مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۷، شماره‌ی ۱، بهار و تابستان ۱۳۹۵، صص ۸۳– ۱۰۲

کاربرد درخت‌های رده‌بندی در پیش‌بینی ترجیح جنسی فرزندآوری زنان در آستانه‌ی ازدواج ایران

مهسا سعادتی* و آرزو باقری

مؤسسه‌ی مطالعات و مدیریت جامع و تخصصی جمعیت کشور

چکیده: استفاده از درخت تصمیم وقتی با وجود مشاهدات و متغیرهای گوناگون نتوان از روش‌های کلاسیک استفاده نمود، اهمیت و جایگاه ویژه‌ای می‌یابد. تحلیل اکتشافی، تشخیص مدل، پیش‌بینی و تصمیم‌گیری در مورد داده‌ها با اندازه‌ی زیاد، از کاربردهای مهم این روش است. درخت تصمیم وقتی متغیر پاسخ رسته‌ای باشد به درخت رده‌بندی معروف است. هدف اصلی این مطالعه، مقایسه‌ی سه روش مختلف استخراج درخت‌های رده‌بندی CART، CHAID و QUEST به منظور پیش‌بینی ترجیح جنسی فرزندان زنان در آستانه‌ی ازدواج با یکدیگر با استفاده از نرم‌افزار SPSS نسخه‌ی ۲۲ است. داده‌های این مطالعه را ۶۳۶۰ زن مراجعه‌کننده به مراکز بهداشتی به منظور دریافت مشاوره‌ی پیش از ازدواج تشکیل دادند که با روش نمونه‌گیری چندمرحله‌ای و به‌صورت تصادفی در سال ۱۳۹۳ در کل کشور گرداوری شد. نتایج نشان داد که هر سه الگوریتم درخت رده‌بندی از دقت تقریباً مشابهی در پیش‌بینی ترجیح جنسی برای فرزندان برخوردارند، اما با توجه به انطباق بیش‌تر مدلCART به فرضیه‌های موجود، این مدل به‌عنوان مدل نهایی در نظر گرفته شد.CART روش ناپارامتری با قابلیت تفسیرپذیری آسان است که امکان محاسبات سریع و حصول به نتایج دقیق را فراهم می‌نماید. بر اساس نتایج حاصل از این درخت، تعداد فرزندان ایده‌آل، سطح تحصیلی و سن زنان متغیرهای تأثیرگذار بر ترجیح جنسی به دست آمد.

واژگان کلیدی: درخت تصمیم؛ درخت رده‌بندی؛ الگوریتم CART؛ الگوریتم CHAID؛ الگوریتم QUEST.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۵/۷/۲۴، پذیرش: ۱۳۹۶/۳/۲۸.

۱- مقدمه

رده‌بندی متغیر پاسخ یکی از مباحث جذاب و چالش برانگیز در روش‌های آماری است. رده‌بندی یکی از انواع روش‌های داده‌کاوی و روشی چندمتغیره است که هدف آن تفکیک مجموعه‌ی مشاهدات متمایز یا تخصیص مشاهدات جدید به رده‌های از پیش تعیین‌شده می‌باشد. بسیاری از پژوهشگران به این منظور از تحلیل تشخیصی، K نزدیک‌ترین همسایگی، رگرسیون لوژستیکی و رگرسیون پروبیتی که از جمله روش‌های متداول در این زمینه هستند، استفاده کرده‌اند [۲۷-۲۹]. با این حال اعتبار نتایج حاصل از این روش‌ها به برقراری پیش‌فرض‌های آن‌ها بستگی دارد. به‌عنوان مثال به منظور انجام تحلیل تشخیصی خطی باید همه‌ی متغیرهای پیش‌بین در هر طبقه دارای توزیع نرمال چندمتغیره و ماتریس‌های واریانس- کوواریانس برابر باشند [۲۳]. با این‌که در این روش پیش‌فرض نرمال بودن برای اعتبار نتایج ضروری است، اغلب بدون توجه به لزوم برقراری آن برای همه‌ی 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 می‌شود. به علاوه همه‌ی متغیرهایی که در ترکیب خطی وارد می‌شوند باید کامل باشند، به عبارت دیگر هر مشاهده‌ای که تنها یک مقدار گمشده داشته باشد از تحلیل حذف می‌شود که اریبی تورم واریانس حاصل از کاهش تعداد نمونه را به دنبال خواهد داشت. همچنین اگر متغیرهای پیش‌بین شامل متغیرهای پیوسته و دوحالتی باشند، این روش نتایج معتبری ندارد [۴۱]. روش‌های رگرسیون لوژستیکی و پروبیتی نیز از دیگر روش‌های پارامتری در مطالعات رده‌بندی می‌باشند که خروجی نهایی آن‌ها، براورد نسبت مواردی است که در طبقات مختلف متغیر وابسته قرار گرفته‌اند. این روش‌ها همانند تحلیل تشخیصی 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 دیگر مدل‌های پارامتری، همه‌ی متغیرهای به کار رفته در تحلیل، توسط پژوهشگر تعیین می‌شوند [۴۱].

با توجه به این‌که در سال‌های اخیر، پژوهشگران علوم مختلف با حجم بزرگی از داده‌ها سرو کار دارند که اغلب نرمال نیستند، روش‌های ناپارامتری توسعه یافته‌اند. یکی از این روش‌ها که در دهه‌های اخیر با پیشرفت نرم‌افزارهای رایانه‌ای برای رده‌بندی داده‌ها به‌کار

می‌رود، درخت تصمیم[1] است که در اکثر روش‌های استخراج آن، هیچ پیش‌فرض آماری وجود ندارد.

درخت تصمیم به‌دلیل انعطاف‌پذیری و ویژگی‌هایی نظیر خروجی آن که به‌صورت گراف می‌باشد و تفسیر نتایج را ساده‌تر می‌نماید، دقت بالا و سرعت الگوریتم‌های آن مقبولیت عام یافته است [۶] و امروزه به‌طور گسترده در علوم جمعیتی، کامپیوتری، پزشکی، روانشناسی، اقتصادی و غیره به کار می‌رود [۵، ۲۲، ۳۲]. این روش‌ها برخلاف سایر روش‌های آماری که ابتدا بر حسب فرضیه‌ها بسط داده شده‌اند، به‌دلیل اندازه‌ی بزرگ داده‌ها با پیشرفت نرم‌افزارهای آماری توسعه یافته‌اند. برای استخراج درخت تصمیم الگوریتم‌های مختلفی مانند ID3[2] ، AID[3]، THAID[4]، CART[5]، CHAID[6] و QUEST[7] وجود دارند که هر یک ویژگی‌های خاص خود دارند [۳۸].

