نشریه علمی ـ پژوهشی
دریافت مقاله: ۹۳/۱۰/۲۴
پذیرش نهایی: ۹۴/۰۲/۱۲
http: //jte.srttu.edu
ISSN: 2345 – 5462
فنـــاوری آمـــوزش
جـــلـد ۹، شـــمـاره ۴، پاییز ۱۳۹۴ صص ۲۴۵ تا ۲۵۸
دانشگاه تربیت دبیر شهید رجائی

پیش‌بینی میزان پیشرفت تحصیلی دانشجویان با روش خوشه‌بندی فازی در محیط‌های آموزشی

محمد رستمی[۱]، سید سعید آیت[۲]، فرید صاغری[۳]و فاطمه یعقوبی[۴]

[۱]عضو باشگاه پژوهشگران و نخبگان جوان دانشگاه آزاد اسلامی واحد دهاقان، اصفهان.
[۲]**دانشیار** گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه پیام نور، (نویسنده مسئول) پست الترونیکی: dr.ayat@pnu.ac.ir
[۳و۴]**کارشناس ارشد** گروه مهندسی کامپیوتر

چکیده: هدف این پژوهش ارائه الگویی جهت پیش‌بینی عملکرد و افزایش کارایی و موفقیت یادگیری دانشجویان در یک محیط آموزشی با استفاده از داده‌کاوی است. با تکیه به روش‌های کتابخانه‌ای و پرسشنامه‌ای و مشاوره با افراد خبره تعدادی از ویژگی‌های تأثیرگذار در یادگیری دانشجویان شناسایی شد و با استفاده از روش انتخاب ویژگی، مؤثرترین آنها انتخاب شدند و برای روشن‌تر شدن روابط بین ویژگی‌های انتخاب شده، خوشه‌بندی فازی بر روی آنها انجام گرفت. در فاز دوم پژوهش با استفاده از تکنیک‌های داده‌کاوی به پیش‌بینی نمرات دانشجویان محیط آموزشی مورد مطالعه پرداخته شد. فیلدهایی که به عنوان متغیر در نظر گرفته شد، نمره میان‌ترم، پایان‌ترم و نمره نهایی (معدل) دروس اخذ شده در یک ترم توسط دانشجویان ورودی ۱۳۸۵ تا ۱۳۹۱ دانشگاه است.
بر مبنای الگوهای به دست آمده می‌توان هر دانشجو را در راستای ویژگی‌های تأثیرگذار بر روی آنها (دانشجویان) از ابتدای ترم راهنمایی و با توجه به نمراتی که در طول ترم کسب می‌کند، او را از محدوده نمره نهایی خود آگاه کرد و بر طبق توانایی‌هایش برنامه‌ریزی مناسب تحصیلی نمود. این الگوها می‌توانند برای کارآمدتر ساختن فرآیند یادگیری در سیستم مؤثر باشند. نتایج آزمایش‌ها حاکی از دقت مطلوب روش پیشنهادی ۰/۹۳۹ نسبت به روش‌های قبلی (کشف قوانین همبستگی،کلاس‌بندی و تشخیص ناهمگونی).

واژگان کلیدی: آموزش الکترونیکی، انتخاب ویژگی، خوشه‌بندی فازی، داده کاوی، کشف قوانین همبستگی.

Applying fuzzy clustering to assess and anticipate students' educational progress in learning environments

Mohammad Rostami[1],Dr. Seyed Saeed Ayat[2], Farid Saghari[3], Fatemeh Yaghoobi[4]

[1]Member of Young Researchers Club, Islamic Azad University, Dehaghan Branch, Isfahan, Iran, mohammadrostami@dehaghan.ac.ir
[2]Associate Professor, Department of Computer Engineering and Information Technology, Payame Noor University, dr.ayat@pnu.ac.ir
[3,4]Software Engineering Department, 7lac.net@gmail.com, fatemehyaghobi20@gmail.com

Abstract

The purpose of this paper is to propose a method to anticipate students' proceed and to enhance their learning efficiency and success in a learning environment, using data mining. Based on library and survey searching methods, as well as consulting with experts, some effective features in students' learning are identified and then using feature selection method, the most efficient ones are chosen. To clarify the relation between selected features, fuzzy clustering is applied to them. In the second phase of the research, scores of the students of Educational environment study, are predicted, using data mining. Variables taken are midterm and final scores and the average score of selected units in one semester by students studying there between 2006 (1385) and 2012 (1391).

According to the achieved methods we can guide each student from the beginning of the semester in line with their effective features, and based on scores gained during the semester we can inform the student about his range of final score to receive an educational plan based on his/her abilities. These methods can be effective in streamlining learning procedure in a system. Test results show the desired accuracy (0.939) of the proposed method than previous methods (discovery of association rules, classification, and identifying the inconsistencies).

Key Words: Electronic education, Feature selection, Fuzzy clustering, Data mining, Detection of integrity rules.

۱– مقدمه

امروزه در اکثر دانشگاه‌های ایران، بانک‌های اطلاعاتی وسیعی از ویژگی‌های دانشجویان موجود است که حجم بالایی از اطلاعات مربوط به سوابق آموزشی، تحصیلی و غیره را شامل می‌شود. نرم‌افزارهای رایانه‌ای به کار گرفته شده برای این منظور، غالبا فقط برای مکانیزه کردن وضع موجود، اجرای پرس‌وجوهای معمولی و برنامه‌ریزی کوتاه مدت اداری پاسخ‌گو هستند. درحالی که در عمق درون این حجم داده‌ها، الگوها و روابط بسیار جالبی میان پارامترهای مختلف به صورت پنهان باقی می‌ماند. داده‌کاوی[۱] یک تکنیک میان‌رشته‌ای برای اکتشاف این الگوها است، که از علوم یادگیری ماشین، تشخیص الگو[۲]، آمار، پایگاه داده و بصری‌سازی[۳] به منظور استخراج اطلاعات از پایگاه‌های داده بزرگ بهره‌مند می‌شود[۱]. دانش قابل کشف از طریق داده‌کاوی در حوزه آموزش نه تنها قابل استفاده صاحبان سیستم یعنی مدرسین و مسئولین آموزشی بلکه قابل استفاده کاربران سیستم یعنی دانشجویان نیز است[۲ و۳].

مطابق[۴] در حلقه‌های تکرار داده‌کاوی، دانش از داده-های خام استخراج می‌شود و این دانش به مرور، پالایش شده و فیلتر می‌گردد.

آموزش برخط به معنای ارائه اطلاعات و مطالب آموزشی به مخاطبانی در فاصله‌های دور با استفاده از فناوری‌های رایانه‌ای و شبکه جهانی است. این آموزش دو مشخصه اصلی دارد:

- مربی یا استاد و دانشجو در مکان‌های متفاوتی قرار دارند.
- از شبکه جهانی برای از بین بردن فاصله مکانی استفاده می‌شود[۵ و ۶].

سیستم‌های آموزشی هوشمند، برنامه‌های رایانه‌ای هستند که به افراد در حال آموزش، روش هوشمندی را ارائه می‌دهند. در آموزش به صورت حضوری به دلیل تعامل مستقیم استاد و دانشجو، استاد می‌تواند میزان یادگیری دانشجو را ارزیابی کند و پیشنهادهای لازم برای راهنمایی به وی را ارائه دهد، در حالی که در سیستم‌های آموزشی غیر حضوری به دلیل عدم وجود این رابطه، وجود سیستم‌هایی برای ارزیابی یادگیری دانشجو و همچنین تشخیص عوامل مؤثر بر یادگیری از اهمیت ویژه‌ای برخوردار است. لذا با توجه به پیشرفت روز افزون فناوری اطلاعات و ارتباطات و تمایل یادگیرندگان به آموزش برخط، استفاده از روش‌هایی که بتواند عوامل مؤثر بر یادگیری اشخاص را شناسایی و آنها را در جهت پیشرفت در یادگیری مشاوره کند، ضروری و لازم است در جهت فراهم کردن زیرساخت‌ها برای مکانیزه کردن این روش ها اقداماتی انجام گیرد.

تحقیق حاضر با درک اهمیت این موضوع و همچنین به دلیل نیاز ساختار سازمانی آموزش‌وپرورش و وزارت علوم در مقطع دانشجویان دانشگاه در تلاش است تا سؤال اصلی تحقیق مبنی بر چگونگی استفاده از روش داده کاوی برای پیش‌بینی موفقیت یادگیری در یک محیط آموزشی را پاسخ گوید.

