طبقه بندی تعداد فرزندان زنده به دنیا آمده با استفاده از مدل کارت (CART)

آرزو باقری[۱]، مهسا سعادتی[۱*]

۱. استادیار، مؤسسه مطالعات و مدیریت جامع و تخصصی جمعیت کشور، آمار کاربردی، تهران، ایران

چکیده

زمینه و هدف: تحلیل ممیزی و رگرسیون لجستیک از روش های کلاسیک در زمینه طبقه بندی داده ها در بسیاری از مطالعات هستند. با این حال، استفاده از این روش ها گاه به دلیل عدم برقراری پیش فرض های لازم، نتایج کارایی به دنبال ندارد. هدف از پژوهش حاضر، استفاده از مدل درخت تصمیم برای طبقه بندی تعداد فرزندان زنده به دنیا آمده به منظور ارائه روشی کارا در طبقه بندی داده های جمعیتی می باشد.

روش بررسی: در این پژوهش از مدل درختی کارت (CART) با استفاده از شاخص افراز جینی برای طبقه بندی تعداد فرزندان زنده به دنیا آمده، استفاده شد. داده های حاصل از بررسی رفتارهای ازدواج و باروری زنان حداقل یکبار ازدواج کرده، ۴۹ -۱۵ ساله در استان سمنان- ۱۳۹۱ استفاده شد. ۴۰۵ زن ۴۹ -۱۵ ساله حداقل یکبار ازدواج کرده، نمونه پژوهش را تشکیل دادند.

یافته ها: زنان کوهورت های موالید اول و دوم که در سنین پایین ازدواج کرده اند،۳ فرزند زنده و زنانی که در سن بالاتر ازدواج کرده اند، ۲ فرزنده زنده به دنیا آورده اند. زنان کوهورت موالید سوم که در سنین پایین ازدواج کرده و شاغلند ۲ فرزنده زنده و غیرشاغلین، بسته به نوع ازدواجشان که خویشاوندی یا غیر خویشاوندی بوده به ترتیب ۰ و ۱ فرزند زنده به دنیا آورده اند. زنان کوهورت موالید سوم که در سن بالاتر ازدواج کرده اند ۱ فرزند زنده به دنیا آورده اند .

نتیجه گیری: از مزایای مهم مدل کارت، سادگی در تفسیر نتایج، آزاد توزیع بودن متغیرهای به کار رفته در ساخت درخت و نحوه برخورد آن با داده های گمشده و دور افتاده می باشد که استفاده از آن را افزایش داده است. در نتیجه روشی مناسب برای طبقه بندی داده های جمعیتی در مقایسه با سایر روش های مدل سازی کلاسیک در شرایط عدم برقراری پیش فرض های لازم می باشد.

کلمات کلیدی: طبقه بندی، درخت تصمیم، مدل کارت، شاخص افراز جینی

نویسنده مسئول: مهسا سعادتی

آدرس: ایران، تهران، مؤسسه مطالعات و مدیریت جامع و تخصصی جمعیت کشور

ایمیل: mahsa.saadati@gmail.com

مقدمه

مطالعات جمعیت‌شناسی، بررسی تغییرات یک جمعیت در طول زمان را با توجه به پدیده‌های باروری، مرگ و میر و مهاجرت تعیین می‌کنند. در این مطالعات، باروری جایگاه ویژه‌ای دارد و نقش آن به عنوان مهمترین پدیده تعیین‌کننده نوسانات جمعیتی سبب شده است که مطالعات مربوط به آن نسبت به سایر پدیده‌های جمعیتی از اهمیت فراوانی برخوردار باشد و بررسی عوامل مختلف اقتصادی، اجتماعی و فرهنگی مؤثر بر آن، سهم بزرگی از پژوهش‌های اجتماعی را به خود اختصاص دهد (۱). در سرشماری‌های بسیاری از کشورها، از زنان در مورد «تعداد فرزندانی که تاکنون به دنیا آورده‌اند» سؤال می‌شود؛ پاسخ به این سؤال، داده‌های مربوط به تعداد فرزندان زنده به دنیا آمده هر زن را تولید می‌کند که جمعیت‌شناسان اغلب از این داده‌ها برای مدل سازی آماری باروری استفاده می‌کنند. تعداد فرزندان زنده به دنیا آمده هر زن پیامدهای مهمی بر روی بهداشت عمومی، اقتصاد، آب و هوا و ساختار جمعیت دارد. این تعداد بر روی مرگ و میر نوزادان، کودکان و بزرگسالان، خدمات بهداشت و درمان مادر و کودک، رشد (یا کاهش) اقتصادی، بار تکفل، جمعیت فعال نیروی کار و ساختار سنی جمعیت مؤثر است (۲).

کشور ما در دهه‌های اخیر، تحولات اقتصادی، اجتماعی و جمعیتی بسیاری را تجربه کرده و همزمان با این تحولات، رفتار باروری نیز تغییر یافته است. براساس گزارش دفتر جمعیت سازمان ملل متحد در سال ۲۰۱۱، ایران در میان ۲۰ کشور پر جمعیت دنیا قرار دارد (۳). با این‌حال، در سه دهه اخیر، نرخ باروری کل (TFR) به شکل قابل توجهی، از ۷ تولد برای هر زن در سال ۱۳۵۸ به ۱/۹ تولد در سال ۱۳۸۵ (۴) و ۱/۸ تولد در سال ۱۳۹۰ (۵)، کاهش یافته است. کاهش باروری نه تنها در مناطق شهری، بلکه در مناطق روستایی هم مشاهده می‌شود. هم اکنون ایران کمترین نرخ باروری را در خاورمیانه دارد (۶).

میزان باروری متأثر از عوامل اقتصادی، اجتماعی، فرهنگی، جمعیت‌شناسی، بیولوژیکی و حتی جغرافیایی است. متون مختلفی به انتقال باروری در ایران و چگونگی تأثیر افزایش سطح تحصیلی، کاهش مرگ و میر کودکان، شهرنشینی، دسترسی گسترده به برنامه‌های پیشگیری از بارداری و اهمیت کیفیت در مقابل کمیت تعداد فرزندان بر روی کاهش آن پرداخته‌اند (۱۱-۷). با وجود پژوهش‌های گوناگونی که در گوشه و کنار جهان به بررسی عوامل مؤثر بر روی باروری پرداخته‌اند، همچنان بسیاری از این عوامل و یا میزان تأثیرگذاری آن‌ها به‌طور کامل مشخص نیست و این نشان‌دهنده ضرورت انجام مطالعات بیشتر در این زمینه می‌باشد.

در برخی از مطالعات، جمعیت‌شناسان به منظور بررسی تأثیر عوامل مختلف بر روی تعداد فرزندان زنده به دنیا آمده، از مدل‌های آماری استفاده می‌کنند. بسیاری از آن‌ها در مطالعات خود برای مدل سازی تعداد فرزندان زنده به دنیا آمده از مدل‌هایی که بر روی داده‌های شمارشی برازش می‌یابند، استفاده کرده‌اند. Wang و Famoye (۱۹۹۷) از مدل رگرسیون پواسن تعمیم یافته به منظور غلبه بر کم و بیش پراکندگی به جای مدل رگرسیون پواسن استاندارد استفاده کردند (۱۲). دو مدل با داده‌های صفر متورم با استفاده از مدل‌های پواسن و گاما توسط Melkerson و Rooth (۲۰۰۰) با یکدیگر مقایسه شدند (۱۳). Olfa و El-Lagha (۲۰۰۲) مدل رگرسیون چندجمله‌ای و مدل رگرسیون پواسن را بعد از تصحیح کم پراکندگی روی داده‌ها برازش دادند (۱۴). Hondroyiannis (۲۰۰۴)، مدل پواسن با روش برآورد شبه درست نمایی ماکسیمم را به کار برد (۱۵).

با این‌حال یکی از مسائل مورد علاقه پژوهشگران در مطالعات جمعیتی، پیش‌بینی طبقه تعداد فرزندان زنده به دنیا آمده می‌باشد. طبقه‌بندی یکی از انواع روش‌های داده‌کاوی و روشی چند متغیره است که با جدا کردن مجموعه‌های متمایز اشیاء (مشاهدات) و یا تخصیص دادن اشیاء (مشاهدات) جدید به طبقات از پیش تعیین شده سروکار دارد. بسیاری از پژوهشگران بدین منظور از تحلیل ممیزی، K نزدیکترین همسایگی، رگرسیون لجستیک و رگرسیون پروبیت که از جمله روش‌های متداول در این زمینه هستند و به شکل وسیعی به کار می‌روند، استفاده کرده‌اند (۱۸-۱۶).

به منظور انجام تحلیل ممیزی خطی پیش‌فرض‌های زیر باید برقرار باشد:

- کلیه متغیرهای پیش‌بین در هر طبقه باید دارای توزیع نرمال چندمتغیره باشند.
- ماتریس‌های واریانس- کوواریانس در هر طبقه باید با یکدیگر برابر باشند (۱۹).

