مجله محاسبات نرم، جلد 14، شماره 2 (پاییز و زمستان 1404)، صفحه 122_137
https://doi.org/10.22052/scj.2025.255049.1257

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

University of Kashan

بررسی عوامل موثر بر موفقیت دانش‌آموزان پایه دهم با استفاده از روش‌های داده‌کاوی: مطالعه موردی آموزشگاه‌های شهر کاشمر ✧

اعظم علیپور فرگی[1]، کارشناسی ارشد، حمید سعادت‌فر[1*]، دانشیار

[1] گروه مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت 28 تیر ماه 1403
پذیرش 10 اسفند ماه 1403

کلمات کلیدی:
داده‌کاوی
انتخاب ویژگی
پیش‌بینی معدل
چرخ تعادل زندگی
موفقیت تحصیلی

چکیده

موفقیت دانش‌آموزان در هر پایه تحصیلی یکی از اهداف مهم و قابل توجه در وزارت آموزش و پرورش است. در این میان، انجام برنامه‌ریزی مناسب برای موفقیت تحصیلی دانش‌آموزان پایه دهم، آن هم به دلیل انتخاب رشته تحصیلی در این پایه، از اهمیت بیشتری برخوردار است. همین‌طور تصمیم‌گیری صحیح، مستلزم گردآوری اطلاعات جامعی از تمام ابعاد زندگی یک دانش‌آموز است. از اینرو ابزار چرخ تعادل زندگی در شش بخش جسمی، مالی، فکری، عاطفی، اجتماعی و معنوی می‌تواند راهکار مناسبی برای این هدف باشد. در این مقاله ابتدا مجموعه داده‌ای با کمک اطلاعات چرخ تعادل زندگی، جمعیت‌شناختی و پیشینه آموزشی دانش‌آموزان پایه دهم آموزشگاه‌های شهر کاشمر از طریق پرسش‌نامه ایجاد می‌شود. در ادامه با استفاده از روش‌های بسته‌بندی و فیلتر، ویژگی‌های موثر انتخاب و الگوریتم‌های شبکه عصبی پرسپترون چندلایه، درخت تصمیم J48، جنگل تصادفی ، طبقه‌بند بیز ساده، طبقه‌بندی SVM و KNN با هدف پیش‌بینی موفقیت دانش‌آموزان در پایه دهم مورد آموزش قرار می‌گیرند. مقایسه نتایج نشان می‌دهد که الگوریتم بسته‌بندی توانسته ویژگی‌های تاثیرگذارتری را نسبت به سایر الگوریتم‌ها انتخاب کند به طوری که ویژگی‌های پول‌توجیبی، میزان مطالعه، جنسیت، رشته تحصیلی، تحصیلات پدر، تحصیلات مادر و بعد روحی بیشترین تاثیر را دارند. همین‌طور الگوریتم پرسپترون چندلایه با دقت 85/62 درصد و الگوریتم بیز ساده با دقت 86/25 درصد، به ترتیب قبل و بعد از فرآیند انتخاب ویژگی بهترین عملکرد را از خود نشان داده‌اند.

1. مقدمه

امروزه با توجه به پیشرفت‌های چشمگیر در زمینه ذخیره‌سازی و پردازش حجم زیادی از داده‌ها، نقش داده‌کاوی بیش از پیش پررنگ‌تر شده است [1]. به طوری که محققان از داده‌کاوی برای استخراج دانش و اطلاعات، کشف الگوهای پنهان از مجموعه داده‌های بسیار بزرگ و پیچیده [2]، در حوزه‌های مختلف مانند بازاریابی، ارتباط با مشتری [3]، مخابرات، صنعت، امور بهداشتی و پزشکی [4]، استفاده می‌کنند. یکی دیگر از کاربردهای گسترده داده‌کاوی، در زمینه آموزشی است [5]. در این میان می‌توان به کارهایی همچون پیش‌بینی موفقیت دانش‌آموزان [6]، بررسی

✧ نوع مقاله: پژوهشی
* نویسنده مسئول
پست(های) الکترونیک: azamalipour@birjand.ac.ir (علیپور فرگی)
saadatfar@birjand.ac.ir (سعادت‌فر)

نحوه ارجاع به مقاله: اعظم علیپور فرگی، حمید سعادت‌فر، «بررسی عوامل موثر بر موفقیت دانش‌آموزان پایه دهم با استفاده از روش‌های داده‌کاوی: مطالعه موردی آموزشگاه‌های شهر کاشمر»، مجله محاسبات نرم، جلد 14، شماره 2، ص122_137، پاییز و زمستان 1404.

عملکرد دانش‌آموزان [7]، [8] و همچنین بررسی عوامل موثر در موفقیت دانش‌آموزان اشاره کرد. با استفاده از نتایج پیش‌بینی‌های انجام شده می‌توان برنامه‌ریزی‌های لازم برای افزایش کیفیت آموزشی مدارس [9]، شناسایی نقاط قوت و نقاط ضعف دانش‌آموزان [10]، را انجام داد و برای رسیدن به موفقیت در جهت درستی گام برداشت.

از طرفی دیگر، وجود مجموعه داده مناسب در زمینه داده‌کاوی حوزه آموزش از اهمیت بالایی برخوردار است. برای این منظور محققان ویژگی‌های مختلفی نظیر جمعیت‌شناختی [11]، [12]، تحصیلی [10]، [13]، خانوادگی [11]، [14]، اجتماعی [11]، [14] و غیره را مورد پژوهش قرار داده و برای پیش‌بینی‌های خود استفاده می‌کنند.

همین‌طور روش‌های مختلفی برای تجزیه و تحلیل و پردازش داده‌های جمع‌آوری شده در داده‌کاوی وجود دارد که از جمله می‌توان به خوشه‌بندی، طبقه‌بندی و استخراج قوانین اشاره کرد [15]. با بررسی‌های انجام شده می‌توان متوجه شد که روش طبقه‌بندی در اکثر مطالعات انجام شده مورد استفاده قرار گرفته است [13]. در این روش یک نمونه یا یک رکورد به کلاس‌های از پیش‌تعریف‌شده نسبت داده می‌شود و برای این منظور از الگوریتم‌هایی نظیر درخت تصمیم، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان و نزدیک‌ترین همسایه استفاده می‌کند [12].

در این مقاله هدف بررسی عوامل موثر بر موفقیت دانش‌آموزان پایه دهم آموزشگاه‌های شهر کاشمر با استفاده از الگوریتم‌های داده‌کاوی است. همین‌طور برای جمع‌آوری داده‌های مورد نیاز از پرسش‌نامه‌ای حاوی اطلاعات چرخ تعادل زندگی، جمعیت-شناختی و پیشینه تحصیلی دانش‌آموزان استفاده شده است. در نهایت، تجزیه و تحلیل داده‌ها و ساخت مدل‌ها با الگوریتم‌های بیز ساده[1]، پرسپترون چندلایه[2]، SVM[3]، KNN[4]، درخت تصمیم J48 و جنگل تصادفی[5] در محیط برنامه WEKA انجام می‌شود.

نوآوری‌های این مقاله به شرح زیر است:

- این پژوهش با بهره‌گیری از مدل چرخ تعادل زندگی، تلاش کرده است تا ابعاد مختلف زندگی دانش‌آموزان (از جمله وضعیت خانوادگی، اقتصادی، تحصیلی و روانی) را در یک چارچوب منسجم بررسی نماید و از این طریق، تصویری چندوجهی از عوامل تاثیرگذار بر موفقیت تحصیلی ارائه دهد.
- استفاده از چهار روش انتخاب ویژگی که جزو دسته‌های رپر و فیلتر هستند برای شناسایی مهم‌ترین عوامل تاثیرگذار بر موفقیت تحصیلی.استفاده از رویکرد یادگیری ماشین کمک می‌کند تا تاثیر متقابل یا هم‌افزایی ویژگی‌ها به خصوص در مسائل پیچیده بهتر کشف شود.
- پیش‌بینی معدل پایانی دانش‌آموزان پایه دهم با بهره‌گیری از ویژگی‌های استخراج شده و مدل‌های متداول یادگیری ماشین.

این مقاله از قسمت‌های زیر تشکیل شده است. بخش 2 مروری بر ادبیات داده‌کاوی تحصیلی ارائه می‌کند. بخش 3 به مروری بر تحقیقات مرتبط با موضوع می‌پردازد. بخش 4 و 5 به ترتیب روش مورد استفاده و نتایج را بیان می‌کند. در نهایت، بخش 6 نتیجه‌گیری و تفسیر نتایج را نشان می‌دهد.

2. داده‌کاوی تحصیلی

فرآیند تبدیل داده‌های خام تحصیلی دانش‌آموزان به اطلاعات مفید که می‌تواند تاثیر زیادی بر تحقیقات و شیوه‌های آموزش دانش‌آموزان و محصلان داشته باشد، داده‌کاوی تحصیلی گفته می‌شود [16]. در این فرآیند ابتدا داده‌های لازم از سیستم‌های آموزشی دریافت و سپس از روش‌های داده‌کاوی برای استخراج دانش، کشف عوامل موثر بر عملکرد، در مواردی پیش‌بینی برخی پارامترهای مهم و در کل بهبود کیفیت آموزشی استفاده می‌شود [17]. داده‌کاوی در حوزه آموزش و پرورش، بیشتر بر تجزیه و تحلیل فرآیند یادگیری و رفتار دانش‌آموزان، استراتژی‌ها و برنامه‌های آموزشی، پیش‌بینی عملکرد، موفقیت و ترک تحصیل

[1] Naïve Bayes
[2] Multilayer Perceptron
[3] Support vector machine
[4] K-nearest neighbors
[5] Random Forest

دانش‌آموزان تمرکز دارد [18] که در این میان پیش‌بینی عملکرد تحصیلی دانش‌آموزان از اهمیت ویژه‌ای برخوردار است.

