آکادمی داده
dataacademy.ir
مجتمع آموزش عالی بم
۱۵-۱۷ بهمن ماه ۱۳۹۲
انجمن سیستم های هوشمند ایران
Intelligent Systems Scientific Society of Iran
وزارت علوم تحقیقات و فناوری
مجتمع آموزش عالی بم

پیش‌بینی وضعیت مشروطی دانشجویان با استفاده از تکنیک‌های داده‌کاوی(موردکاوی: دانشگاه قم)

فرید قنبری[۱]، وحید قنبری[۲]

[۱] گروه مهندسی کامپیوتر، دانشگاه قم، قم
ghanbari.f.70@gmail.com
[۲] گروه مهندسی فناوری اطلاعات، دانشگاه قم، قم
vghanbari@stu.qom.ac.ir

چکیده

مراکز آموزشی به ویژه دانشگاه‌ها همواره به ذخیره‌سازی داده‌هایی در مورد دانشجویان می‌پردازند. اما این داده‌ها در عین ارزشمند بودن به طور معمول تنها در فعالیت‌های اداری همچون ثبت نام، گزارش‌گیری و غیره استفاده می‌شوند. در صورتی که، این داده‌ها می‌توانند با روش‌های کشف دانش از پایگاه داده همچون داده‌کاوی، در مشاورهٔ تحصیلی دانشجویان مورد استفاده قرار گیرند.

در این مقاله تکنیک‌های داده‌کاوی بر پایگاه دادهٔ سامانهٔ آموزش دانشگاه قم اعمال گردیده تا وضعیت مشروطی یک دانشجو در دوران تحصیلش، مورد پیش‌بینی قرار گیرد. در مرحله پیش‌پردازش، فعالیت‌هایی همچون نرمال‌سازی داده، گسسته‌سازی داده و کاهش ابعاد داده انجام گرفته است. همچنین به علت نامتوازن بودن داده‌ها برای پیش‌بینی صحیح‌تر از روش متوازن‌سازی زیرنمونه-گیری استفاده شده است. در نهایت از بین مدل‌هایی که از ترکیب حالت‌های مختلفِ گسسته‌سازی و متوازن‌سازی داده ایجاد شده بود، مدلی که معیار صحت وزن‌دار آن حدود ٦٩ درصد بود، به عنوان مدل برتر شناخته‌شد. دانشگاه قم می‌تواند با استفاده از این مدل دانشجویانی را احتمال مشروطی آنان بالاست، شناسایی کرده و با اخطار به این دانشجویان، آن‌ها را به تلاش بیشتر دعوت نماید.

کلمات کلیدی

داده‌کاوی، داده‌کاوی آموزشی[۱]، داده‌های نامتوازن، گسسته‌سازی داده، رده‌بندی[۲]

۱- مقدمه

در دهۀ گذشته داده‌کاوی بر روی داده‌های آموزشی، به عنوان یک حوزۀ در حال رشد در تحقیقات علمی مربوط به علوم کامپیوتری، مطرح شده است. در واقع داده‌کاوی توانسته به توسعۀ روش‌هایی برای کاوش و استخراج دانش از داده‌های منحصر بفرد سامانه‌های آموزشی بپردازد. اکثر سامانه‌های موجود در مراکز آموزشی، به نگهداری اطلاعات مختلفی از دانشجویان می‌پردازند، اما این اطلاعات بدون ساختار، به دلیل حجم زیاد، ضعف و فقدان ابزارهای تحلیل و گزارش‌گیری، کارایی چندانی برای اساتید و مدیران این مراکز ندارند. حوزۀ داده‌کاوی و استخراج دانش از پایگاه داده برای مقابله با این مشکلات از کارایی مناسبی برخوردار است و همین امر موجب گردیده تا تحقیقات زیادی، در حوزۀ بکارگیری ابزارها و روش‌های داده‌کاوی و استخراج دانش بر روی داده‌های آموزشی در ده سال گذشته صورت پذیرد.

