

## عنوان: تحلیل الگوهای سلامت در بین معلمان یزد: کاربرد خوشه بندی داده کاوی برای بررسی شاخص های فیزیکی و بالینی

حامد فلاح تفتی<sup>۱</sup>، سعیده راستجو<sup>۲</sup>، بتول زیدآبادی<sup>۳</sup>، سمیه کارگر<sup>۴</sup>

### اطلاعات مربوط به مقاله

### چکیده

صنعت سلامت به طور مستقیم در حال تولید میزان زیادی از داده ها است و افرادی که با این نوع داده ها مواجه هستند دریافته اند که بین جمع آوری تا تفسیر آنها شکاف وسیعی وجود دارد. هدف اصلی مطالعه کاربرد خوشه بندی داده کاوی برای بررسی شاخص های فیزیکی و بالینی در بین معلمان یزد می باشد. این پژوهش از نوع توصیفی، اکتشافی و کاربردی می باشد. داده های مورد بررسی از اردیبهشت ۱۴۰۰ تا اردیبهشت ۱۴۰۲ از مراکز بهداشت و درمان و وزارت آموزش و پرورش استان یزد جمع آوری شد. در این مطالعه ۶۲۶ مورد از نتایج شاخص های سلامت فیزیکی و بالینی معلمان استان یزد، استفاده شد که این داده ها شامل ۲۹ مشخصه می باشند. تجزیه و تحلیل اطلاعات از روش های آمار توصیفی و روش داده کاوی می باشد و از زبان برنامه نویسی پایتون استفاده شد. نتایج پژوهش نشان داد که بهترین روش خوشه بندی روش KMeans با ۳ خوشه است که برای انتخاب بهترین روش خوشه بندی و تعداد خوشه ها به ترتیب از شاخص های silhouette و el bow استفاده شد. در نهایت ویژگی های هر یک از این سه خوشه مشخص گردید که با استفاده از آن می توان خوشه نمونه های جدید را براساس آن به دست آورد. این مطالعه نشان داد که روش های داده کاوی می تواند برای شناسایی الگوها و روندها در داده های مربوط به سلامت معلمان استفاده شود. این اطلاعات می تواند برای برنامه ریزی بهتر برنامه های سلامتی و پیشگیری از بیماری ها در معلمان مفید باشد.

کلید واژگان  
خوشه بندی،  
معلمان، شاخص  
های سلامت،  
پایتون، داده کاوی

<sup>۱</sup>. دانشیار گروه مدیریت، دانشکده علوم انسانی، دانشگاه علم و هنر، یزد، ایران

<sup>۲</sup>. دانشجوی دکتری تخصصی آموزش بهداشت و ارتقاء سلامت، گروه آموزش بهداشت و ارتقای سلامت، دانشکده بهداشت، دانشگاه علوم پزشکی شهید

صدققی، یزد، ایران

<sup>۳</sup>. نویسنده مسئول، دانشجوی دکتری تخصصی آموزش بهداشت و ارتقاء سلامت، دانشکده علوم پزشکی سیرجان، سیرجان، ایران

zeidabadi.b@gmail.com

<sup>۴</sup>. دانش آموخته کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه علم و هنر، یزد، ایران

## مقدمه

یکی از شاخصهای مهم توسعه یافتگی کشورها، سطح سلامت و تندرستی افراد جامعه است (۱). آینده‌ی هر جامعه مبتنی بر پویایی و سلامت روانی و جسمانی نوجوانان و جوانان آن جامعه است، به طوری که وجود افراد سالم و توانمند بزرگترین سرمایه‌ی ملی یک جامعه محسوب می‌شود (۲). هدف بخش بهداشت و درمان در هر جامعه‌ای، تأمین سلامت برای تمامی اعضای آن جامعه است. معلمان در کشور ما، به دلیل تشکیل دادن درصد بالایی از جمعیت، آسیب‌پذیری بیشتر و بالا بودن اثر بخشی مداخلات بهداشتی-درمانی در آن گروه به نسبت مداخلات، از گروههای ویژه و مورد توجه می‌باشند، بنابراین در اولویت پژوهشی و اجرایی قرار دارند (۳). بهبود کیفیت در صنعت سلامت را می‌توان به واسطه‌ی نیروهای محرکی که بر آن تأثیرگذار است بهتر تعریف نمود و از جمله‌ی این نیروهای محرک، داده‌های سلامت است؛ به عبارت دیگر در هر نوع برنامه‌ی بهبود کیفیت متمرکز بر بیمار، داده‌ها قلب آن برنامه به حساب می‌آید (۴). حوزه‌ی به نسبت جوان و در حال رشد داده‌کاوی در سلامت از جمله شیوه‌هایی است که می‌تواند این صنعت را از تحلیل عمیق این داده‌ها بهره‌مند سازد و به توسعه‌ی تحقیقات پزشکی و تصمیم‌گیری علمی در زمینه تشخیص و درمان منتج شود (۵، ۶). پایگاه داده‌ها در حوزه سلامت حاوی میزان وسیعی از داده‌های بالینی است که کشف ارتباطات و الگوها در آن می‌تواند به دانش جدید پزشکی منجر شود داده‌کاوی از جمله پیشرفت‌های فن آوری در راستای مدیریت داده‌ها است (۷). داده‌کاوی یک رویکرد کاربردی برای کشف الگوهای جدید و پنهان در داده‌ها می‌باشد. اطلاعات زیادی در سیستم بهداشت و درمان وجود دارد و تکنیکهای داده‌کاوی برای انواع برنامه‌های کاربردی در حوزه بهداشت و درمان نقش مؤثری دارد (۸، ۹).

مدرسه و معلمان جایگاه اساسی برای سازماندهی و ارائه خدمات سلامت روان در بین دانش‌آموزان هستند. مدارس و معلمان در ارتقای سلامت روان، شناسایی کودکان در معرض خطر اختلالات روان‌شناختی و ارجاع آنان برای دریافت کمک‌های تخصصی، نقش‌هایی اساسی دارند. با توجه به ارتباط تنگاتنگ معلمان با دانش‌آموزان و نقش آنها در ارتقای سلامت روانی، بررسی سلامت عمومی در این قشر جامعه ضروری بنظر می‌رسد. لذا هدف پژوهش حاضر خوسه بندی معلمان بر اساس شاخص‌های سلامت فیزیکی و بالینی با روش‌های داده‌کاوی است.

## روش پژوهش

مطالعه حاضر، پژوهش توصیفی، اکتشافی و کاربردی می‌باشد. داده‌های مورد بررسی از اردیبهشت ۱۴۰۰ تا اردیبهشت ۱۴۰۲ از مراکز بهداشت و درمان و وزارت آموزش و پرورش استان یزد جمع‌آوری شد. در این مطالعه ۶۲۶ مورد از نتایج شاخص‌های سلامت فیزیکی و بالینی معلمان استان یزد، استفاده شد. که این داده‌ها شامل ۲۹ شاخصه می‌باشند. شاخص‌های مورد بررسی عبارتند از: تاریخ تولد، سن، سال تولد، جنسیت، آخرین مدرک تحصیلی، نوع بیمه پایه درمان، وضعیت استفاده از بیمه مکمل، وضعیت تأهل، تعداد فرزندان دختر، تعداد فرزندان پسر، تعداد کل فرزندان، Platelet, MCH, MCV, Hematocrit, Hb, TSH, Creatinine, BUN, TG, LDL, Cholesterol, FBS, Lymph, WBC, RDW. تعداد مسافرت زیارتی خارج از کشور در تمام عمر و تعداد مسافرت غیرزیارتی (گردشگری) خارج از کشور می‌باشند.

