Journal of New Approach to Children's Education (JNACE) **Original Article**

Winter-2024, Volume 5, No 4 **DOI: 10.22034/NAES.2024.451241.1385**

ISSN: 2716-9766

GrossMark

Presenting Approaches based on Traditional Machine Learning and Regression on Predicting the Performance of Students of Higher Institutions

Shahram Ranjdoust [1*], Zinat Khezridenkhe [2]

1 Associate Professor Department of curriculum planinng,Marand branch,Islamic Azad University, Marand, Iran.

2 doctoral student of crriculum planning department of islamic azad university marand branch

*** Corresponding author:** Dr.Ranjdoust@gmail.com

Received: 2023-12-08 **Accepted:** 2024-03-29

Abstract

Introduction: Predicting student performance has become an urgent demand in most educational and higher education institutions. This is essential to help at-risk students and ensure their retention, provide excellent learning resources and experiences, and improve the ranking and reputation of institutions. However, this may be difficult to achieve for start-up organizations with small records to analyze. The purpose of the current research was to provide approaches based on traditional machine learning and regression on predicting students' performance.

Method: The current research was of the qualitative research type and applied in terms of purpose and experimental analytical research in terms of method. Linear regression, decision tree, random forest and support vector machine methods were used in this research. In this section, after introducing the implementation environment, simulation parameters were introduced. In the following, by introducing the efficiency evaluation criteria of the proposed method, based on the described evaluation criteria, the findings were compared with other similar methods. For these comparisons, deep learning approach based on deep convolutional network and other deep learning approaches are used. In this research, the data collection of Dr. Hasht Roudi boys' school, which is among the top 10 institutions in Tehran, was used. The data of this institution are publicly available and can be downloaded through the GitHub site. Further investigation has been done on these data. The figure below shows the frequency of features in the dataset. that these features are considered as and on models.

Keywords: Prediction of student performance, Regression system, Machine learning, Higher institutions

How to Cite This Article: Ranjdoust , Sh. (2024). Presenting Approaches based on Traditional Machine Learning and Regression on Predicting the Performance of Students of Higher Institutions *JNACE*, 5(4): 31-44.

فصلنامه رویکردی نو بر آموزش کودکان | مقاله پژوهشی

زمستان ۱۴۰۲، دوره ۵، شماره ۴

DOI: 10.22034/NAES.2024.451241.1385

ISSN: 2716-9766

GrossMark

ارائه رویکردهای مبتنی بر یادگیری ماشین سنتی و رگرسیونی روی پیش بینی عملکرد دانش آموزان مؤسسات عالی

شهرام رنجدوست*[۱]، زینت خضری دنخه[۲]

[۱] دانشیارگروه برنامه ریزی درسی، واحد مرند؛ دانشگاه آزاد اسلامی؛مرند؛ایران.

[۲] دانشجو دکتری تخصصی گزوه برنامه ریزی درسی دانشگاه آزاد اسلامی واحد مرند

* نویسنده مسئول: Dr.Ranjdoust@gmail.com

تاریخ دریافت مقاله: ۱۴۰۲/۰۹/۱۷ **تاریخ پذیرش مقاله:** ۱۴۰۳/۰۱/۱۰

چکیده

مقدمه: پیش‌بینی عملکرد دانش‌آموزان به یک خواسته مبرم در اکثر نهادها و مؤسسات آموزشی و آموزشی عالی تبدیل شدهاست. این مسئله برای کمک به دانشآموزان در معرض خطر و اطمینان از حفظ آنها، ارائه منابع و تجربیات عالی یادگیری و بهبود رتبه و شهرت مؤسسات ضروری است. با این حال، دستیابی به آن برای مؤسسات استارت‌آپی که سوابق کوچکی برای تجزیه و تحلیل دارند، ممکن است دشوار باشد. هدف از پژوهش حاضر ارائه رویکردهای مبتنی بر یادگیری ماشین سنتی و رگرسیونی روی پیش بینی عملکرد دانش آموزان بود.

روش: پژوهش حاضر از نوع پژوهش‌های کیفی بوده و از لحاظ هدف کاربردی و از لحاظ روش از نوع پژوهش‌های تحلیلی آزمایشی بود. در این پژوهش از روش‌های رگرسیون خطی، درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبانی استفاده شد. در این بخش پس از معرفی محیط پیاده‌سازی، پارامترهای شبیه‌سازی معرفی شد. در ادامه نیز با معرفی معیارهای ارزیابی کارایی روش پیشنهادی بر اساس معیارهای ارزیابی موصوف بررسی و یافته‌ها با دیگر روش‌های مشابه مقایسه شد. که برای این مقایسات از رویکرد یادگیری عمیق مبتنی بر شبکه کانولوشنی عمیق و دیگر رویکردهای یادگیری عمیق استفاده می‌شود. در این تحقیق همچنین از مجموعه داده‌های مدرسه پسرانه دکتر هشت رودی که جزو ۱۰ مؤسسه برتر در تهران می‌باشد استفاده شد.

نتیجه‌گیری: نتایج اصلی این مطالعه کارایی جنگل تصادفی را در آموزش داده‌های کوچک و در تولید نرخ آزمون دقیق نشانمیدهد.

واژگان کلیدی: پیش بینی عملکرد دانش آموزان، سیستم رگرسیونی، یادگیری ماشین، مؤسسات عالی

شیوه استناد به این مقاله: رنجدوست، ش (۱۴۰۲) ارائه رویکردهای مبتنی بر یادگیری ماشین سنتی و رگرسیونی روی پیش بینی عملکرد دانش آموزان مؤسسات عالی. *فصلنامه رویکردی نو بر آموزش کودکان* ، ۵(۴): ۳۱-۴۴.

مقدمه

پیش بینی عملکرد دانش‌آموزان برای هر مؤسسه آموزشی با هدف بهبود عملکرد و پایداری دانش‌آموزان دارای ارزش است. براساس پیش بینی‌های انجام‌شده دانش آموزانی که در معرض ترک تحصیل هستند قابل پیش بینی هستند و همچنین عملکرد آنها نیز قابل پیش بینی است. در سیستم‌های آموزشی کنونی پیش بینی عملکرد دانش‌آموزان روز به‌روز رو به وخامت می‌رود. پیش بینی عملکرد دانش‌آموزان از قبل می‌تواند به دانش‌آموز و استاد آنها کمک کند تا پیشرفت دانش‌آموز را پیگیری کنند. امروزه بسیاری از مؤسسات سیستم ارزیابی مداوم را به تصویب رسانده‌اند. چنین سیستم‌هایی در بهبود عملکرد دانش‌آموز برای

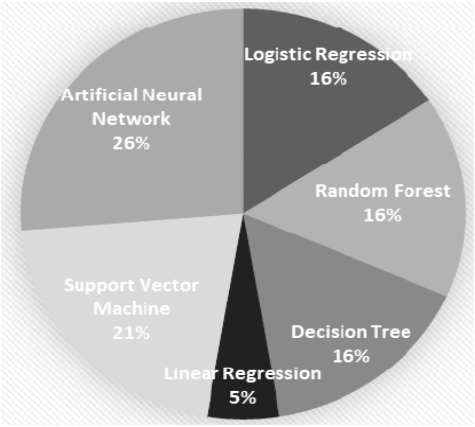
دانش‌آموزان مفید است. هدف از سیستم ارزیابی مداوم کمک به دانش‌آموزان عادی است.

در سال‌های اخیر، مدل‌های پیش بینی رگرسیونی عملکرد و موفقیت‌های زیادی را در طیف گسترده‌ای از برنامه‌های داده کاوی کسب کرده‌اند، که اغلب از سایر طبقه بندها نیز دقت بالاتری را به‌دست آورده‌اند. هدف از این مطالعه بررسی اینکه آیا مدل‌های سنتی یادگیری ماشین پیش بینی کن نده‌های مناسبی برای پیش بینی عملکرد دانش‌آموزان از داده‌های سیستم مدیریت یادگیری در زمینه داده کاوی آموزشی هستند یا خیر.