از طرفی دیگر، عملکرد تحصیلی دانش‌آموزان متکی به عوامل متعددی مانند عوامل اجتماعی، اقتصادی، فردی، آموزشی و غیره است. در نظر گرفتن این عوامل در آموزش ممکن است عملکرد دانش‌آموزان را در مباحث آموزشی به سرعت بهبود بخشد [19]. از همین رو در این زمینه می‌توان از داده‌کاوی تحصیلی استفاده و الگوهای پنهانی را برای پیش‌بینی عملکرد تحصیلی دانش‌آموزان استخراج کرد. نتایج این پیش‌بینی می‌تواند در انجام اقدامات اولیه برای دانش‌آموزان در معرض خطر و همین‌طور بررسی عملکرد مربیان نیز مفید باشد [10].

3. تحقیقات مرتبط

با نگاهی به مطالعات انجام شده می‌توان دریافت که محققان در حوزه‌های مختلفی از داده‌کاوی تحصیلی استفاده کرده‌اند. به طوری که مراحل کار آنها، شامل جمع‌آوری داده‌ها، پیش‌پردازش، مدل‌سازی و ارزیابی و تفسیر داده‌ها است.

با نگاهی کلی داده‌کاوی با جمع‌آوری داده شروع می‌شود و به همین دلیل محققان از ویژگی‌های مختلفی استفاده کرده‌اند که آنها را می‌توان در گروه‌های اطلاعات جمعیت‌شناختی [11]، [12]، پیشینه خانوادگی [11]، [14]، اطلاعات تحصیلی [10]، [13]، ویژگی‌های رفتاری [12]، [20] و فعالیت‌های اجتماعی [11]، [14] طبقه‌بندی کرد. در این بین محققان از پرسش‌نامه [14]، [21]، بانک اطلاعاتی سیستم‌های دانش‌آموزی و داده‌های سیستم مدیریت یادگیری در زمینه جمع‌آوری داده‌های گفته شده استفاده می‌کنند.

بعد از جمع‌آوری داده‌ها لازم است تا با انجام پیش‌پردازش، داده‌ها برای بکارگیری در الگوریتم‌های داده‌کاوی آماده‌سازی شوند [18]. در مرحله پیش‌پردازش ابتدا می‌بایست به پاک‌سازی داده‌ها پرداخت. این فرآیند عموما شامل حذف داده‌های گمشده یا ناقص است که به طور معمول در همه پژوهش‌ها دیده می‌شود. حذف ویژگی‌های نامرتبط از دیگر روش‌های پیش‌پردازش است. در این خصوص محققین ویژگی‌هایی نظیر نام دانشجو یا دانش‌آموز، ملیت، نام دانشگاه [10]، نام شهر، معدل ترم [13]، دروس انتخابی [8] و جنسیت [22]، را از بین سایر ویژگی‌ها بر اساس هدف مورد مطالعه خود حذف می‌کنند.

در گام بعد از پیش‌پردازش، سعی می‌شود تا با انتخاب ویژگی‌های موثر، ابعاد فضای ویژگی کاهش پیدا کند. در واقع فضای ویژگی اشاره به تعداد ویژگی‌های نمونه‌ها به جز ویژگی هدف اشاره دارد. کاهش ابعاد ویژگی‌ها باعث می‌شود تا پیچیدگی‌های محاسباتی کاهش و کارایی الگوریتم‌های داده‌کاوی افزایش پیدا کند. از طرفی دیگر تفسیرپذیری نتایج آسان‌تر و از پیش‌برازش جلوگیری می‌شود [14]. در این مرحله، انتخاب ویژگی می‌تواند بر اساس همبستگی [18]، [23]، فیلتر کردن و رتبه‌بندی [14]، [24] انجام شود. همین‌طور در مطالعات [6]، [7]، [15] و [18] از روش بسته‌بندی و در مطالعات [1]، [18] و [19] از الگوریتم ژنتیک نیز استفاده شده است.

در بعضی از مقالات بعد از آنکه داده‌های نامرتبط حذف و ویژگی‌های مرتبط انتخاب شده‌اند، از روش گسسته‌سازی و دسته‌بندی ویژگی‌ها [7]، [14]، [15]، [19]، نرمال‌سازی [19]، [23]، جایگزینی [8]، [13]، تبدیل ویژگی‌های عددی به اسمی [19]، تبدیل ویژگی‌های اسمی به عددی [25] و کدگذاری [26] استفاده کرده‌اند. از روش نمونه‌گیری نیز در برخی مطالعات استفاده شده است [18].

پس از جمع‌آوری داده‌ها و پیش‌پردازش، نوبت به انتخاب روش داده‌کاوی می‌رسد تا در نهایت مدل مورد نظر ایجاد شود. در این مرحله روش‌های داده‌کاوی روی مجموعه داده‌های مرحله قبل برای ساخت مدل و استخراج الگوهای مهم و جالب توجه بکار می‌روند. درخت تصمیم یکی از روش‌هایی است که در این زمینه به دلیل درک و تفسیر آسان، شناسایی سریع متغیرهای تاثیرگذار و روابط بین آنها مورد استفاده قرار می‌گیرد [27]. در مطالعات [2] و [23] از روش‌های درخت تصمیم به ترتیب روی اطلاعات 133 و 175 دانش‌آموز استفاده شده است و نتایج آنها نشان می‌دهد که درخت تصمیم با دقتی بین 65 تا 75 درصد، عملکرد بهتری دارد. در مطالعه [22] نیز تکنیک‌های مدل‌سازی KNN، بیز ساده، درخت تصمیم و رگرسیون لجستیک بکار

گرفته شده است و نتایج نشان می‌دهد بیز ساده با دقت 89 درصد در مقایسه با سایر الگوریتم‌ها، بهتر عمل می‌کند.

از سویی دیگر، از الگوریتم‌های بیز ساده، پرسپترون چندلایه، درخت تصمیم J48 و جنگل تصادفی برای سنجش میزان اهمیت ویژگی‌های سابقه تحصیلی، فعالیت‌های اجتماعی و پیشرفت تحصیلی در بین 395 دانش‌آموز استفاده شده است [28]. در این میان نتایج نشان می‌دهد که سابقه تحصیلی و فعالیت‌های اجتماعی در پیش‌بینی عملکرد دانش‌آموز بیشترین تاثیر را داشته‌اند.

همچنین مطالعات اخیر نشان می‌دهد که الگوریتم‌های شبکه عصبی، نتایج خوبی در پیش‌بینی عملکرد دانش‌آموزان از خود نشان داده‌اند [27]. به طوری که در مرجع [26] با اجرای الگوریتم‌های شبکه عصبی روی اطلاعات 3518 دانش‌آموز، دقت پیش‌بینی عملکرد 80/47 درصد ثبت شده است. همین‌طور در مرجع [24] پیش‌بینی عملکرد با دقتی بالای 91 درصد مشاهده می‌شود. در مرجع [19] نوع دیگری از شبکه‌های عصبی به نام شبکه‌های عصبی کانولوشنی در کنار الگوریتم K-نزدیک‌ترین همسایه، بیز ساده و درخت‌های تصمیم (C4.5) استفاده شده است. در این تحقیق از الگوریتم ژنتیک دودویی به عنوان یک رویکرد انتخاب ویژگی استفاده شده است و نتایج نشان می‌دهد که عملکرد همه طبقه‌بندی‌کننده‌ها به جز طبقه‌بندی بیز ساده پس از اعمال الگوریتم BGA[1]، 2 الی 3 درصد بهبود می‌یابد. لازم به ذکر است که عملکرد شبکه‌های عصبی کانولوشنی با دقت 95 درصد از همه روش‌های گفته شده بهتر بوده است.

از طرف دیگر پیاده‌سازی الگوریتم‌های شبکه عصبی روی 241 نمونه با استفاده از نمرات امتحانات و داده‌های سیستم‌های برخط مدیریت یادگیری، موفقیت بیشتری را نسبت به الگوریتم‌های درخت تصمیم به دست آورده است [18]. همین‌طور در مجموعه داده‌ای متشکل از 480 نمونه، شبکه عصبی با دقت 76 درصد از بین الگوریتم‌های J48، SVM، جنگل تصادفی و بیز ساده عملکرد بهتری داشته است [12].

