مؤسسه آموزش عالی غیرانتفاعی مجازی
نور طوبی

پایان نامه کارشناسی ارشد

رشته:تجارت الکترونیک

عنوان:

غنی سازی محیط های آموزش الکترونیکی در ایران با استفاده از تکنیک های داده کاوی

استاد راهنما:

جناب آقای دکتر حسن احمدی ترشیزی

استاد مشاور:

جناب آقای مهندس علیرضا قنادان

نگارش: آلاله رنگ ریز

آذر ۸۹

مؤسسه آموزش عالی غیرانتفاعی مجازی
نور طوبی

پایان نامه کارشناسی ارشد

رشته:تجارت الکترونیک

عنوان:

غنی سازی محیط های آموزش الکترونیکی در ایران با استفاده از تکنیک های داده کاوی

استاد راهنما:

جناب آقای دکتر حسن احمدی ترشیزی

استاد مشاور:

جناب آقای مهندس علیرضا قنادان

نگارش: آلاله رنگ ریز

آذر ۸۹

فرم شماره چهار

شماره.....

تاریخ.....

پیوست.....

مؤسسه آموزش عالی غیردولتی غیرانتفاعی مهاجر
شهر دلوپی

بسمه تعالی

صورتجلسه دفاعیه پایان نامه کارشناسی ارشد

باعنایت به آئین نامه آموزشی دوره کارشناسی ارشد ناپیوسته ، جلسه دفاعیه پایان نامه کارشناسی ارشد خانم / آقای [illegible] دانشجوی رشته مهندسی عمران اجرا گرایش مدیریت اجرا به شماره دانشجویی ۸۷۰۶۰۲۰۲ باعنوان [illegible] در تاریخ ۸۹/۹/۲ در محل مؤسسه با حضورهیأت داوران تشکیل شد و براساس کیفیت پایان نامه ، ارائه دفاعیه و نحوه پاسخ به سؤالات ، رأی نهایی به شرح ذیل اعلام گردید.

پایان نامه مورد قبول می باشد ☑ پایان نامه با اصلاحات مورد قبول می باشد ☐ پایان نامه مورد قبول نمی باشد ☐

تعداد واحد پایان نامه ۶ نمره نهایی پایان نامه به عدد ۱۹ به حروف نوزده درجه پایان نامه عالی

ردیف	مشخصات هیات داوران	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضاء
۱	استاد راهنما استاد راهنمای دوم(حسب مورد)	دکتر حسن احدی ترشیزی	استادیار	دانشگاه آزاد اسلامی	
۲	استاد مشاور اول استاد مشاور دوم (حسب مورد)	علیرضا تاوران	مربی	دانشگاه نورضیع	
۳	استاد داور	دکتر مهدی یعقوبی	استادیار	دانشگاه آزاد اسلامی	
۴	معاون آموزشی و تحصیلات تکمیلی مؤسسه یا نماینده وی	دکتر فرهاد وارث			

نام و نام خانوادگی مدیر گروه یا سرپرست تحصیلات تکمیلی گروه

امضاء تاریخ ۸۹/۹/۲

نام و نام خانوادگی معاون آموزشی و تحصیلات تکمیلی مؤسسه

امضاء تاریخ ۸۹/۹/۲

بسم الله الرحمن الرحيم

بسمه تعالی

اظهارنامه(اصالت اثر)

اینجانب آلاله رنگ ریز دانشجوی رشته فناوری اطلاعات گرایش تجارت الکترونیک مؤسسه آموزش عالی غیرانتفاعی مجازی نور طوبی اظهار می کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده ام ، نشانی دقیق و مشخصات کامل آن را نوشته ام. همچنین اظهار می کنم که تحقیق و موضوع پایان نامه ام تکراری نیست و تعهد می نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر ندهم. کلیه حقوق این اثر مطابق با آیین نامه مالکیت فکری و معنوی متعلق به مؤسسه آموزش عالی نورطوبی است.

نام و نام خانوادگی: آلاله رنگ ریز

تاریخ و امضاء: ۸۹/۹/۲

به نام خدا

غنی سازی محیط های آموزش الکترونیکی در ایران با استفاده از تکنیک های داده کاوی

نگارنده

آلاله رنگ ریز

پایان نامه

ارائه شده به تحصیلات تکمیلی مؤسسه آموزش عالی نورطوبی به عنوان بخشی از فعالیت های تحصیلی لازم برای اخذ در جه کارشناسی ارشد

در رشته ی

فناوری اطلاعات (تجارت الکترونیک)

از مؤسسه آموزش عالی نورطوبی

تهران

جمهوری اسلامی ایران

ارزیابی کمیته پایان نامه ، با درجه ی ...عالی...

۱- استاد راهنما : دکتر حمید امیر تیموری

۲- استاد مشاور :

۳- استاد داور :

سپاسگزاری

اکنون که این رساله به پایان رسیده است بر خود فرض می دانم که از استاد ارجمند جناب آقای دکتر حسن احمدی که تجربیات گرانقدر خویش را در اختیارم نهادند کمال تشکر را به جای آورم.

چکیده

غنی سازی محیط های آموزش الکترونیکی در ایران با استفاده از تکنیک های داده کاوی

نگارنده

آلاله رنگ ریز

هدف از انجام این پژوهش، غنی سازی محیط های آموزش الکترونیکی از طریق پیش بینی نتیجه تحصیلی دانشجویان بود. تعیین دانشجویانی که محتمل به عملکرد ضعیف در امتحاناتشان هستند مفید واقع می شود. در این پژوهش از نرم افزار داده کاوی متن باز weka جهت تحلیل ویژگی ها به منظور پیش بینی نتیجه تحصیلی دانشجویان استفاده نمودیم. مجموعه داده ها شامل ۱۵۴ رکورد از دانشجویان دانشگاه مجازی نور طوبی تهران که بین سالهای ۲۰۰۶ تا ۲۰۰۹ ثبت نام نموده اند بود. ما از پنج طبقه بندی کننده (MLP, RandomForest, Logistic, OneR, Naïve Bayes, J48) استفاده نمودیم و میزان صحت پیش بینی نتایج تحصیلی دانشجویان به دو کلاس موفق و ناموفق را بدست آوردیم. نتایج تحصیلی دانشجویان بوسیله دانش کشف شده از پایگاه داده های موجود که شامل تجربیات گذشته است قابل پیش بینی می باشد. روش ارزیابی متقاطع با ده دسته (Fold) برای ارزیابی صحت پیش بینی بکار رفت. ۶۶٪ داده ها را برای آموزش گرفتن(Train) و ۳۴٪ آنها را برای تست (Test) تخصیص دادیم. نتایج نشان داد که روش RandomForest با ۹۱.۶٪ بیشترین میزان درستی پیش بینی را داراست.MLP با ۸۴٪، Logistic با ۸۰٪، OneR با 75.7% ، Naïve Bayes با ۷۴.۴٪ و روش J48 با ۷۳.۳٪ در رتبه های بعدی قرار گرفتند.

فهرست

عنوان صفحه

فهرست جدول ها

عنوان و شماره صفحه

فهرست شکل ها

عنوان و شماره صفحه

مقدمه

عصر ارتباطات و فناوری اطلاعات است. رفته رفته مرزهای دسترسی به دانش و اطلاعات از بین می‌رود و همه چیز جهانی می‌شود. یادگیری الکترونیکی فراهم کردن شرایط یادگیری در هر مکان و هر زمان با بهره‌گیری از انواع ابزارها و رسانه‌های مدرن صوتی و تصویری و چندرسانه ای می‌باشد. اینک دیگر دیوارهای مدرسه چارچوبی برای یادگیری شما نیستند؛ همه جا مکان یادگیری است. هر زمان که اراده نمایید تنها با داشتن یک رایانه می توانید به دانش دسترسی پیدا کنید. یادگیری الکترونیکی فراهم کردن شرایط یادگیری در هر مکان و هر زمان با بهره‌گیری از انواع ابزارها و رسانه‌های مدرن صوتی و تصویری و چندرسانه ای می‌باشد. تغییر روند در تولید وعرضه و دسترسی به دانش ها، چگونگی پردازش، بهره وری، ذخیره سازی و انبار کردن آن، بی شک در همه جوانب مرتبط با دانش تغییراتی را در بر خواهد داشت و از آن جمله تغییرات در سیستم تعلیم و انتقال آن به نوآموزان فضای آموزشی، معلمان، استادان و کادر تعلیم و تربیت است. سیستم و ارگان تعلیم و تربیت و آموزش باید به گونه ای مجازی و سایبرنتیک یا کنترل از راه دور و فرمانی شدن دستگاه ها و با تکیه برکامپیوتر و سیستم ارتباطاتی یا اینترنت شکل بگیرد و هیچ نیازی به تجمع و صرف وقت برای هماهنگی میان اجزاء نداشته و در هر لحظه از زمان و در هر نقطه از مکان در دسترس باشد.

سیستم آموزشی هوشمند و سازگار در محیط هایی که نیاز به یادگیری شخصی غنی تری احساس می شود راهگشا خواهد بود. سیستم آموزشی هوشمند و شخصی شده تلاش می کند بوسیله طرحی از ویژگی های شخصی، علایق و معلومات فراگیران، شکلی از آموزش اختصاصی را به فراگیران ارائه دهد. داده کاوی یا دانش استخراج خودکار از الگوهای مفید از مجموعه داده های بزرگ، می تواند برای دستیابی به مدل فرآیند یادگیری و مدل یابی فراگیران به کار رود و با جستجو و یافتن اطلاعات آموزشی سودمند بر مبنای اسناد آموزشی در ارزیابی و بهبود سیستم آموزش نیز استفاده شود.

هدف از ارائه این پایان نامه بررسی نحوه استفاده از روش های داده کاوی و یادگیری ماشین در پیشگویی رفتارهای دانشجویان سیستم های آموزش الکترونیک در دانشگاههای مجازی ایران بوده است. در همین راستا، در این پژوهش بعد از تشریح مفصل روش های داده کاوی و بررسی دقیق کارهای انجام شده در حوزه داده کاوی آموزشی و مدل کردن فراگیران به پیشگویی نتایج تحصیلی فراگیران سیستم آموزش الکترونیک با استفاده از تکنیک های یادگیری ماشین پرداخته شده است. بدین منظور از یک مجموعه داده حاوی اطلاعات ۱۵۴ نفر از دانشجویان موسسه آموزش عالی غیرانتفاعی مجازی نور طوبی تهران استفاده شده و با اعمال روشهای مختلف نظیر Naïve Bayes، درخت تصمیم J48، Logistic Regression، OneR و MLP و RandomForest به پیشگویی نتایج تحصیلی فراگیران پرداخته شده است که در بین این شش روش، روش RandomForest با ۹۱.۶٪ پیش بینی درست در رتبه اول قرار دارد. رتبه های بعدی به ترتیب عبارتند از MLP با ۸۴٪، Logic Regression با ۸۰٪ ، OneR با ۷۵.۷ ٪، روش Naïve Bayes با۷۴.۴ ٪ و روش J48 با۷۳.۷ ٪ .

در این پایان نامه بعد از مقدمه در فصل اول، در فصل دوم کاربردهای وب کاوی را در سیستم های آموزش الکترونیکی شرح داده ایم و در فصل سوم مروری بر تحقیقات انجام شده در حوزه وب کاوی در آموزش الکترونیک در ایران و سایر نقاط جهان صورت گرفته است. در فصل چهارم قابلیت های نرم افزار وکا مورد بحث قرار گرفته است و در فصل پنجم تحلیل داده ها موضوع مورد بررسی است. نهایتا در فصل ششم نتیجه گیری شده است.

فصل اول

مقدمه ای بر آموزش الکترونیکی

با گسترده شدن فناوری اطلاعات و نفوذ وسایل ارتباط از راه دور به عمق جامعه ، ابزار ها و روش های آموزش نیز دچار تحول شدند. تحول این ابزار ها و روش ها در جهتی است که هر فرد در هر زمان و هر مکان بتواند با امکانات خودش و در بازه زمانی که خودش مشخص می کند مشغول یادگیری شود.

در سال های نه چندان دور آموزش از راه دور مطرح شد. این نوع آموزش ویژگی های خودش را دارد و دارای مزایا و معایبی است. ابتدا آموزش به صورت مکاتبه ای بود و تنها راه ارتباط استفاده از نامه بود. با پیشرفت تکنولوژی و از همه مهمتر ارزان تر شدن هزینه استفاده از تکنولوژی، استفاده از ابزار های جدید تر برای انتقال دانش مطرح شد. با به وجود آمدن و گسترش اینترنت این پدیده جدی تر دنبال شد و ابزار ها و روش ها و استاندارد هایی برای آموزش الکترونیک مطرح شد و هر روز اصلاحات جدیدتری در این زمینه انجام می شود. در واقع می توان گفت آموزش الکترونیکی استفاده از ابزار های انتقال اطلاعات به صورت الکترونیکی (مثل اینترنت) برای انتقال اطلاعات و دانش است. Training به معنی آموزش دادن و learning به معنی یاد گرفتن است. در رویکرد سنتی از روش آموزش دادن استفاده می شود که چون همراه با زور و اجبار است باعث افت تحصیلی دانش آموز می شود و حتی ممکن است مشکلاتی در خارج از محیط آموزشی برای دانش آموز پیش بیاورد. اما در رویکرد جدید از روش آموزش گرفتن استفاده می شود. چون در این روش فرد، خود می خواهد یاد بگیرد بنابر این زور و اجبار در کار نیست و در نتیجه مشکلات قبلی به وجود نخواهد آمد.

۱ – ۱ چند تفاوت آموزش سنتی و مدرن

۱ - فرآیند آموزش

رویکرد سنتی: یک روش و محتوای آموزشی برای همه

رویکرد مدرن: روش و محتوی آموزشی سفارشی شده برای هر آموزش

۲ - زمان آموزش

رویکرد سنتی : یک بازه زمانی برای همه.

رویکرد مدرن : بازه زمانی که توسط آموزش گیرنده تعیین می شود.

۳ - محل آموزشی

رویکرد سنتی : یک مکان خاص برای همه.

رویکرد مدرن : هر مکانی که آموزش گیرنده بتواند به مطالب آموزشی دسترسی پیدا کند.

۱ – ۲ چند دلیل برای استفاده از آموزش الکترونیکی

- معرفی تکنولوژی های جدید به دانش آموزان و آموزش چگونگی دست یابی به تکنولوژی های جدید.
- تشویق دانش آموزان برای بدست آوردن اطلاعات از طریق تکنولوژی های جدید.
- حضور در اتاق های بحث و گفتگو برای تبادل نظر در مورد مطالب که منجر به تولید دانش می شود.
- برقرای ارتباط با مدیر قسمت آموزشی در هر زمان.
- سودمند برای کسانی که به علت داشتن کار یا معلولیت و ... قادر به حضور در کلاس درس فزیکی نیستند.
- سودمند برای افرادی که امکان خروج از کشورشان را برای ادامه تحصیل ندارند.
- دستیابی به جدید ترین اطلاعات روز دنیا در رشته مورد نظر .

۱ – ۳ دامنه آموزش الکترونیک

بعد از اینکه با مفاهیم مقدماتی آموزش الکترونیک آشنا شدیم بهتر است دامنه یادگیری آن را نیز بدانیم یا به عبارت بهتر، آموزش الکترونیک در چه محیط هایی و با چه ابزاهایی ارائه می شود. آموزش الکترونیک دامنه گسترده ای دارد و بسته به نوع استفاده و امکانات به چند دسته تقسیم می شود.

۱ - آموزش بر پایه وب: این روش آموزش از طریق اینترنت خواهد بود. در اکثر موارد آزمون ها و ارائه مدرک هم از طریق وب است. کلاس های درس ، یادداشت های درس، جزوه ها، اتاق بحث، پست الکترونیکی و ... جزء ویژگی های این روش هستند و همگی برروی وب ذخیره می شوند. البته به علت انعطاف پذیری فوق العاده آموزش الکترونیک می توانید نحوه آموزش را به طریق دلخواه، مناسب با فعالیت خودتان، شرایط موجود و امکانات طراحی و پیده سازی کنید؛ از این رو بعضی از ویژگی های گفته شده می توانند در سیستم شما وجود نداشته باشند یا ویژگی های دیگری جایگزین آن ها باشند.
۲ - آموزش مبتنی بر کامپیوتر: در این روش احتیاجی به اتصال به اینترنت و حتی به شبکه نیز نیست. مگر در موارد خاص. در این روش اطلاعات بروی یک واسط الکترونیکی ذخیره می شود و کاربر با استفاده از یک کامپیوتر یا ابزار خواننده آن واسط الکترونیکی می تواند از آن استفاده کند. یک مثال متعارف آن استفاده از CD آموزشی است که در کشور خودمان بسیار از آن استفاده می شود.

۳ - آموزش از طریق وسایل و ابزار دیجیتال همراه: آموزشی است که از طریق و سایل و ابزارهای دیجیتالی همراه از جمله PDA و Tablet PC و ... ارائه می شود.

۴ - آموزش از طریق تلفن همراه: آموزشی است که کاملا جدید است و تقریبا می تواند در گروه بالا قرار گیرد اما به علت افزایش تعداد افراد دارنده تلفن همراه و تمرکز بروی این روش که به m-learning شهرت یافته دسته ای جدا برای آن در نظر گرفته می شود. به خوبی می توان این روش را پیاده سازی کرد ولی لازمه آن ایجاد زیر ساخت های مخابراتی است. خیلی از مردم مخصوصا در جوامع شهری وقت زیادی از خود را در انتظار وسایل نقلیه، ترافیک و ... از دست می دهند. تا چند سال پیش افراد برای استفاده از این وقت، کتاب هایی همراه خود جابه جا می کردند که مشکات خودش را داشت مثلا در را ه های ناهموار

آسیب زیادی به چشم وارد می شد و ... . با ارائه ذخیره ساز های دیجیتالی صدا، مطالعه به سمت آموزش از طریق صدا حرکت کرد؛ ولی همچنان آموزش از طریق چشم بالاترین کارایی را دارد. بنابراین استفاده از ابزارهای دیجیتالی تصویری بهتر از همه است زیرا:

- می توان به جای نوشته از تصاویر ثابت و متحرک و یا فیلم استفاده کرد.
- می توان اندازه نوشته ها را بسته به موقعیت و کاملا دلخواه بزرگ و کوچک کرد.
- می توان از خوانندگان متن یا صدای ضبط شده همراه متن و حتی موسیقی در کنار آموزش استفاده کرد.
- در صورت خستگی از مطالعه می توان از وسایل تفریحی داخل این وسایل استفاده کرد.

۱ – ۴ دسته بندی نوع یادگیری

در مجموع نوع یادگیری را می توان به سه دسته تقسیم کرد:

۱ - یادگیری شخصی

۲ - یادگیری جمعی

۳ - کلاس های مجازی

۱ - یا دگیری شخصی : در این دسته فرد رشته مورد علاقه خود را انتخاب می کند و در محیط اطراف خود مخصوصا اینترنت به دنبال اطلاعات مرتبط با آن می گردد و در آن زمینه تحقیق می کند سپس سوالات خود را از اساتید آن رشته به صورت offline می پرسد.

۲ - یادگیری جمعی : در این دسته شرایطی برای افراد مهیا می شود تا با یکدیگر و اساتید خود ارتباط بر قرار کنند. از جمله این ابزار Forum و chat و... است. در این روش معمولا زمان شروع و خاتمه دوره آموزشی و امتحانات برای همه آن گروه یکسان است.

۳ - کلاس های مجازی : در این دسته شرایط کاملا مانند کلاس درس است و حتی در بعضی از موارد در

کلاس های فزیکی برگزار می شود. در این جا از ویدئو کنفرانس و به جای تخته سیاه از یک ویدئو پروژکتور استفاده می شود. گاهی از اوقات برای هر فرد یک کامپیوتر در نظر گرفته می شود و ارتباط ویدئویی از طریق صفحه نمایشگر و دوربین یا وب کم خواهد بود و هر کسی می تواند از طریق کامپیوتر با استاد ارتباط برقرار کند. این روش مخصوصا برای برگزاری کلاس هایی که استاد مربوط به آن درس به تعداد کافی موجود نیست و امکان جابه جایی اساتید هم وجود ندارد مفید می باشد به ویژه برا ی دانشگاه ها. شاخه ای از این دسته در پزشکی از راه دور نیز استفاده می شود.

چند نکته که در آموزش الکترونیکی باید مورد توجه قرار گیرد:

۱ - زیر ساخت های مخابر اتی: زیر ساخت های مخابرتی در ایران هنوز مهیا نیست اما باعث نمی شود از آموزش الکترونیک صرف نظر شود.

۲ - مفاد آموزشی: مفاد آموزشی به صورت آموزش الکترونیک با روش های آموزش سنتی متفاوت است و باید هم فرق داشته باشد. در آموزش الکترونیک ۴۰ تا ۵۰ درصد متن آموزشی از طریق استاد و بقیه از همکاری و ارتباط دانشجویان تعیین می شود.

۳ - آموزش الکترونیک باید دوطرفه باشد تا کاربر خسته نشود، مانند CD های آموزشی که فقط باید بیننده باشد نباشد. کاربر باید با آن ارتباط داشته باشد.

فصل دوم

وب کاوی آموزشی

تاکنون مطالعات زیادی در مورد پیش بینی نتیجه تحصیلی دانشجویان به منظور غنی سازی محیط های آموزشی و یادگیری صورت گرفته است که شامل تکنیک های مختلفی از جمله تحلیل های آماری، یادگیری ماشین و داده کاوی می باشد.

در [10] با استفاده از تکنیک های تحلیل های آماری به پیش بینی نتایج تحصیلی دانشجویان براساس یک درس خاص پرداخت. وی نشان داد که ارتباط عمیقی بین دروس مختلف جهت پیش بینی نتیجه تحصیلی وجود دارد، بدین معنا که براساس نمره نهایی دانشجو در یک درس خاص می توان نمره وی را در درس مرتبط دیگری در ترم آینده پیش بینی کرد.

در [65] و [66] نشان دادند که عوامل مختلفی مانند حضور در کلاس، سوابق تحصیلی و مهارت های مطالعه در نتیجه تحصیلی دانشجویان نقش بسیار مهمی را بازی می کنند. آنها در پیش بینی نتیجه تحصیلی دانشجویان از روشهای تحلیلی درخت تصمیم گیری (Decision Tree) و شبکه های عصبی (Neural Networks) استفاده نمودند. البته میزان دقت پیش بینی چندان خوب نبود و این بدلیل دشوار بودن طبقه بندی دانشجویان به سه گروه یعنی ریسک بالا (HighRisk)، ریسک متوسط (MediumRisk) و ریسک پایین (LowRisk) قبل از اولین امتحان دانشگاه بود.

در [22] نشان دادند که استفاده از عامل عملکرد در درس علوم کامپیوتر سال اول دانشگاه برای دانشجویان رشته کامپیوتر عامل مناسبی جهت پیش بینی نتیجه تحصیلی دانشجویان است. آنها همچنین نشان دادند که جنسیت (Gender) و سن (Age) نقش مهمی بعنوان عوامل پیش بینی کننده ندارند.

در [43] با استفاده از الگوریتم های داده کاوی به کشف الگوهای پیش بینی کننده جالبی نائل شدند. هدف آنها کمک به معلمان جهت مدیریت کلاسهایشان، درک رفتار دانشجو و حمایت از واکنش یادگیرنده از طریق بازخورد پرواکتیو (Proactive Feedback) بود. آنها پایگاه داده های پژوهش خود را از یک محیط آموزشی تحت وب کسب نمودند.

در [33] به مقایسه پیش بینی نتیجه تحصیلی دانشجویان که با شش طبقه بندی کننده (Naive Bayes, decision tree, feed-forward neural network, support vector machine, 3-nearest neighbor, logistic regression) صورت گرفته بود پرداختند. پایگاه داده های آنها شامل ۳۵۰ رکورد حاوی داده های دموگرافیک (Demographic Data)، اولین نمره نگارش آنها و شرکت در فعالیت های گروهی بود. بهترین طبقه بندی کننده آنها Naïve Bayes و Neural Network با میزان پیش بینی ۸۰٪ بود.

در [46] به پیش بینی نمرات نهایی دانشجویان از طریق داده های لاگ شده سیستم یادگیری و با مقایسه شش طبقه بندی کننده (quadratic Bayesian classifier, 1-nearest neighbors, k-nearest neighbors, Parzen window, feed-forward neural network , decision tree) پرداختند. داده های آنها شامل ۲۵۰ رکورد که حاوی ویژگی هایی از قبیل انجام تکالیف و شرکت در فعالیت های گروهی و مطالعه بود. بهترین طبقه بندی کننده آنها k-nearest neighbors با میزان صحت (Accuracy) ۸۰٪ در شرایطی که نتایج نهایی تنها شامل دو کلاس قبول و مردود بود.

در [45] با استفاده از طبقه بندی کننده های داده کاوی بعنوان ابزاری برای تحلیل و مقایسه به پیش بینی نتیجه تحصیلی دانشجویانی که یک درس فنی از طریق وب گذرانده بودند، پرداختند. نتایج آنها نشان داد که با تلفیقی از چند طبقه بندی کننده میزان دقت پیش بینی نتیجه تحصیلی دانشجویان بالا می رود.

در [2] پژوهشی صورت گرفته است برای بررسی کاربرد تکنیک های توصیفی داده کاوی در سیستم آموزش عالی در دو زمینه: (۱) دسته بندی دانشجویان به لحاظ سطح علمی و (۲) استخراج روابط و روند های موجود در داده های سیستم آموزشی دانشگاه با مطالعه موردی بر روی دانشگاه علم و صنعت ایران. دسته بندی دانشجویان بر اساس معدل کل، میانگین تعداد واحد گذرانده شده در هر نیمسال، سوابق مشروطی و سوابق

ممتازی انجام شده است. این دسته بندی با استفاده از تکنیکهای خوشه بندی Two Step و k-means صورت گرفته است. نتایج حاصل از این تحقیق می تواند برای شناسایی دانشجویان با ریسک بالا و دانشجویان ممتاز مورد استفاده قرار گیرد. استفاده از این نتایج می تواند کمک زیادی برای مدیران آموزشی دانشگاه ها در جهت بهینه کردن فرایندهای آموزشی باشد.

داده کاوی می تواند بر روی داده هایی که از دو نوع سیستم آموزشی استخراج می شود اعمال شود: کلاسهای درس سنتی و آموزش الکترونیکی. با توجه به تفاوت در منابع داده و اهداف هر یک از این دو نوع سیستم آموزشی، ضرورت دارد که کاربردهای مربوط به اعمال تکنیک های داده کاوی در هر یک از این دو نوع سیستم به صورت جداگانه مورد بررسی قرار بگیرد. تنها کار جدی انجام شده در زمینه بررسی کاربردهای مختلف تکنولوژی داده کاوی در آموزش عالی توسط ،رومرو و ونچرو در سال ۲۰۰۶ تمرکز بر روی تکنیک های وب کاوی و متن کاوی در حوزه آموزش الکترونیکی انجام شده است. در [7] تمرکز اصلی بر روی کاربرد سایر تکنیک های داده کاوی درحوزه آموزش عالی و مرتبط با سیستم های آموزش غیر الکترونیکی می باشد . این مقاله ضمن تحلیل فرایندهای سیستم آموزش دانشگاهها، به بررسی کارهای انجام شده و همچنین قابل انجام در زمینه کاربرد عملیات های مختلف داده کاوی از قبیل خوشه بندی، قوانین انجمنی، دسته بندی و پیش بینی جهت پیش بینی ثبت نام در یک رشته یا درس خاص، بررسی ترکیب واحدهای انتخابی هر دانشجو برای زمانبندی مناسب واحدها، شناخت انواع دانشجویان، تحلیل ماندگاری دانشجویان در ترم های آتی، پیش بینی وضعیت تحصیلی و میزان موفقیت دانشجویان در نیمسال های بعدی پرداخته است. در حال حاضر در اکثر دانشگاه های ایران، بانک های اطلاعاتی وسیعی از ویژگی ها، سوابق آموزشی و تحصیلی دانشجویان موجود است. پژوهش مذکور می تواند راهنمای مفیدی برای استفاده از تکنیکهای داده کاوی بر روی داده های این سیستمها باشد.

با توجه به این که، آموزش و پرورش، با در دست داشتن حجم عظیمی از اطلاعات دانش آموزی، همواره با رشد روز افزون داده همراه می باشد ، همواره بطور جدی مورد توجه مسئولین بوده است. لذا در [3] با استفاده از فرایند داده کاوی از پایگاه داده دانش آموزی و گروه بندی اطلاعات آماری دانش آموزان، با استفاده از خوشه بندی بر اساس مشخصه های ورودی، به استخراج دانش پرداخته شده است ودر نهایت معیارهای موفقیتهای آموزشی در رشته های مختلف تحصیلی با توجه به جنسیت دانش آموزان و محل مدرسه آنها تعیین گردیده است. محدوده داده های مورد بحث این تحقیق، اطلاعات دانش آموزان برای

سالهای ۷۸ تا ۸۱ در رشته های هنرستانهای فنی و حرفه ای و کارودانش میباشد. ایجاد پایگاه اطلاعاتی داده جهت فرایند داده کاوی، پیاده سازی راه حل پروژه داده کاوی،تنظیم منابع داده مورد استفاده، اجرای الگوریتم های کاوش، با استفاده از پایگاه داده مایکروسافت انجام گردیده است. بر اساس نتایج این تحقیق میتوان به نقش دو عامل جنسیت و محل مدرسه در موفقیت تحصیلی دانش آموزان در رشته های مختلف پی برد و مسئولین محترم آموزش و پرورش قادر خواهند بود با ایجاد برنامه های ساختار یافته، به نحو موثرتری جهت پیشبرد اهداف آموزشی و توسعه رشته های خاص در مناطق مختلف تصمیم گیری نمایند.

یکی از چالشهای جدی در مدیریت امور آموزشی دانشگاهها پیش بینی وضعیت تحصیلی دانشجویان در نیمسال های آینده به منظور شناسایی دانشجویانی است که دچار افت تحصیلی شده و ادامه تحصیل آنها با مشکل روبرو خواهد شد. در [6] با استفاده از تکنیک- های داده کاوی وضعیت تحصیلی آتی دانشجویان شامل معدل نیمسال آینده، معدل کل در زمان فارغ التحصیلی، و وضعیت فارغ التحصیلی پیش بینی شده است. برای ساخت مدلهای مورد نظر از تکنیک های مختلفی نظیر شبکه های عصبی، درخت های تصمیم و SVM استفاده گردیده است. این مدلها برای داده های سیستم آموزشی دانشگاه علم و صنعت ایران پیاده سازی شده اند. عملکرد هر یک از مدلها، مورد بررسی قرار گرفته و نتایج بدست آمده با یکدیگر مقایسه گردیده اند. اعتبار سنجی انجام شده بر روی مدلها اثبات می کند که نتایج بدست آمده دقیق و قابل اعتماد بوده اند. با بکارگیری این مدلها، مدیران آموزشی می توانند مشاوره های لازم را برای پیشگیری از رسیدن دانشجویان به وضعیت بحرانی بکار گیرند. همچنین این مدلها می توانند به عنوان یک ابزار پشتیبان تصمیم گیری در سیستم های آموزشی مورد بهره برداری قرار گرفته و نقش مهمی را در ارتقاء سطح علمی دانشگاهها داشته باشند.

در [1] نیز به بررسی نحوه استفاده از روش های طبقه بندی داده که یکی از روش های داده کاوی می باشد ، در غنی کردن سیستمهای آموزش الکترونیک پرداخته شده است. دراین مقاله پیشگویی نتیجه تحصیلی کاربران سیستم آموزش الکترونیک با استفاده از روش های یادگیری ماشین نظیر روشهای ماشین بردار پشتیبان ، درخت تصمیم، بیز ساده و K همسایه نزدیکترین مورد استفاده واقع شده است که در بین این چهار روش، دو روش درخت تصمیم و بیز ساده با ۸۰.۱۱۳ بیشترین دقت را در پیشگویی در میان روشهای استفاده شده داشته اند . روش های ماشین بردار پشتیبان وK همسایه نزدیکترین نیز با دقتهای ۷۸.۴۰۹ و ۷۲.۴۴۳ در مرتبه های بعدی قرار گرفتند.

بعنوان خلاصه، تاکنون مطالعات زیادی جهت پیش بینی نتیجه تحصیلی دانشجویان انجام شده است که هر یک جنبه های مختلفی از عوامل تاثیر گذار را مد نظر داشته اند.

2-۱ توسعه یادگیری الکترونیک مبتنی بر وب

تاکنون مباحث زیادی برای از بین بردن دیوارهای مدرسه در بین نظریه پردازان آموزشی شکل گرفته است. عده ای دیوارهای مدرسه را محدود کننده دانش آموزان می دانند و معتقدند آموزش های رسمی باید خارج از دیوارهای مدرسه و در هر جایی در اختیار دانش آموزان قرار گیرد. این نظریه پردازان فناوری را کلید از بین بردن دیوارها می دانند. آن ها با تکیه به فناوریهای رایانه ای و سیار و همچنین امکانات مکاتبه ای و رسانه های جمعی دیوارهای مدرسه را زاید و مرزهای آموزش را بی انتها تصور می کنند. گروهی دیگر نگاهی معتدل تر به موضوع می اندازند. آن ها معتقدند که مدرسه باید از شکل حاضر خود خارج شود ولی ماهیت خود را به عنوان یک سازمان رسمی از دست ندهد. دانش آموزان باید در چارچوب کلاس های رسمی ولی با فناوری مدرن بیاموزند. به این صورت نه تنها کلاس بلکه همه جای جامعه مکان یادگیری می شود. در این دیدگاه مدیریت متمرکز آموزشی همچنان وجود دارد و دانش آموزان ملزم هستند در کلاس های درس شرکت کنند. دیدگاه سوم که همان دیدگاه گذشته نگر یا سنتی است بر توسعه آموزش در کلاس های سنتی تأکید می کند. مدرسه ها باید تجهیز شوند و کلاس های درس استانداردهای لازم را داشته باشند. آنچه امروز بیشتر کشورها آن را پذیرفته اند دیدگاه دوم است. هیچ کس نمی تواند در مفید بودن مدرسه شک کند زیرا مدرسه جایگاهی است که انسان ها در آن مهم ترین نیاز خود یعنی زندگی اجتماعی رفع می کنند. ارتباط از طریق فناوری هرگز نمی تواند جای ارتباط رودر رو را بگیرد.

استفاده از فناوری های جدید اطلاعی و ارتباطی، نظام های آموزشی را قادرساخته است تا بتوانند آموزش خود را در گستره ی وسیع تری عرضه کنند. یادگیری یا آموزش الکترونیکی با تاکید بر رسانه های الکترونیکی است.

یادگیری الکترونیکی مبتنی بر وب یکی از انواع مهم یادگیری الکترونیکی به شمار می رود که به عنوان یکی از پدیده های دنیای مدرن در عصر اطلاعات و در جامعه ی مبتنی بر دانش پا به عرصه ی وجود گذاشته و در تاریخچه ی کوتاه مدت خود از گسترش قابل ملاحظه ای برخوردار بوده است. [50]

۲-۲ یادگیری الکترونیک چیست؟

یادگیری الکترونیکی آن نوع از یادگیری است که در محیط شبکه و با ابزار های شبکه اتفاق می افتد و مستلزم به کارگیری رسانه های آموزشی همزمان و غیر هم زمان است. هرنوع مراجعه به فضای شبکه که به یادگیری شما کمک کند یادگیری الکترونیکی محسوب می گردد. یادگیری الکترونیکی در پرتو آموزش الکترونیکی تحقق می یابد. آموزش الکترونیکی هر نوع محتوای آموزشی است که به صورت الکترونیکی ارائه شود.

یادگیری الکترونیکی هر نوع آموزشی است که با کامپیوتر به وسیله سی دی رام، اینترنت یا اینترانت با ویژگی های زیر جریان می یابد:

- شامل محتوایی می باشد که مربوط به اهداف یادگیری است .
- از روش های آموزشی مانند مثال ها و تمرین ها برای کمک به فرایند یادگیری استفاده می شود .
- از عناصر چند رسانه ای مانند کلمات و عکس ها برای انتقال محتوا و روش ها استفاده می شود .
- ممکن است که این یادگیری به طور همزمان یا غیر همزمان صورت پذیرد .
- فراهم سازی مهارت ها و دانش های جدیدی که بتوانند به اهداف یادگیری انفرادی مختص خود یادگیرنده مربوط شوند . [8]

۲-۳ انواع یادگیری الکترونیک

به طور کلی یادگیری الکترونیک در سه مقوله جای می گیرد:

یادگیری الکترونیکی مبتنی بر رایانه : در این روش، از رایانه ها و ابزار های مرتبط با آن از قبیل لوح فشرده ،دی وی دی و ویدئو استفاده می شود. در این روش از شبکه ی اینترنت خبری نیست و نوع اولیه ی یادگیری الکترونیک می باشد.

یادگیری مبتنی بر وب (اینترنت) : این روش یادگیری شبیه روش یادگیری مبتنی بر رایانه است. با این تفاوت که از طریق اتصال به شبکه ی جهانی، تبادلات بالا انجام می شود.

یادگیری الکترونیکی مبتنی بر شبکه ی محلی (اینترانت) : این شیوه با روش مبتنی بر وب شباهت هایی دارد و تفاوت آن دو در این است که در روش حاضر به جای شبکه ی جهانی اینترنت، مبادلات یاد شده از طریق اتصال به شبکه های داخلی سازمان ها و شرکت ها (اینترانت) صورت می گیرد.

دو نوع یادگیری مبتنی بر اینترنت و اینترانت خود نیز هر یک ، به دو دسته تقسیم می شوند:

۱) یادگیری الکترونیکی هم زمان: منظور نوعی از یادگیری الکترونیکی است که در آن افراد (اعم از استادان، دانشجویان، متخصصان ، مشاوران و ...) به صورت زنده و هم زمان می توانند با هم ارتباط بر قرار کنند و به صورت چهره به چهره (از طریق شبکه ی رایانه های شخصی) با یکدیگر به تبادل افکار و دیدگاه ها ی ا اطلاعات بپردازند.

۲) یادگیری الکترونیکی نا همزمان: نوعی دیگر از یادگیری الکترونیک است که در آن زمان یا مکان مشخص مطرح نیست و افراد در هر زمان و هر مکانی که بخواهند می توانند وارد شبکه شوند و تبادل اطلاعات کنند. [36]

۲-۴ یادگیری الکترونیک مبتنی بر وب

در نظام آموزشی، شبکه ی گسترده ی جهانی (WWW) به عنوان یکی از مشهور ترین روش های انتقال آموزش از راه دور می باشد. تعداد بسیار زیادی از سایت های اینترنتی، به ویژه برای انتقال آموزش طراحی گشته اند. در ابتدا، صفحات وب بدین منظور طراحی می گشتند که تمامی محتوای دوره ی آموزشی را به یادگیرندگان منتقل کنند. اما امروزه، این صفحات توسط اساتید طراحی می گردند که از این نوع از آموزش به عنوان مکمل آموزشی که در کلاس های درس حضوری انجام می گیرد استفاده می کنند.

دوره های آموزش الکترونیک موفقیت آمیزی که برای محیط های یادگیری مبتنی بر وب طراحی گردیده است، به معنای چیزی بیش از استفاده از اسناد آپلود شده و لینک هایی می باشد که به طور الکترونیکی به یکدیگر وصل شده اند. محتوای دوره ی آموزشی باید طوری طراحی گردد که قابلیت استفاده از طریق رسانه های الکترونیک و تعاملی را داشته باشد تا بتوان از آن طریق انواع متفاوتی از اطلاعات دیداری – شنیداری را به یادگیرندگان عرضه نمود. این محتوا می تواند شامل کلیپ های ویدئویی، انیمیشن، جلوه های صوتی، موسیقی،

تصاویر، نقاشی ها و صفحاتی که ممکن است به صفحات دیگر متصل باشد یا نباشد گردد. دانشجویان در این نوع از یادگیری الکترونیک، می توانند به حل مشکلاتی بپردازند که که در دنیای واقعی با آن ها مواجه می شوند و در نتیجه مسئولیت یادگیری خود را بر عهده گیرند. [63]

۲-۵ ویژگی های یادگیری الکترونیک مبتنی بر وب

یادگیری و تدریس الکترونیکی که مبتنی بر وب است از انعطاف پذیری زیادی برخورد دار می باشد. ویژگی های کلیدی یادگیری الکترونیکی عبارتند از:

تعاملی، چند رسانه ای، سیستم آموزشی باز، امکان دست یابی جهانی، انتشار منابع یادگیری به صورت الکترونیکی، جهانی، یک سانی، منابع یادگیری آنلاین، انتشار در سطح وسیع، تعاملات میان فرهنگی، کسب مهارت در چندین زمینه، یادگیرنده محور، قابلیت دست یابی آسان، خود کفایی یادگیرنده، در بردارنده ی تکالیف اصیل و واقعی، برخوردار از امنیت بالا، محیط یادگیری دوستانه، محیط یادگیری غیر تبعیض آمیز، اثر بخشی بالا، یادگیری مشارکتی، محیط های یادگیری رسمی و غیر رسمی، ارزشیابی از یادگیری به صورت آنلاین، برخورداری از فرهنگ مجازی و ...

