دوفصلنامه محاسبات و سامانه های توزیع شده
سال پنجم، شماره۲، شماره پیاپی ۱۰، صفحه ۷۵ تا ۸۷، سال ۱۴۰۱

توسعه یک روش بهبود یافته جهت پیش بینی عملکرد آموزشی و تحصیلی دانشجویان، مبتنی بر تکنیک های داده کاوی و یادگیری ماشین

آرش خسروی[۱]، مرتضی رجب زاده*[۲]، محمد نوری خضرآبادی[۳]
[۱]استادیار، دانشکده مهندسی، مرکز آموزش عالی محلات، محلات، ایران
[۲]استادیار، دانشکده مهندسی، مرکز آموزش عالی محلات، محلات، ایران
[۳]کارشناس ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، مؤسسه آموزش عالی پویش، قم، ایران

چکیده

دانشگاه‌ها و مؤسسات آموزشی، حجم عظیمی از داده‌ها، از قبیل اطلاعات فردی و آموزشی دانشجویان را جمع‌آوری و ذخیره می‌کنند. رشد بسیار زیاد داده‌های الکترونیکی در دانشگاه‌ها، به این واقعیت اشاره دارد که با استفاده از روش‌های تحلیل داده می توان به نتایج مطلوب در حوزه های آموزشی و پژوهشی دست یافت. یکی از چالش های اصلی محیط آموزشی میزان موفقیت دانشجویان است. این مسئله وجود دارد که مهمترین ویژگی های دانشجویان برای پیش‌بینی پیشرفت تحصیلی آنها چیست و کدام الگوریتم برای انجام این پیش‌بینی مناسب‌تر است و در صورت رسیدن به نتایج مناسب در تحلیل پیشرفت تحصیلی، مدیران چگونه می‌توانند برنامه‌ریزی بهتری براساس آن انجام دهند. در این مقاله تمام ویژگی‌های امکان‌پذیر دانشجویان در یک موسسه آموزشی، جمع‌آوری و برخی از الگوریتم‌های داده‌کاوی و نیز یک روش پیشنهادی روی داده‌ها اجرا شده اند و نتایج به دست آمده، بررسی و براساس معیارهای دقت، صحت و بازیابی با یکدیگر مقایسه شده اند. درخت تصمیم با ۰/۸۶۴ کمترین دقت و روش پیشنهادی با ۰/۹۳۵ بالاترین دقت را نشان داد. همچنین مهمترین ویژگی های مؤثر در پیشرفت تحصیلی دانشجویان شناسایی شدند. با استفاده از این پیش‌بینی، مدیران نیز می‌توانند موانع پیش رو را رفع نموده و زمینه را برای پیشرفت دانشجویان فراهم نمایند.

کلمات کلیدی: ویژگی، داده کاوی، داده کاوی آموزشی، طبقه‌بندی، آمار

Development of an Improved Method for Predicting Educational and Academic Performance of Students, Based on Data Mining and Machine Learning

Arash Khosravi[1], Morteza Rajabzadeh*[2], Mohammad Nouri Khezrabadi[3]
[1] Assistant Professor, Faculty of Engineering, Mahallat Institute of Higher Education, Mahallat, Iran.
[2] Assistant Professor, Faculty of Engineering, Mahallat Institute of Higher Education, Mahallat, Iran.
[3] Master of Science, Faculty of Computer Engineering and Information Technology, Pooyesh Institute of Higher Education, Qom, Iran.

Abstract

Universities and educational institutions collect and store a huge amount of data, such as personal and educational information of students. The huge growth of electronic data in universities points to the fact that by using data analysis methods, it is possible to achieve desirable results in the fields of education and research. One of the main challenges of the educational environment is the success rate of students. There is the issue of what are the most important characteristics of students to predict their academic progress and which algorithm is more suitable for making this prediction, and if appropriate results are obtained in the analysis of academic progress, how can managers plan better based on it. In this article, all the possible characteristics of students in an educational institution, collection and some data mining algorithms as well as a proposed method have been implemented on the data and the results have been obtained, checked and compared with each other based on the criteria of Accuracy, Recall and Precision. The decision tree showed the lowest accuracy with 0.864 and the proposed method showed the highest accuracy with 0.935. Also, the most important features that are effective in the academic progress of students were identified. By using this prediction, managers can also remove the obstacles and provide the ground for the progress of students.

تاریخچه مقاله:

تاریخ ارسال: ۱۴۰۱/۱۰/۱۸
تاریخ اصلاحات: ۱۴۰۱/۱۲/۱۵
تاریخ پذیرش: ۱۴۰۱/۱۲/۲۵
تاریخ انتشار: ۱۴۰۱/۱۲/۲۹

Keywords:

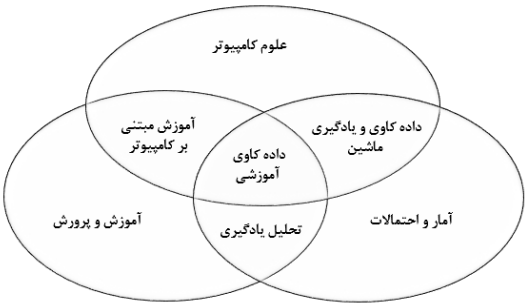
Characteristic
Data Mining
Educational Data Mining
Classification
Statistics

* ایمیل نویسنده مسئول:
rajabzadeh.m@gmail.com

۱ – مقدمه

محیط دانشگاه ها و مؤسسات آموزشی به طور گسترده شامل سه نوع بازیگر، یعنی استاد، دانشجو و محیط آموزشی می‌باشند. تعامل بین این سه بازیگر، داده‌های گسترده‌ای را تولید می‌کند، داده‌هایی که برگرفته از اطلاعات فردی و آموزشی است. رشد بسیار زیاد داده‌های الکترونیکی در دانشگاه‌ها و مؤسسات آموزشی، به این واقعیت اشاره دارد که تقطیر انبوه داده‌ها، به مجموعه پیشرفته تری از الگوریتم‌ها نیاز دارد. این موضوع منجر به ظهور زمینه داده‌کاوی آموزشی[1] شده است.

داده‌کاوی آموزشی، فرآیندی است که در آن داده‌های خام حاصل از سیستم‌های آموزشی را به اطلاعات مفیدی تبدیل می‌کند که به صورت بالقوه می‌تواند تأثیر بیشتری در تحقیقات و تمرین‌های آموزشی بگذارد. به طور سنتی، محققان، روش‌های داده‌کاوی، مانند طبقه‌بندی[2]، خوشه‌بندی[3]، استخراج قانون وابستگی، ارتباط و استخراج متن را در منابع آموزشی استفاده نموده اند. داده‌کاوی آموزشی، یک زمینه تحقیقاتی میان رشته‌ای است و حوزه ای است میان آموزش و پرورش، علوم کامپیوتر، آمار، داده‌کاوی و یادگیری ماشین، تحلیل یادگیری و آموزش مبتنی بر کامپیوتر. ارتباط میان این رشته ها در شکل ۱ نمایش داده شده است [1].



(شکل –۱): مفهوم بین رشته ای داده کاوی آموزشی

انسان‌ها قادرند حجم کمی از اطلاعات را پردازش و بررسی نمایند، اما در صورتی که حجم داده‌ها زیاد باشد، انسان هر اندازه هم توانایی داشته باشد، نمی‌تواند این حجم از داده را بررسی و نتایج آن را استخراج کند. در این حالت است که داده‌کاوی می‌تواند الگوهای مناسبی را از حجم عظیم داده‌ها استخراج کند. داده‌کاوی یکی از موضوعات فعال و جوان در علوم کامپیوتر است و در زمینه‌های بسیاری اعمال شده است. یکی از این زمینه‌ها، داده‌کاوی آموزشی است.

