مجله محاسبات نرم، جلد ۹، شماره ۲ (پاییز و زمستان ۱۳۹۹)، صفحه ۹۴ـ۱۱۳
https://dx.doi.org/10.22052/scj.2021.242837.0

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

دانشگاه کاشان
University of Kashan

پیش‌بینی وضعیت تحصیلی متقاضیان پذیرش‌شده دانشگاه، مبتنی بر داده‌های آموزشی و پذیرشی با استفاده از تکنیک‌های داده کاوی✧

آرش خسروی[1*]، استادیار، هادی عبدالمالکی[2]، کارشناس ارشد، مهری فیاضی[3]، کارشناس ارشد
[1] گروه مهندسی کامپیوتر، دانشکده مهندسی، مرکز آموزش عالی محلات، محلات، ایران.
[2] گروه مهندسی کامپیوتر، دانشکده برق، کامپیوتر و مهندسی پزشکی، موسسه آموزش عالی شهاب دانش، قم، ایران.
[3] گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه قم، قم، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت ۲۰ بهمن ماه ۱۳۹۹
پذیرش ۲۲ شهریور ماه ۱۴۰۰

کلمات کلیدی:
داده‌کاوی
داده‌های آموزشی و پذیرشی
وضعیت تحصیلی دانشجویان
طبقه‌بندی

چکیده

داده کاوی آموزشی در چند سال اخیر بسیار مورد توجه قرار گرفته است. مراکز و مؤسسات آموزشی دارای حجم زیادی از اطلاعات دانشجویان هستند که می‌تواند به عنوان ابزاری برای ارتقا سطح کیفی آموزش مورد استفاده قرار گیرد. دانش استخراج شده به مؤسسات کمک می‌کند تا روش های تدریس، فرآیند یادگیری و تصمیم‌گیری‌های خود را بهبود بخشند. هدف این پژوهش پیش‌بینی وضعیت تحصیلی دانشجویانی است که قرار است از مقطع کاردانی به مقطع کارشناسی ادامه تحصیل دهند. با توجه به اینکه وزارت علوم قصد دارد آزمون ورودی (کنکور) را حذف کند؛ دانشگاه‌ها با این مشکل مواجه خواهند شد که دانشجویان را براساس چه معیارهایی انتخاب کنند. در این پژوهش سعی بر آن است تا با استفاده از تکنیک‌های داده کاویِ درخت تصمیم، نیو بیز، شبکه عصبی، ماشین بردار پشتیبان، جنگل تصادفی، Bagging و Boosting اطلاعات آموزشی دانشجویان تازه وارد تحلیل شود و با مقایسه آن‌ها با اطلاعات دانشجویان فارغ‌التحصیل، انصرافی و اخراجی مقطع کارشناسی، روشی برای انتخاب بهتر دانشجویان ارائه کند. با توجه به نتایج این تحقیق، جنگل تصادفی با ۹۲/۲۸٪ بیشترین دقت و نیو بیز با ۶۱/۰۹٪ کمترین دقت پیش‌بینی را دارند.

۱. مقدمه

با پیشرفت سریع فناوری اطلاعات و همچنین افزایش تجهیزات و ظرفیت‌های جمع‌آوری و ذخیره‌سازی داده‌ها در حوزه‌های مختلف، بشر شاهد یک رشد بسیار زیاد در تولید «داده» است. تحلیل این حجم از داده‌ها به‌صورتی که قابل درک و کاربردی باشد یک مسئله چالش برانگیز است. امروزه با توجه به این حجم از اطلاعات و ارزشمند شدن آن به عنوان یک سرمایه بالقوه، کشف الگوهای پنهان این اطلاعات می‌تواند تصمیم‌گیری‌های مختلف دارندگان آن‌ها را متحول و بهبود ببخشد. ژیاوی هان[1]، دانشمند داده و نویسنده کتاب «داده‌کاوی، مفاهیم و روش‌ها»[2] در این رابطه می‌گوید: «... در نتیجه،

✧ نوع مقاله: پژوهشی
* نویسنده مسئول
پست(های) الکترونیک: khosravi.280@gmail.com (خسروی)
h.abdulmaleki@gmail.com (عبدالمالکی)
mehri.feysi@gmail.com (فیاضی)

[1] Jiawei Han
[2] Data Mining: Concepts and Techniques

نحوه ارجاع به مقاله: خسروی، آرش، عبدالمالکی، هادی، فیاضی، مهری، «پیش‌بینی وضعیت تحصیلی متقاضیان پذیرش‌شده دانشگاه، مبتنی بر داده‌های آموزشی و پذیرشی با استفاده از تکنیک‌های داده کاوی»، مجله محاسبات نرم، جلد ۹، شماره ۲، ص۹۴ـ۱۱۳، پاییز و زمستان ۱۳۹۹.

داده‌های گردآوری شده در مخازن داده به گورهای داده مبدل شده‌اند، ...، شکاف در حال افزایش میان داده و اطلاعات، توسعه سیستماتیک ابزارهای داده‌کاوی را می‌طلبد که می‌توانند گورهای داده را به شمش‌هایی از طلا مبدل کنند» [۱].

با توجه به این‌که نظام آموزشی همواره با داده‌ها و اطلاعات بسیار زیادی در مورد مراکز آموزشی، دانش‌آموزان و دانشجویان، اساتید، پرسنل و... روبرو است و نیز توجه به عملکرد آموزشی آن‌ها، جز لاینفک این نظام است، این داده‌ها دارای اطلاعات با ارزشی هستند؛ ازاین‌رو استفاده از داده کاوی در نظام آموزشی کاربرد بسیاری می‌تواند داشته باشد. در نظام‌های آموزشی، داده‌کاوی را می‌توان به سه گروه تقسیم کرد: ۱) داده‌کاوی اطلاعات دانشجویان ۲) داده‌کاوی اطلاعات آموزش‌دهندگان و ۳) داده‌کاوی اطلاعات مسئولان و مدیران آموزشی [۲].

با استفاده از کشف الگوها و استخراج دانش از پایگاه داده دانشجویان فارغ‌التحصیل، اطلاعات ارزشمندی در مورد کیفیت تحصیلی دانشجویان و عوامل مؤثر بر موفقیت یا عدم موفقیت آن‌ها به‌دست می‌آید که از این دانش به‌دست آمده، می‌توان در انتخاب دانشجویان، پیش از ورود به مقطع بالاتر دقت بیشتری کرد. این کار، رویکردی اثرگذار در برنامه‌ریزی آموزشی سازمان‌ها داشته و به منظور بهبود کارایی و عملکرد دانشجویان به کار می‌آید. داده کاوی آموزشی، نیازمند تشریح سلسله مراتب چند سطحی از داده‌های آموزشی و وابستگی ضمنی میان آنهاست. داده کاوی آموزشی به عنوان یک حوزه جدید پژوهشی، در سال‌های اخیر توجه اغلب محققان حوزه‌های مختلف آموزشی و پرورشی را به خود جلب نموده است.

وب سایت انجمن داده‌کاوی آموزشی، داده‌کاوی آموزشی (EDM)[1] را این چنین تعریف می‌کند: «داده کاوی آموزشی، یک شاخه نوظهور از علم داده کاوی است که به توسعه روش‌هایی برای درک بهتر اطلاعات محیط‌های آموزشی می‌پردازد. این حوزه از علم داده کاوی، برای کشف دانش نهفته درفرآیندهای آموزشی و تحصیلی، استفاده می‌شود» [۳].

۱،۱. مسئله تحقیق

تحقیقات انجام شده در گذشته نشانگر این بوده است که درصد قابل توجهی از افراد مشغول به تحصیل که به طور جدی دچار افت تحصیلی شده‌اند متعلق به خانواده‌هایی از سطح اقتصادی و اجتماعی پایین بوده که انگیزه تحصیلی بسیار پایینی داشته‌اند [۴]. موفقیت تحصیلی دانشجویان یکی از مهم‌ترین امور در حوزه آموزش است. استفاده از روابط پنهان موجود بین داده‌ها برای پیش‌بینی فاکتورهای موفقیت دانشجویان، امری ضروری است. همواره مسئله افت تحصیلی برای اولیاء و مسئولان آموزش مبهم و پیچیده بوده است. پاسخگویی به این سوال که چه عواملی باعث بروز و ظهور این مشکل اجتماعی می‌شوند اهمیت بسیار دارد [۵]. کیفیت تعامل دانشجو و استاد، مشارکت دانشجو و فرآیندهای تعامل، شاخص‌های مهمی هستند که بر عملکرد کلی دانشجو و نرخ انصراف از تحصیل تاثیر می‌گذارند. به منظور آموزش بهتر و برای این‌که نظام آموزشی بتواند رویه‌ها و شکاف‌های خود را بهبود ببخشد، لازم است ابزارهای گردآوری داده‌های بازخوردی مرتبط را داشته باشد [۶] و ارزیابی نظام‌مندی را برای شناسایی ضعف‌ها در ارائه آموزش به عموم دانشجویان به کار بگیرد. داده کاوی آموزشی یکی از روش‌هایی است که می‌تواند برای تشخیص الگو مورد استفاده قرار گیرد [۷].

مشکل اصلی در آموزش عالی، کاهش میزان موفقیت دانشجویان است. برای افزایش نرخ موفقیت دانشجویان، روش پیش‌بینی اولیه به مدیریت کمک خواهد کرد تا به دانشجویان ضعیف در زمان مناسب کمک کند. برای کشف الگوهای جدید از داده‌های مختلف، روش داده کاوی به‌طور گسترده‌ای مورد استفاده قرار می‌گیرد [۸].

۱،۲. فواید تحقیق

وضعیت اقتصادی، اجتماعی، موقعیت خانواده، شاغل بودن یا نبودن، سن، جنسیت، دانشکده، سهمیه‌های بومی و ارائه امکانات رفاهی و... حکایت از وجود یک رابطه مثبت در وضعیت

[1] Educational data mining

تحصیلی دارد. بیشتر تلاش‌ها برای شناسایی عوامل مؤثر بر پیشرفت تحصیلی دانشجویان انجام گرفته است و این عوامل می‌تواند برای دانشجویان، با توجه به تفاوت‌های سنی و دیگر شرایط آن‌ها متفاوت باشد. یکی از دلایل مهم بررسی این موضوع، اثرات مخرب ناشی از تداوم مشکلات تحصیلی بر نیروی کار کشور است. در صورت ادامه این شرایط خسارات متعددی به اقتصاد کشور وارد خواهد شد و سطح علمی دانشجویان فارغ‌التحصیل از کیفیت مناسبی برای برازش توسط کارفرمایان برخوردار نخواهد بود. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزش عالی کمک شایانی کند. استفاده از تکنیک‌های پیشرفته داده‌کاوی مانند خوشه‌بندی [۹]، طبقه‌بندی و... می‌تواند در طبقه‌بندی دانشگاه‌ها، یافتن الگوهای خاص و با ارزش در مورد دانشجویان موفق، یافتن یک برنامه یا روش موفق تدریس، یافتن نقاط بحرانی در مدیریت مالی دانشگاه‌ها و موارد دیگر کاربرد داشته باشد.شناسایی فاکتورهای مؤثر در وضعیت تحصیلی دانشجویان و بذل توجه به آن‌ها، می‌تواند گامی به سوی توسعه پایدار باشد. با نگاهی به روند تحولات جاری نظام آموزش عالی، می‌توان دریافت که آموزش عالی باید ضمن توجه به بحران افزایش کمی و تنگناهای مالی، به حفظ و بهبود ارتقای کیفیت نیز بپردازد. دانشگاه‌ها و مراکز آموزشی با استفاده از فرایند آموزشِ یکپارچه و داده‌محور خود می‌توانند عملکرد دانشجویان را پیشاپیش برآورد کنند و استراتژی‌های مداخله‌گر برای آموزش بهتر به آن‌ها عرضه نمایند. اساتید با استفاده از تکنیک‌های داده‌کاوی می‌توانند میزان پیشرفت دانشجویان را با دقت بالا پیش‌بینی کرده و متوجه شوند که کدام دانشجو به توجه بیشتری نیاز دارد. پیدا کردن الگوها و دانش نهفته درسیستم‌های آموزشی می‌تواند به تصمیم‌گیرندگان عرصه آموزش در جهت ارتقاء و بهبود فرآیندهای آموزشی نظیر برنامه‌ریزی، ثبت نام، ارزیابی و مشاوره کمک شایانی نماید.