یادگیری الکترونیکی دارای مزایا و محاسن بسیاری است که از جمله آن‌ها می‌توان به موارد زیر اشاره نمود :

- امکان ارتقای سریع و موثر سطح علمی دانش‌آموزان؛
- سهولت دسترسی به منابع مختلف آموزشی؛
- امکان دسترسی فراگیر به آموزش در هربیست و چهار ساعت و هفت روز هفته؛
- کاهش هزینه‌های آموزش، امکان ثبت و ضبط فعالیت‌ها و پی‌گیری مستمر پیشرفت تحصیلی؛
- بهره‌گیری از مدل‌ها و روش‌های متنوع آموزش و یادگیری؛
- امکان بهره‌مندی از بهترین معلم‌های کشور با بهره‌گیری از امکانات و تسهیلات فراوان یادگیری الکترونیکی؛

از دو نوع یادگیری الکترونیکی بهره‌گرفته می شود :

یادگیری الکترونیکی به صورت برون خطی (Offline) در قالب سی‌دی‌های آموزشی
یادگیری الکترونیکی به صورت درون خطی (Online) از طریق سایت اینترنتی [8]

۲- ۶ انواع شیوه های به کار گیری یادگیری الکترونیک مبتنی بر وب در نظام آموزشی

نظام آموزشی می تواند به سه شکل از یادگیری الکترونیک مبتنی بر وب استفاده کند که این سه نوع عبارتند از:

۱) استفاده از آموزش مبتنی بر وب به عنوان مکمل آموزش حضوری ۲) استفاده از آموزش مبتنی بر وب به همراه آموزش حضوری ۳) استفاده از آموزش مبتنی بر وب به عنوان جایگزینی برای آموزش حضوری

۲-۷ استراتژی طراحی دوره های آموزشی مبتنی بر وب

به طور کلی، ما پیشنهادی می کنیم که طراحان را برای ارائه محتوای آموزشی درسیستم آموزش الکترونیکی فراهم می سازد. دراین شیوه از آموزش، یادگیرندگان با دریافت محتوای چاپی و آموزش به شیوه آنلاین به یادگیری نائل می شوند . آگاهی از ویژگی ها و مشخصه های این شیوه از آموزش(به ویژه برای یاددهندگان و نیز یادگیرندگانی که این شیوه آموزشی را جهت کسب علم و دانش برگزیده اند) یکی از مولفه های اساسی موفقیت یادگیرندگان راه دور به شمار می آید. اگرچه این شیوه آموزشی می تواند به عنوان مکمل، همراه و یا جایگزینی برای آموزش حضوری تلقی گرددد ، اما .امروزه، آنچه اکثرمتخصصان آموزشی بدان اشاره می نمایند بهره گیری از این شیوه آموزش، در کنار کلاس های درس حضوری جهت نیل به یادگیری موثر وعمیق و کارامد است.

به طور کلی، در طراحی هر دوره ی آموزشی مبتنی بر وب، باید به موارد زیر توجه نماییم: مدیریت و نظارت بر برنامه ی زمانی دوره، توجه به ملزومات اجرایی دوره، اهداف آموزشی و انتظارات دوره، تهیه ی محتوای آموزشی دوره که شامل : مواد نوشتاری، مواد شنیداری، مواد دیداری _ شنیداری، عکس ها و تصاویر

و ... می باشد، برقراری تعامل میان دانشجویان با یکدیگر و با استاد ، فراهم سازی منابع یادگیری اضافی و پشتیبان ، نظارت بر چگونگی پیشترفت دانشجویان در یادگیری (که می تواند این مواد یادگیری توسط خود دانشجویان نیز تهیه گردد.) ، ارزشیابی نهایی از میزان دست یابی دانشجویان به اهداف آموزشی .

۲-۸ ویژگی های اساتید و دانشجویان در یادگیری الکترونیک مبتنی بر وب

۲-۸-۱ ویژگی های استادان:

این آمادگی را کسب کنید که به دانشجویان یاد بدهید که چگونه با یکدیگر به طور آنلاین به تعامل بپردازند، برای عملکرد دانشجویان بازخورد اطلاعاتی ارائه دهید و پیشنهادات اصلاحی خود را برای آن ها فراهم سازید، از این امر اطمینان یابید که دانشجویان شما می توانند هم از طریق فکس و هم به صورت آنلاین در زمان های مشخصی به شما دسترسی داشته باشند و راهنمایی های لازم را از شما دریافت دارند. این امر در مراحل اولیه ی یادگیری الکترونیک از اهمیت بالایی برخور دار است. به گفته هایی که دانشجویان با یکدیگر رد و بدل می کنند گوش فرا دهید و آن ها را به این امر تشویق کنید که با یکدیگر به طور مشارکتی به یادگیری بپردازند. در ابتدای دوره ی آموزشی، اهداف کلی، جرئی و عینی را برای آن ها به طور روشن مشخص سازید تا آن ها بدانند که در پایان دوره ی آموزشی از آنها چه انتظاراتی می رود. سعی کنید مشکلات احتمالی که ممکن است دانشجویان با آنها رو به رو شوند را تشخیص دهید و نحوه ی برطرف کردن آنها را بدانید. در صورت امکان، برای دانشجوان کنفرانس های آنلاین برگزار نمایید. از روش ها و فنون تدریس و یادگیری متنوع استفاده نمایید. در صورتی که شما به طور مداوم از روش سخنرانی استفاده نمایید دانشجویان رغبت خود را برای ادامه ی یادگیری از دست می دهند اما اگر شما در آموزش خود از روش های متنوع تری استفاده نمایید می توانید اثر بخشی و کارآیی آموزش خود را تضمین نمایید.

۲-۸-۲ ویژگی های دانشجویان:

در ابتدای شروع دوره ی آموزشی، از این امر اطمینان حاصل کنید که شما توانایی کار کردن با تکنولوژی های به کار رفته در این نوع آموزش را دارید. بدانید که چگونه باید به راهنمایی آنلاین و همچنین راهنمایی کننده ی آنلاین دسترسی یابید. از این امر اطمینان حاصل کنید که برنامه ی مرور گر اینترنت شما از ویژگی های لازم برای دریافت تمامی مواد آموزشی دوره برخوردار است. همچنین بدانید که در صورت وقوع مشکل، از ویژگی های مرورگر خود به نحوی استفاده نمایید که آن مشکلات رفع گردند. یک حساب صندوق پست الکترونیکی برای خود ایجاد کنید. بدانید که چگونه باید فایل ها و اسناد را در بین سایر دانشجویان و استاد خود تبادل نمایید. از اهداف کلی، جرئی و عینی دوره ی آموزشی اطلاع یابید. اگر این اهداف برای شما ابهام دارند برای رفع این ابهام در مورد آنها از استاد خود سوال کنید. در صورتی که فکر می کنید نسبت به سایر دانشجویان عقب افتاده اید یا اینکه محتوای دروه ی آموزشی را به خوبی درک نمی کنید فورا با استاد خود تماس حاصل نمایید. به فرهنگ ارتباطات اینترنتی واقف باشید. این را بدانید که ایجاد تعامل از طریق متن در یک محیط اینترنتی بسیار مشکل تر از تعاملی است که در کلاس های درسی چهره به چهره انجام می پذیرد. مانند این که شما در محیط وب نمی توانید از زبان بدن برای درک معانی گفته ها استفاده نمایید. کلماتی را که در این نوع ارتباط به کار می برید باید واضح و روشن باشد و سوء تفاهم ها را به حد اقل میزان ممکن کاهش دهد. به این امر واقف باشید که کلاس درس آنلاین ، زمانی مشابه یا حتی بیشتر از کلاس های درس سنتی برای پرداختن به یادگیری نیاز دارد. سعی کنید که آن زمان مورد لزوم را در اختیار داشته باشید.

امروزه تکنولوژی های جدید اطلاعاات و ارتباطات نقش مهمی درارائه آموزش و یادگیری برعهده دارد که البته به طرق مختلفی از آن ها یاد می شود.یکی از این فناوری ها ، آموزش و یادگیری الکترونیکی مبتی بر وب است.این شیوه از آموزش، امکان استفاده از صفحات وب را برای ارائه محتوای آموزشی درسیستم آموزش الکترونیکی فراهم می سازد. دراین شیوه از آموزش، یادگیرندگان با دریافت محتوای چاپی و آموزش به شیوه آنلاین به یادگیری نائل می شوند .آگاهی از ویژگی ها و مشخصه های این شیوه از آموزش(به ویژه برای یاددهندگان و نیز یادگیرندگانی که این شیوه آموزشی را جهت کسب علم و دانش برگزیده اند) یکی از مولفه های اساسی موفقیت یادگیرندگان راه دور به شمار می آید. اگرچه این شیوه آموزشی می تواند به عنوان مکمل، همراه و یا جایگزینی برای آموزش حضوری تلقی گرددد ، اما امروزه، آنچه اکثرمتخصصان آموزشی بدان اشاره می نمایند بهره گیری از این شیوه آموزش، در کنار کلاس های درس حضوری جهت نیل به یادگیری موثر وعمیق و کارامد است. [37]

۲-۹ تحلیل آموزشی دیک-کری

دیک، کری و کری در سال ۲۰۰۱ نوشته اند که فرایند شناسایی مهارت ها و دانشی که باید آموزش دربرگیرد پیچیده است. آن ها این فرایند را تحلیل آموزشی می نامند:" تحلیل آموزشی مجموعه ای از روندهاست که وقتی یک هدف آموزشی به کاربسته می شود، موجب شناسایی گام های مرتبط برای تحقق هدف و مهارت های مادون مورد نیاز برای دانش آموز تا به هدف برسند می شود". آنها این فرایند را به دوقسمت تقسیم می کنند . در قسمت اول، طراح آموزشی اجزای اصلی هدف آموزشی را بواسطه تحلیل هدف تعیین می کند .قسمت دوم این را شامل می شود که چگونه هر گام از هدف آموزشی می تواند بیشتر تحلیل شود تا جایی که مهارت های مادونی که یادگیرندگان باید در جهت مواجه با هدف های آموزشی داشته باشند شناسایی شوند. فرایند تحلیل آموزشی می تواند با هدف های آموزشی که شناسایی شده اند آغاز شود. دیک و دیگران در سال ۲۰۰۱ نوشته اند که طراح آموزشی باید فرایند تحلیل آموزشی را با این سئوال آغاز کند که "دقیقاً یادگیرندگان چه باید انجام دهند تا نشان دهند که از قبل می توانند هدف را تحقق بخشند". این سوال به جای تمرکز صرف بر روی محتوای آموزش، بر روی آنچه یادگیرندگان باید قادر باشند انجام دهند جهت مشارکت در آموزش تأکید می کند. این تمرکز روند تحلیل هدف خوانده می شود. نتیجه تحلیل هدف طبقه بندی نوع یادگیری که اتفاق خواهد افتاد و یک ارائه بصری مثلاً فلوچارت که گام های ویژه و گام های فرعی را که یادگیرنده جهت رسیدن به هدف های آموزشی باید بردارد نشان می دهد. روندهای تحلیل هدف گوناگونی در دسترس طراح آموزشی است. قسمت دوم از فرایند تحلیل آموزشی به تحلیل مهارت های مادون منسوب است. هدف از این تحلیل شناسایی مجموعه مناسبی از مهارت های فرعی است که یادگیرنده به منظور انجام یک گام ویژه نیاز خواهد داشت تا به او در مواجه با هدف آموزشی کمک کند .دیک و دیگران فنون مختلفی را برای اجرای تحلیل مهارت های مادون شناسایی کرده اند، نظریه سلسله مراتبی و تحلیل خوشه ای دو نمونه از این فنون هستند. طراح آموزشی باید فنی ویژه را بر مبنای نوع دانشی که در مواجه با هدف های آموزشی مورد نیاز است انتخاب کند . [9],[62]

۲-۱۰ یادگیری سیار : نیاز امروز جهان

یادگیری سیار به عنوان راه حلی برای برطرف سازی چالش های امروز بسیاری از جوامع در نظر گرفته می شود. اروپا از این فناوری جهت بهبود وضعیت جوانان کم سواد استفاده کرده است. آفریقایی ها در پی حل مشکلات بهداشتی در کشورهای خود با بهره گیری از یادگیری سیار هستند، بیماری ایدز و مالاریا از جمله مهم ترین مواردی هستند که افریقایی ها فعالیت هایی را از طریق یادگیری سیار برای آگاه سازی مردم و مقابله با آن ها آغاز کرده اند . آسیایی ها از جمله کره ای ها بر یادگیری رسمی با بهره گیری از فناوری های سیار تاکید نموده اند و استرالیایی ها حتی در پی این هستند که خدمات دانشجویی خود را از طریق فناوری های سیار کنترل کنند. در کشور ما تاکنون فعالیت های بسیار اندکی در زمینه یادگیری سیار صورت گرفته است. یادگیری سیار در ایران نیز می تواند در کنار یادگیری الکترونیکی مشکل گشا باشد. ولی ابتدا لازم است حداقل امکانات نرم افزاری و سخت افزاری برای آن فراهم شود که این حداقل مستلزم یک سرمایه گذاری وسیع است. امروز این پتانسیل در کشور ما وجود دارد که از یادگیری سیار برای آموزش های عمومی مثل آموزش فرهنگ شهروندی، راهنمایی و رانندگی، آموزش حجاج و بسیاری موارد دیگر استفاده شود. همچنین در زمینه آموزش رسمی نیز اکنون دانشگاه های ما با توسعه سیستم های آموزش مجازی دارند به نقطه ای می رسند که کم کم با احساس کمبودهای یادگیری الکترونیکی به فکر یادگیری سیار بیفتند.

۲-۱۰-۱ یادگیری سیار در کره جنوبی

شون در سال ۲۰۰۶ در مطالعه ای به بررسی پروژه یادگیری سیار در کره جنوبی پرداخت. او در بررسی های اولیه خود بیان داشت که امروزه کشورهای فرانسه، کره جنوبی و آلمان تمایل زیادی به توسعه سیستم های یادگیری سیار دارند. پروژه یادگیری سیار کره جنوبی در دو سطح آموزش رسمی و غیر رسمی، تحت پشتیبانی کمیته اس. سی. ۳۶ کره جنوبی در اپریل ۲۰۰۶ تکمیل شد. مدیریت این پروژه را کمیته ای ۱۰ نفری شامل ۵ دانشمند، ۲ پژوهشگر و ۳ عضو صنعتی تشکیل می دادند. اهداف اصلی این پروژه: تحلیل روند توسعه فناوری یادگیری سیار، تحلیل نیازهای موجود در زمینه استانداردسازی فناوری اطلاعات و ارتباطات برای یادگیری سیار

و در نهایت فراهم سازی استانداردهای لازم بود. شون از جمله چالش های موجود در سر راه یادگیری سیار در کره جنوبی را شامل محدودیت های سخت افزاری و نرم افزاری، تغییر پهنای باند هنگام اتصال و قطع مکرر و پیش بینی نشده اتصال به شبکه می دانست. کره جنوبی برای توسعه یادگیری سیار برای کودکان، پروژه یو-لرنینگ را از سال ۲۰۰۴ تحت سرپرستی موهرد و با همکاری شرکت های کریس، کی. ای، ام. اس کره جنوبی و اینتل کره آغاز نمود. برای اجرای این پروژه ۹ مدرسه به صورت آزمایشی انتخاب شدند. دانش آموزان این مدارس از فناوری رایانه های قابل حمل و پی.دی.ای جهت دریافت مواد درسی و ارتباط با همکلاسی ها و معلم استفاده می‌کردند. اهداف اصلی این پروژه شامل شخصی سازی یادگیری، بهره گیری از فناوری های سیار در آموزش، بالا بردن سطح توانایی خودرهبری در یادگیرندگان، افزایش عملکرد تحصیل دانش آموزان با بهره‌گیری از مواد ویدیویی سفارشی و شاد نمودن محیط مدرسه اعلام شد. تحت پروژه یادگیری سیار، پروژه دیگری نیز در سطح آموزش عالی در کره جنوبی به اجرا گذاشته شد. در این پروژه که یو- کامپوس نام داشت ۶۴ دانشگاه مشارکت داشتند. سرویس شناسه سیار، کتابخانه سیار، برقراری خدمات مدیریتی، اجتماعی و خصوصی برای دانشجویان از جمله مواردی بودند که مورد توجه کنسرسیوم دانشگاه سیار کره جنوبی قرار گرفت. در حال حاضر در بیش از ۲۰ دانشگاه کره جنوبی خدمات یادگیری سیار فراهم شده است، که از جمله آن‌ها می توان به دانشگاه های یونسی و دنگسو اشاره نمود. در زمینه یادگیری غیر رسمی در کره جنوبی نیز، تحت پروژه یادگیری سیار برنامه های متنوعی تعریف و اجرا شده است. [28]

۲-۱۰-۲ پروژه توسعه دانشگاه مجازی عرب

توسعه دانشگاه مجازی عرب تاحدود زیادی به پروژه ابوعلی سینا مربوط می شود. پروژه ابوعلی سینا از جمله پروژه‌های جالبی است که تاکنون در یونسکو انجام شده، و ۱۵ کشور حاشیه دریای مدیترانه در این پروژه مشارکت می کنند. این کشورها با همکاری یکدیگر یک ابر دانشگاه مجازی را به وجود آورده اند. اتحادیه اروپا در حدود ۲ هزار میلیون یورو در پروژه ابوعلی سینا سرمایه گذاری کرده است. دانشگاه مجازی عرب یک مؤسسه خصوصی غیرانتفاعی است که عمده سهام تأسیس آن توسط دانشگاه باز بریتانیا تأمین شده است. این دانشگاه فعالیت خود را در اکتبر۲۰۰۲آغاز نمود. شعبه های منطقه ای دانشگاه مجازی عرب بر طبق برنامه ریزی های صورت گرفته در کشورهای عربستان، بحرین، اردن، لبنان، مصر و کویت تأسیس خواهد شد. در دانشگاه مجازی

عرب واحدهایی در زمینه مهارت در زبان عربی و انگلیسی، مهارت های پایه در ریاضیات، مهارت های استفاده از اینترنت و آموزش برخط، دوره هایی در علوم انسانی و علوم اجتماعی و محیط ارائه می شود. دانشگاه مجازی عرب قصد دارد تنوعی از رسانه های آموزشی مدرن و متناسب شامل مواد چاپی، ویدیویی، لوح فشرده و پشتیبانی های مبتنی بر اینترنت را توسعه دهد. در مجموع بیشتر پشتیبانی های آموزشی، بوسیله مراکز یادگیری از طریق شبکه ماهواره اختصاصی فراهم می شود. در دانشگاه مجازی عرب ترکیبی از مطالعه مستقل و کلاس های درسی تحت حمایت معلم به عنوان یک قالب عمومی برنامه ریزی شده است. این دانشگاه هماهنگ با یونسکو، بر روی ساخت شبکه ارتباطات ماهواره ای کار می کند. این شبکه می‌تواند تمام شعبه های دانشگاه را در کشورهای مختلف عربی به هم پیوند زند. در مجموع شبکه ارتباطات ماهواره ای این امکان را به وجود می آورد که کنفرانس های تحت وب بطور همزمان در تمامی شعبه های دانشگاه دریافت و در صورت نیاز توسط دانشجویان برای مرور مجدد ضبط شود. در دانشگاه مجازی عرب دانشجویان موظف هستند پس از پایان دوره در امتحانات نهایی این دانشگاه شرکت کنند. این آزمون در همه شعب دانشگاه بطور همزمان و زیر نظر اداره سرپرستی دانشگاه در کویت برگزار می شود. نمره آزمون نهایی حدود ۵۰ درصد نمره دانشجویان را تشکیل خواهد داد و۵۰ درصد باقیمانده به فعالیت های دانشجویان در طول نیمسال تحصیلی، شامل نمرات آزمون های کلاسی و تکالیف انجام شده، اختصاص داده می شود. برنامه آموزشی معلمان قبل از آغاز ترم تحصیلی سازمان داده می شود [51] ,[47],[16] ,[15].

فصل سوم

مروری بر تحقیقات انجام شده

۳-۱ توصیف داده ها در داده کاوی

۳-۱ -۱ خلاصه سازی و به تصویر در آوردن داده ها

قبل از اینکه بتوان روی مجموعه ای از داده ها ،داده کاوی انجام بدهیم و یک مدل پیش بینی مناسب ابجاد کنیم ، باید بتوان داده ها را به خوبی شناخت که برا ی شروع این کار می توان از پارامترهایی مثل میانگین , انحراف معیار و.... استفاده کنیم.

ابزارهای تصویرسازی داده ها و گراف سازی برای شناخت داده ها بسیار مفید می باشند و نقش آنها در آماده سازی داده ها بسیار مفید و غیر قابل انکار است ، مثلا با استفاده از این ابزار می توان توزیع مقادیر مختلف داده ها را در یک نمودار مشاهده کرد و میزان داده های دارای خطا را به طور تقریبی حدس زد.

مهمترین مشکل این ابزار این است که معمولا تحلیل ها دارای تعداد زیادی پارامتر هستند که به هم مربوطند و باید رابطه این پارامترها را که چند بعدی می باشد در دو بعد نمایش دهند که این کار اگر هم عملی باشد برای استفاده از آنها نیاز به افراد خبره می باشد. [18]

۳-۱-۲ خوشه بندی (Clustering)

هدف از خوشه بندی این است که داده های موجود را به چند گروه تقسیم کنند و در این تقسیم بندی داده های گروه های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده های موجود در یک گروه باید بسیار به هم شبیه باشند . [42]

برخلاف کلاس بندی (که در ادامه خواهیم دید) در خوشه بندی ، گروه ها از قبل مشخص نمی باشند و همچنین معلوم نیست که بر حسب کدام خصوصیات گروه بندی صورت می گیرد. درنتیجه پس از انجام خوشه بندی باید یک فرد خبره خوشه های ایجاد شده را تفسیر کند و در بعضی مواقع لازم است که پس از بررسی خوشه ها بعضی از پارامترهایی که در خوشه بندی در نظر گرفته شده اند ولی بی ربط بوده یا اهمیت چندانی ندارند حذف شده و جریان خوشه بندی از اول صورت گیرد. [42],[39]

پس از اینکه داده ها به چند گروه منطقی و توجیه پذیر تقسیم شدند از این تقسیم بندی می توان برای کسب اطلاعات در مورد داده ها یا تقسیم داده ها جدید استفاده کنیم.

از مهمترین الگوریتم هایی که برای خوشه بندی استفاده می شوند می توان Kohnen و الگوریتم K-means را نام برد. [24]

۳-۱-۳ تحلیل لینک (Link Analysis)

تحلیل داده ها یکی از روش های توصیف داده هاست که به کمک آن داده ها را بررسی کرده و روابط بین مقادیر موجود در بانک اطلاعاتی را کشف می کنیم. از مهمترین راههای تحلیل لینک کشف وابستگی (Association discovery) و کشف ترتیب (Sequence discovery) می باشد. [24],[23]

منظور از کشف وابستگی یافتن قوانینی در مورد مواردی است که با هم اتفاق می افتند مثلا اجناسی که در یک فروشگاه احتمال خرید همزمان آنها زیاد است.

کشف ترتیب نیر بسیار مشابه می باشد ولی پارامتر زمان نیز در آن دخیل می باشد.

۳-۲ مدل های پیش بینی داده ها

۳-۲-۱ طبقه بندی (Classification)

در مسائل classification هدف شناسایی ویژگیهایی است که گروهی را که هر مورد به آن تعلق دارد را نشان دهند. از این الگو می‌توان هم برای فهم داده‌های موجود و هم پیش‌بینی نحوه رفتار مواد جدید استفاده کرد. داده‌کاوی مدلهای classification را با بررسی داده‌های دسته‌بندی شده قبلی ایجاد می‌کند و یک الگوی پیش‌بینی کننده را بصورت استقرایی می‌یابند. این موارد موجود ممکن است از یک پایگاه داده تاریخی آمده باشند. [42],[24]

۳-۲-۲ رگرسیون (Regression)

واژه رگرسیون در فرهنگ لغت به معنی بازگشت است و اغلب جهت رساندن مفهوم "بازگشت به یک مقدار متوسط یا میانگین"به کار می رود. بدین معنی که برخی پدیده ها به مرور زمان از نظر کمی به طرف یک مقدار متوسط میل می کنند.

بیش از ۱۰۰ سال پیش در سال ۱۸۷۷ فرانسیس گالتون (Francis Galton) در مقاله ای که در همین زمینه منتشر کرد اظهار داشت که متوسط قد پسران دارای پدران قد بلند ، کمتر از قد پدرانشان می باشد . به نحو مشابه متوسط قد پسران دارای پدران کوتاه قد نیز بیشتر از قد پدرانشان گزارش شده است. به این ترتیب گالتون پدیده بازگشت به طرف میانگین را در داده هایش مورد تأکید قرار داد . برای گالتون رگرسیون مفهومی زیست شناختی داشت اما کارهای او توسط کارل پیرسون (Karl Pearson) برای مفاهیم آماری توسعه داده شده . گرچه گالتون برای تأکید بر پدیده بازگشت به سمت مقدار متوسط از تحلیل رگرسیون استفاده کرد، اما به هر حال امروزه واژه تحلیل رگرسیون جهت اشاره به مطالعات مربوط به روابط بین متغیرها به کار برده می شود.

Regression از مقادیر موجود برای پیش‌بینی مقادیر دیگر استفاده می‌کند. در ساده‌ترین فرم، regression از تکنیک‌های آماری استاندارد مانند linear regression استفاده می‌کند. متاسفانه، بسیاری مسائل دنیای واقع تصویرخطی ساده‌ای از مقادیر قبلی نیستند. بنابراین تکنیک‌های پیچیده‌تری(logistic regression، درخت‌های تصمیم، یا شبکه‌های عصبی) ممکن است برای پیش‌بینی مورد نیاز باشند.

در تحقیقاتی که از تحلیل رگرسیون استفاده می شود، هدف معمولا پیش بینی یک یا چند متغیر ملاک از یک یا چند متغیر پیش بین است. چنانچه هدف پیش بینی یک متغیر ملاک از چند متغیر پیش بین باشد از مدل رگرسیون چندگانه استفاده می‌شود. در صورتی که هدف، پیش بینی همزمان چند متغیر ملاک از متغیرهای پیش بین یا زیر مجموعه ای از آنها باشد از مدل رگرسیون چند متغیری استفاده می‌شود. در تحقیقات رگرسیون چندگانه هدف پیدا کردن متغیرهای پیش بینی است که تغییرات متغیر ملاک را چه به تنهایی و چه مشترکا پیش بینی کند. ورود متغیرهای پیش بین در تحلیل رگرسیون به شیوه های گوناگون صورت می‌گیرد. در این جا سه روش اساسی مورد بحث قرار می‌گیرد:

- روش همزمان
- روش گام به گام
- روش سلسله مراتبی

در روش همزمان تمام متغیرهای پیش بین با هم وارد تحلیل می‌شود. در روش گام به گام اولین متغیر پیش بین بر اساس بالاترین ضریب همبستگی صفرمرتبه با متغیر ملاک وارد تحلیل می‌شود. از آن پس سایر متغیرها پیش بین بر حسب ضریب همبستگی تفکیکی (جزئی) و نیمه تفکیکی (نیمه جزئی) در تحلیل وارد می‌شود. در این روش پس از ورود هر متغیر جدید ضریب همبستگی نیمه تفکیکی یا تفکیکی ، تمام متغیرهایی که قبلا در معادله وارد شده اند به عنوان آخرین متغیر ورودی مورد بازبینی قرار می‌گیرد و چنانچه با ورود متغیر جدید معنی داری خود را از دست داده باشد، از معادله خارج می‌شود. به طور کلی در روش گام به گام ترتیب ورود متغیرها در دست محقق نیست.

در روش سلسله مراتبی ترتیب ورود متغیرها به تحلیل بر اساس یک چارچوب نظری یا تجربی مورد نظر محقق صورت می‌گیرد. به عبارت دیگر پژوهشگر شخصا درباره ترتیب ورود متغیرها به تحلیل تصمیم گیری می‌کند. این تصمیم گیری که قبل از شروع تحلیل اتخاذ می‌شود می‌تواند بر اساس سه اصل عمده زیر باشد:

- رابطه علت و معلولی.

- رابطه متغیرها در تحقیقات قبلی.

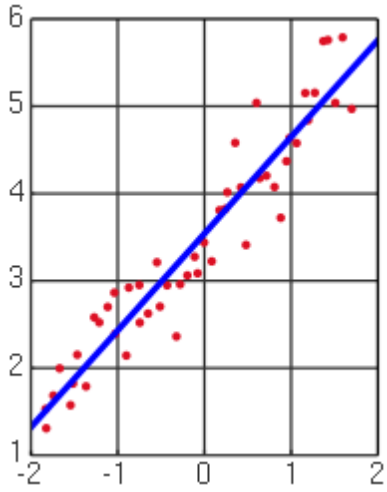
۳ – ۲ – ۲ – ۱ رگرسیون خطی

در رگرسیون خطی، متغیر وابسته y_i ترکیب خطی از ضرایب (پارامترها) است (لازم نیست که نسبت به متغیرهای مستقل خطی باشد). مثلاً تحلیل رگرسیونی ساده زیر با N نقطه، متغیر مستقل x_i و ضرایب β_0 و β_1 خطی است:

معادله ۳ -۱ (خط راست):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, N$$

شکل ۳ -۱ نمودار یک رگرسیون خطی با یک متغیر مستقل را نمایش می دهد.



شکل ۳ - ۱ : رگرسیون خطی با یک متغیر مستقل

۳ – ۲ – ۲ – ۲ رگرسیون چندگانه

در رگرسیون چندگانه، بیش از یک متغیر مستقل وجود دارد:

معادله ۳ – ۲ (سهمی) : $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,\ \ i = 1, \ldots, N$

این همچنان رگرسیون خطی است، زیرا y_i همچنان ترکیب خطی پارامترها (β_0 و β_1) است، هرچند که نسبت به متغیر مستقل (x_i) خطی نیست.

در هر دو حالت، ε_i مقدار خطاست و پانویس i شماره هر مشاهده (هر جفت x_i و y_i) را نشان می‌دهد. با داشتن مجموعه‌ای از این نقطه‌ها می‌توان مدل را به دست آورد:

معادله ۳ – ۳ : $$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i$$

معادله ۳ – ۴ (عبارت e_i مانده نام دارد): $$e_i = y_i - \widehat{y}_i$$

روش رایج برای به‌دست‌آوردن پارامترها، روش کمترین مربعات است. در این روش پارامترها را با کمینه‌کردن تابع زیر به دست می‌آورند:

معادله ۳ – ۵ : $$\mathrm{SSE} = \sum_{i=1}^{N} e_i^2$$

در مورد رگرسیون ساده، پارامترها با این روش برابر خواهند بود با:

معادله ۳ – ۶ : $$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

معادله ۳ – ۷ : $$\hat{\beta}_0 = \bar{y} - \widehat{\beta_1}\bar{x}$$

که در آن $\bar{x}$ و $\bar{y}$ میانگین x و y هستند.

انواع مدل یکسانی را می‌توان هم برای regression و هم برای classification استفاده کرد. برای مثال الگوریتم درخت تصمیم CART را می‌توان هم برای ساخت درخت‌های classification و هم درخت‌های

regression استفاده کرد. شبکه‌های عصبی را نیز می‌توان برای هر دو مورد استفاده کرد. [35],[40],[41],[46]

۳-۲-۳ سری های زمانی (Time series)

پیش‌بینی های Time series مقادیر ناشناخته آینده را براساس یک سری از پیش‌بینی- گرهای متغیر با زمان پیش‌بینی می‌کنند و مانند regression، از نتایج دانسته شده برای راهنمایی پیش‌بینی خود استفاده می‌کنند. مدلها باید خصوصیات متمایز زمان را در نظر گیرند و بویژه سلسله‌مراتب دوره‌ها را. یک سری زمانی مجموعه مشاهداتی است که بر حسب زمان مرتب شده است به عبارت دیگر می توان گفت یک سری زمانی عبارت است از سری داده هایی که از مشاهده یک پدیده در طول زمان بدست آمده اند.

ما یک سری زمانی را به عنوان دنباله ای از مشاهدات بر روی یک متغیر مورد توجه در نظرمی گیریم. متغیر در نقاط گسسته ای از زمان که معمولا فاصله های مساوی دارند، مشاهده می شود.

در یک تقسیم بندی کلی سری های زمانی را به پیوسته و گسسته تقسیم می کنیم. یک سری زمانی را پیوسته گوییم هرگاه مشاهدات به طور پیوسته در زمان ایجاد شده باشند. یک سری زمانی را گسسته گوییم هرگاه مشاهدات فقط در زمانهای معینی که معمولا به فواصل مساوی از یکدیگر قرار دارند اخذ شده باشند.

هدف از تجزیه و تحلیل سری زمانی : به طور خلاصه می توان دو هدف زیر را برای تجزیه و تحلیل سری های زمانی بر شمرد:

۱ – کشف و شناسایی مدل احتمالی مولد داده ها

۲ – پیش بینی مقادیر آینده

در یک سری زمانی با بررسی رفتار گذشته سری مدل احتمالی که می تواند مولد داده ها باشد را شناسایی کرده و سپس با فرض اینکه داده ها در آینده نیز رفتاری مشابه خواهند داشت و از مدل برازش داده تبعیت خواهند نمود، سعی می کنیم مقادیر آینده سری را پیش بینی کنیم.

۳ – ۲ – ۳ – ۱ نمودار سری زمانی

اولین گام در تجزیه و تحلیل یک سری زمانی رسم نمودار آن سری می باشد. از این نمودار می تـوان اطلاعـات مفیدی در مورد طبیعت داده ها بدست آورد. برای رسم نمودار سری زمانی، مشـاهدات یـک سـری زمـانی را در برابر زمان وقوع آنها رسم می کنیم.

۳ – ۲ – ۳ – ۲ اجزاء یک سری زمانی

بسیاری از محققین برای توصیف رفتار یک سری زمانی ، اجزا زیر را برای سری زمانی در نظر می گیرند:

۱. روند یا تمایل بلند مدت(trend)
۲. تغییرات دوره ای(cyclical variation)
۳. تغییرات فصلی(seasonal variation)
۴. تغییرات نامنظم(irregular variation)

به اختصار به معرفی این اجزا می پردازیم:

- روند(T)

روند یا تمایل بلند مدت عبارت است از تحول متغیر مورد مطالعه در یـک دوره طـولانی بـدون در نظـر گـرفتن تغییرات دوره ای، فصلی و نا منظم، به عبارت دیگر می توان گفت روند عبارتست از حرکات رو بـه بـالا و پـایین یک سری زمانی که نشان دهنده کاهش یا افزایش بلند مدت یک سری زمانی است.

- تغییرات دوره ای یا سیکل (C)

تغییرات دوره ای عبارت است از تکرار حرکات رو به بالا و پایین حول سطوح روند. این نوع تغییـرات دارای دوره نوسان بیشتر از یک سال می باشد. نوسانات دوره ای ممکن است دقیقا از طرح های مشابهی بعد از فواصل زمانی مساوی پیروی کنند ولی همیشه این طور نیست. یک دوره کامل را که معمولا ۷ تا ۹ سال طول می کشد اصطلاحا یک "دوره" می نامند. یکی از معمولی ترین نوسانات سیکلی داده های سری زمـانی ، سـیکل تجـاری است . سیکل تجاری وقوع مکرر دوره های رونق و رکود است .

- تغییرات فصلی(S)

تغییرات فصلی تغییراتی هستند که در دوره های تناوبی کوتاه پیش می آیند. این تغییرات مربوط به عواملی هستند که به طریقی منظم و چرخه ای روی یک دوره کمتر از یک سال عمر می کند. در واقع تغییرات فصلی رفتار دوره ای متغیر را نشان می دهد، یعنی رفتاری که معمولا هر سال در همان فصل تقریبا با همان شدت روی می دهد.

- تغییرات نامنظم(I)

تغییرات نامنظم عبارت است از حرکات پراکنده در یک سری زمانی که از الگوی منظم و مشخصی پیروی نمی کنند. در واقع این حرکات بیان می کنند که پس از محاسبه روند، تغییرات دوره ای و تغییرات فصلی چه چیز دیگری در سری زمانی بجا می ماند.

نوسانات نامنظم معمولا ناشی از وقایع غیر معمولی هستند که قابل پیش بینی نیستند. مانند زمین لرزه، اعتصاب، طوفان، جنگ، تصادفات و ...

۳ – ۲ – ۳ – ۳ همبستگی بین مشاهدات سری زمانی

نظریه آمار بیشتر در مورد نمونه های تصادفی که از مشاهدات مستقل ناشی شده اند بحث می کند، اما در سری های زمانی ویژگی مهم این است که معمولا مشاهدات متوالی مستقل نیستند و دقیقا این وابستگی است که می خواهیم آن را بررسی کنیم و به مدل در آوریم. برای بررسی این وابستگی از تابع خود همبستگی و تابع خود همبستگی جزئی استفاده می کنیم.

تعریف خود همبستگی در تاخیر K: عبارت است از همبستگی بین مشاهداتی که K واحد زمانی با یکدیگر فاصله دارند. تابع خود همبستگی نظری که آن را با ρ_k نشان می دهیم، به شکل زیر تعریف می شود:

معادله ۳ – ۸ :
$$\rho_k = \frac{\mathrm{cov}(x_t, x_{t+k})}{\mathrm{var}(x_t)}$$

تعریف ضریب خود همبستگی جزئی:همبستگی بین X_{t+k} و X_t بعد از حذف اثر متغیرهای $x_{t+1}, x_{t+2}, \dots, x_{t+k-1}$ را ضریب خود همبستگی جزئی می نامند.

۳ – ۲ – ۳ – ۴ مدل سازی سری های زمانی به روش باکس – جنکینز(ARIMA)

مدل های ARIMA برای توصیف رفتار بسیاری از سری های زمانی مفید می باشند. برای ساختن یک مدل ARIMA از یک روش سه مرحله ای تکراری استفاده می شود. به این ترتیب که ابتدا یک مدل آزمایشی از طبقه مدلهای ARIMA از طریق تجزیه و تحلیل داده های تاریخی مشخص می شود. سپس پارامترهای نامعلوم مدل تخمین زده می شود. در نهایت آزمون خطا اجرا می شود تا شایستگی مدل را تعیین کند. چنانچه مدل مورد تائید قرار گرفت می توان آن را مبنای پیش بینی رفتار آینده سری قرار داد.

۳ – ۲ – ۳ – ۵ استراتژی مدل سازی

مدل سازی یک سری زمانی بطور کلی شامل سه مرحله می باشد :

۱. تشخیص مدل آزمایشی
۲. تخمین پارامترهای مدل (برازش مدل)
۳. بررسی مناسب بودن مدل

۳ – ۲ – ۳ – ۶ تشخیص مدل آزمایشی

مرحله اول: رسم acf و pacf نمونه ای

مرحله دوم: آزمون وجود روند قطعی در مدل

جدول ۳-۱ : مدل های آزمایشی سری زمانی

	ACF	PACF
AR(p)	بصورت یک تنزل نمائی یا موج سینوسی میرا به سمت صفر میل می کند .	بعد از تاخیر P قطع می شود .
MA(q)	بعد از تاخیر q قطع می شود .	به صورت یک تنزل نمائی یا موج سینوسی به سمت صفر میل می کند .
ARMA(p,q)	بعد از تاخیر (p-q) به سمت صفر میل می کند .	بعد از تاخیر (p-q) به سمت صفر میل می کند .

با توجه به جدول ۳-۱ مدل اولیه اجرا می شود و به معنی داری پارامتر ها توجه می شود. در صورتی که پارامترهای اولیه معنی دار باشد می توان مدل های جامعتر را آزمون کرد . همچنین به جمله ثابت(constant) توجه می کنیم در صورتی که معنی دار نباشد مدل را بدون جمله ثابت برازش می کنیم. بنابراین با آزمون و خطا با کاهش و افزایش پارامترهای مدل ، مدل مناسب را پیدا می کنیم. برای بررسی مناسب بودن مدل از بررسی باقی مانده ها (اختلاف مقدار واقعی و مقدار پیش بینی شده) استفاده می کنیم مثلا اگر مدل مناسب باشد باقی مانده ها باید توزیع نرمال داشته باشند و همچنین ضریب خود همبستگی و ضریب خود همبستگی جزئی باقی مانده ها در هیچکدام از تاخیر ها معنی دار نباشد. حال اگر یک مدل مناسب برای سری زمانی خود پیدا کرده باشیم می توانیم مقادیر آینده را از روی آن پیش بینی کنیم که بطور حیرت آوری نزدیک به واقعیت هستند.

۳-۳ مدل ها و الگوریتم های داده کاوی

در این بخش قصد داریم مهمترین الگوریتم ها و مدل های داده کاوی را بررسی کنیم. بسیاری از محصولات تجاری داده کاوی از مجموعه از این الگوریتم ها استفاده می کنند و معمولا هر کدام آنها در یک بخش خاص

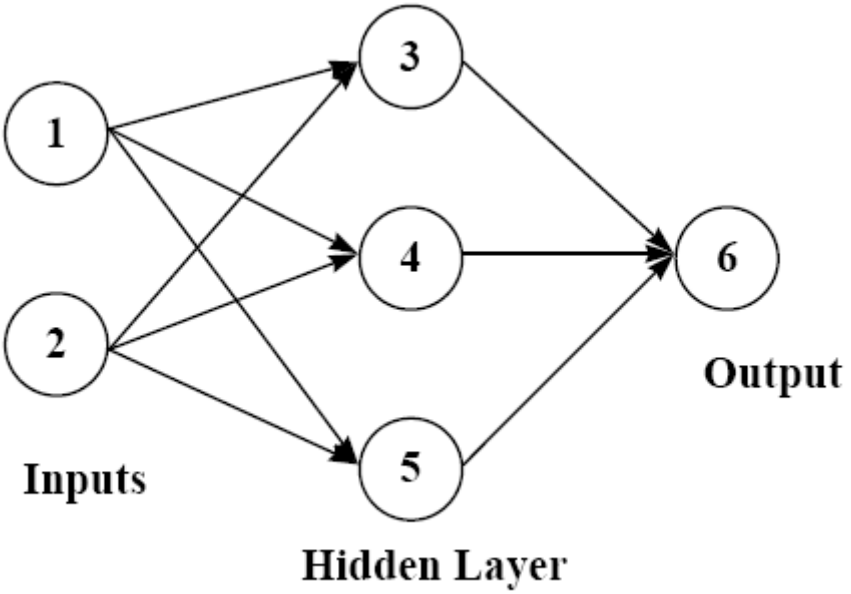
قدرت دارند وبرای استفاده از یکی از آنها باید بررسی های لازم در جهت انتخاب متناسب ترین محصول توسط گروه متخصص در نظر گرفته شود.

