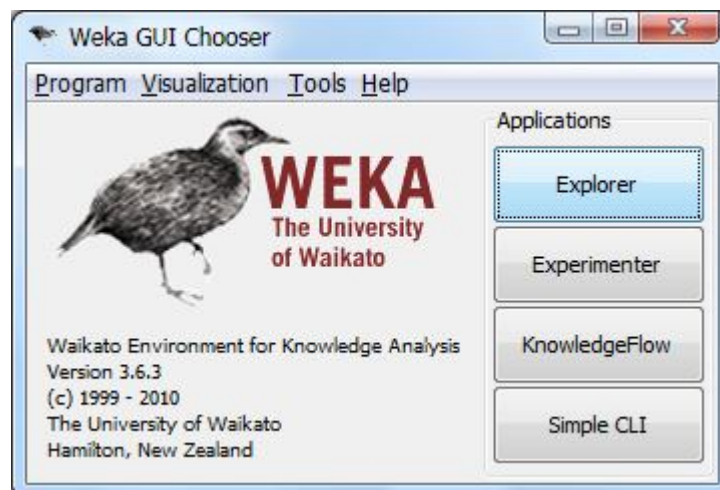


استفاده از WEKA در برنامه‌های دیگر

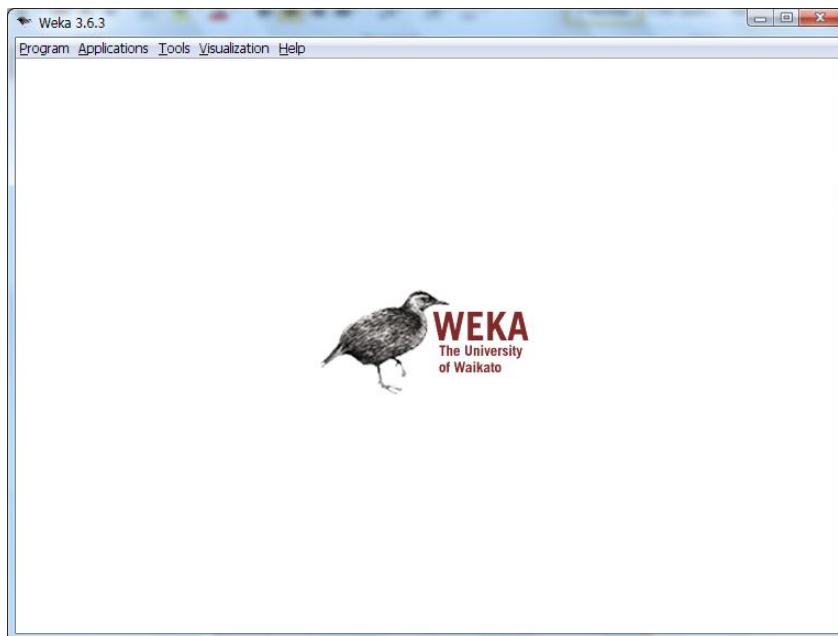
با توجه به اینکه امروزه روند استفاده از ابزارهای داده‌کاوی در نرم‌افزارها رو به رشد است و بسیاری از سازمان‌ها از تیم‌های برنامه‌نویسی می‌خواهند چنین امکاناتی را در برنامه‌ها قرار دهند، می‌توان به راحتی از امکانات WEKA در سایر پروژه‌ها نیز استفاده کرد. اگر از زبان برنامه‌نویسی Java استفاده می‌کنید این کار به سادگی هر چه تمام‌تر خواهد بود. اما اگر از زبان‌هایی مانند C# و یا VB.NET استفاده می‌کنید ابتدا باید باید کاری کرد که بتوان کدهای جاوا را در این برنامه‌ها اجرا کرد. برای حل این مشکل نیز کتابخانه‌ای ارائه شده است.

پنجره اصلی WEKA

در ادامه این قسمت از واسط گرافیکی WEKA استفاده خواهیم کرد و در انتها نیز نگاهی به قسمت خط فرمان این نرم‌افزار خواهیم داشت. همانطور که قبلاً گفته شد WEKA دارای 2 مدل واسط کاربری است. در تصویرهای زیر می‌توانید این مدل‌ها را مشاهده کنید.



تصویر 1. پنجره WEKA در حالت عادی.



تصویر 2. پنجره WEKA در حالت والد و فرزندی

توجه داشته باشید که در این دو حالت فقط نحوه نمایش قسمت اصلی WEKA فرق می‌کند و هیچ اثری در امکانات WEKA نخواهد داشت.

برای تغییر پنجره اصلی WEKA کافیسیت در جایی که این نرم‌افزار را نصب کرده اید فایلی به نام RunWeka.ini را با استفاده از ویرایشگر ساده باز کنید و متن زیر را پیدا کنید:

```
# The MDI GUI
```

```
mainclass=weka.gui.Main
```

```
# TheGUIChooser
```

```
#mainclass=weka.gui.GUIChooser
```

علامت # در ابتدای هر خط بیانگر توضیحات است. در حال حاضر مدلی که در تصویر 2 مشاهده می‌کنید فعال است. چراکه خط 2 که از حالت comment خارج شده است بیانگر این موضع می‌باشد. در صورتی که ترجیح می‌دهید حالت اول را به عنوان حالت پیش‌فرض برای کار انتخاب کنید کافیسیت خط 2 را کامنت کرده و خط 4 را از کامنت خارج کنید.

در ادامه ما از مدل ارائه شده در تصویر 1 استفاده می‌کنیم.

قسمت‌های اصلی WEKA

در پنجره اصلی WEKA امکانات متعددی قرار داده شده است. اولین موردی که کاربران را متوجه خود می‌کند دکمه‌های کنار صفحه است. در واقع با استفاده از این دکمه‌ها می‌توان به اصلی‌ترین بخش نرم‌افزار دسترسی پیدا کرد. در ادامه توضیح مختصری بر روی این گزینه‌ها می‌دهیم و بررسی کامل را به بعد موکول می‌کنیم.

1. **Explorer**: شاید مهمترین بخش نرم‌افزار WEKA مخصوصاً برای کاربرانی که تازه قصد دارند با WEKA کار کنند همین بخش باشد. این بخش امکانات متعددی برای Association Rule، Clustering، Classification و ... در اختیار شما قرار می‌دهد. یکی از مهمترین مشکلات این بخش به نحوه بارگذاری اطلاعات ورودی در حافظه مربوط می‌شود. این بخش هنگامی که یک Dataset را برای کار انتخاب می‌کنید، آن را به صورت کامل در حافظه Load می‌کند. بنابراین این مدل برای داده‌های بسیار حجیم چندان کارآیی ندارد.
2. **Experimenter**: معمولاً ابزارهای Explorer و KnowledgeFlow برای تعیین میزان کیفیت عملکرد یک مدل انتخاب شده بر روی داده‌ها را دارند. اما در عمل بسیار پیش می‌آید که باید چندین الگوریتم یادگیری را بر روی دیتاست‌های متفاوت اجرا کنیم. استفاده از ابزار Experimenter برای این کار بهترین گزینه است. همچنین با استفاده از Experimenter می‌توان انجام عملیات را در چند کامپیوتر انجام داد و مجدداً در زمان صرفه‌جویی کرد. برای انجام این کار بایستی از Java RMI استفاده شود. همچنین این ابزار به ما این امکان را می‌دهد که تا حدودی روند انجام کار را خودکار کنیم. به عبارت دیگر می‌توان Classifierهای متعددی را می‌توان با پارامترهای متفاوت بر روی داده‌ها ست کرده و خروجی آن‌ها را با هم مقایسه کرد.
3. **Knowledge Flow**: برخلاف Experimenter که هدف اصلی غلبه بر محدودیت زمانی بود، با استفاده از این ابزار می‌توان بر محدودیت فضایی غلبه کرد. در این ابزار نیازی نیست که کلیه دیتاست یکباره در حافظه بارگذاری شوند. یکی دیگر از اهداف این ابزار کار با داده‌های جریانی می‌باشد. توجه داشته باشید که در Explorer امکانی برای کار با داده‌های جریانی وجود ندارد.
4. **Simple CLI**: این قسمت حاوی یک کنسول ساده برای اجرای دستی الگوریتم‌های موجود در WEKA می‌باشد.

علاوه بر دکمه‌های کناری صفحه چندین منو در بالای واسط کاربری نیز وجود دارد. تعدادی از مهمترین ابزارهای موجود در این منوها عبارتند از:

1. **ArffViewer** → Tools: این منو امکاناتی را برای مشاهده و تغییر فایل‌های ورودی WEKA که در قالب فرمت ARFF می‌باشند فراهم می‌کند.
2. **SqlViewer** → Tools: با استفاده از این گزینه امکاناتی برای وصل شدن به بانک‌های اطلاعاتی که JDBC از آن‌ها پشتیبانی می‌کند قرار داده شده است. پس از اتصال می‌توانید به راحتی با استفاده از دستور SELECT اطلاعات را بازیابی کرده و بر روی آن‌ها کار مورد نظر خود را انجام دهید.
3. **Visualization**: در این منو امکاناتی برای کار با نمودارهای متفاوت در WEKA قرار داده شده است.

