نامهٔ آموزش عالی، دورهٔ جدید
سال سوم، شمارهٔ دوازدهم، زمستان 1389، صص 124- 107

تاریخ دریافت مقاله 89/03/29
تاریخ پذیرش 89/06/08

# دسته‌بندی و پیش‌بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک‌های داده‌کاوی

**سمیّه حیدری** *
**مسعود یقینی** **

## چکیده

داده‌کاوی و کشف الگوها و دانش نهفته در داده‌های سیستم‌های آموزشی کمک شایانی به تصمیم‌گیرندگان عرصه آموزش عالی جهت بهبود فرآیندهای آموزشی نظیر برنامه‌ریزی، ثبت‌نام، ارزیابی و مشاوره می‌نماید. هدف مقاله حاضر، دسته‌بندی و پیش‌بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک‌های داده‌کاوی است. در این مقاله سعی شده با استفاده از داده‌های دموگرافیک و سوابق تحصیلی دانشجویان و آماده‌سازی مناسب داده‌ها و با کمک تکنیک‌های دسته‌بندی درخت تصمیم، رگرسیون لجستیک، نزدیکترین همسایگی و شبکه‌های عصبی مدل‌های مختلفی برای پیش‌بینی وضعیت تحصیلی دانشجویان در نیمسال آینده ارائه شود. در نهایت مقایسه‌ای میان نتایج حاصل از تکنیک‌های مختلف صورت گرفته و بهترین مدل‌ها در دسته‌بندی صحیح دانشجویان مدل نزدیکترین همسایگی و سپس شبکه‌های عصبی شناخته شده‌اند. بر همین اساس می‌توان مدل‌های پیشنهادی را به عنوان یک ابزار پشتیبان تصمیم‌گیری در سیستم‌های آموزشی مورد استفاده قرار داد.

**واژگان کلیدی**: آموزش عالی، داده‌کاوی آموزشی، دسته‌بندی، پیش‌بینی، موفقیت تحصیلی دانشجو، روند تحصیلی دانشجو

---

* کارشناسی ارشد مهندسی صنایع، دانشگاه علم و صنعت ایران- مدرس دانشکده مهندسی صنایع دانشگاه آزاد اسلامی واحد پرند (مسئول مکاتبات: Somayeh.Heydari@Gmail.com)
** عضو هیئت علمی دانشگاه علم و صنعت ایران

**مقدمه**

امروزه در اکثر دانشگاه‌های ایران اسلامی بانک‌های اطلاعاتی وسیعی از ویژگی‌های دانشجویان موجود است که حجم بالایی از اطلاعات مربوط به سوابق آموزشی، تحصیلی و ... را شامل می‌شود. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزش عالی در جهت ارتقاء و بهبود فرآیندهای آموزشی نظیر برنامه‌ریزی، ثبت‌نام، ارزیابی و مشاوره کمک شایانی نماید. نرم‌افزارهای کامپیوتری بکار گرفته شده برای این منظور، غالباً فقط برای مکانیزه کردن وضع موجود، اجرای پرس‌وجوهای معمولی و برنامه‌ریزی کوتاه مدت اداری جوابگو هستند. در حالیکه در عمق درون این حجم داده‌ها، الگوها و روابط بسیار جالبی میان پارامترهای مختلف به صورت پنهان باقی می‌ماند (رومرو[1] و ونتورا[2]، 2007). الگوهای پنهان کشف شده، سیستم‌های آموزش عالی را در تصمیم‌گیری بهتر و داشتن طرح پیشرفته‌تری در هدایت دانشجویان کمک می‌کند(یانگ[3]، 2006؛ لوان[4]، 2001). داده‌کاوی یک تکنیک میان رشته‌ای برای اکتشاف این الگوها می‌باشد که از علوم یادگیری ماشین، تشخیص الگو، آمار، پایگاه داده و بصری‌سازی به منظور استخراج اطلاعات از پایگاه‌های داده بزرگ بهره‌مند می‌شود (کابانا[5] و همکاران، 1998).

دسته‌بندی از عملیات‌های نظارتی می‌باشد که در داده‌کاوی بسیار رایج و مورد استفاده است. دسته‌بندی عبارتست از تخصیص داده‌ها بر اساس ویژگی‌هایشان به دسته‌هایی که نام آنها از قبل مشخص می‌باشد. دسته‌بندی برای یادگیری قواعد و یا ساختن مدلی به منظور پیش‌گویی دستهٔ داده‌های جدید بکار می‌رود. به عنوان مثال اگر دانشجویان یک دانشکده را بر اساس خصوصیاتشان به دو گروه دانشجویان سریع و کند در گذراندن دروس تقسیم کرده باشیم برای تعیین گروه دانشجویان جدید از دسته‌بندی استفاده می‌کنیم. دسته‌بندی داده‌ها فرآیندی دو مرحله‌ای است. اولین مرحله ساخت مدل و دومین مرحله پیش‌گویی برای داده‌های جدید با استفاده از مدل ساخته شده می‌باشد.

1. Romero
2. Ventura
3. Yang
4. Luan
5. Cabana

در این مقاله، با بررسی تکنیک‌های دسته‌بندی و پیش‌بینی شامل شبکه‌های عصبی، درخت تصمیم، نزدیکترین همسایگی و رگرسیون لجستیک به پیدا کردن الگوها و دانش نهفته در داده‌های سیستم آموزش و ارایه مدلی برای پیش‌بینی وضعیت تحصیلی دانشجویان در ترم آینده پرداخته‌ایم. می‌توان مسیر تحصیلی دانشجو و وضعیت تحصیلی وی را در نیمسال‌های بعدی (مشروطی، ممتازی و...) به منظور تسهیل اقدامات زمان‌بندی و برنامه‌ریزی آموزشی پیش‌بینی کرد. با تشخیص دانشجویان تحت ریسک به کمک داده‌کاوی می‌توان از شکست و حذف آنها جلوگیری کرد و با شناسایی دانشجویان مستعد و قوی می‌توان منابع را به نحو مناسب‌تری تخصیص داد تا منافع بیشتری تأمین گردد.

