dataacademy.ir

کاربرد داده کاوی بر روی داده های آموزش عالی دانشجویان دانشکده دندانپزشکی شهر رشت با استفاده از تکنیکهای طبقه بندی و خوشه بندی

علی زواره[۱]، امیر کوچکی[۲] و علی رهنما رودپشتی[۳]
[۱] کارشناس ارشد مدیریت فناوری اطلاعات، دانشکده دندانپزشکی رشت ، a_zavareh@gums.ac.ir
[۲] مدرس بخش فناوری اطلاعات، سازمان فنی و حرفه ای شهر رشت، SecondAuthor@Email
[۳] مدرس، دانشگاه علمی کاربردی رشت، ThirdAuthor@Email

چکیده – هدف بررسی اطلاعات آموزشی دانش آموختگان رشته دندانپزشکی بین سالهای ۱۳۷۵ تا ۱۳۹۰ و بدست آوردن مدلی جهت پیش بینی دانشجویان ممتاز(قوی) و دانشجویان تحت ریسک مشروطی (ضعیف) و عوامل مؤثر بر نمرات و رفتار آموزشی آنها است.

داده کاوی داده ها با استفاده از از یک فرآیند چند مرحله ای بنام *CRIPS* صورت گرفت در این تحقیق از مدل *CHAID* برای پیش بینی از طریق طبقه بندی و مدل های *Two-Step* و *K-Means* برای خوشه بندی استفاده گردید.

نتایج این مطالعه نشان داد فیلدهای جنس، سن، سهمیه، معدل کل، تعداد ترم مشروطی و فیلد *pass* در خوشه بندی بعنوان فیلدهای با اهمیت بدست آمدند. مدل*Two-Step* بهتر از *K-Means* شناخته شد و چهار خوشه شامل: خوشه ۱(بهداشتکاران)، خوشه ۲(دانشجویان متوسط با ۱۰۰٪ جمعیت مرد)، خوشه ۳(دانشجویان ضعیف) و خوشه ۴(دانشجویان قوی) تولید نمود. هر دو مدل در رابطه با خوشه بندی دانشجویان متوسط، ضعیف عمل نمودند. مدل *CHAID* در تکنیک طبقه بندی برای پیش بینی دانشجویان قوی(ممتاز)، دانشجویان متوسط و دانشجویان ضعیف قوانینی با درصد صحت بالا تولید نمود.

داده کاوی آموزشی نشان داد که از مدل *CHAID* در تکنیک طبقه بندی، جهت پیش بینی دانشجویان ضعیف و قوی و دانشجویانی که در خطر مشروطی هستند و از مدل *Two-Step* می توان در تکنیک خوشه بندی جهت بررسی داده ها و رفتار دانشجویان، برای بهبود کیفیت و برنامه ریزی استراتژیک آموزشی استفاده نمود.

کلید واژه- تکنیک خوشه بندی، تکنیک طبقه بندی، داده کاوی آموزشی، Data Mining

۱- مقدمه

با رشد فناوری اطلاعات و روش های تولید و جمع آوری داده ها، داده های مربوط به تبادلات تجاری، کشاورزی، اینترنت، جزئیات مکالمات تلفنی، داده های پزشکی و غیره سریعتر از هر روز جمع آوری و انبارش می شوند. لذا از اواخر دهه ۸۰ میلادی بشر به فکر دست یابی به اطلاعات نهفته در این پایگاه داده های حجیم افتاد، زیرا سیستمهای سنتی قادر به این کار نبودند. به دلیل رقابت در عرصه های سیاسی، نظامی، اقتصادی، علمی و اهمیت دست یابی به اطلاعات در کمترین زمان و بدون دخالت انسان علم تجزیه و تحلیل داده ها یا داده کاوی پا به عرصه گذاشت.

داده کاوی[۱] (DM) فرآیندی است که در آغاز دهۀ ۹۰ مطرح شد و با نگرشی نو، به مسئله استخراج اطلاعات از پایگاه داده ها پرداخت. از سال ۱۹۹۵ داده کاوی به صورت جدی وارد مباحث آمار شد و در سال ۱۹۹۶، اولین شمارۀ مجلۀ کشف دانش و معرفت از پایگاه داده ها[۲](KDD) منتشر شد. محققانی نظیر

[1] Data Mining
[2] Knowledge Discovery in Database

dataacademy.ir

براچمن[3] و آناند[4] کلیه مراحل واقع گرایانه و رو به جلو کشف دانش از پایگاه داده ها را تشخیص دادند.

در حال حاضر، داده کاوی مهمترین فناوری جهت بهره برداری موثر از داده های حجیم است و اهمیت آن رو به فزونی است. به طوریکه تخمین زده شده است که مقدار داده ها در جهان هر ۲۰ ماه به حدود دو برابر می رسد. در یک تحقیق که بر روی گروه های تجاری بسیار بزرگ در جمع آوری داده ها صورت گرفت مشخص گردید که ۱۹ درصد از این گروه ها دارای پایگاه داده هایی با سطح بیشتر از ۵۰ گیگابایت می باشند و ۵۹ درصد از آنها انتظار دارند که در آینده ای نزدیک در چنین سطحی قرار گیرند[۱]. در سالهای اخیر تحقیقات زیادی در زمینه بکارگیری فرآیند داده کاوی در امر آموزش صورت گرفته است.. این زمینه تحقیقاتی جدید، داده کاوی آموزشی[5] (EDM) نامیده می شود که به امر توسعه روشهای کشف دانش از داده های محیط های آموزشی خصوصاً دانشجویان می پردازد.

۱-۱- داده کاوی آموزشی

داده کاوی آموزشی به عنوان ناحیه پژوهش علمی ای تعریف شده است که در حول روش های توسعه برای ایجاد کشفیاتی که در نوع خود منحصر به فرد است، متمرکز گردیده است. استفاده از آن روشی برای درک بهتر دانشجویان و مجموعه ای که آنها در آنجا آموزش می بینند، می باشد.

به عنوان مثال، در داده کاوی بررسی دانشجویان در استفاده کردن از نرم افزارهای آموزشی، ممکن است به طور همزمان داده ها را در نظر گرفت و می تواند سطح زمان فشردن صفحه کلید، سطح پاسخ، سطح جلسه، سطح دانشجویان، سطح کلاس درس، و سطح مدرسه، داده های ارزشمندی باشند. بحث زمان، ترتیب و مفاد نیز نقش مهمی در مطالعه داده های آموزشی ایفا می کند.

داده کاوی آموزشی به عنوان یک زمینه تحقیقاتی مستقل در سال های اخیر پدید آمده است، اوج آن در سال ۲۰۰۸ با ایجاد کنفرانس بین المللی سالانه در مورد داده کاوی آموزشی، و مجله های آموزشی داده کاوی می باشد[۲].

