

## تعیین الگوی توزیع واقعه مرگ با استفاده از تکنیک‌های داده-کاوی در استان گلستان از سال ۱۳۸۶ تا ۱۳۸۸

فاطمه باقری<sup>۱</sup>، فاطمه آهنگری<sup>۲</sup>، ناصر بهنام پور<sup>۳</sup>

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

### چکیده

**زمینه و هدف:** بررسی وضعیت مرگ‌ومیر در یک جمعیت به عنوان یکی از روش‌های مناسب تعیین سلامت مورد توجه بوده است، گرچه با مشکلاتی از قبیل عدم اطمینان به صحت و کیفیت داده‌ها و روش جمع‌آوری آن روبرو بوده است. راه‌اندازی نظام‌های ثبت مرگ‌ومیر با استفاده از کدهای بین‌المللی طبقه‌بندی بیماری‌ها و ادغام اطلاعات مرگ‌ومیر توسط سازمان‌های 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 زیریکسال و بالای یکسال تقسیم شدند سپس این دو گروه داده با استفاده از روش K-means خوشه‌بندی شدند تا گروه‌های مختلف شناسایی شوند. همچنین با استفاده از روش درخت تصمیم به کشف مدل‌ها و الگوهای موثر 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 داده‌ی مربوط به متوفیان استان گلستان پرداخته شده است. داده‌های مورد استفاده در این مقاله مربوط به اول فروردین ۱۳۸۸ تا پایان فروردین ۱۳۸۹ می‌باشند. 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 خوشه‌بندی در این پژوهش از الگوریتم K-means استفاده شد، دسته‌بندی نیز به کمک الگوریتم‌های درخت تصمیم و الگوریتم شبکه‌های عصبی انجام پذیرفت سپس نتایج و قوانین حاصل از الگوریتم‌ها 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 می‌گیرند و بررسی می‌شوند.

**یافته‌ها:** در روش خوشه‌بندی، تعداد بهینه خوشه‌ها با استفاده از اندازه‌گیری شاخص Dunn، هشت خوشه برای داده‌های زیر یکسال و هفت خوشه برای داده‌های بالای یکسال به دست آمد. از میان چهار الگوریتم درخت تصمیم نظیر C5.0، CHAID، QUEST، CART و الگوریتم C5.0 با نرخ تشخیص ۷۷،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 تاثیرگذار 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:** داده‌کاوی، خوشه‌بندی، دسته‌بندی، درخت تصمیم، مرگ‌ومیر

نویسنده مسئول: فاطمه باقری

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

f.bagheri@gu.ac.ir

ایمیل:



## مقدمه

نزولی و به علت بیماری‌های غیرواگیر بویژه سرطان‌ها و سوانح و حوادث، سیر صعودی داشته است. این امر برنامه‌ریزی برای کنترل و پیش‌گیری از بیماری‌های غیر واگیردار را بیش از پیش روشن می‌سازد (۴).

حمید بختیاری در پایان‌نامه خود به تعیین الگو و توزیع واقعه مرگ به روش داده‌کاوی طی ۸۴ ماه از سالهای ۱۳۸۳ تا ۱۳۸۹ در استان فارس پرداخت. این پژوهش با استفاده از داده‌های مرگ‌ومیر ثبت شده در نظام ثبت داده‌های مرگ که از سال ۱۳۸۳ در کشور براساس کدهای بین‌المللی طبقه‌بندی بیماری‌ها (ICD10) راه اندازی شد، انجام پذیرفت. نرم‌افزار مورد استفاده، نرم‌افزار WEKA و OLAP cube بوده است (۵).