با بررسی مطالعات بیشتر می‌توان متوجه شد که از مدل‌های داده‌کاوی برای پیش‌بینی معدل دانش‌آموزان نیز استفاده می‌شود. به طور نمونه مطالعه [10] با بررسی نمرات 236 دانش‌آموز و استفاده از درخت تصمیم J48 برای این منظور تلاش کرده است. همین‌طور استفاده از درخت تصمیم، جنگل تصادفی در کنار بیز ساده، پرسپترون چندلایه و درخت JRip نشان داده است. الگوریتم درخت تصمیم داده‌ها را در 5 کلاس (معدل‌های 20-16 کلاس 1 ، معدل‌های 15-14 کلاس 2، معدل‌های 13-12 کلاس 3 ، معدل‌های 11-10 کلاس 4 و معدل‌های کمتر از 10 کلاس 5) با دقت 85/62 درصد و الگوریتم جنگل تصادفی در 2 کلاس (موفقیت یا شکست) با دقت 93/49 درصد عملکرد خوبی داشتند [15].

دسته دیگری از تحقیقات در زمینه پیش‌بینی معدل به مقایسه 15 الگوریتم داده‌کاوی، روی اطلاعات 5566 نمونه پرداخته‌اند و نشان داده‌اند دو الگوریتم Naive Bayes و درخت تصمیم Hoeffding tree با دقت 91 درصد نتایج بهتری دارند [13]. در مرجع [7] نیز الگوریتم بیز ساده با دقتی بین 63/33 تا 69/67 درصد بهتر از الگوریتم‌های مبتنی بر درخت عمل می‌کند. اما در شناسایی دانش‌آموزان نمونه، جنگل تصادفی با دقتی بین 85/8% تا 92/6% به طور قابل‌توجهی بهتر از بیز ساده عمل می‌کند.

الگوریتم‌های رگرسیون و طبقه‌بندی نظیر درخت تصمیم، جنگل تصادفی، K–نزدیک‌ترین همسایه برای پیش‌بینی معدل نهایی 145 نمونه مورد مقایسه قرار گرفتند. نتایج نشان داد که رگرسیون SMOReg با دقت 96/98 درصد عملکرد بهتری نسبت به سایر الگوریتم‌ها داشته است [8].

تعدادی از محققین نیز به دنبال بهبود دقت الگوریتم‌های داده‌کاوی از ترکیب الگوریتم‌ها برای مدل‌سازی و ارزیابی عملکرد دانش‌آموزان استفاده کرده‌اند. به عنوان مثال، مرجع [20] الگوریتم‌های طبقه‌بندی SVM، بیز ساده، درخت تصمیم، شبکه عصبی و الگوریتم خوشه‌بندی K-میانگین را با یکدیگر ترکیب کرده و به دقت 75 درصد رسیده است. یا در تحقیقی دیگر محققین، الگوریتم‌های پرسپترون چندلایه، درخت تصمیم J48، و درخت تصمیم PART را با سه الگوریتم Bagging (BAG)، MultiBoost (MB) و Voting (VT) ترکیب کرده و 9 مدل

[1] Binary Genetic Algorithm

طبقه‌بندی جدید را ایجاد کرده‌اند. نتایج ارزیابی آنها روی 1227 نمونه نشان داد که بهترین الگوریتم استفاده ترکیبی دو الگوریتم MULTIBOOST و پرسپترون چندلایه با دقت 98/7٪ است [14].

محققین همین‌طور به نقش روش‌های مختلف داده‌کاوی در شناسایی عوامل تاثیرگذار بر افت تحصیلی دانش‌آموزان پرداخته‌اند. در این زمینه، نجفی و همکاران [25] از الگوریتم‌های درخت تصمیم J48 و خوشه‌بندی K-میانگین و استخراج قوانین[2] استفاده کرده است. در این میان J48 با دقت 95 درصد و خوشه‌بندی K-میانگین با ضریب اطمینان 95 درصد بهتر عمل می‌کنند.

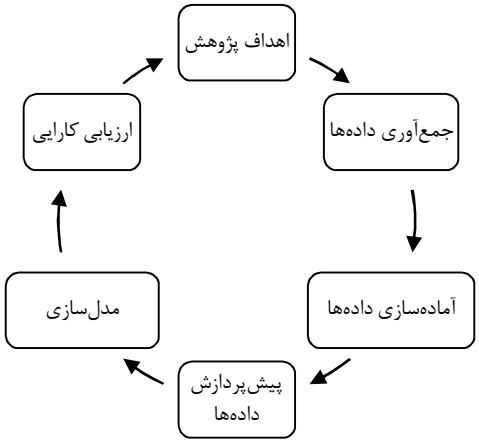
با بررسی مقاله‌های مورد مطالعه می‌توان متوجه شد که محققین در انتخاب ویژگی‌هایی مورد نظر خود معیار جامعی را مشخص نکرده‌اند؛ لذا با توجه به تنوع ویژگی‌ها می‌توان چنین برداشت کرد که می‌بایست ویژگی‌هایی مدنظر قرار گیرد که تمام ابعاد زندگی یک شخص را شامل شوند. از اینرو با جستجوی کلید واژه «ابعاد مهم زندگی یک فرد» به ابزار «چرخ تعادل زندگی» از جی مایر[3] می‌توان رسید. ابزار «چرخ تعادل زندگی» برای اولین بار با مفهوم تبدیل شدن به یک «فرد کامل» توصیف شد و بر شش حوزه حیاتی شامل (1) جسمی و فیزیکی، (2) امور مالی و شغلی، (3) فکری، (4) احساسی، (5) اجتماعی و (6) معنویت تمرکز دارد [29].

در این مطالعه سعی شده است تا میزان موثر بودن ویژگی‌های چرخ تعادل زندگی در موفقیت تحصیلی دانش‌آموزان بررسی شود. برای این منظور ابتدا پرسش‌نامه چرخ تعادل زندگی در اختیار دانش‌آموزان قرار گرفته و برای افزایش دقت، اطلاعات دیگری مانند اطلاعات جمعیت‌شناختی و پیشینه تحصیلی آنها نیز جمع‌آوری شد. سپس معدل به عنوان معیاری برای موفقیت تحصیلی دانش‌آموزان در نظر گرفته شد. درنهایت از الگوریتم‌های داده‌کاوی برای پیدا کردن ارتباطی بین چرخ تعادل زندگی، اطلاعات جمعیت‌شناختی و اطلاعات تحصیلی با موفقیت تحصیلی استفاده شده است. نتایج به دست آمده می‌تواند به پیش‌بینی موفقیت سایر دانش‌آموزان کمک کند.

به طور کلی استفاده از چرخ تعادل زندگی به عنوان چارچوبی منسجم و چندبعدی در داده‌کاوی تحصیلی و همین‌طور استفاده از معدل پایه دهم به عنوان معیاری برای پیش‌بینی موفقیت تحصیلی دانش‌آموزان در رشته مورد نظر را می‌توان جزو نوآوری‌های این مقاله در نظر گرفت.

4. روش پیشنهادی

هدف این مقاله، پیش‌بینی معدل دانش‌آموزان پایه دهم در مقطع متوسطه دوم بر اساس چرخ تعادل زندگی و با استفاده از روش‌های داده‌کاوی است. ابتدا مراحل روش پیشنهادی با توجه به شکل (1) بیان و در ادامه هر کدام از این مراحل توضیح داده می‌شوند.



شکل (1): مراحل روش پیشنهادی

1.4. اهداف پژوهش

مرحله اولیه هر پژوهش بر اهداف و الزامات آن تمرکز دارد تا بتوان درک بهتری از مساله داشت. اهداف این پژوهش را نیز می‌توان شامل موارد زیر برشمرد:

1. بررسی عوامل موثر بر موفقیت تحصیلی دانش‌آموزان پایه دهم متوسطه؛
2. جمع‌آوری داده‌ها با استفاده از پرسش‌نامه استاندارد چرخ تعادل زندگی، اطلاعات جمعیت‌شناختی و پیشینه تحصیلی؛

[2] Apriori
[3] Paul J. Meyer

3. پیش‌بینی معدل دانش‌آموزان با استفاده از الگوریتم‌های مختلف داده‌کاوی و مقایسه نتایج آنها.

نتایج این تحقیق این امکان را می‌دهد تا دانش‌آموزانی که بیشترین احتمال عدم موفقیت در رشته تحصیلی خود را دارند، قبل از ورود به رشته تحصیلی، شناسایی شوند. از اینرو می‌توان برای انتخاب رشته مناسب آنها برنامه‌ریزی‌های بهتری را انجام داد.

2.4. جمع‌آوری داده‌ها

در این مقاله برای جمع‌آوری داده‌ها از پرسش‌نامه استفاده شده است. در تهیه پرسش‌نامه علاوه بر سوالات چرخ تعادل زندگی [30]، با نظر مشاورین مدرسه و متخصصان حوزه آموزش و پرورش، اطلاعات جمعیت‌شناختی و پیشینه تحصیلی دانش‌آموزان (شامل ویژگی‌های جنسیت، وضعیت تاهل، تحصیلات والدین، ساعات مطالعه روزانه، مقدار پول‌توجیبی ماهانه، رشته تحصیلی و معدل) نیز به پرسش‌نامه افزوده شد. بعد از هماهنگی با اداره‌های آموزش و پرورش مناطق استان خراسان رضوی و آموزشگاه‌های شهر کاشمر، پرسش‌نامه‌ها به صورت کاغذی و الکترونیکی در اختیار دانش‌آموزان پایه دهم قرار گرفت. در نهایت اطلاعات 490 دانش‌آموز با 43 ویژگی جمع‌آوری شد. پرسشنامه در پیوست این مقاله قابل مشاهده است.

