مدیریت فناوری اطلاعات
دانشکدهٔ مدیریت دانشگاه تهران

دورهٔ ۷، شمارهٔ ۲
تابستان ۱۳۹۴
ص. ۲۳۸- ۲۱۷

داده‌کاوی دانشجویان انصرافی دانشگاه تهران با تمرکز بر حفظ دانشجویان شهریه‌پرداز (جلوگیری از روی‌گردانی مشتری)

سیدعلی اکبر احمدی[۱]، داوود کریم‌زادگان[۲]، تورج خیراتی کازرونی[۳]

چکیده: انصراف دانشجو یکی از چالش‌های پیش روی آموزش عالی است. مقالهٔ حاضر رویکرد پذیرش دانشجوی شهریه‌پرداز را نوعی کسب‌وکار و انصراف دانشجو را روی‌گردانی مشتری در نظر گرفته است و به‌دنبال بررسی عوامل انصراف دانشجویان و اتخاذ سیاست‌های مداخله‌جویانهٔ بازدارنده است. پژوهش پیش رو کاربردی از نوع توصیفی است که به‌کمک داده‌های کمی و کیفی بر مبنای روش پژوهش کریسپ از داده‌کاوی اطلاعات دانشجویان ورودی شهریه‌پرداز (۲۱۴۲۰ دانشجو دانشگاه تهران طی سال‌های ۱۳۹۲- ۱۳۸۸) استخراج‌شده از بانک‌های اطلاعاتی سیستم آموزش دانشگاه تهران، اجرا شده است. هدف آن، تحلیل رفتار دانشجویان به‌منظور شناسایی دانشجویان در معرض خطر انصراف و ارائهٔ مدل پیش‌بینی احتمال انصراف است. پس از تحلیل داده‌ها و ارائهٔ مدل پیش‌بینی، جدول احتمال انصراف و مدل رگرسیونی انصراف، یافته‌های پژوهش ترم اول و دوم (به‌ویژه ترم اول در دورهٔ سنی ۳۱-۲۴ سال) را به‌منزلهٔ پرخطرترین دورهٔ زمانی، دانشجویان ارشد را مستعدترین مقطع و دورهٔ شبانه را پرخطرترین دورهٔ تحصیلی برای انصراف دانشجو (روی‌گردانی مشتری) شناسایی کرد.

واژه‌های کلیدی: انصراف، داده‌کاوی آموزشی، روی‌گردانی مشتری، مدیریت ارتباط با مشتری.

۱. استاد گروه مدیریت، دانشگاه پیام نور واحد غرب، تهران، ایران
۲. استادیار گروه مهندسی کامپیوتر، دانشگاه پیام نور سازمان مرکزی، تهران، ایران
۳. دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات، دانشگاه پیام نور واحد غرب، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹۳/۱۲/۲۶
تاریخ پذیرش نهایی مقاله: ۱۳۹٤/۰۳/۲۵
نویسندهٔ مسئول مقاله: تورج خیراتی کازرونی
E-mail: tkazerooni@ut.ac.ir

مقدمه

کاهش اعتبارات بخش دولتی آموزش عالی، ورود گستردهٔ بخش خصوصی به حوزهٔ آموزش عالی و آموزش‌های الکترونیکی، تبدیل عرصهٔ آموزش و رقابت‌های آموزشی به رقابت‌های آموزشی ـ تجاری و در رأس همه، استقبال و عطش فراوان ورود به آموزش عالی، مؤسسه‌های آموزش عالی را بر آن داشته است با نگاهی نوین و در قالب دانشجویان شهریه‌پرداز به پذیرش دانشجو بنگرند و بدین نحو نه‌تنها کسب درآمد کنند، بلکه بتوانند برنامه‌های توسعه‌ای خود را نیز تحقق بخشند. اکنون تکیه بر نوآوری و استفاده از فناوری، مزیت رقابتی ایجاد نمی‌کند، گستردگی و استفادهٔ آسان و در دسترس امکانات و توانایی‌های اینترنت، سبب می‌شود که اتکای صرف به اینترنت، خالق مزیت رقابتی نباشد. امروزه خلق مزیت رقابتی، تنها در گرو تداوم و توسعهٔ استفاده از سیستم‌های اطلاعاتی و محصولات نوآورانهٔ مدیریت فناوری اطلاعات است. در آینده، جمع‌آوری داده‌های مشتریان به منبع مزیت رقابتی فزاینده‌ای تبدیل خواهد شد[1]. راهبرد حفظ مشتری، سرمایه‌گذاری ارزشمندی برای کسب درآمد و سود در آینده است (برایان هانیگمن، ۲۰۱۳). از این رو، سازمان‌ها به این نتیجه رسیده‌اند برای دستیابی به موفقیت بلندمدت کسب‌وکار، حفظ مشتری حیاتی است. به گزارش گروه گارتنر، افزایش ۵ درصد در حفظ مشتری می‌تواند سود کسب‌وکار را بین ۲۵ تا ۱۲۵ درصد افزایش دهد (مورفی و مورفی، ۲۰۱۳). هزینهٔ جذب مشتری جدید می‌تواند پنج برابر هزینهٔ جلب رضایت و حفظ مشتری باشد (تنتاکو و یو وای، ۲۰۱۴). روی‌گردانی (ریزش) مشتری از مهم‌ترین چالش‌های پیش روی سازمان‌ها است. برای کاهش هزینه‌های حفظ مشتری باید بر مشتریانی تمرکز کرد که به احتمال بیشتری سازمان را ترک می‌کنند. در این راستا، آموزش عالیِ بسیاری از کشورهای جهان برای مقابله و کاهش زیان‌های ناشی از انصراف دانشجو (ریزش مشتری)، از ابزاری نیرومند و مفیدی با نام «داده‌کاوی آموزشی» برای دستیابی به اهدافی مانند استخراج داده‌های مرتبط با دانشجویان (فردی، آموزشی و...) از حجم زیاد داده‌های ذخیره‌شده در پایگاه‌های اطلاعاتی و تجزیه‌وتحلیل آنها؛ شناسایی دانشجویان در خطر ترک تحصیل؛ ارائهٔ الگوی مناسب برای حفظ دانشجو به‌کمک خدماتی چون مشاوره؛ توصیه به شرکت در کلاس‌های تقویتی و بازآموزی و...، بهره می‌برند. این مسئله، هم به لحاظ مالی و اجتماعی و هم به لحاظ اعتباری، برای مراکز آموزش عالی اهمیت ویژه‌ای دارد. داده‌کاوی آموزشی را می‌توان از جنبه‌های کسب درآمد مالی بیشتر برای مراکز آموزشی، ممانعت

1. Harvard business review (2015). https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust?cm_sp=Magazine%20Archive-_-Links-_-Current%20Issue.

از به هدر دادن منابع مالی و سازمانی، پیشبرد اهداف آموزشی و موفقیت راهکارهای جدید آموزشی، بررسی کرد.

هدف این پژوهش، شناسایی و معرفی عوامل مهم انصراف دانشجویان شهریه‌پرداز دانشگاه تهران است. کدام‌یک از عوامل سن، جنسیت، بومی‌بودن، مقطع تحصیلی، سنوات تحصیلی و دورهٔ تحصیلی، در انصراف دانشجویان از اهمیت بیشتری برخوردار است؟ رویکرد پیش رو بر مبنای جمع‌آوری داده‌های مناسب از داده‌های ذخیره‌شده در پایگاه‌های اطلاعات دانشجویی و تجزیه‌وتحلیل آنها، شناسایی علل انصراف و به‌دنبال آن، ارائهٔ مدل پیش‌بینی برای سیاست‌های مداخله‌جویانهٔ مناسب به‌منظور کاهش درصد انصراف دانشجویان و افزایش میزان بقای آنان (ماندگاری) در راستای راهبردهای مناسب برای پشتیبانی از تصمیم‌های مدیران ارشد دانشگاه تهران (مراکز آموزش عالی) شکل گرفته است.

بیان مسئله

اکنون، بقای دانشجویان از بزرگ‌ترین چالش‌های مؤسسه‌های آموزش عالی به‌شمار می‌رود. به‌طور کلی ماندن دانشجوی بیشتر، به‌معنای بهره‌مندی از برنامه‌های آموزشی بهتر و کسب درآمد بالاتر است (ژانگ، اوسنا، کلارک و کیم، ۲۰۱۰). در انصراف دانشجو، هم مؤسسه‌ها و هم دانشجو درگیر دو نوع هزینهٔ مالی و انسانی خواهند بود (پیتمن، ۲۰۰۸). هزینهٔ تحمیل‌شده بر دانشجو شامل هدر رفتن شهریهٔ تحصیلی (از ورود تا زمان انصراف) و سایر هزینه‌های جانبی، از دست‌دادن فرصت اشتغال و درآمد آتی (پس از اتمام تحصیلات عالی) است. در مراکز آموزش عالی (به‌خصوص مراکز آموزشی دولتی) با کاهش اعتبارات عمومی و کمک‌های دولتی، انصراف دانشجو به‌معنای از دست‌دادن درآمدی است که در گسترش فعالیت‌های آموزشی نقش ارزنده‌ای دارد. گذشته از این، انصراف دانشجو و متعاقب آن کاهش دانشجو، نه‌تنها میزان جذب آتی مؤسسه‌های آموزش عالی را تحت تأثیر قرار می‌دهد، بلکه از اعتبار آنها در جامعه و آموزش عالی می‌کاهد. پیامد این مشکل فقط به دانشجویان و مؤسسه‌ها محدود نمی‌شود، بلکه اثر منفی آن گریبان‌گیر سرمایهٔ انسانی متخصص کشور و توانایی پرورش نیروی ماهر برای پاسخگویی به نیازهای جدید صنایع و حرفه‌های گوناگون خواهد شد (اسکات، شاه، گربنکوف و سینگ، ۲۰۰۸). از این رو، نیاز به طرح‌های مبتکرانه برای خنثی‌کردن این پدیده از اقدام‌های ضروری است.