مطالعات زیادی به مقایسه‌ی روش‌های کلاسیک و درخت تصمیم پرداخته‌اند [۳۱، ۳۸، ۳۹]. مویسن و فرسینو، مدل‌های خطی، مدل‌های جمعی تعمیم یافته و CART را با یکدیگر مقایسه نمودند [۲۵]. دلن و همکاران (۲۰۰۶)، رگرسیون لوژستیکی، درخت تصمیم و شبکه‌های عصبی را در تشخیص سرطان سینه با یکدیگر مقایسه نمودند [۱۰]. استارک و پیفیفر [۳۷]، رگرسیون لوژستیکی و الگوریتم‌های CART، CHAID، ID3 و C4.5 را به منظور حل مسئله‌ی رده‌بندی همه‌گیری بیماری‌های دام با یکدیگر مقایسه کردند. کلومبت و همکاران، نحوه‌ی اجرا و عملکرد درخت تصمیم و شبکه‌های عصبی را به منظور پیش‌بینی خطر بیماری قلبی با یکدیگر مقایسه نمودند [۹]. چائی و دیگران (۲۰۰۱)، رگرسیون لوژستیکی، CHAID و C.5 را برای تشخیص فشار خون بالا به‌کار بردند [۷].

استفاده از درخت تصمیم برای رده‌بندی مشاهدات در مطالعات جمعیتی و علوم اجتماعی زمانی ضرورت می‌یابد که به سه دلیل عمده‌ی زیر، روش‌های آماری متداول نمی‌توانند در تشخیص و پیش‌بینی‌ها در این حوزه‌ها مورد استفاده قرار گیرند:

- در مطالعات اجتماعی معمولاً تعداد متغیرهای پیش‌بین زیاد است که استفاده از روش‌های متداول را با مشکلات و پیچیدگی‌هایی روبرو می‌سازد.
- اعتبار روش‌های آماری متداول به برقراری فرض‌هایی نظیر نرمال بودن توزیع داده‌ها و برابری واریانس‌ها وابسته است که دربیش‌تر داده‌های واقعی معمولاً چنین فرض‌هایی برقرار نخواهد بود.

- نتایج به‌دست آمده از روش‌های آماری متداول در حضور متغیرهای زیاد به سادگی قابل تفسیر نیستند. به‌عنوان مثال در مدل لوژستیکی تعداد اثرات متقابل در مدل با افزایش تعداد پیش‌بین‌ها افزایش می‌یابد که پژوهشگر را در تفسیر آن‌ها با مشکل مواجه می‌کند.

در حالی‌که درخت تصمیم یکی از روش‌هایی است که با وجود مشکلات ذکر شده می‌تواند به خوبی عمل نماید [۲].

یکی از سه مؤلفه‌ی اصلی تأثیرگذار بر پویایی جمعیت که اندازه، ساختار و ترکیب آن‌را تعیین می‌نماید، باروری است که افزایش یا کاهش آن به عوامل مختلفی وابسته است. یکی از این عوامل ترجیح جنسی است که می‌تواند بر رفتار باروری زوج‌ها مؤثر باشد. از این رو مطالعه‌ی اثر ترجیح جنسی بر روی باروری مورد توجه پژوهشگران و سیاستگذاران در این حوزه بوده است [۸، ۱۰، ۱۲، ۱۴، ۳۶، ۴۲]. در نتیجه مطالعه‌ی دقیق عواملی که بر روی ترجیح جنسی تأثیرگذارند نیز از اهمیت بسزایی برخوردار هستند. به این منظور هدف از مقاله‌ی حاضر پیش‌بینی ترجیح جنسی فرزندان زنان در آستانه‌ی ازدواج بر اساس الگوریتم‌های CART، CHAID و QUEST درخت رده‌بندی با استفاده از داده‌های طرح «بررسی نحوه‌ی نگرش جوانان در آستانه‌ی ازدواج و زنان همسردار ۱۵-۴۹ ساله نسبت به فرزندآوری و شناخت عوامل اجتماعی، اقتصادی و فرهنگی مؤثر بر آن» می‌باشد.

تا کنون در ایران، از الگوریتم CART به منظور رده‌بندی تعداد فرزندان زنده بدنیا آمده [۱،۳۳] و تعداد فرزندان ایده‌آل [۲] زنان حداقل یک‌بار ازدواج کرده ۱۵- ۴۹ ساله در استان سمنان در مطالعات جمعیتی، استفاده شده است. بر اساس دانش نویسندگان مقاله، مطالعات محدودی در خصوص مقایسه‌ی الگوریتم‌های مختلف درخت تصمیم در مطالعات جمعیتی انجام گرفته است و این مسئله ضرورت انجام پژوهش حاضر را مشخص می‌نماید. به این منظور در بخش ۲، هر یک از این الگوریتم‌ها معرفی خواهد شد. نتایج به‌دست آمده از استخراج درخت‌های حاصل از هر یک این الگوریتم‌ها در بخش ۳ و بحث و نتیجه‌گیری در بخش ۴ ارائه می‌شوند.

۲- روش بررسی

درخت تصمیم را می‌توان یک روش تحلیل تشخیصی ناخطی در نظر گرفت که مجموعه‌ای از متغیرهای پیش‌بین را برای تقسیم نمونه به زیرگروه‌های کوچک‌تر به کار می‌گیرد. بر حسب این‌که متغیر پاسخ در درخت تصمیم، رسته‌ای یا پیوسته باشد، به ترتیب به آن درخت رده‌بندی[۸] یا درخت رگرسیونی[۹] می‌گویند. یک درخت تصمیم از سه جزء اصلی ریشه[۱۰]، گره‌ی داخلی[۱۱] و گره‌ی خارجی[۱۲] (برگ[۱۳]) تشکیل شده است که این اجزا با توجه به الگوریتم‌های مختلفی تعیین می‌شوند.

