# مقدمه

از بدو تولد انسان همواره مشتاق دانستن اتفاقات و رویدادهای آینده بوده است؛ از این‌رو انسانها همواره به‌دنبال راههای جدیدی بوده‌اند تا بتوانند و قایع آینده را پیش‌بینی کنند و آسایش و راحتی را برای زندگی خویش فراهم نمایند. 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. در سال‌های اخیر پیش‌بینی، نقش مؤثری در سیاستهای دولت دارد. چرا که دولت سیاستهای خود را نه صرفاً بر مبنای وضع موجود، بلکه در مبنای پیش‌بینی‌های کوتاه و بلند مدت از متغیرهای کلیدی اقتصادی تدوین نموده و به مورد اجرا می‌گذارد. بدیهی است که میزان صحت پیش‌بینی از جمله رموز موفقیت این سیاست‌ها می‌تواند به‌شمار آید. چنین اهمیتی موجب شده تا تحقیقات در زمینه مدل‌ها و تکنیک‌های پیش‌بینی در چند دهه اخیر با شتاب بیشتری رشد نماید. یکی از کاربردهای پیش‌بینی، پیش‌بینی در مسائل اقتصادی است که عموماً به سه صورت کوتاه مدت، میان مدت، بلند مدت انجام می‌شود[۵۷].

پیش‌بینی کوتاه مدت ( مثلاً یک روزه ) میان مدت ( هفته‌ای ) بلند مدت ( ماهیانه ) است. ساختار کلی مدل پیش‌بینی در تمام این نوع پیش‌بینی‌ها یکسان بوده و فقط از داده‌ها و متغیرهای ورودی متفاوت استفاده می‌شود. نکته قابل توجه دیگر در پیش‌بینی، «دقت پیش‌بینی» است. در سال‌های اخیر مطالعات متعددی در زمینه چگونگی ارزیابی دقت پیش‌بینی صورت گرفته‌است. خصوصیات محیطی، از قبیل متغیر مورد پیش‌بینی، افق پیش‌بینی، ایدئولوژی پیش‌بینی کننده و فناوری مورد استفاده، عوامل مؤثر در دقت پیش‌بینی هستند.

غالباً شاخص‌هایی برای ارزیابی صحت پیش‌بینی استفاده می‌شوند. که خطای پیش‌بینی، ( اختلاف مقدار واقعی و مقدار پیش‌بینی شده ) یکی از آنها می‌باشد. معیارهای خطای پیش‌بینی هر چه کمتر باشند نمایانگر پیش‌بینی‌های دقیق‌تر هستند. تحقیقات متعدد نشان داده است که دقت پیش‌بینی‌های کوتاه مدت بیشتر از دقت پیش‌بینی‌های بلند مدت است.

تحلیل رگرسیونی فن و تکنیک آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است. رگرسیون تقریباً در هر زمینه‌ای از جمله علوم: مهندسی، فیزیک، اقتصاد، مدیریت، زیست‌شناسی، بیولوژی، اجتماعی و ..... برای برآورد و پیش‌بینی مورد نیاز است. تحلیل رگرسیونی، یکی از کاربردی‌ترین روش‌ها در بین تکنیک‌های آماری دیگر است. اما روش‌های کلاسیک آماری برای مدل‌بندی روابط بین متغیرها دارای تعدادی پیش فرض و محدودیت است. در نظرگرفتن یک توزیع پیش فرض مانند توزیع نرمال برای متغیرهای پاسخ، خطی بودن رابطه‌ی پیشنهادی، یکسان بودن واریانس خطاها و غیره از جمله محدودیت‌های روش‌های کلاسیک هستند که هنگام استفاده‌ی عملی از این روش‌ها، اگر داده‌های واقعی، شرایط مفروض مدل را نداشته باشند استفاده از این روش‌ها امکان پذیر نبوده یا با خطای قابل توجه همراه است. به‌علاوه، هیچ یک از این روش‌ها قابلیت مدل‌بندی روابط پیچیده‌ی غیرخطی و اثر متقابل درجه‌ی بالا را ندارند. حساس بودن بیشتر این مدل‌ها به مشاهده‌های گم شده و داده‌های پرت از دیگر محدودیت‌های این روش‌ها به‌شمار می‌آید. بنابراین، روش‌هایی که با محدودیت‌های کمتری دراین زمینه مواجه باشند، احساس می‌شود. در این بین الگوریتم‌های جدیدی مبتنی بر هوش مصنوعی[1] توانسته تا حدودی نتایج بهتر، مطمئن‌تر و کاراتری را برای پیش‌بینی در اختیار قرار دهد. پیدایش هوش مصنوعی و به تبع آن کشف الگوریتم‌ها و روش‌های جدید در این رشته تأثیر عمده‌ای بر روی بسیاری از علوم دیگر گذاشته و آن‌چنان‌که شاهد هستیم زمینه تعالی و پیشرفت در سایر علوم را فرآهم آورده است. مولفه‌های مهم و اساسی هوش مصنوعی شامل [۱۷] ( محاسبات نرونی )، منطق فازی[۵۲] ( محاسبات تقریبی ) و الگوریتم ژنتیک ( محاسبات ژنتیکی ) بوده و هر کدام به نوعی مغز بشر را الگو قرار داده‌اند. شبکه‌ی عصبی هیچ فرض اولیه‌ای بر توزیع داده‌ها تحمیل نمی‌کند، ضمن این‌که هیچ محدودیتی نیز برای شکل تابعی رابطه‌ی بین متغیرهای مستقل و وابسته اعمال نمی‌کند بلکه

[1] Artificial Neural Network

شبکه‌ی عصبی خود این رابطه‌ی تابعی را کشف می‌کند، که لزوماً این رابطه، یک رابطه‌ی خطی نیست. از دیگرمزایای شبکه‌ی عصبی مصنوعی این است که در آن اطلاعات به صورت ضمنی پردازش می‌شوند. بر این اساس، چنانچه بخشی از سلول‌های شبکه‌ی حذف شوند یا عملکرد غلط داشته باشند باز هم احتمال رسیدن به پاسخ صحیح وجود دارد، ضمن اینکه تعمیم پذیری شبکه‌ی عصبی این امکان را می‌دهد که مدل در ارتباط با یک مشاهده‌ی جدید پاسخ مناسبی ارایه دهد[۴].