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

زینب کرمان‌ساروی و همکارانش در مقاله‌ای سعی کردند به پیش‌بینی علل مرگ‌ومیر نوزادان با استفاده از تکنیک‌های داده‌کاوی در سال ۱۳۹۱ بپردازند. آنها با استفاده از الگوریتم درخت تصمیم C4.5 برای

اطلاعات پیرامون علل مرگ، به عنوان ابزار پایش ارتقاء سطح سلامت جامعه و تعیین اولویت‌های اقدام‌های بهداشتی سال‌هاست 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 ویژگی‌های منطقه‌ای مربوط، صورت گرفته است. متأسفانه در ایران اینگونه مطالعات متأثر از فقدان اطلاعات کافی و بهنگام، تقریباً به فراموشی سپرده شده است. اکنون چندسالی است که اولین داده‌های مربوط به باروری و مرگ‌ومیر 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 مرگ‌ومیر در فضای جغرافیایی یک سرزمین دارای ارزش کاربردی همچون شناسایی نواحی مسئله‌دار و امکان پیش‌بینی دقیق‌تر تحولات آتی خواهد بود. ضمن اینکه متغیرهای مهم مشتق از آنها یعنی امید به زندگی و میزان مرگ‌ومیر اطفال از متغیرهای تاثیرگذار بر سطح توسعه انسانی هر جامعه است (۳).

بررسی روند علت‌های مرگ در دو دهه گذشته نشان می‌دهد که مرگ به علت بیماری‌های واگیردار سیر

از آنجایی که اطلاعات مربوط به سن در فایل اکسل به صورت سال، ماه و روز ثبت شدند، از آنجایی که از نرم افزار ۱۲ Clementine 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.

داده‌های مورد استفاده در این مقاله مربوط به اول فروردین ۱۳۸۸ تا پایان فروردین ۱۳۸۹ می باشند. (بعد از مرحله آماده‌سازی و تصفیه داده‌ها)

ویژگی‌های موجود در این داده‌ها عبارتند از: 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 همه‌ی این اطلاعات مناسب داده‌کاوی، کنکاش و تحلیل و بررسی نمی‌باشند. بنابراین ستون‌های 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.

دسته‌بندی و الگوریتم K نزدیک‌ترین همسایه و دورترین همسایه در نرم‌افزارهای WEKA و SPSS روی حدود ۳۰۰ پرونده، دلایل اصلی مرگ و میر نوزادان پس از تولد و امکان پیش‌بینی عوامل اصلی این رویداد را بررسی و شناسایی کردند (۷).

علیرضا زندی و همکارانش در مقاله‌ای به بررسی و تحلیل مهمترین بیماری‌های منجر به فوت در شهر تهران با استفاده از تکنیک‌های داده‌کاوی پرداختند. برای این منظور داده‌های مربوط به آمار فوت شدگان در شهر تهران در سال ۱۳۸۹ مورد بررسی قرار داده شد. در این مقاله از نرم‌افزار WEKA استفاده شد و الگوریتم‌های 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 داده‌های فوت استان گلستان که در یک فایل Excel ذخیره شده بود، دسترسی پیدا کردیم. حجم داده‌ها ۱۶۳۸۲ مورد بوده که شامل اطلاعات متوفیان از فروردین ۱۳۸۴ تا پایان فروردین ۱۳۸۹ در استان می‌باشد.

## استاندارد ICD10

بیماری‌های مورد بررسی در این پژوهش از مرکز بهداشت استان گلستان گرفته شده که خود فهرستی از ثبت نرم‌افزاری و جدول مرگ و میر برای شرایط اپیدمیولوژیک ایران می‌باشد. در حال حاضر مرگ و میرها مطابق آخرین تجدید نظر در "سیستم طبقه‌بندی بین‌المللی بیماری‌ها" (ICD = International Classification of Diseases) تجدید نظر دهم)، کدگذاری می‌شوند (۹). ICD تاکنون ۱۰ بار تجدید نظر شده است (ICD10). گروه‌های 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 می‌کنیم. این جدول شامل سه لایه است که در لایه اول آن ۲۲ بیماری اصلی قرار دارند. دو لایه دیگر این جدول نیز شامل دیگر بیماری‌ها با ذکر جزئیات می‌باشد که مجموعاً تعداد بیماری‌ها در این سه لایه را می‌توان حدود ۱۵۰ بیماری تخمین زد. به دلیل بالا بودن تعداد بیماری‌ها در لایه‌های دوم و سوم، داده‌های موجود در این پژوهش، با توجه به بیماری‌ها در لایه اول که شامل ۲۲ بیماری می‌باشد، کدگذاری شده است (۱۰).

