مقاله پژوهشی

باقری و همکاران

استفاده از الگوریتم های داده کاوی در بررسی عوامل موثر بر پیش بینی وضعیت بدو تولد نوزادان

فاطمه باقری[۱*]، حکیمه علیزاده مجد[۲]، زهرا مهربخش[۳]، مجید زیارت‌بان[۴]

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. دانشجوی کارشناسی ارشد آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، ایران
۴. استادیار، گروه مهندسی برق، دانشکده فنی و مهندسی، دانشگاه گلستان، گرگان، ایران

دریافت: ۹۳/۶/۱۸ | انجام اصلاحات: ۹۳/۱۲/۱۷ | پذیرش: ۹۴/۴/۱۴

چکیده

زمینه و هدف: پیش بینی وضعیت سلامت نوزادان دربدو تولد و همچنین شناسایی عوامل تاثیر گذار بر آن از اهمیت زیادی برخوردار است. روش های متفاوتی برای این پیش بینی وجود دارد. در این مطالعه با استفاده از الگوریتم های درخت تصمیم به کشف مدل ها و الگوهای موثر پرداخته شده است.

روش بررسی: این مطالعه بر روی ۱۶۶۸ زایمان در سه بیمارستان شهدا، امیدی و مهر شهرستان بهشهر انجام شد. متغیرهای جنسیت، وزن و رتبه تولد نوزاد، سن، سابقه بیماری، بیماری دوران بارداری، نوع زایمان، علت سزارین، سن حاملگی، گروه خونی، شغل، فشار خون و محل زندگی مادر، نسبت فامیلی پدر و مادر، گروه به عنوان متغیر های پیش بینی کننده روش دسته بندی درخت تصمیم و همچنین وضعیت سلامت نوزاد هنگام تولد به عنوان متغیر دو وضعیتی وابسته مورد استفاده قرار گرفتند. تمامی متغیرها در خوشه بندی و قواعد تلازمی نیز استفاده شدند. پیش بینی با ۴ الگوریتم درخت تصمیم انجام و با یکدیگر مقایسه شدند.

یافته ها: در روش خوشه بندی، تعداد بهینه خوشه ها با استفاده از اندازه گیری شاخص Dunn، هشت خوشه تعیین شد. در میان الگوریتم های اجرا شده ی CHAID، QUEST، CART و C۵,۰ ، الگوریتم C۵,۰ با نرخ تشخیص ۹۴,۴۴ درصد، به عنوان بهترین الگوریتم شناسایی شد. با پیاده سازی الگوریتم Apriori نیز قواعد تلازمی استخراج شدند که با در نظر گرفتن حد آستانه برای پشتیبان (Support) و اطمینان (Confidence)، قواعد قوی استخراج شدند. در میان ویژگی ها، سن حاملگی با بیش¬ترین تاثیر و وزن نوزاد و علت سزارین از جمله ویژگی های با اهمیت زیاد در پیش بینی بودند.

نتیجه گیری: با توجه به تفسیر ساده درخت تصمیم و قابل فهم بودن قوانین استخراج شده از آن برای (اغلب افراد) متخصصین و همچنین زنان باردار، می توان در سطوح مختلف از آن استفاده نمود.

کلمات کلیدی: داده کاوی، خوشه بندی، درخت تصمیم، قواعد تلازم، وضعیت سلامت نوزادان

نویسنده مسئول: فاطمه باقری
آدرس: ایران، گرگان، دانشگاه گلستان، دانشکده فنی مهندسی
ایمیل: f.bagheri@gu.ac.ir

مقدمه

تولد پدیده ای زیبا، معجزه آسا و گاهی خطرناک ترین حادثه در طول زندگی فرد است. (تولد، پدیده ای زیبا است که گاهی می تواند به خطرناک ترین حادثه در طول زندگی فرد تبدیل شود). بدن انسان بلافاصله بعد از تولد نیاز به تنظیم و هماهنگی فیزیولوژیک فوق العاده ای دارد، بیشتر از حدی که در ادامه حیات به آن نیاز خواهد داشت (۱). از حدود دویست میلیون حاملگی که هر سال در سطح جهان اتفاق می افتد، قریب به یک سوم به سقط منتهی می شوند و بقیه پدر و مادران ضمن تحمل سختی ها و رنج در انتظار فرزندی سالم و برومند هستند. (بسیاری از) آنان احتمالا از بروز مشکلات احتمالی

بی خبرند و اطلاعات زیادی در مورد حاملگی و زایمان های پرمخاطره ندارند و اصولا نمی دانند که تولد، پرمخاطره ترین اتفاق در زندگی انسان هاست (۳،۲).

باید تمام تلاش ها معطوف پیشگیری از اتفاقات (ممکن) احتمالی گردد و هر جنین و نوزاد در معرض خطر، از قبل شناسایی و مراقبت شود. هیچ مورد حاملگی و یا زایمانی را نمی توان کاملا بدون مخاطره فرض کرد، لذا در مورد هر زایمانی باید پیشاپیش همه آمادگی ها وجود داشته باشد (۱).

حوزه پزشکی و سلامت از بخش های مهم در جوامع صنعتی است. استخراج دانایی از میان حجم انبوه داده های مرتبط با سوابق بیماری و پرونده های پزشکی افراد با استفاده از فرایند داده کاوی می تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسری بیماری ها شود و همچنین اطلاعات ارزشمندی را به منظور شناسایی علل رخداد بیماری ها، پیش بینی،تشخیص و درمان بیماری ها با توجه به عوامل محیطی حاکم در اختیار متخصصان و دست انداران حوزه سلامت قرار دهد که نتیجه آن می تواند منجر افزایش عمر و ایجاد آرامش برای افراد جامعه است.

راحله رضائیان لنگرودی و سلمان خزائی با استفاده از نرم افزار SPSS ۱۶ و با کمک آزمون های آماری توصیفی استنباطی tو χ^2 روی ۱۰۴ زن باردار و نوزادان آن ها، به بررسی شیوع کاهش شاخص های رشد نوزادان در زنان باردار و کاهش سطح سلامت روان آنها پرداختند (۴).