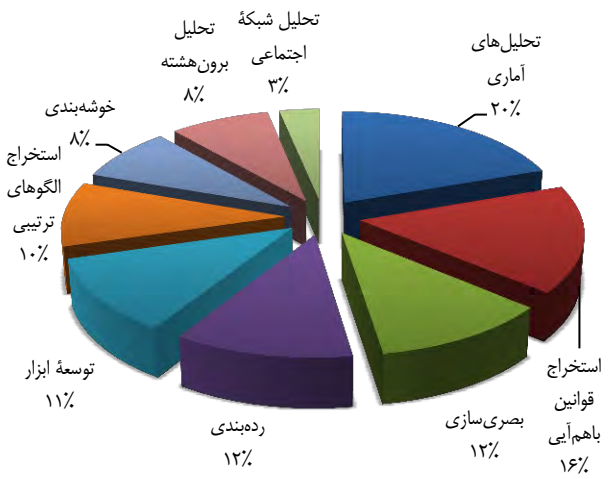
رئوس مطالب مقاله به شرح زیر است: در بخش ۲ به میزان فعالیت‌های انجام شده در حوزه‌های مختلف داده‌کاوی آموزشی پرداخته شده‌است و مروری بر مقالات، در حوزه ردبندی داده‌های آموزشی صورت گرفته است. در بخش ۳ به روش پیشنهادی برای پیش‌بینی مشروط شدن یا نشدن یک دانشجو پرداخته شده است. این بخش شامل پیش‌پردازش داده، اعمال تکنیک‌های داده‌کاوی بر داده‌ها و ارزیابی مدل‌های ایجاد شده با روش‌های مختلف و در نهایت تفسیر مدل برتر است. در بخش آخر نیز نتایج حاصل ارائه شده‌است.

۲- فعالیت‌های انجام‌شده در حوزهٔ داده‌کاوی آموزشی

فعالیت‌ها و پژوهش‌های انجام شده در این حوزه به گروه‌های زیر تقسیم می‌شوند:

- تحلیل‌های آماری
- بصری‌سازی
- استخراج قوانین باهم‌آیی
- رده‌بندی
- توسعهٔ ابزار
- استخراج الگوهای ترتیبی
- خوشه‌بندی
- تحلیل برون‌هشته
- تحلیل شبکهٔ اجتماعی

شکل (۱) میزان فعالیت‌های انجام شده در این حوزه‌ها را به اختصار نشان می‌دهد.



شکل ۱: میزان پژوهش‌های انجام شده در حوزه‌های داده‌کاوی آموزشی [9]

اساس فعالیت‌های صورت گرفته در این مقاله مربوط به حوزهٔ رده‌بندی است. به همین جهت تنها به بررسی فعالیت‌هایی که در این حوزه صورت گرفته‌است، خواهیم پرداخت. در جدول (۱) به چند نمونه از فعالیت‌های صورت گرفته در حوزهٔ رده‌بندی داده‌های آموزشی اشاره شده است.

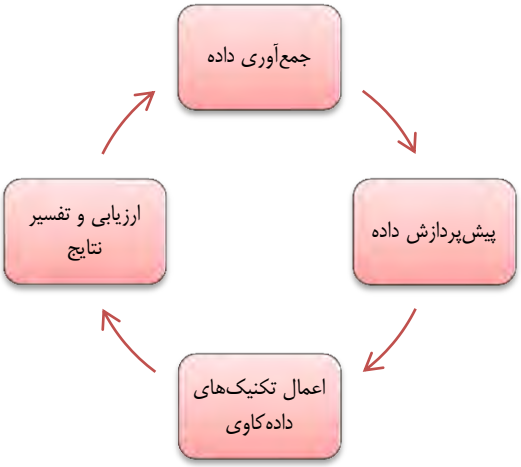
جدول ۱: مروری بر مقالات در حوزهٔ رده‌بندی داده‌های آموزشی

سال	محققان	فعالیت صورت گرفته
۲۰۰۰	چِن[۳] و همکاران	استفاده از روش‌های رده‌بندی برای بررسی واکنش گروه‌های مختلف دانشجویان به راهبردهای مختلف آموزشی [3].
۲۰۰۳	مینایی و پانچ[۴]	پیش‌بینی کارایی دانشجویان و نمرات پایانی آن‌ها با استفاده از ترکیبی از روش‌های رده‌بندی [10].
۲۰۰۴	بِیکر[۵] و همکاران	تشخیص استفاده نادرست دانشجویان از محیط آموزشی و رفتار غیر مسئولانهٔ آن‌ها با استفاده از روش‌های رده‌بندی [2].
۲۰۰۴	کُتسیانتیس[۶] و همکاران	بررسی کارایی دانشجویان در محیط آموزشی در کنار ارتباط ویژگی‌های تأثیرگذار در بهبود آن [8].
۲۰۰۶	یودِلسون[۷]	دانشجویان بر حسب الگوی استفاده به دو گروه متمایل به خطا و متمایل به درستی تقسیم‌بندی شده‌اند و با استفاده از نتایج این دسته‌بندی علل معمول رویداد اشتباه در دانشجویان مورد بررسی قرار گرفته است [13].
۲۰۰۶	کوکِتا[۸] و ویلِزال[۹]	تشخیص دانشجویان با انگیزهٔ پایین و یافتن روش‌هایی برای درمان و پیشگیری از انصراف آن‌ها با استفاده از روش‌های رده‌بندی [4].
۲۰۰۶	هامالینن[۱۰] و وینی[۱۱]	پیش‌بینی میزان موفقیت درس [6].
۲۰۱۰	گورولِر[۱۲] و همکاران	بررسی عوامل موثر بر موفقیت دانشجویان در دانشگاه و رسیدن به این نتیجه که انواع ثبت نام در دانشگاه و سطح درآمد خانواده ها در موفقیت دانشجویان موثر است [5].
۲۰۱۳	پِناآیالا[۱۳]	مروری بر فعالیت‌های اخیر صورت گرفته در حوزه‌های مختلف داده‌کاوی آموزشی از جمله رده‌بندی و خوشه‌بندی [11].

