بررسی عملکرد و کاربرد روش تحلیل درختی در تحقیقات اجتماعی

نعیما محمدی*

تاریخ دریافت: ۹٤/۷/۲
تاریخ پذیرش: ۹۵/۳/۲

چکیده

این مقاله با توجه به اهمیت رویکردهای مقایسه‌ای در تحقیقات اجتماعی به معرفی روش تحلیل درختی پرداخته که در حوزه علوم اجتماعی کمتر شناخته شده است. روش تحلیل درختی یکی از جدیدترین و انعطاف پذیرترین روش‌ها برای تحلیل مجموعه‌های بزرگ، آنالیز اکتشافی، ایجاد طبقه‌بندی ساده و همچنین تفسیر داده‌ها از طریق تقسیم بندی دوتایی است. کارت، یک استراتژی در تحلیل داده‌ها با استفاده رده بندی و رگرسیون درختی ارائه داده است که محققان اجتماعی را برای تعیین زیرگروه‌های همگن و ناهمگن، با ریسک بالا یا پایین با استفاده از آزمون‌های ناپارامتریک یاری می‌رساند. کشف اختلاف و تشابه گروه‌های اجتماعی و نهایتاً پیش بینی وضعیت آن‌ها

* استادیار جامعه‌شناسی و عضو هیئت علمی دانشکده ادبیات و علوم انسانی دانشگاه شهیدباهنر کرمان، ایران.
Naima.mohammadi@uk.ac.ir

در برنامه‌ریزی‌های اجتماعی مهم‌ترین کاربرد این روش تحلیل است. استفاده از روش تحلیل درختی در تحقیقات اجتماعی بر خلاف روش‌های کلاسیک نیازمند فرض‌های کمتر و شامل طیف وسیعی از داده‌هاست. به همین دلیل در دهه‌های اخیر این روش‌ها در برابر مدل‌های کلاسیک تحلیل ممیزی خطی و رگرسیون خطی، مقبولیت عام یافته است. بخصوص که این مدل‌ها برای حجم بالای داده‌ها برازش می‌شود. همچنین مشکل ناشی از داده‌های گم شده در این مدل وجود ندارد. به منظور روشن تر شدن بحث در این مقاله به بررسی مطالبات اجتماعی زنان در چهار کشور عضو کنفرانس اسلامی (ترکیه، مصر، عربستان سعودی و مالزی) که هر یک آن‌ها نمونه ایده آل چهار زیرگروه هستند از طریق این نرم افزار گزارش شده است.

واژه‌های کلیدی: *تحلیل درختی، نرم افزار کارت، روش تحقیق تطبیقی و تحقیقات اجتماعی.*

مقدمه

مقایسه، از قدیمی‌ترین روش‌های تحقیق در علوم انسانی است که همزمان با روشمند شدن تحقیقات علوم اجتماعی توسعه یافته است. انواع مختلفی از روش‌های مقایسه‌ای، اعم از تحلیل کیفی (QCA) چارلز رگین و تحلیل مقایسه‌ای چند ارزشی کیفی کرانکویست (MQCA) مطرح است که همچنان در حال گسترش و تکمیل هستند (کوثری، ۱۳۸٦: ۳۰). ابتدا، روش رگین توسط کرانکویست محقق آلمانی چنان گسترش یافت که بتواند با تعداد زیادی از متغیرها هم به کار گرفته شود. این روش به تحلیل مقایسه‌ای کیفی چند متغیره مشهور است. سپس این روش شناسی توسط خود راگین (۲۰۰٦) با استفاده از منطق فازی گسترش یافت، چنان که به تحلیل مقایس‌های کیفی فازی منجر شد. این دو طرح در اساس همان روش‌های اصلی مقایسه ای میل به نام طرح توافق و تفاوت هستند.

