کلیه حقوق مادی و معنوی این مقاله متعلق به دبیرخانه دائمی کنفرانس داده کاوی ایران است و درج آن فقط در سایت رسمی کنفرانس داده کاوی ایران مجاز بوده و هرگونه انتشار غیر از آن، خلاف عرف، اخلاق و قانون و مستوجب پیگیری است.

شناسایی عوامل موثر بر افت تحصیلی دانشجویان با استفاده از قوانین انجمنی و تحلیل خوشه‌بندی

(مطالعه موردی دانشگاه قم)

بهروز مینایی[۱]، سمیه سادات میرافضل[۲]، سیدحسن هانی[۳]

چکیده

بررسی وضعیت تحصیلی دانشجویان در دانشگاه‌های معتبر کشور در سال‌های اخیر بیانگر آن است که علیرغم توان مناسب تحصیلی دانشجویان و سطح کیفی دانشگاه‌ها، عده‌ای از آن‌ها به ویژه در مقطع کارشناسی در بدو ورود و در ادامه با افت تحصیلی مواجه می‌شوند که این پدیده در نهایت منجر به رکود جایگاه دانشجو می‌شود.

در این پژوهش با استخراج قوانین انجمنی و تحلیل خوشه‌بندی، دانشجویان سال اول دانشگاه قم مورد بررسی تحلیلی قرار گرفته و ارتباط و همبستگی وضعیت تحصیلی دانشجو قبل از ورود به دانشگاه و تأثیر آن بر افت تحصیلی وی بعد از ورود به دانشگاه بررسی شده است. نتایج این مطالعه، در قالب روابط و قوانین کشف شده به مدیران دانشگاه در خصوص شناخت و کنترل بهتر عوامل تأثیرگذار بر موفقیت تحصیلی دانشجویان و بهبود کیفیت برنامه‌ریزی آموزشی و یافتن راه‌کارهایی مناسب به منظور حفظ و تقویت دانشجویان موثر خواهد بود.

کلمات کلیدی

داده‌کاوی، تحلیل خوشه‌بندی، کشف قوانین انجمنی.

کنفرانس داده کاوی ایران

[۱] استادیار دانشگاه علم و صنعت، b_minaei@iust.ac.ir
[۲] دانشجوی کارشناسی ارشد مهندسی فناوری اطلاعات، Somaye_Mirafzal@yahoo.com
[۳] استاد دانشگاه قم، Hani@qom.ac.ir

Exploring the effective factors on the educational regression of the university students using association rules and clustering analysis
(Case Study of the University of Qom)

Behrouz Minaei[۱]

Somaye Sadat Mirafzal[۲]

Seyed Hasan Hani[۳]

Exploring the educational status of students from credible universities in Iran during recent years indicates that in spite of suitable learning capabilities of students and quality level of universities, some students especially upon entering B.A. level and also later face with educational regression and finally this event leads to the decline of the student's position.

Throughout this research by using association rules and clustering analysis, freshman students of the university of Qom are studied and the relation and coherence of the student's learning status before entering the university and its effect on his deterioration after admission is explored. The results of this study can help the university officials to recognize and control the effective factors on the educational success of the students and also improve the quality of programming and seek proper ways to maintain and encourage the educational status of students.

KEYWORDS
Data analysis, Clustering rules, Association rules.

۱. مقدمه

در حال حاضر در اکثر دانشگاه‌ها بانک‌های اطلاعاتی وسیعی از ویژگی‌های دانشجویان موجود است که حجم بالایی از اطلاعات مربوط به سوابق آموزشی، تحصیلی و ... را شامل می‌شود. متأسفانه با وجود انبوه داده‌های موجود در سیستم آموزش دانشگاه‌ها، هیچ‌گاه بررسی عمیق و جامعی برای استخراج اطلاعات و دانش نهفته از این داده‌ها انجام نمی‌شود. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزشی عالی در جهت ارتقاء و بهبود فرایندهای آموزشی نظیر برنامه‌ریزی ، ثبت نام، ارزیابی و مشاوره کمک شایانی کند و بدین ترتیب آن‌ها را در تصیمیم‌گیری بهتر و داشتن طرح پیشرفته‌تری در هدایت دانشجویان کمک کند. در نتیجه، این بهبود مزایای بسیاری از قبیل حداکثر کردن کارایی سیستم آموزشی، کاهش نرخ از دست دادن و حذف دانشجویان، افزایش نرخ گذر دانشجویان، افزایش موفقیت دانشجو، افزایش خروجی یادگیری دانشجو و کاهش هزینه فرایندهای سیستم آموزش عالی را به ارمغان می‌آورد. نرم‌افزارهای کامپیوتری به کار گرفته شده برای این منظور، غالبا فقط برای مکانیزه کردن وضع موجود و اجرای پرس و جوهای معمولی جوابگو هستند. در حالی‌که در عمق این حجم عظیم داده‌ها، الگوها و روابط بسیار جالبی به صورت پنهان باقی می‌ماند[۴،۷،۸،۹].

داده‌کاوی الگوهای قابل فهم، ناشناخته، معتبر و بدیع را از داده‌های آموزشی پایگاه داده‌های بزرگ استخراج می‌کند. الگوهای پنهان کشف شده و دسته‌بندی دانشجویان، سیستم‌های آموزش عالی را در تصمیم‌گیری بهتر و داشتن طرح پیشرفته‌تری در هدایت دانشجویان کمک می‌کند. در نتیجه می‌توان داشتن فرایندهای آموزشی مؤثرتر، کاراتر و دقیق‌تر را در سیستم‌های آموزش عالی در دانشگاه‌ها تسهیل کرد[۵].

داده‌کاوی یک تکنیک میان رشته‌ای برای اکتشاف این الگوها است. داده‌کاوی از اطلاعات پنهانی که برای برنامه‌ریزی‌های استراتژیک و طولانی مدت می‌تواند حیاتی باشد، پرده‌برداری می‌کند. در حقیقت تکنیک‌های مختلف داده‌کاوی مانند: شبکه‌های عصبی، درخت تصمیم، رگرسیون و... به پیدا کردن الگوها، دانش نهفته و خوشه‌بندی دانشجویان در داده‌های سیستم آموزش کمک می‌کنند[۶].

[۱] Assistant professor, Iran university of Science and Technology, b_minaei@iust.ac.ir
[۲] M.A. student of IT, Somaye_Mirafzal@yahoo.com
[۳] Instructor of the university of Qom, Hani@qom.ac.ir

داده‌کاوی فرایندی تحلیلی است که برای کاوش داده‌ها (معمولا حجم عظیمی از داده‌ها) صورت می‌گیرد و یافته‌ها با به‌کارگیری الگوهایی، احراز اعتبار می‌شوند. یکی از اهداف اصلی داده‌کاوی پیش‌بینی است. فرایند داده‌کاوی شامل سه مرحله می‌باشد: ۱. کاوش اولیه ۲. ساخت مدل یا شناسایی الگو با کمک احراز اعتبار/ تایید و ۳. بهره‌برداری[۱۸].

