Tech. Edu. J. 13(1): 49-62, Winter 2019

Technology of Education Journal

(TEJ)

Homepage: jte.sru.ac.ir

ORIGINAL RESEARCH PAPER

Application of data mining techniques to predict students' mental health status to improve educational performance

H. Koosha[1,], S. Dangkoub[2] and A. A. Barzanooni[2]*

[1] *Department of Industrial Engineering, Ferdowsi University of Mashhad, Irann*

[2] *Department of Industrial Engineering and management, Sadjad University of Technology, Mashhad, Iran*

ABSTRACT

Submitted: 09 November 2017
Reviewed: 15 June 2018
Revise: 24 October 2018
Accept: 27 October 2018

KEYWORDS:
Data mining
Classification approach
Prediction
Mental health
Detailed examination requirement

* Corresponding author
koosha@um.ac.ir

Background and Objectives: Student mental health data has been recorded in the information systems of the universities across the country for several years, and due to its high volume, conventional statistical and psychoanalytic methods to predict patterns and factors affecting students' mental health are not effective. This is where data mining technology comes in handy and helps to predict and identify those at high risk based on the recorded data set of students 'physical and especially mental health status, and to make appropriate and timely decisions to improve students' condition. One of the main objectives of every managers of educational centers is making improvements in students' educational performance. Besides the educational factors, physical and mental health is considerable which has a significant effect on students' behavior. Therefore, some rules and patterns are required to make the best decisions, based on the prediction of students' mental health state. This paper proposes a data mining approach for analyzing and extracting patterns in terms of new students' mental health, which means whether they need to visit a psychologist. Our effort was on extracting hidden rules in new students' mental health examination by employing classification approach.

Methods: Techniques used in this study are decision tree, rule based classifier, neural network, logistic regression and support vector machine. Moreover, a parameter tuning process is done for all the techniques mentioned and the results presents the list of symptoms of individuals who need detailed examination.

Findings: The results of the research represent that one can predict the status of students' mental helath based on propsed model. One of the outcomes of decision tree is that if a person severely feels disappointed or seems to be obsessive by others, or feels that life is worthless, definitely a consultaion is needed.

Conclusion: Considering that most of the existing research in the field of health data mining have focused on physical health, it is suggested that for future studies, all levels of health, i.e dimensions of students' health, including physical, social and spiritual health, as well as a combination of these dimensions be considered. In addition, a review of the various approaches and techniques appropriate to the psychological data set should be conducted with the aim of creating an appropriate classification for the existing techniques in this field. It is also suggested that the present data set or similar data sets (student health monitoring information) be examined with other classification techniques and the results be compared with the results of the present study. In general, it is suggested that data mining technology be used to extract hidden patterns in the mental health data set of school students at different levels of education, office workers and organizations. Finally, it is recommended that future research in this field first implement the clustering approach on the psychological data set and then use the classification and forecasting approaches.

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
30	7	3

مقاله پژوهشی

کاربرد فنون داده‌کاوی برای پیش‌بینی وضعیت سلامت روان دانشجویان با هدف بهبود وضعیت آموزش

حمیدرضا کوشا[۱*]، ثناء دنگ‌کوب[۲] و امیرعباس برزنونی[۲]

[۱] گروه آموزشی مهندسی صنایع، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
[۲] گروه آموزشی مهندسی صنایع، دانشکده مهندسی صنایع و مدیریت، دانشگاه صنعتی سجاد، مشهد، ایران

چکیده

پیشینه و اهداف: داده مربوط به سلامت روان دانشجویان چندین سال است که در سیستم‌های اطلاعاتی دانشگاه‌های سراسر کشور ثبت می‌شود و به علت حجم بالای آن روش‌های معمول آماری و روان‌کاوی برای پیش‌بینی الگوها و عوامل مؤثر بر سلامت روان دانشجویان کارایی لازم را ندارد. این‌جاست که فن داده‌کاوی مفید واقع شده و کمک می‌کند بر اساس مجموعه داده ثبت‌شده از وضعیت جسمانی و به‌خصوص روانی دانشجویان، آن‌هایی که در معرض ریسک بالا هستند، پیش‌بینی و شناسایی شده و تصمیم‌گیری‌های مناسب و به‌هنگام برای بهبود وضعیت دانشجویان اتخاذ گردد. بهبود عملکرد دانشجویان همواره یکی از مهم‌ترین اهداف مسئولان و مدیران دانشگاه‌ها و مراکز آموزشی به شمار می‌رود. عوامل متعددی بر عملکرد مناسب دانشجویان تأثیرگذار است. علاوه بر عواملی که در حوزه آموزش و یادگیری دانشجویان است، موضوع سلامت جسمانی و روانی نیز بر نحوه عملکرد آن‌ها تأثیر می‌گذارد. به منظور تصمیم‌گیری به‌موقع و متناسب با وضعیت روانی هر دانشجو نیاز است الگوهایی در دسترس باشد تا بتوان بر اساس آن‌ها وضعیت بهداشت روان هر دانشجو پیش‌بینی شود. در این پژوهش تلاش شده با به‌کارگیری فن داده‌کاوی، وضعیت دانشجویان ورودی جدید دانشگاه، از لحاظ نیاز به مراجعه به مشاوره مورد بررسی قرار گیرد و الگوهای پنهان نهفته در مجموعه داده پایش سلامت روان دانشجویان با به‌کارگیری فنون رویکرد طبقه‌بندی استخراج گردد.

روش‌ها: فنون استفاده‌شده در این پژوهش، شامل درخت تصمیم، طبقه‌بندی بر اساس قانون، شبکه‌های عصبی، رگرسیون لجستیک و ماشین بردار پشتیبان می‌باشد. برای تمامی پارامترهای فنون مذکور، تنظیم انجام شده و نشان‌دهنده علائم نیاز به مشاوره با نرخ صحت ۹۹٪ می‌باشد.

یافته‌ها: نتایج پژوهش نشان داد: می‌توان بر اساس مدل تدوین شده، وضعیت سلامت روانی دانشجویان را پیش بینی نمود. یکی از خروجی‌های کاربرد روش درخت تصمیم، این است که اگر فردی از یک ماه گذشته تا به امروز شدیداً، احساس ناامیدی کند، یا به نظر اطرافیانش فردی وسواسی باشد یا احساس کند زندگی برایش بی‌ارزش است به مشاوره احتیاج دارد.

نتیجه‌گیری: با توجه به این که اکثر پژوهش‌های موجود در زمینه داده‌کاوی سلامت، تمرکز بر سلامت جسمانی داشته‌اند، پیشنهاد می‌شود برای مطالعات آتی تمامی سطوح سلامت یعنی ابعاد سلامت دانشجویان شامل سلامت جسمانی، اجتماعی و معنوی و هم‌چنین ترکیبی از این ابعاد مورد بررسی قرار گیرد. علاوه بر این مطالعه‌ای مروری بر روی انواع رویکردها و فنون مناسب برای مجموعه داده‌های روان‌شناسی با هدف ایجاد یک تقسیم‌بندی مناسب برای فنون موجود در این حوزه انجام شود؛ هم‌چنین پیشنهاد می‌شود، مجموعه داده حاضر و یا مجموعه داده‌های مشابه (اطلاعات پایش سلامت دانشجویان) با فنون دیگر طبقه‌بندی مورد بررسی قرار گرفته و نتایج حاصل با نتایج پژوهش حاضر مقایسه گردد. به طور کلی پیشنهاد می‌شود از فن داده‌کاوی برای استخراج الگوهای پنهان در مجموعه داده سلامت روان دانش‌آموزان مدارس در مقاطع تحصیلی متفاوت، کارمندان ادارات و سازمان‌ها استفاده گردد. در نهایت توصیه می‌گردد پژوهش‌های آتی در این زمینه ابتدا رویکرد خوشه‌بندی را بر روی مجموعه داده روان‌شناسی پیاده کنند و به دنبال آن از رویکردهای طبقه‌بندی و پیش‌بینی استفاده نمایند.

دریافت: ۱۸ آبان ۱۳۹۶
داوری: ۲۵ خرداد ۱۳۹۷
اصلاح: ۰۲ آبان ۱۳۹۷
پذیرش: ۰۵ آبان ۱۳۹۷

واژگان کلیدی:
داده‌کاوی
رویکرد طبقه‌بندی
پیش‌بینی
سلامت روانی
علائم نیاز به مشاوره

*نویسنده مسئول
koosha@um.ac.ir

مقدمه

یکی از مشخصه‌های تأثیرگذار در پیش‌بینی وضعیت عملکرد تحصیلی دانشجویان، وضعیت سلامت جسم و روان آن‌ها است[۱]؛ چرا که در مطالعات پیش از این نشان داده شده است که عدم برخورداری از سلامت جسمانی و روانی در موفقیت یا عدم موفقیت دانشجویان در طول دوران تحصیلشان تأثیر به‌سزایی می‌گذارد؛ اما مسئله‌ای که در این‌جا مطرح است نحوه برخورد با این شرایط است. به همین منظور باید سازوکاری طراحی شود تا این قابلیت برای مسئولین دانشگاه فراهم گردد که بتوانند این دانشجویان را شناسایی کرده و برنامهٔ جداگانه‌ای برای مرتفع ساختن مشکل

ایجادشده طراحی کنند. داده مربوط به سلامت روان دانشجویان چندین سال است که در سیستم‌های اطلاعاتی دانشگاه‌های سراسر کشور ثبت می‌شود و به علت حجم بالای آن روش‌های معمول آماری و روان‌کاوی برای پیش‌بینی الگوها و عوامل مؤثر بر سلامت روان دانشجویان کارایی لازم را ندارد. این‌جاست که فن داده‌کاوی مفید واقع شده و کمک می‌کند بر اساس مجموعه داده ثبت‌شده از وضعیت جسمانی و به‌خصوص روانی دانشجویان، آن‌هایی که در معرض ریسک بالا هستند، پیش‌بینی و شناسایی شده و تصمیم‌گیری‌های مناسب و به‌هنگام برای بهبود وضعیت دانشجویان اتخاذ گردد. سهم بزرگی از مطالعات در زمینه بررسی وضعیت عملکرد تحصیلی دانشجویان، مختص به بررسی عوامل تأثیرگذار و پیش‌بینی وضعیت عملکرد تحصیلی آن‌ها می‌باشد[۲]، در حالی که هر دو گروه مشاوران تحصیلی و بالینی بر این موضوع تأکید دارند که وضعیت بهداشت روان دانشجویان بر بهبود عملکرد تحصیلی آن‌ها تأثیر مستقیم دارد. علی‌رغم اهمیت موضوع، مطالعه‌ای با رویکرد پیش‌بینی وضعیت سلامت روان دانشجویان با استفاده از فن داده‌کاوی به منظور ایجاد فضایی با قابلیت اطمینان بالاتر برای هر دانشجو متناسب با وضعیت بهداشت روان وی، موضوعی است که تا کنون کم‌تر مورد توجه پژوهشگران این عرصه بوده است.

