داده کاوي بر روي پايگاه داده‌ي آموزش دانشگاه کاشان به کمک روش GRI و تحليل نتيجه‌ها

عطیه منعمی بیدگلی[1]؛ احمد یوسفان[2]، ابوالفضل خدمتی[3]

چکیده

داده‌های آموزشی یک دانشگاه از جمله منابع مهم آموزشی برای یک دانشگاه است که به کمک آن می‌توان تصمیم‌گیری‌ها و جهت گیری‌های شایسته‌ای را اتخاذ کرد. با افزایش داده‌های آموزشی و پیچیده‌تر شدن ارتباط میان پارامترهای گوناگون در آن‌ها و همچنین خواسته‌های نوین مدیریتی، روش‌های سنتی و دستی یا روش‌های محاسباتی ساده، برای مدیریت آموزشی چندان کمک کننده نیست؛ بنابراین نیاز به داده کاوی بر روی داده‌های آموزشی احساس می‌شود. برای داده کاوی بر روی پایگاه داده آموزش دانشگاه کاشان، تعداد مناسبی از ویژگی‌های مهم از انبوه ویژگی‌های موجود برگزیده شده و سپس به کمک الگوریتم GRI، داده‌کاوی بر روی این داده‌ها انجام شده است. همچنین این الگوریتم به صورت جداگانه بر روی داده‌های هر کدام از دانشکده‌ها اجرا شده و نتیجه‌های به دست آمده از این الگوریتم‌ها تحلیل شده است.

کلمات کلیدی

داده کاوی، الگوریتم GRI، پایگاه داده آموزش دانشگاه، مدیریت

Data Mining on Education Database of Kashan University by GRI Method and Analyzing Outcomes

Atiye Monemi Bidgili; Ahmad Yoosofan, Aboulfazl Khedmati

Abstract

One such important source of data for a university is educational data that by help of it, good decisions and orientations can be adopted. According to the increase in the educational data and the complex relationship between various parameters and also increase in the demands of modern management, traditional and handicraft methods or simplified calculation methods can not be helpful for educational management So the need for data mining on the educational data is felt. In purpose of performing data mining on educational database of Kashan university, appropriate number of features through large number of existing feature was selected Then data mining is performed on this data by help of GRI method Furthermore this algorithm was executed on separated data of various faculties. Finally outcomes of algorithm was analyzed.

Keywoards

Data mining, GRI algorithm, Educational database of University, Management

۱. مقدمه

[1] فارغ التحصیل ارشد از دانشگاه صنعتی شریف،مدرس دانشگاه کاشان، monemiatieh@gmail.com

[2] عضو هیئت علمی دانشکده مهندسی برق وکامپیوتر دانشگاه کاشان، yoosofan@kashanu.ac.ir

[3] دانشجوی ارشد هوش مصنوعی دانشگاه آزاد قزوین، akhedmati@gmail.com

در حال حاضر در اکثر دانشگاه‌های ایران، بانک‌های اطلاعاتی وسیعی از ویژگی‌ها سوابق آموزشی و تحصیلی دانشجویان موجود است. 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 نشده است. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزش عالی در جهت ارتقاء و بهبود فرآیندهای آموزشی نظیر برنامه‌ریزی، ثبت‌نام، ارزیابی و مشاوره کمک شایانی کند و بدین ترتیب آن‌ها را در تصمیم‌گیری بهتر و داشتن طرح پیشرفته‌تری در هدایت دانشجویان کمک می‌کند. در نتیجه، این بهبود می‌تواند 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 فرآیندهای سیستم آموزش عالی به ارمغان آورد. نرم‌افزارهای کامپیوتری به کار گرفته شده برای این منظور، غالباً برای مکانیزه کردن وضع موجود و اجرای پرس و جوهای معمولی جوابگو هستند. در حالی که در عمق این حجم عظیم داده‌ها، الگوها و روابط بسیار جالبی به طور پنهان باقی می‌ماند. [۱۳، ۱۴ ،۸، ۱۶]

آنچه در این مقاله مورد بررسی قرار می‌گیرد پیدا کردن قوانین و اطلاعات نهفته موجود در پایگاه داده آموزش دانشگاه کاشان و پاسخ به سؤال‌هایی از جمله:

- آیا بومی و غیر بومی بودن در پیشرفت تحصیلی دانشجو تأثیری دارد؟
- روزانه یا شبانه بودن در وضعیت تحصیلی دانشجویان چه تأثیری دارد؟
- وضعیت تحصیلی دانشجویان در کل دانشگاه، دانشکده‌ها و گروه‌های مختلف آموزشی به چه صورت است؟
- آیا در پایگاه‌داده سیستم آموزشی دانشگاه کاشان اطلاعات نهفته‌ای وجود دارد؟
- آیا نوع ورود به دانشگاه در معدل نهایی دانشجو تأثیری دارد؟

برای انجام این کار در آغاز با سیستم آموزش آشنا شده و بعضی از فیلدهایی که در انجام این کار کمک‌کننده می‌باشند از پایگاه‌داده استخراج و قالب دادها به قالب مناسب تبدیل و سپس الگوریتم GRI روی آن‌ها اعمال شده است و در آخر به تحلیل و ارزیابی نتایج و خروجی الگوریتم پرداخته شده است. نتایج این مطالعه در جهت بهبود کیفیت برنامه‌ریزی آموزشی و یافتن راه‌کارهایی مناسب موثر خواهد بود.

در این مقاله و در بخش دوم، تحقیقات انجام شده در زمینه کاربرد داده‌کاوی در آموزش عالی مرور شده و سپس به شرح دقیق مسئله خواهیم پرداخت؛ در بخش سوم توضیح مختصری در مورد الگوریتم GRI ارائه می‌شود؛ در بخش چهارم به همراه کلیه بخش‌های آن به جزئیات روش انجام تحقیق، قوانین ایجاد شده و نتایج آزمایشات پرداخته می‌شود در بخش پنجم نتیجه‌گیری نهایی و پیشنهادات آتی بیان خواهد گردید و در انتهای مقاله مراجع به کار گرفته شده ارائه می‌شود.

۲. کاربرد داده‌کاوي در آموزش عالي و فرآیندهای سیستم آموزش

با توجه به اینکه آموزش عالی همواره با داده‌ها و اطلاعات بسیار زیادی در مورد دانشگاه‌ها، دانشجویان، اعضای هیئت علمی، پرسنل و منابع مادی روبروست و در اکثر مواقع این داده‌ها می‌تواند حامل اطلاعات و الگوهای با ارزشی باشند، لذا به نظر می‌رسد یکی از مهمترین کاربردهای داده‌کاوی می‌تواند روی داده‌های موجود در آموزش عالی باشد. امروزه بانک‌های اطلاعاتی وسیعی از ویژگی‌های دانشجویان موجود است که اطلاعات مربوط به ویژگی‌های خانوادگی، تحصیلی و ... را شامل می‌شود. پیدا کردن الگوها و دانش نهفته در این اطلاعات می‌تواند به تصمیم‌گیرندگان عرصه آموزش عالی کمک شایانی کند. استفاده از تکنیک‌های پیشرفته داده‌کاوی 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[۲].

بیک‌زاده و دلاوری[۱۶] شش فرآیند اصلی را در هر سیستم آموزش عالی مشخص کرده‌اند که شامل ثبت‌نام، برنامه‌ریزی، ارزیابی، مشاوره، بازاریابی و آزمون می‌باشد. هر فرآیند اصلی می‌تواند به زیر فرآیندهایی تقسیم شود. به عنوان مثال "ارزیابی" یک فرآیند آموزشی می‌باشد و زیر فرآیندهای اصلی آن شامل "ارزشیابی دانشجو"، "ارزشیابی مدرس"، ارزشیابی آموزش"، ارزشیابی واحد درسی" و "ارزیابی ثبت‌نام دانشجو" می‌باشند. مقصود اصلی در داده‌کاوی آموزشی، بهبود فرآیندهای فعلی به فرآیندهای جدید و ارتقا یافته‌ای است که مزایای برتری، نسبت به فرآیندهای قبلی دارد. به عنوان مثال " ارزیابی ثبت‌نام دانشجو" یک زیر فرآیند از زیر فرآیند "ارزیابی" است. با استفاده از بعضی تکنیک‌های پیش‌بینی در داده‌کاوی مانند تحلیل شبکه‌های عصبی، رگرسیون خطی و چندگانه بر روی مجموعه داده‌های سیستم، این فرآیند سنتی سیستم آموزش می‌تواند ارتقا یابد و الگوهای موفقیت کسانی که برای دانشگاه پذیرفته شده‌اند، استخراج شود. فرآیند ارتقا یافته نهایی امکان بازگشت هر دانشجوی ثبت‌نام

شده در دانشگاه در نیمسال‌های آتی را پیش‌بینی می‌کند. "طراح الزامات پذیرش ثبت‌نام"[1] به عنوان یک موجودیت خارجی در دانشگاه می‌تواند از نتایج فرآیند استفاده کند و پیش‌بینی دقیقی از دانشجویان ورودی جدید در هر سال ارائه دهد[۱۶].