## مواد و روش‌ها

نرم‌افزاری که در این پژوهش مورد استفاده قرار گرفته است، نرم‌افزار Clementine 12 می‌باشد. این نرم‌افزار، ابزار توانمندی است که اغلب تکنیک‌های داده‌کاوی را دربر دارد.

## خوشه‌بندی

روش خوشه‌بندی به منظور شناخت بهتر جامعه مورد مطالعه استفاده شده است. روش خوشه‌بندی جزو روش‌های غیر نظارتی محسوب می‌شود چراکه برای الگوریتم‌های خوشه‌بندی ویژگی دسته تعریف نمی‌شود و رکوردها برچسب خاصی ندارند. الگوریتم خوشه‌بندی اطلاعاتی را که ویژگی‌های نزدیک به هم و مشابه دارند را 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 خوشه‌بندی وجود دارد، تعیین تعداد خوشه‌های بهینه است. هرچه شباهت میان داده‌های در یک خوشه (شباهت درون خوشه‌ای) بیشتر و تفاوت آن‌ها با سایر خوشه‌ها (فاصله بین خوشه‌ای) بیشتر باشد خوشه‌بندی دارای کیفیت

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

(1)

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

محاسبه می شوند:

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

(2)

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

(3)

$C_i$  خوشه ی  $\Delta$ ام،  $d(x, y)$  بیانگر فاصله ی بین دو نمونه

از داده هاست و  $diam(C_i)$  حداکثر فاصله ی درون

خوشه  $\Delta$ ام را بیان می کند.

### درخت تصمیم

به منظور پیش بینی اصلی ترین علت فوت در داده

های زیر یکسال و داده های بالای یکسال، از مدل

درخت تصمیم استفاده شده است. ساختار درخت

تصمیم یک ساختار درختی شبیه فلوجارت است.

بالاترین گره در درخت گره ی ریشه است و گره های

برگ، دسته ها یا توزیع دسته ها را نشان می دهد.

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 می شود. طوریکه داده

هایی که در برگ های انتهایی این درخت قرار

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 مهمترین مسائل در خوشه بندی انتخاب تعداد

خوشه های مناسب می باشد. برای ارزیابی کیفیت

خوشه بندی، می توان 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).

فیلدهای هدف و پیش‌بینی‌کننده در نرم‌افزار Clementine 12 می‌توانند از جنس طبقه‌ای یا فاصله‌ای باشند. درک مدل‌های درخت CART نسبت به سایر مدل‌ها آسان‌تر است (تفسیر قواعد استخراج شده از درخت آسان است). برخلاف C5.0، درخت CART علاوه بر فیلدهای خروجی طبقه‌ای، نوع فاصله‌ای را نیز می‌پذیرد (۱۵).

CHAID یا Chi-squared Automatic Interaction Detection یک روش دسته‌بندی برای ساخت درخت‌های تصمیم است. در ابتدا CHAID اهمیت هر یک از متغیرهای پیش‌بینی-کننده را برای پیش‌بینی متغیر هدف تعیین می‌کند. اگر بیش از یک مورد از روابط از لحاظ آماری مهم بودند، CHAID پیش‌بینی‌کننده‌ای را انتخاب می‌کند که از سایرین مهم‌تر باشد. متغیرهای پیش‌بینی‌کننده و هدف در این الگوریتم می‌توانند از نوع طبقه‌ای یا فاصله‌ای باشند (۱۵).

QUEST یا درخت آماری سریع از روش دسته‌بندی دودویی برای ساخت درخت تصمیم استفاده می‌کند. انگیزه اصلی در توسعه این درخت، کاهش زمان پردازش مورد نیاز برای متوقف کردن درخت بزرگ CART با متغیرها و یا نمونه‌های زیاد است. روش QUEST از یک سری قوانین مبتنی بر آزمون‌های

می‌گیرند توسط یکی از مقادیر ویژگی هدف برجسب می‌خورند. این مدل به دلیل سهولت در تفسیر نتایج و غیر خطی بودن، نیاز به پیش‌فرض رابطه خطی بین متغیرهای مستقل و وابسته ندارد (۱۴).

