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

اعظم علیپور فرگی[۱]، کارشناسی‌ارشد، حمید سعادت‌فر[۲*]، دانشیار

[۱] گروه مهندسی کامپیوتر – دانشکده مهندسی برق و کامپیوتر – دانشگاه بیرجند – بیرجند– ایران – azamalipour@birjand.ac.ir

[۲] گروه مهندسی کامپیوتر – دانشکده مهندسی برق و کامپیوتر – دانشگاه بیرجند – بیرجند – ایران – saadatfar@birjand.ac.ir

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. ازاین‌رو 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 مؤثر 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 تأثیرگذارتری را نسبت به سایر الگوریتم‌ها انتخاب کند. به‌طوری که ویژگی‌های پول‌توجیبی، میزان مطالعه، جنسیت، رشته تحصیلی، تحصیلات پدر، تحصیلات مادر و بعد روحی بیشترین تأثیر را دارند. همین‌طور الگوریتم پرسپترون چندلایه با دقت ۸۵,۶۲ درصد و الگوریتم بیز ساده با دقت ۸۶,۲۵ درصد، به ترتیب قبل و بعد از فرایند انتخاب ویژگی بهترین عملکرد را از خود نشان داده‌اند.

واژه‌های کلیدی: داده‌کاوی، انتخاب ویژگی، پیش‌بینی معدل، چرخ تعادل زندگی، موفقیت تحصیلی.

* نویسنده مسئول، **saadatfar@birjand.ac.ir**

Investigating the factors affecting the success of 10th grade students using data mining methods: A Case Study of Educational Institutions in Kashmar

Azam Alipourfargi 1, Master of Science, Hamid Saadatfar 2*, Associate Professor

1 Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran, azamalipour@birjand.ac.ir

2 Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran, saadatfar@birjand.ac.ir

***Abstract*:** The success of students at every educational level is one of the significant goals of the Ministry of Education. In the meantime, proper planning for the academic success of tenth-grade students is more valuable because of the choice of study field in this grade. Likewise, making a correct decision requires gathering comprehensive information from all aspects of a student's life. Therefore, the life balance wheel tool in six physical, financial, intellectual, emotional, social and spiritual parts can be a suitable solution. In this article, First, a dataset is created using the Life Balance Wheel, along with demographic and educational background information of 10th-grade students from educational institutions in Kashmar, collected through a questionnaire. In the following, well-known Wrapper and Filter methods are employed for selection of affective features and data mining algorithms including Multilayer Perceptron neural network, J48 decision tree, Random Forest, Naïve Bayes, SVM and KNN are used to predict the success of students in 10th grade. The results show that the Wrapper algorithm can select more impact features than others. So the features of pocket money, the amount of study, gender, field of study, father's education, mother's education, and spiritual dimension have more effect. Also, the Multilayer Perceptron algorithm with 85.62% accuracy and the Naïve Bayes algorithm with 86.25% accuracy have shown the best performance before and after the feature selection process, respectively.

Keywords: *Data Mining*, *Feature Selection*, *GPA prediction*, *Life Balance Wheel*, *Academic Success.*

*** Corresponding author, saadatfar@birjand.ac.ir**

۱. مقدمه

امروزه با توجه به پیشرفت‌های چشمگیر در زمینه ذخیره‌سازی و پردازش حجم زیادی از داده‌ها، نقش داده‌کاوی بیش از پیش پررنگ‌تر شده است [۱]. به‌طوری‌که محققان از داده‌کاوی برای استخراج دانش و اطلاعات، کشف الگوهای پنهان از مجموعه داده‌های بسیار بزرگ و پیچیده [۲] در حوزه‌های مختلف مانند بازاریابی، ارتباط با مشتری [۳]، مخابرات، صنعت، امور بهداشتی و پزشکی [۴] استفاده می‌کنند.

یکی دیگر از کاربردهای گسترده داده‌کاوی، در زمینه آموزشی است [۵]. در این میان می‌توان به کارهایی همچون پیش‌بینی موفقیت دانش‌آموزان [۶]، بررسی عملکرد دانش‌آموزان [۷, ۸] و هم‌چنین بررسی عوامل مؤثر در موفقیت دانش‌آموزان اشاره کرد. با استفاده از نتایج پیش‌بینی‌های انجام شده می‌توان برنامه‌ریزی‌های لازم برای افزایش کیفیت آموزشی مدارس [۹]، شناسایی نقاط قوت و نقاط ضعف دانش‌آموزان [۱۰] را انجام داد و برای رسیدن به موفقیت در جهت درستی گام برداشت.

از طرفی دیگر وجود مجموعه داده مناسب در زمینه داده‌کاوی حوزه‌ی آموزش از اهمیت بالایی برخوردار است. برای این منظور محققان ویژگی‌های مختلفی نظیر جمعیت‌شناختی [۱۱, ۱۲]، تحصیلی [۱۰, ۱۳]، خانوادگی [۱۱, ۱۴]، اجتماعی [۱۱, ۱۴] 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 [۱۵]. با بررسی‌های انجام شده می‌توان متوجه شد که روش طبقه‌بندی در اکثر مطالعات انجام شده مورد استفاده قرار گرفته است [۱۳]. 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 [۱۲].

در این مقاله نیز هدف بررسی عوامل مؤثر بر موفقیت دانش‌آموزان پایه دهم آموزشگاه‌های شهر کاشمر با استفاده از الگوریتم‌های داده‌کاوی است. همین‌طور برای جمع‌آوری داده‌های موردنیاز از پرسش‌نامه‌ای حاوی اطلاعات چرخ تعادل زندگی، جمعیت‌شناختی 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 تأثیرگذار بر موفقیت تحصیلی ارائه دهد.
- استفاده از چهار روش انتخاب ویژگی که جزو دسته‌های رپر و فیلتر هستند برای شناسایی مهم‌ترین عوامل تأثیرگذار بر موفقیت تحصیلی. استفاده از رویکرد یادگیری ماشین کمک می‌کند تا تاثیر متقابل یا هم‌افزایی ویژگی‌ها بخصوص در مسائل پیچیده بهتر کشف شود.

---

[1] Naïve Bayes

[2] Multilayer Perceptron

[3] support vector machine

[4] k-nearest neighbors

[5] Random Forest

- پیش‌بینی معدل پایانی دانش‌آموزان پایه دهم با بهره‌گیری از ویژگی‌های استخراج‌شده و مدل‌های متداول یادگیری ماشین.