فناوری اطلاعات و ارتباطات وجه تمایز عصر حاضر با دوران گذشته است. آنچه که امروز تحت عنوان شکاف یا فاصله بین کشورها، ملت‌ها، اقشار و افراد مطرح است، با میزان بهره‌مندی و کاربرد فنآوری اطلاعات و ارتباطات تناسب مستقیم دارد. از این‌رو میزان توسعه و کاربرد فنآوری اطلاعات و ارتباطات در امر آموزش مهمترین شاخص پیشرفت به شمار می‌رود.

داده‌کاوی، پایگاه‌ها و مجموعه‌های حجیم داده‌ها را در پی کشف واستخراج دانش، مورد تحلیل و کندوکاوهای ماشینی (و نیمه‌ماشینی) قرار می‌دهد.

امروزه بسیاری از امور مؤسسات آموزشی توسط سامانه‌های اینترنتی انجام می‌پذیرد و دانشجویان اغلب امور ثبت‌نام، حذف و اضافه و ارزشیابی خود را توسط این سیستم‌ها انجام می‌دهند. هرچند این داده‌ها متنوع و متعددند اما در اغلب سیستم‌ها پردازش قابل توجهی روی آنها انجام نمی‌شود و اغلب به صورت خام نگهداری می‌گردند. لذا مدرس، مدیر و سایر کاربران به ابزارهایی جهت مشاهده گزارش‌های مفید برای خودشان نیاز دارند تا بتوانند تصمیم‌های لازم را برای بهبود عملکرد سیستم اتخاذ نمایند. به روند تبدیل داده‌های آموزشی خام به اطلاعاتی که بتواند در تصمیم‌گیری‌های طراحی مفید واقع شود یا پاسخی برای پرسش‌های تحقیقی باشد، داده‌کاوی آموزشی گفته می شود. هدف داده-

کاوی آموزشی کاملا بستگی به این دارد که در نهایت چه کسی قرار است از نتایج آن استفاده کند؛ دانشجو، مدرس، مدیر آموزش و یا سایر مسئولین. تکنیک‌های به کار رفته هم، متنوع و گاه ترکیبی هستند و بنا به اهداف مختلف، طراحی و پیاده‌سازی شده‌اند.

در مقاله[۷]، دانش‌آموزان دو کلاس مورد بررسی قرار گرفته‌اند و مشاهده شده است که اگر نمرات دوره‌های اول و دوم موجود باشند، پیش‌بینی‌های خوبی برای دوره‌های بعدی می‌توان انجام داد. این بدان معنا است که هر چند موفقیت یادگیرندگان خیلی به نتایج قبلی‌شان وابسته است اما جنبه‌های مرتبط دیگری هم وجود دارد، که از جمله آن‌ها می‌توان تعداد غیبت‌ها، شغل و تحصیلات و سلامت اجتماعی خانواده را نام برد.

رومرو و دیگران در[۸]، چند راه برای استفاده از طبقه-بندی در محیط آموزشی مشخص کرده‌اند: کشف گروه-های دانشجویی با ویژگی‌های مشابه، شناسایی زبان آموزان با انگیزه کم، پیشنهاد اقدامات اصلاحی، و پیش‌بینی وطبقه‌بندی دانش‌آموزان با استفاده از سیستم‌های آموزش هوشمند از آن جمله است.

هانگ و زانگ در مقاله[۹]، از تکنیک داده‌کاوی برای کشف الگوهای یادگیری الکترونیکی دانشجویان و حمایت از مدیریت یادگیری الکترونیکی، تسهیلات و طرح های آن استفاده کردند. نتایج مطالعات آنان الگوهای یادگیری و عملکرد دانشجویان را نشان داد که باعث تشخیص دانشجویان فعال از دانشجویان غیر فعال و همچنین پارامترهای مهم جهت پیش‌بینی عملکرد دانشجویان گردید[۱۰ و ۱۱].

در این مقاله با استفاده از تکنیک‌های خوشه‌بندی K-means و درخت تصمیم C&R دانشجویان را گروه‌بندی و الگوهایی برای افزایش کارایی آنها ارائه می‌شود. با استفاده از تکنیک K-means دانشجویان بر اساس نمراتشان تقسیم‌بندی شدند و عملکرد آنها در طول ترم به صورت نمودار نشان داده می شود. با استفاده از درخت تصمیم C&R توانستیم دانشجویان را با توجه به عملکردشان در طول ترم یعنی با توجه به ورودی‌هایی چون نمرات میان‌ترم، پایان‌ترم و تمرین و با هدف تعیین نمره نهایی، کلاس‌بندی و الگوهایی برای پیش‌بینی نمره نهایی آنها ارائه کنیم. از آنجا که خوشه‌بندی یک روش یادگیری بدون ناظر است، بیشتر به دنبال یافتن گروه‌هایی از دانشجویان است که قبلا شناخته نشده‌اند و از قبل پیش‌بینی در مورد شباهت‌های موجود ندارد و تفسیر نتایج آن کمی مشکل است، ولی درخت تصمیم یک روش یادگیری با ناظر است که بر اساس یک برچسب کلاس معلوم، دانشجویان را به گروه‌هایی کلاس‌بندی می‌کند و به پیش‌گویی رفتار دانشجویان جدید بر اساس این کلاس‌بندی می‌پردازد.

برای شخصی‌سازی هرچه بیشتر یک سیستم یادگیری الکترونیکی، بهتر آن است که با استفاده از تکنیک خوشه‌بندی به کشف گروه‌های جدید پرداخت. عملکرد دانشجویان را مورد ارزیابی قرار داد و با کمک درخت تصمیم رفتار دانشجویان جدید را پیش‌بینی نمود و روش یادگیری مناسب را به آنها نشان داد. به دست آوردن اطلاعات و دانش مفید و الگوهای غیر بدیهی از موضوعات کلیدی وابسته به یادگیری الکترونیکی، در آموزش عالی ضروری است. فرآیند داده‌کاوی در این امر موفق بوده است.

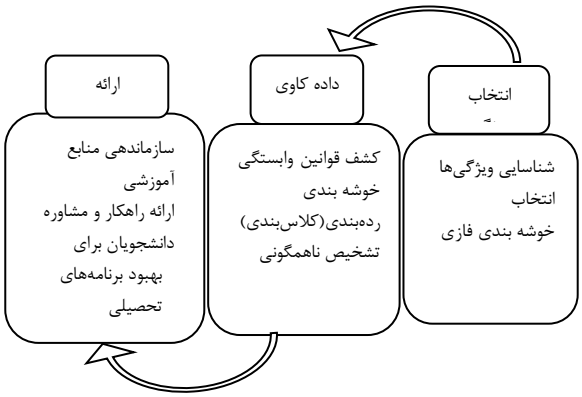
۲– روش تحقیق

امروزه در دانشگاه‌ها آموزش از راه دور مبتنی بر وب و برخط بسیار محبوب و مورد توجه مردم قرار گرفته است. این روش از لحاظ آموزش و یادگیری فرآیندی قابل اطمینان و مطابق با علم آموزش و از لحاظ هزینه مقرون به صرفه است. هر دانشجو با سیستم یادگیری الکترونیکی ارتباط و تعامل برقرار می‌کند و نمرات و فعالیت‌های او در پایگاه‌های داده ذخیره می‌گردد. در بیشتر این سیستم‌ها، این داده‌های ذخیره شده فقط برای گزارش‌های آماری به کار برده می‌شود، که این موقعیت، در حالی که می‌دانیم این مجموعه داده‌ها حاوی اطلاعات مفیدی هستند، بسیار نامناسب است. با کمک تکنیک‌های داده‌کاوی می‌توان این داده‌ها را تحلیل نمود و یک فرآیند یادگیری مؤثر و کارآمد ایجاد نمود. روش پژوهش در این مقاله کتابخانه‌ای و پرسشنامه‌ای است.

یادگیری امر مهمی است که در هر جامعه‌ای توجه

زیادی به آن می‌شود. عوامل متعددی تأثیرگذار در یادگیری افراد است که طبق مطالعات انجام گرفته در این مقاله ۱۱ عامل سن، جنسیت، وضعیت اشتغال، وضعیت خوابگاهی بودن، تأهل، وضعیت سلامت، تعداد افراد خانواده، وضعیت تغذیه، قدرت محاسبات ریاضی، تعداد واحد درسی گذرانده و نمرات میان‌ترم دروس را می‌توان نام برد. در این مقاله، روشی پیشنهاد شده است که با استفاده از الگوریتم‌های انتخاب ویژگی و با استفاده از دانش افراد خبره در این زمینه بتوان از میان این عوامل، آن دسته از ویژگی‌هایی که مهم است و از نظر فرد خبره تأثیرگذاری آنها در فراگیری مطالب بیشتر از بقیه است، را شناسایی کرد.