۳- روش تحقیق

در این بخش به روش پیشنهادی برای پیش‌بینی مشروط شدن یا نشدن یک دانشجو پرداخته خواهد شد.

فرآیند داده‌کاوی، یک چرخهٔ چهار مرحله‌ای تکرارشونده، شامل جمع‌آوری داده، پیش‌پردازش داده، اعمال تکنیک‌های داده‌کاوی و ارزیابی و تفسیر نتایج است. در شکل (۲) این چرخه را مشاهده می‌کنیم.



شکل ۲: فرایند داده‌کاوی [9]

۳-۱- جمع‌آوری داده‌ها

در این مقاله از پایگاه دادهٔ سامانهٔ آموزش دانشگاه قم استفاده شده‌است. این پایگاه داده شامل اطلاعات دانشجویان از سال ۱۳۷۰ تا ۱۳۹۱ در رشته‌های مختلف تحصیلی و در مقاطع کارشناسی، کارشناسی ارشد و دکتری تخصصی می‌باشد. همچنین این پایگاه داده شامل ۹۳۶۹ رکورد و۱۴ ویژگی است. این ویژگی‌ها عبارتند از: دانشکده، جنس، بومی بودن، تاریخ تولد، ترم ورود، رشتهٔ تحصیلی، مقطع، دوره، معدل کل، تعداد واحد گذرانده، تعداد مشروطی، تعداد مشروطی متوالی، آخرین وضعیت دانشجو و نوع سهمیه.

۳-۲- پیش‌پردازش داده‌ها

در پیش‌پردازش داده، دادهٔ خام با استراتژی‌ها و تکنیک‌های مختلف به شکل مناسب برای آنالیز داده تبدیل می‌گردد [11].

۳-۲-۱- افزودن ویژگی جدید

در این قسمت یک ویژگی جدید به مجموعه دادهٔ خود اضافه نموده‌ایم که مشخص می‌سازد که دانشجو تا کنون مشروط شده یا نه؟ مقدار این ویژگی با توجه به ویژگی «تعداد مشروطی» که در دادهٔ خام اولیه وجود دارد، محاسبه می‌شود. بدین ترتیب که اگر تعداد مشروطی بزرگ‌تر از صفر باشد این ویژگی مقدار یک و در غیر اینصورت مقدار صفر به خود می‌گیرد.

۳-۲-۲- نرمال‌سازی داده‌ها

نرمال‌سازی به معنای تغییر مقیاس داده‌ها به گونه‌ای است که آنها را به یک دامنهٔ معین نگاشت کند [7].

یکی از ویژگی‌های موجود در دادهٔ اولیه، ویژگی «تعداد واحد گذرانده» است. با توجه به اینکه داده‌های مورد استفاده در این مقاله، مربوط به سه مقطع کارشناسی، کارشناسی ارشد و دکتری است و حداکثرتعداد واحد گذرانده سه مقطع متفاوت است، بنابراین این ویژگی باید به ازای هر مقطع تحصیلی نرمال‌سازی گردد تا مشخص شود هر دانشجو چه میزان از واحدهای مقطع تحصیلی خود را گذرانده است. در این مقاله از روش نرمال‌سازی Min-max بهره گرفته شده است.

نرمال‌سازی با استفاده از روش Min-max

با فرض اینکه min_A و max_A مینیمم و ماکسیمم مقادیر ویژگی A باشند، روش نرمال‌سازی Min-max با استفاده از رابطهٔ (۱) مقدار v در بازه A را به مقدار v' در بازهٔ $[new_min_A , new_max_A]$ نگاشت می‌کند [7]:

$$v' = \frac{v-min_A}{max_A-min_A}(new_max_A - new_min_A) + new_min_A \qquad (۱)$$

بدین ترتیب ویژگی «تعداد واحد گذرانده» را به ازای هر مقطع تحصیلی به بازهٔ [0,100] نرمال‌سازی نمودیم.