برای استفاده از تکنیک درخت تصمیم از چهار الگوریتم CART، C5.0، QUEST، CHAID، برای مجموعه داده‌های بالای یکسال (بزرگسال) و زیریکسال (نوزادان) بصورت مجزا استفاده شده است.

الگوریتم C5.0 فقط می‌تواند متغیرهای هدف طبقه‌ای را پیش‌بینی کند. الگوریتم C5.0 در مسائلی مثل داده‌های مفقود و تعداد فیلدهای ورودی زیاد خوب عمل می‌کنند. آنها معمولاً نیازی به زمان طولانی برای تخمین ندارند. بعلاوه، درک این الگوریتم‌ها ساده‌تر از برخی الگوریتم‌هاست. همچنین روش قدرتمندی برای افزایش دقت دسته‌بندی دارد. سرعت ساخت مدل C5.0 به علت بهره‌گیری از پردازش‌های موازی بالا است (۱۵).

الگوریتم CART در تجزیه دو زیر گروه را تعیین می‌کند که هر کدام نیز به دو زیر گروه دیگر تقسیم خواهند شد و این روند ادامه می‌یابد تا زمانی که یکی از معیارهای توقف برآورده شود. تمام تجزیه‌ها دودویی هستند. درخت CART یک یا چند فیلد ورودی را به تنها یک فیلد خروجی تبدیل می‌کند.

مشخص برای ارزیابی متغیرهای پیش‌بینی‌کننده در هر گره استفاده می‌کند. حداقل یک تست باید بر روی هر متغیر پیش‌بینی‌کننده در هر گره انجام شود برعکس درخت CART، همه شاخه‌زنی‌ها امتحان نمی‌شوند و برعکس درخت CART و CHAID، ترکیبات متغیرهای طبقه‌ای در ارزیابی متغیرهای پیش‌بین برای انتخاب آزموده نمی‌شوند. این امر باعث سرعت بالای تحلیل می‌شود. فیله‌های پیش‌بینی‌کننده در الگوریتم QUEST می‌تواند عددی باشند ولی متغیر هدف باید حتما طبقه‌ای باشد (۱۵).

کلاس‌بندی داده‌ها با درختان تصمیم یک فرایند دو مرحله‌ای می‌باشد. در مرحله اول که به آن مرحله آموزش گفته می‌شود، مدلی براساس یک الگوریتم کلاس‌بندی منطبق با داده‌کاوی مربوط به مجموعه آموزشی ساخته می‌شود. مجموعه آموزشی به صورت تصادفی از پایگاه داده انتخاب می‌شود. در مرحله دوم یادگیری از طریق یک تابع  $y=f(X)$  انجام می‌شود که می‌تواند برچسب کلاس هر رکورد  $X$  از پایگاه داده را پیش‌بینی کند (۱۲).

به منظور تعیین میزان صحت درخت تصمیم داده‌ها را به دو بخش داده‌های آموزش و آزمون تقسیم شدند. درخت تصمیم با استفاده از داده‌های آموزش مدل را می‌سازد و مدل ساخته شده بر روی داده‌های آزمون

مورد تست قرار می‌گیرد. درصد نمونه‌هایی که ویژگی هدف آن‌ها توسط مدل، درست تشخیص داده شده بود دقت مدل را بیان می‌کند. در این پژوهش ابتدا داده‌های مربوط به متوفیان زیر یکسال بررسی می‌شوند چون تعداد داده‌های این مجموعه کم است، ۷۰ درصد داده‌ها 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.