اغلب به دلیل خطاهای عملیاتی و پیاده‌سازی سیستمها، داده‌های مغشوش و ناسازگار در بین داده‌های جمع‌آوری شده وجود دارد. پردازش اولیه‌ای مورد نیاز است تا مقادیر مفقوده، انحرافات و مقادیر ثبت نشده و مسائلی از این دست را در داده‌های

اولیه بیاید. این پیش پردازش جهت بهبود کیفیت داده های واقعی برای داده کاوی لازم است. جهت داده کاوی از نرم افزار پایتون استفاده می شود. اطلاعاتی که در این پژوهش مورد استفاده قرار گرفت، شامل شاخص های سلامت فیزیکی و بالینی معلمان استان یزد می باشد که، به منظور یکپارچه سازی داده ها، اطلاعات مربوط به شاخص های فیزیکی و روانی با یکدیگر ادغام شد. سپس اطلاعات نویزی و مقادیر تهی در مرحله پاکسازی از فایل ها حذف گردید. شاخص های سلامت فیزیکی و بالینی معلمان از مرکز مدیریت آمار و فناوری اطلاعات وزارت بهداشت اخذ شد. داده کاوی فرآیندی است که از طریق آن دانش مفیدی از حجم عظیمی از داده ها استخراج می شود. این فرآیند شامل مراحل زیر است:

انبارش داده ها: در این مرحله، اطلاعات لازم از منابع مختلف جمع آوری و در یک مخزن واحد ذخیره می شوند (۱۰). انتخاب داده ها: از میان داده های انبار شده، فقط داده هایی که برای کاوش مورد نیاز هستند انتخاب می شوند. پاکسازی و پیش پردازش: در این مرحله، داده ها از نویز و خطاها پاکسازی شده و برای تجزیه و تحلیل آماده می شوند. تبدیل داده ها: داده ها به فرمتی تبدیل می شوند که الگوریتم های داده کاوی می توانند آنها را پردازش کنند (۱۱).

کاوش در داده ها: در این مرحله، از الگوریتم های مختلف داده کاوی از جمله الگوریتم بخش پذیر، الگوریتم های مترکم سازی، دسته بندی بخشی و دسته بندی افزایشی برای استخراج دانش از داده ها استفاده می شود (۱۲). تفسیر نتایج: دانش استخراج شده از داده ها تفسیر و برای حل مسائل و تصمیم گیری استفاده می شود.

### یافته های پژوهش

اولین مرحله در اجرای پژوهش پیش رو به منظور خوشه بندی معلمان مدارس استان یزد بر اساس شاخص های سلامت فیزیکی و بالینی، تعیین شاخص های سلامت فیزیکی و بالینی است. با توجه به بررسی های انجام شده مهم ترین شاخص های سلامت فیزیکی و بالینی بدین شرح می باشد: تاریخ تولد، سن، سال تولد، جنسیت، آخرین مدرک تحصیلی، نوع بیمه پایه درمان، وضعیت استفاده از بیمه مکمل، وضعیت تاهل، تعداد فرزندان دختر، تعداد فرزندان پسر، تعداد کل فرزندان، Platelet.MCH، MCV.Hematocrit.Hb، TSH، Creatinine، BUN، TG، LDL، Cholesterol، FBS، Lymph.WBC.RDW، از کشور در تمام عمر و تعداد مسافرت غیرزیارتی (گردشگری) خارج از کشور می باشند.

### تشکیل بانک داده

اولین گام پس از تعیین شاخص های سلامت فیزیکی و بالینی تهیه یک بانک اطلاعاتی است. برای این منظور، با همکاری مرکز مدیریت آمار و فناوری اطلاعات وزارت بهداشت، داده های مربوط به خوشه بندی معلمان بر اساس شاخص های سلامت فیزیکی و بالینی در قالب دو فایل اکسل تهیه و در اختیار محقق قرار گرفت.

### پیش پردازش و آماده سازی داده ها

در این مرحله، داده های مورد استفاده در پژوهش در قالب دو فایل اکسل از مرکز مدیریت آمار و فناوری اطلاعات وزارت بهداشت دریافت شد که در ادامه به منظور یکپارچه سازی داده ها، اطلاعات موجود در دو فایل را با یکدیگر ادغام کرده و سپس جهت حذف اطلاعات نویزی و مقادیر تهی فایل داده ها مورد بررسی و پاکسازی قرار گرفت. مراحل آماده سازی داده ها در ادامه به صورت مبسوط تشریح می گردد. در

این مرحله جهت پیش‌پردازش داده‌ها ابتدا لازم است تا با کمک کد [1] In فایل داده‌ها در نرم‌افزار jupyter notebook فراخوانی شود.

در ادامه فرآیند پیش‌پردازش داده‌ها، ابتدا اطلاعات و داده‌های مربوط به کلیه شاخص‌ها با کمک کد [2] In مورد بررسی قرار می‌گیرند تا وضعیت و نوع داده‌ها و همچنین داده‌های خراب و نامناسب که منجر به بروز خطا در تحلیل‌ها می‌شوند، شناسایی شوند. نتایج در ادامه کد قبل قابل مشاهده است.

با توجه به بررسی داده‌ها و جدول فوق می‌توان دریافت که داده‌های مربوط به شاخص‌ها به سه دسته اعداد صحیح، اعداد اعشاری و object تقسیم می‌شوند همچنین با توجه به این جدول مشخص می‌شود که داده‌های مربوط به کلیه شاخص‌ها مناسب و کامل هستند به عبارتی هیچ گونه عدد صفر، خط تیره و ... در بین داده‌ها وجود ندارد. در نهایت در این مرحله با بررسی داده‌ها کیفیت آن‌ها جهت انجام فرآیند داده‌کاوی مورد تأیید قرار گرفت یعنی داده‌های مربوط به کلیه شاخص‌ها کامل و مناسب می‌باشند.

با توجه به بررسی شاخص‌ها می‌توان دریافت که شاخص‌ها به دو دسته شاخص‌های کیفی و کمی تقسیم می‌شوند که در کدنویسی شاخص‌های کمی را با unique و شاخص‌های کیفی را value\_counts نشان می‌دهیم.

در شاخص unique کلیه داده‌های منحصر به فرد را و در شاخص‌های value\_counts وضعیت حالت‌های مختلف شاخص کیفی مشخص می‌شود. در این مرحله بررسی می‌شود که آیا شاخص‌های unique به غیر از اعداد مقادیر دیگری و شاخص‌های value\_counts به جز حالت‌های مد نظر آیا مقدار دیگری نیز به خود می‌گیرند.

شاخص سن یک شاخص unique می‌باشد. در این نوع شاخص‌ها، داده‌ها بدون تکرار و به‌صورت منحصر به فرد به نمایش گذاشته می‌شوند یعنی اگر مقدار شاخص مربوطه در بیش از یک سطر برابر با یک عدد برابر باشد در جدول خروجی فقط یک بار تکرار می‌شوند به عبارتی اگر در یک ستون یک یا چند عدد بیش از یک بار تکرار شده باشند در جدول خروجی عدد مربوطه فقط یک بار به نمایش می‌گردد. همچنین لازم به ذکر است که داده‌های مربوط به این شاخص به‌صورت عدد صحیح می‌باشند. همانطور که از خروجی مشخص است داده‌های مربوط به این شاخص در مقدار ۱ نامناسب می‌باشد بنابراین جهت انجام تحلیل‌ها و خوشه‌بندی لازم است تا این داده‌ها از بانک داده‌ها حذف گردند همچنین لازم به ذکر است که تعداد ۱ها برای شاخص سن در بانک داده‌ها برابر با ۴ تا می‌باشد.