۱،۳. سؤالات تحقیق

در ادامه سوالات مطرح در این پژوهش بیان شده است.

۱. چگونه می‌توان وضعیت تحصیلی دانشجویان ورودی را پیش‌بینی کرد؟
۲. چه ابعاد و ویژگی‌هایی در این پیش‌بینی می‌تواند مؤثر باشد؟
۳. چگونه ویژگی‌های خاصی را کم کنیم تا الگوریتم‌ها کاراتر کار کنند؟
۴. چه تکنیکی می‌تواند دقت بالاتری را در طبقه‌بندی به ما بدهد؟

۱،۴. نوآوری‌ها

در مورد نوآوری‌های مهم این مقاله می‌توان به‌طور خلاصه به موارد زیر اشاره کرد:

۱. استفاده کاربردی از داده کاوی برای پیش‌بینی از مقطع کاردانی به کارشناسی ناپیوسته و به همین ترتیب الگوی پیش‌بینی برای مقاطع بالاتر.
۲. جمع‌آوری داده‌های مناسب و واقعی از دانشگاه که معمولاً دسترسی به آنها آسان نیست.
۳. عملیات پاک‌سازی و پیش پردازش مناسب برای آماده‌سازی داده‌ها.
۴. کار روی تعداد بالای رکوردهای داده و فراهم کردن داده تمیز با تعداد رکورد بالا در مقایسه با تحقیقات مشابه.
۵. تحلیل مفهومی و کیفی داده‌ها و همچنین استفاده از معیارهای داده‌کاوی جهت کاهش داده‌ها.
۶. متوازن و متعادل کردن داده‌ها.
۷. استفاده از روش مقایسه‌ای و رسیدن به یک مدل داده کاوی بهینه براساس فاکتورهای ارزیابی جهت پیش‌بینی پیشرفت تحصیلی دانشجویان.

در ادامه سازماندهی مقاله به این صورت خواهد بود: در بخش دوم، مروری بر تحقیقات انجام شده داخلی و خارجی در زمینه‌ی داده کاوی آموزشی آمده است. در بخش ۳ به منظور آشنایی با روش‌های داده کاوی مقایسه‌ایی روی آنها صورت گرفته است. بخش ۴ چهارچوب پژوهش مورد بررسی قرار گرفته است. بخش ۵ به شرح روش پیشنهادی این پژوهش و

مراحل آن و نتیجه هر مرحله پرداخته است. در نهایت نتیجه‌گیری مقاله در بخش ۶ فراهم آمده است.

۲. مروری بر کارهای گذشته

روش‌های داده کاوی آموزشی شامل پیشینه‌ایی در زمینه‌های متعدد هستند از جمله داده کاوی، یادگیری ماشین، روان‌سنجی و سایر زمینه‌های آماری، بصری‌سازی اطلاعات و مدل‌سازی محاسباتی. تحقیقات اخیر در داده کاوی آموزشی، اغلب دانشجو را در سیستم آموزشی مورد بررسی قرار داده است. به عنوان مثال در یک دسته از تحقیقات، ضعف دانشجو، بی‌حوصلی در تحصیل، ناامیدی برای پیش‌بینی عدم موفقیت و عدم حفظ دانشجو در دانشگاه‌ها بررسی شده است [۱۰–۱۲].

هدف مقاله [۱۳] پیش‌بینی عوامل مؤثر در موفقیت تحصیلی دانشجویان دانشگاه پیام‌نور با کمک تکنیک‌های داده‌کاوی می‌باشد. جامعه آماری شامل ۶۰،۰۰۰ رکورد از اطلاعات دانشجویان فارغ التحصیل، در حال تحصیل، انصراف داده و اخراجی تعدادی از مراکز پیام‌نور استان کرمانشاه است. این پژوهش روی داده‌های استخراج شده از سیستم جامع گلستان انجام و پس از گردآوری داده‌ها، از روش simple k-means به‌منظور خوشه‌بندی و تعریف دانشجویان ضعیف، متوسط و موفق استفاده و سپس در نرم افزار Weka طبقه‌بندهای درخت تصمیم‌گیری، نیو بیز، شبکه عصبی و ماشین بردار پشتیبان بر آن اعمال شده است. یافته‌های این پژوهش نشان می‌دهد که بهترین پاسخ‌ها، با بیش از ۹۳/۲ درصد صحت مربوط به درخت تصمیم‌گیری است.

در مقاله [۱۴] گفته شده است که هر نظام آموزشی می‌تواند با به دست آوردن مدل مناسب، راهکارهای مناسبی را جهت افزایش یادگیری و کیفیت نظامش ارائه کند و به این ترتیب در ارتقاء علمی نظام آموزشی خود گام بردارد. در این مقاله، قابلیت‌های بالقوه الگوریتم‌ها و روش‌های داده‌کاوی جهت بهبود کیفیت خدمات آموزشی مدارس و دانشگاه‌ها بررسی شده است.

در مقاله [۱۵] با استفاده از تکنیک‌های داده‌کاوی جهت تحلیل‌های آماری بر روی اطلاعات دانشجویان رشته مهندسی عمران دانشگاه شاهرود، قوانین حاکم و عوامل مؤثر بر موفقیت و یا عدم موفقیت آن‌ها به‌دست‌آمده است. در نهایت، بررسی پارامترهای مؤثر بر موفقیت دانشجویان، مدلی را ارائه می‌کند که به یافتن الگوهای خاص و با ارزش در مورد دانشجویان موفق، یافتن یک برنامه یا روش موفق تدریس، یافتن نقاط ضعف دانشجو و عدم ایجاد شرایط بحرانی کمک می کند و می تواند نقش مهمی را در عرصه پیشرفت آموزش عالی، برنامه‌ریزی‌های درسی و همچنین برنامه‌ریزی‌های دانشجو در جهت بهبود نمرات و فارغ‌التحصیلی بهتر ایفا کند.

مقاله [۱] به بررسی اطلاعات آموزشی دانش‌آموختگان رشته دندانپزشکی بین سال‌های ۱۳۷۵ تا ۱۳۹۰ پرداختند و مدلی جهت تشخیص دانشجویان ضعیف و قوی ارائه داده و عوامل مؤثر بر نمرات و رفتار آموزشی آنان را بیان کردند.

نویسندگان مقاله [۱۶] با نمونه‌برداری ۱۸۰ دانشجو از رشته‌ها و مقاطع مختلف دانشگاه علوم پزشکی تهران، به بررسی عوامل مؤثر بر پیشرفت تحصیلی دانشجویان پرداختند. آن‌ها دریافتند که متغیرهای جنسیت، گروه سنی، مقطع تحصیلی، رشته تحصیلی، علاقه به رشته تحصیلی و سبک یادگیری تفاوت معناداری در پیشرفت تحصیلی شرکت‌کنندگان ایجاد کرده است.

در مقاله پژوهشگران [۱۷] الگوریتم داده‌کاوی را روی بانک اطلاعاتی دانشگاه پیام‌نور استان قم به‌کاربرده و از طریق دسته‌بندی انتخاب واحد دانشجویان با کمک الگوریتم APRIORI، به استخراج الگوهای پنهان پرداخته و از مدل‌های ایجادشده به‌عنوان یک ابزار پشتیبان تصمیم‌گیری در نظام‌های آموزشی به‌منظور ارتقاء سطح علمی دانشگاه‌ها استفاده کردند.

مقاله [۱۸] به مطالعه عوامل مؤثر بر پیشرفت تحصیلی دانشجویان دانشگاه علوم پزشکی تهران (پردیس همت) در سال تحصیلی ۸۸–۸۹ پرداختند. نویسندگان این مقاله در این تحقیق دریافتند که پذیرش تعداد زیاد دانشجو در مقاطع پزشکی عمومی و تحصیلات تکمیلی از جمله دلایل عدم موفقیت این دانشجویان است. کارهای مختلف انجام شده در زمینه داده کاوی آموزشی بین سال های ۲۰۱۰ تا ۲۰۲۰ براساس روش کار و اهداف در جدول (۱) بررسی شده است.

جدول (۱): بررسی مقالات مختلف منتشر شده در داده‌کاوی آموزشی

مرجع	سال	روش کار	اهداف
[۱۹]	۲۰۱۰	مروری	لیست کردن کارهای که در حوزه آموزش با استفاده از داده‌کاوی انجام شده است.
[۲۰]	۲۰۲۰	طبقه‌بندی – رگرسیون لجستیک و درخت تصمیمی	پیش‌بینی کارایی آموزش ترکیبی (حضوری و غیرحضوری) دانشجویان
[۲۱]	۲۰۲۰	طبقه بندی (ترکیبی) K-nearest neighbor (k-NN), random forest (RF), Support Vector machine (SVM), Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Naïve Bayes (NB).	پیش‌بینی کارایی تحصیلی دانشجویان
[۲۲]	۲۰۱۹	طبقه‌بندی، درخت تصمیم	پیش‌بینی کارایی دانش آموزان
[۲۳]	۲۰۱۹	طبقه‌بندی و خوشه‌بندی SVM, Naïve Bayes, Decision tree and Neural Network classifiers	پیش‌بینی کارایی آکادمیک دانشجویان
[۲۴]	۲۰۱۹	طبقه‌بندی و خوشه‌بندی k-means clustering, ZeroR, OneR, ID3, J48, random forest, decision stump, JRip, PART, NBTree, and Prism	پیش‌بینی تعلل و عقب افتادگی دانش آموزان
[۲۵]	۲۰۱۹	قوانین انجمنی و طبقه بندی – ID 3 Decision Tree	پیش‌بینی میزان اشتغال ونوع کار دانشجویان
[۲۶]	۲۰۱۹	طبقه‌بندی و خوشه‌بندی Naive Bayes, Decision Tree,and Artificial Neural Network	پیش‌بینی کارایی تحصیلی دانشجویان
[۲۷]	۲۰۱۹	تحلیل رگرسیون چندگانه Neural Network – Multilayer perceptron algorithms and radial basis function	پیش‌بینی ترک تحصیل در دانشگاه ها
[۲۸]	۲۰۱۹	طبقه‌بندی (ترکیبی) – شبکه عصبی و درخت تصمیم	پیش‌بینی کارایی تحصیلی دانشجویان
[۲۹]	۲۰۱۹	طبقه‌بندی	در کنار صفاتی مانند "نمره" و "غیبت" صفات دیگری مانند "محله"، "مدرسه" و "سن" جزو شاخص‌های موفقیت یا شکست تحصیلی هستند.
[۳۰]	۲۰۱۸	خوشه‌بندی K-Means	تحلیل گرایش تحصیلی دانشجویان
[۳۱]	۲۰۱۸	طبقه‌بندی – Multi-Layer Perceptron (MLP)	پیش‌بینی فارغ التحصیلی دانشجویان
[۳۲]	۲۰۱۸	طبقه‌بندی و خوشه‌بندی J48, Naïve Bayes, Random Forest, and Random Tree, Multiple Linear Regression, binomial logistic regression, Recursive Partitioning and Regression Tree (rpart), conditional inference tree (ctree) and Neural Network (nnet) algorithms	پیش‌بینی جایگاه تحصیلی دانشجویان
[۳۳]	۲۰۱۷	تحلیل رگرسیون چندگانه – Recurrent Neural Network (RNN)	پیش‌بینی نمرات نهایی و پیشرفت تحصیلی دانش آموزان
[۳۴]	۲۰۱۷	طبقه‌بندی (ترکیبی) NaiveBayes, Bayesian Network, ID3, J48 and Neural Network	پیش‌بینی پیشرفت دانشجویان
[۳۵]	۲۰۱۷	درخت تصمیم، ماشین بردار پشتیبان، شبکه عصبی و نیو بیز	دانشجویانی که احتمالاً در مراحل اولیه تحصیل شکست می‌خورد را جهت کاهش نرخ شکست، شناسایی کرده است.