Aksenova با استفاده از روش SVM و مدل‌های پیش‌بینی بر پایه قوانین، به پیش‌بینی ثبت‌نام برای دانشجویان رشته علوم کامپیوترِ دانشگاه ایالتی کالیفرنیا، ساکرمنتو، پرداخته است. انواع داده‌هایی که در طول فرآیند داده‌کاوی مورد استفاده قرار گرفتند شامل: جمعیت، نرخ بیکاری در ناحیه، شهریه و مالیات، درآمد خانواده، نرخ فارغ‌التحصیلی از دبیرستان و داده‌های تاریخی ثبت‌نامِ مربوط به سال‌های گذشته می‌باشد. این روش نسبت به سایر روش‌های پیش‌بینی ثبت‌نام دارای مزایای بسیاری است. از جمله SVM با سیستم‌های پیچیده سازگار است و در برخورد با داده‌های مغشوش سازگار است و در برخورد با داده‌های مغشوش دقیق عمل می‌کند[۸ ، ۳]

داده‌کاوی می‌تواند به هر یک از عاملان فرآیند آموزش کمک کند. 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 می‌باشد. مؤسسات می‌خواهند بدانند که کدامیک از دانشجویان در یک درس خاص ثبت‌نام خواهد کرد، کدامیک از آن‌ها به کمک ویژه و رسیدگی جهت فارغ‌التحصیل شدن نیاز دارند. کدامیک احتمال افتادن در یک درس و یا حذف پیش از فارغ‌التحصیلی را دارند، کدام زیرمجموعه از فارغ‌التحصیلان احتمال بیشتری برای عرضه تعهدات مالی دارند. یک مدیر ممکن است بخواهد به اطلاعاتی نظیر اطلاعات پذیرش دانشجویان پی ببرد و میزان ثبت‌نام دانشجویان در یک کلاس را به منظور برنامه‌ریزی و زمان‌بندی پیش‌بینی کند. دانشجویان ممکن است بخواهند بر اساس پیش‌بینی نحوه عملکردشان بر طبق واحدهای انتخابی خاص به بهترین نحو واحدها را انتخاب کنند.

به طور کلی کاربردهای داده‌کاوی در آموزش عالی می‌تواند به شرح زیر باشد:

- پیش‌بینی ثبت‌نام
- درک ثبت‌نام دانشجویان
- انتخاب دانشجویان مناسب برای شرکت در کلاس‌های جبرانی
- خوشه‌بندی و پیش‌بینی دانشجویان ماندگار و غیر ماندگار
- شناخت انواع دانشجویان برای فهم بهتر و یا استفاده در دسته‌بندی دانشجویان
- برنامه‌ریزی تحصیلی – پیش‌بینی گذراندن دروس
- پیش‌بینی تعهد و منفعت رسانی فارغ‌التحصیلان
- رابطه میان نتایج امتحان ورودی دانشگاه و میزان موفقت دانشجویان
- تحلیل ماندگاری دانشجویان در ترم‌های آتی
- پیش‌بینی ثبت‌نام در یک درس خاص
- بررسی ترکیب واحدهای انتخابی هر دانشجو برای زمان‌بندی مناسب واحدها و جلوگیری از تداخل واحدها در زمان‌بندی
- ارتقاء فرآیند مشاوره دانشجویان
- بررسی رابطه میان واحدهای انتخابی دانشجویان به منظور توسعه واحدهای جدید برای دوره‌های کارشناسی و کارشناسی ارشد

در این مقاله از نرم‌افزار clementine، ساخت شرکت SPSS استفاده شده است. این نرم‌افزار امکان ایجاد مدل‌های متعددی را ، بر اساس تئوری-های آماری، هوش مصنوعی و یادگیری ماشین ارائه می‌دهد[۶،۱۵،۷،۵].

۳. الگوریتم GRI

ساختار قوانین وابستگی، به صورت دو عبارت درست/غلط[2] است که از دو قسمت مقدم و تالی تشکیل شده است. این ساختار برای رسیدگی به داده‌های دسته‌ای بسیار مناسب می‌باشد. نکته‌ی مهم این است که وقتی ویژگی‌ها به صورت عددی باشند و حجم وسیعی از داده‌ها را در اختیار داشته باشیم، قوانین وابستگی چگونه تشخیص داده می‌شوند؟ البته اغلب می‌توان ویژگی‌های عددی را به صورت مجزا درآورد، برای مثال می‌توان درآمد زیر ۳۰۰۰۰$ را پایین، بالای ۷۰۰۰۰$ را بالا، و بین این دو عدد را درآمد متوسط در نظر گرفت. الگوریتم‌های C4.5 و CART (برای ساختن درخت تصمیم پیش‌بینی[3])، نیز به این صورت عمل می‌کنند و ویژگی‌های عددی را گسسته می‌کنند.

عملیات مجزا کردن این ویژگی‌ها باعث از دست رفتن بعضی از اطلاعات می‌شود، بنابراین ممکن است ترجیح دهیم از داده‌های عددی و غیر گسسته به عنوان ورودی استفاده کنیم. برای این کار می‌توان از یک متد تناوبی برای کاوش و یافتن قوانین وابسته‌سازی استفاده کرد: متد GRI برای این عمل بسیار مناسب است. متد GRI، با هر دو نوع داده‌های عددی و دسته‌ای[۴] به عنوان ورودی الگوریتم کار می‌کند اما خروجی آن فقط شامل متغیرهای دسته‌ای می‌باشد. Generalized Rule Induction در سال ۱۹۹۲ توسط Smyth and Goodman معرفی شده است[۹]. علاوه بر استفاده از مجموعه قلم‌های تکراری، GRI یک دیدگاه نظری- اطلاعاتی را نیز برای تشخیص نزدیکی یا میزان وابستگی کاندیداهای قوانین وابستگی به یکدیگر، به کار می‌گیرد.[۱]

۳-۱ معیار J

به طور مشخص الگوریتم GRI از معیار J استفاده می‌کند. معیار J در فرمول ۱ بیان شده است

$$J = P(x)\left[\left[\mathrm{P(y\,|\,x)}\ln\frac{\mathrm{P(y\,|\,x)}}{\mathrm{p(y)}} + [1-\mathrm{P(y\,|\,x)}]\ln\frac{1-\mathrm{P(y\,|\,x)}}{1-P(y)}\right]\right] \qquad (۱)$$

به طوری که :

۱- P(X) احتمال مشاهده شدن X در مقدم، را نشان می‌دهد. این مقیاسی است که طرف مقدم را پوشش می‌دهد. ضمناً می‌توان P(x) را با استفاده از یک توزیع تناوبی برای متغیر مقدم محاسبه کرد. در نرم افزار کلمنتاین ، ضریب پشتیبانِ[۵] محاسبه شده توسط نرم افزار، مقدار P(X) را نشان می‌دهد .

۲- P(Y) احتمال مشاهده شدنY در تالی را نشان می‌دهد. این مقیاس بیانگر تعداد تکرار Y در تالی است. P(Y) با استفاده از یک توزیع تناوبی برای متغیر تالی محاسبه می‌شود.

۳- P(Y | X) احتمال وقوع Y به شرط رخداد X را نشان می‌دهد. در قوانین وابستگی، مقدار P(Y | X)، همان ضریب اطمینانِ[۶] قانون می‌باشد .

۴- Ln تابع لگاریتم طبیعی است که در حقیقت همان log در مبنای e است.

برای قوانینی که بیش از یک مقدم دارند، P(X) احتمال ترکیب همه‌ی مقادیر متغیرهای موجود در مقدم قانون است.

الگوریتم GRI قوانین وابسته‌سازی را با یک مقدم تولید می‌کند و برای قانون پیدا شده، معیار J را محاسبه می‌کند. اگر میزان نزدیکی یا وابستگی قانون جدید، که به وسیله معیار J اندازه‌گیری شده است، از J مینیممی که در جدول قوانین انتخاب شده، بیشتر باشد وارد جدول قوانین می‌شود اما با توجه به اینکه اندازه این جدول ثابت است، در بین قوانین، قانونی که J آن از همه کمتر است از جدول حذف می‌شود. در نهایت قوانین ویژه‌ای با بیش از یک مقدم تولید می‌شود.