این مقاله از قسمت‌های زیر تشکیل شده است. بخش ۲ مروری بر ادبیات داده‌کاوی تحصیلی ارائه می‌کند. بخش ۳ به مروری بر تحقیقات مرتبط با موضوع می‌پردازد. بخش ۴ و ۵ به ترتیب روش مورد استفاده و نتایج را بیان می‌کند. در نهایت، بخش ۶ نتیجه‌گیری و تفسیر نتایج را نشان می‌دهد.

۲. داده‌کاوی تحصیلی

فرآیند تبدیل داده‌های خام تحصیلی دانش‌آموزان به اطلاعات مفید که می‌تواند تأثیر زیادی بر تحقیقات و شیوه‌های آموزش دانش‌آموزان و محصلان داشته باشد، داده‌کاوی تحصیلی گفته می‌شود [۱۶]. در این فرایند ابتدا داده‌های لازم از سیستم‌های آموزشی دریافت و سپس از روش‌های داده‌کاوی برای استخراج دانش، کشف عوامل مؤثر بر عملکرد، در مواردی پیش‌بینی برخی پارامترهای مهم و در کل بهبود کیفیت آموزشی استفاده می‌شود [۱۷]. داده‌کاوی در حوزه آموزش و پرورش، بیشتر بر تجزیه و تحلیل فرایند یادگیری و رفتار دانش‌آموزان، استراتژی‌ها و برنامه‌های آموزشی، پیش‌بینی عملکرد، موفقیت و ترک تحصیل دانش‌آموزان تمرکز دارد [۱۸] 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. درنظرگرفتن این عوامل در آموزش ممکن است عملکرد دانش‌آموزان را در مباحث آموزشی به‌سرعت بهبود بخشد [۱۹]. 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 [۱۰].

۳. تحقیقات مرتبط

با نگاهی به مطالعات انجام شده می‌توان دریافت که محققان در حوزه‌های مختلفی از داده‌کاوی تحصیلی استفاده کرده‌اند. به‌طوری 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 [۱۱, ۱۲]، پیشینه خانوادگی [۱۱, ۱۴]، اطلاعات تحصیلی [۱۰, ۱۳]، ویژگی‌های رفتاری [۱۲, ۲۰] و فعالیت‌های اجتماعی [۱۱, ۱۴] طبقه‌بندی کرد. در این بین محققان از پرسش‌نامه [۱۴, ۲۱]، 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 به‌کارگیری در الگوریتم‌های داده‌کاوی آماده‌سازی شوند [۱۸].

در مرحله پیش‌پردازش ابتدا می‌بایست به پاک‌سازی داده‌ها پرداخت. این فرایند عموماً شامل حذف داده‌های گمشده یا ناقص است که معمولاً در همه پژوهش‌ها دیده می‌شود.

حذف ویژگی‌های نامرتبط از دیگر روش‌های پیش‌پردازش است. دراین خصوص محققین ویژگی‌هایی نظیر نام دانشجو یا دانش‌آموز، ملیت، نام دانشگاه [۱۰]، نام شهر، معدل ترم [۱۳]، دروس انتخابی [۸] و جنسیت [۲۲] را از بین سایر ویژگی‌ها بر اساس هدف موردمطالعه خود حذف می‌کنند.

در گام بعدی از پیش‌پردازش، سعی می‌شود تا با انتخاب ویژگی‌های مؤثر، 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 [۱۴]. در این مرحله، انتخاب ویژگی می‌تواند بر اساس همبستگی [۱۸, ۲۳]، فیلتر کردن و رتبه‌بندی [۱۴, ۲۴] انجام شود. همین‌طور در مطالعات [۶, ۷, ۱۵, ۱۸] از روش بسته‌بندی و در مطالعات [۱, ۱۸, ۱۹] 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 [۷, ۱۴, ۱۵, ۱۹]، نرمال‌سازی [۱۹, ۲۳]، جایگزینی [۸, ۱۳]، تبدیل ویژگی‌های عددی به اسمی [۱۹]، تبدیل ویژگی‌های اسمی به عددی [۲۵] و کدگذاری [۲۶] استفاده کرده‌اند. از روش نمونه‌گیری نیز در برخی مطالعات استفاده شده است [۱۸].

پس از جمع‌آوری داده‌ها و پیش‌پردازش، نوبت به انتخاب روش داده‌کاوی می‌رسد تا در نهایت مدل موردنظر ایجاد شود. در این مرحله روش‌های داده‌کاوی روی مجموعه داده‌های مرحله قبل برای ساخت مدل و استخراج الگوهای مهم و جالب توجه به کار می‌روند.

پیش‌بینی عملکرد تحصیلی دانش‌آموزان یکی از حوزه‌هایی است که همواره موردتوجه محققین قرار گرفته است. درخت تصمیم یکی از روش‌هایی است که در این زمینه به دلیل درک و تفسیر آسان، شناسایی سریع متغیرهای تأثیرگذار و روابط بین آنها مورد استفاده قرار می‌گیرد [۲۷]. در مطالعات [۲, ۲۳] از روش‌های درخت تصمیم به ترتیب روی اطلاعات ۱۳۳ و ۱۷۵ دانش‌آموز استفاده شده است و نتایج آنها نشان می‌دهد که درخت تصمیم با دقتی بین ۶۵ تا ۷۵ درصد، عملکرد بهتری دارد.

در مطالعه [۲۲] نیز تکنیک‌های مدل‌سازی KNN، بیز ساده، درخت تصمیم و رگرسیون لجستیک به کار گرفته شده است و نتایج نشان می‌دهد بیز ساده با دقت ۸۹ درصد در مقایسه با سایر الگوریتم‌ها، بهتر عمل می‌کند.

از سویی دیگر از الگوریتم‌های بیز ساده، پرسپترون چندلایه، درخت تصمیمJ48 و جنگل تصادفی برای سنجش میزان اهمیت ویژگی‌های سابقه تحصیلی، فعالیت‌های اجتماعی و پیشرفت تحصیلی در بین ۳۹۵ دانش‌آموز استفاده شده است [۲۸]. در این میان نتایج نشان می‌دهد که سابقه تحصیلی و فعالیت‌های اجتماعی در پیش‌بینی عملکرد دانش‌آموز بیشترین تأثیر را داشته‌اند.