ادامه جدول (۱): بررسی مقالات مختلف منتشر شده در داده‌کاوی آموزشی

مرجع	سال	روش کار	اهداف
[۳۶]	۲۰۱۶	طبقه‌بندی گروهی[1]	رابطه قوی بین رفتارهای یادگیرندگان و موفقیت تحصیلی آن‌ها اعلام کرده است.
[۳۷]	۲۰۱۴	نیوبیز	رشد تحصیلی و پیشرفت دانشجویان به‌صورت گرافیکی نمایش داده شده است.
[۳۸]	۲۰۱۴	تکنیک خوشه‌بندی	داده‌های متخصصین در حین آموزش در یک شرکت مشاوره، مورد تجزیه‌وتحلیل قرار گرفته است.
[۳۹]	۲۰۱۴	الگوریتم تحلیل انجمنی	مجموعه‌ای از دانشجویان ضعیف بر اساس نمرات فارغ‌التحصیلی مشخص شده است.
[۴۰]	۲۰۱۲	آزمون مربع کای[2]، آزمون One Rule، آزمون InfoGain and Ratio، نیوبیز و درخت تصمیم	یافتن مدل پیش‌بینی کاربرپسند برای عملکرد تحصیلی که برای کاربران حرفه‌ای یا غیرحرفه‌ای مناسب است.
[۴۱]	۲۰۱۲	روش طبقه‌بندی بیزین[3]	دستاوردهای تحصیلی ضعیف در آموزش عالی را تحلیل کرده است.
[۴۲]	۲۰۱۱	مدل‌سازی معادلات ساختاری، ماتریس همبستگی	در ۱۰ هفته اول ثبت‌نام، دو متغیر رویکرد آموزشی و مهارت افزایی مورد بررسی قرار گرفته است.
[۴۳]	۲۰۱۱	تحلیل قوانین انجمنی	راهنمایی‌هایی برای مدیریت علمی و ارزیابی جامع دانشجویان ارائه شده است.
[۴۴]	۲۰۱۱	درخت تصمیم	شانس موفقیت دانشجویان با پیاده‌سازی پروفایل آن‌ها با سیستم Storyboard تخمین زده شده است.
[۴۵]	۲۰۱۱	طبقه‌بندی و خوشه‌بندی	کارایی عملکرد دانشجویان سال آخر تجزیه‌وتحلیل شده است.
[۴۶]	۲۰۱۰	خوشه‌بندی K-Means	رفتار یادگیری دانشجویان جهت بررسی کارایی و پیش‌بینی دانشجویان ضعیف تحلیل شده است.
[۴۷]	۲۰۱۰	مدل طبقه‌بندی درختی	ویژگی‌های مربوط به ثبت‌نام را جهت شناسایی زودهنگام دانشجویان موفق بررسی کرده است.
[۴۸]	۲۰۱۰	خوشه‌بندی، قوانین انجمنی و درخت تصمیم	موفقیت تحصیلی و ترک تحصیل دانشجویان را تحلیل کرده است.
[۴۹]	۲۰۱۰	نظریه مجموعه‌های راف[4]	نمرات دانشجویان تجزیه‌وتحلیل شده است.
[۵۰]	۲۰۱۰	شبکه عصبی، نظریه مجموعه‌های راف	پیش‌بینی دانشجویانی که ترک تحصیل می‌کنند.
[۵۱]	۲۰۱۰	درخت تصمیم	تحلیل Storyboard موفقیت‌آمیز (سیستم یادگیری الکترونیکی) و تحلیل مسیرهای موفقیت دانشجویان.
[۵۲]	۲۰۱۰	درخت تصمیم	نتایج داده‌کاوی می‌تواند تدریس در دبیرستان‌ها را علمی‌تر کند و کیفیت آموزش را بهبود بخشد.
[۵۳]	۲۰۱۰	درخت تصمیم	ایجاد مدل پیش‌بینی برای نمرات دانشجویان به‌منظور شناسایی عادات منفی یادگیری یا رفتارهای منفی دانشجویان.

[1] Ensemble
[2] Chi-Square
[3] Bayesian classification method
[4] Rough set theory

۳. مقایسه روش‌های داده‌کاوی

در این بخش مقایسه‌ای برای روش‌های داده کاوی رایج ارائه خواهد شد. برای این امر جدول (۲) ارائه شده است. این جدول روش‌های داده‌کاوی مختلف اعم از قواعد انجمنی یادگیر، طبقه‌بندی، خوشه‌ای، درخت تصمیم، شبکه عصبی، ماشین بردار پشتیبانی و نیوبیز را از نظر امکانات، مزایا و محدودیت‌ها مورد تحلیل قرار داده و یک دید مقایسه‌ای مناسب در این زمینه ارائه می‌کند.

جدول (۲): امکانات، مزایا و محدویت‌های روش‌های داده‌کاوی [۵۴]

روش	امکانات	مزایای	محدودیت‌ها
قواعد انجمنی یادگیری	الگوهای مشابه از داده را پیدا کرده و قوانینی را تولید می‌کند. رابطه انجمنی را از داده‌ها ایجاد کنید. به فرآیند تصمیم‌گیری کمک می‌کند استفاده از حداقل مقدار پشتیبانی و حداقل مقدار پشتیبانی اطمینان	به یافتن الگوهای ترتیبی کمک می‌کند از روش‌های کنترل اکتساب، یکپارچه‌سازی و تمامیت استفاده می‌کند.	الگوهای منطقی با متغیرهای وابسته را نشان نمی‌دهد. مقدار پشتیبانی و اعتماد پیش‌نیاز است.
طبقه‌بندی	تکنیکی که قوانینی را پیدا می‌کند که داده‌ها را به گروه‌های مختلف تقسیم می‌کند. مشاهده مشابه را از مجموعه داده بزرگ شناسایی کرده و آن‌ها را در یک مجموعه قرار دهید.	دارای بهره‌وری خوب است. داده‌های نویزدار را کنترل می‌کند. بسیار مناسب برای کلاس‌های چند وجهی به زمان‌های محاسباتی کوتاه نیاز دارد.	حساسیت به ساختار محلی داده‌ها نیازمند حافظه زیاد است.
خوشه‌ای	روش جمع‌آوری اشیاء در یک گروه، با پیداکردن ویژگی‌های مشابه. اهداف این روش به شرح زیر است: برای کشف گروه بندی‌های طبیعی. تولید فرضیه از داده‌ها برای یافتن سازماندهی مطمئن داده‌ها.	دیدگاه سطح بالایی را برای کاربر آنچه در پایگاه اطلاعاتی در حال انجام است فراهم می‌کند. تکنیک بسیار کارآمد	زمانی که ادغام یا انشعاب انجام شود، نمی‌توان آن را لغو یا اصلاح کرد.
درخت تصمیم	مدلی که برای پیش‌بینی استفاده می‌شود و می‌تواند به‌عنوان یک درخت در نظر گرفته شود. نمودار جریان مانند ساختار. هر شاخه از درخت نشان دهنده شرایط است و برگها در صورت برآورده شدن شرط نتیجه را نشان می‌دهند. درخت تصمیم‌گیری، داده‌ها را طبق شرایط تقسیم می‌کند و در تصمیم‌گیری کمک می‌کند. درخت تصمیم‌گیری می‌تواند به‌عنوان ابزار پشتیبانی تصمیم‌گیری عمل کند.	درک و تفسیر آن ساده است. قادر به کنترل داده‌های عددی و دسته‌ای است. قدرتمند است. برای داده‌های بزرگ عملکرد خوبی دارد.	گاهی اوقات محاسبات پیچیده است. گاهی اوقات از بیش برازش یک مسئله رنج می‌برد.
شبکه عصبی	برای تشخیص الگوهای مختلف و پیش‌بینی الگوها استفاده می‌شود. خروجی عددی تولید می‌کند. بطور گسترده در تشخیص تقلب، پیش‌بینی واکنش مشتری، درک تصویر و بسیاری دیگر مفید است.	توانایی یادگیری خوب. دارای سرعت خوب.	فقط داده‌های عددی را کنترل می‌کند. بنابراین باید هر داده را به شکل عددی ترجمه کنیم. با توجه به یادگیری، ممکن است مشکل بهینه محلی رخ دهد.