روش تحلیل درختی[1] که در علوم اجتماعی کمتر معرفی شده، یکی دیگر از تکنیک‌های پیشرفته آماری است. اصل مدل‌های درختی[2] که در رده‌بندی و رگرسیون درختی استفاده می‌شوند در دهه ۱۹۶۰ برای بررسی اثر متقابل متغیرها پیشنهاد شد. مدل‌های رگرسیون درختی نیز در سال ۱۹۶۳ توسط مورگان[3] فراهم شد. در سال ۱۹۷۳ تاید برای ارائه دادن رده‌بندی درختی در دانشگاه میشیگان تحقیقاتش را توسعه داد (Banerjee 2000: 408). جنبه‌های نظری و کاربردی این مدل در آمار توسط بریمن و همکارانش در سال ۱۹۸٤ بحث شد. او برای برازش مدل‌های درختی برنامه‌ای با عنوان نرم افزار «کارت[4]» را ابداع کرد (Segal, 1988: 40).

اگر چه تحلیل درختی ابزاری مناسب برای تبیین مقایسه‌ای پدیده‌های اجتماعی (هم با رویکرد کیفی و هم رویکرد کمی) به شمار می‌رود، اما در این جا موضوع بحث تمرکز بر الگوریتم نرم افزار کارت است که در روش‌های کمی مورد استفاده قرار می‌گیرد. شرح کاربرد آن در تحقیقات کیفی، سازوکار خاص خود را می‌طلبد که در این مقاله مجال پرداختن به آن نیست. استفاده از روش درختی برای انجام تحلیل‌های مقایسه‌ای در علوم اجتماعی حتی از برخی رشته‌های علوم انسانی نظیر روان‌شناسی بالینی، مدیریت و ... کمتر است. با توجه به این موضوع، ضرورت دارد پس از معرفی عملکرد، کاربرد و الگوریتم نرم افزار کارت در تحقیقات کمی، با ذکر یک مثالِ اجتماعی به بررسی تطبیقی مطالبات زنان در چهار کشور عضو کنفرانس اسلامی شامل ترکیه، مصر، عربستان و مالزی اشاره شود که در تحلیل نهایی خود از این روش تحلیل استفاده کرده است.

---

1. decision tree
2. Tree Based Model
3. Morgan
4. Classification And Regression Tree

۱- مبانی کلی تحلیل درختی

۱-۱- عملکرد مدل درختی

انجام یک تحلیل رگرسیونی نیاز به بررسی پیش‌فرض‌هایی دارد که در صورت برقرار بودن این پیش‌فرض‌ها می‌توان یک تحلیل رگرسیونی اجرا کرد (Kantardzic, 2003: 78). امتیاز رگرسیون درختی در این است که در آن هیچ نوع پیش‌فرض آماری وجود ندارد (liu, 2004: 230). استفاده از این روش به دلیل قابلیت انعطاف‌پذیری و ویژگی‌های خاص به‌ویژه شکل گرافیکی آن موجب تفسیر ساده‌تر روابط میان متغیرها می‌شود (Hothorn, 2008: 658). از این جهت نسبت به روش‌های آماری قدیمی‌تر نظیر تحلیل تشخیصی، تحلیل خوشه‌ای، تاکسونومی و... دارای امتیاز است.

در سال‌های اخیر محققان با حجم بزرگی از داده‌ها سر و کار دارند که اغلب این داده‌ها در قالب منحنی نرمال نیستند. این امر موجب توسعه روش‌های ناپارامتری شده است (Lamborn, 2004: 231). با توجه به این‌که مدل‌های درختی یکی از روش‌های ناپارامتری هستند، برخلاف روش‌های آماری دیگر که تنها در چارچوب تئوری حجم بزرگی از داده‌ها را تحلیل می‌کنند، قابلیت انعطاف بیشتری دارند. اضافه بر آن‌چه گفته شد شکل گرافیکی نرم افزار کارت ابزار مهمی در داده‌کاوی و تفسیر ساده‌تر داده‌ها در اختیار تحلیل‌گر قرار می‌دهد که بر مقبولیت آن می‌افزاید (Lemon, 2003: 178).

