تعیین الگوی توزیع واقعه مرگ با استفاده از تکنیک های داده کاوی در استان گلستان از سال ۱۳۸۶ تا ۱۳۸۸

فاطمه باقری[۱]، فاطمه آهنگری[۲]، ناصر بهنام پور[۳]

۱. مربی، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه گلستان، گرگان، ایران
۲. دانش آموخته مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه گلستان، گرگان، ایران
۳. استادیار، گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی گلستان، گرگان، ایران

چکیده

زمینه و هدف: بررسی وضعیت مرگ و میر در یک جمعیت به عنوان یکی از روش های مناسب تعیین سلامت مورد توجه بوده است، گرچه با مشکلاتی از قبیل عدم اطمینان به صحت و کیفیت داده ها و روش جمع‌آوری آن روبرو بوده است. راه اندازی نظام های ثبت مرگ و میر با استفاده از کدهای بین المللی طبقه بندی بیماری ها و ادغام اطلاعات مرگ و میر توسط سازمان های مسئول تا حد زیادی مشکلات قبلی را مرتفع ساخته است. در این پژوهش با در نظر گرفتن مجموعه‌ای از پارامترها، ابتدا جامعه مورد بررسی به دو گروه داده‌های مربوط متوفیان زیر یکسال و بالای یکسال تقسیم شدند سپس این دو گروه داده با استفاده از روش K-means خوشه بندی شدند تا گروه‌های مختلف شناسایی شوند. همچنین با استفاده از روش درخت تصمیم به کشف مدل‌ها و الگو های مؤثر پرداخته شد. در آخر از الگوریتم شبکه‌های عصبی که یکی از کارکردهای آن نشان دادن ویژگی‌ها به ترتیب اهمیتشان است، استفاده شد.

روش بررسی: در این پژوهش به بررسی ۱۲۸۶۵ داده ی مربوط به متوفیان استان گلستان پرداخته شده است. داده های مورد استفاده در این مقاله مربوط به اول فروردین ۱۳۸۸ تا پایان فروردین ۱۳۸۹ می باشند. این داده‌ها از مرکز بهداشت استان گلستان تهیه شدند. ویژگی‌های مورد استفاده در این داده‌ها سن متوفی، جنسیت متوفی، علت فوت، منطقه سکونت، محل فوت و وقوع فوت می‌باشند. برای خوشه بندی در این پژوهش از الگوریتم K-means استفاده شد، دسته بندی نیز به کمک الگوریتم های درخت تصمیم و الگوریتم شبکه‌های عصبی انجام پذیرفت سپس نتایج و قوانین حاصل از الگوریتم ها استخراج شدند. در ضمن به جهت تفاوت ماهیت علل مرگ و میر در نوزادان و بزرگسالان از نظر پزشکان، لیست متوفیانی که دارای سن زیر یکسال هستند در فایلی جداگانه قرار می گیرند و بررسی می شوند.

یافته ها: در روش خوشه‌بندی، تعداد بهینه خوشه‌ها با استفاده از اندازه‌گیری شاخص Dunn، هشت خوشه برای داده های زیر یکسال و هفت خوشه برای داده‌های بالای یکسال به دست آمد. از میان چهار الگوریتم درخت تصمیم نظیر C5.0، QUEST، CHAID و CART، الگوریتم C5.0 با نرخ تشخیص۷۷/۳۷ درصد برای زیر یکسال و ۹۶/۸۶ درصد برای بالای یکسال بهترین روش شناخته شد. با اجرای الگوریتم شبکه‌های عصبی ویژگی‌های سن متوفی، محل فوت و جنسیت از جمله ویژگی‌های با اهمیت در پیش‌بینی شناخته شدند.

نتیجه گیری: این پژوهش با در نظر گرفتن عوامل تأثیرگذار در فوت افراد و مبنا قرار دادن استاندارد طبقه بندی بین‌المللی بیماری‌ها، به خوشه بندی داده‌ها پرداخت و به دنبال یافتن الگوی مرگ و میر برای افراد زیر یکسال و بالای یکسال است. با توجه به صریح بودن و قابلیت فهم بالای درخت تصمیم و قوانین استخراج شده توسط آن، می‌تواند برای متخصصین در این حوزه قابل استفاده قرار گیرد.

کلمات کلیدی: داده کاوی، خوشه بندی، دسته بندی، درخت تصمیم، مرگ و میر

نویسنده مسئول: فاطمه باقری

آدرس: ایران، گرگان، دانشگاه گلستان، دانشکده فنی و مهندسی

ایمیل:f.bagheri@gu.ac.ir

مقدمه

اطلاعات پیرامون علل مرگ، به عنوان ابزار پایش ارتقاء سطح سلامت جامعه و تعیین اولویت های اقدام های بهداشتی سال هاست که بکار گرفته شده است (۱). در کشورهای پیشرفته به دلیل قدمت و دقت کافی نظام ثبت آمارهای جمعیتی بویژه آمارهای حیاتی، تحقیقات فراوانی در این زمینه و چگونگی روند تحولات جمعیتی و ویژگی های منطقه ای مربوط، صورت گرفته است. متأسفانه در ایران این گونه مطالعات متأثر از فقدان اطلاعات کافی و بهنگام، تقریباً به فراموشی سپرده شده است. اکنون چندسالی است که اولین داده های مربوط به باروری و مرگ و میر به تفکیک سن، با قانونی شدن لزوم ثبت موالید و متوفیات در مهلت مقرر توسط سازمان ثبت احوال ارائه می‌شود (۲).

شناخت متغیرهای صرفاً جمعیتی به ویژه مرگ و میر در فضای جغرافیایی یک سرزمین دارای ارزش کاربردی همچون شناسایی نواحی مسئله دار و امکان پیش بینی دقیق تر تحولات آتی خواهد بود. ضمن اینکه متغیرهای مهم مشتق از آن ها یعنی امید به زندگی و میزان مرگ و میر اطفال از متغیرهای تأثیرگذار بر سطح توسعه انسانی هر جامعه است (۳).

بررسی روند علت های مرگ در دو دهه گذشته نشان می دهد که مرگ به علت بیماری های واگیردار سیر نزولی و به علت بیماری‌های غیرواگیر بویژه سرطان ها و سوانح وحوادث، سیر صعودی داشته است. این امر برنامه ریزی برای کنترل و پیش گیری از بیماری‌های غیر واگیردار را بیش از پیش روشن می سازد (۴).