3.4. آماده‌سازی و پیش‌پردازش داده‌ها

در مرحله آماده‌سازی، ابتدا تعداد 17 پرسش‌نامه ناقص حذف و در گام بعدی از آنجایی‌که مقادیر برخی از ویژگی‌ها به صورت متن وارد شده‌اند، می‌بایست برای استفاده در الگوریتم‌ها و اندازه‌گیری امتیاز هر بخش، به مقادیر عددی تبدیل شوند. جدول (1) مقادیر عددی اختصاص یافته به هر پاسخ متنی را نشان می‌دهد. البته در بخش رشته تحصیلی، کد 4 مربوط به رشته ریاضی بوده که به دلیل کم بودن نمونه‌های جمع‌آوری شده از استفاده آنها خودداری شده است.

همین‌طور بر اساس نظر کارشناسان خبره داده‌کاوی به دلیل اینکه تعادل بین کلاس ها حفظ شود ویژگی معدل مطابق جدول (2) محدوده‌بندی شده است. این کار با قرار دادن معدل در محدوده‌های مشخص شده می‌تواند به متعادل کردن نمونه‌ها در دسته‌های ویژگی هدف یعنی معدل کمک کند و باعث بهبود عملکرد و کاهش زمان پردازش شود.

جدول (2): دسته‌بندی معدل

معدل	دسته	تعداد نمونه
20-18	1	187
18-16	2	193
16-0	3	93

جدول (1): مقادیر عددی اختصاص‌یافته به پاسخ‌های متنی

ویژگی	توضیحات				
جنسیت	مرد: 0	زن:1			
وضعیت تاهل	مجرد: 0	متاهل:1			
تحصیلات پدر و مادر	سیکل و پایین‌تر: 1	دیپلم: 2	فوق‌دیپلم: 3	کارشناسی: 4	کارشناسی‌ارشد: 5	دکترا: 6
میزان پول توجیبی	کمتر از 50 هزار تومان: 1	بین 50-100 هزار تومان: 2				
	بین 300-100 هزار تومان: 3	بیشتر از 300 هزار تومان: 4				
میزان مطالعه	بیشتر از 4 ساعت: 1	بین 3-1 ساعت: 2	کمتر از 1 ساعت: 3			
وسایل کمک‌آموزشی	هیچ‌کدام: 0	فیلم آموزشی و پادکست: 1	کتاب‌های کمک‌آموزشی: 2	کلاس خصوصی: 3		
رشته تحصیلی	تجربی: 1	انسانی: 2	گرافیک: 3	طراحی دوخت: 5	عکاسی: 6	
	کامپیوتر: 7	مدیریت خانواده: 8	کشت گیاهان دارویی: 9	مکانیک: 10		
پاسخ‌های چرخ تعادل زندگی	کاملا مخالف: 1	مخالف: 2	موافق: 3	کاملا موافق: 4		

بعد از انجام آماده‌سازی، مجموعه داده پیشنهادی با 473 نمونه تهیه شد. از این مجموعه داده می‌توان برای انجام پیش‌پردازش و مدل‌سازی در نرم‌افزار WEKA استفاده کرد. این نرم‌افزار بر پایه جاوا و توسط دانشگاه وایکاتو نیوزلند توسعه یافته است. WEKA، برنامه مناسبی برای انجام پردازش‌ها و تکنیک‌های مختلف داده‌کاوی روی داده‌های بزرگ است.

در پیش‌پردازش، روی مجموعه داده پیشنهادی، فرآیندهایی انجام می‌شود تا داده‌ها، مناسب استفاده برای مدل مورد نظر شوند. از اینرو، در این مقاله از روش‌های استانداردسازی داده و انتخاب ویژگی به ترتیب برای افزایش دقت پیش‌بینی و مشخص کردن ویژگی‌هایی با تاثیر بالا استفاده شده است. الگوریتم‌های استانداردسازی مقادیر عددی ویژگی‌ها را در بازه صفر تا یک مقیاس‌بندی می‌کند و سبب می‌شود تا الگوریتم‌های پیش‌بینی داده‌کاوی فرضیات واضحی از توزیع داده‌ها داشته باشند. همین‌طور از روش‌های انتخاب ویژگی برای دانستن اینکه کدام ویژگی‌ها در بهبود نتایج تاثیر بیشتری دارند، استفاده می‌شود. جدول (3) الگوریتم‌های انتخاب ویژگی استفاده شده را بیان می‌کند.

جدول (3): روش‌های انتخاب ویژگی

روش‌های انتخاب ویژگی	کاربرد
WrapperSubsetEval (Wrapper)	این روش مجموعه را با استفاده از یک طرح یادگیری ارزیابی می‌کند [15].
CfsSubsetEval (Wrapper)	در این روش ویژگی‌هایی که همبستگی بالایی با ویژگی هدف دارند و در عین حال همبستگی کمتری با هم دارند، انتخاب می‌شوند [7].
CorrelationAttributeEval (Filter)	میزان تاثیر ویژگی را با استفاده از اندازه‌گیری همبستگی بین ویژگی و ویژگی هدف ارزیابی می‌کند [7].
InfoGainAttributeEval (Wrapper)	این روش تاثیر یک ویژگی را با اندازه‌گیری میزان سودمندی مرتبط با ویژگی هدف ارزیابی می‌کند [7].

4.4. مدل‌سازی

در این مرحله، روش‌های طبقه‌بندی مختلفی برای پیش‌بینی معدل نهایی دانش‌آموزان، انتخاب و اعمال شدند. روش‌های طبقه‌بندی در این مقاله عبارت‌اند از:

الگوریتم بیز ساده: این الگوریتم از تکنیک‌های آمار و احتمال برای طبقه‌بندی نمونه‌ها در کلاس‌های مختلف استفاده می‌کند. این الگوریتم از قضیه بیز که احتمال وقوع یک پیشامد را بر اساس دانش قبلی توصیف می‌کند، استفاده می‌کند [31]. همچنین این الگوریتم از تمام ویژگی‌های موجود در داده‌ها استفاده می‌کند و فرض می‌کند که هر ویژگی ورودی مستقل است و با استفاده از چند ویژگی که مقدار آنها مستقل از یکدیگر هستند، برای پیش‌بینی مقدار ویژگی هدف استفاده می‌کند [27].

الگوریتم ماشین بردار پشتیبان[4]: یک الگوریتم یادگیری نظارت‌شده می‌باشد که برای طبقه‌بندی و رگرسیون می‌توان از آن استفاده نمود. این روش در واقع ابرصفحه‌ای بهینه را برای جدا کردن نمونه‌های کلاس‌های مختلف از هم در فضای ویژگی می‌یابد. به عبارت ساده‌تر، هدف آن بیشینه کردن حاشیه بین نمونه‌های کلاس‌های مختلف و در عین حال کم کردن خطاهای طبقه‌بندی است. مولفه‌های اصلی این روش شامل ابرصفحه (مرز تصمیم‌گیری که نمونه‌ها را به کلاس‌های مختلف تقسیم می‌کند)، بردارهای پشتیبان (نمونه‌هایی در فضای ویژگی که نزدیک به مرز بوده و موقعیت و جهت آن را تعیین می‌کنند) و حاشیه (فاصله بین مرز تفکیک کلاسی و نزدیک‌ترین بردارهای پشتیبان که روش به دنبال بیشینه کردن آن است) می‌باشد. روش ماشین بردار پشتیبان برای مسائل خطی و غیرخطی موثر بوده و به دلیل استحکام، کارایی برای داده‌ها با ابعاد بالا و توانایی مدیریت روابط پیچیده بین ویژگی‌ها شناخته می‌شود [32].

الگوریتم جنگل تصادفی: یک الگوریتم داده‌کاوی بسیار کاربردی در مجموعه داده‌های بزرگ است و برای مسائل طبقه‌بندی و رگرسیون در یادگیری ماشین استفاده می‌شود. در این الگوریتم چندین درخت تصمیم ساخته می‌شود و میانگین

[4] Support Vector Machine (SVM)

آنها، دقت پیش‌بینی مدل را مشخص می‌کند. استفاده از الگوریتم جنگل تصادفی، خطر پیش‌برازش و زمان آموزش مدل را در کنار ارائه سطح بالایی از دقت، کاهش می‌دهد. الگوریتم جنگل تصادفی با تخمین و مدیریت مقادیر از دست رفته، پیش‌بینی‌های بسیار دقیقی را انجام می‌دهد [15].

الگوریتم K-نزدیک‌ترین همسایه: الگوریتم K-نزدیک‌ترین همسایه، از میزان شباهت یک نمونه به K نمونه از همسایگانش برای تصمیم‌گیری استفاده می‌کند. این الگوریتم با ورود داده جدید، آن را با داده‌های قبلی مقایسه می‌کند و سپس آن را در دسته‌ای که با همسایگانش (داده‌های جدید و قدیم) بیشترین شباهت را داشته باشند، قرار می‌دهد [32]. در این مطالعه الگوریتم KNN با در نظر گرفتن تعداد 7 همسایه، با محاسبه فاصله اقلیدسی بین نمونه‌ها استفاده شده است.