در عصر اطلاعاتی امروز داده ثبت شده در سیستم‌های اطلاعاتی تبدیل به یکی از اصلی‌ترین دارایی‌های موسسات سلامت شده است. هم‌چنین علت اصلی ذخیره الکترونیکی داده این است که بهترین بهره اطلاعاتی در زمان مناسب دریافت گردد چرا که یکی از کاربردهای فناوری اطلاعات در حوزه سلامت فراهم کردن بستری برای پشتیبانی از تصمیمات مدیریتی است[۳]. رشد و توسعه حجم داده‌های ثبت‌شده در سیستم‌های اطلاعاتی، این نیاز را ایجاد کرده است که اطلاعاتی معنادار از میان این انبوه داده استخراج گردد. هم‌چنین امروزه در دنیای فناوری، حجم بسیار زیادی از اطلاعات خام وجود دارد که به تنهایی هیچ کاربردی ندارند، اما می‌توان با استفاده از تکنیک‌های داده‌کاوی بهترین بهره‌برداری را از این اطلاعات خام به عمل آورد و داده‌کاوی یک ابزار مطلوب است که می‌توان با استفاده از آن بهترین الگوها و اطلاعات را از داده خام استخراج کرد. مفهوم کشف دانش از داده بیش از یک دهه است که در محیط‌های مالی- تجاری در حال استفاده می‌باشد و در علوم مدیریت ارتباطات، مهندسی، وب‌کاوی، تحلیل جرایم و پزشکی جای خود را باز کرده است؛ اگر چه کشف دانش با هدف شناسایی اختلاس مالی وارد عرصه‌ی سلامت شد، اما به تدریج در حوزه‌ی بالینی نیز مورد استفاده قرار گرفت. این مهم ناشی از تغییر سریع هوشیاری نسبت به اطلاعات در حوزه‌ی سلامت است[۴].

در این پژوهش، از مجموعه داده پایش سلامت روان دانشجویان ورودی یک سال دانشگاهی در شهر مشهد به منظور پیش‌بینی نیاز یا عدم نیاز به مراجعه به مشاوره استفاده شده است. به همین منظور از رویکرد طبقه‌بندی، و شش روش آن از جمله درخت تصمیم، رگرسیون لجستیک، ماشین بردار پشتیبان (SVM)، شبکه‌های عصبی، نزدیک‌ترین همسایه و طبقه‌بندی بر مبنای قانون (Rule induction) بهره گرفته شد. در ادامه ابتدا مروری بر پیشینه پژوهش با تمرکز بر داده‌کاوی در زمینه تحصیلی و سلامت انجام شده است.

اهمیت و کاربرد داده‌کاوی در حوزه‌های مختلف از جمله بازاریابی، تبلیغات، سلامت، مهندسی و سیستم‌های اطلاعات و غیره بر هیچ کس پوشیده نیست[۱]. یکی از پرکاربردترین حوزه‌ها در مطالعات داده‌کاوی، حوزه سلامت برای بررسی عوارض احتمالی داروهای جدید، کشف ارتباط و علائم بیماری و غیره می‌باشد. پژوهش‌های مرتبط با مطالعه حاضر در دو حوزه تقسیم شده‌اند: داده‌کاوی در حوزه تحصیلی و داده‌کاوی در حوزه سلامت جسمانی و روانی. اهمیت تأثیر سلامت جسمانی افراد بر انجام فعالیت‌های روزانه معمول و به صورت خاص تحصیل بر کسی پوشیده نیست. در این میان مجموعه‌ای از عوامل روانی نیز بر نحوه اجرای فعالیت‌ها اثرگذار بوده، به همین جهت در پیش‌بینی وضعیت تحصیلی دانشجویان باید مورد توجه قرار گیرد.

در راستای رسیدن به موضوع تحقیقاتی مورد نظر، مقالات با رویکردهای داده‌کاوی از میان مقالات موجود با هدف داده‌کاوی تحصیلی و سلامت استخراج شدند و نکات کلیدی آن‌ها در بند پیشینه پژوهش و در جدول یک به طور خلاصه ارائه گردیدند. دسته‌بندی موضوعی پیشینه پژوهش با هدف ترکیب موضوعات داده‌کاوی سلامت روان و داده‌کاوی تحصیلی بود که می‌توان گفت بر اساس جست‌وجوهای انجام گرفته، تحقیقی با این ویژگی انجام نشده است.

داده‌کاوی تحصیلی

در سال‌های اخیر موضوع افت تحصیلی دانشجویان و بررسی عوامل مؤثر بر آن مورد توجه پژوهشگران قرار گرفته است. تجربه نشان داده است که بسیاری از دانشجویان با مشکلات زیادی در طول دوران تحصیل خود مواجه می‌شوند که در نهایت منجر به انصراف و یا عدم موفقیتشان در تکمیل تحصیلات می‌شود. به همین منظور، تحقیقات بسیاری با کمک فن داده‌کاوی انجام شده تا بتوان پیش‌بینی کرد دانشجویان با چه مشخصه‌هایی در معرض شکست تحصیلی هستند. به طور کلی به انجام این فرآیند EDM (Educational Data Mining) یا همان داده‌کاوی تحصیلی گفته می‌شود. در فرآیند داده‌کاوی تحصیلی، تلاش بر این است به این گونه سؤالات پاسخ داده شود؛ به عنوان مثال، دانشجویان در طول دوران تحصیلشان با چه مشکلاتی مواجه می‌شوند؟ چه اقداماتی بر نحوه عملکرد دانشجویان تأثیر بیش‌تری دارد؟ طرح‌های آموزشی مختلف چگونه بر یادگیری دانشجویان اثر می‌گذارد؟ کاستا و همکاران، از ۴ روش طبقه‌بندی قاعده بیزی، درخت تصمیم، شبکه‌های عصبی و ماشین بردار پشتیبان برای پیش‌بینی وضعیت

تحصیلی هر دانشجو استفاده کردند[۱]. نتایج حاصل از داده‌کاوی در مطالعه دیگری نشان داد که یک همبستگی قوی بین عملکرد دانشجویان سال اول و عملکرد کلی آن‌ها در طول دوران تحصیلشان وجود دارد. مشخصه‌هایی که مورد استفاده قرار گرفتند شامل این موارد می‌باشند: سوابق تحصیلی دانشجویان، داده‌های دموگرافیک آن‌ها، وضعیت اقتصادی و وضعیت خانوادگی[۵]. الگوهایی که از به‌کارگیری فنون طبقه‌بندی حاصل می‌شود معمولاً از طریق انطباق آن‌ها با شرایط دانشجویان ورودی سال آینده هر مرکز آموزشی قابل استناد است. به این معنا که در صورتی که نتایج حاصل، از نرخ صحت بالایی برخوردار باشد، به راحتی می‌توان وضعیت تحصیلی دانشجویان جدید را پیش‌بینی نمود. به عنوان مثال، در خروجی یکی از مطالعات در این عرصه با به‌کارگیری روش درخت تصمیم، مشخص شد که مشخصه‌های تحصیلات مادر، عادات متفرقه دانش‌آموز، درآمد خانواده و وضعیت خانوادگی دانش‌آموز بر روی عملکرد بهتر دانشجویان تأثیر می‌گذارد[۶].

رویکردهای توصیفی نیز کاربرد گسترده‌ای در بحث EDM دارد؛ چراکه خروجی آن کمک بزرگی در تسهیل فرآیند پیش‌بینی عملکرد دانشجویان خواهد داشت. به طور مثال در پژوهشی که مجموعه داده دانشگاهی را مورد کاوش قرار داده بود، ابتدا با رویکرد خوشه‌بندی، دانشجویان در سه خوشه دانشجویان فعال و باهوش، دانشجویان متوسط و دانشجویان ضعیف با کمک روش K-means قرار گرفتند و در ادامه برای هر یک از این خوشه‌ها، از فنون طبقه‌بندی برای پیش‌بینی وضعیت عملکرد تحصیلی هر دانشجو در خوشه خودش، استفاده شده است[۷].

داده‌کاوی در حوزه سلامت

یکی از پرکاربردترین حوزه‌های داده‌کاوی، حوزه سلامت است. درصد زیادی از پژوهش‌هایی که تاکنون در این زمینه انجام شده، کاوش داده به منظور استخراج الگو از میان مجموعه داده‌های سلامت جسمانی افراد شامل مشخصه‌های سوابق پزشکی آن‌ها می‌باشد. در مقابل تعداد مطالعاتی که به صورت خاص به داده‌کاوی در حوزه سلامت روان افراد بر اساس مشخصه‌های ثبت‌شده پرداخته باشند کم است. در این بخش به صورت مختصر به بیان خلاصه‌ای از مقالاتی که داده‌کاوی را در این حوزه‌ها به کار بستند پرداخته می‌شود.