حمید بختیاری در پایان‌نامه خود به تعیین الگو و توزیع واقعه مرگ به روش داده‌کاوی طی ۸۴ ماه از سال های ۱۳۸۳ تا ۱۳۸۹ در استان فارس پرداخت. این پژوهش با استفاده از داده های مرگ و میر ثبت شده در نظام ثبت داده های مرگ که از سال ۱۳۸۳ در کشور براساس کدهای بین المللی طبقه بندی بیماری‌ها (ICD10) راه اندازی شد، انجام پذیرفت. نرم افزار مورد استفاده، نرم افزار WEKA و OLAP cube بوده است (۵).

دکتر حمید شایان در مقاله ای به تحلیل وضعیت مرگ و میر بر اساس شاخص های آماری در استان خراسان رضوی پرداختند و ۱۶ شهرستان استان خراسان رضوی را بر اساس آمار متوفیات سال ۱۳۸۵ از نظر مرگ و میر عمومی، مرگ و میر ویژه سنی، امید به زندگی و مرگ و میر استاندارد مورد مطالعه قرار دادند (۶).

زینب کرمان‌ساروی و همکارانش در مقاله‌ای سعی کردند به پیش‌بینی علل مرگ و میر نوزادان با استفاده از تکنیک‌های داده‌کاوی در سال ۱۳۹۱ بپردازند. آن ها با استفاده از الگوریتم درخت تصمیم C4.5 برای دسته بندی و الگوریتم K نزدیک ترین همسایه و دورترین همسایه در نرم افزارهای WEKA و SPSS روی حدود ۳۰۰ پرونده، دلایل اصلی مرگ و میر نوزادان پس از تولد و امکان پیش بینی عوامل اصلی این رویداد را بررسی و شناسایی کردند (۷).

علیرضا زندی و همکارانش در مقاله‌ای به بررسی و تحلیل مهم ترین بیماری‌های منجر به فوت در شهر تهران با استفاده از تکنیک‌های داده‌کاوی پرداختند. برای این منظور داده های مربوط به آمار فوت شدگان در شهر تهران در سال ۱۳۸۹ مورد بررسی قرار داده شد. در این مقاله از نرم افزار WEKA استفاده شد و الگوریتم های Decision Tree، Kmeans و Apriori مورد استفاده قرار گرفت (۸).

روش بررسی

برای انجام این پژوهش، به مرکز بهداشت استان گلستان مراجعه و پس از انجام مراحل اداری به داده های فوت استان گلستان که در یک فایل Excel ذخیره شده بود، دسترسی پیدا کردیم. حجم داده ها ۱۶۳۸۲ مورد بوده که شامل

اطلاعات متوفیان از فروردین ۱۳۸۴ تا پایان فروردین ۱۳۸۹ در استان می باشد.

از آن جایی که اطلاعات مربوط به سن در فایل اکسل به صورت سال، ماه و روز ثبت شدند، از آنج ایی که از نرم‌افزار Clementine۱۲ در این پژوهش استفاده شد، برای ورود به نرم افزار Clementine نیاز به یکپارچه سازی داده است به همین منظور داده‌ها به دو مجموعه داده تقسیم شدند شامل داده‌های زیر یکسال و داده‌های بالای یکسال. تعداد داده‌های زیریکسال، ۱۱۳۶ رکورد و تعداد داده‌های بالای یکسال ۱۱۷۸۶ رکورد بوده است.

داده های مورد استفاده در این مقاله مربوط به اول فروردین ۱۳۸۸ تا پایان فروردین ۱۳۸۹ می باشند. (بعد از مرحله آماده سازی و تصفیه داده ها)

ویژگی های موجود در این داده ها عبارتنداز: نام و نام خانوادگی، نام پدر، تاریخ تولد، سن متوفی، جنسیت متوفی، تاریخ فوت، علت فوت، توضیحات علت فوت، آدرس، محل سکونت، منطقه سکونت، محل فوت، محل ثبت و وقوع فوت. اما همه ی این اطلاعات مناسب داده کاوی، کنکاش و تحلیل و بررسی نمی باشند. بنابراین ستون های نام و نام خانوادگی، نام پدر، تاریخ تولد، توضیحات علت فوت، آدرس، محل فوت، محل ثبت و وقوع فوت که مورد نیاز ما نیستند از مجموعه داده‌ها کنار گذاشته شدند.

استانداردICD10

بیماری های مورد بررسی در این پژوهش از مرکز بهداشت استان گلستان گرفته شده که خود فهرستی از ثبت نرم افزاری و جدول مرگ و میر برای شرایط اپیدمیولوژیک ایران می باشد. در حال حاضر مرگ و میرها مطابق آخرین تجدید نظر در "سیستم طبقه‌بندی بین المللی بیماری ها" (ICD)(= International Classification of Diseases تجدید نظر دهم)، کدگذاری می شوند (۹). ICD تاکنون ۱۰ بار تجدید نظر شده است (ICD1-ICD10). گروه های کدگذاری شده و تنظیمات مربوطه، از یک تجدید نظر به تجدید نظر دیگر تفاوت دارد و ما در این پژوهش از جدول استاندارد ICD10 استفاده می کنیم. این جدول شامل سه لایه است که در لایه اول آن ۲۲ بیماری اصلی قرار دارند. دو لایه دیگر این جدول نیز شامل دیگر بیماری ها با ذکر جزئیات می باشد که مجموعاً تعداد بیماری ها در این سه لایه را می توان حدود ۱۵۰ بیماری تخمین زد. به دلیل بالا بودن تعداد بیماری ها در لایه های دوم وسوم، داده های موجود در این پژوهش، با توجه به بیماری ها در لایه اول که شامل ۲۲ بیماری می باشد، کدگذاری شده است (۱۰).

مواد و روش‌ها

نرم افزاری که در این پژوهش مورد استفاده قرار گرفته است، نرم افزار Clementine 12 می باشد. این نرم افزار، ابزار توانمندی است که اغلب تکنیک های داده کاوی را در بر دارد.