مهم‌ترین فاکتور در استفاده از شبکه‌های عصبی مصنوعی، انتخاب ورودی‌ها (متغیرهای پیش‌بینی) مناسب است. ورودی‌هایی به‌عنوان متغیر مناسب هستند که بتوانند فضای ورودی شبکه را خوب پوشش دهند و در حد امکان اطلاعات اضافی آنها حذف شده باشد[۳۵].

وجود متغیرهای زیادی ( اضافی) موجب می‌شود که در فضای ورودی اطلاعات اضافی زیاد وجود داشته باشد که فرآیند یادگیری را تحت تأثیر قرار داده و موجب شود که شبکه نتواند آن‌طوری‌که باید و شاید کار پیش‌بینی را درست انجام دهد[۳۷].

برای اینکه شبکه‌های عصبی بتوانند به‌خوبی کار پیش‌بینی را انجام دهند می‌بایستی در حد امکان ورودی‌هایی را داشته باشند که دارای بیشترین وابستگی به متغیرهای پیش‌بینی بوده و دارای حداقل اطلاعات اضافی باشد [۳۹].

اعمال ورودی‌های مناسب به شبکه در مسئله با ابعاد بالا علاوه بر کاهش پیچیدگی‌های مسئله، زمان اجرا را نیز کاهش داده که منجر به بهبود نتایج می‌گردد. تحقیقات فراوان در زمینه کاربردهای متعدد شبکه‌های عصبی در حل مسائل تجاری، مالی و ..... کارایی و مزایای آنها را در مقایسه با سایر روش‌های کلاسیک از قبیل رگرسیون خطی [۵۳،۵۴،۵۵] و .... نشان داده‌اند.

در دانشگاه‌ها و یا مؤسسات آموزش عالی که بزرگ‌ترین منبع اصلی بودجه و تأمین مسائل مالی و هزینه‌ای برای آنها، پرداخت شهریه دانشجویان در هر ترم می‌باشد. بررسی و پیش‌بینی میزان افت

تحصیلی و یا ریزش دانشجویان یکی از ارکان مهم و اساسی آموزشی است که به علت خسارتهای مادی وارده بر سیستم به ازاء اخراج و یا عدم ادامه تحصیل و یا به عبارت دیگر ریزش دانشجو از اهمیت بالایی برخوردار و به یک مسالهٔ حیاتی و موجب نگرانی برای آنان تبدیل شده است که علاوه بر افزایش ضررهای مالی منجر به کاهش درجه‌ی معروفیت دانشگاه خواهد شد. در این تحقیق، با استفاده از چندین تکنیک داده‌کاوی، مدل‌های تحلیلی برای پیش‌بینی و توضیح دلایل زمینه‌ساز ریزش دانشجویان جدیدالورود را ارائه دادیم.

در این پایان‌نامه با استفاده از داده‌های مربوط به دانشجویان مقطع کارشناسی ورودی سال ۱۳۸۸ مرکز آموزش الکترونیکی دانشگاه علم و صنعت ایران (جمع آوری از پایگاه داده موجود دانشگاه (سایت گلستان) و سامانه الکترونیک ( LMS ) مرکز آموزش الکترونیکی دانشگاه علم و صنعت ایران ) به‌همراه روش‌های استخراج داده ( به‌صورت انفرادی و جمعی )، مدل‌های تحلیلی را گسترش داده‌ایم تا به پیش‌بینی میزان ریزش دانشجویان بپردازیم.

