
بسم الله الرحمن الرحیم

پردیس بین‌المللی ارس

گروه مهندسی کامپیوتر

عنوان پایان‌نامه

پیش‌بینی فارغ‌التحصیلی دانشجویان در آموزش عالی با استفاده از مدل‌های داده‌کاوی (یک مقایسه)

استاد راهنما

دکتر محمد علی بالافر

پژوهشگر

خاطره امانتی

زمستان ۹۳

سپاس گذاری...

در آغاز از اســتاد گرامیم جناب آقای دکتر محمدعلی بالافر بســیار ســپاســگذارم چرا که بدون راهنماییهای ایشان تامین این پایان نامه بسیار مشکل مینمود.

همچنین از پدر و مادر عزیز ، دلسوز و مهربانم که آرامش روحی و آسایش فکری فراهم نمودند تا با حمایت‌های همه جانبه در محیطی مطلوب، مراتب تحصـیلی و نیز پایان‌نامه درسی را به اتمام برسانم، سپاسگزارم.

خاطره امانتی

۱۳۹۳

نام خانوادگی: امانتی **نام:** خاطره

عنوان پایان‌نامه: پیش‌بینی فارغ‌التحصیلی دانشجویان در آموزش‌عالی با استفاده از مدل‌های داده کاوی: یک مقایسه

استاد راهنما: دکتر محمدعلی بالافر **استاد مشاور:** دکتر لیلی محمدخانلی

مقطع تحصیلی: کارشناسی ارشد **رشته:** مهندسی‌کامپیوتر **گرایش:** نرم‌افزار

دانشگاه: تبریز **دانشکده:** پردیس بین‌المللی ارس

تاریخ فارغ‌التحصیلی: ۱۳۹۳ **تعداد صفحات:**

کلید واژه‌ها: داده‌کاوی آموزشی، پیش‌بینی، درخت‌تصمیم، شبکه‌های عصبی، جنگل‌های تصادفی، رگرسیون لجستیک، فراغت از تحصیل

چکیده:

با توجه به اهمیت بالای آموزش عالی، چالش‌های موجود در این زمینه باعث شده که دست‌اندرکاران آموزش به دنبال روش‌هایی جهت شناسایی زودهنگام دانشجویان در معرض خطر، و بهبود نرخ فارغ‌التحصیلی دانشجویان باشند. دراین پایان‌نامه به منظور بررسی عوامل مؤثر بر فارغ‌التحصیلی دانشجویان از دو مدل داده‌کاوی رگرسیون لجستیکی و درخت تصمیم، استفاده شده سپس با بررسی کارکرد مدلها، در حوزه داده‌های مرتبط با پیش بینی فارغ‌التحصیلی، به اندازه‌گیری خطاهای پیش‌بینی این دو روش پرداخته شده است. با توجه به نتایج به دست آمده و علم به اینکه پارامتر دقت کلی مدل، مهمترین پارامتر برای ارزیابی این روش‌ها است، روش رگرسیون لجستیک دارای بالاترین دقت بود و گزینه مناسب برای پیش‌بینی فارغ‌التحصیلی دانشجویان می‌باشد. مدل رگرسیون لجستیک در رابطه با مجموعه داده‌ی دانشگاه نشان داد که معدل نمرات ترم اول و دیپلم مهم‌ترین متغیرها هستند. بر اساس یافته های این مطالعه، مؤسسات می‌توانند از اطلاعات تحصیلی بخصوص داده‌های نیمسال اول در ایجاد مدل‌های داده‌کاوی استفاده کنند تا متغییرهای معنی داری را برای پیش‌بینی فارغ‌التحصیلی دانشجویان بیابند. نتایج برگرفته از تحلیل‌های داده‌کاوی در توسعه‌ی برنامه‌های مداخله‌ای برای کمک به موفقت دانشجویان در دانشگاه‌ها قابل استفاده هستند.