که پیش‌فرض نرمال بودن ضروری است، با این‌حال این روش بدون توجه به برقراری این پیش‌فرض برای کلیه متغیرها توسط پژوهشگران به‌کار برده می‌شود. از سوی دیگر، این روش تنها برای متغیرهای پیش‌بین (کمکی) پیوسته طراحی شده و متغیرهای پیش‌بین رسته‌ای باید به متغیرهای ظاهری تبدیل شوند که این عمل، منجر به افزایش تعداد متغیرها می‌گردد. به علاوه کلیه متغیرهایی که در ترکیب خطی وارد می‌شوند باید کامل باشند، به عبارت دیگر مشاهده‌ای که یک مقدار گمشده داشته باشد از تحلیل حذف می‌گردد که منجر به اریبی حاصل از کاهش تعداد نمونه می‌گردد. همچنین اگر متغیرهای پیش‌بین شامل هر دو نوع متغیر پیوسته و دوحالتی باشند، این روش نتایج غیر قابل اعتمادی به دنبال خواهد داشت (۲۰).

روش‌های رگرسیون لجستیک و پروبیت نیز از دیگر روش‌های پارامتری در مطالعات طبقه‌بندی می‌باشند. برآمد نهایی این روش‌ها، برآورد نسبت مواردی است که در طبقات مختلف متغیر وابسته قرار گرفته‌اند. این روش‌ها همانند تحلیل ممیزی خطی، آزاد- توزیع نیستند، روشی برای تحلیل نمونه‌ها با مقادیر گمشده در یک متغیر ندارند، تنها برای متغیرهای وابسته رسته‌ای قابل استفاده هستند و همانند کلیه مدل‌های پارامتری، همه متغیرهای به کار رفته در تحلیل توسط پژوهشگر تعیین می‌گردند (۲۰).

روشی که در دهه‌های اخیر با پیشرفت نرم افزارهای رایانه‌ای برای طبقه‌بندی داده‌ها به‌کار می‌رود، درخت تصمیم است. درخت تصمیم نوعی روش ناپارامتری است که در اکثر روش‌های استخراج آن، هیچ پیش‌فرض آماری وجود ندارد. در سال‌های اخیر، پژوهشگران علوم مختلف با حجم بزرگی از داده‌ها سرو کار دارند که اغلب این داده‌ها نرمال نیستند و این امر موجب توسعه روش‌های ناپارامتری شده است. این روش‌ها برخلاف سایر روش‌های آماری که ابتدا بر حسب تئوری بسط داده شده‌اند، به دلیل حجم بزرگ داده‌ها با پیشرفت نرم‌افزارهای آماری توسعه یافته‌اند. درخت تصمیم به دلیل انعطاف‌پذیری و ویژگی‌های خاص، به خصوص خروجی آن که یک گراف می‌باشد و تفسیر آن را ساده‌تر می‌نماید، مقبولیت عام یافته است (۲۱). یک درخت تصمیم از سه جزء اصلی ریشه، گره داخلی و گره خارجی (برگ) تشکیل شده است که این اجزا با توجه به الگوریتم‌های مختلفی تعیین می‌گردند. الگوریتم AID که مخفف تعیین اثرات متقابل خودکار است توسط Morgan و Sonquist در سال ۱۹۶۳ ارائه شد (۲۲). در سال ۱۹۷۳ این الگوریتم *توسط* Morgan و Messenger به AID Theta (THAID) که یک الگوریتم جستجوی دنباله‌ای برای تحلیل متغیرهای وابسته با مقیاس اسمی بود (۲۳) و در سال ۱۹۸۰ به الگوریتم کشف متقابل خودکار کای اسکوئر

(*CHAID*) ، توسط Kass ارتقا یافت (۲۴). این سه روش از تقسیمات چند سطحی برای تولید درخت طبقه‌بندی استفاده می‌کنند. الگوریتم طبقه‌بندی و درخت رگرسیونی کارت، که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌گردد، توسط Breiman و همکارانش در سال ۱۹۸۴ به طور کامل معرفی گردید. این روش که ابتدا تنها برای متغیرهای پیش‌بین کمی طراحی شد، بعدها برای متغیرهای کیفی نیز تعمیم یافت (۲۵).

سه الگوریتم *AID*، *THAID* و *CHAID* برخلاف الگوریتم کارت، آزاد- توزیع نیستند و از آزمون‌های معنی‌داری روی متغیرهای پیش‌بین برای ایجاد تقسیمات و تعیین اندازه درخت استفاده می‌کنند. این روش‌ها در فرایند رشد، هرس درخت و برآورد خطای طبقه‌بندی با یکدیگر تفاوت دارند (۲۰).

با توجه به رویارویی بسیاری از پژوهشگران با مسائلی که هدف آن طبقه‌بندی مجموعه داده‌های موردنظر به روشی دقیق، کارا و قابل درک می‌باشد، هدف از این مقاله معرفی و ارائه درخت تصمیم به عنوان روشی جایگزین برای روش‌های متداول کلاسیک در طبقه‌بندی تعداد فرزندان زنده به دنیا آمده با استفاده از الگوریتم کارت براساس داده‌های طرح «بررسی رفتارهای ازدواج و باروری زنان حداقل یکبار ازدواج کرده، ۴۹-۱۵ ساله در استان سمنان- ۱۳۹۱» (۲۶) است. دلیل انتخاب الگوریتم کارت برای مدل سازی، به دلیل امتیازات این روش در مقایسه با روش‌های کلاسیک و پارامتری متداول و غلبه بر معایبی است که سایر الگوریتم‌ها در مدل سازی دارند. در ادامه این مقاله به معرفی، نحوه استخراج و مزایای الگوریتم کارت پرداخته خواهد شد.

روش بررسی

هدف از مطالعه پیشرو، طبقه‌بندی تعداد فرزندان زنده به‌دنیا آمده با استفاده از الگوریتم کارت و بر اساس داده‌های طرح «بررسی رفتارهای ازدواج و باروری زنان حداقل یکبار ازدواج کرده، ۴۹-۱۵ ساله در استان سمنان- ۱۳۹۱» بود، که در این طرح تغییرات خانواده و شناخت عوامل مؤثر بر آن مطالعه شده بود. داده‌ها از طریق یک طرح پیمایش- مقطعی با استفاده از پرسشنامه ساختاریافته در پاییز سال ۱۳۹۱ جمع‌آوری گردید. ۴۰۵ زن ۴۹-۱۵ ساله متعلق به خانوارهای معمولی ساکن استان سمنان که حداقل یکبار ازدواج کرده‌اند، نمونه طرح را تشکیل داد. در طرح مذکور با توجه به اهداف آن، متغیرهای گوناگونی اندازه‌گیری شد که متغیرهای سن در اولین ازدواج، نوع ازدواج، سطح تحصیلی، وضعیت شغلی، محل تولد و کوهورت موالید با توجه به این‌که می‌توانستند بر روی طبقه‌بندی متغیر تعداد فرزندان زنده به‌دنیا آمده مؤثر باشند، به عنوان متغیرهای پیش‌بین انتخاب شدند.

روش‌های مبتنی بر درخت تصمیم فضای متغیرهای پیش‌بین را به صورت بازگشتی به ناحیه‌های مجزا، افراز کرده و داده‌های متناظر را به طبقات تخصیص می‌دهد (۲۷). این افراز بازگشتی منجر به برازش مدل تکه‌ای ثابت روی ناحیه‌های افراز شده فضای متغیر پیش‌بین خواهد شد (۲۸). برای این‌که هر گره افراز شود، تمام افرازهای ممکن برای هر متغیر پیش‌بین ارزیابی می‌شوند. متغیر و نقطه افراز متناظر با آن به گونه‌ای انتخاب می‌شود که بهترین تفکیک بین دو گره حاصل شود. این روند به صورت بازگشتی ادامه می‌یابد تا این‌که هر گره شامل تعداد محدودی از حالت‌ها شود. بعد از ساخت یک درخت بزرگ، قواعدی برای هرس و تعدیل کردن اندازه درخت به‌کار می‌رود (۲۹ و ۲۷).

الگوریتم کارت یکی از روش‌های مهم است که با استفاده از طبقه‌بندی و رگرسیون درختی به وسیله تقسیم‌بندی دو تایی به تحلیل مجموعه داده‌های بزرگ می‌پردازد. Breiman و همکاران (۱۹۸۴)، Steinberg و Colla

(۱۹۹۵) و Steinberg و همکاران (۱۹۹۸) امتیازات این روش را به صورت زیر بیان نمودند (۲۵، ۳۱-۳۰).

۱. کارت هیچ پیش‌فرضی روی توزیع متغیرهای پیش‌بین و وابسته ندارد و هیچ توزیع آماری برای متغیرهای تحلیل در نظر نمی‌گیرد.
۲. متغیرهای پیش‌بین در کارت می‌توانند آمیخته‌ای از متغیرهای رسته‌ای و پیوسته باشند.
۳. کارت روشی برای مقابله با مقادیر گمشده یک متغیر دارد و در نتیجه، مشاهده‌ای به دلیل یک مقدار گمشده، از تحلیل حذف نمی‌شود.
۴. کارت تحت تأثیر هیچ یک از مقادیر دورافتاده، همخطی یا ساختارهای توزیع خطا که بر روی روش‌های پارامتری اثرگذار است، نمی‌باشد. مقادیر دور افتاده در یک گره مجزا قرار می‌گیرند و هیچ اثری بر روی افرازها ندارند. برخلاف روش‌های پارامتری، کارت از همخطی میان متغیرها برای ساختن افرازهای جانشین استفاده می‌کند.
۵. کارت قابلیت تعیین و آشکار ساختن اثرات متقابل در مجموعه داده‌ها را دارد.
۶. کارت تحت تبدیل‌های یکنوا روی متغیرهای پیش‌بین، بدون تغییر باقی می‌ماند. در نتیجه تبدیل‌های لگاریتمی، توان دوم و جذر روی متغیرهای پیش‌بین، اثری بر روی درخت ندارد.
۷. در صورت عدم وجود تئوری که پژوهشگر را راهنمایی نماید، کارت می‌تواند به عنوان یک روش اکتشافی عمل نماید.
۸. اصلی‌ترین امتیاز کارت، کارایی آن در برخورد با مجموعه داده‌های بزرگ است.
۹. خروجی گرافیکی کارت درک آن را آسان می‌کند.