# فـصل ۱ :

# مروری بر منابع

## ۱-۱- مقدمه

حفظ دانشجویان یکی از قسمت‌های ضروری بسیاری از سیستم‌های مدیریت ثبت‌نام می‌باشد که بر رتبه‌بندی دانشگاه، اعتبار و وضعیت خوب مالی و درآمد آن تأثیر می‌گذارد. حفظ و ابقاء دانشجو به‌عنوان یکی از مهم‌ترین اولویت‌ها برای تصمیم گیرندگان مراکز تحصیلات عالی درآمده است. بهبود حفظ و ابقاء دانشجو با درکی عمیق از دلایل زمینه‌ساز ریزش آغاز می‌شود. چنین درکی پایه و اساس پیش‌بینی دقیق دانشجویان در معرض خطر ریزش بوده و دخالت مناسب مدیریت موجب حفظ آنان می‌شود. در این تحقیق، با استفاده از پنج سال داده‌های مؤسسه‌ای به‌همراه چندین تکنیک داده‌کاوی (انفرادی و جمعی)، مدل‌های تحلیلی برای پیش‌بینی و توضیح دلایل زمینه‌ساز ریزش دانشجویان جدید الورود را ارائه دادیم. نتایج تحلیل‌های مقایسه‌ای نشان داد که مدل‌های جمعی بهتر از مدل‌های فردی عمل می‌کنند. ریزش دانشجو به‌عنوان یکی از چالش برانگیزترین مسائل برای تصمیم گیرندگان در مؤسسات آکادمیک و دانشگاهی در آمده است. بر اساس اطلاعات دایره آموزش و مرکز آمار آموزشی ایالات متحده، علی‌رقم تمام برنامه‌ها و خدمات برای کمک به حفظ دانشجو، تنها نیمی از کسانی که وارد تحصیلات تکمیلی می‌شوند می‌توانند مدرک لیسانس خود را دریافت کنند. مدیریت ثبت‌نام و حفظ دانشجویان یکی از بالاترین اولویت‌ها برای مدیران دانشکده‌ها و دانشگاه‌ها در ایالات متحده و دیگر کشورهای توسعه یافته در جهان شده است. تعداد بالای دانشجویان مردودی معمولاً منجر به افزایش ضررهای مالی، کاهش میزان فارغ‌التحصیلی دانشجویان و کاهش درجه معروفیت دانشگاه در چشم سهام‌داران خواهد شد. قانون‌گذاران و سیاست‌گذاران که مسئول تحصیلات عالیه بوده و منابع مالی را اختصاص می‌دهند، والدینی که هزینه تحصیل فرزندانشان را می‌دهند تا آنها را برای آینده‌ای بهتر آماده کنند و دانشجویانی که به‌دنبال انتخاب دانشگاه در آینده هستند، به‌دنبال مؤسساتی کیفی و معروف می‌گردند تا در فرآیند تصمیم‌گیری به آنها کمک کند. برای بهبود وضعیت حفظ دانشجویان، مؤسسه باید سعی کند تا

دلایل اصلی زمینه‌ساز ریزش را درک کند. برای موفقیت، همچنین مؤسسه باید بتواند به دقت دانشجویانی که در خطر اخراج هستند را شناسایی کند. تاکنونطیف وسیعی از تحقیقات در مورد ریزش دانشجویان معطوف به درک این پدیدهٔ پیچیده و تعیین کننده اجتماعی و حیاتی شده است. هرچند که این مطالعات کیفی، رفتاری و تحقیقی محور، دیدگاه گران‌بهایی را با توسعه و آزمایش طیف وسیعی از تئوری ها فراهم کرده‌اند، در ایجاد ابزاری برای پیش‌بینی دقیق ریزش دانشجویان ناتوان بوده‌اند[۳۱و۴۷].

در این پروژه ما یک رویکرد تحقیقی کمیتی را ارائه می‌دهیم که در آن از اطلاعات دانشجویان، برای توسعه مدل‌هایی که توانایی پیش‌بینی و همچنین توضیح مسئله ریزش دانشجویان را دارد، داشته باشد. گرچه این مفهوم برای تحصیلات تکمیلی در حدود یک دهه، نسبتاً جدید است، و مشکلات مشابه در زمینه بازاریابی با استفاده از روش‌های استخراج داده ( داده‌کاوی ) پیش‌بینی کننده تحت عنوان "تحلیل چرخنده"، مورد مطالعه قرار گرفته است، که در آنها هدف، شناسایی مشتریانی است که به احتمال زیاد از زیر پوشش شرکت خارج خواهند شد تا بدین وسیله فرآیندی برای مشتریانی که ارزش نگهداری را دارند، به‌کار گرفته شود. نگهداری مشتریان کنونی امری بسیار حیاتی است زیرا مطالعات مرتبط نشان داده است که به‌دست آوردن مشتریان جدید تقریبا ده برابر نگهداری مشتری فعلی هزینه دارد[۲۶].

## ۱-۲- بررسی تحقیقات مرتبط

علی‌رقم افزایش نرخ ثبت‌نام در مؤسسات آموزش عالی ایالات متحده، کیفیت پایین دانشگاه‌ها و نرخ بالای اخراج مشکلی اساسی برای دانشجویان جدیدالورود دوره لیسانس باقی مانده است. برای مؤسسات آموزشی نرخ بالای ریزش، برنامه‌ریزی برای ثبت‌نام را پیچیده و دشوار می‌سازد و باری مضاعف روی تلاش برای جذب دانشجویان جدید خواهد بود[۴۴و۷].

برای دانشجویان، اخراج قبل از گرفتن مدرک نشان دهنده پتانسیل انسانی استفاده نشده و بازدهی کمی از توانایی‌های آنها در دانشگاه خواهد بود. کارایی پایین و ضعیف مؤسسات آموزشی اغلب نشان دهنده مشکلات سازگاری با کالج بوده و ترک تحصیل و یا اخراج را افزایش می‌دهد.

ریزش دانشجویان در یک دانشگاه تعداد دانشجویانی که یک مقطع را در آن مؤسسه به اتمام نمی‌رسانند و مدرک خود را در دانشگاه تکمیل نمی‌کنند، تعریف می‌شود[۳۲].

مطالعات نشان داده است که اکثریت دانشجویان در طول اولین سال تحصیلشان نسبت به طول باقی مانده دوران تحصیلشان از درس کناره‌گیری می‌کنند[۲۷].

بنابراین اغلب، اخراج دانشجویان و ترک از تحصیل آنان در انتهای سال اول برای دانشجویان جدیدالورود اتفاق می‌افتد[۱۰و۱۸].

بسیاری از مطالعات در مورد ریزش و یا حفظ دانشجویان روی تعداد اخراج و یا ترک تحصیل دانشجویان در سال اول و یا دانشجویانی که در سال دوم مراجعه نمی‌کنند تمرکز دارند. این تعریف از ریزش تفاوتی میان دانشجویانی که ممکن است به دانشگاه‌های دیگر انتقالی گرفته باشند و در آنجا فارغ‌التحصیل شوند قائل نمی‌شود. در این تحقیق تنها دانشجویان که به‌طور ارادی نه به‌صورت اخراج در پایان اولین سال تحصیلی ترک تحصیل نموده‌اند در نظر گرفته شده اند[۳۶].