فهرست مطالب

عنوان | شماره صفحه

فهرست جداول

عنوان | شماره صفحه

فهرست شکل‌ها

عنوان | شماره صفحه

فهرست معادلات

عنوان | شماره صفحه

فصل اول

مقدمه و شرح مسئله

شکست تحصیلی دانشجویان، ضمن اینکه سرمایه‌های مادی و انسانی جامعه را فرسوده می‌کند، مشکلات زیادی را نیز برای خانواده‌ها ایجاد می‌کند. بدون شک مراکز آموزش‌عالی و دانشگاه‌ها که به طور مستقیم با دانشجویان سروکار دارند و نتیجه زحمات خود را در موفقیت دانشجویان جستجو می‌کنند، دغدغه زیادی را در زمینه پدیده فارغ‌التحصیلی دانشجویان دارند.

۱-۱ انگیزه

مسئله موفقیت یا شکست تحصیلی دانشجویان از مسائل مورد توجه مسئولان آموزشی است. فارغ‌التحصیلی دانشجویان در دانشگاه، وابسته به عوامل مختلفی است، که از آن جمله می‌توان به عوامل فردی، اجتماعی، تحصیلی، آموزشی و روان‌شناختی اشاره کرد. بررسی این عوامل و میزان سهم هریک از آنها در پیشرفت تحصیلی یا شکست تحصیلی دانشجو می‌تواند به تعیین رهیافت‌هایی در جهت شناخت عوامل مؤثر در موفقیت و شکست تحصیلی منجر شده و برنامه‌ریزان نظام دانشگاهی را در راستای اهداف دانشگاه که همانا پرورش نیروهای انسانی متخصص است یاری نموده و از کاهش میزان بازدهی تحصیلی جلوگیری و به افزایش کیفی دوره‌های آموزش عالی بیانجامد. دانشگاه‌ها همچنان تمرکز اساسی بر فارغ‌التحصیلی دانشجویان دارند. این تمرکز شامل دانش‌جویانی است که در دانشگاه ثبت‌نام کرده‌اند تا فرصت‌ها و ابزارهای

مطلوب آموزشی منجر به فارغ‌التحصیلی را در اختیار داشته باشند. کیفیت مؤسسه توسط رده‌بندی ملی مؤسسه ارزیابی می‌شود که با توجه به عواملی مانند دانشجویان دارای نمرات برتر، بورسیه‌ها، دانشجویان غیرانصرافی و دانشجویان فارغ‌التحصیل تعیین می‌شود.

کلید درک مؤثر تعادل پیچیده‌ی بین ثبت‌نام و فراغت از تحصیل در کاربرد الگوریتم‌های بهینه‌سازی یا روش‌هایی مانند داده‌کاوی و مدل‌سازی پیش‌بینانه نهفته است. پرسنل پذیرش و مدیریت باید بتوانند در رابطه با دانشجویانی که فارغ‌التحصیل می‌شوند یا ترک تحصیل می‌کنند معیارهای آینده را پیش‌بینی و به دانشجویان انصرافی کمک کنند. با وجود این پیش‌بینی‌های دقیق، توانایی مدیریت دانشگاه در حفظ تعادل مثبت بین رشد، کیفیت، حفظ (دانشجویان) و فراغت از تحصیل افزایش چشمگیری خواهد یافت.

۱-۲ بیان مسئله

بررسی و شناخت استعدادها، خواستها و مشکلات و مسائل دانشجویان در هر جامعه‌ای از اهمیت ویژه‌ای برخوردار است. به همین دلیل مسئله موفقیت یا شکست تحصیلی دانشجویان از مسائل مورد توجه مسئولان آموزشی است و می‌تواند در حل مشکلات یاریگر باشد.

