رویکرد نیمه‌پارامتری بیزی به خوشه‌بندی رگرسیونی کودکان و نوجوانان ایرانی بر اساس ریسک ابتلا به بیماری‌های قلبی– عروقی و دیابت

الهه شمس [۱]، ریحانه ریخته‌گران [۲]، رویا کلیشادی [۳]

مقاله پژوهشی

چکیده

مقدمه: در این مقاله، خوشه‌بندی رگرسیونی (Regression clustering) با استفاده از فرایند دیریکله (Dirichlet process) جهت به‌دست آوردن بینشی جامع از الگوی سلامت کودکان و نوجوانان بررسی‌شده در مطالعات ملی ایران، درنظر گرفته شد. در این رویکرد به خوشه‌بندی، ضمن این‌که تأثیر عوامل مزاحم، حذف و تحلیل‌های دقیق‌تری از وضعیت افراد به‌دست می‌آید، تعداد خوشه‌ها و الگوهای موجود در داده‌ها نیز تخمین زده می‌شود. در این پژوهش، خوشه‌بندی افراد نمونه از نظر دو شاخص چربی‌خون و قند خون، مدنظر قرار گرفته و میزان تأثیر شاخص تن‌سنجی (Anthropometric)، رده سنی و جنسیت بر نحوه خوشه‌بندی افراد، مورد ارزیابی قرار گرفته است.

روش‌ها: به منظور برآورد پارامترهای مجهول مدل، با رویکرد بیز (Bayesian approach) به مسأله، از روش‌های شبیه‌سازی مونت کارلوی زنجیر مارکفی (Markov chain Monte Carlo) در نرم‌افزار اپن‌باگز (Open Bugs) استفاده شده است. به‌منظور به‌دست آوردن بینش مناسب در رابطه با تعداد الگوهای افراد در معرض خطر بیماری‌های قلبی- عروقی و دیابت، بر مبنای شاخص چربی‌خون و قند خون، از فرایند دیریکله استفاده شده است.

یافته‌ها: تأثیر متغیرهای شاخص تن‌سنجی، رده سنی و جنسیت بر خوشه‌بندی با وارد کردن آن‌ها به‌عنوان متغیرهای توضیحی در مدل رگرسیونی و درنتیجه حذف اثر آن‌ها بر خوشه‌بندی، مورد بررسی قرار گرفت. نتایج منجر به تشکیل سه خوشه شد، به‌طوری که درصد افراد تخصیص‌یافته به خوشه‌های ۱ تا ۳، به‌ترتیب برابر با ۴۷٪ (۱۳۱۰ نفر)، ۴٪(۱۱۲ نفر) و ۴۹٪(۱۳۶۶نفر) شد. با بررسی چارک‌های متغیرهای قند خون و چربی‌خون، نتیجه می‌شود که افراد قرارگرفته در خوشه ۳ که حجم بزرگی از کودکان و نوجوانان نمونه را تشکیل می‌دهند، از لحاظ شاخص چربی‌خون و نیز قند خون، افراد در معرض خطر محسوب می‌شوند. هم‌چنین افراد خوشه اول دارای قند خون و شاخص چربی خون نرمال و افراد خوشه دوم دارای قند خون در محدوده خطر و شاخص چربی خون نرمال هستند.

نتیجه‌گیری: در خوشه‌بندی کودکان و نوجوانان از لحاظ دو متغیر قند خون و شاخص چربی‌خون، متغیرهای جنسیت، رده سنی و شاخص تن‌سنجی عوامل تأثیرگذار بر تعداد و ساختار خوشه‌ها هستند. هم‌چنین، افراد نمونه به سه خوشه، افراد با وضعیت نرمال، افراد در ریسک ابتلا به بیماری دیابت و افراد در ریسک بیماری‌های قلبی- عروقی و نیز دیابت، تقسیم‌بندی شدند.

واژه‌های کلیدی: خوشه‌بندی رگرسیونی، فرایندهای دیریکله، رویکرد بیز، قند خون، چربی خون

ارجاع: شمس الهه، ریخته‌گران ریحانه ، کلیشادی رویا. **رویکرد نیمه‌پارامتری بیزی به خوشه‌بندی رگرسیونی کودکان و نوجوانان ایرانی بر اساس ریسک ابتلا به بیماری‌های قلبی-عروقی و دیابت.** مجله تحقیقات نظام سلامت ۱۳۹۴؛ ۱۱(۲): ۴۳۴-۴۲۲

تاریخ دریافت: ۱۳۹۴/۰۳/۲۹ **تاریخ پذیرش: ۱۳۹۴/۰۶/۰۴**

۱. دانشجوی کارشناسی ارشد آمار، دانشگاه صنعتی اصفهان، دانشکده علوم ریاضی، اصفهان، ایران

۲. استادیار، دانشکده علوم ریاضی، دانشگاه صنعتی اصفهان، ایران (نویسنده مسؤول)

Email: r_rikhtehgaran@cc.iut.ac.ir

۳. استاد متخصص اطفال، دانشکده پزشکی و مرکز تحقیقات رشد و نمو کودکان، دانشگاه علوم پزشکی اصفهان، ایران

مقدمه

در مطالعه انجام‌گرفته توسط انستیتو تحقیقات تغذیه‌ای ایران نشان داده شده است که کیفیت الگوی تغذیه در ایران مناسب نیست به‌طوری‌که میزان مصرف قند، شکر، روغن و چربی در سبد غذایی خانواده‌ها، افزایش یافته و مصرف سبزی و میوه، لبنیات و گوشت محدود شده است. توجه به این مطلب و نیز مشاهده شیوع چاقی در جامعه ایرانی، بیانگر عدم تعادل در الگوی غذایی خانواده‌ها است. این مطلب هم‌چنین در بررسی دیگری که توسط دفتر سلامت روانی اجتماعی و مدارس وزارت بهداشت و درمان انجام شده‌است، مورد تأیید قرار گرفته است، به‌گونه‌ای که نتایج حاصل از آن نشان می‌دهد که درصد بالایی از سبد غذایی مصرفی خانواده‌ها را میان وعده‌های شور، چرب و شیرینی‌ها تشکیل می‌دهد که در این میان، حساسیت موضوع برای کودکان و نوجوانان، بیشتر است (۱).

نوجوانی به‌عنوان مرحله گذر از دوران کودکی به بزرگسالی و همراه با تغییرات خاص فیزیولوژیکی، تغییرات روحی–روانی و نیز جهش رشد است که علاوه بر عامل ژنتیک شدیداً متأثر از وضعیت تغذیه‌ای افراد است (۴–۲). الگوی تغذیه‌ای غلط خانواده‌ها باعث افزایش بیش از حد مصرف گروه چربی‌ها و مواد قندی به‌خصوص در میان کودکان و نوجوانان شده است که این امر یکی از عوامل زمینه‌ساز بیماری‌های قلبی–عروقی، فشارخون بالا و دیابت است. باید به این نکته توجه داشت که قند زیادی در بدن می‌تواند به تری‌گلیسیرید تبدیل‌شده و موجب افزایش وزن و چاقی شود که یکی از مهم‌ترین و مؤثرترین عوامل در افزایش میزان کلسترول است (۵). از طرفی مصرف زیاد چربی‌ها به ویژه چربی‌های حیوانی نیز موجب بالارفتن مقدار کلسترول خون می‌شود. کلسترول بالای خون یکی از علل ابتلا به بیماری‌های غیرواگیر است (۶).

بنابراین به‌طور خلاصه می‌توان چاقی، چربی‌خون و قند خون در کودکان و نوجوانان را به عنوان سه ضلع کاملاً مرتبط مثلث خطرناکی دانست که خود می‌تواند به‌عنوان عاملی برای تشدید سایر بیماری‌ها در بزرگسالی عمل کند (۸–۷). طبق بررسی‌های صورت‌گرفته ارتباط مستقیمی بین چاقی با افزایش قند خون و چربی‌خون وجود دارد (۹) .