الگوریتم کارت شامل ۳ مرحله است (۲۵):

- مرحله اول شامل ساخت درخت است که با استفاده از افرازهای بازگشتی صورت می‌گیرد. به عبارت دیگر این مرحله تنها شامل افراز مجموعه داده‌ها به بخش‌های کوچکتر است. مهم ترین وجه تمایز درخت‌ها، چگونگی انتخاب یک ضابطه برای تقسیم‌بندی مجموعه داده‌ها در هر گره درخت است. انتخاب یک ضابطه افراز به معنی انتخاب یک متغیر از میان متغیرهای پیش‌بین و بهترین سطح افراز است.
- مرحله دوم شامل توقف افراز یا رشد درخت است. در این مرحله بزرگترین درخت ساخته می‌شود. ممکن است در این مرحله، درخت بیش برازش شده باشد.
- مرحله سوم شامل هرس کردن درخت که نتیجه آن به دست آوردن درخت‌های متوالی ساده‌تر است، می‌باشد. این مرحله از طریق قطع افزایشی گره‌های مهم صورت می‌گیرد. در واقع این فرایند راه حلی برای رفع کاستی‌های ضابطه‌های توقف افراز است. پس از هرس درخت، درخت بهینه با توجه به ضوابط و معیارهای خاص، از میان درخت‌های متوالی کاهشی تولید شده بواسطه فرایند هرس انتخاب می‌شود.

ساخت درخت

ساخت درخت از گره ریشه آغاز می‌گردد که شامل کلیه مشاهدات موجود (در نمونه آموزشی) است. فرایند ساخت درخت، افراز را با این گره شروع کرده و بهترین متغیر ممکن برای افراز این گره به دو زیر گره دیگر انتخاب می‌شود. به همین ترتیب برای یافتن بهترین متغیر، باید تمام متغیرهای ممکن و مقادیر آن‌ها برای افراز بررسی شود. در این مورد نرم‌افزارهایی وجود دارند که زمان جستجو را کاهش داده و به دنبال بهترین متغیر مورد نظر هستند. در انتخاب بهترین متغیر، افرازکننده به دنبال ماکسیمم کردن میانگین خلوص (همسانی) در گره‌های تولید شده براساس معیارهای مختلف

برای اندازه‌گیری خلوص است که به آن‌ها ضابطه یا تابع افراز گفته می‌شود. الگوریتم کارت از تابع افراز جینی و دوتایی استفاده می‌کند که در صورت دو حالتی بودن متغیرهای اسمی منجر به نتایج یکسانی می‌شوند (۲۵).

- **شاخص افراز جینی**

شاخص افراز جینی معمولاً در مدل‌های درختی با تقسیمات دوتایی در هر گره، مورد استفاده قرار می‌گیرد که برای گره‌ای مانند t و متغیر وابسته با k طبقه با استفاده از معادله (۱) تعریف می‌شود (۲۵):

(۱)

$$Gini(t) = 1 - \sum_{j=1}^{k} p^2 \left[c = c_j \mid T = t \right]$$

رابطه فوق زمانی‌که مشاهدات فقط متعلق به یک طبقه باشند، برابر صفر و وقتی احتمال تعلق به هر طبقه برابر باشد بیشترین مقدار ممکن را اختیار می‌نماید.

برای متغیرها با مقادیر پیوسته، هر انشعاب محتملی باید در نظر گرفته شود. نقطه میانه بین دو مقدار متوالی (بردارها باید با توجه به مقدار متغیری که شاخص افراز جینی برای آن محاسبه می‌شود، مرتب شوند)، به عنوان نقطه انشعاب محتمل انتخاب می‌شود. نقطه‌ای که شاخص افراز جینی را حداقل کند، نقطه انشعاب برای آن متغیر خواهد بود. به عنوان مثال، C_1 بردارهایی از مجموعه بردارهای C هستند که در آن‌ها «نقطه انشعاب»$\leq A$ و C_2 زیرمجموعه‌ای از C است که «نقطه انشعاب»$>A$ می‌باشد.

- **شاخص افراز دوتایی**

شاخص دیگری که الگوریتم کارت می‌تواند از آن استفاده کند شاخص افراز دوتایی است که برای متغیر پاسخی با k طبقه به صورت زیر تعریف می‌شود:

(۲)

$$\Delta i(t) = \frac{p_l p_r}{4} \left[\sum_{k=1}^{n} \left| p(k \mid t_l) - p(k \mid t_r) \right| \right]^2$$

در معادله (۲)، p_l و p_r به ترتیب احتمال‌های مربوط به گره چپ و راست است. این شاخص، درخت طبقه‌بندی متقارن‌تری را معرفی می‌نماید ولی از سرعت کمتری نسبت به شاخص جینی برخوردار می‌باشد. در حقیقت مزیت شاخص جینی نسبت به این شاخص، سرعت بالاتر آن در انجام محاسبات می‌باشد (۲۵).

با توجه به سرعت بالاتر شاخص افراز جینی و کاربردی‌تر بودن این شاخص (۲۵ و ۲۰)، در این مقاله برای مدل سازی متغیر تعداد فرزندان زنده به دنیا آمده، از شاخص افراز جینی در الگوریتم کارت استفاده شده است.

توقف رشد درخت

فرایند ساخت درخت تا جایی ادامه می‌یابد که تنها یک مشاهده در هر گره وجود داشته باشد، یا تمام مشاهدات درون گره دارای توزیع یکسانی باشند و یا ضابطه افراز بوسیله حدود مشخص شده توسط پژوهشگر متوقف گردد. البته گاه این حدود به اندازه کافی کوچک نیست که این امر موجب می‌شود داده‌ها به خوبی افراز نگردند و درخت، اطلاعات موجود در داده‌ها را به خوبی برازش ندهد. به عبارت دیگر گره‌های پایانی به‌صورت بالقوه قابلیت افراز را خواهند داشت. به‌طور معمول با رسیدن به یک درخت بزرگ فرایند افراز متوقف می‌شود. در این حالت این درخت بیش از اطلاعات موجود در داده‌ها، برازش داده شده است که عموماً به آن بیش برازش نیز می‌گویند. در واقع این درخت همه حالات خاص در مجموعه داده‌ها را که رخداد برخی از آن‌ها نامحتمل است، در نظر می‌گیرد.

هرس درخت

درخت طبقه‌بندی بزرگ لزوماً درخت مناسب‌تری نیست، از این رو باید راهکارهایی در نظر گرفته شود تا بتوان اندازه مناسبی برای یک درخت طبقه‌بندی تعیین نمود به طوری‌که درخت معرفی شده برای نمونه آزمون (مشاهدات جدید) و نمونه آموزشی (مشاهدات موجود) یکسان عمل نماید و همچنین دقت مناسبی داشته باشد. روش‌های به کار گرفته شده بدین منظور را هرس کردن می‌نامند. به عبارت دیگر، درخت بهینه، درختی است که از لحاظ اندازه و میزان خطای طبقه‌بندی بهینه باشد. بنابراین ابتدا باید برآورد خطای طبقه‌بندی بررسی شده و سپس با استفاده از روش‌های مرسوم درخت را هرس کرده و در نهایت، از میان درخت‌های هرس شده با استفاده از قاعده‌ای مناسب درختی با کمترین هزینه و اندازه مناسب انتخاب گردد. هرس کردن یک درخت طبقه‌بندی از دو مرحله تشکیل می‌شود. در مرحله اول روش‌هایی به کار برده می‌شود که موجب توقف توسعه بیش از اندازه درخت طبقه‌بندی به طوری‌که از نظر آماری دقیق یا معنی‌دار نباشد، می‌شود. این روش‌ها به‌عنوان پیش هرس شناخته می‌شوند. در مرحله دوم شاخه‌های غیر مفید درخت طبقه‌بندی حذف می‌شوند که به این مرحله پس هرس می‌گویند. شیوه‌های به کاربرده شده در این مرحله باعث ایجاد درختی کوچکتر با دقتی مناسب می‌گردد (۳۲).

بدیهی است که در تمام روش‌های طبقه‌بندی دستیابی به الگویی که بدون خطا باشد، غیر ممکن است، ولی هدف معرفی الگویی با کمترین خطا می‌باشد. هدف از معرفی یک درخت طبقه‌بندی دستیابی به یک الگوی طبقه‌بندی براساس مشاهدات موجود است به طوری‌که این الگوی معرفی شده از دقت مناسبی برای مشاهدات جدید برخوردار باشد. یعنی این مدل بتواند در مورد یک مشاهده جدید به خوبی و با کمترین خطای ممکن، تصمیم‌گیری نماید.