بررسی حفظ و یا ریزش دانشجویان سابقاً از طریق تحقیق بوده (مثلاً، بررسی گروهی از دانشجویان و دنبال کردن آنها برای یک دوره زمانی خاص برای تعیین اینکه آیا آنها به تحصیلات خود ادامه می‌دهند یا خیر)[۷].

با استفاده از چنین طرحی، محققان روی توسعه و ارزیابی روش‌های تئوریک، ازجمله مدل یکپارچگی دانشجویان که توسط تینتو[1] ایجاد شده بود، کار کرده‌اند[۴۵].

[1] Tinto

با تعمیم تئوری تینتو، دیگران نیز با استفاده از مطالعه تحقیقی مدل‌هایی برای ریزش دانشجویان به‌وجود آورده‌اند[۵و۶].

اگرچه آنها پایه‌های این حوزه را گذاشته‌اند، این مطالعات تحقیقی به‌علت نبود کاربرد عمومی برای دیگر مؤسسات و سختی و هزینه‌دار بودن مدیریت به‌کارگیری این‌چنین ابزار تحقیقی بزرگی مورد انتقاد قرار می‌گیرند[۸].

یک رویکرد جایگزین برای تحقیق‌های سنتی در مورد حفظ، رویکرد تحلیلی است که معمولاً از داده‌های متداول در پایگاه داده‌های مؤسسات استفاده می‌شود. مؤسسات آموزشی معمولاً طیف وسیعی از اطلاعات در مورد دانشجویانشان از جمله آمار جمعیتی، سوابق تحصیلی، مشارکت اجتماعی، وضعیت اجتماعی اقتصادی و پیشرفت تحصیلی را جمع‌آوری می‌کنند. مقایسه بین تحقیقات ریزش و یا حفظ دانشجویان که به‌صورت داده محور و میدانی صورت گرفته شده نشان داد که در بهترین حالت مشابه این دو قابل مقایسه بوده و برای توسعه یک مدل رگرسیون لجستیک داده محور، تحقیقات بر اساس داده از تحقیقات بر پایه مطالعات میدانی بهتر است[۷].

ولی در واقعیت، این دو روش تحقیقی (آنهایی که بر اساس مطالعات میدانی هستند و دیگر تئوری‌های برگرفته شده از داده‌های مؤسساتی و مدل‌های تحقیقاتی) مکمل و کمک کننده یکدیگر هستند[۳۳].

بنابراین، تحقیقات تئوریک می‌تواند کمکی باشند برای شناسایی متغیرهای پیش‌بینی کننده مهم تا از آنها برای مطالعات تحلیلی استفاده شود در حالی‌که مطالعات تحلیلی روابط جدیدی را در مورد متغیرها نشان می‌دهند که می‌تواند منجر به توسعه تئوری‌های جدید و بهبود تئوری‌های موجود می‌شود. تعدادی از عوامل و فاکتورهای آکادمیک، اقتصادی اجتماعی و مرتبط دیگر، در ریزش دانشجویان تأثیرگذار هستند. با توجه به گفته‌ی یوت زل و دوستان[1]،[۴۸] دانشگاه‌هایی که

[1] Wetzell et al

سیاست پذیرش بازتری دارند و در آنها لیست متقاضیان در حال انتظار و انتقالی وجود ندارد نسبت به دانشگاه‌هایی با متقاضیان جایگزین، با مشکلات ریزش دانشجو به‌طور جدی‌تری دست و پنجه نرم می‌کنند.

از طرف دیگر، هرمانویچ[1] [۱۸] دریافت که دانشگاه‌های سخت‌گیرتر در پذیرش لزوماً نرخ فارغ‌التحصیلی بالاتری ندارند، بلکه عوامل و فاکتورهای غیرمستقیم دیگری که با پذیرش در ارتباط هستند می‌توانند نقش کلیدی ایفا کنند. علاوه بر جنبه‌های ساختاری دانشگاه‌ها (مثل پذیرش و تأیید سطح علمی دانشگاه)، بخش و جنبه فرهنگی (مثل هنجارها و ارزش‌ها که راهنمای جامعه است) نیز باید به‌طور یکسان مورد توجه قرار گیرد زیرا نرخ بالاتر حفظ و پایداری دانشجو اغلب زمانی ایجاد می‌شود که دانشجویان متوجه شوند محیط دانشگاه با علائق فردی آنها ارتباط بالایی دارد[۱۸].

در تحقیقی مرتبط، آستین[2] [۳] معین کرد که نرخ ریزش یا پایداری دانشجویان تحت تأثیر عوامل مختلف اجتماعی، اقتصادی، سیاسی و یا حتی روانشناختی قرار دارد که با سطح و کیفیت تعاملات آنها با گروه‌های همانند و همسالان خود در دانشگاه و همچنین اعضای هیات علمی و کارکنان بستگی دارد.

از سوی دیگر تینتو [۴۶] نشان داد که دشواری تحصیل، مشکلات سازگاری، عدم اهداف شغلی و تحصیلی روشن، عدم اطمینان، عدم تعهد، عوامل مؤثر دخیل در اخراج و یا ترک تحصیل دانشجویان می‌باشد. جوزف جی گلین[3] محل اقامت، تأمین کننده هزینه‌های تحصیلی، برنامه‌ریزی شغلی را از عوامل مؤثر در ریزش دانشجویان ذکر می‌کند که این عوامل به‌طور پیچیده‌ای با یکدیگر ارتباط داشته و امکان پیش‌بینی آن را با مشکل مواجه می‌نماید [۱۶].