یکی از ویژگی‌های تحلیل کارت تقسیم دوتایی آن است. در تقسیم دوتایی مجموعه داده‌هایی که در یک گره از درخت قرار دارند، تنها قابل تقسیم[1] به دو قسمت مجزا هستند از این رو آن را تابع منطق ۰ و ۱ می‌دانند (Gimotty, 2007: 1131). در واقع این تشخیص خود سیستم است که بهترین نقطه برای تفکیک بین دو گروه

1. Split

مشخص شده کدام نقطه است (Takashi, 2006: 750). نتیجه اصلی مدل‌های درختی رسیدن به یک تقسیم‌بندی نهایی از افراد بر اساس متغیرهای طبقه‌بندی شده، مهم و تأثیرگذار است. این مدل همچنین قادر به تحلیل و رده‌بندی داده‌ها و زیرگروه‌های همگن است و نقاط افتراق داده‌ها را نیز مشخص می‌کند (Breiman, 2003: 29). این تحلیل‌ها به دلیل منطق اصلی داده‌ها به‌سادگی تفسیر می‌شوند. دقیقاً به همین دلیل تحلیل‌های درختی طرفداران زیادی در سال‌های اخیر پیدا کردند.

ساختار درختی برای پیش‌بینی مقدار متغیر وابسته (پیوسته یا طبقه‌بندی شده) به‌کار می‌رود این مدل‌ها به‌عنوان افزار کننده‌های بازگشتی تعریف می‌شوند، زیرا یک مجموعه n واحدی را به‌صورت تصاعدی به بخش‌های کوچک‌تر تقسیم می‌کند و هدف از این کار تقسیم‌بندی بیشترین تجانس در متغیر وابسته در هر زیرگروه است (Breiman, 2000: 76). معنای واژه بازگشتی، تکرار تقسیم تا رسیدن به بهترین نتیجه و تفکیک مجموعه داده‌ها به مجموعه‌های کوچک‌تر و کوچک‌تر است. به همین ترتیب هر گره به دو گره دیگر تقسیم می‌شود. این روش تحلیل نسبت به مدل لجستیک چندگانه (رگرسیون خطی) دارای تفسیر ساده‌تری است و منطق اصلی چیدمان مجموعه‌ها به سادگی در درخت مشهود است (Chipman, 2000: 19). به همین دلایل در تحقیقاتی که هدف طبقه‌بندی مجموعه‌ها یا افراد بر حسب گونه و نوع آنان است برای تصمیم‌گیری از تحلیل درختی استفاده می‌شود.

۱-۲- انواع مدل‌های درختی

این مدل‌ها بر حسب متغیر پاسخِ مورد بررسی، به سه دسته تقسیم می‌شوند:

۱) مدل‌های رده بندی درختی (اگر متغیر پاسخ، رده بندی شده باشد)؛

۲) مدل‌های رگرسیون درختی (اگر متغیر پاسخ، پیوسته باشد)؛

۳) مدل‌های بقای درختی (اگر متغیر پاسخ، زمان بقا باشد)؛

۱-۳- اجرای ساختار درختی[1]

اجرای تحلیلی درختی در قالب نرم افزار کارت در موارد مشابه شامل چهار مرحله اصلی است. مرحله اول، شامل بنا کردن ساختمان درخت است. ساخت درخت با استفاده از تقسیم‌های بازگشتی صورت می‌گیرد. به عبارت دیگر این مرحله تنها شامل تقسیم مجموعه داده‌ها به بخش‌های کوچک‌تر است و مهم‌ترین جزء متمایز یک درخت، چگونگی انتخاب یک ضابطه برای تقسیم‌بندی مجموعه داده‌ها در هر گره درخت است (Cappelli, 2006: 130).