خوشه‌بندی

روش خوشه‌بندی به منظور شناخت بهتر جامعه مورد مطالعه استفاده شده است. روش خوشه‌بندی جزء روش‌های غیر نظارتی محسوب می‌شود چرا که برای الگوریتم‌های خوشه‌بندی ویژگی دسته تعریف نمی‌شود و رکوردها برچسب خاصی ندارند. الگوریتم خوشه‌بندی اطلاعاتی را که ویژگی‌های نزدیک به هم و مشابه دارند، در دسته‌های جداگانه که به آن خوشه گفته می‌شود قرار می‌دهد (۱۱). در اینجا تمرکز روی گروه‌هایی از اشیا است که به هم شبیه هستند، تا با کشف این شباهت‌ها بتوان رفتارها را بهتر شناسایی کرده و بر مبنای این شناخت بهتر تصمیم‌گیری نمود. از این الگوریتم در مجموعه داده‌های بزرگ و در

مواردی که تعداد ویژگی‌های داده زیاد باشد استفاده می‌شود (۱۲). یکی از مشکلاتی که در روش خوشه بندی وجود دارد، تعیین تعداد خوشه های بهینه است. هرچه شباهت میان داده‌های در یک خوشه (شباهت درون خوشه‌ای) بیشتر و تفاوت آن ها با سایر خوشه‌ها (فاصله بین خوشه‌ای) بیشتر باشد خوشه بندی دارای کیفیت بیشتری خواهد بود. در این پژوهش از الگوریتم K-means برای خوشه بندی داده ها استفاده شده است. الگوریتم K-means به شرح زیر می باشد:

انتخاب K داده به عنوان مرکز خوشه.

تعیین فواصل بقیه داده ها با مراکز خوشه ها.

قرار گیری داده ها در خوشه هایی که به مرکز آن خوشه ها نزدیکترند.

محاسبه میانگین هر خوشه به عنوان مرکز جدید خوشه.

تکرار مرحله دوم تا چهارم تا رسیدن به عدم تغییر در خوشه ها.

روش خوشه بندی K-means بستگی به عواملی چون تعداد خوشه و روش تعیین فاصله بین خوشه ها دارد. یکی از مهم ترین مسائل در خوشه بندی انتخاب تعداد خوشه های مناسب می باشد. برای ارزیابی کیفیت خوشه بندی، می‌توان از شاخص های سنجش کیفیت استفاده کرد. به کمک این شاخص ها می توان تعداد بهینه ی خوشه ها برای خوشه بندی را تعیین کرد. در این پژوهش از شاخص Dunn استفاده شده است که یکی از متداول ترین شاخص های مورد استفاده است. روابط زیر محاسبه شاخص Dunn را نشان می دهند (۱۳).

$$D = \min_{i=1\ldots x_c} \left\{ \min_{j=i+1\ldots x_c} \left(\frac{d(c_i,c_j)}{\max_{k=1\ldots x_c}(diam(c_k))} \right) \right\} \quad (1)$$

که $d(c_i,c_j)$ و $diam(c_k)$ در آن با روابط زیر محاسبه می‌شوند:

$$d(c_i,c_j) = \min_{x\in c_i\, y\in c_i} (d(x,y)) \quad (2)$$

$$diam(c_i) = \max_{x\in c_i\, y\in c_i} (d(x,y)) \quad (3)$$

C_i خوشه‌ی iام، $d(x,y)$ بیانگر فاصله ی بین دو نمونه از داده هاست و $diam(C_i)$ حداکثر فاصله ی درون خوشه iام را بیان می کند.

درخت تصمیم

به منظور پیش‌بینی اصلی ترین علت فوت در داده های زیر یکسال و داده‌های بالای یکسال، از مدل درخت تصمیم استفاده شده است.ساختار درخت تصمیم یک ساختار درختی شبیه فلوچارت است. بالاترین گره در درخت گره‌ی ریشه است و گره‌های برگ، دسته ها یا توزیع دسته‌ها را نشان می‌دهد.

درخت تصمیم یکی از ابزارهای قوی و متداول برای دسته بندی و پیش بینی می باشد. در ساختار درخت تصمیم، پیش بینی به دست آمده از درخت در قالب یکسری قواعد توضیح داده می‌شود. طوری که داده هایی که در برگ‌های انتهایی این درخت قرار می‌گیرند توسط یکی از مقادیر ویژگی هدف برچسب می‌خورند. این مدل به دلیل سهولت در تفسیر نتایج و غیر خطی بودن، نیاز به پیش فرض رابطه خطی بین متغیرهای مستقل و وابسته ندارد (۱۴).

برای استفاده از تکنیک درخت تصمیم از چهار الگوریتم CHAID، QUEST، C5.0، CART برای مجموعه داده های بالای یکسال (بزرگسال) و زیر یکسال (نوزادان) به صورت مجزا استفاده شده است.

الگوریتم C5.0 فقط می تواند متغیرهای هدف طبقه ای را پیش بینی کند. الگوریتم C5.0 در مسائلی مثل داده های مفقود و تعداد فیلدهای ورودی زیاد خوب عمل می کنند. آن ها معمولاً نیازی به زمان طولانی برای تخمین ندارند. به علاوه، درک این الگوریتم ها ساده‌تراز برخی الگوریتم هاست. همچنین روش قدرتمندی برای افزایش دقت دسته‌بندی دارد. سرعت ساخت مدل C5.0 به علت بهره‌گیری از پردازش های موازی بالا است (۱۵).

الگوریتم CART در تجزیه دو زیر گروه را تعیین می کند که هر کدام نیز به دو زیرگروه دیگر تقسیم خواهند شد و این روند ادامه می‌یابد تا زمانی که یکی از معیارهای توقف برآورده شود. تمام تجزیه ها دودویی هستند. درخت CART یک یا چند فیلد ورودی را به تنها یک فیلد خروجی تبدیل می کند. فیلدهای هدف و پیش بینی کننده در نرم‌افزار Clementine12 می توانند از جنس طبقه‌ای یا فاصله‌ای باشند. درک مدل های درخت CART نسبت به سایر مدل ها آسان تر است (تفسیر قواعد استخراج شده از درخت آسان است). برخلاف C5.0، درخت CART علاوه بر فیلدهای خروجی طبقه ای، نوع فاصله ای را نیز می پذیرد (۱۵).

CHAID یا Chi-squared Automatic Interaction Detection یک روش دسته بندی برای ساخت درخت های تصمیم است. در ابتدا CHAID اهمیت هر یک از متغیرهای پیش بینی کننده را برای پیش‌بینی متغیر هدف تعیین می کند. اگر بیش از یک مورد از روابط از لحاظ آماری مهم بودند، CHAID پیش بینی کننده ای را انتخاب می کند که از سایرین مهم تر باشد. متغیرهای پیش بینی کننده و هدف در این الگوریتم می توانند از نوع طبقه ای یا فاصله ای باشند (۱۵).

