

صلى الله عليه وسلم



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 استاد گرامیم جناب آقای دکتر محمدعلی بالافر بسیار سپاسگذارم 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 تحصیلی و نیز پایان نامه درسی را به اتمام برسانم، سپاسگذارم.

خاطره امانتی

۱۳۹۳

نام خانوادگی: امانتی	نام: خاطره
<p>عنوان پایان نامه: پیش بینی فارغ التحصیلی دانشجویان در آموزش عالی با استفاده از مدل های داده کاوی: یک مقایسه</p>	
استاد راهنما: دکتر محمدعلی بالافر	استاد مشاور: دکتر لیلی محمدخانلی
<p>مقطع تحصیلی: کارشناسی ارشد رشته: مهندسی کامپیوتر گرایش: نرم افزار</p> <p>دانشگاه: تبریز دانشکده: پردیس بین المللی ارس</p> <p>تاریخ فارغ التحصیلی: ۱۳۹۳ تعداد صفحات:</p>	
<p>کلید واژه ها: داده کاوی آموزشی، پیش بینی، درخت تصمیم، شبکه های عصبی، جنگل های تصادفی، رگرسیون لجستیک، فراغت از تحصیل</p>	
<p>چکیده:</p> <p>با توجه به اهمیت بالای آموزش عالی، چالش های موجود در این زمینه باعث شده که دست اندرکاران آموزش به دنبال روش هایی جهت شناسایی زود هنگام دانشجویان در معرض خطر، و بهبود نرخ فارغ التحصیلی دانشجویان باشند. در این پایان نامه به منظور بررسی عوامل مؤثر بر فارغ التحصیلی دانشجویان از دو مدل داده کاوی رگرسیون لجستیک و درخت تصمیم، استفاده شده سپس با بررسی کارکرد مدلها، در حوزه داده های مرتبط با پیش بینی فارغ التحصیلی، به اندازه گیری خطاهای پیش بینی 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 روش ها است، روش رگرسیون لجستیک دارای بالاترین دقت بود و گزینه مناسب برای پیش بینی فارغ التحصیلی دانشجویان می باشد. مدل رگرسیون لجستیک در رابطه با مجموعه داده ی دانشگاه نشان داد که معدل نمرات ترم اول و دیپلم مهم ترین متغیرها هستند. بر اساس یافته های این مطالعه، مؤسسات می توانند از اطلاعات تحصیلی بخصوص داده های نیمسال اول در ایجاد مدل های داده کاوی استفاده کنند تا متغیرهای معنی داری را برای پیش بینی فارغ التحصیلی دانشجویان بیابند. نتایج بر گرفته از تحلیل های داده کاوی در توسعه ی برنامه های مداخله ای برای کمک به موفقیت دانشجویان در دانشگاه ها قابل استفاده هستند.</p>	

فهرست مطالب

عنوان	شماره صفحه
۱ فصل اول: مقدمه و شرح مسئله.....	۱
۱-۱ انگیزه.....	۲
۲-۱ بیان مسئله.....	۳
۳-۱ اهداف.....	۴
۴-۱ تعریف اصطلاحات.....	۴
۵-۱ چهارچوب پایان نامه.....	۶
۲ فصل دوم: مفاهیم پایه‌ای.....	۷
۱-۲ داده کاوی.....	۸
۱-۱-۲ مراحل داده کاوی.....	۹
۲-۲ یادگیری ماشین.....	۱۱
۳-۲ الگوریتم‌های طبقه‌بندی.....	۱۲
۱-۳-۲ درخت تصمیم.....	۱۲
۲-۳-۲ جنگل‌های تصادفی.....	۱۳
۳-۳-۲ رگرسیون منطقی.....	۱۶
۴-۳-۲ شبکه‌های عصبی.....	۱۷
۱-۴-۳-۲ شبکه‌های پرسپترون چند لایه.....	۱۸
۴-۲ تکنیک‌های ارزیابی مدل.....	۱۹