در ادامه به بررسی قسمت‌های اصلی نرم‌افزار با ارائه مثال‌هایی می‌پردازیم. اما قبل از آن نگاهی به فرمت ورود اطلاعات در WRKA خواهیم داشت.

فرمت اطلاعات ورودی در WEKA

فرمت پیش فرض WEKA فرمت ARFF¹ است. معمولاً در پوشه‌ی WEKA چندین فایل با این فرمت و حاوی اطلاعاتی برای آموزش وجود دارد. این فایل‌ها در پوشه Data قرار دارند. به عنوان مثال سعی کنید فایل Weather.arff را در یک ویرایشگر متنی ساده مانند Notepad++ باز کنید. با این کار اطلاعات زیر را مشاهده خواهید کرد:

```
@relation weather
```

```
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
```

```
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

لیست 1. اطلاعات فایل weather.arff

این فایل حاوی اطلاعاتی در مورد وضعیت آب و هوا و انجام شدن بازی گلف می‌باشد. خطوطی که با علامت @ شروع می‌شوند در حقیقت قرار است اطلاعاتی را برای WEKA فراهم کنند. اولین خط اطلاعات نامی را برای رابطه فوق تعیین کرده است. این خط چندان کار مهمی را انجام نمی‌دهد و فقط برای تعیین نام رابطه از آن استفاده می‌شود. در ادامه چندین خط اطلاعات با عبارت @attribute وجود دارند که نشان‌دهنده خصیصه‌های موجود در جدول اطلاعاتی می‌باشند. در جلوی نام هر خصیصه اطلاعاتی در مورد نوع آن نیز وجود دارد. اگر خصیصه‌ای از نوع اسمی «Nominal» باشد در جلوی اسم آن خصیصه کلیه مقادیر ممکن برای آن را در بین { } می‌نویسیم (به عنوان مثال خصیصه‌های outlook, windy و play در مثال فوق). در مورد سایر خصیصه-

های که دارای نوعی غیر از اسمی می باشند می توان از انواعی که برای WEKA شناخته شده است استفاده کرد. به عنوان مثال برای داده های عددی اعشاری از نوع real برای رشته ها از string و برای تاریخ و زمان از date استفاده می شود.

بعد از تعیین نوع خصیصه ها نوبت به اضافه کردن اطلاعات ورودی در فایل می شود. این قسمت با عبارت @data آغاز می شود. پس از آن باید در هر سطر اطلاعات مربوط به هر نمونه را به ترتیب خصیصه ها که در قسمت قبل تعریف شدند وارد نمود.

برای تعیین Missing value ها بایر از ؟ استفاده کرد.

گاهی اوقات ایجاد فایل های اطلاعاتی با این فرمت بسیار سخت و زمانگیر است. اگر بار دیگر به فرمت ورود اطلاعات دقت کنید می توان آن را مانند یک فایل صفحه گسترده در Excel در نظر گرفت. با توجه به این مورد می توان از این خصوصیت برای ایجاد آسان تر فایل های ورودی استفاده کرد. برای این کار کفایت ابتدا اطلاعات خود را درون یک فایل Excel وارد کنید. برای نمونه اطلاعات فایل Weather.arff را در فایل Excel زیر وارد کرده ایم:

	A	B	C	D	E
1	outlook	temperatu	humidity	windy	play
2	sunny	85	85	FALSE	no
3	sunny	80	90	TRUE	no
4	overcast	83	86	FALSE	yes
5	rainy	70	96	FALSE	yes
6	rainy	68	80	FALSE	yes
7	rainy	65	70	TRUE	no
8	overcast	64	65	TRUE	yes
9	sunny	72	95	FALSE	no
10	sunny	69	70	FALSE	yes
11	rainy	75	80	FALSE	yes
12	sunny	75	70	TRUE	yes
13	overcast	72	90	TRUE	yes
14	overcast	81	75	FALSE	yes
15	rainy	71	91	TRUE	no
16					

تصویر 3. ورود اطلاعات در Excel.

اکنون میبایست اطلاعات را با فرمت CSV ذخیره نمود. اگر این کار را انجام دهید و سپس فایل CSV مد نظر را در Noptpad++ باز کنید تصویر زیر را مشاهده خواهید کرد.

```

1 outlook,temperature,humidity,windy,play
2 sunny,85,85,FALSE,no
3 sunny,80,90,TRUE,no
4 overcast,83,86,FALSE,yes
5 rainy,70,96,FALSE,yes
6 rainy,68,80,FALSE,yes
7 rainy,65,70,TRUE,no
8 overcast,64,65,TRUE,yes
9 sunny,72,95,FALSE,no
10 sunny,69,70,FALSE,yes
11 rainy,75,80,FALSE,yes
12 sunny,75,70,TRUE,yes
13 overcast,72,90,TRUE,yes
14 overcast,81,75,FALSE,yes
15 rainy,71,91,TRUE,no
16
    