دانشجویان را به‌لحاظ درآمدزایی می‌توان به دو دستهٔ عمدهٔ رایگان و شهریه‌پرداز تقسیم کرد. پذیرش دانشجویان شهریه‌پرداز در قالب دانشجوی شبانه، مجازی، روزانهٔ شهریه‌پرداز و...، رویکرد نوین بسیاری از واحدهای آموزشی دولتی و غیردولتی به‌منظور ایجاد کسب‌وکار جدید، تأمین هزینه‌های نگهداری و توسعه‌ای یا هر دو است. نتایج داده‌کاوی و پیش‌بینی‌های به‌دست‌آمده از

مدل‌های برازش شدهٔ داده‌کاوی، سبب بهبود راهبرد بقای دانشجو شده است (وبلاگ آموزش عالی آمریکا، ۲۰۱۳)[1]. آن چیزی که سبب استفادهٔ محققان از روش‌های داده‌کاوی درخصوص حفظ دانشجویان شده است، بازاریابی هدف مؤثرتر، تمرکز بر بهبود بهره‌وری سازمانی و مدیریت فارغ‌التحصیلان است (کباکچیوا، ۲۰۱۳). یکی از چالش‌های مراکز آموزش عالی، دانشجویان انصرافی است. لذا تحلیل و ارزیابی انصراف دانشجویان شهریه‌پرداز نه‌تنها با تغییر سطح درآمد واحد آموزشی ارتباط دارد، بلکه اصل بسیار حیاتیِ «حفظ مشتری نسبت به جلب مشتری جدید، هزینهٔ کمتری دارد» را یادآوری می‌کند. اگر از دیدگاه عملکرد منابع سرمایه‌گذاری‌شده به این بخش، به موضوع نگریسته شود، شناسایی به‌موقع دانشجویان در معرض خطر انصراف می‌تواند مانع اتلاف منابع فردی، منابع ملی و منابع جهانی شود.

سؤال‌های مطرح‌شده در این پژوهش به شرح زیر است:

- آیا مدلی برای تشخیص الگوی رفتاری دانشجویان در معرض انصراف از سایر دانشجویان، وجود دارد؟ آیا این الگو می‌تواند هزینهٔ ماندگاری مشتری و جذب مشتری جدید را کاهش دهد؟
- آیا می‌توان نقاط عطفی برای گرایش به انصراف تعیین کرد؟
- آیا می‌توان جدول احتمال انصراف دانشجو را استخراج کرد؟
- آیا می‌توان شاخص‌هایی یافت که ضمن کاهش زیان‌های انصراف، درصد ماندگاری دانشجو را افزایش دهد؟
- آیا نتایج پژوهش می‌تواند به مدیران ارشد در برنامه‌ریزی‌های آموزشی کمک کند؟

برای پاسخ به سؤال‌های پژوهش، فرضیه‌های زیر مطرح شده است:

۱. تفاوت رفتار دانشجویان انصرافی تحصیلات تکمیلی با کارشناسی درخور توجه است؛
۲. سال اول تحصیل مرحلهٔ بحرانی انصراف شمرده می‌شود؛
۳. سن، وضعیت اقامت (بومی/ غیربومی) و جنسیت دانشجو، از عوامل عمدهٔ تأثیرگذار در انصراف دانشجو هستند.

هدف پژوهش، آگاه‌کردن کانون‌های تصمیم‌گیری حوزهٔ آموزش دانشگاه تهران، از طریق تعیین میزان اهمیت هریک از صفات بیان‌شده، ارائهٔ مدل انصراف دانشجو، استخراج جدول پیش‌بینی انصراف و دسته‌بندی دانشجویان به لحاظ میزان تهدید خطر انصراف (زیاد، متوسط و کم) است تا از این رهگذر با ایجاد سیستم پشتیبان تصمیم‌گیری برای مدیران ارشد دانشگاه،

1. American Higher Education Blog. available in: http://www.hanoverresearch.com/2012/01/how-11-universities-will-improve-student-retention

زمینهٔ شناسایی دانشجویان در معرض خطر به‌منظور اتخاذ سیاست‌های مداخله‌جویانهٔ مناسب فراهم شود و به هدف غایی یا به بیان دیگر، تسری آنها به آموزش عالی دست یابد.

پیشینهٔ پژوهش

می‌توان بررسی نظام شهریه‌پرداز را از دید نوعی کسب‌وکار در قالب روی‌گردانی مشتری و شناسایی مدل پیش‌بینی انصراف و همچنین معرفی داده‌کاوی آموزشی به پژوهشگران حوزه‌های آموزش عالی را رویکردی جدید در تحلیل مسائل دانشجو و آموزش دانست و اقدامات انجام‌گرفته در این زمینه را نوآوری این پژوهش شمرد.

مدیریت ارتباط با مشتری

مدیریت ارتباط با مشتری نوعی راهبرد کسب‌وکار است که وظایف و فرایندهای درونی را با شبکه‌های بیرونی یکپارچه می‌کند و با ایجاد سود ارزشی آن را در اختیار مشتریان هدف قرار می‌هد (باتل، ۲۰۰۸). چرخهٔ حیات مشتری شامل سه مرحلهٔ جذب، حفظ یا ماندگاری و توسعه است.

برای شرکت‌های تازه‌تأسیس مرحلهٔ جذب مشتری جدید یا معرفی محصول جدید اهمیت حیاتی دارد و احتمال ریزش به‌واسطهٔ جذابیت رقبا، تغییر مکان، مرگ و غیره روی می‌دهد.

در مرحلهٔ ماندگاری مشتری، تمام تلاش‌ها به پایداری روابط طولانی‌مدت و جلوگیری از حرکت به سمت رقبا معطوف می‌شود. هدف ایجاد مشتری وفاداری است که حتی در صورت دریافت پیشنهادی بهتر و پرسودتر از جانب رقبا، وفادار بماند و کسب‌وکار را ترک نکند.

در مرحلهٔ توسعه مشتری، سازمان‌ها سعی در افزایش ارزش مشتریان حفظ شده دارند. دو روش معروف برای توسعهٔ مشتریان روش‌های متقاطع (فروش محصولات بیشتر به مشتریان) و فروش بالاتر (فروش محصولات با قیمت بالاتر) است. این مرحله نیز همانند حفظ مشتری، در بازارهای اشباع‌شده که امکان جذب مشتریان جدید کم است، مفید واقع می‌شود.

روی‌گردانی (ریزش) مشتری[1]

ریزش مشتری به موقعیتی اطلاق می‌شود که مشتری تصمیم می‌گیرد سازمان را ترک کند و مدیریت ریزش مشتری به این مفهوم است که بتوان با پیش‌بینی رفتار مشتریان، از ریزش احتمالی آن پیشگیری کرد. حرکت مشتریان از سازمانی به سازمانی دیگر برای یافتن محصولات و خدمات بهتر و ارزان‌تر را ریزش می‌گویند (هادن، ۲۰۰۷).