- ۲-۴-۱ ماتریس خطا برای طبقه‌بندی دو کلاسه..... ۲۰
- ۲-۴-۱-۱ اصطلاحات و مشتقات ماتریس خطا..... ۲۱
- ۲-۴-۲ منحنی ROC..... ۲۲
- ۵-۲ داده کاوی آموزشی..... ۲۳
- ۳ فصل سوم: بررسی منابع..... ۲۵
- ۴ فصل چهارم: راه کار پیشنهادی ۳۲
- ۱-۴ داده‌های ورودی..... ۳۴
- ۲-۴ مدلسازی پیش‌گویانه..... ۳۵
- ۴-۲-۱ پیش‌بینی با استفاده از درخت تصمیم بدون استفاده از نمونه‌برداری..... ۳۶
- ۴-۲-۲ پیش‌بینی با استفاده از درخت تصمیم با استفاده از نمونه‌برداری..... ۳۷
- ۴-۲-۳ پیش‌بینی با استفاده از رگرسیون لجستیک بدون استفاده از نمونه‌برداری..... ۳۸
- ۴-۲-۴ پیش‌بینی با استفاده از رگرسیون لجستیک با استفاده از نمونه‌برداری..... ۳۸
- ۴-۲-۵ پیش‌بینی با استفاده از شبکه عصبی بدون استفاده از نمونه‌برداری..... ۳۹
- ۴-۲-۶ پیش‌بینی با استفاده از شبکه عصبی با استفاده از نمونه‌برداری..... ۴۱
- ۴-۲-۷ پیش‌بینی با استفاده از جنگل‌های تصادفی بدون استفاده از نمونه‌برداری..... ۴۲
- ۴-۲-۸ پیش‌بینی با استفاده از جنگل‌های تصادفی با استفاده از نمونه‌برداری..... ۴۳
- ۳-۴ نتیجه‌گیری..... ۴۴
- ۵ فصل پنجم: ارزیابی عملی و نتیجه‌گیری..... ۴۵
- ۵-۱ مقایسه روش درخت تصمیم..... ۴۶

۴۸.....	۲-۵ مقایسه روش رگرسیون لجستیک.....
۴۹.....	۳-۵ مقایسه روش شبکه عصبی.....
۵۱.....	۴-۵ مقایسه روش جنگل‌های تصادفی.....
۵۲.....	۵-۵ مقایسه کلی روش‌ها بدون استفاده از نمونه‌برداری.....
۵۵.....	۶-۵ مقایسه کلی روش‌ها با استفاده از نمونه‌برداری.....
۵۸.....	۷-۵ نتیجه‌گیری.....
۵۹.....	مراجع.....

فهرست جداول

عنوان	شماره صفحه
جدول ۱-۲ ماتریس خطای طبقه بندی کننده دودویی (دو کلاسه)	۲۱
جدول ۱-۴ متغیرهای مربوط به اطلاعات دانشجویان	۳۵
جدول ۱-۵ مقایسه روش درخت تصمیم برای معیار دقت	۴۷
جدول ۲-۵ مقایسه روش درخت تصمیم برای معیار جامعیت	۴۷
جدول ۳-۵ مقایسه روش رگرسیون لجستیک برای معیار دقت	۴۸
جدول ۴-۵ مقایسه روش رگرسیون لجستیک برای معیار جامعیت	۴۹
جدول ۵-۵ مقایسه روش شبکه عصبی برای معیار دقت	۵۰
جدول ۶-۵ مقایسه روش شبکه عصبی برای معیار جامعیت	۵۱
جدول ۷-۵ مقایسه روش جنگل‌های تصادفی برای معیار دقت	۵۱
جدول ۸-۵ مقایسه روش جنگل‌های تصادفی برای معیار جامعیت	۵۲
جدول ۹-۵ مقایسه کلی روش‌ها بدون استفاده از نمونه‌برداری برای معیار دقت	۵۳
جدول ۱۰-۵ مقایسه کلی روش‌ها بدون استفاده از نمونه‌برداری برای معیار جامعیت	۵۴
جدول ۱۱-۵ مقایسه کلی روش‌ها با استفاده از نمونه‌برداری برای معیار دقت	۵۵
جدول ۱۲-۵ مقایسه کلی روش‌ها با استفاده از نمونه‌برداری برای معیار جامعیت	۵۶