در زمینه داده‌کاوی آموزشی، اقدامات فراوانی انجام شده است، اما یکی از مهمترین مسائلی که در این زمینه وجود دارد، این است که دانشگاه‌ها و موسسات آموزشی، برای خود اهداف و چشم‌اندازهایی دارند، و در راستای همین چشم‌انداز و اهداف، فعالیت‌های آموزشی و پژوهشی خود را ارائه می‌کنند. حال این مسئله وجود دارد که مؤسسه در راستای رسیدن به اهداف خود چه موانعی را در تحصیل دانشجویان و رسیدن به اهداف خود پیش رو دارد؟

در سال ۲۰۰۷، یک تحقیق توسط رومرُ و ونتورا [2] روی مقالات سال‌های ۱۹۹۵ تا ۲۰۰۵ در حوزه داده‌کاوی آموزشی انجام شد. حاصل این تحقیق نشان می دهد که کاربرد داده‌کاوی آموزشی در واقع برای تبدیل مؤسسات آموزشی سنتی به سیستم های مدیریت یادگیری مبتنی بر وب و سیستم‌های آموزشی هوشمند می باشد.

تکنیک‌های داده‌کاوی[4] به صورت گسترده‌ای در حوزه آموزش در حال افزایش هستند، در بسیاری از بخش‌های آموزش عالی، در حال یافتن تأثیر بالقوه این تکنیک‌ها بر فرآیند یادگیری و حرکت به سمت دوره‌های جدید دانشگاهی می‌باشد. داده‌کاوی آموزشی و تجزیه و تحلیل یادگیری[5] دو منطق خاص هستند که برای نشان دادن استفاده و کاربرد داده‌کاوی استفاده می شوند. گزارش‌ها و کار مداوم با داده‌های دیجیتال به منظور بهبود روند آموزشی و استفاده از داده‌کاوی آموزشی، این پتانسیل را دارند که الگوهای موجود آموزش و یادگیری را شکل دهند و یا راه حل‌های جدیدی را برای مشکلات ارائه دهند [3].

به صورت کلی مهمترین فواید تحقیق در حوزه داده‌کاوی آموزشی را می توان به صورت زیر جمع‌بندی نمود:

- تبدیل مؤسسات آموزشی سنتی، به سیستم های مدیریت یادگیری مبتنی بر وب و سیستم های آموزشی هوشمند.
- استفاده از تکنیک های داده کاوی آموزشی، برای مطالعه دوره‌های آنلاین.
- استفاده از روش‌های پیش‌بینی داده‌کاوی آموزشی، برای توسعه مدل‌های دانشجویی.
- صرف هزینه و منابع دانشگاه و مؤسسه، روی دانشجویان مستعد بر اساس پیش‌بینی‌ها.

مهمترین چالش محیط آموزشی، میزان پیشرفت دانشجویان و عوامل موثر در موفقیت آنها می باشد. هدف اصلی فراهم آوردن محیط آموزشی مطلوب برای رشد بیشتر دانشجویان است. لذا در این تحقیق

[1] Educational data mining
[2] Classification
[3] Clustering
[4] Data mining
[5] Learning analytics

به دنبال تحلیل داده های آموزشی جهت رسیدن به میزان پیشرفت بهینه دانشجویان هستیم.

با توجه به اهمیت داده کاوی آموزشی در این تحقیق به دنبال یافتن جواب سوالات زیر هستیم.

۱) ویژگی‌های مهم در مؤسسات آموزش عالی برای پیش‌بینی تحصیلات دانشجویان کدام هستند؟

۲) کدام الگوریتم یادگیری ماشین، دقت بیشتر و خطای کمتری در پیش‌بینی پیشرفت تحصیلی دانشجویان دارد؟

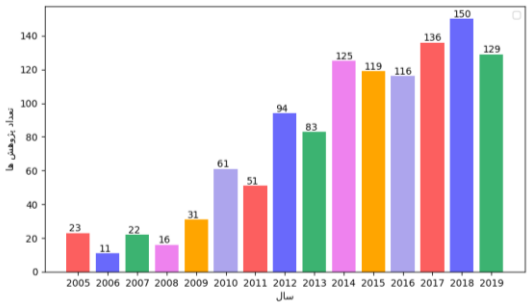
۳) چگونه می‌توان با استفاده از الگوریتم بهینه، جهت پیش‌بینی پیشرفت تحصیلی دانشجویان به مدیریت برنامه‌ریزی بهتر آموزشی کمک کرد؟

داده‌های استفاده شده در این تحقیق، از مجموعه‌ای از پایگاه داده‌های یک مؤسسه آموزشی و پژوهشی می‌باشد، این مؤسسه زیر نظر وزرات علوم، تحقیقات و فناوری است و در سه نوع آموزش حضوری، نیمه حضوری و مجازی، تحصیل در آن انجام می‌شود. در آموزش مجازی، آزمون و پذیرش دانشجو به صورت مستقیم، از طریق کنکور سراسری و سازمان سنجش انجام می شود و در نوع نیمه حضوری و حضوری، از طریق آزمون مؤسسه که با حضور نماینده سازمان سنجش برگزار می‌شود، پذیرش انجام می‌شود. این موسسه در سه مقطع کارشناسی ، ارشد و دکتری دانشجو در رشته های علوم انسانی می پذیرد. همچنین کارهای تحقیقاتی زیادی در این موسسه انجام می شود.

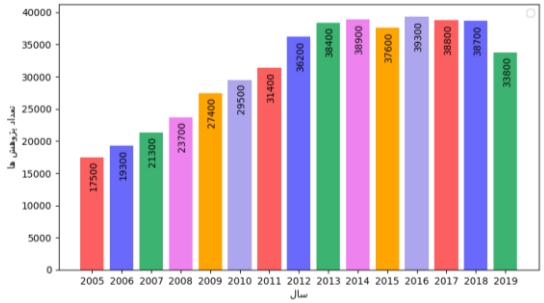
در این تحقیق، از داده‌های آموزشِ نوع حضوری استفاده شده است، از این رو، داده‌های تهیه شده، شامل دانشجویان مقطع کارشناسی است که ورودی آنها در بازه سال‌های ۱۳۸۹ تا ۱۳۹۴ می‌باشد.

۲ – مرور ادبیات

جهت مشخص نمودن موضوع داده‌کاوی آموزشی، جداولی از آمار مقالات منتشر شده در این حوزه تهیه شده است. این آمار و اطلاعات، از پایگاه جستجوی مقالات گوگل اسکولار استخراج شده است. عبارت "Educational Data Mining" یک بار در عناوین مقالات و یک بار هم در کل متن مقالات، مورد جستجو قرار گرفته است، و نتایج به دست آمده در نمودار های شکل ۲ و ۳ نمایش داده شده است.



(شکل -۲): روند پژوهش‌های خارجی در حوزه داده‌کاوی آموزشی بر اساس عنوان پژوهش



(شکل -۳): روند پژوهش‌های خارجی در حوزه داده‌کاوی آموزشی، بر اساس متن پژوهش

هسام السماری و همکاران [3] در مقاله‌ای که در سال ۲۰۱۹ منتشر شد و در راستای روشن ساختن مشکلات یادگیری انجام شده بود، بیان داشتند برخی از تکنیک‌های داده‌کاوی وجود دارند که از ابعاد متنوعی گرفته شده‌اند. آنها این ابعاد را به چهار بعد اصلی تقسیم کرده اند. این ابعاد عبارتند از:

- تجزیه و تحلیل یادگیری پشتیبانی شده توسط رایانه[۶] (CSLA)
- تجزیه و تحلیل پیش‌بینی شده با رایانه[۷] (CSPA)
- تجزیه و تحلیل رفتاری پشتیبانی شده توسط رایانه[۸] (CSBA)
- تجزیه و تحلیل تجسم پشتیبانی شده توسط رایانه[۹] (CSVA)

اینجادات و همکارانش [4] در سال ۲۰۲۰ تحقیقی را ارائه نمودند. هدف آنها از این تحقیق، پیش بینی نمرات نهایی دانشجویان است. این کار، برای شناسایی دانشجویانی است که ممکن است در مراحل بعدی، به کمک نیاز داشته باشند. داده‌های این تحقیق، از سیستم آموزش الکترونیک استخراج شده است. مجموعه داده‌های آنها، شامل سوابق دانشجویانی است که سال اول دوره کارشناسی را در دانشگاه جنوا[۱۰] به پایان رسانده‌اند. ویژگی هدف، که همان معدل کل است به دو دسته "خوب" که نمرات بین ۶۰ تا ۱۰۰ است و دسته "بد" که نمرات

[6] Computer-Supported Learning Analytics
[7] Computer-Supported Predictive Analytics
[8] Computer-Supported Behavioral Analytics
[9] Computer-Supported Visualization Analytics
[10] University of Genoa

کوچکتر مساوی ۵۹ است، طبقه‌بندی شدند. برای انجام عملیات داده‌کاوی در تحقیق آنها، از الگوریتم‌های SVM-RBF، NB، K-NN، MLP و RF استفاده شده است. آنها سه مدل برتر از نظر شاخص جینی[11] را RF، NB و k-NN معرفی کردند.