۱-۲-مرحله۱: کاوش

معمولا این مرحله با آماده‌سازی داده‌ها صورت می‌گیرد که ممکن است شامل پاک‌سازی داده‌ها، تبدیل داده‌ها و انتخاب زیرمجموعه‌هایی از رکوردها با حجم عظیمی از متغیرها (فیلدها) باشد. سپس با توجه به ماهیت مساله تحلیلی، این مرحله به مدل‌های پیش‌بینی ساده یا مدل‌های آماری و گرافیکی برای شناسایی متغیرهای مورد نظر و تعیین پیچیدگی مدل‌ها برای استفاده در مرحله بعدی نیاز دارد[۱۸].

۱-۳-مرحله ۲: ساخت و احراز اعتبار مدل

این مرحله به بررسی مدل‌های مختلف و گزینش بهترین مدل با توجه به کارآیی پیش‌بینی آن می‌پردازد. شاید این مرحله ساده به نظر برسد، اما اینطور نیست. تکنیک‌های متعددی برای رسیدن به این هدف توسعه یافتند. و "ارزیابی رقابتی مدل‌ها" نام گرفتند. بدین منظور مدل‌های مختلف برای مجموعه داده‌های یکسان به‌کار می‌روند تا کارآیی‌شان با هم مقایسه شود، سپس مدلی که بهترین کارآیی را داشته باشد، انتخاب می‌شود[۱۸].

۱-۴-مرحله ۳ : بهره‌برداری

آخرین مرحله مدلی را که در مرحله قبل انتخاب شده است، در داده‌های جدید به کار می‌گیرد تا پیش‌بینی‌های خروجی‌های مورد انتظار را تولید نماید. داده‌کاوی به عنوان ابزار مدیریت اطلاعات برای تصمیم‌گیری، عمومیت یافته است. اخیرا، توسعه تکنیک‌های تحلیلی جدید در این زمینه مورد توجه قرار گرفته است (مانند درخت تصمیم)، اما هنوز داده‌کاوی مبتنی بر اصول آماری نظیرExploratory Data Analysis (EDA) می‌باشد. با این وجود تفاوت عمده‌ای بین داده‌کاوی و EDA وجود دارد. داده‌کاوی بیشتر به برنامه‌های‌کاربردی گرایش دارد تا ماهیت اصلی پدیده، به عبارتی دیگر داده‌کاوی کمتر با شناسایی روابط بین متغیرها سر و کار دارد [۱۸].

۱-۴- تعریف خوشه‌بندی

خوشه‌بندی به معنای دسته‌بندی اعضای مجموعه‌ها بدون نظارت و دخالت است. در این روش، خوشه‌ها یا دسته‌ها از قبل تعیین شده نیستند و به عبارت دیگر، بر چسب خوشه‌ها در دسترس نیست[۱۲].
خوشه‌بندی، به یافتن ساختاری در درون یک مجموعه از داده‌های بدون بر چسب اطلاق می‌شود، خوشه به مجموعه‌ای از داده‌ها گفته می‌شود که به هم شباهت داشته باشند. در خوشه‌بندی سعی می‌شود داده‌ها به خوشه‌هایی تقسیم شوند که شباهت بین داده‌های درون هر خوشه، حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت، حداقل شود[۱۷].

در این پژوهش از نرم‌افزار Clementine۱۲,۰ ساخت شرکت SPSS استفاده شده است. این نرم‌افزار به علت داشتن تمامی تکنیک‌های مورد استفاده در داده‌کاوی، نرم‌افزار بسیار توانمند و کاربردی برای مدل‌سازی در این زمینه می باشد. نحوه کار بدین شکل است که ابتدا بایستی اطلاعات ورودی به مدل داده شود و سپس با اضافه شدن تکنیک‌ها به آن، مدل تکمیل شده و آماده دادن خروجی‌های متنوع می‌باشد. لازم به ذکر است که برای تکنیک‌ها نیز بایستی پارامترهایی را به عنوان ورودی در نظر گرفت که این پارامترها معادل تمامی یا بخشی از اطلاعات ورودی به مدل می‌باشد.
این پژوهش داده‌های دانشجویان دانشگاه قم را مورد مطالعه قرار می‌دهد و بر اساس روش‌های داده‌کاوی آنها را تجزیه و تحلیل می‌نماید. هدف آن کشف الگوهای پنهان در درون داده‌هاست که رفتار دانشجویان سال اول را مورد بررسی قرار می‌دهد.

کنفرانس داده کاوی ایران

۲. کارهای مرتبط

سیستم‌های آموزش عالی از طریق داده‌کاوی قادرند که اثربخشی سیستم‌های آموزشی را حداکثر کنند، پذیرش و مدیریت ثبت نام را بهبود دهند، نرخ حذف دانشجویان را حداقل کنند، نرخ گذر دانشجویان را ارتقاء دهد، موفقیت را افزایش دهند و هزینه فرایندهای سیستم را کاهش دهند. یک موسسه آموزشی از طریق داده کاوی قادر خواهد بود که مزیت رقابتی خود را افزایش دهد و به استانداردهایی بالاتری در سطح علمی برسد[۴].

Han و Kamber در سال ۲۰۰۰ نرم افزار داده‌کاویی را توصیف کردند که به کاربران اجازه تجزیه و تحلیل داده‌ها از ابعاد مختلف، دسته‌بندی آن‌ها و خلاصه روابط در طول فرایند داده‌کاوی را می‌دهد[۱۰].

Hijazi و Nagvi در سال ۲۰۰۶ با انتخاب یک نمونه ۳۰۰ تایی از دانشجویان (۲۲۵ مرد و ۷۵ زن) دانشگاه Punjab روی عملکرد دانشجویان مطالعاتی را انجام دادند با این فرضیه که نسبت حضور در کلاس، ساعت مطالعه به صورت روزانه بعد از اتمام دانشگاه، درآمد خانواده، سن مادر و سطح سواد مادر دانشجویان به میزان عملکرد دانشجو مرتبط می‌باشد[۱۰].

Khan در سال ۲۰۰۵ یک نمونه ۴۰۰ از دانشجویان (۲۰۰ مرد و ۲۰۰ زن) دانشگاه Aligarh Muslim، Aligarh را انتخاب کرده و تأثیر متغیرهای جمعیتی و شخصیتی بر کسب مدارج بالاتر علمی را مورد بررسی قرار دادند. نتیجه این بررسی این بود که دختران با وضعیت اقتصادی-اجتماعی بالاتر و پسران با وضعیت اقتصادی-اجتماعی پایین عملکرد دانشگاهی نسبتاً بالاتری داشته‌اند[۱۰].

Galit در سال ۲۰۰۷ داده‌های دانشجویی را جهت پیش‌بینی رفتار یادگیری آن‌ها مورد بررسی قرار داده تا در زمان مناسب به دانشجویانی که در معرض ریسک هستند هشدار داده شود[۱۰].

