پیش‌بینی عملکرد تحصیلی دانش‌آموزان با استفاده از تکنیک‌های ترکیبی داده‌کاوی: رویکردی جامع با Clustering، Classification و Association Rules

پروژه داده‌کاوی – سالار لطفی , نازنین خاکسبز

تاریخ: شهریور ۱۴۰۴

چکیده

پیش‌بینی عملکرد تحصیلی دانش‌آموزان یکی از چالش‌های مهم در حوزه آموزش است که می‌تواند به شناسایی زودهنگام دانش‌آموزان در معرض خطر افت تحصیلی کمک کند. این پژوهش با هدف توسعه یک سیستم جامع پیش‌بینی عملکرد تحصیلی، از ترکیب سه رویکرد اصلی داده‌کاوی شامل طبقه‌بندی (Classification)، خوشه‌بندی (Clustering) و قوانین همبستگی (Association Rules) استفاده می‌کند. داده‌های مورد استفاده شامل ۱۰۰۰ رکورد از دانش‌آموزان با ۱۷ ویژگی مختلف است. در فاز طبقه‌بندی، ۹ الگوریتم مختلف پیاده‌سازی شد که مدل شبکه عصبی با F۱–Score برابر ۰.۵۲ بهترین عملکرد را داشت. در فاز خوشه‌بندی، الگوریتم GMM با Silhouette Score برابر ۰.۴۱۶۱ بهترین نتیجه را ارائه داد. همچنین ۱۳,۸۸۵ قانون همبستگی استخراج شد که ۴۳۱ قانون قوی با Confidence بالای ۰.۷ شناسایی شدند. نتایج نشان می‌دهد که عوامل کلیدی موثر بر عملکرد تحصیلی شامل نرخ حضور، ساعات مطالعه، عملکرد قبلی و حمایت والدین هستند.

کلمات کلیدی: داده‌کاوی آموزشی، پیش‌بینی عملکرد دانش‌آموزان، طبقه‌بندی، خوشه‌بندی، قوانین همبستگی، یادگیری ماشین

۱. مقدمه

۱.۱ بیان مسئله

سیستم‌های آموزشی در سراسر جهان با چالش شناسایی و پشتیبانی از دانش‌آموزانی که در معرض خطر افت تحصیلی قرار دارند، مواجه هستند. پیش‌بینی زودهنگام عملکرد تحصیلی می‌تواند به مربیان و مدیران آموزشی کمک کند تا مداخلات هدفمند و به‌موقع را طراحی و اجرا کنند. با توجه به حجم عظیم داده‌های آموزشی موجود، استفاده از تکنیک‌های داده‌کاوی می‌تواند الگوهای پنهان و عوامل موثر بر موفقیت یا شکست تحصیلی را آشکار سازد.

۱.۲ اهمیت موضوع

اهمیت این پژوهش از چند منظر قابل بررسی است. از نظر علمی، این مطالعه رویکردی ترکیبی و جامع برای تحلیل داده‌های آموزشی ارائه می‌دهد که می‌تواند به توسعه دانش در حوزه Educational Data Mining کمک کند. از منظر عملی، نتایج این پژوهش می‌تواند به بهبود کیفیت آموزش، کاهش نرخ افت تحصیلی، و بهینه‌سازی تخصیص منابع آموزشی منجر شود. همچنین، شناسایی عوامل کلیدی موثر بر عملکرد تحصیلی می‌تواند در طراحی برنامه‌های آموزشی شخصی‌سازی شده مورد استفاده قرار گیرد.

۱.۳ اهداف و سوالات تحقیق

هدف اصلی این پژوهش، توسعه یک سیستم جامع و دقیق برای پیش‌بینی عملکرد تحصیلی دانش‌آموزان با استفاده از تکنیک‌های مختلف داده‌کاوی است. سوالات اصلی تحقیق عبارتند از:

- کدام الگوریتم طبقه‌بندی بهترین عملکرد را در پیش‌بینی وضعیت قبولی/رد دانش‌آموزان دارد؟
- آیا می‌توان دانش‌آموزان را بر اساس ویژگی‌های تحصیلی و رفتاری در گروه‌های همگن دسته‌بندی کرد؟
- چه قوانین همبستگی قوی بین ویژگی‌های مختلف و عملکرد تحصیلی وجود دارد؟
- کدام عوامل بیشترین تاثیر را بر موفقیت یا شکست تحصیلی دانش‌آموزان دارند؟

۱.۴ ساختار گزارش

این گزارش در ۱۱ بخش سازماندهی شده است. پس از مقدمه، بخش ۲ به مرور ادبیات موضوع می‌پردازد. بخش ۳ داده‌ها و ویژگی‌های آنها را توصیف می‌کند. بخش ۴ روش‌ها و الگوریتم‌های استفاده شده را تشریح می‌کند. بخش‌های ۵ تا ۷ به ترتیب نتایج، تحلیل و بحث را ارائه می‌دهند. بخش ۸ اعتبارسنجی آماری نتایج را بررسی می‌کند و در نهایت، بخش ۹ نتیجه‌گیری و پیشنهادات را ارائه می‌دهد.

