



Comparison of PAM and CLARA Clustering Algorithms for Assessing the Causes of Mortality in the Population Covered by Mashhad University of Medical Sciences, Iran

Maryam Salari^{1*} PhD, Sara Sabbaghian Tousi¹ MA, Sedigheh Rastaghi² MA, Parastoo Golpour¹ MA, Mohammad Vejdani^{3,4} MD, Seyyed Masoud Sadati¹ MD, Mohammad Taghi Shakeri^{1,5} PhD

¹ Department of Epidemiology and Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran

² Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran

³ Department of General Surgery, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

⁴ Cardio-Thoracic Surgery and Transplant Research Center, Emam Reza Hospital, Mashhad University of Medical Sciences, Mashhad, Iran

⁵ Sinus and Surgical Endoscopic Research Center, Mashhad University of Medical Sciences, Mashhad, Iran

*Correspondence to: Maryam Salari, Email: salarimr@mums.ac.ir

Received: April 5, 2022

Revised: July 3, 2022

Accepted: August 16, 2022

Online Published: September 6, 2022

Abstract

Introduction: Analyzing data related to mortality is one of the important tools in health policy for any country. The present study was conducted in order to cluster people who died in 2017 in the area covered by Mashhad University of Medical Sciences.

Methods: An analytical and cross-sectional study was conducted on the number of 21,838 deaths reported in 2017 in the population covered by Mashhad University of Medical Sciences, which were extracted from the death registration system of this university. After cleaning the data and fixing the existing defects, people were clustered with PAM and CLARA algorithms based on 4 variables: age, gender, cause of death and region of residence. The optimal number of clusters and the evaluation of the clustering performance were obtained using the average silhouette criterion.

Results: According to the silhouette criterion, the biggest difference between the clusters was observed with the number of five clusters. In all clusters (except the third cluster), which included people over 65 years old (64.5%), circulatory diseases (43%), neoplasms (17%), diseases of the respiratory system (10%) and Endocrine, nutritional and metabolic diseases (6%) were the most common cause of death. In the third cluster, the most important causes of death were congenital malformations, deformation (44%) and diseases of prenatal origin (19%) for children under 1 year of age (74%).

Conclusion: Age distribution of the causes of death is one of the most important cases. The existence of congenital malformations, deformation, and diseases of prenatal origin for children under 1 year of age, as well as the cause of death due to the presence of cancer at the age of less than 14 years, according to the upcoming policies regarding the youth of the population, should be considered by health care providers. The high percentage of death due to neurological diseases is one of the cases that can be studied carefully to plan a health road map.

Keywords: Cluster Analysis, Mortality, Cause of Death

Highlights

1. In Iran, like developed countries, cardiovascular disease is the first cause of death.
2. Male gender children under 12 are more vulnerable than children with female gender.
3. The prevalence of cancers among Iranians in the mid-year category is a danger to health policy makers.

Citation:

Salari M, Sabbaghian Tousi S, Rastaghi S, Golpour P, Vejdani M, Sadati SM, et al. Comparison of PAM and CLARA clustering algorithms for assessing the causes of mortality in the population covered by Mashhad University of Medical Sciences, Iran. *Iran J Health Insur.* 2022;5(2):108-17.

Copyright © 2022 The Author(s); Published by National Center for Health Insurance Research. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



مقایسه دو الگوریتم خوشه‌بندی PAM و CLARA به منظور بررسی علل مرگ و میر در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد

مریم سالاری^۱ PhD، سارا صباغیان طوسی^۱ MA، صدیقه رستاقی^۲ MA، پرستو گلپور^۱ MA، محمد وجدانی^۳ MD، سید مسعود ساداتی^۱ MSc، محمدتقی شاکری^۱ PhD

^۱ گروه اپیدمیولوژی و آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران
^۲ کمیته تحقیقات دانشجویی، دانشگاه علوم پزشکی مشهد، مشهد، ایران
^۳ گروه جراحی عمومی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران
^۴ مرکز تحقیقات جراحی قلب و قفسه سینه و پیوند، بیمارستان امام رضا (ع)، دانشگاه علوم پزشکی مشهد، مشهد، ایران
^۵ مرکز تحقیقات آندوسکوپی سینوس و جراحی، دانشگاه علوم پزشکی مشهد، مشهد، ایران

* نویسنده مسئول: مریم سالاری، پست الکترونیک: salarimr@mums.ac.ir

انتشار آنلاین: ۱۴۰۱/۰۶/۱۵

پذیرش: ۱۴۰۱/۰۵/۲۵

تصحیح: ۱۴۰۱/۰۴/۱۲

دریافت: ۱۴۰۱/۰۱/۱۶

چکیده

مقدمه: تحلیل اطلاعات مربوط به مرگ و میر یکی از ابزارهای مهم در سیاست‌گذاری سلامت برای هر کشوری است. مطالعه حاضر به منظور خوشه‌بندی افراد فوت‌شده در سال ۱۳۹۶ در منطقه تحت پوشش دانشگاه علوم پزشکی مشهد انجام شد.

روش بررسی: مطالعه تحلیلی و مقطعی بر روی تعداد ۲۱۸۳۸ مرگ گزارش‌شده در سال ۱۳۹۶ در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد که از سامانه نظام ثبت مرگ این دانشگاه استخراج شد، انجام شد. پس از پاک‌سازی داده‌ها و رفع نواقص موجود، افراد با دو الگوریتم PAM و CLARA براساس ۴ متغیر سن، جنس، علت مرگ و منطقه محل سکونت خوشه‌بندی شدند. تعداد بهینه خوشه‌ها و ارزیابی عملکرد خوشه‌بندی با استفاده از معیار میانگین سیلوئت حاصل شد.

یافته‌ها: براساس معیار سیلوئت، بیشترین تفاوت بین خوشه‌ها با تعداد ۵ خوشه مشاهده شد. در تمامی خوشه‌ها (به جز خوشه سوم) که شامل افراد بالای ۶۵ سال (۶۴/۵ درصد) بودند، به ترتیب بیماری‌های گردش خون (۴۳ درصد)، نئوپلاسم‌ها (۱۷ درصد)، بیماری‌های تنفسی (۱۰ درصد) و بیماری‌های غدد درون‌ریز، تغذیه‌ای و سوخت و ساز (۶ درصد) شایع‌ترین علت مرگ و میر بودند. در خوشه سوم، مهم‌ترین علل مرگ و میر ناهنجاری‌های مادرزادی، تغییر شکل و شرایط خاص ناشی از ناهنجاری مادرزادی (۴۴ درصد) و بیماری با منشأ قبل از دوران تولد (۱۹ درصد) برای کودکان زیر یک سال (۷۴ درصد) بود.