**پیشینه تحقیق**

به منظور افزایش استانداردهای آموزشی، نیازمند سیستم داده‌کاوی هستیم که دانش و بصیرت مورد نیاز را برای تصمیم‌گیرندگان در سیستم آموزش عالی فراهم کند. متأسفانه با وجود انبوه داده‌های موجود در سیستم آموزش دانشگاه‌های ایران، هیچ‌گاه بررسی عمیق و جامعی برای استخراج دانش نهفته از این داده‌ها انجام نشده است. کمبود دانش بسنده و کافی در سیستم آموزش عالی از دستیابی مدیریت سیستم به اهداف کیفی‌شان جلوگیری می‌کند. تکنولوژی داده‌کاوی کمک می‌کند تا این شکاف دانشی در سیستم آموزش عالی جبران شود. الگوهای پنهان، وابستگی‌ها و استثناهایی که توسط بعضی از تکنیک‌های داده‌کاوی کشف شده‌اند، می‌توانند برای بهبود کارایی، اثربخشی و سرعت فرآیندها استفاده شوند. در نتیجه، این بهبود مزایای بسیاری از قبیل حداکثر کردن کارایی سیستم آموزشی، کاهش نرخ از دست دادن و حذف دانشجویان، افزایش نرخ ارتقاء دانشجویان، کاهش مدت زمان ماندگاری دانشجویان، افزایش موفقیت دانشجو، افزایش خروجی یادگیری دانشجو و کاهش هزینه فرآیندهای سیستم را برای سیستم آموزش عالی به ارمغان می‌آورد. (حیدری و یقینی، 1387). دانش قابل کشف از طریق داده‌کاوی در حوزه آموزش نه تنها قابل استفاده صاحبان سیستم یعنی مدرسین و مسئولین آموزشی بلکه قابل استفادهٔ کاربران سیستم یعنی دانشجویان نیز می‌باشد (رانجان[1] و مالیک[2]، 2007). این با ارائه توصیه‌هایی

1. Ranjan
2. Malik

می‌تواند به دانشجویان کمک کند تا فرآیند یادگیری را ارتقاء دهند و موفق عمل کنند. از طرفی بازخوردهای عینی به مدرسین ارائه می‌دهد که از طریق آن می‌توانند کارایی فرآیند یادگیری را ارتقا دهند و به مسئولین آموزشی کمک می‌کند تا منابع سازمانی اعم از مادی و انسانی را به نحو مناسب‌تری تخصیص دهند (رومرو و همکاران، 2008).

با توجه به نقش مهم افزایش سرعت کامپیوترها در پیشرفت علم داده‌کاوی می‌توان گفت زمینه‌هایی که در داده‌کاوی داده‌های سیستم آموزش مطرح‌اند، بسیار گسترده می‌باشند. مخصوصاً که در کشور ما، کار خاصی در این زمینه صورت نگرفته است. از طرفی با توجه به اینکه روش‌های پذیرش و تحصیل دانشجو و اهداف مرتبط با پذیرش دانشجو و حفظ و نگهداری وی برای مؤسسات آموزشی داخل و خارج از کشور بسیار با هم متفاوت می‌باشد، لازم است که توانایی‌های بالقوه داده‌کاوی در ارتقاء فرآیندهای سیستم آموزش دانشگاه‌های داخل کشور به نحو جداگانه مورد بررسی و تحلیل قرار بگیرند. در حقیقت، بسیاری از تکنیک‌های داده‌کاوی بکارگرفته شده در سیستم‌های آموزشی خارج از کشور، یا قابل اعمال به داده‌های سیستم‌های آموزشی دانشگاه‌های ما نمی‌باشند و یا در صورت داشتن قابلیت اعمال، نتایج ارزشمندی را جهت ارتقاء فرآیندهای سیستم آموزشی به ارمغان نمی‌آورند.

به منظور پیدا کردن کاربردی کلیدی از داده‌کاوی که از طرفی قابل اعمال بر روی داده‌های سیستم آموزش دانشگاه‌های ایران باشد و از طرفی بتواند نتایج ارزشمندی را برای سیستم به ارمغان بیاورد، بررسی‌های متعددی بر روی داده‌های سیستم مکانیزه آموزش دانشگاه علم و صنعت ایران صورت گرفت. از طرفی، مشورت‌هایی با خبرگان این حوزه صورت گرفت. از داده‌های دانشجویان به شیوه‌های مختلف و در جهت اهداف مختلفی می‌توان استفاده نمود. یکی از استفاده‌هایی که می‌تواند بسیار سودمند واقع شود، دسته‌بندی دانشجویان با کمک ویژگی‌های آنها می‌باشد. این دسته‌بندی می‌تواند بر پایه معیارهای متفاوتی باشد، مثلاً دانشجویان سریع و کند در گذراندن دروس، دانشجویانی که به خدمت جدیدی پاسخ مثبت می‌دهند، دانشجویان ممتاز، مشروط، عادی، ترک تحصیلی، اخراجی و غیره. سرانجام با کمک نظرات خبرگان، دسته‌بندی و پیش‌بینی وضعیت تحصیلی دانشجویان در ترم آتی به عنوان یکی از کاربردهای کلیدی داده‌کاوی قابل اعمال بر روی داده‌های این مرکز مورد توجه قرار گرفت. زمانی که با استفاده از داده‌کاوی دانشجویان تحت ریسک مشخص می‌شوند، می‌توان مشاوره‌های لازم را برای پیشگیری از رسیدن دانشجویان به

وضعیت بحرانی بکار گرفت و از شکست و حذف آنها جلوگیری کرد، قبل از اینکه حتی دانشجویان خودشان از اینکه تحت ریسک هستند، مطلع باشند و یا زمانی که دانشجویان مستعد و قوی شناسایی می‌شوند، می‌توان منابع را به نحو مناسب‌تری تخصیص داد تا منافع بیشتری تأمین شود. می‌توان بررسی نمود که وضعیت تحصیلی دانشجویان در نیمسال بعدی با سوابق آموزشی آنان تا نیمسال جاری رابطه دارد. در حقیقت وضعیت تحصیلی دانشجویان (نظیر مشروطی و ممتازی) در نیمسال بعدی بر اساس مشخصات فردی دانشجو (نظیر سن، جنسیت، وضعیت تأهل) و سوابق آموزشی دانشجویان (نظیر رشته، سهمیه ورودی، تعداد واحد اخذ شده و گذرانده، نمرات کسب شده و ...) قابل پیش‌بینی است.