نکته مهم دیگر این است که در بین این الگوریتم ها و مدل ها ، بهترین وجود ندارد و با توجه به داده ها و کارایی مورد نظر باید مدل انتخاب گردد. [31]

۳-۳-۱ شبکه های عصبی (Neural Networks)

شبکه های عصبی از پرکاربردترین و عملی ترین روش های مدل سازی مسائل پیچیده و بزرگ که شامل صدها متغیر هستند می باشد. شبکه های عصبی می توانند برای مسائل کلاس بندی (که خروجی یک کلاس است) یا مسائل رگرسیون (که خروجی یک مقدار عددی است) استفاده شوند. [68],[31]

هر شبکه عصبی شامل یک لایه ورودی (Input Layer) می باشد که هر گره در این لایه معادل یکی از متغیرهای پیش بینی می باشد. گره های موجود در لایه میانی وصل می شوند به تعدادی گره در لایه نهان (Hidden Layer). هر گره ورودی به همه گره های لایه نهان وصل می شود. گره های موجود در لایه نهان می توانند به گره های یک لایه نهان دیگر وصل شوند یا می توانند به لایه خروجی (Output Layer) وصل شوند. لایه خروجی شامل یک یا چند متغیر خروجی می باشد. [68],[30]



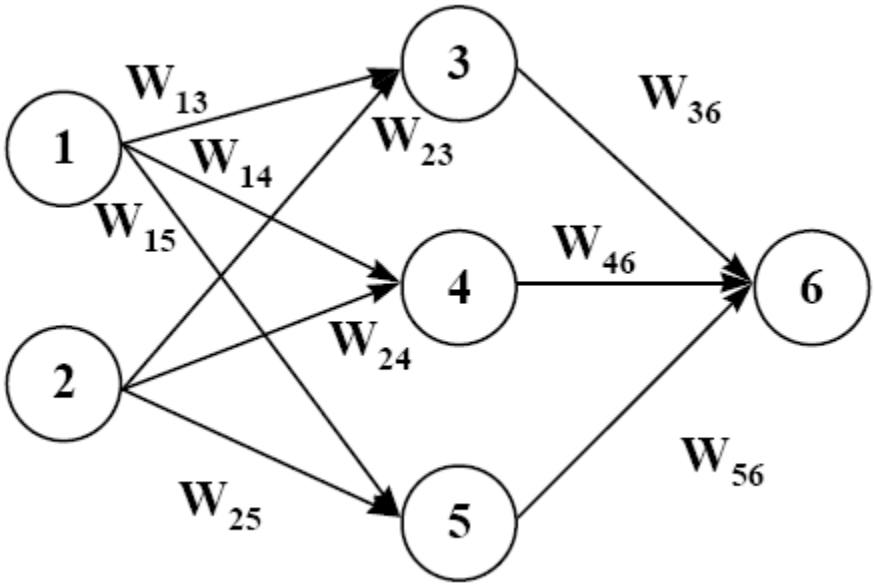
شکل ۳-۲ : شبکه عصبی با یک لایه نهان

هر یال که بین نود های X,Y می باشد دارای یک وزن است که با Wx,y نمایش داده می شود. این وزن ها در محاسبات لایه های میانی استفاده می شوند و طرز استفاده آنها به این صورت است که هر نود در لایه های میانی (لایه های غیر از لایه اول) دارای چند ورودی از چند یال مختلف می باشدکه همانطور که گفته شد هر کدام یک وزن خاص دارند. [31]

هر نود لایه میانی میزان هر ورودی را در وزن یال مربوطه آن ضرب می کند و حاصل این ضرب ها را با هم جمع می کند و سپس یک تابع از پیش تعیین شده (تابع فعال سازی) روی این حاصل اعمال می کند و نتیجه را به عنوان خروجی به نودهای لایه بعد می دهد. [38],[31],[29]

وزن یال ها پارامترهای ناشناخته ای هستند که توسط تابع آموزش (Training method) و داده های آموزشی که به سیستم داده می شود تعیین می گردند. [57],[31]

تعداد گره ها و تعداد لایه های نهان و نحوه وصل شدن گره ها به یکدیگر معماری (توپولوژی) شبکه عصبی را مشخص می کند. کاربر یا نرم افزاری که شبکه عصبی را طراحی می کند باید تعداد نودها ، تعداد لایه های نهان ، تابع فعال سازی و محدودیت های مربوط به وزن یال ها را مشخص کند. [68],[31]



شکل ۳- ۳ : نمایش شبکه عصبی بصورت گراف وزن دار

(Wx,y وزن یال بین X و Y است.)

از مهمترین انواع شبکه های عصبی Feed-Forward Backpropagation می باشد که در اینجا به اختصار آنرا توضیح می دهیم. [68],[31]

Feed-Forward به معنی این است که مقدار پارامتر خروجی براساس پارامترهای ورودی و یک سری وزن های اولیه تعیین می گردد. مقادیر ورودی با هم ترکیب شده و در لایه های نهان استفاده می شوند و مقادیر این لایه های نهان نیز برای محاسبه مقادیر خروجی ترکیب می شوند. [68],[31]

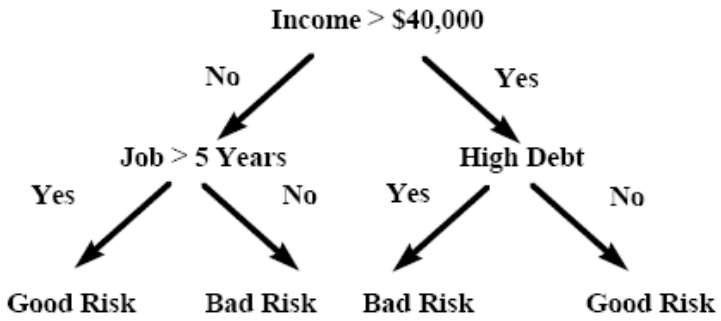
Backpropagation : خطای خروجی با مقایسه مقدار خروجی با مقدار مد نظر در داده های آزمایشی محاسبه می گردد و این مقدار برای تصحیح شبکه و تغییر وزن یال ها استفاده می گردد و از گره خروجی شروع شده و به عقب محاسبات ادامه می یابد. این عمل برای هر رکورد موجود در بانک اطلاعاتی تکرار می گردد. به هر بار اجرای این الگوریتم برای تمام داده های موجود در بانک یک دوره (Epoch) گفته می شود. این دوره ها آنقدر ادامه می یابد که دیگر مقدار خطا تغییر نکند.

از آنجایی که تعداد پارامترها در شبکه های عصبی زیاد می باشد محاسبات این شبکه ها می تواند وقت گیر باشد. ولی اگر این شبکه ها به مدت کافی اجرا گردند معمولا موفقیت آمیز خواهند بود. مشکل دیگری که ممکن است به وجود بیاید Overfitting می باشد و آن بدین صورت است که شبکه فقط روی داده های آموزشی خوب کار می کند و برای سایر مجموعه داده ها مناسب نمی باشد. برای رفع این مشکل ما باید بدانیم چه زمانی آموزش شبکه را متوقف کنیم. یکی از راه ها این است که شبکه را علاوه بر داده های آزمایشی روی داده های تست نیز مرتباً اجرا کنیم و جریان تغییر خطا را در آنها بررسی کنیم. اگر در این داده ها به جایی رسیدیم که میزان خطا رو به افزایش بود حتی اگر خطا در داده های آزمایشی همچنان رو به کاهش باشد آموزش را متوقف کنیم. [29]

از آنجایی که پارامترهای شبکه های عصبی زیاد است یک خروجی خاص می تواند با مجموعه های مختلفی از مقادیر پارامترها ایجاد گردد درنتیجه این پارامترها مثل وزن یالها قابل تفسیر نبوده و معنی خاصی نمی دهند. یکی از مهمترین فواید شبکه های عصبی قابلیت اجرای آنها روی کامپیوترهای موازی می باشد. [68],[61],[57],[37],[29]

۳-۳-۲ درخت های تصمیم گیری (Decision trees)

درخت‌های تصمیم روشی برای نمایش یک سری از قوانین هستند که منتهی به یک طبقه یا مقدار می‌شوند. برای مثال، می‌خواهیم متقاضیان وام را به دارندگان ریسک اعتبار خوب و بد تقسیم کنیم. شکل ۳ - ۴ یک درخت تصمیم را که این مسئله را حل می‌کند نشان می‌دهد و همه مؤلفه‌های اساسی یک درخت تصمیم در آن نشان داده شده است : نود تصمیم، شاخه‌ها و برگ‌ها. [17]



شکل ۳- ۴ : درخت تصمیم گیری

براساس الگوریتم، ممکن است دو یا تعداد بیشتری شاخه داشته باشد. برای مثال، CART درختانی با تنها دو شاخه در هر نود ایجاد می‌کند. هر شاخه منجر به نود تصمیم دیگر یا یک نود برگ می‌شود. با پیمایش یک درخت تصمیم از ریشه به پایین به یک مورد یک طبقه یا مقدار نسبت می‌دهیم. هر نود از داده‌های یک مورد برای تصمیم‌گیری درباره آن انشعاب استفاده می‌کند. درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. [25],[48],[56]

یکی از تفاوت‌ها بین متدهای ساخت درخت تصمیم این است که این فاصله چگونه اندازه‌گیری می‌شود. درخت-های تصمیمی که برای پیش‌بینی متغیرهای دسته‌ای استفاده می‌شوند، درخت‌های classification نامیده می‌شوند زیرا نمونه‌ها را در دسته‌ها یا طبقه ها قرار می‌دهند. درخت‌های تصمیمی که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شوند درخت‌های regression نامیده می‌شوند. [25]

هر مسیر در درخت تصمیم تا یک برگ معمولا قابل فهم است. از این لحاظ یک درخت تصمیم می‌تواند پیش-بینی‌های خود را توضیح دهد، که یک مزیت مهم است. با این حال این وضوح ممکن است گمراه‌کننده باشد. برای مثال، جداسازی های سخت در درخت‌های تصمیم دقتی را نشان می‌دهند که کمتر در واقعیت نمود دارند.

درخت‌های تصمیم تعداد دفعات کمی از داده‌ها گذر می‌کنند(برای هر سطح درخت حداکثر یک مرتبه) و با متغیرهای پیش‌بینی‌کننده زیاد به خوبی کار می‌کنند. درنتیجه، مدلها به سرعت ساخته می‌شوند، که آنها را برای مجموعه داده های بسیار مناسب می‌سازد. اگر به درخت اجازه دهیم بدون محدودیت رشد کند زمان ساخت بیشتری صرف می‌شود که غیرهوشمندانه است، اما مسئله مهمتر این است که با داده‌ها overfit می‌شوند. اندازه درخت‌ها را می‌توان از طریق قوانین توقف کنترل کرد. یک قانون معمول توقف محدود کردن عمق رشد درخت است. [25],[9],[8]

راه دیگر برای توقف هرس کردن درخت است. درخت می‌تواند تا اندازه نهایی گسترش یابد، سپس با استفاده از روش‌های اکتشافی توکار یا با مداخله کاربر، درخت به کوچکترین اندازه‌ای که دقت در آن از دست نرود کاهش می‌یابد. [25]

یک اشکال معمول درخت‌های تصمیم این است که آنها تقسیم‌کردن را براساس یک الگوریتم حریصانه انجام می‌دهند که در آن تصمیم‌گیری اینکه براساس کدام متغیر تقسیم انجام شود، اثرات این تقسیم در تقسیم‌های آینده را درنظر نمی‌گیرد. [17]

بعلاوه الگوریتم‌هایی که برای تقسیم استفاده می‌شوند، معمولا تک‌متغیری هستند: یعنی تنها یک متغیر را در هر زمان در نظر می‌گیرند. درحالیکه این یکی از دلایل ساخت سری مدل است، تشخیص رابطه بین متغیرهای پیش‌بینی کننده را سخت‌تر می‌کند. [25]

۳-۳-۳ Multivariate Adaptive Regression Splines(MARS)

در میانه‌های دهه ۸۰ یکی از مخترعین CART، ـJerome H. Friedman، متدی را برای برطرف‌کردن این کاستی‌ها توسعه داد.

کاستی‌های اساسی که او قصد برطرف کردن آنها را داشت عبارتند از :

- پیش‌بینی های غیرپیوسته(تقسیم سخت)
- وابستگی همه تقسیم‌ها به تقسیم‌های قبلی

به این دلیل او الگوریتم MARS را توسعه داد. ایده اصلی MARS نسبتا ساده است، درحالیکه خود الگوریتم نسبتا پیچیده است. بسیار ساده ایده عبارت است از :

- جایگزینی انشعاب‌های غیرپیوسته با گذر های پیوسته که توسط یک جفت از خط‌های مستقیم مدل می‌شوند. در انتهای فرآیند ساخت مدل، خطوط مستقیم در هر نود با یک تابع بسیار هموار که spline نامیده می‌شود جایگزین می‌شوند.
- عدم نیاز به اینکه تقسیم‌های جدید وابسته به تقسیم‌های قدیمی باشند.

متأسفانه این به معنی اینست که MARS ساختار درختی CART را ندارد و نمی‌تواند قوانینی را ایجاد کند. از طرف دیگر، MARS به صورت خودکار مهم‌ترین متغیرهای پیش‌بینی کننده و همچنین تعامل میان آنها را می‌یابد. MARS همچنین وابستگی میان پاسخ و هر پیش‌بینی کننده را معین می‌کند. نتیجه ابزار رگرسیون اتوماتیک، خودکار و step-wise است. [60],[51]

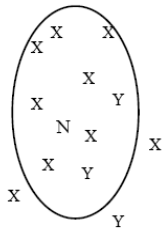
MARS مانند بیشتر الگوریتم‌های شبکه‌های عصبی و درخت تصمیم، تمایل به overfit شدن برای داده‌های آموزش‌دهنده دارد که می‌توان آنرا به دو طریق درست کرد. اول اینکه، cross validation بصورت دستی انجام شود و الگوریتم برای تولید پیش‌بینی خوب روی مجوعه تست تنظیم شود. دوم اینکه، پارامترهای تنظیم متفاوتی در خود الگوریتم وجود دارد که cross validation درونی را هدایت می‌کند. [58]

Rule induction ۴-۳-۳

استنتاج قوانین متدی برای تولید مجموعه‌ای از قوانین است که موارد را طبقه بندی می‌کند. اگرچه درخت‌های تصمیم می‌توانند مجموعه‌ای از قوانین را ایجاد کند، روش های استنتاج قوانین مجموعه‌ای از قوانین مستقل را ایجاد می‌کند که لزوما یک درخت را ایجاد نمی‌کنند. از آنجا که استنتاج‌گر قوانین اجباری به تقسیم در هر سطح ندارد، و می‌تواند به آینده بنگرد، قادر است الگوهای متفاوت و گاها بهتری برای طبقه بندی بیابد. برخلاف درختان، قوانین ایجاد شده ممکن است همه موارد ممکن را نپوشاند. همچنین برخلاف درختان، قوانین ممکن است در پیش‌بینی متعارض باشند، که در هر مورد باید قانونی را برای دنبال کردن انتخاب کرد. یک روش برای حل این تعارضات انتصاب یک میزان اطمینان به هر قانون است و استفاده از قانونی است که میزان اطمینان بالاتری دارد. [58]

۳-۳- ۵ K-nearest neibour and memory-based reansoning(MBR)

هنگام تلاش برای حل مسائل جدید، افراد معمولا به راه‌حل های مسائل مشابه که قبلا حل شده‌اند مراجعه می-کنند. K-nearest neighbor(k-NN) یک تکنیک طبقه بندی است که از نسخه‌ای از این روش استفاده می‌کند. در این روش تصمیم‌گیری، اینکه یک مورد جدید در کدام طبقه قرار گیرد با بررسی تعدادی(k) از شبیه‌ترین موارد یا همسایه‌ها انجام می‌شود. تعداد موارد برای هر کلاس شمرده می‌شوند، و مورد جدید به طبقه ای که تعداد بیشتری از همسایه‌ها به آن تعلق دارند نسبت داده می‌شود.



شکل۳- ۵ : محدوده همسایگی (بیشتر همسایه ها در دسته X قرار گرفته اند)

اولین مورد برای بکاربردن k-NN یافتن معیاری برای فاصله بین ویژگی ها در داده‌ها و محاسبه آن است. در حالیکه این عمل برای داده‌های عددی آسان است، متغیرهای طبقه ای نیاز به برخورد خاصی دارند. هنگامی که فاصله بین موارد مختلف را توانستیم اندازه گیریم، می‌توانیم از مجموعه مواردی که قبلا طبقه‌بندی شده‌اند بعنوان پایه طبقه‌بندی موارد جدید استفاده کنیم، فاصله همسایگی را تعیین کنیم، و تعیین کنیم که خود همسایه‌ها را چگونه بشماریم. [31],[12]

K-NN بار محاسباتی زیادی را روی کامپیوتر قرار می‌دهد زیرا زمان محاسبه بصورت فاکتوریلی از تمام نقاط افزایش می‌یابد. درحالیکه بکار بردن درخت تصمیم یا شبکه عصبی برای یک مورد جدید فرایند سریعی است، K-NN نیاز به محاسبه جدیدی برای هر مورد جدید دارد. برای افزایش سرعت K-NN معمولا تمام داده‌ها در حافظه نگه‌داری می‌شوند. فهم مدلهای K-NN هنگامیکه تعداد متغیرهای پیش‌بینی کننده کم است بسیار ساده است. آنها همچنین برای ساخت مدلهای شامل انواع داده غیر استاندارد هستند، مانند متن بسیار مفیدند. تنها نیاز برای انواع داده جدید وجود معیار مناسب است. [31],[13]

۳-۳-۶ رگرسیون منطقی (Logistic regression)

رگرسیون منطقی یک حالت عمومی تر از رگرسیون خطی می باشد. قبلا این روش برای پیش بینی مقادیر باینری یا متغیرهای دارای چند مقدار گسسته (کلاس) استفاده می شد. از آنجایی که مقادیر مورد نظر برای پیش بینی مقادیر گسسته می باشند نمی توان آنرا به روش رگرسیون خطی مدلسازی کرد [31] . برای این منظور این متغیرهای گسسته را به روشی تبدیل به متغییر عددی و پیوسته می کنیم و برای این منظور مقدار لگاریتم احتمال متغیر مربوطه را در نظر می گیریم و احتمال پیشامد را بدین صورت در نظر می گیریم:

احتمال اتفاق نیفتادن پیشامد/ احتمال اتفاق افتادن پیشامد و تفسیر این نسبت مانند تفسیری است که در بسیاری از مکالمات روزمره در مورد مسابقات یا شرط بندی ها یه موارد مشابه به کار می رود. مثلاً وقتی می گوییم شانس بردن یک تیم در مسابقه ۳ به ۱ است در واقع از همین نسبت استفاده کرده و معنی آن این است که احتمال برد آن تیم ۷۵٪ است.

وقتی که ما موفق شدیم لگاریتم احتمال مورد نظر را بدست آوریم با اعمال لگاریتم معکوس می توان نسبت مورد نظر و از روی آن کلاس مورد نظر را مشخص نمود.

۳-۳-۷ تحلیل تفکیکی (Discriminant analysis)

این روش از قدیمی ترین روش های ریاضی وار طبقه بندی داده ها می باشد که برای اولین بار در سال ۱۹۳۶ توسط فیشر استفاده گردید. روش کار بدین صورت است که داده ها را مانند داده های چند بعدی بررسی کرده و بین داده ها مرزهایی ایجاد می کنند (برای داده ها دو بعدی خط جدا کننده، برای داده های سه بعدی سطح جدا کننده و ..) که این مرزها مشخص کننده کلاس های مختلف می باشند و بعد برای مشخص کردن کلاس مربوط به داده های جدید فقط باید محل قرارگیری آن را مشخص کنیم. [25]

این روش از ساده ترین و قابل رشدترین روش های کلاس بندی می باشد که در گذشته بسیار استفاده می شد.

این روش به سه دلیل محبوبیت خود را از دست داد : اول اینکه این روش فرض می کند همه متغیرهای پیش بینی به صورت نرمال توزیع شده اند که در بسیاری از موارد صحت ندارد . دوم اینکه داده هایی که به صورت عددی نمی باشند مثل رنگها در این روش قابل استفاده نمی باشند. سوم اینکه در این روش فرض می شود که مرزهای جدا کننده داده ها به صورت اشکال هندسی خطی مثل خط یا سطح می باشند حال اینکه این فرض همیشه صحت ندارد. [25]

نسخه های اخیر تحلیل تفکیکی بعضی از این مشکلات را رفع کرده اند به این طریق اجازه می دهند مرزهای جدا کننده بیشتر از درجه ۲ نیز باشند که باعث بهبود کارایی و حساسیت در بسیاری از موارد می گردد.

۳-۳-۸ مدل افزودنی کلی (Generalized Additive Models (GAM))

این روش ها در واقع بسطی بر روش های رگرسیون خطی و رگرسیون منطقی می باشند. به این دلیل به این روش افزودنی می گویند که فرض می کنیم می توایم مدل را به صورت مجموع چند تابع غیر خطی (هر تابع برای یک متغیر پیش بینی کننده) بنویسیم. GAM می تواند هم به منظور رگرسیون و هم به منظور کلاس بندی داده ها استفاده گردد. این ویژگی غیر خطی بودن توابع باعث می شود که این روش نسبت به روشهای رگرسیون خطی بهتر باشد.

۳-۳-۹ Boosting

در این روش ها مبنای کار این است که الگوریتم پیش بینی را چندین بار و هر بار با داده های آموزشی متفاوت (که با توجه به اجرای قبلی انتخاب می شوند) اجرا کنیم و در نهایت آن جوابی که بیشتر تکرار شده را انتخاب کنیم. این روش اگر چه وقت گیر است ولی جواب های آن مطمئن تر خواهند بود. این روش اولین بار در سال ۱۹۹۶ استفاده شد و در این روزها با توجه به افزایش قدرت محاسباتی کامپیوترها بر مقبولیت آن افزوده گشته است. [52]

۳-۴ سلسله مراتب انتخاب‌ها

هدف داده‌کاوی تولید دانش جدیدی است که کاربر بتواند از آن استفاده کند. این هدف با ساخت مدلی از دنیای واقع براساس داده‌های جمع‌آوری شده از منابع متفاوت بدست می‌آید. نتیجه ساخت این مدل توصیفی از الگوها و روابط داده‌هاست که می‌توان آنرا برای پیش‌بینی استفاده کرد. سلسه انتخاب‌هایی که قبل از آغاز باید انجام شود به این شرح است :

- هدف تجاری

- نوع پیش‌بینی
- نوع مدل
- الگوریتم
- محصول

در بالاترین سطح هدف تجاری قرار دارد: هف نهایی از کاوش داده‌ها چیست؟ برای مثال، جستجوی الگوها در داده‌ها ممکن است برای حفظ مشتری‌های خوب باشد، که ممکن است مدلی برای سودبخشی مشتری‌ها و مدل دومی برای شناسایی مشتری‌هایی که ممکن از دست دهیم می‌سازیم. اطلاع از اهداف و نیازهای سازمان ما را در فرموله کردن هدف سازمان یاری می‌رساند.

مرحله بعدی تصمیم‌گیری درباره نوع پیش‌بینی مناسب است: (۱) classification : پیش‌بینی اینکه یک مورد در کدام گروه یا طبقه قرار می‌گیرد. یا (۲) regression : پیش‌بینی اینکه یک متغیر عددی چه مقداری خواهد داشت.

مرحله بعدی انتخاب نوع مدل است: یک شبکه عصبی برای انجام regression ،و یک درخت تصمیم برای classification. همچنین روشهای مرسوم آماری مانند logistic regression، discriminant analysis، و یا مدلهای خطی عمومی وجود دارد.

الگوریتم‌های بسیاری برای ساخت مدلها وجود دارد. می‌توان یک شبکه عصبی را با backpropagation، یا توابع radial bias ساخت. برای درخت تصمیم، می‌توان از میان CART، C5.0، Quest، و یا CHAID انتخاب کرد.

هنگام انتخاب یک محصول داده‌کاوی، باید آگاه بود که معمولا پیاده‌سازی های متفاوتی از یک الگوریتم دارند. این تفاوت‌های پیاده‌سازی می‌تواند بر ویژگیهای عملیاتی مانند استفاده از حافظه و ذخیره داده و همچنین ویژگیهای کارایی مانند سرعت و دقت اثر گذارند.

در مدلهای پیش بینی کننده، مقادیر یا طبقه‌هایی که ما پیش‌بینی می‌کنیم متغیرهای پاسخ، وابسته، یا هدف نامیده می‌شوند. مقادیری که برای پیش‌بینی استفاده می‌شوند متغیرهای مستقل یا پیش‌بینی‌کننده نامیده می‌شوند.

مدلهای پیش بینی کننده با استفاده از داده‌هایی که مقادیر متغیرهای پاسخ برای آنها از قبل دانسته شده است ساخته یا آموزش داده می‌شوند. این نحوه آموزش supervised learning نامیده می‌شود، زیرا که مقادیر

محاسبه شده یا تخمین‌زده شده با نتایج معلومی مقایسه می‌شوند. (در مقابل، تکنیک‌های توصیفی مانند clustering، unsupervised learning نامیده می‌شوند زیرا که هیچ نتیجه از پیش معلومی برای راهنمایی الگوریتم وجود ندارد).

۳-۵ پیش‌گویی نتیجه تحصیلی فراگیران با استفاده از روشهای یادگیری ماشین در داده کاوی

یادگیری ماشین، فرآیند یادگیری از مجموعه‌ای از نمونه‌ها یا مجموعه ای از قوانین یا ایجاد یک طبقه‌بندی کننده برای نمونه‌های جدید است. اهمیت تخمین دقیق نتیجه تحصیلی آتی فراگیران برای تامین یک سیستم آموزشی ، متناسب با توانایی فراگیر و در نتیجه یاری وی در این فرایند برای موفقیت بیشتر ضروری است. اگر مدرس و یا سیستم آموزشی بتواند نتیجه تحصیلی فراگیر را در هر مقطع زمانی از دوره تحصیلی پیش‌گویی کند، قادر خواهد بود که برنامه آموزشی متناسب با فراگیر را به وی ارایه نماید و از افت تحصیلی فراگیران ضعیف جلوگیری نماید و شکوفایی بیشتر فراگیران مستعد را فراهم نماید.بدیهی است که انجام این تخمین و پیش‌گویی نتیجه به صورت دستی و تجربی و استخراج قواعدی برای این امر فرایندی مشکل خواهد بود و به همین دلیل نیاز به شیوه‌هایی است که با روش‌های یادگیری ماشین به صورت خودکار انجام شود. برای این پیشگویی می‌توان از ویژگی‌های مختلف فردی، فرهنگی و اجتماعی مرتبط با فرد استفاده نمود.بدیهی است که استخراج تمام ویژگی‌های مربوط به فراگیران امکانپذیر نمی‌باشد، بنابراین ضروری است که تعدادی از مهم‌ترین ویژگی‌های مربوط به فراگیران که می‌تواند در تعیین کیفیت تحصیلی فراگیران موثرتر باشد انتخاب و استفاده نمود. [18]

پیش‌بینی یک مساله جالب و کاربردی در حوزه سیستم‌های آموزش الکترونیک است که می‌تواند همپوشانی بسیاری با سایر حوزه‌های داده‌کاوی نظیر طبقه‌بندی داشته باشد. امکان پیش‌بینی رفتار و کارایی فراگیران سیستم‌های آموزش الکترونیک باعث افزایش کیفیت کارکردی این سیستم‌ها خواهد شد. [18]

در [60],[61] یک متدلوژی برای ارتقا و توسعه سیستم ارایه دروس ارائه شده است. فایل‌های لاگ استفاده از دروس در یک پایگاه داده ذخیره می‌شود و سپس با اعمال الگوریتم‌های یادگیری ماشین، ارتباطات و الگوهای مهم استخراج و سپس توسط مدرسان قابل استفاده خواهد بود. به عنوان مثال این الگوها می‌تواند رابطه بین

سطح دانش فراگیر با مدت زمان استفاده وی از سیستم و نمرات وی یاشد. در بعضی تحقیقات از رگرسیون برای پیش‌گویی استفاده شده است. در [21] سعی شده است دلایل خطا در پیش‌گویی دانش فراگیران تشخیص داده شود. مدل‌های گرافیکی و روش‌های بیزین نیز در این حوزه استفاده شده‌اند. در [11] نیز از شبکه‌های بیزین دینامیک برای مدل کردن دانش فراگیر و پیش‌گویی نتیجه تحصیلی وی استفاده شده است.

MOODLE یک نرم‌افزار متن باز است که به عنوان یک سیستم مدیریت آموزشی محبوبیت بسیاری پیدا کرده است. این سیستم تعدادی ماژول تعاملی نظیر فروم، اتاق گفتگو و مرکز امتحان دارد که به تسهیل امر یادگیری کمک می‌کند. علاوه بر این ماژول‌های مربوط به یادگیری، این سیستم نیز شامل ماژول‌هایی برای ثبت و ردگیری رفتارهای کاربران نظیر هویت کاربر، IP، زمان دسترسی به سیستم و کنش‌ها و نیز منابع استفاده شده توسط آنها می‌باشد [18].

فصل چهارم

قابلیت های نرم افزار وکا

۴-۱ بسته نرم افزاری weka

تا به امروز نرم افزار های تجاری و آموزشی فراوانی برای داده کاوی در حوزه های مختلف داده ها به دنیای علم و فناوری عرضه شده‌اند. هریک از آنها با توجه به نوع اصلی داده هایی که مورد کاوش قرار می‌دهند، روی الگوریتمهای خاصی متمرکز شده‌اند. مقایسه دقیق و علمی این ابزارها باید از جنبه های متفاوت و متعددی مانند تنوع انواع و فرمت داده های ورودی، حجم ممکن برای پردازش داده ها، الگوریتمها پیاده سازی شده، روشهای ارزیابی نتایج، روشهای مصور سازی ، روشهای پیش پردازش داده ها، واسطهای کاربر پسند ، پلت فرم های سازگار برای اجرا،قیمت و در دسترس بودن نرم افزار صورت گیرد. از آن میان، نرم افزار Weka با داشتن امکانات بسیار گسترده، امکان مقایسه خروجی روشهای مختلف با هم، راهنمای خوب، واسط گرافیگی کارآ و سازگاری با سایر برنامه‌های ویندوزی معرفی می‌شود .

میزکار Weka ، مجموعه‌ای از الگوریتم‌های روز یادگیری ماشینی و ابزارهای پیش پردازش داده‌ها می‌باشد. این نرم‌افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش نمود. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرآیند داده کاوی های تجربی فراهم می‌کند .این پشتیبانی‌ها، آماده سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در بر می‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش پردازش داده‌هاست. این جعبه ابزار متنوع

و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی را که برای مسایل مدنظر مناسب‌تر هستند، تشخیص دهد .

نرم‌افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و اسم آن از عبارت "Waikato Environment for knowledge Analysis" استخراج گشته است. همچنین Weka ، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی‌کند و در نیوزلند، یافت می‌شود. این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر GNU انتشار یافته است Weka .تقریباً روی هر پلت فرمی اجرا می‌شود و نیز تحت سیستم عامل‌های لینوکس، ویندوز، و مکینتاش، و حتی روی یک منشی دیجیتالی شخصی ، آزمایش شده است .

این نرم‌افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش و ارزیابی نتایج طرح های یادگیری روی همه مجموعه های داده موجود، قابل اعمال است .

نرم افزار Weka ، پیاده سازی الگوریتم‌های مختلف یادگیری را فراهم می‌کند و به آسانی می‌توان آنها را به مجموعه های داده خود اعمال کرد .

همچنین، این نرم‌افزار شامل مجموعه متنوعی از ابزارهای تبدیل مجموعه‌های داده‌ها، همانند الگوریتم‌های گسسته سازی می‌باشد. در این محیط می‌توان یک مجموعه داده را پیش پردازش کرد، آن را به یک طرح یادگیری وارد نمود، و دسته‌بندی حاصله و کارآیی‌اش را مورد تحلیل قرار داد. همه این کارها، بدون نیاز به نوشتن هیچ قطعه برنامه‌ای میسر است .

این محیط، شامل روش‌هایی برای همه مسایل استاندارد داده کاوی مانند رگرسیون، طبقه‌بندی، خوشه‌بندی، کاوش قواعد انجمنی و انتخاب ویژگی می‌باشد. با در نظر گرفتن اینکه، داده‌ها بخش مکمل کار هستند، بسیاری از ابزارهای پیش پردازش داده‌ها و مصورسازی آنها فراهم گشته است. همه الگوریتم‌ها، ورودی‌های خود را به صورت یک جدول رابطه‌ای به فرمت ARFF دریافت می‌کنند. این فرمت داده‌ها، می‌تواند از یک فایل خوانده شده یا به وسیله یک درخواست از پایگاه داده‌ای تولید گردد .

یکی از راه‌های به کارگیری Weka ، اعمال یک روش یادگیری به یک مجموعه داده و تحلیل خروجی آن برای شناخت چیزهای بیشتری راجع به آن اطلاعات می‌باشد. راه دیگر استفاده از مدل یادگیری شده برای تولید پیش‌بینی‌هایی در مورد نمونه‌های جدید است. سومین راه، اعمال یادگیرنده‌های مختلف و مقایسه کارآیی آنها به

منظور انتخاب یکی از آنها برای تخمین می‌باشد. روش‌های یادگیری Classifier نامیده می‌شوند و در واسط تعاملی Weka ، می‌توان هر یک از آنها را از منو انتخاب نمود. بسیاری از classifier ها پارامترهای قابل تنظیم دارند که می‌توان از طریق صفحه ویژگی‌ها یا object editor به آنها دسترسی داشت. یک واحد ارزیابی مشترک، برای اندازه‌گیری کارآیی همه classifier به کار می‌رود .

پیاده سازی‌های چارچوب‌های یادگیری واقعی، منابع بسیار ارزشمندی هستند که Weka فراهم می‌کند. ابزارهایی که برای پیش پردازش داده‌ها استفاده می‌شوند filter نامیده می‌شوند. همانند classifier ها، می-توان filter ها را از منوی مربوطه انتخاب کرده و آنها را با نیازمندی‌های خود، سازگار نمود. در ادامه، به روش به کارگیری فیلترها اشاره می‌شود .

علاوه بر موارد فوق، Wekaشامل پیاده سازی الگوریتم‌هایی برای یادگیری قواعد انجمنی، خوشه‌بندی داده‌ها در جایی که هیچ دسته‌ای تعریف نشده است، و انتخاب ویژگی‌های مرتبط در داده‌ها می‌شود. [77],[72],[69]

شکل ۴ - ۱ الگوریتم های یادگیری در وکا را نمایش می دهد.

	Name	Function
Bayes	*AODE*	Averaged, one-dependence estimators
	BayesNet	Learn Bayesian nets
	ComplementNaiveBayes	Build a Complement Naïve Bayes classifier
	NaiveBayes	Standard probabilistic Naïve Bayes classifier
	NaiveBayesMultinomial	Multinomial version of Naïve Bayes
	NaiveBayesSimple	Simple implementation of Naïve Bayes
	NaiveBayesUpdateable	Incremental Naïve Bayes classifier that learns one instance at a time
Trees	*ADTree*	Build alternating decision trees
	DecisionStump	Build one-level decision trees
	Id3	Basic divide-and-conquer decision tree algorithm
	J48	C4.5 decision tree learner (implements C4.5 revision 8)
	LMT	Build logistic model trees
	M5P	M5′ model tree learner
	NBTree	Build a decision tree with Naïve Bayes classifiers at the leaves
	RandomForest	Construct random forests
	RandomTree	Construct a tree that considers a given number of random features at each node
	REPTree	Fast tree learner that uses reduced-error pruning
	UserClassifier	Allow users to build their own decision tree
Rules	*ConjunctiveRule*	Simple conjunctive rule learner
	DecisionTable	Build a simple decision table majority classifier
	JRip	RIPPER algorithm for fast, effective rule induction
	M5Rules	Obtain rules from model trees built using M5′
	Nnge	Nearest-neighbor method of generating rules using nonnested generalized exemplars
	OneR	1R classifier
	Part	Obtain rules from partial decision trees built using J4.8
	Prism	Simple covering algorithm for rules
	Ridor	Ripple-down rule learner
	ZeroR	Predict the majority class (if nominal) or the average value (if numeric)
Functions	*LeastMedSq*	Robust regression using the median rather than the mean
	LinearRegression	Standard linear regression
	Logistic	Build linear logistic regression models
	MultilayerPerceptron	Backpropagation neural network
	PaceRegression	Build linear regression models using Pace regression
	RBFNetwork	Implements a radial basis function network
	SimpleLinearRegression	Learn a linear regression model based on a single attribute
	SimpleLogistic	Build linear logistic regression models with built-in attribute selection
	SMO	Sequential minimal optimization algorithm for support vector classification

	Name	Function
	SMOreg	Sequential minimal optimization algorithm for support vector regression
	VotedPerceptron	Voted perceptron algorithm
	Winnow	Mistake-driven perceptron with multiplicative updates
Lazy	*IB1*	Basic nearest-neighbor instance-based learner
	IBk	*k*-nearest-neighbor classifier
	KStar	Nearest neighbor with generalized distance function
	LBR	Lazy Bayesian Rules classifier
	LWL	General algorithm for locally weighted learning
Misc.	*Hyperpipes*	Extremely simple, fast learner based on hypervolumes in instance space
	VFI	Voting feature intervals method, simple and fast

شکل ۴-۱ : الگوریتمهای طبقه بندی درWeka

۴-۲ شروع کار با weka

در نسخه ویندوز با نصب یکی از نسخه ها ماشین مجازی جاوا (JDK 1.4) همراه با آن نصب می شود. یکی دیگر از نسخه ها تنها weka نصب می شود و ماشین مجازی جاوا نصب نمی شود.

در نسخه لینوکس یک فایل فشرده (zip) شامل weka است. این فایل را باز کرده پس از آن یک شاخه به اسم weka -3-4-6 ایجاد می شود. برای اجرای weka در مسیر این شاخه دستور java –jar weka.jar را اجرا کنید. جاوا باید قبلا در سیستم نصب شده باشد.

قبل از انجام هر کاری با weka بهتر است تنظیمات زیر در سیستم انجام شده باشد. بر روی MyComputer کلیک راست کرده و تنظیمات زیر را در Environment Variables انجام دهید:

- متغیری به نام WEKAHOME ایجاد کرده و مقدار آن برابر مسیر نصب weka قرار داده شود.
- متغیری به نام CLASSPATH ایجاد کرده و مسیر فایل weka.jar را در آن قرار دهید.

۴-۳ قالب فایل های ARFF

استاندارد داده های ورودی در weka باید به قالب ARFF باشد. قبل از استفاده از الگوریتم های موجود در Weka داده ورودی باید به قالب ARFF تبدیل شود. این قالب به صورت (ویژگی- مقدار) می باشد.

خصوصیت فایل های ARFF به شرح زیر است:

یک فایل متنی با پسوند ARFF می باشد.

نام مجموعه داده ای با استفاده از تگ @relation مشخص می شود.

ویژگی با تگ @attribute مشخص می شود و مقدار با تگ @data مشخص می شود.

۴-۴ ارزیابی

انتخاب فاکتورهای موثر برای ارزیابی روش های یادگیری ماشین اعمال شده یکی از مسائل مهم در انجام پایان نامه بوده است. در ادامه به معرفی فاکتورهایی که در ارزیابی روشهای مختلف مورد استفاده واقع شده است می پردازیم.

۴-۴-۱ روش ارزیابی متقاطع(Cross Validation)

در روش ارزیابی متقاطع برای تعیین میزان کیفیت یک طبقه بندی کننده مقداری از داده آموزشی به عنوان مجموعه تست استفاده می شود. به عنوان مثال ۹۰٪ مجموعه داده به عنوان داده آموزشی استفاده می شود و ۱۰٪ بقیه برای تست روش استفاده می شود. تعداد نمونه هایی که به طور صحیح طبقه بندی می شود به عنوان دقت آن روش لحاظ می شود. این فرایند ۱۰ بار تکرار می شود و هر مرتبه با نگاه داشتن ۱۰٪ داده آموزشی اجرا می شود. هنگامی که این فرایند تکمیل می شود میانگین مقادیر دقت مراحل به عنوان میزان دقت نهایی آن روش لحاظ می شود. اگر میزان مقدار داده تست نگاه داشته شده ۱۰ درصد کل داده آموزشی باشد این شیوه ارزیابی متقاطع ده - دسته (ten-fold cross validation) گفته می شود. می توان تعداد دسته ها را به غیر از ده نیز انتخاب نمود. به عنوان مثال اگر ۵٪ از داده به عنوان تست نگاه داشته شود آنگاه ارزیابی متقاطع ۲۰-دسته را خواهیم داشت. [53],[5],[4]

۴-۴-۲ دقت (Precision) و یادآوری (Recall)

بعد از آموزش و تست داده و بعد از اینکه تعیین شد هر نمونه به کدام کلاس تعلق دارد می توان مقادیر دقت و یادآوری هر طبقه بندی کننده را محاسبه نمود. این مقادیر برای هر کلاس نشان می دهد که طبقه بندی کننده در مورد هر کلاس چگونه عمل کرده است. در واقع دقت و یادآوری بیشتر در بازیابی اطلاعات برای ارزیابی دقت الگوریتم های مختلف استفاده می شود.اما این مقادیر برای روش های طبقه بندی نیز می تواند استفاده شود.

اگر a تعداد نمونه که به کلاس صحیح خود طبقه بندی شده باشد و b تعداد نمونه متعلق به یک کلاس و c نیز تعداد نمونه اختصاص داده شده به یک کلاس باشد آنگاه مقادیر دقت و ارزیابی به صورت زیر محاسبه می شود:

معادله ۴ – ۱ : یادآوری=a/b

معادله ۴ – ۲ : دقت =a/c

به عنوان مثال اگر مجموعه داده ما حاوی ۱۰۰۰ نمونه باشد آنگاه اگر یک روش ۸۰۰ نمونه به این کلاس اختصاص دهد که ۶۰۰ نمونه از آن واقعا به این کلاس تعلق داشته باشد آنگاه دقت این روش ۷۵٪ و یادآوری آن ۶۰٪ خواهد بود.