روش‌های مبتنی بر درخت تصمیم فضای متغیرهای پیش‌بین را به‌صورت بازگشتی[۱۴] به ناحیه‌های مجزا افراز[۱۵] کرده و داده‌های متناظر را به رده‌ها تخصیص می‌دهد [۲۸]. این افراز بازگشتی منجر به برازش مدل تکه‌ای ثابت روی ناحیه‌های افراز شده‌ی فضای متغیر پیش‌بین خواهد شد [۲۰]. برای این‌که هر گره افراز شود، تمام افرازهای ممکن برای هر متغیر پیش‌بین ارزیابی می‌شوند. متغیر و نقطه‌ی افراز متناظر با آن به گونه‌ای انتخاب می‌شود که بهترین تفکیک بین دو گره حاصل شود. این روند به‌صورت بازگشتی ادامه می‌یابد تا این‌که هر گره شامل تعداد محدودی از حالت‌ها شود. بعد از ساخت یک درخت بزرگ، قواعدی برای هرس[۱۶] و تعدیل کردن اندازه‌ی درخت به‌کار می‌رود [۳، ۲۸]. ساخت درخت از گره‌ی ریشه آغاز می‌شود که شامل همه‌ی مشاهدات موجود (در نمونه‌ی 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.

الگوریتم AID توسط مورگان و سانکوئیست در سال ۱۹۶۳ ارایه شد [۲۷]. در سال ۱۹۷۳ این الگوریتم توسط مورگان و مسنجر به THAID که یک الگوریتم برای تحلیل متغیرهای وابسته با مقیاس اسمی بود، ارتقا یافت [۲۶] و در سال ۱۹۸۰ توسط کاس به الگوریتم CHAID، تبدیل شد [۱۸]. این سه روش از تقسیمات چندسطحی برای تولید درخت رده‌بندی استفاده می‌کنند. الگوریتم CART، که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌شود، توسط بریمن و همکاران در سال ۱۹۸۴ به طور کامل معرفی شد. این روش که ابتدا تنها برای متغیرهای پیش‌بین کمی طراحی شده بود، بعدها برای متغیرهای کیفی نیز تعمیم یافت [۴]. سه الگوریتم AID، THAID و CHAID برخلاف

الگوریتم کارت، آزاد- توزیع نیستند و از آزمون‌های معنی‌داری روی متغیرهای پیش‌بین برای ایجاد تقسیمات و تعیین اندازه‌ی درخت استفاده می‌کنند. این روش‌ها در فرایند رشد، هرس درخت و براورد خطای رده‌بندی با یکدیگر تفاوت دارند [۴۱].
با توجه به نوع داده‌ها و کاربرد وسیع سه الگوریتم CART، CHAID و QUEST در مطالعات مختلف، در این مقاله به مقایسه‌ی این سه الگوریتم پرداخته شده است.
در ادامه‌ی مقاله فرض می‌شود که N_j تعداد طبقات jام نمونه آموزشی و $\pi(j)$ احتمال پیشین رده‌ی jام باشد. $N_j(t)$ در هر گره‌ی t، تعداد رده‌های jام نمونه آموزشی در t و $p(j|t) = N_j(t)/N_j\pi(j)$ براورد احتمال تعلق یک مشاهده در t به رده‌ی j است.

۱-۲- الگوریتم CART

این روش که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌شود، برای هر دو نوع متغیر پاسخ و مستقل کیفی و کمی کاربرد دارد. معمولاً برای متغیر پاسخ کیفی شاخص جینی[۱۷] را به‌عنوان معیاری برای انتخاب متغیرهای مناسب در این روش که با استفاده از معادله‌ی (۱) قابل محاسبه است، به‌کار می‌برند.

$$i(t) = gini(t) = ۱ - \sum_{i=۱}^{k} p^{۲}\ [c = c_j|T = t] \qquad (۱)$$

در این مدل از شاخص دیگری با عنوان «قانون تقسیم‌بندی دوحالتی[۱۸] نیز استفاده می‌شود که برای متغیر پاسخی با k طبقه به‌صورت زیر تعریف می‌شود:

$$\Delta i(t) = \frac{p_l p_r}{۴}\left[\sum_{k=۱}^{n} |p(k|t_l) - p(k|t_r)|\right]^{۲} \qquad (۲)$$

در معادله‌ی (۲)، p_l و p_r به ترتیب احتمال‌های مربوط به گره‌ی چپ و راست می‌باشند. این شاخص، درخت رده‌بندی متقارن‌تری را ارایه می‌نماید، ولی از سرعت کم‌تری نسبت به شاخص جینی برخوردار می‌باشد. در مدل CART هرس کردن درخت رده‌بندی براساس هزینه- پیچیدگی[۱۹] انجام می‌شود و بررسی دقت درخت معرفی‌شده به کمک نمونه‌ی آزمون بررسی می‌شود. یکی از اشکالات مطرح برای مدل کارت، اریبی این مدل در انتخاب متغیرها است. به‌علاوه زمانی که متغیرهای کیفی دارای بیش از دو سطح باشند، نتایج این

مدل گمراه‌کننده خواهد بود، زیرا ممکن است چند سطح یک متغیر به یک گره تعلق گیرد که در نتیجه نمی‌توان تفسیر ساده‌ای از نتایج به‌دست آورد [۳۵].

۲-۲- الگوریتم CHAID

با این‌که الگوریتم CHAID برای متغیرهای کیفی طراحی شده است، می‌تواند برای متغیرهای کمی رده‌بندی‌شده نیز استفاده شود. در این الگوریتم امکان بیش از دو تقسیم‌بندی نیز در هر گره وجود دارد. در این روش از pمقدار آماره‌ی خی‌دوی آزمون فرض استقلال جداول تقاطعی استفاده می‌شود. از بین متغیرهای موجود متغیری که دارای pمقدار کم‌تری باشد در مرحله‌ی اول برای تقسیمات روی یک گره در نظر گرفته می‌شود [۱۵-۱۷]. به‌منظور تعیین بهترین افراز برای هر گره، هر زوج از رده‌ها متغیر پیش‌بین با یکدیگر ادغام می‌شوند تا زمانی‌که هیچ تفاوت معنی‌داری در هر زوج با توجه به متغیر پاسخ ایجاد نشود. در حقیقت با این روش اثرات متقابل متغیرهای پیش‌بین بررسی می‌شوند. ضعف الگوریتم CHAID عدم توانایی این الگوریتم در ایجاد بهینه‌ی تقسیمات ممکن بر اساس متغیرهای موجود است [۱۳].