## یافته‌ها

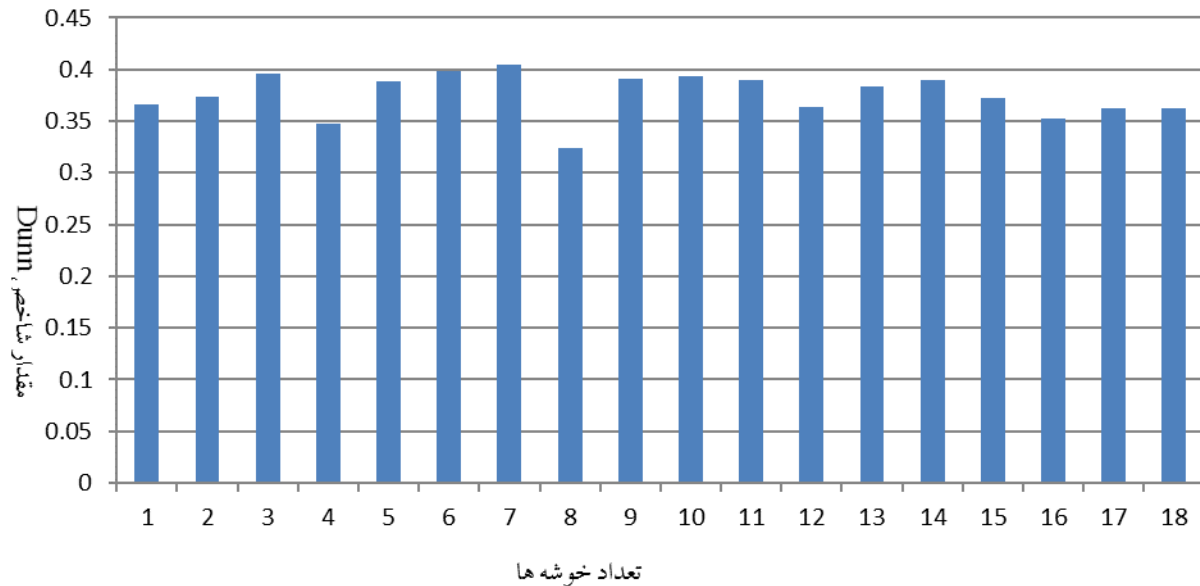
### خوشه‌بندی

الگوریتم خوشه‌بندی K-means را برای خوشه‌های 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 و ۲ 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 خوشه‌بندی ۴ تا ۲۰ تایی برای داده‌های بالای یکسال

### نتایج حاصل از خوشه‌بندی برای داده‌های زیر یکسال

ابتدا نتایج هر خوشه را به صورت مختصر در جدول ۱ مشاهده می‌شود، سپس به تحلیل نتایج حاصل از خوشه‌ها پرداخته می‌شود.

جدول ۱: نتایج حاصل از خوشه‌بندی 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	بیماری‌های دوران حول تولد	۱۰۰٪
۲	۱۹۲	مونث	روستایی	۱۰ روز	بیمارستان	بیماری‌های دوران حول تولد	۱۰۰٪
۳	۸۱	مذکر	روستایی	۱۴۱ روز	منزل	حوادث غیر عمدی	۲۲،۲۲٪

۶۶,۱۸٪	ناهنجاری های مادرزادی و کروموزومی	بیمارستان	۷۳ روز	شهری	مذکر	۶۸	۴
۸۱,۵۱٪	بیماری های دوران حول تولد	بیمارستان	۳۵ روز	شهری	مونث	۱۱۹	۵
۱۰۰٪	ناهنجاری های مادرزادی و کروموزومی	بیمارستان	۴۵ روز	روستایی	مذکر	۲۲۹	۶
۳۱,۹۱٪	بیماری دستگاه تنفسی	منزل	۱۴۷ روز	روستایی	مونث	۴۷	۷
۱۰۰٪	بیماری های دوران حول تولد	بیمارستان	۱۰ روز	روستایی	مذکر	۲۵۶	۸