۲. مرور ادبیات

۲.۱ مطالعات پیشین

در سال‌های اخیر، مطالعات متعددی در زمینه پیش‌بینی عملکرد تحصیلی با استفاده از تکنیک‌های داده‌کاوی انجام شده است. Romero و Ventura (۲۰۲۰) در مطالعه جامع خود، بیش از ۳۰۰ مقاله در حوزه Educational Data Mining را مرور کردند و نشان دادند که الگوریتم‌های طبقه‌بندی، به‌ویژه Decision Trees و Neural Networks، پرکاربردترین روش‌ها در این حوزه هستند. Kumar و Pal (۲۰۲۱) با استفاده از الگوریتم Naive Bayes توانستند با دقت ۷۸٪ عملکرد دانشجویان را پیش‌بینی کنند.

۲.۲ مفاهیم اصلی

۲.۲.۱ الگوریتم‌های طبقه‌بندی

طبقه‌بندی یکی از تکنیک‌های اصلی یادگیری با نظارت است که هدف آن پیش‌بینی برچسب کلاس برای نمونه‌های جدید بر اساس الگوهای آموخته شده از داده‌های آموزشی است. در این پژوهش، از الگوریتم‌های مختلفی شامل:

- **شبکه‌های عصبی مصنوعی (Neural Networks):** مدل‌هایی الهام‌گرفته از مغز انسان که قابلیت یادگیری الگوهای پیچیده غیرخطی را دارند
- **ماشین بردار پشتیبان (SVM):** الگوریتمی که با یافتن ابرصفحه بهینه، داده‌ها را در فضای چندبعدی جدا می‌کند
- **جنگل تصادفی (Random Forest):** روش ensemble که از ترکیب چندین درخت تصمیم برای بهبود دقت استفاده می‌کند
- **رگرسیون لجستیک:** مدل آماری برای پیش‌بینی احتمال وقوع یک رویداد دودویی

۲.۲.۲ الگوریتم‌های خوشه‌بندی

خوشه‌بندی یک تکنیک یادگیری بدون نظارت است که هدف آن گروه‌بندی داده‌ها بر اساس شباهت‌های ذاتی است. الگوریتم‌های استفاده شده شامل:

- **K-Means**: الگوریتمی که داده‌ها را به K خوشه تقسیم می‌کند با هدف حداقل‌سازی واریانس درون‌خوشه‌ای
- **DBSCAN**: روشی مبتنی بر چگالی که قابلیت شناسایی خوشه‌های با شکل دلخواه را دارد
- **GMM (Gaussian Mixture Model)**: مدل احتمالاتی که فرض می‌کند داده‌ها از ترکیب چند توزیع گاوسی تولید شده‌اند
- **Hierarchical Clustering**: روشی که ساختار سلسله‌مراتبی خوشه‌ها را ایجاد می‌کند

۲.۲.۳ قوانین همبستگی

کشف قوانین همبستگی به دنبال یافتن روابط جالب بین متغیرها در پایگاه‌های داده بزرگ است. دو الگوریتم اصلی استفاده شده عبارتند از:

- **Apriori**: الگوریتم کلاسیک که با استفاده از رویکرد bottom-up، مجموعه‌های پرتکرار را شناسایی می‌کند
- **FP-Growth**: الگوریتم بهینه‌تر که با ساخت درخت FP، نیاز به اسکن‌های متعدد پایگاه داده را کاهش می‌دهد

۳. داده‌ها و ویژگی‌های آنها

۳.۱ منبع و توصیف داده‌ها

مجموعه داده مورد استفاده در این پژوهش شامل اطلاعات ۱۰۰۰ دانش‌آموز با ۱۷ ویژگی مختلف است. این داده‌ها از فایل CSV با نام "student_performance_updated_۱۰۰۰.csv" بارگذاری شده‌اند. داده‌ها شامل ترکیبی از ویژگی‌های عددی و دسته‌ای هستند که جنبه‌های مختلف عملکرد تحصیلی و ویژگی‌های دموگرافیک دانش‌آموزان را پوشش می‌دهند.