بسیاری از تکنیک‌های داده‌کاوی که در دنیای تجارت استفاده می‌شود، قابل انتقال به حوزه آموزش عالی می‌باشند. تقریباً تمامی الگوریتم‌ها و مدل‌هایی که در حال حاضر در بخش کسب و کار مورد استفاده قرار می‌گیرند، مستقیماً یا با اندکی تغییرات قابل استفاده برای تحقیق در حوزه آموزش عالی علی‌الخصوص تحقیقات مؤسسه‌ای می‌باشند (شانبران[1] و هیلبرت[2]، 2007). اما استفاده از داده‌کاوی در سیستم‌های آموزشی الزامات ویژه‌ای دارد که در سایر حوزه‌ها مطرح نیست، علی‌الخصوص نیازمند در نظر گرفتن جنبه‌های آموزشی یاد گیرنده (دانشجو) و سیستم می‌باشد. داده‌کاوی می‌تواند بر روی داده‌هایی که از دو نوع سیستم آموزشی استخراج می‌شود 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 اعمال تکنیک‌های داده‌کاوی در هر یک از این دو نوع سیستم به صورت جداگانه مورد بررسی قرار بگیرد (بیک‌زاده[3] و دلاوری[4]، 2004). بررسی مقاله‌های مرتبط با این حوزه نشان می‌دهد که اکثر کارهای انجام شده در زمینه اعمال تکنیک‌های داده‌کاوی در آموزش عالی با تمرکز بر روی تکنیک‌های وب‌کاوی و متن‌کاوی در حوزه آموزش الکترونیکی انجام شده است و تا کنون در حوزه آموزش غیرالکترونیکی کار جدی در زمینه پیش‌بینی وضعیت تحصیلی دانشجویان در نیمسال‌های آتی صورت نگرفته است. تمرکز اصلی این مقاله بر روی کاربرد سایر تکنیک‌های داده‌کاوی در حوزه آموزش عالی و مرتبط با سیستم‌های آموزش غیرالکترونیکی می‌باشد.

1. Schonbrunn
2. Hilbert
3. Beikzadeh
4. Delavari

در این تحقیق، سعی شده است ضمن شناسایی ویژگی‌های مؤثر در پیش‌بینی وضعیت تحصیلی دانشجویان، با اعمال تکنیک‌های مختلف داده‌کاوی مانند شبکه‌های عصبی، درخت تصمیم، نزدیکترین همسایگی و رگرسیون لجستیک بر روی داده‌های مربوط به مشخصات فردی، آموزشی و تحصیلی دانشجویان تا نیمسال کنونی و مقایسه نتایج تکنیک‌های مختلف، بهترین مدل برای پیش‌بینی وضعیت تحصیلی دانشجویان در نیمسال آینده ارایه شود. می‌توان مدل‌های ساخته شده را به عنوان یک ابزار پشتیبان تصمیم‌گیری در سیستم‌های آموزشی مورد بهره‌برداری قرار داد.

## روش‌شناسی پژوهش

پیش‌بینی وضعیت تحصیلی دانشجویان و شناخت دانشجویان تحت ریسک بالا یا دانشجویان مستعد به دست‌اندرکاران عرصه آموزش کمک می‌کند با انجام برنامه‌ریزی‌های تحصیلی مناسب و ارائه مشاوره‌های تحصیلی و مشوق‌های انگیزشی و انجام سایر اقدامات پیشگیرانه از صرف هزینه‌های بی‌مورد و هدر رفتن منابع و استعدادهای جامعه جلوگیری نمایند و سطح علمی دانشگاه را ارتقاء دهند. در این مقاله سعی شده است تا با ارایه مدل‌هایی تا حد امکان دقیق و قابل اعتماد پاسخی مناسب به این نیاز داده شود. برای مدل‌سازی داده‌ها برای دسته‌بندی دانشجویان از متدولوژی کریسپ[1] استفاده شده است. برای کسب اطلاعات بیشتر در مورد این متدولوژی انجام پروژه‌های داده‌کاوی می‌توان به www.crisp-dm.org مراجعه کرد. در این تحقیق، برای مدل‌سازی داده از نرم‌افزار کد باز وکا[2] استفاده شده است که دارای قابلیت‌هایی متناسب با کار ما می‌باشد.

پایگاه داده‌ای که از آن جهت عملیات داده‌کاوی استفاده شده، سیستم مکانیزه آموزش دانشگاه علم و صنعت ایران می‌باشد که شامل کلیه خصوصیات فردی و آموزشی دانشجویان رشته‌ها و مقاطع مختلف می‌باشد. داده‌های استخراجی ما از این سیستم مکانیزه شامل 3782 رکورد از دانشجویان رشته‌های مختلف در مقطع کارشناسی می‌باشند که ورودی سال‌های 82 تا 86 می‌باشند. این داده‌ها شامل 46 خصوصیت فردی و تحصیلی دانشجویان می‌باشد که از نیمسال اول سال تحصیلی 82-83 تا نیمسال دوم سال تحصیلی 86-87 جمع‌آوری شده‌اند.

1. CRISP- DM
2. Weka

کارآیی هر فرآیند داده‌کاوی مستقیماً مرتبط با داده‌های مورد استفاده می‌باشد. آماده‌سازی داده‌ها 50% تا 90% وقت و کار از تمام فرآیند کشف دانش را به خود اختصاص می‌دهد. در اولین مرحله پس از انتخاب داده‌ها، به سراغ انجام تغییرات در داده‌ها برای آماده‌سازی آنها برای دسته‌بندی رفتیم از جمله: ساخت ویژگی‌های جدید، حذف ویژگی‌های همبسته، حذف رکوردهای دارای کلاس پوچ[1]، حذف رکوردهای دارای مقادیر گمشده و یا جایگذاری دستی مقادیر گمشده، حذف ویژگی‌های بی‌تأثیر در ساخت مدل‌ها با استفاده از تعیین ارزش هر ویژگی با محاسبه مقدار آمارهٔ خی دو ($\chi^2$) مرتبط با متغیر کلاس (ارزیابی 10- لایه) و بصری‌سازی داده‌ها، حذف نقاط پرت، گسسته‌سازی برخی ویژگی‌ها نظیر گسسته‌سازی معدل نیمسال‌های مختلف به 5 بازه مجزا و گسسته‌سازی برچسب کلاس (معدل نیمسال 862) به 3 دسته به صورت ذیل.

