

به کارگیری تکنیک‌های داده ناوی در راستای پیتی بینی عوامل موثر در

موفقیت دانشجویان رشته ریاضی کاربردی در فارغ التحصیلی

مریم جلالی شاهرود^۱

^۱ دانشگاه صنعتی شاهرود، jalalim.sut70@gmail.com

چکیده

عصر انفجار اطلاعات هشدار بزرگی برای سازمانها و جامعه اطلاعاتی محسوب می‌شد مبنی بر اینکه سازمان ها علیرغم حجم انبوه داده‌ها با فقر دانش اطلاعاتی مواجه هستند، به همین دلیل علم داده کاوی که از جمله علوم کاربردی ونوین است، جایگاه ویژه‌ای را به خود اختصاص داد. داده کاوی را می‌توان استخراج اطلاعات، دانش و کشف الگوهای پنهان از پایگاه داده‌های بسیار بزرگ تعریف کرد. در این پژوهش نیز در راستای بررسی عوامل مؤثر بر پیشرفت تحصیلی دانشجویان رشته ریاضی کاربردی و کشف مدل‌های پنهان با استفاده از پارامترهای تعیین کننده، از این علم بهره گرفته شده است.

با توجه به اینکه ارتقاء سطح مهارت و پیشرفت تحصیلی دانشجویان که سرمایه‌های آتی هستند، بستر ساز توسعه و پیشرفت در تولید می‌گردد باید عوامل مؤثر بر پیشرفت تحصیلی بررسی گردد تا بتوانیم در جهت بهبود آموزش و یادگیری، سرمایه گذاری‌ها، صرفه جویی در هزینه ها، تقویت پروسه کارآفرینی، ارتقاء رتبه دانشگاه‌ها، و پیشرفت علم وصنعت، برنامه ریزی جامع و کاملی انجام دهیم. طبق نتایج بدست آمده از این پژوهش و با توجه به شناسایی عوامل مؤثر بر پیشرفت تحصیلی و پیش بینی الگوهای پنهان، دانشجویان و مدیران دانشگاهی می‌توانند سیاستهای درستی را در راستای بهبود پیشرفت تحصیلی اتخاذ نمایند.

کلمات کلیدی

داده‌کاوی، درخت تصمیم، پیش بینی معدل، مدل های پنهان

۱. مقدمه

تا حدی از میان برداشته شود تا به سطح موفقیتی مشابه سایر حوزه‌ها نظیر داده‌کاوی پزشکی و داده‌کاوی تجارت الکترونیک دست یابیم [۱۱]. پژوهش‌های متعددی از جمله استخراج قوانین انجمنی وتحلیل خوشه بندی، جهت پیش‌بینی افت تحصیلی دانشجویان [۱۳]، یافتن عوامل مؤثر بر بهبود تحصیل دانشجویان با استفاده از روند تحلیلی درخت تصمیم [۱۲] و...، در راستای تحقق این امر صورت گرفته است. همچنین به سبب این‌که نیروهای تربیتی فارغ التحصیل از مؤسسات آموزش عالی، شالوده ونیروی محرکه سازمانها و ادارات می باشند، به همین دلیل در این پژوهش به کشف الگوهای پنهان در سیستم‌های آموزش عالی پرداخته‌ایم. تسهیل فرآیندهای آموزشی مؤثر، کارا ودقیق وتصمیم‌گیری بهتروسریع، از نتایج به کارگیری این الگوهای کشف شده است.

عوامل متعددی در پیش‌بینی موفقیت یا عدم موفقیت دانشجویان دخیل است که در این پژوهش عواملی چون معدل کل دانشجو، ترم، جنسیت، تعداددروس (پایه وعمومی)، معدل دروس (پایه وعمومی) و... در جهت پیش‌بینی موفقیت یا عدم موفقیت دانشجویان ریاضی کاربردی در

امروزه اکثر دانشگاه‌ها وسازمان‌های کشورهای درحال توسعه، سرشار از پایگاه داده‌های عظیم، حجم انبوه اطلاعات و فقدان سازماندهی اثربخش است. درسال‌های اخیر با توجه به فرآیند دشوارکشف و استخراج دانش از این حجم وسیع منابع اطلاعاتی، تکنیکهای داده‌کاوی وبهره‌گیری مناسب از آن مورد توجه قرار گرفته است. درواقع می‌توان گفت داده کاوی، فرآیند کشف دانش از پایگاه‌های اطلاعاتی به منظور شناسایی الگوهای معتبر، مفید وقابل فهم درداده‌ها می‌باشد. تصمیم‌گیری مناسب وبه موقع از جمله موارد ضروری اجرای فرآیندهای با کیفیت وسازگار با منابع است به همین دلیل، تعریف وشناسایی الگوها ودانش نهفته در این اطلاعات می‌تواند در راستای برنامه‌ریزی، ارتقاء وبهبود سطح کیفی فرآیندهای آموزشی واجرائی، به مسئولین وتصمیم‌گیرندگان سازمانها وآموزش عالی در ارزیابی ومشاورة کمک شایانی نماید [۱].