```

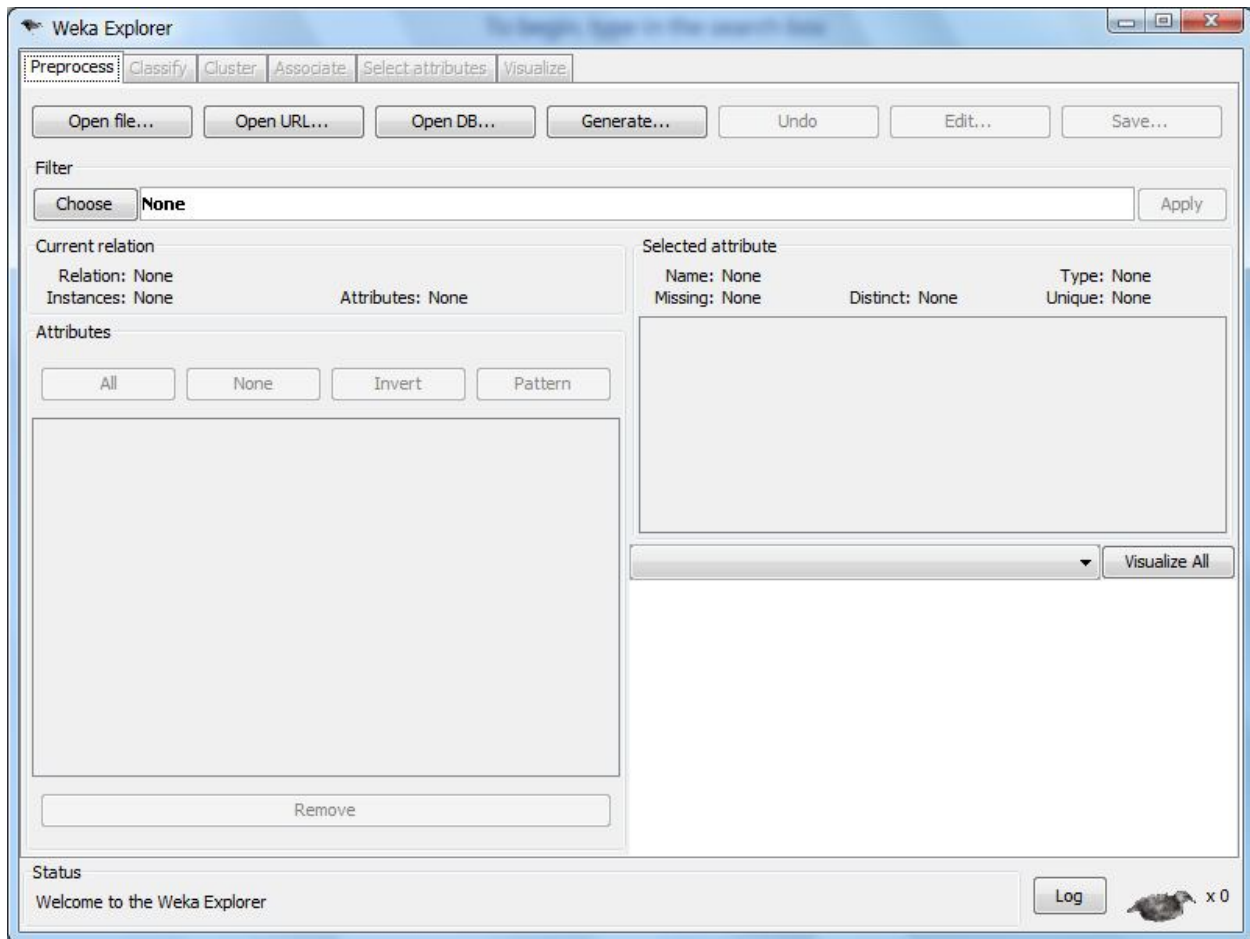
تصویر 4. اطلاعات فایل weather.csv در ویرایشگر متنی.

همانطور که در تصویر مشخص است این اطلاعات بسیار شبیه به فرمت arff می باشند. البته نرم افزار WEKA از فرمت CSV نیز پشتیبانی می کند و شما می توانید به راحتی فایل های CSV را در این نرم افزار باز کنید. نرم افزار با مشاهده این مدل فایل ها ابتدا آنها را به صورت ضمنی به فرمت arff تبدیل نموده و سپس به راحتی آنها را در نرم افزار بارگذاری می کند. بنابراین می توان به راحتی از نرم افزارهای صفحه گسترده مانند Excel برای ایجاد و مدیریت اطلاعات استفاده نمود.

Explorer

با استفاده از امکانات موجود در این صفحه می‌توانید به راحتی بر روی اطلاعات دسته‌بندی کرده، قوانین انجمنی استخراج نمایید، درخت‌های تصمیم ایجاد کنید و کلیه این اعمال را با چند کلیک ساده و در قالب فرم‌ها و منوهای انجام می‌دهید. در این قسمت به بررسی قسمت‌های مختلف این گزینه می‌پردازیم.

با کلیک بر روی دکمه مربوط به این گزینه در صفحه اصلی، پنجره‌ای مانند صفحه زیر باز می‌شود:



تصویر 5. پنجره Explorer.

همانطور که در تصویر فوق مشخص است کلیه قسمت‌های اصلی نرم‌افزار غیرفعال است. برای این که بتوانید از این امکانات استفاده کنید باید ابتدا یک دیتاست را باز کنید. برای بارگذاری اطلاعات در WEKA چندین روش مختلف وجود دارد. 4 دکمه موجود در قسمت Preprocess برای این کار در نظر گرفته شده است. این دکمه‌ها عبارتند از:

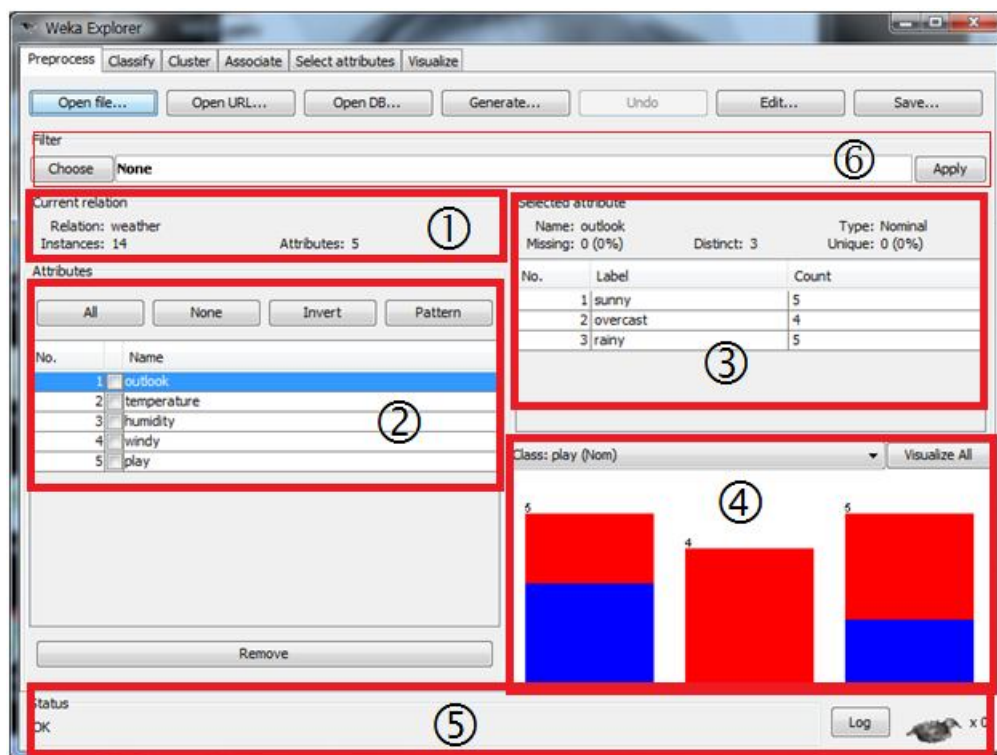
◀ **Open File**: با استفاده از این گزینه می‌توانید یک فایل موجود در کامپیوتر محلی را برای باز کردن انتخاب نمایید.

◀ **Open URL**: با استفاده از این گزینه می‌توانید آدرس ذخیره‌سازی فایل را در یک سرور راه دور وارد کنید و از آن فایل برای انجام مدل‌سازی استفاده نمایید.

◀ **Open DB**: از این گزینه برای اتصال به یک بانک اطلاعاتی و استفاده از اطلاعات موجود در آن استفاده می‌شود.

◀ **Generate**: با کلیک بر روی این دکمه پنجره‌ای باز می‌شود و شما به راحتی می‌توانید الگوریتم‌های متفاوتی که برای ایجاد داده‌ها به صورت خودکار برای دسته‌بندی‌های متفاوت وجود دارد، استفاده کنید.

پس از باز کردن یک فایل حاوی داده در صفحه Preprocess اطلاعات کلی پیرامون داده‌های بارگذاری شده نمایش داده می‌شود. فرض کنید فایل weather.arff را در برنامه بارگذاری کرده‌ایم. با این کار صفحه زیر نمایان می‌شود:



تصویر 6. صفحه Preprocess پس از بارگذاری فایل Weather.arff

قسمت‌های مهم این صفحه پس از بارگذاری اطلاعات عبارتند از:


◀ **قسمت ①**: در این بخش اطلاعات کلی از فایل باز شده ارائه می‌شود. اطلاعاتی همچون نام رابطه، تعداد نمونه‌های موجود در فایل و تعداد خصیصه‌های ارائه شده برای کلیه نمونه‌ها.


◀ **قسمت ②**: در این قسمت نام کلیه خصیصه‌های موجود در فایل را مشاهده می‌کنید. در کنار هر خصیصه یک CheckBox وجود دارد. با استفاده از این کنترل می‌توانید خصیصه‌های مدنظر را انتخاب کرده و آنهایی را که نیاز ندارید حذف نمایید. البته برای انتخاب خصوصیت‌ها در بالای این قسمت چندین دکمه وجود دارد که کارکرد آنها را می‌توان از

