ارائه سیستم تصمیم یار جهت انتخاب رشته تحصیلی بر اساس شاخصه های فردی و محیطی با استفاده از تکنیک‌های یادگیری ماشین
(مطالعه موردی دانش‌آموزان دختر پایه نهم - شهرستان پاکدشت)

نام و نام خانوادگی نویسنده اول
ثریا خداخواه

نام و نام خانوادگی استاد راهنما
جناب آقای دکتر محمد ربیعی

چکیده

از مهم‌ترین دغدغه‌های دانش‌آموزان پس از اتمام دوره تحصیلی مقطع متوسطه اول، انتخاب رشته تحصیلی است. دانش‌آموزان در سن ۱۵سالگی جهت انتخاب رشته تحصیلی نیاز به هدایت و راهنمایی دارند. عوامل مؤثر در انتخاب رشته عمدتاً شامل جنسیت، نگرش، خانواده، معلم، انتخاب شغل و پیشرفت فردی است. تلاش‌های اولیه برای شناسایی این عوامل تأثیرگذار، مهم است؛ بنابراین ارائه یک سیستم پیشنهاددهنده جهت انتخاب رشته تحصیلی دانش‌آموزان پایه نهم ضروری می گردد. به این منظور استفاده از روش‌های یادگیری ماشین و به کار بردن ویژگی های مهم و تاثیر گذار در فرایند انتخاب رشته تحصیلی موجب افزایش کارایی و دقت می گردد .

در این پژوهش از دیتاست دانش‌آموزان دختر دبیرستان پایه نهم شهرستان پاکدشت در سال تحصیلی ۱۴۰۱-۱۴۰۲ استفاده شده است. همچنین از الگوریتم خوشه‌بندی فضایی مبتنی بر چگالی در کاربردهای دارای نویز استفاده می گردد. ابتدا از دیتاست انتخابی ویژگی های موثر در انتخاب رشته جدا می گردد. سپس در مرحله انتخاب ویژگی شاخصه های مهم و تاثیر گذار که از طریق پرسشنامه های مربوطه به دست آمده است به دیتاست اضافه می شود. دقت الگوریتم استفاده شده ۰/۹۲ % می باشد و در مقایسه با الگوریتم k-means دقت بالاتری مشاهده می گردد. در پایان بر اساس نتایج به دست آمده به وسیله تکنیک های مربوطه به ارزیابی نتایج پرداخته شده است.

واژگان کلیدی: سیستم پیشنهاددهنده، یادگیری ماشین، الگوریتم‌های خوشه‌بندی، الگوریتم مبتنی بر چگالی

مقدمه

از مهم‌ترین دغدغه‌های دانش‌آموزان پس از اتمام دوره تحصیلی مقطع متوسطه اول، انتخاب رشته تحصیلی است. دانش‌آموزان در سن ۱۵سالگی مجبور به انتخاب رشته تحصیلی هستند؛ بنابراین نیاز به هدایت و راهنمایی دارند. یکی از اهداف مهم آموزش‌وپرورش فراهم‌کردن شرایط برای هدایت دانش‌آموزان و تربیت انسان‌های کارآمد جهت ایفای نقش در

زندگی فردی و اجتماعی است. این اهداف زمانی محقق خواهد شد که دانش‌آموزان در مسیر درست راهنمایی و هدایت تحصیلی قرار گیرند و از میان رشته‌های موجود، مناسب‌ترین رشته را باتوجه‌به معیارهای ذکر شده انتخاب کنند.
اگر رشته تحصیلی درست و صحیح انتخاب شود در زمینه‌سازی افزایش دانش و مهارت مؤثر خواهد بود. در انتخاب درست، باید ویژگی‌های دانش‌آموز از جمله میزان بهره هوشی، نمرات تحصیلی، نقاط ضعف و قوت در دروس مختلف، استعداد، میزان علاقه‌مندی و امکانات و نیازهای کشور مورد بررسی قرار بگیرد. اگر دانش‌آموز بتواند رشته‌ای را انتخاب کند که با توانایی‌ها و علایقش مطابقت داشته باشد تا حد زیادی توانسته در مسیر رسیدن به موفقیت قدم بردارد؛ بنابراین انتخاب رشته، نیازمند بررسی دقیق و درست حجم وسیعی از داده‌های آموزشی سال‌های گذشته دانش‌آموزان است. این فرایند طولانی برای دانش‌آموز و مشاور مدرسه، به امری دشوار تبدیل گردیده است.
باتوجه‌به اهمیت انتخاب رشته تحصیلی مناسب توسط دانش‌آموزان، نقش يك مدل پیشنهادی و توسعه‌ای در این راستا پررنگ می‌گردد. بهره‌گیری از خوشه‌بندی و استفاده از معیار شباهت در یک سیستم پیشنهاددهنده برای هدایت تحصیلی دانش‌آموزان جهت انتخاب رشته پایه نهم در موفقیت و رضایتمندی وی بسیار مؤثر خواهد بود. نتایج این تحقیق در انتخاب رشته دانش‌آموزان برای رشته‌های مختلف نظری، فنی و حرفه‌ای و کار و دانش کاربرد دارد که با اندکی تغییرات و استفاده از دیتاست‌های مختلف، در زمینه انتخاب رشته دانشگاهی و نیز در مراکز و مؤسسات آموزشی قابل‌استفاده خواهد بود. یک سیستم توصیه‌گر در زمینه انتخاب رشته تحصیلی دانش‌آموزان به دلیل موارد مشروحه ذیل لازم و ضروری است:

- نیاز به ارائه سیستم‌های توصیه‌گر جدید که نسبت به گذشته از دقت بیش‌تری برخوردار باشند.
- نیاز به ارائه سیستم‌های توصیه‌گری که ویژگی‌های جامعه و علایق افراد را مدنظر قرار داده تا توصیه‌های دقیق‌تری داشته باشد.
- باتوجه‌به گسترش روزافزون رشته‌های تحصیلی مختلف و گردآوری داده‌ها در مقیاس بالا باید مدلي ارائه گردد که توانايي مديريت حجم زياد داده‌ها را داشته باشد.

در این پژوهش کوشش می‌شود تا یک سیستم توصیه‌گر هوشمند جدید با الگوریتم یادگیری ماشین بدون نظارت ارائه گردد. هدف اصلي اين سيستم پيشنهاد بهترين آيتم‌ها به كاربران باتوجه‌به علايق آن‌ها و اطلاعات موجود است. همچنین باتوجه‌به حجم زیاد داده‌ها و اهمیت فاکتور زمان، سعی می‌گردد حجم داده‌ها با استفاده از الگوریتم‌های یادگیری ماشین کاهش یابد، بدون این‌که دقت الگوریتم ارائه شده کاهش یابد.