علی‌رغم مشاهده پیشرفت‌های قابل توجه در راستای کاهش میزان مرگ و میر ناشی از بیماری‌های قلبی–عروقی، هنوز این بیماری‌ها نخستین علت مرگ و میر در بسیاری از کشورها محسوب می‌شوند. به‌ویژه شواهد موجود نشانگر افزایش فراوانی عوامل خطرساز در ابتلا به این بیماری‌ها و کاهش سن بروز این بیماری‌ها در ایران است (۱۱–۱۰). در رابطه با شناخت عوامل خطر مؤثر بر بیماری‌های قلبی–عروقی، بررسی‌های زیادی انجام شده است که از آن جمله می‌توان به (۱۵–۱۲) اشاره کرد. از طرف دیگر، یکی از بیماری‌های مزمن دوران کودکی، دیابت است. دیابت نوع ۱ یا دیابت وابسته به انسولین بیشتر در بین سنین ۷ تا ۱۵ سال آغاز می‌شود، هرچند که امکان ایجاد آن در هر سنی وجود دارد (۱۶). از طرف دیگر زمینه بروز دیابت نوع ۲ نیز در دوران کودکی و نوجوانی شکل می‌گیرد. بنابراین به دلیل اهمیت بیماری‌های قلبی– عروقی و دیابت و پایه‌ریزی عوامل ابتلا به آن‌ها در دوران کودکی و نوجوانی، ضرورت خوشه‌بندی کودکان و نوجوانان بر مبنای ریسک ابتلا به این بیماری‌ها، آشکار می‌گردد.

در مطالعه حاضر، از یافته‌های تن‌سنجی دانش‌آموزان در مطالعه کشوری «طرح ملی نظام مراقبت و پیشگیری از بیماری‌های غیرواگیر از دوران کودکی و نوجوانی» استفاده شده است. در این پژوهش، جهت به‌دست آوردن بینشی جامع از وضعیت سلامت کودکان و نوجوانان مورد مطالعه از لحاظ ریسک ابتلا به بیماری‌های قلبی– عروقی و دیابت، از روش خوشه‌بندی افراد نمونه از نظر دو شاخص چربی‌خون و قند خون، استفاده شده و میزان تأثیر شاخص تن‌سنجی و رده سنی کودکان و نوجوانان بر نحوه خوشه‌بندی آن‌ها مورد ارزیابی قرار گرفته است. اهمیت بررسی این موضوع، به‌دلیل وجود ارتباط مستقیم میان افزایش قند خون و چربی‌خون و ظهور بیماری‌های قلبی– عروقی و دیابت آشکار می‌گردد (۱۸–۱۷) . تا کنون تحقیق جامعی در زمینه خوشه‌بندی افراد

جامعه بر اساس ریسک ابتلا به بیماری‌های قلبی- عروقی و دیابت در کشور انجام نشده است و پژوهش حاضر جزو اولین تحقیقات در این زمینه است.

خوشه به مجموعه‌ای از داده‌ها گفته می‌شود که بیشترین شباهت را با هم داشته باشند. در خوشه‌بندی سعی می‌شود داده‌ها به خوشه‌هایی تقسیم شوند که شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت حداقل شود. خوشه‌بندی را می‌توان به‌عنوان ابزاری جهت شناخت بهتر جامعه و مرحله مقدماتی در مطالعات جامعی درنظر گرفت که در آن‌ها جامعه آماری از تنوع و گوناگونی زیادی برخوردار است.

خوشه‌بندی نسبت به روش طبقه‌بندی (Classification) داده‌ها از انعطاف بیشتری برخوردار است. در طبقه‌بندی، هر داده به یک طبقه (دسته) از پیش مشخص شده تخصیص می‌یابد. حال آن‌که در خوشه‌بندی هیچ اطلاعی از دسته‌های موجود درون داده‌ها وجود ندارد و به عبارتی خوشه‌ها و گاها تعداد آن‌ها از داده‌ها استخراج می‌شود. در یک طبقه‌بندی کلی، روش‌های خوشه‌بندی را می‌توان به دو دسته روش‌های غیر مدل- پایه (Non model-based methods) و مدل- پایه تقسیم کرد.

مهم‌ترین و پرکاربردترین الگوریتمی که در بحث خوشه‌بندی غیر مدل- پایه وجود دارد الگوریتم K- میانگین (K-means) است که جزء روش‌های غیر مدل-پایه بوده و در آن، هر داده به خوشه‌ای تخصیص می‌یابد که کم‌ترین فاصله را با مرکز آن خوشه دارد. در این روش، تعداد خوشه‌ها از قبل معلوم است. برای مطالعه بیشتر در این زمینه، می‌توان به مرجع (۱۹) مراجعه کرد.

در روش‌های مدل-پایه، خوشه‌بندی با استفاده از آمیخته‌ای از توزیع‌ها انجام می‌گیرد و به این ترتیب اطلاعات بیشتری از داده‌ها، نسبت به روش‌های غیر مدل-پایه، در مسأله خوشه‌بندی دخیل می‌شوند. در این نوشتار از روش خوشه‌بندی با استفاده از فرایند دیریکله که یکی از انواع روش‌های خوشه‌بندی مدل-پایه است، استفاده می‌شود. از مزایای استفاده از این روش، به این مطلب می‌توان اشاره کرد که هیچ توزیع پارامتری مشخصی برای توزیع متغیر مورد بررسی در خوشه‌ها درنظر گرفته نمی‌شود و ساختار توزیع خوشه‌ها با استفاده از ساختار داده‌ها، تخمین زده می‌شود. هم‌چنین این روش، تعداد خوشه‌ها را نیز مجهول درنظر گرفته و با استفاده از اطلاعات موجود در داده‌ها آن‌ها را برآورد می‌کند.

روش‌ها

توصیف مساله

طرح ملی نظام مراقبت و پیشگیری از بیماری‌های غیرواگیر از دوران کودکی و نوجوانی به صورت مقطعی بر ۵۳۱۱ دانش آموز ۲۸ استان کشور که در محدوده سنی ۱۰ تا ۱۸ سال قرار داشته‌اند، انجام شد. افراد نمونه با روش نمونه‌گیری خوشه‌ای چند مرحله‌ای به روش تصادفی انتخاب شدند. در این مطالعه، علاوه بر ثبت متغیرهای دموگرافیک مانند جنسیت، سن، محل سکونت از نظر شهری یا روستایی و ...، متغیرهای دیگری مانند شاخص توده بدنی (BMI یا Body mass index) ، کلسترول با تراکم پایین (LDL یا Low-density lipoprotein) ، کلسترول با تراکم بالا (HDL یا High-density lipoprotein)، تری‌گلیسیرید، چاقی شکمی به‌صورت نسبت دور کمر به قد، اندازه دور مچ دست طبق روش‌های استاندارد اندازه‌گیری، کلسترول‌مجموع، قند خون و ... نیز به‌ثبت رسیده‌اند. هم‌چنین برخی از عواملی که نشان‌دهنده عادات تغذیه‌ای و فعالیت بدنی افراد بوده است مانند تعداد ساعات تماشای تلویزیون، تعداد ساعات انجام فعالیت ورزشی در هفته، میزان مصرف سبزیجات تازه، میزان مصرف اجیل و تنقلات و ... نیز در این مطالعات ثبت شده است. در مطالعه حاضر، با درنظر گرفتن افراد با اطلاعات کامل، تعداد ۲۷۸۸ نفر جهت تحلیل‌های بعدی درنظر گرفته شده‌اند که ۴۹٪ از این تعداد را پسران تشکیل می‌دهند و ۶۹٪ ساکن شهرها هستند.