نام آنها تشخیص داد. ممکن است گزینه Pattern اندکی ناآشنا باشد. با استفاده از این گزینه می‌توانید خصیصه‌ها را بر اساس عبارات باقاعده که در Perl 5 مورد استفاده قرار می‌گیرد انتخاب نمایید.. در انتها با استفاده از دکمه Remove که در پایین این قسمت قرار دارد می‌توانید خصیصه‌هایی که در ایجاد مدل کمتر اهمیت دارند و آنها را انتخاب کرده‌اید حذف کنید.


◀ **قسمت ③:** با انتخاب هریک از خصیصه‌های موجود در کادر قسمت 2 در این قسمت اطلاعاتی کلی در مورد آن خصیصه نشان داده می‌شود. به عنوان مثال در تصویر فوق خصیصه outlook انتخاب شده است که از نوع nominal است و مقادیر ممکنه آن 3 مورد (sunny, overcast, rainy) بوده و تعداد تکرار هر مقدار روبروی آن ذکر شده است. سایر موارد این قسمت عبارتند از:

- **Unique:** این قسمت تعداد مقادیری از دامنه ورودی که در اطلاعات به صورت یکتا می‌باشند را نشان می‌دهد. با توجه به تعداد تکرار مقادیر ممکن برای خصیصه outlook، مشاهده می‌کنید که هیچ‌یک از این مقادیر یکتا نمی‌باشند؛ بنابراین مقدار این مورد صفر است.
- **Missing:** این مورد برای نشان دادن تعداد نمونه‌هایی است که در آنها برای این خصوصیت مقداری تعیین نشده است. در تصویر فوق این مورد نیز صفر می‌باشد.
- **Distinct:** این مورد برای نشان داده تعداد مقادیر ممکنه برای خصیصه‌های nominal استفاده می‌شود.

برای فهمیده‌هایی از نوع عددی در قسمت میانی مقدار کمینه، بیشینه، میانگین و انحراف از معیار برای داده مورد نظر نشان داده می‌شود. 

معمولاً WEKA آخرین فهمیده را به عنوان Class Variable در نظر می‌گیرد. به عنوان مثال در تصویر فوق متغیر play به عنوان متغیر کلاس‌بندی در نظر گرفته شده است. 

◀ **قسمت ④:** در این قسمت نحوه توزیع داده‌های هر دسته براساس مقدار متغیر کلاس نشان داده می‌شود.

اگر بر روی دکمه VisualizeAll کلیک کنید نحوه توزیع اطلاعات کلیه فهمیده‌ها براساس متغیر کلاس را در یک پنجره مجزا مشاهده خواهید کرد «تصویر 7». البته از لیست کنار این دکمه می‌توانید متغیر کلاس را نیز تغییر دهید. توجه داشته باشید برای اینکه دسته‌بندی به درستی انجام شود باید متغیر کلاس از نوع Nominal باشد 



تصویر 7. نحوه توزیع کلیه خصیصه‌ها براساس متغیر play.

قسمت ⑤: در این قسمت اطلاعاتی در مورد روند اجرای عملیات نشان داده می‌شود. هنگامی که آیکون پرنده نشسته است بدین معناست که هیچ کاری در دست انجام نیست. اگر پرنده در حال حرکت باشد بدین معناست که WEKA هنوز در حال انجام کارهایی است که کاربر برای آن تعیین کرده است و اگر پرنده بدون حرکت ایستاده باشد یعنی اینکه در مشکلی در اجرای عملیات درخواستی به وجود آمده است. برای رفع این مشکل بهترین روش ارائه شده، راه اندازی مجدد نرم‌افزار می‌باشد. در کادر status اطلاعات متنی در حین انجام کار نمایش داده می‌شود و با کلیک راست در این قسمت دو گزینه نمایان می‌شود:

- **Memory Information:** اطلاعاتی در مورد حافظه‌ای که به WEKA اختصاص داده شده، نشان داده می‌شود.

- **Run Garbage Collector:** با استفاده از این گزینه GC را صدا زده و آن را وادار به تمیزسازی حافظه می‌کنیم.

قسمت ⑥: در این قسمت می‌توانید الگوریتم‌هایی برای پالایش داده‌ها انتخاب و استفاده کنید. با کلیک بر روی دکمه Choose لیستی حاوی دو دسته کلی از الگوریتم‌ها نشان داده می‌شود که عبارتند از:

- **:Supervised**

- **:Unsupervised**

هر یک از این لیست‌ها دارای الگوریتم‌هایی برای کار بر روی خصیصه‌ها و نمونه‌ها به صورت مجزا می‌باشند.