روش تحقیق

در این روش، از خوشه‌بندی مبتنی بر چگالی استفاده خواهد شد که یک تکنیک مقیاس‌پذیر برای استخراج خوشه‌ها است و به هیچ پارامتر مشخص شده توسط کاربر نیاز ندارد. مزیت این روش این است که بر روی مجموعه‌داده‌های بزرگ به‌خوبی کار می‌کند و همچنین نگرانی در مورد زمان بسیار طولانی برای تدوین داده‌ها در یک مدل را کاهش می‌دهد .(شکری،۱۳۹٦) در این پژوهش از مجموعه‌داده‌های دانش‌آموزان پایه نهم که از پایگاه‌داده دبیرستان دخترانه در شهر پاکدشت در سال تحصیلی ۱٤۰۲-۱٤۰۱ جمع‌آوری‌شده است به‌منظور ارائه سیستم پیشنهاددهنده جهت انتخاب رشته تحصیلی استفاده گردیده است. همچنین ویژگی‌های فردی و محیطی دانش‌آموزان به‌عنوان متغیرهای پژوهش در انتخاب رشته تحصیلی مورداستفاده قرار گرفته است. به‌منظور افزایش مقیاس‌پذیری و ارائه یک سیستم توصیه‌گر بادقت بالا، از خوشه‌بندی بردار ویژگی کاربران (جهت تسهیل در انتخاب کاربران مشابه) به روش خوشه‌بندی (DBSCAN) استفاده خواهد شد.

یافته ها

تحلیل داده‌ها فرایند جمع‌آوری، مدل‌سازی و تحلیل داده‌ها با استفاده از روش‌ها و تکنیک‌های مختلف آماری و منطقی است. کسب‌وکارها به فرایندها و ابزارهای تجزیه‌وتحلیل تکیه می‌کنند تا دیدگاه‌هایی را استخراج کنند که از تصمیم‌گیری استراتژیک و عملیاتی پشتیبانی می‌کنند. قبل از اینکه به جزئیات در مورد تحلیل همراه با روش‌ها و تکنیک‌های آن پرداخته شود، باید

ظرفیت تجزیه‌وتحلیل داده‌ها را درک نمود. این فرایند شامل مراحل مختلف می‌گردد. شناسایی مرحله‌ای است که سؤالاتی را که باید به آن پاسخ داد، مشخص می‌گردد. جمع‌آوری مرحله‌ای است که جمع‌آوری داده‌های موردنیاز آغاز می‌گردد که از کدام منبع اطلاعاتی و چگونه از آن‌ها استفاده می‌گردد. جمع‌آوری داده‌ها می‌تواند به اشکال مختلفی از قبیل منابع داخلی یا خارجی، نظرسنجی، مصاحبه، پرسش‌نامه و صورت گیرد. در پاکسازی داده همه داده‌هایی که جمع‌آوری می‌شود مفید نیستند. وقتی مقادیر زیادی داده در فرمت‌های مختلف جمع‌آوری شود، به‌احتمال زیاد داده‌های تکراری یا با فرمت بدشکل خواهد گرفت پس رکوردهای تکراری یا با فرمت بد بایستی پاک شود. در مرحله تجزیه‌وتحلیل با استفاده از تکنیک‌های مختلف مانند تحلیل آماری، رگرسیون، شبکه‌های عصبی، تحلیل متن و غیره می‌توانید شروع به تحلیل و دستکاری داده‌ها برای استخراج نتایج مربوطه نمود. در این مرحله روندها، همبستگی‌ها، تغییرات و الگوهایی پیدا می‌شود که می‌توانند به سؤالاتی که اولین‌بار در مرحله شناسایی وجود داشت پاسخ دهد. در مرحله آخر که بایستی به تفسیر نتایج پرداخته شود محقق با بخش‌های عملی بر اساس یافته‌های پژوهش مواجه می‌شود.

در این بخش به سؤالات تحقیق بر اساس داده‌ها و یافته‌های محقق، پاسخ داده می‌شود. داده‌ها با فرمت مناسبی ارائه می‌شوند. مدل (ها) اجرا شده و نتیجه آن مشخص می‌شود.

روش نرمال‌سازی و تبدیل داده ها

در دیتاست دانش‌آموزان، ویژگی معدل از نوع اعشاری بود که به اعداد بدون اعشار تبدیل شد. باتوجه‌به شکل ۱ در دیتاست دانش‌آموزان مقادیر خالی و ازدست‌رفته‌ای وجود نداشت.

شکل ۱ نداشتن مقدار خالی در دیتاست دانش‌آموزان دبیرستان دخترانه

```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   predict        275 non-null    int64
 1   tipology       275 non-null    int64
 2   filed_code     275 non-null    int64
 3   moadel         275 non-null    float64
 4   repeat         275 non-null    int64
 5   sabke_dars     275 non-null    int64
 6   foghe_barname  275 non-null    int64
```

در دیتاست اصلی تمام اعداد از نوع اعشاری بودند که در دیتاست پیش‌بینی به اعداد بدون اعشار تبدیل شدند. همچنین با درجه‌بندی ویژگی‌های دیتاست پیش‌پردازش و نرمال‌سازی داده‌های انجام گردید تا دقت در نتایج به‌دست‌آمده و الگوریتم ها حاصل گردد. تبدیل‌ها مطابق جدول‌های ۱ تا ۶ انجام گردید.

کد ملی دانش‌آموزان دبیرستان دخترانه پایه نهم	
تعداد کد ملی دانش‌آموزان	۲۷۵

جدول ۱ درجه‌بندی ویژگی کد در دیتاست دانش‌آموزان

معدل	
نمره معدل دانش‌آموزان از گستره ۱۳ تا ۱۹	گستره ۱۳ تا ۱۹

جدول ۲ درجه‌بندی ویژگی معدل در دیتاست دانش‌آموزان

هشت گروه شخصیت‌شناسی	
پیکارگر(ENFP)	۱
بازرس(ISTJ)	۲
حامی(ISFJ)	۳
مدیر(ESTJ)	۴

جدول ۳ درجه‌بندی ویژگی تیپولوژی شخصیت در دیتاست دانش‌آموزان

چهار نوع سبک یادگیری	
خواندن متن درس	۱
گوش‌دادن به معلم (ضبط‌صوت)	۲
دیدن فیلم یا تصاویر آموزشی	۳
انجام آزمایش و کار عملی	۴

جدول ٤ درجه‌بندی ویژگی سبک یادگیری در دیتاست دانش‌آموزان

داشتن فعالیت فوق‌برنامه	
بله	۱
خیر	۰

جدول ٥ درجه‌بندی ویژگی داشتن فعالیت فوق‌برنامه در دیتاست دانش‌آموزان

کد رشته	
رشته گرافیک رایانه‌ای کد ۹۹۰۱۱	۱
رشته معماری داخلی کد ۶۱۸۸۱	۲
رشته خیاطی لباس شب و عروس کد ۶۱۴۷۱	۳

رشته مدیریت و برنامه‌ریزی امور خانواده کد ۹۴۱۰۱	۴
رشته خیاطی لباس شب و عروس کد ۹۳۱۵۴	۵
رشته تصویرسازی و جلوه‌های ویژه رایانه‌ای کد ۶۲۳۲۱	۶

جدول ٦ درجه‌بندی ویژگی داشتن فعالیت فوق‌برنامه در دیتاست دانش‌آموزان

الگوریتم‌های به‌کاررفته در پیاده‌سازی

الگوریتم‌های به کار گرفته شده در پیاده‌سازی الگوریتم‌های خوشه‌بندی dbsccan و k-means می‌باشد.