نتیجه‌گیری: توزیع سنی علل مرگ جزو موارد بسیار مهم است. وجود ناهنجاری‌های مادرزادی و بیماری با منشأ قبل از دوران تولد در کودکان زیر یک سال، همچنین علت مرگ به دلیل وجود سرطان در سنین کمتر از ۱۴ سال، با توجه به سیاست‌های پیش رو درمورد جوانی جمعیت باید مدنظر سیاست‌گذاران سلامت قرار گیرد. درصد بالای مرگ به دلیل بیماری‌های عصبی، از جمله مواردی است که می‌توان با مطالعه دقیق درباره آن نقشه راه سلامت را برنامه‌ریزی کرد.

واژگان کلیدی: تحلیل خوشه‌ای، مرگ و میر، علت مرگ

نکات ویژه

۱. در ایران نیز همانند کشورهای توسعه‌یافته، بیماری‌های قلبی- عروقی اولین علت مرگ محسوب می‌شود.
۲. کودکان زیر ۱۲ سال با جنسیت مرد، آسیب‌پذیرتر از کودکان با جنسیت زن هستند.
۳. شیوع سرطان‌ها میان ایرانیانی که در رده میان‌سالی قرار دارند، زنگ خطری برای سیاست‌گذاران سلامت محسوب می‌شود.

مقدمه

تحلیل دقیق داده‌های مرتبط با مرگ و میر، برای ارتقا و ارزیابی سیاست‌های مرتبط با سلامت امری ضروری است. علل مرگ معمولاً براساس تشخیص نهایی و زمینه‌ای براساس استاندارد بین‌المللی ICD-10^۱ گزارش می‌شود. پزشک در تشخیص نهایی علت مستقیم مرگ را گزارش می‌دهد و در تشخیص زمینه‌ای دلایلی را که مستقیماً دلیل مرگ نمی‌شوند اما به طور معناداری بر مرگ تأثیر دارند، گزارش می‌دهد [۱]. در کشور ایران متولی اصلی جمع‌آوری اطلاعات مربوط به مرگ وزارت بهداشت، درمان و آموزش پزشکی است که با هماهنگی و تبادل اطلاعات از سازمان‌های پزشکی قانونی، شهرداری، اطلاعات را جمع‌آوری و صحت‌سنجی می‌کند [۲].

تا سال ۲۰۰۱ مرگ براساس بیماری‌های قلبی و سکته مغزی ۵۰ درصد مرگ‌ها را در هر دو جنس زن و مرد شامل می‌شد. در یک دوره زمانی مشابه، مرگ به دلیل زوال عقلی و آلزایمر در مردان ۶۰ درصد افزایش یافت و در زنان دو برابر شد که سالمندشدن جمعیت را نشان می‌داد. در سال ۲۰۱۵ بیماری‌های قلبی شایع‌ترین دلیل مرگ در مردان بود. مرگ به دلیل بیماری‌های قلبی با ۱/۸ برابر دومین علت مرگ محسوب می‌شد. سرطان ۳۰ درصد مرگ در مردها و ۲۴/۸ درصد مرگ در زن‌ها را تشکیل می‌داد [۳].

اولین تغییر در علت مرگ، در اروپا از قرن ۱۷ تا اوایل قرن ۱۹ میلادی اتفاق افتاد که کاهش و از بین رفتن بیماری‌های واگیر مانند آبله و طاعون بود و در اواخر قرن نوزدهم مرگ به دلیل بیماری‌های اسهال و سل از بین رفت [۴]. بعد از جنگ جهانی دوم سومین دوره تغییر در دلایل مرگ بود. از هر ۱۰ علت مرگ ۴ مرگ به دلیل سرطان بود و برای هر دو جنس زن و مرد سرطان ریه، خون، انقباض و پروستات بیشترین دلایل مرگ بودند. در گروه زنان سرطان سینه و در گروه مردان سرطان پروستات بالاترین نرخ را داشتند. در سال ۲۰۱۵ بیشترین دلیل مرگ در گروه سنی ۵۰ تا ۷۹ سال و به دلیل بیماری‌های قلبی و سرطان بود و در گروه افراد زیر ۳۵ سال دلایل خارجی مانند تصادفات و خودکشی بالاترین نرخ را داشت [۵].

در کشور ایران در سال ۱۳۹۰ بیماری‌های قلبی- عروقی ۴۶/۱۳ درصد، سرطان و تومورها ۱۳/۶۳ درصد و حوادث غیرعمدی ۱۱/۵۵ درصد بود. ۱۰ دلیل نخست مرگ در کل جمعیت در این سال به ترتیب: سکته قلبی، سکته مغزی، حوادث مربوط به حمل و نقل، بیماری‌های قلبی ناشی از پرفشاری خون، سایر بیماری‌های قلبی- عروقی، دیابت، بیماری‌های مزمن ریه و برونشیت، سرطان معده، سایر بیماری‌های قلبی و نارسایی کلیه بوده است [۶].

خوشه‌بندی به روش k-medoids یکی از روش‌های کارآمد خوشه‌بندی است و PAM و CLARA از رایج‌ترین الگوریتم‌های این روش محسوب می‌شوند. با توجه به اینکه در رابطه با به کارگیری و مقایسه این دو الگوریتم مطالعات اندکی انجام شده، مطالعه حاضر با هدف مقایسه دو الگوریتم خوشه‌بندی PAM و CLARA در تجزیه و تحلیل نتایج حاصل از این الگوریتم‌ها برای ۲۱۸۳۸ مرگ گزارش شده در سال ۱۳۹۶ در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد براساس علت اصلی مرگ انجام شد. فرض بر این است که با قرار گرفتن افراد فوت شده در خوشه‌های مختلف علل مرگ براساس سن، جنس و منطقه سکونت مورد بررسی دقیق‌تری قرار گیرند.

روش بررسی

این مطالعه تحلیلی و مقطعی بر روی ۲۱۸۳۸ مرگ گزارش شده در سال ۱۳۹۶ در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد که از سامانه نظام ثبت مرگ این دانشگاه استخراج شد، انجام شد. داده‌های ثبت شده پس از پاک‌سازی و رفع نواقص موجود، به طور کامل بر روی ۴ متغیر سن، جنس، علت مرگ و منطقه محل سکونت مورد مطالعه قرار گرفتند.

خوشه‌بندی یک روش از داده‌کاوی است که باعث استخراج اطلاعات مفید و کشف الگوی پنهان از داده‌های خام می‌شود. الگوریتم‌های خوشه‌بندی مختلفی در دسترس است. یکی از معروف‌ترین الگوریتم‌های خوشه‌بندی k-medoids است که در زمره روش‌های خوشه‌بندی افزایی قرار دارد و برخلاف الگوریتم k-means، در برابر داده‌های پرت و نویزدار مقاوم‌تر است، چون به جای میانگین از خود داده‌ها برای مرکز ثقل و نمایندگی

بهینه، از مدوئیدها تولید شود. این الگوریتم فرآیند نمونه‌گیری و خوشه‌بندی را چندین بار تکرار می‌کند تا آریبی نمونه‌گیری به حداقل برسد [۸].