با توجه به اینکه داده‌کاوی در حوزه آموزش، ازجمله زمینه‌های تحقیقاتی نوینی می‌باشد، نیازمند آن هستیم که در این حوزه تحقیقات جهت‌دار، تخصصی وگسترده‌تری صورت بگیرد و شکاف دانشی در این حوزه

کارگرفته شده در این پژوهش است [۵]. ارائه مدل‌ها و انجام داده‌کاوی با استفاده از نرم‌افزار Microsoft SQL server 2008R2 صورت گرفته و نتایج موردنظر مراحل مختلف داده‌کاوی بیان شده است.

۴. درخت تصمیم

یک درخت تصمیم یک ساختار سلسله مراتبی است که هر گره آن نشان-دهنده تست مقدار مشخصه بوده و هر انشعاب یک خروجی تست را نمایش می‌دهد. برگ‌های درخت نیز دسته‌ها یا توزیع هر دسته را مشخص می‌نمایند. درخت تصمیم قابل تبدیل به قوانین دسته‌بندی یا همان قوانین تصمیم می‌باشد. در برخی موارد نیز افراد ترجیح می‌دهند مقدار یک مشخصه و نه دسته آن را پیشگویی نمایند، که در این حالت به یافتن مقدار مشخصه، پیشگویی اطلاق می‌گردد. در هر حال پیشگویی، تخمین مقدار و برچسب دسته را با هم در برمی‌گیرد [۶ و ۷].

با استفاده از یک درخت تصمیم می‌توانیم برای هریک از مشاهدات X یک مقدار پیش‌بینی شده Y را پیدا کنیم. برای این منظور از ریشه درخت آغاز می‌کنیم، خصوصیات مربوط به ریشه را در نظر می‌گیریم و تعیین می‌کنیم که مقدار مشاهده شده برای خصوصیت معلوم به کدام شاخه تعلق دارد. آنگاه گره ای را در نظر می‌گیریم که شاخه مورد نظر به آن می‌رسد. این کار را برای این گره نیز انجام می‌دهیم و به همین صورت ادامه می‌دهیم تا به یک برگ برسیم. مقدار Y_S منتسب به برگ S ام مقدار پیش‌بینی شده برای X خواهد بود. بنابراین، درخت تصمیم مدل وابستگی T را برای Y از X بصورت $(Y=TX)$ بدست می‌دهد [5].

۵. فرایند آماده سازی اطلاعات

در این پژوهش با استفاده از تکنیک داده‌کاوی درخت تصمیم موجود در محیط Business intelligence development studio. پارامترهای مؤثر در موفقیت یا عدم موفقیت دانشجویان این رشته در طول ترم تحصیلی شناسایی می‌گردد و در نهایت مدل مربوط به این عوامل ارائه شده است. ورودی‌های مربوط به این مدل در جدول (۱) ذکر شده است.

جدول ۱- ورودی‌های مدل تحلیلی برای ایجاد درخت تصمیم

id	شماره دانشجویی
sex	جنسیت
gpa	معدل کل
sysno	دوره
basenum	تعداد واحدهای دروس پایه
generalnum	تعداد واحدهای دروس عمومی
twotrmavg	معدل دو ترم اول
twoagototalpassunit	تعداد واحد پاس شده دو ترم اول
lvlno	مقطع تحصیلات
twogeneralavg	میانگین معدل دروس عمومی دو ترم اول
twobaseavg	میانگین معدل دروس پایه دو ترم اول

فارغ‌التحصیلی در نظر گرفته شده است. همچنین با استفاده از الگوریتم درخت تصمیم، الگویی برای پیش‌بینی این مورد ارائه نموده ایم. در حقیقت عوامل مؤثر بر معدل فارغ‌التحصیلی دانشجوی معرفتی و توصیف می‌شود تا به کمک آگاهی از این عوامل، دانشجوی بتواند مطلوب‌ترین نتیجه را در فارغ-التحصیلی بگیرد [۲].