آسیف و همکاران [5] در سال ۲۰۱۷ مقاله‌ای را منتشر کردند. آنها در این مقاله، سه سوال را مطرح کردند:

سوال اول: تا کنون چندین طبقه‌بندی در دانشگاه برای پیش‌بینی نمرات دانشجویان، تولید شده است. در همه این طبقه‌بندی‌ها، فقط ویژگی‌های آموزشی دانشجویان مانند نمرات پذیرفته شده از امتحانات نهایی دبیرستان، معیار قرار گرفته است، و هیچکدام از ویژگی‌های اجتماعی، جمعیتی و یا اقتصادی، در نظر گرفته نشده است. آیا این پیش بینی فقط از طریق این ویژگی‌های آموزشی امکان پذیر است؟

سوال دوم: آیا می توان دوره‌هایی را شناسایی کرد که به عنوان شاخص خوب یا پایین باشند؟

سوال سوم: آیا می‌توان پیشرفت‌های معمول عملکرد دانشجویان را تشخیص داد و آنها را با دوره‌های شاخص مرتبط کرد؟

آنان برای انجام تحقیق خود، سه رویکرد از پنج رویکرد معرفی شده توسط بیکر [6] در سال ۲۰۱۰ شامل خوشه‌بندی، پیش بینی، و تا حدی تقطیر داده توسط انسان، را ترکیب کردند، نتیجه تحقیق آنان و پاسخ به سوالات به این شرح است:

پاسخ سوال اول: فقط با استفاده از نمرات دبیرستان، می‌توان عملکرد یک دوره چهار ساله را برای دانشجویان، پیش‌بینی کرد.

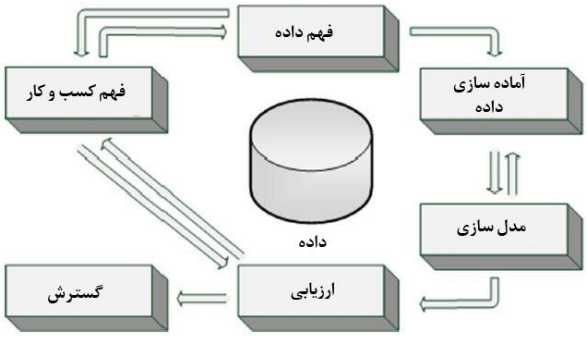
پاسخ سوال دوم: با استفاده از درخت تصمیم، چهار دوره شناسایی شدند.

پاسخ سوال سوم: دانشجویان، در دو گروه عمده قرار می‌گیرند، گروهی از دانشجویان در گروه عملکرد بالا و گروهی دیگر در گروه عملکرد پایین.

رودریگز و همکاران [7] در سال ۲۰۱۸ مقاله‌ای را با عنوان مروری بر روند ارزیابی در آموزش الکترونیک با استفاده از داده‌کاوی آموزشی منتشر کردند. آنها دو سوال را مطرح کردند: اول اینکه چشم اندازها و گرایش‌های اصلی در حوزه داده‌کاوی آموزشی برای آموزش الکترونیک چیست؟ و دوم اینکه موضوعات بالقوه تحقیقاتی که در ارزیابی یادگیری الکترونیک، مورد توجه قرار می‌گیرد، کدام هستند؟ آنها در بررسی های خود از ۷۲ مقاله پیشین در این حوزه استفاده کردند.

اسلاتر و همکاران [8] در سال ۲۰۱۷ مقاله‌ای را منتشر کردند. نویسندگان این مقاله، مبحثی را در مورد اهمیت آشنایی با چندین ابزار را مطرح می‌نمایند، سپس یک جعبه ابزار جهت تجزیه و تحلیل داده‌های حاصل از تحقیقات آنالیز یادگیری و داده‌کاوی آموزشی، ارائه می‌نمایند.

فرناندز و همکاران [9] در سال ۲۰۱۹ مقاله‌ای را با عنوان "تحلیل و پیش بینی عملکرد تحصیلی از دانش‌آموزان مدارس دولتی در پایتخت برزیل" منتشر کردند. آنها، دو مجموعه ویژگی‌ها را فراهم کردند: مجموعه اول، ویژگی‌های مدرسه دانش‌آموزان و مجموعه دوم، ویژگی‌های خود دانش‌آموزان. این جمع‌آوری، از دانش‌آموزان و مدارس پایتخت انجام شد. ویژگی‌ها عبارتند از: منطقه مدرسه، مدرسه، شیفت، محیط استفاده از کلاس، دانشجو، جنسیت، سن، شهر، محله، درجه، غیبت و معدل کل. آنها برای انجام عملیات داده کاوی از روش CRISP-DM[12] استفاده نمودند. نحوه عملکرد روش مذکور در شکل ۴ نمایش داده شده است.



(شکل -۴): نحوه عملکرد روش CRISP-DM [9]

آنها با انجام عملیات داده‌کاوی با استفاده از الگوریتم Gradient Boost Machine روی ویژگی‌های فوق، مهمترین ویژگی‌ها را در سال‌های ۲۰۱۵ و ۲۰۱۶ به ترتیب اهمیت به صورت زیر معرفی نمودند:

- محله دانش‌آموز
- مدرسه
- سن
- شهر
- منطقه مدرسه
- جنسیت

آنها در نتیجه تحقیق خود، بیان داشتند که ویژگی‌های مجموعه اول، مانند محله و مدرسه، اهمیت بیشتری برای موفقیت دانش‌آموزان

[11] Gini index

[12] Cross-Industry Standard Process for Data Mining

دارد. در واقع، میزان دسترسی دانش‌آموزان به مدرسه مناسب که مرتبط با محله سکونت دانش آموز است، تأثیر بالایی بر موفقیت تحصیلی دانش‌آموزان دارد.

فلدمن و همکاران [10] در سال ۲۰۲۱ مقاله‌ای با عنوان پشت صحنه داده‌کاوی آموزشی ارائه نمودند. آنها هدف خود را، تعریف مراحل پیش پردازش داده‌های آموزشی آنلاین و آگاهی محققان و سیاست گذاران آموزشی، در مرحله پیش پردازش بیان نمودند. در جایی دیگر، هدف خود را، به حداقل رساندن داده‌های غیرمرتبط و کاهش اشتباه در تجزیه و تحلیل داده‌ها توسط محققین اعلام نمودند. آنها ماهیت فنی بررسی‌های خود را در چهار مرحله متوالی تنظیم کردند: "جمع آوری داده‌ها"، "تفسیر داده‌ها"، "ایجاد پایگاه داده" و "سازماندهی داده‌ها".

در مرحله اول، آنها دریافتند که چالش اصلی، در مرحله جمع‌آوری داده‌ها آغاز می‌شود، مقاله توصیفی، با داده‌های خام شروع می‌شود، و لذا در مواردی که محققان داده‌های پردازش شده را دریافت می‌کنند، برای اینکه بتوانند داده‌ها را ارزیابی کنند، نیازمند اطلاع از مراحل قبل از پردازش هستند. در مرحله دوم، تاکید بر لزوم بررسی دقیق ویژگی‌های داده‌ها می‌باشد. محققین می‌توانند لیست ویژگی‌ها را مطابق با اهداف تحقیقات خاص و داده‌های موجود، گسترش دهند. در مرحله سوم، ایجاد پایگاه داده، تاکید بر پیروی از مقررات عمومی حفاظت از داده‌های اتحادیه اروپا است. آخرین مرحله، سازماندهی داده‌ها است. در این مرحله، داده‌ها از منابع مختلف، فیلتر و یکپارچه می‌شوند.