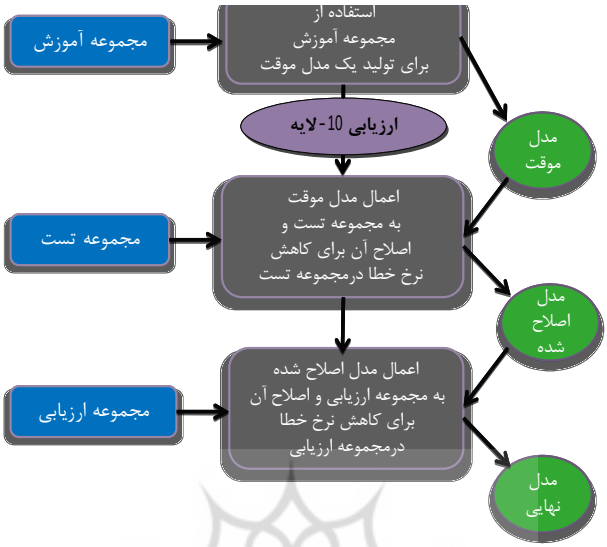
- اگر معدل $\leq$ 17 آنگاه "A"
- اگر 17 > معدل $\leq$ 12 آنگاه "B"
- اگر 12 > معدل آنگاه "C"

برای ساخت مدلی که بتواند رکوردهای مطلوب (کلاس C) را به درستی دسته‌بندی کند، مجموعه آموزش باید حاوی تعداد کافی رکورد مطلوب باشد. با ایجاد نمونه‌ای متوازن[2] از داده‌ها می‌توان اختلاف میان فراوانی کلاس‌های مختلف را کاهش داد و نسبت‌های تقریباً برابری از انواع رکوردها را در مجموعه آموزش ایجاد نمود (لاروز[3]، 2006). یک نمونه متوازن اما کوچک از داده‌ها به یک نمونه بزرگ از داده‌ها که حاوی نسبت کمی از رکوردهای مطلوب می‌باشد، ارجحیت دارد. برای متوازن نمودن داده‌ها می‌توان یا از کلاس‌های مختلف نمونه‌گیری کرد و نسبت رکوردهای مربوط به مقادیر کلاسی رایج را کاهش داد و یا با افزودن یک فاکتور وزنی به رکوردهای با فراوانی نسبی بیشتر وزن کمتری را به آنها القا کرد (بری[4] و لیناف[5]، 2004). با استفاده از نمونه‌گیری تصادفی بدون جایگذاری و کاهش نسبت رکوردهای مربوط به مقادیر کلاسی رایج در میان مجموعه داده‌ها (کاهش تعداد نمونه‌های مربوط

1. null
2. Balanced Sample
3. Larose
4. Berry
5. Linoff

به دسته‌های A و B) یک مجموعه داده متوازن از انواع کلاس‌ها تهیه گردید. در مرحلهٔ بعد بخش‌بندی داده‌ها به این نحو انجام شد که 75% داده‌ها به مجموعه آموزش اختصاص یافت. داده‌های مجموعه آموزش شامل 657رکورد بودند که 226 رکورد آن به کلاس A، 198 رکورد آن به کلاس B و 233 رکورد آن به کلاس C متعلق بود. 25% داده‌ها را نیز به مجموعه ارزیابی اختصاص یافت. داده‌های مجموعه ارزیابی هم شامل 220 رکورد است. ویژگی‌های نهایی مورد نظر برای ساخت مدل‌ها، 14 ویژگی بود که شامل جنسیت، سهمیه ورود به دانشگاه، سال تولد، رشته تحصیلی، مشروطی نیمسال اول، درصد نیمسال‌های مشروطی، درصد نیمسال‌های ممتازی، معدل نیمسال اول، معدل نیمسال کنونی (نیمسال 861)، معدل کل تا نیمسال کنونی (نیمسال 861)، نسبت (درصد) واحدهای گذرانده در نیمسال آخر، نسبت (درصد) کل واحدهای رد شده، میانگین تعداد واحدهای گذرانده در هر نیمسال، میانگین تعداد واحدهای رد شده در هر نیمسال می‌باشند. متدولوژی یادگیری نظارتی و ساخت مدل‌ها در شکل (1) آمده است.

در بخش‌های بعد مختصری از تکنیک‌های دسته‌بندی بکار رفته برای ساخت مدل‌ها و پارامترهای بکار رفته برای ساخت هر مدل، مطرح می‌شوند. در ساخت هر مدل، پارامترهای مختلفی وجود دارد که در نتایج نهایی مدل‌ها تأثیر بسزایی دارند. در ساخت مدل‌ها با توجه به نوع داده‌ها و اهداف ما از ساخت مدل‌ها، در مورد هر الگوریتم تغییرات زیادی بر روی پارامترهای مختلف صورت گرفته است تا در نهایت بهترین مدل استخراج گردیده است. پس از ساخت مدل‌ها نتایج حاصل از ارزیابی آنها در سه حالت مختلف ارائه شده است. علاوه بر این اطلاعات مربوط به ماتریس اغتشاش هر مدل بر روی مجموعه ارزیابی ارائه شده است تا از این طریق مقایسه بهتری میان مدل‌ها در زمینه تشخیص صحیح دسته‌های مختلف صورت گیرد. برای ساخت همهٔ این مدل از نرم‌افزار وکا نسخه 3.5.8 استفاده شده است.




**شکل (1) متدولوژی مدل‌سازی نظارتی**

**الف: استفاده از تکنیک درخت تصمیم C4.5 برای دسته‌بندی دانشجویان**

مدل‌های مختلف درخت تصمیم در داده‌کاوی برای کاوش در داده‌ها و استنتاج قوانینی که می‌توانند برای دسته‌بندی و پیش‌بینی مورد استفاده قرار بگیرند، استفاده می‌شوند. الگوریتم استفاده شده در این تحقیق برای ساخت درخت تصمیم، C4.5 می‌باشد. از جمله خصوصیات الگوریتم C4.5 استفاده از مفهوم کسب اطلاعات یا کاهش آنتروپی برای انتخاب شکست بهینه، استفاده از شکست‌های چندتایی در داده‌های رده‌ای (فقط محدود به شکست‌های دودویی نمی‌باشد)، کاهش خطای هرس کردن و سازگاری با ویژگی‌های پیوسته می‌باشد (لاروز، 2005).