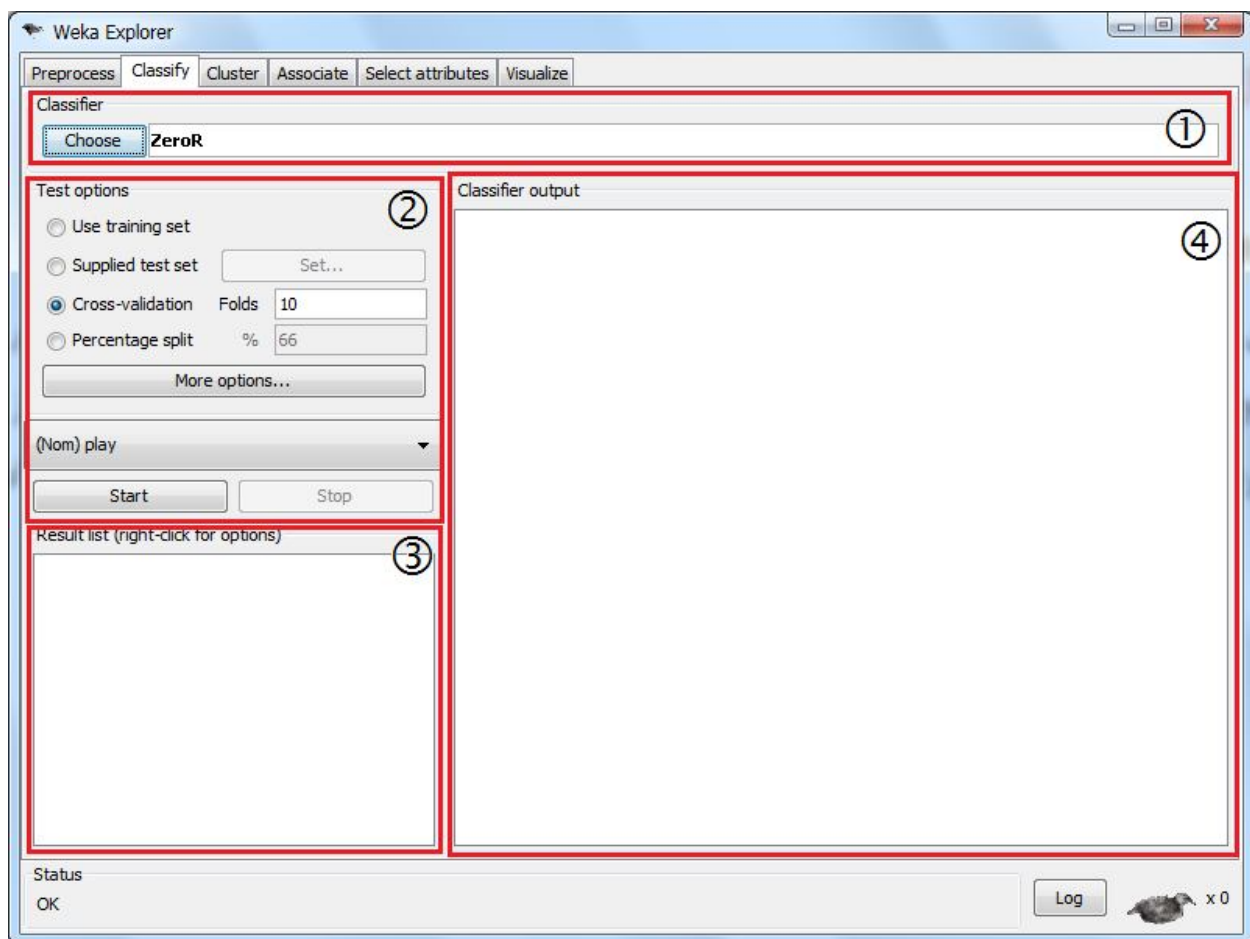
هنگام استفاده از فیلترهای Supervised باید به نمونه ارزیابی هم توجه داشته باشید. چراکه اعمال این فیلترها بر روی مجموعه test معمولاً جانب‌دارانه به جواب تمرکز می‌کند. این مورد برای فیلترهای Unsupervised برقرار نمی‌باشد.



پس از انتخاب فیلتر مد نظر کفایست با استفاده از دکمه Apply که در منتهی الیه سمت راست این قسمت قرار دارد فیلتر انتخاب شده را بر روی داده‌ها اعمال کنید.

برگه Classify

پس از بارگذاری اطلاعات در برگه Preprocess و انجام اصلاحات مد نظر بر روی اطلاعات، می‌توان در این برگه با استفاده از امکانات موجود، به طبقه‌بندی نمونه‌ها پرداخت. به عبارت دیگر یک مدل ارائه نمود که با استفاده از آن بتوان داده‌ها را طبقه‌بندی کرد. همچنین در این قسمت تمهیداتی برای تست کردن مدل ایجاد شده نیز وجود دارد که در ادامه به بررسی آن‌ها می‌پردازیم.



تصویر 8. برگه Classify.

◀ **قسمت 1:** در این قسمت می‌توانید متدهایی برای انجام عملیات کلاس‌بندی داده‌ها انتخاب کنید. به عنوان مثال می‌توانید برای ایجاد مدل‌هایی براساس درخت‌های تصمیم، شبکه Bayes و یا موارد دیگر، الگوریتم مناسبی را انتخاب نمایید. در ادامه همین قسمت نگاهی بر این مورد خواهیم داشت.

◀ **قسمت ②:** در این قسمت می‌توان نحوه تست کردن مدلی که قرار است با استفاده از داده‌های بارگذاری شده در برگه Preprocess و الگوریتم انتخاب شده در قسمت 1 ایجاد شود را تعیین کرد. برای این منظور گزینه‌های متفاوتی ارائه شده است که عبارتند از:

- **Use training set:** با استفاده از این گزینه، الگوریتم انتخاب شده با استفاده از همان مجموعه داده‌های آموزشی ارزیابی می‌گردد. بدیهی است که این گزینه ایده‌آل‌ترین نوع ارزیابی را برای ما به همراه دارد. در این حالت تمامی داده‌های به درستی در کلاس‌های خود طبقه‌بندی می‌شوند. اما سؤالی که مطرح است این است که چرا این گزینه قرار داده شده است؟! در جواب باید به این نکته توجه کرد که یک الگوریتم Classifier ممکن است در بعضی از حالات تصمیم بگیرد، بعضی از نمونه‌ها را طبقه‌بندی نکند. البته این مورد برای اکثر الگوریتم‌های موجود در WEKA معمولاً صادق نیست.
- **Supplied test set:** مدل توسط داده‌هایی که در قسمت Preprocess وارد شده، ایجاد شده و برای ارزیابی مدل باید یک مجموعه تست جدید معرفی نمایید. با کلیک بر روی دکمه مربوط به این گزینه صفحه‌ای حاوی چندین روش جهت بارگذاری اطلاعات تست مدل نمایان می‌شود.
- **Cross validation n folds:** در این گزینه پارامتر n درخواست می‌شود و بر اساس آن n مرتبه و در هر مرتبه 1/n داده‌ها به عنوان مجموعه test برای ارزیابی مدلی که با بقیه داده‌ها ایجاد شده استفاده می‌گردد و در انتها میانگین این n مرتبه اجرا به عنوان خروجی نهایی انتخاب می‌شود.
- **Percentage split:** در این گزینه درصدی درخواست می‌شود که با آن درصد از داده‌های ورودی train انجام شده و با بقیه داده‌ها، مدل ایجاد شده ارزیابی می‌شود.

◀ **قسمت ③:** پس از هر بار اجرا، نتایج در این قسمت نشان داده می‌شود. به عبارت دیگر سابقه اجرای الگوریتم‌ها در این قسمت ذخیره می‌شود و شما می‌توانید با کلیک بر روی هر یک از آنها در قسمت سمت راست خروجی کلی را مشاهده نمایید.

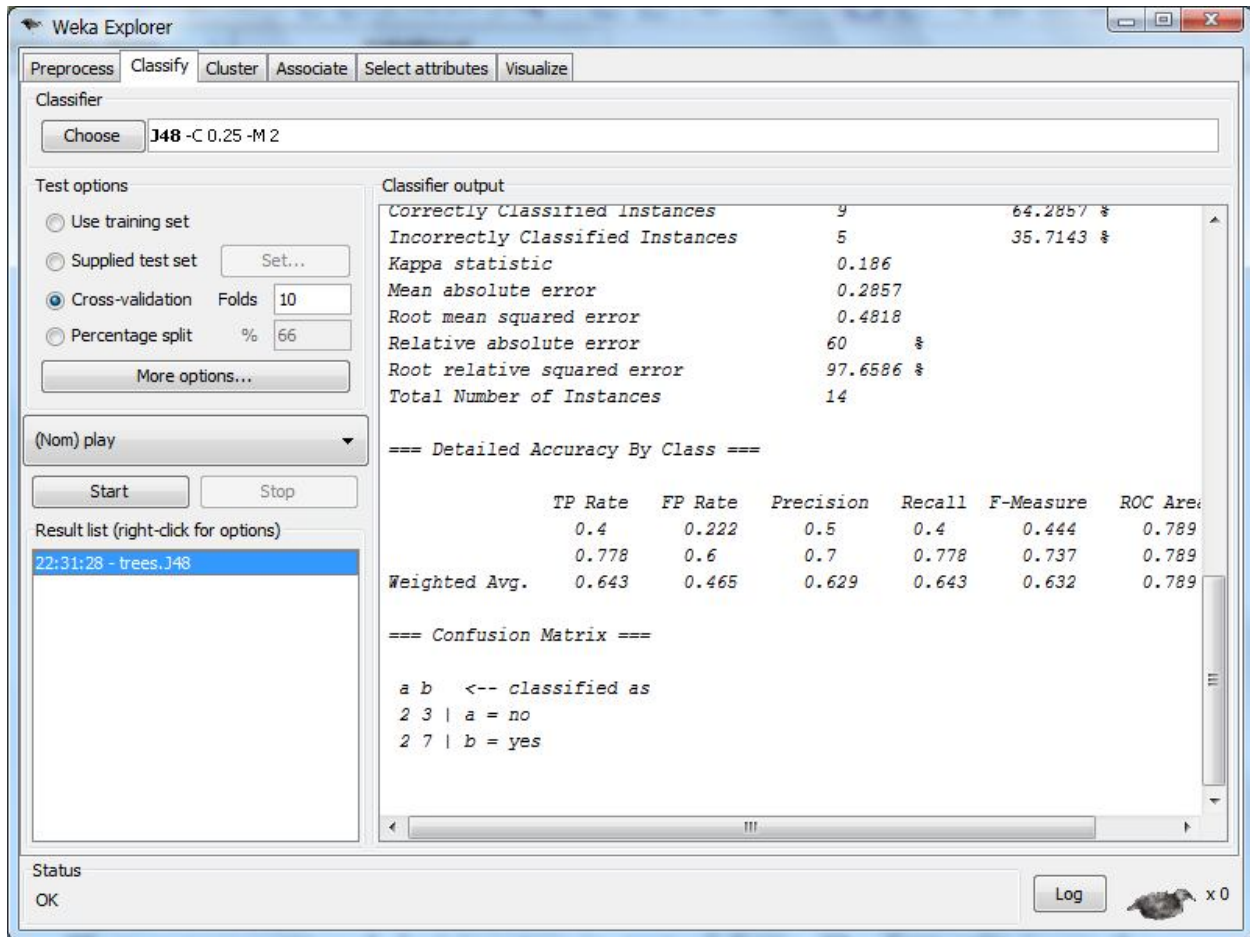
◀ **قسمت ④:** توضیحات تفصیلی اجرای الگوریتم‌ها در این قسمت ارائه می‌گردد.