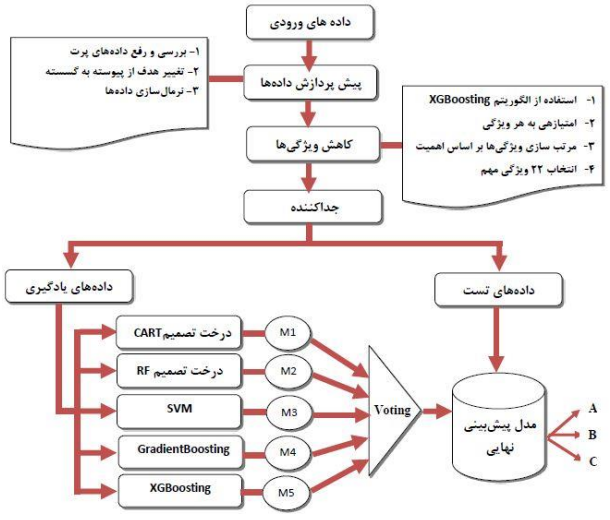
آنها برای رسیدگی به چالش‌ها، سه توصیه اصلی را مطرح کردند: "همکاری"، "اتوماسیون" (فرآیند خودکار) و "تفسیر". در توصیه "همکاری" بیان داشتند: مؤسسات آموزشی، اغلب کنترل محدود بر نوع و قالب داده‌های خود دارند. برای رفع این مشکل، باید نرم افزارها و سیاست مدیران، بروز رسانی شود. ورود موفقیت آمیز به داده‌کاوی آموزشی، مستلزم همکاری افراد، از بخش‌های مختلف دانشگاه و مؤسسات آموزشی است. آنها توصیه "فرآیندهای خودکار" را این چنین شرح دادند: خودکارسازی جنبه‌های مختلف فنی پیش پردازش داده‌ها، می تواند حجم مطالعات مبتنی بر داده‌کاوی را افزایش دهد. این اتوماسیون، از طریق ایجاد گزارش‌های کاربر پسند و قابل اعتماد، می‌تواند به تدوین سیاست‌هایی برای طراحی آموزش بهتر کمک کند. در مورد "تفسیر" این چنین شرح دادند: "تفسیر"، مشکل اساسی در درک داده‌های حاضر است. چندین متغیر، می‌توانند گمراه کننده باشند، به عنوان مثال، برآورد زمان مصرف شده در سیستم و یا تعداد فعالیت‌های آنلاین کاربران.

مارتینز و همکاران [11] در سال ۲۰۲۰ مقاله‌ای را ارائه نمودند. در این مقاله، آنها ابتدا روشی را که در تحقیقات گذشته برای شناسایی مدارس با اثر بخشی زیاد یا کم ارائه شده بود، مورد تایید قرار دادند. سپس یک نظر ابتکاری در زمینه تجزیه و تحلیل تاثیر مدرسه با استفاده از داده‌کاوی آموزشی، ارائه نمودند.

موداسیر و همکاران [12] در سال ۲۰۲۰ مقاله ای را جهت پیش بینی هوشمند برای داده کاوی آموزشی، بر اساس روش Ensemble و فیلتر رویکردها منتشر کردند. در تحقیق آنها، برای انجام داده‌کاوی، از الگوریتم های J48، Naïve Bayes، Random tree و KNN استفاده کرده اند. در این روش، از میان الگوریتم‌ها، چند الگوریتم انتخاب می‌شوند و در نهایت رأی‌گیری روی آنها انجام می شود.

۳ - روش پیشنهادی

در این تحقیق، از روش طبقه‌بندی ترکیبی[13] استفاده شده است. در شکل ۵ چهارچوب و روش کلی تحقیق نمایش داده شده است.



(شکل -۵): چهارچوب تحقیق

۴ - جمع آوری ویژگی ها

داده‌های این تحقیق، از چند پایگاه داده آموزش و نیز سامانه نشریات استخراج شده است. برخی از داده‌های مورد استفاده، به صورت مستقیم استخراج شده‌اند و برخی دیگر نیز با انجام محاسبات به دست

[13] Ensemble classifier

آمده‌اند. داده‌هایی که به روش محاسبه و کدنویسی از داده‌های موجود، استخراج شده‌اند، در نهایت، از آنها، ۸۹ ویژگی شناسایی و استخراج شدند. داده ها برای استفاده سایر محققان بصورت آنلاین به اشتراک گذاشته شده است[۱۴].

۵ - معرفی برچسب[۱۵]

برچسب در نظر گرفته شده در این تحقیق، معدل کل[۱۶] می‌باشد. هدف از بررسی این برچسب، پیش‌بینی معدل دانشجویان و نیز مشخص شدن عواملی (ویژگی هایی) هستند که روی معدل دانشجویان، بیشترین تأثیر را دارند، یا به عبارت دیگر، شناسایی مهمترین عوامل (ویژگی های) موثر بر پیشرفت تحصیلی دانشجویان می‌باشد. در مجموعهٔ داده ها[۱۷] مربوط به این برچسب، نکات زیر در نظر گرفته شده است:

- این مجموعهٔ داده ها، شامل همه دانشجویان مقطع کارشناسی است که ورودی آنها در بازه ۱۳۹۸ تا ۱۳۹۴می‌باشد.
- فقط دانشجویان با وضعیت فارغ التحصیل، برای این مجموعهٔ داده ها انتخاب شده‌اند، تا همه در شرایط یکسانی باشند.

۶ - پیش پردازش[۱۸] داده ها

در این تحقیق، برای انجام پیش پردازش داده‌ها، اقدامات زیر انجام گرفته است:

- بررسی و رفع داده‌های پرت.
- تغییر ویژگی هدف، از نوع متغیر پیوسته به متغیر گسسته.
- کاهش ویژگی‌ها، با استفاده از الگوریتم مناسب.
- نرمال‌سازی داده‌ها، با استفاده از زبان برنامه نویسی پایتون.

۶-۱- تغییر برچسب از حالت پیوسته به حالت گسسته

تغییر داده‌ها از پیوسته به گسسته، باعث می شود تا الگوریتم با گستره کمتری از داده‌ها روبرو شود، بنابراین پیچیدگی عملیات یادگیری ماشین، کاهش خواهد یافت. در این راستا، برچسب "معدل کل" به پنج بازه A B C D E تبدیل شده است. این بازه‌ها در جدول ۱ نمایش داده شده است.

(جدول -۱): تغییر ویژگی هدف از حالت پیوسته به گسسته

۱۸-۲۰	۱۶-۱۷.۹۹	۱۴-۱۵.۹۹	۱۲-۱۳.۹۹	کمتر از ۱۲
A	B	C	D	E

۶-۲- کاهش ویژگی‌ها و انتخاب مهمترین ویژگی‌ها

به دلیل تعداد زیاد ویژگی‌ها، نیاز بود تا ویژگی‌های مهم انتخاب شوند تا عملیات داده‌کاوی با سرعت و دقت بهتری انجام شود، چرا که بعضی از ویژگی‌ها، ممکن است تأثیر چندانی در نتایج نداشته باشند.

یکی از روش‌های کاهش ابعاد و انتخاب مهمترین ویژگی‌ها، استفاده از الگوریتم XGBoost است. با استفاده از این الگوریتم، می‌توان در راستای کاهش نویز[۱۹] ویژگی‌ها و نیز کاهش ابعاد، از طریق افزایش boosting و نیز میانگین gain استفاده نمود. برای نشان دادن اثر بخشی روش (الگوریتم) انتخاب ویژگی XGBoost از کرنل PCA[۲۰] استفاده شده است [13].