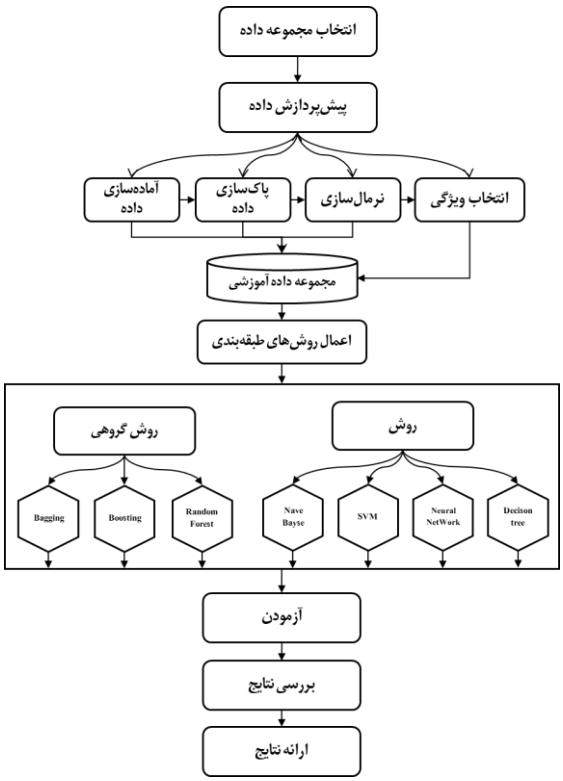
ادامه جدول (۲): امکانات، مزایا و محدویت‌های روش‌های داده‌کاوی [۵۴]

روش	امکانات	مزایای	محدودیت‌ها
ماشین بردار پشتیبانی	تکنیک یادگیری نظارت شده به کاهش ریسک و به حداقل رساندن خطای طبقه‌بندی کمک می‌کند. بخشی از طبقه‌بندی کننده‌های خطی است و می‌تواند تعمیم‌دهنده تکنیک ادراک باشد.	طبقه‌بندی کننده‌های بسیار دقیق تولید می‌کند کاهش بیش برازش و نویز کم‌تر. حافظه فشرده در شناخت حروف دست نوشته مفید است.	سرعت یک قانون برای آموزش و آزمایش است. احتمال حضور داده‌های گسسته. پیچیدگی الگوریتمی بالا.
نیوبیز	تکنیک یادگیری نظارت شده پیش‌بینی بلادرنگ در تشخیص تقلب در کارت اعتباری، پیش‌بینی بیماری‌های قلبی، پیش‌بینی باران و... استفاده می‌شود. فیلتر کردن اسپم، تحلیل احساسات، پشتیبانی از تشخیص نفوذ شبکه سیستم‌های پیشنهاد دهنده	پیاده‌سازی ساده آن. از راندمان محاسباتی و عملکرد طبقه‌بندی بسیار خوبی برخوردار است. مقدار داده‌های مورد نیاز برای این الگوریتم کمتر است، در حالی که سایر الگوریتم‌های پیشرفته برای فرایند یادگیری به داده‌های بیشتری نیاز دارند. نتایج دقیقی برای بسیاری از مشکلات طبقه‌بندی و پیش‌بینی ارایه می‌دهد.	اگر اتفاق خاصی در برچسب کلاس و مقدار ویژگی رخ ندهد، در این الگوریتم احتمال وجود آن صفر در نظر گرفته می‌شود. این مشکل با عنوان "مسئله صفر" خوانده می‌شود. نیاز به پیش‌بینی مستقل دارد. در زندگی واقعی، پیش‌بینی‌ها وابسته هستند، این مانع عملکرد طبقه‌بندی‌کننده می‌شود.

۴. روش پیشنهادی

در شکل (۱) چارچوب این پژوهش نشان داده شده است. در ابتدا مجموعه داده‌ی مورد نیاز مشخص و انتخاب می‌شود و پیش‌پردازش این داده‌ها که شامل مراحل آماده‌سازی، پاک‌سازی، نرمال‌سازی و انتخاب ویژگی است انجام می‌شود تا مجموعه داده‌ی نهایی برای اعمال روش‌های طبقه‌بندی آماده شود. پس از آن هر کدام از طبقه‌بندی‌کننده‌های درخت تصمیم، نیوبیز، ماشین بردار پشتیبان[1]، شبکه عصبی و روش‌های گروهی جنگل تصادفی[2]، Bagging و Boosting به صورت جداگانه برای روی مجموعه داده اعمال خواهد شد و نتایج آنها با یکدیگر مقایسه می‌شود [۵۵].

این پژوهش از نوع کاربردی است و هدف آن بررسی وضعیت تحصیلی دانشجویانی است که مقطع کاردانی را پشت‌سر گذاشته و در مقطع کارشناسی ناپیوسته پذیرفته شده‌اند.



شکل (۱): چارچوب پژوهش

[1] Support Vector Machine (SVM)
[2] Random Forest

در این پژوهش داده‌های آموزشی و جمعیت‌شناختی دانشجویان دانشگاه شهاب دانش بررسی و تحلیل شده و با استفاده از تکنیک‌های داده‌کاوی که در حوزه داده‌کاوی آموزشی پرکاربرد هستند به تشخیص وضعیت تحصیلی دانشجویان پرداخته شده است. مجموعه داده نهایی که در این پژوهش مورد استفاده قرار گرفته شامل ۱۵۵۵ دانشجوی مقطع کارشناسی ناپیوسته دانشگاه شهاب دانش است. دو مجموعه داده برای انجام این پژوهش دریافت شد که اطلاعات آن به شرح زیر است:

۱. مجموعه داده‌ی مشخصات فردی (داده‌های جمعیت‌شناسی) و آموزشی دانشجویان شامل ۲۲۱۷۸ رکورد (به‌عنوان مجموعه داده۱)
۲. مجموعه داده‌ی معدل تمام ترم‌های دانشجویان (داده‌های آموزشی) شامل ۸۲۱۸۳ رکورد (به‌عنوان مجموعه داده ۲)

در روش پیشنهادی از الگوریتم‌های درخت تصمیم، نیوبیز، شبکه عصبی و ماشین بردار پشتیبان [۵۶] و نیز روش‌های گروهی Bagging، Boosting و جنگل تصادفی با این هدف استفاده شده است که تعیین شود دانشجویان پذیرش شده در مقطع کارشناسی در کدام طبقه‌بندی فارغ‌التحصیل خوب، فارغ‌التحصیل متوسط، فارغ‌التحصیل ضعیف، انصرافی و اخراجی قرار خواهند گرفت و وضعیت تحصیلی آنان به چه صورت خواهد بود.

۴،۱. آماده‌سازی داده

همانطور که قبلاً اشاره شد، داده‌ها در دو فایل جداگانه ذخیره شده بود و باید در نهایت این دو جدول پس از اصلاح، در یک جدول ادغام می‌شد؛ اما قبل از این کار، نیاز است تا این دو مجموعه داده آماده‌سازی شوند. از نرم‌افزار SQL Server برای بررسی، تبدیل نوع داده‌ها به نوع مناسب و ادغام دو مجموعه داده استفاده شده است. در ادامه لیست ویژگی‌های مجموعه داده‌ی ۱ (۲۷ ویژگی) و مجموعه داده‌ی ۲ (۷ویژگی) به تفکیک آمده است:

الف) مجموعه‌ی داده ۱: شماره دانشجویی، سال ورود، کدرشته، رشته تحصیلی، نام، نام‌خانوادگی، وضعیت تحصیلی، کد ملی، محل تولد، تاریخ تولد، دین، آدرس دائمی، تاریخ فارغ‌التحصیلی، مقطع تحصیلی، ترم ورود، وضعیت نظام وظیفه، جنسیت، گرایش، معدل کتبی آخرین مدرک، معدل کل آخرین مدرک، بومی/ غیربومی، تبعه کشور، معدل کل پیش دانشگاهی، وضعیت تاهل، رشته دیپلم/ پیش دانشگاهی، معدل کل ترمی، آخرین ترم تحصیلی

در مرحله اولیه‌ی بررسی مجموعه داده ۱، ویژگی‌هایی که دارای مقادیر خالی، ناقص، تعداد نمونه کم یا بدون ارزش بود، حذف شدند. این ویژگی‌ها موارد زیر بوده است:

نام، نام‌خانوادگی، معدل کل پیش دانشگاهی، رشته دیپلم/ پیش‌دانشگاهی، وضعیت نظام وظیفه، معدل کل آخرین مدرک، معدل کتبی آخرین مدرک، کد پاسپورت.

ب) مجموعه‌ی داده ۲: شماره دانشجویی، ترم تحصیلی، واحدهای اخذ شده، واحدهای قبولی، واحدهای مردودی، جمع ارزش، معدل

در مجموعه داده ۲ برای آن که مشخص شود هر دانشجو چند ترم در دانشگاه حضور داشته است یک ویژگی با نام تعداد ترم نیز ایجاد شد و با استفاده از امکانات برنامه اکسل و ویژگی کد دانشجو تعداد ترم حضور دانشجو بدست آمد.

۴،۲. ادغام مجموعه داده‌های ۱ و ۲

برای ادغام، دو مجموعه داده به نرم‌افزار SQL منتقل شده تا با استفاده از امکانات این نرم‌افزار عملیات ادغام انجام پذیرد. پس از ادغام دو مجموعه داده تعداد رکوردها به ۱۸۴۲۴ رسید.

علت کاهش رکوردها از ۲۲۱۷۸ به ۱۸۴۲۴ این است که برخی دانشجویان در ترم اول قبل از این‌که واحدی اخذ کنند انصراف داده‌اند یا توسط دانشگاه لغو ثبت نام یا اخراج شده‌اند و به همین دلیل نمره‌ای در مجموعه داده ۲ برای آن‌ها ثبت نشده است، اما در مجموعه داده ۱ مشخصات اولیه آن‌ها ثبت شده است.

۴،۳. پاکسازی داده‌ها

پس از ادغام و یکی کردن دو مجموعه داده در یک مجموعه

داده، حال نیاز است تا براساس هدف پژوهش کـه پـیش‌بینـی وضعیت تحصیلی دانشجویانی که قرار است از مقطع کاردانی به کارشناسی ناپیوسته وارد شوند، داده‌های اضافی از مجموعه داده حذف شوند. برای انجام اینکار ابتدا دانشجویانی که در مقطـع کارشناسی ناپیوسته هستند از مجموعه داده جداسازی شد و در مجموعه داده‌ای دیگر ذخیره شد. تعداد دانشجویان موجـود در این مقطع ۷۰۲۱ نفر است.

ویژگی «وضعیت تحصیلی» به‌عنوان برچسب[1] مورد استفاده در این پژوهش، برای پیش‌بینی انتخاب شده اسـت. بـا توجـه بـه هدف پژوهش، وضعیت تحصیلی دانشجویان در پنج دسته زیر تقسیم شده است:

۱. فارغ‌التحصیل خوب (نمره معدل بین ۱۷ الی ۲۰)
۲. فارغ‌التحصیل متوسط (نمره معدل بین ۱۴ الی ۱۷)
۳. فارغ‌التحصیل ضعیف (نمره کمتر از ۱۴)
۴. انصرافی
۵. اخراجی

دقت کنید که وضعیت‌های تحصیلی خوب، متوسط و ضعیف با استفاده از ویژگی معدل کل ترم‌ها دسته‌بندی شده است.

۴،۳،۱. کاهش داده‌ها

برای آنکه داده‌های مناسبی برای پیش‌بینی داشته باشیم در چند مرحله داده‌ها کاهش داده شد. در مرحلـه اول، وضـعیت‌هـای مختلف تحصیلی ثبت شده برای دانشجویان بررسـی شـد کـه شامل موارد زیر است:

۱. عـادی؛ ۲. معرفـی بـه اسـتاد؛ ۳. مهمـان از _ پایـان دوره؛ ۴.مرخصی؛ ۵. نامعلوم؛ ۶. انتقالی به؛ ۷. لغو ثبت نام؛ ۸. مهمان به؛ ۹. مهمان از؛ ۱۰. لغو قبولی _ نقص مدرک؛ ۱۱. درشـرف فارغ‌التحصیلی؛ ۱۲. فارغ‌التحصیل؛ ۱۳. انصرافی؛ ۱۴. پایان دوره؛ ۱۵. اخراجی.

موارد ۱ الی ۱۰ به دلیل بی استفاده بودن آنها (پایان نیافتن دوره تحصیلی) در فرایند داده‌کاوی حذف شد. در مرحله دوم داده‌ها ناقص حذف شدند. به عنوان نمونه داده‌های قبل از سال ۱۳۹۰ و بعد از سال ۱۳۹۶ حذف شدند. پس از این مراحل داده‌ها به ۳۱۳۸ رکورد رسید که در جدول (۳) نشان داده شده است.