F-Measure: مقداری است که از تلفیق دو معیار دقت و یادآوری بدست می آید:

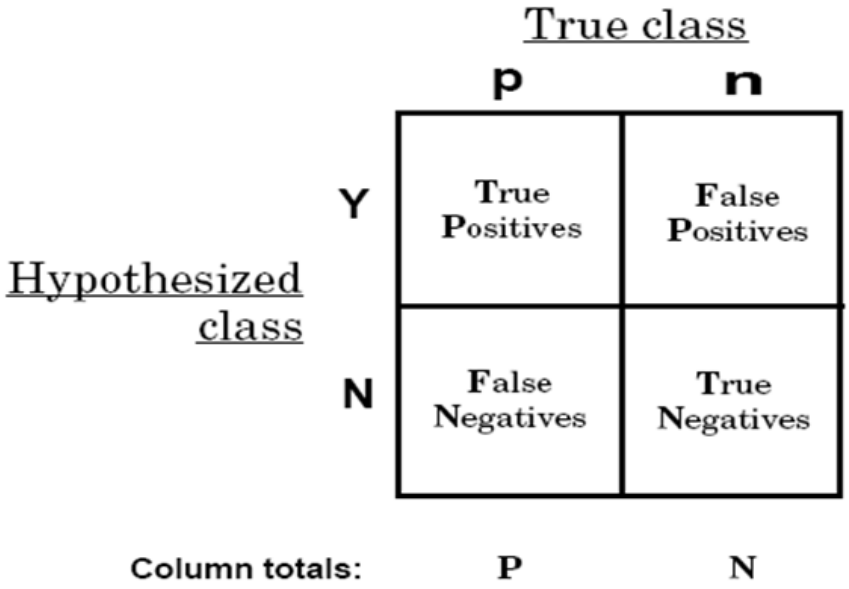
معادله ۴ – ۳ : $F - Measure = \frac{2*recall*precision}{recall+precision}$

F-Measure برای مثال فوق ۶۶٪ است که برای ارزیابی دقت یک طبقه بندی می تواند استفاده شود[5],[4].

۴-۴-۳ ماتریس آشفتگی (Confusion Matrix)

یک ماتریس آشفتگی شامل اطلاعاتی در مورد کلاسه بندی واقعی و پیش بینی شده توسط یک سیستم کلاسه بندی می باشد.[75] [70]

در زمینه پژوهش ما درایه های ماتریس آشفتگی دارای معانی زیر می باشند:



شکل۴-۲: نمایش مفهوم درایه های ماتریس آشفتگی

عبارات استاندارد مختلفی برای یک ماتریس دو کلاسه تعریف شده اند:

نرخ مثبت درست (True Positive rate) : نسبت تعداد حالات مثبت که درست طبقه بندی شده اند.

نرخ مثبت کاذب (False Positive rate) : نسبت تعدادحالات منفی که بصورت اشتباه طبقه بندی شده اند.

نرخ منفی درست (True Negative rate) : نسبت تعداد حالات منفی که درست طبقه بندی شده اند.

نرخ منفی کاذب (False Negative rate) : نسبت تعداد حالات مثبتی که بصورت حالات منفی طبقه بندی شده اند.

صحت (Accuracy) : نسبت تعداد پیش بینی های درست به تعداد کل پیش بینی ها.

معادله ۴ – ۴ : $$accuracy = \frac{TP+TN}{P+N}$$

فراخوان (Recall) یا نرخ مثبت درست (True Positive rate) : نسبت حالات مثبتی که درست تعیین شده اند.

معادله ۴ – ۵ :
$$recall = \frac{TP}{P}$$

دقت (Precision) : نسبت حالات مثبت پیش بینی شده ای که درست بوده اند.

معادله ۴ – ۶ :
$$precision = \frac{TP}{TP+FP}$$

معادله ۴ – ۷ :
$$F-Measure = \frac{2}{{}^{1}/_{precision} + {}^{1}/_{recall}}$$

مقدار صحت (Accuracy) بیان شده در معادله ۴ - ۴ ممکن است معیار کارا و کاملی نباشد اگر تعداد حالات منفی بسیار بیشتر از حالات مثبت باشد. فرض کنید ۱۰۰۰ حالت داریم که ۹۹۵ تای آنها حالات منفی هستند و فقط ۵ تای آنها حالات مثبت هستند. اگر سیستم همه آنها را بعنوان حالت منفی کلاسه بندی کند میزان دقت ۹۹/۵٪ خواهد بود حتی اگر طبقه بندی کننده (classifier) همه حالات مثبت را از دست داده باشد.

معیارهای کارایی دیگری که شامل TP هستند عبارتند از میانگین هندسی (Geometric Mean) همانگونه که در معادلات ۴ - ۸ و ۴ - ۹ نشان داده شده است و همچنین F-Measure که در معادله ۴ - ۱۰ بیان شده است.[34]

معادله ۴ - ۸ :
$$g-mean_1 = \sqrt{TP * P}$$

معادله ۴ - ۹ :
$$g-mean_2 = \sqrt{TP * TN}$$

معادله ۴ - ۱۰ :
$$F-Measure = \frac{(\beta^2+1)*P*TP}{\beta^2*P+TP}$$

در معادله ۴ - ۱۰، ß مقداری بین ۰ تا بینهایت دارد و برای کنترل وزن تخصیص داده شده به TP و P بکار می رود. [61]

راه دیگر برای تست و بررسی عملکرد طبقه بندی کننده استفاده از گراف ROC می باشد. [75],[70]

۴-۴-۴ گراف ROC(Receiver Operating Characteristic)

گراف های ROC روش دیگری برای بررسی عملکرد طبقه بندی کننده می باشند [66] . یک گراف ROC نموداری با نرخ مثبت کاذب روی محور جداکننده X و نرخ مثبت درست روی محور Y می باشد. نقطه (۱و۰) طبقه بندی کننده کامل است یعنی نقطه ای است که تمام حالات مثبت و منفی را درست کلاسه بندی (classification) می کند. مقدار آن (۱و۰) است چون نرخ مثبت کاذب صفر (هیچ) و نرخ مثبت درست یک

(همه) است. نقطه (۰و۰) طبقه بندی کننده ای را نشان می دهد که تمام حالات را به منفی تعبیر می کند در حالیکه نقطه (۱و۱) با طبقه بندی کننده ای تطابق دارد که همه حالات را مثبت تفسیر می کند. نقطه (۰و۱) مبین طبقه بندی کننده ای است که برای تمام کلاسه بندی ها نادرست است.[20] .

نرخ مثبت درست که به آن نرخ ضربه و یادآوری (hit rate and recall) هم گفته می شود برابر است با [20] :

معادله ۴ – ۱۱ : $$Tp\ rate = \frac{TP}{P}$$

معادله ۴ – ۱۲ : $$Tp\ rate \approx \frac{positives\ correctly\ classified}{total\ positives}$$

نرخ مثبت کاذب که به آن نرخ هشدار کاذب (false alarm rate) هم گفته می شود برابر است با [20]:

معادله ۴ – ۱۳ : $$FP\ rate = \frac{FP}{N}$$

معادله ۴ – ۱۴ : $$fp\ \ rate \approx \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$$

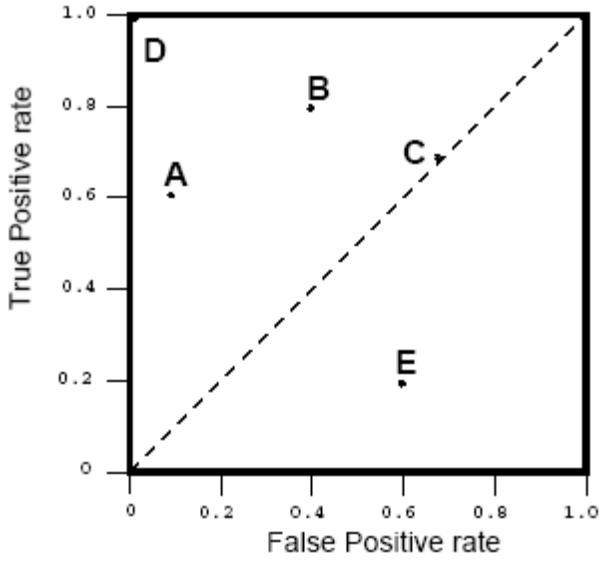
سایر عبارات مرتبط با منحنی های ROC عبارتند از [20] :

معادله ۴ – ۱۵ : Sensitivity = recall

معادله ۴ – ۱۶ : $$\begin{aligned}\text{specificity} &= \frac{\text{True negatives}}{\text{False positives} + \text{True negatives}} \\ &= 1 - fp\ \ rate\end{aligned}$$

معادله ۴ – ۱۷ :

Positive predictive value = precision



شکل ۴ - ۳ : یک گراف ROC ابتدایی با پنج طبقه بندی کننده

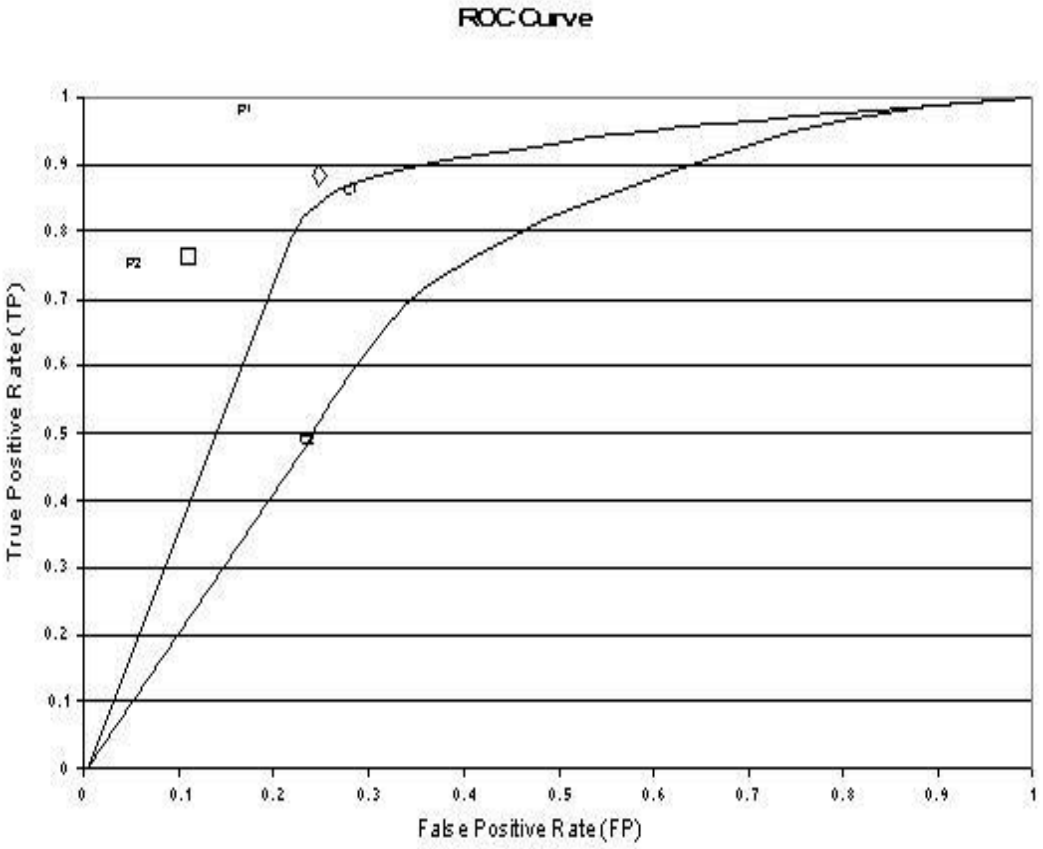
قطر فرعی این محور بیانگر عملکرد تصادفی (Random Performance) است یعنی اگر نمونه ای روی این قطر در گراف قرار گیرد به معنای این است که طبقه بندی کننده اطلاعاتی در مورد کلاس آن نمونه نداشته است. هر چه نمونه ها بیشتر در سمت شمال غرب نمودار باشند عملکرد طبقه بندی کننده بهتر بوده است (بدلیل TP Rate بالاتر و FP Rate پایینتر). هر چه نمونه ها در گوشه سمت چپ پایین و نزدیک به محور X باشند یعنی استراتژی طبقه بندی کننده در برخورد با نمونه ها یک استراتژی محافظه کارانه (Conservative) بوده است چون در این حالت اگرچه نرخ مثبت کاذب را پایین آورده است اما بدنبال آن نرخ مثبت درست هم پایین آمده است . هر چه نمونه ها در گوشه سمت راست بالا متمرکزتر باشند به معنی برخورد لیبرالی (Libral) طبقه بندی کننده است چون اگرچه نرخ مثبت درست را بالا برده است اما بدنبال آن نرخ مثبت کاذب هم بالا رفته است. اگر نمونه هایی داشته باشیم که در مثلث زیر قطر فرعی قرار بگیرند به معنی این است که طبقه بندی کننده در طبقه بندی آنها عملکردی ضعیفتر از عملکرد تصادفی داشته است[20] .

در شکل ۴ – ۳ طبقه بندی کننده A محافظه کارانه تر از طبقه بندی کننده B عمل کرده است و عملکرد C مجازاً تصادفی (random) است. در (۰.۷,۰.۷) می توان گفت که C کلاس مثبت را ۷۰٪ مواقع درست حدس می زند. طبقه بندی کننده E بسیار بدتر از عملکرد تصادفی عمل کرده است و در حقیقت منفی B می باشد.

در بسیاری از حالت ها یک طبقه بندی کننده دارای پارامتری است که می تواند برای افزایش TP تطبیق داده شود با هزینه افزایش FP و یا با کاهش FP با هزینه کاهش TP.

هر پارامتر یک زوج (FP,TP) را فراهم می کند و یک مجموعه از چنین زوج هایی می تواند برای رسم منحنی ROC استفاده شود. در شکل زیر یک طبقه بندی کننده غیر پارامتریک بوسیله یک نقطه ROC منفرد نشان داده شده است که با زوج (FP,TP) خود مطابقت می کند. [69]

شکل ۴ - ۴ مثالی از یک گراف ROC با دو منحنی ROC به نامهای C1 و C2 و دو نقطه ROC به نامهای P1 و P2 را نشان می دهد. الگوریتم های غیر پارامتریک یک نقطه ROC برای گراف تولید می کنند. [69]



شکل ۴- ۴ : یک گراف ROC با دو منحنی و دو نقطه ROC

۴ – ۴ – ۴ – ۱ زمینه های مرتبط با گراف های ROC :

- یک منحنی یا نقطه ROC مستقل از توزیع کلاس یا هزینه خطا می باشد [53].
- یک گراف ROC تمام اطلاعاتی که در ماتریس آشفتگی آمده اند را کپسوله سازی می کند چرا که FN مکمل TP و TN مکمل FP است [66].
- منحنی ROC ابزاری تصویری برای تست تعادل بین توانایی یک طبقه بندی کننده جهت تشخیص درست حالات مثبت و تعداد حالات منفی که بطور اشتباه کلاسه بندی شده اند را فراهم می کند. [71],[70],[69]

۴-۴-۴- ۲ اندازه گیری صحت مبتنی بر ناحیه (Area-based Accuracy Measure)

پیشنهاد شده است که ناحیه زیر منحنی Roc می تواند بعنوان معیار صحت (Accuracy) در بسیاری کاربردها در نظر گرفته شود [66].

پرووست و فاست معتقد بودند که استفاده از میزان صحت طبقه بندی کننده برای مقایسه طبقه بندی کننده کافی نیست مگر آنکه هزینه و توزیع کلاس کاملا ناشناخته باشند و یک طبقه بندی کننده منفرد بایستی برای کنترل هر شرایطی انتخاب شود. آنها مدلی برای ارزیابی طبقه بندی کننده با استفاده از گراف ROC ارائه دادند. [54] [53]

۴ – ۴ – ۵ نمودار خطاهای طبقه بندی کننده (Classifier Errors)

این نمودار نشان دهنده خطاهایی است که توسط مدل ایجاد شده اند. این نمودار می تواند نشان دهد که چگونه این خطاها مرتبط به ویژگی های مختلف هستند. این اطلاعات هنگامی مفید است که بخواهیم تصمیم بگیریم با حالاتی که اشتباه طبقه بندی شده اند چه کنیم. همچنین به کمک این نمودار می توانیم بفهمیم که آیا حالاتی که اشتباه طبقه بندی شده اند در وضعیت معمول قرار دارند و یا اینکه حول مقادیر مشخص یک ویژگی خاص متمرکز شده اند.

محور X نشان دهنده کلاسهای طبقه بندی و محور Y نشان دهنده مقادیر پیش بینی شده برای این کلاسها. نمونه هایی که درست پیش بینی شده اند با علامت ضربدر و نمونه هایی که اشتباه پیش بینی شده اند با علامت مربع کوچک مشخص شده اند.

۴ – ۴ – ۶ منحنی اختلاف (Margin Curve)

کاربرد این منحنی در نشان دادن اختلاف پیش بینی(prediction margin) است. منظور از اختلاف پیش بینی تفاوت بین احتمال پیش بینی شده (predicted probability) برای کلاس واقعی و بالاترین احتمال پیش بینی شده(highest probability predicted) برای سایر کلاسهاست. ما در این پژوهش ۱۵۴ نمونه را مورد بررسی قرار می دهیم. برای هر نمونه ، margin تفاوت بین احتمال پیش بینی شده برای کلاس واقعی و احتمال پیش بینی شده برای کلاس دیگر است. برای هر نمونه تست شده این مقدار اختلاف محاسبه می شود و از پایین ترین اختلاف تا بالاترین اختلاف ذخیره می گردد.

در اینجا با چهار متغیر (margin,current,cumulative,instance number) سر و کار داریم که برای ساخت یک منحنی اختلاف استفاده می شوند.

ویژگی margin شامل مقدار اختلاف پیش بینی و current شامل تعداد نمونه ها با مقدار margin یکسان و نهایتا cumulative شامل تعداد نمونه ها با اختلاف کمتر یا مساوی اختلاف current.

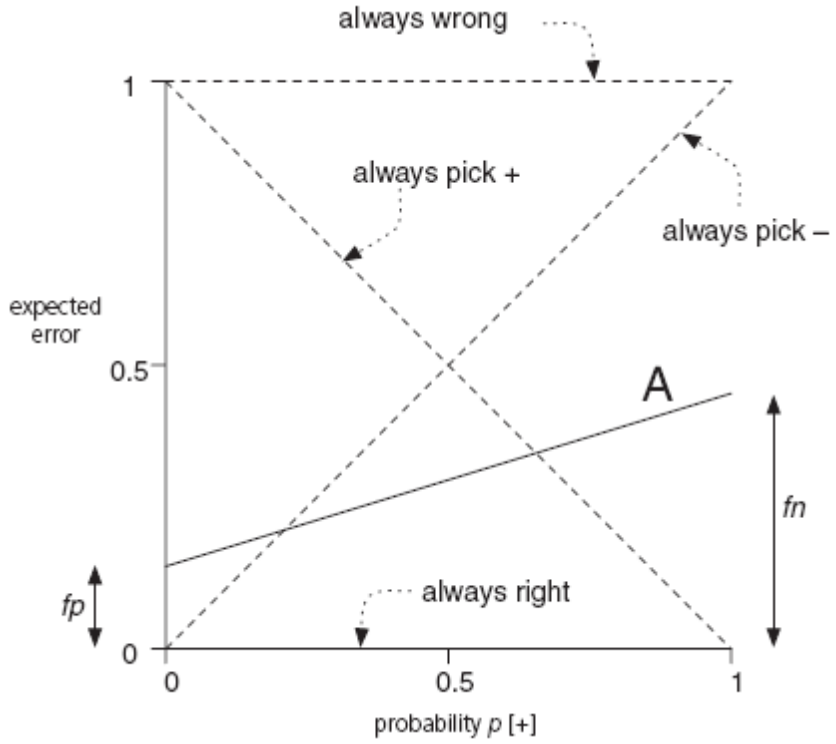
از آنجا که نمونه ها از کمترین اختلاف تا بیشترین اختلاف ذخیره می گردند می توان نتیجه گرفت که شماره نمونه و cumulative اطلاعات یکسانی می دهند.

۴ – ۴ – 7 منحنی هزینه (Cost Curve)

منحنی های هزینه نوع متفاوتی از نمایش هستند که روی آنها یک یک طبقه بندی کننده با خط راستی که نشان می دهد چگونه عملکرد (Performance) با تغییر در توزیع کلاس، تغییر می یابد ، تطابق پیدا می کند. این منحنی ها هنگامی که با یک حالت دو کلاسه مواجه هستیم بهترین کاربرد را از خود نشان می دهند اگر چه هنگامی که با یک حالت چند کلاسه سر و کار داریم می توانیم آنرا با مجزا کردن یک کلاس و ارزیابی آن با کلاسهای باقیمانده به یک حالت دو کلاسه تبدیل کنیم. [26]

شکل۴ - ۵ خطای مورد انتظار (expected error) را نسبت به احتمال (Probability) یکی از کلاسها نشان می دهد. ما کلاسها را با علامت + و – مشخص می کنیم. قطرها عملکرد طبقه بندی کننده ها را نشان می دهند: یکی از آنها همیشه + پیش بینی می کندکه در این حالت خطای مورد انتظار یک می شود اگر مجموعه داده شامل هیچ نمونه + نباشد و صفر می شود اگر همه نمونه های آن مثبت باشند. قطر دیگر همیشه – پیش بینی می کند که عملکرد معکوس دارد. خطوط تیره افقی عملکرد طبقه بندی کننده ای را نشان می دهد که

همیشه اشتباه است و خود محور X طبقه بندی کننده ای را نشان می دهد که همیشه درست است. در عمل هیچ کدام منطقی نیستند. طبقه بندی کننده های خوب نرخ خطای پایینی دارند و این یعنی اینکه آنها تا حد امکان نزدیک به پایین نمودار هستند.[26]



شکل ۴ – ۵ : نمودار خطای مورد انتظار(expected error) بر حسب احتمال(probability)

خطی که با A مشخص شده میزان خطای یک طبقه بندی کننده خاص را نشان می دهد. اگر عملکرد آنرا روی یک مجموعه تست مشخصی در نظر بگیرید نرخ مثبت کاذب (FP) آن در حقیقت خطای مورد انتظار روی یک زیرمجموعه از مجموعه تست که فقط شامل نمونه های منفی است (p[t]=0) می باشد و نرخ منفی کاذب (FN) آن در واقع خطا روی زیر مجموعه ای است که فقط شامل نمونه های مثبت می باشد (p[t]=1). [26]

با دقت در شکل بالا در می یابیم که اگر p[+] کمتر از حدود ۰.۲ باشد آنگاه پیش بینی کننده A بهتر از طبقه بندی کننده ای عمل کرده است که همیشه – پیش بینی می کند و اگر بزرگتر از حدود ۰.۶۵ باشد طبقه بندی کننده دیگر بهتر است.

شکل ۴ - ۶ منحنی هزینه را برای همان طبقه بندی کننده A نشان می دهد (برای سهولت مقیاس عمودی بزرگتر شده است و فعلا خطوط خاکستری را نادیده بگیرید). این نمودار هزینه مورد انتظار استفاده از A را

نسبت به تابع هزینه احتمال (probability cost function) نشان می دهد. این تابع صفر است هنگامی که p[+]=0 و برابر یک است هنگامی که p[+]=1. هزینه پیش بینی مثبت هنگامی که نمونه مورد نظر واقعا منفی باشد با C[+][-] و عکس آن با C[-][+] نشان داده می شود. بنابراین محورهای X و Y شکل زیر عبارتند از :

معادله ۴ – ۱۸ :

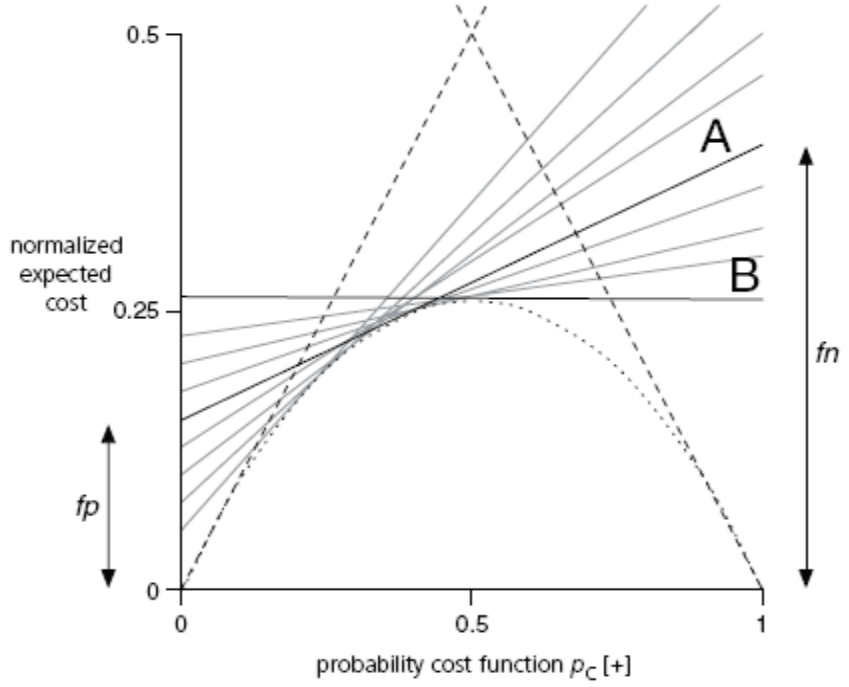
$$Normalized\ excepted\ cost = fn * p_c[+] + fp * (1 - p_c[+])$$

معادله ۴ – ۱۹ :

$$Probability\ cost\ function\ p_c[+] = \frac{p[+]C[+-]}{p[+]C[+-] + p[-]C[-+]}$$

در اینجا فرض نموده ایم که پیش بینی های درست هیچ هزینه ای ندارند یعنی $C[++] = C[--] = 0$ اگر شرایط اینگونه نباشد این فرمول ها اندکی پیچیده تر خواهند شد.[26]

بیشترین مقداری که هزینه مورد انتظاره نرمالیزه می تواند داشته باشد یک است. یک نکته جالب در مورد منحنی هزینه این است که مقادیر هزینه نهایی در سمت چپ و راست گراف fp و fn هستند.



شکل ۴ – ۶ : منحنی هزینه

شکل۴ - ۶ همچنین طبقه بندی کننده B را نشان می دهد که نرخ مثبت کاذب و نرخ منفی کاذب آن با یکدیگر برابر است. همانطور که می بینید نسبت به A عملکرد بهتری از خود نشان می دهد اگر تابع هزینه احتمال از ۰.۴۵ تجاوز کند. در شرایطی که توزیع های مختلف کلاس داریم منحنی های هزینه به راحتی می توانند به این سوال پاسخ دهند که کدام طبقه بندی کننده عملکرد بهتری داشته است.[26]

۴ – ۴ - ۸ ارزیابی پیش بینی عددی (Evaluating Numeric Prediction)

همه معیارهای ارزیابی که توصیف شد بیشتر به شرایط طبقه بندی مربوط هستند تا شرایط پیش بینی عددی. اصول پایه ای _ استفاده از یک مجموعه تست (Test Set) مستقل بجای مجموعه آموزشی (Training Set) برای ارزیابی کارآیی ، روش holdout و ارزیابی متقاطع به خوبی و بطور یکسان برای پیش بینی عددی کار می کنند اما معیار کیفیت پایه که توسط نرخ خطا (error rate) پیشنهاد می شوند چندان مناسب نیستند زیرا خطاها به سادگی حاضر یا غایب نبوده و در اندازه های مختلفی ظاهر می شوند[27].

معیارهای دیگر بسیاری که در جدول۴ – ۱ آمده اند می توانند برای ارزیابی موفقیت پیش بینی عددی استفاده شوند. مقادیر پیش بینی روی نمونه های تست عبارتند از p_1 , p_2 , ... , p_n و مقادیر واقعی عبارتند از $a_1, a_2, \dots, a_n$. در اینجا p_i معنی بسیار متفاوتی با آنچه قبلا آمده دارد. در آنجا p_i بیانگر احتمال(probability) یک پیش بینی خاص در i مین کلاس بود ولی در اینجا به مقدار عددی پیش بینی برای i مین نمونه تست اشاره می کند.

خطای میانگین مربع (Mean Squared error) اصلی ترین و معمولترین معیار مورد استفاده است. بسیاری از تکنیک های ریاضیاتی (مانند رگرسیون خطی) از خطای میانگین مربع استفاده می کنند چرا که آسانترین معیار برای فریبکاری (manipulate)ریاضیاتی می باشد و در اصطلاح ریاضیات به آن خوشرفتار (well - behaved) گفته می شود. به هر حال در اینجا ما آن را بعنوان معیار عملکرد در نظر می گیریم : همه معیارهای عملکرد برای محاسبه آسان هستند بنابراین خطای میانگین مربع مزیت خاصی ندارد[27].

خطای میانگین مطلق (Mean Absolute error) روش دیگری است که اندازه یک خطا را بدون در نظر گرفتن علامت آن در نظر دارد. خطای میانگین مربع تمایل به بزرگنمایی تاثیر outlier ها (نمونه هایی که

خطای پیش بینی آنها بزرگتر از سایرین است) دارد اما خطای مطلق این تاثیر را ندارد و با همه اندازه های خطا بطور یکسان بر طبق بزرگی آنها رفتار می شود.[27] .

گاهی اوقات خطای نسبی مهمتر از خطای مطلق است. مثلا ، اگر یک خطای ۱۰٪ در اینکه آیا یک خطای ۵۰ در یک پیش بینی از ۵۰۰ است یا یک خطای ۰.۲ در یک پیش بینی از ۲ است از اهمیت یکسانی برخوردار باشد ، میانگین های خطای مطلق بی معنی خواهند بود و خطاهای نسبی مناسب واقع خواهند شد. این تاثیر با استفاده از خطاهای نسبی در محاسبات خطای میانگین مربع یا محاسبات خطای میانگین مطلق به کار گرفته می شود.[27].

خطای نسبی مربع (Relative Squared error) در جدول ۴ - ۱ به چیز کاملا متفاوتی اشاره می کند. خطا وابسته به آن چیزی است که اگر یک پیش بینی کننده ساده استفاده می شد ، می بود. پیش بینی کننده ساده میانگین مقادیر واقعی داده های آموزش (Training Data) است. بنابراین خطای نسبی مربع ، خطای مجموع مربع را می گیرد و آن را با تقسیم بر خطای مجموع مربع پیش بینی کننده پیش فرض نرمال می کند[27] .

معیار خطای بعدی خطای نسبی مطلق است و همان خطای مجموع مطلق با همان نرمالیزاسیون می باشد. در این سه معیار خطای نسبی ، خطاها بوسیله خطای پیش بینی کننده ساده که مقادیر میانگین را پیش بینی می کند نرمال می شوند[27] .

آخرین معیار در جدول ۴ - ۱ ضریب همبستگی (Correlation Coefficient) است که همبستگی آماری بین a ها و p ها را اندازه گیری می کند که مقدار یبین ۱ و ۱- دارد. این ضریب ۱ است اگر نتایج کاملا همبستگی داشته باشند و ۰ است اگر هیچ همبستگی وجود نداشته باشد و ۱- است هنگامی که نتایج به صورت منفی کاملا همبستگی داشته باشند. البته مقادیر منفی برای روش پیش بینی معقول نباید ظاهر شوند.
همبستگی کاملا متفاوت از سایر معیارهاست چرا که مستقل از مقیاس (Scale - Independent) می باشد و این یعنی اینکه اگر شما مجموعه ای از پیش بینی های خاصی را داشته باشید ، خطا تغییر نمی کند اگر همه پیش بینی ها ضرب در عامل ثابتی شوند و مقادیر واقعی بدون تغییر می مانند. این عامل در هر عبارت S_{PA} در numerator و در هر عبارت S_P در denominator ظاهر و سرانجام خنثی می شود. (این برای شکل های مختلف خطای نسبی درست نیست مگر نرمالیزاسیون : اگر شما همه پیش بینی ها را ضرب در یک مقدار ثابت بزرگ کنید تفاوت بین مقادیر پیش بینی شده و واقعی تغییر پیدا خواهد کرد مانند خطای درصد). از لحاظ

دیگری نیز متفاوت است و آن این است که عملکرد خوب منجر به مقادیر بزرگ ضریب همبستگی می شود در حالیکه در سایر معیارهای خطا، عملکرد خوب با مقادیر کوچک خطا مشخص می شود[27] .

جدول ۴ – ۱ : معیارهای عملکرد برای پیش بینی عددی

Performance measure	Formula
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i (p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\frac{\sum_i (p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i (a_i-\bar{a})^2}{n-1}$

* *p* are predicted values and *a* are actual values.

۴ - ۵ قوانین پیوندی (Associated Rules)

جدول تصمیم گیری(Decision Table) یک طبقه بندی کننده بر اساس اکثریت جدول تصمیم‌گیری می سازد .این الگوریتم، با استفاده از جستجوی اولین بهترین، زیر دسته های ویژگی‌ها را ارزیابی می کند و می تواند از اعتبارسنجی تقاطعی برای ارزیابی بهره ببرد. [55]

یک امکان این است که به جای استفاده از اکثریت جدول تصمیم‌گیری که بر اساس دسته ویژگی‌های مشابه عمل می‌کند، از روش نزدیکترین همسایه برای تعیین طبقه هر یک از نمونه ها که توسط مدخل جدول تصمیم‌گیری پوشش داده نشده اند، استفاده شود.

از جمله کارهایی که کاوش قوانین پیوندی برای ما انجام می دهد می توان به پیدا کردن وابستگی ها و همبستگی ها و همبستگی های موجود در بین داده ها ، یافتن الگوهایی که غالبا در بین داده ها وجود دارند و همچنین پیدا کردن یک سری ساختار سببی در بین آیتم ها و اشیاء موجود در پایگاه داده های تعاملی و رابطه ای اشاره کرد . قبل ازمعرفی الگوریتم های مربوط به کاوش قوانین پیوندی ، نیاز به معرفی یک سری مفاهیم پایه است .

۱. مجموعه آیتم های موجود در یک پایگاه اطلاعاتی با $ItemSet = \{X_1, X_2, \ldots\}$ نمایش داده می شوند .

۲. برای هر قانون به شکل $X <- Y$ است ، دو مقدار Support و Confidence مشخص می شود .

۳. Support احتمال وجود همزمان X و Y به صورت توام در تراکنش است .

۴. Confidence احتمال شرطی است برای آنکه تراکنش دارای X ، دارای Y نیز باشد .

بنابراین قانون $X <- Y$ با (S=50% و C=66.7%) بدین معنی است که X و Y به صورت توام در ۵۰ درصداز کل تراکنش ها وجود دارند و در ۶۶.۷ درصد از تراکنش ها ، هر جا X در تراکنش حضور داشته ، Y نیز حضور داشته است . کاوش قوانین پیوندی در پایگاه داده ها شامل دو مرحله زیر است :

۱. کشف بزرگ ترین مجموعه آیتم ها (مجموعه آیتم هایی که دارای مقدار Support بالاتر از یک مقدار خاص باشند) .

۲. استفاده از مجموعه آیتم های کشف شده در مرحله قبل و ساخت قوانین پیوندی . به طور کلی بیشتر کارها برای بهینه کردن اجرای مرحله اول یعنی کشف بزرگ ترین مجموعه آیتم انجام می گیرد ، زیرا با داشتن بزرگ ترین مجموعه آیتم ، پیدا کردن قوانین به صورت مستقیم ، ممکن می شود . در ادامه به معرفی الگوریتم های مختلف ارائه شده برای کشف بزرگ ترین مجموعه آیتم می پردازیم .

- الگوریتم Apriori

هدف در این الگوریتم ، پیدا کردن بزرگ ترین مجموعه آیتم است که حداقل Support و Confidence را رعایت کند . دو فرض زیر در این الگوریتم در نظر گرفته می شود :

۱. هر زیر مجموعه از یک مجموعه آیتم تکرار شونده ، تکرار شونده است . (یعنی اگر فرضاً مجموعه { a ،b ،c } تکرار شونده باشد ، آنگاه مجموعه { a ،b } نیز تکرار شونده است .)

۲. هر فوق مجموعه از یک مجموعه آیتم تکرار نشونده ، است . (یعنی اگر فرضاً مجموعه { a ،b } تکرار شونده نباشد ، آنگاه مجموعه { a ،b ،c } نیز تکرار شونده نیست .)

الگوریتم Apriori به این صورت است که در هر بار ، یک سری مجموعه آیتم بزرگ با طول K+1 را از روی مجموعه آیتم های کاندید با طول K می سازد و این کار را تا رسیدن به یک مجموعه آیتم با بیشترین طول انجام می دهد . مجموعه آیتم های کاندید در هر دفعه با ضرب مجموعه کاندید در خودش به دست می آید از مشکلات این روش می توان به حجم بسیار بالای تراکنش های موجود در پایگاه داده ، طولانی بودن زمان جست و جوی آنها در هر بار و تعداد زیاد کاندیدها در هر مرحله اشاره کرد . ایده های مطرح شده برای بهینه سازی الگوریتم Apriori عبارتند از :

۱. کاهش تعداد دفعات جست و جو در پایگاه داده تراکنشی .

۲. کاهش تعداد کاندیدها .

۳. ساده کردن شمارش برای Support .

- الگوریتم DHP

این روش مشابه با الگوریتم Apriori بوده و تنها تفاوت آن در ایجاد مجموعه کاندید در هر مرحله است . در روش Apriori مجموعه کاندید ، با ضرب مجموعه آیتم بزرگ به دست آمده تا این مرحله در خودش ، به وجود آمد . اما در روش DHP برای ساخت مجموعه کاندید در هر مرحله ، از یک جدول hash استفاده می شود و تنها یک سری از مجموعه آیتم های موجود در حاصلضرب به عنوان مجموعه کاندید پذیرفته می شود (مجموعه آیتم هایی که دارای Support بالاتری هستند). الگوریتم DHP با استفاده از کاهش تعداد کاندیدها ، الگوریتم Apriori را بهبود می بخشد . روش های دیگری نیز برای بهبود الگوریتم Apriori مطرح شده اند ؛ روش DlC تعداد جست و جو ها را کاهش می دهد ، روش Partition پایگاه داده را به دو قسمت تقسیم کرده و درهر کدام به دنبال بزرگ ترین مجموعه آیتم محلی می گردد و در نهایت بر اساس آنها بزرگ ترین مجموعه آیتم کلی را پیدا می کند .

روش Sampling هر دفعه یک مجموعه آیتم را به عنوان نمونه انتخاب کرده و بزرگ ترین مجموعه آیتم را پیدا می کند و سپس مرزهای مجموعه را بررسی کرده و پایگاه را برای مجموعه آیتم های بزرگ تر ، جست و جو می کند .

- خلاصه سازی و کلی نگری داده ها در سطوح مختلف

یکی دیگر از روش های داده کاوی ، خلاصه سازی و کلی نگری داده ها در سطوح مختلف است . به طور کلی اغلب داده های موجود در پایگاه داده دارای جزئیات فراوانی هستند . برای مثال در پایگاه داده فروش ، رابطه کالا شامل فیلدهای اطلاعاتی نظیر شماره کالا ، نام کالا ، سال ساخت ، قیمت و غیره است . جزئیات زیاد سبب پایین آمدن سطح ادراکی می شود و برای تصمیم گیری بر اساس اطلاعات قبلی ، نیاز به سطوح ادراکی بالاتری است .

داده کاوی با انجام خلاصه سازی و کلی نگری در داده ها در سطوح مختلف به کمک شتافته و سطوح ادراکی بالاتری را ایجاد می کند . در ادامه رهیافت های مختلف ارائه شده برای این کار ، مورد بررسی قرار گرفته است .

۱. رهیافت هرم داده ها

ایده اصلی در این رهیافت ، جمع آوری نتایج محاسبات پرهزینه ای است که اغلب مورد درخواست بوده و نگهداری از آنها در یک ساختار چند بعدی به نام هرم داده ها صورت می گیرد . این محاسبات

معمولا شامل توابعی نظیر مجموع ، میانگین ، ماکزیمم و غیره بر روی مجموعه ویژگی ها خاصه است . هرم داده ها معمولا بر روی پایگاه داده تحلیلی ایجاد می گردد . برای کلی نگری و ویژه نگری داده ها به ترتیب می توان Roll_Up و Drill_Down را بر روی هرم داده ها انجام داد . در بیشتر مواقع هرم داده ها دارای سه بعد بوده که دو بعد آن معمولاً زمان و مکان و بعد دیگر یک آیتم اطلاعاتی است .

۲. رهیافت استنتاج بر اساس ویژگی خاصه

در رهیافت هرم داده ها ، محاسبات بر روی پایگاه داده تحلیلی به صورت offline انجام می گیرد . برای حل این مشکل ، رهیافت استنتاج بر اساس ویژگی خاصه ابتدا یک درخواست داده کاوی را به صورت DMQL مشخص می کند . سپس یک پرس و جو از روی DMQL داده شده می سازد و از پایگاه داده به صورت ONLINE درخواست می کند . سپس بر روی داده های به دست آمده ، تکنیک های کلی نگری داده ها نظیر حذف ویژگی خاصه ، بالا رفتن از درخت ادراک و غیره را اعمال می کند .

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

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

M5Rules، به کمک M5 از روی درخت های مدل، قواعد رگرسیون استخراج می کند.