هدف این تحقیق مقایسه بین الگوریتم‌های یادگیری سنتی برای پیش بینی عملکرد دانش‌آموزان و پیش بینی نمرات آنها براساس نمرات کسب کرده در گذشته‌است. مؤسسات آموزش عالی باید با ارائه کیفیت آموزشی به دانش‌آموزان انگیزه دهند زیرا آموزش موضوع مهمی برای توسعه یک کشور است. برای بهبود کیفیت و رقابت دانش‌آموزان، مؤسسات باید برای دستیابی به اهداف خود استراتژی‌های خاصی داشته‌باشند و برای اجرای این راهکارها، مؤسسات نیاز دارند که توجه خود را به‌سمت یافتن راه حل برای مشکلات موجود جلب کنند. برای این موضوع، عوامل مؤثر در موفقیت دانش‌آموزان باید زودهنگام شناسایی شوند. موفقیت یک دانش‌آموزان خاص را می‌توان با نمره نهایی در موضوع خاص مشاهده کرد. با توجه به اینکه مسئله مورد نظر یک مسئله رگرسیونی است باید از روش‌های یادگیری رگرسیون برای پیش بینی استفاده شود. الگوریتم‌های یادگیری سنتی ماشین مبتنی بر انتخاب ویژگی‌ها هستند. یعنی هرچقدر ویژگی‌ها بهتری انتخاب شوند نتایج بهتری به‌دست می‌آید. انتخاب ویژگی‌های دستی یکی از چالش‌های این الگوریتم‌ها می‌باشد. زمانی که حجم داده‌ها کم باشد این الگوریتم‌ها پیش برازش می‌شوند. هدف ما در این پژوهش مقایسه بین این الگوریتم‌ها می‌باشد.

- اعمال الگوریتم‌های سنتی یادگیری ماشین نظیر درخت تصمیم رگرسیونی، ماشین بردار رگرسیونی و جنگل تصادفی رگرسیونی برای پیش بینی نمرات دانش‌آموزان.
- مقایسه بین الگوریتم‌های یادگیری سنتی ماشین
- شناسایی عوامل مؤثر در پیش بینی نمرات دانش‌آموزان

در حالت کلی تحقیقات زیادی بر روی تصمیم گیری استراتژیک در مؤسسات آموزش عالی انجام‌شده است نویسندگان در [۱] کارهای بررسی شده را به‌صورت نموداری نشان داده‌اند.

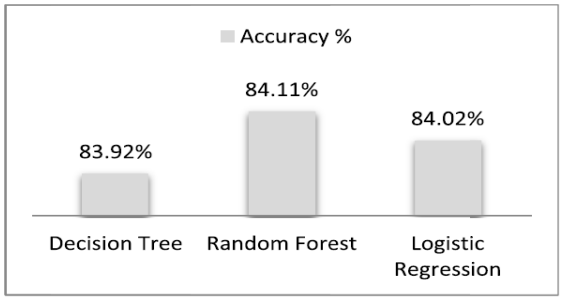


شکل ۱: درصد رویکردهای یادگیری ماشین در پیش بینی عملکرد دانش‌آموزان

با توجه به این نمودار اکثر کارهای انجام‌شده در ادبیات با استفاده از شبکه‌های عصبی بوده است. و درصد کمی از آنها را الگوریتم‌های یادگیری ماشین سنتی در بر گرفته است. شبکه‌های عصبی به‌دلیل انتخاب ویژگی‌ها به‌صورت سنتی انتخاب مناسبی برای این مورد می‌باشند.

نویسندگان در [۲] برای پیش بینی عملکرد دانشجویان در دانشگاه‌های اردن از دو طبقه بند درخت تصمیم و شبکه‌های عصبی مصنوعی استفاده کردند. نوع درخت تصمیم مورد استفاده j48 می‌باشد و برای پیاده‌سازی آنها از نرم‌افزار وکااستفاده کردند. آنها برای ارزیابی مدل روش اعتبارسنجی k فولد را که مقدار k را برابر ۱۰ در نظر گرفتند استفاده کردند. دقت‌های به‌دست آمده نشان دهنده این است که شبکه عصبی مصنوعی بهتر از درخت تصمیم نتایج را به‌دست آورده‌است.

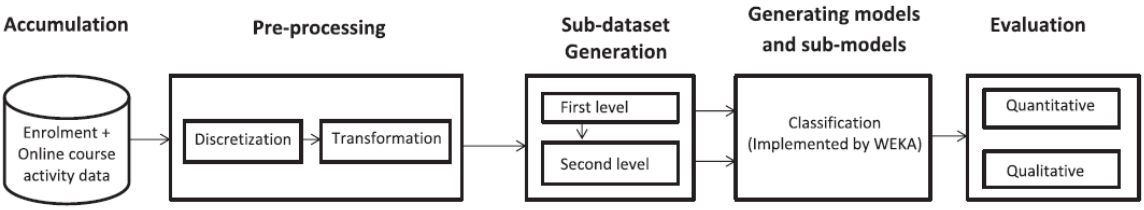
در [۱] الگوریتم‌ها برای پیش بینی نرخ فارغ‌التحصیل در مورد دانشجویان مهندسی کارشناسی‌ارشد در آمریکای جنوبی مورد استفاده قرار گرفتند. سه الگوریتم درخت تصمیم ,sجنگل تصادفی ,رگرسیون‌های لجستیک در این تحقیق مورد استفاده قرار گرفت. با توجه به نوع داده‌های الگوریتم رگرسیون‌های لجستیک بالاترین دقت را به‌دست آورده‌است نمودار زیر دقت به‌دست آمده بر روی این مجموعه داده‌ها می‌باشد.



شکل ۲: دقت‌های به‌دست آمده توسط رویکردهای مختلف

در [۳] نویسندگان با استفاده از داده‌های جمع آوری شده از دانشگاه‌های استرالیا مدل‌های طبقه بندی متفاوتی را برای پیش بینی عملکرد دانشجویان ایجاد کردند. این داده‌ها شامل جزئیات ثبت‌نام دانشجویان همچنین فعالیت تولید شده در سیستم مدیریت دانشگاه LMS می‌باشد. اطلاعات ثبت‌نام شامل اطلاعات دانشجویی از جمله ویژگی‌های جمعیت‌شناختی اجتماعی، پایه پذیرش دانشگاهی (مثلاً از طریق آزمون یا از طریق استعداد درخشان) و نوع حضور و غیاب (تمام وقت یا پاره‌وقت) می‌باشد. داده‌های LMS نیز فعالیت‌های آنلاین دانشجو را در بر می‌گیرد. مهم‌ترین بخش این تحقیق توجه به ناهمگونی دانشجویان در ساختار مدل پیش بینی است.

مدل پیشنهادی آنها در شکل ۱ آورده شده‌است. مدل پیشنهادی شامل ۵ جزء اصلی است که در ادامه به هر یک از آنها به‌صورت جزئی پرداخته می‌شود.



شکل ۳: رویکرد پیشنهادی در [1] برای پیش بینی عملکرد دانشجویان

۱. جمع آوری داده‌های ثبت‌نام و LMS: این مطالعه شامل داده‌های اجتماعی و جمعیت‌شناختی و آکادمیک جمع آوری شده در هنگام ثبت‌نام دانشجویان و داده‌های فعالیت به‌دست آمده از دانشگاه LMS-Moodle است. مجموعه داده‌های ثبت‌نام شامل ویژگی‌های اجتماعی جمعیت‌شناختی (سن، جنسیت و وضعیت اقتصادی) و دانشگاهی (نوع حضور) یک دانشجو است. عملکرد دانشجو توسط میانگین نمره در تمامی دوره‌ها که در یک سال گذرانده است نشان داده شده می‌شود[۳].

۲. گام پیش پردازش داده‌ها: پردازش داده‌ها مرحله‌ی مهم برای تهیه داده‌ها قبل از استفاده از روش‌های داده کاوی است. پیش پردازش در دو مرحله تفسیر و تبدیل انجام‌می‌شود.

- تفسیر: تمامی فعالیت‌های دانشجویان به چهار طبقه Q1, Q2, Q3, Q4 تبدیل می‌شوند که Q1 کمترین مشارکت و Q4 بیشترین مشارکت را نشان‌می‌دهد.
- تبدیل: اجرای مدل‌های طبقه بندی نیازمند داده‌ها در قالب عددی می‌باشند. برای این منظور تمامی داده‌ها به فرمت عددی تبدیل می‌شوند

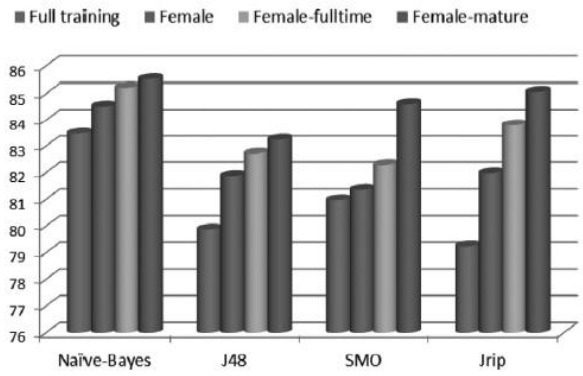
۳. تولید زیر مجموعه داده‌ها: مجموعه داده‌ها به چندین زیر مجموعه داده تقسیم می‌شوند تا زیر گروه‌های دانشجویی تشکیل شوند پارتیشن بندی مجموعه داده‌ها در دو مرحله به شرح زیر تقسیم می‌شود.