الگوریتم درخت تصمیم J48: یک پیاده‌سازی از الگوریتم درخت تصمیم C4.5 در ابزار داده‌کاوی WEKA است و قابلیت‌هایی نظیر محاسبه مقادیر از دست رفته، هرس درخت تصمیم و استخراج قوانین را دارد [12]. این الگوریتم با یک مجموعه اصلی مانند S به عنوان گره ریشه شروع می‌کند. سپس در هر تکرار از الگوریتم، مقدار آنتروپی یا مقدار بهره اطلاعات برای ویژگی‌های مجموعه S که تا آن مرحله مورد استفاده قرار نگرفته‌اند، محاسبه و در این میان ویژگی با کمترین مقدار آنتروپی یا بیشترین بهره اطلاعات انتخاب می‌شود. سپس بر اساس ویژگی انتخاب شده، زیرمجموعه‌هایی از داده‌ها تولید می‌شود. الگوریتم همچنان با استفاده از ویژگی‌هایی که تاکنون انتخاب نشده‌اند به صورت بازگشتی به فرآیند تقسیم کردن ادامه می‌دهد تا درخت مورد نظر تشکیل شود. در درخت تصمیم J48، برگ‌های درخت تصمیم، ویژگی هدف را نشان می‌دهند [15]. در این مطالعه از الگوریتم J48 با در نظر گرفتن ضریب اطمینان 0/9 جهت ارائه بهتر رویدادهای ممکن با انجام کمترین هرس و در نظر گرفتن حداقل 20 نمونه در هر برگ بکار گرفته می‌شود.

الگوریتم پرسپترون چندلایه: الگوریتم‌های شبکه عصبی مصنوعی انواع و کاربردهای مختلفی دارد که الگوریتم پرسپترون چندلایه از جمله الگوریتم‌های شبکه عصبی عمیق است. این نوع شبکه عصبی مصنوعی از چندین لایه، از جمله یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل شده است و هر لایه به لایه بعدی متصل است. این الگوریتم از خروجی‌های لایه اول (ورودی)، به عنوان ورودی‌های لایه بعدی (لایه پنهان) استفاده می‌کند تا زمانی که پس از تعداد خاصی از لایه‌ها، خروجی‌های آخرین لایه پنهان به عنوان ورودی‌های لایه خروجی مورد استفاده قرار می‌گیرد و لایه خروجی وظایفی مانند طبقه‌بندی و پیش‌بینی را انجام می‌دهد [24].

تعداد لایه‌های شبکه عصبی پرسپترون چندلایه در این مقاله برابر با مجموع تعداد کلاس‌ها و تعداد ویژگی‌ها در نظر گرفته شده است و بعد از بر زدن و نامرتب کردن داده‌ها، آموزش را با 100 مرتبه تکرار انجام می‌دهد.

5.4. ارزیابی

در ابتدا باید بیان کرد که روایی پرسشنامه از طریق روایی محتوایی و پایایی پرسشنامه از طریق آزمون آلفای کرونباخ با مقدار 0/82 تایید گردید. در ادامه به بیان شاخص‌های ارزیابی مدل‌های داده‌کای می‌پردازیم.

از آنجایی که در داده‌کاوی به طور معمول چندین مدل ساخته می‌شوند، ارزیابی و انتخاب مناسب‌ترین آنها از اهمیت ویژه‌ای برخوردار است. با بررسی مطالعات انجام شده می‌توان معیارها و ابزارهای زیر را استخراج کرد:

- دقت، صحت ، پوشش و معیار F1 [12]، [14]، [19]، [33]
- ماتریس درهم‌ریختگی [14]، [17]، [19]، [23]
- منحنی ROC [18]، [33]

ارزیابی کارایی، یکی از مهم‌ترین کارهایی است که برای اندازه‌گیری و مقایسه میزان موفقیت مدل‌ها در پیش‌بینی، کاربرد دارد. برای ارزیابی عملکرد الگوریتم‌ها از روش اعتبارسنجی متقابل k-بخشی[5] استفاده شده است. در این روش مجموعه داده‌ها به طور تصادفی به k زیرمجموعه با حجم یکسان تفکیک

[5] K-fold cross validation

شده و در هر مرحله، تعداد $k-1$ از این زیرمجموعه‌ها به عنوان داده‌های آموزش و یک زیرمجموعه به عنوان داده آزمایش در نظر گرفته می‌شود [14]. از اینرو ارزیابی مدل‌ها با استفاده از نمادهای زیر انجام می‌شود:

- TP: تعداد نمونه‌هایی که در واقعیت، متعلق به کلاس درست هستند و الگوریتم طبقه‌بندی نیز آنها را در کلاس درست تشخیص داده است.
- TN: تعداد نمونه‌هایی که در واقعیت، متعلق به کلاس اشتباه هستند و الگوریتم طبقه‌بندی نیز آنها را در کلاس اشتباه تشخیص داده است.
- FN: تعداد نمونه‌هایی که در واقعیت، متعلق به کلاس درست هستند؛ اما الگوریتم طبقه‌بندی آنها را در کلاس اشتباه تشخیص داده است.
- FP: تعداد نمونه‌هایی که در واقعیت، متعلق به کلاس اشتباه هستند؛ اما الگوریتم طبقه‌بندی آنها را در کلاس درست تشخیص داده است.

بر این اساس چهار معیار دقت، صحت، پوشش و معیار F1 از تعداد نمادهای معرفی شده در بالا استفاده می‌کنند. معیار دقت الگوریتم طبقه‌بند به صورت زیر مجاسبه می‌شود:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

که برابر با تعداد موارد درست پیش‌بینی شده تقسیم بر تعداد کل پیش‌بینی‌های انجام شده است. این معیار نشان‌دهنده این است که چند درصد از کل تعداد نمونه‌ها به درستی دسته‌بندی شده‌اند. معیار صحت با رابطه زیر محاسبه می‌شود:

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

که نسبت نمونه‌های مثبت واقعی به همه نمونه‌هایی که مدل مثبت پیش‌بینی کرده است را نشان می‌دهد.

رابطه زیر معیار پوشش یا حساسیت را محاسبه می‌کند:

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

این معیار نسبت تعداد نمونه‌های مثبت واقعی بر تعداد کل مواردی که در واقعیت در کلاس درست هستند را نشان می‌دهد و برخلاف معیار صحت، این معیار خطای پیش‌بینی را در نظر می‌گیرد. در نهایت معیار F1 عبارتست از:

$$F1\_measure = \frac{2 \times (Percision \times Recall)}{Percision + Recall} \qquad (4)$$

با توجه به اینکه دو معیار صحت و پوشش عکس یکدیگر عمل می‌کنند به طوری که با افزایش یکی از آنها، معیار دیگر کاهش پیدا می‌کند. بنابراین یک معیار دیگری به نام معیار F1 برای محاسبه میانگین هندسی دو معیار ذکر شده معرفی می‌شود. در این مقاله مقدار هر کدام از معیارها برای هر مدل اندازه‌گیری و سپس با مقایسه نتایج آنها با یکدیگر، بهترین مدل انتخاب می‌شود.

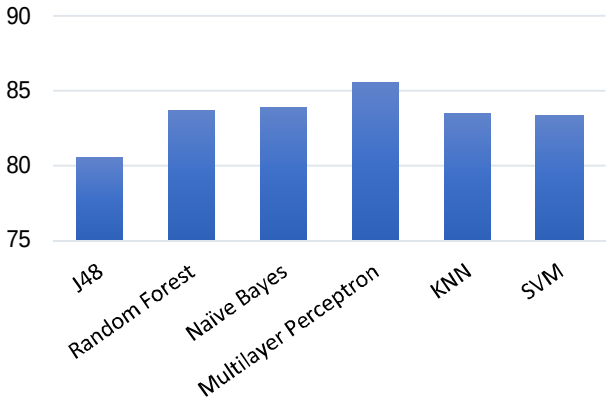
5. نتایج

در این قسمت ابتدا الگوریتم‌های گفته شده روی مجموعه داده پیشنهادی بدون انتخاب ویژگی‌های موثر، اجرا و سپس نتایج مدل‌های به دست آمده با معیارهای گفته شده در بالا اندازه‌گیری می‌شوند. همین‌طور در ادامه فرآیند انتخاب ویژگی انجام و ضمن بررسی ویژگی‌های تاثیرگذار، به عملکرد الگوریتم‌های داده‌کاوی روی ویژگی‌های انتخاب شده در هر یک پرداخته شده است. هدف از این کار نشان دادن اهمیت فرآیند انتخاب ویژگی در بهبود نتایج پیش‌بینی است. جدول (4) نتایج الگوریتم‌های داده‌کاوی را قبل از فرآیند انتخاب ویژگی نشان می‌دهد.