1. Customer Churn

ریـزش مشـتری از دیـد انگیـزه و نـوع، بـه دو دسـتهٔ اختیـاری (داوطلبانـه) و غیراختیـاری (غیرداوطلبانه) تقسیم می‌شود و از لحاظ رویکرد مدیریت، در دو گروه غیرهدفمند و هدفمند جای می‌گیرد. در رویکرد غیرهدفمند، راهبردها بر پایهٔ افزایش هزینه‌های جابه‌جایی مشتری (رفتن بـه سمت رقیب)، کمک به بهبود محصول، تبلیغات جذاب‌تر یا پیاده‌سازی برنامه‌های خاص وفاداری مشتری به‌منظور رضایت مشتری پی‌ریزی می‌شـود. رویکـرد هدفمنـد تـلاش دارد مشـتریان در معرض خطر را شناسایی کند و با اعمال اقدامات مداخله‌جویانه، وفاداری آنان را تحریـک کنـد و زمینهٔ ماندگاری آنان را فراهم آورد. این رویکرد شـامل دو نـوع انفعـالی (واکنشـی) و پیشـگیرانه (فعال) است (بارز و ون‌دن پل، ۲۰۰۷). در رویکرد انفعالی سازمان تا زمـانی کـه مشـتری ارتبـاط خود را قطع‌نکرده و علایم ریزش مشتری هنوز نمایان نشده است، منتظر می‌ماند، اما در رویکـرد پیشگیرانه، سازمان قبل از انجام هر اقدامی از جانب مشتری، بـه‌دنبـال شناسـایی و یـافتن علـل روی‌گردانی مشتری است و قصد دارد با اتخاذ سیاست‌های مناسب، آنان را به ماندن ترغیب کند. رویکرد این پژوهش، انتخاب هدفمند پیشگیرانه است که براساس تحلیل داده‌ها اجـرا مـی‌شـود. مدل مدیریت پیشگیرانه، چهار مرحلهٔ اصلی را دربرمی‌گیرد:

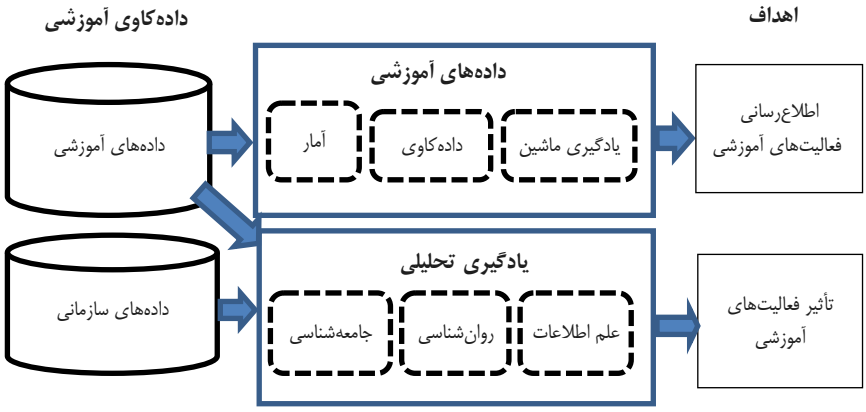
۱. تشخیص ریزش‌گران بالقوه؛
۲. درک و شناخت علل امکان ریزش؛
۳. طراحی راهبرد برقراری ارتباط یا ارائهٔ پیشنهاد مناسب برای ریزش‌گران؛
۴. نظارت و ارزیابی نتایج.

مرحلهٔ اول و دوم بر اساس داده‌های در دسترس تحقق می‌یابد و ماحصل آن تشخیص زمان ریزش و شناسایی دانشجویان در معرض انصـراف و علـل مـؤثر در آن اسـت. مرحلـهٔ سـوم بـه خلاقیت دانشگاه و توانایی تحریک دانشجو و ایجاد انگیـزه بـرای بقـای دانشـجو بسـتگی دارد. (سیاست‌های حمایتی دانشگاه تهران). مسلم است مرحلـهٔ چهـارم کارآمـدی مـدل پـیش‌بینـی و اقدامات اصلاحی را محک خواهد زد.

ونتورا و رومرو برای اولین‌بار بحث داده‌کاوی آموزشی را مطـرح کردنـد و فعالیـت تخصصـی آنان به‌طور رسمی از سال ۲۰۰۷ میلادی با برگزاری اولین کنفرانس جهانی داده‌کاوی آموزشـی توسط انجمن بین‌المللی داده‌کاوی آموزشی آغاز شد. اولین اقدام در این زمینه را کمپل و ابلینگـر (۲۰۰۷) با معرفی روش جدید «تجزیه‌وتحلیـل آموزشـی» و بـا اسـتفاده از روش‌هـای آمـاری و به‌کارگیری تکنیک‌های داده‌کاوی انجام دادند. هـدف آنهـا کمـک بـه اعضـای هیئـت علمـی و مشاوران آموزشی به‌منظور شناسایی و کمک به دانشجویان در معرض خطر بود. آنـان بـه نـوعی پایه‌گذار شاخه‌ای از داده‌کاوی بودند که در سال‌های بعـد «داده‌کـاوی آموزشـی یـا بـه اختصـار EDM» نامیده شد.

بیکر و یاسف در سال ۲۰۰۹ ضمن ارائهٔ تعریف رسمی برای داده‌کاوی آموزشی، آن را رشته‌ای در حال ظهور معرفی کردند که با تمرکز بر توسعهٔ روش‌هایی به‌منظور تحقیق و بررسی مبتنی بر داده‌های برگرفته از فعالیت‌های آموزشی، برای درک و شناسایی بهتر رفتار دانشجو و تنظیم روش یادگیری مطلوب به‌کار می‌رود و بر اهمیت تجزیه‌وتحلیل داده‌های آموزشی برای توسعهٔ مدل‌های بهبود فعالیت‌های یادگیری و همچنین بهبود اثربخشی سازمانی، تأکید دارد.

بینکوفسکی در کنفرانس بین‌المللی جاسیک ساکای سال ۲۰۱۲ اظهار کرد، می‌توان از داده کاوی آموزشی به‌منزلهٔ «یادگیری تحلیلی» به‌کمک مجموعه‌ای از داده‌های حجیم آموزشی یاد کرد. شکل ۱ نظریهٔ بینکوفسکی را به تصویر کشیده است. داده‌کاوی آموزشی از روش‌ها و تکنیک‌های آماری، یادگیری ماشین و داده‌کاوی، به‌منظور تجزیه‌وتحلیل داده‌های جمع‌آوری‌شده برای توسعه و بهبود بهره می‌برد و بر ایجاد مخازن اطلاعات ساختارمند به‌منزلهٔ منابع داده در داده‌کاوی تمرکز می‌کند.



شکل ۱. نظریهٔ بینکوفسکی در کنفرانس جاسیک ساکای

داده‌کاوی زمینه‌هایی را پوشش می‌دهد که به شرح زیر فهرست شده است:

۱. یافتن علل احتمالی تداوم یا انصراف تحصیل دانشجو؛
۲. ساخت مدل رفتار دانشجو به‌منظور تبیین تفاوت‌های رفتاری دانشجویان و پیش‌بینی مبتنی بر آن؛
۳. فراهم‌آوری محیط‌های فراگیر فردی و سیستم‌های پیشنهاددهنده؛
۴. فراهم‌آوری زمینهٔ کار با حجم زیادی از داده‌های ذخیره‌شده توسط کاربران غیرمتخصص و کسب دانش عمیق دربارهٔ متدها، فرایندها و الگوریتم‌های یادگیری (گارسیا، رومرو، ونتورا و کاسترو، ۲۰۱۱)؛

۵. پشتیبانی مؤثرتر آموزشی از دانشجویان، به‌کمک نرم‌افزارهای یادگیری فردی و یادگیری مشارکتی (پچنیزکی، کالدر، واسیلی اوا و برا، ۲۰۰۸)؛
۶. بررسی و مطالعهٔ تئوری‌های آموزشی؛
۷. بررسی، مطالعه و گسترش تئوری‌ها و نظریه‌های آموزشی برای یافتن شواهد تجربی به‌منظور ادراک عمیق از عوامل کلیدی مؤثر در یادگیری (بک و موستاوو، ۲۰۰۸).

در جدول ۱ فراوانی الگوریتم‌های استفاده‌شده در مطالعات مختلف طی سال‌های ۲۰۰۳ تا ۲۰۱۲، همراه با اسامی محققان و سال پژوهش درج شده است.

جدول ۱. فراوانی الگوریتم‌های استفاده‌شده در مطالعات داده‌کاوی آموزشی

الگوریتم	تعداد	نام محقق (سال)
شبکه‌های عصبی	۱۳	نپلاوا و اسنورک (۲۰۰۳)، بیکر، ترفالیس و رهاودس (۲۰۰۴)، هرزوگ (۲۰۰۶)، ساپربای، وندامی و مسکنز (۲۰۰۶)، ساجیتپاراپیتا (۲۰۰۶)، وانگ (۲۰۰۵)، پیتمن (۲۰۰۸)، رومرو،ونتورا، بارنز و دسماریس (۲۰۰۹)، لیکورنزو، جیان نوکوس، نیکولوپوس، امپاردیس و لاموس (۲۰۰۹)، وای یو، دیگانگی، جانش پنل و کارپولت (۲۰۱۰)، دلن (۲۰۱۰)، کبکچیوا (۲۰۱۱)، بوگارد، جیمز، هلبیگ و هاف (۲۰۱۲).
درخت تصمیم‌گیری	۱۳	هرزوگ (۲۰۰۶)، ساپربای و همکاران (۲۰۰۶)، ساجیتپاراپیتا (۲۰۰۶)، وانگ (۲۰۰۵)، پیتمن (۲۰۰۸)، رومرو و همکاران (۲۰۰۹)، ناندشوار و چادری (۲۰۰۹)، دکر، پچینزکیو و ولیشوور (۲۰۰۹)، ژانگ، اوسنا، کلارک و کیم (۲۰۱۰)، وای یو، دیگانگی، جانش پنل و کارپولت (۲۰۱۰)، کواچیچ (۲۰۱۰)، دلن (۲۰۱۰)، کبکچیوا (۲۰۱۳).
رگرسیون لجستیک	۸	هرزوگ (۲۰۰۵ و ۲۰۰۶)، ساجیتپاراپیتا (۲۰۰۶)، وانگ (۲۰۰۵)، پیتمن (۲۰۰۸)، دکر و همکاران (۲۰۰۹)، دلن (۲۰۱۰)، بوگارد و همکاران (۲۰۱۲).
ماشین بردار پشتیبان	۳	بیکر و همکاران (۲۰۰۴)، لیکورنزو و همکاران (۲۰۰۹)، ژانگ و همکاران (۲۰۱۰)،
بیزین	۳	پیتمن (۲۰۰۸)، دکر و همکاران (۲۰۰۹)، ژانگ و همکاران (۲۰۱۰).
جنگل تصادفی	۲	ساپربای و همکاران (۲۰۰۶)، دکر و همکاران (۲۰۰۹).
One R	۲	دکر و همکاران (۲۰۰۹)، کبکچیوا و همکاران (۲۰۱۲).
تحلیل تشخیص خطی	۱	ساپربای و همکاران (۲۰۰۶).
ژنتیک	۱	رومرو و همکاران (۲۰۰۹).
کا نزدیک‌ترین همسایه	۱	کبکچیوا (۲۰۱۲).
مارس (MARS)	۱	وای یو و همکاران (۲۰۱۰).