- ثبت‌نام، فعالیت‌ها و مجموعه داده‌های داده شده با توجه به جنسیت دانشجویان (زن و مرد) و نوع حضور و غیات (تمام وقت و پاره‌وقت) و نحوه حضور (داخلی و خارجی) تقسیم می‌شود. از این رو ۸ زیر مجموعه داده ایجاد شده‌است.
- زیر مجموعه داده‌های زن و مرد با توجه به سن دانشجو نوع حضور و نحوه حضور به ۶ زیر مجموعه تقسیم شد.

۴. تولید مدل و زیر مدل‌ها: نویسندگان از چهار مدل طبقه بندی برای پیش بینی عملکرد دانشجویان استفاده کردند که شامل Naïve Bayes, SMO, J48, jRip بودند برای این کار استفاده کردند. برای پیاده‌سازی این مدل‌ها از مدل‌های ایجاد شده توسط نرم‌افزار وکاااستفاده شد.

- Naïve Bayes: یک روش احتمالی براساس تئوری بیزین است
- SMO: این طبقه بند از یک الگوریتم بهینه سازی برای آموزش SVM استفاده می‌کند
- J48: این متد درخت تصمیم که شامل سه نوع گره متفاوت (ریشه، میانی و برگ) را تولید می‌کند
- jRip: یک روش طبقه بندی است که بر اساس if-then-else عمل طبقه بندی را انجام می‌دهد.

آنالیز نتایج به‌دست آمده توسط رویکردهای پیشنهادی در شکل ۲ آورده شده‌است



شکل ۴: نتایج به‌دست آمده در[۳] برای مجموعه داده‌های جمع آوری شده

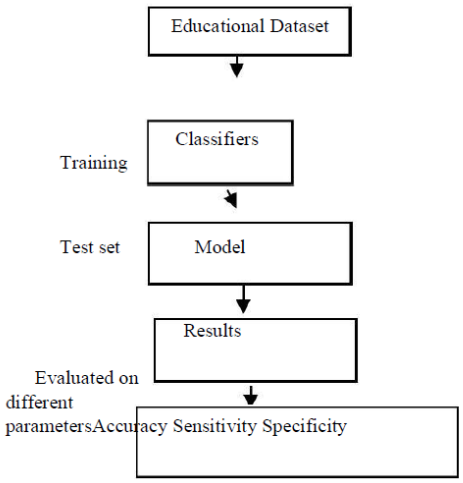
بررسی نتایج حاکی از این است که مدل Naïve Bayes دقت بالاتری را نسبت به مدل‌های دیگر به‌دست آورده‌است. از طرفی مدل Jrip نیز نتایج قابل قیاسی را تولید کرده است.

در [۴] نویسندگان از رویکردهای یادگیری عمیق برای پیش بینی عملکرد دانشجویان استفاده کردند.

طبقه بند داده کاوی و یادگیری عمیق بر روی مجموعه داده‌های جمع آوری شده از محیط آموزشی داده‌می‌شود. داده‌ها پیش پردازش شده و مقادیر گمشده‌ی آنها پیش بینی می‌شوند. برای ساخت مدل‌ها از مجموعه‌ای از طبقه بندها استفاده شده‌است. مدل‌ها با داده‌های آزمون برای پیش بینی عملکرد دانشجویان مورد آزمایش قرار می‌گیرند و بهترین مدل‌هایی که دقت بالایی دارند، در نظر گرفته می‌شوند. متدولوژی پیشنهادی در این تحقیق دارای مراحل مهمی است که در زیر شرح داده‌شده‌است.

- در مرحله اول، طبقه بندی شبکه عصبی MLP-a و طبقه بندی کن نده‌های داده کاوی یعنی Bayes Net، SVM، جنگل تصادفی، درخت تصمیم و طبقه بندی چند طبقه بر روی مجموعه داده‌های آموزشی اعمال می‌شو ندو یک مدل به‌دست می‌آید.
- مدل به‌دست آمده از مرحله اول با مجموعه داده‌های آزمون تهیه شده مورد آزمون قرار می‌گیرد و نتایجی به‌دست می‌آید. در این گام 10-Fold cross validation مورد استفاده قرار می‌گیرد
- نتایج به‌دست آمده بر روی پارامترهایی از قبیل دقت، ویژگی، حساسیت، کاپا- آماری و منحنی ROC ارزیابی شده و بهترین مدل با نتایج بالا در مرحله سوم انتخاب می‌شود.

چارچوب پیش بینی عملکرد دانشجویان در زیر آورده شده‌است.



شکل ۵: مدل پیشنهادی برای پیش بینی عملکرد دانشجویان [۴]

نتایج به‌دست آمده توسط رویکردهای پیشنهادی آنها در جدول زیر آمده است.

جدول ۱. رویکردهای پیشنهادی برای پیش بینی عملکرد دانشجویان

روش	میانگین	TP	FP	ضریب کاپا	ROC
MLP	۹۹/۴۵	۱	۰	۰/۹۹	۱
طبقه بندی چند کلاسه	۹۹/۸۱	۱	۰	۰/۹۹	۱
SVM	۹۳/۹۰	۱	۰/۱۰	۰/۸۹	۰/۹۴
Naïve Bayes	۹۷/۴۵	۰/۹۸	۰	۰/۹۵	۰/۹۹
IBK	۷۹/۸۱	۰/۹۱	۰/۱۶	۰/۶۳	۰/۸۷
Lazy LWL	۸۶/۷۲	۱	۰	۰/۷۵	۱
جنگل تصادفی	۱۰۰	۱	۰	۱	۱
درخت تصمیم	۱۰۰	۱	۰	۱	۱

رویکردهای مبتنی بر شبکه‌های عصبی همچنین در [۵] مورد استفاده قرار گرفته‌اند. این مقاله یک مطالعه جامع در مورد پیش بینی عملکرد دانشجویان در درس برنامه نویسی R است که با استفاده از یادگیری عمیق (که بخش کوچکی از شبکه عصبی مصنوعی است) سعی در پیش بینی عملکرد دانشجویان تحصیلات تکمیلی دارد. هدف از مطالعه ۱) بررسی میزان دقت پیش 2) تحلیل عوامل مؤثر بر پیشرفت تحصیلی است که در پیش بینی عملکرد تحصیلی دانشجویان نقش دارند. محققان ارزیابی روش‌های پیشنهادی را در یک مجموعه داده متشکل از ۳۹۵ سوابق دانشجویی با ۳۰ ویژگی پس از پردازش، پاک‌سازی و فیلتر با استفاده از برنامه نویسی R که از مخازن UCI جمع آوری شده بود مورد تست و ارزیابی قرار دادند. در این مطالعه رویکردهای مختلفی نظیر یادگیری عمیق، جنگل تصادفی و رگرسیون خطی مورد استفاده قرار گرفتند. نتایج حاصل از ماتریس سردرگمی رویکردهای پیشنهادی آنها در جدول زیر آورده شده‌است.

جدول ۲. نتایج ماتریس سردرگمی و رویکردهای پیشنهادی آنها

Linear Regression	جنگل تصادفی	Deep learning (Artificial Neural Networks)			
		Total Data	Test data	Training Data	
12.339%	28.101%	97.429%	94.872%	97.749%	Accuracy
21.918%	43.874%	98.698%	97.368%	99.023%	Precision
22.018%	43.874%	98.698%	97.368%	98.701%	Recall
21.968%	43.874%	98.698%	97.368%	98.862%	F-Measure

نتایج ارزیابی آنها نشان‌می‌دهد که رویکردهای یادگیری عمیق و شبکه‌های عصبی دقت بالاتری را نسبت به سایر روش‌ها به‌دست آورده‌اند. در حالت کلی می‌توان تلاش‌های مختلف در ادبیات را برای حل این مسئله در جدول زیر خلاصه کرد:

جدول ۳. نتایج ارزیابی ادبیات پژوهش در رویکردهای یادگیری عمیق

ML architecture	Study
K-means	Liu & Li [6]
CNN+RNN	.[۷]Wang et al
DT	Al-Shabandar et al. [8]
DT	Al-Shabandar et al. [9]
DNN	Whitehill et al. [10]
DT+LR	Nagrecha et al. [11]
Bayesian Net.+DT	Xing et al. [12]
NLP	Robinson et al. [13]
SVM+LR	Qiu et al. [12]
LR+SVM	Liang et al. [14]
NLP	Crossley et al. [15]
Multinomal LR	Whitehill et al. [16]
LR	Boyer and Veeramachaneni [17]
ANN	Chaplot et al. [18]
NLP	[۱۹]Coleman et al.
LR	Kizilcec et al. [20]
RNN+HMM	Fei & Yeung [21]
SVM	Kloft et al. [22]