همچنین مطالعات اخیر نشان می‌دهد که الگوریتم‌های شبکه عصبی، نتایج خوبی در پیش‌بینی عملکرد دانش‌آموزان از خود نشان داده‌اند [۲۷]. به‌طوری‌که در [۲۶] با اجرای الگوریتم‌های شبکه عصبی روی اطلاعات ۳۵۱۸ دانش‌آموز ، دقت پیش‌بینی عملکرد ۸۰,۴۷ درصد ثبت شده است. همین‌طور در [۲۴] پیش‌بینی عملکرد با دقتی بالای ۹۱ درصد مشاهده می‌شود.

در [۱۹] 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 به‌عنوان 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[۶] ، ۲ الی ۳ درصد بهبود می‌یابد. لازم به ذکر

[۶] Binary Genetic Algorithm

است که عملکرد شبکه‌های عصبی کانولوشنی با دقت ۹۵ درصد از همه روش‌های گفته شده بهتر بوده است.

از طرفی دیگر پیاده‌سازی الگوریتم‌های شبکه عصبی روی ۲۴۱ نمونه با استفاده از نمرات امتحانات و داده‌های سیستم‌های آنلاین مدیریت یادگیری، موفقیت بیشتری را نسبت به الگوریتم‌های درخت تصمیم به‌دست آورده است [۱۸]. همین‌طور در مجموعه داده‌ای متشکل از ۴۸۰ نمونه، شبکه عصبی با دقت ۷۶ درصد از بین الگوریتم‌های J48، SVM، جنگل تصادفی و بیز ساده عملکرد بهتری داشته است [۱۲].

با بررسی مطالعات بیشتر می‌توان متوجه شد که از مدل‌های داده‌کاوی برای پیش‌بینی معدل دانش‌آموزان نیز استفاده می‌شود. به‌طور نمونه مطالعه [۱۰] با بررسی نمرات ۲۳۶ 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 ۵ کلاس(معدل‌های ۲۰-۱۶ کلاس ۱ ، معدل‌های ۱۵-۱۴ کلاس ۲، معدل‌های ۱۳-۱۲ کلاس ۳ ، معدل‌های ۱۱-۱۰ کلاس ۴ و معدل‌های کمتر از ۱۰ کلاس ۵) با دقت ۸۵,۶۲ درصد و الگوریتم جنگل تصادفی در ۲ کلاس (موفقیت یا شکست) با دقت ۹۳,۴۹ درصد عملکرد خوبی داشتند [۱۵].

دسته دیگری از تحقیقات در زمینه پیش‌بینی معدل به مقایسه ۱۵ الگوریتم داده کاوی، روی اطلاعات ۵۵۶۶ نمونه پرداخته اند و نشان داده‌اند دو الگوریتم Naive Bayes و درخت تصمیم Hoeffding tree با دقت ۹۱ درصد نتایج بهتری دارند [۱۳].

در [۷] نیز الگوریتم بیز ساده با دقتی بین ۶۳,۳۳ تا ۶۹,۶۷ درصد بهتر از الگوریتم‌های مبتنی بر درخت عمل می‌کند. اما در شناسایی دانش‌آموزان نمونه، جنگل تصادفی با دقتی بین ٪۸۵,۸ تا ٪۹۲,۶ 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 ۱۴۵ نمونه مورد مقایسه قرار گرفتند. نتایج نشان داد که رگرسیون SMOReg با دقت ۹۶,۹۸ درصد عملکرد بهتری نسبت به سایر الگوریتم‌ها داشته است [۸].

تعدادی از محققین نیز به دنبال بهبود دقت الگوریتم‌های داده‌کاوی از ترکیب الگوریتم‌ها برای مدل‌سازی و ارزیابی عملکرد دانش‌آموزان استفاده کرده‌اند. به‌عنوان مثال [۲۰] الگوریتم‌های طبقه‌بندی SVM، بیز ساده، درخت تصمیم، شبکه عصبی و الگوریتم خوشه‌بندی K-میانگین را با یکدیگر ترکیب کرده و به دقت ۷۵ 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 ۹ مدل طبقه‌بندی جدید را ایجاد کرده‌اند. نتایج ارزیابی آنها روی ۱۲۲۷ نمونه نشان داد که بهترین الگوریتم استفاده ترکیبی دو الگوریتم MULTIBOOST و پرسپترون چندلایه با دقت ٪۹۸,۷ است [۱۴].

محققین همین‌طور به نقش روش‌های مختلف داده‌کاوی در شناسایی عوامل تأثیرگذار بر افت تحصیلی دانش‌آموزان پرداخته‌اند. در این زمینه، آقای نجفی و همکاران [۲۵] از الگوریتم‌های درخت تصمیم J48 و خوشه‌بندی K-میانگین و استخراج قوانین[۷] استفاده کرده است. در این میان J48 با دقت ۹۵ درصد و خوشه‌بندی K-میانگین با ضریب اطمینان ۹۵ درصد بهتر عمل می‌کنند.

---

[۷] Apriori

با بررسی مقاله‌های مورد مطالعه می‌توان متوجه شد که محققین در انتخاب ویژگی‌هایی موردنظر 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. ازاین‌رو با جستجوی کلیدواژه «ابعاد مهم زندگی یک فرد» به ابزار «چرخ تعادل زندگی» از آقای جی.مایر[8] می‌توان رسید. ابزار «چرخ تعادل زندگی» برای اولین‌بار با مفهوم تبدیل‌شدن به یک «فرد کامل» توصیف شد و بر شش حوزه حیاتی ۱- جسمی و فیزیکی ۲- امور مالی و شغلی، ۳- فکری ۴-احساسی ۵-اجتماعی ۶- معنویت تمرکز دارد [۲۹].