فهرست شکل‌ها

عنوان	شماره صفحه
شکل ۱-۲ فرآیند داده‌کاوی (مندونکا، 1999).....	۹
شکل ۲-۲ مدل نرون (دموث و بیل، 2000).....	۱۸
شکل ۳-۲ شبکه پیشخور دولایه (دموث و بیل، 2000).....	۱۸
شکل ۴-۲ شبکه پرسپترون چندلایه (حسینی، ۱۳۸۸).....	۱۹
شکل ۵-۲ نمونه منحنی ROC.....	۲۲
شکل ۶-۲ مثال‌هایی از نمونه منحنی ROC.....	۲۳
شکل ۱-۴ قسمتی از درخت تصمیم مربوط به اجرای برنامه.....	۳۶
شکل ۲-۴ بردار کارایی مربوط به درخت تصمیم بدون استفاده از نمونه‌برداری.....	۳۷
شکل ۳-۴ بردار کارایی مربوط به درخت تصمیم با استفاده از نمونه‌برداری.....	۳۷
شکل ۴-۴ بردار کارایی مربوط به رگرسیون لجستیک بدون استفاده از نمونه‌برداری.....	۳۸
شکل ۵-۴ بردار کارایی مربوط به رگرسیون لجستیک با استفاده از نمونه‌برداری.....	۳۹
شکل ۶-۴ قسمتی از مدل شبکه عصبی.....	۴۰
شکل ۷-۴ بردار کارایی مربوط به شبکه عصبی بدون استفاده از نمونه‌برداری.....	۴۰
شکل ۸-۴ مدل تولید شده توسط شبکه عصبی.....	۴۱
شکل ۹-۴ بردار کارایی مربوط به شبکه عصبی با استفاده از نمونه‌برداری.....	۴۱
شکل ۱۰-۴ قسمتی از مدل جنگل‌های تصادفی.....	۴۲
شکل ۱۱-۴ بردار کارایی مربوط به جنگل‌های تصادفی بدون استفاده از نمونه‌برداری.....	۴۲
شکل ۱۲-۴ بردار کارایی مربوط به جنگل‌های تصادفی با استفاده از نمونه‌برداری.....	۴۳
شکل ۱-۵ نمودار مقایسه روش درخت تصمیم برای معیار دقت.....	۴۷

- شکل ۲-۵ نمودار مقایسه روش درخت تصمیم برای معیار جامعیت ۴۸
- شکل ۳-۵ نمودار مقایسه روش رگرسیون لجستیک برای معیار دقت ۴۸
- شکل ۴-۵ نمودار مقایسه روش رگرسیون لجستیک برای معیار جامعیت ۴۹
- شکل ۵-۵ نمودار مقایسه روش شبکه عصبی برای معیار دقت ۵۰
- شکل ۶-۵ نمودار مقایسه روش شبکه عصبی برای معیار جامعیت ۵۱
- شکل ۷-۵ نمودار مقایسه روش جنگل‌های تصادفی برای معیار دقت ۵۱
- شکل ۸-۵ نمودار مقایسه روش جنگل‌های تصادفی برای معیار جامعیت ۵۲
- شکل ۹-۵ نمودار مقایسه کلی روش‌ها بدون استفاده از نمونه‌برداری برای معیار دقت ۵۳
- شکل ۱۰-۵ نمودار مقایسه کلی روش‌ها بدون استفاده از نمونه‌برداری برای معیار جامعیت ... ۵۴
- شکل ۱۱-۵ نمودار ROC مربوط به مقایسه بدون استفاده از نمونه‌برداری ۵۵
- شکل ۱۲-۵ نمودار مقایسه کلی روش‌ها با استفاده از نمونه‌برداری برای معیار دقت ۵۶
- شکل ۱۳-۵ نمودار مقایسه کلی روش‌ها با استفاده از نمونه‌برداری برای معیار جامعیت ۵۷
- شکل ۱۴-۵ نمودار ROC مربوط به مقایسه روش‌ها با استفاده از نمونه‌برداری ۵۷

فهرست معادلات

عنوان	شماره صفحه
معادله ۱-۲ روش خودراه انداز	۱۴
معادله ۲-۲ مدل رگرسیون منطقی	۱۶
معادله ۳-۲ احتمال رخداد در مدل رگرسیون منطقی	۱۶
معادله ۴-۲ محاسبه دقت در ماتریس خطا	۲۱
معادله ۵-۲ محاسبه معیار جامعیت ماتریس خطا	۲۲

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 آموزش عالی 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 هر یک از آنها در پیشرفت تحصیلی یا شکست تحصیلی دانشجویان 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.

۲-۱ 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 (سهامیه 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).

۳-۱ 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 دانشجوی 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 دانشجوی منجر به فارغ‌التحصیلی دارای آثار اجتماعی، فردی و اقتصادی است.