استفاده از داده کاوی در سیستم های آموزشی را با توجه به گرایش آن به سه گروه می توان تقسیم نمود:

۱-گرایش به سمت دانشجویان: هدف یادگیری فعال، بهبود یادگیری و پیشنهاد تجارب یادگیری خوب و پیشنهاد کوتاه کردن مسیر یادگیری به یادگیرنده و موفقیت آنها است.

۲- گرایش به سمت آموزش دهندگان: هدف بازخورد نتیجه آموزش، مدرسین. ارزیابی ساختار مطالب دوره و اثربخشی آن در روند آموزش، طبقه بندی یادگیرنده ها، کشف اطلاعات مفید که سبب برنامه ریزی در سازماندهی دروس میشود.

۳-گرایش به سمت مسئولان و مدیران دانشگاهی: هدف داشتن اقداماتی در مورد چگونگی سازماندهی بهتر منابع رسمی و سازمانی و پیشنهاد آموزشی می باشد.

هدف ما در این تحقیق گرایش نوع اول یعنی گرایش به سمت دانشجویان می باشد. کشف دانش از داده های ذخیره شده توسط داده کاوی، در واقع یک فرآیند پشت سرهم هست که دارای مراحل زیر می باشد:

۱-۲- مراحل داده کاوی

تعیین اطلاعات گذشته

پاک سازی داده ها و پردازش اولیه- در این مرحله خطاهای داده ها تصحیح می شوند و داده های اشتباه جایگزین میشوند. این مرحله ممکن است تا ۶۰ درصد از زمان داده کاوی را در برگیرد.

یکپارچه سازی داده ها- معمولاً داده ها از منابع متفاوتی جمع آوری می شوند باید به صورتی در آیند که یک مخزن از داده های مناسب ایجاد شود تا بتوان عملیات داده کاوی را بهتر انجام داد.

انتخاب مجموعه داده های هدف.

یافتن ویژگیهای مورد استفاده و تعیین ویژگی های جدید

نمایش داده ها به صورتی که بتوان برای داده کاوی استفاده نمود.

انتخاب عملیات داده کاوی (طبقه بندی، خوشه بندی، پیش بینی و غیره)

انتخاب روش داده کاوی (شبکه های عصبی، درخت تصمیم و نظایر آن).

داده کاوی و جستجو برای یافتن الگوی مناسب.

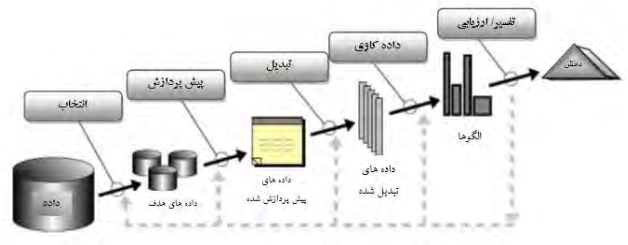
ارزیابی و تحلیل الگوی به دست آمده و حذف الگوهای نامناسب.

تفسیر نتایج داده ها و استنتاج از اطلاعات با ارزش.

[3] Brachman
[4] Anand
[5] Education Data Mining

dataacademy.ir

در شکل ۱ میتوان مراحل داده کاوی را به اختصار نشان داد[۳].



شکل ۱- مراحل داده کاوی

Jan M. و Mykola Pechenizkiy ، Gerben W. Dekker Vleeshouwers (سال ۲۰۰۹) در دومین کنفرانس داده کاوی آموزشی در کشور اسپانیا مقاله ای با عنوان « پیش بینی دانشجویانی که ترک تحصیل می نمایند: مطالعه موردی» ارائه نمودند. آنها بر روی دانشجویان مهندسی برق این مطالعه را انجام دادند و بر روی کسانی که پس از ترم اول ترک تحصیل می نمایند.

در این مطالعه اطلاعات طول سالهای ۲۰۰۰ - ۲۰۰۹ که شامل اطلاعات مربوط به تمام دانشجویان که در برنامه رشته مهندسی برق درگیر جمع آوری شده بودند. در انتخاب مجموعه داده های هدف از اطلاعات ۶۴۸ دانشجو استفاده نمودند.

در مطالعه، از تکنیکهای اثربخش طبقه بندی توسط نرم افزار WEKA استفاده شد. الگوریتم درخت تصمیم به نظر می رسد کمی بهتر از الگوریتم های دیگر : الگوریتمهای CART و C4.5 تنها مواردی بودند که قادر به بهبود مدل خط پایه بصورت قابل توجهی بودند.

هنگامی که به مقایسه بخش های مختلف مجموعه داده پرداخته شد، مشخص شد که پیش بینی نمرات دانشگاه نسبت به داده های پیش دانشگاهی بهتر است. به نظر می رسد ویژگی های داده های پیش دانشگاهی یا خیلی بی ربط و یا به شدت در ارتباط است، و در نتیجه دقت پیش بینی با استفاده از میانگین نمرات دروس علوم حدود ۷۰ درصد است. مقایسه نتایج مدل های مختلف نشان میدهد که پیش بینی نمره درس جبر خطی بدنبال پیش بینی نمره درس حساب دیفرانسیل و انتگرال امکان پذیر است. نتیجه دیگری که گرفته شد مشخص گردید اگر در مدل مورد بررسی مجموعه داده ها را بطور جداگانه بررسی کنیم مفید نمی باشد بلکه بهتر است مجموعه داده ها را بصورت ترکیبی بکار برد[۴].

ونوس شکورنیاز، آرش حاجی علی اکبری (۱۳۸۷ دومین کنفرانس انجمن داده کاوی ایران) در مقاله ای با عنوان «خوشه بندی داده های آماری دانشجویان دانشگاه علم و صنعت و استخراج نمایه ساز توصیفی برای دانشجویان موفق» از اطلاعات آماری ۳۱۱۱ دانشجوی ورودی ۸۲ تا ۸۵ برای سالهای ۸۲ تا ۸۵ استفاده نمودند و از الگوریتهای آماری و میانگین کا توسط نرم افزار SQL با توجه به معیارهای خاص، دانشجویان را از لحاظ موفقیت، پیشرفت، آینده شغلی و ... خوشه بندی نمودند و بر روی تعداد ترم مشروطی تمرکز نمودند و ویژگی های دانشجویانیکه مشروط میشوند را مشخص کردند، برای مشخص نمودن ارتباطات میان مشخصه های ورودی تاثیر گذار به منظور تعیین قدرت وابستگی مشخصه ها در پیش بینی هدف، از تکنیک درخت تصمیم به منظور پیش بینی و تحلیل معدل کل دانشجویان و همچنین تعداد ترمهای مشروطی آنها استفاده گردید. نتایج بدست آمده از تکنیکهای درخت تصمیم و خوشه بندی عبارت است:

دو خوشه شماره ۱ و ۲ در بردارنده ممتازترین دانشجویان با بهترین معدل ها می باشند. معدل ترم اول دانشجویان موفق دانشگاه علم و صنعت، با تفاوت بسیار زیاد، بالاتر ازسایر دانشجویان می باشد. این موارد به پیگیری و تلاش آنها از سالهای اولیه تحصیل اشاره دارد. این خوشه ها پایین ترین درصد دانشجویان مربوط به مهندسی سازه های ریلی را دارا می باشند و این مطلب احتمال دارد به سختگیری بیش از حد استادان این رشته مربوط گردد. با بررسی دانشجویان خوشه های ۵ و ۶ نیز که بعنوان خوشه دانشجویان ضعیف نام برده می شوند، می توان به صحت این موضوع پی برد. میانگین تعداد واحدهای اخذ شده در هر ترم بین ۱۶ و ۱۷ واحد درسی می باشد. این می تواند بعنوان یک راهکار مناسب جهت راهنمایی دانشجویان بکار گرفته شود. شاید بهترین نحوه انتخاب واحد، داشتن تعداد واحد ۱۷ واحد در هر ترم را پیشنهاد نماید. این دو خوشه بیشترین تعداد دختران را نسبت به سایر خوشه ها در خود جای داده اند و این مطلب، علاقمندی بیشتر دانشجویان دختر به مطالعه نسبت به پسران را نشان میدهد[۵].

ابراهیم صحافی زاده (۱۳۸۸،در سومین کنفرانس انجمن داده کاوی ایران) در مقاله خود با عنوان «تحلیل عوامل مؤثر بر نمرات دانشجویان دانشگاه پیام نور بوشهر با استفاده از داده کاوی» بر روی بیش از ۲۵۰ هزار نمره دانشجویان دانشگاه پیام نور بوشهر در بین سالهای ۸۱ تا ۸۷ با استفاده از خوشه بندی به

dataacademy.ir

روش k-mean ، نمرات خوشه بندی شده و تاثیر عواملی از قبیل نوع آزمون، نوع ورود به آموزش عالی، نوع ارائه درس (خود خوان یا حضوری)، واحدهای درس (عملی – نظری)، تعداد جلسات ارائه، دوره، نیمسال ارائه درس، جنسیت و سن بر روی هر خوشه بررسی شده و قوانین تداعی موجود بین این فیلدها استخراج شده است.

در تحلیل بر اساس معدل تعداد گروهی از دانشجویان ورودی بین سالهای ۸۱ تا ۸۶ که معدل آنها خوشه بندی شد ۸۵۰۱ دانشجو بود. این دانشجویان براساس معدل کل آنها تا نیمسال اول سال تحصیلی ۸۸-۸۷ در ۵ خوشه طبقه بندی شدند. خوشه ۱ بیشترین فراوانی را دارد و خوشه ۳ که بالاترین معدل ها را دارد دارای کمترین فراوا نی است و فقط حدود ۸ درصد از دانشجویان را شامل می شود. با توجه به اینکه خوشه ۴ بعد از خوشه ۱ بیشترین فراوانی را دارد می توان گفت که نزدیک به ۶۰ درصد از دانشجویان دانشگاه پیام نور استان بوشهر به لحاظ معدل در خوشه های ۱ و ۴ قرار می گیرند. بنابراین معدل کل ۶۰ درصد دانشجویان بین ۱۲ تا ۱۴ می باشد. با استفاده از ابزارهای تولید قانون در نرم افزار Clementine اولین شاخص اثر گذار بر خوشه بندی معدل کل، پارامتر نحوه ورود به آموزش عالی بود. دانشجویانی که با سهمیه خانواده شهدا وارد دانشگاه می شوند از سال ۸۵ به بعد دچار افت تحصیلی شده و در خوشه شماره ۲ قرار گرفته اند که این خوشه دارای معدل پایینی می باشد . آموزگاران زیر ۳۰ سال معدل بهتری نسبت به آموزگاران بالای ۳۰ سال دارند. همچنین در تحلیلی که بر روی دانشجویانیکه نوع ورود به آموزش عالی آنها عادی بود انجام گرفت مشخص شد که دانشجویان دوره فراگیر از نظر معدل وضعیت بهتری نسبت به دانشجویان دوره رسمی دارند. دانشجویان آزمون محور رسمی و فراگیر و الکترونیکی نیز در خوشه دو قرار گرفتند این خوشه وضعیت ضعیفی دارد. بالاترین میانگین معدل متعلق به دانشجویان ورودی سال ۸۴ و پایین ترین میانگین معدل متعلق به دانشجویان ورودی ۸۱ می باشد.

در تحلیل نمرات دانشجویان دانشگاه پیام نور بوشهر مشخص شد که حدود ۶۰ درصد از نمرات دانشجویان بین ۱۰ تا ۱۶ می باشد. وضعیت نمرات دانشجویان دوره فرا گیر بهتر از دانشجویان رسمی است. ارائه درس به صورت خودخوان و حضوری در نتیجه آزمونها تاثیر متفاوتی ندارند. .میانگین نمرات دانشجویان تا ۵ ترم اول به صورت صعودی بوده و از ترم ۵ به بعد روند کاهش نمرات دانشجویان آغاز و تا آخرین ترمها به صورت نزولی افت می کند. میانگین نمرات دانشجویان، از ورودیهای ۷۷ به بعد رشد صعودی داشته برای دانشجویان ورودی ۸۴ به اوج رسیده اما برای دانشجویان ورودی بعد از ۸۴ به صورت نزولی کاهش می یابد. هر چند میانگین سنی دانشجویان دانشگاه پیام نور به شدت کاهش یافته است اما سن دانشجویان تاثیر معناداری در نمره ندارد. جنسیت تاثیری در نتایج ندارد و نمرات هر دو گروه جنسی در طبقات مشابه قرار دارند. دروس عملی بالاترین نمره را در بین دروس دارند. با توجه به اینکه نمره عملی دست استاد می باشد، تاثیر وجود استاد و برگزا ری امتحان به صورت غیر متمرکز کاملا مشخص است. هر چند اساتید تصحیح اوراق تشریحی را بر عهده دارند اما تفاوتی در نمرات تمام تستی و تستی- تشر یحی مشاهده نمی شود که این نشان دهنده تصحیح دقیق بخش تشر یحی توسط استاد و مطابق با کلید سازمان مرکزی باشد. نمرات امتحانات تشریحی که بدون منبع هستند و آزمون آنها توسط استاد طرح و تصحیح می شود جزء بالاترین نمرات هستند و نهایتا اینکه میانگین نمرات نیمسال بهار بهتر از نیمسال پاییز است[۶].