جدول (4): مقایسه معیارهای ارزیابی در الگوریتم‌های مطرح شده

الگوریتم	دقت (%)	صحت	پوشش	F1
J48	80/54	0/811	0/805	0/808
جنگل تصادفی	83/72	0/842	0/837	0/839
بیز ساده	83/93	0/839	0/839	0/839
پرسپترون چندلایه	85/62	0/860	0/856	0/858
KNN	83/50	0/837	0/835	0/834
SVM	83/29	0/835	0/833	0/834

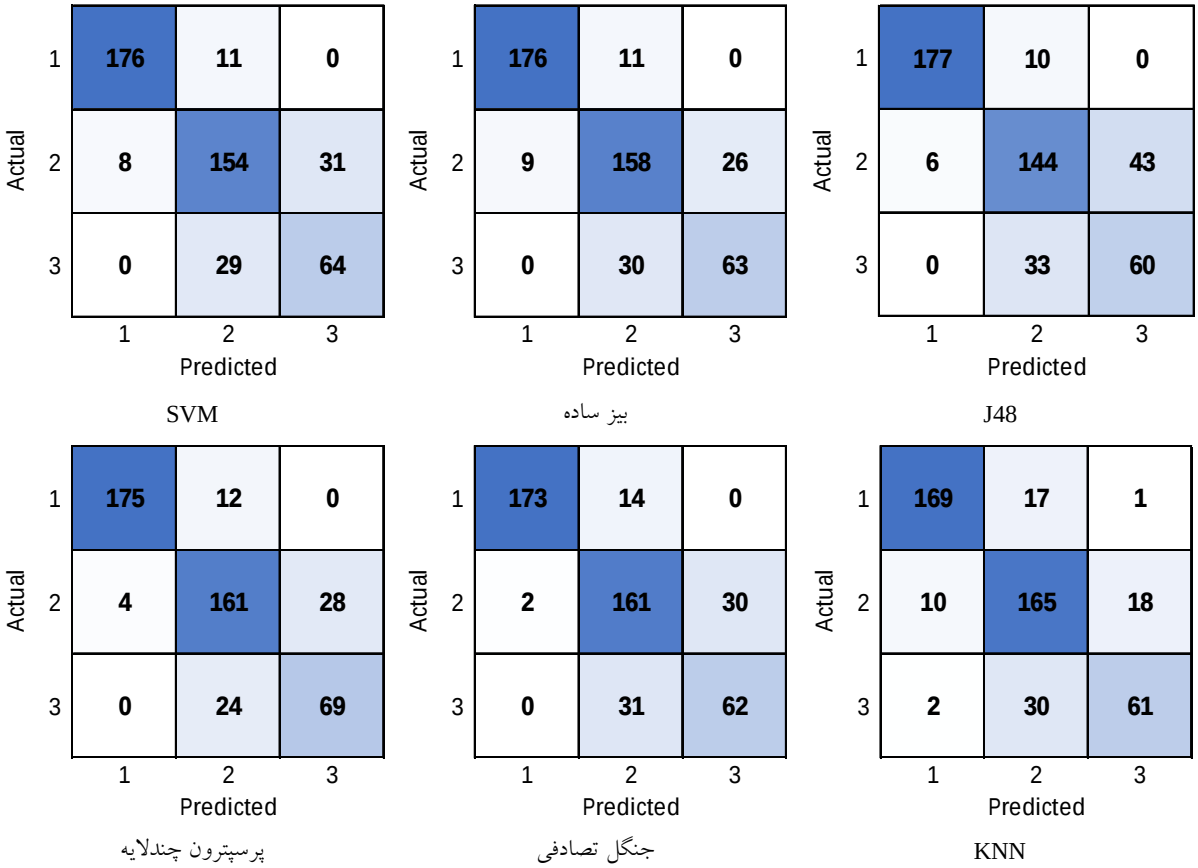
بررسی نتایج نشان می‌دهد که الگوریتم شبکه عصبی با دقت نزدیک به 85 درصد، عملکرد بهتری نسبت به سایر الگوریتم‌ها دارد. این یافته را می‌توان در شکل (3) بهتر مقایسه و بررسی کرد.



شکل (3): مقایسه دقت الگوریتم‌ها

ماتریس درهم‌ریختگی یکی از ابزارهایی است که برای نمایش عملکرد الگوریتم‌های مورد نظر استفاده می‌شود. سطرها کلاس‌های واقعی و ستون‌ها کلاس‌های پیش‌بینی شده توسط الگوریتم طبقه‌بند را نشان می‌دهد. این در حالی است که عناصر روی قطر اصلی تعداد نمونه‌های پیش‌بینی شده درست را نشان می‌دهند و به دنبال آن اعداد خارج از قطر اصلی، بیانگر تعداد نمونه‌هایی هستند که به درستی دسته‌بندی نشده‌اند [6]. شکل (4) ماتریس‌های درهم‌ریختگی هر کدام از مدل‌های مطرح شده در این مقاله را نشان می‌دهد.

یکی دیگر از ابزارهای تحلیل عملکرد مدل‌ها استفاده از مقدار زیر منحنی ROC است. منحنی‌های ROC در هر کلاس، میزان نمونه‌های درست دسته‌بندی شده را در برابر نمونه‌های مثبت کاذب نشان می‌دهد. بهترین حالت برای یک مدل، زمانی رخ می‌دهد که مساحت زیر منحنی ROC آن مدل برابر یک باشد [34]. جدول (5) مساحت زیر منحنی‌های ROC را برای هر کلاس در مدل‌های مطرح شده در این مقاله نشان می‌دهد.



شکل (4): ماتریس‌های درهم ریختگی مدل‌ها قبل از انتخاب ویژگی

جدول (5): مقدار AUC برای هر کلاس در مدل‌ها

کلاس	J48	Native Bayes	SVM	KNN	Random Forest	Multilayer Perceptron
1	0/98	0/98	0/98	0/97	0/98	0/98
2	0/91	0/92	0/83	0/89	0/92	0/92
3	0/92	0/94	0/92	0/93	0/94	0/94

همان‌طور که در جدول بالا مشاهده می‌شود، در کلاس یک مدل‌های J48، بیز ساده، SVM، جنگل تصادفی و پرسپترون چندلایه بهترین نتیجه و به همین ترتیب در کلاس‌های دو و سه الگوریتم‌های بیز ساده، جنگل تصادفی و پرسپترون چندلایه بهترین عملکرد را داشته‌اند. به طوری که تعداد نمونه‌های درست پیش‌بینی شده در آن به واقعیت نزدیک‌تر و تعداد نمونه‌های مثبت کاذب در آن بسیار کم است.

1.5. انتخاب ویژگی‌های تاثیرگذار

هدف از این بخش انتخاب تاثیرگذارترین ویژگی‌ها و به دنبال آن انتخاب بهترین الگوریتم انتخاب ویژگی است. همان‌طور که پیش‌تر گفته شد، بعد از انجام پیش‌پردازش‌های لازم جهت آماده‌سازی داده‌های ورودی، از الگوریتم‌های انتخاب ویژگی استفاده می‌شود. انتخاب ویژگی می‌تواند پیچیدگی محاسباتی را کاهش داده و در دقت عملکرد پیش‌بینی تاثیر بگذارد. از اینرو، از چندین الگوریتم انتخاب ویژگی مشهور به منظور انتخاب ویژگی‌های تاثیرگذار در پیش‌بینی معدل پایه دهم استفاده شد. جدول (6) نتایج الگوریتم‌های انتخاب ویژگی را نشان می‌دهد. همان‌طور که در جدول (3) گفته شد، هر کدام از الگوریتم‌های مورد مقایسه، ویژگی‌ها را بر اساس معیارهای مشخصی انتخاب می‌کنند. از اینرو، الگوریتم‌های گفته شده در جدول (6) به ترتیب ویژگی‌ها را از نظر ارزیابی با یک طرح یادگیری، همبستگی ویژگی با کلاس با در نظر گرفتن همبستگی کمتر با سایر ویژگی‌ها، همبستگی بین ویژگی‌ها با ویژگی‌های هدف و میزان سودمندی مرتبط با ویژگی هدف انتخاب می‌کنند. در همین راستا می‌توان ویژگی‌های قرار گرفته در جدول (6) را از نظر ابعاد گفته شده بررسی کرد.

جدول (6): ویژگی‌های انتخاب شده در الگوریتم‌های انتخاب ویژگی

الگوریتم	ویژگی‌های انتخاب شده		
WrapperSubsetEval (Wrapper)	پول توجیبی میزان مطالعه جنسیت رشته تحصیلی	تحصیلات پدر تحصیلات مادر بعد روحی	
CfsSubsetEval (wrapper)	پول توجیبی میزان مطالعه جنسیت	رشته تحصیلی تحصیلات پدر بعد فکری	
CorrelationAttributeEval (Filter)	پول توجیبی میزان مطالعه جنسیت	رشته تحصیلی وسایل کمک آموزشی	
CfsSubsetEval (wrapper)	پول توجیبی میزان مطالعه رشته تحصیلی	بعد فکری بعد مالی بعد معنوی	

با توجه به نتایج جدول (6)، ویژگی‌های «پول توجیبی»، «میزان مطالعه»، «جنسیت» و «رشته تحصیلی» در تمام الگوریتم‌ها به عنوان ویژگی‌های موثر انتخاب شده‌اند. این نشان می‌دهد که ویژگی‌های گفته شده از تاثیر بیشتری در پیش‌بینی ویژگی هدف یعنی معدل برخوردار هستند. این عوامل در مطالعات قبلی نیز مورد توجه بوده‌اند، به عنوان مثال، مطالعات پیشین نیز مدت زمان مطالعه را در رابطه مستقیم با موفقیت تحصیلی می‌دانند. همچنین در بسیار موارد میزان تحصیلات والدین را دارای نقش بالایی در موفقیت تحصیلی فرزندان می‌دانند. وضعیت اقتصادی خانواده از دیگر عوامل موثر بر موفقیت تحصیلی در تحقیقات گذشته گزارش شده است. همچنین با توجه به جدول (6)، در میان ویژگی‌های مرتبط با چرخ زندگی نیز می‌توان نتایج زیر را به دست آورد:

- بعد روحی با ویژگی هدف بر اساس یک طرح یادگیری تطبیق بهتری داشته است.
- بعد فکری بیشترین همبستگی را با کلاس هدف داشته و در عین حال همبستگی کمتری با سایر ویژگی‌ها دارد.