در یکی از مطالعات در زمینه بیماری‌های قلبی، نویسندگان مقاله، پس از بیان اهمیت افزایش بیماری‌های قلبی در میان مردم جهان، از مجموعه داده شامل ۱۰۰۰ بیمار مراجع با یک سری از مشخصه‌های بالینی تعیین‌شده، الگوهایی استخراج شود که در ادامه بتوان در پیش‌بینی این که بیمار مراجع دارای بیماری قلبی هست یا خیر کمک کند[۸]. در پژوهش پرکاربرد دیگری، مشخصه‌های جسمانی مادر و جنین، مورد تجزیه و تحلیل قرار گرفت، تا نوع عقب‌ماندگی احتمالی نوزاد وی پیش‌بینی شود. چراکه تشخیص زودهنگام این اختلال به پیشگیری و درمان به‌موقع آن کمک می‌کند[۹]. در پژوهش ینگ و همکاران در سال ۲۰۱۴، از دیتاستی دارای ۳۹۷۰ رکورد در کشور چین استفاده شده است. کلاس‌های لیبل پیشنهادی برای پیش‌بینی وضعیت بیماران، به این ترتیب بود: شدید، متوسط، نرمال. مشخصه‌های به‌کارگرفته‌شده نیز در چهار حوزه نمود جسمانی، نمود روانی، سازگاری اجتماعی و زندگی زناشویی طبقه‌بندی شده‌اند که هر کدام دارای مشخصه‌های منحصربه‌فرد خود بوده و در پنج وضعیت قرار گرفته‌اند. به طور کلی با توجه به گستردگی موضوع سلامت روان، مطابق با نظر روان‌کاو، لیبل‌های متنوع و جذابی می‌توان تعریف کرد. به عنوان مثال در پژوهش کاسترومن و همکاران، تلاش شده با به‌کارگیری رویکرد طبقه‌بندی و روش قاعده بیزی، ارتباط بین عوامل ریسک‌زا و تلاش‌های مکرر خودکشی افراد پیدا شود. در پژوهش دیگری با موضوع مشابه، از رویکرد خوشه‌بندی به منظور شناسایی عوامل مؤثر بر ریسک خودکشی مکرر افراد استفاده شد. به طور دقیق‌تر نویسندگان این مقاله، به دنبال کشف الگوهایی بودند تا این عوامل به درستی شناسایی شوند. عوامل شناخته شده در این مطالعه به این ترتیب به دست آمد: اختلال شخصیتی، شکایات بی‌خوابی و سردرد، گزارش‌هایی از اتفاقات و حوادث ناگوار در زندگی مانند بی‌کاری، طلاق و اختلاف، تجربیات احساسات منفی و مصرف الکل. شناسایی این عوامل کمک بزرگی به روان‌کاوان در شناسایی بیماران بالقوه خودکشی مجدد می‌کند[۱۰].

در جدول شماره ۱ شماری از مطالعاتی که با موضوع وضعیت تحصیلی دانشجویان و کاربرد داده‌کاوی در حوزه سلامت انجام گرفته، به تفکیک هدف، رویکرد و روش مورد استفاده، ارائه شده است.

همان‌گونه که در بند ۱.۱ مطرح شد، مقالاتی که برای پیشینه پژوهش مورد بررسی قرار گرفتند در دو حوزه کلی سلامت و آموزش تقسیم‌بندی شده‌اند. در جدول ۱ تلاش شده است، بخشی از مقالاتی که از رویکردهای مختلف داده‌کاوی و فنون آن برای رسیدن به هدف مشخصی استفاده کرده‌اند به اختصار مطرح گردد. رویکردهای داده‌کاوی شامل رویکردهای خوشه‌بندی، طبقه‌بندی و پیش‌بینی است. با توجه به هدف پژوهش، متغیر هدفی که برای رویکردهای طبقه‌بندی و پیش‌بینی مورد استفاده پژوهشگران قرار گرفته است نیز متناسب با مجموعه داده تعیین گردیده است. فنون داده‌کاوی نیز با توجه به رویکرد انتخابی تفاوت می‌کند که برخی از مقالات از چندین روش و با هدف مقایسه نتایج از آن‌ها استفاده کرده‌اند.

با بررسی و جست و جو در میان مقالات با محوریت داده‌کاوی در زمینه آموزش و عوامل مؤثر بر عملکرد تحصیلی دانش‌آموزان و دانشجویان، مطالعات زیادی در زمینه EDM که بیش‌تر شامل مشخصه‌های سوابق تحصیلی دانشجویان بود مشاهده گردید. در میان مشخصه‌های مورد استفاده، در برخی از پژوهش‌ها، مشخصه‌های سلامت جسمانی و روانی دانشجویان نیز یافت

جدول ۱: مرور ادبیات
Table 1: Literature review

Reference	Target	Classification/ Prediction/ Clustering	Target varilable	Technique					
				Bayes' theorem	Decision tree	Support vector machine	Accossiation rule mining	Neural network	Regression
15	Prediction of students' performance by extracting hidden patterns in information systems	Prediction	Final score		*				
7	Discovering hidden rules of students' performance and clustering	Clustering							
16	Prediction students' success or failure during education	Prediction	Final score		*				
1	Effectiveness evaluation of classification techniques in predicting higher education students	Prediction	Leaving school	*	*	*		*	
14	Knowledge discovery from students' educational background to predict E-learning performance	Prediction	Leaving school						*
17	Determining key performance indicators for quality evaluation of each course at university	Classification			*	*		*	
18	Students' performance evaluation by predicting success and improvement trend while studying	Prediction	Instructors' behavior		*				
6	Establishing a model for predicting and recognizing effective factors on students' performance	Prediction	Need for educational consultation	*	*			*	*
19	Establishing an evaluation model to assess students' skills at designing and conducting experiments within a complex systems micro world	Prediction	Students' ability in design of experiment		*				
20	Performance prediction for high school students with low comprehensive power	Prediction	Students' quality level	*	*				
21	Detemination and classification of psychopathy	Classification/ Clustering	Type of psychopathy		*	*			
22	Association between completed suicide and environmental temperature	Classification					*		
23	Predict diabetes type in young and old patients	Prediction	Type of diabetes	*		*			*
24	Determinants of theory of mind performance in Alzheimer's disease	Clustering							
9	Predict early intervention for developmentally-delayed children	Prediction	Type of developmental delay		*		*		
25	Detect and assessment Insomnia symptoms in patients	Classification	Duration of using CPAP		*				
26	Study on efficiency of data mining in diabetes research	Classification/ Prediction	Type of diabetes			*		*	
27	Study on sub-mentally healthy state	Prediction	Predict level of mental health					*	
28	Prediction of educational progress with fuzzy clustering in educational centers	Clustering							
29	Improvement in the quality of electronic educational systems using educational data maining	Prediction	Student unit selection basket analysis				*		

شد، اما تعداد مطالعاتی که به صورت خاص، به کشف الگوهای جذاب از مجموعه داده سلامت جسمانی و مخصوصاً سلامت روانی افراد پرداخته باشد، در ادبیات موضوع بسیار اندک است. با توجه به شکاف تحقیقاتی مشاهده شده در این زمینه و اهمیت تأثیر سلامت روان دانشجویان بر وضعیت عملکرد تحصیلی آن‌ها، تلاش شده است در این پژوهش، با کمک مجموعه داده پایش سلامت روانی دانشجویان، الگوهای جذابی بر اساس متغیر هدف تعیین‌شده ارائه گردد.

ورود به دانشگاه با بروز تغییرات زیادی در روابط اجتماعی و انسانی همراه است. در چنین شرایطی که اغلب با فشار و نگرانی توأم است، عملکرد و بازدهی افراد تحت تأثیر قرار می‌گیرد. آشنا نبودن بسیاری از دانشجویان با محیط دانشگاه در بدو ورود، جدایی و

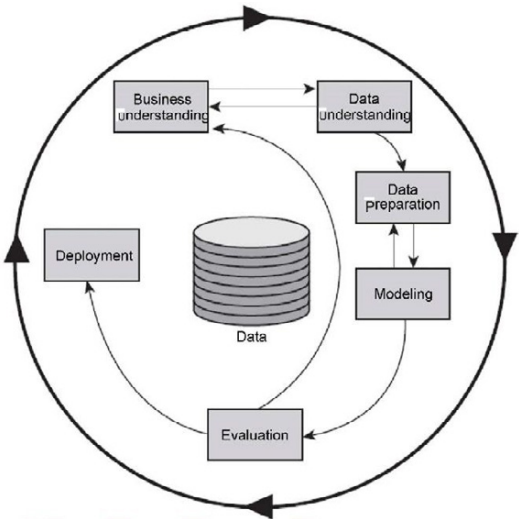
دوری از خانواده، عدم علاقه به رشته قبولی، ناسازگاری با سایر افراد در محیط زندگی و کافی نبودن امکانات رفاهی و اقتصادی، ازجمله شرایطی هستند که می‌توانند موجبات مشکلات و ناراحتی‌های روانی و افت عملکرد را موجب شوند. بنابراین گریز از عوامل استرس‌زا و دوری از مواجهه با تغییرات شتابنده امکان‌پذیر نیست. بنابراین در وهله اول شناخت عواملی که منجر به سلامت روان افراد به خصوص دانشجویان می‌گردد از اهمیت ویژه‌ای برخوردار است. خروجی‌ها در تحلیل‌های آماری صرفاً، بررسی فراوانی و کمیّت‌های مربوطه می‌باشد و به دنبال شناسایی عوامل مؤثر بر سلامت روحی افراد نمی‌باشد. بنابراین با توجه به عدم کارایی تحلیل‌های آماری، تحلیل‌های بیش‌تر و دقیق‌تری نیاز است تا دقیقاً وضعیت هر دانشجو در مورد نیاز یا عدم نیاز به مراجعه به مشاور مشخص گردیده و الگوهایی استخراج شود تا بتوان بر اساس آن فرآیند پیش‌بینی این نیازمندی را نیز انجام داد.

روش تحقیق

هدف فرآیند کاوش داده این است که دانش معنادار، استخراج شود. به صورت خاص داده‌کاوی در حوزه تحصیلی، که از آن به عنوان EDM یاد می‌شود، به معنای توانایی پیش‌بینی عملکرد تحصیلی دانشجو با توجه به عوامل شخصی، اجتماعی، روان‌شناسی و سایر مشخصه‌های محیطی است[۱]. با توجه به مساله حاضر و مرور ادبیات موجود، فن داده‌کاوی برای برآورد نیاز مساله انتخاب گردید. هدف این مطالعه، بررسی وضعیت بهداشت روان دانشجویان می‌باشد و این تحقیق از حیث هدف، تحقیقی کاربردی است و از حیث نحوه گردآوری اطلاعات، از نوع مطالعات همبستگی می‌باشد؛ چرا که در این مطالعه از تجزیه و تحلیل ارتباط بین متغیرها، به منظور پیش‌بینی متغیر هدف (نیاز و یا عدم نیاز به مراجعه به مشاور) بهره گرفته شده است.