۴ - ۵ - ۱ کاربرد قوانین پیوندی در وب کاوی

۴ – ۵ – ۱ - ۱ آماده کردن داده ها (preparing the data)

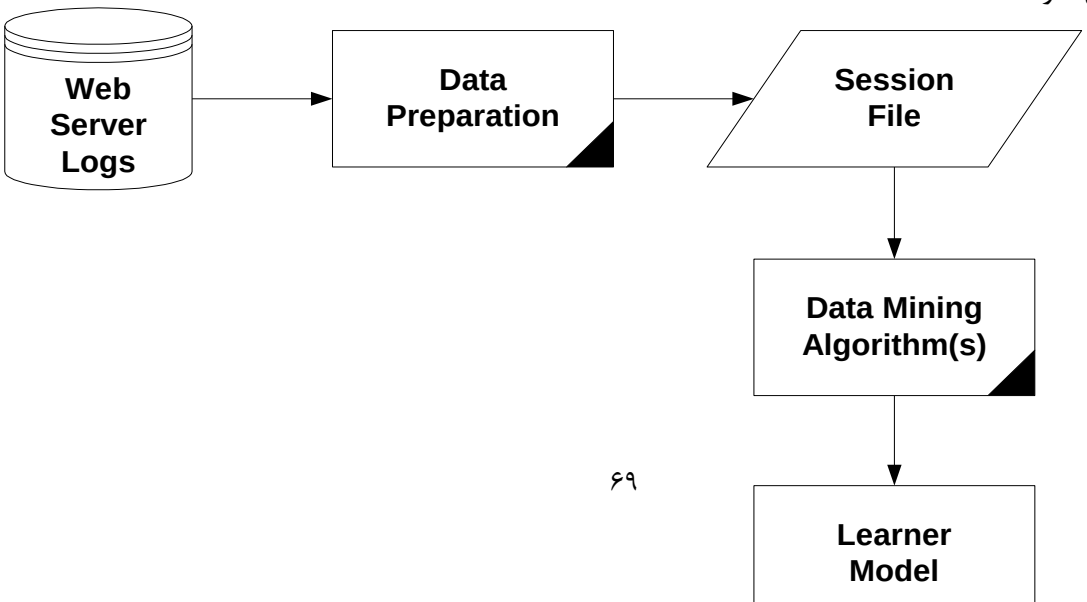
داده های درفایل های لاگ وب سرورها ذخیره شده اند. لاگ فایلهای سرور اطلاعات گسترش یافته مشترک رادر قالب فایل لاگ دراختیار قرار می دهند.

قالب مشترک فایل لاگ :

- آدرس میزبان (Host Address)
- داده / زمان (Date/Time)
- پاسخ (Request)
- وضعیت (Status)
- بایت ها (Bytes)
- صفحه ارجاع کننده (Referring Page)
- نوع مرورگر (Browser Type)

۴ – ۵ – ۱ – ۲ آماده کردن فایل جلسه (preparing the session file)

فایل جلسه فایلی است که بوسیله فرایند آماده سازی داده ایجاد می شود. هرنمونه ازفایل جلسه نشان دهنده جلسه یک کاربر است. هرجلسه کاربر مجموعه ای از درخواست های بازدید صفحه هایی است که توسط یک کاربر از یک وب سرور شده است. هر بازدید صفحه شامل یک یا چند فایل صفحه است که شکل های نمایش پنجره را در یک کاوشگر وب نمایش می دهد. هر باز دید صفحه بوسیله یک url یکتا علامت گذاری می شود.



شکل ۴- ۷ : یک مدل عمومی کاربرد وب

در مورد ایجاد فایل جلسه چند مساله مطرح است:

- ایجادکردن فایل جلسه مشکل است.
- کاربران یک لاگ فایل راشناسایی کنید.
- آدرس های میزبان (host addresses) کمک محدودی می کنند.
- ترکیب آدر س میزبان با صفحه های پالایش شده سودمند است.
- یک درخواست صفحه کاربر ممکن است فایل لاگ دارای ورودی های مختلف ازچند نوع سرور تولید کند.
- ساده ترین حالت هنگامی است که سایت ها مجاز به استفاده از کوکی ها (cookies) هستند.
- تکنیک های سنتی ساخت قوانین انجمنی یا روش های خوشه بندی می توانند به کار برده شوند.
- کاوش پی درپی که بطور خاص توسط الگوریتم های داده کاوی صورت می گیرد برای یافتن فرکانس دستیابی به یک صفحه دریک ترتیب مشخص استفاده می شود.

چهارنمونه بازدید از صفحه رادرنظر بگیرید:

- P5 → P4 → P10→ P3→ P15→ P2→ P1
- P2 → P4 → P10→ P8→ P15→ P4→ P15→ P1
- P4 → P3 → P7→ P11→ P14→ P8 → P2→ P10
- P1 → P3 → P10→ P11→ P4→ P15 → P9

مولد های قوانین پیوندی قانونی مانند موارد زیررا از فایل جلسه تولید می کنند:

IF $\rightarrow P_4 \& P_{10}$

THEN P_{15} {3/4}

این قانون وضعیت p4,p10,p15 راکه درنمونه های فایل جلسه نمایش داده شده اند مشخص می کند. همچنین چهارتا از نمونه ها شامل p4,p10 یک فایل جلسه هستند.

۴ – ۵ – ۱ - ۳ ارزیابی نتایج (خوشه بندی بدون ناظر(unsupervised clustering))

مشابهت نمونه ها بوسیله تقسیم مجموع تعداد صفحات بازدید شده برهر نمونه بازدید شده برای هرنمونه محاسبه می شود. خوشه بندی تجمعی (agglomerative clustering) برای قرار دادن نمونه ها در دسته استفاده شود.

۴ – ۶ ویژگی های منتخب (selected attributes)

وکا این قابلیت را دارد که با تحلیل داده ها، به کاربر در انتخاب ویژگی هایی که بیشترین تاثیر را در نتیجه درست و منطقی دارند کمک کند. این امر از طریق یک انتخاب کننده ویژگی (Attribute Evaluator) و یک روش جستجو (Search Method) انجام می شود. پس از تعیین ویژگی های انتخاب شده ، می توان سایر ویژگی ها را فیلتر کرد و طبقه بندی را با ویژگی ها انتخاب شده انجام داد که قطعا نتیجه بهتری حاصل خواهد شد.

۴ – ۷ نحوه نمایش چگونگی توزیع ویژگی ها ی مختلف در یکدیگر

چگونگی توزیع ویژگی های مختلف در یکدیگر توسط جدولی نمایش داده می شود که محور x و y این جدول می تواند هر یک از ویژگی ها باشد. تعداد این جداول از فرمول $n \times n$ بدست می آید که در آن n تعداد ویژگی ها است.

۴ – ۸ مجموعه آموزش (Training Set)

طبقه بندی کننده می تواند روی همان مجموعه داده ای که از آن آموزش گرفته است (Train) عمل تست (Test) و پیش بینی را انجام دهد. در این پژوهش ما ۶۶٪ داده ها را برای آموزش گرفتن و ۳۴٪ آنها را برای تست گرفتن اختصاص دادیم. در این قسمت به هر نمونه (instance) از داده های مورد تحلیل شماره ای نسبت داده می شود و نتیجه پیشگویی انجام شده با آنچه که واقعا باید باشد مقایسه می گردد. به ازای هر پیشگویی خطا یک علامت + در ستون error کنار آن نمونه گذاشته شده است.

فصل پنجم

تحلیل داده ها

۵-۱ مجموعه داده مورد استفاده

مجموعه داده مورد استفاده ما اطلاعات ۱۵۴ دانشجوی درگیر در فعالیت های سیستم آموزش موسسه آموزش عالی غیرانتفاعی مجازی نور طوبی تهران می باشد که در این بین، ۱۰۶ نفر از آنان دانشجویان موفق و ۴۸ نفر دانشجویان ناموفق بوده‌اند. دانشجویانی که معدل آنها بیشتر مساوی ۱۵ بوده است به عنوان دانشجوی موفق و دانشجویانی که معدل آنها کمتر از ۱۵ بوده است به عنوان دانشجوی ناموفق در نظر گرفته شده‌اند.

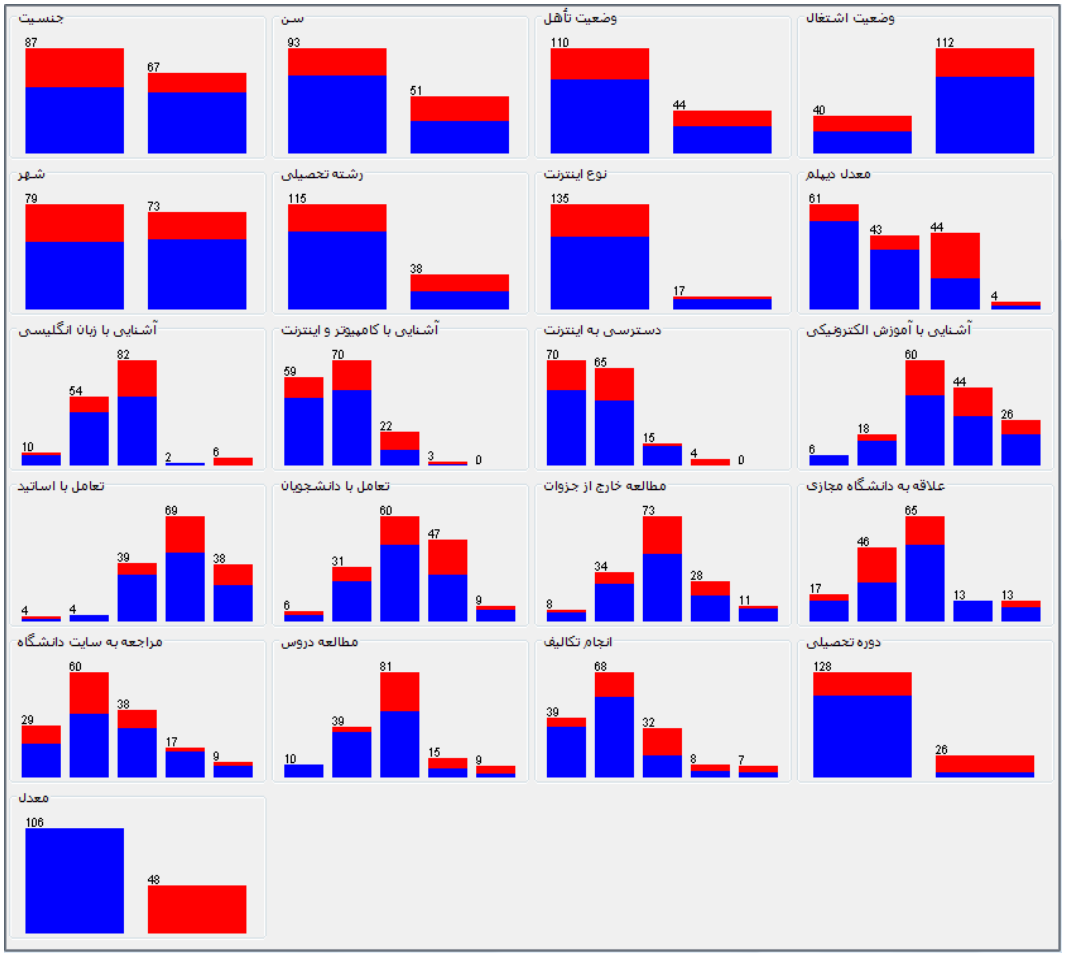
جدول ۵ – ۱ ویژگی های موجود در مجموعه داده و مقادیر معتبر هر یک را نشان می دهد.

جدول ۵- ۱ : ویژگی های موجود در مجموعه داده و مقادیر معتبر آنها

ویژگی	مقادیر
جنسیت	زن، مرد
سن	معمولی، بیش از معمول
وضعیت تأهل	مجرد، متأهل
وضعیت اشتغال	بیکار، شاغل
شهر	تهران- شیراز، دیگر شهرها
رشته تحصیلی	مبتنی بر کامپیوتر، دیگر رشته‌ها
نوع اینترنت	Dial up ،ADSL
معدل دیپلم	۱۴-۰، ۱۶-۱۴، ۱۸-۱۶، ۲۰-۱۸
آشنایی با زبان انگلیسی	خبلی زباد، زباد، متوسط، کم، خیلی کم
آشنایی با کامپیوتر و اینترنت	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
دسترسی به اینترنت	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
آشنایی با آموزش الکترونیکی	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
تعامل با اساتید	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
تعامل با دانشجویان	خیلی زیاد، زباد، متوسط، کم، خیلی کم
مطالعه خارج از جزوات	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
علاقه به دانشگاه مجازی	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
مراجعه به سایت دانشگاه	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
مطالعه دروس	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
انجام تکالیف	خیلی زیاد، زیاد، متوسط، کم، خیلی کم
دوره تحصیلی	کارشناسی ارشد، کارشناسی
معدل	بیشتر مساوی ۱۵، کمتر از ۱۵

سن ۱۸ تا ۲۴ را برای کارشناسی، و سن بین ۲۲ تا ۲۷ را برای کارشناسی ارشد بعنوان سن معمول در نظر گرفته شده است. دانشجویان ساکن شهرهای تهران و شیراز را از آن جهت با دانشجویان ساکن سایر شهرها جدا در نظر گرفتیم که تمرکز موجودیت دانشگاههای مجازی در ایران در این شهرها بیشتر از سایر شهرهاست و تعبیر نمودیم که به این دلیل و همچنین بدلیل میزان آشنایی بیشتر دانشجویان ساکن تهران و شیراز با دانشگاههای مجازی، موفقیت این دانشجویان در آموزش مجازی بیشتر از سایر دانشجویان باشد که البته نادرستی این تعبیر در همان مرحله پیش پردازش داده ها (PreProcess) اثبات شد. همچنین بدلیل مهارت بیشتر دانشجویان رشته های کامپیوتر و فناوری اطلاعات (رشته های مرتبط با کامپیوتر) در کار با ابزارهای آموزش الکترونیک پیش بینی نمودیم که میزان موفقیت این دسته از دانشجویان در آموزش الکترونیک بیشتر از سایر دانشجویان است که درستی این ادعا ثابت شد. همچنین تحلیل نمودیم که دانشجویانی که از اینترنت سرعت بالا (ADSL) استفاده می کنند از موفقیت بیشتری برخوردارند که حداقل این فرضیه در مورد داده های ما و در حوزه این پژوهش رد شد. یکی دیگر از ویژگی هایی که در نظر گرفتیم معدل دیپلم دانشجو بود. این ویژگی در حقیقت پیشینه و سابقه تحصیلی دانشجو را نشان می دهد وتاثیر بسیاری در نتیجه پیش بینی عملکرد تحصیلی دانشجویان از خود نشان داد. برخلاف نوع اینترنت که تاثیر چندانی در پیش بینی نتیجه تحصیلی دانشجویان نداشت، میزان دسترسی به اینترنت بسیار تاثیر گذار بود تا حدی که تمامی دانشجویانی که دسترسی کم به اینترنت داشتند دانشجویان ناموفق بودند. بسیاری از فعالیت های جانبی و اختیاری و بسیاری از کتب معرفی شده توسط اساتید جهت مطالعه بیشتر توسط دانشجویان و همچنین بسیاری از تمرین هایی که اساتید مطرح می کنند در سایت دانشگاه مجازی منتشر می شود و بدین دلیل تعبیر نمودیم که هر چه میزان ملاقات از سایت دانشگاه بیشتر باشد دلیلی بر میزان فعالیت بیشتر دانشجوست و احتمالا او موفق تر است که درستی این ادعا نیز ثابت شد. سایر ویژگی های مطرح شده در جدول ۵ – ۱ نیز هر یک به تعبیری مناسب در جهت پیش بینی نتیجه تحصیلی دانشجویان تحت آموزش مجازی منظور شده اند.

توزیع مقادیر مختلف مربوط به هر کدام از ویژگی ها نیز در شکل ۵-۱ نمایش داده شده است.



شکل ۵ - ۱ : توزیع مقادیر مختلف مربوط به هر کدام از ویژگی ها

با نگاهی اجمالی به نمودار بالا که مرحله پیش پردازش (PreProcess) را نشان می دهد اینطور برداشت می شود که ویژگی های سن ، جنسیت و شهر تاثیر چندانی در نتیجه نخواهند داشت و ویژگی های دوره تحصیلی ، میزان مطالعه دروس و معدل دیپلم بیشترین تاثیر را در نتیجه پیش بینی خواهند داشت. همچنین آخرین نمودار در گوشه سمت چپ پایین گویای این مطلب است که رنگ آبی نشان دهنده میزان موفقیت و رنگ قرمز نشان دهنده میزان عدم موفقیت است.

در شکل۵-۲ نیز نمونه ای از مقادیر مجموعه داده مورد استفاده نشان داده شده است.

sex	age	marital-status	job-status	city	major
female	normal	single	employed	small	computer-based
male	normal	single	employed	small	others
female	normal	single	employed	big	computer-based
male	overage	single	employed	big	others
female	normal	single	employed	small	computer-based
female	overage	married	employed	small	others
male	normal	single	employed	small	others
female	normal	single	employed	small	computer-based
female	normal	single	unemployed	big	others
female	normal	single	unemployed	big	others
male	normal	single	employed	small	computer-based
male	normal	single	employed	small	computer-based
male	normal	single	employed	small	computer-based
female	normal	single	employed	small	computer-based
male	overage	married	employed	big	computer-based
female	normal	single	employed	big	computer-based
male	overage	single	employed	big	computer-based
male	normal	single	unemployed	big	computer-based
female	normal	single	employed	small	computer-based
male	overage	married	employed	big	others
male	overage	married	employed	small	others
male	normal	married	employed	small	others
male	normal	married	employed	small	computer-based

شکل ۵ – ۲ : نمونه ای از مقادیر مجموعه داده مورد استفاده

در شکل ۵ - ۳ نیز بخشی از فایل arff پایگاه داده های مورد نظر نمایش داده شده است.

'relation 'EL@

{'attribute Sex {'male','female@

{'attribute Age {'normal','overage@

{'attribute 'Marital-Status' {'single','married@

{'attribute 'Job-Status' {'unemployed','employed@

{'attribute City {'big','small@

{'attribute Major {'computer-based','others@

{'attribute 'Internet-Type' {'ADSL','dial-up@

{'attribute 'School-Average' {'A','B','C','D@

{'attribute 'English-Proficiency' {'very-high','high','medium','low','very-low@

attribute 'Computer-Internet-Proficiency' {'very-high','high','medium','low','very-@

{'low

{'attribute 'Internet-Accessibility' {'very-high','high','medium','low','very-low@

{'attribute 'ELearning-Familiarity' {'very-high','high','medium','low','very-low@
attribute 'Interaction-with-Professors' {'very-high','high','medium','low','very-@
{'low
{'attribute 'Interaction-with-Students' {'very-high','high','medium','low','very-low@
{'attribute 'Extra-Study' {'very-high','high','medium','low','very-low@
{'attribute 'VU-Interest' {'very-high','high','medium','low','very-low@
{'attribute 'University-Site-Visit' {'very-high','high','medium','low','very-low@
{'attribute 'Study' {'very-high','high','medium','low','very-low@
{'attribute 'Homework' {'very-high','high','medium','low','very-low@
{'attribute 'Period' {'MS','BS@
{'attribute 'Previous-Average' {'A','C@
data@
female,normal,single,employed,small,computer-based,dial-up,A,medium,low,medium,high,low,low,medium,medium,low,medium,high,MS,A
male,normal,single,employed,small,others,ADSL,B,high,very-high,very-high,medium,low,medium,very-low,medium,high,low,medium,MS,A

شکل ۵-۳ : بخشی از فایل arff پایگاه داده های استفاده شده

۵-۲ سیستم عامل مورد استفاده

در تمام موارد پیاده سازی از سیستم عامل ویندوز اکس پی استفاده شده است.

۵-۳ مشخصات سخت افزاری

مشخصات سخت افزاری سیستم مورد استفاده به شرح زیر است:

پردازنده پنتیوم ۴ با سرعت ۲G Hz با حافظه نهان ۵۱۲K و حافظه ۱۰۲۴ MB.

۵-۴ ارزیابی

در این پژوهش ما از روش ارزیابی متقاطع ده - دسته (10 fold) استفاده نمودیم و ۶۶٪ داده ها را برای آموزش دیدن (Train) و ۳۴٪ آنها را برای تست گرفتن (Test) اختصاص دادیم. برای انجام داده‌کاوی نیز روش های درخت تصمیم J48 ، Naïve Bayes ، OneR ، Logistic Regression، و Multi Layer Perceptron (MLP) استفاده شده است .

۵ – ۵ نتایج بدست آمده توسط طبقه بندی کننده ها و الگوریتم های مختلف

۵ – ۵ – ۱ نتایج حاصل از روش درخت تصمیم J48

شکل ۵ – ۴ نتایج بدست آمده توسط روش J48 را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         114               74.026  %
Incorrectly Classified Instances        40               25.974  %
Kappa statistic                          0.3805
Mean absolute error                      0.3022
Root mean squared error                  0.4738
Relative absolute error                 70.2462 %
Root relative squared error            102.2489 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.83      0.458      0.8       0.83      0.815      0.727    A
                 0.542     0.17       0.591     0.542     0.565      0.727    C
Weighted Avg.    0.74      0.368      0.735     0.74      0.737      0.727

=== Confusion Matrix ===

  a  b   <-- classified as
 88 18 |  a = A
 22 26 |  b = C
```

شکل ۵ – ۴ : نتایج ایجاد شده توسط روش J48

همانطور که از شکل ۵ – ۴ پیداست روش درخت تصمیم J48 تعداد ۱۱۴ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۷۴.۰۲۶٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی حاصل از این روش در می یابیم که ۸۸ نمونه (TP) از ۱۱۰ نمونه (TP + FP = ۸۸+۲۲) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس ۸۰٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{88}{110} * 100\% = 80\%$$

همچنین تعداد ۲۶ نمونه (TP) از ۴۴ نمونه (۱۸ + ۲۶ = TP + FP) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با ۵۹.۱٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{26}{44} * 100\% = 59.1\%$$

میزان میانگین وزنی دقت برابر با ۷۳.۵٪ می باشد که جواب نسبتا خوبی است.

تعداد ۱۸ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 18) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر ۸۳٪ می باشد که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{88}{88 + 18} = 83\%$$

و تعداد ۲۲ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 22) بنابراین مقدار یادآوری برای کلاس C برایر است با ۵۴.۲٪ که نمی توان آنرا بعنوان یک جواب خوب در نظر گرفت.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{26}{26 + 22} = 54.2\%$$

همچنین میزان میانگین وزنی برای یادآوری برابر است با ۷۴٪ که جواب نسبتا خوبی است.

میزان F – Measure برای کلاس A برابر است با ۸۱.۵٪ که جواب بسیار خوبی است.

$$\text{F–measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*88}{2*88+22+18} * 100\% = 81.5\ \%$$

میزان F - Measure برای کلاس C برابر است با ۵۶.۵٪ که جواب ضعیفی است .

$$\text{F – Measure}=\frac{2*TP}{2*TP+FP+FN}*100\% = \frac{2*recall*precision}{recall+precision}*100\%$$

$$=\frac{2*26}{2*26+18+22}*100\% = 56.5\ \%$$

میزان میانگین وزنی F – Measure برابر است با ۷۳.۷٪ که جواب نسبتا خوبی است

تا به اینجا متوجه شدیم که این روش برای کلاس دانشجویان موفق (A) بسیار خوب عمل کرده است اما برای کلاس دانشجویان ناموفق (C) نتایج چشمگیری ایجاد نکرده است. درخت‌های تصمیم از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شوند و هدف در این فرآیند افزایش فاصله بین گروه‌ها در هر جداسازی است. درخت‌های تصمیم تعداد دفعات کمی از داده‌ها گذر می‌کنند(برای هر سطح درخت حداکثر یک مرتبه) درنتیجه، مدلها به سرعت ساخته می‌شوند، که آنها را برای مجموعه داده های بسیار مناسب می‌سازد. در کلاس C چون تعداد نمونه ها بسیار کمتر از تعدادنمونه های کلاس A می باشد نتایج ایجاد شده در کلاس C ضعیف تر از نتایج ایجاد شده در کلاس A می باشد.

شکل ۵ – ۵ درخت تصمیم J48 را نمایش می دهد. اندازه درخت مساله مورد تحقیق ما ۴۴ و تعداد برگ های آن ۳۳ می باشد.

```
J48 pruned tree
------------------

Period = MS
|   Sex = male
|   |   VU-Interest = very-high: A (10.0)
|   |   VU-Interest = high
|   |   |   Interaction-with-Students = very-high: C (2.0)
|   |   |   Interaction-with-Students = high: A (4.0)
|   |   |   Interaction-with-Students = medium
|   |   |   |   Study = very-high: A (1.0)
|   |   |   |   Study = high: A (2.0)
|   |   |   |   Study = medium: C (3.0/1.0)
|   |   |   |   Study = low: C (4.0)
|   |   |   |   Study = very-low: C (0.0)
|   |   |   Interaction-with-Students = low
|   |   |   |   Age = normal: A (2.0)
|   |   |   |   Age = overage
|   |   |   |   |   City = big: A (2.0)
|   |   |   |   |   City = small: C (6.0)
|   |   |   Interaction-with-Students = very-low: A (1.0)
|   |   VU-Interest = medium
|   |   |   Interaction-with-Professors = very-high: A (0.0)
|   |   |   Interaction-with-Professors = high: A (2.0)
|   |   |   Interaction-with-Professors = medium: A (10.0)
|   |   |   Interaction-with-Professors = low
|   |   |   |   Computer-Internet-Proficiency = very-high: A (4.0)
|   |   |   |   Computer-Internet-Proficiency = high
|   |   |   |   |   Interaction-with-Students = very-high: C (0.0)
|   |   |   |   |   Interaction-with-Students = high: C (0.0)
|   |   |   |   |   Interaction-with-Students = medium
|   |   |   |   |   |   Extra-Study = very-high: A (0.0)
```

```
|   |   |   Interaction-with-Professors = high: A (2.0)
|   |   |   Interaction-with-Professors = medium: A (10.0)
|   |   |   Interaction-with-Professors = low
|   |   |   |   Computer-Internet-Proficiency = very-high: A (4.0)
|   |   |   |   Computer-Internet-Proficiency = high
|   |   |   |   |   Interaction-with-Students = very-high: C (0.0)
|   |   |   |   |   Interaction-with-Students = high: C (0.0)
|   |   |   |   |   Interaction-with-Students = medium
|   |   |   |   |   |   Extra-Study = very-high: A (0.0)
|   |   |   |   |   |   Extra-Study = high: C (2.0)
|   |   |   |   |   |   Extra-Study = medium: A (2.0)
|   |   |   |   |   |   Extra-Study = low: A (0.0)
|   |   |   |   |   |   Extra-Study = very-low: A (0.0)
|   |   |   |   |   Interaction-with-Students = low: C (6.0)
|   |   |   |   |   Interaction-with-Students = very-low: C (0.0)
|   |   |   |   Computer-Internet-Proficiency = medium: C (2.0)
|   |   |   |   Computer-Internet-Proficiency = low: C (0.0)
|   |   |   |   Computer-Internet-Proficiency = very-low: C (0.0)
|   |   |   Interaction-with-Professors = very-low: A (1.0)
|   |   VU-Interest = low: A (4.0)
|   |   VU-Interest = very-low: A (5.0)
|   Sex = female: A (53.0/4.0)
Period = BS: C (26.0/6.0)

Number of Leaves  :     33

Size of the tree :      44
```

شکل ۵ – ۵ : درخت تصمیم J48 مساله مورد پژوهش

شکل۵ – ۶ نتایج حاصل از training set روش J48 را نشان می دهد. مسلما چون روی همان نمونه هایی که عمل آموزش گرفتن (train) انجام شده عمل تست انجام می شود انتظار نتایج بهتری را داریم که این بدلیل کاهش انواع خطاها می باشد.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         143               92.8571 %
Incorrectly Classified Instances        11                7.1429 %
Kappa statistic                          0.8363
Mean absolute error                      0.1166
Root mean squared error                  0.2415
Relative absolute error                 27.1231 %
Root relative squared error             52.1342 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.934     0.083      0.961     0.934     0.947      0.961    A
                 0.917     0.066      0.863     0.917     0.889      0.961    C
Weighted Avg.    0.929     0.078      0.93      0.929     0.929      0.961

=== Confusion Matrix ===

  a  b   <-- classified as
 99  7 |  a = A
  4 44 |  b = C
```

شکل ۵ – ۶ : نتایج حاصل از training set با روش J48

جدول۵ – ۲ مقایسه ای بین خطاهای ایجاد شده توسط روشهای cross validation و training set حاصل از روش J48 را نشان می دهد. همانطور که انتظار می رفت میزان خطاها در روش training set به مراتب کمتر از روش cross validation است.

جدول ۵ - ۲ : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش J48

Method	Mean Absolute Error	Root Mean Squared	Relative Absolute Error	Root Relative Squared Error
Cross Validation	0.3022	0.4738	70.2462%	102.2489%
Training Set	0.1166	0.2415	27.1231%	52.1342%

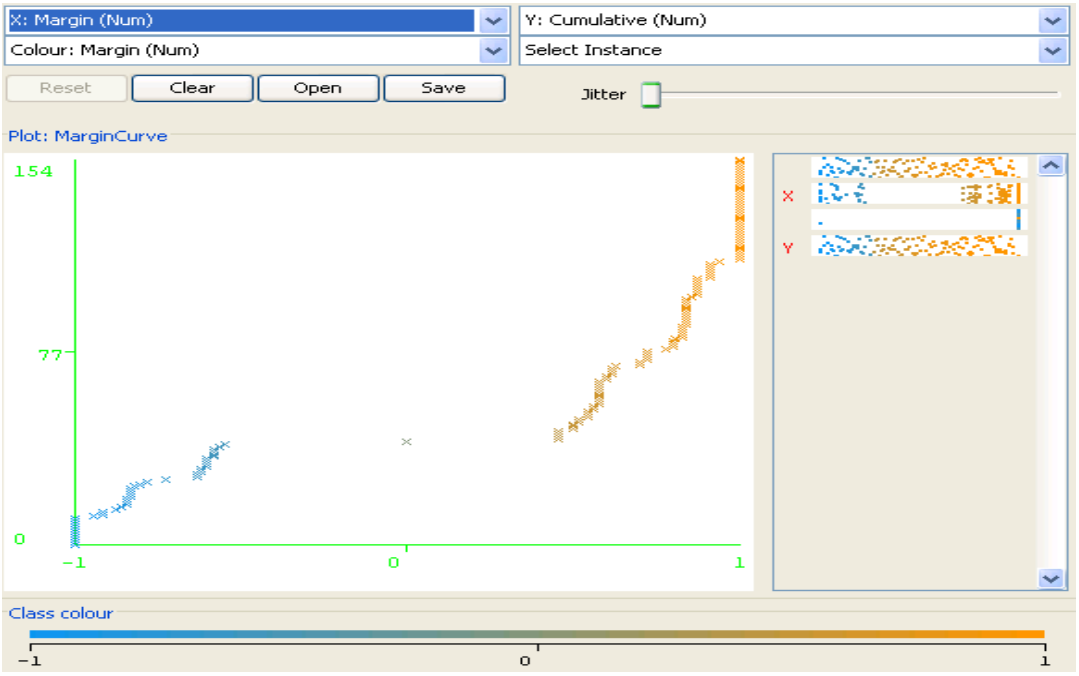
```
=== Predictions on test data ===

 inst#,     actual, predicted, error, probability distribution
     1        2:C        1:A      +  *0.936  0.064
     2        2:C        2:C          0.217 *0.783
     3        2:C        2:C          0.217 *0.783
     4        2:C        2:C          0     *1
     5        2:C        1:A      +  *0.889  0.111
     6        1:A        1:A         *0.936  0.064
     7        1:A        1:A         *0.936  0.064
     8        1:A        1:A         *0.936  0.064
     9        1:A        1:A         *0.936  0.064
    10        1:A        1:A         *1      0
    11        1:A        1:A         *0.936  0.064
    12        1:A        1:A         *0.905  0.095
    13        1:A        1:A         *0.889  0.111
    14        1:A        2:C      +   0.2    *0.8
    15        1:A        2:C      +   0      *1
    16        1:A        2:C      +   0.217 *0.783
     1        2:C        1:A      +  *0.922  0.078
     2        2:C        1:A      +  *0.922  0.078
     3        2:C        1:A      +  *0.922  0.078
     4        2:C        2:C          0.25  *0.75
     5        2:C        2:C          0.25  *0.75
     6        1:A        1:A         *0.956  0.044
     7        1:A        1:A         *0.922  0.078
     8        1:A        1:A         *0.922  0.078
     9        1:A        1:A         *0.922  0.078
    10        1:A        1:A         *0.922  0.078
    11        1:A        1:A         *0.922  0.078
    12        1:A        1:A         *1      0
    13        1:A        1:A         *1      0
    14        1:A        1:A         *0.922  0.078
    15        1:A        1:A         *0.922  0.078
    16        1:A        1:A         *0.922  0.078
     1        2:C        2:C          0     *1
     2        2:C        2:C          0     *1
     3        2:C        1:A      +  *0.957  0.043
     4        2:C        2:C          0     *1
     5        2:C        1:A      +  *0.957  0.043
     6        1:A        1:A         *0.75   0.25
     7        1:A        2:C      +   0      *1
     8        1:A        1:A         *0.957  0.043
     9        1:A        1:A         *1      0
    10        1:A        1:A         *0.957  0.043
    11        1:A        1:A         *0.957  0.043
```

شکل ۵ - ۷ : بخشی از نمودار حاصل از پیش بینی J48 روی نمونه های تست با روش Cross Validation

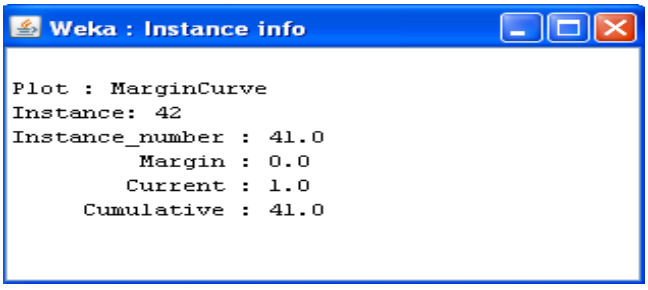
```
=== Predictions on training set ===

 inst#,     actual, predicted, error, probability distribution
     1        1:A        1:A          *0.925  0.075
     2        1:A        1:A          *1      0
     3        1:A        1:A          *0.925  0.075
     4        1:A        1:A          *1      0
     5        1:A        1:A          *0.925  0.075
     6        1:A        1:A          *0.925  0.075
     7        1:A        1:A          *1      0
     8        1:A        1:A          *0.925  0.075
     9        2:C        2:C           0.231 *0.769
    10        2:C        2:C           0.231 *0.769
    11        1:A        1:A          *1      0
    12        1:A        1:A          *1      0
    13        1:A        1:A          *1      0
    14        1:A        1:A          *0.925  0.075
    15        1:A        1:A          *1      0
    16        1:A        1:A          *0.925  0.075
    17        1:A        1:A          *1      0
    18        2:C        2:C           0.231 *0.769
    19        1:A        1:A          *0.925  0.075
    20        2:C        2:C           0     *1
    21        1:A        1:A          *1      0
    22        2:C        2:C           0.333 *0.667
    23        1:A        1:A          *1      0
    24        2:C        2:C           0     *1
    25        1:A        1:A          *1      0
    26        1:A        1:A          *0.925  0.075
    27        1:A        1:A          *1      0
    28        1:A        1:A          *0.925  0.075
    29        1:A        1:A          *1      0
    30        1:A        1:A          *0.925  0.075
    31        1:A        1:A          *0.925  0.075
    32        2:C        2:C           0     *1
    33        1:A        1:A          *0.925  0.075
    34        1:A        1:A          *0.925  0.075
    35        1:A        1:A          *1      0
    36        2:C        2:C           0.231 *0.769
    37        1:A        1:A          *0.925  0.075
    38        1:A        1:A          *0.925  0.075
    39        1:A        1:A          *1      0
    40        1:A        1:A          *1      0
    41        1:A        1:A          *0.925  0.075
    42        1:A        1:A          *1      0
    43        1:A        1:A          *0.925  0.075
```

شکل ۵ - ۸ : بخشی از نمودار حاصل از Training Set با روش J48

شکل۵ - ۹ نمودار اختلاف (margin curve) بدست آمده توسط روش J48 را نشان می دهد. هر چه مقدار margin برای هر نمونه به یک نزدیکتر باشد به معنای عملکرد بهتر طبقه بندی کننده می باشد. در شکل۵ - ۹ می بینیم یک نمونه دارای margin برابر صفر می باشد(نمونه شماره ۴۱). ما این اطلاعات را از Instance Information کسب نمودیم. از آنجایی که نمونه ها از کمترین اختلاف تا بیشترین اختلاف ذخیره می گردند (شماره نمونه و cumulative یکسان است) تعداد ۱۱۴ نمونه (154-41+1=114) دارای margin بیشتر از صفر بوده اند و این نتیجه نسبتا مطلوبی است.(این موضوع توسط شکل ۵ – ۴ نیز اثبات می شود.) شکل۵ - ۱۰ نیز اطلاعات مربوط به نمونه ای که margin آن صفر است را نشان می دهد.

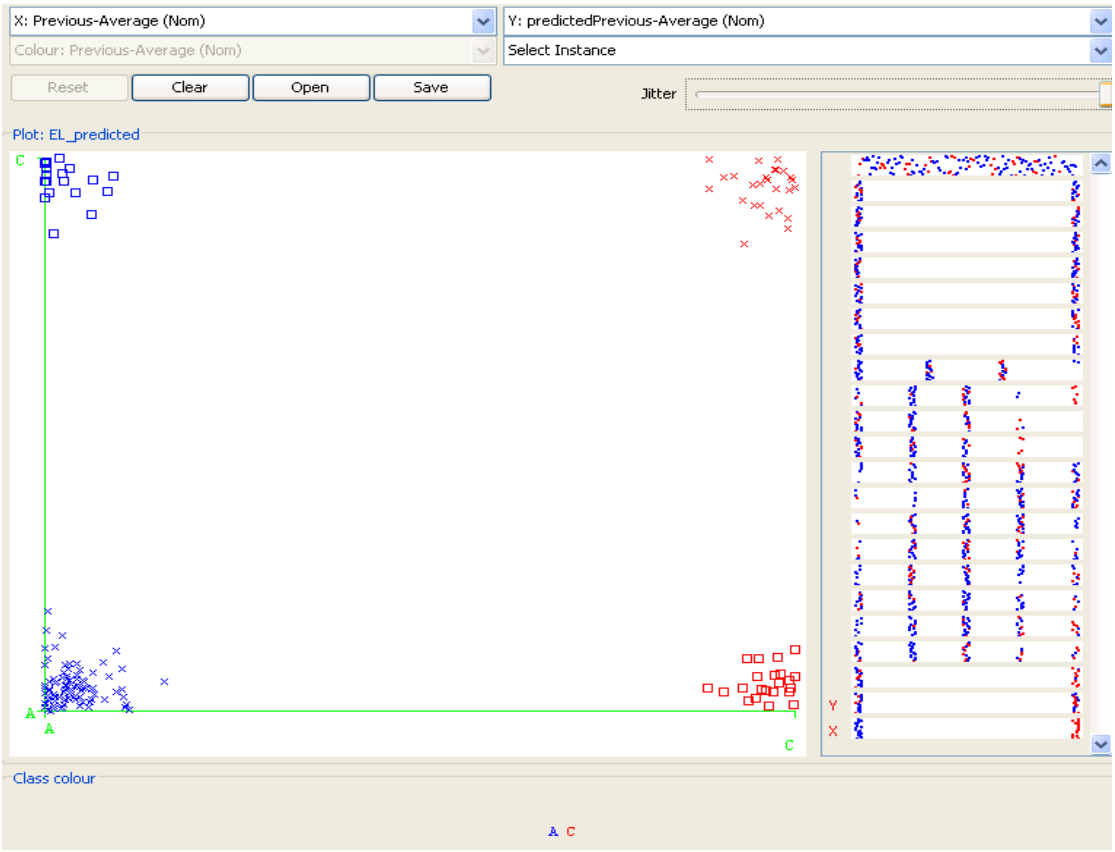


شکل ۵ – ۹ : نمودار اختلاف(Margin) حاصل از روش J48



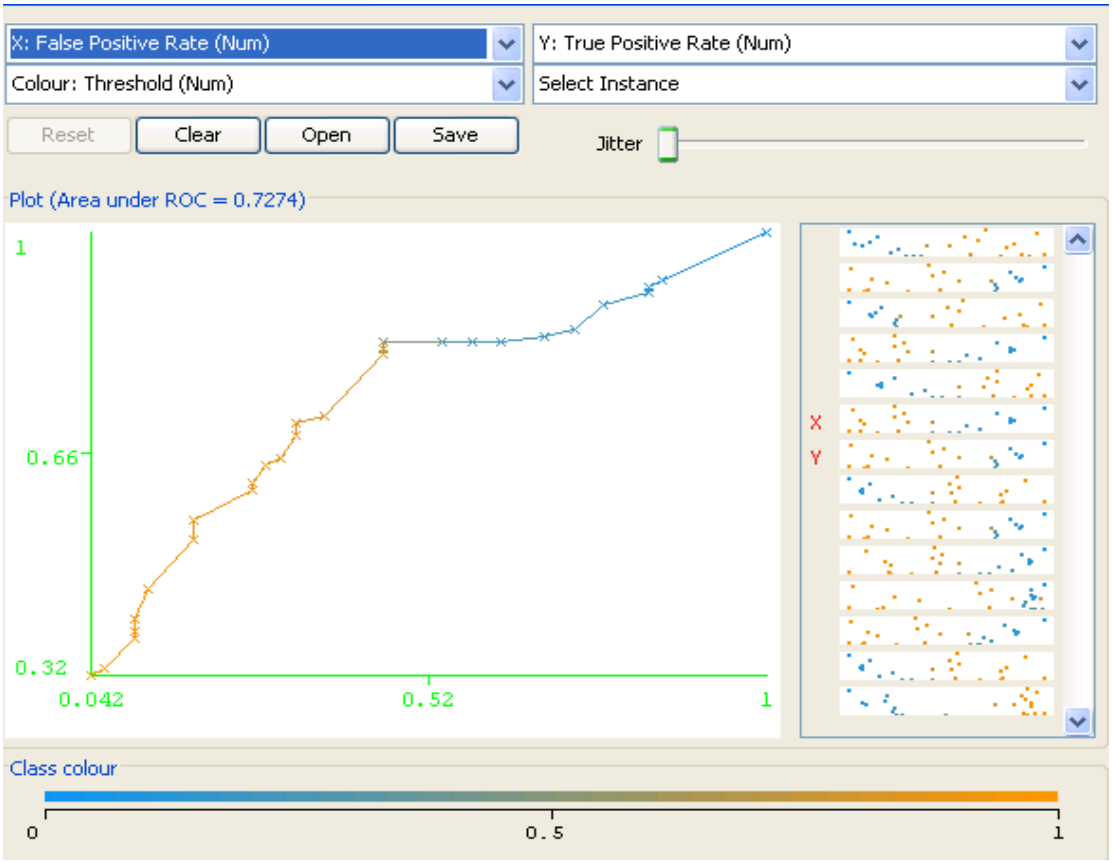
شکل ۵ - ۱۰ : اطلاعات نمونه دارای اختلاف (margin) صفر

شکل ۵ - ۱۱ نمودار خطاهای طبقه بند ی کننده J48 را نشان می دهد. محور X این نمودار بیانگر کلاس طبقه بندی (Previous Average) و محور Y این نمودار بیانگر پیش بینی صورت گرفته برای این کلاس (Predicted Previous Average) می باشد. نقاطی که با ضربدر آبی مشخص شده اند بیانگر نمونه هایی هستند که به درستی به کلاس A طبقه بندی شده اند و در حقیقت TP کلاس A را بدست می آورند. نقاطی که با ضربدر قرمز مشخص شده اند بیانگر نمونه هایی هستند که به درستی به کلاس C طبقه بندی شده اند و TP کلاس C را حاصل می کنند. نقاطی که با مربع کوچک آبی متمایز شده اند مبین نمونه هایی هستند که در واقع متعلق به کلاس A بوده اند اما بطور اشتباه به کلاس C طبقه بندی شده اند(FN کلاس A) و نقاطی که با مربع کوچک قرمز نشان داده شده اند نشانگر نمونه هایی هستند که در واقع متعلق به کلاس C بوده اند اما بطور اشتباه در کلاس A قرار گرفته اند(FP کلاس A و همچنین FN کلاس C). پس در حالت کلی علامت های مربع کوچک نشان دهنده نوعی خطا در پیش بینی طبقه بندی کننده هستند. با توجه به اینکه این علامات بیشتر قرمزند نتیجه می گیریم که عملکرد طبقه بندی کننده J48 برای کلاس C ضعیف تر بوده است.