۳.۲ ویژگی‌های داده‌ها

نام ویژگی	نوع	توضیحات	محدوده/مقادیر
StudentID	عددی	شناسه یکتای دانش‌آموز	۱–۱۰۰۰

نام ویژگی	نوع	توضیحات	محدوده/مقادیر
Gender	دسته‌ای	جنسیت دانش‌آموز	Male/Female
AttendanceRate	عددی	نرخ حضور در کلاس	۰–۱۰۰%
StudyHoursPerWeek	عددی	ساعات مطالعه هفتگی	۰–۴۰ ساعت
PreviousGrade	عددی	نمره دوره قبلی	۰–۱۰۰
ExtracurricularActivities	عددی	میزان فعالیت‌های فوق‌برنامه	۰–۱۰
ParentalSupport	دسته‌ای	سطح حمایت والدین	Low/Medium/High
FinalGrade	عددی	نمره نهایی	۰–۱۰۰
Pass_Status	دسته‌ای (هدف)	وضعیت قبولی/رد	Pass/Fail

۳.۳ پیش‌پردازش داده‌ها

فرآیند پیش‌پردازش داده‌ها شامل مراحل زیر بود:

۱. **پردازش مقادیر گمشده:** مقادیر گمشده در ویژگی‌های عددی با میانگین و در ویژگی‌های دسته‌ای با مد (پرتکرارترین مقدار) جایگزین شدند

۲. **ایجاد متغیر هدف:** بر اساس میانه نمرات نهایی، دانش‌آموزان به دو گروه قبول و رد تقسیم شدند

۳. **کدگذاری متغیرهای دسته‌ای:** با استفاده از LabelEncoder، متغیرهای دسته‌ای به مقادیر عددی تبدیل شدند

۴. **نرمال‌سازی:** ویژگی‌های عددی با استفاده از StandardScaler نرمال‌سازی شدند تا میانگین صفر و انحراف معیار یک داشته باشند

۵. **تقسیم داده‌ها:** داده‌ها با نسبت ۸۰–۲۰ به مجموعه‌های آموزشی (۸۰۰ نمونه) و تست (۲۰۰ نمونه) تقسیم شدند

> **نتیجه پیش‌پردازش:** پس از اتمام پیش‌پردازش، تمامی مقادیر گمشده پردازش شدند و توزیع متوازنی از کلاس‌های قبول (۵۰.۱%) و رد (۴۹.۹%) حاصل شد.

۴. روش‌ها و الگوریتم‌ها

۴.۱ فاز طبقه‌بندی (Classification)

در این فاز، ۹ الگوریتم مختلف طبقه‌بندی پیاده‌سازی و مقایسه شدند:

۴.۱.۱ شبکه عصبی مصنوعی (MLP Classifier)

شبکه عصبی پیاده‌سازی شده از نوع Multilayer Perceptron با پارامترهای زیر بود:

- تعداد لایه‌های پنهان: ۲ لایه با ۱۰۰ نورون
- تابع فعال‌سازی: ReLU
- الگوریتم بهینه‌سازی: Adam
- نرخ یادگیری: ۰.۰۰۱
- حداکثر تکرار: ۱۰۰۰

۴.۱.۲ ماشین بردار پشتیبان (SVM)

دو نوع SVM پیاده‌سازی شد:

- **SVM خطی:** با 'kernel='linear برای جداسازی خطی داده‌ها
- **SVM با کرنل RBF:** برای مدل‌سازی روابط غیرخطی پیچیده

۴.۱.۳ روش‌های Ensemble

- **Random Forest:** با ۱۰۰ درخت تصمیم و عمق نامحدود

- **Gradient Boosting:** با ۱۰۰ estimator و نرخ یادگیری ۰.۱

۴.۲ فاز خوشه‌بندی (Clustering)

۴.۲.۱ تعیین تعداد بهینه خوشه‌ها

برای تعیین تعداد بهینه خوشه‌ها از دو روش استفاده شد:

۱. **روش Elbow:** با بررسی WCSS (Within-Cluster Sum of Squares) برای K از ۲ تا ۱۰

۲. **Silhouette Score:** برای ارزیابی کیفیت خوشه‌بندی

شکل ۱: نمودار Elbow و Silhouette Score

بر اساس تحلیل‌ها، تعداد بهینه خوشه‌ها برابر با ۲ تعیین شد

۴.۲.۲ الگوریتم‌های خوشه‌بندی

الگوریتم	پارامترهای کلیدی	ویژگی‌های خاص
K-Means	K=۲, n_init=۱۰	سرعت بالا، خوشه‌های کروی
DBSCAN	eps=۱.۹, min_samples=۵	شناسایی نقاط نویز، خوشه‌های با شکل دلخواه
GMM	n_components=۲	مدل احتمالاتی، انعطاف‌پذیری بالا
Hierarchical	'linkage='ward	ساختار سلسله‌مراتبی، dendrogram