QUEST یا درخت آماری سریع از روش دسته بندی دودویی برای ساخت درخت تصمیم استفاده می کند. انگیزه اصلی در توسعه این درخت، کاهش زمان پردازش مورد نیاز برای متوقف کردن درخت بزرگ CART با متغیرها و یا نمونه های زیاد است. روش QUEST از یک سری قوانین مبتنی بر آزمون های مشخص برای ارزیابی متغیرهای پیش‌بینی‌کننده در هر گره استفاده می کند. حداقل یک تست باید بر روی هر متغیر پیش بینی کننده در هر گره انجام شود برعکس درخت CART، همه شاخه زنی ها امتحان نمی شوند و برعکس درخت CART و CHAID، ترکیبات متغیرهای طبقه ای در ارزیابی متغیرهای پیش بین برای انتخاب آزموده نمی شوند. این امر باعث سرعت بالای تحلیل می شود. فیلدهای پیش بینی کننده در الگوریتم QUEST می توانند عددی باشند ولی متغیر هدف باید حتماً طبقه ای باشد (۱۵).

کلاس بندی داده ها با درختان تصمیم یک فرایند دو مرحله ای می باشد.در مرحله اول که به آن مرحله آموزش گفته می شود، مدلی براساس یک الگوریتم کلاس بندی منطبق با داده کاوی مربوط به مجموعه آموزشی ساخته می شود. مجموعه آموزشی به صورت تصادفی از پایگاه داده انتخاب می شود. در مرحله دوم یادگیری از طریق یک تابع $y=f(X)$ انجام می شود که می تواند برچسب کلاس هر رکورد x از پایگاه داده را پیش بینی کند (۱۲).

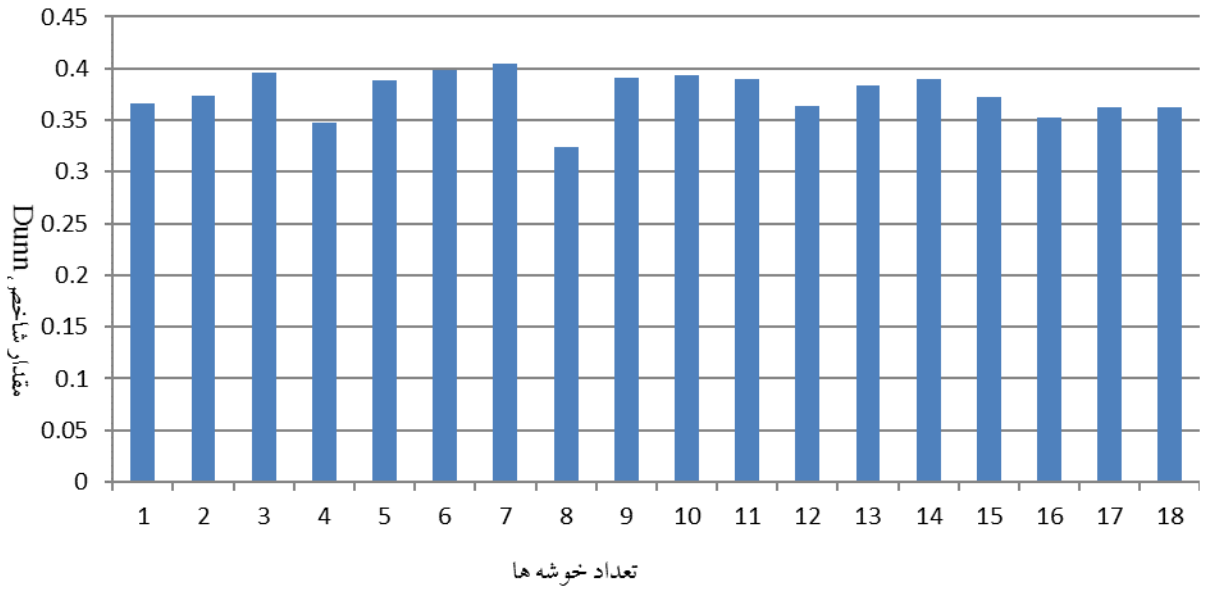
به منظور تعیین میزان صحت درخت تصمیم داده‌ها را به دو بخش دادهای آموزش و آزمون تقسیم شدند. درخت تصمیم با استفاده از داده‌های آموزش مدل را می‌سازد و مدل ساخته شده بر روی داده های آزمون مورد تست قرار می‌گیرد.

درصد نمونه‌هایی که ویژگی هدف آن ها توسط مدل، درست تشخیص داده شده بود دقت مدل را بیان می‌کند. در این پژوهش ابتدا داده های مربوط به متوفیان زیر یکسال بررسی می شوند چون تعداد داده های این مجموعه کم است، ۷۰درصد داده ها به آموزش و ۳۰ درصد بقیه به آزمون اختصاص یافت. برای داده‌های بالای یکسال چون تعداد داده‌ها مقدار قابل قبول هستند، ۶۰ درصد داده‌ها به آموزش و ۴۰ درصد بقیه به آزمون یافت.

یافته ها

خوشه‌بندی

الگوریتم خوشه بندی K-means را برای خوشه های مختلف از ۲ خوشه تا ۲۰ خوشه، هم برای داده‌های زیر یکسال و هم داده‌های بالای یکسال اجرا شد و برای هر دو گروه داده به صورت مجزا شاخص Dunn اندازه‌گیری شد. شکل ۱ و۲ نمودار شاخص Dunn برای داده‌های زیر یکسال و داده‌های بالای یکسال را نشان می‌دهد. برای داده‌های زیر یکسال بیشترین مقدار این شاخص مربوط به خوشه‌بندی با ۸ خوشه است، بنابراین تعداد بهینه خوشه ها برای داده‌های زیر یکسال ۸ خوشه می‌باشد. در ادامه به تحلیل قوانین حاصل از این خوشه‌ها می‌پردازیم.



شکل ۱: نمودار شاخص Dunn خوشه بندی ۴ تا ۲۰ تایی برای داده‌های بالای یکسال

نتایج حاصل ازخوشه‌بندی برای داده‌های زیر یکسال

ابتدا نتایج هر خوشه را به صورت مختصر در جدول۱ مشاهده می‌شود، سپس به تحلیل نتایج حاصل از خوشه‌ها پرداخته می‌شود.