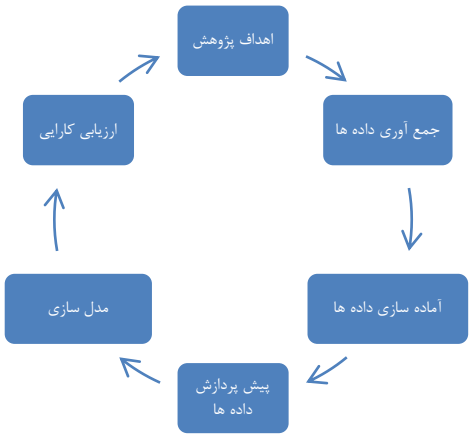
در این مطالعه سعی شده است تا میزان مؤثربودن ویژگی‌های چرخ تعادل زندگی در موفقیت تحصیلی دانش‌آموزان بررسی شود. برای این منظور ابتدا پرسش‌نامه چرخ تعادل زندگی در اختیار دانش‌آموزان قرار گرفته و برای افزایش دقت، اطلاعات دیگری مانند اطلاعات جمعیت‌شناختی و پیشینه تحصیلی آنها نیز جمع‌آوری شد. سپس معدل به‌عنوان معیاری برای موفقیت تحصیلی دانش‌آموزان در نظر گرفته شد. درنهایت از الگوریتم‌های داده‌کاوی برای پیداکردن ارتباطی بین چرخ تعادل زندگی، اطلاعات جمعیت‌شناختی و اطلاعات تحصیلی با موفقیت تحصیلی استفاده شده است. نتایج به‌دست‌آمده 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 به‌عنوان معیاری برای پیش‌بینی موفقیت تحصیلی دانش‌آموزان در رشته موردنظر را می‌توان جزو نوآوری‌های این مقاله در نظر گرفت.

[8] Paul J. Meyer

۴. 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 ۱ بیان و در ادامه هر کدام از این مراحل توضیح داده می‌شوند:



شکل ۱ - مراحل روش پیشنهادی

۱,۴. اهداف پژوهش

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

۱. بررسی عوامل مؤثر بر موفقیت تحصیلی دانش‌آموزان پایه دهم متوسطه؛
۲. جمع‌آوری داده‌ها با استفاده از پرسش‌نامه استاندارد چرخ تعادل زندگی، اطلاعات جمعیت‌شناختی و پیشینه تحصیلی؛
۳. 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. ازاین‌رو می‌توان

برای انتخاب رشته مناسب آنها برنامه‌ریزی‌های بهتری را انجام داد.

۲,۴. جمع‌آوری داده‌ها

در این مقاله برای جمع‌آوری داده‌ها از پرسش‌نامه استفاده شده است. در تهیه پرسش‌نامه علاوه بر سؤالات چرخ تعادل زندگی [۳۰]، با نظر مشاورین مدرسه و متخصصان حوزه‌ی آموزش و پرورش، اطلاعات جمعیت‌شناختی و پیشینه تحصیلی دانش‌آموزان (شامل ویژگی‌های جنسیت، وضعیت تأهل، 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 به‌صورت کاغذی و الکترونیکی در اختیار دانش‌آموزان پایه دهم قرار گرفت. در نهایت اطلاعات ۴۹۰ دانش‌آموز با ۴۳ ویژگی جمع‌آوری شد. پرسشنامه پیوست می‌باشد.

۳,۴. آماده‌سازی و پیش‌پردازش داده‌ها

در مرحله آماده‌سازی، ابتدا تعداد ۱۷ پرسش‌نامه ناقص حذف و در گام بعدی ازآنجایی‌که مقادیر برخی از ویژگی‌ها به‌صورت متن وارد شده‌اند، می‌بایست برای استفاده در الگوریتم‌ها و اندازه‌گیری امتیاز هر بخش، به مقادیر عددی تبدیل شوند. جدول ۱ مقادیر عددی اختصاص یافته به هر پاسخ متنی را نشان می‌دهد. البته در بخش رشته تحصیلی، کد ۴ 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 ۲ 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 ۲ – دسته‌بندی معدل

معدل	دسته	تعداد نمونه
۲۰–۱۸	۱	۱۸۷
۱۸–۱۶	۲	۱۹۳
۱۶–۰	۳	۹۳

بعد از انجام آماده‌سازی، مجموعه داده پیشنهادی با ۴۷۳ 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 توسعه‌یافته است. WEKA، برنامه مناسبی برای انجام پردازش‌ها و تکنیک‌های مختلف داده‌کاوی روی داده‌های بزرگ است.

جدول ۱ – مقادیر عددی اختصاص‌یافته به پاسخ‌های متنی

ویژگی	توضیحات
جنسیت	مرد:۰ زن:۱
وضعیت تأهل	مجرد:۰ متأهل:۱
تحصیلات پدر و مادر	سیکل و پایین‌تر:۱ دیپلم:۲ فوق‌دیپلم:۳ کارشناسی:۴ کارشناسی‌ارشد:۵ دکترا:۶
میزان پول‌توجیبی	کمتر از ۵۰ هزار تومان:۱ بین ۵۰–۱۰۰ هزار تومان:۲ بین ۳۰۰–۱۰۰ هزار تومان:۳ بیشتر از ۳۰۰ هزار تومان:۴
میزان مطالعه	بیشتر از ۴ ساعت:۱ بین ۳–۱ ساعت:۲ کمتر از ۱ ساعت:۳
وسایل کمک‌آموزشی	هیچ‌کدام:۰ فیلم آموزشی و پادکست:۱ کتاب‌های کمک‌آموزشی:۲ کلاس خصوصی:۳
رشته تحصیلی	تجربی:۱ انسانی:۲ گرافیک:۳ طراحی دوخت:۵ عکاسی:۶ کامپیوتر:۷ مدیریت خانواده:۸ کشت گیاهان دارویی:۹ مکانیک:۱۰

در پیش‌پردازش، روی مجموعه داده پیشنهادی، فرایندهایی انجام می‌شود تا داده‌ها، مناسب استفاده برای مدل موردنظر شوند. ازاین‌رو در این مقاله از روش‌های استانداردسازی داده و انتخاب ویژگی به ترتیب برای افزایش دقت پیش‌بینی و مشخص‌کردن ویژگی‌هایی با تأثیر 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 تأثیر بیشتری دارند، استفاده می‌شود. جدول ۳ الگوریتم‌های انتخاب ویژگی استفاده شده را بیان می‌کند.

جدول ۳– روش‌های انتخاب ویژگی

روش‌های انتخاب ویژگی	کاربرد
WrapperSubsetEval (Wrapper)	این روش مجموعه را با استفاده از یک طرح یادگیری ارزیابی می‌کند [۱۵].
CfsSubsetEval (Wrapper)	در این روش ویژگی‌هایی که همبستگی بالایی با ویژگی هدف دارند و درعین‌حال همبستگی کمتری با هم دارند، انتخاب می‌شوند [۷].
CorrelationAttributeEval (Filter)	میزان تأثیر ویژگی را با استفاده از اندازه‌گیری همبستگی بین ویژگی و ویژگی هدف ارزیابی می‌کند [۷].
InfoGainAttributeEval (Wrapper)	این روش تأثیر یک ویژگی را با اندازه‌گیری میزان سودمندی مرتبط با ویژگی هدف ارزیابی می‌کند [۷].