واضح است هر چقدر مقدارP(X) (P(X) ای که خارج از کروشه قرار دارد) بیشتر باشد مقـــدارِ معیارِ J نیز بیشتر خواهد بود. مقدار معیار J بیشتر به سمت J قوانینی میل می‌کند که مقدار مقدم آن‌ها بیشتر تکرار شده است و رایج‌تر هستند. همچنین، اگر P(Y) و P(Y | X) به سمت مقادیر نزدیک به صفر یا نزدیک به یک میل کنند مقدار J افزایش پیدا می‌کند.. در مورد معیار J نیز، برای ما مقادیر نزدیک به یک یا صفر بیشتر اهمیت دارند. به عبارت دیگر در محاسـبه معیار J اگر مقادیر احتمال تالی، P(Y)، و ضریب اطمینان قانون P(Y | X)، در همسایگی صفر یا یک باشند، مقدار معیار J برای ما مطلوب‌تر است.

معیار J ، قوانینی را تایید می‌کند که ضریب اطمینان آن‌ها بسیار بالا یا بسیار پایین باشد. حال سؤال این است که چرا قوانین وابستگی با ضریب اطمینان بسیار پایین برای ما ارزشمند هستند؟

برای مثال فرض کنید که قانون R را به شرح زیر داریم :

اگر خرید آب میوه آنگاه خرید قیچی با ضریب اطمینان برابر 0.01%

که احتمالاً مقیاس J، این قانون را می‌پذیرد زیرا ضریب اطمینان آن بسیار پایین است. اینجاست که تحلیلگر، فرم منفی این قانون را با ضریب اطمینان بسیار بالا نتیجه گیری می‌کند :

اگر خرید آب میوه آنگاه عدم خرید قیچی با ضریب اطمینان برابر 99.99%

اگر چه این قوانین منفی اغلب جالب هستند (در چیدمان محل فروش قیچی، از قسمت مربوط به آب میوه جدا می‌شود) اما در کل، مستقیماً قابل اجرا نیستند.[۴]

٤. روش انجام پژوهش

این تحقیق در طی انجام چند فاز اصلی صورت گرفته است:

- انتخاب داده‌های مناسب، پیش‌پردازش و آماده‌سازی آن‌ها
- اعمال الگوریتم GRI
- تحلیل نتایج

در مبحث داده‌کاوی مهمترین مسئله دستیابی به داده‌هایی است که بتوان بر اساس آن به نتایج مفیدی دست یافت. پژوهش حاضر یک مطالعه توصیفی-تحلیلی بوده و از لحاظ زمانی مقطعِ خاصی را مورد بررسی قرار می‌دهد[۱]. الگوریتم‌های داده‌کاوی اغلب به خصوصیات ویژه داده‌ها حساس هستند لذا با بررسی اولیه داده‌ها، فیلدهای دربردارنده اطلاعات لازم برای شناسایی هویت دانشجو انتخاب نشدند. همچنین از ستون‌هایی که با یکدیگر تغییر می‌کنند یکی انتخاب شده‌اند به طور مثال از دو ستون ترم تحصیلی و سال ورود تنها سال ورود انتخاب شده است. با تحلیل آماری فیلدهایی مانند سن جزء داده‌های پرت محسوب شدند لذا از ورود آن‌ها برای آنالیز و کشف الگوها توسط تکنیک‌های داده‌کاوی اجتناب شده است. فایل داده‌ای حاوی اطلاعات کل دانشجویان راکد دانشگاه کاشان از سال ۱۳۷۰ تا سال ۱۳۸۶ می‌باشد. فیلدهایی که از سیستم آموزش انتخاب شده است در جدول ۱ نشان داده شده است.

جدول۱- فیلدهای استفاده شده در داده‌کاوی

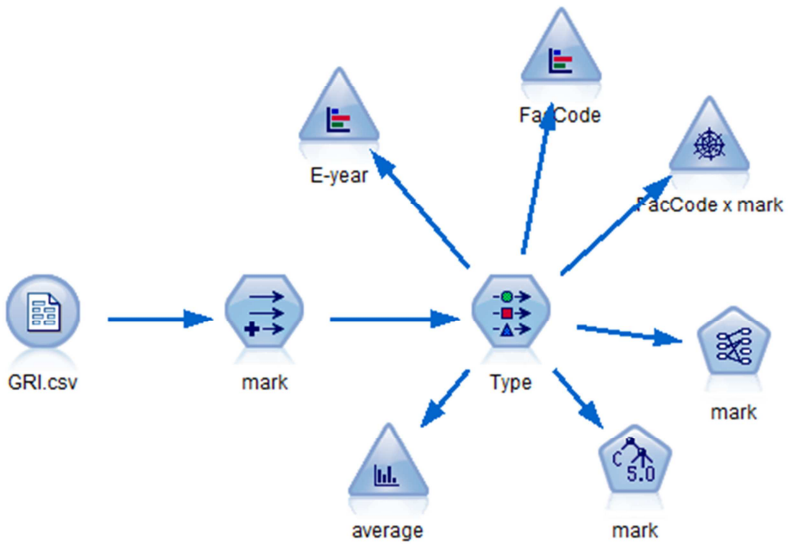
ردیف	نام فیلد	مجموعه مقادیر
۱	معدل کل	عدد اعشاری [۰-۲۰] تا شش رقم اعشار
۲	سال ورود	[۱۳۷۰...۱۳۸۶]
۳	نوع دوره	روزانه، شبانه
۴	موقعیت جغرافیایی	بومی، غیربومی
۵	جنسیت	زن، مرد
۶	رشته	ریاضی، فیزیک، شیمی، مهندسی کامپیوتر،.....
۷	مقطع تحصیلی	کارشناسی، ارشد، دکتری
۸	دانشکده	علوم پایه، شیمی، مهندسی، علوم انسانی، معماری و هنر
۹	سهمیه ورود به دانشگاه	منطقه ۱، منطقه ۲، منطقه ۳، آزادگان، خانواده شهدا،
۱۰	معدل ترم‌های متوالی	ترم ۱، ترم ۲، ترم ۳، ترم ۴، ....

به منظور این که فیلد معدل به عنوان خروجی الگوریتم استفاده شود باید از نوع دسته‌ای باشد . به این دلیل فیلد معدل به چهار دسته تقسیم شده است که در جدول ۲ نشان داد شده است.

جدول ۲- دسته‌بندی معدل

محدوده معدل	نام دسته	مقدار دسته
معدل بالای ۱۷	ممتاز	۱
معدل بین ۱۵ تا ۱۷	عادی	۲
معدل بین ۱۲ تا ۱۵	ضعیف	۳
معدل پایین‌تر از ۱۲	خیلی ضعیف	۴

برای اجرای الگوریتم GRI ، در نرم‌افزار کلمنتاین ابتدا باید یک استریم[۷] تعریف شود. استریم تولید شده برای اجرای الگوریتم GRI در شکل ۱ نشان داده شده است.



شکل ۱- نمایش استریم ایجاد شده برای اجرای الگوریتم GRI

نوع الگوریتم و مدل به کار گرفته شده از تب مدل‌سازی[۸] انتخاب می‌شود. این نود (GRI) قوانین وابستگی را از بین داده‌های خام به روشی که در بخش ۳ به صورت مفصل توضیح داده شده است کشف می‌کند و به صورت قوانین نمایش می‌دهد. قوانین تولید شده با ضریب اطمینان بالا در جدول ۳ نشان داده شده است.

٤-١ نتیجه‌هاي بدست آمده و تحلیل آن‌ها

خروجی‌های بدست آمده از اعمال الگوریتم GRI به علت حجیم بودن در بخش ضمائم آورده شده است و در این قسمت تنها نمونه‌ای از آن آورده شده است و به تحلیل قوانین تولیدی با ضریب اطمینان بالا پرداخته می‌شود. نتایج به دست آمده به این صورت است که سه ستون اول از سمت راست مقدم‌های هر قانون و ستون آخر تالی را نشان می‌دهد. فقط فیلد معدل در قسمت تالی ظاهر شده است. این نتایج بر اساس ضریب اطمینان مرتب شده‌اند.