پیاده سازی الگوریتم‌های خوشه‌بندی

در این مرحله پیاده‌سازی الگوریتم‌های مختلف در داخل داده‌های دیتاست پژوهش بایستی انجام گردد. الگوریتم‌های به کار گرفته شده شامل الگوریتم‌های خوشه‌بندی dbsccan و k-means می‌باشد که dbscsn به‌عنوان الگوریتم اصلی و k-means به‌عنوان الگوریتم مقایسه‌ای با الگوریتم اصلی استفاده می‌گردد. جهت پیاده‌سازی الگوریتم‌ها از نرم‌افزار پایتون استفاده می‌گردد. برای این منظور در ابتدا باید دیتاست به نرم‌افزار وارد شود. پس از واردکردن دیتاست به پایتون، دستورهای مربوط به شرح کلی دیتاست اجرا می‌گردد. نتایج در شکل ۲ دیده می‌شود.

شکل ۲ نمایش شرح کلی داده‌های دیتاست

tipology	filed_code	moadel	repeat	sabke_dars	foghe_barname
4	61471	16.77	1	2	0
3	62321	17.83	1	4	0
4	61471	16.17	1	1	1
6	61881	18.66	1	3	1
1	61881	14.64	1	3	1

این خروجی مربوط به اجرای الگوریتم t-SNE است که به‌منظور کاهش ابعاد داده‌ها و تصویرسازی آن‌ها در فضای دو یا سه‌بعدی استفاده می‌شود. این خروجی نشان می‌دهد که الگوریتم t-SNE در ارتباط با داده‌ها و محاسبات خود در حال انجام مراحل مختلف است. در ادامه به تشریح هر خط از خروجی پرداخته می‌شود .

[t-SNE] Computing 91 nearest neighbors...

در این خط، الگوریتم t-SNE در حال محاسبه ۹۱ همسایه نزدیک‌تر برای هر نمونه است. این همسایگان نزدیک در محاسبات t-SNE استفاده می‌شوند.

[t-SNE] Indexed 275 samples in 0.000s...

در این خط، الگوریتم t-SNE تعداد ۲۷۵ نمونه را در ۰.۰۰۰ ثانیه (زمان ناچیز) پردازش کرده است. این نمونه‌ها برای اجرای الگوریتم t-SNE استفاده می‌شوند.

[t-SNE] Computed neighbors for 275 samples in 0.019s...

در این خط، الگوریتم t-SNE همسایگان نزدیک را برای ۲۷۵ نمونه محاسبه کرده است. این همسایگان نزدیک در محاسبات t-SNE استفاده می‌شوند.

[t-SNE] Computed conditional probabilities for sample 275 / 275

در این خط، الگوریتم t-SNE احتمال شرطی برای نمونه ۲۷۵ را محاسبه کرده است. این احتمالات شرطی در محاسبات t-SNEاستفاده می‌شوند.

[t-SNE] Mean sigma: 3.734338

در این خط، الگوریتم t-SNE میانگین مقدار sigma را برابر با ۳.۷۳۴۳۳۸ محاسبه کرده است. این مقدار نشان می‌دهد که همسایگان نزدیک در محاسبات t-SNE چقدر دور از هم قرار می‌گیرند.

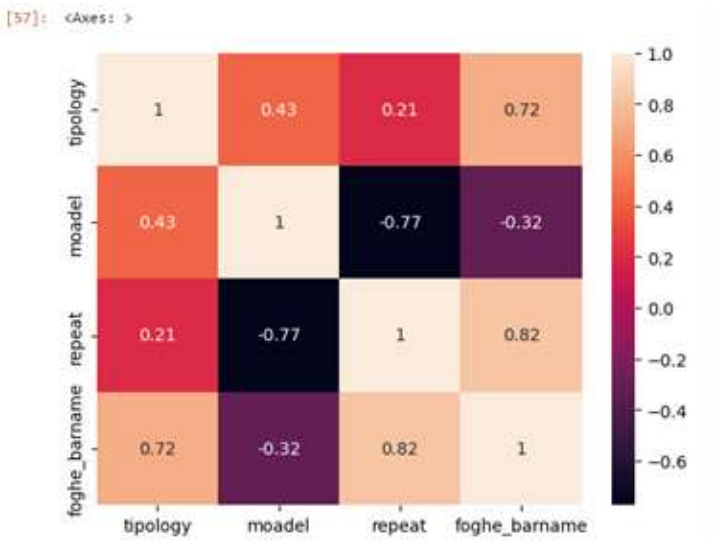
[t-SNE] KL divergence after 1000 iterations: 1.713743

در این خط، الگوریتم t-SNE مقدار KL divergence پس از ۱۰۰۰ تکرار را برابر با ۱.۷۱۳۷۴۳ محاسبه کرده است .

در دیتاست ۸ نوع تیپولوژی شخصیتی داریم که هر نوع حالات و رفتارهای منحصربفرد و حتی مشترک با یکدیگر دارند. از پرسشنامه مهمترین سوال که گویای نوع آن تیپولوژی است را بر اساس محور x که فعالیت های فوق برنامه و در محور y سبک یادگیری درس را برای هر تیپولوژی و مهمترین و نزدیکترین پرسش در پرسشنامه برای آن نوع شخصیت را نمایش می دهد. این پلات برای یافتن خطا های احتمالی، مقایسه و اندازه گیری توزیع متغیرها استفاده شده است.

ماتریس همبستگی دیتاست: در این قسمت، داده‌ها بر اساس ستون 'repeat' گروه‌بندی می‌شوند و میانگین مقادیر هر گروه محاسبه می‌شود. سپس با استفاده از iloc[:,:۴] ، ستون‌های اول تا چهارم این مجموعه‌داده‌ها برگشت داده می‌شوند. نتیجهٔ این عملیات یک DataFrame است که شامل میانگین مقادیر هر گروه برای ستون‌های انتخاب شده است. این قسمت از کتابخانه seaborn استفاده می‌کند تا یک نمودار Heatmap از ضرایب همبستگی بین ستون‌های اول تا چهارم مجموعه‌داده‌ها را رسم کند. ابتدا داده‌ها بر اساس ستون 'sabke_dars' گروه‌بندی شده و میانگین مقادیر هر گروه محاسبه می‌شود. سپس با استفاده از iloc [:,:۴]، ستون‌های اول تا چهارم این مجموعه‌داده‌ها برگشت داده می‌شوند. سپس تابع corr() بر روی این مجموعه‌داده‌ها فراخوانی می‌شود تا ضرایب همبستگی بین ستون‌ها محاسبه شود. در نهایت، با استفاده از sns.heatmap()، نمودار Heatmap از این ضرایب همبستگی رسم می‌شود. ضریب همبستگی هر دو ستون در هر خانه از Heatmap با استفاده از اعداد و رنگ‌ها نشان داده می‌شود. با تنظیم آرگومان annot=True، اعداد ضرایب همبستگی در داخل هر خانه نمایش داده می‌شوند. این ماتریس در شکل ۳ دیده می‌شود.