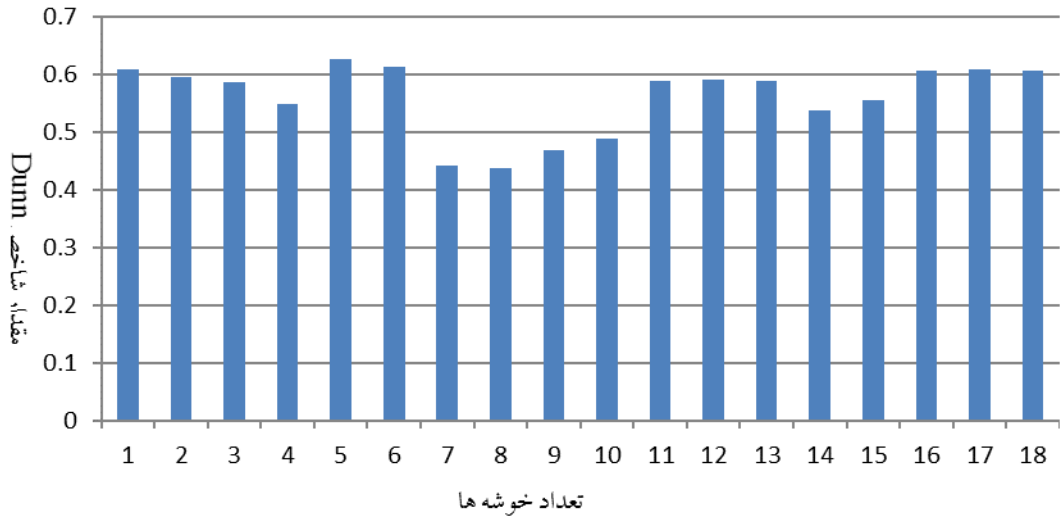
جدول ۱: نتایج حاصل ازخوشه‌بندی برای داده‌های زیر یکسال

شماره خوشه	تعداد رکوردهای خوشه	جنسیت	منطقه سکونت	میانگین سنی	محل فوت	علت فوت	درصد فراوانی علت فوت در خوشه
۱	۱۴۳	مذکر	شهری	۳ روز	بیمارستان	بیماری های دوران حول تولد	۱۰۰ درصد
۲	۱۹۲	مونث	روستایی	۱۰ روز	بیمارستان	بیماری های دوران حول تولد	۱۰۰ درصد
۳	۸۱	مذکر	روستایی	۱۴۱ روز	منزل	حوادث غیر عمدی	۲۲/۲۲ درصد
۴	۶۸	مذکر	شهری	۷۳ روز	بیمارستان	ناهنجاری های مادرزادی و کروموزومی	۶۶/۱۸ درصد
۵	۱۱۹	مونث	شهری	۳۵ روز	بیمارستان	بیماری های دوران حول تولد	۸۱/۵۱ درصد
۶	۲۲۹	مذکر	روستایی	۴۵ روز	بیمارستان	ناهنجاری های مادرزادی و کروموزومی	۱۰۰ درصد
۷	۴۷	مونث	روستایی	۱۴۷ روز	منزل	بیماری دستگاه تنفسی	۳۱/۹۱ درصد
۸	۲۵۶	مذکر	روستایی	۱۰ روز	بیمارستان	بیماری های دوران حول تولد	۱۰۰ درصد

تحلیل نتایج حاصل ازخوشه‌بندی برای داده‌های زیر یکسال

- ✓ اصلی‌ترین علت فوت در نوزادان در روزهای اول تولد، بیماری‌های دوران حول تولد است، درحالی که اگر سن نوزاد به بیشتر از ۹۰ روز برسد عللی مانند حوادث عیرعمدی می‌تواند علت فوت نوزاد باشد.
- ✓ بیشتر فوت‌ها در بیمارستان اتفاق افتاده است.
- ✓ از نظر منطقه سکونت، بیشتر فوت‌ها در مناطق روستایی در مقایسه با مناطق شهری، اتفاق می‌افتد.
- ✓ از نظر جنسیت، بیشتر فوت‌ها مربوط به جنس مذکر است. در مقایسه با جنسیت مونث در واقع حدود ۷۰ درصد فوت‌ها در نوزادان پسر اتفاق افتاده است.
- ✓ نتایج نشان می‌دهد که اصلی‌ترین علت فوت در نوزادان، بیماری‌های دوران حول تولد است. ناهنجاری های مادرزادی و کروموزومی علت بعدی است که درصد کمتری دارد. بیماری‌های دوران حول تولد شامل مشکلاتی که حین زایمان بوجود می آید. بنابراین مشکل اصلی حفظ نوزاد حین و بعد از تولد است.

شکل ۲ نمودار شاخص Dunn را برای خوشه بندی های مختلف برای داده‌های بالای یکسال را نشان می دهد. همانطور که در نمودار زیر کاملاً مشهود است، خوشه بندی با تعداد خوشه ۷، دارای بیشترین مقدار شاخص Dunn است.



شکل ۲ نمودار شاخص Dunn خوشه بندی ۴ تا ۲۰ تایی برای داده‌های بالای یکسال

نتایج حاصل ازخوشه‌بندی برای داده‌های بالای یکسال

پس از خوشه بندی با تعداد بهینه ۷ خوشه برای داده های بالای یکسال، نتایج آن در جدول ۲ آورده شده است.

جدول ۲: نتایج حاصل ازخوشه‌بندی برای داده‌های بالای یکسال

شماره خوشه	تعداد رکوردهای خوشه	جنسیت	منطقه سکونت	میانگین سنی	علت فوت	درصد فراوانی علت فوت در خوشه
۱	۳۴۰۸	مذکر	روستایی	۵۳ سال	حوادث غیر عمدی	۳۰/۷۲ درصد
۲	۲۲۵	مونث	شهری	۶۲ سال	بیماری دستگاه تنفسی	۱۰۰ درصد
۳	۲۷۷۸	مونث	روستایی	۷۰ سال	بیماری های قلبی و عروقی	۱۰۰ درصد
۴	۹	مذکر	شهری	۶۸ سال	بیماری های قلبی و عروقی	۵۵/۵۶ درصد
۵	۳۶۲	مذکر	شهری	۶۱ سال	بیماری های دستگاه گوارش	۱۰۰ درصد
۶	۱۶۶۰	مونث	روستای	۵۷ سال	سرطان ها و تومورها	۲۷/۱۱ درصد
۷	۳۳۲۵	مذکر	شهری	۶۸ سال	بیماری های قلبی و عروقی	۱۰۰ درصد

تحلیل نتایج حاصل ازخوشه‌بندی برای داده‌های بالای یکسال

کمترین میانگین سن، مربوط به افراد در خوشه ی اول است با میانگین سن ۵۳ سال که اصلی ترین علت فوت در آن مربوط به حوادث غیر عمد می باشد. از آن جایی که این گروه از جدول ICD10 شامل بیماری‌هایی از قبیل حوادث ترافیکی، غرق شدگی، سقوط، سوختگی و گزیدگی است این میانگین سن با توجه به دلایل فوت برای این گروه سنی منطقی است. از طرفی تعداد رکوردهای این خوشه از دیگر خوشه ها بیشتر است. همچنین حوادث غیر عمدی درصد٪ این گروه را تشکیل می دهند که نشان دهنده این موضوع است که این علت یکی از عوامل مهم در فوت افراد است.