به‌دلیل وجود همبستگی بالا بین برخی از متغیرهای ثبت‌شده و به‌منظور تلخیص آن‌ها چند شاخص به نحوی که در ادامه

توضیح داده می‌شود، درنظر گرفته شده‌اند. با استفاده از روش تحلیل مؤلفه‌های اصلی (Principle component analysis)، شاخص تن‌سنجی از ترکیب متغیرهای شاخص توده بدنی، چاقی شکمی و اندازه دور مچ دست حاصل شده است. از آن‌جا که همبستگی شاخص جدید با متغیرهای مذکور به ترتیب برابر ۰/۹۱۱، ۰/۹۱۸ و ۰/۹۵ است، بنابراین اطلاعات موجود در این سه متغیر به خوبی در شاخص تن‌سنجی خلاصه شده است. هم‌چنین از ترکیب متغیرهای LDL، HDL، تری‌گلیسیرید و کلسترول‌مجموع، شاخص چربی به‌دست آمد که همبستگی آن با این متغیرها به ترتیب ۰/۹۲۹، ۰/۰۳، ۰/۹۶۸ و ۰/۳۷۹ به‌دست آمد. به این ترتیب، خوشه‌بندی کودکان و نوجوانان نمونه از لحاظ دو شاخص چربی و قند خون و بررسی عوامل مؤثر بر خوشه‌بندی آن‌ها، مورد توجه قرار گرفته است.

خوشه‌بندی به روش آمیخته‌ای از توزیع‌ها

همان‌گونه که بیان شد، با استفاده از آمیخته‌ای از توزیع‌ها به عنوان یک روش مدل–پایه، اطلاعات بیشتری از داده‌ها جهت خوشه‌بندی، مورد استفاده قرار می‌گیرد. در روش خوشه‌بندی آمیخته‌ای از توزیع‌ها، هر فرد با احتمالی به یک خوشه تخصیص می‌یابد. در این روش، فرض می‌شود مجموعه n–تایی از داده‌ها، مشاهداتی از متغیرهای تصادفی $Y_1, Y_2, \dots, Y_n$ هستند که هر یک از این متغیرها توزیع‌های احتمال $f_{\theta_1}, f_{\theta_2}, \dots, f_{\theta_K}$ را با احتمال‌های $\pi_1, \pi_2, \dots, \pi_K$ اختیار می‌کنند، به‌طوری که مجموع این احتمال‌ها، یک است. بنابراین طبق قانون احتمال کل، داریم

$$f_\theta(y_i) = \sum_{i=1}^{K} \pi_i \times f_{\theta_j}(y_i), \qquad i = 1,2, \dots, n$$

که در آن $\sum_{i=1}^{K} \pi_i = 1$. معمولاً توزیع‌های $f_{\theta_1}, f_{\theta_2}, \dots, f_{\theta_K}$ از یک خانواده از توزیع‌ها درنظر گرفته می‌شوند. به‌منظور خوشه‌بندی داده‌ها، متغیر پنهان w_i را به‌عنوان برچسب مشاهده i–ام، درنظر می‌گیریم. مقادیر این متغیر در مجموعه $\{1, \dots, K\}$ که در آن K تعداد مؤلفه‌های توزیع آمیخته یا به عبارتی تعداد خوشه‌ها است، اختیار می‌شود.

به این ترتیب در صورت معلوم بودن مقدار متغیر w_i برای i–امین مشاهده، توزیع Y_i به صورت زیر مشخص می‌شود

$$f_\theta(y_i|w_i = j) = f_{\theta_j}(y_i)$$

پس از برآورد پارامترهای θ_j برای $j = 1, \dots, K$ و w_i برای $i = 1, \dots, n$، توابع درستنمایی محاسبه شده و سپس مشاهده موردنظر به دسته‌ای تعلق می‌گیرد که در آن دسته دارای درستنمایی بیشتری است. به منظور برآورد پارامترهای مدل بالا، می‌توان از الگوریتم (Expectation Maximization (EM)) در آمار کلاسیک و یا با رویکرد بیز به مسأله و درنظر گرفتن تابع زیان مربع خطا، از برآوردگر بیز استفاده کرد. در این مطالعه، برآورد پارامترها با استفاده از نرم افزار اپن‌باگز (۲۰) و با رویکرد بیز، به‌دست آمده‌اند.

خوشه‌بندی رگرسیونی با استفاده از آمیخته‌ای از توزیع‌های نرمال دومتغیره

اغلب دسته‌بندی افراد جامعه، بدون درنظر گرفتن عوامل مؤثر بر خوشه‌بندی منجر به تحلیل نادرست از افراد قرار گرفته در خوشه‌ها می‌شود. به بیان دیگر گاهی وجود برخی عوامل مزاحم، موجب ایجاد خوشه و یا تخصیص نادرست افراد به خوشه‌هایی می‌شود به‌طوری‌که در شرایط مشابه ولی بدون حضور این عوامل مزاحم، نتایج خوشه‌بندی متفاوتی حاصل می‌شد. این مسأله با حذف اثر برخی عوامل تأثیرگذار بر خوشه‌بندی و نحوه تغییر در تعداد و ساختار خوشه‌ها بهتر آشکار می‌شود.

از این رو، در این مطالعه، با وارد کردن متغیرهای توضیحی چون جنسیت، شاخص تن‌سنجی و ... در مدل‌های رگرسیونی، به بررسی اثرگذاری آن‌ها بر خوشه‌بندی پرداخته و ویژگی‌های مشترک افراد قرارگرفته در یک خوشه را مشخص می‌کنیم، به‌گونه‌ای که در نهایت بتوان به تحلیل دقیق‌تری از ساختار جامعه بر مبنای متغیرهای پاسخ درنظر گرفته‌شده، دست‌یافت.

فرض‌کنید بردار $\boldsymbol{Y}_i$ شامل متغیرهای قند خون و شاخص چربی و بردار $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ شامل مشاهداتی از p متغیر توضیحی مانند جنسیت، رده سنی و... بر روی i–امین فرد

از نمونه باشند. ازاین‌رو، مدل رگرسیونی مورد نیاز برای خوشه‌بندی داده‌ها، به صورت زیر ارائه می‌شود:

$$y_i = \mu_{i,w_i} + \varepsilon_i,$$

که در آن

$$\mu_{i,w_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \alpha_{w_i},$$

برای $i = 1, \ldots, n$ و α_{w_i} را اثر تصادفی مربوط به خوشه فرد i–ام درنظر می‌گیریم. در مدل بالا، ε_i مانده‌های مدل و $\beta_0, \beta_1, \ldots, \beta_p$ ضرایب رگرسیونی دومؤلفه‌ای هستند که در آن، هر مولفه متناظر با تأثیرگذاری متغیر توضیحی مربوطه بر یکی از متغیرهای قند خون و شاخص چربی است. متغیر w_i نیز برچسب و به عبارت دیگر نشانگر خوشه فرد i –ام است. معمولاً متغیر تصادفی α_{w_i} و مانده‌های مدل، مستقل از هم و نیز مستقل از متغیرهای توضیحی فرض می‌شوند. به این ترتیب، با معلوم بودن مقدار متغیر w_i برای i–امین مشاهده و درنظر گرفتن توزیع نرمال برای مانده‌های مدل، توزیع Y_i به صورت زیر مشخص می‌شود.

$$f_{(\mu,\Sigma)}(y_i | w_i = j) = N_2(\mu_{ij}, \Sigma)$$

که در آن $P(w_i = j) = \pi_j$ برای j=1,…,K. اکنون با استفاده از الگوریتم نمونه‌بردار گیبز (Gibbs sampler) به‌عنوان یکی از روش‌های شبیه‌سازی مونت کارلوی زنجیر مارکفی، برآورد پارامترها به دست خواهند آمد و سپس هر فرد به خوشه‌ای تخصیص می‌یابد که بیشترین درستنمایی را دارد.