مسئولان آموزشی همیشه به پاسخ به سؤالات خاصی علاقه دارند:

هریک از عوامل فردی (سن، جنسیت، تأهل، ...)، اقتصادی(میزان درآمد خانواده، شاغل بودن، ...)، تحصیلی (سهمیه قبولی، معدل دیپلم، ...)، روانی-اجتماعی (بهداشت روانی، میزان اضطراب، افسردگی، ...) و غیره چه سهمی در پیش‌بینی میزان فارغ‌التحصیلی دانشجویان دارد؟ چرا دانشجویان فارغ‌التحصیل نمی‌شوند؟ چرا به دانشگاه دیگر منتقل می‌شوند؟ چرا بعضی از آنها پیش از سایرین فارغ‌التحصیل می‌شوند؟ چرا دوره‌ی تحصیلی بعضی نسبت به دیگران طولانی‌تر می‌شود؟ کدام دانشجویان در معرض خطر هستند؟ پاسخ این سؤالات به مدیران موسسات کمک می‌کند تا اقدامات مقتضی را برای افزایش نرخ ثبت نام و

فراغت از تحصیل اتخاذ کنند (مانند برنامه‌های مداخله‌ای مؤثر).

۱-۳ اهداف

هدف اصلی این پروژه، مقایسه الگوریتم‌های مختلف داده‌کاوی از طریق مقایسه دقت الگوریتم‌ها، جهت انتخاب دقیقترین مدل برای پیش‌بینی فارغ‌التحصیلان است به عبارت دیگر هدف این پژوهش مقایسه‌ی روش‌های داده‌کاوی در تحلیل آن دسته از متغیرهای دانشجویی است که به فارغ‌التحصیلی دانشجو از دانشگاه منجر می‌شوند. این تحقیق مدل‌های داده‌کاوی پیش‌بینانه‌ی آماری را ایجاد و مقایسه خواهد کرد که مدل‌هایی مانند رگرسیون لجستیک، درخت تصمیم، جنگل‌های تصادفی و شبکه‌های عصبی از آن جمله‌اند. هر یک از این مدل‌ها متناسب با داده‌های حفظ دانشجو بهینه‌سازی و ارزیابی می‌شوند تا بهترین مدل داده‌کاوی تعیین شود. این تحقیق خصوصیات مهم دانشجویانی که فارغ‌التحصیل می‌شوند را در مقابل خصوصیات دانشجویانی که فارغ‌التحصیل نمی‌شوند بررسی خواهد کرد. نهایتاً این مطالعه به غنای اندک تحقیقات انجام شده پیرامون اثرگذاری تکنیک‌های داده‌کاوی استفاده شده در آموزش عالی کمک خواهد کرد. همچنین به مؤسسات آموزشی کمک خواهد کرد تا در تامین اطلاعات برای راهبردهای فراغت از تحصیل دانشجو استفاده‌ی مؤثرتری از تکنیک‌های داده‌کاوی داشته باشند. از دیدگاه مؤسسات (دانشگاهی)، افزایش حفظ دانشجویِ منجر به فارغ‌التحصیلی مدیریت ثبت‌نام را بهبود می‌بخشد، هزینه‌های استخدام را کاهش می‌دهد و همچنین باعث ترفیع جایگاه دانشگاه می‌شود. از دیدگاه دانشجویان نیز حفظ دانشجویِ منجر به فارغ‌التحصیلی دارای آثار اجتماعی، فردی و اقتصادی است.

۱-۴ تعریف اصطلاحات

پیش‌بینی: فرآیند برآورد موقعیت‌های ناشناخته است. یک پیش‌بینی یک پیش‌گویی در مورد رویدادهای

آینده در اختیار می‌گذارد و می‌تواند تجارب گذشته را به پیش‌بینی حوادث آینده بدل سازد.

صفت: صفت در این پژوهش به متغیری واحد اشاره دارد، از جمله نژاد یا جنسیت. صفات برای ایجاد مدل‌های آماری استفاده می‌شوند. متغیر، واژه‌ای معادل صفت است.