افراد خوشه سوم همگی زن هستند و علت اصلی مرگ در آن ها بیماری های قلبی و عروقی با فراوانی ۱۰۰ درصد بوده است. از طرفی در خوشه هفتم نیز اصلی ترین علت فوت بیماری های قلبی و عروقی است. فراوانی آن نیز ۱۰۰ درصد است با این تفاوت که همه‌ی افراد این خوشه را مردان تشکیل می‌دهند. همانطور که مشاهده می شود از هفت خوشه، دو خوشه مربوط به بیماری های قلبی و عروقی است که مجموعاً ۶۱۰۳ نفر را تشکیل می دهد که نمایانگر مهم تر بودن این عامل نسبت به دیگر علت های منجر به فوت است.

در خوشه دوم اصلی ترین علت منجر به فوت در افراد آن که همگی زن بوده اند، بیماری های دستگاه تنفسی بوده است. در خوشه سوم نیز همگی زن بوده با علت اصلی فوت بیماری های قلبی و عروقی اما در خوشه چهارم که افراد آن مردان هستند. اولین علت بیماری های قلبی و عروقی و علت بعدی بیماری های دستگاه تنفسی است. علت آن می تواند این باشد که بیماری های مربوط به دستگاه تنفسی در مردان نسبت به زنان کمتر است.

در مجموع همانند داده های مربوط به نوزادان، مرگ و میر در مردان بیشتر از زنان است. سهم مرگ مردان ۶۱،۰۴ درصد و در زنان ۳۸،۹۶ درصد می باشد.همچنین بیشتر فوت ها در مناطق روستایی و در بیمارستان ها رخ می دهد.

تکنیک درخت تصمیم

جدول ۳ درصد درستی داده های آموزش و آزمون الگوریتم های درخت تصمیم را برای مجموعه ی داده های متوفیان زیر یکسال نشان می دهد. از میان الگوریتم‌های موجود در جدول ۳، الگوریتمی برای بیش‌بینی انتخاب می‌شود که درصد خطای کمتری داشته باشد. هدف پیش‌بینی علت مرگ با استفاده از ویژگی‌های مانند سن، جنسیت، محل فوت و منطقه سکونت است.

جدول ۳: درصد درستی قوانین به دست آمده از الگوریتم‌های درخت تصمیم در مجموعه داده‌های متوفیان زیر یکسال

الگوریتم	میزان صحت آموزش (درصد)	میزان صحت آزمون (درصد)
C 5.0	۸۲/۴۱	۷۷/۳۷
CHAID	۷۲/۰۴	۶۷/۲۶
CART	۷۴/۸۹	۶۵/۹۲
QUEST	۶۷/۶۵	۶۵/۴۳

تحلیل قانون های ایجاد شده توسط الگوریتم C5.0 برای داده‌های زیر یکسال

قانون اول: برای نوزادان زیر ۳۹ روز اگر جنسیت مذکر و محل فوت بیمارستان، آنگاه علت فوت، بیماری های دوران حول تولد است.

قانون دوم: برای نوزادان زیر ۳۹ روز اگر جنسیت مذکر و محل فوت منزل، آنگاه علت فوت، ناهنجاری های مادرزادی و کروموزمی است.

قانون سوم: اگر سن نوزاد زیر ۳۹ روز و جنسیت مؤنث باشد، آنگاه علت فوت، بیماری های دوران حول تولد می باشد.

قانون چهارم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر اگر سن کمتر از ۱۲۰ روز باشد، آنگاه علت فوت، ناهنجاری های مادرزادی و کروموزمی می باشد.

قانون پنجم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر اگر سن کمتر از ۲۱۰ و بیشتر از ۱۲۰ روز (۲۱۰>سن>۱۲۰) باشد و محل فوت بیمارستان، آنگاه علت فوت، ناهنجاری های مادرزادی و کروموزمی می باشد.

قانون ششم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن کمتر از ۲۷۰روز و بیشتر از ۲۱۰ روز (۲۷۰>سن>۲۱۰) باشد و محل فوت بیمارستان، آنگاه علت فوت، حوادث غیر عمدی است.

قانون هفتم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن کمتر از ۲۴۰ روز و بیشتر از ۲۱۰ روز (۲۴۰>سن>۲۱۰) باشد و محل فوت منزل، آنگاه علت فوت، ناهنجاری های مادرزادی و کروموزمی می باشد.

قانون هشتم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن کمتر از ۲۷۰ روز و بیشتر از ۲۴۰ روز (۲۷۰>سن>۲۴۰) باشد، محل فوت منزل و منطقه سکونت روستایی، آنگاه علت فوت، بیماری های دستگاه گوارش است.

قانون نهم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن کمتر از ۲۷۰ روز و بیشتر از ۲۴۰ روز (۲۷۰>سن>۲۴۰) باشد، محل فوت منزل و منطقه سکونت شهری، آنگاه علت فوت، مرگ با علت نامعلوم است.

قانون دهم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن بیشتر از ۲۷۰ روز باشد، آنگاه علت فوت ناهنجاری های مادرزادی و کروموزمی می باشد.

قانون یازدهم: برای نوزادان بالای ۳۹ روز و جنسیت مؤنث: اگر سن کمتر از ۱۸۰ روز باشد، آنگاه علت فوت، بیماری های دوران حول تولد است.

قانون دوازدهم: برای نوزادان بالای ۳۹ روز و جنسیت مؤنث: اگر سن بیشتر از ۱۸۰ روز باشد، آنگاه علت فوت، حوادث غیر عمدی است.