### تحلیل نتایج حاصل از خوشه بندی برای داده های زیر یکسال

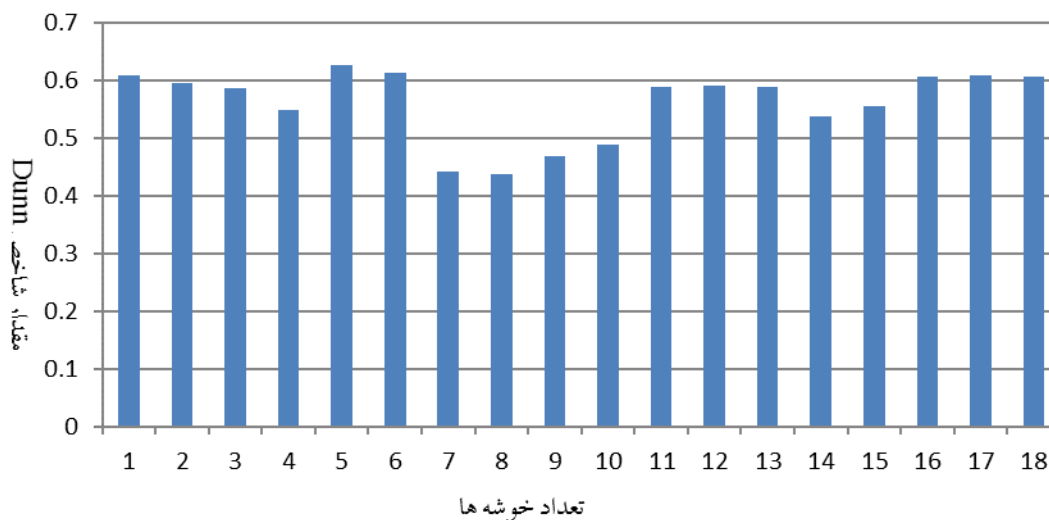
✓ نتایج نشان می دهد که اصلی ترین علت فوت در نوزادان، بیماری های دوران حول تولد است. ناهنجاری های مادرزادی و کروموزومی علت بعدی است که درصد کمتری دارد. ، بیماری های دوران حول تولد شامل مشکلاتی که حین زایمان بوجود می آید بنابراین مشکل اصلی حفظ نوزاد حین و بعد از تولد است.

شکل ۲ نمودار شاخص Dunn را برای خوشه بندی های مختلف برای داده های 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.

✓ اصلی ترین علت فوت در نوزادان در روزهای اول تولد، بیماری های دوران حول تولد است، درحالی که اگر سن نوزاد به بیشتر از ۹۰ روز برسد عللی مانند حوادث غیرعمدی می تواند علت فوت نوزاد باشد.

✓ بیشتر فوت ها در بیمارستان اتفاق افتاده است. ✓ از نظر منطقه سکونت، بیشتر فوت ها در مناطق روستایی در مقایسه با مناطق شهری، اتفاق می افتد.

✓ از نظر جنسیت، بیشتر فوت ها مربوط به جنس مذکر است. در مقایسه با جنسیت مونث در واقع حدود ۷۰٪ فوت ها در نوزادان پسر اتفاق افتاده است.



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

### نتایج حاصل از خوشه‌بندی برای داده‌های بالای یکسال

پس از خوشه‌بندی با تعداد بهینه ۷ خوشه برای داده‌های بالای یک سال، نتایج آن در جدول ۲ آورده شده است.

جدول ۲: نتایج حاصل از خوشه‌بندی 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٪
۲	۲۲۵	مونث	شهری	۶۲ سال	بیماری دستگاه تنفسی	۱۰۰٪
۳	۲۷۷۸	مونث	روستایی	۷۰ سال	بیماری های قلبی و عروقی	۱۰۰٪
۴	۹	مذکر	شهری	۶۸ سال	بیماری های قلبی و عروقی	۵۵٫۵۶٪
۵	۳۶۲	مذکر	شهری	۶۱ سال	بیماری های دستگاه گوارش	۱۰۰٪
۶	۱۶۶۰	مونث	روستای	۵۷ سال	سرطان ها و تومورها	۲۷٫۱۱٪
۷	۳۳۲۵	مذکر	شهری	۶۸ سال	بیماری های قلبی و عروقی	۱۰۰٪

تنفسی بوده است. در خوشه سوم نیز همگی زن بوده با علت اصلی فوت بیماری‌های قلبی و عروقی اما در خوشه چهارم که افراد آن مردان هستند. اولین علت بیماری‌های قلبی و عروقی و علت بعدی بیماری‌های دستگاه تنفسی است. علت آن می‌تواند این باشد که بیماری‌های مربوط به دستگاه تنفسی در مردان نسبت به زنان کمتر است.