با استفاده از الگوریتم XGBoost ارزش هر یک از ویژگی‌ها محاسبه شد. الگوریتم برای هر ویژگی، سه مقدار weight، gain و cover را محاسبه و سپس بر اساس میانگین آنها مرتب سازی کرده است که نتایج آن در نمودار شکل ۶ نمایش داده شده است. این عملیات با استفاده از زبان برنامه نویسی پایتون انجام شد که برنامه‌نویسی آن را در پیوست ۱ می توان مشاهده نمود.

برای انتخاب مهمترین ویژگی‌ها، از مقدار ۰.۰۴۰۸ بررسی دقت شروع شده است، به این صورت که همه ویژگی‌هایی که مقدار میانگین[۲۱] آنها از ۰.۰۴۰۸ بالاتر می‌باشد، به الگوریتم، داده شد و خروجی چهار معیار Accuracy، F1-score، Recall و Precision محاسبه و ثبت گردید. این روند ادامه یافت تا جایی که مقدار دقت افزایش و سپس کاهش یافت. بنابراین، این نقطه، قلهٔ انتخاب و ویژگی‌هایی که مقدار میانگین آنها از قله بیشتر هستند، انتخاب شدند.

با توجه به مقادیر جدول ۲، ویژگی‌های انتخاب شده، شامل همه ویژگی‌هایی هستند که مقدار آنها بیشتر از ۰.۱۱۷۳ می‌باشد، لیست این ویژگی‌ها در جدول شماره ۳ نمایش داده شده است.

۶-۳- الگوریتم‌های داده‌کاوی مورد استفاده

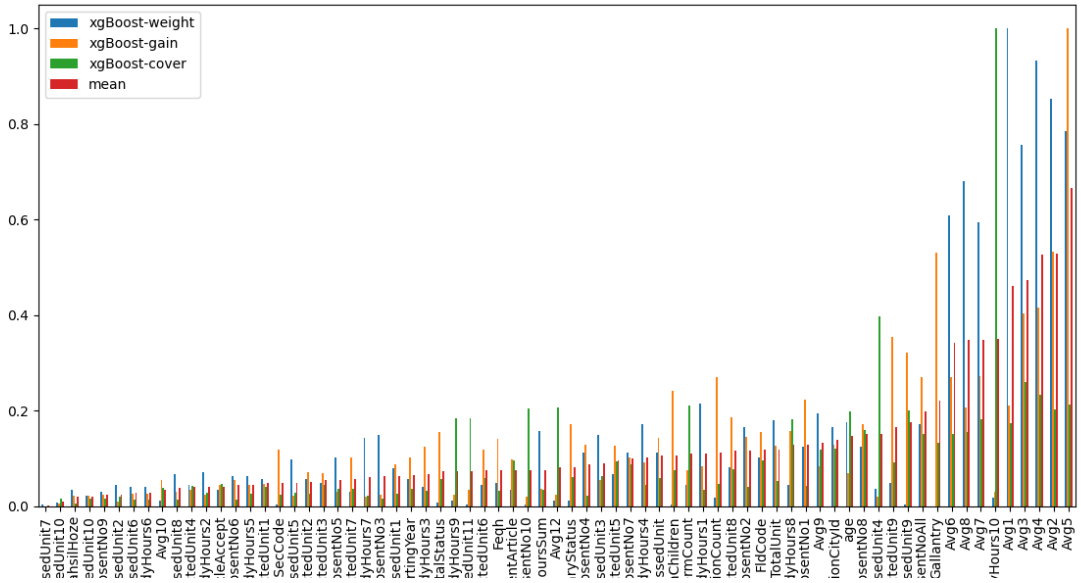
از آنجایی که حداقل نمره قبولی در مقطع کارشناسی موسسه نمره ۱۲ است و معدل ترم کمتر از ۱۴ نیز مشروط می‌باشد و بر اساس قوانین مشروطی، اخراج خواهند شد و از آنجایی که لیست داده های ما، شامل دانشجویان فارغ التحصیل است، بنابراین معدل

[14] https://github.com/khosravi280/Academic-Performance-
[15] Lable
[16] Total average
[17] Data set
[18] Preprocessing
[19] Noise
[20] Principal Component Analysis
[21] Mean

کمتر از ۱۲ در لیست داده‌ها وجود ندارد، پس برچسب ما شامل موارد A B C می باشد.

تعداد الگوریتم‌های استفاده شده در روش Ensemble باید فرد باشد، و بیشتر از تعداد برچسب ها باشد. بنابراین در این حالت که تعداد برچسب‌ها سه مورد هستند، پس تعداد الگوریتم‌های مورد استفاده نیز باید پنج مورد باشد. در این مقاله، از الگوریتم‌های داده‌کاوی زیر استفاده شده است:

- درخت تصمیم CART[۲۲]
- درخت تصمیم RF[۲۳]
- الگوریتم SVM[۲۴]
- الگوریتم Boosting Gradient
- الگوریتم Boosting XG[۲۵]



(شکل -۶): مقادیر اهمیت برای ویژگی‌ها بر اساس الگوریتم XG Boosting

(جدول -۲): مقادیر معیار بدست آمده در بازه های مختلف داده‌های با اهمیت

بیشتر از	0.0408	0.0542	0.0666	0.0747	0.082	0.1007	0.1104	0.1173	0.1196	0.1296
تعداد	61	52	45	39	35	30	26	22	20	19
Accuracy	0.8915	0.8983	0.9050	0.9016	0.9186	0.9118	0.9220	0.9254	0.9254	0.9118
F1-score	0.8836	0.8879	0.9067	0.8893	0.9146	0.9027	0.9186	0.9149	0.9241	0.9067
Recall	0.8915	0.8983	0.9050	0.9016	0.9186	0.9118	0.9220	0.9254	0.9254	0.9118
Precision	0.8930	0.9039	0.9093	0.8980	0.9162	0.9044	0.9188	0.9334	0.9276	0.9030

۷- ارزیابی مدل ها

۷-۱- درخت تصمیم CART

تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۴ و مقادیر ماتریس درهم ریختگی را در شکل ۷ می‌توان مشاهده نمود.

۷-۲- درخت تصمیم RF

تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۵ و مقادیر ماتریس درهم ریختگی را در شکل ۸ می توان مشاهده نمود.

۷-۳- الگوریتم SVM

معیارهای استخراج شده، تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۶ و مقادیر ماتریس درهم ریختگی را در شکل ۸ می‌توانید مشاهده نمایید.

[22] Classification And Regression Tree
[23] Random Forest
[24] Support Vector Machines
[25] Extreme Gradient Boosting

(جدول -۳): ویژگی های نهایی جهت انجام عملیات داده کاوی

ردیف	نام ویژگی	شرح ویژگی	میانگین
۱	Avg5	معدل ترم ۵	0.6659
۲	Avg2	معدل ترم ۲	0.5297
۳	Avg4	معدل ترم ۴	0.5272
۴	Avg3	معدل ترم ۳	0.4731
۵	Avg1	معدل ترم ۱	0.461
۶	StudyHours10	میزان ساعت مطالعه ترم ۱۰	0.3495
۷	Avg7	معدل ترم ۷	0.3492
۸	Avg8	معدل ترم ۸	0.3475
۹	Avg6	معدل ترم ۶	0.3431
۱۰	Gallantry	ایثارگری (۰.عادی، ۱.آزاده، ۲.فرزند آزاده، ۳.جانباز، ۴.فرزند جانباز، ۵.فرزند شهید)	0.2215
۱۱	No Absent All	تعداد کل غیبت‌های کلاسی	0.1979
۱۲	PassedUnit9	تعداد واحد پاس شده ترم ۹	0.1754
۱۳	SelectedUnit9	تعداد واحد انتخابی ترم ۹	0.1652
۱۴	PassedUnit4	تعداد واحد پاس شده ترم ۴	0.1519
۱۵	AbsentNo8	تعداد غیبت‌های کلاسی ترم ۸	0.1516
۱۶	age	سن	0.148
۱۷	Location Id City	شهر محل سکونت	0.1382
۱۸	Avg 9	معدل ترم ۹	0.1324
۱۹	AbsentNo1	تعداد غیبت‌های کلاسی ترم ۱	0.1296
۲۰	Study Hours 8	میزان ساعت مطالعه ترم ۸	0.1284
۲۱	Total Les Unit	تعداد واحد انتخاب شده کل	0.1196
۲۲	Code Fld	کد رشته تحصیلی	0.118