۲-۳- الگوریتم QUEST

الگوریتم QUEST توسط لو و شی در سال ۱۹۹۷ برای متغیرهای پاسخ اسمی طراحی شد [۲۱]. درخت رده‌بندی حاصل از این الگوریتم نظیر مدل CART دارای تقسیمات دوتایی است. ملاک تصمیم‌گیری برای انتخاب متغیرها با مقایسه‌ی pمقدار مربوط به آماره‌ی F در آزمون ANOVA برای متغیرهای کمی و pمقدار آماره‌ی خی‌دوی مربوط به جداول تقاطعی برای متغیرهای کیفی انجام می‌شود. این الگوریتم با توجه به این‌که از pمقدار برای تصمیم‌گیری استفاده می‌نماید، موجب تشکیل درختی نااریب برای متغیرها می‌شود. QUEST علاوه بر این‌که از سرعت بالاتری در معرفی یک درخت رده‌بندی نسبت به CART برخوردار است، دارای دقت خوبی نیز می‌باشد [۱۳].

۴-۲- محاسبه‌ی دقت درخت تصمیم

بدیهی است که در تمام روش‌های رده‌بندی دستیابی به الگویی که بدون خطا باشد، غیر ممکن است؛ هدف در این روش‌ها معرفی الگویی با کم‌ترین خطا است. از این رو یکی از مهم‌ترین نکات مورد توجه در معرفی یک الگوی درختی، بررسی دقت و کارایی آن است؛ بدان معنی که این الگوریتم بتواند در مورد یک مشاهده‌ی جدید به خوبی و با کم‌ترین خطای ممکن، تصمیم‌گیری کند. با توجه به این‌که در نظر گرفتن میزان رده‌بندی اشتباه بر اساس نمونه‌ای که درخت رده‌بندی با توجه به آن ساخته شده است، به‌عنوان میزان خطا مناسب نیست، روش‌هایی برای برآورد میزان خطای یک درخت رده‌بندی مبتنی بر روش‌های اعتبارسنجی متقاطع[۲۰] مورد استفاده قرار می‌گیرند. اعتبارسنجی متقاطع k مرتبه‌ای برای زمانی‌که اندازه‌ی مشاهدات موجود با توجه به متغیرهای پیش‌بین کافی نباشد، مناسب است. در این روش مشاهدات موجود به‌طور تصادفی به k قسمت با اندازه‌های برابر تقسیم می‌شوند و در هر مرحله $k-1$ قسمت به‌عنوان نمونه‌ی آموزشی و قسمت دیگر به‌عنوان نمونه‌ی آزمون در نظر گرفته می‌شود. با توجه به نمونه‌ی آموزشی معرفی‌شده درخت مناسب ایجاد می‌شود و خطای مربوط به آن بر اساس نمونه‌ی آزمون برآورد می‌شود. از آن‌جا که مشاهدات به k قسمت تقسیم شده‌اند این اقدام نیز k بار انجام می‌شود و k اندازه‌ی خطا به‌دست می‌آید که میانگین این خطاها به‌عنوان برآوردی از خطای مدل درختی با توجه به کل مشاهدات به‌دست می‌آید. این روش از دقت مناسبی در برآورد خطا برخوردار است و برای نمونه‌های کوچک مناسب می‌باشد [۳۴].

در محاسبه خطا معمولاً از ماتریس اغتشاش[۲۱] استفاده می‌شود. این ماتریس که برای یک متغیر دو سطحی، الگویی مشابه جدول ۱ دارد و مقادیر واقعی رده‌ی متغیر مورد نظر را با مقادیر پیش‌بینی‌شده‌ی رده‌ی آن متغیر توسط مدل مقایسه می‌کند.

جدول ۱- ماتریس اغتشاش برای رده‌بندی یک متغیر دو سطحی

	مقدار پیش‌بینی رده	۰	۱
مقدار واقعی رده	۰	**A**	**B**
	۱	**C**	**D**

معادله‌ی (۳) نحوه‌ی محاسبه دقت را برای جدول ۱ نشان می‌دهد.

$$\text{دقت} = \frac{A+D}{A+B+C+D} \quad (۳)$$

هر چه رده‌بندی پیش‌بینی‌شده و واقعی به یکدیگر نزدیک‌تر باشند، مدل از دقت بالاتری برخوردار خواهد بود. به عبارتی دیگر در این ماتریس هر چه مقادیر روی قطر فرعی قرار گیرند، دقت مدل و اعتبار آن مدل در پیش‌بینی متغیر مورد نظر بیش‌تر می‌شود. در ادامه به مدل‌سازی ترجیح جنسی با استفاده از الگوریتم‌های معرفی‌شده در این بخش و با استفاده از نرم افزار SPSS نسخه‌ی ۲۲ پرداخته شده است. لازم به ذکر است که برای مقایسه‌ی مدل‌ها از معادله‌ی (۳) استفاده شد.

۳- یافته‌ها

در این مطالعه جامعه‌ی هدف، زنان مراجعه‌کننده به مراکز مشاوره‌ی قبل از ازدواج وزارت بهداشت، درمان و آموزش پزشکی واقع در همه‌ی استان‌های کشور بودند.

داده‌ها طی طرح پژوهشی با عنوان «بررسی نحوه‌ی نگرش جوانان در آستانه‌ی ازدواج و زنان همسردار ۴۹-۱۵ ساله نسبت به فرزندآوری و شناخت عوامل اجتماعی، اقتصادی و فرهنگی مؤثر بر آن»، به‌صورت مقطعی و بر اساس پرسشنامه‌ی ساختاریافته در سال ۱۳۹۳ گرداوری شدند [۱۹]. با توجه به اهداف مطالعه‌ی حاضر ۶۳۶۰ زن در آستانه‌ی ازدواج به روش تصادفی انتخاب و متغیرهای سن، سطح تحصیلی، وضعیت شغلی، محل سکونت، تعداد فرزندان ایده‌ال و ترجیح جنسی فرزندان در مورد آنان اندازه‌گیری شد. با توجه به نتایج به‌دست آمده بیش‌تر زنان در سن ۲۹-۲۰ سالگی (۵۸/۳ درصد) دارای تحصیلات دبیرستان و دیپلم (۳۸/۸ درصد) و فوق دیپلم/ لیسانس (۳۸/۸ درصد)، غیر شاغل (۷۶ درصد) و دارای ۱ یا ۲ فرزند (۷۷/۹ درصد) بودند.