انتخاب یک ضابطه تقسیم به معنای انتخاب یک متغیر پیش‌بین از میان متغیرها و انتخاب بهترین سطح تقسیم‌بندی است. مرحله دوم، شامل توقف تقسیم یا رشد درخت است (Hothor, 2008: 670). در این مرحله بزرگ‌ترین درخت ساخته می‌شود که ممکن است بیش از اطلاعات مجموعه داده‌های اولیه، داده‌ها را برازش[2] داده باشد (Schittgen, 1999: 954). زیرا تا زمانی‌که گره‌های پایانی کاملاً متجانس نشده باشند، فرآیند تقسیم ادامه می‌یابد. مرحله سوم، شامل هرس کردن درخت است که در نتیجه آن درخت‌های متوالی ساده‌تر به‌دست می‌آیند (Green, 2003). این مرحله به‌واسطه قطع کردن افزایشی گره‌های مهم صورت می‌گیرد. در واقع این فرآیند راه حلی برای رفع کاستی‌های ضابطه‌های توقف تقسیم است تا سهم مهم‌ترین مجموعه متغیرها در ایجاد معلول مشخص شود (Huang, 1999: 110). مرحله چهارم انتخاب درخت بهینه است. در این مرحله با توجه به ضوابط و معیارهای خاص، درختی از میان درخت‌های متوالی که بواسطه فرآیند هرس درصد خطای کمتری دارد، انتخاب می‌شود (Dean, 2007: 54).

---

1. Step in CART
2. Over fetes

۱-٤- ساخت مدل درختی[1]

ساخت درخت از گره ریشه‌ای شروع می‌شود که شامل تمام داده‌های طبقه‌بندی شده است. برای یافتن بهترین متغیر تمام متغیرهای ممکن و مقادیر آن‌ها را باید برای تقسیم چک کرد. در این مورد نرم افزارهایی وجود دارند که زمان جستجو را کاهش داده و به دنبال بهترین متغیر مورد نظر هستند (Oberc, 1993: 34). در حالتی که متغیر وابسته طبقه‌بندی شده باشد، تعداد تقسیم‌های ممکن با افزایش تعداد سطوح متغیر به سرعت بالا می‌روند. در انتخاب بهترین متغیر تقسیم‌کننده، برنامه به دنبال بیشترین میانگین در زیر گره‌های تولید شده می‌رود. تعدادی از معیارهای مختلف بر اساس چارچوب نظری می‌توانند انتخاب شوند که ضابطه تقسیم نامیده می‌شوند. رایج‌ترین نوع تقسیم تابع «جینی» است که به دنبال تابع «دوتایی» ایجاد شده است (Green, 2003: 230). اگر متغیر به صورت دوتایی باشد این دو تابع نتایج یکسانی ارائه می‌دهند. همان‌طور که در قسمت بعد ذکر خواهد شد هر گره تعیین‌کننده یک طبقه برآمد پیش‌بینی شده است (Wang, 2002: 65). فرآیند تقسیم گره که تخصیص طبقه‌های پیش‌بینی شده است، در هر زیر گره تکرار می‌شود و به‌طور بازگشتی ادامه می‌یابد تا وقتی که ادامه آن غیر ممکن باشد (Callaghan, 2008: 98).

۱-۵- اجزای رشد درخت

بطور کلی می‌توان گفت که چهار جزء برای فرایند اولیه رشد درخت لازم است که عبارت‌اند از:

۱) مجموعه ای از سؤالات دوتایی؛

۲) نیکویی ضابطه افراز که برای هر افراز s از گره t قابل ارزیابی است؛

۳) ضابطه توقف افراز و انتخاب درخت بهینه؛

---

1. Tree Building

٤) قاعده ای برای تخصیص هر گره پایانی به یک کلاس (رده بندی درختی).

۱-٦- افراز بندی و قاعده توقف افرازها

نیکویی افراز از تابع ناخالصی مشتق گرفته می‌شود. منظور از ناخالصی در این قسمت بیان نامتجانس بودن داده‌ها در یک گره است.

تابع ناخالصی، تابعی است که روی یک مجموعه از اعداد J تایی تعریف می‌شود.

$\sum_j p_j = 1$ و j=1,2,.....J.

افراز باید تا جایی ادامه یابد که تمام گره‌های پایانی کوچک شده و امکان هیچ افراز دیگری وجود نداشته باشد، به عبارت دیگر بزرگ‌ترین درخت ممکن ساخته شود، سپس درخت حاصل به سمت گره ریشه ای هرس خواهد شد تا به درخت‌های کوچک‌تری برسد. درخت‌های کوچک‌تر ناشی از هرس بزرگ‌ترین درخت «درخت بیشینه[1]» با استفاده از براوردهای حاصل از تکنیک‌های اعتبار سنجی مدل[2] و برآورد نمونه مستقل[3] به دست می‌آیند. به این ترتیب بهترین زیر درخت تولید شده با کمترین برآورد رده بندی اشتباه انتخاب خواهد شد.