۲- روش تحقیق

ساختار اجرایی این تحقیق بر اساس دو رویکرد دانش محور[۶] و داده محور[۷] بنا گردیده است که در ادامه به آن اشاره گردیده است. ارزیابی قوانین حاصله از مدلسازی داده کاوی(تکنیک خوشه بندی و طبقه بندی) ماهیت دانش محور دارد، زیرا براساس دانش خبرگان دانشگاهی شامل مدیران، اساتید و کارشناسان آموزشی صورت می پذیرد و با استفاده از تایید صحت قوانین اخذ شده از طرف ایشان ارزیابی مدلهای تولید شده انجام می پذیرد.

ماهیت اصلی تحقیق داده محور بوده و پایه اصلی تحقیق حاضر بر کشف دانش پایگاه داده های بانک مورد مطالعه نهاده شده است.

از فرآیند CRIPS[۸] و از نرم افزار SPSS Clementine 12 به منظور داده کاوی استفاده شده است.

[6] Knowledge Oriented
[7] Data Oriented
[8] CRoss-Industry Standard Process for Data Mining

dataacademy.ir

۲-۱- فرآیند CRIPS-DM

محصول کنسرسیومی متشکل از سه شرکت(سال ۱۹۹۶) دایملر -کرایسلر، SPSS و NCR است تاکید بر چرخشی بودن فرآیند دارد.

شامل فازهای اصلی:

درک کسب و کار

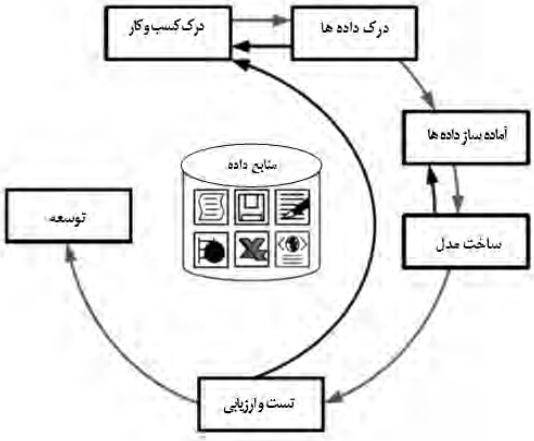
شناسایی و درک داده ها

آماده سازی داده ها

مدلسازی

تست و ارزیابی

توسعه



شکل ۲- فرآیند CRIPS-DM

۲-۲- مرحله درک کسب و کار

در این مرحله شناسایی پس زمینه فعالیت ، اهداف و معیارهای موفقیت فعالیت کسب وکار انجام گرفت. دانشکده دندانپزشکی گیلان در سال ۱۳۷۵ اولین گروه دانشجویان خود را از طریق کنکورسراسری پذیرش نمود. از ابتدای شروع بکار دارای مشکلات متعددی بوده است. یکی از مهمترین مشکلات آن نداشتن فضای فیزیکی مناسب است. هدف کسب و کار در دانشکده دندانپزشکی گیلان عبارت است از « تربیت دانش پژوی دکتری دندانپزشکی است که پس از گذراندن ۶ سال تحصیلی بعنوان فردی متعهد و مجرب بتواند در جامعه دندانپزشکی کشور قادر به خدمت رسانی به ملت عزیز ایران باشد».

در این مرحله اهداف استخراج داده ها نیز مشخص گردید، مشکلاتی که در حین مطالعه سبب مشخص شدن اهداف تحقیق گردید عبارت بودند «چرا عده ای از دانشجویان دارای تعداد ترم های مشروطی بیشتری هستند؟» ، « چرا میزان مردودی دانشجویان در بعضی از دورس بیشتر از دروس دیگر است؟»، «چرا دانشجویان ورودی یک سال خاص نسبت به سالهای ورودی دیگر قویتر می باشند»، از اهداف استخراج داده ها، مجزا نمودن داده هایی که محقق را در شناسایی عوامل مشروطی و مردودی افراد در ترم و درس و همچنین رفتار آموزشی دانشجویان کمک نماید. معیار موفقیت پروژه داده کاوی کشف قوانینی که به اهداف پروژه نزدیک باشد و توسط خبرگان آموزش دانشکده مورد تایید باشد.

۲-۳- مرحله درک داده ها

درک داده ها به بررسی داده های مورد نیاز می پردازد. این مرحله می تواند شامل جمع آوری داده های اولیه، توصیف داده ها، اکتشاف داده ها، و تایید کیفیت داده ها باشد. اکتشاف داده ها از قبیل مشاهده اطلاعات آماری خلاصه (که شامل نمایش تصویری از طبقه بندی متغیرها است) می تواند در پایان این مرحله رخ دهد.

جامعه آماری تحقیق شامل ۵۷۷ دانشجو که از این تعداد ۳۳۸ نفر دانشجو شاغل به تحصیل(فعال)، و ۲۳۹ دانشجو از نظر تحصیلی، فارغ التحصیل، اخراج یا جابجایی به دانشگاه های دیگر داشته اند(غیرفعال). منتهی پس از بررسی نهایی و حذف رکوردهایی که انتقال، اخراج یا مهمان از دانشگاه های دیگر بودند تعداد رکوردها به ۵۱۹ عدد رسید. البته جهت فرآیند داده کاوی با توجه به طرح در نظر گرفته برای الگوهای پیش بینی، تعداد ۳۵۲ رکورد که بیش از ۱۰ ترم تحصیلی را گذرانده بودند به عنوان یک منبع جداگانه برای تکنیکهای پیش بینی طبقه بندی و خوشه بندی تفکیک گردیدند.

فیلدهای دانشجویان شامل دو قسمت اطلاعات شخصی و اطلاعات آموزشی بودند. این فیلدها ۲۱ عدد و شامل:

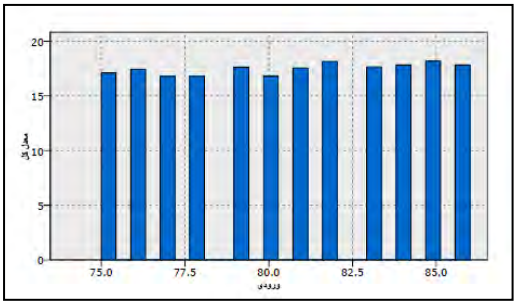
شماره دانشجویی، سال شروع به تحصیل، جنسیت، سهیمه ثبت نامی، تاریخ تولد، معدل کل، تعداد واحدهای گذرانده، تعداد واحدهای مردود شده، تعداد ترم مشروطی، معدل ترم اول، تعداد واحدهای پاس شده در ترم اول، تعداد واحدهای پاس نشده در ترم اول، معدل ترم دوم، تعداد واحدهای پاس شده در ترم دوم، تعداد واحدهای پاس نشده در ترم دوم، معدل ترم سوم، تعداد واحدهای پاس شده در ترم سوم، تعداد واحدهای پاس نشده در ترم سوم، معدل ترم چهارم، تعداد واحدهای پاس شده در ترم چهارم، تعداد واحدهای پاس نشده در چهارم بودند.

dataacademy.ir

برای پی بردن رابطه بین معدل کل با معدل ترم ۱، ورودی، سن، تعداد واحد قبولی، تعداد ترم مشروطی و از طریق ضریب پیرسون متوجه شدیم که بجز ورودی که دارای رابطه ضعیفی است بقیه دارای رابطه قوی با معدل کل می باشند.