جدول ۲ اطلاعات اجرای داده‌کاوی انصراف در برخی از دانشگاه‌ها را نشان می‌دهد. در ستون دوم، عوامل مؤثر بر انصراف، الگوریتم‌های به‌کاررفته و دستاورد با نشانهٔ / از یکدیگر جدا شده‌اند. جای خالی به‌معنای در دسترس نبودن اطلاعات است.

جدول۲. اطلاعات پیشینهٔ داده‌کاوی آموزشی انصراف دانشجو

محل پژوهش (سال)	علل مؤثر بر انصراف / الگوریتم‌های استفاده‌شده / دستاورد
دانشگاه ویچیتا (۱۹۸۶)	بی‌تأثیربودن معدل دورهٔ متوسطه در انصراف / / مشخص‌شدن علل انصراف
چند دانشگاه مشترک (۱۹۹۵)	/ رگرسیون خطی / تأثیر نمره‌های آزمون و ساعات حضور در کلاس
چند دانشگاه مشترک (۱۹۹۵)	/ / شناسایی عوامل مؤثر در جذب دانشجوی بیشتر
دانشگاه جنوا (۱۹۹۹)	/ / ارائهٔ مدلی برای شناسایی دانشجویان پرخطر
دانشگاه مهندسی برق چک (۲۰۰۳)	/ شبکهٔ عصبی/ مدل پیش‌بینی، بی‌تأثیربودن معدل دبیرستان، بی‌تأثیری مستقیم سن
دانشگاه میشیگان (۲۰۰۳)	/ / ارائهٔ مدل پیش‌بینی، نیاز به توصیه‌هایی برای جلوگیری از انصراف
دانشگاه اوکلاهاما (۲۰۰۴)	/ شبکهٔ عصبی، ماشین بردار پشتیبان/ استفاده از داده‌های محیطی در کنار داده‌های هویتی و جمعیت‌شناسی
دانشگاه فلوریدا (۲۰۰۵)	کیفیت یادگیری دورهٔ متوسطه، انگیزه‌های شغلی و علایق شخصی/ رگرسیون لجستیک، درخت تصمیم و شبکهٔ عصبی /
چند مؤسسه آموزشی مشترک (۲۰۰۵)	/ درخت تصمیم‌گیری، شبکه عصبی و رگرسیون لجستیک / ارائه مدل پیش‌بینی
دانشگاه‌های فرانسه و بلژیک (۲۰۰۶)	سوابق تحصیلی و وضعیت خانوادگی ـ اجتماعی / درخت تصمیم‌گیری، شبکهٔ عصبی و جنگل تصادفی / شناسایی دانشجویان زیاد، متوسط و کم‌خطر
بانک دادهٔ ملی دانشجویان آمریکا (۲۰۰۶)	/ درخت تصمیم‌گیری C۵/۰، شبکهٔ عصبی و رگرسیون لجستیک / تغییر در شیوه‌های یادگیری، تغییر در کمپوس دانشگاه و تغییر در فرهنگ دانشجو و والدین
دانشگاه‌های مجازی آمریکا (۲۰۰۷)	/ / ساعات مطالعهٔ دروس دانشگاهی، عوامل مکانی (عوامل مالی و عاطفی بر آنها تأثیر می‌گذارند)
دانشگاه گلارمورگان بریتانیا	/ درخت پاسخ SPSS / ارائهٔ مدل پیش‌بینی
دانشگاه اورگان جنوبی (۲۰۰۸)	/ درخت‌تصمیم‌گیری، شبکهٔ عصبی، رگرسیون لجستیک و شبکهٔ بیز/ مدل پیش‌بینی
دانشگاه کوردوبا اسپانیا (۲۰۰۸)	/ درخت تصمیم‌گیری، شبکهٔ عصبی و درخت ژنتیک / طبقه‌بندی دانشجویان
دانشگاه‌های دولتی آمریکا (۲۰۰۸-۲۰۰۴)	/ درخت تصمیم‌گیری، ماشین بردار پشتیبان، شبکهٔ عصبی و رگرسیون لجستیک/ ارائهٔ مدل پیش‌بینی
دانشگاه فنی آتن (۲۰۰۹)	/ شبکهٔ عصبی پیش‌رو، ماشین بردار پشتیبان / ارائهٔ مدل پیش‌بینی
دانشگاه ویرجینیا (۲۰۰۹)	/ درخت تصمیم‌گیری / ارائهٔ مدل پیش‌بینی
دانشگاه آیندهوون (۲۰۱۰- ۲۰۰۹)	/ درخت تصمیم‌گیری، شبکهٔ بیز، رگرسیون لجستیک و جنگل تصادفی/ ارائهٔ پیش‌بینی
دانشگاه ولینگتون (۲۰۰۹ - ۲۰۰۶)	/ درخت تصمیم، سبد خرید / ارائهٔ مدل پیش‌بینی
دانشگاه آریزونا (۲۰۱۰)	/ درخت تصمیم‌گیری، رگرسیون چندمتغیرهٔ تطبیقی (MARS) و شبکهٔ عصبی / اقامت و قومیت
دانشگاه اقتصاد صوفیه (۲۰۱۲)	/ درخت تصمیم‌گیری، شبکهٔ عصبی و K نزدیک‌ترین همسایه / نمرهٔ کنکور
دانشگاه کنتاکی غربی (۲۰۱۲)	معدل دیپلم/ رگرسیون لجستیک، درخت تصمیم‌گیری و شبکهٔ عصبی/ ارائهٔ پیش‌بینی
کالج تگزاس جنوبی (۲۰۱۲)	/ / تعیین نرخ انصراف (دسته‌بندی دانشجویان)
دانشگاه آلاباما (۲۰۱۲)	/ / تأثیر تعدد سفرهای درون‌شهری و برون‌شهری برای حضور درکلاس درس
دانشگاه سانی (۲۰۱۲)	/ / شناسایی از طریق داشتن ۵ از ۷ مشخصه در ارزیابی عملکرد تحصیلی دانشجو
دانشگاه تبفین (۲۰۱۲)	/ / خوشه‌بندی دانشجویان انصرافی به دسته‌های پرخطر، متوسط‌الخطر و کم‌خطر