مورد مطالعه

جامعه آماری مورد مطالعه، تمامی دانشجویان دانشگاه مورد مطالعه، و نمونه مورد نظر جهت اجرای تجزیه و تحلیل، دانشجویان ورودی سال ۱۳۹۵ (مهرماه) می‌باشد. بنابراین، مجموعه داده مورد استفاده، شامل داده سلامت روان دانشجویان نمونه آماری بوده که شامل ۷۷۷ دانشجو و ۱۰۶ مشخصه است. مشخصه‌های مورد بررسی، در دو بخش کلی شامل مشخصات عمومی دانشجویان و مشخصه‌های سلامت روان آن‌ها قرار گرفته است. پرسش‌نامه ارائه شده به دانشجویان شامل سه بخش می‌شود که این ابعاد مبتنی بر ابعاد پیشنهادی دفتر مشاوره و سلامت سازمان امور دانشجویان وزارت علوم، تحقیقات و فناوری است و به صورت مدون در مراکز آموزش عالی سراسر کشور مورد بررسی قرار می‌گیرد. ابعادی که در بخش اول مورد ارزیابی قرار می‌گیرد از آن‌ها به عنوان ابعاد سلامت‌نگر نام برده شده و شامل عواطف مثبت و بهزیستی روانی می‌شود. بخش دوم پرسش‌نامه که مربوط به احساسات، نگرش‌ها و رفتار یک ماه اخیر دانشجویان می‌باشد بعد بیماری‌نگر نام دارد و شامل این موارد می‌گردد: افسردگی، اضطراب، وسواس، اضطراب اجتماعی، اختلال خواب، افسردگی تحصیلی، اضطراب تحصیلی و تمایلات خودکشی. در بخش سوم نیز سایر مشخصه‌های روانی دانشجویان تحت عنوان سازه‌های سلامت روان شامل این موارد ارزیابی شده است: کمال‌گرایی، جو خانوادگی، جهت‌گیری مذهبی، خودکارآمدی و حمایت اجتماعی.

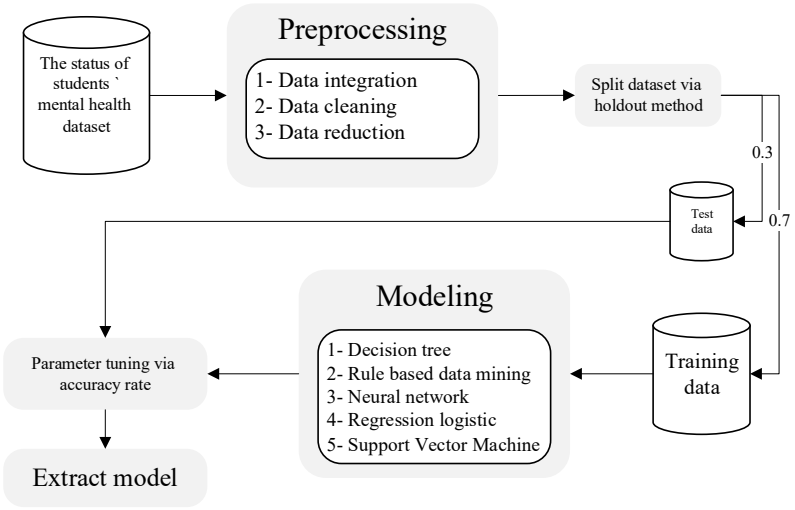


شکل ۱: مراحل داده‌کاوی بر اساس استاندارد CRISP-DM [۱۱]
Fig. 1: Data mining process based on CRISP-DM [11]

داده‌کاوی

داده‌کاوی فرآیند کشف الگو از حجم بالایی از یک داده ثبت شده در سیستم‌های اطلاعاتی است[۱۱]. داده‌کاوی فن بسیار ارزشمندی است که در سال‌های اخیر به طور گسترده‌ای برای استخراج اطلاعات، جست‌وجوی روابط و الگوها بین حجم عظیمی از داده استفاده شده است. داده‌کاوی از ترکیب چندین علم نشأت می‌گیرد: آمار، یادگیری ماشین، روش‌های تنظیم، روش‌های تشخیص و شناخت الگو، مدل‌های ریاضی و غیره فنونی هستند که داده‌کاوی از آن‌ها بهره می‌برد[۱۲]. مراحل انجام پژوهش حاضر بر اساس الگویی است که استاندارد جهانی فرآیند داده‌کاوی در صنعت ارائه کرده و مراحل آن در شکل ۱ نمایش داده شده است. الگوریتم کریسپ یکی از روش‌های تحلیلی متفاوت برای فرآیند داده‌کاوی است؛ این واژه از عبارت «Cross industry standard process for data mining» و به معنی "فرآیندهای استاندارد صنعت متقابل برای داده‌کاوی" تشکیل شده است.

در پژوهش حاضر، شش مرحله‌ای که در الگوریتم مذکور وجود دارد مورد استفاده قرار گرفته است و جزئیات اقدامات هر یک از مراحل در ادامه توضیح داده شده است. در گام فهم تجاری، نیازسنجی اجرای تحقیق از طریق بررسی نیازمندی‌های سازمان و مصاحبه با کارشناسان مرتبط انجام می‌گیرد. در گام درک داده،



شکل ۲: ارچوب پیشنهادی پژوهش
Fig. 2: Suggested research framework

مجموعه داده جهت تجزیه و تحلیل، انتخاب و گردآوری می‌شود. گام‌های ۳، ۴ و ۵ شامل پیش‌پردازش (آماده‌سازی داده)، مدل‌سازی و ارزیابی نتایج می‌باشد که در شکل ۲ ارائه شده است. در گام نهایی، یا فاز توسعه، از نتایج به دست آمده، در راستای بهبود سلامت روان و به دنبال آن وضعیت تحصیلی دانشجویان استفاده خواهد شد.

رویکردهای داده‌کاوی شامل سه دسته یادگیری نظارتی (پیش‌بینی‌کننده)، غیرنظارتی (توصیفی) و شبه‌نظارتی می‌باشد. در الگوریتم‌های یادگیری نظارتی هدف از داده‌کاوی مشخص است و تصمیم‌گیرنده می‌داند که به دنبال چه نوع دانشی است. یادگیری نظارتی شامل دو دسته طبقه‌بندی و رگرسیون می‌باشد.

در الگوریتم‌های یادگیری بدون نظارت، هدف کاملاً تعریف‌شده نیست و خروجی حاصل نوعی دسته‌بندی اشیا بر اساس مشخصه‌های مورد نظر می‌باشد.؛ یکی از روش‌های آن خوشه‌بندی می‌باشد. هدف مطالعه حاضر، پیش‌بینی نیاز به مشاور روان‌شناسی با توجه به مشخصه‌ها و لیبل مورد نظر می‌باشد که توسط روش‌های یادگیری بانظارت قابل دست‌یابی است.

مراحل اجرای فرآیند کاوش داده و به دنبال آن، استخراج الگوهای پنهان مجموعه داده وضعیت بهداشت روان، در شکل ۲ نشان داده شده است. که در ادامه به بیان شرح مختصری از هر یک از گام‌های پیاده‌سازی آن پرداخته خواهد شد.

با توجه به مورد مطالعه، چارچوبی برای داده‌کاوی و کاوش داده پیشنهاد می‌شود که در شکل ۲ مشاهده می‌شود. مراحل در نظر گرفته شده، شامل ۷ فاز اصلی می‌باشد.

مرحله اول: پیش‌پردازش؛ در این مرحله داده‌ی مغشوش و ناسازگار حذف می‌شود و در ادامه در صورت نیاز داده‌ای که در چند منبع مختلف قرار دارد تجمیع و یکپارچه می‌شوند.

مرحله دوم: تقسیم‌بندی مجموعه داده؛ در این مرحله داده به دو مجموعه تست و آزمون تقسیم می‌شوند.

مرحله سوم: مدل‌سازی؛ مرحله اساسی کار که در آن با روش‌های هوشمند، الگوها از داده استخراج می‌گردند

مرحله چهارم: ارزیابی الگوها؛ تعیین الگوهای جالبِ نشان‌دهنده دانش.

مرحله پنجم: تنظیم پارامتر؛ پارامترهای فنون استفاده شده بر اساس نرخ صحت و دقت پیش‌بینی تنظیم می‌شوند.

مرحله ششم: ارائه دانش فنون مختلفی که برای نمایش و بصری‌سازی دانش وجود دارد در این مرحله به کار گرفته می‌شود و دانش برای کاربران ارائه می‌گردد.

پیش‌پردازش

مطابق چارچوب پیشنهادی انجام داده‌کاوی، در مرحله اول برای مجموعه داده وضعیت بهداشت روان دانشجویان پیش‌پردازش انجام گردید. پیش‌پردازش داده اولین گام در فرآیند داده‌کاوی و یکی از گام‌های مهم آن به شمار می‌رود. معمولاً مجموعه‌داده‌هایی در دسترس، دارای سه نقص بزرگ وجود داده‌های پرت و خطا، داده‌های گمشده و وجود ناسازگاری بین داده‌ها هستند. به همین منظور فنونی ارائه شده تا با اجرای آن‌ها قبل از آغاز فرآیند داده‌کاوی از تولید خروجی‌های نامناسب تا حد امکان جلوگیری شود. این پیش‌پردازش شامل سه بخش یک‌پارچه‌سازی داده، پاک‌سازی و کاهش داده می‌باشد.