تعیین تعداد بهینه خوشه‌ها با استفاده از معیار سیلوئت: قبل از اجرای الگوریتم‌ها، با استفاده از روش میانگین معیار سیلوئت تعداد بهینه خوشه‌ها محاسبه می‌شود. این معیار، پراکندگی داده‌ها را در خوشه‌ها مشخص می‌کند. هر چه مقدار سیلوئت برای K مورد نظر بیشتر باشد، کیفیت خوشه‌بندی نیز بیشتر خواهد بود. این روش به ازای مقادیر مختلف K (به طور مثال بازه ۲ تا ۱۰) اجرا می‌شود و به ازای هر اجرا، مقدار سیلوئت برای تک تک مشاهدات و برای هر مقدار K محاسبه می‌شود و سپس از سیلوئت‌های به دست آمده، میانگین گرفته می‌شود. مقدار میانگین سیلوئت بر حسب مقادیر مختلف K رسم می‌شود. نقطه ماکزیمم نمودار رسم شده، تعداد خوشه بهینه را مشخص می‌کند.

ارزیابی عملکرد الگوریتم‌های خوشه‌بندی با استفاده از معیار سیلوئت: معیار سیلوئت، هم به پیوستگی درون خوشه‌ها و هم به میزان تفکیک‌پذیری آن‌ها بستگی دارد و مقدار آن برای هر مشاهده، میزان تعلق آن را به خوشه‌اش در مقایسه با خوشه مجاور اندازه‌گیری می‌کند.

فرض کنید مشاهده‌ای مانند x_i در میان داده‌هایی که خوشه‌بندی شده، وجود دارد و طی مراحل خوشه‌بندی نیز، k خوشه $(C_1, C_2, \dots, C_K)$ ایجاد شده است. برای محاسبه معیار سیلوئت احتیاج به آشنایی با دو مفهوم اصلی است: میانگین فاصله یک مشاهده از خوشه با نقاط دیگر آن خوشه: این مقدار با $a(i)$ مشخص می‌شود و ملاکی برای ارزیابی تعلق مشاهده در خوشه‌اش است. هر چه این مقدار کوچک‌تر باشد، میزان تعلق مشاهده x_i به خوشه‌اش بیشتر است.

$$a(i) = \frac{1}{n_i} \sum_{l=1}^{n_i} d(x_i, x_l)$$

حداقل میانگین فاصله مشاهده با خوشه‌های دیگر: فرض کنید مشاهده x_i به خوشه C_j تعلق دارد. میانگین فاصله این مشاهده با نقاط خوشه‌های دیگر C_k اندازه می‌شود. خوشه‌ای که دارای کمترین میانگین فاصله برای مشاهده x_i باشد، به عنوان خوشه

خوشه‌ها استفاده می‌کند [۷]. همچنین مجموع مربعات خطا در این الگوریتم نسبت به الگوریتم k -means پایدارتر است، چون به جای مجموع مربعات فواصل اقلیدسی، مجموع جفت‌های فاقد تشابه را به حداقل می‌رساند [۸].

الگوریتم PAM: یکی از معروف‌ترین الگوریتم‌های k -medoids، الگوریتم PAM است که برای مجموعه داده با حجم کم مناسب است چون در غیر این صورت زمان محاسبه زیاد می‌شود [۸]. در این الگوریتم برای ایجاد k خوشه، ابتدا k داده به عنوان نماینده خوشه‌ها انتخاب می‌شود. هر نماینده دارای بیشترین تشابه با سایر اعضای خوشه است. این داده که مدوئید نامیده می‌شود در مرکزی-ترین نقطه خوشه قرار دارد. پس از مشخص شدن مدوئیدها سایر داده‌ها به خوشه‌ای تعلق می‌گیرند که با مدوئید آن خوشه تشابه بیشتری داشته باشند. بدین ترتیب خوشه‌های اولیه تشکیل می‌شوند. سپس در یک فرآیند تکراری تمامی داده‌ها برای جانشینی با مدوئیدها آزمایش می‌شوند تا مجموعه مدوئیدی که بهترین خوشه‌بندی را دارد، انتخاب شود. در واقع این روش براساس اصل حداقل‌سازی فقدان شباهت‌ها میان هر شیء با شیء مرجع (مدوئید) عمل می‌کند. در اینجا معیار فقدان شباهت، فاصله است، که معرف نبود تجانس است و به ما کمک می‌کند تا در فضای داده‌ای حرکت کنیم و خوشه‌ها را تشکیل دهیم [۸، ۹]. در الگوریتم PAM می‌توان از دو فاصله جذر مجموع مربعات تفاوت‌ها و مجموع قدرمطلق فواصل استفاده کرد [۱۰]. اما در صورتی که که مجموعه داده شامل متغیرهای عددی، طبقه‌ای و باینری باشد، می‌توان برای ساخت ماتریس فقدان تشابه از فاصله Gower استفاده کرد [۱۱].

الگوریتم CLARA: این الگوریتم، تعمیمی بر الگوریتم PAM است و برای داده‌هایی که شامل تعداد زیادی از مشاهدات (بیش از چندین هزار مشاهده) است، کاربرد دارد. این الگوریتم پایدارتر از الگوریتم PAM است و با استفاده از روش نمونه‌گیری، زمان محاسبات و ذخیره‌سازی RAM، را کاهش می‌دهد.

الگوریتم CLARA به جای پیدا کردن مدوئیدها برای کل مجموعه داده، یک نمونه از داده‌ها را در نظر می‌گیرد و الگوریتم PAM را برای نمونه انتخاب شده اجرا می‌کند تا مجموعه‌ای

یافته‌ها

از تعداد کل ۲۱۸۳۸ مرگ گزارش شده در سال ۱۳۹۶ در جمعیت تحت پوشش دانشگاه علوم پزشکی مشهد، ۴۴ درصد افراد زن و ۵۶ درصد مرد بودند. ۲۰ درصد افراد در مناطق روستایی، ۷۸ درصد در مناطق شهری و بقیه افراد محل سکونت نامشخص داشتند. همان‌طور که در نمودار ۱ مشاهده می‌شود بالاترین درصد مرگ در سنین ۶۵ سال به بالا و حدود ۵۹ درصد مرگ‌ها را شامل می‌شد. ۷ درصد مرگ، مرتبط با گروه سنی زیر یک سال بود.

با توجه به تعداد مشخصه‌های افراد مورد بررسی، ابتدا خوشه‌بندی به روش k-medoids به ترتیب برای ۱۰، ...، ۲، K خوشه مورد بررسی قرار گرفت. با توجه به نمودار ۲ و با استفاده از معیار میانگین سیلوئت و برای هر دو الگوریتم PAM و CLARA، ۵ و K=۴ بیشترین پهنای سیلوئت را به خود اختصاص دادند. تعداد بهینه خوشه‌ها K=۵ در نظر گرفته شد، زیرا گویای بیشترین تفاوت بین مشاهدات بود.