برای پی بردن رابطه بین تعداد ترم مشروطی با معدل کل، معدل ترم ۱، ورودی، سن، تعداد واحد قبولی از طریق ضریب پیرسون متوجه شدیم که تعداد واحد قبولی که دارای رابطه ضعیفی است بقیه دارای رابطه قوی می باشند.

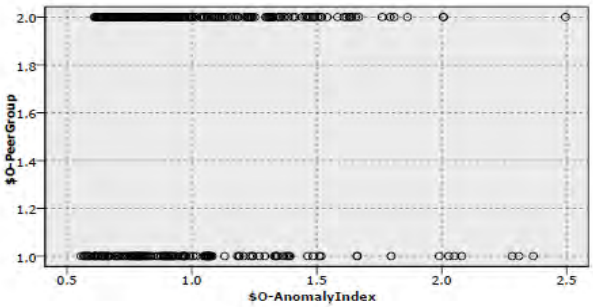
همانطور که در ابتدای این بخش گفته شد. اکتشاف داده ها از طریق مشاهده اطلاعات آماری خلاصه (که شامل نمایش تصویری از طبقه بندی متغیرها است) نیز صورت می گیرد. این اکتشاف اولیه جالب است که قبل از عمل مدل سازی انجام می پذیرد. بعنوان مثال نمودار ۱ نشان میدهد اختلاف فاحشی بین کسانی که دارای بالاترین معدل کل در هر ورودی بودند مشاهده نمیشود و ورودیهای ۸۲ و ۸۵ تقریبا نسبت به سایر ورودیها بالاترین معدل کل را دارند.



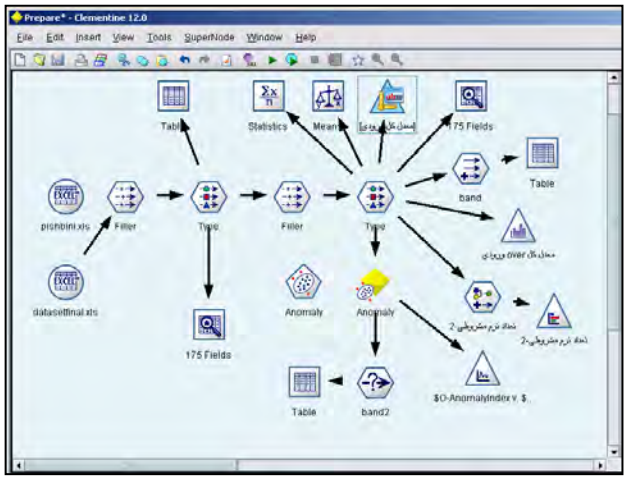
نمودار ۱- نمودار میله ای ارتباط بین ورودیها و حداکثر معدل کل هر ورودی

در این مرحله همچنین کیفیت داده ها مورد بررسی قرار گرفت که شامل، بررسی عدم انطباق داده ها، مقادیر پرت و نقاط گمشده بود. برای اینکار از نود Data Audit نرم افزار SPSS Clementine استفاده گردید. ابتدا از سربرگ Audit کیفیت داده ها از لحاظ صحت مقادیر کمینه و بیشینه و همچنین مراجعه به ساختار نمودار آماده شدهٔ آنها بررسی شد و در صورت مشاهدهٔ مورد اختلاف رکورد مربوطه را یافته و اختلاف را بررسی و در صورت لزوم اصلاح گردید. به عنوان مثال در بررسی تعداد واحدهای پاس شده ترم دوم مواردی عدم انطباق مشاهده گردید بطوریکه رکوردهایی دارای مقادیر بالای ۲۰ واحد پاس شده داشتند، رکوردها مربوط به دانشجویان بهداشتکار می باشد و مقادیر بالای ۲۰ واحد مربوط به واحدهای تطبیقی این دانشجویان که در دوره فوق دیپلم خود آنها را پاس نموده اند، می باشد و اشتباه در ورود اطلاعات داشتیم که برطرف گردید.

در خصوص مقادیر پرت و گمشده به سربرگ Quality مراجعه گردید در فیلدهای تعداد ترم مشروطی و تعداد واحد مردودی چون دارای مقادیر خالی بودند، مقادیر گمشده مشاهده گردید که مشکل با دادن مقادیر صفر در داخل مقادیر خالی برطرف شد. با استفاده از نود Anomaly نرم افزار SPSS Clementine رکوردهای پرت شناسایی شدند همانطورکه در شکل ۳ مشاهده میشود رکوردهایی که مقدار AnomalyIndex آنها از ۲/۲ بیشتر بود بعنوان رکورد پرت شناسایی شدند و مجزا گردیدند پس از بررسی این رکوردها، این نتیجه حاصل شد که با توجه به اینکه از لحاظ آماری نقاط پرت مشاهده شدند ولی از لحاظ آموزشی این مقادیر پرت منطقی و توجیه پذیر بودند و با توجه به اینکه شناسایی رفتار آنها جزء اهداف این تحقیق می باشد بنابراین آنها حذف نگردیدند.



شکل ۳- نمایش نقاط پرت با استفاده از نود Anomaly



شکل ۴- نمایی از کارهای انجام گرفته برای درک داده ها در نرم افزار SPSS Clementine

dataacademy.ir

۲-۴- آماده سازی داده ها

هنگامی که منابع داده ای در دسترس شناسایی شدند، نیاز به آن دارند که انتخاب شوند، پاکسازی شوند، و به شکل مورد نظر ساخته ، و ساختار بندی شوند.*پاکسازی داده ها* و *تبدیل داده ها* در هنگام آماده سازی مدل داده ها، باید در این مرحله رخ دهد. هدف از آماده سازی داده ها تمیز کردن داده های انتخاب شده برای کیفیت بهتر است.