جنسیت یک شاخص value\_counts می‌باشد. با توجه به خروجی مربوط به این شاخص مشخص می‌شود که از ۶۲۶ نفر، جنسیت ۵۲۵ نفر زن و ۱۰۱ نفر مرد می‌باشد همچنین مشخص می‌شود کلیه داده‌های مربوط به این شاخص مناسب و درست می‌باشند. بنابراین هیچ داده خراب و ناقصی که تحلیل‌ها را با خطا مواجه کنند در این شاخص وجود ندارد.

وضعیت مدرک تحصیلی نشان می‌دهد. ۳۹۵ نفر لیسانس، ۲۱۳ نفر فوق‌لیسانس، ۹ نفر فوق‌دپلم، ۶ نفر دکتری و ۳ نفر دپلم دارند. تمام داده‌ها معتبر و بدون خطا هستند و به صورت عدد صحیح هستند.

شاخص نوع بیمه پایه درمان یک شاخص value\_counts می‌باشد. با توجه به خروجی مربوط به این شاخص مشخص می‌شود که از ۶۲۶ نفر، ۵۲۶ نفر از بیمه تأمین اجتماعی، ۴۲ نفر از بیمه سلامت، ۴۶ نفر از سایر بیمه‌ها استفاده می‌کنند و ۱۲ نفر نیز از هیچ بیمه درمانی استفاده نمی‌کنند. همچنین مشخص می‌شود کلیه داده‌های مربوط به این شاخص مناسب و درست و به‌صورت عدد صحیح می‌باشند. شاخص نوع بیمه مکمل یک شاخص value\_counts می‌باشد که داده‌های مربوط به این شاخص مطابق جدول پیوست است. با توجه به خروجی مربوط به این شاخص مشخص می‌شود که از ۶۲۶ نفر، ۲۵۲ نفر از بیمه مکمل استفاده می‌کنند و ۳۷۴ نفر

باقی‌مانده از هیچ بیمه مکملی استفاده نمی‌کنند. همچنین مشخص می‌شود کلیه داده‌های مربوط به این شاخص مناسب و درست و به‌صورت عدد صحیح می‌باشند.

شاخص وضعیت تأهل یک شاخص `value_counts` می‌باشد. با توجه به خروجی مربوط به این شاخص مشخص می‌شود که از ۶۲۶ نفر، ۵۵۳ نفر متأهل، ۶۱ نفر مجرد (ازدواج نکرده)، ۱۰ نفر مطلقه و نفر بیوه (فوت همسر) می‌باشند. همچنین مشخص می‌شود کلیه داده‌های مربوط به این شاخص مناسب و درست می‌باشند بنابراین هیچ داده خراب و ناقصی که تحلیل‌ها را با خطا مواجه کنند در این شاخص وجود ندارد. همچنین لازم به ذکر است که داده‌های مربوط به این شاخص نیز به‌صورت عدد صحیح می‌باشند.

در خصوص تعداد فرزندان نتایج نشان داد که (۴۲،۴٪) دو فرزند، ۱۳۸ نفر (۲۲،۰٪) یک فرزند، ۹۰ نفر (۱۴،۴٪) سه فرزند، ۱۴ نفر (۲،۲٪) چهار فرزند، ۱ نفر (۰،۲٪) پنج فرزند و ۱ نفر (۰،۲٪) شش فرزند داشتند.

۱۱۷ نفر (۱۸،۷٪) هیچ فرزندی نداشتند. برای داده‌های مربوط به این شاخص و ۱۵ شاخص خونی هیچ داده گمشده یا نامعتبری یافت نشد و همه داده‌های مربوط به شاخص‌های خونی به صورت عدد صحیح بودند. بررسی داده‌های مربوط به سفرهای زیارتی و غیرزیارتی نشان داد که هیچ داده گمشده یا نامعتبر وجود ندارد. لذا نتایج این مطالعه نشان داد که داده‌های مربوط به تمامی شاخص‌ها از کیفیت بالایی برخوردار است.

جدول ۱: فراوانی و درصد شاخص‌های سلامت فیزیکی و بالینی معلمان

|       | سن        | سال تولد    | تعداد فرزند پسر | تعداد فرزند دختر | تعداد کل فرزندان | Hb         | Hematocrit | MCV        | MCH        |
|-------|-----------|-------------|-----------------|------------------|------------------|------------|------------|------------|------------|
| count | 626.0000  | 626.000000  | 626.000000      | 626.000000       | 626.000000       | 626.000000 | 626.000000 | 626.000000 | 626.000000 |
| mean  | 37.924920 | 1363.075080 | 0.873802        | 0.731629         | 1.605431         | 13.382268  | 40.552396  | 84.329233  | 27.8619    |
| std   | 6.831161  | 6.831161    | 0.834295        | 0.759909         | 1.042721         | 1.567711   | 3.459989   | 5.779804   | 2.7122     |
| min   | 1.000000  | 1342.000000 | 0.000000        | 0.000000         | 0.000000         | 8.900000   | 29.500000  | 32.600000  | 18.4000    |
| 25 %  | 34.000000 | 1359.000000 | 0.000000        | 0.000000         | 1.000000         | 12.500000  | 38.400000  | 82.000000  | 26.7000    |
| 50 %  | 39.000000 | 1362.000000 | 1.000000        | 1.000000         | 2.000000         | 13.450000  | 40.800000  | 85.200000  | 28.3000    |
| 75 %  | 42.000000 | 1367.000000 | 1.000000        | 1.000000         | 2.000000         | 14.300000  | 42.600000  | 87.900000  | 29.7000    |
| max   | 59.000000 | 1400.000000 | 4.000000        | 3.000000         | 6.000000         | 18.100000  | 50.500000  | 97.200000  | 39.2000    |

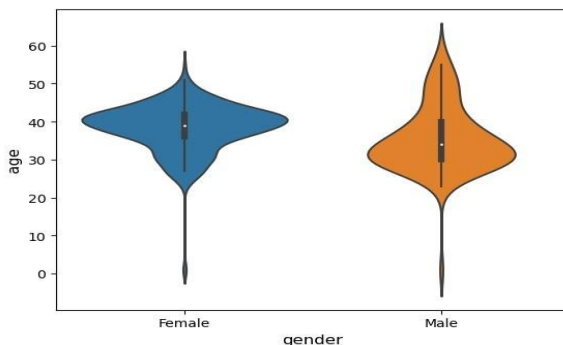
بر اساس نمودار رابطه بین دو شاخص سن و جنسیت افراد مشخص می‌گردد که در بین زنان سن ۴۰ سال بیشترین فراوانی و در بین مردان سن ۳۰ سال بیشتر

| تعداد<br>مسافرت<br>گردشگ<br>ری | تعداد<br>مسافرت<br>زیارتی | TSH          | Creati<br>nine | BUN           | TG             | LDL           | HDL           | Cholest<br>erol | FBS           | Platelet       |           |
|--------------------------------|---------------------------|--------------|----------------|---------------|----------------|---------------|---------------|-----------------|---------------|----------------|-----------|
| 626                            | 626                       | 626          | 626            | 626           | 626            | 626           | 626           | 626             | 626           | 626            | cou<br>nt |
| 0.265<br>176                   | 0.899<br>361              | 2.706<br>535 | 0.8668<br>37   | 25.952<br>077 | 136.306<br>709 | 96.246<br>006 | 51.861<br>022 | 174.573<br>482  | 93.958<br>466 | 269.257<br>188 | me<br>an  |
| 1.330<br>252                   | 1.655<br>372              | 1.973<br>54  | 0.0965<br>74   | 6.1515<br>61  | 57.4272<br>84  | 33.521<br>005 | 10.052<br>097 | 32.5247<br>26   | 17.697<br>499 | 69.2538<br>01  | std       |
| 0                              | 0                         | 0.1          | 0.64           | 10            | 40             | 9             | 30            | 100             | 70            | 120            | mi<br>n   |
| 0                              | 0                         | 1.592<br>5   | 0.8            | 22            | 100            | 72            | 46            | 151             | 85            | 223            | 25<br>%   |
| 0                              | 0                         | 2.295        | 0.8            | 25            | 120            | 92            | 51            | 170             | 91            | 261            | 50<br>%   |
| 0                              | 1                         | 3.2          | 0.9            | 30            | 153.75         | 114           | 58            | 192             | 98            | 309            | 75<br>%   |
| 20                             | 20                        | 22.47        | 1.3            | 44            | 770            | 303           | 222           | 289             | 276           | 617            | ma<br>x   |