روش‌های پژوهش

پژوهش حاضر از نوع پژوهش‌های کیفی بوده و از لحاظ هدف کاربردی و از لحاظ روش از نوع پژوهش‌های تحلیلی آزمایشی بود. در این پژوهش از روش‌های رگرسیون خطی، درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبانی استفاده شد. در این بخش پس از معرفی محیط پیاده‌سازی، پارامترهای شبیه‌سازی معرفی شد. در ادامه نیز با معرفی معیارهای ارزیابی کارایی روش پیشنهادی بر اساس معیارهای ارزیابی موصوف بررسی و یافته‌ها با دیگر روش‌های مشابه مقایسه شد. که برای این مقایسات از رویکرد یادگیری عمیق مبتنی بر شبکه کانولوشنی عمیق و دیگر رویکردهای یادگیری عمیق استفاده می‌شود. در این تحقیق همچنین از مجموعه داده‌های مدرسه پسرانه دکتر هشت رودی که جزو ۱۰ مؤسسه برتر در تهران می‌باشد استفاده شد. داده‌های این مؤسسه به‌صورت عمومی در دسترس می‌باشند و از طریق سایت گیت هاب قابل دانلود می‌باشند. در ادامه بررسی بر روی این داده‌ها انجام‌شده است. در شکل زیر فراوانی ویژگی‌های موجود در مجموعه داده‌ها آورده شده‌است. که این ویژگی‌ها به‌عنوان و روی مدل‌ها در نظر گرفته می‌شوند.

رگرسیون خطی

رگرسیون خطی یک رویکرد مدل خطی بین متغیر پاسخ با یک یا چند متغیر توصیفی است. فرم مدل رگرسیون خطی ساده به‌صورت زیر است[۲۳]:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

همان طور که دیده‌می‌شود این رابطه، معادله یک خط است که جمله خطا یا همان ϵ به آن اضافه‌شده. پارامترهای این مدل خطی عرض از مبدا ($\beta 0$) و شیب خط ($\beta 1$) است. شیب خط در حالت رگرسیون خطی ساده، نشان‌می‌دهد که میزان حساسیت متغیر وابسته به متغیر مستقل چقدر است. به این معنی که با افزایش یک واحد به مقدار متغیر مستقل چه میزان متغیر وابسته تغییر خواهد کرد. عرض از مبدا نیز بیانگر مقداری از متغیر وابسته است که به ازاء مقدار متغیر مستقل برابر با صفر محاسبه می‌شود [۲۳].

درخت تصمیم

درخت تصمیم مدل‌های طبقه‌بندی کاتیون را در قالب یک ساختار درختی می‌سازد [۱۵۶]. این یک مجموعه داده را به زیر مجموعه‌های کوچک‌تر و کوچک‌تر تجزیه می‌کند در حالی که در همان زمان یک درخت تصمیم مرتبط به‌طور تدریجی توسعه می‌یابد. نتیجه نهایی درختی با گره‌های تصمیم گیری و گره‌های برگ است. یک گره تصمیم دارای دو یا چند شاخه است و یک گره برگ نشان دهنده یک طبقه بندی یا تصمیم است. بالاترین گره تصمیم گیری در یک درخت که با بهترین پیش بینی کنند به نام گره ریشه مطابقت دارد.

جنگل تصادفی

جنگل تصادفی یک الگوریتم طبقه بندی آسان برای استفاده است که حتی بدون تنظیم فرا پارامتر، نتایج عالی برای بسیاری از مشکلات ایجاد می‌کند [۲۵]. جنگل‌های تصادفی یک روش یادگیری مجموعه‌ای هستند که می‌توانند برای مسائل طبقه‌بندی و رگرسیون استفاده شوند. این الگوریتم با ساخت انبوهی از درخت‌های تصمیم در زمان آموزش و خروجی کلاسی که حالت کلاس‌ها (طبقه‌بندی) یا پیش‌بینی میانگین (رگرسیون) هر درخت است، عمل می‌کند. جنگل تصادفی همزمان با رشد درختان، تصادفی بیشتری را به مدل اضافه می‌کند. به‌جای جستجوی مهم‌ترین ویژگی در حین تقسیم یک گره، بهترین ویژگی را در میان زیرمجموعه‌ای تصادفی از ویژگی‌ها جستجو می‌کند.

ماشین بردار پشتیبانی

ماشین بردار پشتیبان (SVM) یک طبقه بندی متمایز است که به‌طور رسمی توسط یک ابر صفحه جداکننده تعریف شده‌است [۲۷]. به‌عبارت دیگر، با توجه به داده‌های آموزشی برچسب‌گذاری شده (یادگیری تحت نظارت)، الگوریتم یک ابر صفحه بهینه را خروجی می‌دهد که نمونه‌های جدید را دسته‌بندی می‌کند. الگوریتم SVM در عمل با استفاده از یک هسته پیاده‌سازی می‌شود. هسته شباهت یا اندازه گیری فاصله بین داده‌های جدید و بردارهای پشتیبانی را تعیین می‌کند. به‌عنوان مثال. یادگیری ابر صفحه در SVM خطی / شعاعی با تبدیل مسئله با استفاده از جبر خطی / شعاعی انجام‌می‌شود. SVM معمولاً در فضاهای با ابعاد بالا مؤثر است.

شیوه اجرا. در این بخش پس از معرفی محیط پیاده‌سازی، پارامترهای شبیه‌سازی معرفی شد. در ادامه نیز با معرفی معیارهای ارزیابی کارایی روش پیشنهادی بر اساس معیارهای ارزیابی موصوف بررسی و یافته‌ها با دیگر روش‌های مشابه مقایسه شد. که برای این مقایسات از رویکرد یادگیری عمیق مبتنی بر شبکه کانولوشنی عمیق و دیگر رویکردهای یادگیری عمیق استفاده می‌شود.

محیط پیاده‌سازی پژوهش. برای پیاده‌سازی مدل‌های رگرسیونی از زبان برنامه نویسی پایتون استفاده شد. پایتون یک زبان برنامه نویسی محبوب است که توسط گیدو ون روسوم ساخته شد و در سال ۱۹۹۱ منتشر شد. این زبان برای:

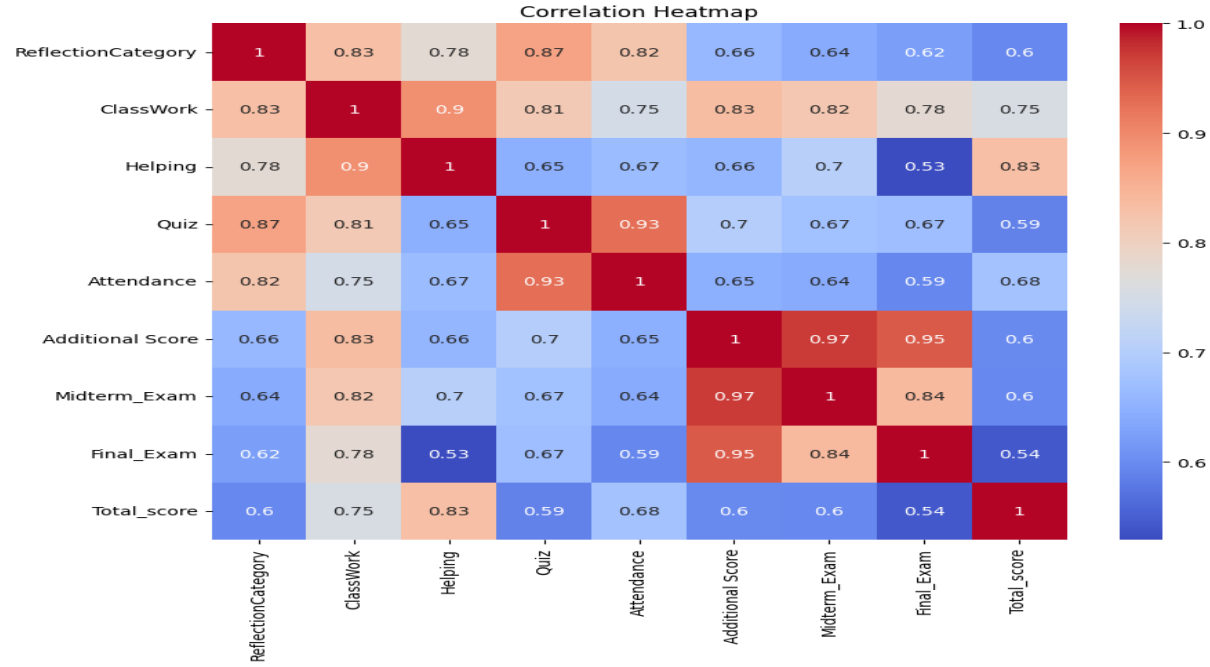
- توسعه وب (سمت سرور)
- توسعه نرم‌افزار
- ریاضیات
- برنامه نویسی سیستم