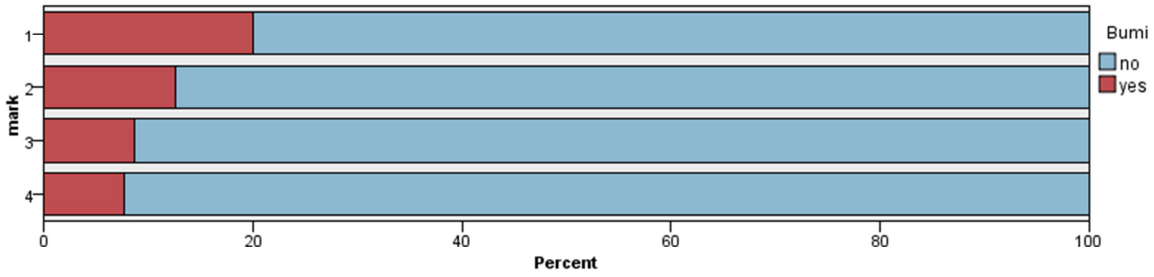
Consequent	Antecedent 1	Antecedent 2	Antecedent 3
mark = 1	Level = PH.D		
mark = 1	Course = daily	Level = PH.D	
mark = 1	Course = daily	Bumi = yes	Level = MS
mark = 3	Course = nightly	Bumi = no	Level = BS
mark = 3	Course = nightly	Level = BS	
mark = 1	Bumi = yes	Level = MS	
mark = 3	Course = nightly	Bumi = no	
mark = 3	Bumi = no	Level = BS	

شکل ۲- نمونه‌ای از خروجی الگوریتم GRI روی فیلدهای مقطع تحصیلی،نوع دوره و موقعیت جغرافیایی

قانون اول با ضریب اطمینان بالا بیان می‌کند که تمام دانشجویانِ مقطعِ دکتری دانشگاه کاشان معدل بالای ۱۷ دارند و دانشجویان ممتازی هستند؛ علاوه بر این با توجه به قوانین تولیدی هر چه سطح مقطع دانشجو بیشتر باشد، دانشجو در وضعیت تحصیلی بهتری قرار دارد دانشجویان ارشد در دسته ۲ و دانشجویان کارشناسی معمولاً در دسته ۳ قرار دارند که نشان از این دارد که دانشجویانی که در مقاطع تحصیلی بالاتر درس می‌خوانند با توجه به این که با رشته خود آشنایی بهتری نسبت به دانشجویان مقطع پایین‌تر دارد با انگیزه بیشتری تلاش می‌کنند و معدل بهتری دارند. علاوه بر این نسبت به دانشجویان سطح کارشناسی و کاردانی که ممکن است بدون آشنایی با رشته مورد نظر و تنها به علت اینکه از سد کنکور عبور کنند انتخاب رشته کرده و قبول شوند.

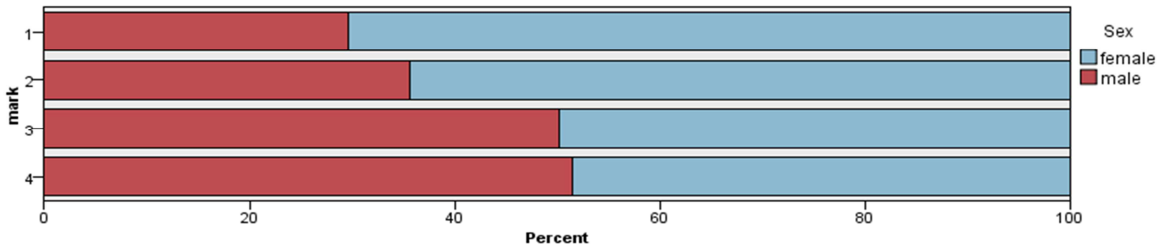
عامل موثر دیگری در قوانین بدست آمده و مورد بررسی قرار گرفته است؛ تأثیر بومی و غیر بومی بودن دانشجو در وضعیت تحصیلی آن می‌باشد. دانشجویان بومی نسبت به دانشجویان غیر بومی کارایی و معدل بهتری داشتند. که نمودار شکل ۳ نیز بیان کننده این مطلب می‌باشد. همان

طور که نشان داده شده است. با توجه به اینکه تعداد دانشجویان بومی کمتر است ولی درصد دانشجویان بومی که در دسته ۱ و ۲ قرار دارند بیشتر می‌باشد.



شکل ۳- نمودار درصد توزیع فراوانی فیلد بومی روی فیلد معدل

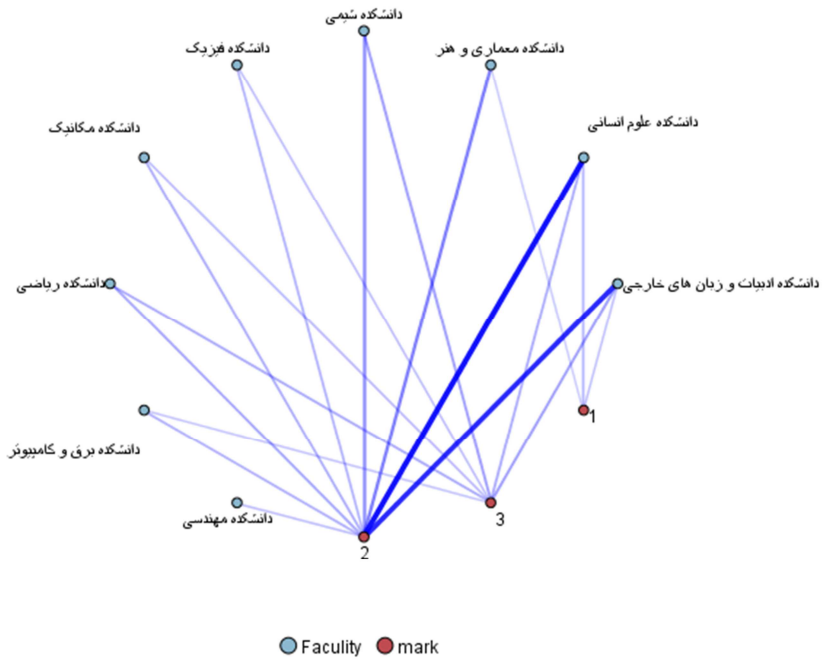
همان‌طور که در نمودار شکل ۴ نشان داده شده است عامل جنسیت در وضعیت تحصیلی تأثیری معناداری نداشت ولی به طور کلی خانم‌ها نسبت به آقایان در وضعیت تحصیلی بهتری قرار داشتند.



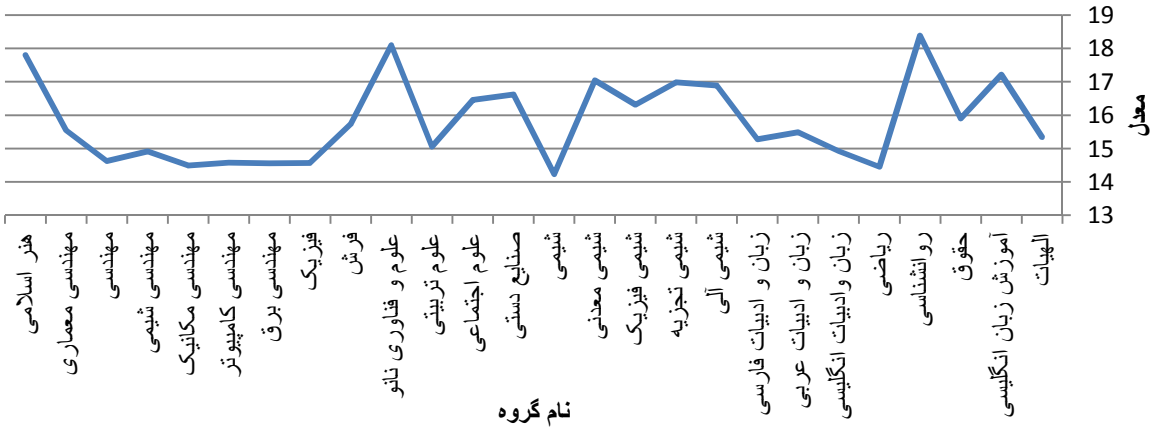
شکل ۴- نمودار توزیع فراوانی فیلد جنسیت روی فیلد معدل

با استفاده از نود وب[۹] می‌توان مقادیری که با هم ارتباط قوی دارند را به صورت بسیار گویا نمایش داد. به منظور بررسی وضعیت تحصیلی دانشجویان در دانشکده‌های مختلف نود وب برای دو فیلد معدل دسته‌ای و دانشکده رسم شده است. گراف ایجاد شده به عنوان خروجی نود وب در شکل ۲ نشان داده شده است. در این نوع گراف، فیلدهایی که در بیشتر مواقع با هم آمده‌اند یعنی ارتباط قوی‌تری دارند با خط مشکی ضخیم‌تر و فیلدهایی که ارتباط ضعیف‌تری دارند به صورت نقطه چین نمایش داده می‌شوند. شکل ۵ نشان می‌دهد که معدل دانشجویان دانشکده مهندسی و برق و کامپیوتر بیشتر در دسته ۳ قرار می‌گیرند و دانشجویان دانشکده انسانی و ادبیات و زبان‌های خارجی در دسته‌های ۱ و ۲ قرار می‌گیرند که نشان از سخت‌تر بودن رشته تحصیلی آن‌ها نسبت به رشته‌های دیگر می‌باشد علاوه بر این اگر سخت بودن درجه امتحان را در نظر بگیریم و با میانگین نمره در دانشگاه‌های دیگر مقایسه کنیم متوجه می‌شویم که اساتید این گروه‌ها امتحانات به نسبت سخت‌تر و دور از انتظار از دانشجویان خود می‌گیرند و این باعث شده تا معدل آن‌ها نسبتاً پایین باشد.