در این مطالعه به منظور تعیین ترجیح جنسی از پرسش «خود شما دوست دارید چند فرزند داشته باشید؟» استفاده شده است. در پاسخ به این پرسش سه گزینه «پسر»، «دختر» و «فرقی نمی‌کند» وجود داشت که بر اساس آن همه‌ی زنانی که در پاسخ به این پرسش، پاسخ«فرقی نمی‌کند» را انتخاب کرده بودند به‌عنوان بدون ترجیح جنسی و در غیر این صورت با ترجیح جنسی در نظر گرفته شدند. داده‌ها نشان داد که اکثریت زنان این مطالعه ترجیح جنسی نداشتند (۵۲/۵ درصد).

جدول ۲ ماتریس اغتشاش سه الگوریتم را که از روی آن می‌توان میزان صحت پیش‌بینی درخت‌ها را با یکدیگر مقایسه نمود، نشان می‌دهد. همان‌گونه که ملاحظه می‌شود هر سه درخت با دقتی تقریباً برابر، حدود ۶۵ درصد (دقت الگوریتم‌های CART، CHAID و QUEST به ترتیب برابر با ۶۵/۶، ۶۵/۲ و ۶۴/۹ درصد) مشاهدات را به‌درستی پیش‌بینی نموده‌اند.

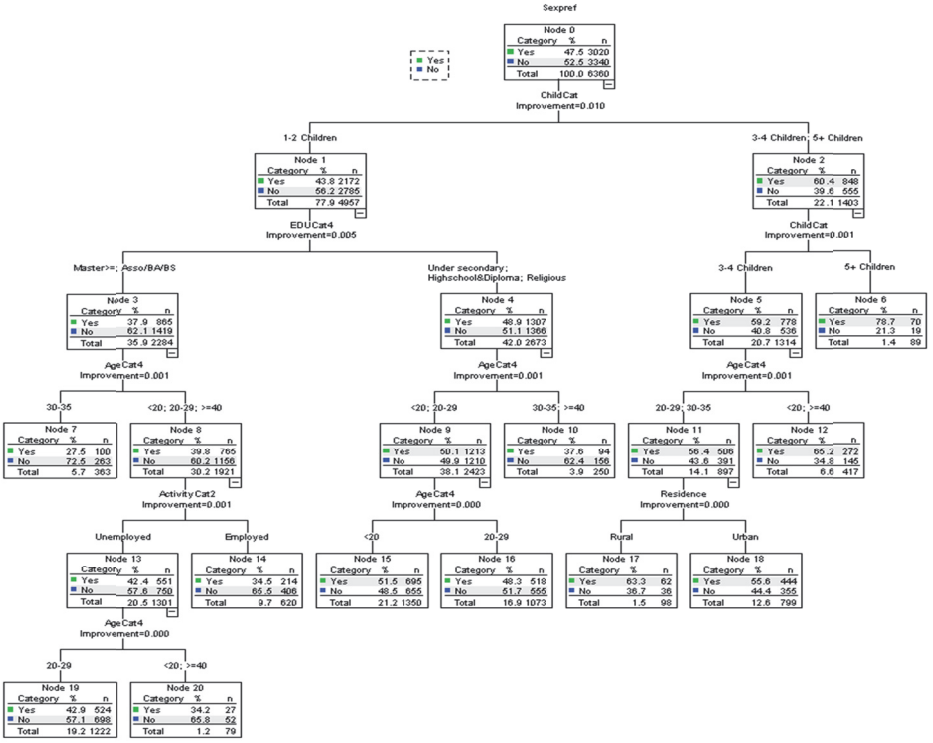
جدول ۲- ماتریس اغتشاش درخت‌های CART، CHAID و QUEST

الگوریتم	رده‌های مشاهده‌شده	رده‌های پیش‌بینی‌شده		
		ترجیح جنسی دارد	ترجیح جنسی ندارد	درصد صحیح
CART	ترجیح جنسی دارد	۲۰۴۳	۹۷۷	۴۷/۵
	ترجیح جنسی ندارد	۱۲۱۰	۲۱۳۰	۵۲/۵
	درصد کل	۵۱/۱	۴۸/۸	۶۳/۶
CHAID	ترجیح جنسی دارد	۲۰۷۴	۹۴۶	۴۵/۵
	ترجیح جنسی ندارد	۱۲۶۳	۲۰۷۷	۵۲/۵
	درصد کل	۴۴/۶	۴۷/۵	۶۵/۲
QUEST	ترجیح جنسی دارد	۲۰۳۶	۹۸۴	۴۷/۵
	ترجیح جنسی ندارد	۱۲۵۰	۲۰۹۰	۵۲/۵
	درصد کل	۵۱/۷	۴۸/۳	۶۴/۹

شکل ۱ درخت CART با ۳۰ گره را نشان می‌دهد. همان‌گونه که ملاحظه می‌شود همه‌ی متغیرهای پیش‌بین مطالعه شامل سن، سطح تحصیلی، وضعیت شغلی، محل سکونت، و تعداد فرزندان ایده‌ال در این درخت ظاهر شده‌اند.