(جدول -۴): مقادیر معیار بدست آمده برای الگوریتم درخت تصمیم CART

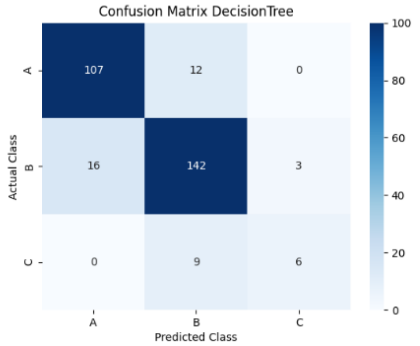
Accuracy	۰.۸۶۴۴۰۶۷۷۹۶۶۱۰۱۷
F1-score	۰.۸۶۰۵۲۴۹۵۵۰۸۰۶۸۱۸
Recall	۰.۸۶۴۴۰۶۷۷۹۶۶۱۰۱۷
Precision	۰.۸۶۰۲۶۴۳۸۶۱۵۲۲۳۲
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۵۵
تعداد تشخیص های نادرست از ۲۹۵ رکورد تست	۴۰

(جدول -۵): مقادیر معیارهای بدست آمده برای درخت تصمیم RF

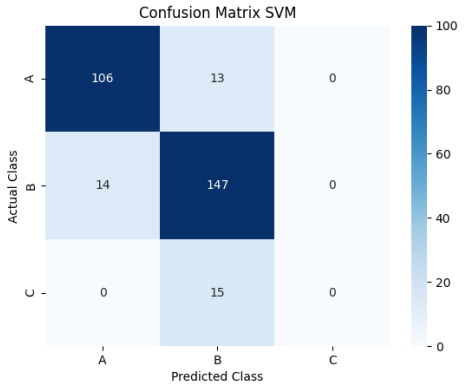
Accuracy	۰.۹۱۱۸۶۴۴۰۶۷۷۹۶۶۱
F1-score	۰.۸۹۳۱۵۲۷۵۰۸۶۲۱۵۴۴
Recall	۰.۹۱۱۸۶۴۴۰۶۷۷۹۶۶۱
Precision	۰.۹۱۹۹۳۹۴۷۲۶۳۹۶۲۷۶
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۶۹
تعداد تشخیص های نادرست از ۲۹۵ رکورد تست	۲۶

(جدول -۶): مقادیر معیارهای بدست آمده برای الگوریتم SVM

Accuracy	۰.۸۵۷۶۲۷۱۱۸۶۴۴۰۶۷۸
F1-score	۰.۸۳۵۳۶۰۹۶۷۳۰۷۲۸۳
Recall	۰.۸۵۷۶۲۷۱۱۸۶۴۴۰۶۷۸
Precision	۰.۸۶۵۶۱۵۸۱۹۲۰۹۰۳۹۶
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۵۳
تعداد تشخیص های نادرست از ۲۹۵ رکورد تست	۴۲



(شکل -۷): ماتریس درهم ریختگی بدست آمده برای درخت تصمیم CART



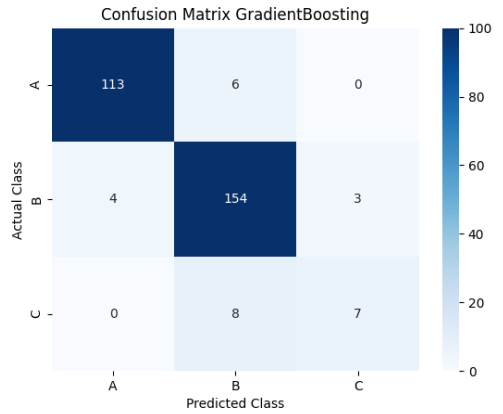
(شکل -۸): ماتریس درهم ریختگی بدست آمده برای الگوریتم SVM

۴-۷- الگوریتم Gradient Boosting

معیارهای استخراج شده، تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۷ و مقادیر ماتریس درهم ریختگی را در شکل ۹ می‌توانید مشاهده نمایید.

(جدول -۷): مقادیر معیارهای بدست آمده برای الگوریتم Gradient Boosting

Accuracy	۰.۹۲۸۸۱۳۵۵۹۳۲۲۰۳۳۹
F1-score	۰.۹۲۵۶۹۸۴۱۱۴۳۷۱۶۳۴
Recall	۰.۹۲۸۸۱۳۵۵۹۳۲۲۰۳۳۹
Precision	۰.۹۲۵۴۷۴۴۳۱۴۰۶۶۳۴۹
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۷۴
تعداد تشخصی های نادرست از ۲۹۵ رکورد تست	۲۱



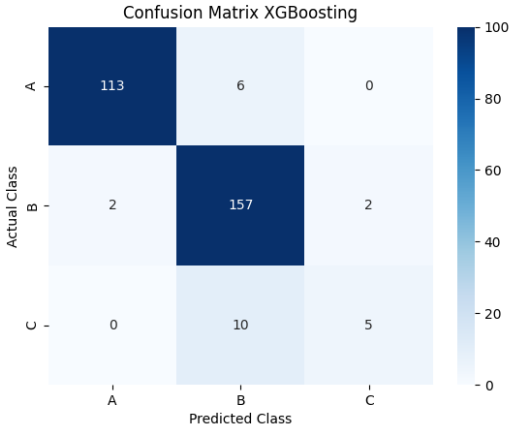
(شکل -۹): ماتریس درهم ریختگی بدست آمده برای الگوریتم Gradient Boosting

۷-۵- الگوریتم XG Boosting

معیارهای استخراج شده، تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۸ و مقادیر ماتریس درهم ریختگی را در شکل ۱۰ می‌توانید مشاهده نمایید.

(جدول -۸): مقادیر معیارهای بدست آمده برای الگوریتم XG Boosting

Accuracy	۰.۹۳۲۲۰۳۳۸۹۸۳۰۵۰۸۴
F1-score	۰.۹۲۵۷۹۳۵۱۵۸۸۷۹۰۳۴
Recall	۰.۹۳۲۲۰۳۳۸۹۸۳۰۵۰۸۴
Precision	۰.۹۲۷۹۸۱۵۱۵۵۴۷۴۲۳۱
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۷۵
تعداد تشخصی های نادرست از ۲۹۵ رکورد تست	۲۰



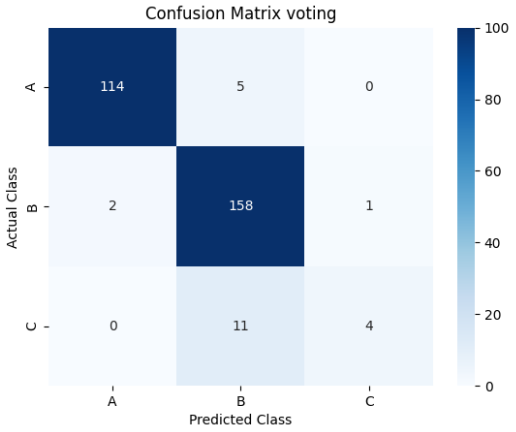
(شکل -۱۰): ماتریس درهم ریختگی بدست آمده برای روش XG Boosting

۷-۶- روش Voting

معیارهای استخراج شده، تعداد تشخیص‌های درست و نادرست برای این الگوریتم را در جدول ۹ و مقادیر ماتریس درهم ریختگی را در شکل ۱۱ می‌توانید مشاهده نمایید.