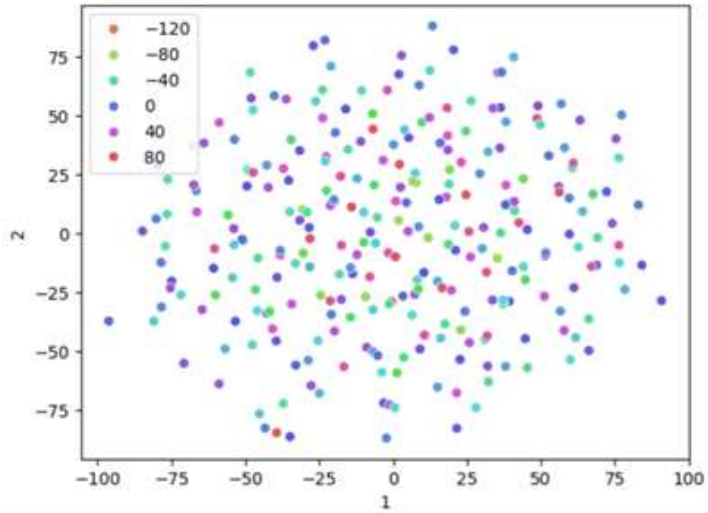
شکل ۳ ماتریس همبستگی ویژگی‌های دیتاست



تجزیه‌وتحلیل شکل ۴ به این صورت است که محور افقی شامل مؤلفه اول در t-SNE، محور عمودی شامل مؤلفه دوم در t-SNE و رنگ شامل مؤلفه سوم در t-SNE می‌باشد. هر نقطه نشان‌دهنده بیانگر یک پاسخ‌دهنده به پرسش‌نامه است. رنگ هر نقطه نشان‌دهنده مقدار مؤلفه سوم t-SNE برای آن پاسخ‌دهنده است. نقاط با رنگ‌های مشابه تمایل به خوشه شدن در کنار یکدیگر دارند. این نشان می‌دهد که پاسخ‌دهندگانی که در یک خوشه قرار دارند، به طور مشابه به سؤالات پرسش‌نامه پاسخ داده‌اند؛ بنابراین پاسخ‌دهندگان به پرسش‌نامه به ۳ گروه اصلی تقسیم می‌شوند:

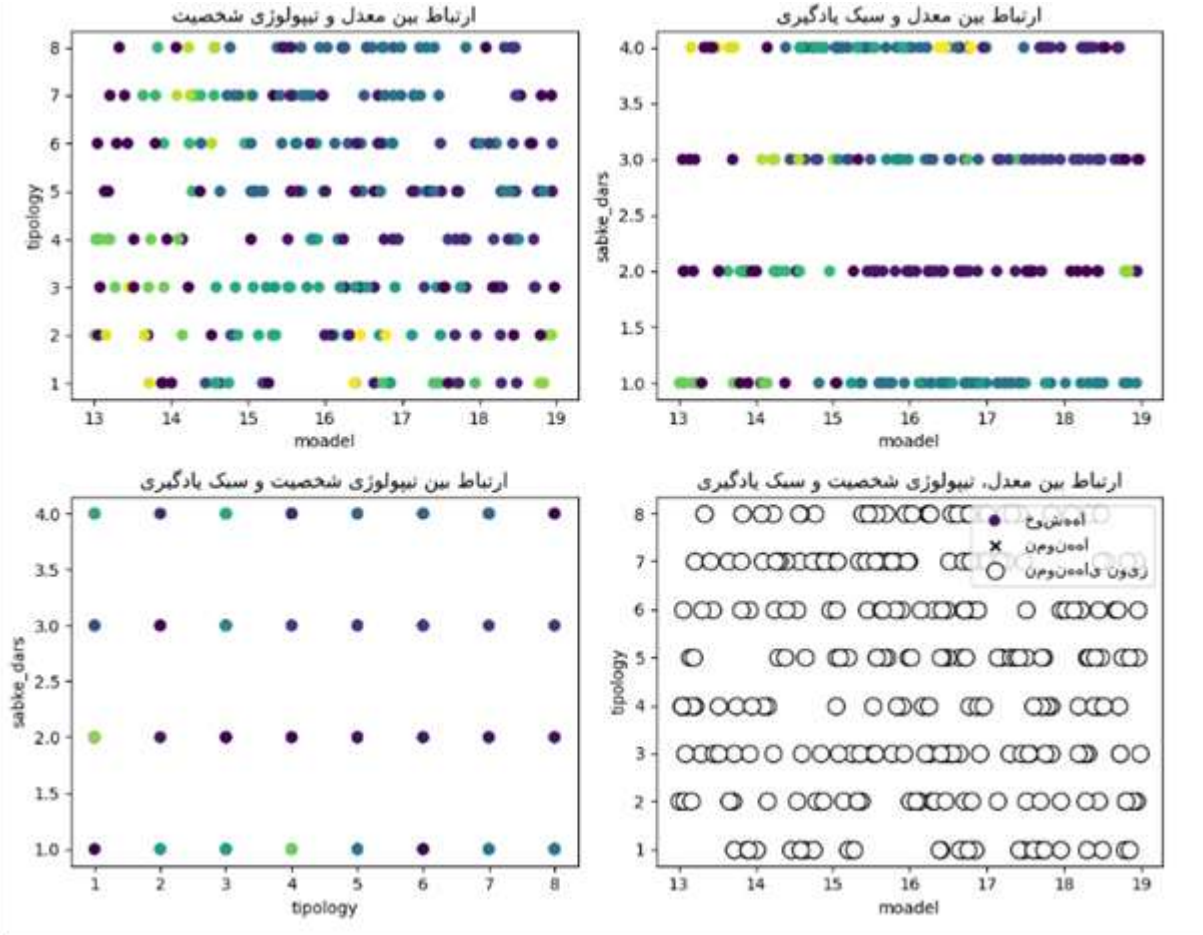
گروه اول (سبز) تمایل به سبک یادگیری "واقعی" و گرایش به فعالیت‌های فوق‌برنامه "اجتماعی" دارند. گروه دوم (آبی) تمایل به سبک یادگیری "تجربی" و گرایش به فعالیت‌های فوق‌برنامه "هنری" دارند. گروه سوم (قرمز) تمایل به سبک یادگیری "انتزاعی" و گرایش به فعالیت‌های فوق‌برنامه "علمی" دارند. این تفسیر فقط بر اساس ۳ مولفه اول t-SNE است. ممکن است مولفه های دیگر اطلاعات بیشتری در مورد پاسخ دهندگان ارائه دهند.