پس از ورود داده‌های مربوط به متوفیان بالای یکسال در الگوریتم‌های درخت تصمیم، نتیجه آنالیز قابل قبولی مشاهده نشد. این موضوع منطقی به نظر می‌رسد، چرا که داده‌ها به ۲۲ گروه و حدود ۳۵۰ زیرگروه مختلف تقسیم‌بندی شده‌اند. همین امر موجب کاهش دقت الگوریتم ها و افزایش خطا در داده‌های آموزش و تست می شود. به نظر می‌رسد بهترین راه، استفاده از مجموعه داده‌ها پس از خوشه‌بندی آن هاست چرا که داده‌ها به جای تقسیم شدن به ۲۲ گروه در ۷ خوشه اصلی قرار می گیرند که داده ها در هر خوشه دارای شباهت بالا و داده ها در خوشه های مختلف دارای کمترین شباهت به یکدیگر می باشند از طرفی این هفت خوشه را می‌توان با نام‌های دیگر نیز نام‌گذاری کرد، از آن جایی که هر یک از این خوشه‌ها دارای ویژگی‌های مشخصی هستند، می‌توانند با یک Lable (برچسب) مجزا نام‌گذاری شوند. بنابراین این روش علاوه بر منطقی بودن موجب ارائه مدلی مفید و به دست آمدن الگوهای مناسب می شود. جدول ۴ درصد درستی داده های آموزش و آزمون الگوریتم های درخت تصمیم را برای مجموعه ی داده های متوفیان بالای یکسال(بزرگسالان) نشان می دهد.

جدول ۴: درصد درستی قوانین به دست آمده از الگوریتم‌های درخت تصمیم در مجموعه داده‌های متوفیان بالای یکسال

الگوریتم	میزان صحت آموزش (درصد)	میزان صحت آزمون (درصد)
C 5.0	۹۷/۴۳	۹۶/۸۶
CHAID	۹۲/۳۳	۹۲/۴۳
CART	۹۵/۰۳	۹۵/۱۹
QUEST	۹۴/۶۲	۹۴/۵۴

تحلیل قانون های ایجاد شده توسط الگوریتم $C_{5.0}$ برای داده‌های بالای یکسال

قانون اول: اگر جنسیت مذکر باشد، افراد ساکن مناطق شهری باشند و دارای سن کمتر از ۴۱ سال باشند، آنگاه این افراد در خوشه اول قرار می‌گیرند که اصلی‌ترین علت فوت در این خوشه، حوادث غیر عمدی گزارش شده است.

قانون دوم: اگر جنسیت مذکر باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۴۱ سال و کمتر از ۵۴ سال باشند، آنگاه این افراد متعلق به خوشه اول هستند.

قانون سوم: اگر جنسیت مذکر باشد، افراد ساکن مناطق روستایی باشند و از توابع شهرستان گنبد باشند و دارای سن بیشتر از ۵۴ سال و کمتر از ۷۰ سال باشند، آنگاه این افراد متعلق به خوشه هفتم هستند که که اصلی‌ترین علت فوت در این خوشه، بیماری‌های قلبی و عروقی با فراوانی ۱۰۰ درصد است.

قانون چهارم: اگر جنسیت مذکر باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۷۰ سال باشند، آنگاه این افراد متعلق به خوشه اول هستند.

قانون پنجم: اگر جنسیت مذکر باشد، افراد ساکن مناطق شهری باشند، ساکن یکی از شهرستان‌های آق‌قلا، آزادشهر، گنبد و یا گرگان باشند و دارای بیشتر از ۵۴ سال و کمتر از ۶۲ سال باشند، آنگاه این افراد متعلق به خوشه اول هستند.

قانون ششم: اگر جنسیت مذکر باشد، افراد ساکن مناطق شهری باشند، ساکن یکی از شهرستان‌های آق‌قلا، آزادشهر، گنبد و یا گرگان باشند و دارای بیشتر از ۶۲ سال باشند، آنگاه این افراد متعلق به خوشه هفتم هستند.

قانون هفتم: اگر جنسیت مؤنث باشد، افراد ساکن مناطق روستایی باشند و دارای سن کمتر از ۲۵ سال باشند، آنگاه این افراد متعلق به خوشه دوم هستند که که اصلی‌ترین علت فوت در این خوشه، بیماری دستگاه تنفسی است.

قانون هشتم: اگر جنسیت مؤنث باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۲۵ سال و کمتر از ۳۷ سال باشند، آنگاه این افراد متعلق به خوشه ششم هستند که که اصلی‌ترین علت فوت در این خوشه، سرطان‌ها و تومورها می‌باشد.

قانون نهم: اگر جنسیت مؤنث باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۳۷ سال باشند، آنگاه این افراد متعلق به خوشه سوم هستند که که اصلی‌ترین علت فوت در این خوشه، بیماری‌های قلبی و عروقی است.

بحث و نتیجه گیری

در چند سال اخیر گسترش بکارگیری برنامه های نرم افزاری همراه با تجهیزات سخت افزاری در برنامه ثبت مرگ معاونت سلامت گرچه به کارایی، دقت، صحت وکیفیت داده های مرگ و میر افزوده است ولی با توجه به بالا رفتن حجم داده ها در اثر ثبت فیلدهای متعدد مورد نیاز و همچنین نیاز به آموزش نرم افزاری برنامه ثبت مرگ در سطح استان، امکان بهره برداری از این داده ها برای تصمیم گیرندگان سیاست های بهداشت ودرمان مشکل تر به نظر می رسد. لذا یافتن راهی که بتواند با افزایش حجم روزافزون داده های مرگ ومیر و تغییرات شاخص های آن، دانش مدیریت برنامه های

سلامت جامعه را با بررسی موشکافانه داده های به دست آمده، ارتقاء دهد می تواند بسیار ارزشمند باشد و داده کاوی یکی از ده علم برتر قرن اخیر پاسخ بسیار خوبی به این نیاز می باشد.

در این پژوهش، داده های مورد استفاده، با روش K-Means، خوشه بندی شدند و با بکار گیری شاخص Dunn، تعداد بهینه خوشه‌ها به دست آمد سپس با استفاده از درخت تصمیم، دسته بندی داده ها انجام شد.