شکل ۵ – ۱۱ : نمودار خطاهای طبقه بندی کننده J48

شکل۵ - ۱۲ منحنی ROC بدست آمده توسط روش J48 را نشان می دهد.



شکل ۵ – ۱۲ : نمودار ROC حاصل از روش J48

قطر فرعی این محور بیانگر عملکرد تصادفی (Random Performance) است یعنی اگر نمونه ای روی این قطر در گراف قرار گیرد به معنای این است که طبقه بندی کننده اطلاعاتی در مورد کلاس آن نمونه نداشته است. هر چه نمونه ها بیشتر در سمت شمال غرب نمودار باشند عملکرد طبقه بندی کننده بهتر بوده است (بدلیل TP Rate بالاتر و FP Rate پایینتر) . هر چه نمونه ها در گوشه سمت چپ پایین و نزدیک به محور X باشند یعنی استراتژی طبقه بندی کننده در برخورد با نمونه ها یک استراتژی محافظه کارانه (Conservative) بوده است چون در این حالت اگرچه نرخ مثبت کاذب را پایین آورده است اما بدنبال آن نرخ مثبت درست هم پایین آمده است . هر چه نمونه ها در گوشه سمت راست بالا متمرکزتر باشند به معنی برخورد لیبرالی (Libral) طبقه بندی کننده است چون اگرچه نرخ مثبت درست را بالا برده است اما بدنبال آن نرخ مثبت کاذب هم بالا رفته است. اگر نمونه هایی داشته باشیم که در مثلث زیر قطر فرعی قرار بگیرند به معنی این است که طبقه بندی کننده در طبقه بندی آنها عملکردی ضعیفتر از عملکرد تصادفی داشته است که

خوشبختانه در نمودار بالا هیچ نمونه ای در این وضعیت نیست. در مورد نمونه های بالا می توان گفت که طبقه بندی کننده در طبقه بندی آنها بیشتر رویکرد محافظه کارانه داشته است. از آنجایی که اکثر نمونه ها در بالای قطر فرعی قرار گرفته اند می توان گفت عملکرد طبقه بندی کننده نسبتا خوب بوده است.

بیشترین مقدار صحت در (0.45,0.83) اتفاق افتاده که این امر اثبات می کند که عملکرد طبقه بندی کننده برای کلاس A بهتر بوده است. همچنین بیشترین صحت در Threshold=0.5 اتفاق افتاده که نشان می دهد توزیع کاملا متوازن (Balanced) بوده است. لازم به یادآوری است که نمودار ROC مستقل از توزیع کلاس است یعنی اگر نسبت TP و FP تغییر کند نمودار ROC تغییر نمی کند چون نمودار ROC وابسته به FP Rate و TP Rate است نه به FP و TP .

همچنین سطح زیر نمودار ROC (AUC) به عنوان معیاری در عملکرد طبقه بندی کننده در نظر گرفته می شود. در اینجا سطح زیر نمودار ROC برای هر دو کلاس برابر 0.7274 می باشد که نشان می دهد عملکرد آن خوب بوده است.

5 – ۵ – ۲ نتایج حاصل از روش Naïve Bayes

شکل ۵ – ۱۳ نتایج حاصل از روش Naïve Bayes را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         115               74.6753 %
Incorrectly Classified Instances        39               25.3247 %
Kappa statistic                          0.3995
Mean absolute error                      0.2661
Root mean squared error                  0.4176
Relative absolute error                 61.8556 %
Root relative squared error             90.13   %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.83      0.438      0.807     0.83      0.819      0.825    A
                 0.563     0.17       0.6       0.563     0.581      0.825    C
Weighted Avg.    0.747     0.354      0.743     0.747     0.744      0.825

=== Confusion Matrix ===

  a  b   <-- classified as
 88 18 |  a = A
 21 27 |  b = C
```

شکل ۵ – ۱۳ : نتایج ایجاد شده توسط روش Naïve Bayes

همانطور که از شکل ۵ – ۱۳ پیداست روش Naïve Bayes تعداد ۱۱۵ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۷۴.۶۷۵۳٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی حاصل از این روش در می یابیم که ۸۸ نمونه (TP) از 109 نمونه (TP + FP = ۸۸+21) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس ۸۰.۷٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{88}{109} * 100\% = 80.7\%$$

همچنین تعداد ۲۷ نمونه (TP) از ۴۵ نمونه (۱۸ + ۲۷ = TP + FP) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با ۶۰٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{27}{45} * 100\% = 60\%$$

میزان میانگین وزنی دقت برابر با ۷۴.۴٪ می باشد که نسبت به میانگین وزنی بدست آمده توسط روش J48 جواب بهتری است.

تعداد ۱۸ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 18) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر ۸۳٪ می باشد که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{88}{88 + 18} = 83\%$$

و تعداد ۲۱ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 21) بنابراین مقدار یادآوری برای کلاس C برایر است با ۵۶.۳٪ که نمی توان آنرا بعنوان یک جواب خوب در نظر گرفت .

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{27}{27 + 21} = 56.3\%$$

همچنین میزان میانگین وزنی برای یادآوری برابر است با ۷۴.۷٪ که جواب نسبتا خوبی است و نسبت به میانگین وزنی روش J48 نیز بهتر می باشد.

میزان F – Measure برای کلاس A برابر است با ۸۱.۹٪ که جواب بسیار خوبی است.

$$\text{F–measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*88}{2*88+21+18} * 100\% = 81.9\ \%$$

میزان F - Measure برای کلاس C برابر است با ۵۶.۵٪ که جواب ضعیفی است .

$$\text{F-Measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*27}{2*27+18+21} * 100\% = 56.5\ \%$$

میزان میانگین وزنی F – Measure برابر است با ۷۴.۴٪که جواب نسبتا خوبی است و نسبت به میانگین وزنی بدست آمده توسط روش J48 بهتر می باشد. سطح زیر نمودار ROC برای هر دو کلاس برابر ۸۲.۵ است و از آنجا که یکی از روشهای ارزیابی عملکرد طبقه بندی کننده همین سطح زیر نمودار ROC می باشد نتیجه می گیریم عملکرد طبقه بندی کننده Naïve Bayes خوب بوده است.

شکل ۵ – ۱۴ نتایج حاصل از training set با روش Naïve Bayes را نشان می دهد.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         125               81.1688 %
Incorrectly Classified Instances        29               18.8312 %
Kappa statistic                          0.5636
Mean absolute error                      0.2002
Root mean squared error                  0.3354
Relative absolute error                 46.5574 %
Root relative squared error             72.4012 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.858     0.292      0.867     0.858     0.863      0.922    A
                 0.708     0.142      0.694     0.708     0.701      0.922    C
Weighted Avg.    0.812     0.245      0.813     0.812     0.812      0.922

=== Confusion Matrix ===

  a  b   <-- classified as
 91 15 |  a = A
 14 34 |  b = C
```

شکل ۵ – ۱۴ : نتایج حاصل از training set با روش Naïve Bayes

مسلما چون روی همان نمونه هایی که عمل آموزش گرفتن (train) انجام شده عمل تست انجام می شود انتظار نتایج بهتری را داریم که این بدلیل کاهش انواع خطاها می باشد. جدول ۵ – ۳ به مقایسه ای اجمالی بین خطاهای موجود در دو روش cross validation و training set می پردازد.

جدول ۵ - ۳ : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش Naïve Bayes

Method	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Cross validation	0.2661	0.4176	61.8556%	90.13%
Training set	0.2002	0.3354	46.5574%	72.4012%

=== Predictions on test data ===

inst#,	actual,	predicted,	error,	probability distribution	
1	2:C	1:A	+	*0.932	0.068
2	2:C	2:C		0.221	*0.779
3	2:C	2:C		0	*1
4	2:C	1:A	+	*0.746	0.254
5	2:C	2:C		0.019	*0.981
6	1:A	1:A		*0.991	0.009
7	1:A	1:A		*0.861	0.139
8	1:A	1:A		*0.995	0.005
9	1:A	1:A		*0.986	0.014
10	1:A	1:A		*0.997	0.003
11	1:A	1:A		*0.991	0.009
12	1:A	1:A		*0.996	0.004
13	1:A	2:C	+	0.268	*0.732
14	1:A	1:A		*0.953	0.047
15	1:A	1:A		*0.931	0.069
16	1:A	2:C	+	0.13	*0.87
1	2:C	2:C		0.317	*0.683
2	2:C	1:A	+	*0.677	0.323
3	2:C	2:C		0.247	*0.753
4	2:C	2:C		0.23	*0.77
5	2:C	2:C		0.23	*0.77
6	1:A	2:C	+	0.462	*0.538
7	1:A	1:A		*0.996	0.004
8	1:A	1:A		*0.999	0.001
9	1:A	1:A		*0.989	0.011
10	1:A	1:A		*0.976	0.024
11	1:A	1:A		*0.974	0.026
12	1:A	1:A		*0.99	0.01
13	1:A	1:A		*0.927	0.073
14	1:A	1:A		*0.978	0.022
15	1:A	2:C	+	0.497	*0.503
16	1:A	1:A		*0.996	0.004
1	2:C	1:A	+	*0.509	0.491
2	2:C	2:C		0.123	*0.877
3	2:C	1:A	+	*0.964	0.036
4	2:C	2:C		0.397	*0.603
5	2:C	1:A	+	*0.908	0.092
6	1:A	1:A		*0.549	0.451
7	1:A	2:C	+	0.247	*0.753
8	1:A	1:A		*0.971	0.029
9	1:A	1:A		*0.979	0.021
10	1:A	1:A		*0.994	0.006
11	1:A	1:A		*0.998	0.002

شکل ۵ – ۱۵ : بخشی از نمودار حاصل از پیش بینی Naïve Bayes روی نمونه های تست با روش Cross Validation

```
=== Predictions on training set ===

 inst#,     actual, predicted, error, probability distribution
     1        1:A        1:A          *0.984  0.016
     2        1:A        1:A          *0.769  0.231
     3        1:A        1:A          *0.995  0.005
     4        1:A        1:A          *0.992  0.008
     5        1:A        1:A          *0.997  0.003
     6        1:A        1:A          *0.995  0.005
     7        1:A        1:A          *0.989  0.011
     8        1:A        1:A          *0.946  0.054
     9        2:C        2:C           0      *1
    10        2:C        2:C           0      *1
    11        1:A        1:A          *0.998  0.002
    12        1:A        2:C      +    0.471 *0.529
    13        1:A        2:C      +    0.471 *0.529
    14        1:A        1:A          *0.98   0.02
    15        1:A        1:A          *0.67   0.33
    16        1:A        1:A          *0.998  0.002
    17        1:A        1:A          *0.968  0.032
    18        2:C        2:C           0.036 *0.964
    19        1:A        1:A          *1      0
    20        2:C        1:A      +   *0.67   0.33
    21        1:A        1:A          *0.917  0.083
    22        2:C        1:A      +   *0.648  0.352
    23        1:A        2:C      +    0.35  *0.65
    24        2:C        2:C           0.087 *0.913
    25        1:A        1:A          *0.52   0.48
    26        1:A        1:A          *0.977  0.023
    27        1:A        2:C      +    0.462 *0.538
    28        1:A        1:A          *0.976  0.024
    29        1:A        1:A          *0.828  0.172
    30        1:A        1:A          *0.997  0.003
    31        1:A        1:A          *0.905  0.095
    32        2:C        2:C           0.01  *0.99
    33        1:A        1:A          *0.999  0.001
    34        1:A        2:C      +    0.279 *0.721
    35        1:A        2:C      +    0.477 *0.523
    36        2:C        2:C           0.049 *0.951
    37        1:A        1:A          *0.997  0.003
    38        1:A        1:A          *0.882  0.118
    39        1:A        1:A          *0.985  0.015
    40        1:A        1:A          *0.995  0.005
    41        1:A        1:A          *0.999  0.001
    42        1:A        1:A          *0.986  0.014
    43        1:A        1:A          *0.997  0.003
```

شکل ۵ – ۱۶ : بخشی از نمودار حاصل از Training Set روش Naïve Bayes

نمودارهایmargin و خطاهای طبقه بندی کننده و ROC حاصل از روش Naïve Bayes در پیوست یک آمده است.

۵ – ۵ – ۳ نتایج حاصل از روش OneR

شکل ۵ – ۱۷ نتایج ایجاد شده توسط روش OneR را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         120               77.9221 %
Incorrectly Classified Instances        34               22.0779 %
Kappa statistic                          0.4117
Mean absolute error                      0.2208
Root mean squared error                  0.4699
Relative absolute error                 51.3191 %
Root relative squared error            101.4117 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.943     0.583      0.781     0.943     0.855      0.68     A
                 0.417     0.057      0.769     0.417     0.541      0.68     C
Weighted Avg.    0.779     0.419      0.778     0.779     0.757      0.68

=== Confusion Matrix ===

   a   b   <-- classified as
 100   6 |   a = A
  28  20 |   b = C
```

شکل ۵ – ۱۷ : نتایج ایجاد شده توسط روش OneR

همانطور که شکل ۵ – ۱۷ نشان می دهد، روش Naïve Bayes تعداد ۱۲۰ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۷۷.۹۲۲۱٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی حاصل از این روش در می یابیم که ۱۰۰ نمونه (TP) از ۱۲۸ نمونه (۲۸ + ۱۰۰ = TP + FP) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس ۷۸.۱٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{100}{100+28} * 100\% = 78.1\%$$

همچنین تعداد ۲۰ نمونه (TP) از ۲۶ نمونه (۶ + ۲۰ = TP + FP) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با ۷۶.۹٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{20}{20+6} * 100\% = 76.9\%$$

میزان میانگین وزنی دقت برابر با ۷۷.۸٪ می باشد که نسبت به میانگین وزنی بدست آمده توسط روش J48 و Naïve Bayesجواب بهتری است.

تعداد ۶ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 6) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر 94.3% می باشد که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{100}{100 + 6} = 94.3\%$$

و تعداد ۲۸ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 28) بنابراین مقدار یادآوری برای کلاس C برایر است با ۴۱.۷٪که جواب ضعیفی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{20}{20 + 28} = 41.7\%$$

همچنین میزان میانگین وزنی برای یادآوری برابر است با ۷۷.۹٪که جواب نسبتا خوبی است و نسبت به میانگین وزنی روش های قبلی نیز بهتر می باشد.

میزان F – Measure برای کلاس A برابر است با ۸۵.۵٪ که جواب بسیار خوبی است.

$$\text{F–measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*100}{2*100+28+6} * 100\% = 85.5\ \%$$

میزان F - Measure برای کلاس C برابر است با ۵۴.۱٪ که جواب ضعیفی است .

$$\text{F – measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\%$$

$$= \frac{2*20}{2*20+6+28} * 100\% = 54.1\ \%$$

میزان میانگین وزنی F – Measure برابر است با ۷۵.۷٪ که جواب نسبتا خوبی است و نسبت به میانگین وزنی بدست آمده توسط روش های قبلی نیز بهتر می باشد اما سطح زیر نمودار ROC برای هر دو کلاس برابر ۶۸٪ است و این مقدار کمتر از سطح زیر نمودار دو روش قبلی می باشد.

شکل ۵ – ۱۸ نتایج بدست آمده از Training Set با روش OneR را نشان می دهد.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         120               77.9221 %
Incorrectly Classified Instances        34               22.0779 %
Kappa statistic                          0.4117
Mean absolute error                      0.2208
Root mean squared error                  0.4699
Relative absolute error                 51.3456 %
Root relative squared error            101.4426 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.943     0.583      0.781     0.943     0.855      0.68     A
                 0.417     0.057      0.769     0.417     0.541      0.68     C
Weighted Avg.    0.779     0.419      0.778     0.779     0.757      0.68

=== Confusion Matrix ===

   a   b   <-- classified as
 100   6 |   a = A
  28  20 |   b = C
```

شکل ۵ – ۱۸ : نتایج حاصل از Training Set با روش OneR

همانطور که مشاهده می کنید در روش OneR نتایج حاصل از Training Set با نتایج حاصل از Cross Validation تفاوتی ندارد.

جدول ۵ – ۴ : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش OneR

Method	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Cross validation	0.2208	0.4699	51.3191%	101.4117

Training set	0.2208	0.4699	51.3456%	101.4426

```
=== Predictions on test data ===

 inst#,     actual, predicted, error, probability distribution
     1        2:C        1:A      +  *1      0
     2        2:C        2:C         0     *1
     3        2:C        2:C         0     *1
     4        2:C        1:A      +  *1      0
     5        2:C        1:A      +  *1      0
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        1:A         *1      0
    14        1:A        1:A         *1      0
    15        1:A        1:A         *1      0
    16        1:A        2:C      +   0     *1
     1        2:C        1:A      +  *1      0
     2        2:C        1:A      +  *1      0
     3        2:C        1:A      +  *1      0
     4        2:C        2:C         0     *1
     5        2:C        2:C         0     *1
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        1:A         *1      0
    14        1:A        1:A         *1      0
    15        1:A        1:A         *1      0
    16        1:A        1:A         *1      0
     1        2:C        2:C         0     *1
     2        2:C        2:C         0     *1
     3        2:C        1:A      +  *1      0
     4        2:C        2:C         0     *1
     5        2:C        1:A      +  *1      0
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
```

شکل ۵ – ۱۹ : بخشی از نمودار حاصل از پیش بینی OneR روی نمونه های تست با روش Cross Validation

```
=== Predictions on training set ===

 inst#,     actual, predicted, error, probability distribution
      1        1:A        1:A         *1      0
      2        1:A        1:A         *1      0
      3        1:A        1:A         *1      0
      4        1:A        1:A         *1      0
      5        1:A        1:A         *1      0
      6        1:A        1:A         *1      0
      7        1:A        1:A         *1      0
      8        1:A        1:A         *1      0
      9        2:C        2:C          0     *1
     10        2:C        2:C          0     *1
     11        1:A        1:A         *1      0
     12        1:A        1:A         *1      0
     13        1:A        1:A         *1      0
     14        1:A        1:A         *1      0
     15        1:A        1:A         *1      0
     16        1:A        1:A         *1      0
     17        1:A        1:A         *1      0
     18        2:C        2:C          0     *1
     19        1:A        1:A         *1      0
     20        2:C        1:A      +  *1      0
     21        1:A        1:A         *1      0
     22        2:C        1:A      +  *1      0
     23        1:A        1:A         *1      0
     24        2:C        1:A      +  *1      0
     25        1:A        1:A         *1      0
     26        1:A        1:A         *1      0
     27        1:A        1:A         *1      0
     28        1:A        1:A         *1      0
     29        1:A        1:A         *1      0
     30        1:A        1:A         *1      0
     31        1:A        1:A         *1      0
     32        2:C        1:A      +  *1      0
     33        1:A        1:A         *1      0
     34        1:A        1:A         *1      0
     35        1:A        1:A         *1      0
     36        2:C        2:C          0     *1
     37        1:A        1:A         *1      0
     38        1:A        1:A         *1      0
     39        1:A        1:A         *1      0
     40        1:A        1:A         *1      0
     41        1:A        1:A         *1      0
     42        1:A        1:A         *1      0
     43        1:A        1:A         *1      0
```

شکل ۵ – ۲۰ : بخشی از نمودار حاصل از Training Set با روش OneR

نمودارهایmargin و خطاهای طبقه بندی کننده و ROC حاصل از روش OneR در پیوست دو آمده است.

۵ – ۵ – ۴ نتایج حاصل از روش Logistic

شکل ۵ – ۲۱ نتایج حاصل از روش Logistic را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         122               79.2208 %
Incorrectly Classified Instances        32               20.7792 %
Kappa statistic                          0.574
Mean absolute error                      0.214
Root mean squared error                  0.4557
Relative absolute error                 49.7416 %
Root relative squared error             98.3522 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.736     0.083      0.951     0.736     0.83       0.753    A
                 0.917     0.264      0.611     0.917     0.733      0.761    C
Weighted Avg.    0.792     0.14       0.845     0.792     0.8        0.755

=== Confusion Matrix ===

  a  b   <-- classified as
 78 28 |  a = A
  4 44 |  b = C
```

شکل ۵ – ۲۱ : نتایج ایجاد شده توسط روش Logistic

همانطور که شکل ۵ – ۳۳ نشان می دهد، روش Logistic تعداد ۱۲۲ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۷۹.۲۲۰۸٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی حاصل از این روش در می یابیم که ۷۸ نمونه (TP) از ۸۲ نمونه (TP + FP = 78 + 4) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس 95.1% است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{78}{78+4} * 100\% = 95.1\%$$

همچنین تعداد ۴۴ نمونه (TP) از ۷۲ نمونه (۲۸ + ۴۴ = TP + FP) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با 61% است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{44}{44+28} * 100\% = 61\%$$

میزان میانگین وزنی دقت برابر با ۸۴.۵٪ می باشد که نسبت به میانگین وزنی بدست آمده توسط روش های قبلی جواب بهتری است.

تعداد ۲۸ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 28) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر ۷۳.۶٪ می باشد که جواب خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{78}{78 + 28} = 73.6\%$$

و تعداد ۴ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 4) بنابراین مقدار یادآوری برای کلاس C برایر است با ۹۱.۷٪که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{44}{44 + 4} = 91.7\%$$

همچنین میزان میانگین وزنی یادآوری برابر است با ۷۹.۲٪که جواب نسبتا خوبی است و نسبت به میانگین وزنی روش های قبلی نیز بهتر می باشد.

میزان F – Measure برای کلاس A برابر است با ۸۳٪ که جواب بسیار خوبی است.

$$\text{F–measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*78}{2*78+4+28} * 100\% = 83\%$$

میزان F - Measure برای کلاس C برابر است با ۷۳.۳٪ که جواب خوبی است .

$$\text{F – measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\%$$

$$= \frac{2*44}{2*44+28+4} * 100\% = 73.3\ \%$$

میزان میانگین وزنی F – Measure برابر است با ۸۰٪ که جواب بسیار خوبی است و نسبت به میانگین وزنی بدست آمده توسط روش های قبلی نیز بهتر می باشد.

اما نکته جالب در مورد این روش این است که برخلاف روشهای پیشین، سطح زیر نمودار ROC برای هر دو کلاس برابر نیست و برای کلاس A، ۷۵.۳٪ و برای کلاس C، ۷۶.۱٪ می باشد. میزان میانگین وزنی سطح زیر نمودار ROC نیز ۷۵.۷٪ است.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         154               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0.0003 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        A
                 1         0          1         1         1          1        C
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 106   0 |   a = A
   0  48 |   b = C
```

شکل ۵ – ۲۲ : نتایج بدست آمده از Training Set با روش Logistic

جدول ۵ – ۵ : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش Logistic

Method	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Cross validation	0.214	0.4557	49.7416%	98.3522%
Training set	0	0	0	0.0003%

```
=== Predictions on test data ===

 inst#,    actual, predicted, error, probability distribution
     1        2:C        2:C          0     *1
     2        2:C        2:C          0     *1
     3        2:C        2:C          0     *1
     4        2:C        2:C          0     *1
     5        2:C        2:C          0     *1
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        2:C      +   0     *1
    14        1:A        1:A         *1      0
    15        1:A        2:C      +   0     *1
    16        1:A        2:C      +   0     *1
     1        2:C        2:C          0     *1
     2        2:C        2:C          0     *1
     3        2:C        2:C          0     *1
     4        2:C        2:C          0.003 *0.997
     5        2:C        2:C          0.003 *0.997
     6        1:A        2:C      +   0     *1
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        2:C      +   0     *1
    14        1:A        1:A         *1      0
    15        1:A        2:C      +   0     *1
    16        1:A        1:A         *1      0
     1        2:C        2:C          0     *1
     2        2:C        2:C          0     *1
     3        2:C        2:C          0     *1
     4        2:C        2:C          0     *1
     5        2:C        2:C          0     *1
     6        1:A        1:A         *1      0
     7        1:A        2:C      +   0     *1
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
```

شکل ۵ – ۲۳ : بخشی از نمودار حاصل از پیش بینی Logistic روی نمونه های تست با روش Cross Validation

```
=== Predictions on training set ===

 inst#,     actual, predicted, error, probability distribution
     1        1:A        1:A         *1      0
     2        1:A        1:A         *1      0
     3        1:A        1:A         *1      0
     4        1:A        1:A         *1      0
     5        1:A        1:A         *1      0
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        2:C        2:C          0     *1
    10        2:C        2:C          0     *1
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        1:A         *1      0
    14        1:A        1:A         *1      0
    15        1:A        1:A         *1      0
    16        1:A        1:A         *1      0
    17        1:A        1:A         *1      0
    18        2:C        2:C          0     *1
    19        1:A        1:A         *1      0
    20        2:C        2:C          0     *1
    21        1:A        1:A         *1      0
    22        2:C        2:C          0     *1
    23        1:A        1:A         *1      0
    24        2:C        2:C          0     *1
    25        1:A        1:A         *1      0
    26        1:A        1:A         *1      0
    27        1:A        1:A         *1      0
    28        1:A        1:A         *1      0
    29        1:A        1:A         *1      0
    30        1:A        1:A         *1      0
    31        1:A        1:A         *1      0
    32        2:C        2:C          0     *1
    33        1:A        1:A         *1      0
    34        1:A        1:A         *1      0
    35        1:A        1:A         *1      0
    36        2:C        2:C          0     *1
    37        1:A        1:A         *1      0
    38        1:A        1:A         *1      0
    39        1:A        1:A         *1      0
    40        1:A        1:A         *1      0
    41        1:A        1:A         *1      0
    42        1:A        1:A         *1      0
    43        1:A        1:A         *1      0
```

شکل ۵ – ۲۴ : بخشی از نمودار حاصل از Training Set با روش Logistic

نمودارهایmargin و خطاهای طبقه بندی کننده و ROC حاصل از روش Logistic در پیوست سه آمده است

۵ – ۵ – ۵ نتایج حاصل از روش MLP

شکل ۵ – ۲۵ نتایج حاصل از روش MLP را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         129               83.7662 %
Incorrectly Classified Instances        25               16.2338 %
Kappa statistic                          0.6362
Mean absolute error                      0.1792
Root mean squared error                  0.3969
Relative absolute error                 41.6557 %
Root relative squared error             85.658  %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.849     0.188      0.909     0.849     0.878      0.847    A
                 0.813     0.151      0.709     0.813     0.757      0.847    C
Weighted Avg.    0.838     0.176      0.847     0.838     0.84       0.847

=== Confusion Matrix ===

  a  b   <-- classified as
 90 16 |  a = A
  9 39 |  b = C
```

شکل ۵ – ۲۵ : نتایج ایجاد شده توسط روش MLP

همانطور که شکل ۵ – ۲۵ نشان می دهد، روش MLP تعداد ۱۲۹ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۸۳.۷۶۶۲٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی حاصل از این روش در می یابیم که ۹۰ نمونه (TP) از ۹۹ نمونه (TP + FP = 90 + 9) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس ۹۰.۹٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{90}{90+9} * 100\% = 90.9\%$$

همچنین تعداد ۳۹ نمونه (TP) از ۵۵ نمونه (۱۶ + ۳۹ = TP + FP) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با ۷۰.۹٪ است .

$$Precision = \frac{TP}{TP+FP} * 100\% = \frac{39}{39+16} * 100\% = 70.9\%$$

میزان میانگین وزنی دقت برابر با ۸۴.۷٪ می باشد که نسبت به میانگین وزنی بدست آمده توسط روش های قبلی جواب بهتری است.

تعداد ۱۶ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 16) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر ۸۴.۹٪ می باشد که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{90}{90 + 16} = 84.9\%$$

و تعداد ۹ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 9) بنابراین مقدار یادآوری برای کلاس C برایر است با ۸۱.۳٪که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{39}{39 + 9} = 81.3\%$$

همچنین میزان میانگین وزنی یادآوری برابر است با ۸۳.۸٪که جواب بسیار خوبی است و نسبت به میانگین وزنی روش های قبلی نیز بهتر می باشد.

میزان F – Measure برای کلاس A برابر است با ۸۷.۸٪ که جواب بسیار خوبی است.

$$\text{F–measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\% = \frac{2*90}{2*90+9+16} * 100\% = 87.8\%$$

میزان F - Measure برای کلاس C برابر است با ۷۵.۷٪ که جواب خوبی است .

$$\text{F – measure} = \frac{2*TP}{2*TP+FP+FN} * 100\% = \frac{2*recall*precision}{recall+precision} * 100\%$$

$$= \frac{2*39}{2*39+16+9} * 100\% = 75.7\ \%$$

میزان میانگین وزنی F – Measure برابر است با ۸۴٪ که جواب بسیار خوبی است و نسبت به میانگین وزنی بدست آمده توسط روش های قبلی بهتر می باشد .سطح زیر نمودار ROC برای هر دو کلاس A و C برابر است با ۸۴.۷٪ که جواب بسیار خوبی است.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         152               98.7013 %
Incorrectly Classified Instances         2                1.2987 %
Kappa statistic                          0.9694
Mean absolute error                      0.0166
Root mean squared error                  0.1141
Relative absolute error                  3.8566 %
Root relative squared error             24.6341 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.042      0.981     1         0.991      0.967    A
                 0.958     0          1         0.958     0.979      0.967    C
Weighted Avg.    0.987     0.029      0.987     0.987     0.987      0.967

=== Confusion Matrix ===

   a   b   <-- classified as
 106   0 |   a = A
   2  46 |   b = C
```

شکل۵ - ۲۶ : نتایج حاصل از Training Set با روش MLP

جدول۵- ۶ : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش MLP

Method	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Cross validation	0.1792	0.3969	41.6557%	85.658%
Training set	0.0166	0.1141	3.8566%	24.6341%

```
=== Predictions on test data ===

 inst#,     actual, predicted, error, probability distribution
     1        2:C        1:A      +  *1      0
     2        2:C        2:C          0.01  *0.99
     3        2:C        2:C          0.002 *0.998
     4        2:C        2:C          0.02  *0.98
     5        2:C        2:C          0     *1
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        2:C      +   0.097 *0.903
    14        1:A        1:A         *0.805  0.195
    15        1:A        1:A         *1      0
    16        1:A        2:C      +   0.001 *0.999
     1        2:C        2:C          0.013 *0.987
     2        2:C        2:C          0.02  *0.98
     3        2:C        2:C          0.011 *0.989
     4        2:C        1:A      +  *0.806  0.194
     5        2:C        1:A      +  *0.806  0.194
     6        1:A        1:A         *0.928  0.072
     7        1:A        1:A         *1      0
     8        1:A        1:A         *1      0
     9        1:A        1:A         *0.999  0.001
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
    12        1:A        1:A         *1      0
    13        1:A        1:A         *0.61   0.39
    14        1:A        1:A         *0.983  0.017
    15        1:A        1:A         *0.945  0.055
    16        1:A        1:A         *1      0
     1        2:C        2:C          0.003 *0.997
     2        2:C        2:C          0.015 *0.985
     3        2:C        1:A      +  *1      0
     4        2:C        2:C          0.012 *0.988
     5        2:C        1:A      +  *1      0
     6        1:A        1:A         *0.96   0.04
     7        1:A        2:C      +   0.002 *0.998
     8        1:A        1:A         *1      0
     9        1:A        1:A         *1      0
    10        1:A        1:A         *1      0
    11        1:A        1:A         *1      0
```

شکل۵ – ۲۷ : بخشی از نمودار حاصل از پیش بینی MLP روی نمونه های تست با روش Cross Validation

```
=== Predictions on training set ===

 inst#,    actual, predicted, error, probability distribution
     1        1:A        1:A         *1      0
     2        1:A        1:A         *0.997  0.003
     3        1:A        1:A         *1      0
     4        1:A        1:A         *1      0
     5        1:A        1:A         *1      0
     6        1:A        1:A         *1      0
     7        1:A        1:A         *1      0
     8        1:A        1:A         *0.996  0.004
     9        2:C        2:C          0.005 *0.995
    10        2:C        2:C          0.005 *0.995
    11        1:A        1:A         *1      0
    12        1:A        1:A         *0.989  0.011
    13        1:A        1:A         *0.989  0.011
    14        1:A        1:A         *1      0
    15        1:A        1:A         *1      0
    16        1:A        1:A         *1      0
    17        1:A        1:A         *0.999  0.001
    18        2:C        2:C          0.009 *0.991
    19        1:A        1:A         *1      0
    20        2:C        2:C          0.014 *0.986
    21        1:A        1:A         *0.992  0.008
    22        2:C        2:C          0.012 *0.988
    23        1:A        1:A         *0.99   0.01
    24        2:C        2:C          0.001 *0.999
    25        1:A        1:A         *0.992  0.008
    26        1:A        1:A         *1      0
    27        1:A        1:A         *0.996  0.004
    28        1:A        1:A         *0.995  0.005
    29        1:A        1:A         *1      0
    30        1:A        1:A         *1      0
    31        1:A        1:A         *0.999  0.001
    32        2:C        2:C          0.001 *0.999
    33        1:A        1:A         *1      0
    34        1:A        1:A         *0.997  0.003
    35        1:A        1:A         *0.988  0.012
    36        2:C        2:C          0.001 *0.999
    37        1:A        1:A         *1      0
    38        1:A        1:A         *0.994  0.006
    39        1:A        1:A         *1      0
    40        1:A        1:A         *1      0
    41        1:A        1:A         *1      0
    42        1:A        1:A         *1      0
    43        1:A        1:A         *1      0
```

شکل ۵ – ۲۸ : بخشی از نمودار حاصل از Training Set با روش MLP

نمودارهایmargin و خطاهای طبقه بندی کننده و ROC حاصل از روش MLP در پیوست چهار آمده است.

۵ – ۵ – ۶ نتایج حاصل از توسط روش RandomForest

شکل ۵ – ۲۹ نتایج بدست آمده توسط روش RandomForest را نشان می دهد.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         141               91.5584 %
Incorrectly Classified Instances        13                8.4416 %
Kappa statistic                          0.8066
Mean absolute error                      0.2113
Root mean squared error                  0.2932
Relative absolute error                 49.1072 %
Root relative squared error             63.282  %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.925     0.104      0.951     0.925     0.938      0.94     A
                 0.896     0.075      0.843     0.896     0.869      0.94     C
Weighted Avg.    0.916     0.095      0.918     0.916     0.916      0.94

=== Confusion Matrix ===

  a  b   <-- classified as
 98  8 |  a = A
  5 43 |  b = C
```

شکل ۵ – ۲۹ : نتایج بدست آمده توسط روش RandomForest

همانطور که از شکل ۵ – ۲۹ پیداست روش RandomForest تعداد ۱۴۱ نمونه از ۱۵۴ نمونه کل و یا بعبارتی ۹۱.۵۵۸۴٪ را به درستی طبقه بندی کرده است. با دقت در ماتریس آشفتگی این روش در می یابیم که ۹۸ نمونه (TP) از ۱۰۳ نمونه (TP + FP = ۹۸ + ۵) تخصیص داده شده به کلاس A درست طبقه بندی شده اند یعنی میزان دقت پیش بینی برای این کلاس ۹۳.۲٪ است .

$$precision = \frac{TP}{TP + FP} * 100\% = \frac{98}{103} = 95.1\%$$

همچنین تعداد ۴۳ نمونه (TP) از ۵۱ نمونه (TP + FP = ۴۳ + ۸) تخصیص داده شده به کلاس C درست طبقه بندی شده اند و این یعنی اینکه میزان دقت پیش بینی برای این کلاس برابر با ۸۴.۳٪ است .

$$precision = \frac{TP}{TP + FP} * 100\% = \frac{43}{51} = 84.3\%$$

میزان میانگین وزنی دقت برابر با ۹۱.۸ ٪ می باشد که جواب بسیار خوبی است.

تعداد ۸ نمونه از کلاس A بطور غلط در کلاس C طبقه بندی شده اند (FN = 8) بنابراین مقدار یادآوری (Recall) برای کلاس A برابر ۹۲.۵٪ می باشد که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{98}{98 + 8} = 92.5\%$$

و تعداد ۵ نمونه از کلاس C بطور غلط در کلاس A طبقه بندی شده اند (FN = 5) بنابراین مقدار یادآوری برای کلاس C برایر است با ۸۹.۶٪ که جواب بسیار خوبی است.

$$Recall = \frac{TP}{TP + FN} * 100\% = \frac{43}{43 + 5} = 89.6\%$$

همچنین میزان میانگین وزنی برای یادآوری برابر است با ۹۱.۶٪ که جواب بسیار خوبی است.

میزان F – Measure برای کلاس A برابر است با ۹۳.۸٪که جواب بسیار خوبی است.

$$F - Measure = \frac{2 * TP}{2 * TP + FP + FN} * 100\% = \frac{2 * recall * precision}{recall + precision} * 100\%$$
$$= \frac{2 * 98}{2 * 98 + 5 + 8} * 100\% = 93.8\%$$

میزان F - Measure برای کلاس C برابر است با ۸۶.۹٪که جواب بسیار خوبی است .

$$F - Measure = \frac{2 * TP}{2 * TP + FP + FN} * 100\% = \frac{2 * recall * precision}{recall + precision} * 100\%$$

$$= \frac{2 * 43}{2 * 43 + 8 + 5} * 100\% = 86.9\%$$

میزان میانگین وزنی F – Measure برابر است با ۹۱.۶٪ که بسیار عالی است.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         154               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0479
Root mean squared error                  0.089
Relative absolute error                 11.1327 %
Root relative squared error             19.2245 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        A
                 1         0          1         1         1          1        C
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

   a   b   <-- classified as
 106   0 |   a = A
   0  48 |   b = C
```

شکل ۵ – ۳۰ : نتایج حاصل از training set با روش RandomForest

شکل۵ – ۳۰ نتایج حاصل از training set روش RandomForest را نشان می دهد. چون روی همان نمونه هایی که عمل آموزش گرفتن (train) انجام شده عمل تست انجام می شود انتظار نتایج بهتری را داریم که این بدلیل کاهش انواع خطاها می باشد. همانطور که مشاهده می شود روش RandomForest هنگام Training Set اگرچه میزان خطاها را به صفر نمی رساند اما میزان دقت پیش بینی آن ۱۰۰٪ است.

جدول۵ – ۷ مقایسه ای بین خطاهای ایجاد شده توسط روشهای cross validation و training set روش RandomForest را نشان می دهد.