جدول (۳): تعداد رکورد هر وضعیت تحصیلی

وضعیت تحصیلی	تعداد
فارغ‌التحصیل متوسط	۱۳۷۴
انصرافی	۵۱۴
فارغ‌التحصیل خوب	۳۶۱
فارغ‌التحصیل ضعیف	۶۱۴
اخراجی	۲۷۵
مجموع	۳۱۳۸

دقت داشته باشد کـه قبـل از اسـتفاده از تکنیـک‌هـای EDM، پیش‌پردازش هر منبع داده بطور جداگانه انجام شد تا بتوانیم با دو مشکل مهم که اغلب در داده‌های آموزشی وجود دارد مقابله کنیم [۵۷]. این دو مشکل عبارتند از:

۱. ابعاد بالا، یعنی تعداد زیادی از ویژگی‌هـا. تعـداد زیـاد ویژگی‌ها ممکن است مـانع پـیش‌بینـی الگـوریتم‌هـای پیش‌بینی برای رسیدن به نتایج جالب در یک زمان کوتاه شوند.
۲. داده‌های نامتعادل (نامتوازن).

الگوریتم‌های یادگیری ماشین در مواجهه با مجموعه داده‌هـای نامتوازن، طبقه‌بندی‌های نامناسبی را ایجـاد مـی‌کننـد. در یـک مجموعه داده‌های نامتوازن اگر رویدادی که می‌خـواهیم پـیش بینی کنیم به کلاسی که تعـداد کمـی نمونـه را دارد و نـرخ آن رویداد کم‌تر از ۵ درصد باشد، معمولا یک رویداد نادر محسوب می‌شود. به عبارت دیگر هنگامی کـه تعـداد نمونـه‌هـای یـک کلاس، بسیار بیشتر از تعداد نمونه‌های کلاس‌های دیگر اسـت [۵۸]، الگوریتم‌های پیش‌بینی تمایل دارند که روی یادگیری از کلاس‌هایی که تعداد بیشتری نمونه دارند تمرکز کنند.

همانطور که در جدول (۳) مشخص است وضـعیت تحصـیلیِ اخراجی کمترین تعداد (۲۷۵ نمونه) و فارغ‌التحصـیل متوسـط بیشترین تعداد (۱۳۷۴ نمونه) را دارند که این مسئله نکته دوم را نقص می‌کند. برای متـوازن کـردن داده‌هـا، تعـداد نمونـه‌هـای کلاس‌هایی که از وضعیت تحصیلیِ اخراجـی بیشـتر بودنـد را

[1] Lable

کاهش دادیم تا توازن بین داده‌ها برابر شود. در جدول(۴) این توازن تا حدودی رعایت شده است.

جدول (۴): تعداد رکورد وضعیت‌های تحصیلی

وضعیت تحصیلی	تعداد
فارغ‌التحصیل متوسط	۳۲۰
انصرافی	۳۳۰
فارغ‌التحصیل ضعیف	۳۳۰
فارغ‌التحصیل خوب	۳۲۲
اخراجی	۲۵۳
مجموع	۱۵۵۵

۴،۴. تبدیل داده‌ها

برای اجرای بهتر الگوریتم‌های داده‌کاوی، مقادیر متنی به مقادیر عددی تبدیل شد. به عنوان مثال جنسیت مرد و زن به اعداد یک و صفر تبدیل شد.

۴،۵. انتخاب ویژگی براساس وزن

انتخاب ویژگی از مهم‌ترین بخش‌های پیش پردازش است و با انجام این مرحله، حجم داده‌های پردازشی کمتر و عملیات داده‌کاوی سریع‌تر و دقت الگوریتم‌های یادگیری بیشتر می‌شود. روش‌های انتخاب ویژگی‌ها از لحاظ نحوه انتخاب به دو نوع انتخاب مجموعه‌ای و رتبه‌بندی ویژگی‌ها تقسیم می‌شوند. در جدول (۵) مقایسه سه الگوریتم وزن‌دهی زیر ذکر شده است.

- وزن بر اساس نسبت افزایش اطلاعات[1]
- وزن با افزایش اطلاعات[2]
- وزن بر اساس اهمیت درخت[3]

در انتخاب ویژگی براساس وزن در داده‌های موجود، سه ویژگی معدل کل، تعداد ترم، معدل ترم چهارم بیشترین وزن را به خود اختصاص داده‌اند. همانطور که در جدول مشخص است الگوریتم وزن بر اساس اهمیت درخت، وزن‌دهی متفاوت‌تری نسبت به دو الگوریتم دیگر داده است و مجموعه نمرات ترم سوم نیز وزن بیشتری به خود اختصاص داده است.

[1] Weight by Information Gain Ratio
[2] Weight by Information Gain
[3] Weight by Tree Importance

جدول (۵): انتخاب ویژگی براساس وزن

ردیف	ویژگی	وزن بر اساس نسبت افزایش اطلاعات	وزن با افزایش اطلاعات	وزن بر اساس اهمیت درخت
۱	میانگین کل ترم‌ها	۰/۹۷۳	۰/۸۳۳	۳/۸۲۴
۲	تعداد کل واحدها	۰/۸۴۹	۰/۷۸۴	۱/۴۶۲
۳	میانگین ترم ۴	۰/۷۳۴	۰/۶۷۶	۱/۲۰۸
۴	تعداد ترم‌ها	۰/۷۲۱	۰/۶۲۸	۰/۵۶۲
۵	مجموع نمرات ترم ۴	۰/۷۱۶	۰/۶۱۹	۰/۸۷۱
۶	تعداد واحد ترم ۴	۰/۷۰۷	۰/۶۰۸	۰/۷۶۷
۷	میانگین ترم ۲	۰/۶۳۰	۰/۵۵۳	۱/۰۶۴
۸	میانگین ترم ۳	۰/۶۱۳	۰/۵۸۶	۰/۸۲۳
۹	مجموع نمرات ترم ۳	۰/۶۰۰	۰/۵۹۲	۱/۲۸۲
۱۰	مجموع نمرات ترم ۲	۰/۵۹۶	۰/۵۸۰	۱/۰۸۵
۱۱	تعداد واحد ترم ۳	۰/۵۴۰	۰/۴۹۷	۰/۶۵۵
۱۲	میانگین ترم ۱	۰/۵۱۰	۰/۴۵۴	۱/۰۹۹
۱۳	مجموع نمرات ترم ۱	۰/۴۵۷	۰/۴۱۸	۱/۱۷۷
۱۴	تعداد واحد ترم ۲	۰/۴۳۸	۰/۴۳۲	۰/۶۵۷
۱۵	میانگین ترم ۶	۰/۴۲۱	۰/۲۴۳	۰/۵۲۳
۱۶	مجموع نمرات ترم ۶	۰/۳۹۳	۰/۲۳۹	۰/۴۱۰
۱۷	تعداد واحد ترم ۶	۰/۳۹۲	۰/۲۴۰	۰/۴۸۴
۱۸	میانگین ترم ۵	۰/۳۳۱	۰/۳۲۳	۰/۵۴۹
۱۹	مجموع نمرات ترم ۵	۰/۳۲۲	۰/۳۱۶	۰/۴۲۲
۲۰	تعداد واحد ترم ۵	۰/۳۱۸	۰/۳۱۳	۰/۴۴۶
۲۱	سن	۰/۲۱۸	۰/۰۳۰	۰/۲۴۵
۲۲	مسافت تا دانشگاه	۰/۲۰۳	۰/۰۳۱	۰/۱۸۰
۲۳	تعداد واحد ترم ۱	۰/۱۸۲	۰/۱۳۱	۰/۸۶۸
۲۴	جنسیت	۰/۱۱۷	۰/۱۰۴	۰/۰۳۰
۲۵	بومی / غیربومی	۰/۰۳۶	۰/۰۳۰	۰/۰۵۹
۲۶	وضعیت تأهل	۰/۰۰۳	۰/۰۰۱	۰/۰۵۹

۵. ارزیابی نتایج پژوهش

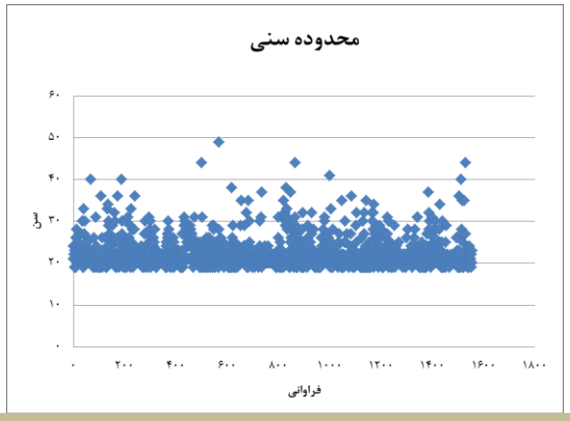
مجموعه داده‌ی نهایی توسط برنامه Rapidminer نسخه ۹،۴ [۵۹] تجزیه و تحلیل شده و نتایج حاصل از روش پیشنهادی در قالب جدول و نمودار آمده است.

با توجه به مجموعه داده‌ی نهایی، نمودارهای جمعیت‌شناسی[4] به دست آمد. در شکل (۲) تعداد دانشجویان مرد حدود ۶۸/۸۸ درصد از کل جمعیت داده‌های انتخاب شده را تشکیل داده است. شکل (۳) محدوده سنی دانشجویان را نمایش می دهد، که بر اساس نتایج به دست آمده، بیشتر دانشجویان سنی بین ۱۹ الی ۳۰ سال دارند. در نهایت با توجه به شکل (۴) بیشتر

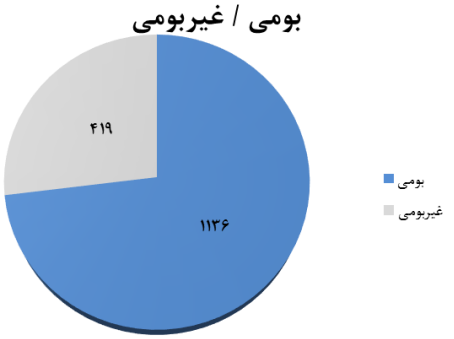
[4] Demography

دانشجویان بومی هستند و این مقدار به ۷۳ درصد از کل دانشجویان می‌رسد.

شکل (۲): جنسیت دانشجویان



شکل (۳): محدوده سنی دانشجویان



شکل (۴): وضعیت بومی / غیربومی بودن دانشجویان

۱،۵. معیار ارزیابی

فرآیند ارزیابی نحوه عملکرد مدل‌های داده‌کاوی در مقابل داده‌های واقعی را اعتبارسنجی[1] می‌گویند. نکته بسیار مهم قبل از پیاده‌سازی مدل‌های داده‌کاوی، تایید این مدل‌ها با درک کیفیت و ویژگی‌های آن‌هاست. در این پژوهش از ماتریس درهم‌ریختگی[2] استفاده می‌شود.

برای اعتبار سنجی نتایج از روش اعتبارسنجی متقابل k-Fold استفاده شده است. اگر مجموعه داده‌های آموزشی را به طور تصادفی به k زیرنمونه با حجم یکسان تفکیک کنیم، می‌توان در هر مرحله از فرایند اعتبارسنجی، تعداد $k-1$ از این لایه‌ها را به عنوان مجموعه داده آموزشی و یکی را به عنوان مجموعه داده اعتبارسنجی در نظر گرفت. در این تحقیق مقدار k را عدد ۱۰ در نظر گرفته‌ایم.