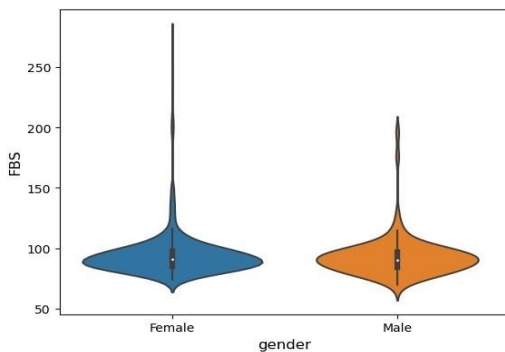
فراوانی را دارد. بنابراین می‌توان نتیجه گرفت که بیشترین میزان فراوانی سن در بین زنان ۱۰ سال بیشتر از بیشترین میزان فراوانی سن

در بین مردان است بنابراین میانگین سنی مردان باید کمتر

از مقدار میانگین سنی زنان باشد. همچنین رابطه بین دو شاخص جنسیت و قند خون ناشتا افراد نشان داد که بیشترین فراوانی در پارامتر قند خون ناشتا در زنان و مردان تقریباً ۹۰ می‌باشد. بنابراین بر این اساس می‌توان دریافت که تفاوت چندانی بین میزان فراوانی قند خون ناشتا در گروه زنان و مردان وجود ندارد.

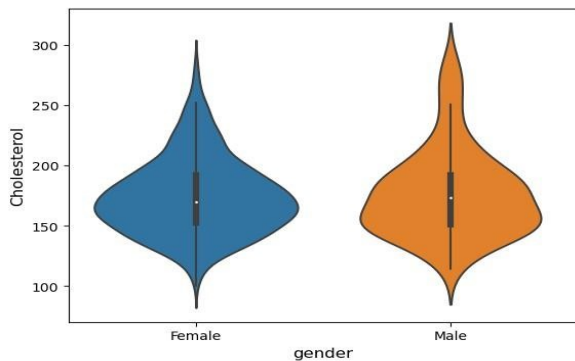


شکل ۱- نمودار رابطه بین دو شاخص جنسیت و سن

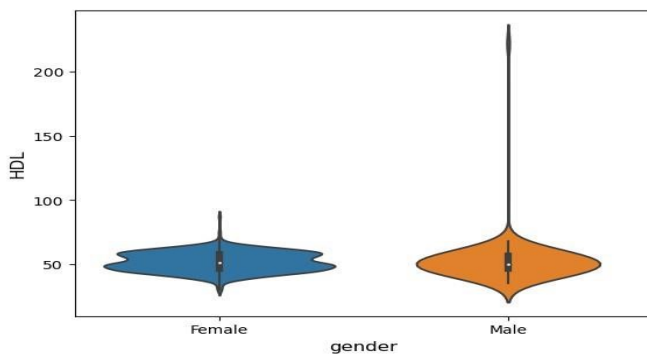


شکل ۲- نمودار رابطه بین دو شاخص جنسیت و قندخون

همچنین رابطه بین دو شاخص جنسیت و کلسترول نشان داد که بیشترین میزان فراوانی در پارامتر کلسترول افراد بین ۱۵۰ تا ۲۰۰ می‌باشد. بنابراین بر این اساس می‌توان بیان کرد که تفاوت چندانی بین میزان قند خون ناشتا بر حسب بیشترین فراوانی در دو گروه زنان و مردان وجود ندارد.

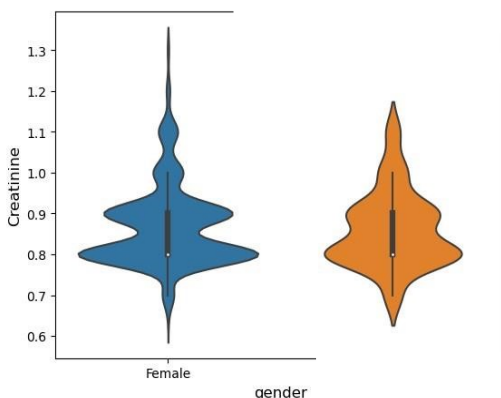


شکل ۳- نمودار رابطه بین دو شاخص جنسیت و کلسترول افراد



شکل ۴- نمودار رابطه بین دو شاخص جنسیت و HDL

در ادامه رابطه بین دو شاخص جنسیت و HDL افراد نیز مبین آن است که بیشترین فراوانی در پارامتر HDL در بین مردان ۵۰ و در بین زنان بیشترین فراوانی مربوط به مقادیر ۵۰ و ۶۰ می‌باشد. بنابراین می‌توان نتیجه گرفت بیشترین فراوانی در شاخص HDL در مقدار ۵۰ بین زنان و مردان مشترک می‌باشد اما در مردان علاوه بر مقدار ۵۰ در مقدار ۶۰ نیز بیشترین فراوانی وجود دارد. و در آخر رابطه بین دو شاخص جنسیت و کراتینین خون افراد مورد بررسی قرار گرفت که بر اساس نمودار رابطه بین این دو شاخص مشخص می‌گردد که بیشترین فراوانی در پارامتر کراتینین خون افراد در بین زنان به ترتیب مربوط به مقادیر ۰٫۸ و ۰٫۹ و در بین مردان نیز بیشترین فراوانی مربوط به مقدار ۰٫۸ می‌باشد. بنابراین می‌توان نتیجه گرفت بیشترین فراوانی در مقدار ۰٫۸ در زنان و مردان مشترک است اما در بین زنان علاوه بر بیشترین فراوانی در مقدار ۰٫۸ در مقدار ۰٫۹ این پارامتر نیز بیشترین فراوانی وجود دارد.



شکل ۵- نمودار رابطه بین دو شاخص جنسیت و کراتینین افراد

| z_سن     | z_Hb:    | z_Hematocrit: | z_MCV:   | z_MCH:   | z_Platelet: | z_RDW:   | z_WBC:   | z_TSH:   |
|----------|----------|---------------|----------|----------|-------------|----------|----------|----------|
| -1.32191 | 1.540618 | 1.25576       | 0.115984 | 0.751184 | -1.11845    | -0.72736 | 0.626119 | -0.30429 |
| -1.15996 | 0.137689 | -0.27731      | 1.04997  | 1.193748 | -0.80123    | -0.64907 | 1.401267 | 0.228105 |
| 1.755174 | -1.07393 | -1.02938      | 0.202464 | -0.31835 | -0.15237    | -0.41421 | -1.10306 | -0.10148 |
| -0.67411 | -0.3087  | 0.011952      | -1.11204 | -1.12971 | -0.61378    | 0.055505 | 0.745373 | 2.225871 |
| 0.297606 | 1.221771 | 1.892127      | -0.00509 | -0.17082 | -0.90217    | -0.41421 | 0.685746 | 0.557686 |
| ...      | ...      | ...           | ...      | ...      | ...         | ...      | ...      | ...      |
| -0.83606 | 0.392767 | 0.416913      | 0.115984 | 0.161099 | -0.19563    | -1.11879 | 1.34164  | -0.4057  |

|          |          |          |          |          |          |          |          |          |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1.269318 | 0.073919 | -0.27731 | 0.115984 | 0.419262 | -0.88775 | -0.17935 | -0.56642 | -0.13697 |
| -0.18825 | -1.26524 | -1.23186 | 0.21976  | -0.35523 | -0.31098 | -0.02278 | -1.34156 | 0.760505 |
| 1.269318 | 0.520306 | 0.50369  | 0.444609 | 0.493022 | 0.395551 | -0.25764 | 0.506866 | -0.36007 |
| 0.135654 | -0.11739 | 0.011952 | 0.807826 | 0.382381 | 1.664429 | 0.055505 | 1.759027 | -0.56796 |