با توجه به این‌که در نظر گرفتن میزان طبقه‌بندی اشتباه براساس نمونه‌ای که درخت طبقه‌بندی با توجه به آن ساخته شده است، به عنوان میزان خطا مناسب نمی‌باشد، روش‌هایی برای برآورد میزان خطای یک درخت طبقه‌بندی مبتنی بر روش‌های روایی متقاطع مورد استفاده قرار می‌گیرند که می‌توان به روش روایی متقاطع k مرتبه‌ای اشاره کرد.

این روش زمانی‌که اندازه مشاهدات موجود با توجه به متغیرهای پیش‌بین کافی نباشد، مناسب است. در این روش مشاهدات موجود به طور تصادفی به k قسمت با اندازه های مساوی تقسیم و در هر مرحله k-1 قسمت به عنوان نمونه آموزشی و قسمت دیگر به عنوان نمونه آزمون در نظر گرفته می‌شود. براساس نمونه آموزشی معرفی شده، درخت مناسب ایجاد و خطای مربوط به آن با توجه به نمونه آزمون برآورد می‌گردد. از آن‌جا که مشاهدات به k قسمت تقسیم شده‌اند این اقدام نیز k بار صورت می‌پذیرد و k اندازه خطا که میانگین این خطاها به عنوان برآوردی از خطای درخت تصمیم با توجه به کل مشاهدات است، به‌دست می‌آید. این روش از دقت مناسبی در برآورد خطا برخوردار است و برای نمونه‌های کوچک مناسب می‌باشد (۳۳ و ۲۵، ۲۱). با توجه به این نکته، در مقاله پیشرو از این روش برای محاسبه دقت درخت تصمیم استفاده شده است.

در بخش یافته‌ها الگوریتم درختی کارت به منظور پیش‌بینی تعداد فرزندان زنده بدنیا آمده به وسیله متغیرهای پیش‌بین شامل سن در اولین ازدواج، نوع ازدواج، سطح تحصیلات، وضعیت شغلی، محل تولد و کوهورت موالید در طرح «بررسی رفتارهای ازدواج و باروری زنان حداقل یکبار ازدواج کرده، ۴۹-۱۵ ساله در استان سمنان- ۱۳۹۱» (۲۶)، معرفی می‌شود. در این بخش، درخت تصمیم برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده، با استفاده از نرم افزار STATISTICA نسخه ۷ ارائه شده است. با وجود چندین الگوریتم مختلف برای معرفی و ساخت درخت

تصمیم، با توجه به امتیازات الگوریتم کارت که به آن‌ها اشاره شد، در این پژوهش از این الگوریتم برای ساخت درخت استفاده گردید.

یافته‌ها

متغیرهای این مطالعه به صورت زیر تعریف می‌شوند که تعداد فرزندان زنده به دنیا آمده متغیر پاسخ و سایر متغیرها به عنوان متغیر پیش‌بین در نظر گرفته شده‌اند:

تعداد فرزندان زنده به دنیا آمده: بیانگر تعداد فرزندانی است که پاسخگویان، زنده به دنیا آورده‌اند و سؤال مورد نظر نیز «چند فرزند زنده به دنیا آورده‌اید؟» می‌باشد. این متغیر، در این مطالعه در چهار سطح ۰ و ۱ و۲ و ۳ و بیشتر اندازه‌گیری شده‌اند.

سن در اولین ازدواج: این متغیر بیانگر سن پاسخگو در زمان اولین ازدواج و برحسب سال اندازه‌گیری شده است.

نوع ازدواج: این متغیرکه نسبت فامیلی بین زن و شوهر را نشان می‌دهد با دو گزینه خویشاوند و غیرخویشاوند مشخص شده است.

سطح تحصیلی: این متغیر در دو سطح دیپلم و بالاتر و زیر دیپلم تعیین شده است.

وضعیت شغلی: اشتغال انجام هر گونه فعالیتی است که پاسخگو به عنوان شغل خود اعلام و بابت آن مزد دریافت می‌کند و یا به نوعی در تولید درآمد خود سهیم است (اعم از اشتغال در منزل و یا خارج از منزل) (۳۴). این متغیر به‌صورت دو سطحی با دو گزینه شاغل و غیرشاغل سنجیده شده است. غیرشاغلان کلیه زنان خانه‌دار، محصل و مستمری بگیر را در بردارند.

محل تولد: این متغیر به‌صورت دو سطحی با دو گزینه شهر و روستا مشخص شده است.

کوهورت موالید: کوهورت موالید عبارت است از موالید یک گروه از افراد که در دوره یکسانی متولد شده‌اند و زمان فردی مشابهی را در بستر تاریخی یکسانی تجربه کرده‌اند (۳۵). نمونهٔ بررسی‌شده شامل زنان ۴۹ـ۱۵ ساله هستند که متولدان سال‌های ۱۳۴۰ تا ۱۳۷۵ را شامل می‌شوند و در این مطالعه در سه **کوهورت موالید** قرار گرفته‌اند. متولدان سال‌های ۱۳۴۰ تا ۱۳۴۹ به‌عنوان کوهورت موالید دهه ۱۳۴۰، متولدان سال‌های ۱۳۵۰ تا ۱۳۵۹ به‌عنوان کوهورت موالید دهه ۱۳۵۰ و متولدان سال‌های ۱۳۶۰ تا ۱۳۷۵ به‌عنوان کوهورت موالید بعد از سال ۱۳۶۰ در نظر گرفته شده‌اند.

در این مطالعه میانه و نمای تعداد فرزندان زنده به دنیا آمده با یکدیگر مساوی و برابر با ۲، همچنین دامنه آن برابر با ۳ فرزند به‌دست آمد. ۴۵/۲ درصد زنان مطالعه، ۲ فرزند زنده به دنیا آورده‌اند و ۱۱/۴ درصد زنان مطالعه، بی‌فرزندی را تجربه کرده‌اند. از کوهورت‌های موالید مختلف، تقریباً به‌صورت مساوی در مطالعه شرکت داشته‌اند. ۸۰/۲ درصد از زنان غیرشاغل و ۶۷/۱ درصد از آنان دارای تحصیلات زیر دیپلم بوده‌اند. نوع ازدواج ۴۰/۷ درصد از زنان خویشاوندی و محل تولد اکثر زنان (۷۷/۳ درصد)، شهر بوده است.

مدل کارت برای طبقه‌بندی تعداد فرزندان زنده به دنیا (مدل ۱)

با وارد نمودن متغیرهای پیش‌بین سن در اولین ازدواج، نوع ازدواج، سطح تحصیلی، وضعیت شغلی، محل تولد و کوهورت موالید با فرض استفاده از شاخص افراز جینی و احتمالات پیشین برآوردی درخت شکل (۱) که شامل ۱۱ گره و ۱۲ برگ (تعداد قوانین استخراج شده) است، ساخته شد. همان‌گونه که ملاحظه می‌شود در این مدل کلیه متغیرهای پیش‌بین به عنوان گره در مدل درختی کارت وارد شده، متغیر کوهورت موالید در ریشه درخت قرار گرفته و مؤثرترین متغیر برای تقسیم‌بندی در نظر گرفته شده است.

جدول (۱) ماتریس اغتشاش را نشان می‌دهد که براساس آن می‌توان دقت مدل (۱) را محاسبه نمود.

جدول ۱- ماتریس اغتشاش مدل (۱)

طبقه مشاهده شده		طبقات پیش‌بینی شده مدل (۱)				جمع
		۰	۱	۲	۳+	
۰	تعداد	۱۸	۱۲	۱۱	۵	۴۶
	درصد کل	۴/۴۴	۲/۹۶	۲/۷۲	۱/۲۳	۱۱/۳۶
۱	تعداد	۱۲	۴۹	۱	۲۴	۸۶
	درصد کل	۲/۹۶	۱۲/۱۰	۵/۹۳	۰/۲۵	۲۱/۲۳
۲	تعداد	۶	۷	۱۴۳	۲۷	۱۸۳
	درصد کل	۱/۴۸	۱/۷۳	۳۵/۳۱	۶/۶۷	۴۵/۱۹
۳ و بیشتر	تعداد	۰	۰	۳۵	۵۵	۹۰
	درصد کل	۰	۰	۸/۶۴	۱۳/۵۸	۲۲/۲۲
تعداد کل		۳۶	۶۸	۲۱۳	۸۸	۴۰۵
درصد کل		۸/۸۹	۱۶/۷۹	۵۲/۵۹	۲۱/۷۳	۱۰۰

سلول‌های مشخص شده در این جدول، طبقه‌بندی درست مدل کارت (۱) را نشان می‌دهند. با توجه به یافته‌های این جدول، دقت مدل (۱) برابر با معادله (۳) است:

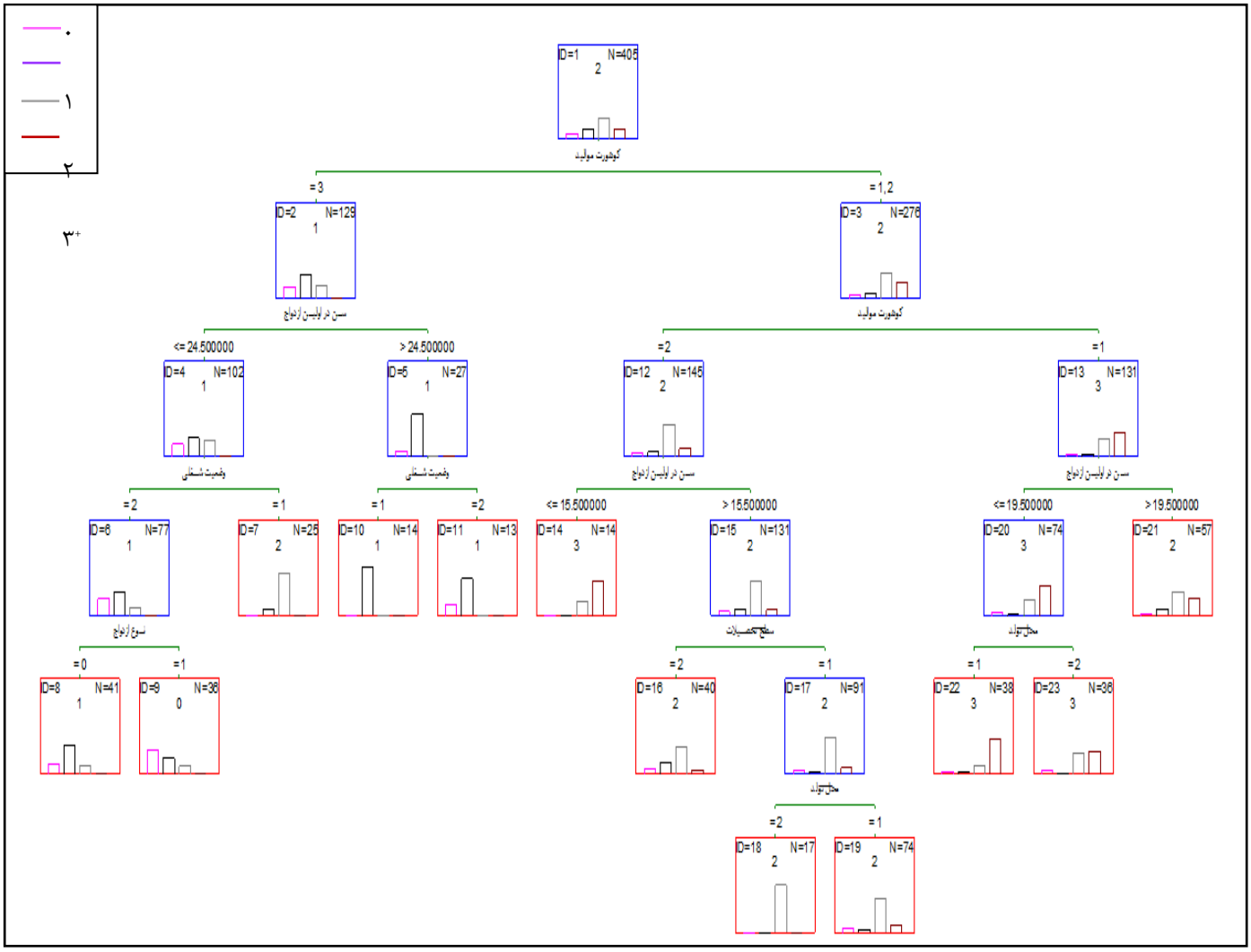
(۳) دقت مدل (۱) (۵۵+۱۴۳+۴۹+۱۸)/(۴۰۵) =۰/۶۵

یعنی ۶۵ درصد موارد توسط این مدل به درستی طبقه‌بندی شده‌اند (۳۵ درصد خطا وجود دارد).

جدول (۲) مخاطره و خطای استاندارد مدل (۱) را برای دو مجموعه داده آموزشی و آزمون نشان می‌دهد. با توجه به این‌که این مقادیر برای هر دو مجموعه داده تقریباً با هم برابر است، می‌توان گفت که مدل از اعتبار مناسبی برخوردار می‌باشد.

جدول۲- مخاطره و خطای استاندارد برای دو مجموعه داده آموزشی و آزمون مدل (۱)

	مدل (۱)	
	مخاطره	خطای استاندارد
مجموعه آموزشی	۰/۳۴۶	۰/۰۲۴
مجموعه آزمون (براساس k اعتبارسنجی متقابل مرتبه‌ای)	۰/۳۶۸	۰/۰۲۶



کوهورت موالید: دهه ۱۳۴۰=۱، دهه ۱۳۵۰=۲، دهه ۱۳۶۰=۳/ **وضعیت شغلی:** شاغل=۱، غیرشاغل=۲ / **سطح تحصیلی:** زیر دیپلم=۱، دیپلم و بالاتر=۲/ **نوع ازدواج:** غیرخویشاوندی=۰، خویشاوندی=۱/ **محل تولد:** شهر =۱، روستا=۲

شکل ۱- مدل کارت برای طبقه‌بندی تعداد فرزندان زنده به‌دنیا آمده (مدل ۱)

مدل کارت برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده با حذف محل تولد (مدل ۲)

در این بخش با توجه به این‌که در مدل (۱)، متغیر پیش‌بین محل تولد بر روی طبقه‌بندی تعداد فرزندان زنده به دنیا آمده اثرگذار نبود، به منظور به‌دست آوردن مدلی با پیچیدگی کمتر (تعداد کمتر گره و برگ)، مدل (۱) با حذف این متغیر، مجدداً ساخته شد.

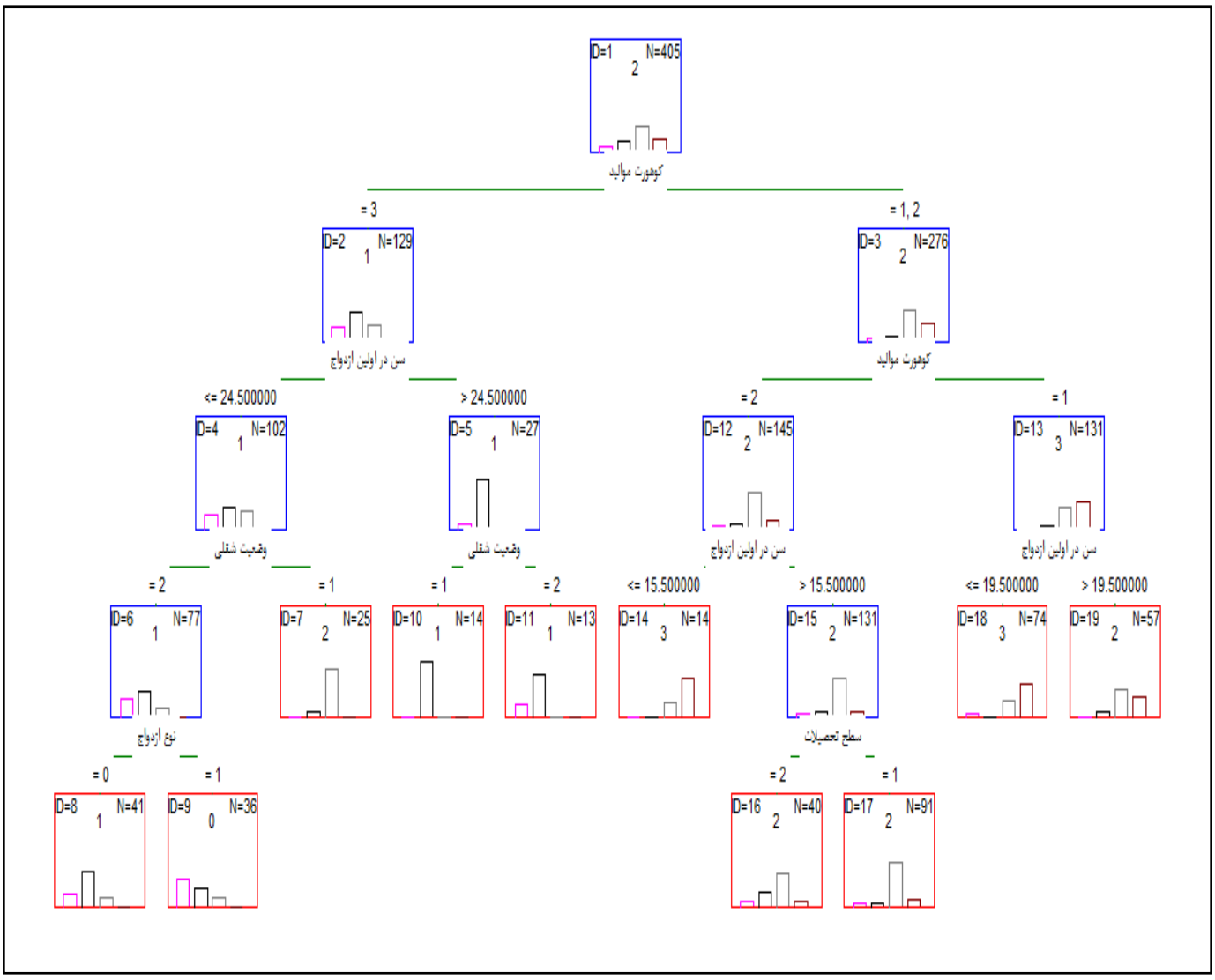
با وارد نمودن متغیرهای پیش‌بین سن در اولین ازدواج، نوع ازدواج، سطح تحصیلی، وضعیت شغلی و کوهورت موالید با فرض استفاده از شاخص افراز جینی با احتمالات پیشین برآوردی، درخت طبقه‌بندی شکل (۲) که شامل ۹ گره و ۱۰ برگ بود، ساخته شد. همان‌گونه که ملاحظه می‌شود کلیه متغیرهای پیش‌بین به عنوان گره در مدل کارت وارد شده‌اند.