- ابعاد فکری، مالی و معنوی نیز در پیش‌بینی ویژگی هدف بیشترین سودمندی را داشته‌اند.

بعد از ارزیابی ویژگی‌های موثر، نوبت به مشخص کردن بهترین الگوریتم انتخاب ویژگی با کمک شبکه عصبی می‌رسد، چرا که بر اساس بخش‌های قبلی، الگوریتم شبکه عصبی بهترین نتایج را به دنبال داشته است.

ابتدا، ویژگی‌هایی که تاثیرگذاری بیشتری دارند، توسط الگوریتم‌های انتخاب ویژگی مشخص (جدول (6)) و به دنبال آن سایر ویژگی‌ها از مجموعه داده حذف می‌شوند. در نهایت شبکه عصبی با روش اعتبارسنجی متقابل 10-بخشی روی ویژگی‌های باقیمانده یادگیری را انجام می‌دهد. این فرآیند برای هر الگوریتم انتخاب ویژگی انجام می‌شود. جدول (7) نتایج آنچه گفته شد را نشان می‌دهد.

آنچه که از جدول‌های (6) و (7) برمی‌آید، الگوریتم انتخاب ویژگی WrapperSubsetEval توانسته مجموعه ویژگی‌های تاثیرگذارتری را نسبت به سایر الگوریتم‌ها انتخاب کند.

جدول (7): عملکرد شبکه عصبی پس از انتخاب ویژگی

الگوریتم انتخاب ویژگی	عملکرد شبکه عصبی			
	Acuuracy	Precision	Recall	F1
WrapperSubsetEval	%85,20	%85,3	%85,2	%85,3
CfsSubsetEval	%84,14	%84,3	%84,1	%84,2
CorrelationAttributeEval	%83,50	%83,8	%83,5	%83,5
CfsSubsetEval	%81,39	%81,6	%81,4	%81,5

همین‌طور که از مقایسه جدول‌های (4) و (7) مشاهده می‌شود، عملکرد مدل شبکه عصبی بر اساس شاخص‌های مختلف تغییر چشمگیری نداشته است. این موضوع نشان می‌دهد که مجموعه ویژگی‌های محدود انتخاب شده نیز به خوبی می‌توانند موفقیت دانش‌آموز را پیش‌بینی نمایند.

6. نتیجه‌گیری

امروزه کاربرد داده‌کاوی در مراکز آموزشی از اهمیت زیادی برخوردار است. دانش به دست آمده ممکن است بتواند باعث اتخاذ تصمیمات بهتر و در نتیجه بهبود و پیشرفت عملکرد دانش‌آموزان شود. در این مطالعه یک مجموعه داده پیشنهادی با کمک ابزار چرخ تعادل زندگی، اطلاعات جمعیت‌شناختی، خانوادگی و تحصیلی دانش‌آموزان آموزشگاه‌های شهر کاشمر و شامل 473 نمونه ایجاد شد. همین‌طور شش الگوریتم طبقه‌بندی‌کننده شبکه عصبی پرسپترون چندلایه، درخت تصمیم J48، جنگل تصادفی، بیز ساده، SVM و KNN با هدف طبقه‌بندی معدل دانش‌آموزان در پایه دهم بکار گرفته شدند. برای پیاده‌سازی الگوریتم‌های گفته شده از نرم‌افزار WEKA استفاده و دقت آنها محاسبه شد. نتایج آزمایش‌ها نشان می‌دهند که الگوریتم پرسپترون چندلایه با دقت 85 درصد، بهترین عملکرد را در میان الگوریتم‌های دیگر داشته است و می‌تواند برای پیش‌بینی معدل و به دنبال آن اخذ تصمیمات بهتر برای انتخاب رشته دانش‌آموزان قبل از ورود آنها به رشته تحصیلی استفاده شود. در کارهای آینده، سعی خواهد شد تا نمونه‌های بیشتری در جهت ایجاد یک مجموعه داده مناسب جمع‌آوری شوند. چرا که در کنار داشتن یک مجموعه داده مناسب، می‌توان سایر روش‌های داده‌کاوی را نیز مورد بررسی قرار داد. استفاده از روش‌های مختلف، تاثیرات مثبت داده‌کاوی تحصیلی در مدارس را روشن‌تر و تصمیم‌گیری‌های کلان آموزشی را آسان‌تر خواهد کرد.

سپاسگزاری

این تحقیق حاصل پایان‌نامه کارشناسی ارشد بوده و نویسندگان از تمامی کسانی که در انجام این تحقیق یاری رساندند، به ویژه اداره آموزش و پرورش شهرستان کاشمر تشکر و قدردانی می‌نمایند.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[1] S. Hussain and M. Q. Khan, “Student-Performulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning,” Ann. Data Sci., vol. 10, no. 3, pp. 637-655, 2023, doi: 10.1007/s40745-021-00341-0.

[2] E. C. Abana, “A Decision Tree Approach for Predicting Student Grades in Research Project Using Weka,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 7, pp. 314-319, 2019, doi: 10.14569/IJACSA.2019.0100739.

[3] O. Khalaf Beigi, S. A. Bashiri Mosavi, and S. Gharloghi, “Applying Character-Level Neural Network-Based Sentiment Analysis Model on Persian Comments of the Social Media-Online Store Platforms,” Soft Comput. J., vol. 11, no. 2, pp. 118-133, 2023, doi: 10.22052/scj.2023.248311.1094 [In Persian].

[4] M. Eftekharian and A. Nodehi, “Breast Cancer Diagnosis and Classification Improvement Based on Deep Learning and Image Processing,” Soft Comput. J., vol. 12, no. 1, pp. 22-26, 2023, doi: 10.22052/scj.2023.246416.1067.

[5] J. D. Dresser, K. M. Whitfield, L. J. Kremer, and K. J. Wilby, “Exploring How Postmillennial Pharmacy Students Balance Life Priorities and Avoid Situations Known to Deplete Resilience,” Amer. J. Pharm. Educ., vol. 85, no. 4, p. 8410, 2021, doi: 10.5688/ajpe8369.

[6] E. Alyahyan and D. Duştegor, “Predicting Academic Success in Higher Education: Literature Review and Best Practices,” Int. J. Educ. Technol. High. Educ., vol. 17, p. 3, 2020, doi: 10.1186/s41239-020-0177-7.

[7] S. Alturki and N. Alturki, “Using Educational Data Mining to Predict Students’ Academic Performance for Applying Early Interventions,” J. Inf. Technol. Educ. Innov. Pract., vol. 20, pp. 121-137, 2021, doi: 10.28945/4835.

[8] B. Al Breiki, N. Zaki, and E. A. Mohamed, “Using Educational Data Mining Techniques to Predict Student Performance,” in Proc. Int. Conf. Elect. Comput. Technol. Appl. (ICECTA), 2019, pp. 1-5, doi: 10.1109/ICECTA48151.2019.8959676.

[9] A. E. Tatar and D. Dustegor, “Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?,” Appl. Sci., vol. 10, no. 14, p. 4967, 2020, doi: 10.3390/app10144967.

[10] M. A. Al-Barrak and M. Al-Razgan, “Predicting Students Final GPA Using Decision Trees: A Case Study,” Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528-533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[11] H. Turabieh, “Hybrid Machine Learning Classifiers to Predict Student Performance,” in Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS), 2019, pp. 1-6, doi: 10.1109/ICTCS.2019.8923093.

[12] C. Jalota and R. Agrawal, “Analysis of Educational Data Mining Using Classification,” in Proc. Int. Conf. Mach. Learn. Big Data Cloud Parallel Comput. (COMITCon), 2019, pp. 243-247, doi: 10.1109/COMITCon.2019.8862214.

[13] N. Alangari and R. Alturki, “Predicting Students Final GPA Using 15 Classification Algorithms,” Romanian J. Inf. Sci. Technol., vol. 23, no. 3, pp. 238-249, 2020.

[14] A. Siddique et al., “Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers,” Appl. Sci., vol. 11, no. 24, p. 11845, 2021, doi: 10.3390/app112411845.

[15] M. Kumar, C. Sharma, S. Sharma, N. Nidhi, and N. Islam, “Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance,” in Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA), 2022, pp. 1013-1017, doi: 10.1109/DASA54658.2022.9765236.

[16] D. Hooshyar, M. Pedaste, and Y. Yang, “Mining Educational Data to Predict Students’ Performance Through Procrastination Behavior,” Entropy, vol. 22, no. 1, p. 12, 2019, doi: 10.3390/e22010012.

[17] M. H. B. Roslan and C. J. Chen, “Predicting Students’ Performance in English and Mathematics Using Data Mining Techniques,” Educ. Inf. Technol., vol. 28, no. 2, pp. 1427-1453, 2023, doi: 10.1007/s10639-022-11259-2.