مثالی از Classifier

در این قسمت قصد داریم برای اطلاعات موجود در فایل weather.arff یک درخت تصمیم ایجاد کنیم. معروف‌ترین الگوریتم‌های موجود برای این کار عبارتند از ID3، C4.5 و J4.8.

هر یک از این الگوریتم‌ها در حقیقت بهبود یافته الگوریتم قبل می‌باشد. در WEKA برای ایجاد یک درخت تصمیم از الگوریتم J48 استفاده می‌شود. برای استفاده از این الگوریتم پس از بارگذاری اطلاعات در برگه Preprocess به برگه Classify آمده و در آنجا بر روی دکمه Choose کلیک کنید. در دیالوگی که باز می‌شود از گزینه‌های موجود در قسمت trees گزینه J48 را انتخاب نمایید. این الگوریتم دارای پارامترهایی با مقادیر پیش‌فرض مناسب برای کار می‌باشد. در قسمت Test Option نیز گزینه Cross

Validation با مقدار 10 را انتخاب نمایید. در انتها بر روی دکمه Start کلیک نمایید. پس از گذشت مدت کمی خروجی ایجاد شده و صفحه اصلی برنامه به فرم زیر خواهد بود:



تصویر 9. خروجی Classifier.

در تصویر فوق قسمت انتهایی خروجی این الگوریتم نشان داده شده است. خروجی کامل این الگوریتم به صورت زیر

است:

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
 Relation: weather
 Instances: 14
 Attributes: 5
 outlook
 temperature
 humidity
 windy
 play
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

 outlook = sunny
 | humidity <= 75: yes (2.0)
 | humidity > 75: no (3.0)
 outlook = overcast: yes (4.0)
 outlook = rainy
 | windy = FALSE: yes (3.0)
 | windy = TRUE: no (2.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 9
 64.2857 %
 Incorrectly Classified Instances 5
 35.7143 %
 Kappa statistic 0.186
 Mean absolute error 0.2857
 Root mean squared error 0.4818
 Relative absolute error 60 %
 Root relative squared error 97.6586 %
 Total Number of Instances 14

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure
no	0.4	0.222	0.5	0.4	0.444
yes	0.789	0.778	0.6	0.7	0.737
Weighted Avg.	0.643	0.465	0.629	0.643	0.632

=== Confusion Matrix ===

a b <-- classified as
 2 3 | a = no
 2 7 | b = yes

①

③

②

④

⑤

◀ **قسمت ①:** در این قسمت اطلاعات کلی در مورد روشی که برای ایجاد درخت تصمیم انتخاب شده است، نام رابطه، نام خصیصه‌ها، تعداد نمونه‌ها و نحوه تست کردن مدل ایجاد شده نوشته شده است.

◀ **قسمت ②:** در این قسمت نحوه ایجاد درخت تصمیم به صورت متنی نشان داده شده است. همانطور که در بالا مشخص است ابتدا با استفاده از متغیر outlook به ایجاد درخت پرداخته است و سپس در مرحله بعد از متغیرهای humidity و windy برای این کار استفاده کرده است. در انتها نیز تعداد برگ‌های درخت و همچنین تعداد کل گره‌های آن و زمان سپری شده برای ایجاد مدل را ذکر کرده است.

نوع انتقاب فصیصهها براساس روشهای *Information Gain*، *Gain Ratio* و *Gini Index* می باشد. روشهای مختلف ممکن است از هر یک از این موارد استفاده کنند. به عنوان مثال در این روش از *Information Gain* برای انتقاب فصیصههایی به منظور ایبار درخت تصمیم استفاده می شود.



◀ قسمت ③: در این قسمت اطلاعات آماری در مورد درخت ایجاد شده ارائه می گردد. به عنوان مثال می توان به میزان طبقه بندی درست نمونه ها، طبقه بندی اشتباه نمونه ها و چندین اطلاعات آماری دیگر اشاره نمود. ضریب Kappa برای تعیین میزان تطبیق میان پیش بینی و مشاهدات را بیان می نماید. البته معمولاً این فاکتور در نظر گرفته نمی شود.