۴،۴. 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 می کند.؛ استفاده می‌کند[۳۱]. 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 [۲۷].
- **الگوریتم ماشین بردار پشتیبان**[9]: یک الگوریتم یادگیری نظارت‌شده می‌باشد که برای طبقه‌بندی و رگرسیون می توان از آن استفاده نمود. این روش در واقع ابرصفحه‌ای بهینه را برای جدا کردن نمونه‌های کلاس‌های مختلف از هم در فضای ویژگی می‌یابد. به‌عبارت 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 بشینه کردن آن

[9]support vector machine (SVM)

است) می‌باشد. روش ماشین بردار پشتیبان برای مسائل خطی و غیرخطی مؤثر بوده و به دلیل استحکام، کارایی برای داده‌ها با ابعاد بالا و توانایی مدیریت روابط پیچیده بین ویژگی‌ها شناخته می‌شود [۳۲].

- **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 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 ازدست‌رفته، پیش‌بینی‌های بسیار دقیقی را انجام می‌دهد [۱۵].
- **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 [۳۲]. در این مطالعه الگوریتم KNN با درنظر گرفتن تعداد ۷ 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 ازدست‌رفته، هرس درخت تصمیم و استخراج قوانین را دارد [۱۲]. این الگوریتم با یک مجموعه اصلی مانند S به‌عنوان 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 نگرفته اند، محاسبه و در این میان ویژگی با کمترین مقدار آنتروپی یا بیشترین بهره اطلاعات انتخاب می‌شود. سپس براساس ویژگی انتخاب شده، زیرمجموعه‌هایی از داده‌ها تولید می‌شود. الگوریتم همچنان با استفاده از ویژگی‌هایی که تاکنون انتخاب نشده‌اند به‌صورت بازگشتی به فرایند تقسیم کردن ادامه می‌دهد تا درخت مورد نظر تشکیل شود. در درخت تصمیم j48 ، برگ‌های درخت تصمیم، ویژگی هدف را نشان می‌دهند [۱۵]. در این مطالعه از الگوریتم J48 با درنظر گرفتن ضریب اطمینان ۰,۹ جهت ارایه بهتر رویدادهای ممکن با انجام کمترین هرس و درنظر گرفتن حداقل ۲۰ نمونه در هر برگ به‌کار 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)، به‌عنوان ورودی‌های لایه بعدی (لایه پنهان) استفاده می‌کند تا زمانی که پس از تعداد خاصی از لایه‌ها، خروجی‌های آخرین لایه پنهان به‌عنوان ورودی‌های لایه خروجی مورد استفاده قرار می‌گیرد و لایه خروجی وظایفی مانند طبقه‌بندی و پیش‌بینی را انجام می‌دهد [۲۴].

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 ۱۰۰ مرتبه تکرار انجام می‌دهد.

۵،۴. ارزیابی

در ابتدا باید بیان کرد که روایی پرسشنامه از طریق روایی محتوایی، و پایایی پرسشنامه از طریق آزمون آلفای کرونباخ با مقدار ۰،۸۲ تایید گردید. در ادامه به بیان شاخص‌های ارزیابی مدل‌های داده‌کای می‌پردازیم.

ازآنجایی‌که در داده‌کاوی معمولاً 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 معیارF1 [۱۲, ۱۴, ۱۹, ۳۳]
- ماتریس درهم‌ریختگی [۱۴, ۱۷, ۱۹, ۲۳]
- منحنی ROC [۱۸, ۳۳]

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 [۱۰]استفاده شده است. در این روش مجموعه داده‌ها به طور تصادفی به k زیرمجموعه با حجم یکسان تفکیک شده و در هر مرحله، تعداد k-1 از این زیرمجموعه‌ها به‌عنوان داده‌های آموزش و یک زیرمجموعه به‌عنوان داده آزمایش در نظر گرفته می‌شود [۱۴]. ازاین‌رو 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 به‌صورت زیر مجاسبه می‌شود:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ , \qquad (۱)$$

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

$$Precision = \frac{TP}{TP+FP}\ , \qquad (۲)$$

که نسبت نمونه‌های مثبت واقعی به همه نمونه‌هایی که مدل مثبت پیش‌بینی کرده است را نشان می‌دهد. رابطه‌ی زیر معیار پوشش یا حساسیت را محاسبه می‌کند:

$$Recall = \frac{TP}{TP+FN}\ , \qquad (۳)$$

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. معیار F1 عبارتست از:

$$F1\_measure = \frac{2*(Percision\times Recall)}{Percision+Recall}\ , \qquad (۴)$$

---

[۱۰] K-fold cross validation

با توجه به اینکه دو معیار صحت و پوشش عکس یکدیگر عمل می‌کنند به‌طوری‌که 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.

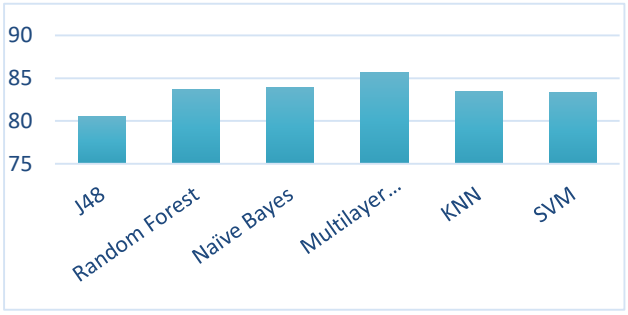
۵. نتایج

در این قسمت ابتدا الگوریتم‌های گفته شده روی مجموعه داده پیشنهادی بدون انتخاب ویژگی‌های مؤثر، اجرا و سپس نتایج مدل‌های به‌دست‌آمده با معیارهای گفته شده در بالا اندازه‌گیری می‌شوند. همین‌طور در ادامه فرایند انتخاب ویژگی انجام و ضمن بررسی ویژگی‌های تأثیرگذار، به عملکرد الگوریتم‌های داده‌کاوی روی ویژگی‌های انتخاب شده در هر یک پرداخته شده است. هدف از این کار نشان‌دادن اهمیت فرایند انتخاب ویژگی در بهبود نتایج پیش‌بینی است. جدول ۴ نتایج الگوریتم‌های داده‌کاوی را قبل از فرایند انتخاب ویژگی نشان می‌دهد.