۴.۳ فاز قوانین همبستگی (Association Rules)

۴.۳.۱ آماده‌سازی داده‌ها

برای استخراج قوانین همبستگی، ابتدا ویژگی‌های پیوسته به دسته‌ای تبدیل شدند:

- **AttendanceRate**: به سه دسته Low، Medium و High تقسیم شد
- **StudyHoursPerWeek**: به سه سطح Low_Study، Medium_Study و High_Study
- **PreviousGrade**: به Poor، Average و Good تبدیل شد
- **ExtracurricularActivities**: به Low، Medium و High Activities

۴.۳.۲ پارامترهای الگوریتم‌ها

- **حداقل Support**: ۰.۰۱ (حداقل ۱% از تراکنش‌ها)
- **حداقل Confidence**: ۰.۵ (۵۰% اطمینان)
- **حداکثر طول قانون**: ۵ آیتم

۴.۴ معیارهای ارزیابی

۴.۴.۱ معیارهای طبقه‌بندی

- **Accuracy**: نسبت پیش‌بینی‌های صحیح به کل پیش‌بینی‌ها
- **Precision**: نسبت پیش‌بینی‌های مثبت صحیح به کل پیش‌بینی‌های مثبت
- **Recall**: نسبت پیش‌بینی‌های مثبت صحیح به کل موارد واقعاً مثبت
- **F۱-Score**: میانگین هارمونیک Precision و Recall
- **AUC-ROC**: سطح زیر منحنی ROC

$$F_1\text{-}Score = ۲ \times (Precision \times Recall) / (Precision + Recall)$$

۴.۴.۲ معیارهای خوشه‌بندی

- **Silhouette Score**: معیاری بین -۱ تا ۱ که کیفیت خوشه‌بندی را نشان می‌دهد
- **Davies-Bouldin Index**: نسبت شباهت درون‌خوشه‌ای به بین‌خوشه‌ای (کمتر بهتر)
- **Calinski-Harabasz Score**: نسبت پراکندگی بین‌خوشه‌ای به درون‌خوشه‌ای (بیشتر بهتر)

۴.۴.۳ معیارهای قوانین همبستگی

- **Support**: فراوانی نسبی آیتم‌ست در کل تراکنش‌ها
- **Confidence**: احتمال وقوع consequent به شرط وقوع antecedent
- **Lift**: نسبت Confidence واقعی به Confidence مورد انتظار در حالت استقلال

۵. نتایج

۵.۱ نتایج طبقه‌بندی

مدل	Accuracy	Precision	Recall	F_1-Score	AUC	CV Mean ± Std
Neural Network	**۰.۵۲۰**	**۰.۵۲۰**	**۰.۵۲۰**	**۰.۵۲۰**	**۰.۵۱۹**	**± ۰.۵۱۳ ۰.۰۱۷**
SVM (RBF)	۰.۵۲۰	۰.۵۲۰	۰.۵۲۰	۰.۵۲۰	۰.۵۴۵	± ۰.۵۱۶ ۰.۰۴۷
Random Forest	۰.۵۱۰	۰.۵۱۰	۰.۵۱۰	۰.۵۱۰	۰.۴۸۸	± ۰.۵۳۳ ۰.۰۱۵
Logistic Regression	۰.۴۹۰	۰.۴۹۰	۰.۴۹۰	۰.۴۸۹	۰.۵۰۸	± ۰.۵۰۹ ۰.۰۴۲
Gradient Boosting	۰.۴۸۰	۰.۴۸۰	۰.۴۸۰	۰.۴۷۹	۰.۴۷۸	± ۰.۵۲۶ ۰.۰۰۹
SVM (Linear)	۰.۴۶۵	۰.۴۶۵	۰.۴۶۵	۰.۴۶۴	۰.۴۷۴	± ۰.۵۱۹ ۰.۰۲۷

مدل	Accuracy	Precision	Recall	F۱–Score	AUC	CV Mean ± Std
Decision Tree	۰.۴۶۰	۰.۴۶۰	۰.۴۶۰	۰.۴۶۰	۰.۴۶۰	± ۰.۴۸۵ ۰.۰۳۶
K–Nearest Neighbors	۰.۴۶۰	۰.۴۶۰	۰.۴۶۰	۰.۴۵۹	۰.۴۵۵	± ۰.۵۱۹ ۰.۰۱۶
Naive Bayes	۰.۴۶۵	۰.۴۶۳	۰.۴۶۵	۰.۴۵۸	۰.۴۷۰	± ۰.۴۸۳ ۰.۰۳۰

نتیجه کلیدی: مدل شبکه عصبی با F۱–Score برابر ۰.۵۲۰ بهترین عملکرد را داشت، اگرچه این دقت نشان‌دهنده نیاز به بهبود بیشتر مدل‌ها است.