۲،۵. ماتریس درهم‌ریختگی

ماتریس درهم‌ریختگی ماتریسی است که عملکرد الگوریتم‌ها را نشان می‌دهد (جدول ۶). چنین نمایشی بیشتر برای الگوریتم‌های یادگیری با ناظر[3] استفاده می‌شود، هرچند در یادگیری بدون ناظر[4] نیز کاربرد دارد. این ماتریس یک ماتریس $N \times N$ است که N تعداد کلاس‌هاست. هر ستون از ماتریس، نمونه‌ای از مقدار پیش‌بینی شده را نشان می‌دهد. در صورتی که هر سطر نمونه‌ای واقعی (درست) را در بر دارد. این ماتریس کمک می‌کند تا آسان‌تر، اشتباه و تداخل بین نتایج را مشاهده کرد. توجه داشته باشید که *دقت*، نسبت تعداد کل پیش‌بینی‌ها است که در آن به درستی محاسبه شده‌اند.

جدول (۶): ماتریس درهم ریخته

		تشخیص‌داده‌شده	
		مثبت	منفی
واقعی	مثبت	درست مثبت (TP)	منفی کاذب (FN)
	منفی	اشتباه مثبت (FP)	منفی درست (TN)

۳،۵. نتایج حاصل از روش درخت تصمیم

درخت تصمیم با دقت ۸۵ درصد، دقت قابل قبولی را ارائه نداده است. طبق بخش ضمائم جدول (۷)، علت کاهش درصد پیش‌بینی، برچسب‌های انصرافی و اخراجی است. در مقابل

[1] Validation

[2] Confusion matrix
[3] Supervised learning
[4] UNSupervised

برچسب‌های فارغ‌التحصیل متوسط، خوب و ضعیف در کلاس فراخوانی ۱۰۰ درصد است که بسیار مناسب است.

در شکل (۵شکل () نمایی از خروجی درخت تصمیم آمده است که در آن ویژگی‌های تأثیرگذار در نتیجه نهایی مشخص است. در این تصویر FinalAvg (معدل کل) به عنوان ریشه در نظر گرفته شده است.

۵,۴. نتایج حاصل از روش نیوبیز

همان‌طور که در جدول (۸) در بخش ضمایم مشخص است، الگوریتم نیوبیز نتایج بسیار ضعیفی را ارائه داده است که میزان دقت آن به ۶۱/۰۹٪ رسیده است و برای این پژوهش چندان قابل استفاده نیست.

در این روش با تغییر درصد داده‌های مورد آزمون و آموزش میزان دقت، تغییر کمی داشته است که باز هم قابل استناد و استفاده در این پژوهش نیست. نتایج این تغییر در جدول (۹) بخش ضمائم آمده است.

۵,۵. نتایج حاصل از روش ماشین بردار پشتیبان

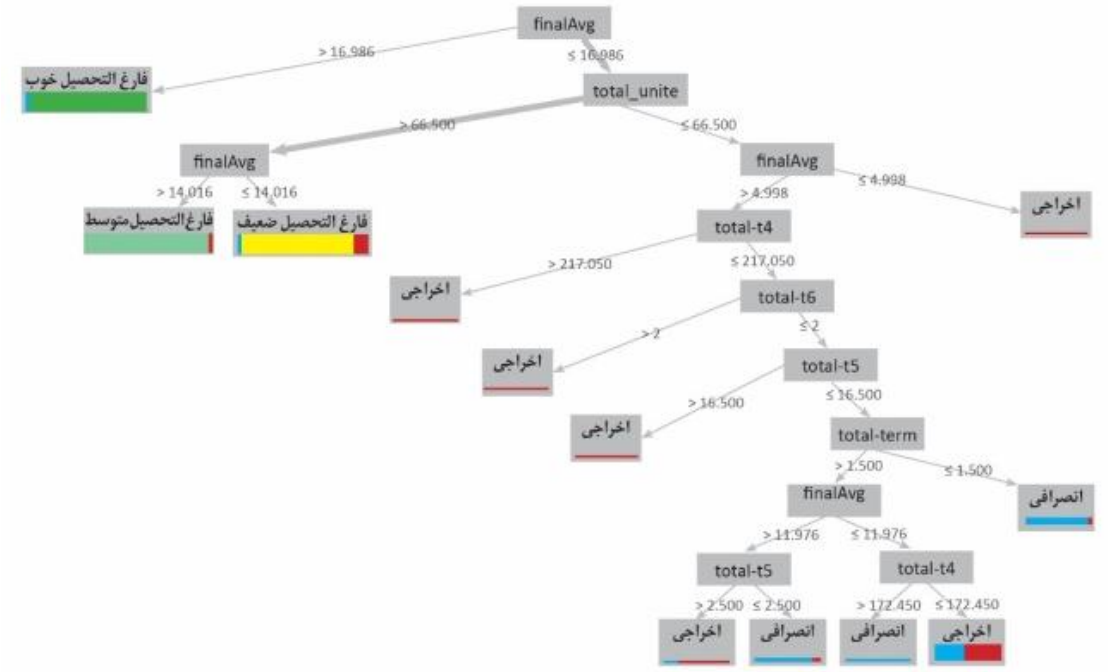
روش SVM نیز همانند نیوبیز نتایج قابل قبولی را ارائه نداده است و میزان دقت آن بر اساس جدول (۱۰) بخش ضمائم مقدار ۶۸/۳۹ درصد است. درصدهای ارائه شده در کلاس‌های فراخوانی و صحت نیز نشان از نامناسب بودن این روش در این تحقیق است.

۵,۶. نتایج حاصل از روش شبکه عصبی

نتایج حاصل از مدل شبکه عصبی نسبت به سه مدل درخت تصمیم، نیوبیز و SVM بهتر است، اما هنوز درصد قابل قبولی را ارائه نداده است و دقت آن هنوز کافی نیست. نتایج این روش در جدول (۱۱) بخش ضمائم قابل مشاهده است.

۵,۷. نتایج حاصل از روش Boosting

با توجه به اینکه مدل Boosting از ترکیب مدل‌های دیگر ساخته می‌شود انتظار می‌رود که میزان دقت بیشتری داشته باشد؛ اما با توجه به برچسب های مشخص شده و نیز ویژگی‌های دیگر، دقت آن کمتر از مدل شبکه عصبی است. همان‌طور که در جدول (۱۲) بخش ضمائم مشخص است در کلاس فراخوانی برای برچسب اخراجی، مقدار بسیار پایینی به‌دست آمده که در نتیجه کلی تأثیر بسزایی داشته است.



شکل (۵): نمای از ساختار خروجی درخت تصمیم

۵،۸. نتایج حاصل از روش Bagging

مدل Bagging نیز همانند مدل Boosting از ترکیب مدل‌های دیگر ساخته می‌شود. این مدل نسبت مدل Boosting نتایج بهتری ارائه داده است و دقتی نزدیک به مدل شبکه عصبی داشته است. در جدول (۱۳) بخش ضمائم این نتایج قابل مشاهده است. همان‌طور که مشخص است کلاس فراخوانی مقادیر بهتری نسبت به کلاس دقت ارائه داده است.

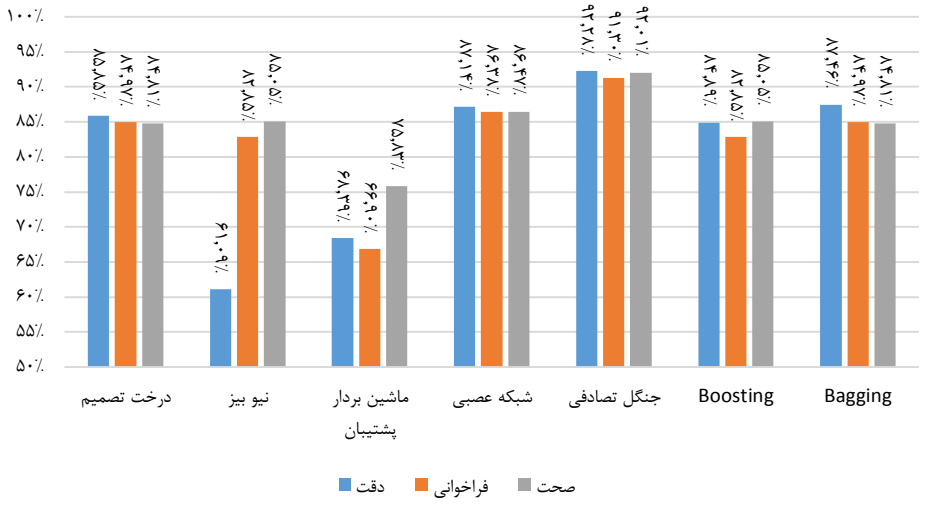
۵،۹. نتایج حاصل از روش جنگل تصادفی

در جدول (۱۴) بخش ضمائم نتایج مدل جنگل تصادفی آمده است. این مدل با دقت ۹۲/۲۸ درصد نسبت به تمام مدل های قبلی نتیجه بهتری ارائه داده است. در این جدول از تعداد ۳۵ داده انتخابی برای برچسب اخراجی در کلاس فراخوانی ۷ مورد و در کلاس دقت ۱۶ مورد به اشتباه پیش‌بینی شده است که در نتیجه کلی تأثیرگذار بوده است.

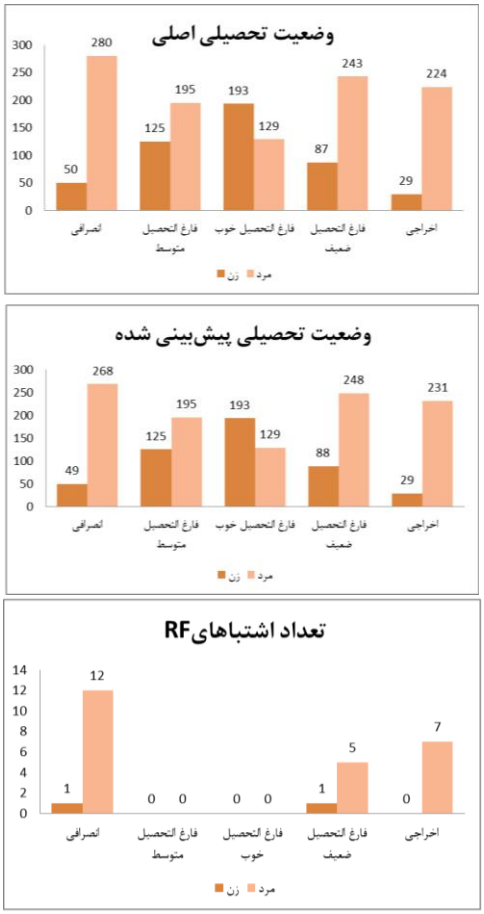
۵،۱۰. مقایسه روش‌های استفاده شده

مقادیر به دست آمده در شکل (۶) میانگین فراخوانی و صحت برچسب‌ها در روش‌های مذکور است. در این شکل به صورت نموداری، مقایسه‌ی روش‌های استفاده شده در این پژوهش نمایش داده شده است. همان‌طور که مشخص است می‌توان دید که جنگل تصادفی با ۹۲/۲۸٪ بیشترین دقت و نیوبیز با ۶۱/۰۹٪ کمترین دقت پیش‌بینی را دارند. با توجه به نکته گفته شده در خصوص توازن داده‌ها و نتایج به دست آمده می‌توان نتیجه گرفت علت کاهش درصد دقت روش‌های استفاده شده تعداد کم داده‌های برچسب اخراجی است.