گاهی این امکان وجود دارد که مشخصه‌های مورد نظر، در فایل‌های مختلف و با فرمت‌های متفاوتی ذخیره شده باشند، در این شرایط برای نیل به یک مجموعه داده برای پیاده‌سازی فرآیند داده‌کاوی، انجام یک‌پارچه‌سازی ضروری می‌باشد. پاک‌سازی داده‌ها شامل چهار قسمت جایگذاری مقادیر گم‌شده، از بین بردن خطای داده، شناسایی و از بین بردن داده‌های پرت و برطرف کردن

ناسازگاری‌ها می‌باشد. در پژوهش حاضر، دو قسمت جایگذاری مقادیر گم‌شده و از بین بردن خطا داده مورد استفاده قرار گرفته است. وجود داده‌های گم‌شده در مجموعه داده، چالش‌های مهمی را به دنبال دارد که مطلوبیت پیش‌بینی نهایی را تحت تأثیر قرار می‌دهد. روش‌های مختلفی نیز برای برطرف کردن مشکل وجود داده گم‌شده مانند جایگزین کردن مقادیر گم‌شده با مد، میانه و میانگین، حذف رکورد مربوطه و جایگذاری آن‌ها با یک مقدار ثابت وجود دارد. هم‌چنین برای رفع مشکل خطای داده، می‌توان از انواع روش‌های ظرف‌بندی، خوشه‌بندی و رگرسیون استفاده کرد. مجموعه داده انتخابی، دارای مقادیر گم‌شده و تعدادی رکورد خطا می‌باشد که با توجه به ماهیت مجموعه داده، روش مناسب برای رفع مشکل مقادیر گم‌شده، جایگزینی آن‌ها با مقدار میانگین مشخصه مربوطه است.

در حجم بالای داده ممکن است بعضی از داده‌های غیرمفید هم وجود داشته باشد و نیازی به پردازش تمامی مجموعه داده نباشد. به همین دلیل برای کاهش حجم محاسبات و افزایش سرعت پردازش نیاز است تا مشخصه‌های غیرضروری که تأثیری در پیش‌بینی نهایی ندارند، حذف شود. در مجموعه داده مفروض نیز مشخصه‌هایی وجود داشتند که در نظر گرفتن آن‌ها با توجه به این که تمامی مقادیر آن یکسان بود، ضروری نمی‌باشد.

پیش‌پردازش برای تمامی فنون ذکر شده در شکل ۲، یکسان است و تنها در مقدار مشخصه لیبل (نیاز به مشاوره) تفاوت وجود دارد. تمامی مشخصه‌های موجود در مجموعه داده دارای طیف لیکرت هستند و در پیش‌پردازش نیز از اعداد همین طیف یعنی، ۱ الی ۵ استفاده شده است. به منظور مقداردهی لیبل برای دو روش اول یعنی درخت تصمیم و طبقه‌بندی بر اساس قانون از بله و خیر و برای دیگر فنون از ۰ و ۱ استفاده گردید.

مدل‌سازی

در الگوریتم‌های طبقه‌بندی مجموعه داده اولیه به دو مجموعه با عنوان‌های آموزشی و آزمایشی تقسیم می‌شود، با استفاده از مجموعه داده آموزشی مدل ساخته می‌شود و از مجموعه داده آزمایشی برای اعتبارسنجی و محاسبه دقت مدل ساخته شده استفاده می‌شود[۱۳].

الگوریتم‌های نظارتی شامل دو مرحله با عنوان مرحله آموزش (یادگیری) و مرحله ارزیابی هستند. در مرحله آموزش، مجموعه داده آموزشی به یکی از الگوریتم‌های دسته‌بندی اختصاص داده می‌شود. پس از ساخت مدل، ارزیابی آن توسط مجموعه داده آزمایشی انجام می‌شود؛ که این مجموعه داده آزمایشی در مدل ساخته شده، استفاده نشده‌اند. از مجموعه داده آزمایشی در مرحله آموزش و ساخت مدل استفاده نمی‌شود.

همان‌گونه که پیش از این اشاره شد، رویکردهای طبقه‌بندی و رگرسیون برای مجموعه داده حاضر مناسب می‌باشد. در این قسمت، شش روش داده‌کاوی بر روی مجموعه داده پیاده‌سازی و اجرا شده است که هر کدام دارای مزایا، معایب و مطلوبیت‌های متفاوتی می‌باشند. در ادامه فنون درخت تصمیم، طبقه‌بندی بر اساس قانون، شبکه عصبی، رگرسیون لجستیک، ماشین بردار پشتیبان و نزدیک‌ترین همسایه معرفی و پارامترهای آن تنظیم شده‌اند. فنون ذکرشده و تنظیم پارامترهای آن با استفاده از نرم‌افزار Rapid miner ۷.۱.۱ پردازش شده‌اند. از آن جا که لازمه محاسبه نرخ صحت و دقت مدل ارائه شده تقسیم مجموعه داده به دو قسمت داده آموزش و تست می‌باشد، مجموعه داده حاضر نیز توسط روش Hold-out[۳۰] با اختصاص ۰.۳ کل داده به داده آزمایش تقسیم شد.

درخت تصمیم

معروف‌ترین و شاید بتوان گفت جذاب‌ترین روش پیش‌بینی‌کننده رویکرد طبقه‌بندی، درخت تصمیم و الگوریتم‌های آن است. بیش‌تر محققین به خاطر سهولت کاربرد و فراگیر بودن آن از درخت تصمیم برای استخراج الگوهای پنهان مجموعه داده خود استفاده می‌کنند. خروجی این الگوریتم، در قالب یک درخت نمایش داده می‌شود که به راحتی می‌توان قوانین اگر- آن‌گاه را بر اساس آن به دست آورد. به طور کلی درخت تصمیم یک نوع فلوچارت غیر حلقوی است که شامل مجموعه‌ای از گره‌های داخلی مربوط به یک آزمون منطقی بر روی یک مشخصه و شاخه‌های متصل‌کننده که نشان‌دهنده خروجی آزمون بوده، می‌باشد. گره‌ها و شاخه‌ها یک مسیر را در درخت تصمیم ایجاد می‌کنند که در نهایت به برگ می‌رسد؛ منظور از گره برگ در درخت تصمیم همان مشخصه لیبل می‌باشد. هر گره در یک درخت تصمیم نشان‌دهنده زیرمجموعه‌ای از مجموعه داده است. یک درخت ایده‌آل، درختی است که تمامی عناصر آن مقدار یکسانی برای متغیر هدف داشته باشند[۵]. در ادامه به توضیح برخی از پارامترهای مربوط پرداخته خواهد شد.

به منظور محاسبه شاخص بهره اطلاعاتی، در ابتدا باید آنتروپی هر یک از مشخصه‌ها محاسبه شود. به طور کلی می‌توان گفت شاخص بهره اطلاعاتی، مجموعه‌ای از اطلاعات است که از طریق درک مقدار هر مشخصه حاصل می‌شود؛ که مقدار آن از اختلاف عدد آنتروپی قبل و بعد از تفکیک به دست می‌آید. منظور از آنتروپی نیز مقدار ناخالصی موجود در رکوردهای یک مشخصه است.

شاخص Gain-ratio [۳۰] نوع دیگری از شاخص بهره اطلاعاتی است. بهره اطلاعاتی هر مشخصه را از حیث یکنواختی و وسعت تنظیم نماید. به بیان دیگر این شاخص برای کاهش میزان اریبی مشخصه‌های چندمقداره مورد استفاده قرار می‌گیرد.

شاخص Gini به طور مستقیم میزان ناخالصی رکوردهای یک مشخصه را نمایش می‌دهد. هر چه مقدار آن به صفر نزدیک‌تر باشد، بیانگر خالص بودن رکوردهای مشخصه مورد نظر است.

به طور کلی منظور از هرس یک درخت تصمیم، حذف یک گره و تبدیل آن به برگ است. گاهی ممکن است با حذف یک سری از گره‌ها، نرخ خطا افزایش یابد، اما از ایجاد برخی از قانون‌های

ناکارآمد و بزرگ شدن بیش از اندازه مطلوب درخت جلوگیری می‌کند.

طبقه‌بندی بر اساس قانون

روش طبقه‌بندی بر اساس قانون، یکی از زمینه‌های یادگیری ماشین به شمار می‌رود که ایده اصلی آن، این است که یک سری قانون یا الگوهای پنهان نهفته در مجموعه‌ای از مشاهدات ارائه دهد. قوانین استخراج‌شده ممکن است یک مدل علمی کامل از مجموعه داده را تشکیل دهد و یا یک سری الگو با توجه به برخی از مشخصه‌ها ایجاد کند. به طور کلی یک قانون شامل دو قسمت اگر (مقدم) و آن‌گاه (تالی) می‌باشد. هم‌چنین از آن‌جا که این روش جزو رویکردهای یادگیری نظارتی به شمار می‌رود، غالباً در پژوهش‌هایی که هدف پیش‌بینی دارند مورد استفاده قرار می‌گیرد.

شبکه عصبی

شبکه عصبی یک الگوی ریاضی مبتنی بر سیستم زیستی موجودات زنده می‌باشد و روش شبکه‌های عصبی مصنوعی به دنبال یادگیری بر اساس ساختار یادگیری در موجودات زنده است. شبکه‌های عصبی با توجه به توانایی استنتاج معانی از داده پیچیده و مبهم، روشی قدرتمند برای استخراج الگوها و پیش‌بینی در داده‌کاوی می‌باشند.

در این شبکه‌ها ابتدا وزن‌هایی تصادفی برای ورودی‌ها تعیین می‌شود و توسط این وزن‌ها مقدار لایه‌های پنهان محاسبه می‌شود و نهایتا به گره انتهایی می‌رسیم. پس از محاسبه گره نهایی مقادیر اختلاف مقدار محاسبه شده با مقدار واقعی محاسبه شده و با توجه به مقدار اختلاف مقدار جریمه‌ای برای بهبود وزن‌های تصادفی اعمال شده در نظر گرفته می‌شود و دوباره به گره اول باز می‌گردد. این روند تا آن جایی ادامه دارد که یکی از دو شرط حداقل اختلاف و یا حداکثر تکرار برقرار شوند.

رگرسیون لجستیک

پیش‌بینی مقدار یک متغیر پیوسته بر اساس مقادیر سایر متغیرها بر مبنای یک مدل وابستگی خطی یا غیرخطی رگرسیون نامیده می‌شود. در واقع یک بردار $\vec{X}$ به عنوان ورودی وجود دارد که یک متغیر خروجی y نگاشته شده است. هدف، محاسبه y یا همان $F(\vec{X})$است که از روی تخمین تابع مقدار آن محاسبه می‌شود. در این جا می‌بایست به ازای یک بردار $\vec{X}$ ،مقدار دقیق y قابل محاسبه باشد.