برای ساخت مدل درخت تصمیم در ابتدا به رتبه‌بندی ویژگی‌ها با استفاده از معیار کسب اطلاعات (با کمک اعتبارسنجی 10-لایه) پرداختیم. سپس به حذف برخی ویژگی‌های با رتبه کمتر جهت افزایش دقت مدل و کاهش پیچیدگی آن پرداخته شد. سرانجام 9 ویژگی برای ساخت درخت نهایی استفاده شد که شامل: سهمیه ورود به دانشگاه، سال تولد، درصد نیمسال‌های مشروطی، درصد نیمسال‌های ممتازی، معدل نیمسال کنونی (نیمسال 861)، معدل کل تا نیمسال کنونی (نیمسال 861)، نسبت (درصد) کل واحدهای رد شده، میانگین تعداد واحدهای گذرانده در هر نیمسال،

میانگین تعداد واحدهای رد شده در هر نیمسال می‌باشند. یکی دیگر از پارامترهای تأثیرگذار در ساخت درختی قدرتمند، حداقل تعداد نمونه‌ها در هر برگ می‌باشد که با توجه به تعداد رکوردهای آموزشی باید حداقلی را برای آن تعیین نمود تا از بیش برازش مدل جلوگیری شود (ادلستین[1]، 1998). حداقل تعداد نمونه‌ها برای هر برگ در مدل نهایی 20 در نظر گرفته شد تا هم مدل ما دچار پیچیدگی و بیش برازش نگردد و هم دقت آن کاهش قابل ملاحظه‌ای نیابد. پس از ساخت درخت با استفاده از هرس C4.5 و ارزیابی 3- لایه، درخت ساخته شده هرس شد. خصوصیات درخت نهایی ایجاد شده به قرار زیر است:

- تعداد ویژگی‌های استفاده شده: 9 ویژگی؛
- تعداد برگ‌ها: 22؛
- اندازه درخت: 30؛
- عمق درخت: 6؛
- حداقل تعداد نمونه‌ها در هر برگ: 20.

**جدول (1) ماتریس اغتشاش حاصل از آزمایش مدل ساخته شده بر روی مجموعه ارزیابی(تکنیک درخت تصمیم C4.5)**

| | | کلاس پیش‌بینی شده | | |
|---|---|---|---|---|
| | | A | B | C |
| کلاس واقعی | A | 64 | 9 | 3 |
| | B | 22 | 21 | 24 |
| | C | 2 | 5 | 70 |

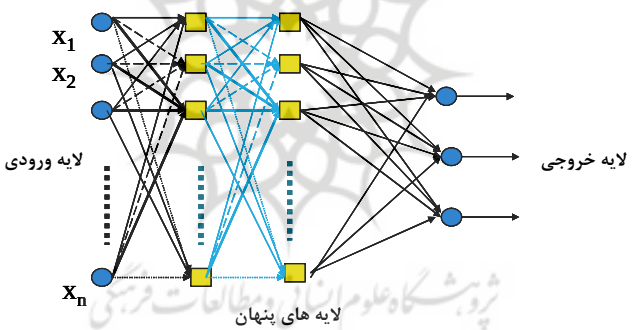
درخت ساخته شده کلاس‌های C و در مراحل بعدی به ترتیب کلاس‌های A و B را به خوبی تشخیص می‌دهد. دقت این درخت در مجموعه ارزیابی و اعتبارسنجی 10-لایه در مقایسه با مدل‌های ساخته شده با سایر تکنیک‌ها، دقت مناسبی می‌باشد.

**ب: استفاده از تکنیک شبکه‌های عصبی برای دسته‌بندی دانشجویان**

مطالعه شبکه‌های عصبی مصنوعی متأثر از سیستم‌های یادگیری طبیعی است که در آنها یک مجموعه پیچیده از نرون‌های به هم متصل در کار یادگیری دخیل هستند. این

1. Edelstein

شبکه‌ها در لایه‌هایی سازماندهی شده‌اند که هر لایه متشکل از تعدادی نرون می‌باشد. نرون‌های لایه‌های مختلف توسط یک سری اتصالات وزن‌دار به هم متصل شده‌اند. اولین لایه، لایه ورودی می‌باشد که در آن هر نرون منسوب به یک متغیر ورودی می‌باشد. آخرین لایه هم لایه خروجی می‌باشد که منسوب به متغیر یا متغیرهای پیش‌بینی‌کننده می‌باشد. در این میان تعدادی لایه میانی یا پنهان وجود دارد. گره‌های ورودی به یک تعداد از گره‌ها در لایه پنهان متصل می‌شوند. گره‌ها در لایه پنهان می‌توانند به گره‌هایی در یک لایه پنهان دیگر یا به لایه خروجی متصل شوند. تعداد گره‌های ورودی به تعداد و نوع ویژگی‌های مجموعه داده‌ها و تعداد گره‌های خروجی به نوع عملیات دسته‌بندی بستگی دارد. لیکن تعداد گره‌های لایه پنهان یک مفهوم ابتکاری است و با سعی و خطا حاصل می‌شود. نمونه‌ای از یک شبکه عصبی با دو لایه پنهان در شکل (2) آمده است (هند[1]، 1998؛ بری و لینف، 2004؛ هن[2] و کمبر[3]، 2006).