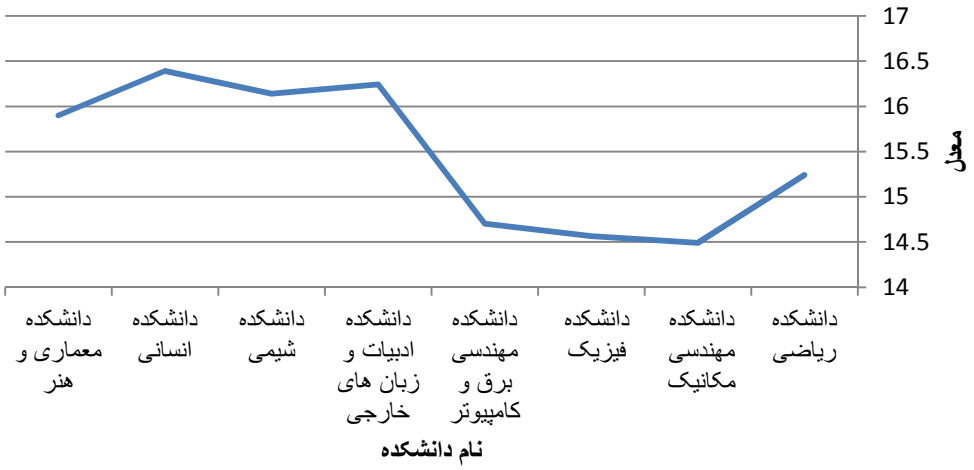
در نمودار شکل ۶ میانگین نمرات در گروه‌های آموزشی مختلف نشان داده شده است که تفاوت زیادی در بعضی از گروه‌های آموزشی وجود دارد که می‌توان به صورت جداگانه به تحلیل آن‌ها پرداخت و برای مقایسه و ارزیابی گروه‌های مختلف آموزشی از آن استفاده کرد. به طور کلی می‌توان گفت که عواملی از جمله تعداد آزمایشگاه‌ها، تعداد اعضای هیأت علمی، فضای فیزیکی ، قدمت گروه، میانگین سنی اعضای هیأت علمی، نسبت تعداد دانشجو به اعضای هیأت علمی در رتبه بندی گروه‌های مختلف آموزشی از جهت معدل دانشجویان تأثیر گذار باشد. در این پژوهش تنها به تأثیر یکی از این عوامل که تعداد اعضای هیأت علمی می‌باشد پرداخته شده است. همان‌طور که در نمودار شکل‌های ۷ و ۸ نشان داده شده است رابطه معناداری بین میانگین نمره دانشجویان دانشکده و تعداد اعضای هیأت علمی وجود دارد و هر چه تعداد اعضای هیأت علمی بیشتر باشد میانگین معدل دانشجویان بیشتر است.



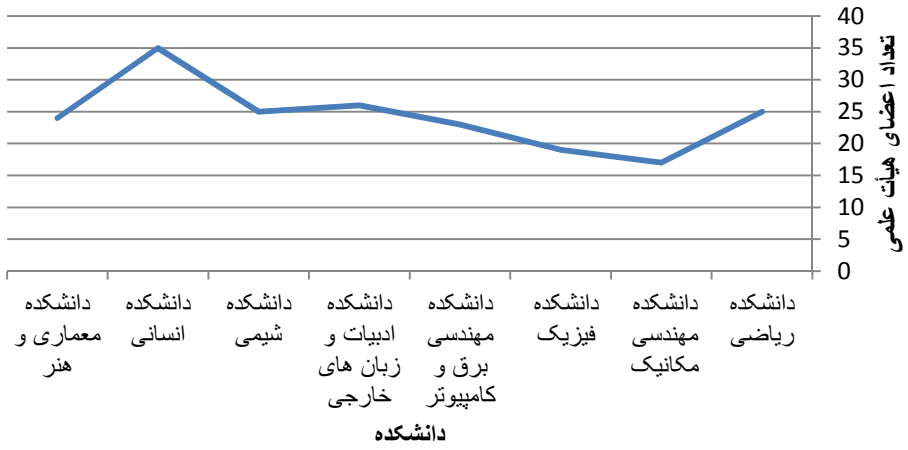
شکل ۵- ارتباط فیلد دانشکده با فیلد معدل دسته‌ای