مورد استفاده قرار می‌گیرد

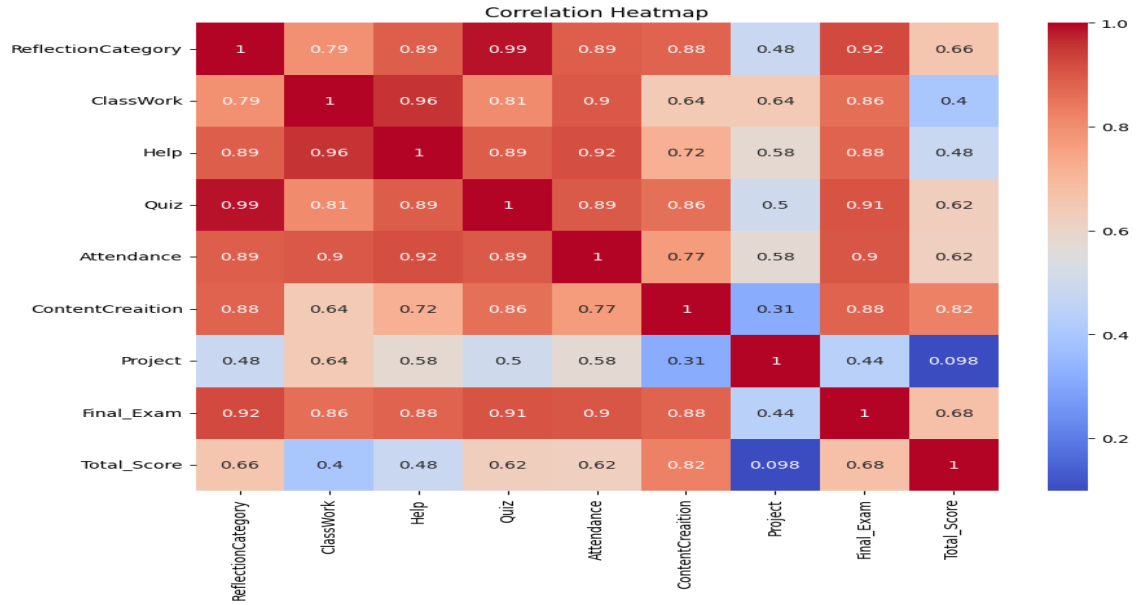
پایتون روی پلتفرم‌های مختلف (ویندوز، مک، لینوکس، رزبری پای و غیره) کار می‌کند. این برنامه نویسی یک نحو ساده شبیه به زبان انگلیسی دارد. پایتون دارای نحوی است که به توسعه دهندگان اجازه می‌دهد برنامه‌هایی را با خطوط کمتری نسبت به سایر زبان‌های برنامه نویسی بنویسند. همچنین این زبان رو یک سیستم مفسر اجرا می‌شود، به این معنی که کد را می‌توان به‌محض نوشتن اجرا کرد. این بدان معنی است که نمونه سازی می‌تواند بسیار سریع باشد. این زبان دارای توابع و کتاب خانه‌های زیادی برای کنترل و دست‌کاری داده‌ها و آموزش مدل‌ها می‌باشد که در این تحقیق ما از کتابخانهscikit-learn آن بهره بردیم. scikit-learn یک ماژول پایتون برای یادگیری ماشینی است که بر روی SciPy ساخته شده‌است و تحت مجوز clause BSD–۳ توزیع شده‌است. این پروژه در سال ۲۰۰۷ توسط دیوید کورناپو به‌عنوان یک پروژه تابستانی کد گوگل آغاز شد و از آن زمان بسیاری از داوطلبان در این پروژه مشارکت داشته‌اند.

جامعه پژوهش

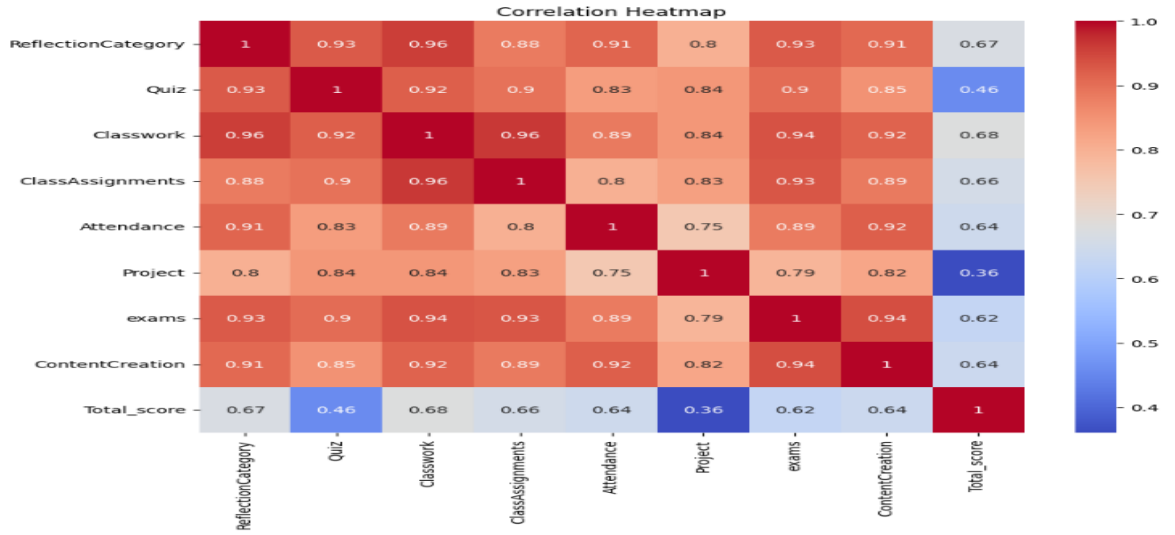
در این تحقیق همچنین از مجموعه داده‌های مدرسه پسرانه دکتر هشت رودی که جزو ۱۰ مؤسسه برتر در تهران می‌باشد استفاده شد. داده‌های این مؤسسه به‌صورت عمومی در دسترس می‌باشند و از طریق سایت گیت هاب قابل دانلود می‌باشند. در ادامه بررسی بر روی این داده‌ها انجام‌شده است. در شکل زیر فراوانی ویژگی‌های موجود در مجموعه داده‌ها آورده شده‌است. که این ویژگی‌ها به‌عنوان و روی مدل‌ها در نظر گرفته می‌شوند.



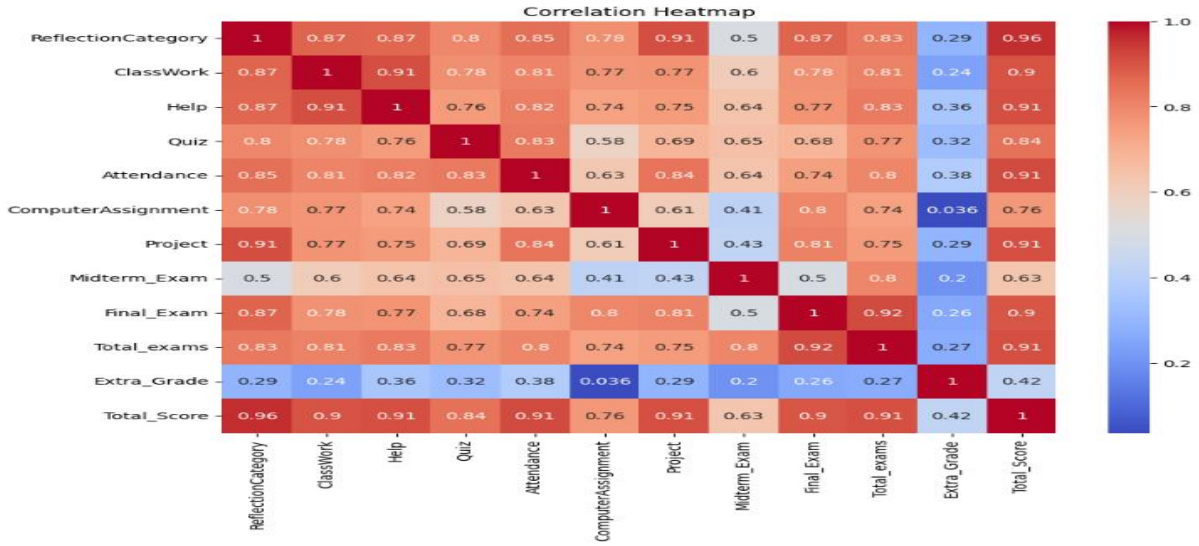
شکل ۶: وابستگی ویژگی‌های مختلف به یک دیگر (داده‌های ریاضی)