۳-۲-۳- گسسته‌سازی داده‌ها

برخی از الگوریتم‌های داده‌کاوی مانند الگوریتم‌های رده‌بندی، برای کارایی بهتر نیاز دارند که در ورودی، دادهٔ گسسته‌سازی شده دریافت کنند. البته باید به این موضوع توجه داشت که در برخی از الگوریتم‌های رده‌بندی، گسسته-سازی با تعداد دستهٔ بسیار زیاد باعث کاهش کارایی الگوریتم می‌شود [12].

ویژگی «درصد واحد گذرانده» که در بخش نرمال‌سازی به دست آمد، یک ویژگی پیوسته است. برای گسسته‌سازی این ویژگی از سه روش مقادیر برابر، خوشه‌بندی و هیستوگرام با بخش‌بندی به بازه‌های با تعداد اعضای برابر[۱۴]، استفاده شده‌است.

گسسته‌سازی با استفاده از مقادیر برابر

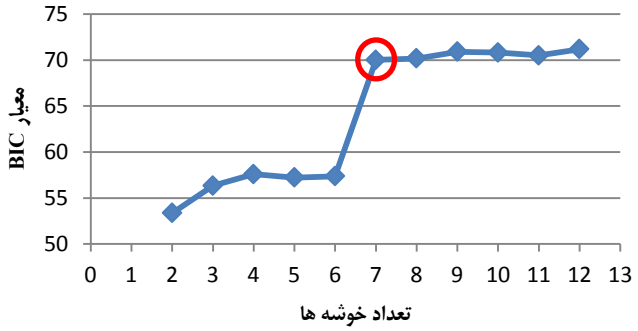
در این روش مقادیر برابر از ویژگی ای که قصد گسسته‌سازی آن را داریم، در یک دسته قرار می‌گیرند.

با این روش ویژگی «درصد واحد گذرانده» به ۱۶۱ دستهٔ مجزا تقسیم شد.

گسسته‌سازی با استفاده از خوشه‌بندی[۱۵]

تحلیل خوشه‌ای روش رایجی برای شناسایی گروه‌های مشابهی از اشیا به نام خوشه‌ها است، به طوری که اشیای در یک خوشه با هم مشابه و مرتبط و در عین حال متفاوت از اشیا در خوشه‌های دیگر هستند [12]. این روش یکی از روش‌های رایج در گسسته‌سازی داده‌هاست.

در این مقاله برای خوشه بندی از الگوریتم k-means استفاده شده است. همچنین برای به دست آوردن تعداد خوشهٔ بهینه، از نمودار تعداد خوشه در مقابل معیارBIC استفاده شده است. در این روش، نقطه‌ای که نشان‌دهندهٔ تغییر ناگهانی شیب است، به عنوان خوشهٔ بهینه انتخاب می‌شود. این نمودار در شکل (۳) قابل مشاهده است. همانطور که از این شکل پیداست، تعداد خوشهٔ بهینه، ۷ خوشه است. بنابراین با استفاده از این روش مقادیر ویژگی «درصد واحد گذرانده» به ۷ خوشه گسسته‌سازی می‌گردد.



شکل ۳: تعیین تعداد خوشه بهینه با استفاده از معیار BIC

گسسته‌سازی با استفاده از هیستوگرام با بخش‌بندی به بازه‌های با تعداد اعضای برابر

در این روش ابتدا نمودار هیستوگرام ویژگی مورد نظر رسم می‌گردد. سپس باید با استفاده از این نمودار، مقادیر ویژگی را به گونه‌ای بخش‌بندی می‌کنیم که هر بخش شامل تعداد برابری از رکوردها باشد [7]. البته در این روش ممکن است رکوردهایی وجود داشته باشند که مقادیر این ویژگی‌شان برابر باشد ولی در دو بخش مجزای متوالی قرار گرفته باشند. برای حل این مشکل رکوردهای با مقادیر برابر را فقط در یکی از این دو بخش (بخشی که مقادیر ویژگی آن کوچک‌تر است) قرار دادیم.

در این مقاله برای گسسته‌سازی ویژگی «درصد واحد گذرانده»، بر اساس توصیهٔ خبره، یک بار از اندازهٔ بخش ۲۰۰۰ رکورد (۵ بازه) و بار دیگر از اندازهٔ بخش ۱۰۰۰ رکورد (۹ بازه) استفاده شده‌است.

۳-۲-۴- متوازن‌سازی داده‌ها

داده‌های مورد استفاده در این مقاله نامتوازن بوده و اقلیت داده‌ها شامل دانشجویان مشروط‌شده هستند. در داده‌های نامتوازن که یک کلاس اعضای بیشتری دارد، ممکن است پیش‌بینی‌های مدل، متمایل به کلاس اکثریت شود و مدل در دام بیفتد. لذا برای پیش‌بینی صحیح‌تر و واقع‌بینانه‌تر، متوازن‌سازی امری اجتناب ناپذیر است. با بکارگیری روش‌های متوازن‌سازی کارایی و صحت الگوریتم‌های پیش‌بینی افزایش خواهد یافت [۱].