شکل ۶- نمودار میانگین معدل دانشجویان هر گروه

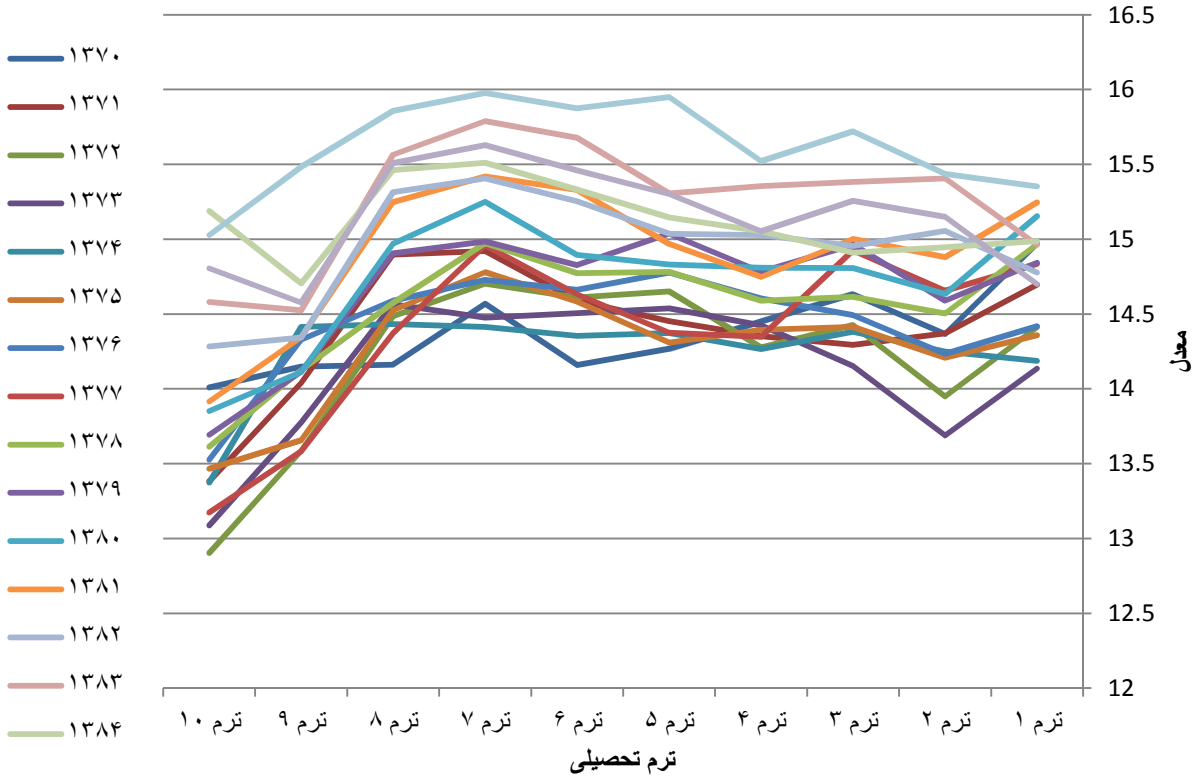


شکل ۷- نمودار میانگین معدل دانشجویان هر دانشکده



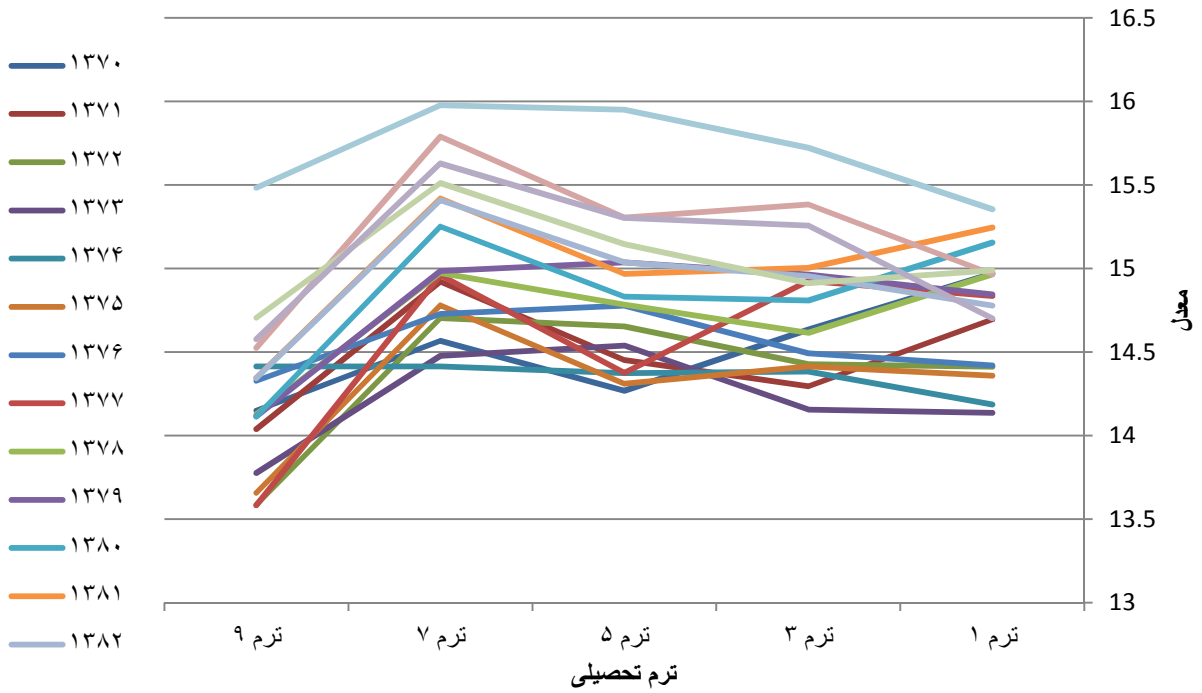
شکل ۸- تعداد اعضای هیأت علمی در دانشکده‌های مختلف

در نمودار شکل شماره ۹ میانگین نمرات دانشجویان دانشگاه کاشان از سال‌های ۱۳۷۰ تا ۱۳۸۶ نشان داده شده است بر طبق این نمودار معدل دانشجویان در ترم‌های اولیه تا ترم ۶ رو به رشد بوده و بعد از آن به علت مشکل شدن دروس اتخاذی روند کاهش را داشته است؛ افت شدیدی در ترم‌های ۹ و ۱۰ به علت این است که اکثر دانشجویان در این دو ترم به نسبت دانشجویان ضعیفی می‌باشند.



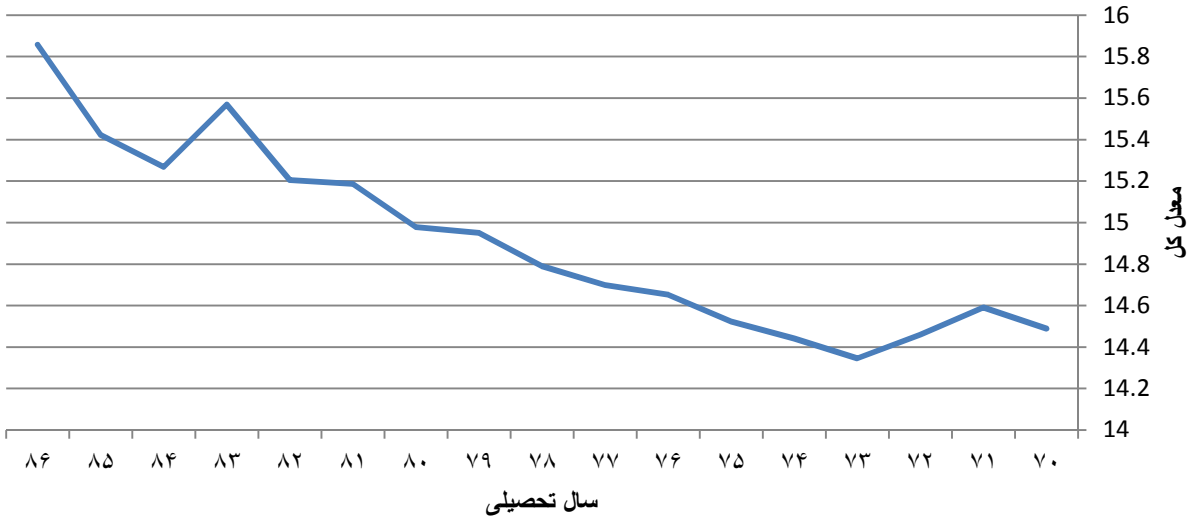
شکل ۹- نمودار میانگین معدل دانشجویان ورودی سال‌های مختلف در ترم‌های تحصیلی متفاوت

نکته دیگری که در این نمودار بدست می‌آید این است که به طور کلی معدل دانشجویان در ترم‌های فرد نسبت به ترم‌های زوج وضعیت بهتری دارد و علت آن این است که در ترم‌های فرد به علت طولانی‌تر بودن ترم تحصیلی دانشجویان کارایی بهتری دارند و علاوه بر آن دانشجویان، ترم فرد را با انرژی بیشتری بعد از تعطیلات تابستان شروع می‌کنند و وضعیت تحصیلی بهتری دارند. علاوه بر این در ترم‌های زوج به علت تعطیلات فراوان دروس به صورت فشرده‌تر ارائه می‌شود. بنابراین در برنامه‌ریزی تحصیلی بهتر است به علت آمادگی بیشتر دانشجویان درس‌های سنگین‌تر در ترم-های فرد برای دانشجویان ارائه شود. نکته قابل ذکر دیگر در این نمودار افت شدیدتر درسی دانشجویان در ترم ۲ می‌باشد که می‌توان علت آن‌را عدم آشنایی با محیط و شرایط جدید و دوری از خانواده ذکر کرد. همان‌طور که در نمودار مشخص است این افت در سال‌های اخیر کاهش پیدا کرده است که می‌توان علت آن را توسعه فناوری ارتباطات در سال‌های اخیر دانست که امکان ارتباط با خانواده بیشتر شده و دانشجویان فشار کمتری را احساس می‌کنند. همان طور که در شکل ۱۰ نشان داده شده است با حذف میانگینِ معدل در ترم‌های زوج، میانگین معدل رشد یکنواختی دارد.

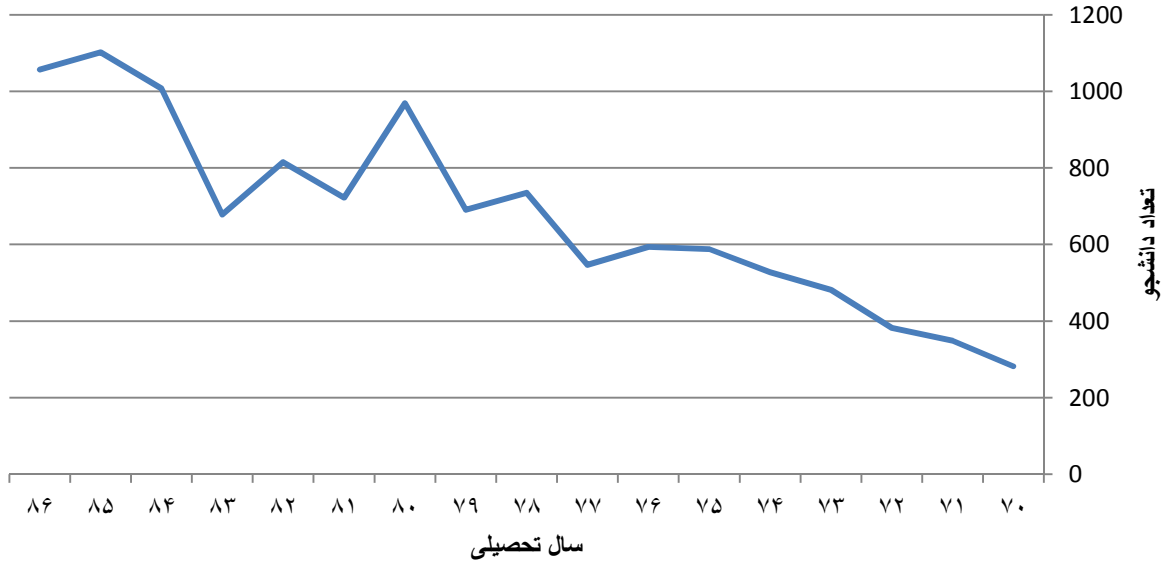


شکل ۱۰- نمودار میانگین معدل دانشجویان سال‌های ورودی مختلف درترم های تحصیلی فرد

نمودارهای شکل‌های ۱۱ و ۱۲ نشان می‌دهد که با وجود اینکه روند جذب دانشجو در سال‌های متمادی در دانشگاه کاشان روبه رشد بوده است میانگین معدل دانشجویان نیز روبه رشد بوده است که نشان می‌دهد دانشگاه کاشان در حال توسعه می‌باشد و سیاست‌های مدیریتی آن هم کارآمد بوده است و علاوه بر آن با توسعه دانشگاه کاشان و افزایش امکانات، دانشجویان قوی‌تری، دانشگاه کاشان را برای ادامه تحصیل انتخاب کرده‌اند. احتمالاً کیفیت و تعداد اعضای هیأت علمی در دانشگاه کاشان نیز روبه افزایش بوده است.