همچنین خوشه‌بندی فازی بر روی عوامل انتخاب شده اجرا شود تا درجه وابستگی هر کدام از این عوامل نسبت به هم مشخص گردد. در نهایت، با استفاده از روش انتخاب ویژگی، از میان ویژگی‌هایی که در افزایش کارایی دانشجویان تأثیرگذار است، ویژگی‌های انتخاب شده و سپس چهار روش داده‌کاوی مبتنی بر دقت مدل‌ها جهت پیش‌بینی عملکرد و موفقیت یادگیری مقایسه شدند و هدف نهایی دستیابی به مدلی با بالاترین میزان دقت است. شکل ۱ فلوچارت روش پیشنهادی را نشان می‌دهد.



شکل ۱- فلوچارت روش پیشنهادی

هدف از انجام این تحقیق این است که از نظر ما نتیجه این کار می‌تواند در دانشگاه‌ها و مراکز آموزشی بر روی افرادی که مشکل یادگیری دارند مطالعه شده و در این زمینه‌ها مشاوره لازم به دانشجویان و افراد ارائه شود. هر چند که این مشکلات مرتفع گردند، افراد دانشجو بیشتر مطالب را یاد می‌گیرند که این امر باعث می‌شود ایده‌های جدیدی به ذهن آنها خطور کند که پیشرفت و آبادانی جامعه و کشور را به همراه دارد.

در این تحقیق با ارائه مدل‌هایی تا حد امکان دقیق و قابل اعتماد پاسخی مناسب به پرسش تحقیق داده می‌شود. در جهت پیش بینی عملکرد، افزایش کارایی و موفقیت یادگیری دانشجویان، تکنیک‌های خوشه‌بندی، کشف قوانین همبستگی، کلاس‌بندی و تشخیص ناهمگونی بر روی نمرات دانشکده کامپیوتر و فناوری اطلاعات دانشگاه امیرکبیر اعمال گردید.

روش تحقیق استفاده شده در این پژوهش کتابخانه‌ای و پرسشنامه‌ای است. برای انتخاب ویژگی[۴] از نرم‌افزار Matlab؛ برای مدل‌سازی داده‌ها و دسته‌بندی دانشجویان از متدولوژی کریسپ استفاده شده، و برای مصورسازی نتایج از نرم‌افزار Clementine 12 استفاده می‌شود که دارای قابلیت‌هایی متناسب با تحقیق است.

۳- نتایج و بحث

در این پژوهش به پیش‌بینی نمرات نهایی (معدل) دانشجویان یکی از دانشگاه‌های کشور، پرداخته شده است. در این تحلیل با استفاده از روش انتخاب ویژگی، از میان ویژگی‌های تأثیرگذار در وضعیت تحصیلی دانشجویان چهار ویژگی انتخاب می‌شود. همچنین در ادامه با استفاده از چهار روش خوشه‌بندی، کشف قوانین وابستگی، رده‌بندی و تشخیص ناهمگونی و متدولوژی کریسپ رفتار دانشجویان در یک محیط آموزشی بررسی می‌گردد.

۳-۱- انتخاب ویژگی

هدف از انتخاب ویژگی انتخاب یک زیر مجموعه از ویژگی‌ها برای افزایش دقت پیش‌گویی است. انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مسأله در بسیاری از کاربردها مانند طبقه‌بندی اهمیت به سزایی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آنها یا بلااستفاده‌اند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها

مشکلی از لحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد مورد نظر بالا می‌برد.

انتخاب ویژگی، زیر مجموعه‌ای از تخمین زننده‌ها (متغیرها یا ویژگی‌ها) را از یک لیست بزرگی از تخمین زننده‌های کاندید، بدون فرض ارتباط بین آنها و اینکه متغیرهای وابسته یا نتیجه مورد علاقه، خطی یا یکنواخت هستند، انتخاب می‌نماید. زیرا تمام ویژگی‌ها در ساخت خوشه‌ها مفید نیستند. در ادامه پس از انتخاب چندین ویژگی، باید خوشه‌بندی انجام شود تا بتوان به وسیله آن، نمرات و وضعیت درسی دانشجویان را ارزیابی نمود. یکی از معیارهای شناخته شده برای انتخاب ویژگی در آمار، مربع میانگین باقیمانده[۵] پیش‌بینی است، که برای مدلی با p متغیر به صورت رابطه (۱) تعریف می‌شود:

$$\mathrm{RMS_p} = \frac{\mathrm{SSE_p}}{\mathrm{n-p}} \qquad (۱)$$

که SSE مجموع باقیمانده مربع خطاها[۶] (انحراف معیار یک توزیع نمونه برداری آماری که تخمین انحراف معیار از نمونه های مجموعه آموزشی می باشد.)، و n تعداد نقاط داده است، که در مقایسه دو مدل، مدلی با کمترین RMS انتخاب می‌شود [۶] پس از نوشتن کد در نرم‌افزار Matlab برای ۱۱ ویژگی، روش انتخاب ویژگی در چند بار اجرای مختلف، از میان ویژگی‌ها، ویژگی‌های مشترک سن، وضعیت تاهل، شاغل بودن و قدرت محاسبات ریاضی را انتخاب کرده است.

گام‌های الگوریتم پیشنهادی انتخاب ویژگی شامل:

گام اول: در این قسمت میزان اهمیت هر ویژگی نسبت به سایر ویژگی‌ها را مشخص می‌کنیم. برای انجام این کار ابتدا از تجربه‌های افراد خبره در این زمینه استفاده کرده-ایم و در ادامه به کمک روش انتخاب ویژگی با استفاده از نرم‌افزار Matlab ویژگی‌های مؤثرتر انتخاب می‌شوند. بدیهی است اگر ضریب یک ویژگی صفر در نظر گرفته شود، آنگاه احتمال انتخاب آن بسیار کمتر خواهد بود.

```
1. b = [1 1 2 1 1 1 0.7 1 0.5 0.2 2];
```

گام دوم: در این بخش دیتاستی از مقادیر به دست آمده از ویژگی‌ها که شامل ۱۴۶۹ داده برای هر ویژگی است، در متغیری به نام X قرار می‌گیرد، سپس میزان اهمیت هر ویژگی بر روی داده‌ها اعمال می‌شود.

```
2. X = [X1; X2; X3; X4; X5; X6; X7; X8; X9;
X10; X11];
3. Z = X';
4. Zb = Z * b';
```

گام سوم: در این قسمت مقادیر اولیه را برای اجرای الگوریتم مشخص می‌کنیم. این مقادیر شامل مشخص کردن تابع توزیع داده‌ها، تعداد داده‌ها و تولید کردن اعداد تصادفی با استفاده از توزیع دو جمله‌ای برای استفاده در گام چهارم جهت اعمال تابع لجستیک است:

```
5. p = 1. / (1+ exp (-Zb));
6. N = 1469;
7. y = binornd (N,p);
```

گام چهارم: در این بخش توزیع لجستیک بر روی داده‌ها اعمال می‌شود، این توزیع ارزش‌ها را بین صفر و یک قرار می‌دهد. برای این منظور از تابع gmlfit که مدل تعمیم خطی است، استفاده می-شود. لازم به ذکر است که model0در این قسمت ضریب نمایش و خطاهای استاندارد را نشان می‌دهد:

```
%fits a logistic model to the data
8. Y = [y N*ones(size(y))];
9. [b0, dev0, stats0] = glmfit (Z, Y,
'binomial');
% Display Coefficient estimates and their
standard errors
10. model0 = [b0 stats0.se];
```

گام پنجم: در این قسمت انحراف به دست آمده از گام چهارم نشان داده می‌شود:

```
11. dev0;
```

گام ششم: تنظیمات مربوط به اجرای الگوریتم انتخاب ویژگی را در این مرحله مشخص می‌شود. در کد زیر Displayبیانگر مقدار اطلاعات نمایش داده شده توسط الگوریتم است، که گام به گام تعیین شده است. Tolfun بیانگر تلرانس خاتمه برای مقدار تابع هدف است، که در این روش به دلیل جلو رونده بودن الگوریتم این مقدار برابر 1e-6 است و TolfunType تعیین کننده نوع استفاده تلرانس تابع هدف است که یکی از حالات مطلق یا نسبی باید انتخاب شود که حالت مطلق انتخاب شده است:

```
12. maxdev = chi2inv(.95,1);
```

```
13. opt = statset ( 'display', 'iter', ... 'TolFun',
maxdev, ... 'TolTypeFun', 'abs' );
```

گام هفتم :روش پیشنهادی برای انتخاب ویژگی، انتخاب ترتیبی جلو رونده است. نوع تابع ارزیابی استفاده شده در این روش معیارهای مبتنی بر خطای طبقه‌بندی کننده بوده است. در این کد cv بیانگر روش اعتبارسنجی برای محاسبه معیار ارزیابی نسبت به هر زیر مجموعه‌ای از ویژگی‌های کاندید است، که با انتساب none مشخص شده و مجموعه آموزشی و تست در این روش وجود ندارد. سپس تنظیمات بیان شده در گام ششم اعمال و در نهایت، جهت شروع الگوریتم را مشخص می‌شود.