مهشید مظاهری و همکارانش به بررسی و مقایسه درصد وقوع بارداری بالای ۳۵ سال یا دیگر علل بارداری پرخطر از دیدگاه سلامت پرداختند. نتیجه این که بارداری در سن بالای ۳۵ سال رتبه دوم علل بارداری پرخطر در طی سال های ۸۷ تا ۸۹ را به خود اختصاص می دهد (۵).

Michael P.Harris (Mpharris) , با استفاده از روش های داده کاوی J۴۸، Naïve Bayes، OneR، PART، DecisionTable در نرم افزار weka به پیش بینی وزن پایین نوزادان هنگام تولد پرداختند که روشNaïve Bayes با دقت ۷۲/۳۷ درصد بهترین مدل را ایجاد کرده است (۶).

در این مقاله عوامل تاثیرگذار بر وضعیت نوزاد بررسی و با استفاده از برخی فرایندهای داده کاوی وضعیت نوزاد تازه متولد شده پیش بینی شد تا بتوان با کمک به رفع بعضی ناهنجاری های تاثیرگذار، به افزایش سلامت نوزاد هنگام تولد که عاملی مهم و و حیاتی می باشد کمک شایانی نمود.

روش بررسی

در راستای این مقاله، ۱۶۶۸ پرونده پزشکی مربوط به زایمان زنان باردار بیمارستان های شهدا، امیدی و مهر شهرستان بهشهر به طور تصادفی انتخاب شدند. پرونده های پزشکی شامل برگه ها و فرم های مختلفی بودند که اطلاعات مربوط به مادر و نوزاد در آن ثبت می شد. این پرونده ها حاوی داده های گوناگونی بودند. پس از مشورت با پرسنل بیمارستان شهدا (که آشنا با مبحث زایمان و زنان بودند) و در نهایت پزشک جراح و متخصص زنان و زایمان و نازایی، ویژگی ها انتخاب شدند و مورد بررسی قرار گرفتند.

از آنجایی که اطلاعات مربوط به نوزادان دوقلویی به شکلی متفاوت از نوزادان تک قل در اکسل ثبت شده برای این که اختلالی در جواب نهایی ایجاد نشود، جداگانه مورد بررسی قرار گرفتند، یعنی داده ها به دو مجموعه داده تقسیم شدند که مجموعه داده های نوزادان تک قل N_A و مجموعه داده های نوزادان دوقلویی N_B نام گرفتند. تعداد داده های N_A، ۱۶۴۰ رکورد و تعداد داده های N_B ۲۸ رکورد بوده است.

ویژگی های استفاده شده برای دو مجموعه جنسیت نوزاد، وزن نوزاد، ترتیب فرزند متولد شده (فرزند چندم خانواده)، سن مادر، تحصیلات مادر، سابقه بیماری مادر، بیماری دوران بارداری، نوع زایمان، علت سزارین، سن حاملگی، نسبت فامیلی پدر و مادر، گروه خونی مادر، شغل مادر، فشارخون مادر (فقط برای مجموعه داده N_A)، محل زندگی مادر و وضعیت نوزاد به هنگام تولد می باشند که ویژگی آخر به عنوان ویژگی هدف در نظر گرفته شد.

برای مجموعه داده N_B ویژگی فشار خون مادر به علت تک مقداره بودن در جامعه مورد بررسی و برای جلوگیری از ایجاد اختلال در جواب نهایی فیلتر شد. در مرحله بعد، ویژگی های مربوط به علت سزارین و سابقه بیماری مادر مورد بررسی قرار گرفتند. به علت وجود چندین دلیل برای سزارین برخی مادران و همچنین مادرانی در لیست داده ها بودند که بیش از یک مورد سابقه بیماری داشتند و از آنجایی که از نرم افزار clementine ۱۲ در این پژوهش استفاده شده است، برای این که بتوان از این داده ها در نرم افزار برای تحلیل استفاده کرد، دلایل سزارین و همچنین سابقه ی بیماری ها به صورت دسته های مختلف از هم تفکیک شده اند.

در رابطه با علت سزارین، دلایلی که مورد تایید بیمه هستند به عنوان دسته های مختلف علت سزارین در نظر گرفته شد. دلایل دیگری نیز وجود داشتند که مورد تایید بیمه نیستند ولی در پرونده های زایمان دیده می شدند. این دلایل به عنوان ویژگی «سایر موارد» در نظر گرفته شد. همچنین سابقه بیماری های مادر نیز به دسته بندی های مختلف انواع بیماری تقسیم شد (که به صورت دودویی تایید و یا عدم تایید ثبت شده است). ویژگی علت سزارین برای مجموعه N_A، ۳۴ علت مختلف و برای مجموعه N_B، ۵ علت مختلف، همچنین سابقه بیماری مادر برای مجموعه N_A، ۱۹ سابقه متفاوت و برای مجموعه N_B، ۶ سابقه متفاوت بوده است.

خوشه بندی

ابتدا به منظور شناخت جامعه مورد مطالعه از روش خوشه بندی استفاده شده است. زمانی که شناخت زیادی از مجموعه داده وجود ندارد، برای شناختن گروه های مختلف مجموعه مورد مطالعه می توان از الگوریتم های مختلف خوشه بندی استفاده نمود. از آنجایی که برای الگوریتم های خوشه بندی ویژگی دسته تعریف نمی شود و رکورد ها برچسب خاصی ندارند، جزء روش های غیر نظارتی محسوب می شوند. خوشه بندی داده ها را طوری گروه بندی می کند که داده های در یک خوشه بسیار شبیه به هم و همچنین متفاوت از سایر گروه ها هستند. هر چه شباهت میان داده های در یک خوشه (شباهت درون خوشه ای) بیشتر و تفاوت آن ها با سایر خوشه ها (فاصله بین خوشه ای) بیشتر باشدخوشه بندی دارای کیفیت بیشتری خواهد بود. در این مطالعه از الگوریتم K-mean برای خوشه بندی داده ها استفاده شده است.

خوشه بندی k-mean در سال ۱۹۶۵ توسط Forgy ارائه شد که یک روش طبقه بندی بدون نظارت است. این روش علی‌رغم سادگی آن، یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر (مانند خوشه‌بندی فازی) محسوب می‌شود. این روش، روشی انحصاری و مسطح محسوب می‌شود. در اجرای الگوریتم k-means، اولین کار مشخص کردن K یا همان تعداد خوشه ها است. برای این الگوریتم شکل‌های مختلفی بیان شده است. ولی همهٔ آنها دارای روالی تکراری هستند. الگوریتم زیر الگوریتم پایه برای این روش محسوب می‌شود (۷):

۱. در ابتدا K نقطه به عنوان مراکز خوشه‌ها انتخاب می‌شوند.