۵.۲ نتایج خوشه‌بندی

الگوریتم	Silhouette Score	Davies–Bouldin	Calinski–Harabasz	تعداد خوشه‌ها
GMM	**۰.۴۱۶۱**	**۱.۴۹۹**	۵۶.۴۳	۲
K–Means	۰.۱۳۹۷	۲.۴۹۸	**۱۴۵.۱۲**	۲
DBSCAN	۰.۱۳۴۲	۲.۶۴۳	۸۷.۴۹	۲ (+ ۳۷ نویز)
Hierarchical	۰.۱۲۲۹	۲.۷۰۴	۱۱۹.۴۱	۲

تحلیل خوشه‌ها

خوشه	تعداد دانش‌آموزان	درصد	نرخ قبولی	ویژگی‌های برجسته
خوشه ۱	۴۵۱	۴۵.۱%	۵۲.۳%	حضور متوسط، مطالعه متوسط، نمره قبلی بالاتر
خوشه ۲	۵۴۹	۵۴.۹%	۴۸.۳%	حضور متوسط، مطالعه متوسط، نمره قبلی پایین‌تر

نکته مهم: Adjusted Rand Index (۰.۰۰۰۰۶) و Adjusted Mutual Information (۰.۰۰۰۰۵) پایین نشان می‌دهند که خوشه‌های شناسایی شده ارتباط ضعیفی با برچسب‌های واقعی قبول/رد دارند.

۵.۳ نتایج قوانین همبستگی

خلاصه نتایج:

- تعداد کل قوانین استخراج شده: ۱۳,۸۸۵
- قوانین قوی (Confidence ≥ ۰.۷, Lift > ۱.۲): ۴۳۱
- قوانین منتهی به قبولی: ۱,۶۵۷
- قوانین منتهی به رد: ۱,۶۳۳
- بالاترین Lift: ۴.۷۱۷

قوی‌ترین قوانین کشف شده

قانون با بالاترین Lift (۴.۷۱۷):

IF: Attendance_Level=Low AND Previous_Performance=Poor AND Gender=Female

THEN: Activities_Level=Low AND Cluster_۱ AND Pass_Result=Fail

Support: ۰.۰۱۱ | **Confidence:** ۰.۵۰۰ | **Lift:** ۴.۷۱۷

عوامل کلیدی موثر بر قبولی

- **عوامل مثبت:** حضور بالا، ساعات مطالعه زیاد، عملکرد قبلی خوب، حمایت بالای والدین
- **عوامل منفی:** حضور پایین، مطالعه کم، عملکرد قبلی ضعیف، حمایت پایین والدین

۶. تحلیل و بحث

۶.۱ تحلیل عملکرد مدل‌های طبقه‌بندی

نتایج نشان می‌دهد که مدل‌های طبقه‌بندی عملکرد متوسطی دارند با حداکثر F۱–Score برابر ۰.۵۲. این عملکرد نسبتاً پایین می‌تواند به دلایل زیر باشد:

- **پیچیدگی ذاتی مسئله:** عملکرد تحصیلی تحت تأثیر عوامل متعدد و پیچیده‌ای است که ممکن است همه آنها در داده‌ها موجود نباشند
- **کیفیت داده‌ها:** احتمال وجود نویز در داده‌ها یا عدم کفایت ویژگی‌های موجود
- **توزیع متوازن کلاس‌ها:** با وجود توزیع ۵۰–۵۰، ممکن است الگوهای تمایز واضحی بین دو کلاس وجود نداشته باشد
- **روابط غیرخطی پیچیده:** حتی شبکه عصبی که قابلیت مدل‌سازی روابط پیچیده را دارد، نتوانست دقت بالایی کسب کند

۶.۲ تحلیل نتایج خوشه‌بندی

الگوریتم GMM با Silhouette Score برابر ۰.۴۱۶۱ بهترین عملکرد را داشت که نشان‌دهنده کیفیت متوسط خوشه‌بندی است. نکات کلیدی:

- تعداد بهینه خوشه‌ها (۲) ممکن است برای تمایز کامل دانش‌آموزان کافی نباشد

- عدم ارتباط قوی خوشه‌ها با برچسب‌های قبول/رد (ARI=۰.۰۰۰۶) نشان می‌دهد که الگوهای طبیعی در داده‌ها لزوماً با وضعیت قبولی/رد منطبق نیستند
- GMM به دلیل انعطاف‌پذیری در شکل خوشه‌ها، عملکرد بهتری نسبت به K-Means داشت

۶.۳ تحلیل قوانین همبستگی

استخراج ۱۳,۸۸۵ قانون نشان‌دهنده روابط پیچیده بین ویژگی‌ها است. تحلیل قوانین قوی نشان می‌دهد:

- **اهمیت حضور در کلاس:** حضور پایین قوی‌ترین پیش‌بینی‌کننده شکست تحصیلی است
- **تأثیر عملکرد قبلی:** عملکرد ضعیف قبلی با احتمال بالای رد در آینده همراه است
- **نقش حمایت والدین:** حمایت پایین والدین در ترکیب با سایر عوامل منفی، احتمال رد را افزایش می‌دهد
- **تفاوت‌های جنسیتی:** برخی قوانین نشان می‌دهند که الگوهای موفقیت/شکست در دختران و پسران متفاوت است

۶.۴ محدودیت‌های پژوهش

۱. **محدودیت داده‌ها:**
   - حجم نسبتاً کم داده‌ها (۱۰۰۰ نمونه)
   - احتمال عدم پوشش همه عوامل موثر (مثل وضعیت اقتصادی، سلامت روانی)

۲. **محدودیت‌های روش‌شناسی:**
   - استفاده از میانه برای تعیین آستانه قبولی ممکن است واقع‌بینانه نباشد
   - عدم استفاده از روش‌های Deep Learning پیشرفته‌تر

۳. **محدودیت‌های تعمیم‌پذیری:**
   - نتایج ممکن است به سایر محیط‌های آموزشی قابل تعمیم نباشد

۷. اعتبارسنجی آماری نتایج

۷.۱ آزمون Friedman برای مقایسه مدل‌ها

برای بررسی معناداری تفاوت بین عملکرد مدل‌های طبقه‌بندی، از آزمون Friedman استفاده شد:

- **Friedman Statistic:** ۳۰.۹۸۷۳
- **P-value:** ۰.۰۰۰۰۱۴۱
- **نتیجه:** با ۰.۰۵ > p-value، تفاوت معناداری بین عملکرد مدل‌ها وجود دارد ✓

۷.۲ فاصله اطمینان ۹۵% برای معیارهای ارزیابی

معیار	میانگین	فاصله اطمینان ۹۵%
Accuracy	۰.۴۸۵۶	[۰.۵۰۴۰ ,۰.۴۶۷۱]
Precision	۰.۴۸۵۲	[۰.۵۰۳۹ ,۰.۴۶۶۶]
Recall	۰.۴۸۵۶	[۰.۵۰۴۰ ,۰.۴۶۷۱]
F۱-Score	۰.۴۸۴۲	[۰.۵۰۳۳ ,۰.۴۶۵۰]

۷.۳ رتبه‌بندی نهایی روش‌ها

بر اساس امتیاز ترکیبی محاسبه شده:

رتبه	روش	امتیاز ترکیبی	وضعیت
۱ 🥇	Clustering	۰.۵۳۱۵	خوب
۲ 🥈	Classification	۰.۴۸۵۹	متوسط
۳ 🥉	Association Rules	۰.۴۱۸۶	متوسط

۸. پیشنهادات برای بهبود

۸.۱ بهبود مدل‌های طبقه‌بندی

۱. **استفاده از روش‌های Ensemble Learning:** پیاده‌سازی Stacking یا Voting Classifier برای ترکیب نقاط قوت مدل‌های مختلف

۲. **بهینه‌سازی عمیق‌تر Hyperparameters:** استفاده از Bayesian Optimization یا Grid Search گسترده‌تر

۳. **مهندسی ویژگی:** ایجاد ویژگی‌های ترکیبی جدید و استفاده از تکنیک‌های انتخاب ویژگی

۴. **تنظیم Threshold:** بهینه‌سازی آستانه تصمیم‌گیری برای متوازن کردن Precision و Recall

۸.۲ بهبود خوشه‌بندی

۱. **کاهش ابعاد موثرتر:** استفاده از t-SNE یا UMAP به جای PCA

۲. **آزمایش معیارهای فاصله مختلف:** Manhattan Distance یا Cosine Similarity

۳. **خوشه‌بندی سلسله‌مراتبی:** بررسی تعداد خوشه‌های بیشتر برای تمایز بهتر

۴. **Semi-supervised Clustering:** استفاده از برچسب‌های موجود برای هدایت خوشه‌بندی

۸.۳ بهبود قوانین همبستگی

۱. **تنظیم پارامترها:** کاهش حد آستانه Support برای کشف قوانین نادر اما مهم

۲. **Pruning قوانین:** حذف قوانین redundant و نگهداری قوانین با بیشترین ارزش اطلاعاتی