در این روش الگوریتم انتخاب ویژگی کار خود را بدون ویژگی آغاز می‌کند و با روش جلو رونده در هر تکرار سعی در انتخاب ویژگی‌های مؤثر را دارد. هدف اصلی در حالت جلو رونده، کاهش نیافتن معیار ارزیابی است که هر بار تکرار الگوریتم و انتخاب ویژگی جدید باعث کاهش این معیار می‌شود. تکرار الگوریتم تا زمانی که معیار ارزیابی از معیار به دست آمده در گام پنجم بزرگتر باشد، ادامه می‌یابد:

```
14. inmodel = sequentialfs ( @critfun, Z, Y, ...
'cv', 'none',
'nullmodel', true, ...
'options', opt, ...
'direction', 'forward' );
% Display Coefficient estimates and their
standard errors
15. [b, dev, stats] = glmfit (Z ( :, inmodel), Y,
'binomial');
16. model = [b stats.se]
```

بعد از شناسایی ویژگی‌های تأثیرگذار در یادگیری دانشجویان، داده‌های به دست آمده برای هر ویژگی در نرم‌افزار Matlab وارد می‌شود و سپس الگوریتم برای به دست آوردن ویژگی‌های مهم‌تر اجرا می‌شود.

برای یازده ویژگی ذکر شده، روش انتخاب ویژگی در چند بار اجرای مختلف برای ما ویژگی‌های مشترک سن، وضعیت تأهل، شاغل بودن و قدرت محاسباتی را انتخاب می‌کند که این ویژگی‌ها می‌توانند در یادگیری افراد مؤثر واقع شوند.

۲-۳- خوشه‌بندی فازی

فازی نظریه‌ای است برای اقدام در شرایط عدم اطمینان، این نظریه قادر است بسیاری از مفاهیم و متغیرها و سیستم‌هایی را که نادقیق هستند، صورت‌بندی ریاضی ببخشد و زمینه را برای استدلال، استنتاج، کنترل و تصمیم‌گیری در شرایط عدم اطمینان فراهم آورد.

در مرحله آخر از فاز اول باید بر روی ویژگی‌های به دست آورده از مرحله انتخاب ویژگی، به دلیل عدم قطعیت ویژگی‌ها، خوشه‌بندی فازی بر روی آنها انجام داد. زیرا ممکن است برخی از ویژگی‌ها نسبت به ویژگی‌های دیگر در هدف تحقیق تأثیر کمتری داشته باشند. هدف از انجام خوشه‌بندی فازی، مشخص کردن درجه وابستگی ویژگی‌های انتخاب شده نسبت به یکدیگر است. با مشخص شدن این وابستگی می‌توان سایر ویژگی‌های تأثیرگذار در یادگیری دانشجویان را نیز شناسایی و با بیشتر شدن درجه ارتباط این ویژگی‌ها نسبت به هم می‌توان دانشجویانی که دارای این ویژگی‌ها نیستند را در راستای تحصیل بهتر مشاوره نمود.

یکی از مهمترین و پرکاربردترین الگوریتم‌های خوشه‌بندی، الگوریتم C میانگین است. در این الگوریتم نمونه‌ها به C خوشه تقسیم می‌شوند و تعداد C از قبل مشخص شده است. در نسخه فازی این الگوریتم نیز تعداد خوشه‌ها از قبل مشخص شده است که تابع هدف آن نیز به صورت رابطه (۲)است [۱۲ و ۱۳ و ۱۴]:

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}{}^{m} d_{ik}{}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}{}^{m} \| x_k - v_i \|^2 \qquad (۲)$$

در فرمول فوق m یک عدد حقیقی بزرگتر از ۱ است که در اکثر موارد برای خوشه‌بندی فازی داده‌ها برای m عدد ۲ انتخاب می‌شود. x_k نمونه kام و v_i نماینده یا مرکز خوشه iام و n تعداد نمونه‌هاست. u_{ik} میزان تعلق نمونه iام در خوشه kام را نشان می‌دهد. علامت || میزان تشابه (فاصله) نمونه با (از) مرکز خوشه است، که می‌توان از هر تابعی که بیانگر تشابه نمونه و مرکز خوشه باشد، استفاده کرد. از روی u_{ik} می‌توان یک ماتریس U تعریف کرد که دارای c سطر و n ستون است و

مؤلفه‌های آنها مقداری بین ۰ تا ۱ را می‌توانند اختیار کنند. مجموع مؤلفه‌های هر یک از ستون‌ها باید برابر ۱ باشد که در رابطه (۳) ارائه شده است [۱۳]:

$$\sum_{i=1}^{c} u_{ik} = 1, \ \forall k = 1, \dots, n \qquad (۳)$$

برای به دست آوردن فرمول‌های مربوط به u_{ik} و v_i باید تابع هدف تعریف شده را مینیمم کرد. با استفاده از شرط فوق و برابر صفر قرار دادن مشتق تابع هدف، رابطه (۴) را خواهد بود[۱۳]:

$$v_i = \frac{\sum_{k=1}^{n} {u_{ik}}^m x_k}{\sum_{k=1}^{n} {u_{ik}}^m}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/m-1}} \qquad (۴)$$

با استفاده از دو فرمول محاسبه شده مراحل الگوریتم خوشه‌بندی C میانگین فازی به صورت زیر است:

۱. مقدار دهی اولیه برای c، m و U0. خوشه‌های اولیه حدس زده شود.
۲. مراکز خوشه‌ها محاسبه شود (محاسبه v_iها).
۳. محاسبه ماتریس تعلق از روی خوشه‌های محاسبه شده در ۲.
۴. اگر $\|Ul+1-Ul\| \leq \varepsilon$ الگوریتم خاتمه می‌یابد و در غیر این صورت رفتن به مرحله ۲.

گام‌های اجرای الگوریتم خوشه‌بندی فازی بر روی ویژگی‌های به دست آمده از مرحله انتخاب ویژگی به شرح زیر است:

گام اول: ویژگی‌ها جهت استفاده در خوشه‌بندی و قرار گرفتن در بازه [0,1] نرمال‌سازی می‌شود. برای نرمال‌سازی از رابطه زیر استفاده می‌شود.

(X – min) / (max – min)

که در این فرمول X نمونه داده، max بزرگترین عضو مجموعه داده و min کوچکترین عضو مجموعه داده است:

```
% Feature Value Normalized
1. NX2 = ( X2 – min(X2) ) / ( max(X2) – min(X2) );
2. NX10 = ( X10 – min(X10) ) / ( max(X10) – min(X10) );
```

گام دوم: ویژگی‌های به دست آمده از مرحله انتخاب ویژگی، دو به دو در یک مجموعه قرار می‌گیرد:

```
% input features
3. X = [NX2; NX10];
4. Z = X';
5. Plot ( Z( :, 1), Z( :, 2), '0');
```

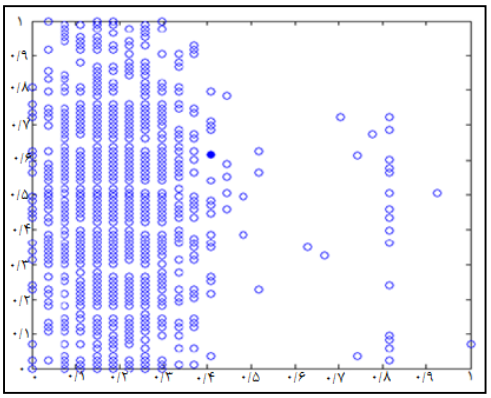
گام سوم: برای مشخص شدن ماتریس U و مراکز اولیه خوشه‌ها عملیات زیر انجام می‌شود:

```
% fuzzy c-means clustering
6. [center, U, obj_fcn] = fcm(Z, 2);
7. figure
8. Plot (obj_fcn)
9. title ('Objective Function Values')
10. xlabel ('Iteration Count')
11. ylabel ('Objective Function Values')
```