**شکل (2) نمونه‌ای از ساختار یک شبکه عصبی با دو لایه پنهان**

در این تحقیق، برای دسته‌بندی داده‌ها با استفاده از شبکه‌های عصبی از الگوریتم پس انتشار خطا استفاده شده است و تابع فعال‌سازی همهٔ گره‌ها سیگموئید می‌باشد. برای ساخت شبکه‌ها از 14 ویژگی استفاده شده است. زیرا حذف ویژگی‌ها باعث کاهش دقت شبکه می‌شد. لازم به ذکر است که برای ساخت مدل‌های شبکه عصبی لازم است تمامی متغیرهای اسمی به مقادیر عددی تبدیل شوند. از طرفی لازم است

1. Hand
2. Han
3. Kamber

کلیه متغیرها از طریق نرمال‌سازی به بازهٔ (1+، 1-) بروند. با دادن وزن‌های اولیه تصادفی اولیه و بهم آمیختن رکوردها به صورت تصادفی توانستیم دقت مدل را افزایش دهیم. در صورتی که شبکه از جواب مورد نظر واگرا شد، با شروع مجدد آموزش مدل از ابتدا با نرخ یادگیری کمتر از شکست آموزش شبکه جلوگیری شد. زمان آموزش مدل 500 دور و نرخ یادگیری 0.3 در نظر گرفته شده است. اما برای جلوگیری از ساخت شبکه‌ای با نرخ خطای بالا از یک مجموعه ارزیابی به اندازهٔ 25% داده‌های آموزش استفاده شده است. آموزش مدل ادامه پیدا می‌کند تا اینکه یا نرخ خطای مجموعهٔ ارزیابی به طور مداوم بدتر شود و یا زمان آموزش پایان یابد. برای ساخت مدل شبکه عصبی بهینه از صفر لایه پنهان شروع کردیم و به تدریج تعداد لایه‌های پنهان را افزایش دادیم. در حقیقت با افزایش تعداد لایه‌ها غیرخطی بودن و پیچیدگی مدل افزایش می‌یابد و این افزایش پیچیدگی بیش از حد به قیمت بیش برازش و کاهش دقت مدل در مجموعه ارزیابی تمام می‌شود. در مورد داده‌های ما با افزایش لایه‌ها، دقت مدل مرتباً افزایش یافت اما از چهار لایه به بعد، دقت بر روی مجموعه ارزیابی و اعتبارسنجی 10-لایه شروع به کاهش کرد. تعداد لایه‌های پنهان در مدل نهایی 3 لایه می‌باشد.

نرخ دقت این مدل نسبت به درخت تصمیم، در مجموعه آموزش و اعتبارسنجی 10-لایه افزایش یافت، اما در مجموعه ارزیابی کاهش یافت. بیشترین دقت مدل شبکه عصبی ساخته شده در تشخیص صحیح دانشجویان مربوط به دسته A و پس از آن به ترتیب دسته‌های C و B می‌باشد. نرخ تشخیص صحیح دسته B در این مدل نسبت به درخت تصمیم افزایش یافت.

**جدول (2) ماتریس اغتشاش حاصل از آزمایش مدل ساخته شده بر روی مجموعه ارزیابی (تکنیک شبکه عصبی)**

| | | کلاس پیش‌بینی شده | | |
|---|---|---|---|---|
| | | A | B | C |
| کلاس واقعی | A | 64 | 10 | 2 |
| | B | 19 | 29 | 19 |
| | C | 1 | 16 | 60 |

**ج: استفاده از تکنیک نزدیکترین همسایگی برای دسته‌بندی دانشجویان**

این تکنیک نمونه‌ای از یادگیری مبتنی بر نمونه است. در این تکنیک داده‌های مجموعه آموزش در جایی ذخیره می‌شوند و دسته‌بندی برای یک داده جدید از طریق مقایسه داده جدید با داده‌های ذخیره شدهٔ مجموعه آموزش و پیدا کردن شبیه‌ترین k دادهٔ موجود در مجموعه آموزش به آن داده تعیین می‌شود. نکات مورد توجه در رابطه با ساخت یک مدل نزدیکترین همسایگی شامل این موارد می‌باشد:

- چه تعداد نقطه به عنوان همسایه در نظر بگیریم (مقدار k)؛
- چگونگی اندازه‌گیری شباهت بین نقاط؛
- چگونگی ترکیب اطلاعات چندین مشاهده (تابع ترکیب)؛
- آیا باید همه نقاط وزن یکسانی در تعیین دسته رکورد جدید داشته باشند و یا بعضی نقاط باید تأثیر بیشتری از دیگر نقاط داشته باشند؟ (هند، 1998؛ لاروز، 2006)

ویژگی‌های استفاده برای یادگیری این مدل همان 14 ویژگی تعیین شده در قبل می‌باشد. حذف هر یک از این ویژگی‌ها باعث کاهش دقت مدل در سه مجموعه می‌شد. برای استفاده از این تکنیک با نرمال‌سازی متغیرها و بردن آنها به بازهٔ (1+ ، 0) از تأثیر بیهوده ویژگی‌های با مقیاس بالا جلوگیری نمودیم. یکی از پارامترهای مهم در روش نزدیکترین همسایگی تعداد نقاط همسایه یا مقدار k می‌باشد. برای تعیین مقدار مناسب k از روش آزمون و خطا و اعتبارسنجی 10-لایه استفاده شد. مقادیر مختلفی برای k آزمایش شد تا به حداقل خطای دسته‌بندی در دو مجموعه آموزش و ارزیابی برسیم. بهترین حالت برای k برابر با 11 شد. مقادیر کمتر یا بیشتر برای k باعث کاهش دقت مدل ساخته شده می‌شد، زیرا مقادیر کوچکتر متأثر از داده‌های مغشوش می‌شد و مقادیر بزرگتر رفتارهای محلی را نادیده می‌گرفت. در مورد تابع فاصله نیز دو تابع فاصله اقلیدوسی و فاصله منهتن آزمایش شدند که تابع فاصله اقلیدوسی جواب بهتری داد. همچنین در مقایسهٔ نتایج مربوط به نرخ خطا و سطح زیر منحنی **آر-او-سی** [1] مربوط به دو مدل که درآنها فاصله وزن‌دار و غیروزن‌دار در نظر گرفته شده بود به این نتیجه رسیدیم که در حالت در نظر گرفتن معکوس فاصله‌ها به عنوان وزن‌ها در تابع ترکیب، نتایج بهتری در تعیین کلاس‌ها حاصل شد.