Al-Radaideh و همکاران در سال ۲۰۰۵ مدل درخت تصمیمی را برای پیش‌بینی نمره نهایی دانشجویان در درس C++ در دانشگاه‌های Yarmouk و Jordan مورد مطالعه قرار دادند و بیان کردند که مدل درخت تصمیم نسبت به سایر مدل‌های پیش‌بینی بهتر می‌باشد[۱۰].

Pandey و Pal عملکرد ۶۰ دانشجو را در سال ۲۰۱۱ برای پیدا کردن دانشجویان علاقه‌مند در کلاس آموزش زبان مورد بررسی قرار داده‌اند[۱۰].

Khan ، Ayesh ، Mustafa و Satar در سال ۲۰۱۰ با استفاده از الگوریتم‌های خوشه‌بندی فعالیت‌های یادگیری دانشجویان را شرح دادند که این اطلاعات تولید شده پس از اجرای تکنیک‌های داده‌کاوی برای اساتید و هم‌چنین خود دانشجویان مفید بود[۱۰].

Bary در سال ۲۰۰۷ مطالعه خود را بر روی تدریس خصوصی و پیامدهای آن متمرکز کرده و تصریح می‌کند درصد دانشجویانی که در هند از تدریس خصوصی استفاده می‌کنند نسبت به دانشجویان مالزی، سنگاپور، ژاپن، چین و سریلانکا بیشتر است. هم‌چنین مشاهده می‌شود که بهبود عملکرد تحصیلی به شدت با تدریس خصوصی و شرایط اقتصادی-اجتماعی بستگی دارد[۱۰].

Pal و Bhardwaj در سال ۲۰۱۱ عملکرد ۳۰۰ دانشجوی مهندسی کامپیوتر از پنج دانشگاه متفاوت را مورد مطالعه قرار داده و با استفاده از تکنیک رده‌بندی بیزین عواملی مانند محل زندگی دانشجو، معتاد بودن دانشجو، درآمد سالانه خانواده و وضعیت خانوادگی به شدت با عملکرد دانشجویان در ارتباط می‌باشد[۱۰].

A.Silva و P.Cortez در سال ۲۰۰۸ روی اطلاعات دانشجویان دوره متوسطه برای پیش‌بینی وضعیت آن‌ها در سیستم آموزش و پرورش کار کرده‌اند. عملکرد گذشته دانشجویان مانند اطلاعات اجتماعی و اقتصادی به شدت با عملکرد دانشجویان در دوره‌های بعدی مرتبط می‌باشد[۱۱].

Z.J.Kovacic در سال ۲۰۱۰ داده‌کاوی را برای شناسایی پیش‌بینی موفقیت‌آمیز ثبت نام دانشجویان مورد بررسی قرار داد. در این تحقیق از الگوریتم‌های CHAID و CART برای رده‌بندی دانشجویان موفق و ناموفق استفاده شد[۱۱].

کنفرانس داده کاوی ایران

M.Ramaswami و R.haskaran در سال ۲۰۱۰ وابستگی‌های متقابلی را بین متغیرهای عملکرد در مدارس آموزش و پرورش و عملکرد دانشجویان پیدا کرده‌اند ویژگی‌های مانند محل مدرسه، محل زندگی و نوع آموزش مهم‌ترین شاخص‌ها برای عملکرد بهتر دانشجو در آموزش عالی می‌باشد[۱۱].

N.Ghisoiu، M.resfelean و V.P.resfelean در سال ۲۰۰۸ نشان دادند که موفقیت دانشجویان به مسیر آموزش و پرورش آن‌ها و هم چنین تحصیلات دانشگاهی و دیگر تخصص‌ها بستگی دارد[۱۱].

Jing Luan در سال ۲۰۰۲، با به‌کارگیری الگوریتم‌های C۵٫۰، CART، شبکه عصبی و Two Step به خوشه‌بندی و پیش‌بینی دانشجویان ماندگار و غیر ماندگار پرداخته است. بدین منظور از اطلاعاتی نظیر جنسیت، نژاد، ساعات کاری برنامه‌ریزی شده، محل اقامت، تعداد کل واحدهای اخذ شده (واحدهای اصلی، تخصصی، پایه، کارآموزی) و معدل استفاده کرده‌اند[۱۴،۱۵].

در یک مطالعه توسط Ming Yang در سال ۲۰۰۶، از الگوریتم K-means با تعداد خوشه‌های مختلف از ۲ تا ۶ به منظور پیش‌بینی ماندگار دانشجویان سال اول استفاده شده است. داده‌های به‌کار گرفته شده در این تحقیق شامل داده‌های ثبت‌نام کلاسی دانشگاه تگزاس از پاییز ۲۰۰۰ تا پاییز ۲۰۰۴ است. از جمله متغیرهای مورد استفاده در این تحقیق می‌توان به جنسیت، نژاد، سن، سطح تحصیلی دانشجو، دانشکده و فاصله مکانی زندگی دانشجو تا دانشگاه اشاره کرد. بررسی‌های این مطالعه بیان کرد که دانشجویان با نمره بالا از امتحان ورودی دانشگاه، که واحدهای بیشتری را گذرانده‌اند با احتمال بیشتری در ترم آینده ثبت‌نام خواهند کرد و بالعکس[۱۶].

با توجه به اینکه داده‌کاوی در حوزه آموزش، حوزه تحقیقاتی کاملاً جدیدی می‌باشد، نیازمند آن هستیم که در این حوزه تحقیقات جهت‌دار، تخصصی و گسترده‌تری صورت بگیرد و شکاف دانشی در این حوزه تا حدی از میان برداشته شود تا به سطح موفقیتی مشابه سایر حوزه‌ها نظیر داده‌کاوی پزشکی و داده‌کاوی تجارت الکترونیک برسیم[۴] .