۲. هر نمونه داده به خوشه‌ای که مرکز آن خوشه کمترین فاصله تا آن داده را داراست، نسبت داده می‌شود.

۳. پس از تعلق تمام داده‌ها به یکی از خوشه‌ها برای هر خوشه یک نقطه جدید به عنوان مرکز محاسبه می‌شود (میانگین نقاط متعلق به هر خوشه).

۴. مراحل ۲ و ۳ تکرار می‌شوند تا زمانی که دیگر هیچ تغییری در مراکز خوشه‌ها حاصل نشود.

به منظور ارزیابی کیفیت خوشه بندی، می توان از شاخص های سنجش کیفیت خوشه ها استفاده کرد. با اندازه گیری شاخص کیفیت برای هر بار اجرای خوشه بندی با تعداد خوشه مختلف، و با در نظر گرفتن بهترین نتیجه، می توان به تعداد بهینه خوشه ها پی برد. بهترین نتیجه زمانی اتفاق می افتد که شباهت داده های درون خوشه و همچنین عدم شباهت داده ها بین خوشه های مختلف به حداکثر ممکن برسد (۸).

شاخصی که در این پژوهش مورد استفاده قرار گرفته است شاخص Dunn است که یکی از متداول ترین شاخص های مورد استفاده جهت اعتبار سنجی خوشه ها است. بزرگترین مقدار D در معیار Dunn بیانگر بهترین روش خوشه بندی است. این شاخص به صورت زیر محاسبه می شود (۹):

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

که $diam(c_K)$ و $d(ci,cj)$ در آن با روابط زیر محاسبه می‌شوند:

$$support = \frac{n(A \cup B)}{N}$$

$$confidence(A \rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{n(A \cup B)}{n(A)}$$

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

یافتن قواعد تلازمی

از دیگر اهدافی که در این مطالعه مورد بررسی قرار گرفت، یافتن قواعد تلازمی است. داده کاوی انجمنی و استخراج قوانین انجمنی از مجموعه داده ها از جنس قوانین احتمالی هستند که اولین بار توسط Agrawal و همکاران در سال ۱۹۹۴ ارائه شد (۱۰). در واقع برای مشخص شدن وابستگی های مهم و پنهان رکوردهای یک پایگاه داده از قوانین انجمنی استفاده می شود. این قوانین وابستگی اتفاق و وقوع یک شی را بر اساس وقوع سایر اشیا پیش بینی می کند. اغلب مشاهده می شود که وابستگی نزدیک بین مجموعه ای از داده های معین وجود دارد. بنابراین الگوریتم های یافتن قوانین وابستگی، تمامی قواعد تلازمی ممکن را درون پایگاه داده پیدا می کنند. خروجی مهم در این روش، عبارت است از مجموعه ای از قوانین اگر آن گاه که بیانگر ارتباطات میان رخداد توامان مجموعه ای از اشیا با یکدیگر هستند.

لازم است که با چند اصطلاح پایه آشنا شویم:

Item Set: مجموعه آیتم های موجود در یک پایگاه داده.

Support: این پارامتر احتمال وجود همزمان A و B را در قانون A→B بیان می کند.

Confidence: این پارامتر احتمال شرطی است برای آنکه تراکنش A دارای B نیز باشد.

دو پارامتر اخیر(Confidence و Support) جهت ارزیابی قانون های تولید شده استفاده می شود.

در رابطه A→B ضریب اطمینان و پشتیبان از دو فرمول زیر محاسبه می گردند:

$$support = \frac{n\,(A \cup B)}{N}$$

$$confidence\,(A \rightarrow B) = \frac{support\,(A \cup B)}{support(A)} = \frac{n(A \cup B)}{n\,(A)}$$

n بیانگر تعداد رکوردهای دارای شرایط و N تعداد کل رکوردها می باشند. یکی از الگوریتم های تکامل یافته کشف قوانین تلازمی الگوریتم Apriori می‌باشد که در این مطالعه استفاده شده است. این الگوریتم در سال ۱۹۹۶ توسط Cheung و همکارانش ابداع شد (۱۱).

درخت تصمیم

به منظور پیش بینی ویژگی هدف وضعیت سلامت نوزاد هنگام تولد، از مدل درخت تصمیم استفاده شده است. الگوریتم درخت تصمیم قادر است علاوه بر متغیرهای کمی، متغیرهای کیفی را نیز پیش بینی کند. این روش اولین بار توسط Breman و همکاران ارائه شد (۱۲). نتیجه پیاده سازی الگوریتم درخت تصمیم مجموعه ای از شرط های منطقی (if-then conditions) با ساختار درختی است که برای پیش بینی یک ویژگی به کار می رود. طوری که داده هایی که در برگ های انتهایی این درخت قرار می گیرند توسط یکی از مقادیر ویژگی هدف برچسب می خورند. این مدل به دلیل سهولت در تفسیر نتایج و ناپارامتری و غیر خطی بودن، نیاز به پیش فرض رابطه خطی بین متغیر های مستقل و وابسته ندارد (۱۳).

الگوریتم درخت تصمیم به گونه ای عمل می کند که سعی دارد گوناگونی و یا تنوع (از نظر ویژگی هدف) را در گره ها به حداقل ممکن برساند. این عدم یکنواختی در گره ها با استفاده از معیارهای عدم خلوص (Impurity measure) قابل اندازه گیری است که مهمترین و پرکاربرد ترین آن شاخص جینی می باشد (۱۴).

اغلب تفاوت انواع درخت های تصمیم در همین معیار اندازه گیری عدم خلوص، شیوه شاخه بندی (Splitting) و هرس کردن گره های درخت می باشد. در این پژوهش از ۴ نوع الگوریتم درخت تصمیم CHAID، QUEST، CART و C۵,۰ استفاده شده است.