از آن‌جا که در عمل، معمولا توزیع متغیر مورد بررسی در خوشه‌ها نرمال نیست و می‌تواند دارای ساختارهای چوله و یا دم‌کلفتی باشد که توزیع نرمال نمی‌تواند به درستی آن‌ها را پوشش دهد، بنابراین نیاز به استفاده از توزیع‌های منعطف‌تر از نرمال به خوبی احساس می‌شود. از طرف دیگر در روش خوشه‌بندی با استفاده از آمیخته‌ای از توزیع‌ها، مانند بسیاری از روش‌های خوشه‌بندی دیگر، لازم است که تعداد خوشه‌ها از قبل، معلوم باشد، حال آن‌که در عمل، معمولا تعداد خوشه‌ها مجهول است و نیاز است که به همراه سایر پارامترهای توزیع، تعداد خوشه‌ها نیز تخمین زده شود. به همین دلیل در ادامه، به معرفی فرایند دیریکله به عنوان یک روش نیمه‌پارامتری و تعمیمی از روش خوشه‌بندی به‌وسیله آمیخته‌ای از توزیع‌ها، می‌پردازیم.

خوشه‌بندی رگرسیونی با استفاده از فرآیند دیریکله

در تحقیقات اخیر، به منظور دستیابی به خوشه‌بندی دقیق‌تر، توجه به خانواده‌هایی از توزیع‌های با ساختار منعطف‌تر از توزیع نرمال برای توزیع مانده‌ها و یا اثرات تصادفی مدل‌ها، بیشتر شده است. در راستای تلاش‌های صورت‌گرفته جهت به‌کارگیری توزیع‌های منعطف، استفاده از رویکرد نیمه‌پارامتری به‌نام فرایند دیریکله مورد توجه است. در این شیوه، ابتدا توزیع احتمالی G که مشاهدات از آن تولید شده‌اند نامعلوم فرض‌شده و سپس اندازه احتمالی به‌عنوان توزیع پیشین برای توزیع احتمالی نامعلوم G، درنظرگرفته می‌شود. این اندازه احتمال، حول اندازه احتمال پایه G_0 متمرکز است و مقدار این تمرکز با پارامتر مثبت M تعیین می‌شود، به‌طوری که هرچه مقدار M بزرگ‌تر باشد، G به G_0 شبیه‌تر است.

مزیت اصلی استفاده از این شیوه، عدم تحمیل توزیع پارامتری مشخص به مدل است. علی‌رغم ویژگی‌های برجسته این روش، به‌دلیل وجود پیچیدگی‌های محاسبات در کاربرد عملی آن در خوشه‌بندی، تنها در سال‌های اخیر و با پیشرفت نرم‌افزارهای آماری مانند اپن‌باگز، خوشه‌بندی با استفاده از فرایند دیریکله مورد استفاده فراوان قرار گرفته‌است.

فرایندهای دیریکله، برای اولین‌بار توسط فرگوسن و پس از آن توسط افرادی مانند بلکول و مک‌کویین و ستارامن بسط داده شد (۲۳-۲۱). این فرآیند، توزیع تصادفی گسسته G را به‌صورت زیر تولید می‌کند

$$G(.) = \sum_{j=1}^{\infty} \pi_j \delta_{\alpha_j}(.),$$

که در آن $\delta_{\alpha_j}(.)$ برای مقادیر برابر با α_j، مقدار ۱ و درغیراینصورت مقدار صفر را اختیار می‌کند. هم‌چنین $\alpha_j \sim G_0$ و $\pi_j = \vartheta_j \prod_{l=1}^{j-1}(1 - \vartheta_l)$ که در آن

$\vartheta_j \sim Beta(1, M)$. فرایند دیریکله را با نماد $DP(M, G_0)$ نمایش می‌دهیم. معمولا در عمل، مجموع بالا را تا مقدار C محاسبه می‌کنند.

به‌منظور استفاده از فرایند دیریکله در خوشه‌بندی رگرسیونی داده‌ها، مدل قبل را به‌صورت زیر تغییر می‌دهیم.

$$\boldsymbol{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i,$$

که در آن

$$\boldsymbol{\mu}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \cdots + \boldsymbol{\beta}_p x_{ip} + \boldsymbol{\alpha}_i,$$

دقت کنید که این‌بار، $\boldsymbol{\alpha}_i$ معرف اثر تصادفی مرتبط با فرد i–ام است. حال فرایند دیریکله را به‌عنوان پیشین موردنیاز برای توزیع مجهول اثرات تصادفی افراد، درنظر می‌گیریم. از آنجا که فرایند دیریکله، تولید توزیع تصادفی گسسته می‌کند، بنابراین تنها چندین مقدار متمایز (K مقدار) برای α_i ها تولید می‌شوند که هر از آن‌ها می‌توانند معرف یک خوشه باشند. به‌عبارت دیگر، تمام افراد با اثرات تصادفی یکسان، به یک خوشه تعلق می‌گیرند. به این ترتیب بسته به تنوع و ساختار موجود در داده‌ها، تعداد خوشه‌ها بدون این‌که از پیش تعیین شده باشد، مشخص می‌شود.

حال مشابه قبل، با استفاده از الگوریتم نمونه‌بردار گیبز و استفاده از نرم‌افزار اپن‌بگز، برآورد پارامترها به‌دست می‌آید و سپس می‌توان خوشه‌بندی را بر مبنای محاسبه توابع درستنمایی انجام داد.

مدل نیمه‌پارامتری جهت خوشه‌بندی رگرسیونی داده‌های طرح ملی نظام مراقبت و پیشگیری از بیماری‌های غیرواگیر از دوران کودکی و نوجوانی

در این مطالعه، خوشه‌بندی افراد بر اساس دو عامل قند خون و شاخص چربی‌خون، به‌عنوان متغیرهای پاسخ، مدنظر قرار گرفته است. برای این منظور از مدل رگرسیونی نیمه‌پارامتری زیر استفاده می‌شود.

$$y_{ij} = \beta_{0j} + \beta_{1j} Sex_i + \beta_{2j} C\_age_i + \beta_{3j} Anth_i + \alpha_{ij} + \varepsilon_{ij}, \quad j = 1, 2$$

که در آن $j = 1$ نشان‌دهنده مدل رگرسیونی مرتبط با متغیر قند خون و $j = 2$ نشان‌دهنده مدل رگرسیونی مرتبط با شاخص چربی خون است. توزیع بردار مانده‌های $(\varepsilon_{i1}, \varepsilon_{i2})'$، نرمال دومتغیره با میانگین‌های صفر و ماتریس واریانس–کوواریانس قطری با اعضای روی قطر $\sigma^2_{\varepsilon_1}$ و $\sigma^2_{\varepsilon_2}$ و توزیع بردار اثرات تصادفی $(\alpha_{i1}, \alpha_{i2})'$، توزیعی نامشخص با پیشین فرایند دیریکله درنظر گرفته می‌شود که در آن توزیع پایه، نرمال دومتغیره با میانگین‌های صفر و ماتریس واریانس–کوواریانس قطری با اعضای روی قطر $\sigma^2_{\alpha_1}$ و $\sigma^2_{\alpha_2}$ است. متغیرهای توضیحی موجود در این مدل، عبارت هستند از: متغیر رده سنی (C_age) که به‌صورت دو رده کودکان و نوجوانان تعریف شده است، متغیر شاخص تن‌سنجی (Anth) و نیز متغیر جنسیت (Sex) است.