روش‌شناسی پژوهش

پژوهش پیش رو کاربردی و توصیفی است که با بهره‌مندی از داده‌های کمی ـ کیفی برمبنای روش کریسپ داده‌کاوی، به‌کمک نسخهٔ دوازدهم نرم‌افزار کلمنتاین از مجموعهٔ نرم‌افزار SPSS، روی اطلاعات دانشجویان شهریه‌پرداز (غیر رایگان) دانشگاه تهران طی سال‌های ۱۳۸۸ تا ۱۳۹۲ اجرا شده است. تأکید این پژوهش بر استفاده از داده‌های داده‌کاوی است؛ زیرا خبرگان داده‌کاوی به صراحت می‌گویند: «داده‌کاوی فرایند کشف دانش جالب از بین داده‌های حجیم ذخیره‌شده در پایگاه داده، انبار داده، یا سایر مخازن اطلاعات است» (هان و کمبر، ۲۰۰۶). بنابراین روش پژوهش پر پایهٔ استخراج و جمع‌آوری داده از بانک‌های اطلاعات دانشجویی استوار است. مجموعهٔ داده‌ها از ۲۱ صفت سن (۶۴ – ۱۸)، جنسیت (زن، مرد)، وضعیت اقامت (بومی/ غیربومی)، تأهل (مجرد، متأهل)، ملیت، آخرین ترم، جمع واحد گذرانده، تعداد مشروطی متوالی، تعداد مشروطی، تعداد ترم (۱۳–۱)، اولین ترم، آخرین وضعیت (شاغل، انصراف، و...)، جمع واحد گرفته‌شده، رشتهٔ تحصیلی، ترم ورود، گروه تحصیلی، مقطع تحصیلی (کارشناسی، ارشد، دکتری حرفه‌ای، دکتری تخصصی)، دورهٔ تحصیلی (روزانه، شبانه، روزانهٔ شهریه‌پرداز)، دانشکده، نام و نام خانوادگی و شمارهٔ دانشجویی شکل گرفته است. البته برخی از آنها ماهیت تحلیلی نداشتند و جنبهٔ کنترلی یا نقش ارتباط بین دو بانک اطلاعاتی را ایفا کردند. مهم‌ترین مسئلهٔ انتخاب صفت، هدف است. با توجه به عنوان پژوهش، از صفت آخرین وضعیت دانشجو، متغیر جدیدی به‌منزلهٔ صفت هدف با نام Attrition و مقادیر ۱ برای انصراف و صفر برای انصراف‌نداده، ایجاد شد. مدیران آموزشی و دانشجویی بر تأثیر وضعیت اقامت در انصراف اصرار داشتند و احتمال انصراف در کارشناسی را کم، اما برای کارشناس‌ارشد شهریه‌پرداز زیاد می‌دانستند. از آنجاکه سن دانشجویان پذیرفته‌شده در مقاطع و دوره‌های مختلف تحصیلی متفاوت است، هر سه صفت سن، مقطع و دورهٔ تحصیلی، صفت مؤثر در نظر گرفته شده است. پیشینهٔ پژوهش رأی به تأثیر ترم اول و دوم داده بود، بنابراین تعداد ترم نیز صفت مؤثر دیگر مد نظر قرار گرفت. با توجه به توضیحاتی که بیان شد، هفت صفت دورهٔ تحصیلی، مقطع، تعداد ترم، وضعیت اقامت، وضعیت تأهل، جنسیت و سن، متغیرهای مستقل یا صفات پیش‌بینی‌کننده انتخاب شدند و به تجزیه‌وتحلیل آنها پرداخته شد (صفت هشتم، صفت هدف است که پیش از این معرفی شد). میان داده‌ها، سن ۲۰۱ دانشجو مشخص نبود. برای رفع مشکل، ابتدا از شمارهٔ دانشجویی برای ارتباط بین بانک اطلاعات فردی و دانشجویی استفاده شد و با محاسبهٔ تفاوت سال تولد و سال ورود (قسمتی از ترم ورود) سن ۱۴۴ دانشجو به‌دست آمد. در ادامه برای برآورد سن ۵۷ دانشجو باقی‌مانده، از رایج‌ترین معیار، یعنی میانگین استفاده شد و برای افزایش دقت با توجه به ارتباط

مقطع و دورهٔ تحصیلی، وضعیت تأهل و جنسیت با سن، داده‌ها بر اساس این ویژگی‌ها دسته‌بندی شدند و میانگین هر دسته برای مقادیر متناظر مفقوده ثبت شد. یکی از آموخته‌های این پـژوهش آن است که باید رویه‌هایی اتخاذ شود که ثبت و ذخیرهٔ اطلاعات اساسی تضمین شود.

یافته‌های پژوهش

در الگوریتم‌هایی که پیش‌بینـی و ارائـهٔ مـدل مطـرح اسـت، داده‌هـا بـه دو مجموعـهٔ یـادگیری (داده‌هایی که الگوریتم جهت الگوبرداری از آنها یاد می‌گیرد) و تسـت (داده‌هـایی شـامل بـرآورد الگوریتم برمبنای شـناخت از مجموعـهٔ یـادگیری) تقسـیم مـی‌شـوند. از کـل مجموعـهٔ داده‌هـا (۲۱۴۲۰)، ۸۰ درصد به مجموعهٔ یادگیری و ۲۰ درصد به مجموعهٔ تست اختصاص یافتـه اسـت. تجزیه‌وتحلیل پژوهش طی سه مرحله انجام گرفت:

مرحلهٔ ۱. رتبه‌بندی اهمیت پیش‌بینی‌کننده‌ها و ارائهٔ جدول احتمال انصراف

از آنجاکه شبکهٔ بیزین، نمایشی از معناداری رابطه‌های نامشخص میـان پارامترهـای یـک حـوزه است، در این مرحله با مقایسهٔ الگوریتم‌های گونـاگون شـبکهٔ بیـزین، دقیـق‌تـرین آنهـا انتخـاب می‌شود. شبکهٔ بیزین، گراف جهت‌دار غیرحلقوی اسـت کـه از نودهـا بـرای نمـایش متغیرهـای تصادفی استفاده می‌کند و با کمان‌ها روابط احتمالی میان متغیرها را نشان می‌دهـد. کلمنتـاین دو روش برای ساخت مدل شبکه بیزین ارائه کرده است؛ یکی مدل شـبکهٔ بیـزین سـاده و دیگـری شبکه‌های بیزین توسعه‌یافته (ترکیبی). به‌منظور بهبود مدل در شبکه‌های ترکیبی، خروجـی یـک الگوریتم، ورودی الگوریتم دیگری قرار می‌گیرد. در جدول ۳ دقت چهار مدل شبکهٔ بیزین کـه از اجرای جریان‌های[1] کلمنتاین به‌دست‌آمده، نشان داده شده است.

جدول ۳. مقایسهٔ دقت الگوریتم‌های شبکهٔ بیزین

الگوریتم	درصد پیش‌بینی صحیح مجموعهٔ یادگیری	درصد پیش‌بینی صحیح مجموعهٔ تست
بیز ساده	۹۵/۴۵	۹۵/۵۲
TAN	۹۷/۴۶	۹۷/۳۴
مارکوف	۹۷/۴۶	۹۷/۲۷
ترکیبی مارکوف FS	۹۷/۵۱	۹۷/۴۳

1. Stream

با توجه به جدول ۳، الگوریتم ترکیبی مارکوف FS دقیق‌ترین الگوریتمی است که می‌تواند سبب بهبود مدل شود. پوشش مارکوف برای متغیر هدف در شبکهٔ بیزین، مجموعه‌ای از نودها است که والدین، فرزندان آنها و والدین فرزندان آنها را دربرمی‌گیرد. پوشش مارکوف تمام متغیرهایی که در شبکه برای پیش‌بینی متغیر هدف لازم است را تعریف می‌کند. جدول ۴ اهمیت پیش‌بینی‌کننده‌ها در انصراف (صفت هدف) را نشان می‌دهد (الگوریتم، سن را فاقد اهمیت تشخیص داده است). جدول‌های ۵، ۶ و ۷ احتمال شرطی انصراف بر مبنای گراف الگوریتم ترکیبی مارکوف FS را نشان می‌دهند.

جدول ٤. رتبه‌بندی اهمیت پیش‌بینی‌کننده‌های مدل الگوریتم مارکوف FS در انصراف

شرح	رتبهٔ ۱	رتبهٔ ۲	رتبهٔ ۳	رتبهٔ ٤	رتبهٔ ٥	رتبهٔ ٦
پیش‌بینی‌کننده	ترم	دورهٔ تحصیلی	جنسیت	مقطع تحصیلی	بومی/غیربومی	وضعیت تأهل
میزان اهمیت	۰/۶۸۷	۰/۰۸۱	۰/۰۶۸	۰/۰۶۴	۰/۰۶۳	۰/۰۳۸

جدول ٥. احتمال انصراف در ترم‌های مختلف به تفکیک وضعیت تأهل و جنسیت

پیش‌بینی‌کننده		احتمال انصراف در ترم					
وضعیت تأهل	جنسیت	اول	دوم	سوم	چهارم	پنجم	ششم
مجرد	مرد	۰/۵۹۵	۰/۲۳۱	۰/۰۶۹	۰/۰۴۴	۰/۰۲۲	۰/۰۲۹
مجرد	زن	۰/۶۸۶	۰/۱۳۹	۰/۰۳۳	۰/۰۵۰	۰/۰۲۱	۰/۰۶۷
متأهل	مرد	۰/۷۵۱	۰/۱۶۵	۰/۰۴۱	۰/۰۲۵	۰/۰۱۰	۰/۰۰۵
متأهل	زن	۰/۸۳۵	۰/۱۱۳	۰/۰۱۲	۰/۰۱۲	۰	۰/۰۲۵

جدول ٦. احتمال انصراف در مقاطع تحصیلی به‌تفکیک وضعیت تأهل و جنسیت

پیش‌بینی‌کننده		احتمال انصراف مقطع تحصیلی				
وضعیت تأهل	جنسیت	کارشناسی	ارشد	دکتری تخصصی	دکتری حرفه‌ای	دکتری دامپزشکی
مجرد	مرد	۰/۰۱۱	۰/۹۴۸	۰/۰۲۹	۰/۰۱۱	۰/۰
مجرد	زن	۰/۰۱۲	۰/۸۷۷	۰/۰۲۱	۰/۰۸۸	۰
متأهل	مرد	۰	۰/۹۵۳	۰/۰۴۷	۰	۰
متأهل	زن	۰/۰۳۷	۰/۹۴۹	۰/۱۰۳	۰	۰

جدول ۷. احتمال انصراف در دوره‌های تحصیلی به تفکیک وضعیت تأهل و جنسیت

پیش‌بینی‌کننده		احتمال انصراف دورهٔ تحصیلی		
وضعیت تأهل	جنسیت	روزانهٔ شهریه‌پرداز	شبانه	مجازی
مجرد	مرد	۰/۱۶۹	۰/۵۵۵	۰/۲۷۵
مجرد	زن	۰/۱۶۹	۰/۵۹۷	۰/۲۳۳
متأهل	مرد	۰/۲۱۲	۰/۱۴۵	۰/۶۴۲
متأهل	زن	۰/۲۲۷	۰/۱۵۱	۰/۶۲۰

مرحلهٔ ۲. یافتن مدل پیش‌بینی انصراف و ارتباط صفات

ارتباط بین متغیرها به‌کمک رگرسیون بررسی شد و به‌دلیل دودویی‌بودن صفت هدف، از رگرسیون لجستیک و کاکس در کلمنتاین استفاده شد. جدول ۸ مبین دقت بیشتر رگرسیون لجستیک نسبت به کاکس است.