شکل ٤ تحلیل متغیرهای موردتوجه در فرضیه بیان مساله پرسش‌نامه و سبک یادگیری و گرایش به فعالیت



در شکل ۵ چهار پلات نشان داده شده است که در ارتباط بین معدل و تیپولوژی شخصیت، نقاط بر اساس تیپولوژی شخصیت رنگی می شوند. به نظر می رسد رابطه ی قوی بین معدل و تیپولوژی شخصیت وجود ندارد. با این حال، می توان مشاهده کرد که ISTJها تمایل به معدل های بالاتر و ESFPها تمایل به معدل های پایین تر دارند. در ارتباط بین معدل و سبک یادگیری نقاط بر اساس سبک یادگیری رنگی می شوند. به نظر می رسد رابطه ی قوی بین معدل و سبک یادگیری وجود ندارد. با این حال، می توان مشاهده کرد که دانش آموزانی که سبک یادگیری "واقعی" دارند تمایل به معدل های بالاتر دارند. درارتباط بین تیپولوژی شخصیت و سبک یادگیری نقاط بر اساس تیپولوژی شخصیت رنگی می شوند. رابطه ی قوی بین تیپولوژی شخصیت و سبک یادگیری وجود دارد. برای مثال، ISTJها تمایل به سبک یادگیری "واقعی" و ESFPها تمایل به سبک یادگیری "تجربی" دارند. در ارتباط بین معدل، تیپولوژی شخصیت و سبک یادگیری نقاط بر اساس خوشه ها رنگی می شوند. این نشان می دهد که چگونه معدل، تیپولوژی شخصیت و سبک یادگیری با هم مرتبط هستند. می توان مشاهده کرد که ISTJها با سبک یادگیری "واقعی" تمایل به بالاترین معدل ها دارند. ESFPها با سبک یادگیری "تجربی" تمایل به پایین ترین معدل ها دارند. نمونه های نر با رنگ سفید و لبه های سیاه نمایش داده می شوند.

شکل ۵ ارتباط بین ویژگی های مسئله با یکدیگر



بر اساس خروجی‌ها و مفاهیم روان‌شناسی بر اساس تکنیک MBTI توانستیم مفاهیم بین ویژگی‌های افزوده شده یعنی تیپولوژی شخصیتی برای هر نوع شخصیت با شناسه خودش را بیابیم. برای بررسی شناسایی تیپولوژی شخصیتی بر اساس MBTI و ارتباط آن با رشته‌های تحصیلی، باید توجه داشت که هر فرد می‌تواند با هر تیپولوژی شخصیتی در هر رشته‌ای موفق باشد. اما برخی از تیپولوژی‌ها معمولاً با برخی رشته‌ها بیشتر سازگاری دارند.

جدول ۷ رابطه میان رشته های هنرستان های شاخه کاردانش و گرایش تیپولوژی ها به رشته های تحصیلی می باشد. که از عدد ۱ تا ۵ وزن دهی شده است که عدد ۱ کمترین گرایش و تمایل و عدد ۵ بیشترین تمایل آن تیپولوژی به رشته تحصیلی می باشد. این نتایج حاصل نگرش تحلیلی و نظر مشاور تحصیلی می باشد.

رشته تحصیلی	بازرس (ISTJ)	مجری (ESTJ)	روشنفکر (INTJ)	فرمانده (ENTJ)	واسطه‌گر (INFP)	پیکارگر (ENFP)	حامی (ISFJ)	سفیر (ESFJ)
گرافیک رایانه‌ای	2	3	4	3	4	5	3	4
معماری داخلی	3	4	4	4	3	2	4	5
خیاطی لباس شب و عروس	1	2	1	1	5	4	5	5

مدیریت و برنامه‌ریزی امور خانواده	4	5	2	3	5	4	5	5
تصویرسازی و جلوه‌های ویژه رایانه‌ای	2	3	4	3	4	5	3	4

جدول ۷ سازگاری و انطباق رشته‌ها با تیپولوژی شخصیت

تحلیل دقت الگوریتم‌های بکار گرفته شده در دیتاست دانش‌آموزان

باتوجه‌به تجزیه‌وتحلیل دادها و بر اساس شکل ۶ می توان این گونه بیان کرد که محور افقی مولفه اول t-SNE محور عمودی مولفه دوم t-SNE و رنگ تیپولوژی شخصیت می باشد. هر نقطه نشان دهنده یک پاسخ دهنده به پرسشنامه است. رنگ هر نقطه نشان دهنده تیپولوژی شخصیت آن پاسخ دهنده است. نقاط با رنگ های مشابه تمایل به خوشه شدن در کنار یکدیگر دارند.این نشان می دهد که پاسخ دهندگانی که در یک خوشه قرار دارند، تیپولوژی شخصیت مشابهی دارند. پاسخ دهندگان به پرسشنامه به ۴ گروه اصلی تقسیم می شوند:

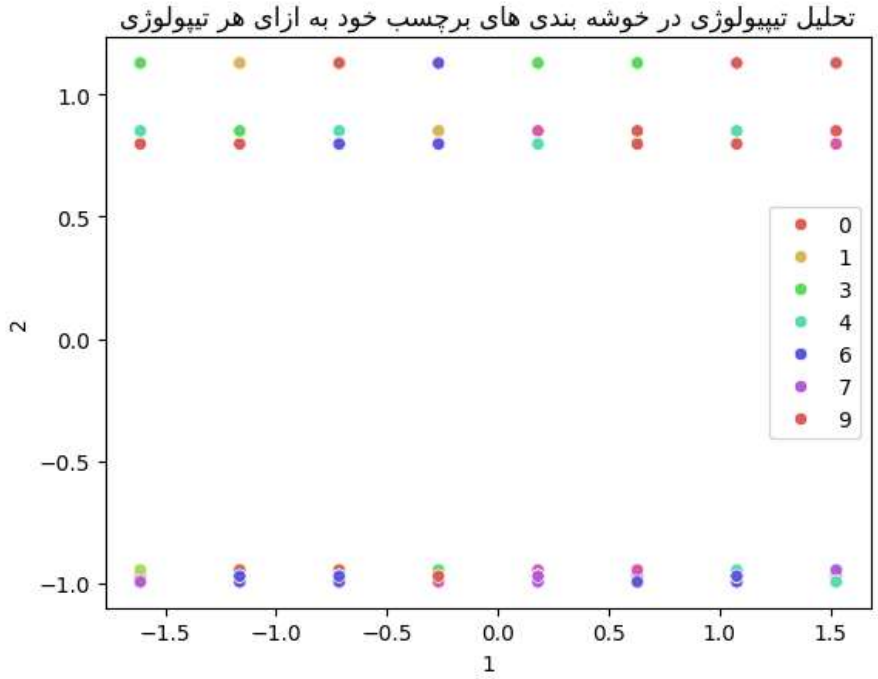
گروه اول (سبز): این گروه شامل تیپولوژی‌های ISTJ و ISFJ است.