۲. روند اجرایی داده کاوی

باتوجه به اینکه داده‌کاوی علم کشف و استخراج اطلاعات از پایگاه داده‌های عظیم است، به همین دلیل کشف دانش داده‌ها دارای مراحل مختلفی می-باشد، که این تحقیق طی انجام چند فاز اصلی صورت گرفته است [۳ و ۴].

- جمع آوری داده‌ها، بررسی و تحلیل آن‌ها و در نهایت انتخاب داده‌های مناسب
- آماده سازی داده‌ها
- طراحی با درخت تصمیم
- کشف عوامل مؤثر بر پیشرفت و موفقیت دانشجویان و مقایسه نتایج

در این مقاله برای پیش‌بینی معدل نهایی دانشجوی در فارغ‌التحصیلی ۴ حالت مختلف براساس ۲ ترم اول، ۳ ترم اول، ۲ ترم و ۳ و ۴ در نظر گرفته شده است و در نهایت نیز مقایسه و اندازه گیری درصد خطا صورت می‌گیرد.

۲.۱. جامعه آماری

جامعه آماری به کار گرفته شده در این پژوهش، دانشجویان ریاضی کاربردی نیم سال اول ۱۳۷۴ تا نیمسال اول ۱۳۹۰ دانشگاه شاهرود می‌باشد که شامل ۱۰۴۴ رکورد است و بهترین حالت در نظر گرفته شده برای پیش‌بینی معدل (ترم ۲) شامل ۵۳۴ رکورد است که ۴۹.۷۷٪ از این دانشجویان شامل فارغ‌التحصیلی، اخراج و انصراف و ۵۰.۲۱٪ مشغول به تحصیل هستند.

۳. فنون داده کاوی

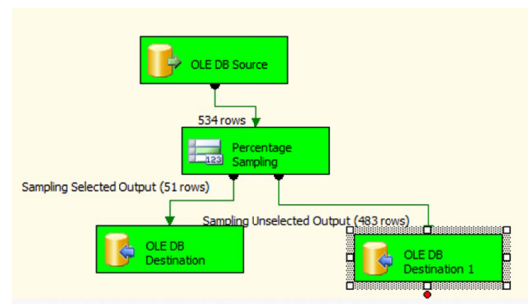
برخی از فنون رایج در داده‌کاوی عبارتند از:

- ابزارهای پرس‌جو
- فنون آماری
- مصورسازی
- پردازش تحلیلی پیوسته
- یادگیری مبتنی بر مورد
- درختان تصمیم گیری
- قوانین وابستگی
- شبکه‌های عصبی
- الگوریتم ژنتیکی

که ابزارهای پرس‌وجو، فنون آماری و درختان تصمیم از جمله فنون به










































































۶. تحلیل داده‌ها و مدل‌سازی

همان‌طور که گفته شد با استفاده از الگوریتم‌ها و ابزارهای داده‌کاوی محیط نرم‌افزار *business intelligence development studio* می‌توانیم در رابطه با حل پروژه‌های گسترده و مختلف راه‌حل مناسب و جامعی را ارائه دهیم. ابتدا با تعریف پروژه جدید از نوع *integration service* داده‌ها را به دو بخش آزمایشی و آموزشی تقسیم می‌کنیم. که پیش‌بینی مدل براساس این داده‌ها صورت می‌گیرد. طبق شکل (۲) با استفاده از سرویس خدمات یکپارچه‌سازی به همراه ابزار نمونه‌گیری درصدی (انتخاب درصد تصادفی نمونه‌ها)، ۹۰٪ داده‌ها برای آموزش (*train*) و ۱۰٪ داده‌ها برای تست (*test*) و آزمودن صحت و دقت مدل در نظر گرفته شده است [۸].



شکل ۱- عملیات تفکیک‌سازی داده‌ها برای پیش‌بینی معدل در فارغ التحصیلی

برای ارتباط با پایگاه داده، با استفاده از پایگاه داده‌ای که اطلاعات کلی در آنجا قرار دارد، یک *Data source* ایجاد می‌کنیم پس برای مشاهده دیدگاه داده کاوی از داده‌ها یک *Data source view* ایجاد می‌کنیم. حال باید از قسمت *Mining structure* یک مدل داده‌کاوی با استفاده از درخت تصمیم ایجاد نماییم [۹ و ۱۰]. در هنگام ایجاد درخت ستون شماره دانشجویی به عنوان کلید، معدل کل به عنوان ستون قابل پیش‌بینی و بقیه پارامترهای عنوان شده در شکل (۲) به عنوان پارامترهای ورودی انتخاب شده‌اند.