جدول ۸. تجزیه‌وتحلیل دقت پیش‌بینی الگوریتم‌های رگرسیون لجستیک و کاکس

وضعیت پیش‌بینی (مدل)	مجموعهٔ دادهٔ یادگیری		مجموعهٔ دادهٔ تست		دقت کل مدل (درصد)
	تعداد	دقت (درصد)	تعداد	دقت (درصد)	
صحیح (رگرسیون کاکس)	۱۳۶۴۹	۹۵/۴۱	۴۰۹۴	۹۵/۵۴	۹۵/۴۰۷
صحیح (رگرسیون لجستیک)	۱۶۶۵۱	۹۷/۱۸	۴۱۶۳	۹۷/۱۵	۹۷/۱۷۵

روش اجرای رگرسیون لجستیک پیش‌رو گام‌به‌گام است. در این روش متغیرها به‌ترتیب از لحاظ معناداری وارد مدل می‌شوند. این عمل تا زمانی ادامه می‌یابد که خطای آزمون معناداری به ۵ درصد (یا ۹۵ درصد اطمینان) برسد. در هر مرحله استقلال پیش‌بینی‌کننده‌ها با صفت هدف (انصراف) بررسی می‌شود. جدول ۹ برخی اطلاعات به‌دست‌آمده از اجرای مدل رگرسیون لجستیک استقلال متغیرها را نشان می‌دهد (حذف صفت سن به‌معنای خطای آزمون بیش از ۵ درصد برای این صفت است که آنچه در شبکهٔ بیزین مطرح شد را تأیید می‌کند).

جدول ۹. تجزیه‌وتحلیل استقلال متغیرها و صفت هدف (انصراف) در رگرسیون لجستیک

گام	عامل واردشده (متغیر مستقل)	Chi-Square (a,b)	درجهٔ آزادی	Sig*	مقدار خی‌دو در سطح		استقلال/ وابستگی
					٪۵	٪۱	
۱	ترم	۲۴۱۳/۱۸۵	۱	۰/۰۰۰	۳/۸۴	۶/۶۳	وابسته
۲	دورهٔ تحصیلی	۲۵۸/۹۹۳	۲	۰/۰۰۰	۵/۹۹	۹/۲۱	وابسته
۳	مقطع تحصیلی	۱۳۸/۰۷۱	۴	۰/۰۰۰	۹/۴۹	۱۳/۲۸	وابسته
۴	وضعیت تأهل	۱۸/۰۸۴	۱	۰/۰۰۰	۳/۸۴	۶/۶۳	وابسته
۵	بومی/ غیربومی	۱۸/۰۱۹	۱	۰/۰۰۰	۳/۸۴	۶/۶۳	وابسته
۶	جنسیت	۸/۳۸۹	۱	۰/۰۰۴	۳/۸۴	۶/۶۳	وابسته

* با جای‌گذاری Sig در رابطهٔ ۱۰۰ × ((Sig×۲)− ۱) درصد معناداریِ اثر متغیر مستقل بر متغیر وابسته به‌دست می‌آید. مقادیر صفر به‌معنای آن است که با ۹۹/۹۹۹ درصد اطمینان، انصراف متأثر از عامل (یا عامل‌های) واردشده (تا آن گام) است. مقدار ۰/۰۰۴ برای جنسیت نشان می‌دهد با اطمینان ۹۹/۲ درصد، انصراف متأثر از جنسیت است.

برازش نهایی مدل که مناسب‌بودن آن را نشان می‌دهد، در جدول ۱۰ مشاهده می‌شود.

جدول ۱۰. اطلاعات برازش مدل در رگرسیون لجستیک

مدل	معیارهای برازش مدل			آزمون‌های نسبت Likelihood		
	AIC	BIC	-۲Log likelihood	Chi-Square	df	Sig.
شروع	۶۴۸۷/۳۴۸	۶۴۹۵/۳۲۰	۶۴۸۵/۳۴۸			
نهایی	۲۵۰۱/۷۰۸	۲۵۸۹/۴۰۰	۱۹۱۰/۲۱۶	۴۵۷۵/۱۳۲	۱۰	۰/۰۰۰

در سطح ۹۹ درصد اطمینان با توجه به مقدار زیر فرض استقلال متغیرها با صفت هدف (انصراف) رد می‌شود.

$\text{Chi-Square}_{(\text{مدل نهایی})}$ آمارهٔ خی‌دو $> \text{۴۵۷۵/۱۳۲} = \text{df} = ۱۰$ و $\alpha = ۰/۰۱$ (df = ۱۰) = ۲۳/۲۰۹۳

به بیانی دیگر، با اطمینان ۹۹ درصد می‌توان گفت شش صفت مندرج در جدول ۹ (متغیرهای مستقل) هدف را توجیه می‌کنند. معادلهٔ رگرسیونی خطر انصراف در زیر آمده است:

خطر انصراف =۳/۵ (کارشناسی) −۱۷/۴ + (دکتری دامپزشکی) ۲/۰ + (ارشد) ۲/۲ + (دکتری حرفه‌ای)

۱/۰ (روزانهٔ شهریه‌پرداز) −۱/۷ (مجازی) −۲/۳ (تعداد ترم) −۰/۴ (متأهل) −۰/۳ (غیربومی) −۲/۵ + (زن)

قواعد استنتاج به قرار زیر است:

درصد انصراف برای دانشجویان با افزایش تعداد ترم کاهش می‌یابد، برای دانشجویان مجازی کاهش نشان می‌دهد، برای شهریه‌پرداز روزانه افزایش می‌یابد، برای دانشجویان زن، غیربومی و

متأهل کاهش نشان می‌دهد، برای دکتری دامپزشکی کاهش می‌یابد و برای کارشناسی، کارشناس‌ارشد و دکتری حرفه‌ای افزایش نشان می‌دهد.

مقادیر کای‌اسکوئر جدول ۹ نشان‌دهندهٔ تأثیر شایان توجه ترم در تغییرات درصد انصراف (مؤثربودن) است. برای اینکه تشخیص داده شود تأثیر کدام ترم بیشتر است، در اجرای رگرسیون تغییری ایجاد می‌شود. در رگرسیون انجام‌گرفته متغیر ترم، متغیر عددی در نظر گرفته شده بود؛ حال اگر نوع این متغیر به‌صورت مجموعه[1] تعریف شود، مدل رگرسیون مقدار عددی هر ترم را به‌منزلهٔ یک متغیر وارد رگرسیون می‌کند و تحلیل را انجام می‌دهد. یکی از خروجی‌های اجرای جدید آمارهٔ نسبت برتری یا همان Exp(B) (به‌معنای اینکه یک واحد تغییر در انحراف معیار متغیر مستقل، سبب تغییر انحراف معیار متغیر وابسته به اندازه Exp(B) می‌شود) است. جدول ۱۱ این مقادیر را نشان داده است و جدول ۱۲رتبه‌بندی انصراف به تفکیک ترم براساس نسبت برتری را نشان می‌دهد.