جدول ۴ – مقایسه معیارهای ارزیابی در الگوریتم‌های مطرح شده

الگوریتم	دقت	صحت	پوشش	F1
J48	٪۸۰,۵۴	۰,۸۱۱	۰,۸۰۵	۰,۸۰۸
جنگل تصادفی	٪۸۳,۷۲	۰,۸۴۲	۰,۸۳۷	۰,۸۳۹
بیز ساده	٪۸۳,۹۳	۰,۸۳۹	۰,۸۳۹	۰,۸۳۹
پرسپترون چندلایه	٪۸۵,۶۲	۰,۸۶۰	۰,۸۵۶	۰,۸۵۸
KNN	٪۸۳,۵۰	۰,۸۳۷	۰,۸۳۵	۰,۸۳۴
SVM	٪۸۳,۲۹	۰,۸۳۵	۰,۸۳۳	۰,۸۳۴

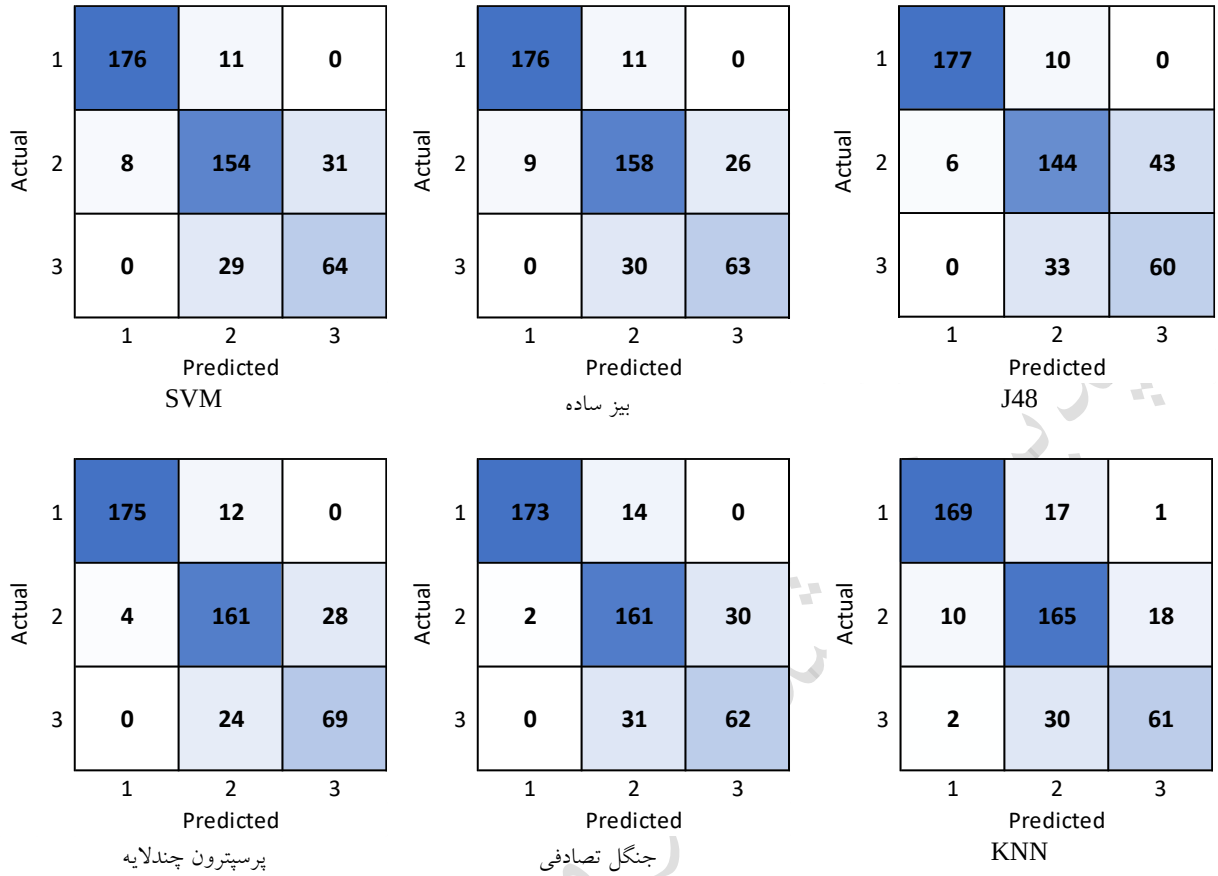
بررسی نتایج نشان می‌دهد که الگوریتم شبکه عصبی با دقت نزدیک به ۸۵ درصد، عملکرد بهتری نسبت به سایر الگوریتم‌ها دارد. این یافته را می‌توان در شکل ۳ بهتر مقایسه و بررسی کرد.



شکل ۳- مقایسه دقت الگوریتم‌ها

ماتریس درهم‌ریختگی یکی از ابزارهایی است که برای نمایش عملکرد الگوریتم‌های موردنظر 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 به‌درستی دسته‌بندی نشده‌اند [۶]. شکل ۴ 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 [۳۴].



شکل ۴ – ماتریس‌های درهم ریختگی مدل‌ها قبل از انتخاب ویژگی

جدول ۵ مساحت زیر منحنی‌های ROC را برای هر کلاس در مدل‌های مطرح شده در این مقاله نشان می‌دهد:

جدول ۵ – مقدار AUC برای هر کلاس در مدل‌ها

	J48	Native Bayes	SVM	KNN	Random Forest	Multilayer Perceptron
کلاس ۱	۰,۹۸	۰,۹۸	۰,۹۸	۰,۹۷	۰,۹۸	۰,۹۸
کلاس ۲	۰,۹۱	۰,۹۲	۰,۸۳	۰,۸۹	۰,۹۲	۰,۹۲
کلاس ۳	۰,۹۲	۰,۹۴	۰,۹۲	۰,۹۳	۰,۹۴	۰,۹۴

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. به‌طوری‌که تعداد نمونه‌های درست پیش‌بینی شده در آن به واقعیت نزدیک‌تر و تعداد نمونه‌های مثبت کاذب در آن بسیار کم است.

۱,۵. انتخاب ویژگی‌های تأثیرگذار

هدف از این بخش انتخاب تأثیرگذارترین 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 تأثیر بگذارد. ازاین‌رو، از چندین الگوریتم انتخاب ویژگی مشهور به‌منظور انتخاب ویژگی‌های تأثیرگذار در پیش‌بینی معدل پایه دهم استفاده شد. جدول ۶ نتایج الگوریتم‌های انتخاب ویژگی را نشان می‌دهد.