جدول ۱ | تفسیر میانگین سیلوئت در ارزیابی عملکرد خوشه‌بندی

میانگین شاخص سیلوئت	تفسیر
۰/۷۱ - ۱/۰۰	ساختار قوی
۰/۵۱ - ۰/۷۰	ساختار منطقی (مناسب)
۰/۲۶ - ۰/۵۰	ساختار ضعیف
$\geq ۰/۲۵$	هیچ ساختار قابل توجهی وجود ندارد.

مجاور با این مشاهده نامیده می‌شود. مقدار میانگین فاصله مشاهده با مشاهدات خوشه مجاور با $b(i)$ نشان داده می‌شود.

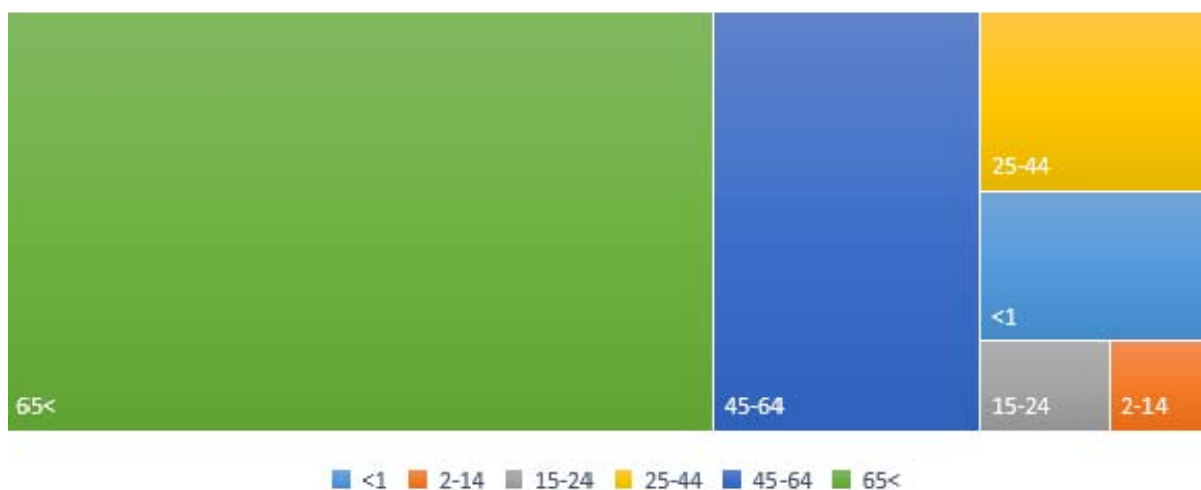
$$b(i) = \min_{1 \leq l \leq k} \frac{1}{n_l} \sum_{y_m \in C_l} (d(x_i, y_m))$$

میزان معیار سیلوئت برای مشاهده به وسیله رابطه زیر اندازه‌گیری می‌شود:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

اگر $a(i)$ کوچک‌تر از $b(i)$ باشد، مقدار معیار سیلوئت مثبت خواهد بود و اگر $b(i)$ کوچک‌تر از $a(i)$ ، مقدار این معیار منفی می‌شود که نشانگر خوشه‌بندی ضعیف است چون مشاهده بیش از آن که شبیه خوشه خودش باشد به خوشه مجاور شباهت دارد. با توجه به رابطه بالا مقدار این معیار بین -۱ تا +۱ متغیر است. مقدار نزدیک به ۱ بیانگر انطباق مناسب، بین مشاهده و خوشه‌اش نسبت به خوشه مجاور است. اگر معیار سیلوئت برای همه نقاط درون خوشه‌ها نزدیک به ۱ باشد، خوشه‌بندی به درستی انجام شده است. در حالی که کوچک بودن این مقدار برای خوشه‌ها، بیانگر ضعیف بودن نتایج خوشه‌بندی است که ممکن است به علت انتخاب نامناسب تعداد خوشه‌ها (k) نیز باشد [۸] (جدول ۱).

در این مطالعه تجزیه و تحلیل با استفاده از نرم افزار R نسخه ۴،۱،۳ انجام شد.



نمودار ۱ | درصد فراوانی مرگ در ۶ گروه عمده سنی

گروه سنی ۲۵ تا ۴۴ بیشترین جمعیت را دارا بود اما به ازای هزار نفر ۹۹ مرگ در این بازه مشاهده شد. کمترین میزان مرگ در رده سنی ۲ تا ۱۴ سال با ۳۷ نفر مشاهده شد. رتبه یک علت مرگ مرتبط با بیماری‌های گردش خون بود که ۴۳ درصد از علل مرگ را در بر داشت. پس از آن به ترتیب، نتوپلاسماها با ۱۷ درصد و بیماری‌های دستگاه تنفسی با ۱۰

باتوجه به اینکه تعداد مرگ خام تابعی از جمعیت است، به همین دلیل شاخص مرگ در ۱۰۰۰۰۰ نفر محاسبه شد. به این صورت که تعداد مرگ در هر گروه سنی در عدد ۱۰۰۰۰۰ ضرب و بر جمعیت متعلق به آن گروه سنی تقسیم شد. در گروه سنی زیر یک‌سال به ازای هر ۱۰۰۰۰۰ نفر ۶۸۸ فوت و در گروه سنی بالای ۶۵ سال به ازای هر صد هزار نفر ۴۷۰۵ مرگ مشاهده شد.