جدول ۱۱. مقادیر نسبت برتری ترم‌های مختلف

آماره	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲
Exp(B)	۱۹E+۹	۶۵E+۷	۲۴E+۶	۲۸E+۷	۷۷E+۵	۱۶E+۶	۱۳E+۶	۴	۷	۸	۹	۳

جدول ۱۲. دسته‌بندی و رتبه‌بندی دانشجویان به لحاظ میزان خطر انصراف

رتبه‌بندی دانشجویان پرخطر					رتبه‌بندی دانشجویان بسیار کم‌خطر				
رتبهٔ ۱	رتبهٔ ۲	رتبهٔ ۳	رتبهٔ ۴	رتبهٔ ۵	رتبهٔ ۱	رتبهٔ ۲	رتبهٔ ۳	رتبهٔ ۴	رتبهٔ ۵
ترم ۱	ترم ۲	ترم ۶ و ۷	ترم ۳ و ۴	ترم ۵	ترم ۱۱	ترم ۱۰	ترم ۹	ترم ۸	ترم ۱۲

مرحلهٔ ۳. خوشه‌بندی بر مبنای خطر انصراف

چالشی که پرکاربردترین الگوریتم خوشه‌بندی، یعنی K میانگین[2] دارد، تعداد خوشه‌هایی است که کاربر باید مشخص کند. این الگوریتم به‌طور مکرر داده‌ها را درون خوشه‌ها قرار می‌دهد و هر بار مرکز خوشه (همان میانگین نقاط متعلق به خوشه) را تعیین می‌کند تا جایی که مجموع تشابه بین مرکز خوشه و همهٔ اعضای خوشه حداکثر شود و مجموع تشابه بین مراکز خوشه‌ها به حداقل

1. Set
2. K-means

برسد و دیگر امکان بهبود بیشتری وجود نداشته باشد. برای رفع مشکل از الگوریتم دوگام[1] استفاده شده است که در گام اول به ساختن خوشه‌های کوچک از رکوردها می‌پردازد و در گام دوم توسعهٔ خوشه به خوشه‌های بزرگتر را انجام می‌دهد. بدین ترتیب تعداد بهینهٔ خوشه‌ها به‌دست می‌آید (۳ خوشه). جدول ۱۳ نتایج K میانگین برای ۳ خوشه را نشان می‌دهد.

جدول ۱۳. اطلاعات مدهای صفات هر خوشه در الگوریتم K میانگین

شمارهٔ خوشه	درصد کل انصراف	میانگین سن	مقطع	درصد انصراف	بومی/ غیربومی	درصد انصراف	وضعیت تأهل	درصد انصراف	جنسیت	درصد انصراف	دوره	درصد انصراف	تعداد ترم	درصد انصراف
۱	۳۵	۳۰	ارشد	۹۴	بومی	۱۰۰	مجرد	۶۴	مرد	۱۰۰	شبانه	۴۴	۱	۶۱
۲	۳۵	۳۰	ارشد	۹۷	غیربومی	۱۰۰	مجرد	۵۷	زن	۶۷	مجازی	۵۳	۱	۷۸
۳	۳۰	۲۷	ارشد	۸۴	بومی	۹۳	متأهل	۷۹	زن	۱۰۰	شبانه	۵۵	۱	۶۹

قواعد استنتاج‌های الگوریتم K میانگین بر اساس اطلاعات جدول ۱۳ به شرح زیر است:

– میانگین سن ۳۵ درصد دانشجویان انصرافی ۳۰ سال است. ۶۱ درصد در ترم اول انصراف داده‌اند که ۹۴ درصد آنان کارشناس‌ارشد، ۱۰۰ درصد بومی، ۶۴ درصد مجرد و ۱۰۰ درصد مرد هستند و ۴۴ درصد نیز دورهٔ تحصیلی شبانه را می‌گذرانند.

– میانگین سن ۳۵ درصد دانشجویان انصرافی ۳۰ سال است. ۷۸ درصد در ترم اول انصراف داده‌اند که ۹۷ درصد آنان کارشناس‌ارشد، ۱۰۰ درصد غیربومی، ۵۷ درصد مجرد و ۶۷ درصد زن هستند و ۵۳ درصد دورهٔ تحصیلی مجازی را می‌گذرانند.

– میانگین سن ۳۰ درصد دانشجویان انصرافی ۲۷ سال است. ۶۹ درصد در ترم اول انصراف داده‌اند که ۸۴ درصد آنان کارشناس‌ارشد، ۹۳ درصد بومی، ۷۹ درصد متأهل و ۱۰۰ درصد زن هستند و ۵۵ درصد دورهٔ تحصیلی شبانه را سپری می‌کنند.

جدول ۱۴ دسته‌بندی دانشجویان را از دیدگاه خطر انصراف نشان می‌دهد.

جدول ۱٤. خوشه‌های مدل به تفکیک میزان خطر پیش‌رو

در معرض خطر زیاد انصراف	در معرض خطر متوسط انصراف	در معرض خطر کم انصراف
ارشد، مجرد، مرد، شبانه، ترم ۱	ارشد، مجرد، زن، مجازی، ترم ۱	ارشد، بومی، متأهل، زن، شبانه، ترم۱

1. Two Step

دانشجویان کارشناسی ارشد ترم اول و دوم، به‌ویژه دانشجویان ترم اول، در معرض بیشترین خطر انصراف قرار دارند. برای بهبود اطلاعات، الگوریتم‌های درخت تصمیم‌گیری در کلمنتاین با یکدیگر مقایسه شدند (دقت چاید ۹۵/۵، C۵/۰ ۹۷/۴۶ و C&R ۹۷/۴۶ درصد محاسبه شد). چون الگوریتم چاید در مقایسه با C۵/۰ و C&R احتمال مشروط صفات را نشان می‌دهد با در نظر گرفتن دقت کمتر (۱/۹۶درصد) از آن استفاده شد. نتایج در جدول ۱۵ نشان داده شده است.

جدول ۱۵. رتبه‌بندی درصد انصراف شاخه‌های درخت تصمیم چاید

میزان خطر	شرح خوشه	درصد انصراف
زیاد	دانشجوی ترم اول یا دوم با ۲۴ ≤ سن	۱۹/۷
	دانشجوی ترم اول یا دوم با ۳۳ ≤ سن > ۳۱	۱۷/۵
	دانشجوی ترم اول یا دوم با ۳۱ ≤ سن > ۲۴ متأهل بومی	۱۴/۷
	دانشجوی ترم اول یا دوم با ۳۷ ≤ سن > ۳۳	۱۴/۴
جمع		۶۶/۳
متوسط	دانشجوی ترم اول یا دوم با ۳۱ ≤ سن > ۲۴ متأهل	۸/۷
	دانشجوی ترم اول یا دوم با ۳۱ ≤ سن > ۲۴ مجرد غیربومی	۸/۴
	دانشجوی ترم سوم و سوم به بعد مجرد بومی	۷/۹
جمع		۲۵
کم	دانشجوی ترم اول یا دوم با ۳۷ < سن	۳/۸
	دانشجوی ترم سوم و سوم به بعد مجرد غیربومی	۲/۶
	دانشجوی ترم سوم و سوم به بعد متأهل	۲/۳
جمع		۸/۷

خلاصهٔ قواعد استنتاج را می‌توان چنین بیان کرد: ۸۷ درصد دانشجویان در ترم اول و دوم و ۱۳ درصد در ترم سوم و سوم به بعد انصراف می‌دهند. پس همچنان مخاطره‌آمیزترین دوره برای انصراف، ترم اول و دوم است. این دانشجویان به توجه و مساعدت بیشتری برای ادامهٔ تحصیل و ماندگاری دارند.

نتیجه‌گیری و پیشنهادها

درجهٔ اهمیت صفات پیش‌بینی‌کننده در انصراف دانشجو با دقت ۹۷ درصد، به‌ترتیب عبارت‌اند از: ترم (۰/۶۹)، دورهٔ تحصیلی (۰/۰۸)، جنسیت (۰/۰۷)، مقطع تحصیلی (۰/۰۶)، بومی/غیربومی (۰/۰۶) و وضعیت تأهل (۰/۰۴).

تقسیم‌بندی داده‌های الگوریتم‌های چاید و C۵/۰ به تفکیک ترم، تأییدی بر اهمیت حیاتی ترم است. در این راستا با دقت ۹۵/۵ درصد، ترم اول و دوم پرمخاطره‌ترین ترم برای انصراف معرفی می‌شود و دانشجویان کارشناسی ارشد ترم اول و دوم (به‌ویژه ترم اول)، در معرض بیشترین خطر از بُعد مقطع تحصیلی هستند و در این بین دانشجویان مرد مجرد شبانه بالاترین درصد انصراف را به خود اختصاص داده‌اند. احتمال انصراف دانشجویان ۲۴ سال یا کمتر و دانشجویان ۳۱ تا ۳۳ سال در ترم اول یا دوم ۰/۳۷ درصد، دانشجویان متأهل بومی ۲۴ تا ۳۱ سال در ترم اول یا دوم ۰/۱۵ درصد و بین ۳۳ تا ۳۷ سال در ترم اول یا دوم ۰/۱۵ درصد است. در مجموع احتمال انصراف دانشجویان ترم اول و دوم کمابیش ۰/۸۷ درصد و سایر ترم‌ها ۰/۱۳ درصد است که دانشجویان ارشد، مجرد، مرد، شبانه، ترم ۱ پرخطر؛ دانشجویان ارشد، مجرد، زن، مجازی، ترم ۱ در سطح متوسط و دانشجویان ارشد، بومی، متأهل، زن، شبانه، ترم ۱ کم‌خطر شناسایی شدند.