پس از تشکیل بانک داده در مرحله قبل و آماده‌سازی و پیش‌پردازش داده‌ها و همچنین بررسی میزان فراوانی و رابطه بین شاخص‌های متفاوت در این مرحله به اجرای فرآیند خوشه‌بندی پرداخته می‌شود. برای این منظور از نرم افزار jupyter notebook استفاده شده است. قبل از شروع فرآیند خوشه‌بندی ابتدا با بررسی شاخص‌ها، داده‌های شاخص‌هایی نظیر تاریخ تولد، سن تولد، آخرین مدرک تحصیلی، نوع بیمه پایه درمان، وضعیت بیمه مکمل، وضعیت تاهل، تعداد فرزندان دختر، تعداد فرزندان پسر و کل فرزندان، تعداد مسافرت زیارتی و تعداد مسافرت گردشگری که تأثیری روی فرآیند خوشه‌بندی ندارد را از بانک داده جدا کرده و فرآیند خوشه‌بندی با کمک داده‌های سایر شاخص انجام می‌دهیم. زیرا این شاخص‌ها هیچ تأثیری روی سلامت فیزیکی و بالینی افراد ندارند و همچنین به عنوان مثال لازم به ذکر است زمانی که سن افراد را داشته باشیم دیگر به تاریخ و سال تولد افراد نیازی نیست. سپس از آنجایی که مقیاس داده‌ها با یکدیگر متفاوت است ابتدا لازم است تا داده‌ها بی‌مقیاس شوند. با استفاده از کد [37] داده‌های مربوط به شاخص‌های باقی‌مانده بی‌مقیاس می‌شوند.

همچنین، از آنجایی که متغیر جنسیت یک متغیر کیفی است داده‌های مربوط به این شاخص قابل بی‌مقیاس‌سازی نمی‌باشند. بنابراین داده‌های این شاخص (عدد ۰ و ۱ به ترتیب بیانگر جنسیت زن و مرد می‌باشند) از مرحله بی‌مقیاس‌سازی جدا شده و پس انجام بی‌مقیاس‌سازی روی سایر داده‌ها، داده جنسیت به این پایگاه داده اضافه می‌شود؛ پایگاه داده نهایی در جدول ۲ قابل مشاهده است. مقادیر بی‌مقیاس‌شده داده‌ها با استفاده از رابطه زیر محاسبه می‌شود.

$$\tilde{x}_j = \frac{X - \bar{X}}{\delta}$$

جدول ۲: پایگاه داده نهایی

| z_Lymph: | z_FBS:   | z_Cholesterol: | z_HDL:   | z_LDL:   | z_TG:    | z_BUN:   | z_Creatinine: | جنسیت |
|----------|----------|----------------|----------|----------|----------|----------|---------------|-------|
| 0.012243 | 0.282697 | 0.898941       | -0.58162 | 0.852131 | 0.464414 | 1.626909 | 1.374641      | 0     |
| -0.84248 | -0.05545 | -0.69869       | -0.28403 | -0.9667  | 1.038077 | -0.48197 | -0.69164      | 0     |
| -0.42169 | -0.50632 | -0.23784       | -0.08564 | -0.19146 | -0.09187 | 0.004695 | 0.341501      | 0     |
| 0.209485 | 1.917085 | -0.26856       | -1.67277 | -0.13183 | 1.05546  | 0.004695 | -0.69164      | 1     |
| -0.84248 | 0.057264 | 0.345914       | -1.17679 | 0.643413 | 0.081972 | 2.600238 | 1.374641      | 1     |
| ...      | ...      | ...            | ...      | ...      | ...      | ...      | ...           | ...   |

|          |          |          |          |          |          |          |          |   |
|----------|----------|----------|----------|----------|----------|----------|----------|---|
| -0.2902  | -0.67539 | -0.0535  | -0.28403 | -0.31073 | 0.951158 | 1.789131 | 0.341501 | 0 |
| 1.353491 | -0.78811 | -0.39146 | 0.707924 | -0.40018 | -0.66553 | -0.80641 | -0.69164 | 0 |
| 1.787424 | -0.84447 | 0.038676 | 1.104706 | -0.07219 | -0.71768 | 0.65358  | -0.69164 | 0 |
| -0.67153 | -0.33724 | 0.161571 | 0.80712  | 0.106708 | -0.63076 | 1.140245 | -0.69164 | 0 |
| -0.97397 | 2.480667 | 1.175455 | -0.58162 | 1.388835 | -0.2657  | -0.64419 | -0.69164 | 0 |

622 rows  $\times$  18 columns

در ادامه داده‌های train و test بر اساس جدول ۲ مشخص می‌شوند. در جدول ۳ قسمتی از داده‌های train قابل مشاهده است.

جدول ۳: داده‌های train

|     | Z_age     | Z_Hb      | z_Hematocrit | z_MCV     | z_MCH     | z_Platel et | z_RDW     | z_WBC     | z_Lymph   | z_FBS    |
|-----|-----------|-----------|--------------|-----------|-----------|-------------|-----------|-----------|-----------|----------|
| 550 | -1.483865 | -2.349322 | -2.070705    | -1.890358 | -2.420523 | 1.794200    | 2.560660  | -0.089402 | 0.077990  | 0.90263  |
| 222 | 2.079077  | 0.137689  | 0.156581     | 0.807826  | 0.566783  | -1.868243   | -0.179353 | -0.506789 | 0.091140  | -0.28088 |
| 320 | -1.807769 | -2.285553 | -2.388888    | 0.358128  | -0.613387 | 0.222522    | 0.838366  | 1.639774  | -1.828687 | -1.06990 |
| 559 | 0.783462  | -1.073932 | -1.116154    | 0.980786  | 0.308621  | 0.626256    | -0.414211 | 1.639774  | -1.052866 | 0.11362  |
| 27  | -0.350202 | -0.818854 | -0.682267    | -0.005089 | -0.428986 | 1.736524    | 0.525222  | -0.626042 | -0.789877 | 0.00090  |

سپس داده‌های test فراخوانی می‌شوند که در جدول ۴ قسمتی از داده‌های test قابل مشاهده است.