در مجموع همانند داده‌های مربوط به نوزادان، مرگ و میر در مردان بیشتر از زنان است. سهم مرگ مردان ۶۱٫۰۴٪ و در زنان ۳۸٫۹۶٪ می‌باشد. همچنین بیشتر فوت‌ها در مناطق روستایی و در بیمارستان‌ها رخ می‌دهد.

### تکنیک درخت تصمیم

جدول ۳ درصد درستی داده‌های آموزش و آزمون الگوریتم‌های درخت تصمیم را برای مجموعه‌ی داده‌های متوفیان زیر یکسال نشان می‌دهد. از میان الگوریتم‌های موجود در جدول ۳، الگوریتمی برای پیش‌بینی 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.

### تحلیل نتایج حاصل از خوشه‌بندی برای داده‌های بالای یکسال

کمترین میانگین سن، مربوط به افراد در خوشه‌ی اول است با میانگین سن ۵۳ سال که اصلی‌ترین علت فوت در آن مربوط به حوادث غیر عمد می‌باشد. از آنجایی 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 خوشه‌ها بیشتر است. همچنین حوادث غیر عمدی ۳۰٪ این گروه را تشکیل می‌دهند 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 آنها بیماری‌های قلبی و عروقی با فراوانی ۱۰۰٪ بوده است. از طرفی در خوشه هفتم نیز اصلی‌ترین علت فوت بیماری‌های قلبی و عروقی است. فراوانی آن نیز ۱۰۰٪ است با این تفاوت که همگی افراد این خوشه را مردان تشکیل می‌دهند. همانطور که مشاهده می‌شود از هفت خوشه، دو خوشه مربوط به بیماری‌های قلبی و عروقی است که مجموعاً ۶۱۰۳ نفر را تشکیل می‌دهد که نمایانگر مهم‌تر بودن این عامل نسبت به دیگر علت‌های منجر به فوت است.

در خوشه دوم اصلی‌ترین علت منجر به فوت در افراد آن که همگی زن بوده اند، بیماری‌های دستگاه

جدول ۳: درصد درستی قوانین بدست آمده از

الگوریتم‌های درخت تصمیم در مجموعه داده‌های متوفیان زیر

یکسال

الگوریتم	میزان صحت آموزش %	میزان صحت آزمون %
C 5.0	۸۲,۴۱	۷۷,۳۷
CHAID	۷۲,۰۴	۶۷,۲۶
CART	۷۴,۸۹	۶۵,۹۲
QUEST	۶۷,۶۵	۶۵,۴۳

علت فوت، ناهنجاری‌های مادرزادی و کروموزومی می‌باشد.

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، ناهنجاری‌های مادرزادی و کروموزومی می‌باشد.

قانون ششم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن کمتر از ۲۷۰ روز 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، ناهنجاری‌های مادرزادی و کروموزومی می‌باشد.

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، بیماری‌های 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.

### تحلیل قانون‌های 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، بیماری‌های دوران حول تولد است.

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

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

قانون چهارم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر اگر سن کمتر از ۱۲۰ روز باشد، آنگاه

روش علاوه بر منطقی بودن موجب ارائه مدلی مفید و بدست آمدن الگوهای مناسب می‌شود. جدول ۴ درصد درستی داده‌های آموزش و آزمون الگوریتم-های درخت تصمیم را برای مجموعه‌ی داده‌های متوفیان بالای یکسال (بزرگسالان) نشان می‌دهد.

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

الگوریتم	میزان صحت آموزش.٪	میزان صحت آزمون.٪
C 5.0	۹۷,۴۳	۹۶,۸۶
CHAID	۹۲,۳۳	۹۲,۴۳
CART	۹۵,۰۳	۹۵,۱۹
QUEST	۹۴,۶۲	۹۴,۵۴

### تحلیل قانون‌های ایجاد شده توسط الگوریتم C5.0

#### 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 دهم: برای نوزادان بالای ۳۹ روز و جنسیت مذکر: اگر سن بیشتر از ۲۷۰ روز باشد، آنگاه علت فوت ناهنجاری‌های مادرزادی و کروموزومی می‌باشد.