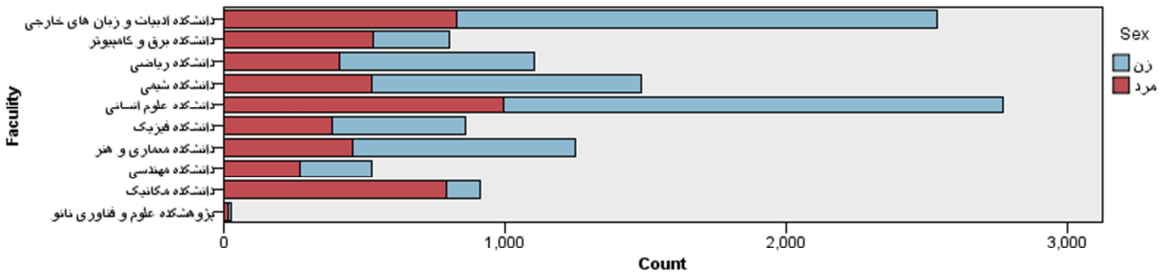


شکل ۱۱- معدل کل دانشجویان ورودی سال‌های مختلف



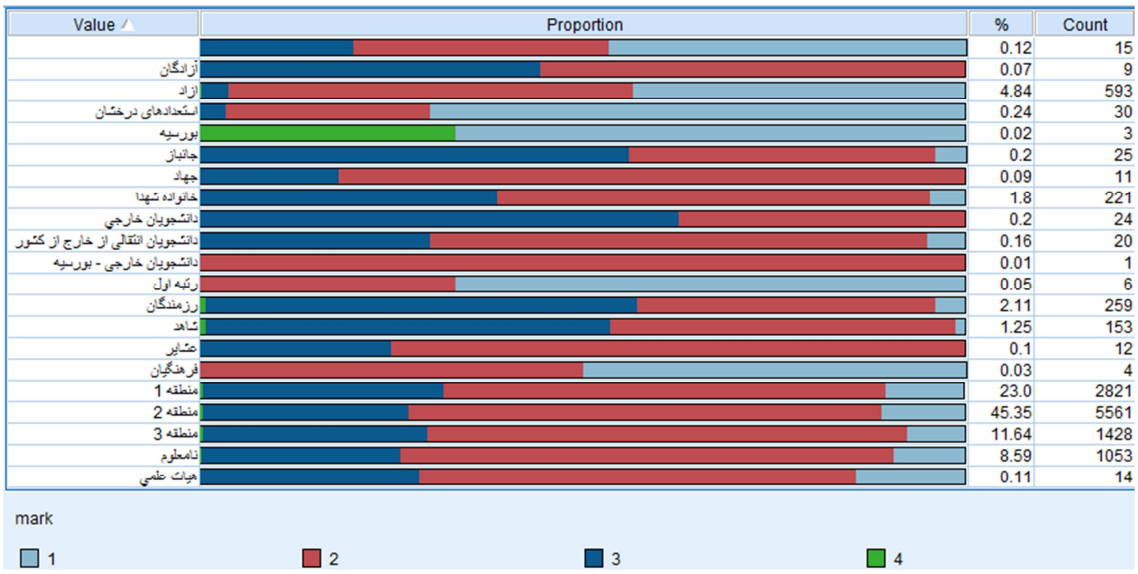
شکل ۱۲- تعداد دانشجوی ورودی در سال‌های متمادی

نمودار شکل ۱۳ میزان توزیع دانشجویان دختر و پسر در دانشکده‌های مختلف را نشان می‌دهد. همان‌طور که مشخص است درصد دخترها در رشته‌های علوم انسانی نسبت به پسرها بیشتر و در رشته‌هایی مثل مهندسی مکانیک که رشته‌ای تقریباً مردانه است شانس ورود خانم‌ها در این رشته کم‌تر است.



شکل ۱۳- نمودار توزیع فراوانی جنسیت در دانشکده‌های مختلف

همان‌طور که در شکل ۱۴ نشان داده شده است سهمیه ورود به دانشگاه تأثیر معناداری در وضعیت تحصیلی دانشجویان دانشگاه کاشان ندارد.



شکل ۱۴- نمودار درصد توزیع معدل دانشجویان سهمیه‌های مختلف

۵. نتیجه‌گیری کلی و کارهای آتی

سیستم‌های آموزش عالی از طریق داده‌کاوی قادرند که اثربخشی سیستم‌های آموزشی را حداکثر کنند، پذیرش و مدیریت ثبت‌نام را بهبود دهند، نرخ حذف دانشجویان را حداقل کنند، موفقیت دانشجویان را افزایش دهند و هزینه فرآیندهای سیستم را کاهش دهند. یک موسسه آموزش از طریق داده‌کاوی قادر خواهد بود که مزیت رقابتی خود را افزایش دهد و به استانداردهای بالاتری در سطح آکادمیک برسد.

در این مقاله به بررسی وضعیت تحصیلی دانشجویان دانشگاه کاشان از سال ۱۳۷۰ تا ۱۳۸۶ پرداخته شده است. با اعمال الگوریتم GRI روی داده‌های آموزش دانشگاه کاشان به قوانینی دست یافتیم که می‌تواند به مدیران آموزشی سطح دانشگاه، دانشکده‌ها و گروه‌های آموزشی مختلف کمک کرده و به آن‌ها در اتخاذ تصمیم، حذف و تغییر قانون و سیاست‌ها . برنامه‌ریزی‌های آموزشی در جهت بهبود هر چه بهتر فرآیندهای آموزشی کمک کند.

به طور کلی می‌توان با توجه به خروجی‌های حاصل از داده‌کاوی به نتایج زیر رسید:

- تعداد اعضای هیأت علمی یا میانگین معدل دانشجویان ارتباط مستقیم دارد.
- بومی بودن برای دانشجویان کارشناسی در شهر کاشان، عاملی می‌باشد تا دانشجو در وضعیت تحصیلی خوبی قرار بگیرد.
- دانشجویانی که در مقطع روزانه تحصیل می‌کنند نسبت به دانشجویان شبانه در وضعیت تحصیلی بهتری قرار دارند.
- دانشجویان در ترم‌های فرد وضعیت تحصیلی بهتری دارند و بهتر است مسئولانِ برنامه‌ریزی آموزشی درس‌های سنگین‌تر را در ترم مهر ارائه کنند.
- توسعه وسایل ارتباطی تأثیر مستقیم بر ارتقاء کیفیت آموزشی دانشجویان دارد.
- سهمیه ورود به دانشگاه تأثیر معناداری در وضعیت تحصیلی دانشجو ندارد.
- جنسیت در وضعیت تحصیلی دانشجو تأثیر چندانی ندارد ولی به طور کلی دانشجویان دختر در وضعیت بهتری نسبت به دانشجویان پسر قرار دارند.
- با وجود روند روبه رشد جذب دانشجو در دانشگاه کاشان، وضعیت تحصیلی دانشجویان نیز روبه رشد است.
- هرچه مقطع تحصیلی دانشجو بیشتر باشد دانشجو در وضعیت تحصیلی بهتری قرار دارد به طوری که با ضریب اطمینان بالا دانشجویان دکتری در دسته ۱ دانشجویان ارشد در دسته ۲ و دانشجویان کارشناسی در دسته ۳ قرار گرفتند.
- درصد بیشتری از دانشجویان دانشکده انسانی و ادبیات و علوم انسانی را خانم‌ها تشکیل داده در حالی که در گروه مهندسی مکانیک درصد بسیار کمی را خانم‌ها تشکیل می‌دهند.
- دانشجویان دانشکده شیمی و انسانی و ادبیات و زبان‌های خارجی اکثراً دانشجویان ممتازی می‌باشند.