متغیر: متغیر به عنوان خصوصیت یا صفت دانشجو تعریف می‌شود. مانند جنسیت، سن، و GPA[1] (معدل نمرات پایه).

داده: تعریف لغتنامه‌ی آکسفورد از داده به عنوان «حقایق و آمار جمع‌آوری شده جهت ارجاع و تفسیر» مورد پذیرش این پژوهش خواهد بود.

داده‌کاوی: فراولی[2] و همکاران (۱۹۹۱) [3] داده‌کاوی را به عنوان استخراج غیر بدیهی اطلاعات ضمنی، از پیش نامعلوم و بالقوه سودمند از داده تعریف کرده‌اند.

مدل‌سازی: در این مطالعه مدل‌سازی به عمل ایجاد معادلاتی اشاره دارد که از داده‌های مشاهده شده برای پیش‌بینی نمونه‌های آتی توسط داده‌های مشاهده نشده‌ی آتی استفاده می‌کنند.

موفقیت دانشجو: موفقیت دانشجو بر مبنای فارغ‌التحصیل شدن دانشجو تعریف می‌شود. دانشجوی موفق تدریجاً در درجه‌ی تحصیلی خود پیشرفت می‌کند و نهایتاً ظرف مدت ۴ سال پس از ثبت نام فارغ‌التحصیل می‌شود.

نرخ طبقه‌بندی نادرست: نرخ طبقه‌بندی اشتباه نشانگر خطا در پیش‌بینی تعداد واقعی فارغ‌التحصیلان است.

[1] - Grade Point Average
[2] - Frawley

۱-۵ چهار چوب پایان‌نامه

در فصل بعدی چهار الگوریتم داده‌کاوی و روش‌های ارزیابی و مقایسه استفاده شده بر روی این الگوریتم‌ها به صورت مفصل شرح داده می‌شود، در ادامه فصل اصطلاح داده‌کاوی تعریف و تشریح شده و در انتهای فصل کارهای مرتبط در این زمینه با عنوان کارهای پیشین مورد بررسی قرار می‌گیرد. در فصل ۳ به راهکار پیشنهادی برای پیش‌بینی فارغ‌التحصیلی دانشجویان خواهیم پرداخت و مدل‌های ارائه شده برای پیش‌بینی فارغ‌التحصیلی دانشجویان پیاده‌سازی می‌گردد و راهکار پیشنهادی را روی یک نمونه واقعی بررسی می‌کنیم. در فصل ۴ به صورت کلی مقایسه جامعی بین مدل‌های ارائه شده انجام می‌دهیم تا کارایی و دقت هر کدام از این مدل‌ها برای پیش‌بینی فارغ‌التحصیلی دانشجویان مشخص شود. در نتیجه بعد از مقایسه روش‌های ارائه شده بهترین و دقیقترین مدل برای پیش‌بینی معرفی می‌گردد.