گروه دوم (آبی): این گروه شامل تیپولوژی‌های ENFJ و INFJ است.

گروه سوم (قرمز): این گروه شامل تیپولوژی‌های ESTP و ESFP است.

گروه چهارم (بنفش): این گروه شامل تیپولوژی‌های INTJ و ENTJ است.

شکل ٦ تحلیل تیپولوژی‌ها درباره خوشه‌بندی‌های برچسب خودش به‌ازای هر کدام از آن‌ها



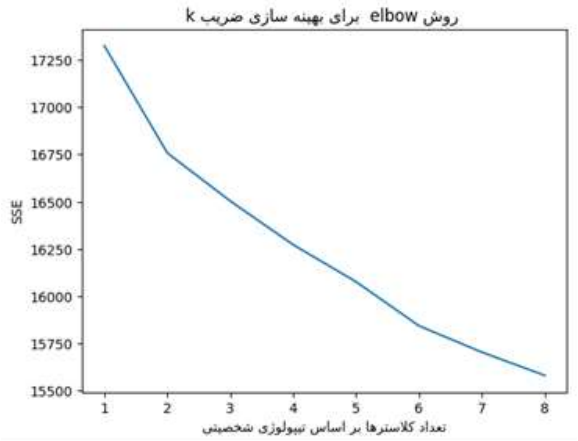
بر اساس دیتاست و داده‌های پرسش‌نامه و اجرای الگوریتم پیشنهادی این تحقیق یعنی DBSCAN باید بادقت مناسبی تیپولوژی‌های ۸ گانه در این تحقیق دست به انتخاب های زیر بزنند؛ بنابراین با مقایسه همبستگی پیش بینی تحلیلی در این پژوهش و انتخاب رشته واقعی دانش‌آموزان می توان صحت و دقت این تحقیق را بررسی کرد.

تجزیه‌وتحلیل داده‌ها با مشاهده و بررسی که هر مقدار در جدول نشان‌دهنده تعداد دانش‌آموزان با تیپولوژی خاص است که آن رشته را انتخاب کرده‌اند. محبوب‌ترین رشته برای ISTJها، ISFJها، INTJها و ENTJها است. این نشان می‌دهد که این تیپولوژی‌ها به کارهای عملی و فنی علاقه دارند.معماری داخلی محبوب‌ترین رشته برای ENFJها و INFJها است که نشان می‌دهد این تیپولوژی‌ها به کارهای خلاقانه و هنری علاقه دارند. خیاطی لباس شب و عروس محبوب‌ترین رشته برای ESTPها و ESFPها است که نشان می‌دهد این تیپولوژی‌ها به کارهای خلاقانه و اجتماعی علاقه دارند. در مورد مدیریت و برنامه‌ریزی امور خانواده هیچ تیپولوژی خاصی به طور قابل توجه به این رشته علاقه نشان نمی‌دهد. تصویرسازی و جلوه‌های ویژه رایانه‌ای این رشته برای ISTJها، ISFJها، ENFJها و INFJها جذابیت دارد که نشان علاقه این تیپولوژی‌ها به کارهای خلاقانه و بصری است.

روش elbow برای کاهش مقدار ضریب k:

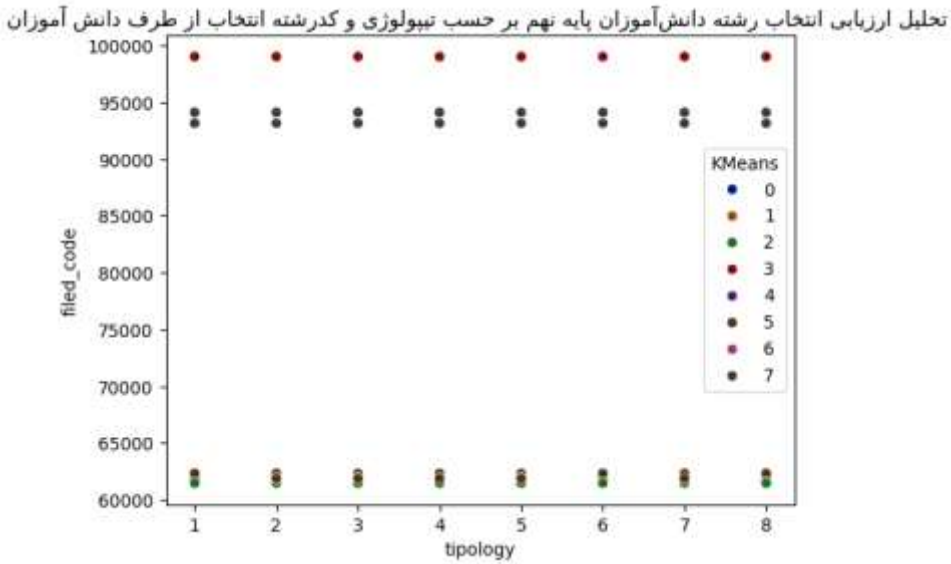
محور افقی تعداد خوشه‌ها(k)، محور عمودی مجموع مربعات خطا (SSE) و خط نشان‌دهنده SSE برای هر مقدار k است. نقطه "آرنج" در نمودار نقطه‌ای است که در آن شیب‌خط به طور قابل‌توجهی تغییر می‌کند و نشان‌دهنده تعداد خوشه‌های بهینه است. آرنج در k=4 قرار دارد که نشان می‌دهد تعداد خوشه‌های بهینه برای دیتاست ما ۴ است. روش Elbow یک روش تقریبی برای انتخاب تعداد خوشه‌ها است. روش‌های دیگری مانند Silhouette Coefficient و Gap Statistic نیز وجود دارند. برای انتخاب بهترین تعداد خوشه‌ها، باید از چندین روش استفاده کرد و نتایج را با هم مقایسه نمود. به‌طورکلی به نقطه آرنج در k=4 به‌عنوان تعداد خوشه‌های بهینه اشاره شده است. انتخاب k=4 به‌عنوان تعداد خوشه‌ها به این معنی است که ۴ تیپولوژی شخصیتی اصلی در دیتاست وجود دارد که بیشترین و پراهمیت‌ترین نقش را در تحلیل ایفا می‌کنند. این تیپولوژی‌ها ممکن است با تیپولوژی‌های MBTI مانند ISTJ، ISFJ، ENFJ و INFJ مطابقت داشته باشند. این تحلیل بر اساس یک مطالعه کوچک انجام شده است و برای تعمیم آن به جمعیت‌های بزرگ‌تر به تحقیقات بیشتری نیاز است. پلات elbow جهت بهینه‌سازی ضریب k در شکل ۷ آورده شده است.