الگوریتم CART متغیرهای ورودی را برای یافتن بهترین تجزیه می‌آزماید تا شاخص ناخالصی حاصل از تجزیه کمترین مقدار باشد. در تجزیه دو زیر گروه تعیین می شود و هر کدام در مرحله بعد به دو زیر گروه دیگر تقسیم خواهند شد و این روند ادامه می یابد تا زمانی که یکی از معیار های توقف برآورده شود. درخت CART بازگشتی دو دویی است، که گره های والدین را دقیقا به دو گروه فرزند منشعب می کند و به طور بازگشتی منشعب کردن را تا زمانی که انشعاب دیگری نتواند ساخته شود ادامه می دهد (۱۵).

(Quick Unbiased and Efficient Statically Tree) الگوریتم جدید دو دویی گسترش درخت است که فقط فیلد خروجی سمبلیک را مورد استفاده قرار می دهد. در تقسیم داده ها به صورت بازگشتی به زیر گروه ها فقط دو زیر گروه را پشتیبانی می کند. یک الگوریتم سریع است که هرس نمودن آن رو به عقب است(۱۵).

CHAID مخفف شده کلمات Chi-square Automatic Interaction Detector می باشد. به طور کلی روش CHAID، برای ساخت یک درخت تصمیم گیری، داده ها را متناوباً به زیر

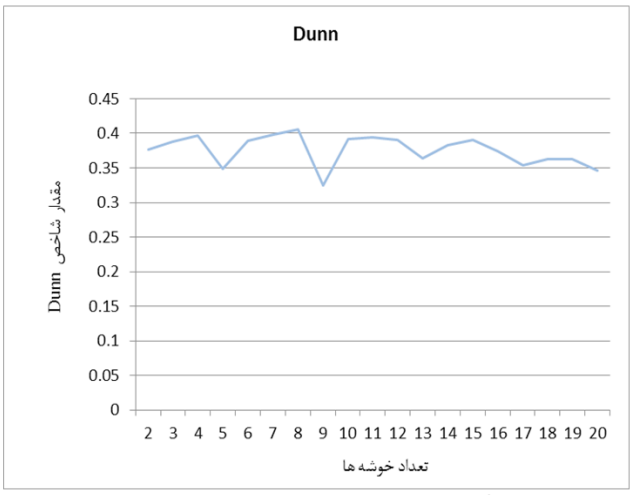
مجموعه های مشابه افراز می کند تا آنجا که هر زیر مجموعه دارای تعداد مشخصی نمونه شود. این الگوریتم می تواند درختی تولید کند که در برخی از مواقع به صورت غیر دو دویی عمل کند، در واقع از روش جدا کردن چند تایی به جای جدا کردن دو دویی استفاده می کند. به این صورت که می توان نود پدر را به بیش از دو تقسیم نماید. این الگوریتم از آزمون «کای دو» برای تصمیم گیری در هر تقسیم برای مشخص کردن نود های فرزند استفاده می کند. سپس شاخه های درخت ساخته شده تا تحقق معیار توقف یا رسیدن به سطح پیچیدگی خواسته شده، هرس می شوند. به بیان دیگر، CHAID ابتدا تفاوت های هر نمونه را با سایر نمونه ها می یابد و درخت مورد نظر را تولید می کند. هرس کردن درخت از طریق یافت تفاوت های مشابه انجام می شود (۱۵).

الگوریتم C۵,۰ یک نوع درخت تصمیم گیری تک متغیره و بهبود یافته الگوریتم C۴,۵ است. این الگوریتم مشابه با CART ابتدا درختی تقریبا پر ایجاد می کند ولی استراتژی هرس آن کاملا متفاوت است. این الگوریتم دسته بندی را با تقسیم کردن داده ها به زیر مجموعه هایی که شامل رکورد های همگن تر از والد خود هستند انجام می دهد. در C۵,۰ تقسیم کردن نمونه ها براساس فیلدی که بیشترین بهره اطلاعات را دارد صورت می گیرد. این الگوریتم روشی افزایشی از هرس کردن درخت را به کار می گیرد تا خطای طبقه بندی کردن ناشی از نویز یا جزئیات خیلی زیاد را در داده های آموزشی کاهش دهد. هرس کردن با جایگزینی گره داخلی با گره برگ رخ می دهد که بدان وسیله درصد یا میزان خطا کاهش می یابد (۱۵).

به منظور اعتبارسنجی مدل درختی نیز داده ها به دو بخش داده های آموزش و آزمون تقسیم شدند. مدل درختی با استفاده از داده های آموزش ساخته شدو مدل ساخته شده بر روی داده های آزمون مورد تست قرار گرفت. درصد نمونه هایی از داده های آزمون که ویژگی هدف آن ها توسط مدل، درست تشخیص داده شده بود دقت مدل را بیان می کند (۱۶). برای هر ۴ الگوریتم درخت تصمیم ۷۰ درصد داده ها به صورت تصادفی به عنوان داده های آموزش انتخاب شدند و ۳۰ درصد مابقی به عنوان داده های آزمون مورد آزمایش قرار گرفتند.

یافته ها:

الگوریتم خوشه بندی K-mean برای تعداد مختلف خوشه ها از ۲ تا ۲۰ خوشه انجام شد و برای هر حالت معیار کیفیت Dunn اندازه گیری شد که در شکل شماره ۱ آمده است. با توجه به این که بالاترین میزان این شاخص مربوط به خوشه بندی با ۸ خوشه است، تعداد بهینه خوشه ها ۸ خوشه می باشد. به عبارت دیگر زمانی که داده ها با ۸ خوشه گروه بندی شده باشند حالتی است که فاصله بین داده ها در یک خوشه حداقل و فاصله بین داده ها در خوشه های مختلف حداکثر می شود.



شکل ۱: مقادیر شاخص Dunn برای تعداد مختلف خوشه ها

در زیر به اختصار تحلیلی از نتایج خوشه بندی آورده شده است.

– نوع زایمان داوطلبانه سزارین نسبت به طبیعی بین زنان باردار تحصیل کرده و شاغل بیش تر است.

– میانگین سن مادران باردار شهرنشین (۲۸و۲۹ سال) در اولین بارداری، نسبت به زنان باردار روستانشین (۲۵ و ۲۶ سال) بیش تر است.