1. ROC (Receiver Operating Characteristics)

**جدول (3) ماتریس اغتشاش حاصل از آزمایش مدل ساخته شده بر روی مجموعه ارزیابی (تکینک نزدیکترین همسایگی)**

| | | کلاس پیش‌بینی شده | | |
|---|---|---|---|---|
| | | A | B | C |
| کلاس واقعی | A | 67 | 8 | 1 |
| | B | 24 | 24 | 19 |
| | C | 0 | 6 | 71 |

در مقایسه نتایج می‌توان گفت دقت این مدل نسبت به دو مدل قبل در سه حالت ارزیابی ارجحیت دارد. بیشترین دقت مدل نزدیکترین همسایگی در تشخیص دسته‌های C و پس از آن بترتیب در تشخیص دسته‌های A و B می‌باشد.

**د: استفاده از تکنیک رگرسیون لجستیک برای دسته‌بندی دانشجویان**

رگرسیون از قدیمی‌ترین و معروف‌ترین تکنیک‌هایی است که در داده‌کاوی بکار می‌رود. این روش یکی از روش‌های آماری داده‌کاوی می‌باشد. اساساً رگرسیون یک مجموعه اطلاعات و داده را در اختیار گرفته و یک فرمول ریاضی متناسب با آن داده‌ها به عنوان مدل رگرسیون ایجاد می‌کند و زمانیکه بخواهیم به کمک آن، نتیجه را برای داده جدیدی پیش‌بینی کنیم، کافی است داده جدید خود را به مدل رگرسیون داده و نتایج حاصل که همان پیش‌بینی‌های مورد نظر است را دریافت کنیم. رگرسیون خطی برای تخمین روابط میان یک متغیر هدف پیوسته و مجموعه‌ای از متغیرهای پیش‌بینی کننده بکار می‌رود که این متغیرها می‌توانند عددی، اسمی، باینری و ... باشند. در حالیکه رگرسیون لجستیک شیوه‌ای برای توصیف روابط میان یک متغیر اسمی و مجموعه‌ای از متغیرهای پیش‌بینی‌کننده است. این تکنیک زمانی که متغیر خروجی محدود به دو یا چند مقدار اسمی باشد، مفید است. رگرسیون لجستیک فرمولی را بدست می‌آورد که احتمال اتفاقی را به عنوان تابعی از متغیرهای مستقل پیش‌بینی می‌کند (گوایدیسی[1]، 2003؛ هند و مانیلا[2] و اسمیت[3]، 2001). اگر فرض کنیم که $C_1=1$ نشانه A بودن وضعیت تحصیلی دانشجو در ترم آینده و $C_2=0$ به نشانه A نبودن وضعیت تحصیلی دانشجو در ترم آینده باشد، رگرسیون لجستیک

1. Giudici
2. Mannila
3. Smyth

می‌تواند $P(C_1|X)$ را پیش‌بینی کند که نشان‌دهنده احتمال تعلق نمونه X به دسته A می‌باشد. برای استفاده از این تکنیک متغیرهای اسمی به مقادیر عددی تبدیل شده‌اند. در ساخت این مدل از 14 ویژگی تعیین شده استفاده شده است.

**جدول (4) ماتریس اغتشاش حاصل از آزمایش مدل ساخته شده بر روی مجموعه ارزیابی (تکنیک رگرسیون لجستیک)**

| | | کلاس پیش‌بینی شده | | |
|---|---|---|---|---|
| | | A | B | C |
| کلاس واقعی | A | 63 | 13 | 0 |
| | B | 20 | 28 | 19 |
| | C | 2 | 15 | 60 |

دقت مدل رگرسیون ساخته‌شده در اعتبارسنجی 10-لایه از مدل شبکه عصبی بیشتر ، اما از دو مدل دیگر کمتر شد. در آزمایش روی مجموعه ارزیابی نیز دقت این مدل از سه مدل مراحل قبل کمتر شد. بیشترین قدرت این مدل در تشخیص صحیح دسته A و کمترین دقت آن همانند سایر مدل‌ها در تشخیص دسته B است. این نشان می‌دهد که نظم خاصی در ویژگی‌های استخراجی از دانشجویان متعلق به کلاس B وجود ندارد، به عبارتی نقاط پرت در دانشجویان مربوط به این دسته زیاد می‌باشند.

**هـ: مقایسه و ارزیابی نتایج مدل‌های مختلف**

برای مقایسه مدل‌های مختلف و انتخاب بهترین آنها با توجه به اهداف داده‌کاوی می‌توان از روش‌ها و معیارهای مختلفی استفاده نمود. در این تحقیق، برای مقایسه نتایج دسته‌بندی از معیار نرخ خطای حاصل از روش اعتبارسنجی 10-لایه و خطای حاصل از آزمایش مدل‌ها بر روی مجموعهٔ ارزیابی استفاده شده است. اطلاعات مربوط به میزان دقت مدل‌های مختلف ساخته شده در 3 حالت مجموعه آموزش، اعتبارسنجی 10-لایه و آزمایش بر روی مجموعه ارزیابی در جدول (5) آمده است.

**جدول (5) مقایسه‌ای از میزان دقت مدل‌های مختلف ساخته شده**

| | نرخ دسته‌بندی صحیح در مجموعهٔ آموزشی | نرخ دسته‌بندی صحیح در ارزیابی 10-لایه | نرخ دسته‌بندی صحیح در مجموعهٔ ارزیابی |
|---|---|---|---|
| **درخت تصمیم C4.5** | % 68.4932 | % 61.9482 | % 70.4545 |
| **شبکه عصبی** | % 74.5814 | % 64.688 | % 69.5455 |
| **نزدیکترین همسایگی** | % 100 | % 71.0807 | % 73.6364 |
| **رگرسیون لجستیک** | % 72.2983 | % 64.9924 | % 68.6364 |

استفاده از مدل نزدیکترین همسایگی برای پیش‌بینی وضعیت تحصیلی دانشجویان، نتایج رضایت‌بخشی از خود نشان داد. متوسط نرخ خطا در اعتبارسنجی 10-لایه و آزمایش بر روی مجموعه ارزیابی از سایر روش‌ها کمتر می‌باشد. از آنجایی که هدف اصلی از ساخت مدل‌های دسته‌بندی، تشخیص صحیح دانشجویانی است که معدل آنها در ترم آینده در زمرهٔ C یا A قرار می‌گیرد، لازم است مقایسه مدل‌ها بر اساس دسته‌بندی صحیح این نمونه‌ها صورت بگیرد. بهترین مدل در دسته‌بندی صحیح دسته A و دسته C مدل نزدیکترین همسایگی می‌باشد و بهترین مدل در دسته‌بندی صحیح دسته B مدل شبکه عصبی می‌باشد.