جدول ۵ - 7 : مقایسه ای بین خطاهای ایجاد شده روشهای cross validation و training set روش RandomForest

Method	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Cross validation	0.2113	0.2932	49.1072%	63.282%
Training set	0.0479	0.089	11.1327%	19.2245%

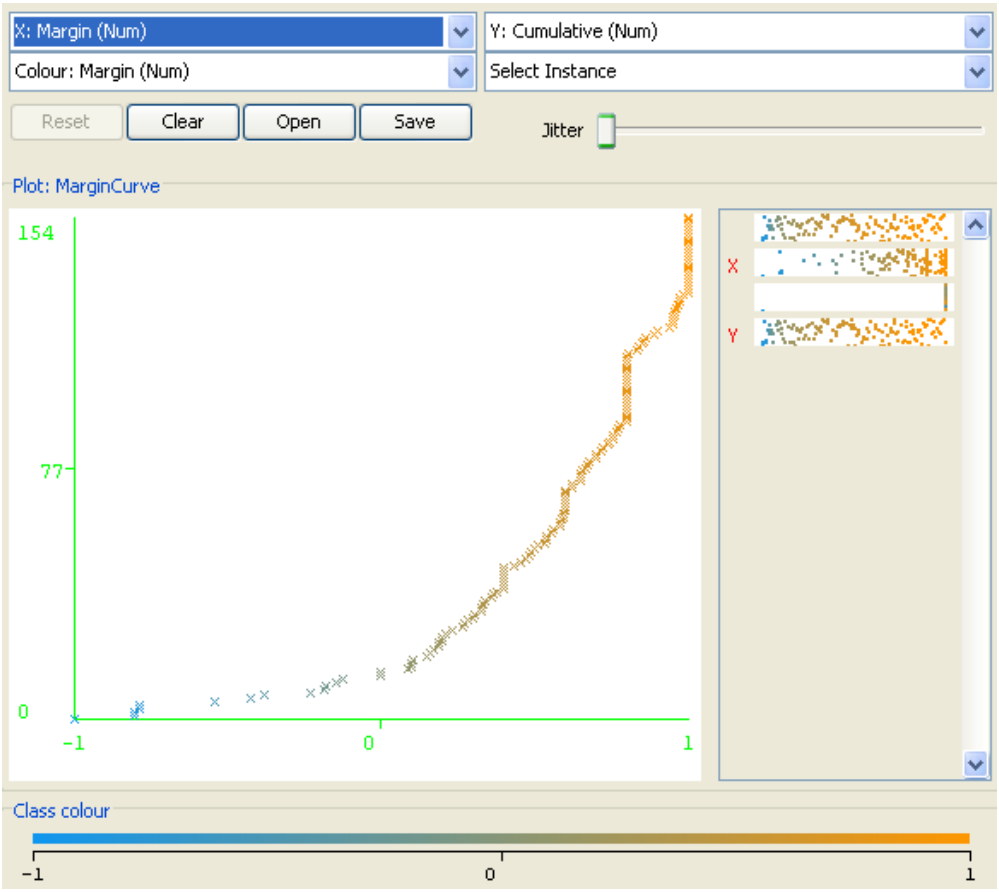
```
=== Predictions on test data ===

 inst#,     actual, predicted, error, probability distribution
     1        2:C        2:C          0.1   *0.9
     2        2:C        2:C          0     *1
     3        2:C        2:C          0     *1
     4        2:C        2:C          0.3   *0.7
     5        2:C        2:C          0.2   *0.8
     6        1:A        1:A         *0.9    0.1
     7        1:A        1:A         *0.593  0.407
     8        1:A        1:A         *0.95   0.05
     9        1:A        1:A         *1      0
    10        1:A        1:A         *0.888  0.112
    11        1:A        1:A         *0.9    0.1
    12        1:A        1:A         *0.971  0.029
    13        1:A        1:A         *0.574  0.426
    14        1:A        1:A         *0.9    0.1
    15        1:A        2:C      +   0.429 *0.571
    16        1:A        2:C      +   0.41  *0.59
     1        2:C        2:C          0.3   *0.7
     2        2:C        2:C          0.1   *0.9
     3        2:C        2:C          0.1   *0.9
     4        2:C        1:A      +  *0.893  0.107
     5        2:C        1:A      +  *0.893  0.107
     6        1:A        1:A         *0.718  0.282
     7        1:A        1:A         *0.8    0.2
     8        1:A        1:A         *1      0
     9        1:A        1:A         *0.85   0.15
    10        1:A        1:A         *0.917  0.083
    11        1:A        1:A         *1      0
    12        1:A        1:A         *0.9    0.1
    13        1:A        1:A         *0.768  0.232
    14        1:A        1:A         *0.883  0.117
    15        1:A        2:C      +   0.438 *0.562
    16        1:A        1:A         *1      0
     1        2:C        2:C          0.1   *0.9
     2        2:C        2:C          0.15  *0.85
     3        2:C        2:C          0.1   *0.9
     4        2:C        2:C          0.25  *0.75
     5        2:C        2:C          0.33  *0.67
     6        1:A        1:A         *0.836  0.164
     7        1:A        2:C      +   0.23  *0.77
     8        1:A        1:A         *0.978  0.022
     9        1:A        1:A         *0.863  0.137
    10        1:A        1:A         *0.987  0.013
    11        1:A        1:A         *1      0
```

شکل ۵ - ۳۱ : بخشی از نمودار حاصل از پیش بینی RandomForest روی نمونه های تست با روش Cross Validation

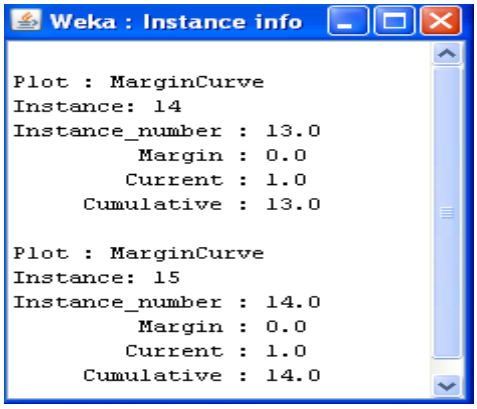
```
=== Predictions on training set ===

 inst#,     actual, predicted, error, probability distribution
     1        1:A        1:A          *0.929  0.071
     2        1:A        1:A          *0.94   0.06
     3        1:A        1:A          *1      0
     4        1:A        1:A          *1      0
     5        1:A        1:A          *1      0
     6        1:A        1:A          *0.914  0.086
     7        1:A        1:A          *1      0
     8        1:A        1:A          *0.9    0.1
     9        2:C        2:C           0     *1
    10        2:C        2:C           0     *1
    11        1:A        1:A          *1      0
    12        1:A        1:A          *1      0
    13        1:A        1:A          *1      0
    14        1:A        1:A          *0.9    0.1
    15        1:A        1:A          *1      0
    16        1:A        1:A          *1      0
    17        1:A        1:A          *0.982  0.018
    18        2:C        2:C           0.1   *0.9
    19        1:A        1:A          *1      0
    20        2:C        2:C           0     *1
    21        1:A        1:A          *1      0
    22        2:C        2:C           0     *1
    23        1:A        1:A          *0.95   0.05
    24        2:C        2:C           0     *1
    25        1:A        1:A          *1      0
    26        1:A        1:A          *1      0
    27        1:A        1:A          *1      0
    28        1:A        1:A          *1      0
    29        1:A        1:A          *1      0
    30        1:A        1:A          *1      0
    31        1:A        1:A          *0.817  0.183
    32        2:C        2:C           0     *1
    33        1:A        1:A          *0.935  0.065
    34        1:A        1:A          *1      0
    35        1:A        1:A          *0.732  0.268
    36        2:C        2:C           0.1   *0.9
    37        1:A        1:A          *1      0
    38        1:A        1:A          *1      0
    39        1:A        1:A          *1      0
    40        1:A        1:A          *1      0
    41        1:A        1:A          *1      0
    42        1:A        1:A          *0.9    0.1
    43        1:A        1:A          *0.8    0.2
```

شکل ۵ – ۳۲ : بخشی از نمودار Training set با روش RandomForest

شکل۵ - ۳۳ نمودار اختلاف (margin curve) بدست آمده توسط روش RandomForest را نشان می دهد. هر چه مقدار اختلاف (margin) برای هر نمونه به یک نزدیکتر باشد به معنای عملکرد بهتر طبقه بندی کننده می باشد. در شکل۵ - ۳۴ می بینیم نمونه دارای اختلاف صفر نمونه شماره ۱۳ است. ما این اطلاعات را از Instance Information کسب نمودیم. از آنجایی که نمونه ها از کمترین اختلاف تا بیشترین اختلاف ذخیره می گردند (شماره نمونه و cumulative یکسان است) تعداد ۱۴۲ نمونه (154-13+1=142) دارای اختلاف بیشتر از صفر بوده اند و ۱۳ نمونه اشتباه طبقه بندی شده اند و این نتیجه بسیار مطلوبی است.(این موضوع توسط شکل ۵ – ۴۸ نیز اثبات می شود.) شکل۵ - ۳۳ نیز اطلاعات مربوط به نمونه ای که اختلاف آن صفر است را نشان می دهد.



شکل ۵ – ۳۳ : نمودار اختلاف (margin)حاصل از روش RandomForest

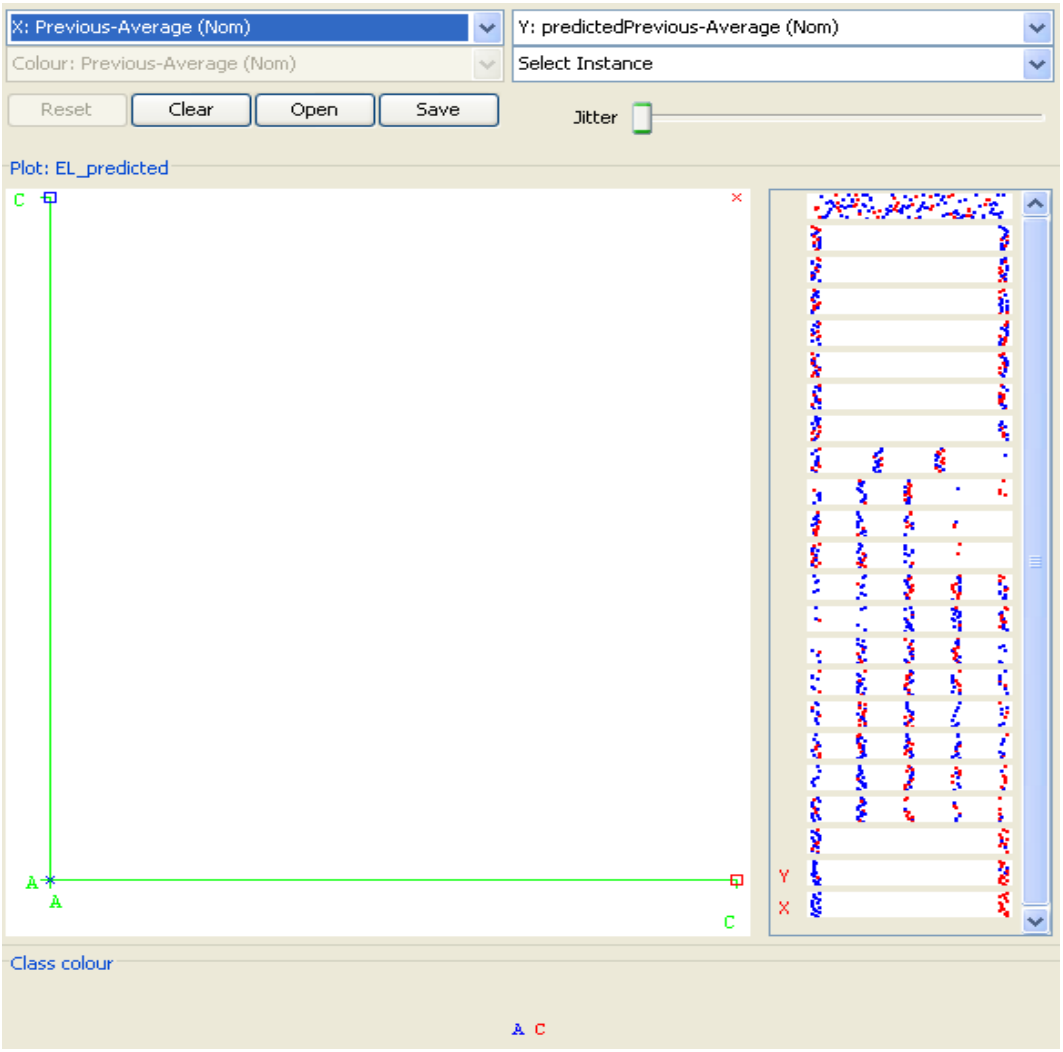


شکل ۵ – ۳۴ : اطلاعات نمونه دارای اختلاف (margin) صفر

همانطور که دیدیم در روش RandomForest مقدار اختلاف نمونه ها عددی است بین صفر و یک و برای همین است که در شکل های ۵ – ۳۱ و ۵ – ۳۲ در ستون احتمال (probabality) اعدادی بین صفر و یک دیده می شود.

شکل ۵ - ۳۵ نمودار خطاهای طبقه بند ی کننده RandomForest را نشان می دهد. محور X این نمودار بیانگر کلاس طبقه بندی (Previous Average) و محور Y این نمودار بیانگر پیش بینی صورت گرفته برای این کلاس (Predicted Previous Average) می باشد. نقاطی که با ضربدر آبی مشخص شده اند بیانگر مجموعه نمونه هایی هستند که به درستی به کلاس A طبقه بندی شده اند و در حقیقت TP کلاس A را بدست می آورند. نقاطی که با ضربدر قرمز مشخص شده اند بیانگر مجموعه نمونه هایی هستند که به درستی به کلاس C طبقه بندی شده اند و TP کلاس C را حاصل می کنند. نقاطی که با مربع کوچک آبی متمایز شده اند مبین مجموعه نمونه هایی هستند که در واقع متعلق به کلاس A بوده اند اما بطور اشتباه به کلاس C طبقه بندی شده اند(FN کلاس A) و نقاطی که با مربع کوچک قرمز نشان داده شده اند نشانگر مجموعه نمونه هایی هستند که در واقع متعلق به کلاس C بوده اند اما بطور اشتباه در کلاس A قرار گرفته اند(FP کلاس A و همچنین FN کلاس C). پس در حالت کلی علامت های مربع کوچک نشان دهنده نوعی خطا در پیش بینی طبقه بندی کننده هستند. از Instance Information استفاده نمودیم و تعداد نمونه های مشمول در مربع های کوچک آبی و قرمز را شمردیم. مربع کوچک آبی شامل ۸ نمونه و مربع کوچک قرمز شامل ۵ نمونه در درون خود بود که این نشان می دهد تعداد ۸ نمونه که واقعا متعلق به کلاس A بوده اند بطور اشتباه در کلاس C قرار گرفته اند و همچنین تعداد ۵ نمونه که واقعا به کلاس C متعلق بوده اند به کلاس A بطور اشتباه طبقه

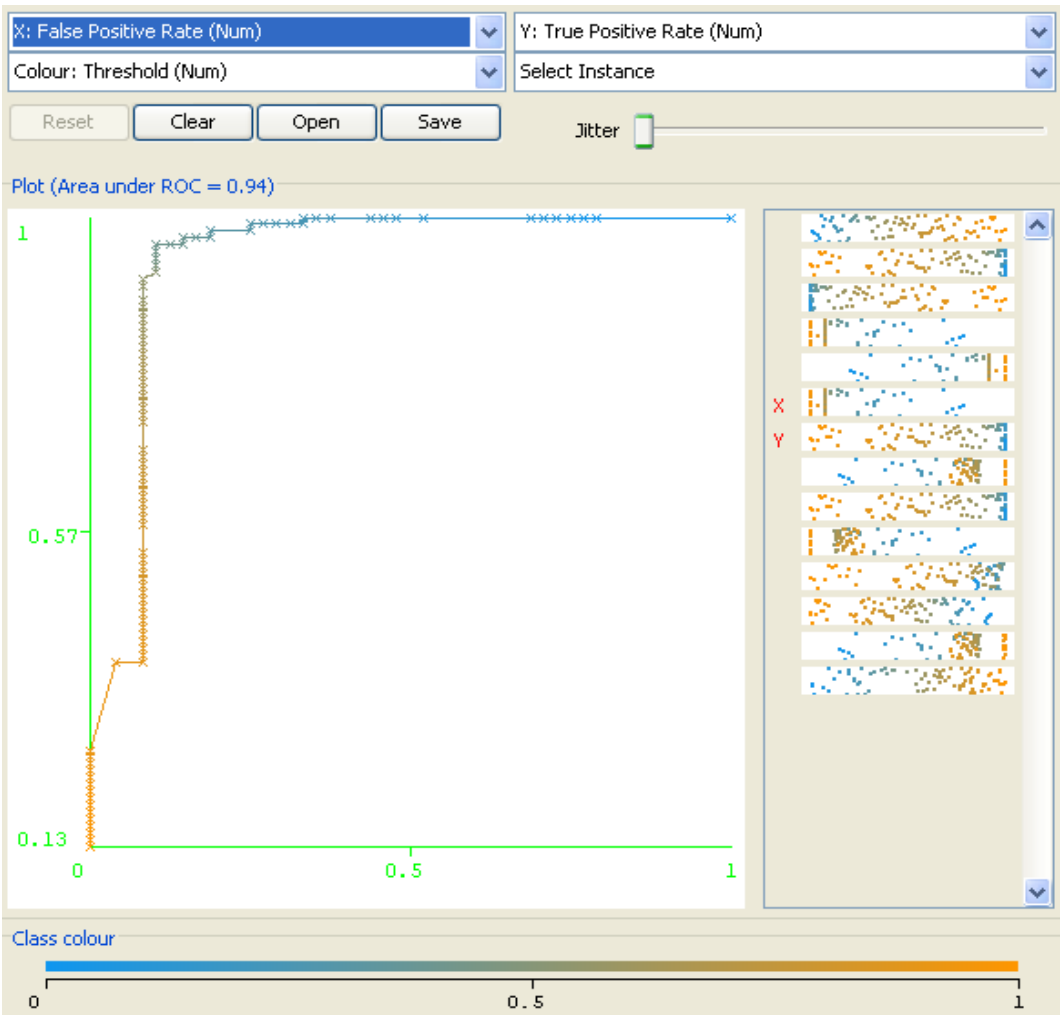
بندی شده اند و این یعنی اینکه FP کلاس A برابر ۵ و FP کلاس C برابر۸ می باشد. این موضوع توسط شکل ۵ – ۲۹ هم قبلا اثبات شده بود.



شکل ۵ – ۳۵ : نمودار خطاهای طبقه بندی کننده RandomForest

شکل ۵ – ۳۶ نمودار ROC حاصل از روش RandomForest را نشان می دهد.

در این روش مقدار Threshold مقداری است بین صفر و یک پس نمونه ها در جایی از نمودار قرار می گیرند که با این مقدار Threshold منطبق باشند. اولین نقطه ای که بیشترین صحت (۹۳.۵۰۶۵٪) در آن مشاهده شده (0.1,0.9) با مقدار Threshold برابر ۰.۴۱ می باشد که نشان می دهد توزیع متوازن نبوده است.



شکل ۵ – ۳۶ : نمودار ROC حاصل از روش RandomForest

همچنین سطح زیر نمودار ROC (AUC) به عنوان معیاری در عملکرد طبقه بندی کننده در نظر گرفته می شود. در اینجا سطح زیر نمودار ROC برای هر دو کلاس برابر ۰.۹۴ می باشد که نشان می دهد عملکرد آن بسیار خوب بوده است.

۵ – ۵ – ۷ مقایسه ای بین عملکرد طبقه بندی کننده های مورد استفاده:

جدول ۵ – 8 به مقایسه ای اجمالی بین عملکرد طبقه بندی کننده های استفاده شده می پردازد.

جدول ۵ – 8 : مقایسه ای بین عملکرد طبقه بندی کننده های استفاده شده

Classifier	TP Rate(Recall)	FP Rate	Precision	F-Measure	ROC Area
J 48	0.74	0.368	0.735	0.737	0.727
NaÏve Bayes	0.747	0.354	0.743	0.744	0.825
OneR	0.779	0.419	0.778	0.757	0.68
Logistic	0.792	0.14	0.845	0.8	0.755
MLP	0.838	0.176	0.847	0.84	0.847
RandomForest	0.916	0.095	0.918	0.916	0.94

از آنجایی که در مساله مورد پژوهش فقط دو کلاس داریم (کلاس A و C) مقدارTP Rate و Recall با یکدیگر برابر است . به منظور جلوگیری از تکرار، در جدول بالا ستون TP Rate و Recall را یکی در نظر گرفتیم.

۵ – ۶ مجموعه تست آماده (Supplied Test Set)

در این قسمت ده نمونه جدید که فرض می کنیم معدل آخرین ترم گذرانده شده (Previous- Average) آنها را نداریم وارد سیستم می کنیم. در حقیقت در فایل arff این ده نمونه ، مقدار ویژگی Previous-Average را ? می گذاریم. (نمونه های ۱ تا ۵ از دانشجویان موفق(A) و نمونه های ۶ تا ۱۰ از دانشجویان ناموفق (C) در نظر گرفته شده اند). در شکل ۵ - ۳۷ بخشی از فایل arff این ده نمونه نشان داده شده است.

```
@attribute Previous-Average {A,C,?}

@data
male,normal,single,unemployed,small,computer-based,dial-
up,D,low,medium,low,very-low,low,low,medium,medium,high,medium,high,MS,?
female,overage,single,employed,small,computer-
based,ADSL,B,high,high,high,medium,very-low,very-
low,medium,medium,medium,medium,high,MS,?
male,overage,married,employed,small,others,ADSL,C,medium,medium,medium,ve
ry-low,very-low,very-low,very-low,low,medium,high,very-high,MS,?
male,overage,single,employed,big,computer-
based,ADSL,C,high,high,high,high,medium,medium,medium,high,high,high,high
,BS,?
male,normal,single,employed,big,others,dial-up,C,medium,very-high,high,
very-low,very-low,very-high,high,medium,very-high,very-high,high,BS,?
female,normal,married,unemployed,small,others,dial-
up,B,low,low,medium,very-low,very-low,very-low,very-low,very-
low,medium,medium,medium,BS,?
male,overage,married,employed,small,others,dial-
up,C,low,high,high,medium,very-low,very-low,medium,medium,high,very-
high,high,BS,?
female,normal,single,employed,big,others,dial-
up,C,medium,medium,low,very-low,very-low,medium,very-low,very-
low,low,medium,medium,BS,?
female,normal,single,employed,big,computer-
based,ADSL,A,high,high,high,very-low,very-low,low,very-high,very-
low,very-low,very-low,very-low,MS,?
male,normal,married,employed,small,computer-
based,ADSL,C,medium,high,high,very-
low,low,low,low,medium,high,medium,medium,MS,?
```

شکل ۵ – ۳۷ : بخشی از فایل arff ده نمونه تست جدید

در شکل ۵ – ۳۸ نیز نتایج تست روی این ده نمونه جدید با روش RandomForest نشان داده شده است.

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability distribution
     1          ?        1:A      +  *0.77   0.23   0
     2          ?        1:A      +  *0.968  0.032  0
     3          ?        1:A      +  *1      0      0
     4          ?        1:A      +  *0.75   0.25   0
     5          ?        1:A      +  *0.8    0.2    0
     6          ?        1:A      +  *0.867  0.133  0
     7          ?        1:A      +  *0.643  0.357  0
     8          ?        2:C      +   0.463 *0.538  0
     9          ?        2:C      +   0     *1      0
    10          ?        2:C      +   0     *1      0
```

شکل ۵ – ۳۸ : نتایج تست روی ده نمونه جدید با روش RandomForest

همانطور که گفته شد پنج نمونه اول از دانشجویان موفق(A) و بقیه دانشجوی ناموفق(C) هستند. با توجه به شکل ۵ – ۳۸ در می یابیم که طبقه بندی کننده RandomForest بجز نمونه های ۶ و ۷ سایر نمونه ها را درست طبقه بندی کرده است. البته باید این نکته را مد نظر داشت که در اینجا به این دلیل در ستون error جلو همه نمونه ها علامت + گذاشته شده است که مقدار معدل ترم گذشته که همان اساس طبقه بندی است و در اینجا با actual نشان داده شده است برای همه نمونه ها مجهول می باشد.

۵ – ۷ قوانین پیوندی

شکل ۵ - ۳۹ نمونه ای از مجموعه قوانین پیوندی تولید شده را در موضوع مورد پژوهش نشان می دهد.

```
Best rules found:

 1. Job-Status=employed Major=computer-based 87 ==> Period=MS 86    conf:(0.99)
 2. Job-Status=employed Major=computer-based Internet-Type=ADSL 78 ==> Period=MS 77    conf:(0.99)
 3. Major=computer-based Previous-Average=A 85 ==> Period=MS 83    conf:(0.98)
 4. Job-Status=employed Previous-Average=A 82 ==> Period=MS 80    conf:(0.98)
 5. Job-Status=employed Internet-Type=ADSL 99 ==> Period=MS 94    conf:(0.95)
 6. Job-Status=employed 112 ==> Period=MS 106    conf:(0.95)
 7. Internet-Type=ADSL Previous-Average=A 93 ==> Period=MS 88    conf:(0.95)
 8. Previous-Average=A 106 ==> Period=MS 100    conf:(0.94)
 9. Major=computer-based Internet-Type=ADSL 100 ==> Period=MS 92    conf:(0.92)
10. Major=computer-based 115 ==> Period=MS 105    conf:(0.91)
```

شکل ۵ - ۳۹ : نمونه ای از قوانین پیوندی تولید شده توسط وکا با الگوریتم Apriori

۵ – ۸ ویژگی های منتخب

در شکل زیر پیاده سازی این عمل با استفاده از ارزیابی کننده ویژگی CfsSubsetEval و با روش Best First نمایش داده شده است. همانطور که از شکل پیداست سه ویژگی Diploma Average وStudyHours وGraduationSection بیشترین تاثیر را در نتیجه دارند.

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 134
        Merit of best subset found:    0.19

Attribute Subset Evaluator (supervised, Class (nominal): 21 Previous-Average):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 8,18,20 : 3
                     Diploma Average
                     StudyHours
                     GraduationSection
```

شکل۵ – ۴۰ : نمونه ای از انتخاب ویژگی های موثرتر درپیش بینی

همچنین با استفاده از ارزیابی کننده ویژگی InfoGainAttributeEval و روش Ranker می توان تمام ویژگی ها را به ترتیب اولویت تاثیر در نتیجه مشاهده کرد.

```
Attribute Evaluator (supervised, Class (nominal): 21 Previous-Average):
        Information Gain Ranking Filter

Ranked attributes:
 0.13361   20 GraduationSection
 0.12336   18 StudyHours
 0.11651    8 Diploma Average
 0.08852   19 Homework
 0.08709    9 English Skills
 0.07547   16 Interests in VU
 0.05788   11 Internet Accessibility
 0.04004   10 IT Skills
 0.03183   13 Interaction with Professors
 0.0307    17 Optional Activities
 0.02748   12 ELearning Familiarity
 0.02689    6 Major
 0.01983    2 Age
 0.01901   14 Interaction with Students
 0.01404   15 Extra-Study
 0.01397    1 Sex
 0.01078    4 Job Status
 0.00531    5 City
 0.00357    3 Marital Status
 0.00194    7 Internet Type

Selected attributes: 20,18,8,19,9,16,11,10,13,17,12,6,2,14,15,1,4,5,3,7 : 20
```

شکل ۵ – ۴۱ : انتخاب ویژگی ها به ترتیب تاثیر در نتیجه با الگوریتم Ranker

در شکل ۵ – ۴۱ مشاهده می شود که ویژگی GraduationSection بیشترین تاثیر و ویژگی Internet Type کمترین تاثیر را در نتیجه نهایی دارد.

در اینجا تعداد نه ویژگی (Internet-Type , Marital Status , City , Job Status , Sex , Extra-Study , Age , Major , Elearning Familiarity)که طبق الگوریتم Ranker کمترین تاثیر را در نتیجه نهایی دارند حذف شده است. همانطور که از شکل ۵ – ۴۲ پیداست با حذف ویژگی های کم اهمیت و زاید به پیش بینی دقیق تری می رسیم.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         121               78.5714 %
Incorrectly Classified Instances        33               21.4286 %
Kappa statistic                          0.4672
Mean absolute error                      0.2978
Root mean squared error                  0.4251
Relative absolute error                 69.226  %
Root relative squared error             91.7439 %
Total Number of Instances              154

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.896     0.458      0.812     0.896     0.852      0.771    A
                 0.542     0.104      0.703     0.542     0.612      0.771    C
Weighted Avg.    0.786     0.348      0.778     0.786     0.777      0.771