– در مادران دارای گروه خونی A+، احتمال افزایش فشار خون و بیماری دوران بارداری نسبت به سایر گروه خونی بیش تر است.

– زنان بارداری که فرزند اول خود را طبیعی به دنیا آورده اند، در صورتی که هیچ یک از علت های سزارین در بارداری دوم آنها اتفاق نیفتد میل دوباره به زایمان نوع طبیعی در آنها وجود دارد.

– میانگین وزن نوزاد زنان بارداری که شاغل و شهرنشین هستند و تحصیلات دانشگاهی دارند، نسبت به زنان خانه دار و دارای تحصیلات دیپلم و روستانشین بیش تر است.

– بیماری دوران بارداری در زنانی که جنسیت نوزاد آنها دختر است، چه شهرنشین و چه روستانشین نسبت به جنسیت نوزاد پسر بیش تر است.

برای به دست آوردن قوانین انجمنی الگوریتم Apriori روی دو مجموعه داده N_A و N_B پیاده سازی شده است. از آنجایی که هدف یافتن روابط بین متغیرهای ورودی و متغیر هدف (در این مقاله یعنی وضعیت نوزاد در هنگام تولد) است، فقط روابط مربوطه بررسی شده است. مینیمم Support برای قوانین، ۲۰ درصد و مینیمم Confidence، ۸۰ درصد در نظر گرفته شده است. همچنین به دلیل تعداد زیاد مدل های ایجاد شده، فقط تعدادی از آنها بیان شده اند.

تحلیل قواعد تلازمی برای مجموعه داده N_A در ادامه توضیح داده شده است.

- با پشتیبان ۱۱/۷۲۲ درصد مادر شاغل و نوزاد فرزند اول خانواده و سالم است و پدر و مادر نسبت فامیلی ندارند. به احتمال ۹۶/۳۵۴ درصد اگر مادر شاغل باشد و پدر و مادر نسبت فامیلی نداشته باشند و نوزاد فرزند اول خانواده باشد، آنگاه نوزاد سالم است.

- با پشتیبان ۱۶/۶۰۶ درصد گروه خونی مادر O+ و فشار خون مادر نرمال و نوزاد فرزند اول و سالم است و پدر و مادر نسبت فامیلی ندارند. به احتمال ۹۱/۵۴۴ درصد اگر گروه خونی مادر O+ و فشار خون مادر نرمال و نوزاد فرزند اول باشد و پدر و مادر نسبت فامیلی نداشته باشند، آنگاه نوزاد سالم است.

- با پشتیبان ۱۰/۵۰۱ درصد نوع زایمان طبیعی و محل زندگی مادر روستا و فشار خون مادر نرمال و نوزاد سالم است. به احتمال ۹۱/۲۷۹ درصد اگر نوع زایمان طبیعی و محل زندگی مادر روستا و فشار خون مادر نرمال باشد، آنگاه نوزاد سالم است.

- با پشتیبان ۱۲/۲۷ درصد مادر بیماری دوران بارداری و زایمان دارد و نوزاد فرزند اول و سالم و فشار خون مادر نرمال است و مادر و پدر نسبت فامیلی ندارند. به احتمال ۸۵/۰۷۵ درصد اگر مادر بیماری دوران بارداری و زایمان داشته باشد و نوزاد فرزند اول و فشار خون مادر نرمال باشد و پدر و مادر نسبت فامیلی نداشته باشند، آنگاه نوزاد سالم است.

همچنین تحلیل قواعد تلازمی برای مجموعه داده N_B در ادامه آورده شده است.

- با پشتیبان ۱۰/۷۱۴ درصد گروه خونی مادر B+ و نوزادان فرزند دوم و جنسیت نوزادان دختر است و نوزادان بستری می شوند و با احتمال ۱۰۰ درصد اگر گروه خونی مادر B+ باشد و نوزادان فرزند دوم باشند و جنسیت نوزادان دختر باشد، آنگاه نوزادان بستری می شوند.

- با پشتیبان ۱۰/۷۱۴ درصد گروه خونی مادر B+ و نوزادان، فرزند دوم و محل زندگی مادر روستا است و نوزادان بستری می شوند و با احتمال ۱۰۰ درصد اگر گروه خونی مادر B+ و نوزادان، فرزند دوم و محل زندگی مادر روستا باشد، آنگاه نوزادان بستری می شوند.

- با پشتیبان ۱۴/۲۸۶ درصد مادر بیماری دوران بارداری و زایمان دارد و جنسیت نوزادان دختر و تحصیلات مادر دیپلم است و نوزادان بستری می شوند و با احتمال ۱۰۰ درصد اگر مادر بیماری دوران بارداری و زایمان داشته باشد و جنسیت نوزادان دختر و تحصیلات مادر دیپلم باشد، آنگاه نوزادان بستری می شوند.

- با پشتیبان ۱۷/۸۵۷ درصد مادر بیماری دوران بارداری و زایمان دارد و محل زندگی مادر روستا است و پدر و مادر نسبت فامیلی ندارند و نوزادان بستری می شوند و با احتمال ۱۰۰ درصد اگر مادر بیماری دوران بارداری و زایمان داشته باشد و محل زندگی مادر روستا و پدر و مادر نسبت فامیلی نداشته باشند، آنگاه نوزادان بستری می شوند.

برای استفاده از تکنیک درخت تصمیم از چهار الگوریتم C۵,۰، QUEST، CHAID، CART برای مجموعه داده های N_A و N_B استفاده شده است. بعد از ورود و تعیین نوع داده ها و تقسیم بندی آن ها به دو بخش آموزش و آزمون، این ۴ الگوریتم بر روی ۷۰ درصد داده ها (داده های آموزش) اجرا شده و مدل حاصل از آنها بر روی ۳۰ درصد مابقی داده ها (داده های آزمون) آزمایش شده است. جدول شماره ۱ و ۲ به ترتیب نتایج حاصل از الگوریتم های درخت تصمیم برای دو مجموعه داده ی N_A و N_B را نشان می دهند.