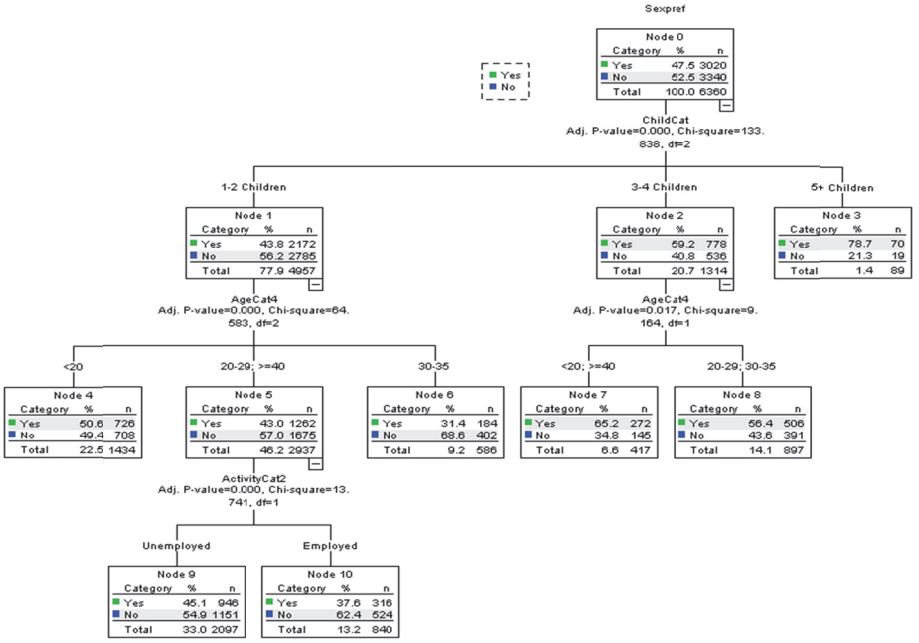
بر اساس این شکل زنان با ایده‌آل ۱ و ۲ فرزند که دارای تحصیلات دانشگاهی هستند ترجیح جنسی ندارند. اگرچه وضعیت شغلی و سن آنان در این شاخه وارد شده اما برای همه‌ی رده‌ها این متغیرها در زیرشاخه‌ی تحصیلات دانشگاهی، ترجیح جنسی وجود ندارد. از سویی دیگر زنان با ایده‌آل ۱ و ۲ فرزند که دارای تحصیلات کم‌تر از دانشگاه و تحصیلات حوزوی می‌باشند بسته به گروه سنی «کم‌تر از ۲۰ سال» یا «۲۰ ساله و بیش‌تر» به‌ترتیب ترجیح جنسی دارند یا ندارند. همه‌ی زنان با ایده‌ال ۳ فرزند و بیش‌تر دارای

ترجیح جنسی هستند و حضور متغیرهای سن و محل سکونت تأثیری بر روی ترجیح جنسی آنان ندارد.



شکل ۱- درخت CART برای رده‌بندی ترجیح جنسی

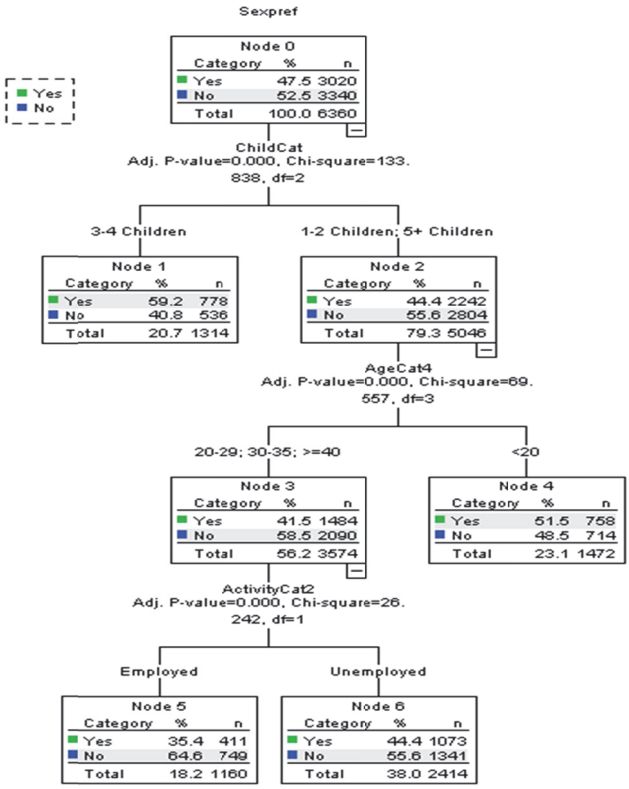
شکل ۲ درخت CHAID با ۱۸ گره را نشان می‌دهد. همان‌گونه که ملاحظه می‌شود در این درخت تنها سه متغیر سن، وضعیت شغلی و تعداد فرزندان ایده‌ال وارد مدل شده‌اند. بر اساس این شکل زنان با ایده‌آل ۱ و ۲ فرزند در گروه «زیر ۲۰ سال» دارای ترجیح جنسی و زنان «۲۰ ساله و بیش‌تر» بدون ترجیح جنسی می‌باشند. اگرچه وضعیت شغلی در زیرشاخه گروه سنی «۲۰-۲۹ و ۴۰ ساله و بیش‌تر» وارد شده، اما بر روی ترجیح جنسی آنان تأثیرگذار نیست. همه‌ی زنان با ایده‌ال ۳ فرزند و بیش‌تر بدون تأثیر سایر متغیرها، دارای ترجیح جنسی هستند.



شکل ۲- درخت CHAID برای رده‌بندی ترجیح جنسی

شکل ۳ درخت QUEST با ۱۱ گره را نشان می‌دهد. در این درخت همانند درخت CHAID تنها سه متغیر سن، وضعیت شغلی و تعداد فرزندان ایده‌ال وارد مدل شده‌اند. نتایج مستخرج از این درخت نشان می‌دهد که زنان با ایده‌آل ۳ و ۴ فرزند دارای ترجیح جنسی هستند. زنان با ایده‌آل ۱، ۲ و ۵ فرزند و بیش‌تر که «کم‌تر از ۲۰ سال» سن دارند دارای ترجیح جنسی و زنان با ایده‌آل ۱، ۲ و ۵ فرزند و بیش‌تر که «۲۰ سال و بیش‌تر» سن دارند ترجیح جنسی ندارند. با وجود حضور متغیر وضعیت شغلی برای گروه‌های سنی «۲۰ سال و بیش‌تر»، این متغیر در تشخیص ترجیح جنسی بی‌تأثیر است.

مقایسه‌ی هر ۳ درخت حاکی از آن است که متغیر تعداد فرزندان ایده‌آل به‌عنوان با اهمیت‌ترین متغیر در ریشه آن‌ها قرار گرفته است.