=== Confusion Matrix ===

  a  b   <-- classified as
 95 11 |  a = A
 22 26 |  b = C
```

شکل ۵ – ۴۲ : نتایج بدست آمده توسط روش J48 پس از حذف ویژگیهای کم اهمیت

فصل ششم

نتیجه گیری

در این پایان نامه به نحوه استفاده از روش های داده کاوی در موسسه آموزش عالی غیرانتفاعی مجازی نور طوبی تهران جهت پیشگویی نتیجه آتی دانشجویان پرداخته شده است. سیستم های آموزشی، داده های غنی و پر محتوا درمورد رفتار دانش آموزان و دانشجویان در طی یادگیری بدست می آورند. هدف از استفاده از روش های داده کاوی غنی کردن و توسعه دادن این محیط هاست. برای ساخت مدل های مورد نظر از تکنیک های مختلفی نظیر Naïve Bayes، درخت تصمیم J48، Logistic Regression، OneR و Multi Layer Perceptron (MLP) و RandomForest استفاده گردیده است. عملکرد هر یک از مدلها، مورد بررسی قرار گرفته و نتایج بدست آمده با یکدیگر مقایسه گردیده اند. اعتبار سنجی انجام شده بر روی مدلها اثبات می کند که نتایج بدست آمده دقیق و قابل اعتماد بوده اند. با بکارگیری این مدلها، مدیران آموزشی می توانند مشاوره های لازم را برای پیشگیری از رسیدن دانشجویان به وضعیت بحرانی بکار گیرند. همچنین این مدلها می توانند به عنوان یک ابزار پشتیبان تصمیم گیری در سیستم های آموزشی مورد بهره برداری قرار گرفته و نقش مهمی را در ارتقاء سطح علمی دانشگاهها داشته باشند.

برای انجام ارزیابی از روش Cross Validation با ده دسته (fold)استفاده شده است. ۶۶٪ داده ها را برای آموزش گرفتن (Train) و ۳۴٪ آنها را برای تست گرفتن (Test) اختصاص دادیم. برای انجام داده کاوی نیز روش‌های Naïve Bayes، درخت تصمیم J48، Logistic، OneR و Multi Layer Perceptron (MLP) استفاده شده است که در بین این پنج روش، روش RandomForest با ۹۱.۶٪ بیشترین دقت را در پیشگویی داشته است. روش MLP با ۸۴٪ در رتبه دوم و روش Logic

Regression با ۸۰٪ در رتبه سوم قرار دارد. روش OneR با ۷۵.۷ ٪، روش Naïve Bayes با ۷۴.۴ ٪ و روش J48 با ۷۳.۷ ٪ در رتبه های بعدی قرار گرفتند.

برای ایجاد نتایج دقیق تر می توان از تلفیق الگوریتم های یادگیری بهره جست. در [41] نشان دادند که با تلفیقی از چند طبقه بندی کننده میزان دقت پیش بینی نتیجه تحصیلی دانشجویان بالا می رود. همچنین می توان با انتخاب دقیق تر ویژگی ها و معیارهای مناسب تر، نتایج کار را بهبود بخشید. مثلاً در [41] نشان دادند که جنسیت (Gender) و سن (Age) نقش مهمی بعنوان عوامل پیش بینی کننده ندارند. می توان با انتخاب مجموعه ای از ویژگی های مذکور و سپس استفاده از الگوریتم Ranker در نرم افزار وکا، مهمترین ویژگی ها که بیشترین تاثیر را در پیش بینی نتایج تحصیلی دانشجویان دارند شناسایی و استخراج نمود و سپس به پیش بینی نتایج پرداخت. از بهینه سازی ویژگی ها استفاده شد و ویژگی های زاید حذف گردید و با استفاده از این تکنیک پیش بینی روش J48 به ۷۷.۷٪ رسید.

یکی از مسائلی که در آینده می توان روی آن تحقیق کرد" شخصی سازی آموزش سیار با استفاده از سیستم های آگاه از زمینه (Context Aware Systems) " است.

وسایل سیار دارای نمایشگرهای عریض و لمسی، دوربین دیجیتال ، گیرنده سیستم موقعیت یاب جهانی ، RFID و غیره می باشدکه وجود این تکنولوژی ها باعث کاربرد مفید وسایل سیار در آموزش سیار آگاه از متن شده است. استفاده از این نوع آموزش میزان موفقیت دانشجویان را افزایش می دهد و باعث پدید آمدن انواع جدیدی از سیستم های آموزشی شده است. با توجه به افزایش استفاده از آموزش الکترونیک در ایران ، استفاده از سیستم های آگاه از زمینه می تواند باعث افزایش کارآیی و بازدهی آموزش سیار شود.

فهرست منابع

[1] احمدی،حسن و شاکری اسکی،بهارک و علیشاهی،محمد (۱۳۸۷) . ،"غنی سازی محیط های آموزش الکترونیکی با استفاده از تکنیک های طبقه بندی داده"،دومین کنفرانس داده کاوی ایران .

[2] ایرجی،اعظم و مینایی،بهروز و شکورنیاز،ونوس،(و۱۳۸۷). ،"استخراج قوانین تصمیم با استفاده از الگوریتم درخت تصمیم جهت هدایت تحصیلی دانش آموزان به کمک دسته بندی داده ها ی آموزش و پرورش "،دومین کنفرانس داده کاوی ایران.

[3] ایرجی،اعظم و مینایی،بهروز و شکورنیاز،ونوس،(۱۳۸۷). ،" بکارگیری داده کاوی برای کشف تاثیرعامل جنسیت و مدرسه در موفقیت تحصیلی رشته های مختلف "،دومین کنفرانس داده کاوی ایران.

[4] پاینده فر، هومن و سید رضی، حسن ورهگذر،مسعودوفراهی،احمد (۱۳۸۷). ،"مقایسه تکنیک های داده کاوی جهت شخصی سازی مطلوبتر در آموزش الکترونیک"،دومین همایش ملی مهندسی کامپیوتر، برق و فناوری اطلاعات.

[5] شاکری اسکی،بهارک،(۱۳۸۷). *،شخصی سازی محیط های آموزش الکترونیک با استفاده از تکنیک های یادگیری ماشین و کلاسه بندی*،دانشگاه آزاد اسلامی مشهد:پایان نامه کارشناسی ارشد.

[6] یقینی،مسعود واکبری،امین و شریفی،سیدمحمدمهدی،(۱۳۸۷). ،" پیش بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک های داده کاوی "،دومین کنفرانس داده کاوی ایران.

[7] یقینی،مسعود وحیدری،سمیه ،(۱۳۸۷). ،" داده کاوی جهت ارتقاء و بهبود فرآیندهای سیستم آموزش عالی "،دومین کنفرانس داده کاوی ایران

[8] Allen, I . E., Seaman.(2003)., Sizing the Opportunity : the Quality and Extent of Online Education In the united states.

[9] Buell, cindy.,"learning theories and instructional design",(2004).

[١٠]Chamillard, A.T., (2006). Using student performance predictions in a computer science curriculum. Proceeding of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, June 26-28, Bologna, Italy, 260-264.

[1١] Chang, K., Beck, J., Mostow, J., Corbett, A.,(2006)., A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems., In Ikeda, M., et al. , 8th International Conference on Intelligent Tutoring Systems, ITS2006. LNCS Vol. 4053. Springer-Verlag, Berlin Heidelberg New York.

[1٢] Chernoff, H. (1973).,The use of face to represent points in k-Dimensional space Graphically., Journal of American Statistical Assosiation.

[1٣] Comon, P.Independent Component Analysis – a New Concept? , Signal Processing.(1994).

[1٤] Cooley, R., B ., et al. Web mining: Informaion and Pattern Discovery on the World Wide Web. In Proc of IEEE International Confrence on Tools with Artificial

[1٥] Cornford, James. and Pollock, Neil. (2007)."Theory and Practice of the Virtual University report on UK universities use of new technologies".

[1٦] Cornford, James. and Pollock, Neil. (2007)." The Theory and Practice of the Virtual University: Working Through the Work of Making Work Mobile".

[17] Dames, M., Marsala, C., Dang, T., Bouchon-Meunier, B.(2005)., Fuzzy decision tree for user modeling from human-computer interactions., In International Conference on human system learning: who is in control.

[18] Duda, R. O., Hart, P.E.(1973)., Pattern Classification and Sense Analysis. John Wiley and Sons, Inc., New York NY.

[19] Fawcett, T., Provost, F., (1997) ., Adaptive fraud detection. Data Mining and Knowledge Discovery 1 (3), 291–316.

[20]Fawcett. Tom , An introduction to ROC analysis,Pattern Recognition Letters 27 (2006) 862–863

[21] Feng, M., Heffernan, N., Koedinger, K. (2005).,"Looking for Sources of Error in Predicting Student's Knowledge"., in The Twentieth National Conference on Artificial Intelligence by the American Association for Artificial Intelligence, AAAI'05, Workshop on Educational Data Mining. July 9-13, Pittsburgh, Pennsylvania.

[22]Golding, P. and O. Donaldson. (2006).,Predicting academic performanc., Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference T1D-21, Oct. 28-31, San Diego, CA., 1-6.

[23]Guha,S.,Rastogi, R. , and Shim, k., CURE .(1998)., An efficient clustering algorithm for large database.,In processing of the 1998 ACM SIGMOD International conference on Management of data engineering,.

[24] Han, j.; Kamber , M.; and Tung, A. (2001)., Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, j., eds.,Geographic Data Mining and Knowledge Discovery. Taylor and Francis.

[26] Hand, D.J .(1987).,Discimination and Classification, John Wiley and Sons, Chichester U.K.

[26] Ian H. Witten and Eibe Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005, 173-176.

[27] Ian H. Witten and Eibe Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005, 176-178.

[28] Insook Lee,." What has been going on in e-Learning of South Korea?", Department of Education,Sejong University: South Korea.

[29] Jain, A.K.; and Dubes; R. C.(1998)., Algorithms for Clustering Data., Englewood Cliffs, NJ: Prentice-Hall.

[30] Jain, A.K..,Mao,J., Mohiuddin, K.(1996).,Artificial Neural Networks: A Tutorial, IEEE computer.

[31] Kantardzic M.(2003)., Data Mining: Concepts, Models, Methods, and Algorithms, Wiley publishing.

[32] Kontkanen, P.,Myllymaki, P., and Tirri, H.1996., Predictive data mining with finite mixtures, Proceeding 2 nd international Conference “Knowledge Discovery and Data Mining (KDD,96).

[33] Kotsiantis, S.B., C.J.2003., Pierrakeas and P.E. Pintelas. Preventing student dropout in distance learning using machine learning techniques. Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Oct. 21, Springer Berlin, Heidelberg, pp: 267-274.

[34] Kubat, M., Holte, R.C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Machine Learning 30 (2–3), 195–215.

[35] Lange, R.(1996)., An empirical test of weighted effect approach to generalized prediction using neural nets. In processing 2nd international conference "Knowledge Discovery and Data Mining (KDD,96)".

[36] Li, J., & Zaiane, O. (2004).,Combining usage, content, and structure data to improve web site recommendation. In International conference on ecommerce and web technologies .,.305-315

[37] Lu, J.2004., Personalized e-learning material recommender system in international conference on information technology for application.

[38] Lu, S. Y., and Fu, K.S.1978., A sentence-to-sentence clustering procedure for pattern analysis, IEEE Transactions on Systems, Man and Cybernetics SMC.

[39] Luan, J.Data Mining and its application in Higher Education .New Direction for Institutional Researcher, 2002.

[40] Masand, B. and Piatetsky – Shapiro, G.(1996)., A Comparison of approaches for maximizing business payoff of prediction models . In Proceeding 2nd international Conference "Knowledge Discovery and Data Mining (KDD,96)".

[41] McLachan, G.j. and Krishnan, T., The EM Algorithm and Extentions. Wiley series in probabality and statistics, 1997.

[42] McQueen, J.B.(1967)., Some methods of classification and Analysis of multivariate observations, Proceeding of Fifth Berkeley Symposium on Mathematical Statistics and probability.

[43] Merceron, A. and K. Yacef, 2005. Educational data mining: A Case Study. http://www.it.usyd.edu.au/~kalina/publis/merceron_yacef_aied05.pdf.

[44]Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer and W.F. Punch, 2003. Predicting student performance: An application of data mining methods with an educational web-based system. Proceedings of the 33rd Annual Conference on Frontiers in Education, Nov. 5-8, IEEE Computer Society, Washington, DC, USA., 13-18.

[45]Minaei-Bidgoli, B., G. Kortemeyer and W.F. Punch, 2004. Enhancing online learning performance: An application of data mining method. Proceedings of the 7th IASTED International Conference on Computers and Advanced Technology in Education, August 2004, Kauai, Hawaii, USA., 173-178.

[46] Mitchell, Tom M..(1997). Machine Learning, McGraw-Hill.

[47] Moses O. Oketch,.(2004)., "the african virtual university developments and critique, international higher education".

[48] Murty, M.N. and Krishna, G.(1980). A computentionally efficient technique for data clustering, Pattern Recognition.

[49] Ng, R.T., Han, J.(1994). Effective Clustering Methods for Special Data Mining, In 20nd International Conference On Very Large Data Base.

[50] Nguyen., Thai Nghe, Janecek, Paul., Haddawy.,Peter.2007. A Comparative Analysis of Techniques for predicting Academic Performance, Frontiers in education conference global engineering: knowledge without borders, opportunities without passports.

[51] Nottingham,Boulton, Helen.(2007). "UK Managing e-Learning: What are the Real Implications for Schools?" Trent University.

[52] Özyer, T., Alhajj, R., and Barker, K.(2007). Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule pre-screening. J. Netw. Comput. Appl. ,99-113. DOI= http://dx.doi.org/10.1016/j.jnca.2005.06.002

[53] Provost, F., Fawcett, T., 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining (KDD-97). AAAI press, Menlo Park, CA, 43–48

[54] Provost, F., Fawcett, T., 1998. Robust classification systems for imprecise environments. In: Proc. AAAI-98. AAAI Press, Menlo Park, CA, pp. 706–713. Available from: http://www.purl.org/NET/tfawcett/ papers/aaai98-dist.ps.gz.

[55] Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimation for comparing induction algorithms. In: Shavlik, J. (Ed.), Proc. ICML-98. Morgan Kaufmann, San Francisco, CA, 445–453. Available from: <http://www.purl.org/NET/tfawcett/papers/ICML98- final.ps.gz>.

[56] Punch, W.F.,Pei, M., Chia-Shun, L., Goodman, E.D., Hovland , P., and Enbody, R.(1993) .Futeher research on Feature Selection and Classification Using Generic Algorithms, In 5th International Conference On Generic Algorithm, Champaigh IL.

[57] Quinlan, J.R.(1986). Induction of decision trees. Machine Learning.

[58] Quinlan, J.R.(1987). Rule induction with statistical data, a comparison with multiple regression. Journal of Operational Research Society.

[59] Romero, C., Ventura, S., De Bra, P .(2004).” Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware”, User Modeling and User-Adapted Interaction 14(5).

[60] Romero, C., Ventura, S., De Bra, P., De Castro, C.(2003).” Discovering Prediction Rules”, in AHA! Courses. In: User Modelling Conference. June 2003, Johnstown, Pennsylvania.

[61] Ruck, D.W., Rogers, S.K., Kabirsky, M., Oxley , M.E., and Suter . B.W.(1990). The Multi-layer Perceptron as an Approximation to a Bayes Optimal Discremanant Function, IEEE Transactions on Neural Networks.

[62] shrock,s.a.(no date)."A brief history of instructional development ",available :http://uttc-med.utb.edu/6320/chapters/summary_ch2.html

[63] Srivasta, J. and Cooley, R., Deshpande, M. and Tan, P.N.(2000). Web Usage Mining: Discovery and Application of Usage Patterns from Web Data. SIGKDD Explorations.

[64] Stren M., Beck J. and Woolf B.(1999). Naïve Bayes Classifiers For User Modeling. Center for Knowledge Communic action, Computer Science Department, University of Massachusetts.

[65]Superby, J.F., J.P. Vandamme and N. Meskens.(2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS’06), Jhongali, Taiwan, 37-44.

[66] Swets, J.(1988). Measuring the accuracy of diagnostic systems. Science 240,1285–1293.

[67]Vandamme, J.P., N. Meskens and J.F. Superby, 2007. Predicting academic performance by data mining methods. Educ. Econ., 15, 405-419.

[68] Weiss, S. M. and Kulikowski C. A.(1991). Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems, Morgan Koufman.

[69] www2.cs.uregina.ca/~dbd/cs831/notes/ROC/ROC.html

[70]www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion _matrix/ confusion _matrix .html

[71]www.cs.iastate.edu/~cs573x/weka.html
[72] www.cs.waikato.ac.nz/ml/weka/

[73] www.gim.unmc.edu/dxtests/ROC3

[74]www.ieeexplore.ieee.org/xpl/freeabs_all.jsp?reload=true& arnumber =5451108

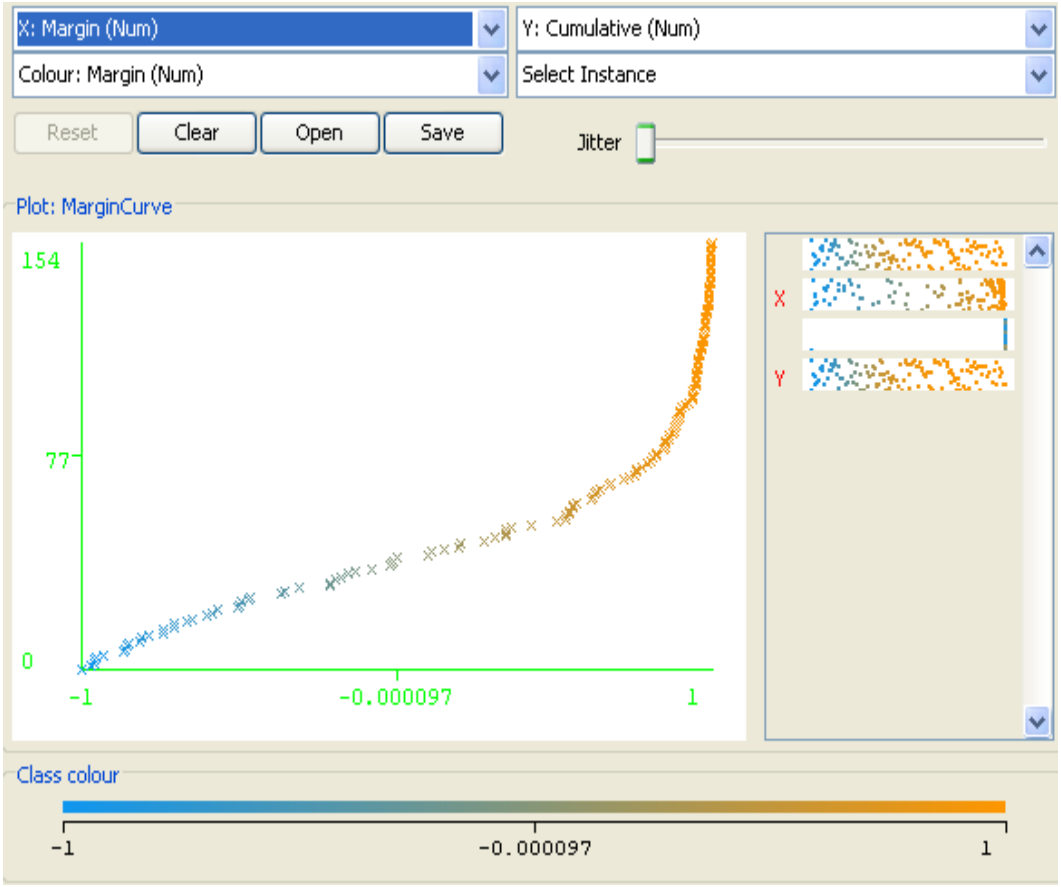
[75]www.improvedoutcomes.com/docs/WebSiteDocs/Plots Classification_ and _Prediction/Confusion _Matrix.htm

[76] www.publib.boulder.ibm.com/infocenter/ db2luw/v8/index.jsp ?topic=/com .ibm.im.visual.doc/idmu0mst125.html

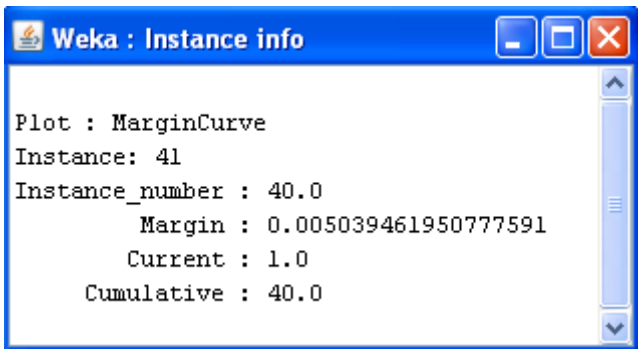
[77] www.weka.net.nz

پیوست یک

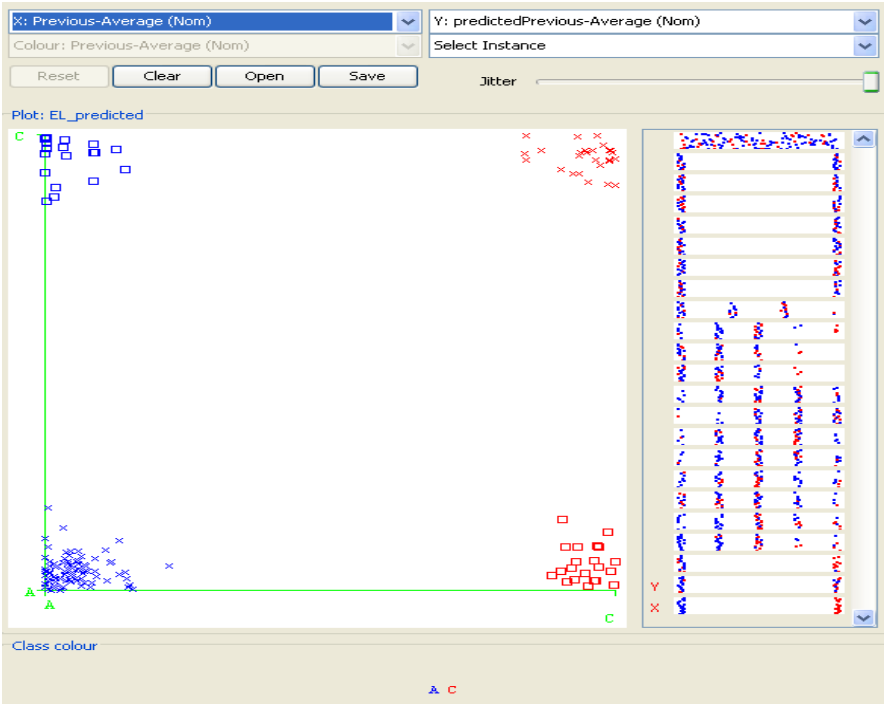
نمودارهای روش Naïve Bayes



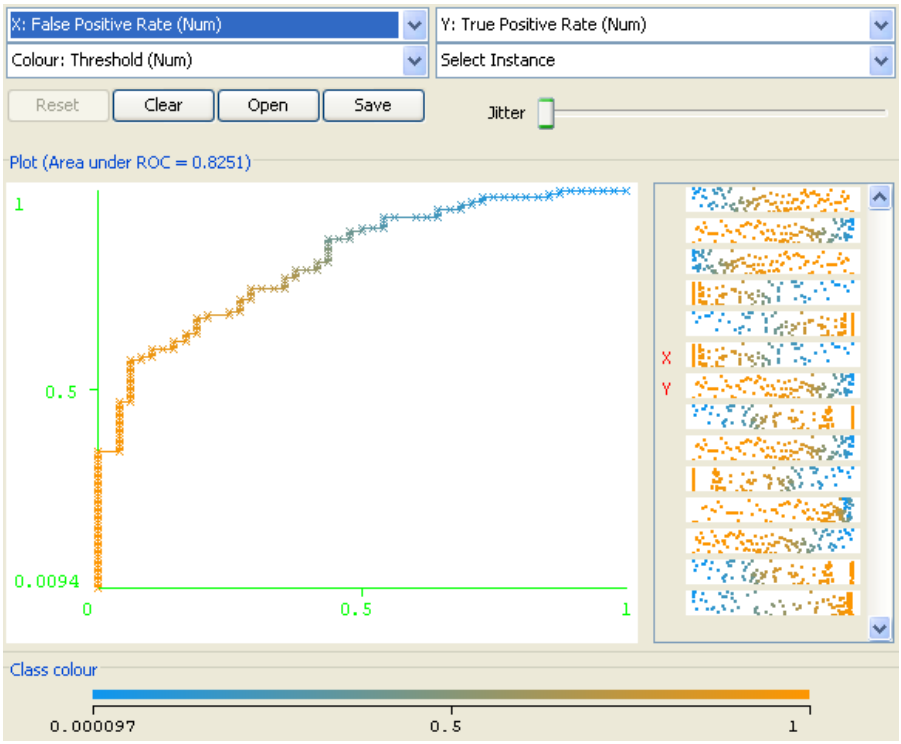
شکل ۱ : نمودار اختلاف (margin) حاصل از روش Naïve Bayse



شکل ۲ : اطلاعات نمونه دارای اختلاف (margin) تقریبا صفر



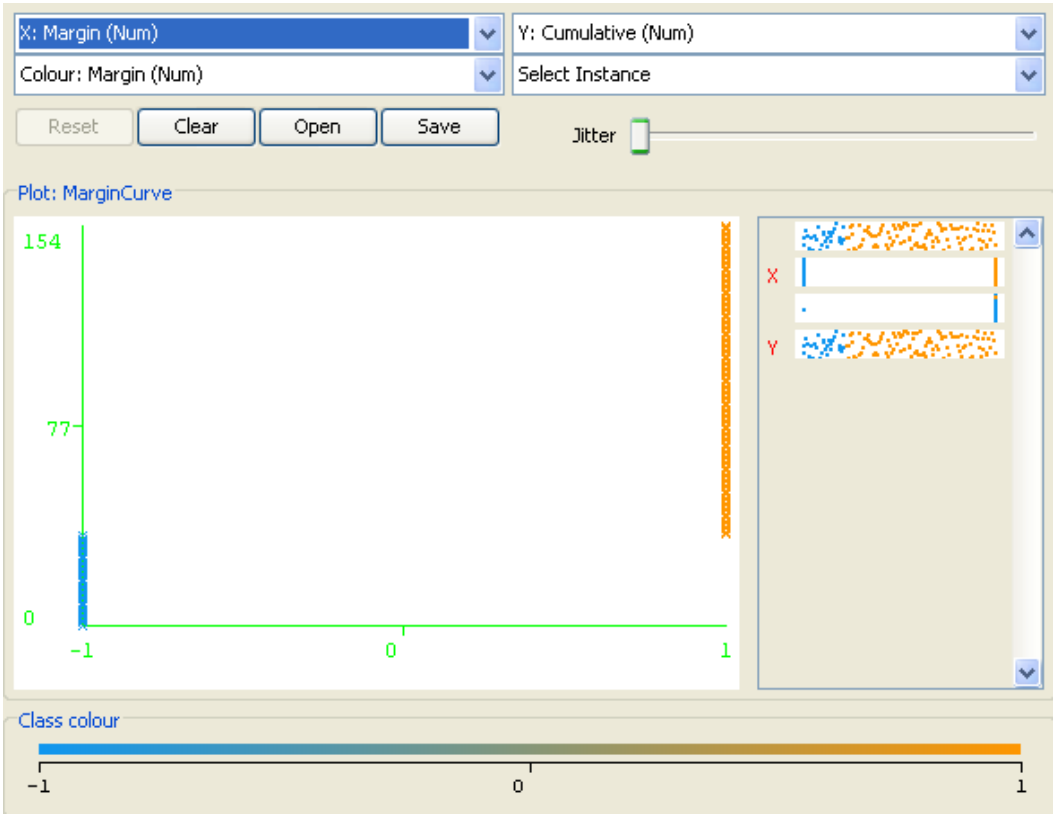
شکل۳ : نمودار خطاهای طبقه بندی کننده Naïve Bayes



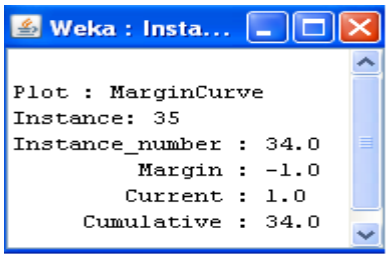
شکل ۴ : نمودار ROC حاصل از روش Naïve Bayes

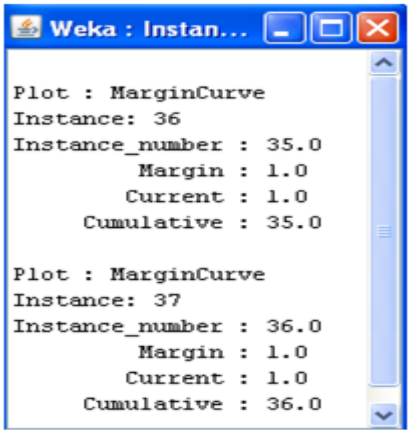
پیوست دو

نمودارهای روش OneR

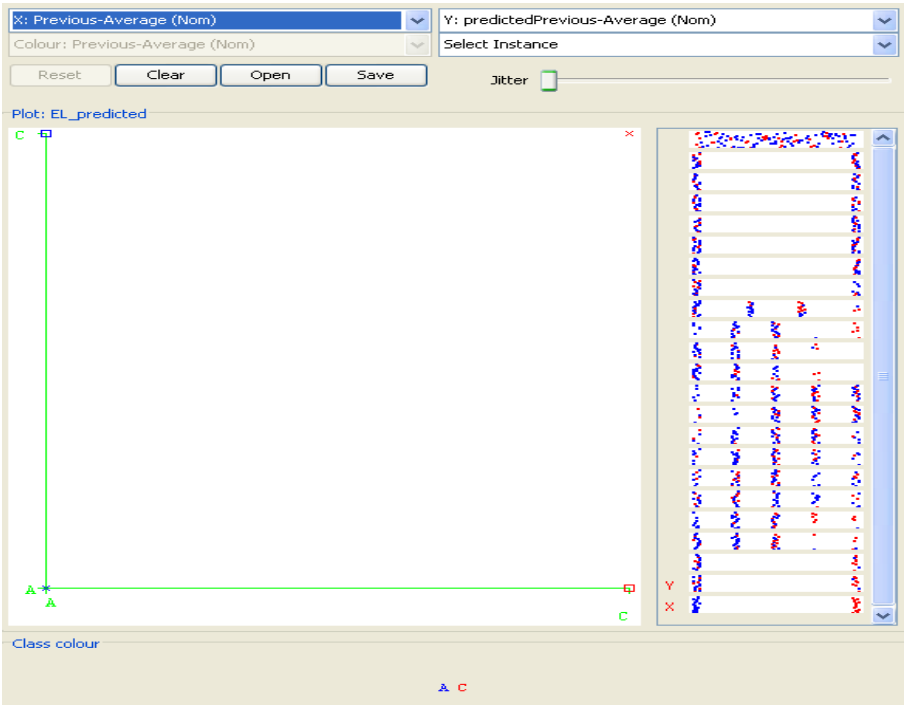


شکل ۱ : نمودار اختلاف(margin) حاصل از روش OneR

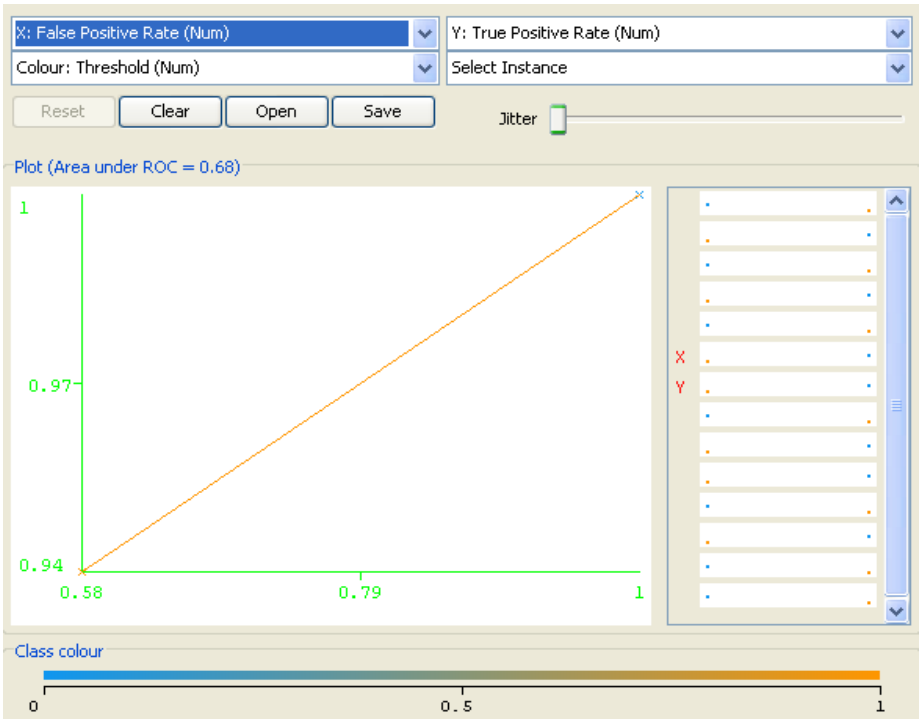




شکل ۲ : اطلاعات اولین نمونه های دارای نزدیکترین اختلاف (margin) به صفر (نمونه های مرزی نمودار اختلاف)



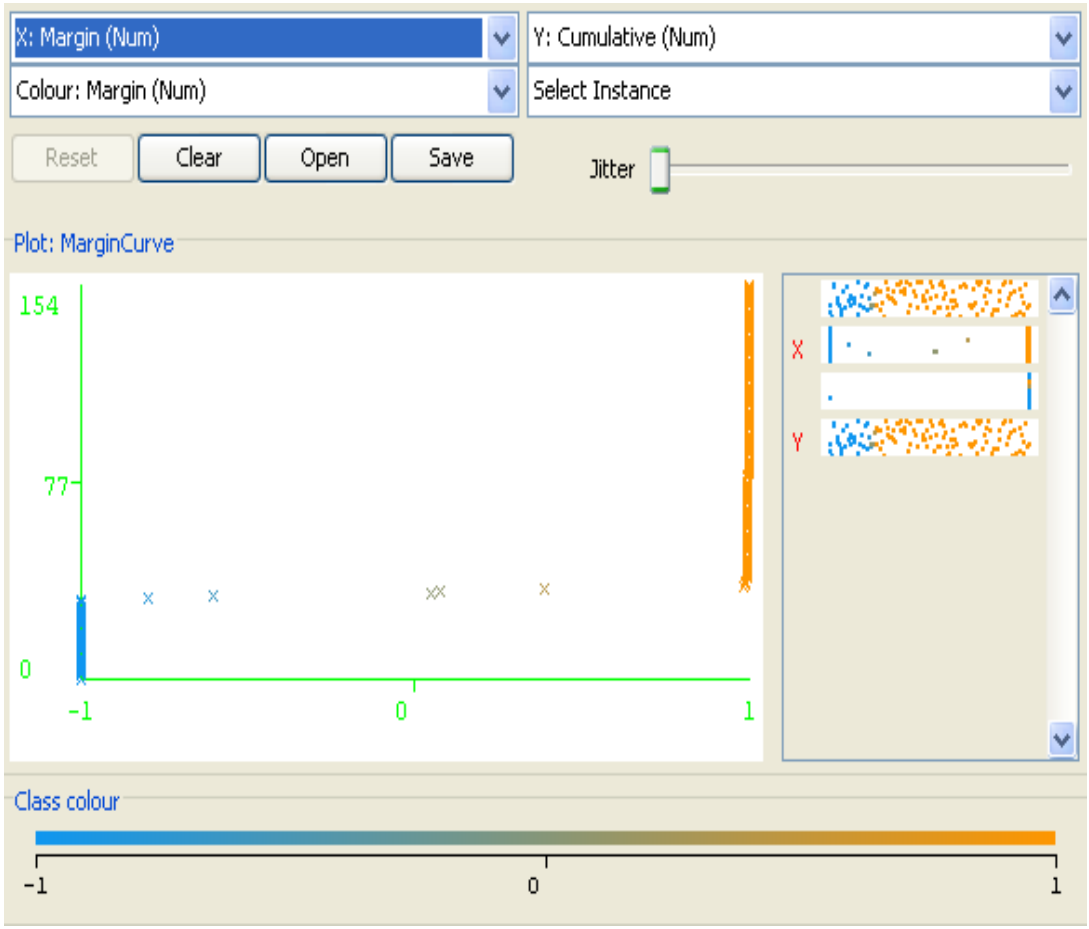
شکل ۳ : نمودار خطاهای طبقه بندی کننده OneR



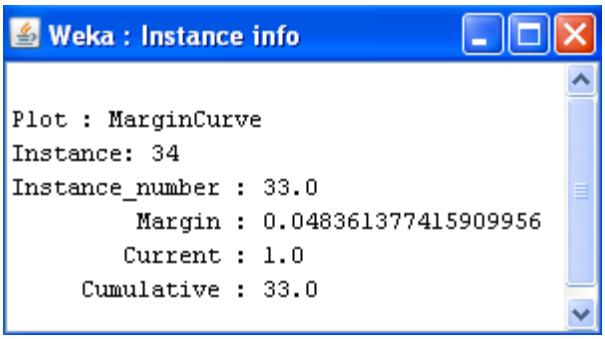
شکل ۴ : نمودار ROC حاصل از روش OneR

پیوست سه

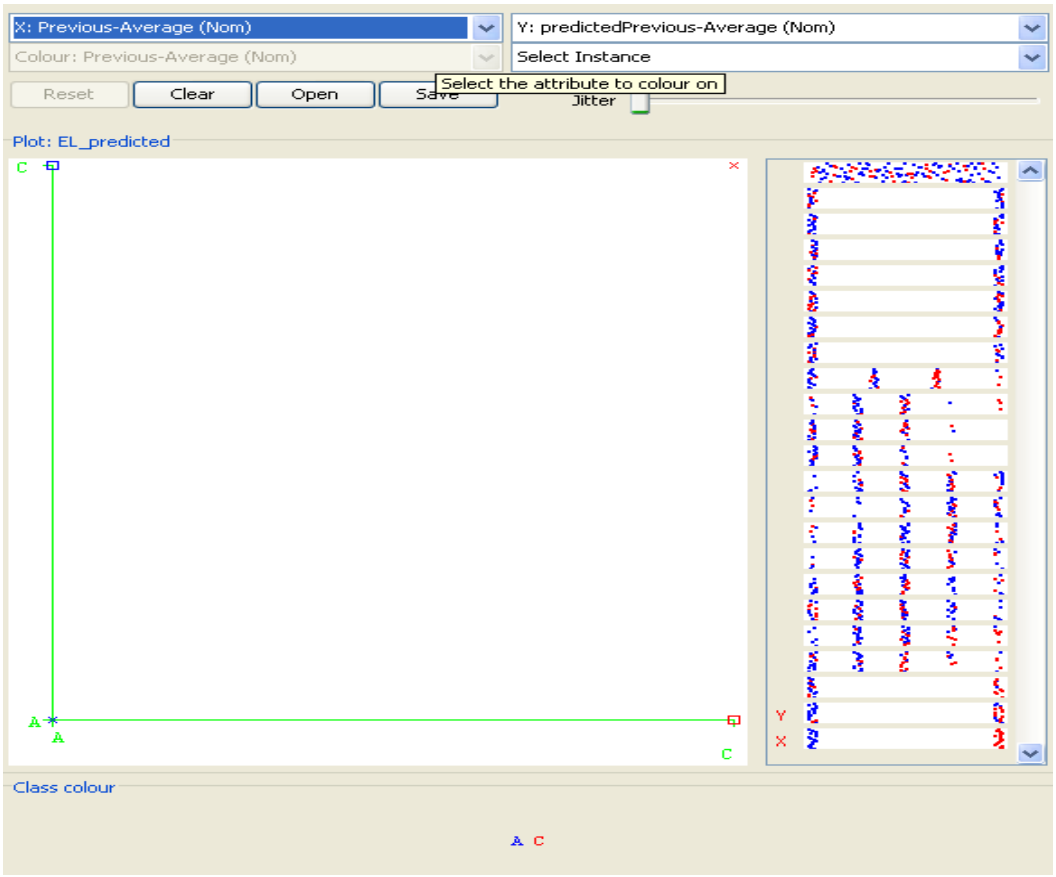
نمودارهای روش Logistic



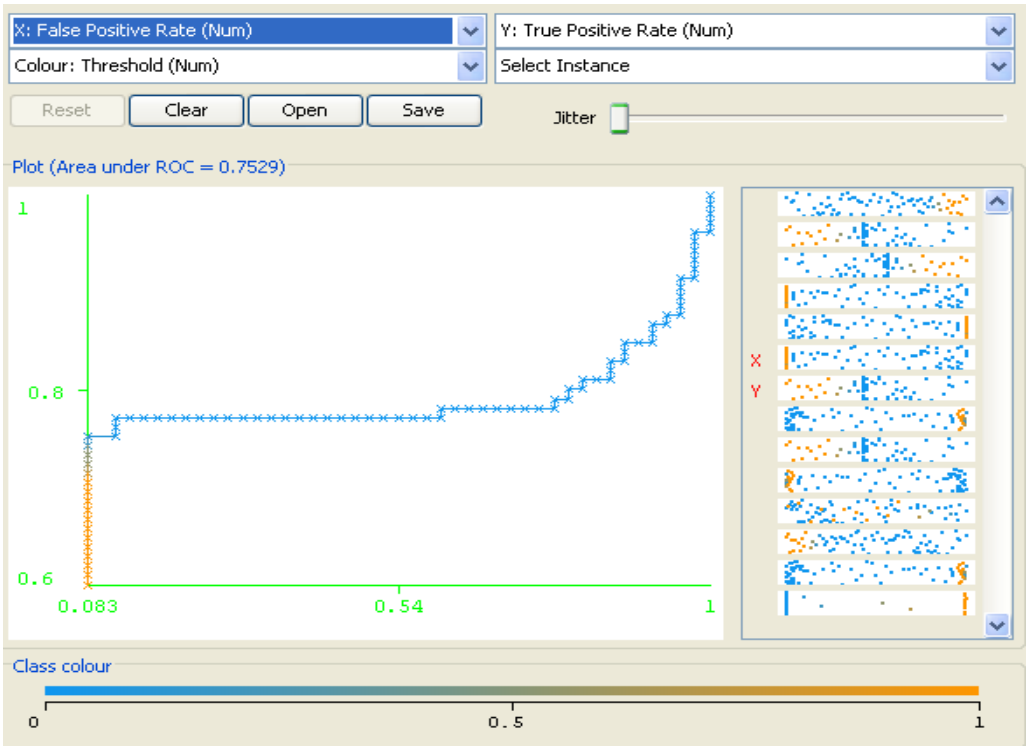
شکل ۱ : نمودار اختلاف (margin) حاصل از روش Logistic



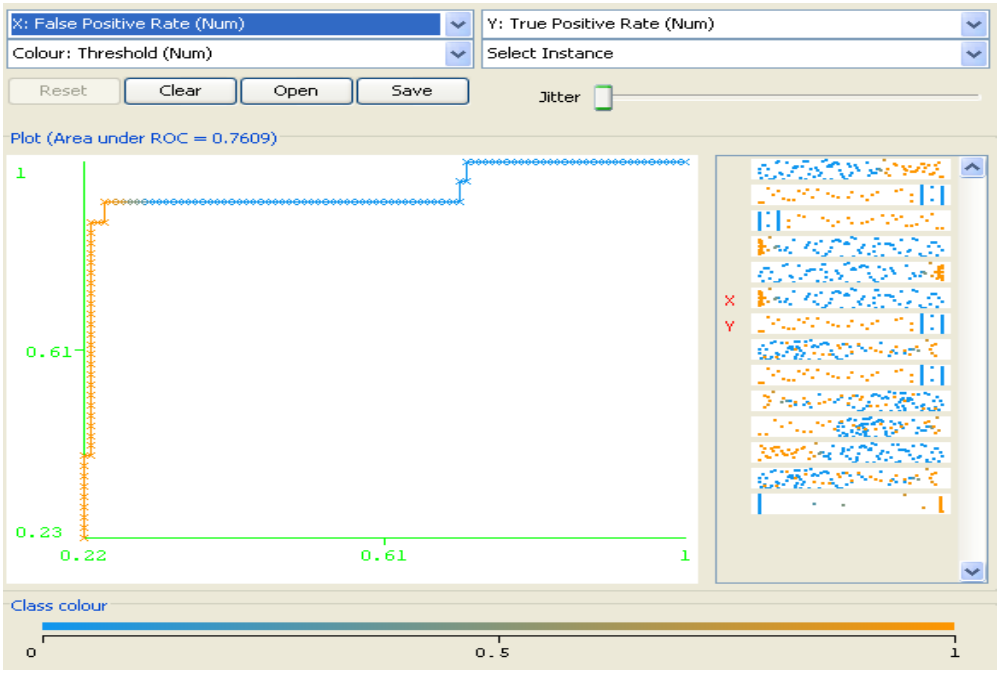
شکل ۲ : اطلاعات نمونه دارای نزدیکترین اختلاف (margin) به صفر



شکل ۳ : نمودار خطاهای طبقه بندی کننده Logistic



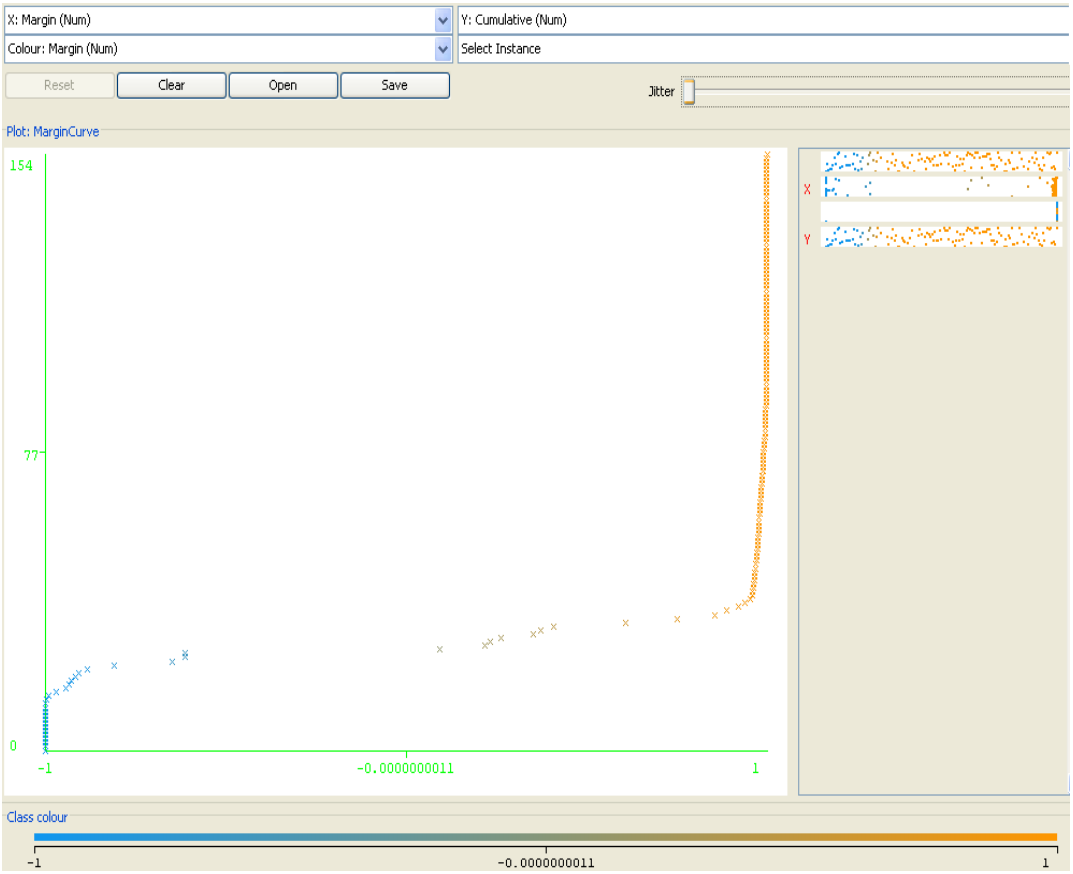
شکل ۴ : نمودار ROC کلاس A حاصل از روش Logistic



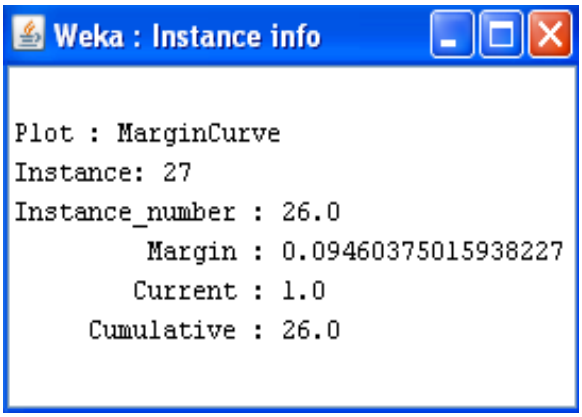
شکل ۵ : نمودار ROC کلاس Cحاصل از روش Logistic

پیوست چهار

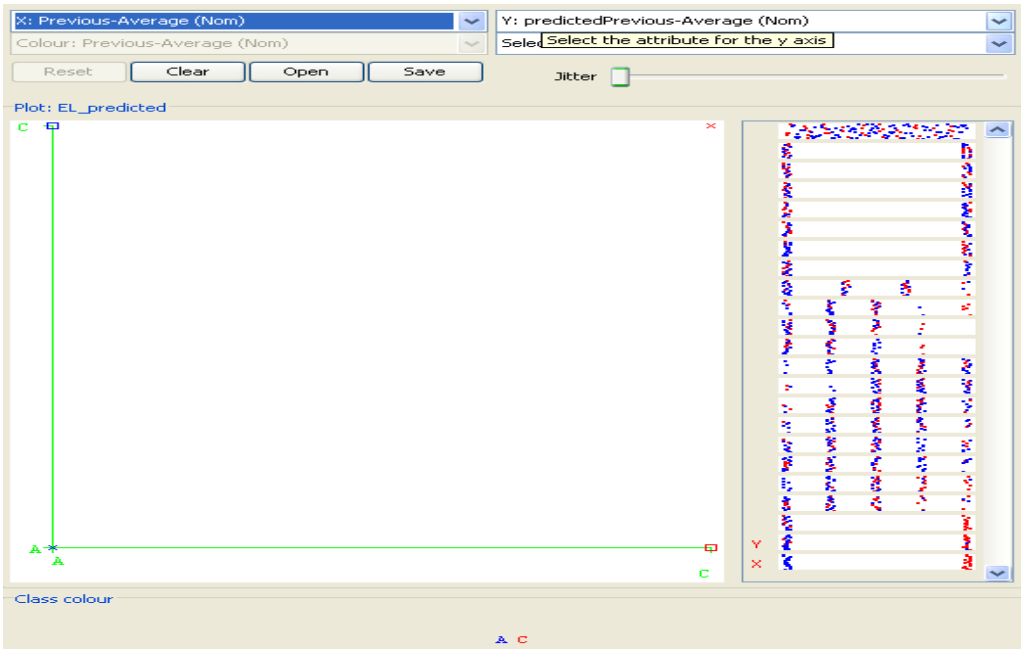
نمودارهای روش MLP



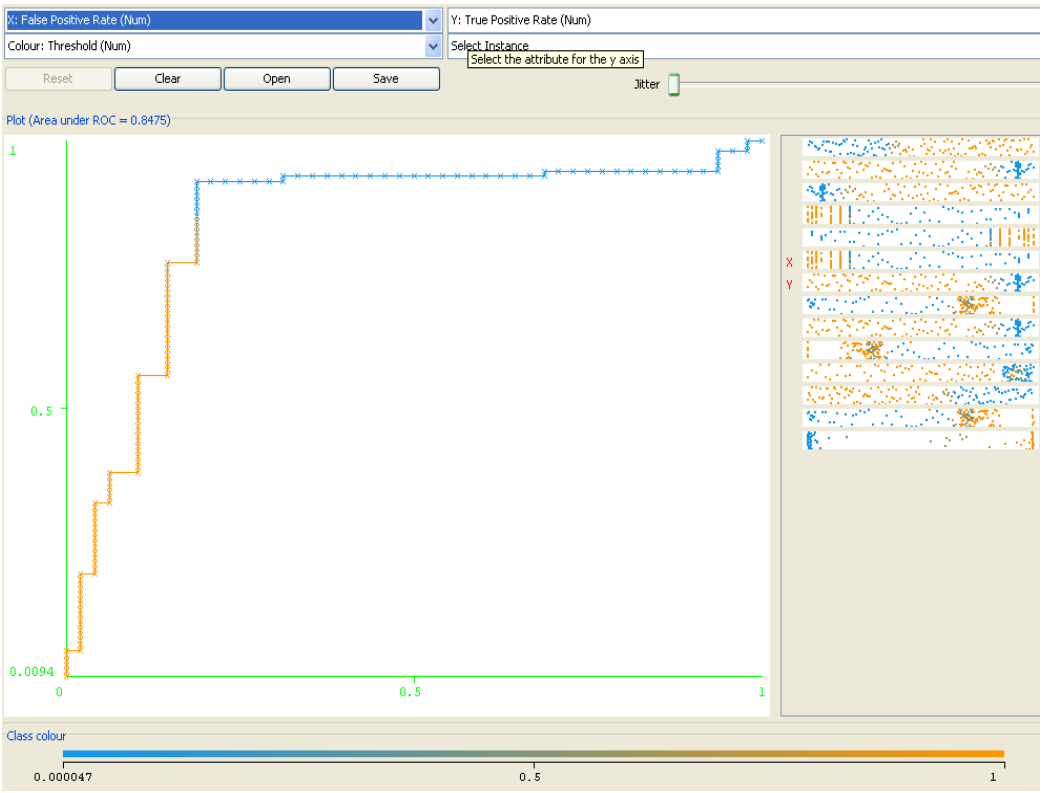
شکل ۱ : نمودار اختلاف (margin) حاصل از روش MLP



شکل ۲ : اطلاعات نمونه دارای اختلاف (margin) تقریبا صفر



شکل ۳ : نمودار خطاهای طبقه بندی کننده MLP



شکل ۴ : نمودار ROC حاصل از روش MLP

IN THE NAME OF GOD

Enrichment Of E-Learning Environment In Iran By Using Data Mining Techniques

BY

Alaleh Rangriz

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MESTER OF SCIENCE (MSc.)

IN

INFORMATION TECHNOLOGY

(ELECTRONIC COMMERCE)

NOORETOUBA VIRTUAL UNIVERSITY

TEHRAN

ISLAMIC REPUBLIC OF IRAN

EVALUATED AND APPROVED BY THETHESIS COMMITTEE AS:

……..…(CHAIRMAN)

……..(CONSULTANT)

…….....(ARBITRATOR)

ABSTRACT

Enrichment Of E-Learning Environment In Iran By Using Data Mining Techniques

By

Alaleh Rangriz

The aim of this study was to enrich the E-learning environment by predicting the students' academic performance. It is useful in identifying weak students who are likely to perform poorly in their studies. In this study, we used WEKA open source data mining tool to analyze attributes for predicting students' academic performance. The data set comprised of 154 number of student records of NooreTouba Virtual University Of Tehran and 21 attributes of students registered between year 2006 and 2009. Preprocessing includes attribute importance analysis. We applied the data set to six classifiers (RandomForest,MLP, Logistic Regression, OneR, Naïve Bayes, J48) and obtained the accuracy of predicting the students' performance into either successful unsuccessful class. the student's academic performance can be predicted by using past experience knowledge discovered from the existing database. A cross-validation with 10 folds was used to evaluate the prediction accuracy. We used 66% of data for training and 34% for testing. In addition, the result showed that RandomForest classifiers scored the higher percentage of prediction of 91.6%.MLP, Logistic, OneR, Naïve Bayes, J48 classifiers orderly scored the percentage of prediction accuracy of 84%,80% , 75.7% , 74.4% , 73.7% .

Nooretouba University

A thesis for the degree of M.S.

E-Commerce

Title of thesis:

Enrichment Of E-Learning Environment In Iran By Using Data Mining Techniques

Supervised by:

Dr. Hassan Ahmadi Torshizi

Advised by:

Eng. Alireza Ghannadan

By:

Alaleh Rangriz

November 2010