دقت حاصل از ماتریس اغتشاش مدل (۲) برابر با دقت حاصل از ماتریس اغتشاش مدل (۱) در معادله (۲)، برابر با ۶۵ درصد به‌دست می‌آید. با توجه به این‌که دقت مدل‌های (۱) و (۲) با یکدیگر برابر و مدل (۲) با ۱۹ گره و برگ، ساده‌تر از مدل (۱) با ۲۳ گره و برگ است، برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده از این مدل استفاده شد. جدول مخاطره و خطای استاندارد برای دو مجموعه آموزشی و آزمون مدل (۲) منطبق بر جدول (۲) برای مدل (۱) است، زیرا قوانین حاصل از دو درخت منطبق بر یکدیگر است و در حقیقت محاسبه دقت درخت مدل (۲) به منظور تأیید انطباق دقت دو مدل انجام گرفته است.

لازم به ذکر است که در مدل (۲) سطح تحصیلی بر روی تعداد فرزندان زنده به دنیا آمده اثرگذار نیست؛ با حذف این متغیر و اجرای الگوریتم کارت مدل پیچیده‌تری حاصل شد که در نهایت از مدل (۲) برای تفسیر نتایج استفاده گردید.

در مدل (۲) متغیر کوهورت موالید در ریشه درخت قرار گرفته و مؤثرترین متغیر برای تقسیم‌بندی در نظر گرفته شده است. قوانین زیر از این مدل قابل استخراج است:

- زنان در ***کوهورت موالید اول*** (دهه ۱۳۴۰) که ***سن ازدواجشان بیشتر از ۱۹/۵ سال*** است ۲ فرزند زنده و افرادی که ***سن ازدواجشان کمتر یا مساوی ۱۹/۵ سال*** است ۳ فرزند زنده و بیشتر به دنیا آورده‌اند.
- برای زنان در ***کوهورت موالید دوم*** (دهه ۱۳۵۰) که ***سن ازدواجشان بیشتر از ۱۵/۵ سال*** است، سطح تحصیلی اثری روی تعداد فرزندان زنده به دنیا آمده نداشته و کلیه افراد ۲ فرزند زنده به‌دنیا آورده‌اند.
- زنان در ***کوهورت موالید دوم*** (دهه ۱۳۵۰) که ***سن ازدواجشان کمتر یا مساوی ۱۵/۵ سال*** است، ۳ فرزند زنده و بیشتر به دنیا آورده‌اند.
- برای زنان در ***کوهورت موالید سوم*** (دهه ۱۳۶۰) که ***سن ازدواجشان بیشتر از ۲۴/۵ سال*** است، وضعیت شغلی در تعداد فرزندان زنده به دنیا آمده آن‌ها، اثری نداشته و برای هر دو وضعیت شغلی (شاغل و غیرشاغل)، این زنان ۱ فرزند زنده به دنیا آورده‌اند.
- زنان در ***کوهورت موالید سوم*** (دهه ۱۳۶۰) که ***سن ازدواجشان کمتر یا مساوی ۲۴/۵ سال*** می‌باشد و شاغل هستند، تعداد ۲ فرزند زنده به‌دنیا آورده‌اند. در حالی‌که تعداد فرزندان غیرشاغلان به نوع ازدواج آن‌ها بستگی داشته است؛ زنانی که با غیرخویشاوند خود ازدواج کرده‌اند ۱ فرزند زنده به دنیا آورده‌اند، اما زنانی که ازدواج خویشاوندی داشته‌اند، تاکنون فرزندی به دنیا نیاورده‌اند.



کوهورت موالید: دهه ۱۳۴۰=۱، دهه ۱۳۵۰=۲، دهه ۱۳۶۰=۳/ **وضعیت شغلی:** شاغل=۱، غیرشاغل=۲ / **سطح تحصیلات:** زیر دیپلم=۱، دیپلم و بالاتر=۲/ **نوع ازدواج:** غیرخویشاوندی=۰، خویشاوندی=۱/ **محل تولد:** شهر =۱، روستا=۲

کوهورت موالید: دهه ۱۳۴۰=۱، دهه ۱۳۵۰=۲، دهه ۱۳۶۰=۳/ **وضعیت شغلی:** شاغل=۱، غیر شاغل=۲/ **سطح تحصیلات:** زیر دیپلم= ۱، دیپلم و بالاتر= ۲/ **نوع ازدواج:** غیرخویشاوندی=۰، خویشاوندی=۱/ **محل تولد:** شهر=۱، روستا=۲

شکل ۲- مدل درختی کارت برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده با حذف متغیر محل تولد (مدل ۲)

بحث و نتیجه‌گیری

پژوهش‌های جمعیت‌شناسی در سایه پیشرفت فناوری، از امکان جمع‌آوری مجموعه داده‌های بزرگ برخوردار شدند. حجم بزرگ داده‌ها از یک سو اطلاعات وسیعی را در اختیار پژوهشگران قرار می‌دهد و از سوی دیگر آنان را با چالش چگونگی استفاده از داده‌ها مواجه می‌نماید. در نتیجه، به‌منظور بهره‌مندی از اطلاعات، نیاز به مدیریت و سازمندهی صحیح آن‌ها وجود دارد. عدم مدیریت صحیح داده‌ها و تحلیل نادرست نتایج، علاوه بر هدر رفتن هزینه و بودجه صرف شده در این مطالعات، می‌تواند منجر به تصمیم‌گیری‌های نادرست شود. یکی از مباحث مطرح در جمعیت‌شناسی باروری است که در مطالعات مختلف از جهات گوناگون به آن پرداخته شده است و تعداد فرزندان زنده به دنیا آمده یکی از متغیرهای کلیدی برای سنجش آن می‌باشد.

در بررسی عوامل تأثیرگذار بر باروری، تحقیقات و بررسی‌های بسیاری در سرتاسر دنیا صورت گرفته است. در سال ۱۹۷۶، Caldwell در طرحی با عنوان ضرورت انتقال و تجدید خانواده، بیان داشت که تغییرات باروری که عموماً به صنعتی شدن، شهرنشینی و عقلگرایی نسبت داده شده است، تقریباً موضوع روشن و ثابت شده‌ای است (۳۶). Freed man در مطالعه‌ای در سال ۱۹۷۰ در هنگ کنگ با عنوان «کاهش باروری در هنگ کنگ»، به بررسی شاخص‌های اقتصادی و اجتماعی مؤثر بر باروری و انگیزه کاهش باروری در آن کشور پرداخت. حاصل مطالعه فریدمن این بود که کاهش قابلیت باروری در سنین بالای ۳۰ سال بیشتر است (۳۷). Abedini و Rahman (۲۰۱۰)، عواملی که نقش اساسی در افزایش و جلوگیری از باروری زنان ازدواج کرده بنگلادشی دارند را بررسی نمود. آن‌ها نشان دادند که ازدواج زودهنگام یکی از عوامل اصلی مؤثر روی باروری می‌باشد، اما تنها عامل نیست، اثر متقابل میان آگاهی و شرایط اجتماعی می‌تواند بر روی کاهش باروری حتی زمانی‌که سن ازدواج پایین است، مؤثر باشد (۳۸). Dey و Goswami (۲۰۰۹)، الگوهای باروری را تحلیل و همبستگی‌های میان آن‌ها را در شمال شرقی هند بررسی نمودند. سطح تحصیلی، مذهب، وضعیت شغلی، وضعیت اقتصادی، مرگ و میر کودکان، سن ازدواج، سن زنان و استفاده از وسایل جلوگیری از بارداری به عنوان عوامل مؤثر روی این الگوها در نظر گرفته شد (۳۹). Rahman (۲۰۰۸)، تعداد فرزندان زنده به دنیا آمده را با استفاده از مدل لجستیک دوحالتی با نمونه‌گیری از زنان ۴۹-۱۵ ساله بررسی نمود. سطح تحصیلات زوجین، میانگین هزینه و درآمد خانوار، سن ازدواج و فاصله تولد از عوامل مؤثر بر روی تعداد فرزندان زنده به دنیا آمده بود (۴۰). Kannan و Nagarajan (۲۰۰۸)، با استفاده از مدل رگرسیون چند متغیره عوامل تأثیرگذار روی باروری را بررسی نمودند (۴۱).