قانون یازدهم: برای نوزادان بالای ۳۹ روز و جنسیت مؤنث: اگر سن کمتر از ۱۸۰ روز باشد، آنگاه علت فوت، بیماری‌های 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 یک سال 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 الگوریتم‌ها و افزایش خطا در داده‌های آموزش و تست می‌شود. به نظر می‌رسد بهترین راه، استفاده از مجموعه داده‌ها پس از خوشه‌بندی آنهاست چرا که داده‌ها به جای تقسیم شدن به ۲۲ گروه در ۷ خوشه اصلی قرار می‌گیرند که داده‌ها در هر خوشه دارای شباهت بالا و داده‌ها در خوشه‌های مختلف دارای کمترین شباهت به یکدیگر می‌باشند از طرفی این هفت خوشه را می‌توان با نام‌های دیگر نیز نام‌گذاری کرد، از آنجایی که هر یک از این خوشه‌ها دارای ویژگی‌های مشخصی هستند، می‌توانند با یک Label (برچسب) مجزا نام‌گذاری شوند. بنابراین این

خوشه ششم هستند که که اصلی ترین علت فوت در این خوشه، سرطان‌ها و تومورها می‌باشد. قانون نهم: اگر جنسیت مونث باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۳۷ سال باشند، آنگاه این افراد متعلق به خوشه سوم هستند که اصلی ترین 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 و کیفیت داده های مرگ و میر افزوده است ولی با توجه به بالا رفتن حجم داده ها در اثر ثبت فیلهای 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 سیاستهای بهداشت و درمان مشکل تر بنظر میرسد. لذا یافتن راهی که بتواند با افزایش حجم روزافزون داده های مرگ و میر و تغییرات شاخص های آن، دانش مدیریت برنامه های سلامت جامعه را با بررسی موشکافانه داده های بدست آمده، ارتقاء دهد میتواند بسیار ارزشمند باشد و داده کاوی یکی از ده علم برتر قرن اخیر پاسخ بسیار خوبی به این نیاز می باشد.

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 اصلی ترین علت فوت در این خوشه، بیماری های قلبی و عروقی با فراوانی ۱۰۰٪ 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 شهرستان های آق قلا، 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 شهرستان های آق قلا، 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 مونث باشد، افراد ساکن مناطق روستایی باشند و دارای سن کمتر از ۲۵ سال باشند، آنگاه این افراد متعلق به خوشه دوم هستند که که اصلی ترین علت فوت در این خوشه، بیماری دستگاه تنفسی است.

قانون هشتم: اگر جنسیت مونث باشد، افراد ساکن مناطق شهری باشند و دارای سن بیشتر از ۲۵ سال و کمتر از ۳۷ سال باشند، آنگاه این افراد متعلق به

### 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.

در این پژوهش، داده های مورد استفاده، با روش K-Means، خوشه بندی شدند و با بکار گیری شاخص Dunn، تعداد بهینه خوشه ها بدست 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 بیماری ها در معاونت بهداشتی و درمان دانشگاه علوم پزشکی، می توان استفاده نمود. همچنین با راه اندازی کلاسهای داده کاوی برای دانشجویان و کارشناسان رشته های بهداشتی و درمانی، گام مهمی برای دستیابی سریع به عواملی که نقش بسزایی در پیشگیری مهم ترین 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 و اگیر بیشترین درصد مرگ و میر را تشکیل می داد که بهبود آگاهی و دانش بهداشتی، آب سالم و واکسیناسیون موجب کنترل، بیماری های و اگیر شده است.



**References:**

- [1] Greenberg, Raymond S. Medical Epidemiology, 3<sup>rd</sup>ed. 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 6<sup>th</sup> Iran Data Mining Conference, Tehran, 2012.
- [8] Zandi, A.; Aliabadi, P.; Sheikahmadi, O.; " Data mining on concrete sampling and concrete resistance examination "; The 6<sup>th</sup> 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](http://www.statsoft.com) 2005 October 12 Available from: URL: [www.statsoft.com/textbook/stcart.html](http://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.