◀ قسمت ④: در این قسمت اطلاعاتی در مورد نحوه طبقه بندی اطلاعات در دسته های متفاوتی که در متغیر کلاس تعیین شده است ارائه می شود. در این مثال متغیر کلاس حاوی دو مقدار *yes* و *no* می باشد. تعدادی از موارد تعیین شده در این قسمت عبارتند از:

- **TP Rate**: True Positive مخفف True Positive است و به معنای میزان دسته بندی درست داده ها می باشد. این میزان برای هر کلاس به صورت جداگانه مشخص شده است.
- **FP Rate**: False Positive مخفف False Positive است و به معنای نمونه هایی است که به صورت اشتباه دسته بندی شده اند
- **Recall**: نسبت میزان کل مشاهدات طبقه بندی شده و مرتبط به به نسبت کل مشاهدات مرتبط بیان می کند.
- **Precision**: نسبت میزان کل مشاهدات طبقه بندی شده و مرتبط به به نسبت کل مشاهدات بیان می کند.
- **F-Measure**: این مورد از طریق فرمول زیر بدست می آید:

$$\frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

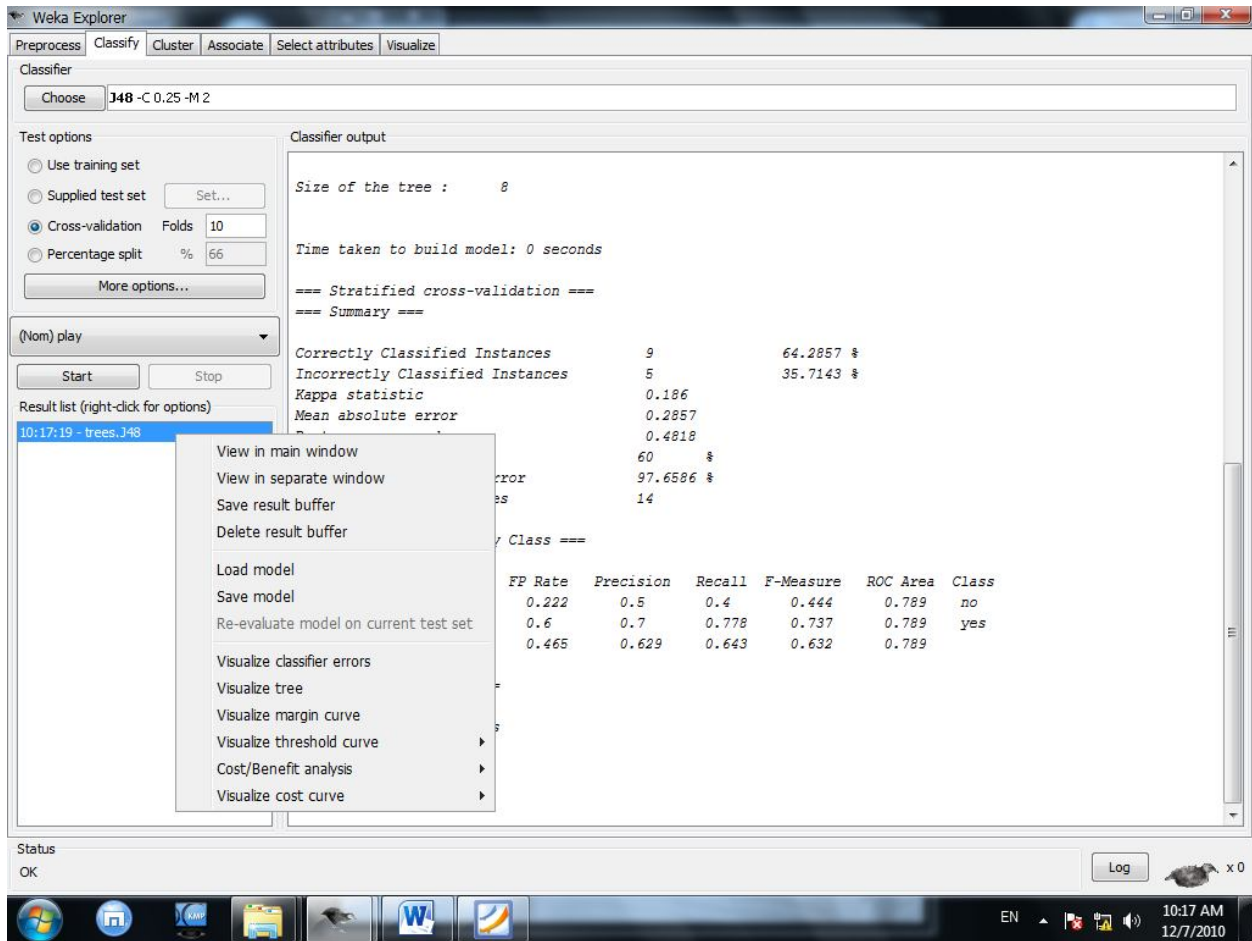
◀ قسمت ⑤: در این قسمت نحوه دسته بندی اطلاعات در کلاس های متفاوت را نشان می دهد.

د صورتیکه در جایی از صفحه کلیک سمت چپ ماوس را در حالی که کلیدهای *alt + ctrl + shift* را پایین نگه داشته ایبر را فشار دهید صفحه ای باز شده و به شما امکان می دهد فریبی ایبار شده را در قالب فایل تصویر ذخیره نمایید.



نمایش درخت تصمیم

پس از اینکه یک الگوریتم را بر روی داده ها کارش با موفقیت به اتمام رسید در قسمت History یک مورد برای آن ثبت می شود. با کلیک راست بر روی گزینه مربوط به آن منویی باز شده و به شما چندین گزینه برای مشاهده نمودارهای متفاوت را می دهد. در تصویر زیر این کار بر روی خروجی الگوریتم J48 انجام شده است:

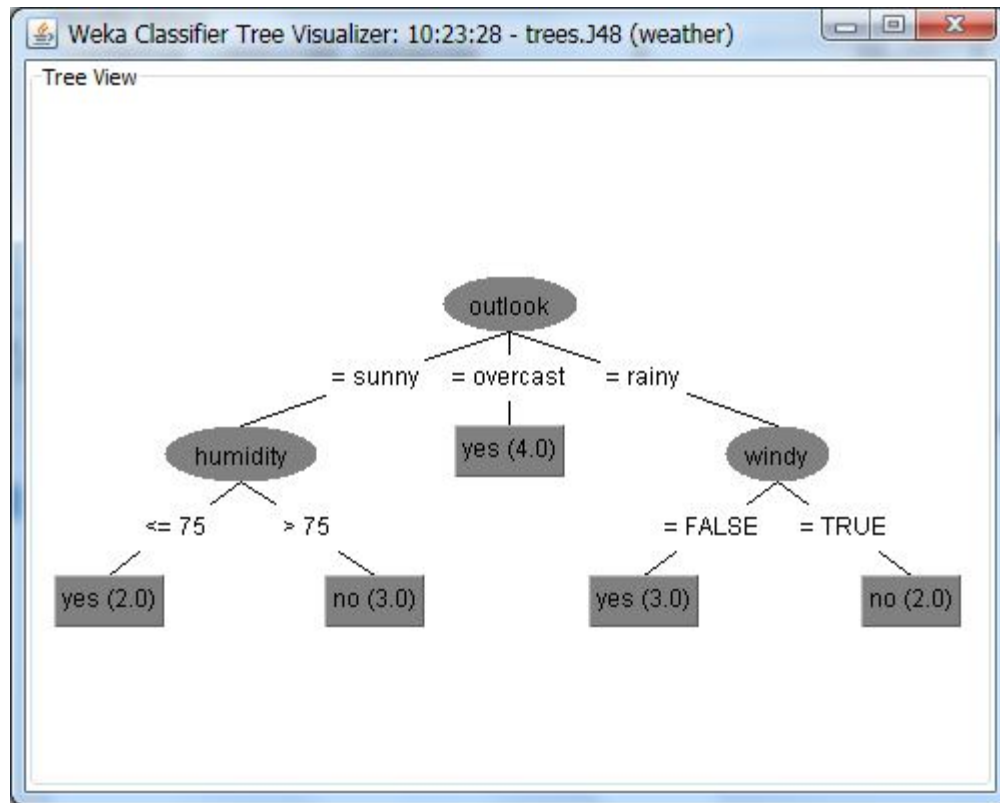


تصویر 10. منوی کلیک راست برای گزینه‌های موجود در سابقه اجرایی.

دو گزینه اول برای نحوه نشان دادن خروجی در یک پنجره جدا و یا در پنجره اصلی برنامه است. گزینه بعد به منظور ذخیره‌سازی خروجی در قالب یک فایل متنی گزینه چهارم برای حذف این خروجی از لیست History می‌باشد.

در قسمت بعد منو چندین گزینه برای ذخیره‌سازی خروجی مدل برحسب فرمتی است که برای جاوا شناخته شده است.

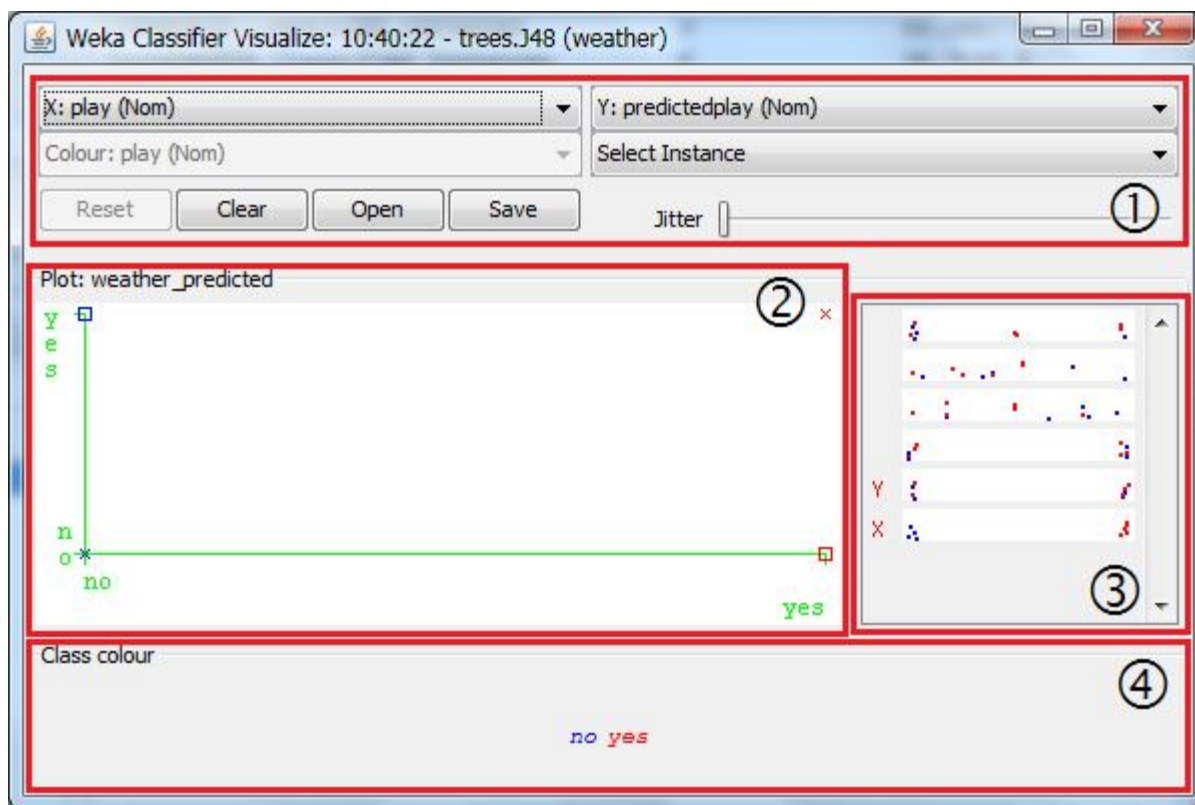
در قسمت بعد امکاناتی برای مشاهده نموداری خروجی وجود دارد. یکی از مهمترین گزینه‌ها Visualize tree می‌باشد. با انتخاب این گزینه شما می‌توانید خروجی مدل را به صورت یک درخت تصمیم مشاهده نمایید. خروجی این گزینه به صورت زیر است



تصویر 11. درخت تصمیم ایجاد شده براساس الگوریتم J48.

شما می‌توانید با استفاده درگ کردن به وسیله کلیک چپ ماوس نمودار را حرکت دهید. با کلیک راست بر روی صفحه نیز می‌توانید به گزینه‌هایی برای تعیین مختصات و سایز نمودار دسترسی پیدا کنید.

گزینه مهم دیگر در این قسمت Visualize classifier error می‌باشد در این نمودار وضعیت دسته‌بندی کردن اطلاعات و خطاهای آنها را به صورت نموداری مشاهده خواهید کرد. با انتخاب گزینه تصویر زیر نمایان خواهد شد:



تصویر 12. نمودار Visualize Classifier error.

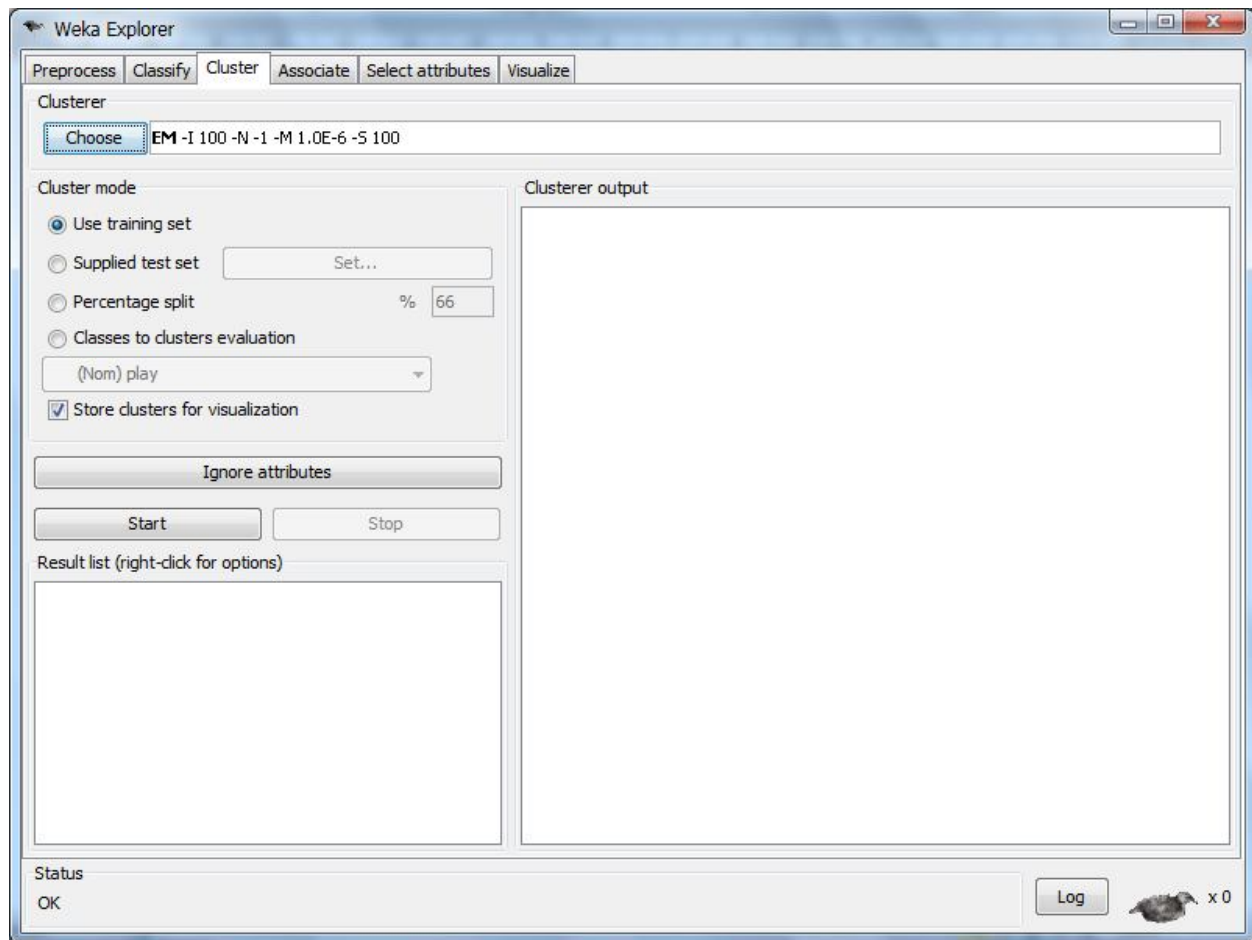