فیلد سال ورود به تحصیل که حاوی اطلاعات سال، ماه و روز ورود به تحصیل دانشجو بود به فیلد ورودی تبدیل شد و فقط سال ورود به این فیلد وارد گردید. از فیلد تاریخ تولد دانشجو سال تولد دانشجو گرفته شد. همچنین در فیلدهای تعداد واحد های پاس شده و پاس نشده ترمهای ۱ تا ۴ ، بجای اینکه در نتیجه گیری تعداد واحد های پاس شده و پاس نشده چهار ترم اول در قانون های تولید شده در تصمیم گیری دخیل باشد با استفاده از فیلدهای جدید Pass1 تا Pass4 نتیجه گیری بهتری خواهیم داشت. که بعنوان مثال فرمول (۱) زیر روش بدست آمدن فیلد Pass1 را نشان می دهد که به همین صورت برای سه فیلد دیگر محاسبه می شود.

(۱)

(تعداد واحد های پاس شده ۱)/ (تعداد واحد های پاس شده ۱+ تعداد واحدهای پاس نشده ۱) =pass1

فرمت تمام داده های کمی بجز معدل از حالت اعشاری به عدد صحیح تبدیل شد.

۲-۵- مدل سازی

پس از آماده سازی داده ها نوبت به تشخیص این است که برای رسیدن به اهداف داده کاوی خود از چه تکنیک ها و الگوریتمهایی باید استفاده کنیم. تکنیکهای داده کاوی بر اساس اهداف داده کاوی به دو گروه تقسیم می شوند الف) توضیح گذشته ب) پیش بینی آینده

با توجه به هدف، در این تحقیق ما از تکنیکهای پیش بینی آینده استفاده کردیم، تکنیکهای پیش بینی آینده که به چهار دسته ۱) طبقه بندی ۲) رگرسیون ۳) خوشه بندی ۴) قوانین انجمنی تقسیم میشوند. که ما در تحقیق خود از تکنیک طبقه بندی و خوشه بندی استفاده نمودیم. که طبقه بندی جهت پیشگویی مقادیر گسسته و اسمی مورد استفاده قرار می گیرد، خوشه بندی در واقع یک عملیات غیر نظارتی می باشد. این عملیات هنگامی استفاده می شود که ما به دنبال یافتن گروه هایی از داده های مشابه می باشیم بدون اینکه از قبل، پیش بینی در مورد شباهت های موجود داشته باشیم. جهت انجام تکنیک طبقه بندی از الگوریتم های معروف CHAID ، CRT و C5 استفاده گردید. جهت انجام تکنیک خوشه بندی از الگوریتم های معروف K-Means و TwoStep استفاده گردید.

۲-۶- تست و ارزیابی

پس از مدل سازی مدل ها مورد تست و ارزیابی قرار گرفت و چندین بار اصلاح گردید تا مدل مورد قبول با اهداف ما بدست آمد.

۲-۷- توسعه

این مرحله بعد از تولید مدل و تایید دانش بدست آمده، شروع میشود. نیازِ به تغییر مدلهای تولید شده در شرایط عملیاتی باید بررسی شود، زیرا ممکن است قوانین تولید شده توسط مدلها در سالهای بعد درست نباشند. اگر تغییرات پیش آمده در محیط کسب و کار قابل توجه باشند مدل ها باید دوباره بازسازی شوند.

۳- نتایج

۳-۱- نتایج خوشه بندی

کشف دانش موجود در اطلاعات ورودی و مشخص کردن ورودی های با اهمیت از نتایجی است که از طریق استفاده از این تکنیک بدست می آید. ورودیهای با اهمیت الگوریتم های Two-Step و K-Means یکسان بوده و عبارت بودند از فیلدهای pass، تعداد ترم مشروطی، جنس، سن، سهمیه و معدل کل. این امر اشاره به آن دارد متغیرهایی که ما جهت فرآیند داده کاوی درنظرگرفته ایم با متغیرهای تشخیص داده شده مدل داده کاوی مطابقت داشته و انتخاب صحیحی نموده ایم. نتایج بدست آمده در هر دو تکنیک در جدول ۱ نشان داده شده است.

dataacademy.ir

جدول ۱- مقایسه دو مدل خوشه بندی K-Means و TwoStep

		مدل K-Means
خوشه ۱	دانشجویان متوسط	معدل کل ۱۵/۱٤– ۱۰۰٪ مرد– تعداد ترم مشروطی صفر– همه سهمیه وجود دارند بالاترین درصد فراوانی منطقه ۲– کمترین شاهد
خوشه ۲	دانشجویان ضعیف	معدل کل ۱۳/٦۷ – ٦۸٪ زن و ۳۲٪ مرد هستند– ۵۲٪ ۱ یا ۲ مشروط شده اند و ٤۸٪، ۳ بار یا بیشتر– بجز بهداشتکار بقیه سهمیه دیده میشوند، بالاترین درصد فراوانی شاهد و کمترین بهداشتکار(۰٪)
خوشه ۳	دانشجویان ضعیف	معدل کل ۱۳/۷٦٪– ۱۰۰٪ مرد– ۹۷/۵٦٪ تعداد ترم مشروطی ۱ یا ۲ بار و ۲/٤٤٪، ۳ بار یا بیشتر– همه سهیمه ها بجز منطقه ۳ وجود دارند بالاترین درصد بهداشتکار و کمترین منطقه ۳ (۰٪)
خوشه ۴	دانشجویان قوی	معدل کل ۱۵/۹۷٪– ۱۰۰٪ زن– تعداد ترم مشروطی صفر– همه سهیمه های وجود دارند. بالاترین درصد مربوط به منطقه ۱ و کمترین رزمندگان و جانبازان
		مدل Two-Step
خوشه ۱	دانشجویان بهداشتکار	معدل کل ۱٤/٤۰– ۳۱/۲۵٪ دانشجو بهداشتکار که ۱ یا ۲ ترم مشروط شدند جز ضعیف محسوب میشوند و٦۸/۷۵٪ دانشجو بهداشتکار که تعداد ترم مشروطی آنها صفر است که جزء دانشجو متوسط محسوب می شوند
خوشه ۲	دانشجویان متوسط	معدل کل ۱۵/۳۳ – ۱۰۰٪ مرد– تعداد ترم مشروطی صفر– همه سهیمه ها بجز بهداشتکار هستند–بالاترین درصد منطقه ۲ و کمترین شاهد ، فرزند هیئت علمی
خوشه ۳	دانشجویان ضعیف	معدل کل ۱٤/۰٦– ۵۲/۱۷٪ زن و ٤۷/۸۳٪ مرد می باشند. ۱۸/۸٤٪ ، ۳ترم یا بیشتر مشروط شده اند و ۷٤/۸۳٪ یک یا دو ترم مشروط شده اند و ۳۳/۳۳٪ مشروط نشده اند همه سهیمه ها در آن شرکت دارند بجز بهداشتکاران، بالاترین درصد مربوط به شاهد در این سهیمه است و کمترین بهداشتکار
خوشه ۴	دانشجویان قوی	معدل کل ۱٦/۱۲– ۹۲/۹۱٪ زن–تعداد ترم مشروطی صفر، سهمیه منطقه ۱،۲ و ۳ . ۰/۷۹ رزمنده یا جانباز بیشترین مقدار مربوط به منطقه ۲ و کمترین رزمنده و جانباز