کنفرانس داده کاوی ایران

خلاصه‌ای از کارهای انجام شده در زمینه اعمال تکنیک‌های داده‌کاوی بر روی سیستم آموزش عالی را در جدول(۱) زیر مشاهده می‌کنید:

جدول ۱-کارهای انجام شده در زمینه تکنیک‌های داده کاوی بر روی سیستم آموزش عالی[۴]

محقق	سال	کاربرد	تکنیک داده کاوی	نوع مدل و الگوریتم
Song and Chissom	۱۹۹۳	پیش‌بینی ثبت نام	پیش بینی	شبکه‌های عصبی
Sanjeev and Zytkow	۱۹۹۵	درک ثبت نام	----------	----------
Ma et al	۲۰۰۰	انتخاب دانشجویان مناسب برای شرکت در کلاس‌های جبرانی	قواعد انجمنی	تابع امتیاز SBA و مبنای قواعد انجمنی
Jing Luan	۲۰۰۲	خوشه‌بندی و پیش‌بینی دانشجویان ماندگار و غیر ماندگاری	خوشه‌بندی و پیش‌بینی	Twostep،شبکه‌های عصبی، C۵,۰ و C&RT
Jing Luan	۲۰۰۴	شناخت انواع دانشجویان برای فهم بهتر و یا استفاده در رده‌بندی	خوشه‌بندی	Twostep وK-Means
Jing Luan	۲۰۰۴	برنامه‌ریزی تحصیلی-پیش‌بینی گذراندن دروس	رده‌بندی	شبکه‌های عصبی،C&RT وC۵,۰
Jing Luan	۲۰۰۴	پیش‌بینی تعهد و منفعت رسانی فارغ‌التحصیلان	رده‌بندی	----------
Erdogan And Timor	۲۰۰۵	رابطه میان نتایج امتحان ورودی دانشگاه و میزان موفقیت دانشجویان	خوشه‌بندی	K-Means
Akseova et al	۲۰۰۶	پیش‌بینی ثبت نام	رگرسیون	SVM
Ming Yang	۲۰۰۶	تحلیل ماندگاری دانشجویان در ترم‌های آتی	خوشه‌بندی	K-Means
Ming Yang	۲۰۰۶	پیش‌بینی ثبت نام در یک درس خاص	رگرسیون	رگرسیون پواسون
Schonbrunn and Hilbert	۲۰۰۷	بررسی ترکیب واحدهای انتخابی هر دانشجو برای زمان‌بندی مناسب واحدها و جلوگیری از تداخل واحدها در زمان‌بندی	قواعد انجمنی	----------
Jayanthi and Malik	۲۰۰۷	ارتقاء فرایند مشاوره دانشجویان	قواعد انجمنی و پیش‌بینی	مدل‌های بیزی و درخت تصمیم
Schonbrunn and Hilbert	۲۰۰۷	بررسی رابطه میان واحدهای انتخابی دانشجویان به منظور توسعه واحدهای جدید برای دوره کارشناسی و کارشناسی ارشد	تحلیل لینک	----------

۳. مدلهای مورد استفاده شده

درخت CHAID می‌تواند درختی تولید کند که در برخی موارد به صورت غیر دودویی عمل کند، یعنی یک گروه آن به سه زیر گروه و یا بیشتر شکسته شود. صفات پیش‌بینی کننده و هدف می‌توانند هم از نوع بازه و هم از نوع رده‌ای باشند[۱].

گره K-Means داده را به گروه‌های(یا خوشه‌های) مجزا خوشه‌بندی می‌کند. این روش تعداد خوشه‌های ثابتی را تعیین می‌کند، به طور تکراری رکوردها را به خوشه‌ها تخصیص می‌دهد، و مراکز خوشه‌ها را تنظیم می‌کند تا هنگامی که اصلاح بیشتر نتواند مدل را بهبود بخشد. در عوض تلاش برای پیش‌بینی یک خروجی، K-Means از یک فرایند به نام یادگیری بدون نظارت برای کشف الگوها در مجموعه از فیلدهای ورودی استفاده می‌کند. روش K-Means معمولا سریع‌ترین روش برای خوشه‌بندی داده‌های بزرگ است.[۱]

۴. پردازش و ورود داده‌ها به مدل

داده‌ها شامل اطلاعات ۱۶۰۶ نفر دانشجوی ورودی سال تحصیلی ۱۳۹۰ که اقدام به ثبت نام در دانشگاه کرده‌اند می‌باشد البته این داده‌ها دارای فیلدها زیر می‌باشد:

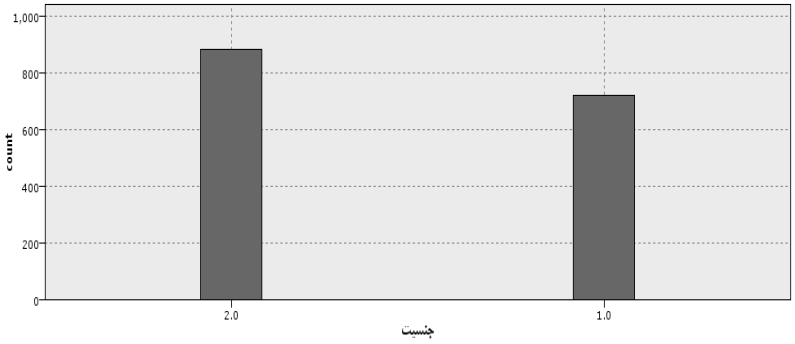
۱. داده‌های شناسنامه‌ای دانشجو مانند نام، نام خانوادگی، شماره شناسنامه، کد ملی، کد وضعیت نظام وظیفه، کد بخش (تولد-پیش دانشگاهی-دیپلم و اقامت)، کد شهر (تولد-پیش دانشگاهی-دیپلم و اقامت)، کد استان (تولد-پیش دانشگاهی-دیپلم و اقامت)، سال تولد، جنسیت، کد استان بومی، سهمیه قبولی، معلولیت، کد عنوان دیپلم، دین، اتباع خارجی، زبان، سال اخذ پیش دانشگاهی، سال اخذ دیپلم، رتبه در سهمیه و کد رشته محل قبولی در دانشگاه.

۲. داده‌های مربوط به دیپلم دانشجو شامل معدل دیپلم، معدل پیش‌دانشگاهی، نوع دیپلم و نوع پیش دانشگاهی.

۳. داده‌های مربوط به زمان دانشجو شدن همچون معدل ترم اول دانشجو، تعداد واحد گذرانده شده، تعداد واحد اخذ شده دانشجو و دوره تحصیلی دانشجو (شبانه یا روزانه بودن آن).

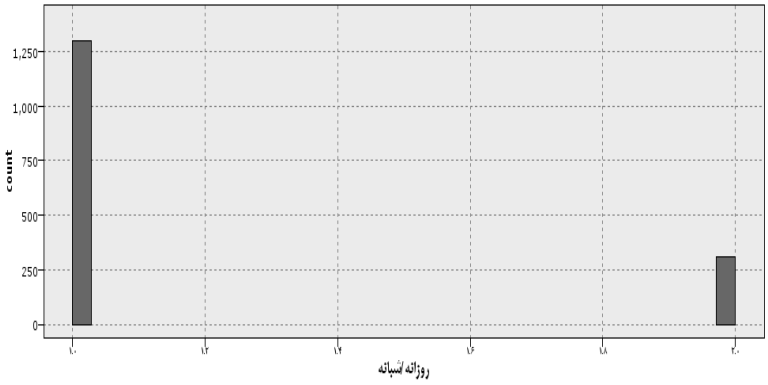
تعداد دانشجویان به تفکیک جنسیت در مقطع کارشناسی در شکل ۱ نمایش داده شده است که از این تعداد ۷۲۲ نفر زن (که با کد یک نمایش داده شده) و تعداد ۸۸۴ نفر مرد (که با کد دو نمایش داده شده) می‌باشد:

شکل ۱-توزیع فراوانی بر حسب تفکیک جنسیتی در مقطع کارشناسی



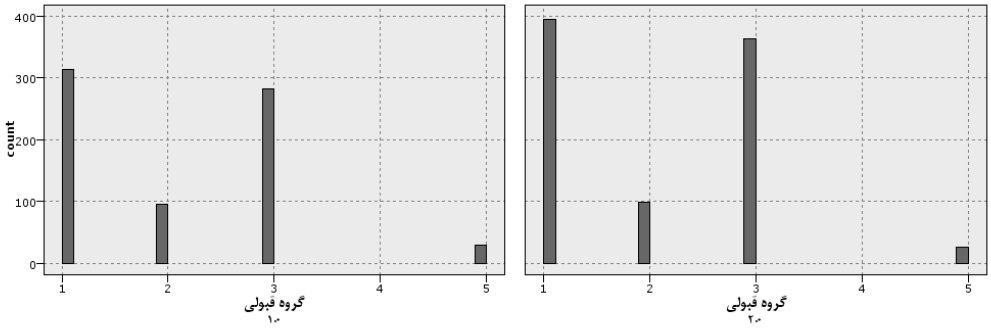
تعداد افراد پذیرش شده را به تفکیک دوره تحصیلی (روزانه/شبانه) در شکل ۲ نمایش می‌دهیم که تعداد ۱۲۹۸ در کد یک (دوره تحصیلی روزانه) و ۳۰۸ نفر در کد دو (دوره تحصیلی شبانه یا نوبت دوم) ثبت نام نموده‌اند:

شکل ۲- توزیع فراوانی بر حسب تفکیک دوره تحصیلی



تفکیک گروه قبولی و جنسیتی دانشجویان را در شکل ۳ نمایش می‌دهیم بیشترین میزان قبولی بین پسران و دختران در گروه آزمایشی ریاضی و فنی در دانشگاه قم می‌باشد که این آمار در جدول شماره ۲ نیز نمایش داده شده است:

شکل ۳-توزیع فراوانی بر حسب گروه قبولی دانشجویان به تفکیک جنسیتی

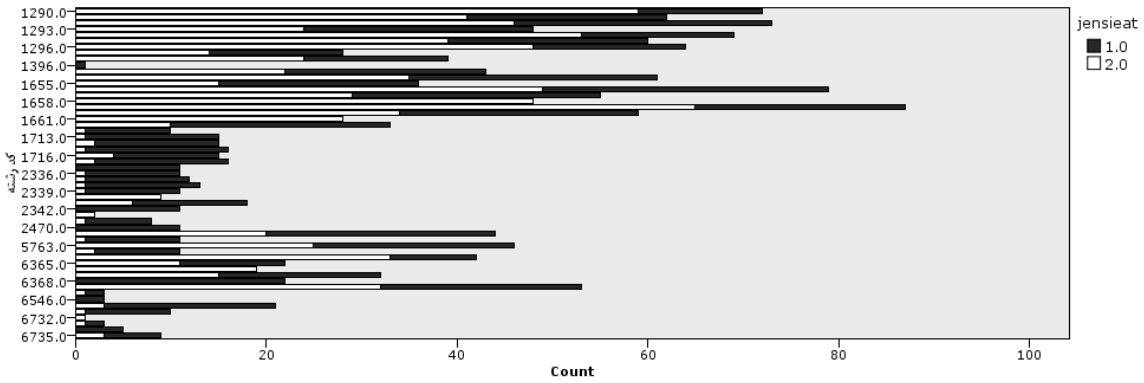


جدول ۲-جدول فراوانی گروه‌های قبولی به تفکیک جنسیت

ردیف	نام گروه قبولی	گروه قبولی	تعداد کل افراد ثبت نام شده	جنسیت	
				زن	مرد
۱	گروه آزمایشی علوم ریاضی و فنی	۱	۷۰۸	۳۱۴	۳۹۴
۲	گروه آزمایشی علوم تجربی	۲	۱۹۵	۹۶	۹۹
۳	گروه آزمایشی علوم انسانی	۳	۶۴۶	۲۸۲	۳۶۴
۴	گروه آزمایشی هنر	۴	پذیرفته شده نداریم	--------	
۵	گروه آزمایشی زبان‌های خارجی	۵	۵۷	۳۰	۲۷

در شکل شماره ۴ رشته‌های پذیرش شده سال تحصیلی ۱۳۹۰ دانشگاه قم را که افراد در آن ثبت نام داشته‌اند را به صورت گرافیکی مشاهده می-نمایید:

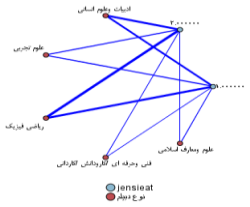
شکل ۴-توزیع فراوانی افراد ثبت نام شده به تفکیک رشته‌های تحصیلی و جنسیتی



همان‌طور که در شکل شماره ۴ مشاهده می‌نمایید بیشترین میزان ثبت نامی برای کد رشته ۱۶۵۹(مهندسی عمران) با تعداد ۸۷ نفر که ۲۲ نفر زن و ۶۵ نفر مرد می‌باشد.

در شکل شماره ۵ نمودار گرافیکی نوع دیپلم و جنسیت ثبت نام شدگان را مشاهده می‌نمایید که خط پر رنگ برای دیپلم ریاضی فیزیک می‌باشد که بیشترین تعداد ثبت نام شده در دانشگاه را دارا می‌باشد:

شکل ۵-نمودار گرافیکی نوع دیپلم و جنسیت افراد ثبت نام شده

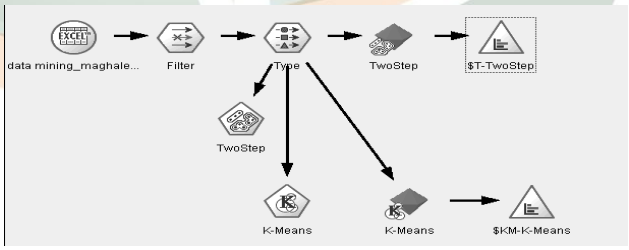


گروه TwoStep از یک روش خوشه‌بندی دو مرحله‌ای استفاده می‌کند. مرحله اول با یک بار گذر از داده‌ها، آنها را در یک مجموعه قابل مدیریتی از زیر خوشه‌ها فشرده می‌کند. قدم دوم از یک روش خوشه‌بندی سلسله مراتبی، به منظور ادغام تکاملی این زیر خوشه‌ها به خوشه‌های بزرگ‌تر و بزرگ‌تر، بهره می‌برد. روش TwoStep مزیت تخمین خودکار تعداد بهینه خوشه‌ها را داراست. او می‌تواند فیلدهای با نوع مختلف و مجموعه داده-های بزرگ را به خوبی مدیریت کند[۱].