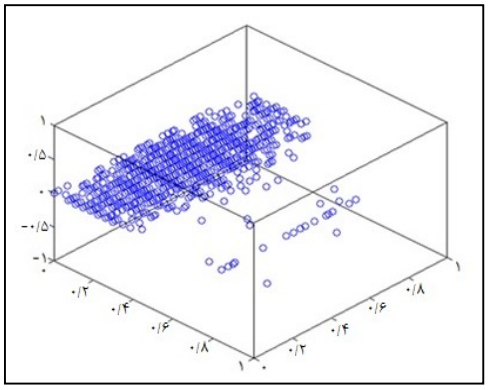
گام چهارم: در هر بار اجرای خوشه‌بندی فازی داده‌ها نسبت تعلق هر داده نسبت به مرکز خوشه‌ها محاسبه می‌شود و در صورتی که $\|Ul+1-Ul\| \leq \varepsilon$ صادق باشد، الگوریتم خاتمه پیدا می‌کند:

```
12. maxU = max(U);
13. index1 = find (U (1, : ) == maxU);
14. index2 = find (U (2, : ) == maxU);
% Clusters Style & Color
15. figure
16. line ( Z (index1, 1), Z (index1, 2),
'linestyle', 'none', 'marker', '0', 'color', 'g' );
17. line ( Z (index2, 1), Z (index2, 2),
'linestyle', 'none', 'marker', 'x', 'color', 'r' );
% Clusters Center
18. hold on
19. plot ( center (1,1), center(1,2), 'ko',
'markersize', 15, 'LineWidth', 2)
20. plot ( center (2,1), center(2,2), 'kx',
'markersize', 15, 'LineWidth', 2)
```

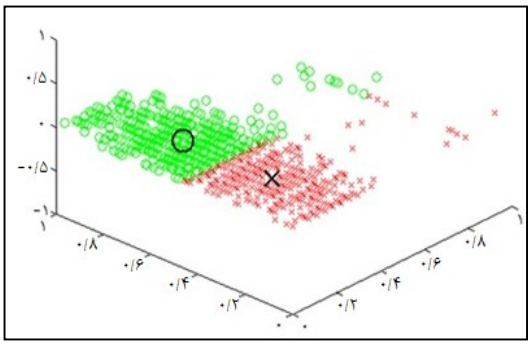
بعد از به دست آوردن ویژگی‌ها در بخش قبلی آنها را برای خوشه‌بندی فازی وارد نرم‌افزار Matlab می‌شود. در این بخش ورود اطلاعات داده‌های هر ویژگی انتخاب شده نیز مورد نیاز است، بعد از اجرای خوشه‌بندی نتایج حاصل در شکل‌های ۲ و ۳ و ۴ مشاهده می‌شود.



شکل ۲- نمای یک بعدی داده‌های توزیع شده برای خوشه‌بندی



شکل ۳- نمای سه بعدی داده‌های توزیع شده برای خوشه‌بندی



شکل ۴- نمای سه بعدی خوشه‌بندی فازی داده‌ها

برای مثال همان طور که در شکل ۲ مشاهده می‌شود، داده‌هایی که به صورت توپر نشان داده شده است، مطابق محور عمودی ۰/۶ متعلق به یک خوشه و مطابق محور افقی ۰/۴ متعلق به خوشه دیگری است. شکل ۵ مورد (از ۱۴۶۹ مورد) از میزان تعلق هر نمونه از داده‌ها نسبت به خوشه‌های موجود آمده است که همان طور که مشاهده می‌شود، مجموع تعلق هر نمونه به c خوشه برابر ۱است.

Columns 1 through 11

0.8112	0.8673	0.8387	0.1164	0.8782	0.0056	0.6171	0.0525	0.8552	0.0273	0.7354
0.1888	0.1327	0.1613	0.8836	0.1218	0.9944	0.3829	0.9475	0.1448	0.9727	0.2646

Columns 12 through 22

0.9917	0.8470	0.8303	0.1568	0.8668	0.0476	0.5564	0.9425	0.8978	0.6454	0.5564
0.0083	0.1530	0.1697	0.8432	0.1332	0.9524	0.4436	0.0575	0.1022	0.3546	0.4436

Columns 23 through 33

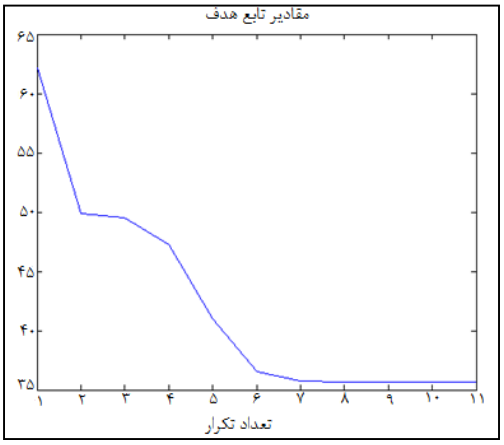
0.1859	0.7950	0.0024	0.0970	0.9207	0.9054	0.3696	0.8100	0.9524	0.9322	0.0024
0.8141	0.2050	0.9976	0.9030	0.0793	0.0946	0.6304	0.1900	0.0476	0.0678	0.9976

Columns 34 through 44

0.9061	0.4924	0.2785	0.8303	0.1315	0.0482	0.6760	0.9286	0.2406	0.0147	0.9340
0.0939	0.5076	0.7215	0.1697	0.8685	0.9518	0.3240	0.0714	0.7594	0.9853	0.0660

شکل ۵- میزان تعلق داده‌ها در خوشه‌بندی فازی

شکل ۶ نیز کاهش مقدار تابع هدف با هر بار تکرار خوشه‌بندی را نشان می‌دهد. این تکرار تا زمانی که شرط $\varepsilon \leq ||U_{l+1}-U_l||$ صدق نکند، ادامه می‌یابد.



شکل ۶- مقادیر تابع هدف با هر بار تکرار خوشه‌بندی

۳-۳- داده‌های ثبت نام دانشجویان

پیش پردازش داده‌ها به تنهایی ۶۰ درصد از حجم کار داده‌کاوی است. مرحله پیش پردازش شامل پاک‌سازی داده‌ها و تلفیق جداول به منظور استخراج داده‌های موردنظر و ایجاد جداول خلاصه به صورت مفصلی انجام شد. ناسازگاری‌ها و مقادیر ذکر نشده زیادی وجود داشت، که به نحو مناسبی حل و فصل گردید. به عنوان مثال لازم بود، جدول داده‌های اصلی با جدول تعداد واحد درس تلفیق شود، تا بتوان به میزان اهمیت دروس پی برد و در صورت نیاز موارد کم اهمیت‌تر که باعث به ثمر نرسیدن تحقیق می‌شدند حذف گردید. گاهی لازم بود که جداول خلاصه ایجاد شوند و گرفتن گزارش‌های گرافیکی سراسری به کار روند که جداول میانگین از این دسته‌اند.

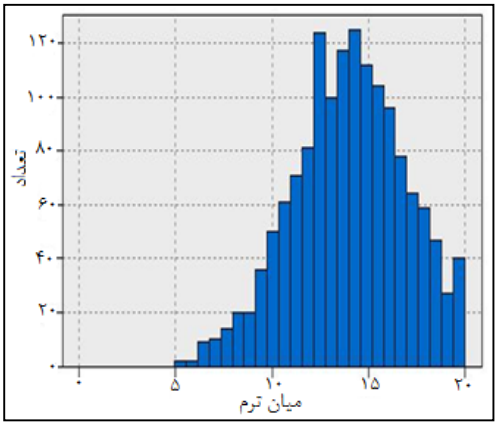
از آنجایی که در محیط‌های آموزشی چه در کشور ایران و چه در سایر کشورها شیوه ارزیابی دانشجویان نسبت

به همدیگر، بر مبنای نمره است که آنها در یک درس به دست می‌آوردند، بنابراین برای این مطالعه نمرات افرادی مورد نیاز است که یک درس را در طول چند سال گذرانده.

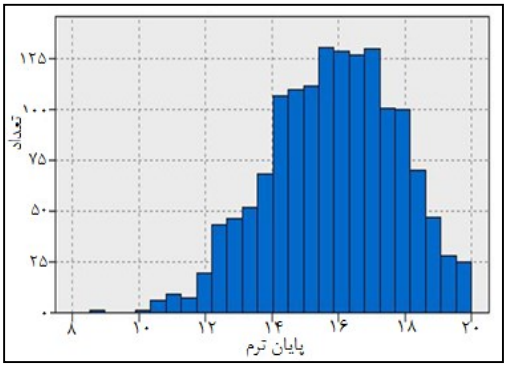
پارامترهایی که برای این بخش از تحقیق در نظر گرفته شده‌است، نمرات میان‌ترم، پایان‌ترم و نمره نهایی افراد است که با استفاده از نرم‌افزارهای خاص در این زمینه مطالعه بر روی این نمرات با چند روش مختلف صورت گرفته و نتایج نیز با هم بررسی می‌شوند.