فصل دوم

مفاهیم پایه‌ای

۲-۱ داده‌کاوی

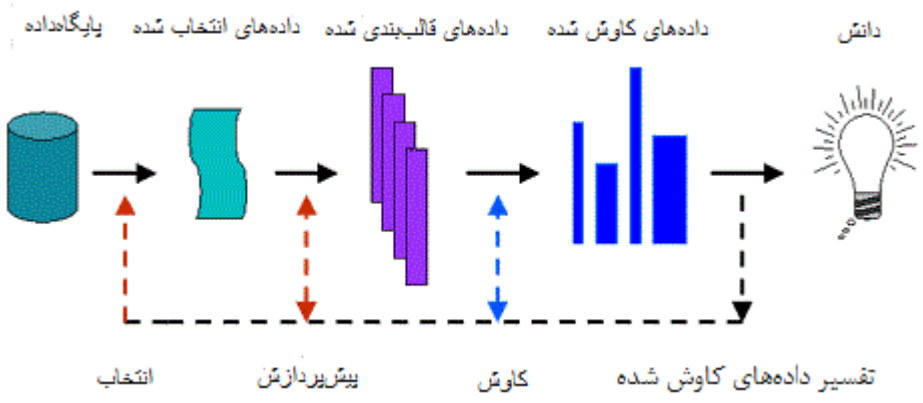
ظهور فنـاوری اطلاعـات در زمینـه‌هـای مختلف باعث ذخیره‌ســـازی حجم زیادی از داده‌ها در قالب‌های مختلفی مانند: سـوابق، فایل‌ها، اسناد، تصاویر، صدا، ویدئو و بسیاری دیگر از قالب‌های داده‌ای جدید شده اسـت. داده‌های جمع‌آوری شـده از برنامه‌های کاربردی مختلف نیازمند روش‌هایی مناسب برای استخراج دانش جهت تصـمیم‌گیری بهتر دارد. کشـف دانش در پایگاه داده که اغلب داده‌کاوی نامیده می‌شـود، با هدف کشف اطلاعات مفید از مجموعه بزرگ داده‌ها انجام می‌شود.

داده‌کاوی رشـته نسـبتا جدیدی در آمار محسوب می‌شود. یکی از تعاریف اولیه داده‌کاوی از جانب جان[1] (۱۹۹۷) ارائه شده است که داده‌کاوی را نام جدیدی برای یک فرآیند کشف الگوهای مفید از داده‌ها بیان کرده اسـت. اما تعریفی که در اکثر مراجع به اشـتراک ذکر شـده عبارت است از استخراج اطلاعات و دانش و کشـــف الگوهای پنهان از پایگاه داده‌های بســـیار بزرگ و پیچیده [4]. تعاریف اولیه داده‌کاوی محدود به فرایند مدلسـازی بودند، اما بعدها به ارزیابی مدل بسط یافتند. یافتن الگوهای موجود در داده‌ها فرایند پیچیده‌ایی اسـت که می‌تواند توسـط الگوریتم‌های آماری انجام گیرد، همچنین می‌توان روابط بین متغیرها را با این الگوریتم‌ها ارزیابی کرد. تکنیک‌های داده کاوی اغلب در رشـــته های گوناگون با نام‌های مختلف استفاده می‌شوند.

[1] John

۲-۱-۱ مراحل داده‌کاوی

همان گونه که در شکل ۲-۱ نشان داده شده است، فرآیند داده‌کاوی عمومی شامل چهار مرحله است (مندونکا[1]، ۱۹۹۹) [5].



شکل ۲-۱: فرآیند داده‌کاوی (مندونکا، ۱۹۹۹)

اولین مرحله گردآوری یا انتخاب پایگاه‌داده‌ای است که توسط الگوریتم داده‌کاوی مورد استفاده قرار خواهد گرفت. مجموعه داده‌های خام معمولاً شامل داده‌های گوناگون است که همه آنها برای نایل شدن بـه اهداف داده‌کاوی، ضروری نیست. برای مثـال فرض کنیـد پایگاه‌داده‌ای از داده‌های دانشجویان یک موسسه را در اختیار داریم که اطلاعات ثبت‌نام، نمرات و جمعیتی دانشجویان در آن موجود باشد و می‌خواهیم با استفاده از این اطلاعات موفقیت یا شکست تحصیلی دانشجویان را پیش‌بینی کنیم. حال فرض کنید در کنار این اطلاعات شماره شناسنامه دانشجو نیز آورده شود. مسلماً شماره شناسنامه دانشجو برای اهداف داده‌کاوی ما ضروری و یا اصلاً مناسب نمی‌باشد. بنابراین تحلیل‌گر داده، مکان داده‌های مورد نظر را شناسایی و انتخاب می‌کند و آن‌ها را از مکان اصلی به مخزن داده‌ها منتقل می‌کند که با آن کار خواهد نمود. این

[1] Manoel Mendonca