متوازن‌سازی با روش زیرنمونه‌گیری[۱۶]

در روش زیرنمونه‌گیری به طور تصادفی نمونه‌هایی از کلاس اکثریت حذف می‌شود تا زمانی که کلاس اقلیت درصدی از کلاس اکثریت شود [۱].

در این مقاله برای متوازن‌سازی داده‌ها از روشی مشابه روش زیرنمونه‌گیری استفاده شده‌است. به این ترتیب که ابتدا کل داده به دو دسته دانشجویان مشروط‌شده (کلاس اقلیت) و دانشجویان مشروطنشده (کلاس اکثریت) تقسیم و سپس با استفاده از روش نمونه‌گیری تصادفی، از دو کلاس ذکر شده، نمونه‌های تصادفی جدیدی ایجاد شد که در آن‌ها نسبت دانشجویان مشروطنشده به دانشجویان مشروط‌شده نسبت ۷۰ درصد به ۳۰ درصد بود. با همین روش نمونه‌های تصادفی دیگری با نسبت ۶۰ درصد به ۴۰ درصد ایجاد شد. لازم به ذکر است که ایجاد مجموعه‌های با نسبت ۷۰ به ۳۰ و ۶۰ به ۴۰ برای تمامی حالت‌های گسسته‌سازی شده، به صورت جداگانه و با استفاده از نمونه‌گیری تصادفی، انجام گرفته است. در ادامه از هر یک از این مجموعه‌های ایجاد شده، به عنوان ورودی الگوریتم رده‌بندی استفاده خواهد شد.

۳-۲-۵- کاهش ابعاد داده

با جمع‌بندی ویژگی پیشنهاد SQL Server Analysis و مشورت با خبره به این نتیجه رسیدیم که از بین ویژگی‌های معرفی شده در بخش ۳-۱ و ویژگی‌های جدیدی که در قسمت‌های قبلی ایجاد شده، موارد ذکر شده در جدول (۲) تاثیرگذارترین ویژگی‌ها در مشروطی یک دانشجو هستند.

جدول ۲: ویژگی‌های تاثیرگذار در مشروطی یک دانشجو

جنس
بومی بودن
رشته تحصیلی
مقطع
دوره
نوع سهمیه
درصد واحد گذرانده (گسسته‌سازی‌شده)

۳-۳- اعمال تکنیک‌های داده‌کاوی و ارزیابی و تفسیر نتایج

در این مقاله به منظور ایجاد مدل پیش‌بینی، از درخت تصمیم که یک ابزار قوی و متداول برای رده‌بندی و پیش‌بینی است، استفاده شده است.

۳-۳-۱- اجرای رده‌بندی بدون متوازن‌سازی

ابتدا ۵ بار الگوریتم درخت تصمیم C4.5 را بر روی دادهٔ نامتوازن اجرا نمودیم. لازم به ذکر است که روش گسسته‌سازی داده‌ها در تمام این ۵ بار، روش مقادیر برابر بود. از بین این ۵ مدل، مدلی با صحت[۱۷] حدود ۸۱ درصد، بالاترین صحت را داشت. با وجود داشتن صحت حدود ۸۱ درصد، نتایج حاصل از محاسبه سایر معیارها نظیر معیار F[۱۸] که در جدول (۴) آورده‌شده‌است، نشان می‌دهد که این میزان صحت قابل قبول نمی‌باشد. علت این امر این است که معیار صحت در مورد داده‌های نامتوازن که یک کلاس اعضای بیشتری دارد، متمایل به کلاس اکثریت است و حتی ممکن است تمام داده‌ها را متعلق به کلاس اکثریت معرفی کند.

لذا برای ارزیابی مدل‌ها نیاز به یک معیار دیگر داشتیم. نتایج مشورت ما با خبره سیستم مشخص کرد که این معیار باید بتواند ویژگی‌های زیر را ارضاء نماید:

۱. اگر دانشجویی مشروط شود ولی وضعیت مشروطی وی، «مشروطنشده» پیش‌بینی گردد هزینه زیادی در برخواهد داشت بنابرین مدلی مناسب است که تعداد دانشجویانی که مشروط شده‌اند ولی وضعیت مشروطی آن‌ها، «مشروطنشده» پیش‌بینی شده، مینیمم گردد که به تبع آن تعداد دانشجویانی که مشروط شده‌اند و وضعیت مشروطی آن‌ها، «مشروط‌شده» پیش‌بینی شده ماکسیمم می‌گردد.