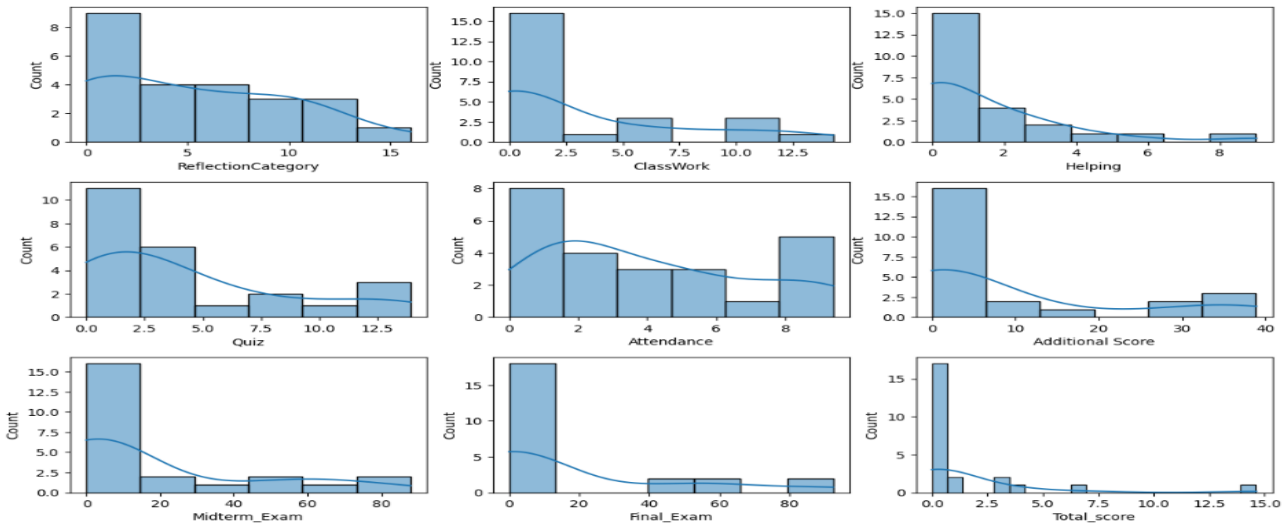
شکل ۷: وابستگی ویژگی‌های مختلف به یک دیگر (داده‌های کامپیوتر)



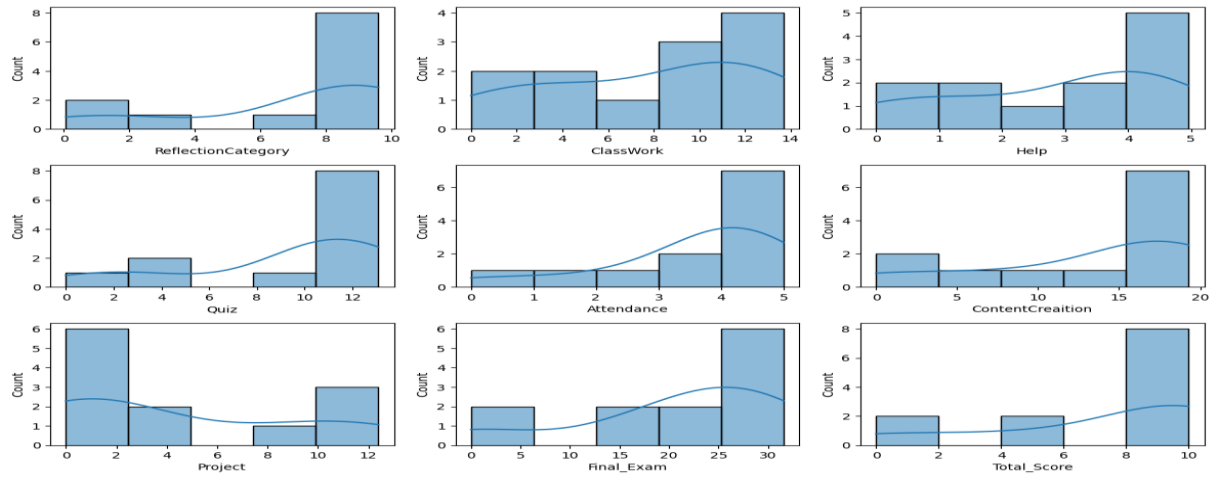
شکل ۸: وابستگی ویژگی‌های مختلف به یک دیگر (داده‌های الکترونیک)



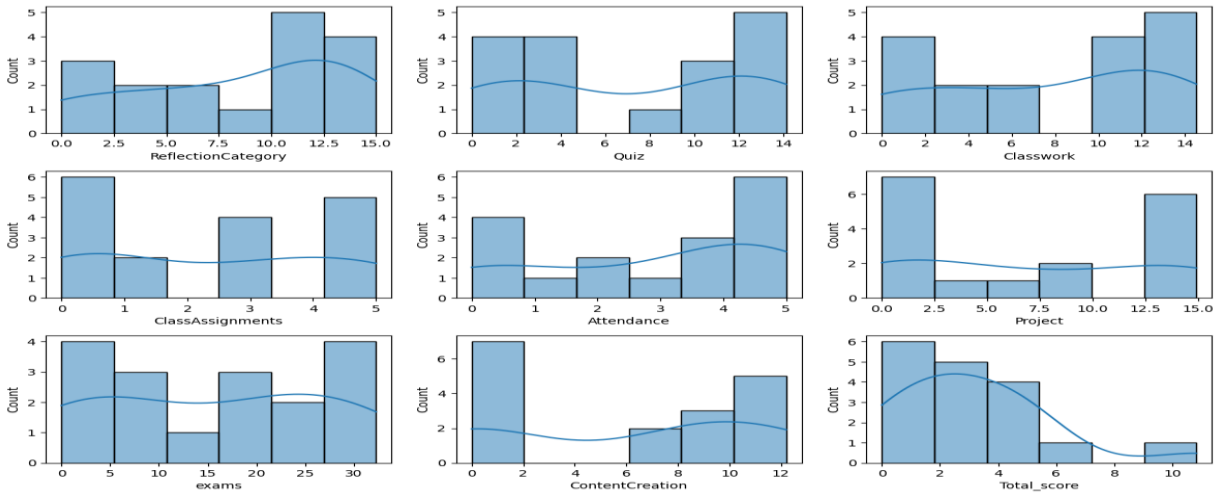
شکل ۹: وابستگی ویژگی‌های مختلف به یک دیگر (داده‌های شیمی)



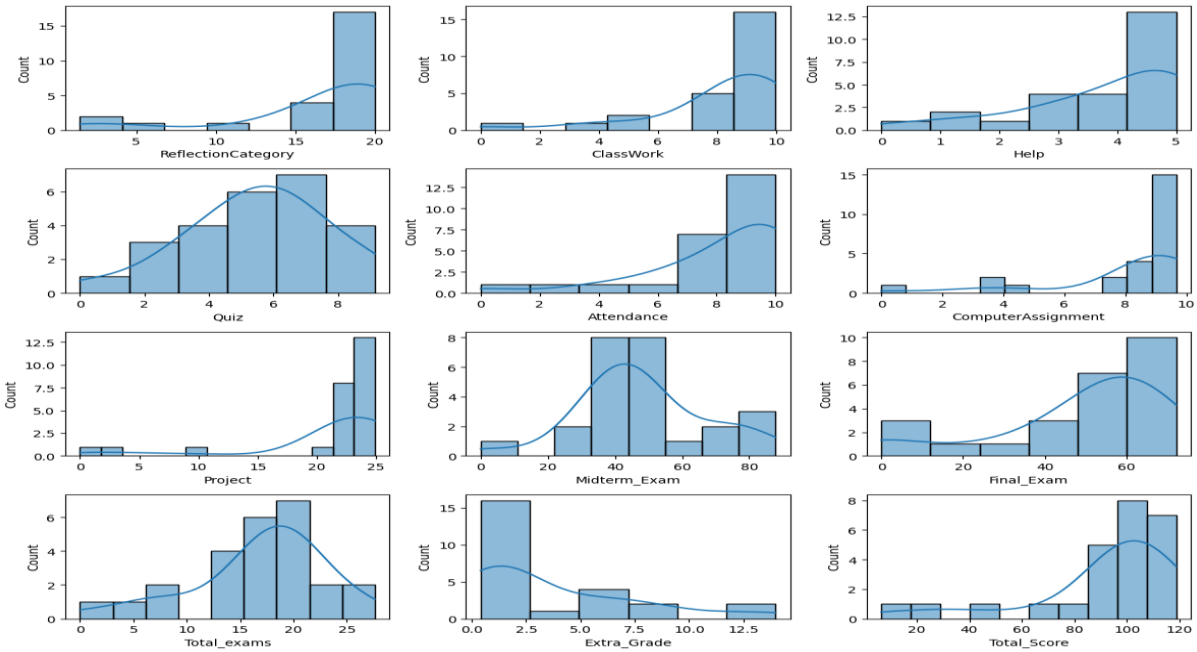
شکل ۱۰: نمودار فراوانی ویژگی‌های مختلف بر روی داده‌های ریاضی