استفاده از رگرسیون خطی برای مدل سازی تعداد فرزندان زنده به دنیا آمده زمانی مناسب است که میانگین این متغیر بزرگ باشد، زیرا در این شرایط توزیع این متغیر، تقریباً نرمال می‌شود. اما اگر میانگین فرزندان زنده به دنیا آمده بزرگ نباشد، مانند جوامع با نرخ پایین باروری، استفاده از رگرسیون خطی برای تحلیل تعداد فرزندان زنده به دنیا آمده، مناسب نخواهند بود (۴۲). Barmby و Cigno (۱۹۹۰) الگوهای باروری را با استفاده از یک مدل احتمالی برآورد نمودند (۴۳). Sobel و Arminger (۱۹۹۲) به‌طور همزمان از مدل پروبیت و مدل غیرخطی برای برآورد این الگوها استفاده کردند (۴۴). Caudill و Mixon (۱۹۹۵) رگرسیون‌های سانسور شده را برای داده‌های باروری معرفی کردند (۴۵).

درخت تصمیم به عنوان روشی سودمند در برخورد با حجم زیاد داده ها و به عنوان فرایندی برای تحلیل اکتشافی داده‌ها مورد توجه پژوهشگران علوم مختلف از جمله

پزشکی، جمعیت‌شناسی، جامعه‌شناسی، روان‌شناسی، مدیریت، هواشناسی و غیره به‌منظور تشخیص مدل، پیش‌بینی و تصمیم‌گیری قرار گرفت (۴۸-۴۶). استفاده از درخت تصمیم زمانی‌که با وجود مشاهدات و متغیرهای گوناگون نتوان از روش‌های کلاسیک استفاده نمود، اهمیت و جایگاه ویژه‌ای می‌یابد، زیرا امکان محاسبات سریع و حصول نتایج دقیق را فراهم می‌کند. این روش در مقایسه با روش‌های آماری دیگر دارای مزایایی به شرح ذیل می‌باشد (۲۱):

- در اکثر روش‌های استخراج درخت تصمیم از روش‌های ناپارامتری استفاده می‌شود که نیازی به داشتن توزیع خاصی برای مشاهدات ندارد. به‌ویژه زمانی‌که هدف، طبقه‌بندی متغیر پاسخ براساس تعداد زیاد متغیرهای پیش‌بین است، این روش بسیار مناسب می‌باشد.
- از آن‌جا که درخت تصمیم یک گراف قابل درک و ساده برای طبقه‌بندی معرفی می‌نماید، دیگر نیازی به معرفی یک ساختار تصمیم‌گیری نمی‌باشد.
- درخت تصمیم به خصوص برای متغیرهای پاسخ کیفی بسیار مناسب می‌باشد.
- در الگوهای درختی چون در هر مرحله طبقه‌بندی از یک متغیر استفاده می‌نماید، می‌تواند از تمام اطلاعات موجود مربوط به متغیر استفاده نماید. در نتیجه نحوه برخورد با مقادیر گمشده نسبت به سایر روش‌های آماری مناسب‌تر است.
- در الگوهای درخت تصمیم، نیازی به در نظر گرفتن اثرات متقابل نمی‌باشد.

طبقه‌بندی مجموعه داده‌های مورد نظر پژوهشگران جمعیت‌شناسی و علوم اجتماعی نیاز به ارائه روشی دقیق، کارا و قابل درک برای آنان را ضروری نمود. هدف از این مقاله معرفی و ارائه مدل کارت به عنوان روشی جایگزین برای روش‌های کلاسیک متداول نظیر تحلیل ممیزی و رگرسیون لجستیک برای طبقه‌بندی بود.

کارت، قواعد ساده‌ای برای تعیین گروه‌های با ریسک بالا یا پایین نسبت به متغیر مورد نظر فراهم می‌کند (۵۰-۴۹). در واقع، کارت زیرگروه‌های همگن را با استفاده از روش‌های ناپارامتری استخراج می‌کند (۵۱، ۲۸). در نتیجه این الگوریتم برای آنالیز اکتشافی داده‌ها و ایجاد طبقه‌بندی ساده و قابل تفسیر، همچنین تعریف ضوابط پیش‌بینی با کمترین پیش‌فرض‌ها مفید است (۵۱). سادگی در تفسیر نتایج، آزاد توزیع بودن شاخص‌های به کار رفته در ساخت درخت و نحوه برخورد آن با داده‌های گمشده و دورافتاده از مزایای مهم مدل کارت است که استفاده از آن را تا حد زیادی افزایش داده است. در این مقاله نیز با توجه به این موارد از مدل کارت برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده استفاد شد.

جدول (۳)، خلاصه نتایج دو مدل کارت به‌دست آمده برای طبقه‌بندی تعداد فرزندان زنده به دنیا آمده را نشان می‌دهد. همان‌گونه که ملاحظه می‌شود در مدل (۲) می‌توان با پیچیدگی کمتر و اندازه کوچکتر (تعداد کمتر گره و برگ) به همان دقت مدل (۱) دست یافت.

جدول ۳- مقایسه مدل‌های کارت (۱) و (۲)

	اندازه درخت (مجموع گره و برگ)	دقت مدل (درصد)	دقت مدل در پیش‌بینی هر طبقه			
			۰	۱	۲	۳+
مدل (۱)	۲۳	۶۵	۳۹	۵۷	۷۸	۶۱
مدل (۲)	۱۹	۶۵	۳۹	۵۷	۷۸	۶۱

بر اساس مدل کارت (۲) می‌توان نکات زیر را بیان نمود:

- ✓ زنان در ***کوهورت موالید اول و دوم***، بسته به سن در اولین ازدواجشان که به ترتیب کمتر یا مساوی ۱۹/۵ سال و کمتر یا مساوی ۱۵/۵ سال بوده، ۳ فرزند زنده و بیشتر به دنیا آورده‌اند. این در حالی است که زنان واقع در ***کوهورت موالید اول*** که در سنین بیشتر از ۱۹/۵ سال و زنان واقع در ***کوهورت موالید دوم*** که در سنین بیشتر از ۱۵/۵ سال ازدواج کرده‌اند، ۲ فرزند زنده به دنیا آورده‌اند.
- ✓ ***نوع ازدواج*** روی تعداد فرزندان زنده به دنیا آمده ***کوهورت موالید اول و دوم*** تأثیری نداشته است و این کوهورت‌ها بدون توجه به نوع ازدواج و تنها با توجه به سن در اولین ازدواج ۲ یا ۳ فرزند زنده و بیشتر به دنیا آورده‌اند.
- ✓ در ***کوهورت موالید سوم***، زنانی که در سنین بیشتر از ۲۴/۵ سال ازدواج کرده‌اند، بدون تأثیر هیچ متغیری ۱ فرزند زنده به دنیا آورده‌اند.
- ✓ در ***کوهورت موالید سوم***، زنانی که در سنین کمتر یا مساوی ۲۴/۵ سال ازدواج کرده‌اند، چنان‌چه شاغل باشند، ۲ فرزند زنده و در صورت غیرشاغل بودن برحسب این‌که نوع ازدواج آن‌ها خویشاوندی یا غیرخویشاوندی بوده به ترتیب ۰ و ۱ فرزند زنده به دنیا آورده‌اند.

در تحلیل نتایج به‌دست آمده از این مدل می‌توان نکات زیر را نتیجه گرفت:

- ✓ تأثیر نوع ازدواج روی تعداد فرزندان زنده به دنیا آمده کوهورت موالید سوم و عدم تأثیر آن روی تعداد فرزندان زنده به دنیا آمده کوهورت موالید اول و دوم، می‌تواند ناشی از افزایش آگاهی نسل جوان نسبت به مخاطراتی باشد که فرزندان را در این ازدواج‌ها تهدید می‌نماید.
- ✓ تعداد فرزندان بیشتر زنان شاغل نسبت به زنان غیرشاغل در کوهورت موالید سوم از جنبه‌های مختلفی قابل بحث و بررسی است؛ از جمله این‌که زنان شاغل ممکن است به دلیل داشتن امنیت اقتصادی، امکان داشتن فرزند بیشتر را داشته باشند. همچنین این زنان به دلیل حضور در اجتماع با مخاطرات تک فرزندی بیشتر از زنان غیرشاغل آشنا هستند (۵۲).

تقدیر و تشکر

این مقاله مستخرج از طرح کاوش داده های جمعیتی با استفاده از درخت تصمیم (ابلاغ شماره ۲۰/۱۵۲۸۳ مورخ ۹۳/۱۱/۵) است که با حمایت مالی مؤسسه مطالعات و مدیریت جامع و تخصصی جمعیت کشور در سال ۱۳۹۳ انجام شده است. نویسندگان مقاله بر خود لازم می‌دانند از همکاری سرکار خانم دکتر رازقی نصرآباد برای در اختیار قرار دادن داده‌های طرح «بررسی رفتارهای ازدواج و باروری زنان حداقل یکبار ازدواج کرده، ۱۵-۴۹ ساله در استان سمنان- ۱۳۹۱» کمال تشکر را داشته باشند.

References

1. Agha H. The study fertility of women in Iran and its relationship with socio-economic indicators.Research Report. Shiraz: Population Studies Center. Shiraz University; 1985: 10-12. (Persian).

2. Cleland, JG. Trends in Human Fertility. In H. K. Heggenhougen (Ed.), International Encyclopedia of Public Health. Oxford: Academic Press; 2008: 364-371.

3. United Nations, department of economic and social affairs, population division, population estimates and projections section. world population prospects, revision 2011 revision. (Cited 28 March 2013).