جدول‌های پیش‌بینی احتمال انصراف دانشجو (جدول‌های ۵، ۶ و ۷) تأیید می‌کنند که باید بر دانشجویان ترم اول و دوم در مقطع کارشناسی ارشد تمرکز بیشتری کرد تا بتوان زمینهٔ ماندگاری آنان را فراهم آورد. شایان توجه اینکه مشتری ارضاشده نه‌تنها مبلغ خوبی است و می‌تواند در جذب مشتری جدید کمک کند، بلکه می‌تواند اعتبار کسب‌وکار در بازار را افزایش دهد. از سویی دیگر مشتری روی‌گردان، سرمایهٔ از دست رفته‌ای است که می‌تواند زمینهٔ ترک مشتریان بیشتری را نیز فراهم کند (در جهت خلاف اولویت حفظ مشتری موجود).

ارائهٔ پیشنهاد و توصیه، نیازمند اطلاع از اهداف سازمانی، خط‌مشی و راهبردها، دیدگاه‌ها و انگیزهٔ مدیران ارشد، وضعیت منابع مختلف و شدت تغییرپذیری محیط درون و برون‌سازمانی است. از آنجاکه اطلاعی از این موارد در دست نیست، ارائهٔ پیشنهاد آسان نخواهد بود، اما با توجه به یافته‌های پژوهش و دستاوردهای جهانی (مطروح در پیشینه) می‌توان به موارد زیر اشاره کرد:

- به‌منظور استخراج دانش برای بهبود فعالیت‌ها و فرایندهای کاری، داده‌های بانک‌های اطلاعاتی به‌طور دائم تحلیل شوند؛
- ضمن ارزیابی عملکرد دانشجویان باید این فرصت را برای دست‌اندرکاران آموزش و اعضای هیئت علمی فراهم آورد که شرایط ویژه‌ای را در اختیار دانشجویان در معرض خطر قرار دهند و زمینهٔ ماندگاری آنان را فراهم کنند. شرایط ویژه می‌تواند مشاورهٔ بهتر، کمک به انتخاب دروس، کمک در انجام تکالیف آنان به‌منظور ایجاد انگیزه و برنامهٔ کاربردی متناسب با شغل آتی را دربرگیرد (تمرکز بیشتر بر دانشجوان ترم اول و دوم کارشناسی ارشد)؛

- درصد ریزش مشتری دانشگاه تهران ۴/۶ درصد (۹۷۸ انصرافی از ۲۱۴۲۰ دانشجو شهریه‌پرداز)، به‌معنای از دست‌دادن ۴/۶ درصد درآمد جدید است؛ درآمدی که می‌تواند برخی از مشکلات کمبود اعتباری را حل کند یا پیاده‌سازی برنامهٔ توسعه‌ای را پی‌ریزی و بهبود بخشد.

آنچه این پژوهش را از سایر پژوهش‌ها متمایز می‌کند، بیان شیوهٔ نگرش مدیریت فناوری اطلاعات به آموزش شهریه‌پرداز به‌منزلهٔ کسب‌وکار، طرح داده‌کاوی آموزشی، معرفی الگوریتم‌های مناسب در تحلیل مسائل آموزشی و تأکید بر استفاده از داده‌های ذخیره‌شده به‌مثابهٔ منبع ارزشمند در تجزیه‌وتحلیل است. باید آینده‌نگری در ساختار بانک‌های اطلاعاتی مد نظر قرار گیرد و هر نوع داده‌ای به‌منظور استفاده در تحلیل‌های آینده ذخیره شود و با در نظر گرفتن تمهیداتی از داده‌های ذخیره‌شده، نگهداری شود.

ضرورت این مطالعه و مطالعات مشابه را می‌توان کاهش هزینه‌های تولید، افزایش درآمد و درنتیجه، بهبود کیفیت محصول (فارغ‌التحصیل) و پیشگیری از اتلاف منابع، کنترل و کاهش نرخ ریسک ریزش و افزایش یا حفظ اعتبار مؤسسهٔ آموزش عالی دانست.

References

Baker, R. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1 (1): 3-17.

Barker, K., Trafalis, T. & Rhoads, T. (2004). *IEEE Systems and Information Engineering Design Symposium,* University of Oklahoma.

Beck, J. & Mostow, J. (2008). A Case Study Empirical Comparsion of Three Methods. *9th international conference on Intelligent Tutoring Systems*, June 23-27, Montral Canada.

Bienkowski, M., Feng M. & Means B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*, Available in: http://www.cra.org/ccc/files/docs/learning-analytics-ed.pdf.

Bogard, M. (2013). *A Data Driven Analytic Strategy for Increasing Yield and Retention at Western Kentucky University Using SAS Enterprise BI and SAS Enterprise Miner*. Available in: http://support.sas.com/resources/papers/proceedings13/044-2013.pdf.

Bogard, M., James, C., Helbig, T. & Huff, G. (2012). Using SAS® Enterprise BI and SAS®Enterprise Miner TM to Reduce Student Attrition. *SAS Conference Proceedings: SAS Global Forum 2012*, April 22-25, Orlando, Florida.

Burez, J. & Van den Poel, D. (2007). Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications,* 32(2): 277–288.

Buttle, F. (2008). *Customer Relationship Management*, UK, Routledge.

Campbell, J.P., Oblinger, D.G. (2007). *Academic Analytics*. Available in: http://net.educase.edu/ir/library/pdf/PUB6101.pdf.

Dekker, G.W., Pechenizkiy, M. & Vleeshouwers J.M. (2009). Predicting student drop out: A case study. *2nd International Educational Data Mining Conference*, July 1-3, Cordoba Spain.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49 (4): 498-506.

García, E., Romero, C., Ventura, S. & Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education,* 14 (2): 77-88.

Hadden, J., Tiwari, A., Roy, R. & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers and Operations Research,* 34(10): 2902–2917.

Herzog, S. (2005). Measuring Determinants of Student Return VS. Dropout/ Stopout VS. Transfer: A First-to-Second Year Analysis of New Freshmen. *Research in Higher Education*, 46(8): 883–928.

Honigman, B. (2013). *5 Secrets to Increasing Customer Retention -- and Profits*, Available in: http://www.entrepreneur.com/article/227946

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1): 61–72.

Kabakchieva, D., Stefanova, K. & Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Predicting Performance, *4th International Conference on Educational Data Mining,* Eindhoven, July 6-8, The Netherlands.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3): 950-965.

Murphy, E.C. & Murphy, M.A. (2013). *Leading on the Edge of Chaos, John Wiley & Sons, Canada.* Available in: http://www.impactlearning.com/resources /metrics /customer-retention.

Nandeshwar, A. & Chaudhari, S. (2009). *Enrollment prediction models using data mining*. Avialable in: http://nandeshwar.info/wp-content/uploads/2008/11/DM WVU_Project.pdf.

Nandeshwar, A., Menzies, T. & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications*, 38(12): 14984–14996.

Náplava, P. & Šnorek, M. (2003). Modeling of Student's Quality by Means of GMDH Algorithms. *System Analysis Modeling Simulation,* 43(10): 1415-1426.

Pechenizkiy, M., Calders, T., Vasilyeva, E. & Bra, P. D. (2008). Mining the student assessment data: Lessons drawn from a small scale case study. *in 1st International Educational Data Mining Conference (EDM2008).* Montreal Canad, June 20-21 2008.

Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention.* PhD thesis, Nova Southeastern University, USA.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1): 135–146.

Romero, C. (2008). Data mining algorithms to classify students. *in 1st International Educational Data Mining Conference (EDM2008),* June 20-21 Canada.

Romero, C., Ventura, S., Barnes, T. & Desmarais, M. (2009). Recommendation in higher education using data mining techniques, *2nd International Conference on Educational Data Mining,* EDM 2009, July 1-3, Cordoba Spain.

Scott, G., Shah, M., Grebennikov, L. & Singh, H. (2008). Improving student retention: A University of Western Sydney case study. *Journal of Institutional Research*, 14(1): 9–23.

Sujitparapitaya, S. (2006). Considering student mobility in retention outcomes. *New Direction for Institutional Research*, 2006(131): 35-51.

Superby, J., Vandamme, J. & Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. *Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006),* Hong kong, 37-44.

Tantuco N. & Uy R. (2014). *Creating Long-term Loyalty Relationships*. Available in:http://catalogue.pearsoned.co.uk/assets/hip/gb/hip_gb_pearsonhighered/samplechapter/0273755021.pdf.

Wang, M.C. (2005). Using Data Mining Techniques to Predict Student Development and Retention. *in 2005 National Student Affairs Assessment and Retention Conference*, June 3, USA.

Yu, C. H., Digangi, S., Jannasch-pennell, A. & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8(2): 307–325.

Yu, Ch. H., Digang, S., Jannasch, A., Kaprolet, Ch. (2010). A Data Mining Approach for Identifying Predictors of Student from sophomore to Junior Year. *Journal of Data Science*, 8 (2): 307-325.

Zhang, Y., Oussena, S., Clark, T. & Kim, H. (2010). Use Data Mining To Improve Student Retention in Higher Education - A Case Study. *in 12th International Conference on Enterprise Information Systems (ICEIS),* June 8-1, Madeira Portugal.