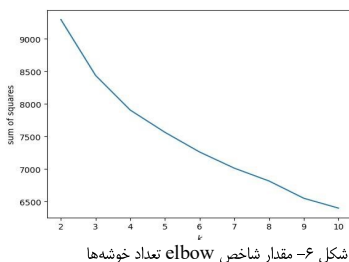
جدول ۴: قسمتی از داده‌های test

|     | Z_age     | Z_Hb      | z_Hematocrit | z_MCV    | z_MCH    | z_Platel et | z_RDW     | z_WBC     | z_Lymph   | z_FBS    |
|-----|-----------|-----------|--------------|----------|----------|-------------|-----------|-----------|-----------|----------|
| 392 | 1.107366  | 0.073919  | 0.214432     | 0.617569 | 0.308621 | 0.136008    | -0.414211 | 0.685746  | -0.526887 | -0.5063  |
| 200 | -1.159961 | -0.372467 | -0.566564    | 1.015378 | 0.751184 | -1.161709   | -0.179353 | -0.506789 | 0.617119  | -0.28088 |

|         |                   |                   |           |                   |                   |                   |                   |              |                   |                  |
|---------|-------------------|-------------------|-----------|-------------------|-------------------|-------------------|-------------------|--------------|-------------------|------------------|
| 39<br>6 | 0.62151<br>0      | 0.20145<br>8      | 0.301209  | -<br>0.22993<br>7 | -<br>0.20770<br>4 | 1.79420<br>0      | 0.21207<br>7      | 1.34164<br>0 | -<br>0.31649<br>5 | 0.0572<br>6      |
| 75      | -<br>1.15996<br>1 | -<br>1.45654<br>9 | -1.174005 | 0.13328<br>0      | -<br>0.72402<br>8 | -<br>0.54168<br>9 | -<br>0.41421<br>1 | 0.56649<br>2 | -<br>0.77672<br>7 | -<br>0.5063      |
| 25<br>9 | 0.62151<br>0      | 0.52030<br>6      | 0.214432  | 0.44460<br>9      | 0.75118<br>4      | 0.04949<br>3      | -<br>0.33592<br>5 | 0.44723<br>9 | -<br>1.02656<br>8 | -<br>0.2808<br>8 |

در ادامه جهت خوشه‌بندی افراد با استفاده از روش K-means ابتدا لازم است تا تعداد خوشه‌های بهینه (K) محاسبه شود. در پژوهش حاضر به منظور تعیین تعداد بهینه خوشه‌ها و مناسب‌ترین روش خوشه‌بندی به ترتیب، از شاخص *silhouette* و *elbow* طبق کد [41] و [42] استفاده شده است. مقدار شاخص *elbow* برای تعداد ۲ تا ۱۰ خوشه محاسبه شده است در هر حالت که در شکل ۷ شیب نمودار حالت کاهشی پیدا کند مقدار *k* در آن نقطه مناسب‌ترین است.

امتیاز *silhouette* معیاری از میزان تناسب خوشه‌بندی با داده‌ها است. به طور معمول، امتیاز برای هر داده به طور جداگانه محاسبه می‌شود و میانگین به عنوان معیاری برای ارزیابی تطابق مدل به طور کلی با کل مجموعه داده مورد استفاده قرار می‌گیرد. دو مؤلفه اصلی برای این امتیاز وجود دارد؛ مؤلفه اول میزان تناسب داده با خوشه‌ای که به آن اختصاص داده شده است که به عنوان فاصله متوسط بین آن و سایر اعضای همان خوشه تعریف می‌شود. مؤلفه دوم میزان تناسب داده با نزدیکترین خوشه بعدی را اندازه گیری می‌کند. به همین ترتیب با اندازه گیری فاصله متوسط بین داده و تمام داده‌های اختصاص داده شده به نزدیکترین خوشه بعدی محاسبه می‌شود. تفاوت بین این دو عدد را می‌توان به عنوان معیاری برای ارزیابی داده در خوشه اختصاص داده شده در مقابل خوشه‌های مختلف در نظر گرفت. به عبارت بهتر، نقطه داده چقدر در خوشه اختصاص داده شده تناسب دارد. امتیاز *silhouette* عددی بین ۱- و ۱ است. نمره منفی به این معنی است که این داده در واقع به طور متوسط به خوشه دیگر نزدیکتر است، در حالی که نمره مثبت به این معنی است که با خوشه اختصاص داده شده تناسب بسیار بهتری دارد. باید توجه داشت که امتیاز *silhouette* معیاری کلی است که نشان می‌دهد خوشه‌بندی چقدر با داده‌ها مطابقت دارد. بنابراین طبق خروجی شکل ۶ تعداد بهینه خوشه برابر ۳ می‌شود. ( $K=3$ ).



در ادامه به دنبال تعیین بهترین روش خوشه‌بندی مقدار شاخص *silhouette* را برای سه روش خوشه‌بندی Mean-*K*-means و shift و K-modes محاسبه کرده که مقادیر این شاخص برای سه روش خوشه‌بندی فوق به شرح جدول ۵ می‌باشد.

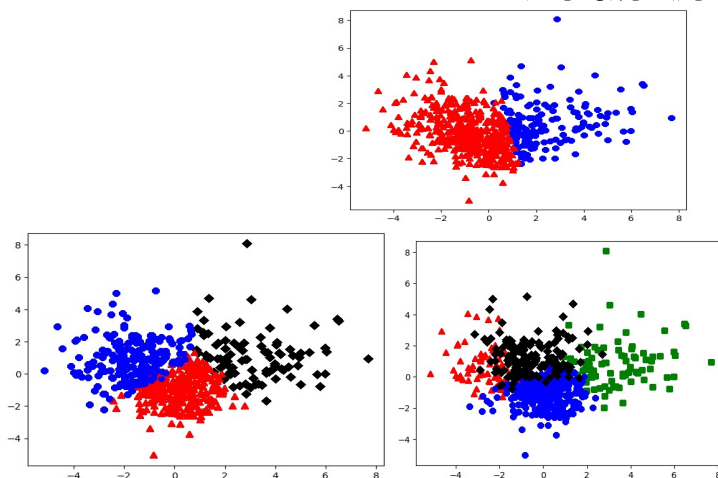
جدول ۵: مقادیر شاخص silhouette روش‌های خوشه‌بندی K-means، Mean-shift و K-modes

| روش خوشه‌بندی         | K-means | Mean-shift | K-modes |
|-----------------------|---------|------------|---------|
| مقدار شاخص silhouette | ۰,۱۴۰   | ۱۳۶        | ۰,۴۰    |

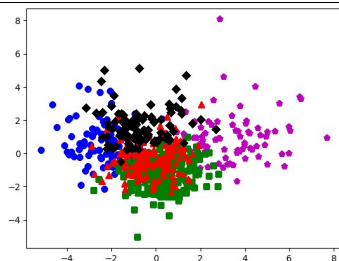
با توجه به مقادیر شاخص silhouette در جدول فوق می‌توان نتیجه گرفت که روش خوشه‌بندی K-means با توجه به اینکه بیشترین مقدار را در شاخص silhouette دارا می‌باشد مناسب‌ترین روش جهت خوشه‌بندی افراد می‌باشد بنابراین در ادامه جهت خوشه‌بندی افراد از روش خوشه‌بندی K-means استفاده شد.

در این مرحله از پژوهش با روش خوشه‌بندی (K-means) بر اساس داده‌های بی‌مقیاس شده، خوشه بندی در خوشه‌های دوتایی، سه تایی، چهارتایی و پنج تایی برای بررسی نحوه خوشه‌بندی در حالت‌های مختلف انجام شد.

سپس، در این مرحله از پژوهش جهت نمایش توزیع خوشه‌بندی داده‌ها بر روی نمودار دو بعدی لازم است تا تعداد شاخص‌ها کاهش یابد. در تحلیل آماری استنباطی مفهومی به نام تحلیل مولفه اصلی یا تحلیل عاملی (PCA) وجود دارد. این تحلیل یک رویکرد چندگانه است که به منظور کاهش ابعاد یک مجموعه داده و در عین حال حفظ تا حد امکان اطلاعات از داده‌ها استفاده می‌شود. بنابراین در پژوهش حاضر نیز با کمک از رویکرد تحلیل عاملی یا تحلیلی مولفه اصلی<sup>۱</sup> (PCA) اقدام به کاهش تعداد ابعاد نموده و ابعاد را به دو بعد PC1 و PC2 کاهش می‌دهیم. سپس با استفاده از دو بعد شناسایی شده نمایش خوشه‌بندی در خوشه دوتایی، سه تایی، چهارتایی و پنج تایی انجام شد.