Mining model structure:		Tables/Columns		Key	Input	Predict
	 train123					
	entrtm					
	 gfa					
	 id					
	 sex					
	ststatus					
	 syno					
	threeagototalpassunit					
	 threebaseavg					
	 threebasenum					
	 threegeneralavg					
	 threegeneralnum					
	 threebasemavg					
	threebasenum					
	 threeemavg					
	 threeemnum					

Recommend inputs for currently selected predictable:

Submit

همچنین در شکل (۵) وابستگی‌های مدل ایجاد شده و رابطه بین ستون‌ها و پارامترهای تأثیرگذار مشاهده می‌شود.

- معدل سه ترم اول
- تعداد واحد دروس عمومی اخذ شده
- تعداد واحد دروس پایه اخذ شده
- تعداد واحد سه ترم اول

نتایج مقایسه در جدول (۳) بیان شده است، بدین صورت که حدود ۴۹.۲۶٪ از معدل کل فارغ‌التحصیلی دانشجویان با اختلاف کمتر از ۰.۵ نمره پیش‌بینی شده است. درصد تفاضل بین ۰.۵ تا ۱ نمره حدود ۲۴.۲۱٪ است و درصد تفاضل بین ۱ تا ۲ نمره ۲۲.۱۲٪ و در نهایت درصد تفاضل معدل واقعی و معدل پیش‌بینی با اختلاف بیش‌تر از ۲ نمره حدود ۴.۳۸٪ است.

جدول ۳- درصد تفاضل مقادیر واقعی معدل و پیش‌بینی شده

درصد	تعداد دانشجویان	تفاضل معدل واقعی ترم و معدل پیش‌بینی شده
49.26%	236	$X \leq 0.5$
24.21%	116	$0.5 < x < 1$
22.12%	106	$1 \leq x \leq 2$
4.38%	21	$x > 2$

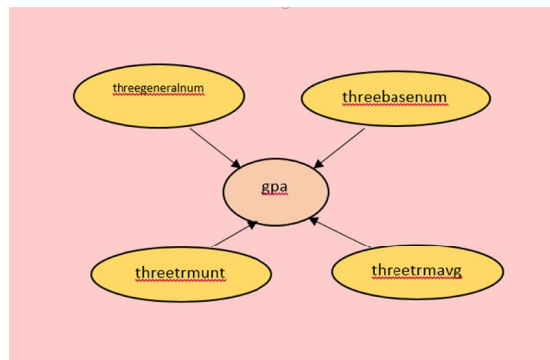
خطای نسبی مدل پیش‌بینی شده با محاسبه قدر مطلق تفاضل بین مقدار واقعی و مقدار پیش‌بینی معدل کل و تقسیم آن بر معدل واقعی، برای هر دانشجو محاسبه شد و با گرفتن میانگین از این مقادیر، در نهایت میانگین درصد خطای نسبی ۶.۵۱٪ بدست آمد.

۹. نتیجه‌گیری

در این پژوهش با یک روند تحلیلی و با استفاده از تکنیک‌های داده‌کاوی، عوامل مؤثر بر بهبود تحصیلی دانشجویان ریاضی کاربردی بررسی شد و با توجه به نتایج بدست آمده، بهترین حالت پیش‌بینی معدل، براساس معدل دو ترم اول است و دریافتیم که تعداد واحد دروس عمومی و پایه اخذ شده در سه ترم اول، تعداد واحد اخذ شده سه ترم اول و معدل سه ترم اول، از عوامل اثرگذار بر بهبود و پیشرفت تحصیلی دانشجویان است. براساس این یافته‌ها دانشجویان و مسئولان مؤسسات آموزش عالی می‌توانند تصمیم‌گیری‌های درست و به موقعی را در راستای بهبود تحصیلی و کاهش هزینه‌ها اتخاذ نمایند. در نهایت این نتایج می‌تواند زمینه‌ای برای شروع تحقیقات گسترده‌تر و بررسی مدل‌های دیگر باشد که با در نظر گرفتن پارامترهای جدید نظیر میزان تحصیلات والدین، بومی یا غیر بومی بودن و... می‌توان به نتایج دقیق و جامع‌تری دست یافت.