شکل ۳- درخت QUEST برای رده‌بندی ترجیح جنسی

۴- بحث و نتیجه‌گیری

درخت تصمیم به‌عنوان یک روش ناپارامتری، ابزاری قدرتمند در داده‌کاوی است. نتایج مطالعات شبیه‌سازی‌شده نشان می‌دهد که این روش، زمانی‌که داده‌ها کشیدگی و چولگی بالایی داشته باشند یا زمانی‌که درصد بالایی از متغیرها کیفی باشند (همانند مطالعه‌ی حاضر)، روشی مناسب برای کاوش داده‌ها است [۳۰]. مهم‌ترین مزیت این روش، قابلیت تفسیرپذیری بسیار بالای آن به واسطه‌ی ساختار درختی است. لذا زمانی‌که به جای تمرکز بر دقت بالای روش پیشگویی، به تفسیر نتایج علاقمند باشیم، یکی از بهترین گزینه‌ها خواهد بود [۱۶، ۳۳]. در این مطالعه، سه الگوریتم پرکاربرد CART، CHAID و

QUEST در ساخت درخت رده‌بندی در پیش‌بینی ترجیح جنسی با یکدیگر مقایسه شدند. بر اساس مطالعات پیشین، بهترین روش رده‌بندی از مجموعه داده‌ای به مجموعه داده‌ای دیگر متفاوت است [۴۰]. نتایج مطالعه‌ی حاضر نشان می‌دهد که هر سه الگوریتم از دقت تقریباً مشابهی در پیش‌بینی رده‌بندی ترجیح جنسی برخوردارند. به ترتیب CART، CHAID و QUEST بزرگ‌ترین تا کوچک‌ترین درخت را تولید نموده‌اند. در هر سه درخت، تعداد فرزندان ایده‌آل و سن به‌ترتیب پر اهمیت‌ترین متغیرها در پیش‌بینی بوده‌اند. این یافته را می‌توان بر اساس نتایج سایر مطالعات نیز تأیید نمود؛ فرمپونگ و کودجوئی (۲۰۱۳)، جنس، سن، محل سکونت، منطقه مسکونی، مذهب، وضعیت شغلی و سطح تحصیلی را به‌عنوان متغیرهای پیش‌بین در مطالعه‌ی ترجیح جنسی والدین غنایی با رگرسیون لوژستیکی در نظر گرفتند [۱۱]، که در آن بر نقش متغیر سن بر ترجیح جنسی والدین تأکید نمودند. در نتایج یافته‌های منصوریان و خوشنویس (۱۳۸۵) نیز به اهمیت نقش سن و تعداد فرزندان ایده‌آل زنان تهرانی بر ترجیح جنسی و رفتار باروری آنان اشاره شده است [۲۴].

در الگوریتم‌های CART و CHAID، زنان با تعداد فرزندان ایده‌آل بالاتر ترجیح جنسی داشتند. در CART، متغیر تحصیلات زنان نیز بر روی ترجیح جنسی آنان بسیار تأثیرگذار است، این در حالی است که در دو درخت دیگر این متغیر ظاهر نشده است. با توجه به یافته‌های ونگ بونیسین و رفالو [۴۰]، نیز متغیرهای تأثیرگذار بر ترجیح جنسی شامل متغیرهای خرد (شامل مشخصه‌های فردی والدین) و کلان (شامل سیاستگذاری‌های جمعیتی، مدرنه شدن، فرهنگ و مسائل اجتماعی- اقتصادی و سیاسی) هستند که در بین متغیرهای خرد، سطح تحصیلی متغیر بسیار تأثیرگذاری به دست آمد. این یافته را سایر پژوهشگران نیز تأیید می‌نمایند [۱۱، ۱۲، ۲۴، ۳۶]. بنا بر این با توجه به این‌که در مطالعات پیشین نقش تحصیلات به‌عنوان یک عامل مهم در تعیین ترجیح جنسی به اثبات رسیده است علی‌رغم پیچیدگی درخت CART نسبت به دو درخت دیگر، از آن برای پیش‌بینی ترجیح جنسی استفاده شد.

سپاس‌گزاری

این مقاله مستخرج از طرح کاوش داده‌های جمعیتی با استفاده از درخت تصمیم (ابلاغ شماره۲۰/۱۵۲۸۳ مورخ ۱۳۹۳/۱۱/۵ است که با حمایت مالی مؤسسه‌ی مطالعات و مدیریت جامع و تخصصی جمعیت کشور در سال ۱۳۹۳ انجام شده است). نویسندگان مقاله برخود لازم می‌دانند که از سرکار خانم دکتر شهلا کاظمی‌پور برای در اختیار قراردادن اطلاعات طرح بررسی نحوه‌ی نگرش جوانان در آستانه‌ی ازدواج و زنان همسردار ۱۵-۴۹ ساله نسبت به فرزندآوری و شناخت عوامل اجتماعی، اقتصادی و فرهنگی مؤثر بر آن که با حمایت پژوهشکده‌ی آمار انجام شده است، کمال تشکر و قدردانی را داشته باشند.

توضیحات

1. Decision Tree
2. Interaction Detector
3. Automatic Interaction Detector
4. Theta Automatic Interaction Detector
5. Classification and Regression Tree
6. Chi-Square Automatic Interaction Detector
7. Quick Unbiased Efficient Statistical Tree
8. Classification Tree
9. Regression Tree
10. Root
11. Internal Node
12. External Node
13. Leaf
14. Recursive
15. Split
16. Pruning
17. Gini Index

18. Towing Splitting Rule
19. Cost-complexity
20. Cross Validation
21. Confusion Matrix

مرجع‌ها

[1] Bagheri, A. and Saadati, M. (1394). Classification of Children ever Born by CART Model. *Hakim Seyed Esmail Jorjani Journal*. **3**, DOI: 10.17485/ijst/2015/v8i30/90251.

[2] Baheri, A., Saadati, M. and Razeghi Nasrabad, H.B.B. (2014). Introduction and Application of CART Model for Classifying 15-49 Year Old Women Ideal Number of Children in Semnan Province, *Journal of Population Association of Iran,* **9**, 77-111.

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

[4] Bhuyar, V. (2014). Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for District INDIA, *International Journal of Emerging Trends & Technology in Computer Science*, **3**, 200-203.

[5] Breiman, L., Friedman, J., Olsen, R. and Stone, C. (1984). *Classification and Regression Trees*. London, Chapman & Hall.

[6] Bringman, B.Z. (2005). *Tree Decision Tree for Tree Structured Data*. New York, Springer, 46-58.

[7] Chae, Y.M., Ho, S.H., Cho, K.W., Lee, D.H. and Ji, S.H. (2001). Data Mining Approach to Policy Analysis in a Health Insurance *Domain. International Journal of Medical Informatics*. **62**, 103–111.