جدول شماره ۱: درصد درستی قوانین بدست آمده برای داده های آموزش و آزمون در مجموعه داده ی N_A

نام الگوریتم	میزان صحت آموزش٪	میزان صحت آزمون٪
C 5.0	۹۳/۵۶	۹۴/۴۴
CHAID	۹۲/۶۸	۹۲/۸۶
CART	۹۲/۰۶	۹۳/۴۵
QUEST	۹۰/۳۹	۹۳/۰۶

جدول شماره ۲:درصد درستی قوانین بدست آمده برای داده های آموزش و آزمون در مجموعه داده ی N_B

نام الگوریتم	میزان صحت آموزش٪	میزان صحت آزمون٪
CART	۱۰۰	۱۰۰
QUEST	۶۶/۶۷	۲۸/۵۷

از آنجایی که درصد درستی قوانین ایجاد شده برای مجموعه داده N_A توسط الگوریتم درخت تصمیم C ۵,۰ بهتر از سایر الگوریتم ها است، در ادامه تحلیل قانون های ایجاد شده توسط این الگوریتم آورده شده است.

قانون اول: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۲۶ هفته باشد و وزن نوزاد کم تر و یا مساوی ۲/۶۷۰ کیلوگرم باشد، آن گاه نوزاد مرده است.

قانون دوم: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۳۶ هفته و بیش تر از ۲۶ هفته باشد و وزن نوزاد کم تر و یا مساوی ۲/۶۷۰ کیلوگرم باشد، آنگاه نوزاد بستری می شود.

قانون سوم: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۳۶ هفته باشد و وزن نوزاد بیش تر از ۲/۶۷۰ کیلوگرم باشد و علت سزارین نازایی باشد، آنگاه نوزاد بستری می شود.

قانون چهارم: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۳۶ هفته باشد و وزن نوزاد بیش تر از ۲/۶۷۰ کیلوگرم و کم تر و یا مساوی ۳/۸۰۰ کیلوگرم باشد و علت سزارین نازایی نباشد و نوزاد اولین و دومین فرزند خانواده باشد، آن گاه نوزاد سالم است.

قانون پنجم: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۳۶ هفته باشد و وزن نوزاد بیش تر از ۲/۶۷۰کیلوگرم و کم تر و یا مساوی ۳/۱۲۰ کیلوگرم باشد و علت سزارین نازایی نباشد و نوزاد فرزند سوم و چهارم و پنجم خانواده باشد، آن گاه نوزاد سالم است.

قانون ششم: اگر علت سزارین IUFD نباشد و سن حاملگی کم تر و یا مساوی ۳۶ هفته باشد و وزن نوزاد بیش تر از ۳/۱۲۰کیلوگرم و کم تر و یا مساوی ۳/۸۰۰ کیلوگرم باشد و علت سزارین نازایی نباشد و نوزاد فرزند سوم و چهارم و پنجم خانواده باشد، آن گاه نوزاد بستری است.

قانون هفتم: اگر علت سزارین IUFD نباشدو سن حاملگی کم ترو یا مساوی ۳۶ هفته باشد و وزن نوزاد بیش تر از ۳/۸۰۰ کیلوگرم باشد و علت سزارین نازایی نباشد، آن گاه نوزاد بستری می شود.

قانون هشتم: اگر علت سزارین IUFD نباشد و سن حاملگی بیش تر از ۳۶ هفته باشد، آنگاه نوزاد سالم است.

از میان الگوریتم های استفاده شده بر روی داده های N_B، لگوریتم CART با میزان درستی ۱۰۰ درصد بهتر از سایر الگوریتم ها بوده است. بنابراین در ادامه تحلیل قانون های ایجاد شده توسط این الگوریتم برای مجموعه داده ی N_B آورده شده است.

قانون اول: اگر وزن نوزادان به صورت «۱۰۰۰-۱۰۰۰»،«۱۰۰۰-۳۰۰۰»، «۲۰۰۰-۱۵۵۰»، «۲۰۰۰-۱۷۵۰»، «۲۰۰۰-۱۸۵۰»، «۲۱۰۰-۲۰۶۰»، «۲۱۰۰-۱۹۰۰»، «۳۶۰۰-۲۱۰۰»، «۲۳۰۰-۱۷۰۰»، «۲۱۰۰-۲۵۰۰»، «۲۸۰۰-۲۵۰۰»، «۲۹۰۰-۲۵۰۰»، «۲۵۵۰-۲۵۵۰»، «۲۷۵۰-۲۴۵۰»، «۳۲۰۰-۲۵۰۰» باشد، آن گاه نوزادان بستری می شوند.

قانون دوم: اگر وزن نوزادان به صورت «۲۲۰۰-۲۵۵۰»، «۲۳۵۰-۲۰۰۰»، «۲۲۵۰-۲۴۵۰»، «۲۴۷۰-۲۱۰۰»، «۲۶۰۰-۲۵۰۰»، «۲۵۵۰-۲۶۰۰»، «۲۸۵۰-۲۶۰۰»، «۲۷۵۰-۲۵۰۰»،«۲۷۵۰-۲۵۵۰»، «۲۷۶۰-۲۵۹۰»، «۲۹۰۰-۲۶۵۰» باشد، آن گاه نوزادان سالم اند.

قانون سوم: اگر وزن نوزادان به صورت «۲۴۰۰-۲۶۰۰»، «۶۰۰-۶۵۰» باشد و جنسیت نوزادان دختر باشد، آن گاه نوزادان می میرند.

قانون چهارم: اگر وزن نوزادان به صورت «۲۴۰۰-۲۶۰۰»، «۶۰۰-۶۵۰» باشد و جنسیت نوزادان پسر باشد، آن گاه یکی از نوزادان سالم و دیگری بستری می شود.

نوزادان دوقلویی اکثرا به علت وزن پایین هنگام تولد بستری می شوند. در قانون سوم، علت مرگ نوزادان را وزن «۲۶۰۰-۲۴۰۰» کیلوگرم بیان داشته در صورتی که این وزن برای نوزادان وزن مناسبی است و دلیل مناسبی برای مرگ نوزادان تلقی نمی شود. دلیل اینکه این الگوریتم این وزن را علت مرگ دانسته این است که تعداد داده های دوقلویی برای انجام پروژه کم بوده، و موردی در بین داده ها وجود داشته که نوزادان با این وزن، مرده متولد شده اند و علت مرگ آنها می تواند آنومالی مادرزادی مثل بیماری قلبی، مشکلات مغزی، هیدروپس و یا سایر علل باشد. پس قانون سوم در رابطه با این وزن رد می شود اما وزن «۶۰۰-۶۵۰» کیلوگرم برای این قانون رد نمی شود و منطقی می باشد.