مراجع

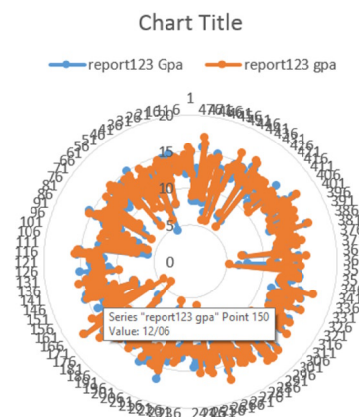
[۱] شکورنیا، ونوس؛ حاجی علی اکبری، آرش؛ "خوشه بندی داده‌های آماری دانشجویان دانشگاه علم و صنعت و استخراج نمایه ساز توصیفی برای دانشجویان موفق"، سومین کنفرانس داده‌کاوی، دانشگاه صنعتی امیرکبیر، ۱۳۸۸.



شکل ۵- نمودار وابستگی‌ها

۸. اعتبارسنجی نتایج

با استفاده از گزارش‌گیری توسط report server project wizard می‌توانیم میزان خطا در پیش‌بینی را بررسی نماییم برای این از جدول train12 که شامل ۴۸۲ دانشجو است برای ارزیابی خطا استفاده نموده‌ایم سپس با توجه به شکل (۶)، ستون معدل کل پیش‌بینی شده در هنگام فارغ-التحصیلی (Gpa)، را با ستون معدل کل واقعی (gpa) موجود در دو جدول مقایسه نموده‌ایم.



شکل ۶- نمودار راداری میزان خطای پیش‌بینی معدل کل

[۲] یقینی، مسعود؛ اکبری، امین؛ شریفی، سید محمد مهدی؛ "پیش‌بینی وضعیت تحصیلی دانشجویان با استفاده از تکنیک‌های داده‌کاوی"، دومین کنفرانس داده‌کاوی ایران، دانشگاه صنعتی امیرکبیر، ۱۳۸۷.

[۳] غضنفری، مهدی؛ علیزاده، سمیه؛ تیمورپور، بابک؛ "داده‌کاوی و کشف دانش"، تهران، دانشگاه علم و صنعت ایران، ۱۳۸۷.

[۴] شهرابی، جمال؛ شکورنیا، ونوس؛ "داده‌کاوی در SQL Server"، انتشارات جهاد دانشگاهی واحد صنعتی امیرکبیر، چاپ اول، ۱۳۸۸.

[۵] فرتوت، رقیه؛ شهرابی، جمال؛ "ارائه مدلی به منظور شناسایی تائمین کنندگان مناسب با استفاده از داده‌کاوی" پنجمین کنفرانس داده‌کاوی، دانشگاه صنعتی امیرکبیر، ۱۳۹۰.

[۹] احمدی، سعید؛ فرهادی، محسن؛ رضایی، سمیه؛ "پیش‌بینی وضعیت نمره دانشجویان در درس معماری کامپیوتر" پنجمین کنفرانس داده-کاوی، دانشگاه صنعتی امیرکبیر، ۱۳۹۰.

[۱۰] شهرابی، جمال؛ شکورنیا، ونوس؛ مفاهیم داده‌کاوی در اوراکل ۱۱g، تهران متالون، تهران، ۱۳۸۷.

[۱۱] یقینی، مسعود؛ حیدری، سمیه؛ "داده‌کاوی جهت ارتقاء وب‌بهدود فرآیندهای سیستم آموزش عالی"، کنفرانس بین‌المللی داده‌کاوی ۲۰۰۸.

[۱۲] فرهادی، محسن؛ تقوی، مطهره السادات؛ نوروزی، مهدیه؛ "پیش‌بینی موفقیت یا عدم موفقیت دانشجویان رشته عمران در فارغ التحصیلی با به کارگیری تکنیک‌های داده‌کاوی"، ششمین کنفرانس داده‌کاوی، تهران، ۱۳۹۱.

[۱۳] امینایی، بهروز؛ میرافضل، سمیه السادات؛ هانی، سید حسن؛ "شناسایی عوامل مؤثر بر افت تحصیلی دانشجویان با استفاده از قوانین انجمنی و تحلیل خوشه‌بندی"، ششمین کنفرانس داده‌کاوی، تهران، ۱۳۹۱.

[۶] Ramanathan naryan, Daniel Honbo, Gokhan Memic, Alok choudhary, Joseph Zambreno "An FPGA Implementation of decision Tree classification" IEEE, 2007

[۷] Jing-Feng Gou, Jingli, Wei-Feng Bian, "An Efficient relational Decision Tree Classification Algorithm", IEEE, 2007.

[۸] Data Mining Tutorials, Microsoft SQL Server 2008 Analysis Services (SSAS), Microsoft Corporation, Inc 2008

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.