- اساتید گروه‌های مهندسی و ریاضی و شیمی برای ارزیابی دانشجویان خود نسبت به گروه‌های دیگر از آزمون مشکل‌تری استفاده می‌کنند.

با توجه به قابلیت‌های دستاورد داده‌کاوی، می‌توان از تکنیک‌های موجود در این علم به منظور بهبود برنامه‌ریزی و حل مسائل آموزشی استفاده کرد. از جمله زمینه‌هایی که می‌توان به عنوان تحقیقات آتی به آن اشاره کرد، عبارتند از:

- بررسی تأثیر عامل‌هایی از جمله تعداد آزمایشگاه‌ها، فضای فیزیکی ، قدمت گروه، میانگین سنی اعضای هیأت علمی، نسبت تعداد دانشجو به اعضای هیأت علمی در رتبه بندی گروه‌های مختلف آموزشی از جهت معدل دانشجویان
- اثر فاکتورهایی نظیر نوع سهمیه ورود به دانشگاه، استفاده از خوابگاه دانشجویی، مسافت محل سکونت دانشجو تا دانشگاه، نوع دانشگاه (دولتی، آزاد)، هزینه تحصیل (خوابگاه، مسافت، تهیه مقالات و کتب درسی، هزینه ثبت‌نام‌ترم جدید)، وضعیت تأهل، زمان ازدواج، وضعیت و رشته تحصیلی همسر و نیز تعداد فرزندان را می‌توان بر روی تحصیل و علاقه‌مندی به ادامه تحصیل بررسی کرد.
- اعمال کاربردهای دیگر داده‌کاوی (ازجمله پیش‌بینی ثبت‌نام دانشجویان در یک درس) در آموزش عالی روی داده‌های آموزشی دانشگاه کاشان
- اعمال الگوریتم‌های داده‌کاوی روی داده‌های آموزشی دیگر دانشگاه‌ها برای بررسی و مقایسه گروه‌های مختلف درسی در دانشگاه‌های مختلف از جمله دانشگاه‌های آزاد و پیام نور

می‌توان طرح پژوهشی مطرح شده در مقاله را بر روی جامعه دانشجویان سایر رشته‌های آموزشی عالی اعمال کرد و با بررسی نتایج بدست آمده در جهت حل مسائل آموزشی برآمد.

٦. ضمائم

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2	Antecedent 3
25	0.200	88.000	8.000	mark = 1	Level = PH.D		
25	0.200	88.000	8.000	mark = 1	Course = daily	Level = PH.D	
67	0.550	75.000	6.818	mark = 1	Course = daily	Bumi = yes	Level = MS
2420	19.730	65.000	1.300	mark = 3	Course = nightly	Bumi = no	Level = BS
2882	23.500	63.000	1.260	mark = 3	Course = nightly	Level = BS	
115	0.940	63.000	5.727	mark = 1	Bumi = yes	Level = MS	
2661	21.700	60.000	1.200	mark = 3	Course = nightly	Bumi = no	
9861	80.410	56.000	1.120	mark = 3	Bumi = no	Level = BS	

شکل ۱۵- خروجی الگوریتم GRI بر روی فیلدهای مقطع تحصیلی، موقعیت جغرافیایی و دوره

Instances	Support	Confidence	Lift	Consequent	Antecedent 1	Antecedent 2	Antecedent 3
23	0.190	91.000	8.273	mark = 1	FacCode = 31		
25	0.200	88.000	8.000	mark = 1	Level = PH.D		
14	0.110	86.000	7.818	mark = 1	E-year = 1385	Level = PH.D	
963	7.850	76.000	1.520	mark = 3	Level = BS	FacCode = 11	
1247	10.170	74.000	1.480	mark = 3	Level = BS	FacCode = 15	
514	4.190	68.000	1.360	mark = 3	E-year = 1373	Level = BS	
518	4.220	68.000	1.360	mark = 3	E-year = 1373		
1102	8.990	67.000	1.340	mark = 3	FacCode = 11		
416	3.390	66.000	1.320	mark = 3	E-year = 1372		
30	0.240	63.000	5.727	mark = 1	E-year = 1385	Level = MS	FacCode = 12
116	0.950	61.000	5.545	mark = 1	Level = MS	FacCode = 12	
220	1.790	60.000	1.579	mark = 2	Level = MS	FacCode = 15	
272	2.220	58.000	1.526	mark = 2	Level = kardani	FacCode = 14	
272	2.220	58.000	1.526	mark = 2	Level = kardani		
232	1.890	56.000	5.091	mark = 1	E-year = 1386	Level = MS	
78	0.640	55.000	5.000	mark = 1	E-year = 1384	Level = MS	
1251	10.200	54.000	1.421	mark = 2	FacCode = 14		
974	7.940	54.000	1.421	mark = 2	Level = BS	FacCode = 14	
11118	90.660	54.000	1.080	mark = 3	Level = BS		
86	0.700	51.000	4.636	mark = 1	E-year = 1383	Level = MS	

شکل ۱۶- خروجی الگوریتم GRI بر روی فیلدهای مقطع تحصیلی، موقعیت جغرافیایی، دوره، دانشکده و سال تحصیلی

۷. منابع

[۱] منعمی بیدگلی، عطیه؛ یوسفان، احمد؛**"کشف قوانین موجود در پایگاه داده آموزش با استفاده از الگوریتم‌های GRI ،CART و NaiveBayse"**، شماره ۱۴۷۸،۶۰۵، دانشکده مهندسی، دانشگاه کاشان، ۱۳۸۹

[۲] سعیدی، احمد؛ **"داده کاوی، مفهوم و کاربرد آن در آموزش عالی"** ، نامه آموزش عالی، شماره ۸،۱۸، اسفند ۱۳۸۴

[۳] یقینی مسعود؛ حیدری سمیه؛ **"داده‌کاوی جهت ارتقاء و بهبود فرآیندهای سیستم آموزش عالی"**، دومین کنفرانس داده‌کاوی ایران، دومین کنفرانس، دانشگاه صنعتی امیرکبیر، ۷- ۱۰، ۱۳۸۷.

[۴] Tan, P.-N; Steinbach, M; Kumar, V*; "Introduction to Data Mining*" , Addison-Wesley, 2005.

[۵] "*Welcome to Clementine*"; piano.dsi.uminho.pt/disciplinas/LIGIA/.../tutorial/clemtut.htm

[۶] "PSPP"; *http://www.gnu.org/software/pspp/manual/pspp.html*, 2005

[۷] Pallant, J. F; " *SPSS Survival Manual: A Step By Step Guide to Data Analysis Using SPSS for Windows*"*(Version 12)*, Crows Nest, Australia, Allen & Unwin,2005.

[۸] Aksenova, S.S; Zhang ,Du; Meiliu, Lu; "*Enrollment Prediction through Data Mining"* IEEE International Conference on Information Reuse and Integration, Sept2006, pp. 510-515, 2006.

[۹] Smyth P, Goodman RM. "*An information theoretic approach to rule induction from databases*" IEEE Transactions on Knowledge and Data Engineering.;4(4):301–316. doi: 10.1109/69.149926; 1992.

[۱۰] Berry, Michael; Berry, Gordon; *"Data Mining Techniques*" (For Marketing, Sales, and Customer Relationship Management), Second Edition, Wiley, Inc, 2004.

[۱۱] Larose, Daniel; "*Discovering Knowledge In Data"*, Wiley & Sons, Inc, 2005.

[۱۲] Luan, Jing; "*Data Mining and Knowledge Management in Higher Education*", Knowledge and Data Management White Papers, Presentation at AIR Forum in CabrilloCollege,Toronto, Canada, 6- 16, 2002.

[۱۳] Romero,C; Ventura,S; "*Educational data mining: A survey from 1995 to 2005*"' , Elsevier, Expert Systems with Applications , 33, 135- 146, 2007.

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

SPSS Inc . "*Clementine 12". (Algorithms Guide”*. http://www.spss.com) [۱۵]
Mohammad Reza Beikzadeh, Naeimeh Delavari; "*A New Analysis Model for Data Mining Processes in Higher Educational Systems*",Proceedings of M2USIC, 2004 [۱۶]

[1] Acceptance Requirement Designer
[2] Boolean
[3] Predictive Decision Tree
[4] Categorical
[5] Support
[6] Confidence
[7] Stream
[8] Modeling
[9] Web