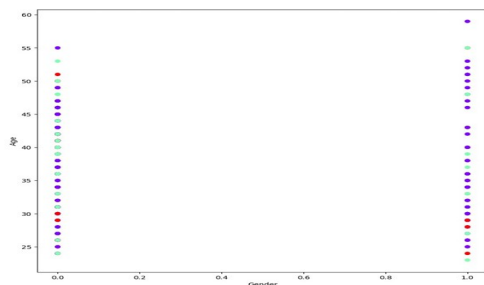


<sup>۱</sup>. Principal Component Analysis



شکل ۷: توزیع داده‌ها در خوشه‌های دوتایی، سه تایی، چهارتایی و پنج تایی بر اساس روش PCA

همانطور که شاخص elbow و شکل‌های خوشه‌بندی نشان داد، خوشه‌بندی سه تایی از نظر این شاخص و هم از نظر نمایشی به دلیل ایجاد مرز مناسب بین خوشه‌ها بهترین تعداد خوشه می‌باشد. در ادامه در خوشه‌بندی در سه خوشه با توجه به دو شاخص سن و جنسیت، خوشه فرد (نمونه) جدید را می‌توان پیش‌بینی کرد در نمودار زیر رنگ بنفش، قرمز و سبز به ترتیب بیانگر خوشه ۱، ۲ و ۳ می‌باشد. بر اساس نمودار زیر به عنوان مثال اگر نمونه جدید یک فرد با جنسیت زن و سن ۴۰ سال باشد این نمونه در خوشه ۲ قرار می‌گیرد به عبارت بهتر خصوصیات این نمونه به خصوصیات افراد خوشه ۲ نزدیکتر می‌باشد. همچنین اگر همین یک نمونه جدید با جنسیت مرد و سن ۴۰ سال باشد در خوشه ۳ قرار می‌گیرد. در نهایت، همانطور که در قبل بیان شد می‌توان براساس خوشه‌بندی انجام شده خوشه نمونه جدید را یافت و ویژگی‌های نمونه‌ی جدید را براساس خوشه اختصاص داده شده، پیش‌بینی کرد. از همین رو، براساس ویژگی‌های هر یک از سه خوشه را فراخوانی نموده که بر همین اساس می‌توان مقادیر مرتبط با نمونه جدید را پیش‌بینی کرد.



شکل ۸: خوشه‌بندی براساس جنسیت و سن

### بحث و نتیجه‌گیری

پژوهش حاضر با هدف خوشه‌بندی معلمان مدارس استان یزد بر اساس شاخص‌های سلامت فیزیکی و بالینی انجام شده‌است. پس از تشکیل بانک داده و آماده‌سازی و پیش‌پردازش داده‌ها، به اجرای فرآیند خوشه‌بندی پرداخته شد. قبل از شروع فرآیند خوشه‌بندی ابتدا با بررسی

شاخص‌ها، آن دسته از شاخص‌هایی که تأثیری روی فرآیند خوشه‌بندی ندارد را از بانک داده جدا کرده و فرآیند خوشه‌بندی با کمک داده‌های سایر شاخص انجام شد. پس از جداسازی شاخص‌های بدون تأثیر بر فرآیند خوشه‌بندی، داده‌های مربوط به شاخص‌های باقی‌مانده بی‌مقیاس شدند. در ادامه جهت خوشه‌بندی افراد با استفاده از روش‌های خوشه‌بندی تعداد خوشه‌های بهینه (K) محاسبه شد. در پژوهش حاضر به منظور تعیین تعداد بهینه خوشه‌ها، از شاخص silhouette استفاده شد و مقدار این شاخص برای تعداد ۲ تا ۱۰ خوشه محاسبه شده است در هر حالت که میانگین شاخص silhouette بیشتر باشد به عنوان تعداد خوشه بهینه انتخاب گردید. پس از محاسبه این شاخص برای K برابر با ۲ تا ۱۰، تعداد بهینه خوشه برابر با ۳ حاصل شد.

در ادامه با محاسبه مقادیر شاخص silhouette برای سه روش خوشه‌بندی K-means، Mean-shift، K-modes بهترین روش خوشه‌بندی تعیین گردید. که با توجه به مقادیر شاخص silhouette می‌توان نتیجه گرفت که روش خوشه‌بندی K-means با توجه به اینکه بیشترین مقدار را در شاخص silhouette دارا می‌باشد مناسب‌ترین روش جهت خوشه‌بندی افراد می‌باشد بنابراین جهت خوشه‌بندی افراد از روش خوشه‌بندی K-means استفاده شد. همچنین در این پژوهش با توجه به دو شاخص سن و جنسیت، خوشه فرد (نمونه) جدید را نیز می‌توان پیش‌بینی کرد.

در همین راستا، در گام اول خوشه بندی، نیاز است که پیش پردازش داده ها صورت گیرد؛ هدف از انجام این گام آن است که داده های دارای خطا و نامناسب شناسایی شوند و از سایر داده ها جدا شوند که در مرحله خوشه بندی تأثیرگذار نباشند. بطوریکه، پس انجام پیش پردازش نتایج نشان داد که در داده های مربوط به سن، سن یک سال وجود دارد که این مقدار برای این ویژگی جز داده های خطا محسوب می شود. بنابراین، در ویژگی سن اطلاعات مربوط به داده های با مقدار یک حذف گردید و سایر ویژگی ها دارای داده های مناسب و بدون خطا بودند. پس آن برای بررسی نحوه توزیع داده ها در هر یک از ویژگی ها نمودار متناسب با هر ویژگی رسم گردید و مورد بررسی قرار گرفتند. پس انجام گام پیش پردازش، در ابتدا با استفاده از شاخص silhouette مناسب ترین روش برای خوشه بندی بین روش های k-means، mean-shift، k-modes مشخص گردید. این شاخص از بین روش های بیان شده، روش k-means را به عنوان بهترین روش معرفی کرد. پس از آن که روش k-means به عنوان روش مناسب انتخاب شد، تعیین تعداد خوشه مناسب برای این روش با توجه به داده های موجود از اهمیت زیادی برخوردار است؛ از همین رو، برای مشخص کردن تعداد خوشه مناسب از شاخص elbow استفاده گردید. این شاخص از بین خوشه های دو تایی تا ده تایی، خوشه سه تایی را به عنوان بهترین تعداد خوشه مشخص کرد. در همین حال در ادامه برای اطمینان از تعداد خوشه مناسب، خوشه بندی دو تایی الی پنج تایی نیز انجام شد و این خوشه بندی ها به صورت گرافیکی نشان داده شدند. در نمایش گرافیکی نیز با توجه به اینکه خوشه سه تایی بهترین و مشخص ترین مرز را بین خوشه ها ایجاد کرده بود، به عنوان بهترین تعداد خوشه انتخاب شد. در نهایت با استفاده از سه خوشه، داده ها خوشه بندی شدند و بر همین اساس نموداری بر مبنای سن و جنسیت شکل گرفت که براساس آن بتوان خوشه نمونه جدید را براساس این دو ویژگی مشخص نمود ویژگی های هر خوشه در هر یک از شاخص ها نیز در جدول پیوست قابل مشاهده است.

مولایی فر (۲۰۲۳) (۱۳) در تحقیق خود با عنوان ارائه یک سیستم توصیه گر برای صنعت گردشگری سلامت با استفاده از روش های داده کاوی نشان داد که طبق تحقیقات صورت گرفته خوشه بندی داده ها با استفاده از الگوریتم DBSCAN، امتیاز کارایی ۹۹٪ را بدست آورد که بالاترین امتیاز کارایی در بین الگوریتم های موجود می باشد، همچنین روش SVM در بخش دقت، امتیاز ۹۵٪ و در بخش فراخوانی، امتیاز ۹۹٪ را بدست آورد که نشان از دقت بالای پیش بینی نتایج را دارد و روش پیشنهادی به صورت کلی تا ۸۰٪ می تواند مکان های مورد نیاز گردشگر را به درستی تشخیص داده و مکان مناسب را تا حدود زیادی به درستی پیشنهاد دهد.