به‌منظور بررسی تأثیرگذاری متغیرهای توضیحی بر خوشه‌بندی افراد، مدل‌های زیر را درنظر می‌گیریم. مدل ۱، که شامل هیچ متغیر توضیحی نیست و تنها بر اساس مقادیر مشاهده شده برای متغیرهای پاسخ، خوشه‌بندی را انجام می‌دهد. مدل ۲ که شامل متغیرهای توضیحی جنسیت و شاخص تن‌سنجی است. مدل ۳ که متغیرهای توضیحی جنسیت و رده سنی را درنظر می‌گیرد و در نهایت مدل ۴ که هر سه متغیر جنسیت، شاخص تن‌سنجی و رده سنی را به‌عنوان متغیرهای توضیحی لحاظ می‌کند. به این ترتیب با مقایسه نتایج حاصل از خوشه‌بندی‌های به‌دست آمده از مدل‌های ۱ تا ۴، می‌توان تأثیرگذاری عوامل جنسیت، رده سنی و شاخص تن‌سنجی را بر خوشه‌بندی افراد، مورد ارزیابی قرار داد.

به منظور به‌دست آوردن برآوردهای بیز پارامترهای مدل‌ها، پیشین‌های مزدوج و آگاهی‌نابخش را برای پارامترها درنظر می‌گیریم. به عبارتی توزیع N (۰/۰۰۰۱) برای ضرایب رگرسیونی مدل و توزیع گامای معکوس IG (۰/۰۱، ۰/۰۱) برای مؤلفه‌های واریانس مدل‌ها، درنظر گرفته شده‌اند. سپس با استفاده از نرم‌افزار اپن‌باگز تعداد ۱۵۰۰۰۰ نمونه پس از حذف ۲۵۰۰۰۰ دورریز و همگرا شدن توزیع‌های پسین پارامترها، تولید شده است. به این ترتیب، متوسط مقادیر تولیدشده به‌عنوان برآوردهای بیز پارامترها درنظر گرفته می‌شوند (به

جدول ۱ توجه کنید) و در نهایت با استفاده از پارامترهای برآورد شده، درستنمایی مرتبط با هر خوشه، برای هر داده محاسبه شده و سپس داده به خوشه‌ای تعلق پیدا کرده که درستنمایی بیشتری داشته است.

جدول ۱. برآوردهای بیز (واریانس برآوردها) برای پارامترهای مدل‌های ۱ تا ۴.

مدل	K	β_{01}	β_{02}	β_{11}	β_{12}	β_{21}	β_{22}	β_{31}	β_{32}	$\sigma^2_{\varepsilon_1}$	$\sigma^2_{\varepsilon_2}$
مدل ۱	۴	۱/۵۶ (۹/۳۴)	۱/۲۳ (۰/۶۹)	-	-	-	-	-	-	۱۲۵ (۰/۵۴)	۳/۹ (۰/۰۳)
مدل ۲	۵	۴/۳۳ (۹/۶۵)	۱/۳۶ (۰/۷۳)	۱/۳۲ (۰/۴۷)	−۰/۱۴ (۰/۰۴)	-	-	۰/۴۸ (۰/۲۲)	۰/۱۱ (۰/۰۲)	۱۲۰/۹ (۴/۸۹)	۰/۵۱ (۰/۰۴)
مدل ۳	۵	۴/۱۹ (۱۰/۰۶)	۱/۱۹ (۰/۶۰)	۱/۴۱ (۰/۴۵)	−۰/۱۳ (۰/۰۴)	−۰/۵۸ (۰/۴۰)	−۰/۱۷ (۰/۰۴)	-	-	۱۱۸/۸ (۴/۱۴)	۰/۴۸ (۰/۰۴)
مدل۴	۳	۳۹/۳۰ (۱۲/۵۲)	−۲۸/۴۸ (۰/۳۴)	۰/۳۹ (۱۴/۰۸)	−۰/۱۶ (۰/۰۸)	−۰/۹۱ (۳/۶۲)	−۰/۲۵ (۰/۰۴)	۰/۶۱ (۰/۸۲)	۰/۱۵ (۰/۰۲)	۴۹۳/۶ (۶۶۰/۷)	۰/۹۳ (۰/۰۵)

یافته‌ها

پس از به‌کارگیری فرایند دیریکله جهت خوشه‌بندی داده‌های طرح ملی نظام مراقبت و پیشگیری از بیماری‌های غیرواگیر از دوران کودکی و نوجوانی در مدل ۱، نتایج زیر حاصل‌شد. همان‌طور که در شکل ۱ ملاحظه می‌شود، بر اساس مدل ۱، چهار خوشه ایجاد شده و درصد افراد تخصیص‌یافته به خوشه‌های ۱ تا ۴، به‌ترتیب برابر با ۶۳٪ (۱۷۵۹ نفر)، ۳۲٪(۸۸۹ نفر) ۲۴/۳٪(۶۸ نفر) و ۲۴/۷٪(۶۹ نفر) است. با استفاده از این مدل، چارک‌های اول و سوم متغیر قندخون خوشه‌های اول تا چهارم به‌ترتیب برابر با (۹۳، ۷۸)، (۹۶، ۸۱)، (۹۹/۷۵، ۸۴) و (۱۱۰، ۹۰) به‌دست آمده است. بنابراین می‌توان نتیجه گرفت افراد خوشه اول و دوم در دسته افراد دارای قند خون نرمال قرار می‌گیرند، درحالی‌که در خوشه سوم تقریباً ۲۵٪ از افراد دارای قند خون بالای ۹۹ هستند و از لحاظ این کمیت در گروه افراد در معرض خطر قند خون بالا قرار می‌گیرند. افراد خوشه چهارم نیز به‌عنوان افراد در معرض خطر قند خون بحرانی شناخته می‌شوند، زیرا ۷۵٪ از این افراد دارای قند خون بالای ۹۰ و ۲۵٪ نیز دارای قند خون بالای ۱۱۰ هستند. هم‌چنین، چارک‌های اول و سوم متغیر شاخص چربی خوشه‌های اول تا چهارم مدل ۱ به‌ترتیب برابر با (۰/۳۴، ۰/۸۵−)، (۰/۹۵، ۰/۳−)، (۲/۲۵، ۰/۲۵−) و (۰/۷۳، ۰/۶۵−) است که بر مبنای متغیر کلسترول‌مجموع به‌ترتیب معادل است با (۱۵۶/۹۳، ۱۲۲/۱۹)، (۱۷۴/۷۳، ۱۳۸/۲۴)، (۲۱۲/۶۸، ۱۳۹/۷۰) و (۱۶۸/۳۱، ۱۲۸/۰۳). بنابراین می‌توان استنتاج کرد که نیمی از افراد خوشه سوم دارای شاخص چربی ۰/۲۵− تا ۲/۲۵ و ۲۵٪ دارای شاخص چربی بیش‌تر از ۲/۲۵ هستند، ازاین‌رو افراد قرارگرفته در این خوشه، از لحاظ شاخص چربی خون، افراد در معرض خطر چربی خون بالا محسوب می‌شوند، درحالی که سه خوشه دیگر چنین وضعیت بحرانی را ندارند. پس به‌طور خلاصه می‌توان نتیجه‌گیری کرد که افراد خوشه‌های اول و دوم دارای قند خون و شاخص چربی خون نرمال و افراد خوشه سوم دارای چربی خون در محدوده خطر و شاخص قند خون نرمال و افراد خوشه چهارم دارای قند خون در محدوده خطر و شاخص چربی خون نرمال هستند.

به‌منظور تشخیص عوامل پنهان در تشکیل خوشه‌های مدل ۱، متغیرهایی در رابطه با میزان فعالیت‌های جسمی و نوع عادات غذایی در جدول ۲ بررسی شده که تنها اطلاعات به‌ثبت رسیده از نوع عادات غذایی افراد در رابطه با درصد افراد مصرف‌کننده سبزیجات و نیز آجیل و تنقلات دارای تنوع

بیش‌تری در میان خوشه‌ها است و شاید بتوان از آن‌ها به‌عنوان دسته‌ای از عوامل که باعث ایجاد چنین خوشه‌بندی شده‌اند، یاد کرد که البته به‌علت پایین بودن درصد افراد پاسخ‌دهنده به سؤالات مرتبط با آن‌ها، چنین نتیجه‌گیری‌ای چندان معتبر نیست.