[1] Hermanowicz
[2] Astin
[3] Joseph G.Glynn

با توجه به نظریه ادغام دانشجویان تینتو [۴۶]، موفقیت تحصیلی گذشته و جاری جزئی کلیدی در تشخیص ریزش است. نمرات GPA وSAT ( جمع نمرات و معدل مقطع قبل ) دبیرستان دیدگاهی به ارائه آکادمیک بالقوه دانشجویان جدید الورود فراهم می‌کند و نشان داده شده است که اثر مثبت خوبی در پایداری آنها دارد[۴۵].

به‌صورت مشابه و گاهی مهم‌تر، معدل ترم اول همبستگی زیادی با ریزش را نشان می‌دهد. در این تحقیق ما این شاخص‌های موفقیت تحصیلی را مورد استفاده قرار داده‌ایم. به نظر می‌رسد تعهد به مؤسسه و اهداف پیش‌بینی کنندگان تأثیر قابل توجهی برای حفظ و نگهداری دانشجویان داشته باشند[۸].

دانشجویانی که در اهداف شغلی خود مصمم نیستند به اندازه دانشجویانی که در مورد مسیر شغلیشان مطمئن هستند از سطح مناسبی قدرت ذهنی و تعهد به اهداف خود برخوردار نیستند. بنابراین به‌عنوان مقیاسی کاذب از تعهد تحصیلی، در این تحقیق اعلان مدرک لیسانس و ساعت‌های پاس شده در ترم اول در این تحقیق به‌کار گرفته شده‌اند. به‌علاوه وضعیت سکونت دانشجویان (که به ساکن همان ایالت و ساکن ایالات دیگر دسته‌بندی می‌شوند) و یکپارچگی بهتر با فرهنگ مؤسسه، شاید نشانه‌ای از ارتباط عاطفی و اجتماعی باشد[۷].

دانشجویانی که از ایالات دیگر می‌آیند تعاملات ناشناخته‌تری را تجربه می‌کنند که باعث افزایش احساسات غریبگی و انزوا می‌گردد. مطالعات بسیاری نیز در گذشته، اثر کمک مالی را روی ریزش دانشجویان بررسی کرده‌اند[۱۹و۲۰و۴۰].

در این تحقیق مشخص شد که نوع کمک مالی یک عامل تعیین کننده در رفتار ریزشی دانشجویان است. دانشجویانی که بر اساس موفقیت‌های تحصیلی کمک مالی دریافت می‌کنند نرخ فرسایش بیشتری دارند[۴۰].

هچستین و بولتر[1] [۲۰] یافتند که کمک‌های مالی رابطه‌ای مستقیم با ریزش دانشجویان دارند که این در مورد وام منفی است. همین‌طورهرزوگ[2] [۱۹] به این نکته پی برد که بورسیه به باقی ماندن دانشجویان کمک می‌کند در حالی‌که از دست دادن این بورسیه‌ها به خاطر نمرات یا اعتبار نامناسب می‌تواند منجر به افزایش نرخ نقل و انتقال و اخراج دانشجو شود.

بنابراین، هدف‌های اصلی این مطالعه دو قسمتی هستند:

(۱) مدل‌های توسعه برای شناسایی دقیق دانشجویان جدیدالورودی که به احتمال بیشتر بعد از سال اول اخراج می‌شوند.

(۲) شناسایی مهم‌ترین متغیرها با به‌کارگیری تحلیل‌های حساسیت روی مدل‌های توسعه یافته.

مدل‌هایی که ما توسعه داده‌ایم به‌شکلی فرموله شده‌اند تا پیش‌بینی را در انتهای ترم اول انجام دهند (انتهای ترم پاییز) تا بتوان تصمیم گیرندگان را آماده برنامه‌ریزی مناسب برای نگهداری آنها در طول ترم بعد ( ترم بهار ) کنند.

## ۱-۲-۱-الگوسنجی (روش شناسی)

در این تحقیق، از یک الگوسنجی استخراج داده مشهور به نام CRISP-DM (فرآیند استاندارد بین صنعتی برای استخراج داده) استفاده شده که یک فرآیند شش مرحله‌ای است:

(۱) درک حوزه و توسعه اهداف برای مطالعه،

(۲) شناسایی، ارزیابی و درک منابع داده مرتبط،

(۳) پیش فرآیند، پاک‌سازی و نقل و انتقال داده‌های مرتبط،

(۴) توسعه مدل‌ها با استفاده از روش‌های تحلیل مقایسه‌ای،

(۵) ارزیابی و ارزیابی اعتبار و کارایی مدل در برابر هم و در برابر اهداف مطالعه،

(۶) به‌کارگیری مدل‌ها برای استفاده در فرآیندهای تصمیم‌گیری.

---

[1] Hochstein and Butler

[2] Herzog

این روش شناسی مشهور راهی نظام‌مند و ساختاری برای استفاده از مطالعات استخراج داده ایجاد می‌کند و بنابراین احتمال به‌دست آوردن نتایج دقیق و قابل استناد را افزایش می‌دهد. توجه مبذول شده به مراحل اولیه در CRISP-DM (مثلا درک حوزه مطالعه، درک داده و آماده‌سازی‌داده) زمینه را برای مطالعه استخراج داده موفق فراهم می‌کند. تقریبا ۸۰ درصد زمان کلی پروژه صرف این سه مرحله اولیه می‌شود. در این مطالعه برای تخمین عملکرد مدل‌های پیش‌بینی، از یک رویکرد بررسی اعتباری ۱۰ مرحله‌ای استفاده شده (روش خوشه‌بندی کی مینز- اعتبار سنجی)، در نظر گرفته شده است[۲۵].