(جدول -۹): مقادیر معیارهای بدست آمده برای روش Voting

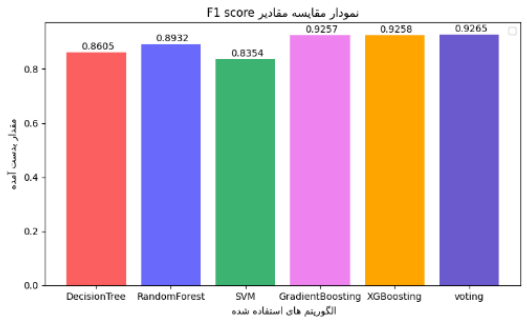
Accuracy	۰.۹۳۵۵۹۳۲۲۰۳۳۸۹۸۳۱
F1-score	۰.۹۲۶۵۲۱۹۵۲۰۸۵۹۴۶۱
Recall	۰.۹۳۵۵۹۳۲۲۰۳۳۸۹۸۳۱
Precision	۰.۹۳۲۶۹۰۴۳۴۴۴۳۷۹۵۱
تعداد تشخیص های درست از ۲۹۵ رکورد تست	۲۷۶
تعداد تشخصی های نادرست از ۲۹۵ رکورد تست	۱۹



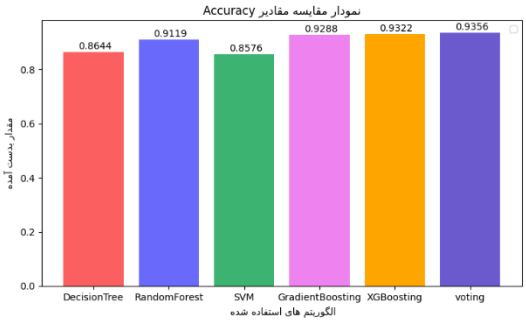
(شکل -۱۱): ماتریس درهم ریختگی بدست آمده برای روش Voting

۷-۷- مقایسه نتایج الگوریتم ها

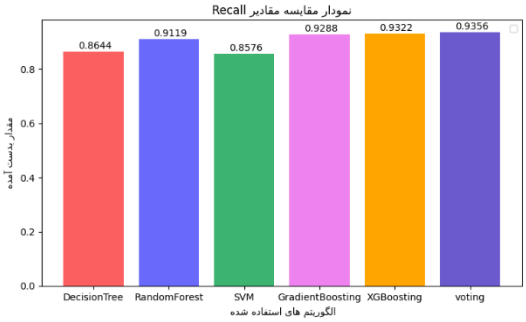
از مقایسه مقادیر نتایج بدست آمده از پنج الگوریتم مورد نظر و نیز روش رای گیری، نمودارهای شکل های ۱۲ تا ۱۵ به دست آمده است. این نمودارها به تفکیک مقادیر Accuracy،F1-score، Precision ،Recall می باشد.



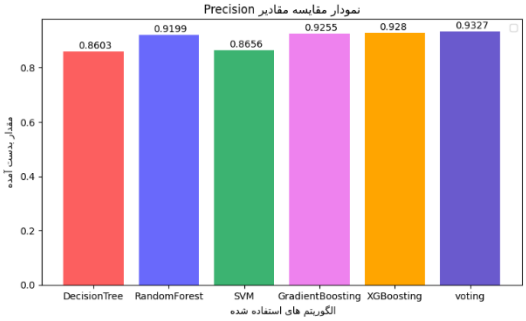
(شکل -۱۲): مقایسه مقدار معیار F1-Score به تفکیک الگوریتم‌ها



(شکل -۱۳): مقایسه مقدار دقت Accuracy به تفکیک الگوریتم‌ها



(شکل -۱۴): مقایسه مقدار معیار Recall به تفکیک الگوریتم‌ها



(شکل -۱۵): مقایسه مقدار معیار Precision به تفکیک الگوریتم‌ها

۸ – جمع بندی نتایج تحقیق

مهمترین ویژگی ها جهت پیش‌بینی معدل کل و یا به عبارتی وضعیت تحصیلی دانشجویان به ترتیب اهمیت در جدول ۱۰ نمایش داده شده است.

(جدول -۱۰): مهمترین ویژگی ها به ترتیب اهمیت

1- معدل ترم ۵	25- تعداد ترم مرخصی	49- میزان ساعت مطالعه ترم ۷
2- معدل ترم ۲	26- میزان ساعت مطالعه ترم ۱	50- تعداد واحد انتخابی ترم ۷
3- معدل ترم ۴	27- تعداد ترم تحصیلی	51- تعداد غیبت های کلاسی ترم ۵
4- معدل ترم ۳	28- تعداد فزرند	52- تعداد واحد انتخابی ترم ۳
5- معدل ترم ۱	29- تعداد واحد پاس شده کل	53- تعداد واحد انتخابی ترم ۲
6- میزان ساعت مطالعه ترم ۱۰	30- میزان ساعت مطالعه ترم ۴	54- تعداد واحد پاس شده ترم ۵
7- معدل ترم ۷	31- تعداد غیبت های کلاسی ترم ۷	55- کد مقطع تحصیلی
8- معدل ترم ۸	32- تعداد واحد انتخابی ترم ۵	56- تعداد واحد انتخابی ترم ۱
9- معدل ترم ۶	33- تعداد واحد پاس شده ترم ۳	57- میزان ساعت مطالعه ترم ۵
10- ایثارگری	34- تعداد غیبت های کلاسی ترم ۴	58- تعداد غیبت های کلاسی ترم ۶
11- تعداد کل غیبت های کلاسی	35- وضعیت نظام وظیفه	59- تعداد مقالات منتشر شده
12- تعداد واحد پاس شده ترم ۹	36- معدل ترم ۱۲	60- میزان ساعت مطالعه ترم ۲
13- تعداد واحد انتخابی ترم ۹	37- جمع ساعات مطالعه	61- تعداد واحد انتخابی ترم ۴
14- تعداد واحد پاس شده ترم ۴	38- تعداد غیبت های کلاسی ترم ۱۰	62- تعداد واحد پاس شده ترم ۸
15- تعداد غیبت های کلاسی ترم ۸	39- تعداد مقالات ارسال شده	63- معدل ترم ۱۰
16- سن	40- آیا در دروه فقه شرکت می کند	64- میزان ساعت مطالعه ترم ۶
17- شهر محل سکونت	41- تعداد واحد انتخابی ترم ۶	65- تعداد واحد پاس شده ترم ۶
18- معدل ترم ۹	42- تعداد واحد انتخابی ترم ۱۱	66- تعداد واحد پاس شده ترم ۲
19- تعداد غیبت های کلاسی ترم ۱	43- میزان ساعت مطالعه ترم ۹	67- تعداد غیبت های کلاسی ترم ۹
20- میزان ساعت مطالعه ترم ۸	44- وضعیت تاهل	68- تعداد واحد پاس شده ترم ۱۰
21- تعداد واحد انتخاب شده کل	45- میزان ساعت مطالعه ترم ۳	69- مقطع تحصیل حوزوی

22- کد رشته تحصیلی	46- سال شروع به تحصیل	70- تعداد واحد انتخابی ترم ۱۰
23- تعداد غیبت های کلاسی ترم ۲	47- تعداد واحد پاس شده ترم ۱	71- تعداد واحد پاس شده ترم ۷
24- تعداد واحد انتخابی ترم ۸	48- تعداد غیبت های کلاسی ترم ۳	

همچنین ویژگی‌هایی که هیچ تاثیری روی معدل کل نداشتند نیز شناسایی شدند که در جدول ۱۱ نمایش داده شده است.

(جدول -۱۱): ویژگی هایی که فاقد اهمیت شناخته شده اند

ترم شروع به تحصیل	تعداد واحد انتخابی ترم ۱۲	نحوه ورود
ملیت	تعداد واحد پاس شده ترم ۱۱	میزان ساعت مطالعه ترم ۱۱
وضعیت بومی	تعداد واحد پاس شده ترم ۱۲	میزان ساعت مطالعه ترم ۱۲
وضعیت جسمانی	تعداد غیبت های کلاسی ترم ۱۱	
معدل ترم ۱۱	تعدادغیبت های کلاسی ترم ۱۲	

از میان پنج الگوریتمی که استفاده شده و نیز روش رأی‌گیری، روش رأی‌گیری بر اساس همه معیارها، دقت بیشتر و خطای کمتری را نسبت به پنج الگوریتم استفاده شده دارد و از این رو یک روش بهبود یافته می‌باشد. ترتیب دقت الگوریتم‌های استفاده شده در مجموع به صورت زیر می باشد.