۴-۱ 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 GPA¹ (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: فراولی² 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 دانشجوی: موفقیت دانشجوی 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

۵-۱ چهار چوب پایان نامه

در فصل بعدی چهار الگوریتم داده کاوی و روش های ارزیابی و مقایسه استفاده شده بر روی این الگوریتم ها به صورت مفصل شرح داده می شود، در ادامه فصل اصطلاح داده کاوی تعریف و تشریح شده و در انتهای فصل کارهای مرتبط در این زمینه با عنوان کارهای پیشین مورد بررسی قرار می گیرد. در فصل ۳ به راهکار پیشنهادی برای پیش بینی فارغ التحصیلی دانشجویان خواهیم پرداخت و مدل های ارائه شده برای پیش بینی فارغ التحصیلی دانشجویان پیاده سازی می گردد و راهکار پیشنهادی را روی یک نمونه واقعی بررسی می کنیم. در فصل ۴ به صورت کلی مقایسه جامعی بین مدل های ارائه شده انجام می دهیم تا کارایی و دقت هر کدام از این مدل ها برای پیش بینی فارغ التحصیلی دانشجویان مشخص شود. در نتیجه بعد از مقایسه روش های ارائه شده بهترین و دقیقترین مدل برای پیش بینی معرفی می گردد.

فصل دوم

مفاهیم پایه‌ای

۱-۲ داده‌کاوی

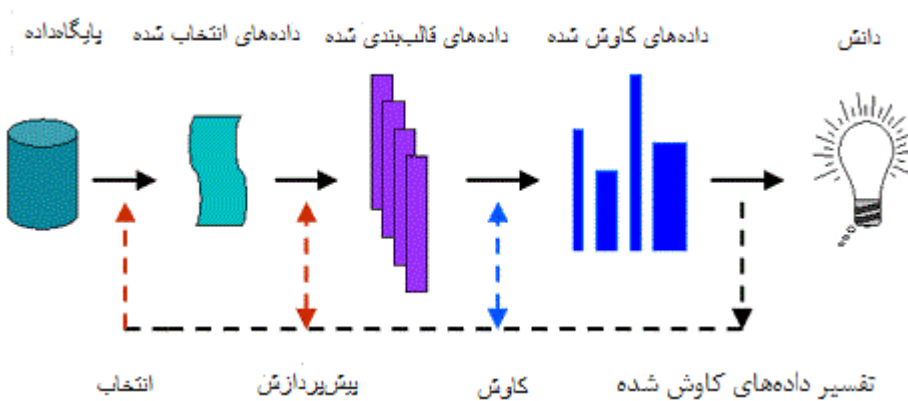
ظهور فناوری اطلاعات در زمینه‌های مختلف باعث ذخیره‌سازی حجم زیادی از داده‌ها در قالب‌های مختلفی مانند: سوابق، فایل‌ها، اسناد، تصاویر، صدا، ویدئو و بسیاری دیگر از قالب‌های داده‌ای جدید شده است. داده‌های جمع‌آوری شده از برنامه‌های کاربردی مختلف نیازمند روش‌هایی مناسب برای استخراج دانش جهت تصمیم‌گیری بهتر دارد. کشف دانش در پایگاه داده که اغلب داده‌کاوی نامیده می‌شود، با هدف کشف اطلاعات مفید از مجموعه بزرگ داده‌ها انجام می‌شود.

داده‌کاوی رشته نسبتاً جدیدی در آمار محسوب می‌شود. یکی از تعاریف اولیه داده‌کاوی از جانب جان^۱ (۱۹۹۷) ارائه شده است که داده‌کاوی را نام جدیدی برای یک فرآیند کشف الگوهای مفید از داده‌ها بیان کرده است. اما تعریفی که در اکثر مراجع به اشتراک ذکر شده عبارت است از استخراج اطلاعات و دانش و کشف الگوهای پنهان از پایگاه داده‌های بسیار بزرگ و پیچیده [4]. تعاریف اولیه داده‌کاوی محدود به فرایند مدلسازی بودند، اما بعدها به ارزیابی مدل بسط یافتند. یافتن الگوهای موجود در داده‌ها فرایند پیچیده‌ایی است که می‌تواند توسط الگوریتم‌های آماری انجام گیرد، همچنین می‌توان روابط بین متغیرها را با این الگوریتم‌ها ارزیابی کرد. تکنیک‌های داده‌کاوی اغلب در رشته‌های گوناگون با نام‌های مختلف استفاده می‌شوند.

¹ John

۱-۱-۲ مراحل داده‌کاوی

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



شکل ۱-۲: 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 به اهداف داده‌کاوی، ضروری نیست. برای مثال فرض کنید 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