۳. **قوانین چندسطحی:** بررسی قوانین با طول بیشتر برای کشف روابط پیچیده‌تر

۹. نتیجه‌گیری

۹.۱ خلاصه نتایج

این پژوهش با هدف توسعه یک سیستم جامع پیش‌بینی عملکرد تحصیلی دانش‌آموزان، از ترکیب سه رویکرد اصلی داده‌کاوی استفاده کرد. نتایج کلیدی عبارتند از:

- مدل شبکه عصبی با F۱–Score برابر ۰.۵۲ بهترین عملکرد را در طبقه‌بندی داشت
- الگوریتم GMM با Silhouette Score برابر ۰.۴۱۶۱ بهترین نتیجه خوشه‌بندی را ارائه داد
- ۱۳,۸۸۵ قانون همبستگی استخراج شد که ۴۳۱ قانون قوی شناسایی شدند
- عوامل کلیدی موثر بر عملکرد تحصیلی شامل حضور در کلاس، ساعات مطالعه، عملکرد قبلی و حمایت والدین هستند

۹.۲ دستاوردهای علمی

این پژوهش نشان داد که:

۱. پیش‌بینی عملکرد تحصیلی یک مسئله پیچیده است که نیازمند رویکردهای چندگانه است

۲. ترکیب روش‌های مختلف داده‌کاوی می‌تواند دیدگاه جامع‌تری از عوامل موثر ارائه دهد

۳. قوانین همبستگی می‌توانند بینش‌های قابل تفسیری برای مربیان فراهم کنند

۴. خوشه‌بندی می‌تواند در شناسایی گروه‌های همگن دانش‌آموزان برای برنامه‌ریزی آموزشی هدفمند مفید باشد

۹.۳ کاربردهای عملی

نتایج این پژوهش می‌تواند کاربردهای عملی زیر را داشته باشد:

- **سیستم هشدار زودهنگام:** شناسایی دانش‌آموزان در معرض خطر در ابتدای ترم تحصیلی
- **برنامه‌ریزی آموزشی شخصی‌سازی شده:** طراحی برنامه‌های حمایتی متناسب با نیاز هر گروه
- **تخصیص بهینه منابع:** اولویت‌بندی تخصیص منابع آموزشی به دانش‌آموزان نیازمند
- **مشاوره تحصیلی هدفمند:** ارائه توصیه‌های مبتنی بر داده به دانش‌آموزان و والدین

۹.۴ پیشنهادات برای تحقیقات آینده

۱. **جمع‌آوری داده‌های بیشتر:** افزایش حجم نمونه و افزودن ویژگی‌های جدید مانند وضعیت اقتصادی–اجتماعی، سلامت روانی، و سبک یادگیری

۲. **استفاده از Deep Learning:** پیاده‌سازی معماری‌های پیشرفته مانند LSTM برای تحلیل داده‌های سری زمانی عملکرد تحصیلی

۳. **رویکرد Multi-Task Learning:** پیش‌بینی همزمان چندین جنبه از عملکرد تحصیلی

۴. **تحلیل علّی:** استفاده از روش‌های Causal Inference برای شناسایی روابط علت و معلولی

۵. **مطالعات طولی:** پیگیری عملکرد دانش‌آموزان در طول زمان برای درک بهتر روند تغییرات

نکته پایانی: با وجود عملکرد متوسط مدل‌ها (عملکرد کلی سیستم: ۰.۴۷۸۷)، این پژوهش پایه‌ای مناسب برای توسعه سیستم‌های پیشرفته‌تر پیش‌بینی عملکرد تحصیلی فراهم می‌کند. بهبود مستمر این سیستم‌ها می‌تواند نقش مهمی در ارتقای کیفیت آموزش و کاهش نرخ افت تحصیلی داشته باشد.

---

۱۰. منابع

[۱] Romero, C., & Ventura, S. (۲۰۲۰). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, ۱۰(۳), e۱۳۵۵.

[۲] Kumar, M., & Pal, R. (۲۰۲۱). Prediction of student's performance using machine learning algorithms. International Journal of Advanced Computer Science and Applications, ۱۲(۷), ۲۴۳-۲۵۲.

[۳] Breiman, L. (۲۰۰۱). Random forests. Machine learning, ۴۵(۱), ۵-۳۲.

[۴] Cortes, C., & Vapnik, V. (۱۹۹۵). Support-vector networks. Machine learning, ۲۰(۳), ۲۷۳-۲۹۷.

[۵] MacQueen, J. (۱۹۶۷). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. ۱, No. ۱۴, pp. ۲۸۱-۲۹۷).

[۶] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (۱۹۹۶). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. ۹۶, No. ۳۴, pp. ۲۲۶-۲۳۱).

[۷] McLachlan, G., & Peel, D. (۲۰۰۰). Finite mixture models. John Wiley & Sons.