در شکل (۷) میزان دقت روش جنگل تصادفی براساس جنسیت دانشجویان مشخص شده است. همان‌طور که در نمودار «تعداد اشتباهای RF» دیده می‌شود، تشخیص در برچسب انصرافی بیشتر از باقی برچسب‌ها است و در برچسب‌های فارغ التحصیل خوب و متوسط با دقت کامل این کار صورت گرفته است. در این تحقیق تلاش شد تا آنجا که امکان‌پذیر بود داده متوازن شود که مدل داده‌کاوی از صحت و دقت بالاتری برخوردار باشد، اما همان‌طور که مشخص است داده‌ها به‌طور کامل متوازن نیست و این یکی از دلایل عمده پایین بودن دقت روش های نیوبیز و SVM می‌باشد. با بررسی تحقیقات مشابه به طور معمول روش درخت تصمیم در این نوع مجموعه داده ها بهتر عمل می‌کند و روش جنگل تصادفی که بهبود یافته درخت تصمیم است بهترین نتایج را ارائه کرده است.



شکل (۶): نمودار مقایسه میزان دقت روش‌ها استفاده شده



شکل (۷): تعداد اشتباه‌های جنگل تصادفی، با توجه به جنسیت

۶. نتیجه‌گیری

این پژوهش در جهت داشتن سهمی در حوزه داده‌کاوی آموزشی سعی کرده است تا با استفاده از فناوری‌های نوین امروزی، قدمی برداشته و کمکی به بهبود فرآیندهای آموزشی کند. در بخش اول ساختار کلی این پژوهش بیان شد و در بخش دوم کارهای انجام شده در حوزه‌ی داده‌کاوی آموزشی مورد بررسی قرار گرفت است. به علت اینکه مجموعه داده‌های استفاده شده، داده‌های واقعی بود و هیچ‌گونه عملیات آماده‌سازی بر روی آنها انجام نشده بود، بیشترین کار انجام شده در بخش ۴ به بعد، توضیح روش آماده سازی است که توسط برنامه Rapidminer مجموعه داده‌ی نهایی، تجزیه و تحلیل شده و نتایج حاصل از روش پیشنهادی در قالب جدول و نمودار ایجاد شد. پس از بررسی و انجام روش‌های مختلف، روش جنگل تصادفی با ارائه بهترین نتایج به عنوان بهترین روش در این پژوهش انتخاب شد. نتایج به دست آمده برای روش جنگل تصادفی به ترتیب دقت ۹۲/۲۸٪، صحت ۹۲/۰۱٪ و فراخوانی ۹۱/۳۰٪ است. کمترین نتایج نیز برای روش نیوبیز با دقت ۶۱/۰۹٪ و SVM با فراخوانی ۶۶/۹۰٪ و صحت ۷۵/۸۳٪ بوده است. با توجه به کاربردهای وسیع شناخته شده داده کاوی در حوزه آموزش، نظام‌های آموزشی می‌توانند با جمع‌آوری داده‌های رفتاری محصلان مانند وضعیت مالی، شغلی، خانوادگی و ...، در کنار داده‌های آموزشی و بررسی و تجزیه و تحلیل آنها به بهبود فرآیندهای آموزشی و همچنین به بهبود عملکرد محصلان و در نهایت فارغ‌التحصیلی بهتر آنها کمک کنند.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[1] Han J., and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2011.

[۲] زواره ع.، «کاربرد داده‌کاوی بروی داده‌های آموزش عالی دانشجویان دندانپزشکی شهر رشت با استفاده از تکنیک‌های طبقه‌بندی و خوشه‌بندی»، اولین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات ایران، ۱۳۹۲.

[3] Educational Data Mining Group. Recouces, Murch 2021, https://educationaldatamining.org/ recouces.

[٤] افروز غ.، «جامعه فرهنگ و تدوین شخصیت کودکان ونوجوانان»، پیوند نشریه ماهانه آموزشی تربیتی انجمن اولیاء و مربیان جمهوری اسلامی ایران، شماره ۱۸۰، ۱۳۷٤.

[۵] بیابانگرد ا.، «روش‌های تحقیق در روانشناسی و علوم

مهندسین برق و الکترونیک-شاخه غرب، ۱۳۹٤.

[۱۵] فرهادی م.، تقوی م.، نوروزی م.، «پیش‌بینی موفقیت یا عدم موفقیت دانشجویان رشته عمران در فارغ‌التحصیلی با به‌کارگیری تکنیک‌های داده‌کاوی»، هفتمین کنفرانس داده‌کاوی ایران، تهران، ۱۳۹۲.

[۱٦] سنایی‌نسب ه.، رشیدی‌جهان ح.، صفاری م.، «عوامل مؤثر بر پیشرفت تحصیلی دانشجویان»، دو ماهنامه راهبردهای آموزش در علوم پزشکی، سال پنجم، شماره ٤، ص۲٤۳-۲٤۹، ۱۳۸۹.

[۱۷] تاری م.، «اثر بخشی داده‌کاوی در مدیریت آموزش عالی و مطالعه موردی آن در دانشگاه پیام‌نور استان قم»، سومین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات، ۱۳۸۹.

[۱۸] رودباری م.، احمدی ا.، عبادی‌فردآذر ف.، «تعیین عوامل موثر بر پیشرفت تحصیلی دانشجویان دانشگاه علوم پزشکی تهران (پردیس همت)»، نشریه طب و تزکیه، دوره ۱۹، شماره ۳ (مسلسل ۷۸)، ص۳۷-٤۸، ۱۳۸۹.

[19] Romero C. and Ventura S., “Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6): 601-618, 2010.

[20] Hung H-C, Liu I-F, Liang C-T, and Su Y-S., "Applying Educational Data Mining to Explore Students’ Learning Patterns in the Flipped Learning Approach for Coding Education". Symmetry, 12(2):213, 2020.

[21] Injadat M., Moubayed A., Nassif A., and Shami A., "Systematic Ensemble Model Selection Approach for Educational Data Mining", Knowledge-Based Systems, vol. 200, 2020.

[22] Durairaj M. and Vijitha C., "Educational data mining for prediction of student performance using clustering algorithms". Int. J. Comput. Sci. Inf. Technol, 5(4): 5987-5991, 2014.

[23] Francis B.K. and Babu, S.S., "Predicting academic performance of students using a hybrid data mining approach". J. Med. Syst, 43(6): 162, 2019.

[24] Akram A., Fu C., Li Y., Javed M.Y., Lin R., Jiang Y., and Tang Y., "Predicting students’ academic procrastination in blended learning course using

تربیتی»، تهران، انتشارات دوران، چاپ اول، ۱۳۸٤.

[6] Daradoumis T., Marquès Puig J.M., Arguedas M., and Calvet Liñan L.; “Analyzing students’ perceptions to improve the design of an automated assessment tool in online distributed programming”, Journal of Computers & Education, 128: 159-170, 2019.

[7] Adekitan A.I. and Salau O., “The impact of engineering students' performance in the first three years on their graduation result using educational data mining”, Heliyon, vol. 5, 2019.

[8] Thilagaraj T. and Sengottaiyan N., “A Review of Educational Data Mining in Higher Education System”, In: Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, pp. 349-358, 2017.

[۹] نبی‌لو م.، دانشپور ن.، «ارائه یک الگوریتم خوشه‌بندی برای داده‌های دسته‌ای با ترکیب معیارها»، مجله محاسبات نرم، جلد ۵، شماره ۱، ۱۳۹۵.

[10] Dekker G., Pechenizkiy M., and Vleeshouwers J., “Predicting Students Drop Out: A Case Study”, In: Proceedings of the International Conference on Educational Data Mining, 2nd, Cordoba, Spain, pp. 41-50, July. 2009.

[11] Romero C., Ventura S., Espejo P.G., and Hervas C.,“Data Mining Algorithms to Classify Students”, In: Proceedings of the 1st International Conference on Educational Data Mining, pp. 8-17, 2008.

[12] Superby J.F., Vandamme J.-P., and Meskens N., “Determination of factors influencing the achievement of the first-year university students using data mining methods”, In Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), pp. 37-44, 2006.

[۱۳] خاجه‌وند س.، چاله‌چاله ع.، «پیش‌بینی عوامل مؤثر در موفقیت تحصیلی دانشجویان دانشگاه پیام‌نور با کمک تکنیک‌های داده‌کاوی»، دومین کنفرانس ملی فناوری، انرژی و داده با رویکرد مهندسی برق و کامپیوتر، کرمانشاه، انجمن IEEE شاخه دانشجویی کردستان، ۱۳۹۵.

[۱٤] خیرخواه م.، جوانمرد م.، «کاربرد داده‌کاوی در سیستم آموزشی»، کنفرانس ملی فن‌آوری، انرژی و داده با رویکرد مهندسی برق و کامپیوتر، کرمانشاه، انجمن

homework submission data". IEEE Access, 7:102487–102498, 2019.

[25] Rojanavasu P., "Educational data analytics using association rule mining and classification". In: 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), pp. 142–145, 2019.

[26] Fatima S., Siddiqui I.F., and Ali Q., "Analyzing students' academic performance through educational data mining: 3c Tecnol. glosas innovación Apl. a la pyme", International Multi-Topic Conference on Engineering and Science,402–421, 2019.

[27] Alban M. and Mauricio D., "Neural networks to predict dropout at the universities". Int. J. Mach. Learn. Comput. 9(2): 149–153, 2019.

[28] Feng J., "Predicting students' academic performance with decision tree and neural network", PhD Dissertation, 2019.

[29] Daradoumis T., Marquès Puig J.M, Arguedas M., and Calvet Liñan L., "Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming", Journal of Computers & Education, 128:159-170, 2019.

[30] Bharara S., Sabitha S., and Bansal A.,"Application of learning analytics using clustering data Mining for Students' disposition analysis". Educ. Inf. Technol, 23(2): 957–984, 2018.

[31] Nurhayati O.D., Bachri O.S., Supriyanto A., and Hasbullah M., "Graduation prediction system using artificial neural network". Int. J. Mech. Eng. Technol, 9(7): 1051–1057, 2018.

[32] Rao K.S., Swapna N., and Kumar P.P., "Educational data mining for student placement prediction using machine learning algorithms". Int. J. Eng. Technol. Sci, 7(1.2): 43–46, 2018.

[33] Okubo F., Yamashita T., Shimada A., and Ogata H., "A neural network approach for students' performance prediction". In: LAK 2017, pp. 598–599, 2017.

[34] Almarabeh H., "Analysis of students' performance by using different data mining classifiers". Int. J. Mod. Educ. Comput. Sci, 9(8): 9, 2017.

[35] Costa E.B., Fonseca B. , Santana M.A., de Araújo F.F., and Rego J., "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", Computers in Human Behavior, 73:247-256, 2017.

[36] Amrieh E.A., Hamtini T., and Aljarah I., "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods", International Journal of Database Theory and Application, 9 (8): 119-136, 2016.