نتایج خوشه‌بندی حاصل از مدل ۲ یعنی خوشه‌بندی که در آن به جنسیت و شاخص تن‌سنجی افراد توجه شده و به رده سنی افراد توجه نشده است، نشان‌دهنده ۵ الگو برای جامعه مورد مطالعه است به‌طوری که درصد افراد تخصیص‌یافته به خوشه‌های ۱ تا ۵، به‌ترتیب برابر با ۳۸٪ (۱۰۷۲ نفر)، ۴۵٪(۱۲۴۹نفر)، ۱۴٪(۳۸۶نفر)، ۲٪(۵۰ نفر) و ۱٪(۲۸ نفر) است. بر طبق نتایج حاصل از بررسی چارک‌های اول و سوم متغیرهای قند خون و چربی خون برای این خوشه‌ها، این‌طور نتیجه شد که افراد خوشه‌های اول و دوم دارای قند خون و چربی خون نرمال، افراد خوشه سوم دارای قند خون نرمال و چربی خون نسبتا بالا هستند. دو خوشه چهارم و پنجم از لحاظ وضعیت سلامت در خطر و به‌ترتیب مستعد بیماری دیابت و بیماری‌های قلبی–عروقی به‌نظر می‌رسند.

خوشه‌بندی بر مبنای مدل ۳ یعنی خوشه‌بندی که در آن به جنسیت و رده سنی افراد توجه شده و به شاخص تن‌سنجی آن‌ها توجه نشده است، نیز منجر به تشکیل ۵ خوشه شده است. درصد افراد تخصیص‌یافته به خوشه‌های ۱ تا ۵، برابر با ۲۳٪(۶۳۷ نفر)، ۴۶٪(۱۲۸۸نفر)، ۲۸٪(۷۵۹نفر)، ۲٪(۵۸ نفر) و ۲٪(۴۳ نفر) است. تحلیل خوشه‌های ایجاد شده نیز مشابه تحلیل‌های حاصل از مدل ۲ است.

در نهایت در مدل ۴، تمامی متغیرهای جنسیت، رده سنی و شاخص تن‌سنجی وارد مدل شده و درنتیجه اثر آن‌ها بر خوشه‌بندی، حذف گردیده است. بر اساس این مدل سه خوشه ایجاد شده و درصد افراد تخصیص‌یافته به خوشه‌های ۱ تا ۳، به‌ترتیب برابر با ۴۷٪ (۱۳۱۰ نفر)، ۴٪(۱۱۲ نفر) و ۴۹٪(۱۳۶۶نفر) است. با استفاده از این مدل، چارک‌های اول و سوم متغیر قندخون خوشه‌های اول تا سوم به‌ترتیب برابر با (۷۶،۸۹)، (۸۸،۹۹) و (۸۴،۹۸) به‌دست آمده است. بنابراین می‌توان نتیجه گرفت که نیمی از افراد خوشه اول دارای قند خون ۷۶ تا ۸۹ هستند و از این‌رو در دسته افراد دارای قند خون نرمال قرار می‌گیرند، درحالی که در دو خوشه دوم و سوم تقریباً ۲۵٪ از کودکان و نوجوانان دارای قند خون بالای ۹۹ هستند و از لحاظ این کمیت در گروه افراد در معرض خطر بیماری دیابت قرار می‌گیرند. هم‌چنین، چارک‌های اول و سوم شاخص چربی خوشه‌های اول تا سوم به‌ترتیب برابر با (۰/۴، ۰/۹–)، (۰/۳، ۱–) و (۰/۸، ۰/۵–) است که به‌ترتیب معادل است با (۱۵۸/۶۷، ۱۲۰/۷۳)، (۱۵۵/۷۶، ۱۱۷/۸۱) و (۱۷۰/۳۵، ۱۳۲/۴۱) برای متغیر کلسترول‌مجموع. بنابراین می‌توان استنتاج کرد که نیمی از افراد خوشه سوم دارای شاخص چربی ۱۳۲/۴۱ تا ۱۷۰/۳۵ و ۲۵٪ دارای شاخص چربی بیش‌تر از ۱۷۰/۳۵ هستند، از این‌رو افراد قرارگرفته در این خوشه که حجم بزرگی از افراد مورد مطالعه را تشکیل می‌دهند، از لحاظ شاخص چربی خون، افراد در معرض خطر چربی خون بالا محسوب می‌شوند، درحالی که دو خوشه ۱ و ۲ از لحاظ چربی خون تقریباً دارای وضعیت مشابه هستند و در گروه افراد دارای چربی خون نرمال قرار می‌گیرند. پس به‌طور خلاصه می‌توان نتیجه‌گیری کرد که افراد خوشه اول دارای قند خون و شاخص چربی خون نرمال و افراد خوشه دوم دارای قندخون در محدوده خطر و شاخص چربی خون نرمال و افراد خوشه سوم، دارای دارای قند خون و شاخص چربی خون در محدوده خطر هستند.

به‌منظور تشخیص عوامل احتمالی در ایجاد خوشه‌ها، لازم است که سایر متغیرهای مورد بررسی در این مطالعه نیز از نقطه‌نظر تاثیرگذاری در ایجاد خوشه‌ها، مورد بررسی قرار گیرند. از آن جمله می‌توان به متغیرهایی اشاره نمود که در رابطه با میزان فعالیت‌های جسمی و نوع عادات غذایی دانش‌آموزان، ثبت شده‌اند و در جدول ۳ گزارش شده‌اند. از میان این متغیرها می‌توان به درصد افراد مصرف‌کننده سبزیجات و نیز آجیل و تنقلات اشاره کرد که نسبت آن در سه خوشه ایجاد شده، متفاوت است. البته به‌علت پایین بودن

درصد افراد پاسخ‌دهنده، نتایج حاصل از آن چندان قابل استناد نیست. سایر متغیرهای درنظر گرفته شده در جدول۳ نیز در هر سه خوشه، تقریبا مشابه بوده و نشان‌دهنده عدم تأثیر چشم‌گیر این عوامل در خوشه‌بندی افراد است.

قابل ذکر است که اثر حذف متغیر جنسیت از مدل ۴ نیز مورد بررسی قرار گرفت که نتایج خوشه‌بندی حاصل از آن، تفاوت چندانی با مدل ۴ نداشت و لذا نتایج گزارش نشده است. در شکل ۱، نتایج کلی خوشه‌بندی متناظر با مدل‌های ۱ تا ۴، خلاصه شده است.

جدول ۲. درصد فراوانی نسبی (درصد پاسخ‌دهندگان) برای هر متغیر، تحت خوشه‌های به‌دست آمده در مدل ۱.