جدول ۶ – ویژگی‌های انتخاب شده در الگوریتم‌های انتخاب ویژگی

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

همان‌طور که در جدول ۳ گفته شد هر کدام از الگوریتم‌های بالا، ویژگی‌ها را براساس معیارهای مشخصی انتخاب می‌کنند. از این رو الگوریتم‌های گفته شده در جدول ۶ به ترتیب ویژگی‌ها را از نظر ارزیابی با یک طرح یادگیری، همبستگی ویژگی با کلاس با درنظر 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 ۶ را از نظر ابعاد گفته شده بررسی کرد.

با توجه به نتایج جدول ۶، ویژگی‌های «پول‌توجیبی»، «میزان مطالعه»، «جنسیت» و «رشته تحصیلی» در تمام الگوریتم‌ها به عنوان ویژگی‌های مؤثر انتخاب شده‌اند. این نشان می‌دهد که ویژگی‌های گفته شده از تأثیر 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 ۸، در میان ویژگی‌های مرتبط با چرخ زندگی نیز می‌توان نتایج زیر را به دست آورد:

- بعد روحی با ویژگی هدف براساس 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 مؤثر، نوبت به مشخص‌کردن 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 تأثیرگذاری بیشتری دارند، توسط الگوریتم‌های انتخاب ویژگی مشخص (جدول ۶) و به دنبال آن سایر ویژگی‌ها از مجموعه داده حذف می‌شوند. در نهایت شبکه عصبی با روش اعتبارسنجی متقابل ۱۰–بخشی روی ویژگی‌های باقیمانده یادگیری را انجام می‌دهد. این فرایند برای هر الگوریتم انتخاب ویژگی انجام می‌شود. جدول ۷ نتایج آنچه گفته شد را نشان می‌دهد.

آنچه که از جدول‌های ۶ و ۷ برمی‌آید الگوریتم انتخاب ویژگی

WrapperSubsetEval توانسته مجموعه ویژگی‌های تأثیرگذارتری را نسبت به سایر الگوریتم‌ها انتخاب کند.

جدول ۷- عملکرد شبکه عصبی پس از انتخاب ویژگی

الگوریتم انتخاب ویژگی	عملکرد شبکه عصبی			
	Acuuracy	Precision	Recall	F1
WrapperSubsetEval	٪۸۵,۲۰	٪۸۵,۳	٪۸۵,۲	٪۸۵,۳
CfsSubsetEval	٪۸۴,۱۴	٪۸۴,۳	٪۸۴,۱	٪۸۴,۲
CorrelationAttributeEval	٪۸۳,۵۰	٪۸۳,۸	٪۸۳,۵	٪۸۳,۵
CfsSubsetEval	٪۸۱,۳۹	٪۸۱,۶	٪۸۱,۴	٪۸۱,۵

همین‌طور که از مقایسه جدول ۴ و جدول ۷ مشاهده می‌فرمایید، 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.

نتیجه‌گیری

امروزه کاربرد داده‌کاوی در مراکز آموزشی از اهمیت زیادی برخوردار است. دانش به‌دست‌آمده 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 ۴۷۳ 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 به‌کار 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 ۸۵ 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، تأثیرات مثبت داده‌کاوی تحصیلی در مدارس را روشن‌تر و تصمیم‌گیری‌های کلان آموزشی را آسان‌تر خواهد کرد.

سپاسگزاری

این تحقیق حاصل پایان نامه کارشناسی ارشد می باشد و نویسندگان از تمامی کسانی که در انجام این تحقیق یاری رساندند بویژه اداره آموزش و پرورش شهرستان کاشمر تشکر و قدردانی می نمایند. آنها همچنین از اعضای هیئت تحریریه مجله علمی محاسبات نرم به خاطر بینش و همکاری ارزشمندشان تشکر می کنند.

مراجع

[1] S. Hussain and M. Q. Khan, "Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning," *Annals of data science*, 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," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019, doi: 10.14569/IJACSA.2019.0100739.

[۳] خلف بیگی، امید، بشیری موسوی، سید علیرضا، قارلقی، سینا، «بکارگیری مدل تحلیل احساسات در سطح حروف مبتنی بر شبکه عصبی روی نظرات فارسی ثبت شده در شبکه‌های اجتماعی و فروشگاههای اینترنتی»، مجله محاسبات نرم، جلد ۱۱، شماره ۲، ص ۱۱۸-۱۳۳، پاییز و زمستان ۱۴۰۱.

[14] A. Siddique, A. Jan, F. Majeed, A .I. Qahmash, N. N. Quadri, and M. O. A. Wahab, "Predicting academic performance using an efficient model based on fusion of classifiers," *Applied Sciences*, 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 *2022 International Conference on Decision Aid Sciences and Applications (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," *Education and Information Technologies*, 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," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 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 *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 2018, pp. 1-5: IEEE.

[20] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of medical systems*, vol. 43, pp. 1-15, 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," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 3, pp. 75-82, 2020.

[22] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, 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.

[4] M. Eftekharian, 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," *American journal of pharmaceutical education*, vol. 85, no. 4, 2021.

[6] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, pp. 1-2,2020, 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," *Journal of Information Technology Education: JITE. Innovations in Practice: IIP*, 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 *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2019, pp. 1-5: IEEE.

[9] A. E .Tatar and D. Düştegör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?," *Applied sciences*, 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," *International journal of information and education technology*, vol. 6, no. 7, p. 528, 2016, doi: 10.7763/IJIET.2016.V6.745.

[11] H. Turabieh, "Hybrid machine learning classifiers to predict student performance," in *2019 2nd international conference on new trends in computing sciences (ICTCS)*, 2019, pp. 1-6: IEEE.

[12] C. Jalota and R. Agrawal, "Analysis of educational data mining using classification," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, ۲۰۱۹ ,2 pp. 243-247: IEEE.

[13] N. Alangari and R. Alturki, "Predicting students final GPA using 15 classification algorithms," *Romanian Journal of Information Science and Technology*, vol. 23, no. 3, pp. 238-249, 2020.

[23] F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1-5: IEEE.

[24] T. Ahajjam, M. Moutaib, H. Aissa, M. Azrour, Y. Farhaoui, and M. Fattah, "Predicting students' final performance using artificial neural networks," *Big Data Mining and Analytics*, vol. 5, no. 4, pp. 294-301, 2022, doi: 10.26599/BDMA.2021.9020030.