مطالعات تجربی نشان داده‌اند که ۱۰ حداکثر مقدار ممکن برای مراحل می‌باشد (زمان مورد استفاده را برای تکمیل تست بهبود می‌بخشد در حالی که خطا و واریانس مرتبط با فرآیند ارزیابی را به حداقل می‌رساند). در بررسی اعتبار ۱۰ مرحله‌ای، کل دسته داده به ۱۰ زیر مجموعه خاص همانند تقسیم می‌شوند (یا ده مرحله) و هر مرحله تنها یک‌بار برای آزمون عملکرد پیش‌بینی مدل استفاده می‌شود و از داده‌های مرکب نه مرحله باقی مانده ساخته می‌شود.

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \qquad (۱-۱)$$

## ۱-۲-۲-توصیف داده‌ها

داده‌ها برای این تحقیق تنها از یک مؤسسه (یک دانشگاه دولتی فراگیر در منطقه غرب میانی ایالات متحده) گرفته شده است که به‌طور متوسط در آن ۲۳۰۰۰ دانشجو ثبت‌نام می‌کنند که تقریبا ۸۰ درصد آنها ساکنین همان ایالت هستند و تقریبا ۱۹ درصد ایشان بر اساس دسته‌بندی‌های اقلیتی ثبت شده‌اند. تفاوت معنی‌داری بین دو جنس در تعداد ثبت‌نام وجود ندارد. نرخ متوسط ماندن دانشجویان جدید الورود برای این مؤسسه ۸۰ درصد است و میانگین نرخ فارغ‌التحصیلی۶ ساله حدود ۶۰ درصد می‌باشد.

در این مطالعه ما از پنج سال داده‌های مؤسسه استفاده کردیم که شامل ۱۶۰۶۶ دانشجوی جدید الورودی بود که بین سال های ۲۰۰۴ تا ۲۰۰۸ ثبت‌نام کرده بودند. داده‌ها از پایگاه‌های داده‌های مختلف دانشجویی جمع‌آوری و مرتب شده است. خلاصه کوتاهی، از عددهای ثبت شده بر اساس سال در جدول (۱-۱) آمده است.

**Table 1**
Five-year freshmen student data used in this study.

| Year | Total number of freshmen students | Returned for the 2nd fall | Freshmen attrition (%) |
|---|---|---|---|
| 2004 | 3249 | 2541 | 21.79% |
| 2005 | 3306 | 2604 | 21.23% |
| 2006 | 3234 | 2576 | 20.35% |
| 2007 | 3207 | 2445 | 23.76% |
| 2008 | 3070 | 2391 | 22.12% |
| | Total: 16066 | Total: 12557 | Average: 21.84% |

جدول (۱-۱) دانشجویان جدیدالورود بر اساس سال

داده‌ها شامل متغیرهایی مرتبط با خصوصیات تحصیلی، مالی و جمعیتی دانشجویان می‌باشد. لیست کاملی از داده‌های به‌دست آمده از پایگاه‌های داده و پرونده دانشجویی در جدول(۱-۲) آمده است.

**Table 2**
Variables obtained from student records.

| No | Variables | Data type |
|---|---|---|
| 1 | College | Multi nominal |
| 2 | Degree | Multi nominal |
| 3 | Major | Multi nominal |
| 4 | Concentration | Multi nominal |
| 5 | Fall hours registered | Number |
| 6 | Fall earned hours | Number |
| 7 | Fall GPA | Number |
| 8 | Fall cumulative GPA | Number |
| 9 | Spring hours registered | Number |
| 10 | Spring earned hours | Number |
| 11 | Spring GPA | Number |
| 12 | Spring cumulative GPA | Number |
| 13 | Second fall registered (Y/N) | Nominal |
| 14 | Ethnicity | Nominal |
| 15 | Sex | Binary nominal |
| 16 | Residential code | Binary nominal |
| 17 | Marital status | Binary nominal |
| 18 | SAT high score comprehensive | Number |
| 19 | SAT high score English | Number |
| 20 | SAT High score Reading | Number |
| 21 | SAT High score Math | Number |
| 22 | SAT High score Science | Number |
| 23 | Age | Number |
| 24 | High school GPA | Number |
| 25 | High school graduation year and month | Date |
| 26 | Starting term as new freshmen | Multi nominal |
| 27 | TOEFL score | Number |
| 28 | Transfer hours | Number |
| 29 | CLEP earned hours | Number |
| 30 | Admission type | Multi nominal |
| 31 | Permanent address state | Multi nominal |
| 32 | Received fall financial aid | Binary nominal |
| 33 | Received spring financial aid | Binary nominal |
| 34 | Fall student loan | Binary nominal |
| 35 | Fall grant/tuition waiver/scholarship | Binary nominal |
| 36 | Fall federal work study | Binary nominal |
| 37 | Spring student loan | Binary nominal |
| 38 | Spring grant/tuition waiver/scholarship | Binary nominal |
| 39 | Spring federal work study | Binary nominal |

جدول (۱-۲) بیوگرافی دانشجویان

بعد از تبدیل داده‌های چند بعدی دانشجویی به فایل یک سطحی (یک تک فایل با ستون‌هایی که متغیرها را نشان می‌دهد و سطرهایی که داده‌های دانشجویی را نشان می‌دهد)، داده‌ها ارزیابی و پیش فرآیند شدند تا داده‌های غیر متعارف و غیر قابل استفاده شناسایی و حذف گردد.