۲. اگر دانشجویی مشروط نشود ولی وضعیت مشروطی وی، «مشروط‌شده» پیش‌بینی شود هزینه زیادی در برنخواهد داشت بنابراین نسبت به حالت قبلی از اهمیت کمتری برخوردار است.

اما معیاری که بتواند این ویژگی‌ها را ارضاء کند، معیارهایی همچون معیار صحت، فراخوانی[۱۹]، دقت[۲۰] و معیار F نیست چرا که، ارزش پارامترهای مد نظر خبره با یکدیگر متفاوت است. بنابراین برآن شدیم که از معیار صحت وزن‌دار[۲۱] استفاده نماییم. با استفاده از ماتریس اغتشاش[۲۲] که در جدول (۳) ارائه شده، معیار صحت وزن‌دار را با استفاده از رابطه (۲) محاسبه می‌کنیم: (وزن‌های معیار صحت وزن‌دار، با مشورت خبره تعیین شده است.)

جدول ۳: ماتریس اغتشاش

واقعی / پیش‌بینی شده	کلاس صفر (مشروط نشده)	کلاس یک (مشروط شده)
کلاس صفر (مشروط نشده)	a (وزن: ۲)	b (وزن: ۶)
کلاس یک (مشروط شده)	c (وزن: ۱)	d (وزن: ۶)

$$\text{صحت وزن‌دار} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d} = \frac{2a+6d}{2a+6b+c+6d} \qquad (۲)$$

جدول ٤: نتیجه ارزیابی درخت تصمیم C4.5 ، بدون متوازن‌سازی

صحت	معیار F	صحت وزن‌دار
۸۱.۲۶ درصد	۰.۱۱	۶۰.۴۶ درصد

بنابراین قبل از اجرای مدل، به منظور پیش‌بینی صحیح‌تر و واقع‌بینانه‌تر باید نسبت به متوازن‌سازی داده‌ها اقدام نمود.

۳-۳-۲- اجرای رده‌بندی با متوازن‌سازی

ابتدا به ازای هر یک از سه روش گسسته‌سازی معرفی شده در بخش ۳-۲-۳ مجموعه‌های آموزش[۲۳] و آزمون[۲۴] ایجاد می‌کنیم. همان‌گونه که در بخش ۳-۲-۴ توضیح داده‌شد، در تمامی مجموعه‌های آموزش و آزمون، نسبت دانشجویان مشروطنشده به دانشجویان مشروط‌شده دارای دو حالت ۷۰ به ۳۰ و ۶۰ به ۴۰ است. هر یک از حالت‌های ذکر شده در جدول (۵) را ، ۵ بار توسط الگوریتم درخت تصمیم C4.5 اجرا نموده و بهترین مدل از بین این ۵ مدل را، با توجه به معیار صحت وزن‌دار انتخاب می‌کنیم. نتایج حاصل از حالت‌های ذکرشده در جدول (۵) قابل مشاهده است.

جدول ٥: اجرای درخت تصمیم C4.5 با متوازن‌سازی

نسبت غیرمشروط-ها به مشروط‌ها	روش گسسته‌سازی	تعداد دسته‌ها در ویژگی گسسته‌سازی شده	ماکسیمم صحت وزن‌دار(مدل برتر)
۷۰ به ۳۰	مقادیر برابر	۱۶۱	۶۰.۲۳ درصد
	خوشه‌بندی	۷	۶۶.۱۳ درصد
	هیستوگرام	۵	۶۰.۶۳ درصد
		۹	۶۱.۱۲ درصد
۶۰ به ۴۰	مقادیر برابر	۱۶۱	۶۴.۰۲ درصد
	خوشه‌بندی	۷	۶۵.۵۴ درصد
	هیستوگرام	۵	۶۱.۷۵ درصد
		۹	۶۹.۱۹ درصد

با توجه به جدول(۵)، مدل با دقت ۶۹.۱۹ درصد، بیشترین میزان دقت وزن‌دار را دارد. برای اینکه بتوان به دقت به دست آمده در این مدل اعتماد کرد از روش ارزیابی k-fold cross validation با k=10 استفاده شده است. نتایج حاصل از این ارزیابی در جدول (۶) قابل مشاهده است.