[8] Chung, W. and Gupta, M.D. (2007). The Decline of Son Preference in South Korea: The Roles of Development and Public Policy. *Population and Development Review*, **33,** 757-783.

[9] Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P. and Jaulent, M.C. (2000). Models to Predict Crdiovascular Risk: Comparison of CART, Multilayer Perceptron and Logistic Regression. Proceedings AMIA Symposium, 156–160.

[10] Delen, D., Walker, G. and Kadam, A. (2005). Predicting Breast Cancer Survivability: a Comparison of Three Data Mining Methods. *Artificial Intelligence in Medicine*, **34**, 113–27.

[11] Frempong, G.A. and Codjoe, S.N.A. (2013). Education and Sex Preferences of Children in Ghana, Regional Institute for Population Studies, University of Ghana. P.O. Box LG 96. Legon, in XXVII IUSSP International Population Conference, Session 171: Poster Session on Fertility. Busan, Korea, 26–31.

[12] Fuse, K. (2010). Variations in Attitudinal Gender Preferences for Children across 50 Less-developed Countries. *Demographic Research*, **23**, 1031–1048.

[13] Gilbert, R. (2010). CHAID and Earlier Supervised Tree Methods, http://www.unige.ch/ses/metri/.2010.

[14] Hasan, M. and Sabiruzzaman, (2008). Factors Affecting Fertility Behavior in Bangladesh: A Probabilistic *Approach. Research Journal of Applied Sciences*, **3**, 70–76.

[15] Hoare, R. (2004). Using CHAID for Classification Problem, New Zealand Statistical Association Conference, Wellington.

[16] Hosseini, M., Tazhibi, M., Amini, M., Zareei, A.S. and Jahani-Hashemi, H. (2010). Using Classification Tree for Prediction of Diabetic Retinopathy on Type ii Diabetes. *Journal of Isfahan Medical School*, **28**, 15–24.

[17] Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics*, **15**, 651–674.

[18] Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, **29**, 119–127.

[19] Kazemipour, S.H. (2014). *Study of Pre-marriage Youths and 15-49 Year-old Married Women Attitudes Toward Childbearing and Its Influential Social, Economical and Cultural Factors*, Statistical Research Center, Tehran, Iran.

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

[21] Loh, W.Y. and Shih Y.S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*, **7**, 815-840.

[22] Loris, N., Alessandra, L. and Claudio, M. (2011). A Data Mining Approach for Predicting the Pregnancy Rate in Human Assisted Reproduction, **326**, 97-111.

[23] Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econom et rics*, Cambridge University Press, Cambridge.

[24] Mansurian, M.K. and Khushnevis, A. (2005). Sex Preference and Married Women Attitudes to Fertility Behavior, Tehran case study, *Shiraz University Human and Social Science Journal*, **24**, 129-146.

[25] Moisen, G.G. and Frescino, T.S. (2002). Comparing Five Modeling Techniques for Predicting Forests Characteristics. *Ecological Modelling*, **157**, 209–225.

[26] Morgan, J.N. and Messenger, R.C. (1973). 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.

[27] Morgan, J.N. and Sonquist, J.A. (1963). Problems in the Analsis of Survey Data, and a Proposal. *Journal of American Statistical Association*, **58**, 415–434 .

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

[29] Nwakeze, N.M. (2007). The Demand for Children in Anambra State of Nigeria: A Logit Analysis. *African Population Studies*, **22**, 175-201.

[30] Paul, R. and Harper. A. (2005). A Review and Comparison of Classification Algorithms for Medical Decision Making. *Health Policy*, **71**, 315-331.

[31] Rahman, M., Ahmad, T. and Hoque, A. (2008). Factors Affecting Children ever Born in Slum Areas of Rajshahi City Corporation, Bangladesh. *Middle East Journal of Nursing*, **2**, 5-10.

[32] Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F. and Hadaegh, F. (2014). Applying Decision Tree for Identification of a Low Risk Population for Type 2 Diabetes. Tehran Lipid and Glucose Study. *Diabetes Res. Clin. Pract*, **105**, 391-8.

[33] Saadati, M. and Bagheri, A. (2014). *Demographical Data Mining by Decision Tree*. National Population Studies & Comprehensive Management Institute, Tehran, Iran.

[34] Schaffer, C. (1993). Selecting a Classification Method by Cross-Validation. *Machine Learning*, **13**, 135-143.

[35] Scoh, C., Willett, R. and Nowak, R (2003). CART: Classification or Regression Trees, IEEE International Conference.

[36] Shahbaziyan, S., Gholami, A. and Shahbazi, S. (2015). The Role of Gender Preference in Reproductive Behavior of Women in the City of Kangavar. *Ilam University Medical Science Journal*, **22**, 133-142.

[37] Stark, K.D.C. and Pfeiffer, D.U. (1999). The Application of Non-parametric Techniques to Solve Classification Problems in Complex Data Sets in Veterinary Epidemiology-an Example. *Intelligent Data Analysis*, **3**, 23–35.

[38] Timofeev, R. (2004). Classification and Regression Trees (CART) Theory and Applications, Doctoral Dissertation, Humboldt University, Berlin.

[39] Ture, M., Kurt, I., Kurum, A.T. and Ozdamar, K. (2005). Comparing Classification Techniques for Predicting Essential Hypertension. *Expert Systems with Applications*, **29**, 583-588.

[40] Wongboonsin, K. and Ruffolo, V.P. (1995). Sex Preference for Children in Thailand and Some Other South-East Asian Countries. *Asia-Pacific Population Journal*, **10**, 43-62.

[41] Yohannes, Y. (2003). *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. Washington, D.C., USA.

[42] Zhu, W.X., Lu, L. and Hesketh, T. (2009). China's Excess Males, Sex Selective Abortion, and one Child Policy: Analysis of Data from 2005 National Intercensus Survey. BMJ, 338, b1211.

مهسا سعادتی
دکترای آمار زیستی
تهران، خیابان شهید بهشتی، خیابان پاکستان، خیابان دوم، شماره‌ی ۵.
رایانشانی: mahsa.saadati@psri.ac.ir

آرزو باقری
دکترای آمار کاربردی
تهران، خیابان شهید بهشتی، خیابان پاکستان، خیابان دوم، شماره‌ی ۵.
رایانشانی: abagheri_000@yahoo.com