## ۱-۲-۳- مدل‌های پیش‌بینی

در این تحقیق، دو روش دسته‌بندی مشهور (شبکه‌های عصبی مصنوعی و رگرسیون لجستیک) با استفاده از دقت پیش‌بینی در نمونه‌های بسط یافته مقایسه شدند. بسیاری از محققان به مقایسه

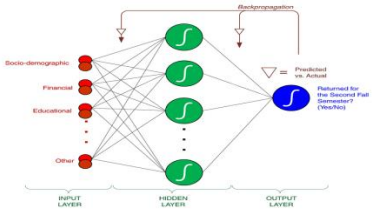
روش‌های استخراج داده در بسترهای مختلف به تحقیق پرداخته‌اند. بیشتر این تحقیقات روش یادگیری ماشینی (شبکه‌های عصبی مصنوعی) را نسبت به همتاهای آماری خود (رگرسیون لجستیک) در زمینه‌ی محدودیت‌های کمتر به‌وسیله‌ی فرضیات و تولید نتایج پیش‌بینی کننده‌ی بهتر، برتر دانسته اند[۱۲و۱۳و۲۱و۲۹و۴۱].

یافته‌های ما در این تحقیق این نتایج را تأیید می‌کند. در زیر توصیفی کوتاه از مدل‌های تکی و جمعی استفاده شده ارائه می‌گردد:

### ۱-۲-۳-۱- شبکه‌های عصبی مصنوعی(ANN)

تکنیک‌های تحلیلی برگرفته از زیست شناسی هستند و می‌توانند توابع غیرخطی بسیار پیچیده را مدل‌سازی کنند[۲۲].

در این تحقیق ما ساختاری معروف از شبکه‌های عصبی را به نام پرسپترون چند لایه را به‌همراه یک الگوریتم یادگیری back-propagation تحت نظارت مورد استفاده قرار داده‌ایم. پرسپترون چند لایه، یک تابع تقریب زننده قوی برای پیش‌بینی و دسته‌بندی مشکلات است و تقریباً رایج‌ترین ساختار شبکه عصبی مصنوعی مورد استفاده و مطالعه شده می‌باشد. هورنیک[1] [۲۳] به‌صورت تجربی نشان داد که پرسپترون چند لایه با اندازه و ساختار درست می‌تواند تابع‌های غیرخطی پیچیده دلخواه را تا سطح دقتی دلخواه فراگیرد. یک ارائه تصویری از ساختار شبکه عصبی مصنوعی مورد استفاده در این مطالعه در شکل (۱-۱) نشان داده شده است.




شکل ( ۱-۱ ) شبکه عصبی چند لایه

[1] Hornik et al

### ۱-۲-۳-۲- رگرسیون لجستیک

رگرسیون لجستیک تعمیم یافته‌ی رگرسیون خطی است. ازآن برای پیش‌بینی متغیرهای غیر مستقل دوتایی و چند طبقه‌ای استفاده می‌شود. از آنجا که متغیر پاسخ مجزا است، آن‌را نمی‌توان به‌صورت مستقیم با رگرسیون خطی مدل‌سازی کرد. بنابراین، به‌جای پیش‌بینی تخمین نقطه‌ای خود رویداد، مدلی می‌سازد که تا شانس روی دادن آنرا پیش‌بینی کند. در حالی‌که رگرسیون یک ابزار آماری معمول برای دسته‌بندی مسائل است، فرضیات محدود کننده‌ی آن روی استقلال و بهنجاری منجر به افزایش معروفیت و استفاده از تکنیک‌های یادگیری ماشینی برای مسائل پیش‌بینی دنیای واقعی می‌گردد.

## ۱-۳- بررسی نتایج

در این دسته از تجارب، ما از دسته داده اصلی استفاده کردیم که مرکب از ۱۶۰۶۶ داده ثبت شده بود. بر این اساس بهترین نتایج را بر اساس نرخ پیش‌بینی شبکه عصبی مصنوعی و رگرسیون لجستیک با نرخ پیش‌بینی ۸۶,۴۵ و ۸۶,۱۲ درصد قرار داشتند:

**Table 3**
Prediction results for *10*-fold cross validation with unbalanced dataset.

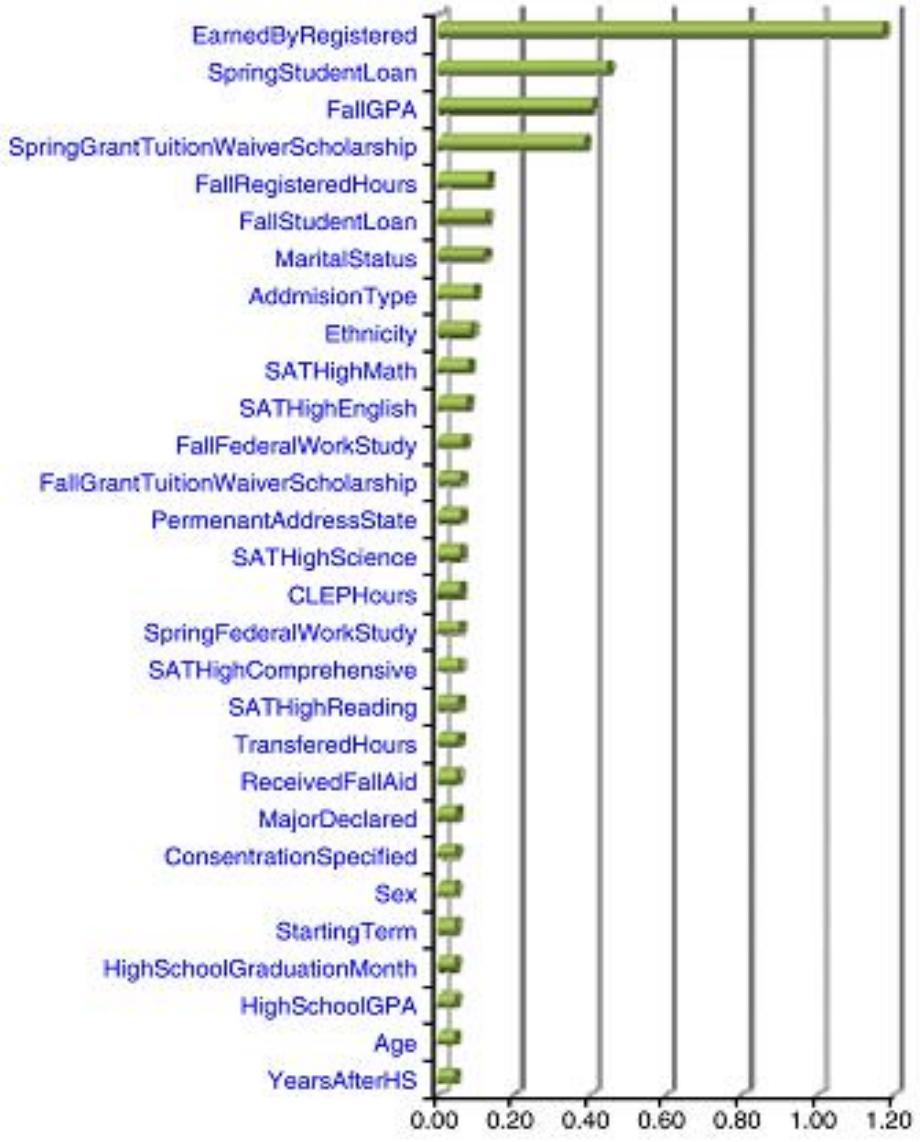
| | ANN(MLP) | | DT(C5) | | SVM | | LR | |
|---|---|---|---|---|---|---|---|---|
| | No | Yes | No | Yes | No | Yes | No | Yes |
| No | 1494 | 384 | 1518 | 304 | 1478 | 255 | 1438 | 376 |
| Yes | 1596 | 11142 | 1572 | 11222 | 1612 | 11271 | 1652 | 11150 |
| SUM | 3090 | 11526 | 3090 | 11526 | 3090 | 11526 | 3090 | 11526 |
| Per-class accuracy | 48.35% | 96.67% | 49.13% | 97.36% | 47.83% | 97.79% | 46.54% | 96.74% |
| Overall accuracy | 86.45% | | 87.16% | | 87.23% | | 86.12% | |

جدول (۱-۳) نتایج پیش‌بینی مدل غیر جمعی

آزمون محتاطانه این نتایج این موضوع را آشکار می‌کند که دقت پیش‌بینی برای طبقه "بله" به صورت قابل توجهی بیشتر از دقت پیش‌بینی طبقه "نه" است. در واقع، دانشجویانی که امکان بازگشت آنها در سال دوم محتمل‌تر است (ادامه تحصیل می‌دهند) را با دقتی بیشتر از ۹۰ درصد

پیش‌بینی کرده‌اند. در کل، تکنیک یادگیری ماشینی به‌صورت قابل توجهی بهتر از همتایان آماری خود و رگرسیون لجستیک است.

در شکل (۱-۲)، بردار y متغیرهای مستقل را بر اساس اهمیت حساسیت از بالا(اهمیت بیشتر) بـه پایین(اهمیت کمتر) لیست می‌کند در حالی‌که بردار X تراکم اهمیت نسـبی هـر متغیـر را نشـان می‌دهد.




شکل ( ۱-۲ ) اهمیت نسبی متغیرها

## ۱-۴- بحث و نتیجه‌گیری

نتایج نشان می‌دهد که، با ارائه داده کافی با متغیرهای مناسب، روش‌های استخراج داده می‌توانند ریزش دانشجویان جدیدالورود را با دقت تقریبی ۸۰ درصد پیش‌بینی کنند. از مدل‌های پیش‌بینی استفاده شده در این تحقیق، شبکه‌های عصبی از رگرسیون لجستیک بهترین عملکرد (پیش‌بینی) را داشته‌اند.

کاربرد اختصاصی این تحقیق دو دسته‌اند:

اول، مطالعه نشان داد که مؤسسات می‌توانند با استفاده از پایگاه داده موجود خود به‌همراه تکنیک‌های تحلیل پیشرفته، دانشجویان در معرض خطر را پیش‌بینی کنند و بنابراین تخصیص منابع محدود خود را برای نگهداری آنها بهبود بخشند

دوم، تحلیل فاکتورهای پراهمیت تعیین کننده علل اصلی ریزش دانشجویان است و بنابراین باید به دقت نظارت و مدیریت شوند.

متغیرهای پراهمیت در ریزش دانشجویان شامل:

(۱) تعاملات اجتماعی دانشجویان

(۲) سوابق تحصیلی دانشجویان ( معدل مقطع قبلی، نمرات دروس در مقاطع قبلی )

(۳) سابقه مالی و تحصیلی والدین دانشجو

(۴) تأمین کننده هزینه‌های تحصیلی

برای بهبود میزان ریزش دانشجویان، مؤسسات شاید بخواهند دانشجویانی را ثبت‌نام کنند که از لحاظ تحصیلی موفق‌تر هستند و برای آنها کمک مالی فراهم کنند. همچنین سوابق تحصیلی دانشجویان جدیدالورود در اولین ترم تحصیلی با دقت به ترکیب میانگین نمرات آنها مفید می‌باشد.