جدول ۱ | آمار توصیفی افرادی که به خدمات بستری نیاز داشتند

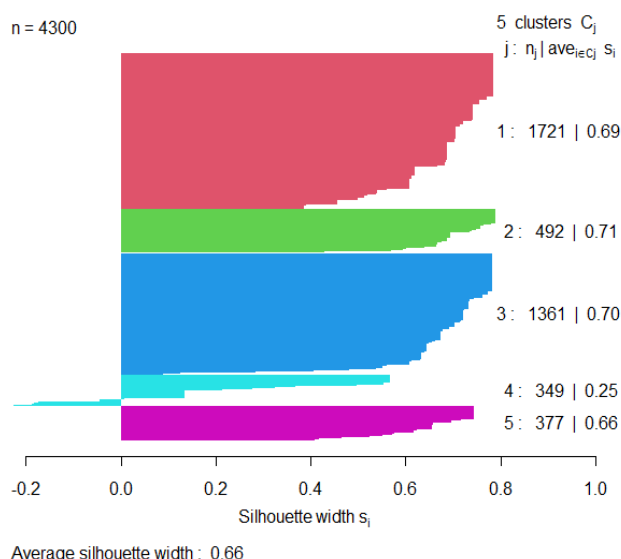
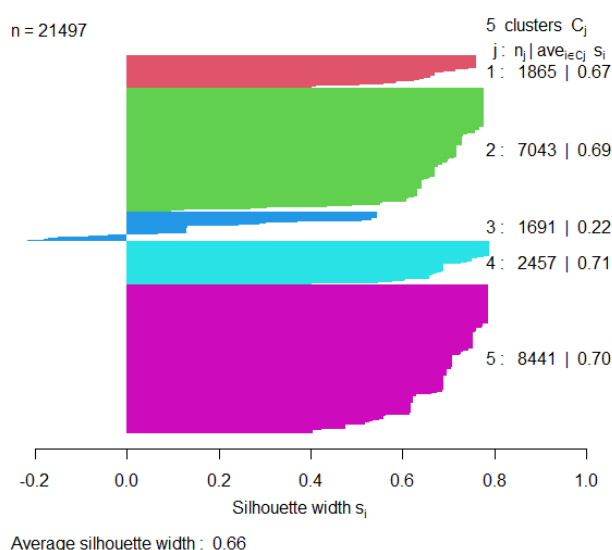
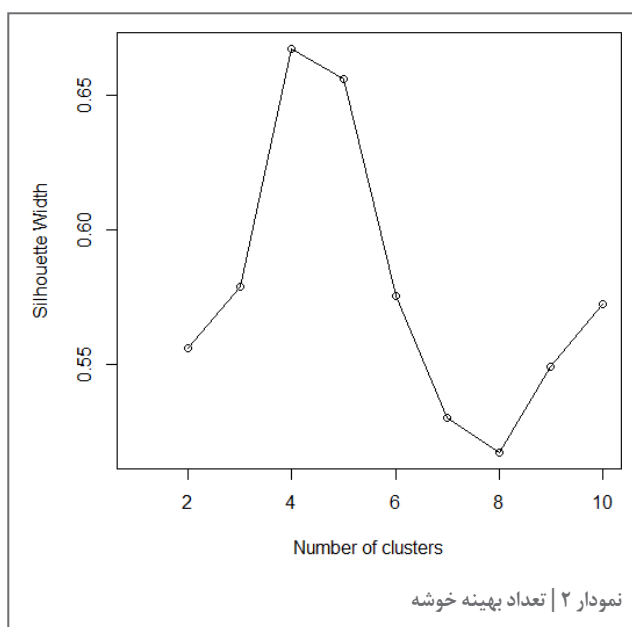
متغیر		PAM (درصد)					CLARA (درصد)				
		خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴	خوشه ۵	خوشه ۱	خوشه ۲	خوشه ۳	خوشه ۴	خوشه ۵
جنسیت	زن	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۲۴	۱۰۰
	مرد	۰	۰	۰	۰	۰	۰	۰	۰	۷۶	۰
سن (سال)	زیر یک‌سال	۸	۱	۷۴	۱	۰	۰	۰	۱	۷۳	۷
	۱-۲	۲	۲	۱۱	۲	۰	۰	۰	۲	۱۱	۳
	۳-۱۴	۱	۲	۱۶	۳	۰	۰	۰	۳	۱۶	۱
	۱۵-۲۴	۵	۷	۰	۹	۱۱	۱۲	۸	۷	۰	۶
	۲۵-۴۴	۱۶	۲۱	۰	۲۱	۲۹	۲۸	۲۴	۲۰	۰	۱۴
	۴۵-۶۴	۶۸	۶۷	۰	۶۴	۵۹	۶۰	۶۳	۶۸	۰	۶۸
وضعیت سکونت	بیش از ۶۵ سال	۱۰۰	۰	۹	۹۱	۱۰۰	۰	۰	۱۰۰	۷	۱۰۰
	روستایی	۰	۱۰۰	۰	۹۱	۱۰۰	۰	۰	۱۰۰	۹۳	۰
علت مرگ	شهری	۰	۱۰۰	۰	۹۱	۱۰۰	۰	۰	۱۰۰	۹۳	۰
	اختلالات روانی و رفتاری	۱	۰	۰	۱	۱	۱	۱	۱	۰	۱
	آسیب‌های ناشی از مسمومیت و سایر عواقب قطعی و علل خارجی آن	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
	بارداری زایمان دوران نفاس	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱
	بیماری‌های غدد درون‌ریز، تغذیه‌ای و سوخت و ساز	۵	۹	۲	۴	۷	۷	۴	۹	۱	۶
	بیماری‌های اندام‌های خون‌ساز و اختلالات خاص دیگر	۰	۰	۰	۰	۰	۱	۰	۰	۱	۰
	بیماری‌های پوست و بافت زیرجلدی	۰	۰	۰	۰	۰	۰	۰	۰	۰	۱
	بیماری‌های دستگاه تناسلی و ادراری	۲	۳	۱	۳	۳	۲	۲	۴	۱	۲
	بیماری‌های دستگاه تنفس	۱۰	۱۲	۳	۸	۱۱	۱۱	۸	۱۱	۳	۱۰
	بیماری‌های دستگاه عصبی	۲	۳	۳	۲	۲	۲	۳	۳	۳	۱
	بیماری‌های دستگاه عضلانی، اسکلتی و بافت همبند	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
	بیماری‌های دستگاه گوارش	۲	۳	۱	۲	۳	۳	۱	۴	۲	۳
	بیماری‌های عفونی و انگلی	۲	۳	۱	۲	۳	۲	۲	۳	۲	۳
	بیماری‌های گردش خون	۴۷	۴۰	۳	۴۵	۴۰	۴۰	۴۴	۴۰	۳	۴۳
	حوادث ترافیکی	۲	۲	۸	۶	۴	۳	۷	۲	۷	۱
	(حوادث غیرمترقبه (سایر عوامل خارجی	۳	۳	۷	۷	۵	۵	۶	۲	۸	۳
	شرایط خاص که منشأ قبل از دوران تولد دارد	۳	۰	۴۴	۰	۰	۰	۰	۰	۴۵	۴
	ناهنجاری‌های مادرزادی تغییر شکل‌ها و ناهنجاری‌های کروموزومی	۳	۰	۱۹	۰	۰	۰	۱	۰	۱۹	۳
	نشانه‌ها و علائم و یافته‌های غیرطبیعی بالینی و آزمایشگاهی که در جایی دیگر طبقه‌بندی نشده‌اند	۲	۳	۳	۱	۳	۴	۱	۳	۲	۲
	نتوپلاسماها	۱۵	۱۷	۴	۱۸	۱۷	۱۸	۱۹	۱۷	۴	۱۸