داده‌های ثبت نام دانشجویان حدود سه هزار رکورد است که نشان می‌دهد یک دانشجو هر درس را با کدام استاد و با چه نمره‌ای به اتمام رسانده است. آماره‌های خلاصه سه متغیر اصلی مقاله، در جدول ۱ نشان داده شده است. در این فاز تکنیک‌های داده‌کاوی بر روی نمرات دانشجویان در راستای پیش‌بینی عملکرد، گروه‌بندی و کشف قوانین با استفاده از نرم‌افزار Clementine اجرا شد و نتایج به دست آمده مورد ارزیابی قرار گرفت.

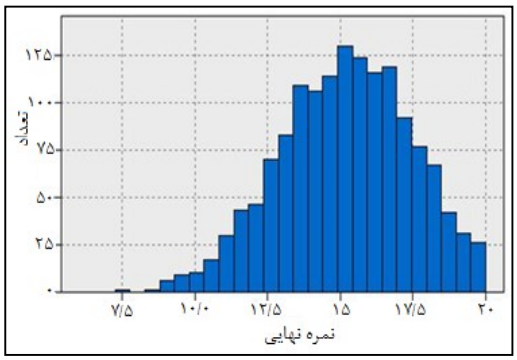
برای این منظور ابتدا داده‌ها مورد بررسی اولیه قرار گرفت. اشکال ۷ و ۸ و ۹ نمودارهای توزیع و هیستوگرام این داده‌ها را بیان می‌کنند:



شکل ۷- نمودار هیستوگرام نمرات میان‌ترم دانشجویان (خروجی نرم‌افزار Clementine)



شکل ۸- نمودار هیستوگرام نمرات پایان ترم دانشجویان (خروجی نرم‌افزار Clementine)



شکل ۹- نمودار هیستوگرام نمرات نهایی دانشجویان (خروجی نرم‌افزار Clementine)

۴-۳- خوشه‌بندی به روش K-means

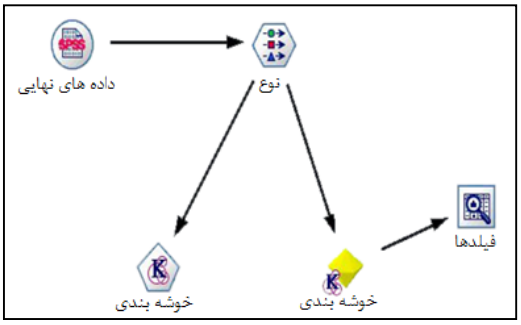
برای این الگوریتم اشکال مختلفی بیان شده، ولی همه آن‌ها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها، سعی در تخمین موارد زیر دارند: به دست آوردن نقاطی به عنوان مراکز خوشه‌ها و نسبت دادن هر نمونه داده به یک خوشه که کمترین فاصله تا مرکز آن خوشه را دارا باشد.

با استفاده از نرم‌افزار Clementine شرایط لازم برای خوشه بندی را مهیا و تنظیمات لازم اعمال می‌شود.

جدول ۱- آماره‌های خلاصه سه متغیر اصلی

	فیلد	نمره کمتر	نمره بیشتر	میانگین	واریانس	انحراف معیار	چولگی	تعداد داده
۱	نمره میان ترم	۴/۹۳۰	۲۰/۰۰۰	۱۳/۹۸۶	۸/۴۷۳	۲/۹۱۱	-۰/۱۸۷	۱۴۶۹
۲	نمره پایان ترم	۸/۵۲۰	۲۰/۰۰۰	۱۵/۸۵۸	۳/۶۹۳	۱/۹۲۲	-۰/۲۵۵	۱۴۶۹
۳	نمره نهایی	۷/۲۶۰	۲۰/۰۰۰	۱۵/۱۷۴	۵/۰۷۰	۲/۲۵۲	-۰/۲۲۳	۱۴۶۹

در شکل ۱۰ می‌توان مشاهده نمود، که تعداد خوشه‌ها مشخص شده و مجموعه داده برای خوشه بندی را می‌توان مسیردهی نمود.



شکل ۱۰- اجرای خوشه بندی k-means

خوشه‌بندی با استفاده از تکنیک k-mean به پنج گروه تقسیم‌بندی شده که در شکل ۱۱ نشان داده شده است.

	خوشه ۳	خوشه ۵	خوشه ۱	خوشه ۲	خوشه ۴	درصد خوشه ها نسبت به هم
نمره پایان ترم	۱۶/۸۶ (۰/۴۵)	۱۵/۳۵ (۰/۴۵)	۱۳/۸۷ (۰/۴۵)	۱۸/۵۳ (۰/۶۱)	۱۲/۱۰ (۰/۷۶)	۱۵/۸۶ (۱/۹۲)
نمره میان ترم	۱۵/۴۰ (۰/۷۰)	۱۳/۱۸ (۰/۶۴)	۱۱/۰۶ (۰/۶۸)	۱۸/۱۵ (۰/۹۴)	۸/۳۱ (۱/۱۴)	۱۳/۹۹ (۲/۹۱)
نمره نهایی	۱۶/۳۱ (۰/۵۳)	۱۴/۵۷ (۰/۵۱)	۱۲/۸۷ (۰/۵۲)	۱۸/۳۴ (۰/۷۳)	۱۰/۷۸ (۰/۸۸)	۱۵/۱۷ (۲/۲۵)

شکل ۱۱- خوشه‌بندی دانشجویان

در هر خوشه ناحیه رنگی در هر دایره درصد تعداد اعضای هر خوشه نسبت به کل اعضا را نشان می‌دهد. در هر سطر میانگین نمرات هر خوشه به عنوان مرکز آن تعیین شده و انحراف از معیار آنها مشخص است. با توجه به این تقسیم‌بندی می‌توان دانشجویان را بر اساس نمراتشان خوشه‌بندی و برحسب توانایی‌هایشان در زمینه یادگیری بهتر راهنمایی نمود.

جدول ۲ نیز مشخصات هر خوشه در خوشه‌بندی نمره نهایی دانشجویان را ارائه می‌کند.

۵-۳- روش کشف قوانین وابستگی

قوانین وابستگی، ارتباطات جالب و پنهان را مابین خصایص یک مجموعه داده نشان می‌دهند. این قوانین می‌توانند، آشکار کنند که یادگیرندگان تمایل دارند، کدام محتواها را با هم دسترسی داشته باشند، یا چه ترکیبی از ابزارها را به کار می‌برند. جهت استفاده از این روش، متغیر هدف به صورت گسسته تبدیل می‌شود: معدل کمتر از ۱۲، معدل ۱۲ تا ۱۴، معدل ۱۴ تا ۱۵/۵، معدل ۱۵/۵ تا ۱۷، معدل ۱۷ تا ۱۸/۵، معدل ۱۸/۵ تا ۲۰.

با توجه به متغیر هدف تحقیق که به صورت گسسته در نظر گرفته شده است، خروجی نرم‌افزار در جدول ۳ مثلا در اولین خط نشان می‌دهد، که اگر نمره میان‌ترم بین ۱۴/۳۴ تا ۱۶/۳۴ باشد، آنگاه نمره نهایی در گروه ۴ (یعنی معدل ۱۵/۵ تا ۱۷) قرار خواهد داشت.

۶-۳- کلاس‌بندی به روش درخت تصمیم‌گیری

کلاس‌بندی یکی از عملیات موجود در داده کاوی است، که عضویت یک نمونه داده را در یکی از گروه‌های از قبل مشخص شده پیش‌بینی می‌کند. در داده‌کاوی آموزشی ممکن است با در نظر گرفتن مقدار تلاش یک دانشجو نمره نهایی او را بتوان پیش‌بینی نمود. کلاس‌بندی تکنیکی است که داده‌ها را به صورت کلاس‌های از پیش تعریف شده جدا می‌کند.

بنابراین کلاس‌بندی بر اساس خصوصیات موجود در داده‌ها انجام می‌شود. نتیجه این کلاس‌بندی توصیف داده‌های موجود و درک بهتر از هر کلاس در پایگاه داده است. در داده‌کاوی آموزشی کلاس‌بندی دانشجویان بر حسب نمرات میان ترم و پایان ترم آنها برای پیش‌گویی نمرات نهایی‌شان است. درخت تصمیم‌گیری در بین الگوریتم‌های کلاس‌بندی روش قدرتمندی است که محبوبیت آن با رشد نام دیگر درختان تصمیم است.