مدل Two-Step را می توان بهتر از K-Means بـدلایل زیـر در نظر گرفت:

خوشه ۱ مختص دانشجویان بهداشتکار بوده و هیچ سـهمیه دیگری در آن جـا نگرفتـه اسـت و ایـن نشـان مـی دهـد رفتـار آموزشی دانشجویان بهداشتکار جدای سـهمیه هـای دیگـر بـوده است. البتـه ۱۰۰٪ دانشـجویان بهداشـتکار در ایـن خوشـه قـرار گرفته اند. در این خوشه ۳۱/۲۵٪ دانشـجویان بهداشـتکار جـزء دانشجویان ضعیف بوده و ۶۸/۷۵٪ جزء دانشجویان متوسـط مـی باشند. میانگین معدل کل آنها ۱۴/۴۰ بـا ضـریب انحـراف معیـار ۰/۸۱۸ می باشد.

در تقسیم بنـدی دانشـجویان قوی(خوشـه ۴) بـا میـانگین معدل کل ۱۶/۱۲، بـا توجـه بـه میـانگین معـدل کـل ۱۶/۴۵ در تحقیق که برای تفیکیک دانشجویان قـوی تعیـین گردیـده بـود، بهتر عمل نموده است و درصد فراوانی جنسـیتی و سـهمیه ای را با تطبیق با مجموعه داده دقیقتر محاسبه نمـوده اسـت(۹۲/۹۱٪ زن و ۷/۰۹٪ مرد).

در مورد دانشجویان ضعیف(خوشه ۳) هر چند کـه میـانگین معـدل کـل آنهـا ۱۴/۰۶ بـا ضـریب انحـراف معیـار ۱/۰۲۱ بـا استانداردی که ما در این تحقیق در نظـر گرفتـه ایـم فاصـله ای مشاهده میشـود، امـا از لحـاظ جنسـیت، سـهمیه و تعـداد تـرم مشروطی و بدون مشروطی تفیکیک صـحیح تـری را انجـام داده است. همچنین در مـدل Two-Step بـر خـلاف مـدل K-Means دانشجویان ضعیفی که در طـول تحصـیل بـدون مشـروطی مـی باشند، مشاهده می شوند.

در مورد دانشجویان متوسط هر دو مدل خوب عمل ننمـوده اند. زیرا ۱۰۰٪ جنسـیت تشـکیل دهنـدهٔ آنهـا مـرد بـوده انـد و دانشـجویان زن متوسـط در هـر دو مـدل بیشـتر در خوشـه دانشجویان ضعیف قرار گرفتند. با توجـه بـه اینکـه هـدف اصـلی تحقیق ما دانشجویان قوی(ممتـاز) و ضـعیف(تحت ریسـک) بـود وضعیت دانشجویان متوسط مورد بررسی ما نمی باشد.

۳-۲- نتایج طبقه بندی

هدف اصلی از تکنیک طبقه بندی پـیش بینـی دانشـجویان ممتاز(قوی) و تحت ریسک مشروطی(ضعیف) می باشد خروجـی مدل را براساس دو فیلد قرار دادیـم یکـی معـدل کـل و دیگـری تعداد ترم مشروطی برای انجام آن از مـدلهای CHAID ، CRT و C5 استفاده نمودیم. بطورکلی در قوانین بوجود آمده، با توجه بـه تعداد رکوردهای شرکت کننده، قوانینی دارای شرط لازم و کافی بـرای انتخـاب شـدند کـه درصـد پشـتیبانی(support) آن دارای حداقل ۵٪ تعداد رکوردهای مورد استفاده در آمـوزش(Training) باشند و درصد اطمینان(confidence) آن حداقل ۸۰٪ باشد.

- قانون بدست آمده مدل CHAID بر اساس معدل کل برای دانشجویان قوی:

*در مجموعه داده ها، از دانشجویانی که معـدل تـرم ۴ آنهـا بزرگتـر از ۱۶/۷۵ و معـدل تـرم ۳ آنهـا بزرگتـر از ۱۶/۵۰ بـود،

dataacademy.ir

۹۴/۴٪ از آنها معدل بالای ۱۶/۴۴ داشتند که طبق تقسیم بندی تحقیق جزء دانشجویان قوی محسوب میشوند.

قانون بدست آمده مدل CHAID بر اساس معدل کل برای دانشجویان متوسط:

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۴ آنها بزرگتر از ۱۳/۶۲ و کمتر از ۱۵/۰۳ باشد و معدل ترم ۳ آنها بزرگتر از ۱۲/۱۰ و کمتر از ۱۷/۳۳ باشد، ۹۰٪ از آنها معدل بین ۱۳/۹۵ و ۱۶/۴۴ را داشتند که طبق تقسیم بندی تحقیق جزء دانشجویان متوسط محسوب میشوند.

قانون بدست آمده مدل CHAID بر اساس معدل کل برای دانشجویان ضعیف:

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۴ آنها کوچکتر یا مساوی ۱۳/۶ باشد و معدل ترم ۱ کوچکتر مساوی ۱۳/۴۷ باشد و مقدار pass1 آنها کوچکتر مساوی ۰/۹۵ باشد، ۷۵/۶۷٪ از آنها معدل کمتر از ۱۳/۹۵ داشتند که طبق تقسیم بندی تحقیق جزء دانشجویان ضعیف محسوب میشوند.