[37] Saranya S., Ayyappan R., and Kumar N., "Student progress analysis and educational institutional growth prognosis using data mining". International Journal of Engineering Sciences & Research Technology (IJESRT), 3(4): 1982-1987, 2014.

[38] Ariouat H., Cairns A.H., Barkaoui K., Akoka J., and Khelifa N., "A two-step clustering approach for improving educational process model discovery". 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, pp. 38-43, 2016.

[39] Arora R.K., Badal D., "Mining association rules to improve academic performance". Int. J. Comput. Sci. Mob. Comput., 3 (1): 428-433, 2014.

[40] Osmanbegovi´c E. and Sulji´c M., "Datamining approach for predicting student performance", Economic Review - Journal of Economics and Business. vol. 1, 2012.

[41] Sukanya M., Biruntha S., Karthik S., and Kalaikumaran T., "Data mining: performance improvement in education sector using classification and clustering algorithm". International Conference on Computing and Control Engineering, 2012.

[42] Torenbeek M., Jansen E.P.W.A., and Hofman W.H.A., "Predicting first-year achievement by pedagogy and skill development in the first weeks at university", Teach. High. Educ, 16(6): 655-668, 2011.

[43] He Y. and Zhang S.; "Application of data mining on students' quality evaluation", 3th International Workshop on Intelligent Systems and Applications, Wuhan, pp. 1-4, 2011.

[44] Sakurai Y., Tsuruta S., and Knauf R., "Success chances estimation of university curricula based on educational history, self-estimated intellectual traits and vocational ambitions". In: IEEE 11th International Conference on Advanced Learning Technologies, Athens, GA, pp. 476-478, 2011.

[45] Aher B.S. and Lobo L.M.R.J., "combination of clustering, classification & association rule-based approach for course recommender system in e-

[۵۶] وثیقی‌ذاکر ا.، جلیلی س.، «پیش‌بینی ژن‌های بیماری با استفاده از دسته‌بند تک‌کلاسی ماشین بردار پشتیبان»، مجله محاسبات نرم، جلد ٤، شماره ۱، ۱۳۹٤.

[57] Marquez-Vera C., Morales C., and Soto S., "Predicting school failure and dropout by using data mining techniques", IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 8: 7-14, 2013.

[58] Gu Q., Cai Z., Zhu L., and Huang B.,"Data mining on imbalanced data sets. In Advanced computer theory and engineering", In: ICACTE '08. International conference, pp. 1020-1024, 2008.

[59] Rapidminer Group. Products, Products/Studio/, Version2020, Murch2020, https://rapidminer.com /Products/Studio/

learning". Int. J. Comput. Appl, 39(7):8-15, 2012.

[46] Ayesha S., Mustafa T., Sattar A.R., and Khan M.I., "Data mining model for higher education system", Eur. J. Sci. Res, 43(1): 24-29, 2010.

[47] Kovačić Z.J., "Early prediction of student success: mining students enrolment data", In: Proceedings of Informing Science and IT Education Conference (InSITE), pp. 647-665. 2010.

[48] Al-shargabi A.A. and Nusari A.N., "Discovering vital patterns from UST student's data by applying data mining techniques", In: 2th International Conference on Computer and Automation Engineering (ICCAE), Singapore, pp. 547-551, 2010.

[49] Yan Z., Shen Q., and Shao B., "The analysis of student's grade based on Rough Sets", In: 3th IEEE International Conference on Ubi-Media Computing, Jinhua, pp. 345-349, 2010.

[50] Ningning G., "Proposing datawarehouse and data mining in teachingmanagement research", International Forum on InformationTechnology andApplications, Kunming, pp. 436-439, 2010.

[51] Knauf R., Sakurai Y., Takada K., and Tsuruta S., "Personalizing learning processes by datamining". In: 10th IEEE International Conference on Advanced Learning Technologies, Sousse, pp. 488-492, 2010.

[52] Xiangjuan B. and Youping G., "The application of data mining technology in analysis of college student's performance", In: The 2nd International Conference on Information Science and Engineering, Hangzhou, China, pp. 5477-5480, 2010.

[53] Liu Z. and Zhang X.; "Prediction and analysis for students' marks based on decision tree algorithm", In: Third InternationalConference on IntelligentNetworks and Intelligent Systems, Shenyang, pp. 338-341, 2010.

[54] Akulwar P., Pardeshi S., and Kamble A., "Survey on Different Data Mining Techniques for Prediction", In: Proceedings of the Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 513-519, 2018.

[۵۵] ویسی ه.، قایدشرف ح.، ابراهیمی م.، «بهبود کارایی الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های قلبی با بهینه‌سازی داده‌ها و ویژگی‌ها»، مجله محاسبات نرم، جلد ۸، شماره ۱، ۱۳۹۸.

ضمائم

جدول (۱): ماتریس درهم‌ریختگی درخت تصمیم

دقت (Accuracy) ٪۸۵/۸۵	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	class precision (صحت)
pred. انصرافی	۴۱	۰	۰	۰	۹	٪۸۲
pred. فارغ‌التحصیل متوسط	۰	۶۴	۰	۰	۱	٪۹۸/۴۶
pred. فارغ‌التحصیل خوب	۱	۰	۶۴	۰	۰	٪۹۸/۴۶
pred. فارغ‌التحصیل ضعیف	۰	۰	۰	۶۶	۹	٪۸۸
pred. اخراجی	۲۴	۰	۰	۰	۳۲	٪۵۷/۱۴
فراخوانی (class recall)	٪۶۲/۱۲	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۶۲/۷۵	

جدول (۲): ماتریس درهم‌ریختگی نیو بیز

دقت ٪۶۱/۰۹	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
pred. انصرافی	۶۱	۰	۰	۰	۲۲	٪۷۳/۴۹
pred. فارغ‌التحصیل متوسط	۰	۶۴	۶	۴	۰	٪۸۶/۴۹
pred. فارغ‌التحصیل خوب	۰	۰	۵۸	۰	۰	٪۱۰۰
pred. فارغ‌التحصیل ضعیف	۰	۰	۰	۶۲	۱۰	٪۸۶/۱۱
pred. اخراجی	۵	۰	۰	۰	۱۹	٪۹۷/۱۷
فراخوانی	٪۹۲/۴۲	٪۱۰۰	٪۹۰/۶۲	٪۹۳/۹۴	٪۳۷/۲۵	

جدول (۳): ماتریس درهم‌ریختگی نیوبیز با تغییر داده تست و آموزش

دقت ٪۶۱/۴۶	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
pred. انصرافی	۹۷	۰	۰	۰	۵۵	٪۶۳/۸۲
pred. فارغ‌التحصیل متوسط	۰	۹۵	۷۹	۳۰	۵	٪۴۵/۴۵
pred. فارغ‌التحصیل خوب	۰	۰	۱۸	۰	۰	٪۱۰۰
pred. فارغ‌التحصیل ضعیف	۰	۱	۰	۶۶	۵	٪۹۱/۶۷
pred. اخراجی	۲	۰	۰	۳	۱۱	٪۶۸/۷۵
فراخوانی	٪۹۸/۹۸	٪۹۸/۹۶	٪۱۸/۵۶	٪۶۶/۶۷	٪۱۴/۴۷	

جدول (۴): ماتریس درهم‌ریختگی SVM

دقت ٪۶۸/۳۹	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
pred. انصرافی	۲۸	۰	۰	۰	۴	٪۸۷/۵۰
pred. فارغ‌التحصیل متوسط	۰	۱۷	۲	۳	۰	٪۷۷/۲۷
pred. فارغ‌التحصیل خوب	۰	۲	۲۱	۰	۰	٪۹۱/۳۰
pred. فارغ‌التحصیل ضعیف	۲	۱۳	۹	۳۰	۱۱	٪۴۶/۱۵
pred. اخراجی	۳	۰	۰	۰	۱۰	٪۷۶/۹۲
فراخوانی	٪۸۴/۸۵	٪۵۳/۱۲	٪۶۵/۶۲	٪۹۰/۹۱	٪۴۰	

جدول (۵): ماتریس درهم‌ریختگی شبکه عصبی

دقت ٪۸۷/۱۴	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
انصرافی .pred	۵۲	۰	۰	۰	۱۵	٪۷۷/۶۱
فارغ‌التحصیل متوسط .pred	۰	۵۹	۳	۰	۰	٪۹۵/۱۶
فارغ‌التحصیل خوب .pred	۰	۱	۶۱	۰	۱	٪۹۶/۸۳
فارغ‌التحصیل ضعیف .pred	۰	۴	۰	۶۴	۰	٪۹۴/۱۲
اخراجی .pred	۱۴	۰	۰	۲	۳۵	٪۶۸/۶۳
فراخوانی	٪۷۸/۷۹	٪۹۲/۱۹	٪۹۵/۳۱	٪۹۶/۹۷	٪۶۸/۶۳	

جدول (۶): ماتریس درهم‌ریختگی BOOSTING

دقت ٪۸۴/۸۹	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
انصرافی .pred	۶۱	۰	۰	۰	۲۲	٪۷۳/۴۹
فارغ‌التحصیل متوسط .pred	۰	۶۴	۶	۴	۰	٪۸۶/۴۹
فارغ‌التحصیل خوب .pred	۰	۰	۵۸	۰	۰	٪۱۰۰
فارغ‌التحصیل ضعیف .pred	۰	۰	۰	۶۲	۱۰	٪۸۶/۱۱
اخراجی .pred	۵	۰	۰	۰	۱۹	٪۷۹/۱۷
فراخوانی	٪۹۲/۴۲	٪۱۰۰	٪۹۰/۶۲	٪۹۳/۹۴	٪۳۷/۲۵	

جدول (۷): ماتریس درهم‌ریختگی BAGGING

دقت ٪۸۷/۴۶	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
انصرافی .pred	۴۱	۰	۰	۰	۹	٪۸۲
فارغ‌التحصیل متوسط .pred	۰	۶۴	۰	۰	۱	٪۹۸/۴۶
فارغ‌التحصیل خوب .pred	۱	۰	۶۴	۰	۰	٪۹۸/۴۶
فارغ‌التحصیل ضعیف .pred	۰	۰	۰	۶۶	۹	٪۸۸
اخراجی .pred	۲۴	۰	۰	۰	۳۲	٪۵۷/۱۴
فراخوانی	٪۶۲/۱۲	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۶۲/۷۵	

جدول (۸): ماتریس درهم‌ریختگی جنگل تصادفی

دقت ٪۹۲/۲۸	True انصرافی	True فارغ‌التحصیل متوسط	True فارغ‌التحصیل خوب	True فارغ‌التحصیل ضعیف	True اخراجی	صحت
انصرافی .pred	۵۹	۰	۰	۰	۱۶	٪۷۸/۶۷
فارغ‌التحصیل متوسط .pred	۰	۶۴	۰	۰	۰	٪۱۰۰
فارغ‌التحصیل خوب .pred	۰	۰	۶۴	۰	۰	٪۱۰۰
فارغ‌التحصیل ضعیف .pred	۰	۰	۰	۶۵	۰	٪۱۰۰
اخراجی .pred	۷	۰	۰	۱	۳۵	٪۸۱/۴۰
فراخوانی	٪۸۹/۳۹	٪۱۰۰	٪۱۰۰	٪۹۸/۴۸	٪۶۸/۶۳	