جدول ۲- مشخصات هر خوشه در خوشه‌بندی نمره نهایی دانشجویان

خوشه بندی	خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴	خوشه ۵
میانگین نمره	۱۲/۹۱۴	۱۸/۳۶۱	۱۶/۳۵۶	۱۰/۸۰۳	۱۴/۶۱۴
انحراف معیار	۰/۵۲۴	۰/۷۲۸	۰/۵۳۳	۰/۸۷۶	۰/۵۱۴

جدول۳- خروجی نرم‌افزار جهت کشف قوانین وابستگی با روش GRI

در صد اعتماد	درصد پشتیبان	قانون	گروه بندی نمرات نهایی
۱۰۰	۲۳/۵۵	۱۶/۳۴> نمره میان ترم و ۱۴/۳۴> نمره میان ترم	۴
۱۰۰	۲۳/۴۹	۱۶/۳۴> نمره میان ترم و ۱۶/۱۸> نمره پایان ترم	۴
۱۰۰	۲۱/۲۴	۹/۹۸ > نمره میان ترم و ۱۴/۸۶> نمره پایان ترم	۲
۱۰۰	۱۵/۳۸	۱۸/۳۷> نمره میان ترم و ۱۷/۴۲> نمره پایان ترم	۵
۱۰۰	۱۵/۳۸	۱۸/۳۷> نمره میان ترم و ۱۶/۳۴> نمره میان ترم	۵
۱۰۰	۸/۸۵	۹/۹۸> نمره میان ترم	۱
۱۰۰	۶/۵۴	۱۸/۳۷> نمره میان ترم	۶
۹۹/۰۵	۲۱/۴۴	۹/۹۸> نمره میان ترم و ۱۲/۴۶> نمره میان ترم	۲
۷۰/۵۹	۳۰/۰۹	۱۴/۸۶> نمره پایان ترم	۲
۷۰/۱۹	۲۱/۹۲	۱۶/۳۴> نمره میان ترم	۵
۷۰/۱۱	۳۰/۲۹	۱۲/۴۶> نمره میان ترم	۲

در این ارزیابی از درخت تصمیم C&R برای ارائه قوانین منطقی برای تعیین نمره نهایی دانشجویان استفاده شده است. در این درخت ورودی‌ها نمره‌های میان‌ترم، پایان-ترم، و هدف تعیین نمره نهایی دانشجو است. از این درخت می‌توان فهمید که نمره پایان‌ترم بیشترین تأثیر را در نمره نهایی دارد. به این دلیل که در دو سطح اول درخت، نمره پایان‌ترم تعیین‌کننده مسیر است.

در کنار گره ریشه، هر زیر درخت میانگین و انحراف از معیار نمره نهایی فرزندان آن زیر درخت نوشته شده است. به طور مثال Midterm <= 13.74 [Ave: 13.196, Effect: -1.978] بدین معنی است که زیر درخت سمت راست درخت اصلی، میانگین نمرات نهایی‌اش برابر ۱۳/۱۹۶ است و این نمرات به فاصله ۱/۹۷۸- قابل تغییر هستند. یعنی نمرات در محدوده [۱۵,۱۷۴، ۱۱,۲۱۸] قرار دارند. هرچه به برگ درخت نزدیک می‌شویم محدوده نمره نهایی کمتر می‌شود و همچنین هرچه تعداد داده‌های ورودی بیشتر باشد نتیجه پیش‌بینی دقیق‌تر خواهد بود.

```
Midterm <= 13.74 [Ave: 13.196, Effect: -1.978]
Midterm <= 10.80 [Ave: 11.469, Effect: -1.726]
Midterm <= 8.62 [Ave: 10.182, Effect: -1.288]
Midterm <= 7.00 [Ave: 9.246, Effect: -0.935] =>
    Midterm > 7.00 [Ave: 10.65, Effect: 0.468]
    Midterm <= 7.88 [Ave: 10.327, Effect: -0.323] =>
10.327
    Midterm > 7.88 [Ave: 10.865, Effect: 0.215] =>
10.865
    Midterm > 8.62 [Ave: 11.995, Effect: 0.526]
    Midterm <= 9.92 [Ave: 11.541, Effect: -0.454]
    Final <= 12.68 [Ave: 11.325, Effect: -0.216] =>
11.325
    Final > 12.68 [Ave: 11.726, Effect: 0.185] =>
11.726
    Continue…
```

۷-۳- تشخیص ناهمگونی

تکنیک تشخیص ناهمگونی، داده‌ها را جستجو می‌کند تا آنهایی را که خیلی متفاوت از بقیه هستند، پیدا کند. در داده‌کاوی آموزشی می‌توان از این تکنیک در جهت پیدا کردن دانشجویانی که مشکلات خاصی از قبیل یادگیری دارند استفاده نمود.

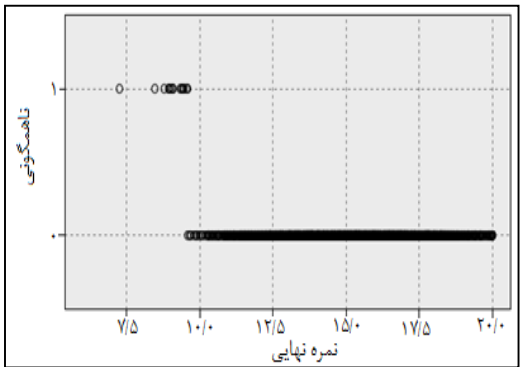
روش پیشنهادی در این زمینه با توجه به حجم داده‌ها، بهره گرفتن از الگوریتم k نزدیکترین همسایه است. این الگوریتم به دنبال k نمونه از نزدیکترین نمونه‌ها می‌گردد. (k نمونه مشابه) نزدیکی دو نمونه با به دست آوردن تشابه و یا فاصله میان این دو نمونه محاسبه می‌شود. شبه کد نشان داده شده بیانگر تشخیص ناهمگونی با استفاده از الگوریتم k نزدیکترین همسایه است:

```
1. Build the normal data set D;
2. for each process X in the data do
3. if X has an unknown system call then
4. X is abnormal;
5. else then
6. for each process in Di data do
7. calculate sim(X, Di);
8. if sim(X, Di) equals 1.0 then
9. X is normal; exit;
10. find k biggest scores of sim(X, D);
11. calculate sim_avg for k-nearest neighbors;
12. if sim_avg is greater than threshold then
13. X is normal;
14. else then
15. X is abnormal;
```

در این شبه کد، به ازای هر داده از مجموعه داده اگر داده‌ای برچسب ناهنجاری داشته باشد به خروجی فرستاده می‌شود در غیر این صورت میزان تشابه آن با مجموعه داده ساخته شده با عنوان D محاسبه می‌شود. اگر میزان تشابه محاسبه شده برابر با ۱ باشد آنگاه داده به

عنوان داده‌ای نرمال شناسایی شده و حلقه خاتمه می‌یابد. سپس به تعداد k تا بیشترین همسایگی مشابه، شناسایی شده و میانگین تشابهات برای k همسایه نزدیک محاسبه می‌گردد. اگر این میانگین بیشتر از حد آستانه باشد، آنگاه داده نرمال و در غیر این صورت غیر نرمال یا ناهنجار تشخیص داده می‌شود.

در مجموعه داده‌های این مقاله ۱۴ مورد انحراف و ناهمگونی تشخیص داده شده است. داده‌های ناهمگون نمره نهایی دانشجویان، که خروجی نرم‌افزار Clementine است، در شکل ۱۲ نشان داده شده است. با تشخیص این موارد مدرس می‌تواند راهکار مناسبی را در جهت رفع آنها به کار گیرد.



شکل ۱۲- داده‌های ناهمگون نمره نهایی

جهت ارزیابی و مقایسه روش‌های داده‌کاوی استفاده شده در این مقاله، دقت روش‌ها با استفاده از نرم‌افزار SPSS محاسبه شده‌اند. در آمار برای به دست آوردن دقت از فرمول $1-(\text{standard division})^2$ استفاده می‌شود. جدول ۴ دقت چهار روش استفاده شده که داده‌ها بر روی تک‌تک آنها مورد بررسی قرار گرفتند را نشان می‌دهد. بر مبنای الگوهای به دست آمده در این مقاله می‌توان هر دانشجو را از ابتدای ترم راهنمایی و با توجه به نمراتی که در طول ترم کسب می‌کند، او را از محدوده نمره نهایی خود آگاه کرد و بر طبق توانایی‌هایش برنامه‌ریزی مناسب تحصیلی نمود. این الگوها می‌توانند برای کارآمدتر ساختن فرآیند یادگیری در سیستم مؤثر باشند.

جدول ۴- مقایسه دقت چهار روش استفاده شده

روش مورد استفاده	دقت روش
روش K-Means	۰/۹۳۹
روش C&R	۰/۹۳۶
روش GRI	۰/۹۱۷
روش ناهمگونی	۰/۹۲۱

بنابراین روش K-means، بیشترین دقت و روش GRI کمترین دقت را در بین این چهار روش داشته است.