با دو الگوریتم PAM و CLARA تقریباً یکسان به دست آمد. به همین دلیل در این مطالعه تنها نتایج خوشه‌بندی با الگوریتم PAM گزارش شد. خوشه اول شامل زنان روستایی که ۶۸ درصد آنها در رده سنی بیش از ۶۵ سال قرار داشتند و علل اصلی مرگ به ترتیب، بیماری‌های گردش خون، نئوپلاسم‌ها، بیماری‌های دستگاه تنفسی و بیماری‌های غدد درون‌ریز، تغذیه‌ای و سوخت و ساز بود. خوشه دوم شامل زنان شهری که ۶۷ درصد آنها در رده سنی بیش از ۶۵ سال قرار داشتند و دلایل اصلی مرگ با همان ترتیب خوشه اول بود. خوشه سوم ۷۴ درصد مردان و ۲۶ درصد زنان را شامل شد و بیشترین تعداد متوفیان در جمعیت شهری و با رده سنی زیر یک سال بودند و به ترتیب علل اصلی مرگ، شرایط خاص با منشأ قبل از دوران تولد و ناهنجاری‌های مادرزادی بودند. در این خوشه نیز ۲۶ درصد رده سنی ۲-۲۴ سال قرار داشتند که علت مرگ حوادث ترافیکی و غیرمترقبه بود. خوشه چهارم شامل مردان روستایی بود که ۶۴ درصد آنها در رده سنی بیش از ۶۵ سال قرار داشتند و علل اصلی مرگ به ترتیب بیماری‌های گردش خون، نئوپلاسم‌ها، بیماری‌های دستگاه تنفسی و حوادث غیرمترقبه بودند. خوشه پنجم که بزرگ‌ترین خوشه این مطالعه بود، شامل مردان شهری بود که ۵۹ درصد آنها در رده سنی بیش از ۶۵ سال قرار داشتند و بیشترین علل مرگ به ترتیب بیماری‌های گردش خون،

درصد در رتبه‌های دوم و سوم قرار داشتند.

در این مطالعه، خوشه‌بندی افراد با ۴ متغیر، سن، جنسیت، محل سکونت و علت مرگ انجام شد. دسته‌بندی علل مرگ به دلیل وسیع بودن طیف آنها، براساس ۲۱ علت اصلی مرگ کتاب ICD-10، صورت گرفت.

الگوریتم PAM کل مجموعه مشاهدات و الگوریتم CLARA تعداد ۴۳۰۰ نمونه از کل مجموعه داده را برای خوشه‌بندی مورد استفاده قرار داد. باتوجه به جدول ۲ نتایج حاصل از خوشه‌بندی



نمودار ۳ | ارزیابی کیفیت خوشه‌بندی با الگوریتم PAM (چپ) و CLARA (راست)

اما PAM کل مشاهدات مجموعه داده را در نظر می‌گیرد [۸]. مطالعه‌ای که توسط اونگکولانن در سال ۲۰۱۷ به منظور بررسی تفاوت‌های مرگ و میر براساس سطوح اجتماعی و اقتصادی مردم تایلند انجام شد، از دو روش خوشه‌بندی K-means و تجزیه و تحلیل مولفه اصلی برای بررسی ۳۱ علت مرگ استفاده شد. با انتخاب ۶ خوشه به ترتیب بیشترین علت مرگ عفونت خونی، بیماری‌های قلبی-عروقی، سرطان کبد و صفرا، التهاب ریه، حوادث ترافیکی و غیرمترقبه و سرطان‌های دیگر و ... بود [۱۳] که با بیشترین علل مرگ و میر در مطالعه حاضر همخوانی دارد.

در مطالعه کاظمی‌پور که در سال ۲۰۰۳ به منظور بررسی علل تعیین‌کننده مرگ و میر در شهر تهران در سال ۱۳۷۵ انجام شد، علل عمده مرگ و میر به ترتیب عبارت بودند از: بیماری‌های قلبی و دستگاه گردش خون (۴۴ درصد)، سرطان و سایر تومورها (۱۰ درصد)، سوانح و مسمومیت‌ها (۹ درصد)، بیماری‌های نوزادان (۶ درصد)، بیماری‌های دستگاه تنفس (۵ درصد) و سایر علل مانند ناهنجاری‌های مادرزادی و ... [۱۴] که با بیشترین علل مرگ و میر در مطالعه حاضر همخوانی دارد.

در این مطالعه، در همه خوشه‌ها به جز خوشه سوم، اولین علت مرگ، بیماری‌های گردش خون بود. در این خوشه‌ها بیشتر افراد بالای ۴۵ سال سن داشتند. همان‌طور که در کشورهای توسعه‌یافته بیماری‌های قلبی-عروقی هنوز هم اولین علت مرگ را به خود اختصاص داده [۱۵] و یک‌سوم افراد بالای ۳۵ سال به دلیل بیماری‌های قلبی-عروقی فوت می‌کنند [۱۶، ۱۷].

در این مطالعه، در همه خوشه‌ها به جز خوشه سوم، دومین علت مرگ، انواع سرطان‌ها بود. باتوجه به توزیع سنی و قرار گرفتن بیش از ۶۰ درصد افراد بالای ۶۵ سال و همچنین بیش از ۲۰ درصد افراد بین ۴۵ تا ۶۵ سال در این خوشه‌ها، شاهد درصد بالای ابتلا به سرطان در سنین میان‌سالی هستیم که زنگ خطری برای سیاست‌گذاران سلامت است. ۴ درصد از علل مرگ در خوشه سوم سرطان‌ها بودند که ۷۴ درصد افراد زیر یک‌سال و ۱۱ درصد افراد زیر ۱۴ سال را شامل می‌شد و نشان‌دهنده وجود این بیماری در سنین پایین است. در مطالعه ساسانی‌پور و همکاران که به‌منظور مشخص شدن میزان از دست‌دادن نیروی کار در کشور انجام

نئوپلاسم‌ها، بیماری‌های دستگاه تنفسی و بیماری‌های غدد درون‌ریز، تغذیه‌ای و سوخت و ساز بودند.

زمان محاسبه پس از اجرای خوشه‌بندی با الگوریتم‌های PAM و CLARA، به ترتیب حدود ۳/۵ دقیقه و ۵۳ ثانیه بود. خوشه‌بندی با الگوریتم CLARA، زمان محاسبه را تقریباً ۶ برابر کاهش داد.

باتوجه به نمودار ۳، میانگین پهنای سیلوئت در دو الگوریتم PAM و CLARA یکسان و برابر با ۰/۶۶ به‌دست آمد که با توجه به جدول ۱ نشان‌دهنده خوشه‌بندی با ساختار منطقی و مناسب بود.