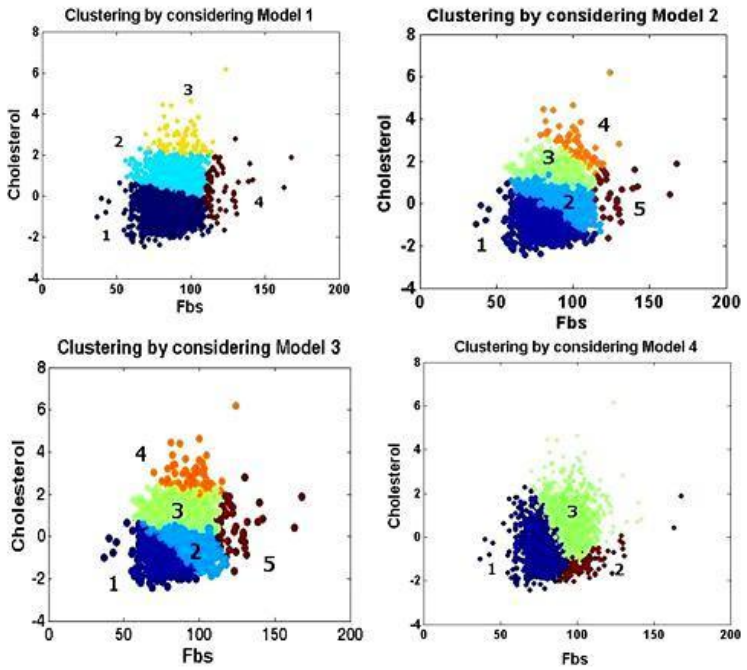
مدل ۱	خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴
سکونت در شهر	۶۷	۷۳/۸	۷۳/۸	۷۱
نوجوانان	۵۳/۲	۵۰/۳	۴۸/۵	۶۰/۹
جنسیت(مرد)	۴۹/۹	۴۶/۶	۵۳/۴	۶۳/۸
داشتن سابقه بیماری قلبی‌عروقی خانوادگی	۸۳/۲	۸۴/۱	۷۹/۴	۸۸/۴
فشار خون انقباضی نرمال	۹۳/۹	۹۶	۹۷	۹۸/۳
فشار خون انبساطی نرمال	۹۵/۸	۹۶/۱	۹۷	۹۸/۳
تماشای بیش‌تر از ۷ ساعت تلویزیون	۶۴/۴	۳۶/۶	۷۰/۸	۴۵/۶
شرکت منظم در کلاس‌های ورزشی	۹۱/۸	۹۱/۸	۹۵/۵	۸۸/۲
استفاده از روغن مایع در غذا	۹۶/۱	۹۷/۳	۹۰/۵	۹۳/۲
مصرف سبزیجات تازه در هفته	۲۳/۳ (۰/۸)	۵۰ (۰/۹)	– (۰)	– (۰)
مصرف آجیل در هفته	۱۸/۴ (۲/۴)	۲۶/۹ (۲/۹)	۵۰ (۲/۹)	– (۰)
مصرف تنقلات در هفته	۷/۹ (۲/۲)	۱۰/۵ (۲/۱)	۱۰۰ (۱/۵)	۰ (۱/۴)
استفاده از شیر مادر در دوران شیرخوارگی	۷۸/۹	۸۰/۱	۷۹/۴	۸۸/۴
وزن بالای ۳ کیلوگرم هنگام تولد	۱۰/۸	۹/۸	۴/۷	۱۱/۹

جدول ۳. درصد فراوانی نسبی (درصد پاسخ‌دهندگان) برای هر متغیر، تحت خوشه‌های به‌دست آمده در مدل ۴.

مدل ۴	خوشه ۱	خوشه ۲	خوشه ۳
سکونت در شهر	۶۲	۷۵/۳۷	۷۶/۶
نوجوانان	۵۳/۹	۵۰/۴۴	۵۷/۷
جنسیت(مذکر بودن)	۴۹/۲	۴۸/۹	۴۹/۵
داشتن سابقه بیماری قلبی‌عروقی خانوادگی	۸۳/۵	۸۳/۷	۸۲
فشار خون انقباضی نرمال	۹۳/۸	۹۵/۹	۹۲/۳
فشار خون انبساطی نرمال	۹۵/۹	۹۶/۹	۹۸
تماشای بیش‌تر از ۷ ساعت تلویزیون	۶۱/۸	۶۱/۸	۶۴/۲
شرکت منظم در کلاس‌های ورزشی	۹۱/۳	۹۲/۲	۹۲/۷

ادامه جدول ۳

استفاده از روغن مایع در غذا	۳۶/۵ (۰/۸)	۲۹/۷ (۲/۷)	۹۸
مصرف سبزیجات تازه در هفته	۳۶/۴ (۰/۸)	۵۶/۵ (۰/۸)	۰ (۰/۹)
مصرف آجیل در هفته	۱۹/۴ (۲/۴)	۲۹/۷ (۲/۷)	۰ (۱/۸)
مصرف تنقلات در هفته	۱۰ (۲/۳)	۷/۴ (۲)	۰ (۱/۸)
استفاده از شیر مادر در دوران شیرخوارگی	۸۱/۴	۷۷/۴	۷۵/۵
وزن بالای ۳ کیلوگرم هنگام تولد	۱۰/۶	۹/۸	۱۴/۲



شکل ۱. نتایج کلی خوشه‌بندی

بحث

با توجه به نتایج به‌دست آمده از مدل‌های ۱ تا ۴، چنین نتیجه می‌شود که در خوشه‌بندی کودکان و نوجوانان از لحاظ دو متغیر قند خون و شاخص چربی‌خون، متغیر رده سنی و شاخص تن‌سنجی دو عامل تأثیرگذار بر تعداد و ساختار خوشه‌ها هستند. یکی از مهم‌ترین مزایای خوشه‌بندی رگرسیونی با استفاده از فرایند دیریکله، در مقایسه نتایج حاصل از مدل‌های ۱ و ۴ نمایان می‌شود. مدل ۴، بر مبنای

خوشه‌بندی رگرسیونی و مدل ۱ بر اساس خوشه‌بندی غیر رگرسیونی عمل می‌کند که در آن نقش عوامل مزاحم بر متغیر پاسخ، نادیده گرفته می‌شود. خوشه‌بندی حاصل از مدل‌های ۱ و ۴، نمایش داده شده در شکل۱ را درنظر بگیرید. خوشه‌بندی بر مبنای مدل۱، خوشه‌بندی اولیه و خام از داده‌ها است و تنها بر اساس محدوده مقادیر قند خون و شاخص چربی خون افراد، خوشه‌ها ایجاد و تخصیص‌ها صورت گرفته است. بنابراین اگر فردی در خوشه شماره ۴ قرار گرفته به این معنی است که این فرد در ریسک ابتلا به بیماری دیابت قرار دارد ولی در خطر ابتلا به بیماری‌های قلبی–عروقی نیست. حال افرادی از خوشه ۴ از مدل ۱ را درنظر بگیرید که در خوشه‌بندی حاصل از مدل ۴، در خوشه ۳، قرار گرفته‌اند. همان‌طور که قبلا عنوان‌شده بود، افراد خوشه ۳ در مدل ۴ را می‌توان افراد در معرض ابتلا به بیماری‌های قلبی–عروقی و نیز دیابت دانست. دلیل تفاوت در خوشه‌های ایجاد شده و درنتیجه، تحلیل‌های متفاوت برای افراد را می‌توان در این دانست که فردی که با مدل ۱، در معرض بیماری‌های قلبی–عروقی نبوده، درصورت توجه به جنسیت، رده سنی و نیز وضعیت تن‌سنجی وی، در خوشه‌بندی حاصل از مدل۴، مستعد چنین بیماری‌هایی می‌شود. در واقع، خوشه‌بندی مدل ۱ را می‌توان همانند تحلیل یک فرد بدون تخصص و خوشه‌بندی مدل ۴ را می‌توان مانند تحلیل یک پزشک باتجربه و متخصص دانست که با توجه به شرایط فرد، وی را مستعد ابتلا به بیماری‌های قلبی–عروقی و یا دیابت می‌داند.

هم‌چنین، حضور سهم زیادی از کودکان و نوجوانان نمونه در خوشه‌های پرخطر ۲ و ۳ در خوشه‌بندی حاصل از مدل ۴ را می‌توان به‌عنوان زنگ خطری برای وضعیت سلامت جامعه ایرانی در سال‌های آتی از لحاظ ریسک ابتلا به بیماری‌های قلبی–عروقی و دیابت دانست که نیازمند به تدوین سیاست‌ها و برنامه‌ریزی‌های اصولی در زمینه غذا و نیز تعدیل در الگوی تغذیه‌ای خانواده‌های ایرانی است.

نتیجه‌گیری

در خوشه‌بندی کودکان و نوجوانان از لحاظ دو متغیر قند خون و شاخص چربی‌خون، متغیرهای جنسیت، رده سنی و شاخص تن‌سنجی عوامل تأثیرگذار بر تعداد و ساختار خوشه‌ها هستند. هم‌چنین، افراد نمونه به سه خوشه، افراد با وضعیت نرمال، افراد در ریسک ابتلا به بیماری دیابت و افراد در ریسک بیماری‌های قلبی– عروقی و نیز دیابت، تقسیم‌بندی شدند.