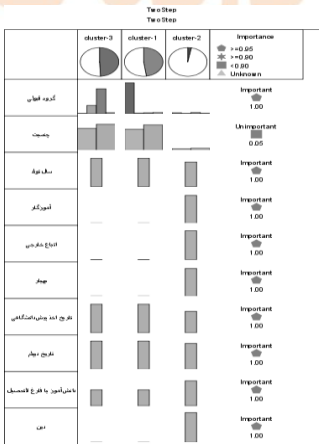
لازم به ذکر است که در خوشه‌بندی سلسله مراتبی با ساختاری از خوشه‌ها مواجه هستیم که هر کدام از این خوشه‌ها، خود مجموعه‌ای از خوشه-های دیگر می‌باشند. در این روش خوشه‌بندی، در مرحله اول دارای تعداد معینی خوشه نمی‌باشیم. به این معنی که خوشه‌بندی در چندین مرحله انجام می‌گیرد. در مرحله اول دارای یک خوشه شامل تمامی عوارض و در مراحل نهایی دارای n خوشه که هر کدام دارای تنها یک عارضه می‌باشد، خواهیم بود[۱۳].

در شکل شماره ۶ مدل ایجاد شده برای خوشه‌بندی را نمایش می‌دهیم در شکل شماره ۷ خوشه‌های به دست آمده با روش TwoStep سه عدد می‌باشد در حالی که در روشK-Means این خوشه‌ها به پنج عدد افزایش می‌یابد:

شکل ۶-نمایش مدل خوشه‌بندی در نرم‌افزار Clementine



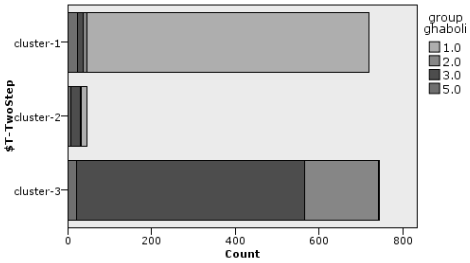
شکل ۷ -نمایش گرافیکی از آمار و توزیع فیلدها بین خوشه‌ها با روش TwoStep



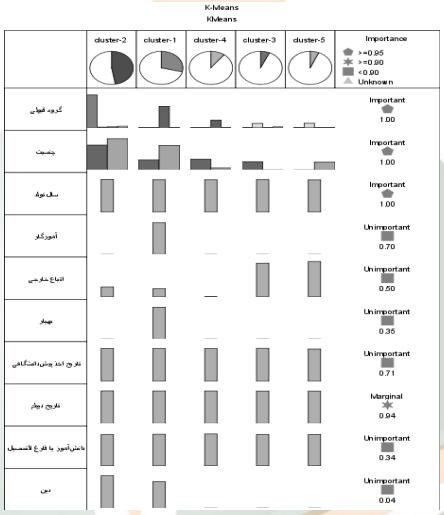
کنفرانس داده کاوی ایران

پس از مشخص شدن خوشه‌ها حال نمودار گرافیکی شکل شماره ۸ به دست می‌آید:

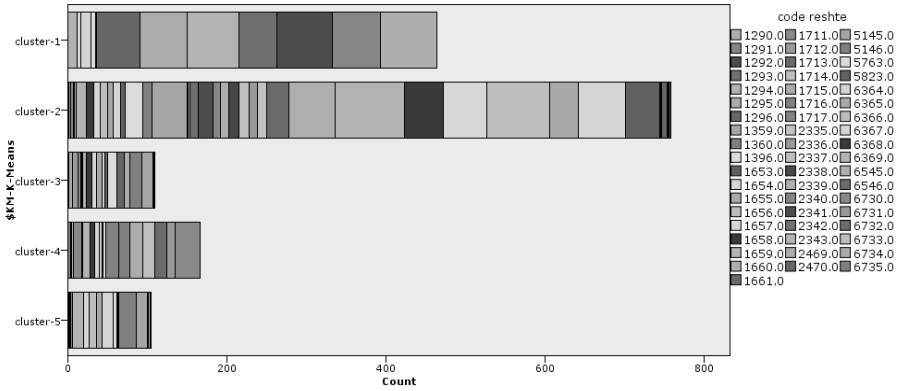
شکل ۸-نمایش گرافیکی خوشه‌ها با روش TwoStep



شکل ۹-نمایش گرافیکی از آمار و توزیع فیلدها بین خوشه‌ها با روش K-Means

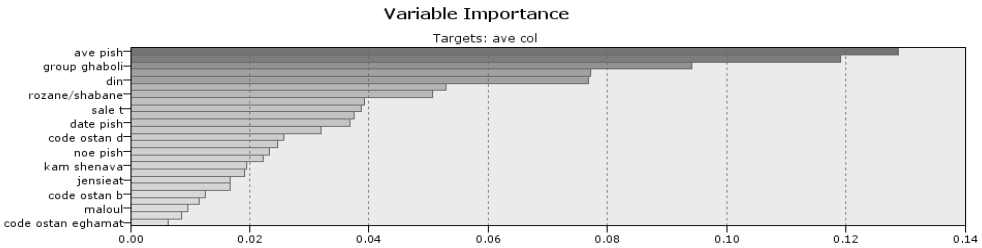


شکل ۱۰-نمایش گرافیکی خوشه‌ها با روش K-Means بر اساس کد رشته قبولی



با توجه به اینکه فیلدهای معدل کل دانشجو بعد از زمان قبول شدن در دانشگاه به عنوان فیلدهای خروجی می‌باشد میزان اهمیت بقیه فیلدها که روی این فیلد خروجی تاثیر می‌گذارد بدین ترتیب می‌باشد:

شکل ۱۱-میزان اهمیت فیلدها به ترتیب با روش Neural Net

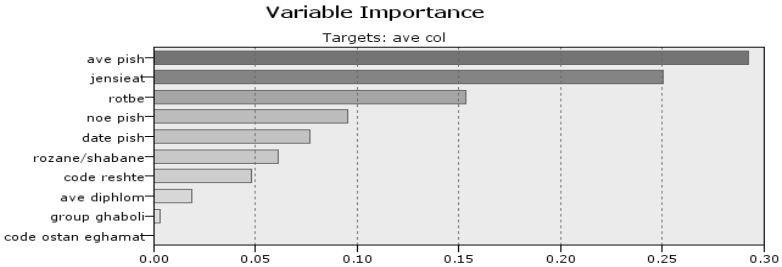


همان‌طور که در شکل شماره ۱۱ مشاهده می‌کنید فیلدهای معدل پیش‌دانشگاهی و گروه قبولی به ترتیب دارای اهمیت می‌باشند که اهمیت تمامی فیلدها را در جدول شماره ۳ نمایش داده‌ایم:

جدول ۳ – جدول تعیین میزان اهمیت مشخصه‌ها با روش Neural Net

V۱	Nodes	Importance	Importance
۱	معدل پیش دانشگاهی	۰,۱۲۸۷	۰,۱۳
۱	معدل دیپلم	۰,۱۱۹۱	۰,۱۲
۱	گروه قبولی	۰,۰۹۴۱	۰,۰۹
۱	رتبه	۰,۰۷۷۲	۰,۰۸
۱	دین	۰,۰۷۶۷	۰,۰۸
۱	نوع دیپلم	۰,۰۵۲۸	۰,۰۵
۱	روزانه/شبانه	۰,۰۵۰۷	۰,۰۵
۱	کد استان تولد	۰,۰۳۹۲	۰,۰۴
۱	سال تولد	۰,۰۳۸۸	۰,۰۴
۱	تاریخ دیپلم	۰,۰۳۷۵	۰,۰۴
۱	تاریخ پیش دانشگاهی	۰,۰۳۶۷	۰,۰۴
۱	اتباع	۰,۰۳۱۹	۰,۰۳
۱	کد استان دیپلم	۰,۰۲۵۶	۰,۰۳
۱	کد استان ماقبل دیپلم	۰,۰۲۴۷	۰,۰۲
۱	نوع پیش دانشگاهی	۰,۰۲۳۲	۰,۰۲
۱	بهیار	۰,۰۲۲۳	۰,۰۲
۱	کم شنوا	۰,۰۱۹۶	۰,۰۲
۱	نظام وظیفه	۰,۰۱۹۲	۰,۰۲
۱	دانش آموز یا فارغ التحصیل	۰,۰۱۶۷	۰,۰۲
۱	جنسیت	۰,۰۱۶۷	۰,۰۲
۱	کد استان بومی	۰,۰۱۲۵	۰,۰۱
۱	کد استان پیش‌دانشگاهی	۰,۰۱۱۴	۰,۰۱
۱	معلول	۰,۰۰۹۶	۰,۰۱
۱	آموزگار	۰,۰۰۸۶	۰,۰۱
۱	کد استان اقامت	۰,۰۰۶۳	۰,۰۱

شکل ۱۲-میزان اهمیت فیلدها به ترتیب با روش CHAID



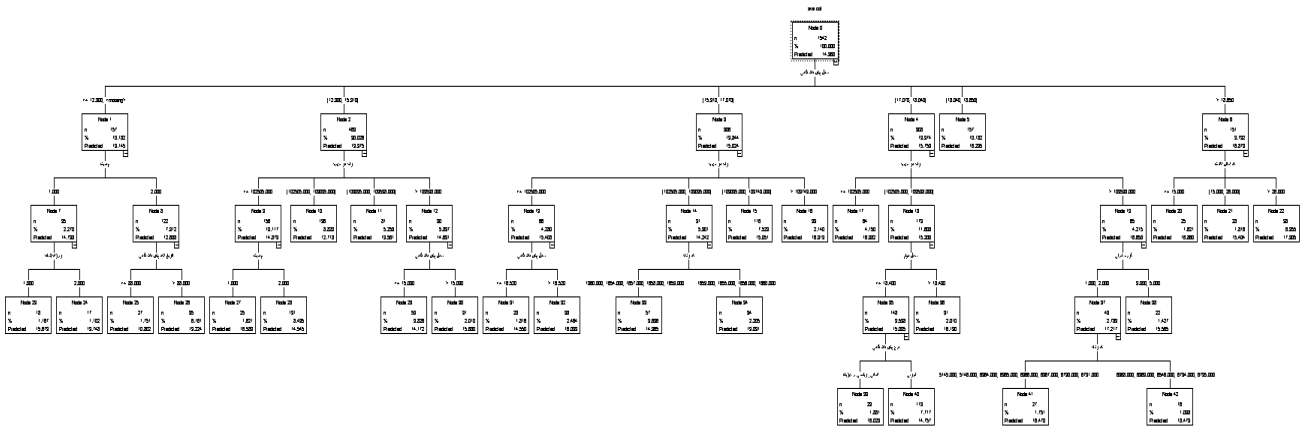
جدول ۴- جدول تعیین میزان اهمیت مشخصه‌ها با روش CHAID

V۱	Nodes	Importance	Importance
۱	معدل پیش دانشگاهی	۰,۲۹۲۶	۰,۲۹
۱	جنسیت	۰,۲۵۰۳	۰,۲۵
۱	رتبه	۰,۱۵۳۳	۰,۱۵
۱	نوع پیش دانشگاهی	۰,۰۹۵۲	۰,۱
۱	تاریخ پیش دانشگاهی	۰,۰۷۷	۰,۰۸
۱	روزانه/شبانه	۰,۰۶۱۳	۰,۰۶
۱	رشته تحصیلی	۰,۰۴۸۱	۰,۰۵
۱	معدل دیپلم	۰,۰۱۹	۰,۰۲
۱	گروه قبولی	۰,۰۰۳۲	۰
۱	استان اقامت	۰	۰

همان‌طور که مشاهده می‌کنید مشخصه‌هایی مانند معدل پیش‌دانشگاهی، گروه قبولی، جنسیت، رتبه و نوع پیش‌دانشگاهی شخص در کسب معدل ترم اول در زمان دانشجو شدن او نقش پر رنگی را ایفا می‌کند. این موضوع را این گونه می‌توان تحلیل کرد که؛ شخصی که در دوران دانش‌آموزی خود دارای معدل بالاتری بوده بعد از قبولی در دانشگاه هم، معدل بالاتری را کسب خواهد نمود. در شکل شماره ۱۳ درخت این الگوریتم نمایش داده شده است.

کنفرانس داده کاوی ایران

شکل ۱۳ -نمایش درختی الگوریتم CHAID



در درخت بالا معدل پیش دانشگاهی باعث تقسیم‌بندی در ریشه شده است.

۵. نتیجه

سهم اصلی در این پژوهش توجه به توانایی‌ها و قدرتمندی‌های تکنولوژی داده‌کاوی در زمینه سیستم‌های آموزش عالی می‌باشد. نتایج این پژوهش مورد استفاده وزارت علوم، تحقیقات و فناوری و دانشگاه‌های معتبر خواهد بود.

این پژوهش تلاشی برای پیاده‌سازی مدل‌های داده‌کاوی پیش‌بینی کننده، به منظور پیش‌بینی وضعیت تحصیلی دانشجویان بر اساس مشخصات فردی و گذشته تحصیلی آنان بوده است. با توجه به نتایج آماری که از ساخت مدل‌های پیش‌بینی کننده وضعیت دانشجو در این پژوهش بدست آمده است، می‌توان با اطمینان بالایی از آینده تحصیلی دانشجویان بر مبنای داده‌های گذشته اطلاع حاصل نمود. اگر چه نتایج این پژوهش منتهای هدف در این حوزه نیست و می‌توان با بسط مدل‌ها و درگیر کردن پارامترهای جدید به نتایج دقیق‌تر و قابل اطمینان‌تری دست یافت. میزان اهمیت این پیش‌بینی‌ها بسیار واضح و روشن است، چنانکه در اکثر مؤسسات آموزشی از دغدغه‌های اصلی مدیران آموزشی است. بنابراین نتایج حاصل از این تحقیق و موارد مشابه می‌تواند به صورت جدی در مراجع ذکر شده پیاده سازی و مورد استفاده قرار گیرد. استفاده از چنین مدل‌هایی دانشگاه را در ارتقاء سطح علمی دانشجویان و هدفمند نمودن فرایندهای آموزشی یاری می‌نماید[۳].