بحث

باتوجه به نتایج به‌دست آمده، مشخص شد به جز خوشه سوم که بیشترین علل مرگ نوزدان با منشأ قبل از دوران تولد و ناهنجاری‌های مادرزادی مرتبط بود، در تمامی خوشه‌ها، در هر دو جنس (زن و مرد) و در هر دو نوع منطقه سکونت (شهری و روستایی)، بیشترین علل مرگ در افراد بالای ۶۵ سال و به ترتیب بیماری‌های گردش خون، نئوپلاسم‌ها و بیماری‌های دستگاه تنفس بود. در این خوشه‌ها حدود ۲۰ درصد از متوفیان نیز در رده سنی ۴۵ تا ۶۴ سال قرار داشتند. هر یک از الگوریتم‌های خوشه‌بندی با بهینه‌سازی پارامترهای از پیش تعریف‌شده، داده‌های مشابه را در گروه‌های مجزا از هم سازمان‌دهی می‌کنند. باوجود پیدایش الگوریتم‌های خوشه‌بندی نوین و کارآمد، روش فراگیری وجود ندارد که بتوان عملکرد آن را در همه داده‌ها یا حتی مجموعه‌ای از داده‌ها با ویژگی‌های یکسان، رضایت‌بخش تلقی کرد. از طرفی عملکرد الگوریتم‌های خوشه‌بندی به شدت تحت تأثیر ساختار پنهان داده‌ها قرار دارد و اجرای الگوریتم‌های مختلف روی مجموعه داده یکسان، می‌تواند منجر به نتایج کاملاً متفاوت شود. حتی ممکن است اجرای یک الگوریتم خوشه‌بندی واحد اما با پارامترهای ورودی مختلف، نتایج متفاوتی را بر روی یک مجموعه داده ایجاد کند [۱۲]. اما با توجه به معیار سیلوئت برای ارزیابی عملکرد الگوریتم‌ها، مشخص شد الگوریتم CLARA نتایجی مشابهی با الگوریتم PAM دارد، مطالعه کوفمن در سال ۲۰۰۹ نیز نشان داد سیلوئت‌های به‌دست آمده از الگوریتم CLARA، مشابه PAM است و تنها تفاوت در این است که CLARA از مشاهدات نمونه استفاده می‌کند

موضوع نشان-دهنده اهمیت بالای مراقبت از مادر در دوران بارداری و مراقبت از مادران پرخطر است.

دومین علل مرگ در خوشه سوم حوادث غیرمترقبه و حوادث ترافیکی بود که در مجموع ۱۵ درصد از علل مرگ را شامل شد. کودکان به دلیل آسیب پذیر بودن و ناتوانی در مراقبت از خود، بیشتر در معرض حوادث غیرمترقبه مانند افتادن از ارتفاع، برخورد با اشیاء و... هستند. مطالعه ویلیام بیکت در مطالعه اپیدمیولوژی بر روی ۹۹۰ کودک آسیب دیده از حوادث غیرمترقبه نشان داد ۶۱ درصد علت حادثه مربوط به افتادن کودک است. افتادن از مبل، صندلی اتومبیل و کالسکه کودک بیشترین فراوانی را داشتند [۲۴].

در خوشه سوم نیز ۲۶ درصد مرگ ها در گروه زنان و ۷۴ درصد در گروه مردان بود. این نشان دهنده آسیب پذیری تر بودن کودکان با جنسیت مرد است. گروه تحقیقاتی کلین در مطالعه بر روی علت فوت کودکان زیر ۱۲ سال نشان داد که ۶۰ درصد فوت ها در جنسیت مرد رخ داده است [۲۳].

نتیجه گیری

این مطالعه کل افراد فوت شده را در ۵ خوشه اصلی دسته بندی کرد که بیشترین فراوانی علل مرگ بیماری های شرایط خاص قبل تولد، بیماری های مربوط به سندروم متابولیک، نئوپلاسم ها، بیماری های گردش خون و بیماری های تنفسی بودند. توزیع سنی همان طور که در متن اشاره شد در این گروه ها متفاوت بود. به جز خوشه سوم که بیماری های با شرایط خاص بدو تولد را شامل می شد، بیشترین فراوانی مرگ ها در در گروه سنی بالای ۶۵ سال بود. در مطالعات بعدی هر کدام از خوشه ها بایستی به دقت بررسی و نوع بیماری های هر خوشه رصد شود. به عنوان مثال در خوشه نئوپلاسم ها پاسخ به این سوال که کدام نوع سرطان در کدام گروه سنی بیشترین فراوانی را دارد یکی از موارد پراهمیت در سیاست گذاری سلامت است. همچنین نتایج این مطالعه نشان داد که دو الگوریتم PAM و CLARA عملکرد مشابه داشتند اما سرعت اجرای الگوریتم CLARA بیشتر از الگوریتم PAM بود که این مسئله با افزایش روزافزون حجم اطلاعات و همچنین محدودیت فضای ذخیره سازی آن، چالشی انکارناپذیر در امر پردازش اطلاعات است. در نتیجه این پژوهش زمینه ساز پژوهش های بیشتر در راستای

شد، سیر صعودی مرگ به دلیل سرطان ها و تومورها در گروه سنی کار یعنی ۱۵ تا ۶۵ سال بود، به طوری که در سال ۹۵ حدود ۷۵ درصد مرگ به علت سرطان در این گروه سنی بوده است [۱۲].

در این مطالعه در همه خوشه ها به جز خوشه سوم، بیماری های تنفسی در رتبه سوم علل مرگ قرار داشت. در سال ۲۰۱۶ و ۲۰۱۷ بیماری های دستگاه تنفسی چهارمین علت مرگ در جهان بود [۱۸، ۱۹]. در مطالعه اپیدمیولوژیک ترکاشوند و همکاران بیماری های دستگاه تنفسی چهارمین علت مرگ سالمندان بالای ۶۵ سال در ایران بود [۲۰]. در مطالعه حاضر افراد بالای ۶۵ سال بیشترین فراوانی را در علت مرگ با بیماری های تنفسی داشتند.

در خوشه اول، دوم و پنجم، چهارمین علت مرگ مربوط به بیماری های غدد درون ریز، تغذیه ای و سوخت و ساز بود که جمعیت شهری در هر دو جنس (زن و مرد) و زنان روستایی را شامل می شد. در این خوشه ها بیشترین متوفیان بالای ۴۵ سال داشتند. بیماری های غدد درون ریز شامل دسته ای از ریسک فاکتورها مانند چاقی، مقاومت به انسولین، فشار خون بالا، اختلال در سطح سرم چربی هاست که ریسک ابتلا به دیابت نوع ۲ و بیماری های قلبی-عروقی را بالا می برد [۲۱]. در مطالعه پاول هس که در سال ۲۰۱۷ به منظور بررسی رابطه بین ابتلا به سندروم متابولیک و مرگ به دلیل بیماری قلبی-عروقی انجام داد، شانس مرگ بر اثر بیماری قلبی-عروقی در افرادی که مبتلا به سندروم متابولیک بودند ۴ و برای سایر افراد ۲/۳ بود [۲۲]. در این مطالعه هم تفاوتی بین ابتلا در زنان و مردان مشاهده نشد. گروه سنی ۴۵ تا ۶۵ سال ۲۴ درصد این گروه را شامل می شد که نشان دهنده درصد بالای ابتلا در این گروه سنی است.