بحث و نتیجه گیری:

داده کاوی، جستجوی خودکار منابع داده ای بزرگ، برای یافتن الگوها و وابستگی هایی است که تحلیل های ساده ومعمول آماری قادر به انجام آن نیستند. یکی از زمینه های استفاده از این ابزار برای تحلیل داده های وسیع و مدل سازی پیشگویانه با روش های محاسباتی جدید، علوم پزشکی و بهداشتی است. هدف از این مقاله یافتن عوامل موثر بر وضعیت بدو تولد نوزاد می باشد. این عوامل شامل جنسیت نوزاد، سن مادر، تحصیلات مادر، شغل مادر، وزن نوزاد، سن حاملگی و ... می باشد. این متغیرها، قابل اندازه گیری از ابتدای دوران بارداری یا در ادامه این دوران می باشند. با شناسایی وکنترل عوامل موثر بر وضعیت های غیر نرمال در بدو تولد نوزادان، می توان از بوجود آمدن این وضعیت ها پیش گیری نمود. در این مقاله به خوشه بندی داده ها و دسته بندی آنها با استفاده از الگوریتم های درخت تصمیم و کشف روابط بین متغیرها با استفاده از قواعد تلازمی پرداخته شده است.

مطالعات مختلفی به منظور پیش بینی وضعیت نوزادان هنگام تولد انجام شده است که تنها اندکی از آن ها با استفاده از داده کاوی و الگوریتم¬های درخت تصمیم بوده است.

Elena Bermd و همکارانش با استفاده از الگوریتم های داده کاوی GRI، Apriori در نرم افزار clementine ۱۱، به بررسی داده هایی برای کشف قوانین انجمنی خاصی روی داده های مواظبت پری ناتال پرداختند (۱۷).

Janie Wilson و همکارانش سلامت نوزادان و سن حاملگی زمان زایمان را با استفاده از داده کاوی بررسی کردند. نتایج مطالعه نشان داد که زایمان های قبل از ۳۹ هفتگی ، زایمان های طولانی تر و پیچیده تر و همچنین القای درد زایمان قبل از ۳۹ هفتگی به طور انتخابی، به احتمال زیاد نوزادان را دچار علائم اختلال تنفسی می کند (۱۸).

جوریان و همکاران در سال ۱۳۹۳ الگوریتم های داده کاوی را به منظور پیش بینی میزان اثربخشی داروهای پره اکلامپسی بر اساس دوز و روش مصرف دارو به کار گرفتند. سه الگوریتم،CART CHAID و C۵,۰ بر روی دادههای پره اکلامپسی اعمال شد، حساسیت ۱۰۰ درصد و ویژگی ۹۹/۵ درصد برتری الگوریتم CART را تاییدکرد. همچنین برای تایید نتایج حاصل از پیش بینی، از روش خوشه بندی استفاده شد (۱۹).

کشتکار و همکاران در سال ۸۵ ۱۳ برای تعیین عوامل موثر بر شدت پره اکلامپسی، نقش برخی عوامل زمینه ای و مراقبتی همراه با پره اکلامپسی شدید را با استفاده از مدل رده بندی درختی و رگرسیونی مورد ارزیابی قرار دادند.

این مطالعه نشان داد که استفاده از متغیرهای قابل سنجش در دوران بارداری، قادر به پیش بینی پیامد پرخطر پره اکلامپسی شدید می باشد(۲۰) .

مطالعه حاضر نشان داد که سن حاملگی بیش ترین تاثیر را برای مجموعه داده ی N_A دارد و بعد از آن به ترتیب وزن نوزاد و علت سزارین (جنین مرده) تاثیر کم تر و سپس ترتیب فرزند متولد شده و نازایی کم ترین تاثیر را دارند و برای مجموعه داده ی N_B وزن نوزاد بیش ترین تاثیر را بر وضعیت نوزاد می گذارد.

گروه خونی مادر، سن حاملگی، محل زندگی مادر، نوع زایمان، علت سزارین (دوقلویی)، سن مادر، ترتیب فرزند متولد شده، جنسیت نوزاد ، تاثیر کم تری را نسبت به وزن نوزاد بر وضعیت بدو تولد نوزادان مجموعه یN_B دارند.

ناپارامتری بودن روش طبقه‌بندی درختی، تفسیر آسان نتایج، عدم نیاز به پیش فرض‌هایی مشابه سایر مدل‌های پیش‌بینی، لحاظ کردن اثر متقابل بین متغیرهای پیش‌بینی کننده از مزایای الگوریتم های درخت تصمیم است. این مطالعه را می‌توان با توجه به تاثیر عوامل محیطی، مانند آلودگی هوا و آلودگی صوتی و سایر عوامل و با داده‌های سایر شهرستان ها نیز انجام داد تا بتوان قدرت تعمیم آن را افزایش داد. با توجه به تفسیر ساده قوانین حاصل از الگوریتم های درخت تصمیم، می توان از این قوانین در سطوح مختلف خدمات درمانی نظیر مراکز بهداشتی، مطب ها، کلینیک های تخصصی زنان و زایمان و مراکز مرتبط دیگر استفاده نمود. در صورت مشاهده علائم ایجاد کننده خطر برای سلامت نوزاد هنگام تولد در دوران بارداری، مادر باید به طور دقیق تحت نظارت قرار گرفته تا درمان های لازم انجام شود و از وقوع سایر علائم خطر زا که امکان وقوع آن ها در ادامه دوران بارداری و یا هنگام زایمان وجود دارد، جلوگیری به عمل آید.

با توجه به اهمیت سلامت زنان باردار و نوزادان آن ها هنگام تولد، مطالعه و پژوهش روی این مساله می تواند بسیار با ارزش باشد. طراحی و تدوین مدل های پیش بینی مناسب و با اعتبار بالا می تواند کمک زیادی به این مهم باشد.

تقدیر و تشکر

نویسندگان مقاله از مسئولین، پرسنل و همچنین متخصصان زنان و زایمان و نازایی بیمارستان های شهدا، امیدی و مهر شهرستان بهشهر به خاطر همکاری های بی دریغشان کمال تشکر و قدردانی را دارند.