[18] A. Alhassan, B. Zafar, and A. Mueen, “Predict

Students' Academic Performance Based on Their Assessment Grades and Online Activity Data," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, pp. 231-239, 2020, doi: 10.14569/IJACSA.2020.0110425.

[19] C.-C. Kiu, "Data Mining Analysis on Student's Academic Performance Through Exploration of Student's Background and Social Activities," in Proc. 4th Int. Conf. Adv. Comput. Commun. Autom. (ICACCA), 2018, pp. 1-5, doi: 10.1109/ICACCAF.2018.8776809.

[20] B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach," J. Med. Syst., vol. 43, no. 6, p. 162, 2019, doi: 10.1007/s10916-019-1295-4.

[21] D. T. Ha, P. T. T. Loan, C. N. Giap, and N. T. L. Huong, "An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques," Int. J. Comput. Sci. Inf. Secur., vol. 18, no. 3, pp. 75-82, 2020.

[22] W. F. W. Yaacob, S. A. M. Nasir, and N. M. Sobri, "Supervised Data Mining Approach for Predicting Student Performance," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, pp. 1584-1592, 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

[23] F. J. Kaunang and R. Rotikan, "Students' Academic Performance Prediction Using Data Mining," in Proc. 3rd Int. Conf. Informatics Comput. (ICIC), 2018, pp. 1-5, doi: 10.1109/IAC.2018.8780547.

[24] T. Ahajjam et al., "Predicting Students' Final Performance Using Artificial Neural Networks," Big Data Min. Anal., vol. 5, no. 4, pp. 294-301, 2022, doi: 10.26599/BDMA.2021.9020030.

[25] M. Najafi, M. Afzali, and M. Moradi, "Use Data Mining to Identify Factors Affecting Students' Academic Failure," Intell. Multimedia Process. Commun. Syst. (IMPCS), vol. 2, no. 1, pp. 23-33, 2021, dor: 20.1001.1.27832570.1399.1.2.4.0 [In Persian].

[26] S. Aydogdu, "Predicting Student Final Performance Using Artificial Neural Networks in Online Learning Environments," Educ. Inf. Technol., vol. 25, no. 3, pp. 1913-1927, 2020, doi: 10.1007/s10639-019-10053-x.

[27] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing Prediction of Student Success: Automated Machine Learning Approach," Comput. Electr. Eng., vol. 89, p. 106903, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[28] Z. Shao, H. Sun, X. Wang, and Z. Sun, "An Optimized Mining Algorithm for Analyzing Students' Learning Degree Based on Dynamic Data," IEEE Access, vol. 8, pp. 113543-113556, 2020, doi: 10.1109/ACCESS.2020.3001749.

[29] Z. M. Taras and V. I. Mesyura, "Application of the Wheel of Life Balance to Time Management Software," in Proc. Int. Conf. Develop. Appl. Syst. (DAS), 2022, pp. 178-185, doi: 10.1109/DAS54948.2022.9786194.

[30] G. M. Robertson, Life Balance Assessment and Action Planning Guide. New York, NY, USA: Robertson Consulting, 2002.

[31] F. Zare Mehrjardi, M. Yazdian-Dehkordi, and A. Latif, "Evaluating Classical Machine Learning and Deep-Learning Methods in Sentiment Analysis of Persian Telegram Message," Soft Comput. J., vol. 11, no. 1, pp. 88-105, 2022, doi: 10.22052/SCJ.2022.246281.1058 [In Persian].

[32] H. Al-Shehri et al., "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor," in Proc. IEEE 30th Can. Conf. Elect. Comput. Eng. (CCECE), 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.

[33] M. Yagci, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," Smart Learn. Environ., vol. 9, no. 1, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[34] I. Sandoval-Palis, D. Naranjo, R. Gilar-Corbi, and T. Pozo-Rico, "Neural Network Model for Predicting Student Failure in the Academic Leveling Course of Escuela Politecnica Nacional," Front. Psychol., vol. 11, p. 515531, 2020, doi: 10.3389/fpsyg.2020.515531.

پیوست: پرسشنامه

1. جنسیت: مرد ☐ زن ☐
2. وضعیت تاهل: مجرد ☐ متاهل ☐
3. تحصیلات پدر:

سیکل و پایین‌تر ☐ دیپلم ☐ کارشناسی (لیسانس) ☐ کارشناسی ارشد (فوق لیسانس) ☐ دکترا ☐

4. تحصیلات مادر:

سیکل و پایین‌تر ☐ دیپلم ☐ کارشناسی (لیسانس) ☐ کارشناسی ارشد (فوق لیسانس) ☐ دکترا ☐

5. میزان پول‌توجیبی دریافتی ماهانه شما از پدر و مادر چقدر است؟

بیشتر از 100 هزار تومان ☐ بین 100-50 هزار تومان ☐ کمتر از 50 هزار تومان ☐

6. به طور متوسط چند ساعت در روز مطالعه می‌کنید؟

بیشتر از 4 ساعت ☐ بین 3-1 ساعت ☐ کمتر از 1 ساعت ☐

7. علاوه بر امکانات درسی مدرسه از چه امکانات کمک‌آموزشی استفاده می‌کنید؟

کلاس‌های خصوصی ☐ کتاب‌های کمک‌آموزشی ☐ فیلم‌های آموزشی و پادکست ☐

8. معدل سال دهم:
9. رشته تحصیلی خود را بنویسید:
10. نام مدرسه خود را بنویسید:

	ردیف	سوال	کاملا موافق	موافق	مخالف	کاملا مخالف
فیزیکی	11	من در رژیم غذایی خود از مواد غذایی سالم نظیر سبزی‌ها، میوه‌ها، غلات و حبوبات، پروتئین و لبنیات به اندازه کافی و متعادل بر اساس هرم غذایی مصرف می‌کنم.				
	12	من در هفته چند بار حرکات ورزشی به مدت حداقل 20 دقیقه (همراه با بالارفتن ضربان قلب) انجام می‌دهم.	یک‌بار	دو بار	سه بار	چهار بار و بیشتر
	13	من بدنی سالم و به دور از ناتوانی جسمی با خواب کافی، رژیم غذایی مناسب، ورزش و فعالیت بدنی، رعایت بهداشت، عدم تنش در زندگی و... دارم.				
	14	من به طور کلی بیماری نگران‌کننده‌ای (چه جسمی چه روحی) ندارم.				
	15	من سلامتی خود را به صورت دوره‌ای و مرتب بررسی می‌کنم.				
	16	من دخانیات و هر چیزی که برای بدن ضرر دارد را مصرف نمی‌کنم.				
مالی	17	من استقلال مالی دارم و به اندازه پول توجیبی خود خرج می‌کنم.				
	18	من برای مواقع ضروری در آینده، مبلغی را پس‌انداز می‌کنم.				
	19	من انسان خوش‌حسابی هستم.				

	20 بین هزینه‌های فعلی و پس‌انداز برای آینده تعادل ایجاد می‌کنم.	
	21 خرج کردن من بر اساس نیازهای اولویت‌های خودم می‌باشد نه تقلید از دیگران.	
فکری	22 من از یادگیری مهارت‌ها و به دست آوردن دانش جدید لذت می‌برم.	
	23 من افکار مثبت دارم.	
	24 من به طور کلی از رشته تحصیلی خود راضی هستم.	
	25 من زمان و انرژی خود را صرف رشد تحصیلی و مهارت‌آموزی خود می‌کنم.	
	26 رشته تحصیلی من با علایق و استعدادهای من مطابقت دارد.	
	27 من تفریحات و سرگرمی‌های مورد علاقه خود را دنبال می‌کنم.	
عاطفی	28 من می‌توانم شرایط زندگی خود را کنترل کرده و با تغییر شرایط کنار بیایم.	
	29 من در مقابل مشکلات ناامید نشده و به دنبال قوی‌تر کردن خودم هستم.	
	30 من می‌توانم در هنگام مشکلات، آرامشم را حفظ کرده و خود را دلگرمی دهم.	
	31 من انسان شادی هستم و می‌توانم به راحتی بخندم.	
	32 دیگران مرا از نظر روحی و روانی باثبات می‌دانند.	
	33 من خود را مسئول احساساتم و نحوه ابراز آنها می‌دانم.	
اجتماعی	34 من چند دوست صمیمی دارم.	یک نفر دو نفر سه نفر چهار نفر و بیشتر
	35 من توانایی حل مشکلات (خانواده و دوستان) را دارم.	
	36 من با مسئولین مدرسه و معلمان روابط خوبی دارم.	
	37 من به حریم خصوصی خود و دیگران احترام می‌گذارم.	
	38 من احساسات دیگران را درک کرده و می‌توانم رفتار مناسبی داشته باشم.	
	39 من احساس تعلق به یک دسته یا گروه خاصی دارم.	
معنوی	40 زندگی من هدفمند است (برای زندگی‌ام اهدافی دارم).	
	41 من به طور کلی در زندگی خود احساس آرامش دارم.	
	42 فعالیت‌هایی جهت رشد فکری مانند مطالعه، تفکر و عبادت را انجام می‌دهم.	
	43 من به توانایی‌های خود اعتماد دارم و بعد از شکست ناامید نمی‌شوم.	