4. Abbasi-Shavazi MJ, McDonald P, Hosseini-Chavoshi M. The fertility transition in Iran: revolution and reproduction. 2st ed. Canbera: Springer. National University Canbera. 2009;48-50.

5. Abbasi-Shavazi MJ, Hosseini-Chavoshi M, Banihashemi F, Khosrvi A. Assessment of the own–children estimates of fertility applied to the 2011 Iran Census and the 2010 Iran-MIDHS. International Population Conference, Busan, Korea, 26-31, August 2013.

6. Haub C, Yanagishita, M. world population data sheet. Population Reference Bureau, Washington, DC; 2011.

7. Aghajanian, A. A new direction in population policy and family planning in the Islamic Republic of Iran. Asia-Pacific population journal/United Nations. 1995;10(1).

8. Aghajanian A, Mehryar A. H. Fertility transition in the Islamic Republic of Iran: 1976-1996. Asia-Pacific population journal/United Nations. 1999;14(1).

9. Salehi-Isfahani D, Abbasi-Shavazi MJ, Hosseini-Chavoshi M. Family planning and fertility decline in rural Iran: the impact of rural health clinics. Health Economics. 2010; 19(S1): 159-180.

10. Torabi F. Marriage postponement and fertility decline in Iran. London School of Hygiene and Tropical Medicine (University of London). 2011.

11. Abbasi-Shavazi MJ, Torabi F. Women's Education and Fertility in Islamic Countries Population Dynamics in Muslim Countries Springer; 2012:43-62.

12.Wang W, Famoye F. Modeling household fertility decisions with generalized Poisson regression; J Popul Econ 1997; 10: 273-83.

13. Melkersson M. , Rooth D-O. Modeling Female Fertility Using Inflated Count Data Models. Journal of Population Economics. 2000;13:189-203.

14. Olfa F. El-Lahga AR. A socioeconomic analysis of fertility determinants with a count data models: the case of Tunisia, 2002.

15. Hondroyiannis G. Modeling household fertility decisions in Greece. The Social Science Journal. 2004;41(3): 477-483.

16. Nwakeze NM. The demand for children in Anambra State of Nigeria: A logit analysis. African Population Studies.2007;22(2):175-201.

17. Rahman M, Ahmad T, Hoque A. Factors affecting children ever born in slum areas of Rajshahi city corporation, Bangladesh. Middle East Journal of Nursing. 2008;2(4): 5-10.

18. Hasan M. and Sabiruzzaman. Factors affecting fertility behavior in Bangladesh: A probabilistic approach. Research Journal of Applied Sciences. 2008; 3(1): 70-76.

19. Maddala GS. Limited dependent and qualitative variables in econometrics. Cambridge: Cambridge University Press,1983.

20. Yohannes Y. Classification and Regression Trees, Cart: A User Manual for Identifying Indicators of Vulneability to Famine and Chronic Food In security. International Food Policy Research Institute, 2003. Washington, D.C, USA.

21. Bringman B, Z.Tree decision tree for tree structured data, Springer. 2005:46-58.

22.Morgan JN, Sonquist JA. Problems in the analsis of survey data, and a proposal. Journal of American Statistical Association. 1963; 58: 415–434 .

23.Morgan JN, Messenger RC. THAID: A sequential search program for the analysis of nominal scale dependent variables. Technical report. Institute for Social Research, University of Michgan, Ann Arbor, Mich., U.S.A, 1973.

24. Kass GV. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 1980;29 (2): 119–127.

25. Breiman L, Friedman J, Olsen R, Stone C. Classification and Regression Trees, Chapman & Hall. 1984.

26.Razeghi-Nasrabad H. Marriage and fertility behavior at least once married women, 15-49 years old in 1391 Semnan-Iran. Research Report, National Population Studies & Comprehensive Management Institute; 2013. (Persian).

27. Noon A M, Banerjee M. Computational Methods in Biomedical Research. In Chow S C, Jones B, Liu P J and Peace K (ed), New York, Chapman& Hall .2008:77-101.

28. LeBlanc M. Handbook of Statistics in Clinical Oncology In J Crowley (ed.), Tree-Based Methods for Prognostic Stratiification, New York, Basel, Marcel Dekker Inc. 2001:457-472.

29. Banerjee M, Noone A M. Advances in the Biomedical Sciences. In Biswas A and et al. (ed), New Jersey, John Wiley& Sons, Inc.2008: 265-285.

30. Steinberg D, Colla P. CART: Tree-structured nonparametric data analysis. San Diego, Calif., U.S.A.: Sa ford Systems. 1995.

31.Steinberg D, Colla P, Martin K. CART—Classification and regression trees: Supplementary manual for Windows. San Diego, Calif., U.S.A.: Sa ford Systems. 1998.

32. Frank E, Pruning Decision trees and lists, Phd thesis, University of Waikato, Hamilton, Newzealand, 2000.

33. Izenman AJ. Modern Multivariate Statistical Techniques. In Casella G, Fienberg S , olkin I(ed.), Philadelphia, Springer, 2008:281-314 .

34.Abbasi-Shavazi MJ, Hosseini-Chavoshi M, McDonald PF, Delavar B. Fertility transition in Iran according to the evidence of four proviences, ministiry of health and medical education, 2004.(Persian).

35. Ryder N. The Cohort as a Concept in the Study of Social Change, American Sociological Review.1965; 30: 843-861.

36. Caldwell JC.Toward a restatement of demographic theory population and development review. 1976.

37. Freedman R, Family Planning Programs in the Third World.1970.

38. Abedin S, Rahman JAM. On the dynamics of high-risk fertility in Bangladesh. International Journal of Human Science: 2010;9.

39. Dey S, Goswami S, Fertility pattern and its correlates in North East India, Journal of Human Ecology, 2009; 26 (2):145–152.

40. Rahman M, Predicting the Number of Children Ever Born Using Logistic Regression Model. Biometrics & Biostatistics International Journal; 2009.

41. Kannan KS, Nagarajan V, Factor and Multiple Regression Analysis for Human Fertility in Kanyakumari District. Anthropologist, 2008;10(3): 211-214.

42. Cameron Colin, Trivedi Pravin. Essentials of Count Data Regression. Theoretical econometrics; Wiley, 2007:
24-29.

43. Barmby T, Cigno AA. sequential probability model of fertility patterns. Journal of population economics; 1990; 3: 31-51.

44. Sobel ME, Arminger G. Modeling household fertility decisions: a nonlinear simultaneous probit model. J Am Stat Assoc. 1992;87(417):38-47.

45. Caudill S, Mixon F. Modeling household fertility decisions: Estimation and Testing of Censored Regression Models for Count Data. Empirical Economics. 1995; 20:183-196.

46. Aurangabad M. Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District INDIA, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 2014; 3(2).

47. Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F. Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study, Diabetes Res Clin Pract, 2014;105(3):391-8.

48. Loris Nanni, Alessandra Lumini, Claudio Manna. A Data Mining Approach for Predicting the Pregnancy Rate in Human Assisted Reproduction, Advanced Computational Intelligence Paradigms in Healthcare. 2011;326:97-111.

49. Holfprd T R. Multivarite Methods in Epidemiology. New York, Oxford University Press. 2002:315-342.

50.Siciliano R, Mola F. Multivariate data analysis and modeling through classification and regression tree.Computational Statistics & Data Analysis. 2000; 32: 285-301.

51. Bacchetti P, segal M. Survival trees with time-dependent covariates: Application to Estimating Changes in the Incubation Period of AIDs. Biostatistics 1994; 39:1-20.

52. Saadati M, Bagheri, A. Mining Demographic Data by Decision Tree. Research Report, National Population Studies & Comprehensive Management Institute; 2014. (Persian).

Classification the Number of Children Ever Born using CART Model

Arezoo Bagheri[1], Mahsa Saadati[1]*

1. Associate Professor of Biostatistics, National Population Studies & Comprehensive Management Institute, Tehran, Iran

Abstract

Background & Objective: Discriminant analysis and logistic regression are classical methods for classifying data in several studies. However, these models do not lead in valid results due to not meeting all necessary assumptions. The purpose of this study was to classify the number of Children Ever Born (CEB) using decision tree model in order to present an efficient method to classify demographic data.

Method: In the present study, CART tree model with Gini splitting rule was fitted to classify the number of CEB in fertility behavior of at least once married 15-49 year-old women, in Semnan-2012. 405 women aged 15-49 years old comprised the survey sample.

Results: Women in first and second birth cohorts who had married at an early age had 3 CEB while women who had married at an older age had 2 CEB. Women in third birth cohort who had married at an early age and were employed, had 2 CEB while unemployed women in this cohort whose type of marriages were familial and non-familial had 0 and 1 CEB respectively. Women in the third birth cohort who were married in older age had 1 CEB.

Conclusion: Among important advantages of CART model are the simplicity in interpretation, using distribution-free measures, considering missing data and outliers for construction trees which has increased the usage of this method. Therefore, this method is a suitable way for classifying demographic data in comparison to other classical modeling methods in the conditions that necessary assumptions are not met.

Key words: Classification, Decision Tree, CART Model, Gini Splitting Rule, Children Ever Born (CEB)

Corresponding Author: Mahsa Saadati

Address: National Population Studies & Comprehensive Management Institute, Tehran, Iran.

E-mail: mahsa.saadati@gmail.com