- قانون بدست آمده مدل CHAID بر اساس تعداد ترم مشروطی:

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها کوچکتر مساوی ۹/۳۵ باشد ، ۱۰۰٪ از آنها تعداد ترم مشروطی مساوی یا بیشتر از ۳ ترم دارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها بزرگتر از ۹/۳۵ و کوچکتر مساوی ۱۱/۳۶ باشد، ۹۰٪ از آنها تعداد ترم مشروطی ۱ یا ۲ ترم دارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها بزرگتر از ۱۲/۴۲ و کوچکتر مساوی ۱۳/۳۸ باشد، ۷۵٪ از آنها تعداد ترم مشروطی ۱ یا ۲ ترم دارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها کوچکتر مساوی ۱۳/۳۸ و بزرگتر از ۱۲/۴۲ باشد و معدل ترم ۴ آنها کوچکتر مساوی ۱۳/۳۸ و بزرگتر از ۱۲/۵۷ باشد، ۶۶/۶۶٪ از آنها تعداد ترم مشروطی ۱ یا ۲ ترم دارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها کوچکتر مساوی ۱۴/۱۰ و بزرگتر از ۱۳/۳۸ و سهمیه آنها فرزند هیئت علمی یا انتقالی از خارج باشد، ۶۰٪ از آنها تعداد ترم مشروطی ۱ یا ۲ ترم دارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها بزرگتر از ۱۴/۱۰۳۵ و کوچکتر مساوی ۱۵/۶۰ و معدل ترم ۴ آنها بزرگتر از ۱۴/۷۹ باشد، ۱۰۰٪ آنها ترم مشروطی ندارند.

*در مجموعه داده ها، از دانشجویانی که معدل ترم ۲ آنها بزرگتر از ۱۵/۶۰ و مقدار فیلد pass3 آنها بزرگتر از ۰/۷۱۴ باشد، ۹۹/۱۳٪ آنها ترم مشروطی ندارند.

۴- نتیجه گیری

در مدل Two-Step، ۴ خوشه بوجود آمد، خوشه ۱ مربوط به دانشجویان بهداشتکار می باشد که با توجه به تعداد آنها و نوع رفتار آموزشی آنها یک خوشه به آنها اختصاص یافت و در سایر خوشه ها دیده نشدند. و در بررسی داده های آموزشی، خوشه ۲ بعنوان خوشه دانشجویان متوسط نام گرفت با توجه به اینکه ۱۰۰٪ جمعیت آن مرد بودند دانشجویان زن متوسط در خوشه های ۳ و ۴ قرار گرفتند و این از نقاط ضعف این مدل می باشد، خوشه ۳، خوشه دانشجویان ضعیف نام گرفت، و خوشه ۴، خوشه دانشجویان قوی که تعداد ترم مشروطی آنها صفر و ۹۲/۹۱٪ آنها زن بودند. که نشان دهندهٔ تلاش و پشتکار و اهمیت دادن به درس خواندن در نزد خانم ها می باشد.

اطلاعات دیگری که از خوشه بندی جهت تحلیل رفتار دانشجویان حاصل شد، شامل:

-افزایش درصد پذیرفته شدگان زن نسبت به مرد از سال ۸۸ به بعد

-بالا بودن میانگین معدل کل خانم ها در تمامی ورودیها نسبت به آقایان که تایید کننده بالا بودن تعداد زن ها در دانشجویان قوی نسبت به مردها می باشد.

-با توجه به وجود دانشجویان با تعداد ترم مشروطی ۳ ، ۴ و ۵ فقط در خوشه ۳ با صحت بیشتری می توان این خوشه را بعنوان خوشه دانشجویان ضعیف نامگذاری کرد.

-با توجه به مشاهده شدن دانشجویان با تعداد ترم مشروطی ۳ ، ۴ و ۵ در سهمیه های شاهد، رزمندگان و فرزندان هیئت علمی، نیاز به بازنگری آموزشی در این سه گروه می باشد، توصیه می شود نقاط ضعف آموزشی این دانشجویان شناسایی و مسئولین جهت بهبود آنها برنامه ریزی نمایند.

-دانشجویان ورودی ۷۸ و ۸۴ دارای کمترین تعداد مشروطی در بین ورودیها بوده اند(ورودی ۷۸ یک نفر و ورودی ۸۴ دو نفر) که نسبت به دانشجویان ورودی سالهای دیگر دارای راندمان آموزشی خوبی بودند.

-با توجه به بالا بودن میانگین معدل خانم ها نسبت به آقایان و قرار گرفتن اکثریت خانم ها در خوشه دانشجویان قوی

dataacademy.ir

می توان ، جنسیت را جزء عوامل موثربر نمرات دانشجویان برشمرد.

-با وجود قرارگرفتن تمامی سهمیه ها در خوشه دانشجویان ضعیف و متوسط می توان گفت متغیر سن تاثیری در این دو گروه ندارد، ولی با قرار گرفتن دانشجویان سهمیه مناطق در خوشه قوی و داشتن میانگین سنی زیر ۲۰ سال در این سهمیه ، می توان گفت متغیر سن در کسب نمرات بهتر در این خوشه تأثیر گذار است.

-نتایج حاصل از تحقیق نشان می دهد نوع سهیمه ورودی دانشجویان نیز بر نمرات کسب شده توسط دانشجویان مؤثر است. بطوریکه دانشجویان سهیمه های شاهد، رزمنده و فرزند هیئت علمی دارای نمرات پایین تری بودند و نیاز به توجه آموزشی دارند.

در مدل CHAID تکنیک طبقه بندی معدل ترم ۴ جزء پراهمیت ترین فیلد در درخت تصمیم می باشد. قوانین برگرفته از آن دارای ضریب اطمینان ۷۵ تا ۱۰۰ درصد بود، قوانین بدست آمده از نظر پیش بینی دانشجویان تحت ریسک و ممتاز بسیار ارزشمند بوده و می توان از آنها بعنوان قوانینی در تحلیل رفتار دانشجویان استفاده نمود.

مراجع

[1] Thearling, K., 2007. An Introduction to Data Mining [Online]. Available at: http://www.thearling.com

[2] Ryan S.J.d. Baker., 2008. Data Mining for Education. [pdf] Available at: < http://users.wpi.edu/~rsbaker/Encyclopedia%20Chapter%20Draft%20v10%20-fw.pdf >

[3] Fayyad, U., Piatetsky-Shapiro G., and Smyth P., 1996. From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, [pdf] Availabel at: www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf

[4] Dekker, G., Pechenizkiy, M. and Vleeshouwers, J., 2009. Predicting Students Drop Out: A Case Study. p.41-50.Conference Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain [pdf]. Available at: < http://www.win.tue.nl/~mpechen/projects/edm/internshipreport_090409.pdf >

[۵] شکورنیاز. ونوس و حاجی علی اکبری. آرش، سال ۱۳۸۷. « خوشه بندی داده های آماری دانشجویان دانشگاه علم و صنعت و استخراج نمایه ساز توصیفی برای دانشجویان موفق»، مقاله دومین کنفرانس انجمن داده کاوی ایران.

[۶] - صحافی زاده. ابراهیم، سال ۱۳۸۸. «تحلیل عوامل مؤثر بر نمرات دانشجویان دانشگاه پیام نور بوشهر با استفاده از داده کاوی»، مقاله سومین کنفرانس انجمن داده کاوی ایران.