## نتیجه‌گیری

در این مقاله سعی شد با استفاده از داده‌های مربوط به سوابق شخصی و تحصیلی دانشجویان و با کمک تکنیک‌های دسته‌بندی مختلف، مدل مناسبی برای پیش‌بینی وضعیت تحصیلی دانشجویان در هر نیمسال، برای نیمسال بعدی ارائه شود. در ساخت هر مدل، پارامترهای مختلفی وجود دارد که در نتایج نهایی مدل‌ها تأثیر بسزایی دارند. در ساخت مدل‌ها با توجه به نوع داده‌ها و اهداف ما از ساخت مدل‌ها، در مورد هر الگوریتم تغییرات زیادی بر روی پارامترهای مختلف صورت گرفته است تا در نهایت بهترین مدل استخراج گردیده است. تمامی روش‌های دسته‌بندی معمولاً دارای نقاط ضعف و قوتی می‌باشند که این امر سبب می‌شود که هر یک در شرایطی خاص خوب عمل کنند. با توجه به ماهیت داده‌های مربوط به سوابق تحصیلی دانشجویان بهترین مدل‌ها در دسته‌بندی صحیح دانشجویان مدل نزدیکترین همسایگی با پارامتر $k=11$ و تابع فاصله اقلیدوسی وزن‌دار و سپس شبکه‌های عصبی با 3 لایه پنهان شناخته شدند.

با استفاده از مدل نهایی ساخته شده به راحتی پیش‌بینی برای داده‌های جدید صورت می‌گیرد. به این ترتیب که با کمک داده‌های دموگرافیک و سوابق تحصیلی یک دانشجوی جدید (همان ویژگی‌هایی که برای ساخت مدل از آنها استفاده شد) و وارد کردن آنها به مدل نهایی می‌توان وضعیت تحصیلی این دانشجو را در نیمسال آینده از جهت تعلق وی به دسته A یا B یا C پیش‌بینی نمود و با توجه به نتیجه این پیش‌بینی اقدامات لازم را صورت داد و از صرف هزینه‌های بی‌مورد و هدر رفتن منابع و استعدادهای جامعه جلوگیری نمود و سطح علمی دانشگاه را ارتقاء داد. نتایج سودمندی از انجام این تحقیق متصور است. زمانیکه دانشجویان تحت ریسک (دسته C) در نیمسال آتی مشخص می‌شوند، می‌توان با انجام برنامه‌ریزی‌های تحصیلی مناسب و ارائه مشاوره‌های تحصیلی و مشوق‌های انگیزشی و انجام سایر اقدامات پیشگیرانه از رسیدن دانشجویان به وضعیت بحرانی و شکست و حذف آنها جلوگیری کرد، قبل از اینکه حتی دانشجویان خودشان از اینکه تحت ریسک هستند مطلع باشند و یا زمانیکه دانشجویان مستعد و قوی شناسایی می‌شوند، می‌توان منابع را به نحو مناسب‌تری تخصیص داد تا منافع بیشتری تأمین شود. می‌توان مدل ساخته شده را به عنوان یک ابزار پشتیبان تصمیم‌گیری در سیستم‌های آموزشی مورد بهره‌برداری قرار داد. سیستم‌های آموزش عالی از این طریق قادرند که اثر بخشی سیستم خود را حداکثر کنند، نرخ حذف دانشجویان را حداقل کنند، نرخ گذر دانشجویان را ارتقاء دهند، موفقیت دانشجویان را افزایش دهند و هزینه فرآیندهای سیستم را کاهش دهند. بدین ترتیب می‌توانند مزیت رقابتی خود را افزایش دهد و به استانداردهایی بالاتری در سطح آکادمیک برسد.

## منابع

حیدری، سمیه و یقینی، مسعود (1387). *داده‌کاوی جهت ارتقاء و بهبود فرآیندهای سیستم آموزش عالی*. دومین کنفرانس داده‌کاوی، دانشگاه امیرکبیر، ( IDMC) 2008.

Beikzadeh, M. R. & Delavari, N. (2004). *A New Analysis Model for Data Mining Processes in Higher Educational Systems*. Proceedings of M2USIC.

Berry, Michael J.A. & Linoff, Gordon S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management, 2nd ed.* Wiley Publishing, Inc.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). *Discovering Data Mining: From Concepts to Implementation.* Upper Saddle River, NJ: Prentice Hall.

Edelstein, Herb (1998). *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation.

Giudici, Paolo (2003). *Applied data mining: statistical methods for business and industry*. John Wiley & Sons Ltd.

Han, Jiawei & Kamber, Micheline (2006). *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining.* MIT Press, Cambridge, MA.

Hand, David J. (1998). *Data Mining: Statistics and More?*. The American Statistician, Vol. 52, No. 2, (112-118).

Larose, Daniel T. (2005). *Discovering knowledge in data: an introduction to data mining.* John Wiley & Sons, Inc.

Larose, Daniel T. (2006). *Data Mining Methods and Models*. John Wiley & Sons, Inc.

Luan, Jing (2001). *Data mining as driven by knowledge management in the higher education.* In Proceedings of SPSS Public Conference, UCSF.

Ranjan, J & Malik, K. (2007). *Effective educational process :a data-mining approach.* VINE: The journal of information and knowledge management systems, Vol. 37, No. 4, (502-515).

Romero, C. & Ventura, S. (2007). *Educational data mining: A survey from 1995 to 2005.* Expert Systems with Applications, No. 33, (135–146).

Romero, C., Ventura, S. & Garcia, E. (2008). *Data mining in course management systems: Moodle case study and tutorial.* Computers & Education, 51, (368–384).

Schonbrunn, K. & Hilbert, A. (2007). *Data Mining in Higher Education.* Advances in Data Analysis, Springer-Verlag, Heidelberg-Berlin, (489 – 496).

Yang, M. (2006). *Data Mining Techniques Applied to Texas Woman's University's Enrollment data – What Can the Data Tell us?*. MS Thesis, Texaz Woman's University.