روش رگرسیون یکی از پرکاربردترین روش‌های برای بررسی ارتباط بین متغیرها و پیش‌بینی به شمار می‌رود و انواع مختلفی بر اساس چندگانه بودن یا نبودن، خطی یا غیرخطی بودن و غیره دارد. پیش از این اشاره شد که یکی از اهداف داده‌کاوی، پیش‌بینی مقدار متغیر هدف یا مشخصه لیبل است؛ به همین منظور می‌توان از روش رگرسیون نیز استفاده کرد. وجه تفاوت رگرسیون و رگرسیون لجستیک در نوع مشخصه لیبل است. در رگرسیون معمولی، نوع مشخصه حتما باید از نوع عددی بوده اما با به‌کارگیری رگرسیون لجستیک می‌توان از داده‌های اسمی برای پیش‌بینی استفاده کرد.

ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از قدرتمندترین ابزارها برای امور طبقه‌بندی و پیش‌بینی به شمار می‌رود. رویکرد SVM به این صورت است که در مرحله آموزش، سعی دارد که مرز تصمیم‌گیری را به گونه‌ای انتخاب نماید که حداقل فاصله آن با هر یک از دسته‌های مورد نظر را بیشینه کند. این نوع انتخاب باعث می‌شود که تصمیم‌گیری در شرایطی که داده‌ها دارای پراکندگی بالا می‌باشند را به خوبی تحمل نموده و هم‌چنین پاسخ‌دهی مناسبی داشته باشد. این نحوه انتخاب مرز بر اساس نقاطی به نام بردارهای پشتیبان انجام می‌شود.

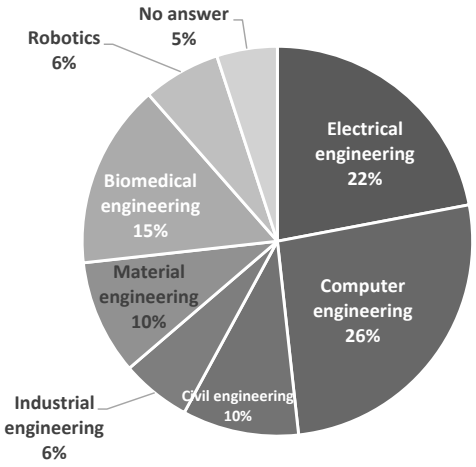
هم‌چنین الگوریتم‌های مبتنی بر ماشین‌های بردار پشتیبان الگوریتم‌هایی هستند که سعی می‌کنند یک حاشیه را بیشینه کنند. این الگوریتم‌ها برای پیدا کردن خط جدا کننده دسته‌ها، از دو خط موازی شروع کرده و این خطوط را در خلاف جهت یک‌دیگر حرکت می‌دهند تا هر کدام از خطوط بهینه به یک نمونه از یک دسته خاص در سمت خود برسد. پس از انجام این مرحله، میان دو خط موازی یک نوار یا حاشیه شکل می‌گیرد. هر چه پهنای این نوار بیش‌تر باشد، به این معناست که الگوریتم توانسته است حاشیه را بیشینه کند و هدف نیز بیشینه نمودن این حاشیه است[۱۴].

نتایج و بحث

آمار توصیفی

مشخصه‌های مورد بررسی، در دو بخش کلی شامل مشخصات عمومی دانشجویان و مشخصه‌های سلامت روان آن‌ها قرار گرفته است. در مورد مشخصات عمومی دانشجویان با توجه به مجموعه داده مورد نظر، ۵۸.۸٪ آن‌ها پسر و ۴۱.۵٪ دختر بوده است. از نظر وضعیت تأهل، ۴.۸٪ متأهل و ۹۲.۵٪ مجرد و ۲.۷٪ بدون و پراکندگی سنی دانشجویان، از ۱۷ سال تا ۳۳ و میانگین سنی نیز ۱۹ سال بوده است. پراکندگی سنی دانشجویان، از ۱۷ سال تا ۳۳ و میانگین سنی نیز ۱۹ سال بوده است. فراوانی دانشجویان بر اساس رشته تحصیلی در شکل ۳ نمایش داده شده است. بر اساس مقطع تحصیلی، ۸۱.۵٪ دانشجویان، کارشناسی، ۱۱.۲٪ کاردانی و ۵.۳٪ کارشناسی ارشد و ۲.۱٪ نیز بدون پاسخ بوده است. ۹۱.۴٪ دانشجویان با خانواده زندگی می‌کنند و سایر افراد در خوابگاه‌های دولتی و و خانه اجاره‌ای ساکن هستند.

بخش دوم مشخصه‌ها، مربوط به سلامت روان دانشجویان است؛ که شامل ۹۳ مشخصه است. پرسش‌نامه ارائه شده به دانشجویان شامل



شکل ۳: فراوانی دانشجویان بر اساس رشته تحصیلی
Fig. 3: Frequency of students by field of study

۳ بخش می‌شود. ابعادی که در بخش اول مورد ارزیابی قرار می‌گیرد از آن‌ها به عنوان ابعاد سلامت‌نگر نام برده شده و شامل عواطف مثبت و بهزیستی روانی می‌شود. بخش دوم پرسش‌نامه که مربوط به احساسات، نگرش‌ها و رفتار یک ماه اخیر دانشجویان می‌باشد بعد بیماری‌نگر نام دارد و شامل این موارد می‌گردد: افسردگی، اضطراب، وسواس، اضطراب اجتماعی، اختلال خواب، افسردگی تحصیلی، اضطراب تحصیلی و تمایلات خودکشی. در بخش سوم نیز سایر مشخصه‌های روانی دانشجویان تحت عنوان سازه‌های سلامت روان شامل این موارد ارزیابی شده است: کمال‌گرایی، جو خانوادگی، جهت‌گیری مذهبی، خودکارآمدی و حمایت اجتماعی.

تنظیم پارامتر

به منظور دسترسی به بهترین سطح عملکرد یک مدل، یکی از اقداماتی که ضروری است مورد نظر قرار گیرد، تنظیم پارامترهای هر روش است. به طور کلی می‌توان گفت منظور از تنظیم پارامتر به‌کارگیری الگوریتمی است تا بتوان به تعداد بهینه پارامتر مورد نظر دست یافت. به عنوان مثال در روش درخت تصمیم، یکی از مهم‌ترین پارامترهایی که تعداد آن باید برای کاربر مشخص باشد، تعداد شاخه‌هاست. بر اساس نرخ صحت به‌دست‌آمده می‌توان تعداد بهینه پارامتر مورد نظر را محاسبه و در نظر گرفت. منظور از نرخ صحت میزان مطابقت شماری از رکوردهای پیش‌بینی‌شده در مقابل کل رکوردهاست. بدیهی است هرچه درصد بالاتری داشته باشد، مطلوب‌تر است. بنابراین به صورت خاص در روش درخت تصمیم، همواره مشخصه‌ای برای شاخه زدن انتخاب می‌شود که بیشترین نرخ صحت را برای کل درخت ایجاد نماید.

همان‌طور که ذکر شد با استفاده از شاخص نرخ صحت، پارامترها تنظیم گردید و نتایج تنظیم پارامتر برای فنون طبقه‌بندی بر اساس قانون و درخت تصمیم به ترتیب در جداول ۲ و ۳ مشاهده می‌شود. با توجه به نرخ صحت به‌دست آمده از شاخص‌های مورد نظر، شاخص accuracy با نرخ صحتی برابر ۹۵٪، شاخص بهتری برای

جدول ۲: پارامترهای تنظیم‌شده برای طبقه‌بندی بر اساس قانون
Table 2: Tuned parameters for rule based data mining

Creation	Sample ratio	Pureness	Accuracy
Accuracy	0.4	0.9	0.95
Information_gain	0.3	0.8	0.94

جدول ۳: نتایج تنظیم پارامترهای روش درخت تصمیم
Table 3: Result of parameter tuning for decision tree

Creation	Apply pruning	Confidence	Apply prepruning	Accuracy
Accuracy	True	0.5	True	0.96
Gain_ratio	True	≈ 0	True	0.95
Gini_index	False	0.5	False	0.95
Information gain	True	0.5	True	0.97

ایجاد قوانین می‌باشد.

در شبکه‌های عصبی پارامتر مهم نرخ یادگیری می‌باشد که همانند روش‌های قبل این پارامتر نیز تنظیم شده است و با استفاده از نرخ یادگیری ۰،۰۴ به بیش‌ترین مقدار نرخ صحت که ۹۹٪ می‌باشد دست می‌یابیم. در شکل ۴ مقادیر مختلف نرخ صحت با توجه به نرخ یادگیری مشاهده می‌شود.

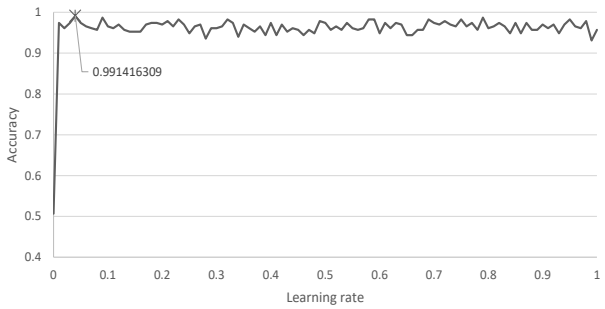
همانند روش‌های قبل در روش رگرسیون لجستیک نیز پارامترهای مربوطه تنظیم شده‌اند که همان‌طور در شکل ۵ مشاهده می‌شود مقدار بهینه برای پارامتر C، ۵۵۳ می‌باشد که نرخ صحت به‌دست آمده از طریق این مقدار ۹۹.۵۷٪ می‌باشد.

پارامتر C مقدار که در روش‌های رگرسیون لجستیک و ماشین بردار پشتیبان استفاده شده است، ثابت پیچیدگی SVM نام دارد که خطای دسته‌بندی را تعیین می‌نماید به این مفهوم که هر چه مقدار C بزرگ‌تر باشد امکان پدیده بیش‌برازش افزایش می‌یابد. مقادیر خیلی کم و خیلی زیاد برای پارامتر C باعث می‌شود تا برآورد مناسب و دقیقی به‌دست نیاید.