شکل ۷ پلات Elbow



در تحلیل ارزیابی انتخاب رشته دانش‌آموزان پایه نهم بر حسب تیپولوژی و کد رشته از طرف دانش‌آموزان که در شکل ۸ نمایش داده شده است. محور افقی مولفه اول t-SNE محورعمودی مولفه دوم t-SNE و رنگ تیپولوژی شخصیتی است. هر نقطه نشان دهنده یک دانش آموز است. رنگ هر نقطه نشان دهنده تیپولوژی شخصیتی آن دانش آموز است. موقعیت هر نقطه نشان دهنده کد رشته انتخابی توسط دانش آموز است. نقاط با رنگ های مشابه تمایل به خوشه شدن در کنار یکدیگر دارند که نشان می دهد که دانش‌آموزان با تیپولوژی شخصیتی مشابه، تمایل به انتخاب رشته های مشابهی دارند. پلات نشان می دهد که دانش‌آموزان با تیپولوژی هایISTJ، ISFJ و INTJ تمایل به انتخاب رشته های تجربی و ریاضی دارند. دانش‌آموزان با تیپولوژی های ENFJ و INFJ تمایل به انتخاب رشته های انسانی و علوم اجتماعی دارند. دانش‌آموزان با تیپولوژی های ESTP و ESFP تمایل به انتخاب رشته های فنی و حرفه ای و کاردانش دارند .این تفسیر فقط بر اساس ۲ مولفه اول t-SNE است. ممکن است مولفه های دیگر اطلاعات بیشتری در مورد دانش‌آموزان ارائه دهند. تیپولوژی های ISTJ ISFJ و INTJ به دلیل ویژگی های شخصیتی مانند وظیفه شناسی، نظم و انضباط و تفکر انتقادی، تمایل به انتخاب رشته های تجربی و ریاضی دارند. تیپولوژی های ENFJ و INFJ به دلیل ویژگی های شخصیتی مانند همدلی، ایده آلیسم و گرایش به کمک به دیگران، تمایل به انتخاب رشته های انسانی و علوم اجتماعی دارند. تیپولوژی های ESTP و ESFP به دلیل ویژگی های شخصیتی مانند برونگرایی، عمل گرایی و لذت بردن از زندگی، تمایل به انتخاب رشته های فنی و حرفه ای و کاردانش دارند.

شکل ۸ تحلیل ارزیابی انتخاب رشته دانش‌آموزان پایه نهم بر حسب تیپولوژی و کد رشته K-Means



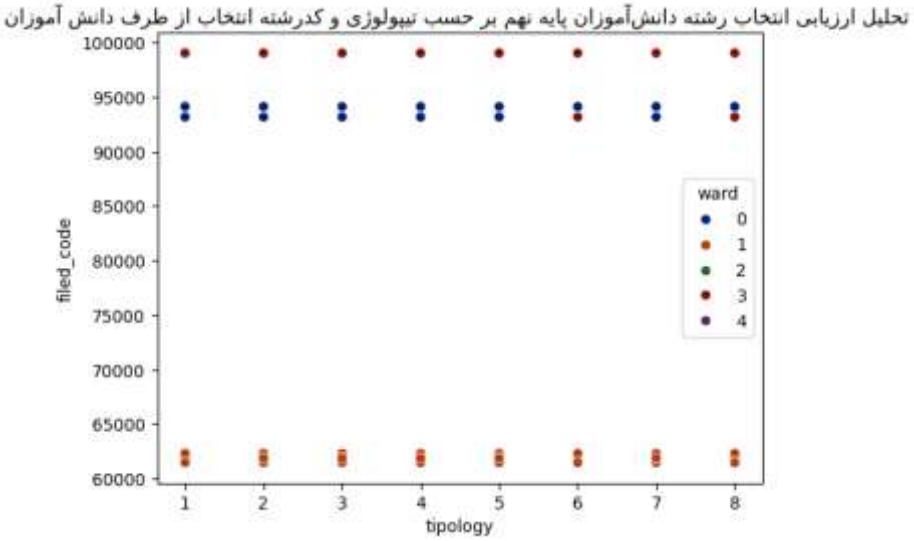
مقایسه با پلات KMeans

پلات KMeans نشان می‌دهد که دانش‌آموزان با تیپولوژی‌های شخصیتی مشابه، تمایل به انتخاب رشته‌های مشابهی دارند. پلات DBSCAN نشان می‌دهد که دانش‌آموزان با الگوهای رفتاری مشابه، تمایل به انتخاب رشته‌های مشابهی دارند.

الگوریتم ward

در خوشه‌های سلسله‌مراتبی الگوریتم Ward که در شکل ۹ آمده است هر نقطه نشان‌دهنده یک دانش‌آموز است. رنگ هر نقطه نشان‌دهنده خوشه سلسله‌مراتبی آن دانش‌آموز است. موقعیت هر نقطه نشان‌دهنده کد رشته انتخابی توسط دانش‌آموز است. نقاط با رنگ‌های مشابه تمایل به خوشه شدن در کنار یکدیگر دارند که نشان می‌دهد دانش‌آموزان با الگوهای رفتاری مشابه، تمایل به انتخاب رشته‌های مشابهی دارند.

شکل ۹ تحلیل ارزیابی انتخاب رشته دانش‌آموزان پایه نهم بر حسب تیپولوژی و کد رشته ward



مقایسه با پلات KMeans و DBSCAN

پلات KMeans نشان می‌دهد که دانش‌آموزان با تیپولوژی‌های شخصیتی مشابه، تمایل به انتخاب رشته‌های مشابهی دارند .پلات DBSCAN نشان می‌دهد که دانش‌آموزان با الگوهای رفتاری مشابه، تمایل به انتخاب رشته‌های مشابهی دارند .پلات Ward نیز نشان می‌دهد که دانش‌آموزان با الگوهای رفتاری مشابه، تمایل به انتخاب رشته‌های مشابهی دارند. در مورد شباهت‌ها نیز می‌توان گفت که هر سه پلات نشان می‌دهد که دانش‌آموزان با الگوهای رفتاری مشابه، تمایل به انتخاب رشته‌های مشابهی دارند. هر سه پلات از t-SNE برای کاهش ابعاد داده‌ها استفاده می‌کنند. همچنین درباره تفاوت‌ها می‌توان بیان کرد که پلات KMeans از یک روش مبتنی بر مرکز برای خوشه‌بندی استفاده می‌کند. پلات DBSCAN از یک روش مبتنی بر چگالی برای خوشه‌بندی استفاده می‌کند. پلات Ward از یک روش سلسله‌مراتبی برای خوشه‌بندی استفاده می‌کند. مزایا و معایب هر کدام از الگوریتم‌ها به این صورت است که روش KMeans ساده و سریع است، اما ممکن است به انتخاب تصادفی مرکز خوشه‌ها حساس باشد. روش DBSCAN می‌تواند خوشه‌های با اشکال نامنظم را شناسایی کند، اما به انتخاب مقادیر مناسب برای پارامترها حساس است. روش Ward می‌تواند خوشه‌های سلسله‌مراتبی را شناسایی کند، اما ممکن است کندتر از روش‌های دیگر باشد.

ward vs k-means : 0.621117

اعتبارسنجی

پس از شکل‌گیری مدل، باید آن را مورد بررسی قرارداد تا کیفیت از دیدگاه آنالیز داده مشخص شود. در واقع قبل از رسیدن به استقرار نهایی مدل، مهم است که آن را به طور کامل ارزیابی و مراحل اجرا شده برای ساخت را بررسی نمود. با این روش می‌توان از درستی اهداف اطمینان حاصل کرد. در پایان این مرحله تصمیم‌گیری در مورداستفاده از نتایج استخراج داده باید انجام گردد. اعتبارسنجی بر روی داده‌های دیتاست با استفاده از روش‌های مختلف صورت می‌گیرد. در ادامه به توضیح و تحلیل پرداخته می‌شود.