جدول ٦: ارزیابی مدل برتر با روش 10-fold

شماره بخش	اندازه بخش	تعداد پیش‌بینی درست	دقت (درصد)
۱	۳۶۹	۲۵۵	۶۹.۱۱
۲	۳۷۰	۲۵۸	۶۹.۷۳
۳	۳۷۱	۲۵۷	۶۹.۲۷
۴	۳۷۱	۲۵۲	۶۷.۹۲
۵	۳۷۱	۲۵۳	۶۸.۱۹
۶	۳۷۱	۲۵۴	۶۸.۴۶
۷	۳۷۱	۲۵۶	۶۹.۰۰
۸	۳۷۱	۲۵۶	۶۹.۰۰
۹	۳۷۱	۲۶۰	۷۰.۰۸
۱۰	۳۶۹	۲۵۹	۷۰.۱۹
میانگین دقت: ۶۹.۰۹ درصد			

همان‌گونه که در جدول (۶) مشاهده می‌کنیم، میانگین دقت این مدل ۶۹.۰۹ درصد است. بنابراین می‌توان گفت که این مدل، یک مدل قابل اعتماد است.

۳-۳-۳- تفسیر مدل برتر

در شکل (۴) برشی از درخت تصمیم ایجاد شده را مشاهده می‌کنیم. در این شکل خط آبی نشان‌دهندهٔ میزان مشروط‌نشده‌ها و خط قرمز نشان دهندهٔ میزان مشروط‌شده‌ها است.

چند نمونه از نتایجی که از شکل (۴) قابل استنباط است به شرح زیر می‌باشد:

۱. قرار گرفتن ویژگی گسسته‌سازی شدهٔ درصد واحد گذرانده در سطوح اولیهٔ مدل برتر، نشان دهندهٔ میزان تاثیر بالای این ویژگی بر مشروطی است.

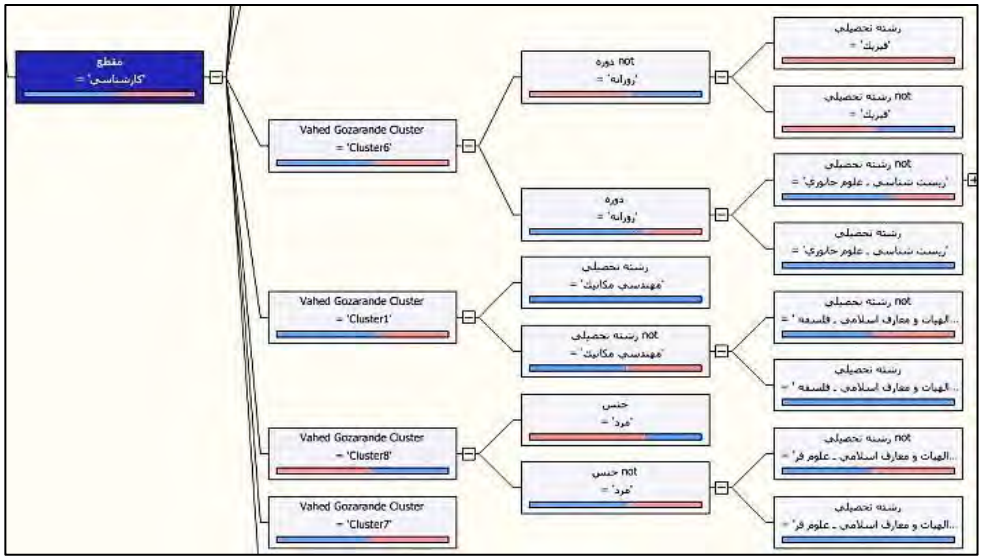
۲. در مقطع کارشناسی دانشجویانی که درصد واحد گذرانده‌شان Cluster6 (بازه ۳۸.۸۸ درصد تا ۵۶.۲۵ درصد) است و دانشجوی دوره روزانه نیستند و رشتهٔ تحصیلی آنها فیزیک است، همواره مشروط شده‌اند.

۳. در مقطع کارشناسی دانشجویانی که درصد واحد گذرانده‌شان Cluster8 (بازه ۷۵ درصد تا ۹۰.۶ درصد) است و جنس آن‌ها زن است و رشتهٔ تحصیلی آنها علوم قرآنی است، هیچ‌گاه مشروط نشده‌اند.

به همین ترتیب سایر نتایج نیز از روی درخت تصمیم به راحتی قابل تفسیر است.