۱) روش Voting
۲) الگوریتم XG Boosting
۳) الگوریتم Gradient Boosting
۴) درخت تصمیم RF
۵) درخت تصمیم CART
۶) الگوریتم SVM

مدیران تصمیم گیرنده در محیط‌های دانشگاهی با شناسایی ویژگی‌های مهم و تاثیرگذار در پیشرفت تحصیلی دانشجویان، می‌توانند با تلاش در راستای رفع موانع، زمینه‌ساز پیشرفت دانشجویان و در نهایت پیشرفت دانشگاه خود شوند. مدیران همچنین با استفاده از پیش‌بینی‌ها می‌توانند دانشجویان مستعد را زودتر شناسایی نمایند و در صورت تمایل، سرمایه‌گذاری بیشتری روی این دانشجویان انجام دهند.

۹ – مراجع

[1] Dutt, A., Ismail, M. A., and Herawan, T. (2017), A Systematic Review on Educational Data Mining, IEEE Access 5(c): pp. 15991-16005.

[2] Romero, C., Ventura, S. (2007), Educational Data Mining: A Survey from 1995 to 2005, Expert Systems with Applications 33(1): pp. 135-146.

[3] Aldowah, W. M. F. H., Al-Samarraie H. (2019), Educational Data Mining and Learning Analytics for 21stcentury Higher Education: A Review and Synthesis, 2019 Jt. Int. Conf. Digit. Arts, Media Technol. with ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. (ECTI DAMT-NCON) 94(6): pp. 142-145.

[4] Injadat M., Moubayed A., Bou A., and Shami A. (2020), Knowledge-Based Systems Systematic Ensemble Model Selection Approach for Educational Data Mining, Knowledge-Based Syst. vol. 200: p. 105992.

[5] Asif, R., Merceron, A., Ali, S. A., and Haider, N. G. (2017), Analyzing Undergraduate Students' Performance Using Educational Data Mining, Comput. Educ. vol. 113: pp. 177-194.

[6] Baker, R., Yacef, K. (2010), The State of Educational Data Mining in 2009: A Review and Future Visions, Journal of Educational Data Mining, 1(11): pp. 3-17.

[7] Rodrigues, M. W., Isotani, S., and Zárate, L. E. (2018), Educational Data Mining: A Review of Evaluation Process in the E-learning, Telemat. Informatics 35(6): pp. 1701-1717.

[8] Slater, S., Baker, R. S. (2016), Tools for Educational Data Mining: A Review, Journal of Educational and Behavioral Statistics XX(X): pp. 1-22.

[9] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. V. (2019), Educational Data Mining: Predictive Analysis of Academic Performance of Public-School Students in the Capital of Brazil, Journal of Business Research Vol. 94: pp. 335-343.

[10] Feldman, Y., Barhoom, S., Blonder. R., & Tuvi., I. (2021), Behind the Scenes of Educational Data Mining, Education and Information Technologies 26: pp. 1455-1470.

[11] Martínez F., Gamazo A., M. Rodríguez M. (2020), Educational Data Mining: Identification of Factors Associated with School Effectiveness in PISA Assessment, Studies in Educational Evaluation 66: pp. 1-10.

[12] Mudasir A., Majid Z., Muheet A (2020)., An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches, Procedia Computer Science 167: pp. 1471-1483.

[13] Chen C., Zhang Q., Yu B., Yu Z., Lawrence P. J., Ma Q., Zhang Y. (2020), Improving Protein-Protein Interactions Prediction Accuracy Using

XGBoost Feature Selection and Stacked Ensemble Classifier, Computers in Biology and Medicine 123: pp. 1-12.

آرش خسروی مدرک کارشناسی خود را در رشته مهندسی نرم افزار در سال ۱۳۸۲ از دانشگاه صنعتی اصفهان، مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات در سال ۱۳۹۲ و مدرک دکتری خود را در رشته مهندسی فناوری اطلاعات، گرایش سیستمهای اطلاعاتی در سال ۱۳۹۶ از دانشگاه صنعتی مالزی اخذ کرده است. ایشان در حال حاضر به عنوان هیات علمی مرکز آموزش عالی محلات مشغول به کار هستند. زمینه های پژوهشی مورد علاقه ایشان عبارتند از: هوش تجاری، سیستم های پیشنهاد دهنده، مدیریت دانش مشتری، داده کاوی، متن کاوی و فناوری اطلاعات در پزشکی.
نشانه رایانامه ایشان عبارتند از:
Khosravi.280@gmail.com

مرتضی رجب زاده مدرک کارشناسی خود را در رشته مهندسی صنایع در سال ۱۳۸۱ از دانشگاه شمال مازندران، مدرک کارشناسی ارشد خود را در رشته مهندسی صنایع در سال ۱۳۸۳ از دانشگاه ملی هوافضای خارکُف اوکراین و مدرک دکترا را در رشته مهندسی صنایع در سال ۱۳۹۲ از دانشگاه ایالتی صومی اوکراین اخذ کرده است. ایشان در حال حاضر عضو هیئت علمی مرکز آموزش عالی محلات می باشد. زمینه های پژوهشی مورد علاقه ایشان عبارتند از: مدیریت کیفیت، مدیریت بهره وری، مدیریت ریسک، استانداردسازی و سیستم های مدیریت یکپارچه.
نشانه رایانامه ایشان عبارتند از:
Rajabzadeh.m@gmail.com

محمد نوری خضرآبادی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر در سال ۱۳۹۷ و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک در سال ۱۳۹۹ از موسسه آموزش عالی پویش اخذ کرده است. ایشان در حال حاضر به عنوان رئیس اداره پشتیبانی فنی پایگاه ها در موسسه آموزشی و پژوهشی امام خمینی (ره) مشغول به کار هستند.
نشانه رایانامه ایشان عبارتند از:
Nooriemail@gmail.com

روش ارجاع به مقاله:
آ. خسروی، م. رجب زاده، م. نوری خضرآبادی. یک روش بهبود یافته جهت پیش بینی عملکرد آموزشی و تحصیلی دانشجویان، مبتنی بر تکنیک های داده کاوی و یادگیری ماشین، دوفصلنامه محاسبات و سامانه های توزیع شده، سال پنجم، شماره ۲، شماره پیاپی ۱۰، صفحه ۷۵ تا ۸۷، سال ۱۴۰۱

How to cite:
Arash Khosravi, Morteza Rajabzadeh, Mohammad Nouri Khezrabadi. Development of an Improved Method for Predicting Educational and Academic Performance of Students, Based on Data Mining and Machine Learning, Journal of Distributed Computing and Systems (JDCS), Vol 5, Issue 2, Page 75-87, 2023.

(پیوست -۱): رتبه‌بندی ویژگی‌ها با استفاده از کدنویسی پایتون

```
#خواندن داده ها و تعیین ویژگی هدف
X_train = pd.read_csv("average.csv")
y = X_train['TotalAve']
X = X_train.drop('TotalAve', axis=1).select_dtypes(include=[np.number])

# ایجاد مدل
clf_xgBoost = xgb.XGBClassifier()
# مقدار دهی مدل
clf_xgBoost.fit(X, y)

# دریافت موارد اهمیت xgBoost
importance_dict = {}
for import_type in ['weight', 'gain', 'cover']:
    importance_dict['xgBoost-'+import_type] = clf_xgBoost.get_booster()
                                              .get_score(importance_type=import_type)
# مقایسه بیشترین و کمترین مقادیر
importance_df = pd.DataFrame(importance_dict).fillna(0)
importance_df = pd.DataFrame(
    preprocessing.MinMaxScaler().fit_transform(importance_df),
    columns= importance_df.columns,
    index=importance_df.index
)

# تولید میانگین برای مقادیر اهمیت
importance_df['mean'] = importance_df.mean(axis=1)

# ترسیم نمودار
importance_df.sort_values('mean').plot(kind='bar', figsize=(20, 7))
plt.show()
```