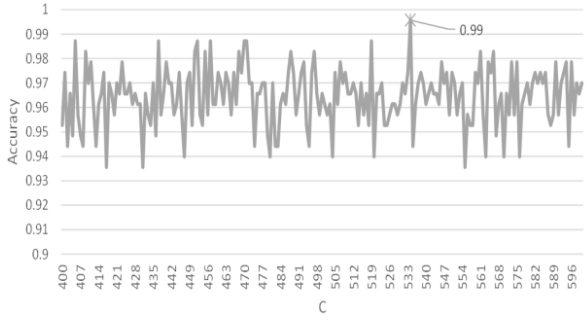
هم‌چنین با استفاده از روش SVM نیز سه مشخصه مهم که وزن آن‌ها بیش‌تر از ۰،۵ بود، مشخص شدند و شامل مشخصه‌های از یک ماه گذشته تا به امروز، احساس ناامیدی می‌کنم با وزن ۰.۷۶، به نظر دیگران من آدم وسواسی هستم با وزن ۰.۶۴ و زندگی ارزشمندی دارم با وزن ۰.۵۴ می‌باشد.

یافته‌های پژوهش

به طور کل تمامی روش‌های استفاده شده از نرخ صحت بالای ۹۰٪ برخوردار بود اما بهترین و بالاترین نرخ صحت مربوط به روش رگرسیون لجستیک با مقدار ۹۹.۵۷٪ بوده است و نتایج و ضرایب



شکل ۴: مقدار بهینه نرخ یادگیری
Fig. 4: Optimum value of learning rate



شکل ۵: مقدار بهینه پارامتر C
Fig. 5: Optimum value of C parameter

هر مشخصه به طور مفصل در بخش قبل بررسی گردیدند. درخت تصمیم و طبقه‌بندی بر اساس قانون ایجاد شده نیز به ترتیب دارای نرخ صحت ۹۷٪ و ۹۵٪ بودند. شکل ۶، نمودار مقایسه دقت و ضریب اطمینان فنون استفاده شده را نشان می‌دهد.

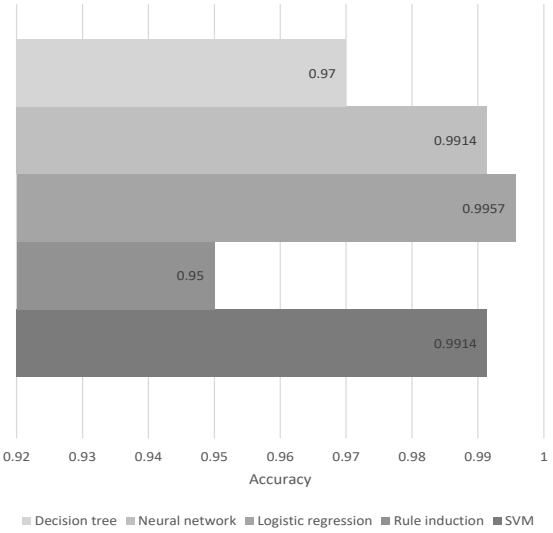
جذاب‌ترین روش پیش‌بینی‌کننده رویکرد طبقه‌بندی، درخت تصمیم و الگوریتم‌های آن است. بیش‌تر محققین به خاطر سهولت کاربرد و فراگیر بودن آن از درخت تصمیم برای استخراج الگوهای پنهان مجموعه داده خود استفاده می‌کنند. خروجی این الگوریتم، در قالب یک درخت نمایش داده می‌شود که به راحتی می‌توان قوانین اگر- آن‌گاه را بر اساس آن به دست آورد. نتیجه درخت تصمیم با توجه به طیف شش‌تایی پاسخ، به شرح زیر می‌باشد و در شکل ۷ مشاهده می‌شود.

• اگر سوال از یک ماه گذشته تا به امروز بدون دلیل در حین خواب وحشت‌زده می‌شوم، مقداری برابر ۴، ۵ یا ۶ بگیرد آن‌گاه:

• اگر سوال از یک ماه گذشته تا به امروز از صحبت در حضور جمع پرهیز می‌کنم، مقداری بیش‌تر از یک بگیرد به مشاوره نیاز دارد. در غیر این‌صورت:

• اگر سوال از یک ماه گذشته تا به امروز زندگی برایم بی‌ارزش است، مقداری برابر ۴، ۵ یا ۶ اتخاذ کند به مشاوره نیاز دارد و در غیر این‌صورت به مشاوره نیاز ندارد.

• اگر سوال از یک ماه گذشته تا امروز بدون دلیل در حین خواب وحشت‌زده می‌شوم، مقداری برابر ۴، ۵ یا ۶ بگیرد آن‌گاه:



شکل ۶: مقایسه ضریب اطمینان فنون
Fig. 6: Comparison of accuracy rate for techniques

• اگر سوال از یک ماه گذشته تا به امروز همیشه نگرانم، مقداری کم‌تر از ۷ اتخاذ کند، به مشاوره نیازی ندارد. در غیر این صورت:

• اگر سوال از یک ماه گذشته تا به امروز از صحبت در حضور جمع پرهیز می‌کنم مقداری برابر یک بگیرد به مشاوره نیاز ندارد. در غیر این‌صورت:

• اگر سوال از یک ماه گذشته تا به امروز احساس ناامیدی می‌کنم، مقداری بیش‌تر از یک بگیرد، آن‌گاه:

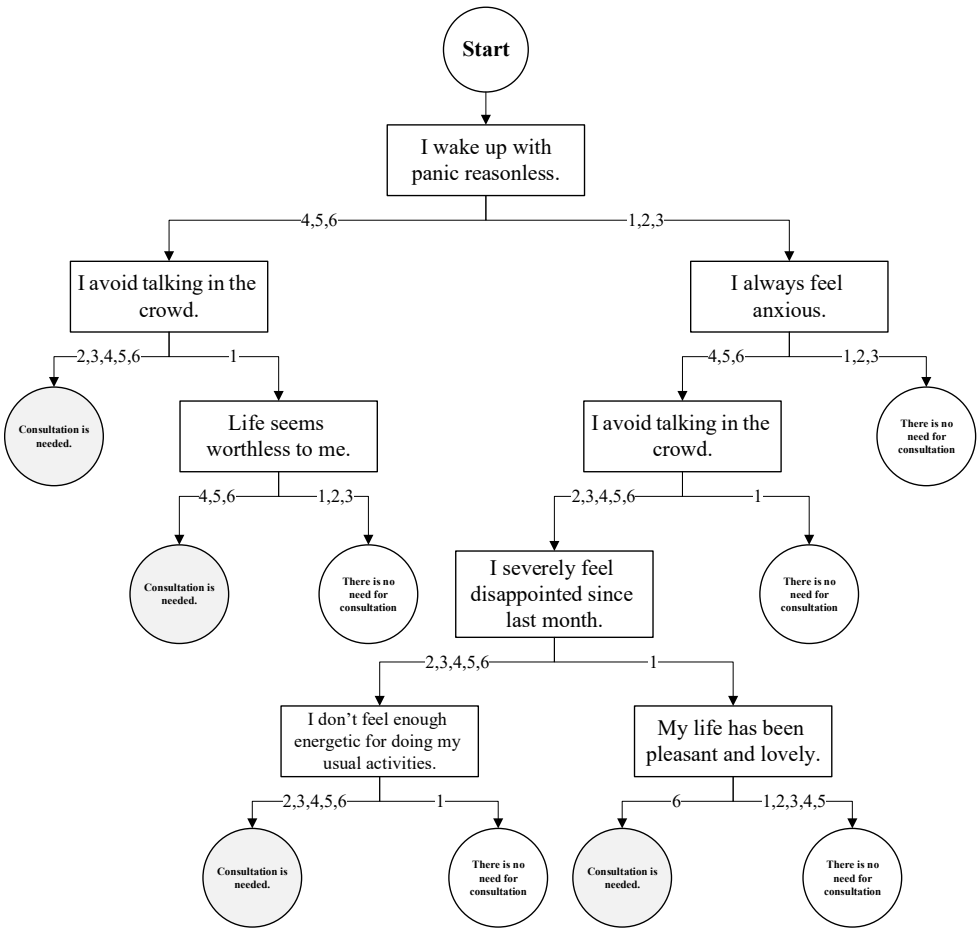
• اگر سوال حال و حوصله‌ی انجام فعالیت‌های معمولم را ندارم، مقداری بیش‌تر از یک بگیرد به مشاوره نیاز دارد در غیر صورت به مشاوره نیاز ندارد.

• اگر سوال از یک ماه گذشته تا احساس ناامیدی می‌کنم، مقداری برابر یک بگیرد، آن‌گاه:

• اگر سوال در یک ماه گذشته، زندگی‌ام پرشور و شوق بوده است، مقداری برابر ۶ بگیرد به مشاوره نیاز دارد در غیر صورت به مشاوره نیاز ندارد.