٤- نتیجه

در این مقاله مدلی برای پیش‌بینی مشروط شدن یا نشدن یک دانشجو در دوران تحصیلش ارائه گردید. برای پیش‌بینی واقع‌بینانه‌تر، از متوازن‌سازی داده‌ها بهره گرفته‌شد. برای ارزیابی مدل‌ها از معیار صحت وزن‌دار استفاده شد، چرا که به دلیل متفاوت بودن ارزش پارامترهای مد نظر خبره، معیارهایی



شکل ٤: برشی از مدل برتر

dataacademy.ir

همچون صحت و معیار F، معیارهای مناسبی نبودند. همچنین بـرای رسـیدن به مدل با بیشترین دقت ترکیبی از روش‌های مختلف گسسته‌سازی و نسبت-های مختلف متوازن‌سازی (۷۰ به ۳۰ و ۶۰ به ۴۰) مورد آزمون قـرار گرفـت. نتایج حاصل از مدل برتر به سادگی از روی ایـن مـدل قابـل تفسـیر هسـتند. نمونه‌ای از این نتایج، به شرح زیر اسـت: در مقطـع کارشناسـی ویژگـی‌هـای میزان واحد گذرانده و رشتهٔ تحصیلی و در مقطع کارشناسی ارشد ویژگی دورهٔ تحصیلی از تاثیرگذارترین ویژگی‌ها بر مشروطی هستند.

در آخر پیشنهاد می‌شود برای متوازن‌سازی داده‌ها از روش‌های ترکیبـی همانند روش Modified bagging بهره گرفته شود. استفاده از این روش‌ها ممکن است باعث افزایش دقت مدل گردد.

مراجع

[۱] فیاضی، مهری، قیاسی، راضیه، دهنوی باقری، ملیحه ، مینائی بیدگلی، بهروز ، **"مدیریت داده های نامتوازن برای پیش بینی مرگ و میر در حوادث هوایی با استفاده از تکنیک های داده کاوی"**، دومین کنفرانس ملی مهندسی نرم افزار، لاهیجان، دانشگاه آزاد اسلامی واحد لاهیجان، ۱۳۹۱.

[2] Baker, R. S., Corbet, A. T., Koedinger, K., "*Detecting Student Misuse of Intelligent Tutoring Systems*", Proceedings of the 7th International Conference on Intelligent Tutoring Systems, pp. 531-540, Maceio, Brazil, 2004.

[3] Chen, G., Liu, C., Ou, K., Liu, B., "*Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology*", *Journal of Educational Computing Research*, Vol. 23, No. 3, pp. 305–332, 2000.

[4] Cocea, M., Weibelzahl, S., "*Can Log Files Analysis Estimate Learners' Level of Motivation?*", Proceedings of the 14th Workshop on Adaptivity and User Modeling in Interactive Systems, pp. 32-35, Hildesheim, Germany, 2006.

[5] Guruler, H., Istanbullu, A., Karahasan, M., "*A new student performance analysing system using knowledge discovery in higher educational databases*", Computers & Education, Vol. 55, pp. 247–254, 2010.

[6] Hamalainen, W., Vinni M., "*Comparison of machine learning methods for intelligent tutoring systems*", Proceedings of the 8th international conference intelligent tutoring systems, pp. 525–534, Taiwan, 2006.

[7] Han, J., Kamber, M., Pei, J., *Data mining: concepts and techniques*, Second Edition, Morgan kaufmann Publishing, 2006.

[8] Kotsiantis, S. B., Pierrakeas, C. J., Pintelas, P. E., "*Predicting Students' Performance in Distance Learning Using Machine Learning Techniques*", Applied Artificial Intelligence, Vol. 18, No. 5, pp. 411–426, 2004.

[9] Minaei-Bidgoli, B., Hani, S.H., Ghanbari, V., "*Data Mining in the E-Learning Systems: A Virtual University Case Study*", International Journal of Information & Communication Technology Research (IJICTR), Vol. 4, No. 2, pp. 71-81, 2012.

[10] Minaei-Bidgoli, B., Punch, B., "*Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System*", Genetic and Evolutionary Computation, Vol. 2, pp. 2252–2263, 2003.

[11] Peña-Ayala, A., "*Educational Data Mining: A Survey and a Data Mining-based Analysis of Recent Works*", Expert Systems with Applications, 2013.

[12] Tan, P.N., Steinbach M., Kumar, V., *Introduction to data mining*, Pearson Addison Wesley, 2006.

[13] Yudelson, M. V., et al., "*Mining Student Learning Data to Develop High Level Pedagogic Strategy in a Medical ITS*", Proceedings of the AAAI Workshop on Educational Data Mining, pp. 1-8, Boston, MA, USA, 2006.

زیرنویس‌ها

[1] Educational data mining
[2] Classification
[3] G. Chen
[4] B. Punch
[5] R. S. Baker
[6] S. Kotsiantis
[7] M. V. Yudelson
[8] M. Cocea
[9] S. Weibelzahl
[10] W. Hamalainen
[11] M. Vinni
[12] H. Guruler
[13] A. Peña-Ayala
[14] Equal frequency histogram
[15] Clustering
[16] Under-sampling
[17] Accuracy
[18] F-measure
[19] Recall
[20] Precision
[21] Weighted accuracy
[22] Confusion matrix
[23] Train set
[24] Test set

آکادمی داده
dataacademy.ir