[۸] Agrawal, R., & Srikant, R. (۱۹۹۴). Fast algorithms for mining association rules. In Proc. .۲۰th int. conf. very large data bases, VLDB (Vol. ۱۲۱۵, pp. ۴۸۷–۴۹۹)

[۹] Han, J., Pei, J., & Yin, Y. (۲۰۰۰). Mining frequent patterns without candidate generation. .ACM sigmod record, ۲۹(۲), ۱–۱۲

[۱۰] Friedman, M. (۱۹۳۷). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association, ۳۲(۲۰۰), .۶۷۵–۷۰۱

[۱۱] Rousseeuw, P. J. (۱۹۸۷). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, ۲۰, .۵۳–۶۵

[۱۲] Davies, D. L., & Bouldin, D. W. (۱۹۷۹). A cluster separation measure. IEEE .transactions on pattern analysis and machine intelligence, (۲), ۲۲۴–۲۲۷

[۱۳] Caliński, T., & Harabasz, J. (۱۹۷۴). A dendrite method for cluster analysis. .Communications in Statistics–theory and Methods, ۳(۱), ۱–۲۷

[۱۴] Hubert, L., & Arabie, P. (۱۹۸۵). Comparing partitions. Journal of classification, ۲(۱), .۱۹۳–۲۱۸

[۱۵] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (۲۰۱۱). Scikit–learn: Machine learning in Python. Journal of machine .learning research, ۱۲(Oct), ۲۸۲۵–۲۸۳۰

---

۱۱. پیوست‌ها

پیوست الف: نمونه کد پایتون برای پیش‌پردازش داده‌ها

```
# بارگذاری کتابخانه‌ها
import pandas as pd
```

```
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split

# بارگذاری داده‌ها
df = pd.read_csv('student_performance.csv')

# پردازش مقادیر گمشده
numerical_columns = df.select_dtypes(include=['float64', 'int64']).columns
for col in numerical_columns:
    df[col].fillna(df[col].mean(), inplace=True)

# کدگذاری متغیرهای دسته‌ای
le = LabelEncoder()
categorical_columns = df.select_dtypes(include=['object']).columns
for col in categorical_columns:
    df[col + '_Encoded'] = le.fit_transform(df[col])

# نرمال‌سازی
scaler = StandardScaler()
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])

# تقسیم داده‌ها
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

پیوست ب: جدول معیارهای ارزیابی تفصیلی

معیار	فرمول	محدوده	تفسیر
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	$[0, 1]$	نسبت پیش‌بینی‌های صحیح

معیار	فرمول	محدوده	تفسیر
Precision	$TP / (TP + FP)$	$[0, 1]$	دقت در پیش‌بینی‌های مثبت
Recall	$TP / (TP + FN)$	$[0, 1]$	پوشش موارد مثبت واقعی
F_1–Score	$2 \times (Precision \times Recall) / (Precision + Recall)$	$[0, 1]$	میانگین هارمونیک دقت و بازخوانی
Silhouette Score	$max(a, b) / (b - a)$	$[-1, 1]$	کیفیت خوشه‌بندی
Support	$freq(X,Y) / N$	$[0, 1]$	فراوانی نسبی آیتم‌ست
Confidence	$freq(X,Y) / freq(X)$	$[0, 1]$	احتمال شرطی
Lift	$Confidence(X \rightarrow Y) / Support(Y)$	$[0, \infty)$	قدرت قانون

پیوست ج: لیست فایل‌های خروجی پروژه

- **processed_student_data.csv**: داده‌های پردازش شده نهایی
- **train_data.csv**: مجموعه داده آموزشی
- **test_data.csv**: مجموعه داده تست
- **best_model.pkl**: بهترین مدل طبقه‌بندی ذخیره شده
- **clustering_results.pkl**: نتایج کامل خوشه‌بندی
- **association_rules_all.csv**: تمام قوانین همبستگی استخراج شده
- **model_comparison_charts.png**: نمودارهای مقایسه مدل‌ها
- **confusion_matrices_all.png**: ماتریس‌های درهم‌ریختگی

- **roc_curves.png**: منحنی‌های ROC
- **elbow_silhouette_analysis.png**: تحلیل Elbow و Silhouette
- **clustering_visualization.png**: نمایش خوشه‌ها
- **final_evaluation_report.txt**: گزارش نهایی ارزیابی

پایان مقاله

این مقاله بر اساس پروژه داده‌کاوی آموزشی با هدف پیش‌بینی عملکرد تحصیلی دانش‌آموزان تهیه شده است.

تاریخ تکمیل: شهریور ۱۴۰۴

حجم داده: ۱۰۰۰ رکورد

تعداد ویژگی: ۱۷ ویژگی

روش‌های استفاده شده: Classification, Clustering, Association Rules