[۲۵] م. نجفی، م. افضلی، م. مرادی، «کاربرد داده کاوی آموزشی جهت شناسایی عوامل موثر بر افت تحصیلی دانش آموزان»، فصلنامه سیستمهای پردازشی و ارتباطی چندرسانهای هوشمند، سال ۱، شماره ه ۱، ص ۳۳-۲۳، بهار۱۴۰۰.

[26] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Education and Information Technologies*, 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," *Computers & Electrical Engineering*, 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 *2022 International Conference on Development and Application Systems (DAS)*, 2022, pp. 178-185: IEEE, doi: 10.1109/DAS54948.2022.9786194.

[30] G. M. Robin Robertson, "Life Balance Assessment and Action Planning Guide," 2001-2002.

[۳۱] زارع مهرجردی، فاطمه، یزدیان دهکردی، مهدی، لطیف، علی محمد، «ارزیابی روش‌های یادگیری کلاسیک و یادگیری عمیق در تجزیه و تحلیل احساسات داده‌های تلگرام فارسی», مجله محاسبات نرم، جلد ۱۱ ، شماره ۱، ص ۶۶-۱۰۵، بهار و تابستان ۱۴۰۱.

[32] H. Al-Shehri *et al.*, "Student performance prediction using support vector machine and k-nearest neighbor," in *2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE)*, 2017, pp. 1-4: IEEE.

[33] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, 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 Politécnica Nacional," *Frontiers in Psychology*, vol. 11, p. 515531, 2020, doi: 10.3389/fpsyg.2020.515531.

پیوست: پرسشنامه

۱. جنسیت: مرد ☐ زن ☐

۲. وضعیت تأهل: مجرد ☐ متأهل ☐

۳. تحصیلات پدر:

سیکل و پایین‌تر☐ دیپلم ☐ کارشناسی (لیسانس)☐

کارشناسی ارشد (فوق لیسانس) ☐ دکترا ☐

۴. تحصیلات مادر:

سیکل و پایین‌تر☐ دیپلم ☐ کارشناسی(لیسانس)☐

کارشناسی ارشد(فوق لیسانس) ☐ دکترا☐

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

بیشتر از ۱۰۰ هزار تومان ☐ بین ۱۰۰-۵۰ هزار تومان☐ کمتر از ۵۰ هزار تومان☐

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

بیشتر از ۴ساعت ☐ بین ۳-۱ ساعت ☐ کمتر از ۱ ساعت☐

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

کلاس‌های خصوصی☐ کتاب‌های کمک‌آموزشی☐ فیلم های آموزشی و پادکست☐

۸. معدل سال دهم:

۹. رشته تحصیلی خود را بنویسید:

۱۰. نام مدرسه خود را بنویسید:

	ردیف	سؤال	کاملاً موافق	موافق	مخالف	کاملاً مخالف
فیزیکی	۱۱.	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.				
	۱۲.	من در هفته چند بار حرکات ورزشی به مدت حداقل ۲۰ دقیقه (همراه با بالارفتن ضربان قلب) انجام می‌دهم.	یک‌بار	دو بار	سه بار	چهار بار و بیشتر
	۱۳.	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.				
	۱۴.	من به‌طورکلی بیماری نگران‌کننده‌ای (چه جسمی چه روحی) ندارم.				
	۱۵.	من سلامتی خود را به‌صورت دوره‌ای و مرتب بررسی می‌کنم.				
	۱۶.	من دخانیات و هر چیزی که برای بدن ضرر دارد را مصرف نمی‌کنم.				
ی	۱۷.	من استقلال مالی دارم و به‌اندازه پول‌توجیبی خود خرج می‌کنم.				

	۱۸.	من برای مواقع ضروری در آینده، مبلغی را پس‌انداز می‌کنم.				
	۱۹.	من انسان خوش‌حسابی هستم.				
	۲۰.	بین هزینه‌های فعلی و پس‌انداز برای آینده تعادل ایجاد می‌کنم.				
	۲۱.	خرج‌کردن من بر اساس نیازهای اولویت‌های خودم می‌باشد نه تقلید از دیگران.				
فکری	۲۲.	من از یادگیری مهارت‌ها و به‌دست‌آوردن دانش جدید لذت می‌برم.				
	۲۳.	من افکار مثبت دارم.				
	۲۴.	من به‌طورکلی از رشته تحصیلی خود راضی هستم.				
	۲۵.	من زمان و انرژی خود را صرف رشد تحصیلی و مهارت‌آموزی خود می‌کنم.				
	۲۶.	رشته تحصیلی من با علایق و استعدادهای من مطابقت دارد.				
	۲۷.	من تفریحات و سرگرمی‌های مورد علاقه خود را دنبال می‌کنم.				
عاطفی	۲۸.	من می‌توانم شرایط زندگی خود را کنترل کرده و با تغییر شرایط کنار بیایم.				
	۲۹.	من در مقابل مشکلات ناامید نشده و به دنبال قوی‌تر کردن خودم هستم.				
	۳۰.	من می‌توانم در هنگام مشکلات، آرامشم را حفظ کرده و خود را دلگرمی دهم.				
	۳۱.	من انسان شادی هستم و می‌توانم به‌راحتی بخندم.				
	۳۲.	دیگران مرا از نظر روحی و روانی باثبات می‌دانند.				
	۳۳.	من خود را مسئول احساساتم و نحوه ابراز آنها می‌دانم.				
اجتماعی	۳۴.	من چند دوست صمیمی دارم.	یک نفر	دو نفر	سه نفر	چهار نفر و بیشتر
	۳۵.	من توانایی حل مشکلات (خانواده و. دوستان) را دارم.				
	۳۶.	من با مسئولین مدرسه و معلمان روابط خوبی دارم.				
	۳۷.	من به حریم خصوصی خود و دیگران احترام می‌گذارم.				
	۳۸.	من احساسات دیگران را درک کرده و می‌توانم رفتار مناسبی داشته باشم.				
	۳۹.	من احساس تعلق به یک دسته یا گروه خاصی دارم.				
معنوی	۴۰.	زندگی من هدفمند است. (برای زندگی‌ام اهدافی دارم).				
	۴۱.	من به‌طورکلی در زندگی خود احساس آرامش دارم.				
	۴۲.	فعالیت‌هایی جهت رشد فکری مانند مطالعه، تفکر و عبادت را انجام می‌دهم.				

	۴۳.	من به توانایی‌های خود اعتماد دارم و بعد از شکست ناامید نمی‌شوم.				