شکل ۱۱: نمودار فراوانی ویژگی‌های مختلف بر روی داده‌های کامپیوتر



شکل ۱۲: نمودار فراوانی ویژگی‌های مختلف بر روی داده‌های الکترونیک



شکل ۱۳: نمودار فراوانی ویژگی‌های مختلف بر روی داده‌های شیمی

٥-٤) پارامترهای شبیه‌سازی

هر یک از رویکردهای بررسی شده در این بخش دارای مدل‌ها و هایپر پارامترهایی می‌باشند که نقش اساسی در به‌دست آوردن نتایج آنها دارند. در ادامه به هر یک از آنها پرداخته شده‌است.

جدول ۴

Model	Hyper parameters
Linear regression	n_estimators=3000, learning_rate=0.05, max_depth=4, max_features='sqrt', min_samples_leaf=15, min_samples_split=10, loss='huber', random_state =5
XGBRegressor	colsample_bytree=0.4603, gamma=0.0468, learning_rate=0.05, max_depth=3, min_child_weight=1.7817, n_estimators=2200, reg_alpha=0.4640, reg_lambda=0.8571, subsample=0.5213, silent=1, nthread = -1
SVR	"C": np.arange(1, 100), "gamma": np.linspace(0.00001, 0.001, 50), "epsilon": np.linspace(0.01, 0.1, 50)
جنگل تصادفی	n_estimators=100
درخت تصمیم	n_estimators=100

معیارهای ارزیابی:

معیارهای عملکرد اعتبار: برچسب امتیاز یک مقدار عددی است، بنابراین تفاوت مقدار پیش‌بینی‌شده و واقعی آن به‌عنوان یک خطا، معیار مهم‌تری برای ارزیابی خواهد بود. معیارهای معروف بسیاری بر اساس این خطاها وجود دارد که به شرح زیر است:

۱- میانگین خطای مطلق (میانگین قدرمطلق خطا): اگر x_i مقدار واقعی و $\hat{x}_i$ مقدار پیش بینی شده باشد، تفاوت آنها مقدار خطا است که حاوی مقادیر مثبت و منفی است و جمع کردن آنها خطاها را خنثی می‌کند، بنابراین مقادیر خطای مطلق آنها را می‌توان جمع کرد. تا مجموع خطاها را بهتر نشان دهند و میانگین آنها نشان دهنده میانگین خطاها است.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|$$

۲- جذر میانگین Root Mean Squared Error (مربعات خطا): مانند میانگین قدرمطلق خطا است اما به‌جای قدر مطلق از مربع برای خنثی کردن خطاهای مثبت و منفی استفاده می‌کند و در نهایت از ریشه برای خنثی کردن اثر مربع استفاده می‌کند. اندازه برداری معمولاً با استفاده از این معادله محاسبه می‌شود، بنابراین اندازه بردار خطا است.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

۳- میانگین درصد مطلق خطا (میانگین درصد قدرمطلق خطا): مانند میانگین قدرمطلق خطا است، اما از خطای نسبی با تقسیم خطا بر مقدار واقعی برای عادی سازی مقادیر خطا استفاده می‌کند.

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{|x_i - \hat{x}_i|}{x_i}$$

یافته‌ها

در ابتدا بر روی داده‌های ریاضی رویکردهای پیشنهادی مورد ارزیابی قرار گرفتند که در این بین رویکرد رگرسیون خطی توانست به میانگین قدرمطلق خطا، ۰/۰۹۲۲= میانگین درصد قدرمطلق خطا ۰/۱۸۳۵= و جذر میانگین مربعات خطا ۰/۱۳۲۷= دست یابد که بدترین رویکرد در بین رویکردهای تست شده بر روی این داده‌ها می‌باشد. رویکرد رگرسیون با ماشین بردار پشتیبان نیز بر روی این داده‌ها به میانگین قدرمطلق خطا ۰/۰۴۰۰= میانگین درصد قدرمطلق خطا ۰/۰۷۹۸= و جذر میانگین مربعات خطا ۰/۰۶۶۱= دست‌یافت. این رویکرد نسبت به ال آر نتایج بهتری را ارائه داد. دو رویکرد جنگل تصادفی و درخت تصمیم نتایج نسبت نزدیک‌تری را به‌هم ارائه دادند این دو رویکرد توانستند به میانگین قدرمطلق خطا ۰/۰۳۲۳= دست یابند. از لحاظ میانگین درصد قدرمطلق خطا رویکرد درخت تصمیم نتایج بهتری را ارائه داد و از لحاظ جذر میانگین مربعات خطا رویکرد جنگل تصادفی بهتر عمل کرده است.

جدول ۵

مدل	جذر میانگین مربعات خطا	میانگین درصد قدرمطلق خطا	میانگین قدرمطلق خطا
Linear regression	۰/۱۳۲۷	۰/۱۸۳۵	۰/۰۹۲۲
SVR	۰/۰۶۶۱	۰/۰۷۹۸	۰/۰۴۰۰
جنگل تصادفی	۰/۰۳۲۴	۰/۱۹۴۳	۰/۰۳۲۳
درخت تصمیم	۰/۱۱۲۴	۰/۱۷۴۳	۰/۰۳۲۳

در ادامه در جدول ۲ نتایج رویکردهای پیشنهادی بر روی داده‌های کامپیوتر آورده شده‌است. در این مجموعه داده‌ها رویکردِ درخت تصمیم بدترین نتایج را به‌دست آورده‌است. این رویکرد بر روی مجموعه داده‌های مورد نظر توانست به میانگین قدرمطلق خطا ۰/۳۰۳۹= به‌دست بیاورد همچنین دومین بدترین نتیجه توسط رگرسیون با ماشین بردار پشتیبان به‌دست آمد این رویکرد به میانگین قدرمطلق خطا ۰/۲۴۷۸= دست‌یافت. رویکردهای جنگل تصادفی در این مجموعه داده‌ها به بهترین عملکرد رسید و توانست به میانگین قدرمطلق خطا ۰/۸۳۵= دست یابد.

جدول ۶

مدل	جذر میانگین مربعات خطا	میانگین درصد قدرمطلق خطا	میانگین قدرمطلق خطا
Linear regression	۰/۲۵۰۶	۰/۴۴۴۶	۰/۲۲۵۲
SVR	۰/۳۸۹۴	۰/۴۹۱۰	۰/۲۴۷۸
جنگل تصادفی	۰/۰۹۹۱	۰/۱۶۴۰	۰/۰۸۳۵
درخت تصمیم	۰/۳۱۰۹	۰/۷۳۰۰	۰/۳۰۳۹

نتایج بر روی داده‌های الکترونیک نیز در ادامه آورده شده‌است. برخلاف دو مجموعه داده‌ی قبلی رویکرد ال آر توانست به کمترین میانگین قدرمطلق خطا بر روی این مجموعه داده‌ها دست یابد. این رویکرد میانگین قدرمطلق خطا، ۰/۱۰۴۷= میانگین درصد قدرمطلق خطا ۰/۲۰۹۰= و جذر میانگین مربعات خطا ۰/۱۱۴۲= دست‌یافت. مدل درخت تصمیم به بدترین نتیجه در این مجموعه داده‌ها دست‌یافت این رویکرد توانست میانگین قدرمطلق خطا، ۰/۳۵۶۹= میانگین درصد قدرمطلق خطا ۰/۶۸۸۱= و جذر میانگین مربعات خطا ۰/۴۳۲۱= را به ثبت برساند.