نتیجه‌گیری

رشد و توسعه حجم داده‌های ثبت‌شده در سیستم‌های اطلاعاتی، این نیاز را ایجاد کرده است که اطلاعاتی معنادار از میان این انبوه داده استخراج گردد. مراکز دانشگاهی و آموزشی نیز از این قاعده مستثنی نیستند و اطلاعاتی در زمینه‌های مختلف آموزشی، روانشناسی و غیره در این مراکز موجود می‌باشند. در این مطالعه نیز به بررسی اطلاعات مربوط به سلامت روان دانشجویان با هدف کاوش علائم نیاز به مشاوره پرداخته شد. برای کاوش، از مجموعه داده مربوط به پرسش‌نامه‌های تکمیل شده توسط دانشجویان استفاده گردیده است. با توجه به این که تحقیق حاضر از نوع کاربردی بوده و مجموعه داده واقعی مورد استفاده قرار گرفت، یکی از چالش‌برانگیزترین محدودیت‌ها، گردآوری مجموعه داده بوده و علت آن هم مسائل امنیتی سازمان در انتشار اطلاعات می‌باشد؛



شکل ۷: درخت تصمیم

Fig. 7: Decision tree

تأثیر غیرقابل اجتناب سلامت روانی دانشجویان بر عملکرد تحصیلی، موضوعی است که مراکز آموزشی همواره باید در نظر داشته باشند. یکی از فوایدی که اجرای چنین پایش‌هایی به همراه دارد، امکان استخراج الگو و پیش‌بینی وضعیت تحصیلی دانشجویان بر اساس سوابق سلامت روان آن‌ها می‌باشد. پایش سلامت روان دانشجویان با هدف بهره‌برداری از مقیاس ملی سلامت روانی دانشجویان از سال ۱۳۸۵ در دانشگاه‌های سراسر کشور تحت نظارت دفتر مشاوره و سلامت سازمان امور دانشجویان در وزارت علوم، تحقیقات و فناوری اجرا شده است. بنابراین با توجه به این که چنین فرآیندی در تمامی مراکز آموزشی پیاده‌سازی می‌شود، اجرای تحقیقاتی مشابه تحقیق حاضر به منظور بهره‌گیری از تلاش مشاورین مراکز آموزشی در راستای رسیدن به نتایج کاربردی و الگوهای معتبر، ضروری به نظر می‌رسد. پیش از کاوش، برای افزایش دقت و بهبود پیش‌بینی‌های انجام شده، عملیات پیش‌پردازش بر روی مجموعه داده انجام و برای یافتن الگوهای پنهان و پیش‌بینی نیاز به مشاوره از نرم‌افزار Rapid miner ۷.۱.۱ استفاده شده است. با توجه به هدف مطالعه که پیش‌بینی نیاز به مشاوره می‌باشد از فنون یادگیری نظارتی و طبقه‌بندی شامل درخت تصمیم، نزدیک‌ترین همسایه، رگرسیون لجستیک، شبکه عصبی مصنوعی، ماشین بردار ویژه و طبقه‌بندی بر پایه قانون استفاده شده است. برای هر یک از روش‌های مذکور مقدار بهینه پارمترها با توجه به نرخ صحت به‌دست آمده و بر اساس مقدار بهینه پیش‌بینی صورت پذیرفته است. دقت پیش‌بینی تمام فنون بیش‌تر از ۰.۹ بوده است و هم‌چنین بیش‌ترین دقت مربوط به روش رگرسیون لجستیک با دقت ۹۹.۵۷٪ بوده است. نتایج نشان می‌دهند که اگر فردی از یک ماه گذشته تا به امروز شدیداً، احساس ناامیدی کند، یا به نظر اطرافیانش فردی وسواسی باشد یا احساس کند زندگی برایش بی‌ارزش است به مشاوره احتیاج دارد.

این مطالعه تنها به مطالعه بعد روانی سلامت پرداخته و علت عدم تحقیق گسترده‌تر در سایر ابعاد، وجود سیاست‌های محرمانگی داده و در دسترس نبودن اطلاعات برای دیگر ابعاد موجود است. اصل محرمانگی در رابطه با چاپ، انتقال، ذخیره‌سازی و افشای اطلاعات پرونده‌ی دانشجویان و عدم در اختیار گذاشتن اطلاعات حتی با وجود شیوه‌های رمزنگاری داده (کدگذاری داده)، از اصلی‌ترین محدودیت‌های پژوهش حاضر است. هم‌چنین در مواردی مانند داده تغذیه، کتابخانه و امور دانشجویی، بر خلاف وجود سیستم‌های

اطلاعاتی، داده موجود ناقص و غیر قابل استفاده بوده است. با توجه به این که اکثر پژوهش‌های موجود در زمینه داده‌کاوی سلامت، تمرکز بر سلامت جسمانی داشته‌اند، پیشنهاد می‌شود برای مطالعات آتی تمامی سطوح سلامت یعنی ابعاد سلامت دانشجویان شامل سلامت جسمانی، اجتماعی و معنوی و هم‌چنین ترکیبی از این ابعاد مورد بررسی قرار گیرد. علاوه بر این مطالعه‌ای مروری بر روی انواع رویکردها و فنون مناسب برای مجموعه داده‌های روان‌شناسی با هدف ایجاد یک تقسیم‌بندی مناسب برای فنون موجود در این حوزه انجام شود؛ هم‌چنین پیشنهاد می‌شود، مجموعه داده حاضر و یا مجموعه داده‌های مشابه (اطلاعات پایش سلامت دانشجویان) با فنون دیگر طبقه‌بندی مورد بررسی قرار گرفته و نتایج حاصل با نتایج پژوهش حاضر مقایسه گردد. به طور کلی پیشنهاد می‌شود از فن داده‌کاوی برای استخراج الگوهای پنهان در مجموعه داده سلامت روان دانش‌آموزان مدارس در مقاطع تحصیلی متفاوت، کارمندان ادارات و سازمان‌ها استفاده گردد. در نهایت توصیه می‌گردد پژوهش‌های آتی در این زمینه ابتدا رویکرد خوشه‌بندی را بر روی مجموعه داده روان‌شناسی پیاده کنند و به دنبال آن از رویکردهای طبقه‌بندی و پیش‌بینی استفاده نمایند.

منابع و مآخذ

[1] Costa EB, Fonseca B, Santana MA, de Araújo FF, Rego J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*. 2017; 73: 247-256.

[2] Angeli C, Howard SK, Ma J, Yang J, Kirschner PA. Data mining in educational technology classroom research: Can it make a contribution. *Computers & Education*. 2017; 113: 226-242.

[3] Bali RK. (Ed.). *Clinical knowledge management: opportunities and challenges*. IGI Global. Published in the United State of America by Idea Group Publishing; 2005.

[4] Moghadasi H, Hosseini A, Asadi F, Jahanbakhsh M. Data mining and health care. *Health Information Management Journal*. 2010; 9(2): 297-304. Persian.

[5] Asif R, Merceron A, Ali SA, Haider NG. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. 2017; 113: 177-194.

[6] Bhardwaj BK, Pal S. Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv*. 2012; 1201: 34-18.

[7] Alfiani AP, Wulandari FA. Mapping Student's Performance Based on Data Mining Approach (A Case Study). *Agriculture and Agricultural Science Procedia*. 2015; 3: 173-177.

[8] Ilayaraja M, Meyyappan T. Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets. *Procedia Computer Science*. 2015; 70: 586-592.

[9] Chang CL. A study of applying data mining to early intervention for developmentally-delayed children. *Expert Systems with Applications*. 2007; 33(2): 407-412.

[10] Choo C, Diederich J, Song I, Ho R. Cluster analysis reveals risk factors for repeated suicide attempts in a multi-ethnic Asian population. *Asian Journal of Psychiatry*. 2014; 8: 38-42.

[11] Paramasivam, V, Yee TS, Dhillon SK, Sidhu AS. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. *Biocybernetics and Biomedical Engineering*. 2014; 34(3): 139-145.

[12] Torkestani MS, Dehpanah A, Taghavifard MT, Shafee SH. A framework for modifying the insurance rate in automobile industry by Neural Network (case: Asia insurance company). *Journal of Information Technology Management*. 2015; 8(4):711-732. Persian.

[13] Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

[14] Burgos, C., Campanario, M. L., de la Pena, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering,* 66, 541-556.

[15] Ahmed ABED, Elaraby IS. Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*. 2014; 2(2): 43-47.

[16] Natek S, Zwilling M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications*. 2014; 41(14), 6400-6407.

[17] Peral J, Maté A, Marco M. Application of Data Mining techniques to identify relevant Key Performance Indicators. *Computer Standards & Interfaces*. 2017; 50: 55-64.

[18] Ahmed AM, Rizaner A, Ulusoy AH. (2016). Using data mining to predict instructor performance. *Procedia Computer Science*. 2016; 102: 137-142.

[19] Gobert JD, Kim YJ, Sao Pedro MA, Kennedy M, Betts CG. Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*. 2015; 18: 81-90.

[20] Kaur P, Singh M, JosanG.S. Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*. 2015; 57: 500-508.

[21] Diederich J, Al-Ajmi A, Yellowlees P. E x-ray: data mining and mental health. *Applied Soft Computing*. 2007; 7(3): 923-928.

[22] Fernández-Arteaga V, Tovilla-Zárate CA, Fresán A, González-Castro TB, Juárez-Rojop IE, López-Narváez L, Hernández-Díaz Y. Association between completed suicide and environmental temperature in a Mexican population, using the Knowledge Discovery in Database approach. *Computer Methods and Programs in Biomedicine*. 2016; 135: 219-224.

[27] Ni H, Yang X, Fang C, Guo Y, Xu M, He Y. (2014). Data mining-based study on sub-mentally healthy state among residents in eight provinces and cities in China. *Journal of Traditional Chinese Medicine*. 2014; 34(4): 511-517.

[28] Rostami M, Ayat SS, Saghari F, Yaghoobi F. (2014). Prediction of educational progress with fuzzy clustering in educational centers. *Journal of Educational Technology*. 2014; 10(1), 23-36. Persian.

[29] Maghsoodi B, Soleilmani S, Amiri A, Afsharchi M. Improvement in the quality of electronic educational systems using educational data maining. *Journal of Educational Technology*. 2012; 6(4): 277-286. Persian.

[30] Kohavi R, Quinlan JR. Data mining tasks and methods: Classification: decision-tree discovery. Willi Klösgen, Willi Klosgen, Jan M. Żytkow, (Eds), In *Handbook of data mining and knowledge discovery* (pp. 267-276). Oxford University Press, Inc; 2002.

[23] Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*. 2013; 25(2): 127-136.

[24] Ramanan S, de Souza LC, Moreau N, Sarazin M, Teixeira AL, Allen Z, Bertoux M. Determinants of theory of mind performance in Alzheimer's disease: A data-mining study. *Cortex*. 2017; 88: 8-18.

[25] Nguyên XL, Chaskalovic J, Rakotonanahary D, Fleury B. Insomnia symptoms and CPAP compliance in OSAS patients: A descriptive study using Data Mining methods. *Sleep Medicine*. 2010; 11(8): 777-784.

[26] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*. 2017; 15: 104-116.

Citation: (Vancoure): Koosha H, Dangkoub S, Barzanooni A.A. [Application of data mining techniques to predict students' mental health status to improve educational performance]. *Tech. Edu. J.* 2019; 13(1): 49-62.

http://dx.doi.org/10.22061/jte.2018.3075.1779