تشکر و قدردانی

در پایان، از تمامی شرکت‌کنندگان و همکاران در اجرای مطالعه طرح ملی نظام مراقبت و پیشگیری از بیماری‌های غیرواگیر از دوران کودکی و نوجوانی تشکر می‌نماییم.

References

1. Safavi M. Modifying consumption patterns and its important nutritional status. The World of Nutrition 2011; 86. [In Persian].
2. Chukwunonso Ejike ECC, Chidi Ugwu E, Lawrence US. Physical growth and nutritional status of a cohort of semi-urban Nigerian adolescents. Pak j Nutr 2010; 9(4):392-397.
3. Malina RM, Physical activity and training: effects on stature and the adolescent growth spurt. Medicine and Science in Sports and Exercise 1994; 26(6): 759-766.
4. Rogal AD, Clark PA, Roemmich JN. Growth and pubertal development in children and adolescents: effects of diet and physical activity. Am J Clin Nutr 2000; 72(2): 521-528.
5. Despres JP. Obesity and lipid metabolism: relevance of body fat distribution. Curr Opin Lipidol 1991; 2: 5–15.
6. Ma H. Cholesterol and Human Health. Nature and Science 2004; 2(4): Supplement.
7. Burke GL, Webbe LS, Srinivasan SR, Radhakrishnamurthy B, Freedman DS, Berenson GS. Fasting plasma glucose and insulin levels and their relationship to cardiovascular risk factors in children: Bogalusa Heart Study. Metabolism 1986; 35(5): 441–446.
8. Fujioka S, Matsuzawa Y, Tokunaga K, Tarui S. Contribution of intra-abdominal fat accumulation to the impairment of glucose and lipid metabolism in human obesity. Metabolism 1987; 36(1): 54–59.

9. Zavaroni I, Dall'Aglio E, Alpi O, Bruschi F, Bonora E, Pezzarossa A, Butturini U, Evidence for an independent relationship between plasma insulin and concentration of high density lipoprotein cholesterol and triglyceride. Atherosclerosis 1985; 55(3): 259-66.
10. Sarraf-Zadegan N, Sayed-Tabatabaei FA, Bashardoost, N, The prevalence of coronary artery disease in an urban population in Isfahan, Iran, Acta Cardiologica, 1999; 54: 257-263.
11. Sarraf-Zadegan, N Boshtam, M, Rafiei, M., Risk factors for coronary artery disease in Isfahan, Iran, European Journal of Public Health, 1999; 9: 20-26.
12. Gavish D., Leibovitz E., Elly I.,Shargorodsky M., Zimlichman R. Follow-up in a lipid clinic improves the management of risk factor in cardiovascular disease patients, Isr Med Assoc J., 2002, 4: 694-7.
13. Yamori L., Liu L., Mu L., Zhao H., Pen Y., Hu Z., et al. Diet-related factors, educational levels and blood pressure in a Chinese population sample: findings from the Japan-China cooperative research project, Hypertens Res, 2002, 25: 559-64.
14. Wagrowska H. Risk of developing coronary disease in relation of the level of education and type of work in a make population of warsaw factories, kardiol Pol, 1989, 32:57-60.
15. Liazarzaburu JL., Palinkas LA. Immigration, acculturation, and risk factors for obesity and cardiovascular disease: a comparison between lations of peruvian descent in peru and in the united states, Ethn Sis: 2002, 12, 342-52.
16. Alemzadeh R, Wayatt DT. Diabetes Mellitus in Children. In: Behrman RE, Kliegman RM, Jenson HB. Nelson Textbook of Peadiatrics. 17th ed. Philadelphia, Saunders 2004: 1947-1972.
17. Yasien N. Jarrah S, Petro-Nutas W, Jaber R, Terzi N, Froelicher ES, Khawajah E. Obesity Indices and their Relationship to Cardiovascular Risk Factors in Young Adult Group. The Bahrain Medical Society 2010; 22(4): 133-137.
18. Grundy SM, Brewer HB Jr, Cleeman JI. Definition of metabolic syndrome: Report of the national heart, lung, and blood institute, American Heart Association conference on scientific issues related to definition. Circulation 2004: 109-433.
19. Kanungo T, Mount D, Netanyahu N, Piatko C, Silverman R, Wu A. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on pattern analysis and machine intelligence 2002; 24(7): 881-892.
20. Lunn D, Spiegelhalter D, Thomas A, Best N, The BUGS project: evolution, critique and future directions (with discussion). Stat Med 2009; 28: 3049–3082.
21. Ferguson TS, A Bayesian analysis of some nonparametric problems. Ann Stat 1973; 1: 209–230.
22. Blackwell D. and MacQueen JB, Discreteness of Ferguson selections. Annals of Statistics 1973; 1: 365-358.
23. Sethuraman J. A constructive definition of Dirichlet priors. Statistica Sinica 1994; 4: 639-650.

A Bayesian semi-parametric approach to regression clustering of Iranian children and adolescence based on risk of cardiovascular disease and diabetes

Elaheh Shams [1], Reyhaneh Rikhtehgaran [2], Roya Kelishadi [3]

Original Article

Abstract

Background: In this paper, we study the Regression clustering using Dirichlet processes to achieve a comprehensive insight about the status of children and adolescence's health based on Iranian national studies. This clustering approach, as well as removing the effects of troublous factors in clustering and obtaining more accurate analysis, estimates the number of clusters and their patterns.In the present study, subjects of the sample are clustered according to the cholesterol index and fast blood sugar (FBS) which are known as risk factors of cardiovascular disease and diabetes. Furthermore, the effect of the anthropometric index, category of age and sex on clustering are evaluated.

Methods: To estimate unknown parameters of model, in a Bayesian approach, we adopt Markov chain Monte Carlo simulation methods using OpenBUGs software. In addition, to get a correct view of the number of patterns on the risk of cardiovascular disease and diabetes based on FBS and the cholesterol index, the Dirichlet Process is applied.

Findings: The anthropometric index, category of age and sex as explanatory variables were entered in the model and effects of them were removed. Then, three clusters were obtained. Percentages of allocated subjects to clusters one to three are 47 %(1310 persons), 4% (112 persons) and 49% (1366 persons), respectively. Analyzing the first and the third quantiles of the FBS and the cholesterol index in each cluster, it is revealed that the allocated subjects to the third cluster, who consist a significant part of the sample, are assumed to be at the risk of both cardiovascular disease and diabetes. Moreover, people in the first cluster have normal levels of FBS and the cholesterol index and the subjects allocated to the second cluster have dangerous level of FBS and normal level of the cholesterol index.

Conclusion: In clustering of children and adolescence based on FBS and the cholesterol index, category of age, sex and the anthropometric index were known as influential factors on estimating the number of clusters and creating their structures. Also, three clusters are obtained based on this study: subjects with the safe status, subjects at the risk of diabetes and subjects at the risk of both cardiovascular disease and diabetes.

Key Words: Regression Clustering, Dirichlet Process, Bayesian Approach, Food Blood Sugar, The Cholesterol Index.

Citation: Shams E, Rikhtehgaran R, Kelishadi R. **A Bayesian semi-parametric approach to regression clustering of Iranian children and adolescence based on risk of cardiovascular disease and diabetes.** J Health Syst Res 2015; 11(2): 422-434

Received date: 19.06.2015 **Accept date: 26.08.2015**

1. MSc Student, Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran
2. Assistant, Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran (Corresponding Author) Email: r_rikhtehgaran@cc.iut.ac.ir
3. Professor of Pediatrics, Faculty of Medicine & Child Growth and Development Research Center, Isfahan University of Medical Sciences, Isfahan, Iran