	Predict	tipology	filed_code	moadel	repeat	sabke_dars	foghe_barname \
0	61471	4	61471	16.77	1	2	0
1	62321	3	62321	17.83	1	4	0
2	61471	4	61471	16.17	1	1	1
3	61881	6	61881	18.66	1	3	1
4	61881	1	61881	14.64	1	3	1
..	...	...	...	...	...	...	...
270	61471	1	61471	14.58	1	4	1
271	61881	5	61881	16.40	1	2	1
272	99011	5	99011	15.08	1	4	0
273	61471	6	61471	15.04	1	4	1
274	99011	5	99011	17.41	1	1	1

خروجی‌هایی که در پایان، جهت‌نمایش پیش‌بینی و دقت قابل‌مشاهده هستند:

DBSCAN Prediction Adjusted Rand Score: ۰.۸۳۸۴۱۲

DBSCAN Prediction Accuracy: ۰.۹۳۰۹۰۹

این اعداد نشان‌دهنده امتیاز تطبیق تصحیح شده و دقت پیش‌بینی الگوریتم DBSCAN روی دیتاست هستند. امتیاز تطبیق تصحیح شده ۰/۸۳۸۴۱۲ است که نشان می‌دهد که DBSCAN با برچسب‌های واقعی در دیتاست به خوبی تطابق دارد. همچنین، دقت پیش‌بینی الگوریتم ۰/۹۳۰۹۰۹ است.

بحث و نتیجه‌گیری

در انتخاب رشته تحصیلی دانش‌آموزان پایه نهم دبیرستان عوامل و معیارهای فردی و محیطی مختلفی نقش دارند. از جمله این موارد می‌توان به نمره معدل، تیپولوژی شخصیت، روش یادگیری دروس و همچنین امکان استفاده از فعالیت‌های فوق‌برنامه از طرف محل تحصیل و حمایت‌های خانواده را نام برد. به این منظور سیستمی طراحی گردید تا با به‌کارگیری تکنیک‌های یادگیری ماشین، فرایند انتخاب رشته با کارایی و دقت بالایی همراه گردد؛ بنابراین از الگوریتم‌های خوشه‌بندی استفاده گردید. جهت ارائه یک سیستم پیشنهاددهنده به‌منظور انتخاب رشته تحصیلی دانش‌آموزان الگوریتم dbscan به کار گرفته شد.

در موضوع انتخاب رشته اگر عوامل مهم در انتخاب در نظر گرفته نشود و یا در این زمینه از کمک مشاور یا روان‌شناس استفاده نشود ممکن است پیامدهای جبران‌ناپذیری در انتظار دانش‌آموزان باشد. یکی از مهم‌ترین آنها بی‌علاقگی و بی‌انگیزگی در رشته است که سبب بی‌معنا شدن تلاش و کوشش فرد می‌شود. این مسئله می‌تواند تمام ابعاد زندگی مانند اعتمادبه‌نفس را تحت‌تأثیر قرار بدهد؛ بنابراین بهتر است برای این انتخاب مهم، تمام عوامل مهم را در کنار یکدیگر در نظر گرفت. از شاخصه‌های مهم در انتخاب رشته دانش‌آموزان، شخصیت‌شناسی آن‌ها در کنار روش‌های یادگیری و استفاده از فعالیت‌های فوق‌برنامه در طول تحصیل می‌باشد. در این پژوهش کوشش می‌شود بر اساس شاخصه‌های ذکر شده و با استفاده از الگوریتم‌های خوشه‌بندی، بهترین دقت مورد بررسی قرار گیرد. با به‌کارگرفتن این روش یک مدل جهت بهینه‌سازی سیستمی که ارائه شده حاصل خواهد شد.

- منابع داخل متن:

مقاله منبع	فارسی
یک نویسنده	(شکری، ۱۳۹۶)

- منابع انتهای مقاله:

صفری، مهدی، فتحی، عبدالله؛ محمدیان، حسین؛ عطایی، محمدرضا؛ "تأثیر آموزش گروهی بر افزایش مهارت‌های اجتماعی دانش‌آموزان دوره متوسطه"، فصلنامه روان‌شناسی تحولی: روان‌شناسان ایرانی، سال هشتم، شماره ۳۱، صفحه ۸۱-۸۹، ۱۳۹۷.

محمودی، محمود، "اثربخشی آموزش هوش هیجانی بر کاهش تعارضات خانوادگی دانش‌آموزان دوره متوسطه"، فصلنامه پژوهشی مطالعات خانواده و ازدواج، سال پنجم، شماره ۱۹، صفحه ۷۱-۸۲، ۱۳۹۵.

شکری، محمدرضا، قهرمانی، فاطمه، "بررسی تأثیر آموزش مدیریت زمان بر بهبود عملکرد تحصیلی دانش‌آموزان دوره دوم متوسطه"، فصلنامه تحقیقات آموزش و یادگیری، سال پنجم، شماره ۱۸، صفحه ۶۳-۷۳، ۱۳۹۶.

رضایی، محمد، عبدالعلی، جواد، "اثربخشی آموزش راهبردهای یادگیری بر عملکرد تحصیلی دانش‌آموزان دوره متوسطه"، فصلنامه تحقیقات آموزش و یادگیری، سال ششم، شماره ۲۲، صفحه ۱۰۹-۱۲۲، ۱۳۹۷.

ابراهیمی، ناصر، محسنی، محمدرضا، "بررسی تأثیر آموزش راهبردهای خودتنظیمی یادگیری بر تسلط تحصیلی دانش‌آموزان دوره متوسطه"، فصلنامه تحقیقات آموزش و یادگیری، سال هفتم، شماره ۲۷، صفحه ۶۷-۷۸، ۱۳۹۸.

کرمی، حسین، موسوی، محمد، "اثربخشی آموزش مهارت‌های ارتباطی بر بهبود رفتار اجتماعی دانش‌آموزان دوره متوسطه"، فصلنامه روان‌شناسی تحولی: روان‌شناسان ایرانی، سال نهم، شماره ۳۶، صفحه ۱۹-۲۸، ۱۳۹۸.

https://www.javatpoint.com/machine-learning-techniques

https://blog.faradars.org/dbscan-algorithm-in-python

https://www. 16personalities.com/free-personality-test

https://raahbord.com/clustering-in-data-mining

https://www.geeksforgeeks.org/collaborative-filtering-ml/

https://blog.faradars.org/recommender-systems-in-python