در اکثر مؤسسات و آموزشگاه‌های داخل و خارج از کشور ملاک اصلی برای ترفیع درجه دانشجویان نمره کسب شده توسط دانشجو است. حال با توجه به داده‌های جمع‌آوری شده از نمرات دانشجویان در ترم‌های مختلف تحصیلی می‌توان، با استفاده از روش‌هایی در جهت گروه‌بندی افراد نسبت به اهداف هر مؤسسه، عملکرد هر گروه را مورد بررسی قرار داد و راهکارهایی جهت افزایش کارایی در یادگیری و مشاوره ارائه نمود.

این تحقیق شامل دو فاز انتخاب ویژگی و خوشه‌بندی فازی بر روی ویژگی‌های انتخاب شده و همچنین اجرای تکنیک‌هایی از داده‌کاوی جهت پیش‌بینی عملکرد دانشجویان و شناسایی آنهایی که متفاوت از بقیه هستند، است. هدف از انجام خوشه‌بندی فازی، مشخص کردن درجه وابستگی ویژگی‌های انتخاب شده نسبت به یکدیگر است. با مشخص شدن این وابستگی می‌توان سایر ویژگی‌های تأثیرگذار در یادگیری دانشجویان را نیز شناسایی کرد. با بیشتر شدن درجه ارتباط این ویژگی‌ها نسبت به هم می‌توان دانشجویانی که دارای این ویژگی‌ها نیستند را در راستای تحصیل بهتر مشاوره نمود. در فاز اول به دلیل عدم وجود دیتاستی معتبر در رابطه با ویژگی‌های تأثیرگذار در یادگیری، با استفاده از روش‌های مختلف اعم از پرس‌وجو، چک لیست، مشاوره با افراد خبره در این زمینه و روش‌های دیگر، یازده ویژگی تأثیرگذار بر روی یادگیری دانشجویان را شناسایی کرد و با اجرای الگوریتم انتخاب ویژگی ترتیبی به صورت جلو رونده با استفاده از نرم‌افزار Matlab ویژگی‌های سن، وضعیت تحصیلی، شاغل بودن و قدرت محاسباتی، به عنوان ویژگی‌های تأثیرگذار انتخاب شدند. بعد از انتخاب ویژگی برای مشخص شدن درجه ارتباط این ویژگی‌ها خوشه‌بندی فازی C میانگین در نرم‌افزار

Matlabانجام گرفت و نتایج، نشان دهنده درجه ارتباط نزدیک در اکثر داده‌های مورد مطالعه است. نتایج حاصل از اجرای خوشه‌بندی تقسیم کردن دانشجویان در گروه‌های مختلف بر حسب کارایی آنهاست.

در فاز دوم نیز تکنیک‌هایی از داده‌کاوی بر روی نمرات دانشجویان با استفاده از نرم‌افزار Clementine صورت گرفت. این تکنیک‌ها در جهت گروه‌بندی دانشجویان از نظر نحوه عملکرد، شناسایی روابط جذاب همبستگی برای به دست آوردن قوانین موجود در میان نمرات دانشجویان در راستای مشخص شدن گروه‌های نمرات، کلاس‌بندی نمرات در جهت تعیین نمره نهایی و در نهایت برای شناسایی دانشجویانی که عملکرد آنها نسبت به سایرین متفاوت است، روش تشخیص ناهمگونی اجرا شد و با به دست آوردن دقت هر روش با استفاده از نرم-افزار Clementine و روش‌های آماری در SPSS، روش خوشه‌بندی K-means بیشترین دقت و روش GRI با کمترین دقت مشخص شدند.

با استفاده از تکنیک K-means دانشجویان بر اساس نمراتشان تقسیم‌بندی شدند و عملکرد آنها در طول ترم به صورت نمودار نشان داده شد. محققان با استفاده از درخت تصمیم C&R توانستند دانشجویان را با توجه به عملکردشان در طول ترم یعنی با توجه به ورودی‌هایی چون نمره‌های میان‌ترم، پایان‌ترم و تمرین و با هدف تعیین نمره نهایی، کلاس‌بندی و الگوهایی برای پیش-بینی نمره نهایی آنها ارائه کنند. از آنجا که خوشه‌بندی یک روش یادگیری بدون ناظر است، بیشتر به دنبال یافتن گروه‌هایی از دانشجویان است، که قبلا شناخته نشده‌اند و از قبل پیش‌بینی در مورد شباهت‌های موجود ندارند و تفسیر نتایج آن کمی مشکل است، ولی درخت تصمیم یک روش یادگیری با ناظر است که بر اساس یک برچسب کلاس معلوم، دانشجویان را به گروه‌هایی کلاس‌بندی می‌کند و به پیش‌گویی رفتار دانشجویان جدید بر اساس این کلاس‌بندی می‌پردازد. برای شخصی‌سازی هرچه بیشتر یک سیستم یادگیری الکترونیکی، بهتر آن است که با استفاده از تکنیک خوشه‌بندی به کشف گروه‌های جدید پرداخت و عملکرد دانشجویان را مورد ارزیابی قرار داد و با کمک درخت تصمیم رفتار دانشجویان جدید را پیش‌بینی نمود و روش یادگیری مناسب را به آنها نشان داد. به دست آوردن اطلاعات و دانش مفید و الگوهای غیر بدیهی از موضوعات کلیدی وابسته به یادگیری الکترونیکی در آموزش عالی ضروری است. فرآیند داده‌کاوی در این امر موفق بوده است.

۴- نتیجه‌گیری

بالا بودن رتبه دانشجویان یک مؤسسه سبب بالا رفتن سطح علمی با ارائه مقالات معتبر بین‌المللی توسط دانشجویان آن مؤسسه خواهد شد و در نهایت موجب افزایش سطح علمی آن مرز و بوم می‌شود. همه این اتفاقات با شناسایی عوامل مؤثر بر یادگیری، مشاوره در جهت افزایش کارایی و غیره امکان‌پذیر می‌شود. در این پژوهش به شناسایی عوامل مؤثر بر یادگیری و ارائه الگویی جهت پیش‌بینی عملکرد و افزایش کارایی و موفقیت یادگیری دانشجویان در یک محیط آموزشی پرداخته شده است.

پی نوشت

[1] Data Mining
[2] Pattern Recognition
[3] Visual Modeling
[4] Feature Selection
[5] Residual Mean Square (RMS)
[6] Residual Sum of Square Errors (SSE)

مراجع

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

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

[3] Shafiepoor, F. Nazari, H.,"Designing an Adjusted Model for Evaluating Electronic Learning Strategies' Efficiency on Students' Academic Achievement", *jte.srttu.edu*, pp. 93-101, (2014). [In Persian]

[4] Romero, C. Ventura, S. and Garcia, E. "Data mining in course management systems: Moodle case study and tutorial", *Computers & Education*, Vol. 51, (2008), pp. 368-384.

[5] Han Binglan "Student Modeling and Adaptively in Web Based Learning Systems" Ms. c. Thesis, Massey University, New Zealand.

[6] Zarghami, E. and Azamati, S., "Considering the Desirability of Campus in Students Viewpoint", *Journal of technology of education*, pp. 287-296, (2013). [In Persian]

[7] Paulo Cortez, Alice Silva, "Using data minig to predict secondraycshool student performance", (2007).

[8] Romero, C., Espejo, G, Zafra, A, Romero and Ventura, J. R., "Web usage mining for predicting final marks of students that use Moodle courses. Computer Applications in Engineering Education". Doi 10.1002/ cae. 20456, (2010).

[9] Hung, J., & Zhang, K. "Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching". MERLOT Journal of Online Learning and Teaching, 2008.

[10] Félix Castro, Alfredo Vellido, ÀngelaNebot, and Francisco Mugica, "Applying Data Mining Techniques to e-Learning Problems".Studies in Computational Intelligence (SCI), (2007), pp. 183-221.

[11] Mehdi, S. "Feature Selection using combination of GA and ACO", Islamic Azad University of Tehran, (2009). [In Persian]

[12] Romero, S. Ventura, "Educational data mining: A survey from 1995 to (2005), in Expert Systems with Applications", (2007).

[13] http://ceit.aut.ac.ir/~shiry/lecture/machinelearning/tutorial/fuzzy%20clustering/introduction/introduction.htm [Accessed June 18, 2014].

[14] Seraji, F., Movahedi, R. M., and Siyahatkhah, "An Investigation of Iranian Virtual Universities Teachers' Skills in Teaching These Courses", *jte.srttu.edu*, pp. 25-37, (2015). [In Persian]

JTE.srttu.edu= journal of technology of education