References:

.1. Neonatal Resuscitation, American Academy of Pediatrics, American Heart association, 5th edition, 2006.

2. Ward M P, Platt; Brown, K.,” Evaluation of advanced neonatal nurse practitioners: confidential enquiry into the management of sentinel cases”, Arch Dis Child Fetal Neonatal Ed, 89:F241, 2004.

3. Chan L C; Hey E; “Can all neonatal resuscitation be managed by nurse practitioners?”Arc Dis Child- Fetal Neonatal Ed, 91:F52-F55, 2006.

4. rezaeyanLangroodi R, Khazaie S, StudyingOn Reduction of Neonatal Growth Indices in Pregnant Women with Reductionof their Mental HealthLevels, 21st conference by Nursing & Midwifery Faculty, Khorasgan Branch, Islamic Azad University, 27-28 November 2013.[Persian]

5. Mazaheri M, MaghzianEsfahani F, Jahangiri M, Parsa A, Ahmadian M, Investigation and comparison of percentage of pregnancies over 35 years old women or other high-risk pregnancyfactors from the reproductivehealth programs point of view in the health center No. 2 in Esfehan, since 1387 until 1391, 21st conference by Nursing & Midwifery Faculty, Khorasgan Branch, Islamic Azad University, 27-28 November 2013.[Persian]

6. P.Harris, Michael; ”Predicting Lowbirth Weight Through Data Mining”, CAS CS 105 Introduction to Database, 2009.

7. Forgy, E.W., “Cluster analysis of multivariate data: efficiency vs interpretability of classifications”, Biometrics 1, (1965), 768–769.

8. Halkidi, M., Batistakis, Y., and Vazirgiannis, M., “Cluster validity methods: part II”, SIGMOD Rec., Vol. 31, No. 3, 19- 7(11).

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

10. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’93), pages 207–216,Washington, DC, May 1993.

11. D. W. Cheung, J. Han, V. Ng, and C. Y. Wong. Maintenance of discovered association

rules in large databases: An incremental updating technique. In Proc. 1996 Int. Conf.Data

Engineering (ICDE’96), pages 106–114, New Orleans, LA, Feb. 1996.

[12]Breiman L, Friedman J, Olshen R, Stone Ch. Classification and regression trees. Boca Raton: Chapman & Hall/CRC. 1984.

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

14. Yoneyama Y, Suzuki S, Sawa R, Yoneyama K, Power GG,Araki T. Increased plasma adenosine concentrations and theseverity of preeclampsia. Obstet Gynecol. 2002;100(6):1266-70

15. Chattamvelli R, Data mining Algorithm, Alpha science, 2011.

16. Gupta G.K,Introduction to Data Miningwith Case Studies, Prentice-Hall of India, second edition, 2011.

17. Dogaru, Roxana; Zaharie, Daniela; Lungeanu, Diana; Bernad, Elena; Bari, Maria; “A Framework For Mining Association Rules In Data On Perinatal Care”, the 8th international conference on technical informatics, Timisoara, Romania, 2008.

18. Wilson, Janie; M.S.; R.N.; “Desinging best practices through Data Mining”http://www.hhnmostwired.com/hhnmostwired_app/jsp/article-display.jsp?dcrpath=HHNMOSTWIRED/PubsNewsArticleMostWired/data/07Summer/070912MW_online_Wilson&domain=HHNMOSTWIRED

20. keshtkar a, mahamad k. Determining factors on preeclampsia severity the application of classification and regression tree. Gorgan University of Medical Sciences. 1385;8(2).

Original Article

Bagheri & Colligues...

Use of data mining algorithms in assessing the affecting factors on predicting the health status of newborns

Fatemeh Bagheri[1]*, Hakimeh Alizadeh Majd[2], Zahra Mehrbakhsh[3], Majid Ziaratban[4]

1. Instructor, Department of Computer Engineering, School of Engineering, University of Golestan, Gorgan, Iran
2. Computer Engineer, School of Engineering, University of Golestan, Gorgan, Iran
3. MSc student in Biostatistics, School of Public Health, Mashhad University of Medical Sciences, Mashhad, Iran
4. Assistant Professor, Department of Electrical Engineering, School of Engineering, University of Golestan, Gorgan, Iran

Received: 2014.9.9 Revised: 2015.3.7 Accepted: 2015.7.5

Abstract

Background & Objective: Prediction of health status in newborns and also identification of its affecting factors is of the utmost importance. There are different ways of prediction. In this study, effective models and patterns have been studied using decision tree algorithm.

Method: This study was conducted on 1,668 childbirths in three hospitals of Shohada, Omidi and Mehr in city of Behshahr. Variables such as baby's gender, birth weight, birth order, maternal age, maternal history of illness, gestational diseases, type of delivery, reason of caesarean section, maternal age, family relationship of father and mother, mother's blood type, mother's occupation and blood pressure and place of residence were chosen as predictive factors of decision tree categorization method. The health status of the baby was used as a dependent dual-mode variable. All variables were used in clustering and correlation rules. Prediction was done and then compared using 4 decision-tree algorithms.

Results: In the clustering method, the optimal number of clusters was determined as 8, using the Dunn index measurement. Among all the implemented algorithms of CART, QUEST, CHAID and C5.0, C5.0 algorithm with detection rate of 94.44% was identified as the best algorithm. By implementing the Apriori algorithm, strong correlation rules were extracted with regard to the threshold for Support and Confidence. Among the characteristics, maternal age, birth weight and reason of caesarean section with the highest impacts were found as the most important factors in the prediction.

Conclusion: Due to the simple interpretation of the decision tree and understandability of the extracted rules derived from it, this model can be used for (most individuals) professionals and pregnant women at different levels.

Keywords: Data Mining, Clustering, Decision Trees, Correlation Rules, Health Status of Babies

Corresponding Author: Fatemeh Bagheri
Address: Iran, Gorgan, University of Golestan, School of Engineering
Email: f.bagheri@gu.ac.ir

Please cite this paper as: Bagheri F, Alizadeh Majd H, Mehrbakhsh Z, Ziaratban M. Use of data mining algorithms in assessing the affecting factors on predicting the health status of newborns. Hakim Jorjani J. 2015; 2(2): 59-68.