در خوشه سوم که ۸۵ درصد افراد کمتر از ۱۴ سال داشتند، بالاترین درصد مرگ به علت بیماری هایی با منشأ قبل از دوران تولد بود. این بیماری ها شامل طیف وسیعی (بیماری هایی که جنین و مادر تحت تاثیر بیماری های عفونی و انگلی، تنفسی، فشار خون، اختلالات تغذیه ای مادران و...) می شوند. در مطالعه گذشته نگر گروه تحقیقاتی کلین در سال ۲۰۱۳ بر روی علت فوت کودکان زیر ۱۲ سال، نشان داد ۶۰ درصد علت مرگ کودکان در این گروه سنی مربوط به بیماری هایی با منشأ قبل از دوران تولد است [۲۳]. این

- Statistical Data Analysis based on the L1 Norm. Y Dodge, Ed. 1987:405-16.
- Gower J. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biometrics*. 1971;27:857-71. doi: [10.2307/2528823](https://doi.org/10.2307/2528823).
 - Sasanipour M, Mohebi M, Shahbazin S. Changes in the Mortality pattern in Iran with an emphasis on the Working age population in recent decade. *Journal of Population Association of Iran*. 2016;11(21):190-210. [Persian]
 - Aungkulanon S, Tangcharoensathien V, Shibuya K, Bundhamcharoen K, Chongsuvivatwong V. Area-level socioeconomic deprivation and mortality differentials in Thailand: results from principal component analysis and cluster analysis. *Int J Equity Health*. 2017;16(1):117. doi: [10.1186/s12939-017-0613-z](https://doi.org/10.1186/s12939-017-0613-z).
 - Kazemipour SH. A review of mortality rates and causes in greater tehran in the Persian calendar year (1996-1997). *Nameh-Ye Olum-E Ejtemai*. 2003;11(1):1-30. [Persian]
 - Roger VL. Epidemiology of myocardial infarction. *Med Clin North Am*. 2007;91(4):537-52; ix. doi: [10.1016/j.mcna.2007.03.007](https://doi.org/10.1016/j.mcna.2007.03.007).
 - Rosamond W, Flegal K, Furie K, Go A, Greenlund K, Haase N, et al. Heart disease and stroke statistics--2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*. 2008;117(4):e25-146. doi: [10.1161/CIRCULATIONAHA.107.187998](https://doi.org/10.1161/CIRCULATIONAHA.107.187998).
 - Nichols M, Townsend N, Scarborough P, Rayner M. Cardiovascular disease in Europe 2014: epidemiological update. *Eur Heart J*. 2014;35(42):2950-9. doi: [10.1093/eurheartj/ehu299](https://doi.org/10.1093/eurheartj/ehu299).
 - Heron M. Deaths: Leading Causes for 2016. *Natl Vital Stat Rep*. 2018;67(6):1-77.
 - Kochanek KD, Murphy SL, Xu J, Arias E. Deaths: final data for 2017. 2019.
 - Torkashvand Moradabadi M, Abbasi M. An Epidemiological study of mortality among elderly in Iran using Years of Life Lost (YLL) index during 2011 to 2017. *Payesh*. 2020;19(1):85-97. [Persian] doi: [10.29252/payesh.19.1.85](https://doi.org/10.29252/payesh.19.1.85).
 - O'Neill S, O'Driscoll L. Metabolic syndrome: a closer look at the growing epidemic and its associated pathologies. *Obes Rev*. 2015;16(1):1-12. doi: [10.1111/obr.12229](https://doi.org/10.1111/obr.12229).
 - Hess PL, Al-Khalidi HR, Friedman DJ, Mulder H, Kucharska-Newton A, Rosamond WR, et al. The metabolic syndrome and risk of sudden cardiac death: the atherosclerosis risk in communities study. *J Am Heart Assoc*. 2017;6(8):e006103. doi: [10.1161/JAHA.117.006103](https://doi.org/10.1161/JAHA.117.006103).
 - Celine T. A study on childhood death at a tertiary care level in ernakulam district. *Annals of medical and health sciences research*. 2013;3(3):407-11. doi: [10.4103/2141-9248.117945](https://doi.org/10.4103/2141-9248.117945)
 - Pickett W, Streight S, Simpson K, Brison RJ. Injuries experienced by infant children: a population-based epidemiological analysis. *Pediatrics*. 2003;111(4):e365-e70. doi: [10.1542/peds.111.4.e365](https://doi.org/10.1542/peds.111.4.e365)

به کارگیری الگوریتم‌های خوشه‌بندی مناسب با توجه به حجم پایگاه داده است.

تشکر و قدردانی: مقاله حاضر حاصل بخشی از طرح تحقیقاتی مصوب در دانشگاه علوم پزشکی مشهد است. بدین وسیله از همکاری این دانشگاه قدردانی می‌شود.

تأییدیه اخلاقی: پژوهش حاضر با کد اخلاق IR.MUMS.REC.1399.587 مورد تأیید کمیته اخلاق دانشگاه علوم پزشکی مشهد قرار گرفته است.

تضاد منافع: نویسندگان این مقاله هیچ گونه تضاد منافی را اعلام نکرده‌اند.

سهم نویسندگان: مریم سالاری؛ متدولوژیست و سرپرست تیم پژوهش (۵۰ درصد)، سارا صباغیان؛ تحلیل داده (۲۰ درصد)، محمدتقی شاکری، سید مسعود ساداتی، محمد وجدانی، پرستو گلپور و صدیقه رستاقی؛ نگارش، جمع‌آوری و پاک‌سازی داده، مشاور بالینی (۳۰ درصد) در این مقاله نقش داشته‌اند.

منابع مالی: این پژوهش با حمایت دانشگاه علوم پزشکی مشهد صورت پذیرفته است.

References

- World Health Organization. International statistical classification of diseases and related health problems: World Health Organization; 2004.
- <http://www.sbm.ac.ir/uploads>.
- Public Health England. Chapter 2: Major causes of death and how they have changed. 2017.
- Vallin J, Schofield R, Reher D, Bideau A. Mortality in Europe from 1720 to 1914: long-term trends and changes in patterns by age and sex. *The Decline of Mortality in Europe*. 1991:38-67.
- Meslé F, Vallin J, Andreyev Z. Mortality in Europe: the divergence between east and west. *Population*. 2002;57(1):157-97. doi: [10.3917/pope.201.0157](https://doi.org/10.3917/pope.201.0157).
- Aghamohammadi S, Kazemi E, Khosravi A, Kazemeini H. The trend of ten leading causes of death in the Islamic Republic of Iran, 2006-2011. *Iranian Journal of Epidemiology*. 2017;12(4):1-11. [Persian]
- Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems. 2011:83-124. doi: [10.1016/B978-0-12-381479-1.00003-4](https://doi.org/10.1016/B978-0-12-381479-1.00003-4).
- Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons; 2009.
- Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*. 2006;5(4):475-504. doi: [10.1007/s10852-005-9022-1](https://doi.org/10.1007/s10852-005-9022-1).
- Kaufman L, Rousseeuw PJ. Clustering by means of medoids.