ممکن است در نمای اول این نمودار قدری برای شما نامفهوم باشد. اما جای نگرانی وجود ندارد. برای درک بهتر این صفحه به توضیحات زیر دقت نمایید:

- ◀ **قسمت ①:** در این قسمت می‌توان به راحتی متغیرهایی که قرار است بر روی محورهای نمودار قرار بگیرند را تعیین نمود.
- ◀ **قسمت ②:** در این قسمت نحوه پراکندگی داده‌ها بر اساس محورهای انتخاب شده در قسمت 1 را مشاهده خواهید کرد. علامت × به معنای دسته‌بندی درست و علامت مربع به معنای دسته‌بندی نادرست است.
- ◀ **قسمت ③:** در این قسمت کلیه متغیرهایی که می‌توان برای محورهای نمودار در نظر گرفت نمایش داده می‌شوند. با کلیک بر روی هر کدام می‌توان به راحتی آن را بر روی نمودار اعمال نمود.
- ◀ **قسمت ④:** در این بخش متغیر کلاس و مقادیر آن و همچنین رنگ‌هایی که برا دسته‌های مختلف در نظر گرفته شده است را نمایش می‌دهد.



برگه Cluster

از این برگه برای دسته‌بندی اطلاعات بدون توجه به متغیر کلاس استفاده می‌شود. در این بخش چندین الگوریتم وجود دارد که با استفاده از آن‌ها می‌توان دسته‌بندی اطلاعات استفاده نمود. یکی از معروف‌ترین الگوریتم‌های این زمینه الگوریتم EM می‌باشد. همچنین الگوریتم CobWeb نیز در میان الگوریتم‌ها موجود است. این الگوریتم هم برای داده‌های عددی و هم برای داده‌های اسمی کار می‌کند. نمای کلی این برگه مانند برگه Classify بوده و در تصویر زیر می‌توانید آن را مشاهده نمایید.

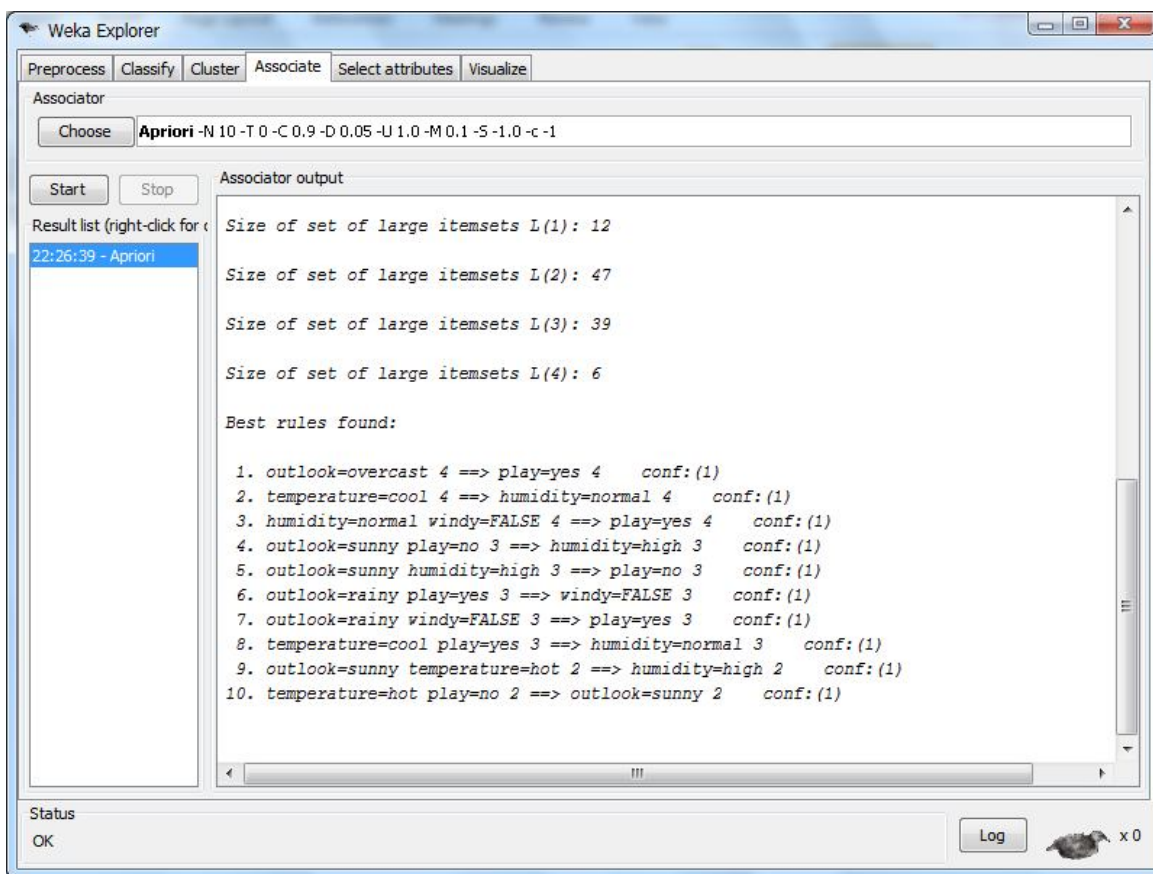


تصویر 13. نمای کلی برگه Cluster.

یکی از تفاوت‌های موجود در قسمت Test option است. 3 گزینه اول این قسمت مانند گزینه‌های مشابه در برگه Classify عمل می‌کنند. اما گزینه چهارم به منظور تطبیق دسته‌بندی انجام شده و کلاسی که از قبل برای داده‌ها تعیین شده است به کار می‌رود. در لیست پایین این گزینه می‌توانید متغیر کلاس را انتخاب نمایید. در اینجا به صورت پیش‌فرض نیز متغیر آخر به عنوان متغیر کلاس مورد استفاده قرار می‌گیرد.

برگه Associate

در این برگه امکاناتی برای انجام عملیات استخراج قوانین انجمنی در نظر گرفته شده است. یکی از معروفترین الگوریتم‌های موجود در این قسمت الگوریتم Apriori است. در این برگه با توجه به اینکه هدف استخراج قوانین انجمنی است دیگر نیاز به قسمت Test Option نیست. نمای کلی این صفحه به صورت زیر است:



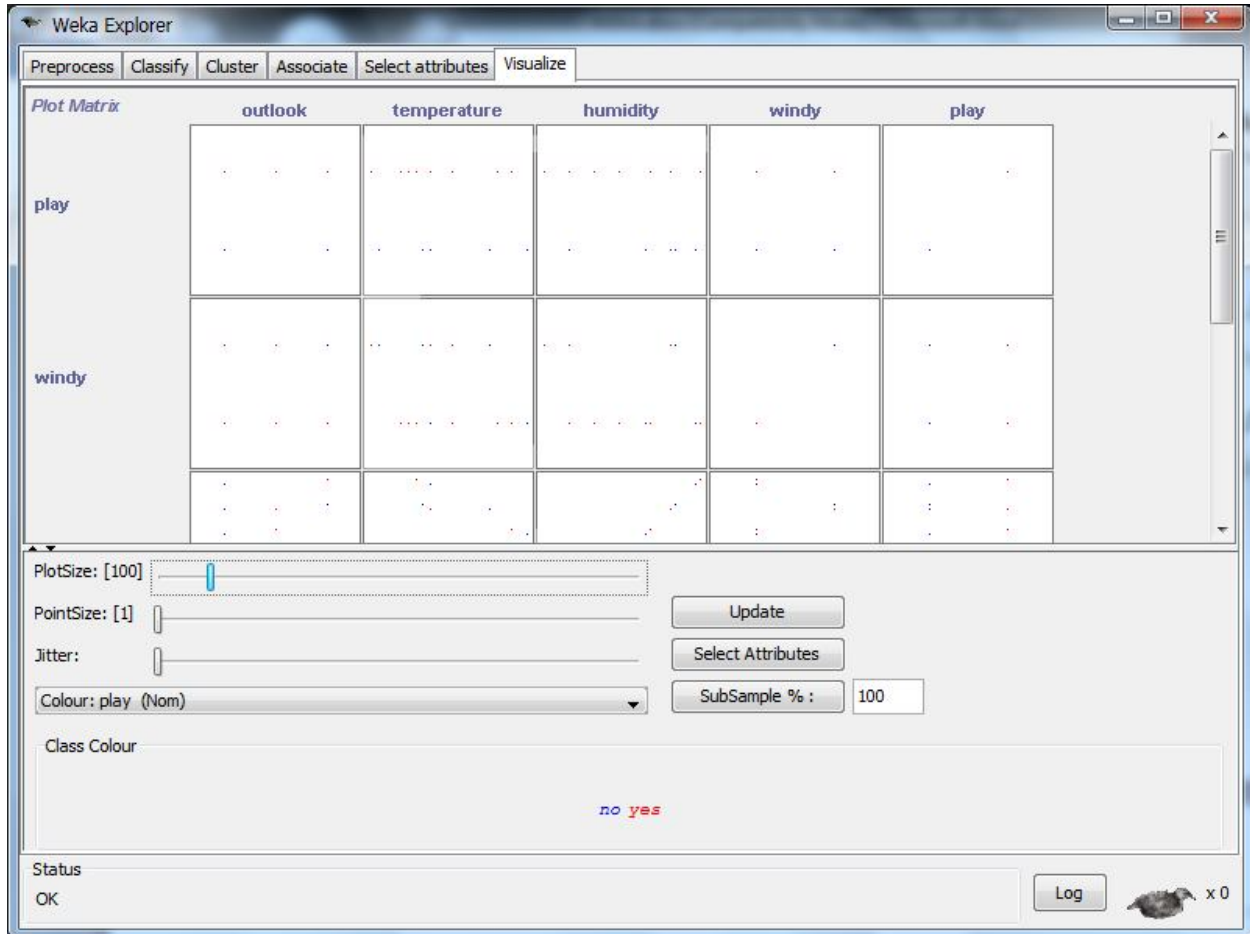
تصویر 14. نمای کلی برگه Associate.

توجه داشته باشید در WEKA باید این الگوریتم را باید بر روی داده‌های اسمی اجرا کنید. این الگوریتم کار را مقدار support معادل 100 درصد برای تمامی Data Itemها آغاز می‌کند. سپس در هر مرحله از تکرار میزان 5% از آن را کم می‌کند تا اینکه حداقل 10 قانون با میزان confidence برابر 0.9 ایجاد شوند و یا اینکه میزان support به 10% برسد. البته کلیه این اعداد و ارقام به صورت پیش فرض تعیین شده‌اند و شما می‌توانید آنها را تغییر دهید.

الگوریتم دیگر در این زمینه الگوریتم PredictiveApriori می‌باشد. این الگوریتم confidence و support را با هم ترکیب کرده و یک واحد اندازه‌گیری به نام Predictive Accuracy ارائه می‌نماید و سپس به ترتیب n قانون مورد نظر را بازایی می‌کند.

برگه Visualize

در این برگه نحوه توزیع داده‌های خصیصه‌های متفاوت بر اساس متغیر کلاس نشان داده شده است. در تصویر زیر نمای کلی این برگه را مشاهده می‌کنید.



تصویر 15. نمای کلی برگه Visualize.