جدول ۷

مدل	جذر میانگین مربعات خطا	میانگین درصد قدرمطلق خطا	میانگین قدرمطلق خطا
Linear regression	۰/۱۱۴۲	۰/۲۰۹۰	۰/۱۰۴۷
SVR	۰/۱۸۶۹	۰/۲۹۶۲	۰/۱۴۸۲
جنگل تصادفی	۰/۱۴۴۴	۰/۲۷۸۰	۰/۱۳۹۴
درخت تصمیم	۰/۴۳۲۱	۰/۶۸۸۱	۰/۳۵۶۹

همانند داده‌های الکترونیک بر روی داده‌های شیمی نیز بهترین و بدترین نتیجه توسط رویکردهای ال آر و درخت تصمیم به‌دست آمد. در این داده‌ها مقادیر میانگین قدرمطلق خطا تمامی رویکردها بسیار نزدیک به یک دیگر می‌باشند. رویکرد ال آر توانست در این مجموعه داده‌ها به میانگین قدرمطلق خطا ۰/۰۱۴۴ دست یابد. از طرفی رویکرد درخت تصمیم در این داده‌ها به میانگین قدرمطلق خطا ۰/۰۴۷۷ دست‌یافت.

جدول ۸

مدل	جذر میانگین مربعات خطا	میانگین درصد قدرمطلق خطا	میانگین قدرمطلق خطا
Linear regression	۰/۰۱۷۵	۰/۰۲۸	۰/۰۱۴۴
SVR	۰/۰۲۸۲	۰/۰۴۸	۰/۰۲۴۸
جنگل تصادفی	۰/۰۲۵۹	۰/۰۴۵	۰/۰۲۳۲
درخت تصمیم	۰/۰۵۲۰	۰/۱۳۸۱	۰/۰۴۷۷

بحث و نتیجه‌گیری

پیش بینی عملکرد دانش‌آموزان برای هر مؤسسه آموزشی مهم است. این امر به‌ویژه برای کسانی که قصد دارند به دانش‌آموزان فرصت‌هایی برای انجام کاری مفید در رشته تحصیلی خود بدهند و کسانی که قصد دارند منابع آموزشی مورد نیاز برای تجربیات یادگیری عالی را به‌خوبی مدیریت کنند مهم است. دانستن عملکرد دانش‌آموزان در هر دوره از قبل یک نیاز اصلی برای کمک به دانش‌آموزان در معرض خطر با کاهش چالش‌هایی است که آنها در سفرهای یادگیری با آن‌ها مواجه هستند و کمک به برتری آنها در فرآیند یادگیری می‌کند. در حالی که، چنین پیش‌بینی‌هایی، به‌ویژه، برای یک دانشگاه جدید یک چالش است، زیرا هیچ رکورد داده کافی برای تجزیه و تحلیل وجود ندارد. طبقه‌بندی‌کننده جنگل تصادفی همانی بود که کارایی خود را (در بین بقیه طبقه‌بندی‌کننده‌ها) در پیش‌بینی عملکرد دانش‌آموزان در تمام نمرات دروس، از جمله ثابت کرد. دلیل اصلی که ممکن است به موفقیت آن طبقه‌بندی‌کننده نسبت داده شود، روش آموزش مدلی است که از آن استفاده می‌کند، که برای ساخت مدل طبقه‌بندی آن تنها به چند نقطه داده یا نمونه متکی است.

مشکل اساسی در پیش بینی مدل‌ها عدم وجود داده‌های زیاد می‌باشد. رویکردهای تولید توزیع داده‌ها مشابه با داده‌های منبع یکی از کارهایی است که می‌توان به‌عنوان پیشنهادات برای کارهای آینده انجام داد.

موازین اخلاقی

در این مطالعه اصول اخلاق در پژوهش شامل اخذ رضایت آگاهانه از شرکت کنندگان و حفظ اطلاعات محرمانه آنها رعایت گردیده است.

تشکر و قدردانی

پژوهشگران مراتب قدردانی و تشکر خود را از کلیه شرکت‌کنندگان این پژوهش که با استقبال و بردباری، در روند استخراج نتایج همکاری نمودند، اعلام می‌دارند.

تعارض منافع

نویسندگان این مطالعه هیچ گونه تعارض منافعی در انجام و نگارش آن ندارند.

فهرست منابع

[1] Nieto Y, Gacía-Díaz V. Montenegro CC. González C, Crespo RG. "Usage of machine learning for strategic decision making at higher educational institutions, " IEEE Access, 2019.

[2] Alsalman YS, Halemah NKA, AlNagi ES, Salameh W. "Using and Decision Tree Artificial Neural Network to Predict Students Academic Performance, " in 2019 10th International Conference on Information and Communication Systems (ICICS), 2019; 104-109: IEEE.

[3] Helal S. "Predicting academic performance by considering student heterogeneity, " Knowledge-Based Systems, 2018; 161: 134-146.

[4] Daud A, Aljohani NR, Abbasi RA, Lytras MD, Abbas F, Alowibdi JS. "Predicting student performance using advanced learning analytics, " in Proceedings of the 26th international conference on world wide web companion, 2017; 415-421: International World Wide Web Conferences Steering Committee.

[5] Ibrahim Z, Rusli D. "Predicting students' academic performance: comparing artificial neural network and linear regression, " in 21st Annual SAS Malaysia Forum, 5th September, 2007.

[6] LIU Ty, Xiu L. "Finding out reasons for low completion in MOOC environment: an explicable approach using hybrid data mining methods," DEStech Transactions on Social Science, Education and Human Science, no. meit, 2017.

[7] Wang W, Yu H, Miao C. "Deep model for dropout prediction in MOOCs, " in Proceedings of the 2nd International Conference on Crowd Science and Engineering, 2017; 26-32: ACM.

[8] Al-Shabandar R, Hussain A, Laws A, Keight R, Lunn J, Radi N. "Machine learning approaches to predict learning outcomes in Massive open online courses, " in 2017 International Joint Conference on Neural Networks (IJCNN), 2017; 713-720: IEEE.

[9] Al-Shabandar R, Hussain A, Laws A, Keight R, Lunn J. "Towards the differentiation of initial and final retention in massive open online courses, " in International Conference

on Intelligent Computing, 2017; 26-36: Springer.

[10] Whitehill J, Mohan K, Seaton D, Rosen Y, Tingley D. "Delving deeper into MOOC student dropout prediction, " arXiv preprint arXiv:1702.06404, 2017.

[11] Nagrecha S, Dillon JZ, Chawla NV. "MOOC dropout prediction: lessons learned from making pipelines interpretable, " in Proceedings of the 26th International Conference on World Wide Web Companion, 2017; 351-359: International World Wide Web Conferences Steering Committee.

[12] Xing W, Chen X, Stein J, Marcinkowski M. "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization, " Computers in human behavior, 2016; 58: 119-129.

[13] Robinson C, Yeomans M, Reich J, Hulleman C, Gehlbach H. "Forecasting student achievement in MOOCs with natural language processing, " in Proceedings of the sixth international conference on learning analytics & knowledge, 2016; 383-387: ACM.

[14] Liang J, Li C, Zheng L. "Machine learning application in MOOCs: Dropout prediction, " in 2016 11th International Conference on Computer Science & Education (ICCSE), 2016; 52-57: IEEE.

[15] Crossley S, Paquette L, Dascalu M, McNamara DS, Baker RS, "Combining click-stream data with NLP tools to better understand MOOC completion, " in Proceedings of the sixth international conference on learning analytics & knowledge, 2016; 6-14: ACM.

[16] Whitehill J, Williams J, Lopez G, Coleman C, Reich J. "Beyond prediction: First steps toward automatic intervention in MOOC student stopout, " Available at SSRN 2611750, 2015.

[17] Boyer S, Veeramachaneni K. "Transfer learning for predictive models in massive open online courses, " in International conference on artificial intelligence in education, 2015; 54-63: Springer.

[18] Chaplot DS, Rhim E, Kim J. "Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks, " in AIED Workshops, 2015; (53): 54-57.

[19] Coleman CA, Seaton DT, Chuang I. "Probabilistic use cases: Discovering behavioral patterns for predicting certification, " in Proceedings of the Second (2015) ACM Conference on Learning@ Scale, 2015; 141-148: ACM.

[20] Kizilcec RF, Halawa S. "Attrition and achievement gaps in online learning, " in Proceedings of the second (2015) ACM conference on learning@ scale, 2015; 57-66: ACM.

[21] Fei M, Yeung DY. "Temporal models for predicting student dropout in massive open online courses, " in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015; 256-263: IEEE.

[22] Kloft M, Stiehler F, Zheng Z, Pinkwart N. "Predicting MOOC dropout over weeks using machine learning methods, " in Proceedings of the EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs, 2014, pp. 60-65.

[23] Murphy KP. Machine learning: a probabilistic perspective. MIT press, 2012.