از نتایج حاصل از این پژوهش در بخش بهداشت به منظور راه اندازی سیستم مراقبت مرگ و میر به موازات سیستم مراقبت بیماری‌ها درمعاونت بهداشتی و درمان دانشگاه علوم پزشکی، می‌توان استفاده نمود. همچنین با راه اندازی کلاس های داده‌کاوی برای دانشجویان وکارشناسان رشته‌های بهداشتی و درمانی، گام مهمی برای دستیابی سریع به عواملی که نقش بسزایی در پیشگیری مهم‌ترین بیمارهای منجر به فوت دارد، برداشته شود.

طبق نتایج حاصل از خوشه ها بیشترین علت فوت همانند وضعیت مرگ و میر در ایران و جهان، بیماری های قلبی و عروقی گزارش شده است. مقایسه این نتیجه با پژوهش های مربوط به دهه ۵۰ و ۶۰ که اصلی ترین علت فوت، بیماری های عفونی و انگلی بوده است، نشان می دهد در گذشته بیماری های کشنده و مهلک واگیر بیشترین درصد مرگ و میر را تشکیل می داد که بهبود آگاهی و دانش بهداشتی، آب سالم و واکسیناسیون موجب کنترل ، بیماری های واگیر شده است.

تشکر و قدردانی

نویسندگان مقاله از مسئولین و پرسنل مرکز بهداشت و همچنین دانشگاه علوم پزشکی گرگان کمال تشکر و قدردانی را دارند و همچنین از سرکار خانم زهرا یوسفی تلوری دانش آموخته کارشناسی ارشد آمار دانشگاه گلستان به خاطر همکاری های بی دریغشان سپاسگزارند.

References

1. Greenberg, Raymond S. Medical Epidemiology, 3rded. The McGraw Hill Company ; 2001: 48.

2. Nabavi,,A.;" Philosophy of Power", Research Institute of Howzah & University,2000.

3. Abbasi, M. J.; Kave, Z.;" Analysis of the sociological thesis ", Journal of Population, National Organization of Civil Registration, Numbers 39 and 40,2003

4. Yavari. P.; Abadi, A.; Mehrabi, Y.; "Epidemiology Causes of death and process of Changes in Iran from 2001 to 2006". Journal of the Hakim; Volume VI;No.3 ,2003.

5. Bakhtiari H, Determination of Model & Distribution of Mortality Data by Data Mining Methods Duration of 84 Months (2004-2009) in Fars Province, MARCH 2012.

6. Shayan, H.; "Analysis of Mortality Data in Khorasan Razavi Province ", Ferdowsi University of Mashhad.

7. Kermansaravi,Z.; Golipoorgoodarzi, B.; Babazade, M.; " Predicting The Cause of Neonatal Mortality Using Data Mining Techniques"; The 6th Iran Data Mining Conference, Tehran, 2012.

8. Zandi, A.; Aliabadi, P.; Sheikhahmadi, O.;" Data mining on concrete sampling and concrete resistance examination "; The 6th Iran Data Mining Conference, Tehran, 2012

9. Jekel James F.,Katz David L. "Epidemiology, Biostatistics and Preventive Medicine" Second Edition, W.B. Saunders Company, 2001.

10. Baggot Rob, "Public Health Policy and Politics". Mac Millan press Ltd, 2000.

11. Palik,H., " Clustering algorithms in data mining", http://www.tejaratgah.com/583323d3cfc162ee3df2efe546177a3d-1.html, 2013.

12. Ghazanfari,M.; Alizadeh, S.; Teymoori, B.;" Data Mining and Knowledge Discovery" , Iran University of Science and Technology Tehran,2008.

13. Dunn, J. C., "Well Separated Clusters and Optimal Fuzzy Partitions", Journal of Cybernetica 4, 95-114 (1974).

14. StatSoft Inc. Classification and Regression Trees (C&RT).www statsoft com 2005 October 12Available from: URL: www.statsoft.com/textbook/stcart.html

15. Alizadeh, Somayeh; Malekmohammadi, Samira;" Data mining and knowledge discovery step by step with Clementine"; K.N.Toosi University of Technology, Tehran,2011.

Determination of the Distribution Pattern of Mortality Using Data Mining Technique in Golestan Province since 2007 to 2009

Fatemeh Bagheri[1*], Fatemeh Ahangari[2], Naser Behnampour[3]

1. M.Sc. in Computer Engineering, Computer Engineering Department, Golestan University, Gorgan, Iran.
2. B.Sc. in Computer Engineering, Computer Engineering Department, Golestan University, Gorgan, Iran.
3. Ph.D. in Biostatistics, Department of Biostatistics, Golestan University of Medical Science, Gorgan, Iran.

Abstract

Background & Objective: Investigatingg the mortality in a population has been considered as one of the appropriate methods of health detection. Although, there are some problems such as lack of confidence in accuracy measurement and quality of data collection. Establishment of death registration systems and using international classification codes of diseases, and also mortality data integrating by responsible organizations have solved great parts of the previous problems. In this study, considering a set of parameters, the study population was divided into two groups: deceased under one year (infants) and over one year (adults). Then both groups were clustered using the *K-means* method to identify different groups. Hidden models and useful patterns were also discovered using decision tree algorithms. Finally, a neural network algorithm was used to show the ranking of attributes in order of their importance.

Method: In this research, data of 12,865 deceased individuals in Golestan province since 2007 to 2009 is studied. The data has been obtained from the Health Center of Golestan province. The main characteristics used in this study are: deceased age, gender, cause of death, place of residence and place of death. K-means algorithm is used to cluster data. The decision tree algorithms and neural networks algorithm were also used for classification. Finally, results and rules were extracted. Due to different natures of causes of death in infants and adults, studying on these different groups is performed separately.

Results: In clustering phase, the optimal number of clusters is obtained by *Dunn index*; eight clusters for infants and seven clusters for adults were obtained. Among four decision-tree algorithms (C5.0, QUEST, CHAID and CART), C5.0 algorithm with high correction rate, 77.37% in infants data and 96.86% in adults data was the best classifier algorithm. Age, gender and place of death were the most important variables that were detected by neural network algorithm.

Conclusion: In the present study, the collected mortality data was clustered by considering the effective factors and the standard of International Classification of Diseases. The hidden patterns of mortality for infants and adults were extracted. Due to the explicit nature and the intelligibility of the decision tree algorithms, the results and extracted rules are very useful for specialists in this field.

Key words: Data Mining, Clustering, Classification, Decision Tree, Mortality

Corresponding Author: Fatemeh Bagheri
Address: Computer Engineering Department, Golestan University, Gorgan, Iran.
E-mail: f.bagheri@gu.ac.ir