مارتین و همکاران (۲۰۱۸) (۱۴) در تحقیق خود با عنوان داده کاوی برای سلامتی: تعیین قلمرو اخلاقی فتوتیپ دیجیتال نشان دادند که از آنجایی که چارچوب های اخلاقی و نظارتی موجود برای ارائه مراقبت های بهداشتی روانی به وضوح در فتوتیپ دیجیتال اعمال نمی شود، بررسی پیام های احتمالی اخلاقی، قانونی و اجتماعی آن ضروری است. این مقاله به چهار حوزه اصلی می پردازد که دستورالعمل ها

و بهترین شیوه‌ها مفید خواهند بود؛ شفافیت، رضایت آگاهانه، حریم خصوصی و مسئولیت‌پذیری. در نظر گرفتن این موضوعات در مراحل اولیه توسعه این رویکرد جدید بسیار مهم خواهد بود تا وعده آن با اثرات مضر یا پیامدهای ناخواسته محدود نشود.

یافته‌های این مطالعه دارای چندین پیامد مهم است. نخست، نتایج نشان می‌دهد که می‌توان از روش‌های داده‌کاوی برای شناسایی الگوها و روندها در داده‌های مربوط به سلامت معلمان به طور مؤثر استفاده کرد. این اطلاعات می‌تواند برای توسعه مداخلات هدفمند جهت ارتقای سلامت و تندرستی معلمان به کار گرفته شود. دوم، این مطالعه چارچوبی برای خوشه‌بندی معلمان بر اساس شاخص‌های سلامت آنها ارائه می‌دهد که می‌تواند در تصمیم‌گیری در مورد تخصیص منابع و توسعه برنامه‌ها مؤثر باشد. سوم، این مطالعه سه خوشه مجزا از معلمان با پروفایل‌های سلامت متفاوت را شناسایی می‌کند که می‌توان از آنها برای ارائه استراتژی‌های ارتقای سلامت و پیشگیری از بیماری به صورت متناسب استفاده کرد. با وجود این، مطالعه حاضر دارای برخی محدودیت‌ها نیز می‌باشد. نخست، داده‌های مورد استفاده در این مطالعه از یک استان واحد در ایران جمع‌آوری شده‌اند، به همین دلیل ممکن است نتایج به سایر جمعیت‌ها تعمیم‌پذیر نباشد. دوم، این مطالعه شامل مؤلفه طولی نیست، بنابراین تعیین اثرات بلندمدت خوشه‌بندی بر سلامت معلمان امکان‌پذیر نیست. سوم، داده‌های مربوط به سایر عواملی که ممکن است بر سلامت معلمان تأثیر بگذارند، مانند سبک زندگی و سطح استرس، در این مطالعه جمع‌آوری نشده‌اند.

نتایج این مطالعه نشان داد که روش‌های داده‌کاوی می‌تواند برای شناسایی الگوها و روندها در داده‌های مربوط به سلامت معلمان استفاده شود. این اطلاعات می‌تواند برای برنامه‌ریزی بهتر برنامه‌های سلامتی و پیشگیری از بیماری‌ها در معلمان مفید باشد. پیشنهاد می‌گردد که در تحقیقات آتی با مطالعه پیشینه پژوهش علاوه بر موارد شناسایی شده موارد دیگری نیز به عوامل ذکر شده اضافه گردیده و ارتباط میان آنها شناسایی گردد.

**تشکر و قدردانی:** مقاله حاضر نتیجه پایان نامه دانشجویی در مقطع کارشناسی ارشد از دانشگاه علم و هنر یزد در سال ۱۴۰۲ می‌باشد. نویسندگان بر خود لازم می‌دانند تا از کلیه کسانی که در انجام این پژوهش ما را یاری رساندند، تشکر و قدردانی نمایند.

**تضاد منافع:** نویسندگان تعهد می‌دهند که در انجام پژوهش هیچگونه تعارض منافی وجود ندارد.

## References

- Paydar S, Reisi Ardali G, Raeesi H. Predicting Emergency Department Admission Using Data Mining (Case Study: Imam Ali Hospital in Shahrekord). *Health Information Management*. 2023;20(4):190-7.
- Engom B, Flerlage J. Blood chemistries and body fluids. *The Harriet Lane Handbook* Saunders Elsevier. 2015:621-33.
- Yang H, Luo Y, Ren X, Wu M, He X, Peng B, et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Information Fusion*. 2021;75:140-9.
- Rogers G, Joyner E. Mining Your Data for Healthcare Quality Improvement [Online]. 2011 [cited 2011 Aug 8]. Available from: URL: <http://www2.sas.com/proceedings/sugi22/EMERGING/PAPER139.PDF>.
- Mazaheri S, Ashoori M, Bechari Z. A model to predict Heart disease treatment using data mining. *Payavard Salamat*. 2017;11(3):287-96.
- Olson DL, Delen D, Meng Y. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*. 2012;52(2):464-73.

- Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*. 2013;25(2):127-36.
- Lakshmi K, Krishna MV, Kumar SP. Utilization of data mining techniques for prediction and diagnosis of tuberculosis disease survivability. *International journal of modern education and computer science*. 2013;5(8):8.
- Saura JR, Palos-Sanchez P, Grilo A. Detecting indicators for startup business success: Sentiment analysis using text data mining. *Sustainability*. 2019;11(3):917.
- Han J, Pei J, Tong H. *Data mining: concepts and techniques*: Morgan kaufmann; 2022.
- Englehardt SP, Nelson R. *Health care informatics: An interdisciplinary approach*. (No Title). 2002.
- Koh HC, Tan G. Data mining applications in healthcare. *Journal of healthcare information management*. 2011;19(2):65.
- Molae Fard R. Developing a recommender system for the health tourism industry using data mining methods. *Engineering Management and Soft Computing*. 2023;8(2):12. ۴۲-۵
- Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. *NPJ digital medicine*. 2018;1(1):68.

## Analysis of health patterns in Yazd teachers: application of data mining clustering for physical and clinical indicators

Hamed Fallah Tafti<sup>1</sup>, Saeede Rastjoo<sup>2</sup>, Batool zeidabadi<sup>3</sup>, Somayeh Kargar<sup>4</sup>

### Abstract

The healthcare industry is directly generating a large amount of data, and those who encounter this type of data have found that there is a wide gap between its collection and interpretation. The main objective of this research is to cluster teachers based on physical and clinical health indicators using data mining methods. The research is descriptive, exploratory and applied. The data were collected from health centers and the Ministry of Education of Yazd province from May 1398 to May 1400. In this study, 626 cases of the results of physical and clinical health indicators of teachers in Yazd province were used, which include 29 features. Descriptive statistics and data mining methods were used to analyze the information. The Python programming language was used in this project. The results of the study showed that the best clustering method is KMeans with 3 clusters. The silhouette and elbow indices were used to select the best clustering method and the number of clusters, respectively. Finally, the characteristics of each of these three clusters were identified, which can be used to obtain the cluster of new samples based on them. This study showed that data mining methods can be used to identify patterns and trends in data related to teacher health. This information can be useful for better planning of health programs and disease prevention in teachers.

**Keywords:** Clustering, Teachers, Physical and Clinical Health Indicators, Data Mining Methods

<sup>1</sup>. Associate Professor, Management department, Humanity Faculty, Science and Arts University, Yazd, Iran.

<sup>2</sup>. Department of Health Education and Health Promotion, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

<sup>3</sup>. Corresponding Author: Ph.D. Student in Health Education and Health Promotion, Sirjan School of Medical Sciences, Sirjan, Iran.

Email:zeidabadi.b@gmail.com.

<sup>4</sup>. Master's degree in Information Technology Engineering, University of Science and Arts, Yazd, Iran.