در پایان لازم است به این نکته اشاره کنیم که کاربردهای داده‌کاوی در آموزش عالی به تازگی مورد توجه قرار گرفته است. دانشگاه‌ها به منظور اینکه در حوزه رقابتی آموزش باقی بمانند نیازمند دانش پایه‌ای اولیه‌ای هستند که می‌توانند آن را از داده‌های تاریخی و عملیاتی‌شان استخراج کنند.

تکنیک‌های داده‌کاوی ابزارهای تحلیلی هستند که برای استخراج دانش معنادار از مجموعه داده‌های بزرگ مورد استفاده قرار می‌گیرند.

با توجه به اینکه داده‌کاوی در حوزه آموزش، حوزه تحقیقاتی کاملا جدیدی می‌باشد، نیازمند آن هستیم که در این حوزه تحقیقات گسترده‌ای صورت بگیرد و شکاف دانشی در این حوزه تا حدی از میان برداشته شود. دقت روشهای پیش‌بینی در سایر حوزه‌ها به تناسب نوع کاربرد ارتقاء زیادی یافته است، ولیکن در این حوزه خاص ارتقاء زیادی در مدل‌ها صورت نگرفته و نیازمند تحقیقات گسترده‌ای در این حوزه می‌باشیم تا از این طریق داشتن فرایندهای آموزشی مؤثرتر، کاراتر و دقیق‌تر را در سیستم‌های آموزش عالی در دانشگاه‌ها تسهیل کرد.

شاید مهم‌ترین نکته این باشد که هیچ مدل یا الگوریتمی نمی‌تواند و نباید به تنهایی استفاده شود. نمی‌توان هیچ مدل یا الگوریتمی را در این زمینه بهترین نامید برای هر مسئله داده شده، طبیعت داده استفاده شده بر روی انتخاب مدل‌ها و الگوریتم‌هایی که برگزیده می‌شود تأثیر خواهد گذاشت.

مراجع

[۱] مینائی، بهروز؛ نصیری، مهدی؛ حسنی، دانیال؛ شناسا، ابراهیم؛ **"آموزش داده‌کاوی با Clementine"**، انتشارات گروه مهندسی-پژوهشی ساحر، تهران، چاپ اول، پاییز ۱۳۹۰

[۲] خادم القرانی، فریبا؛ سرائی، محمدحسین؛مصطفوی، سید ابوالفضل؛ **"کاربرد داده‌کاوی در هدفمند کردن انتخاب رشته دانشگاهی و بهبود کیفیت برنامه ریزی آموزشی"**

[۳] یقینی، مسعود؛ اکبری، امین؛ شریفی، سیدمحمدمهدی؛ **"پیش‌بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک‌های داده‌کاوی"**، کنفرانس بین‌المللی داده‌کاوی، ۲۰۰۸

[۴] یقینی، مسعود؛ حیدری، سمیه؛ **"داده‌کاوی جهت ارتقاء و بهبود فرآیندهای سیستم آموزش عالی"**، کنفرانس بین‌المللی داده‌کاوی، ۲۰۰۸

[۵] یقینی، مسعود؛ اکبری، امین؛ شریفی، سیدمحمدمهدی؛ **"دسته‌بندی دانشجویان استخراج روابط موجود در سیستم آموزشی دانشگاه"**، کنفرانس بین‌المللی داده‌کاوی، ۲۰۰۸

[۶] شکورنیاز، ونوس؛ حاجی علی اکبری، آرش؛ **"خوشه‌بندی داده‌های آماری دانشجویان دانشگاه علم و صنعت و استخراج نمایه ساز توصیفی برای دانشجویان موفق"**، کنفرانس بین‌المللی داده‌کاوی، ۲۰۰۸

[۷] C. Romero, and S. Ventura; **"Educational data mining: A survey from ۱۹۹۵ to ۲۰۰۵",** Expert Systems with Applications ۳۳, ۱۳۵–۱۴۶, ۲۰۰۷.

[۸] J Ranjan, K Malik; **"Effective educational process : a data - mining approach",** VINE: The journal of information and knowledge management systems, Vol. ۳۷ No. ۴, ۵۰۲-۵۱۵, ۲۰۰۷.

[۹] Mohammad Reza Beikzadeh, Naeimeh Delavari; **"A New Analysis Model for Data Mining Processes in Higher Educational Systems"**, Proceedings of M۲USIC, ۲۰۰۴.

[۱۰] Saurabh Pal , Brijesh Kumar Baradwaj; ***"Mining Educational Data to Analyze Student'Perfomance";***International Journal of Advanced Computer Science and Application*;* Vol.۲, No.۶, ۲۰۱۱

[۱۱] R. S. Bichkar, R. R. Kabra; **"*Performance Prediction of Engineering Students using Decision Trees*";** International Journal of Computer Application(۰۹۷۵-۸۸۸۷); Vol۳۶, December ۲۰۱۱

[۱۲] Jain, A., M. Murty and P. Flynn. ۱۹۹۹. "**Data Clustering: A Review. ACM Computing Surveys",**۳۲۳ – ۲۶۴ :(۳)۳۱

[۱۳] عامری، فاطمه؛ ولدان زوج، محمدجواد؛ مختارزاده، مهدی؛ **"بررسی تکنیک‌های خوشه‌بندی به روش نظارت نشده"**

[۱۴] Luan, Jing; **"Data Mining and Knowledge Management in Higher Education"**, Knowledge and Data ManagementWhite Papers, Presentation at AIR Forum in Cabrillo College,Toronto, Canada, ۶-۱۶،۲۰۰۲

[۱۵] [۲] Romero, C; Ventura, S; **"Educational data mining: A survey from ۱۹۹۵ to ۲۰۰۵ "**, Elsevier, Expert Systems with Applications ۳۳,۱۳۵-۱۴۶،۲۰۰۷

[۱۶] Yang, Ming**; "Data Mining Techniques Applied to Texas Woman's University's Enrollment data – What Canthe Data Tell us?"**, MS Thesis, Texaz Woman's University, ۲۰۰۶.

[۱۷] حاج احمدی، امیرحسین، ۱۳۸۵، **مبانی خوشه‌بندی**.دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه امیرکبیر http://ceit.aut.ac.ir/~shiry/lecture/machine-learning/tutorial/clustering/Introduction.html (تاریخ دسترسی ۱۳۸۷/۴/۵)

[۱۸] http://barnamenevis.org/showthread.php

کنفرانس داده کاوی ایران