۱-۷- ضابطه جینی

ضابطه‌های گوناگونی را می‌توان برای انتخاب بهترین افراز روی هر گره به کار برد. در ساختار ضابطه‌های افراز متغیر هزینه رده بندی اشتباه و توزیع پیشین نیز نقش دارند. از طرف دیگر مفاهیم یک ضابطه به اندازه ناخالصی یک گره نیز بستگی دارد. در هر حال نتایج به دست آمده از ضابطه‌های گوناگون منجر به انتخاب درختی می‌شود که

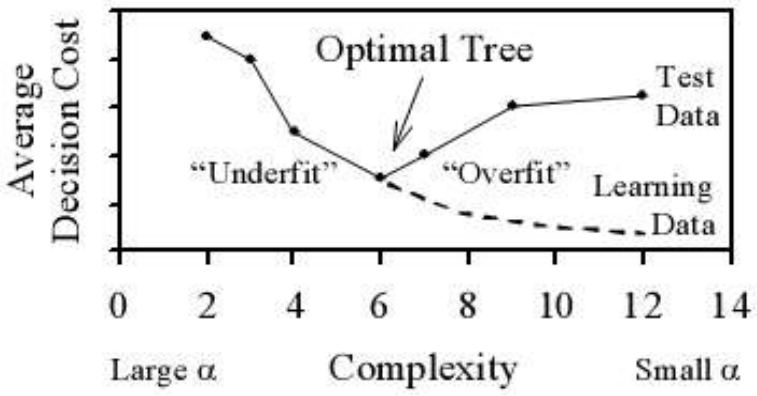
1. Maximum tree
2. Cross Validation
3. Test sample estimate

نسبت به افراز غیرحساس خواهد بود، بنابراین ضابطه‌های مورد استفاده برای هرس و بازترکیب درخت به سمت گره ریشه‌ای مهم‌تر خواهد بود.

۱-۸- توقف ساخت درخت[1]

فرآیند ساختن درخت تا جایی ادامه دارد که تنها یک مشاهده در هر گره وجود داشته باشد یا تمام مشاهدات درونِ گره توزیع یکسانی داشته باشند و یا ضابطه تقسیم به‌وسیله حدود مشخص شده توسط تحلیل‌گر متوقف شود. گاهی این حدود مشخص شده به اندازه کافی کوچک نیست، این امر موجب می‌شود داده‌ها به خوبی تقسیم نشوند و درخت تحلیل نتواند اطلاعات موجود را به خوبی برازش دهد. به‌طور معمول با رسیدن به یک درخت بزرگ، فرآیند تقسیم متوقف می‌شود در این حالت تحلیل درختی قادر به طبقه‌بندی اطلاعات بیشتری است که عموماً به آن بیش برازنده شده[2] می‌گویند (Hess, 1999: 3409). به عبارت دیگر این درخت همه حالات خاص در مجموعه داده‌ها را دنبال می‌کند که رخ دادن بعضی از آن‌ها بسیار غیر محتمل است. تقسیم‌های آخر این درخت نسبت به تقسیم‌های اولیه به احتمال زیاد نشان دهنده بیش برازندگی هستند (Molinaro ،۲۰۰٤: ١٦٣).

1. Stopping tree building
2. Overfit



به منظور انتخاب درخت بهینه[1] باید به اهداف محقق توجه کرد. خوشبختانه این نیاز با استفاده از تکنیک اعتبارسنجی مدل[2] انجام می‌شود. شکل زیر رابطه بین پارامتر پیچیدگی درخت و تعداد گره‌های پایانی و تصمیم برای انتخاب درخت بهینه را نشان می‌دهد (Garzotto, 2005: 4325).

این نمودار نشان دهنده این است که با افزایش تعداد گره‌های پایانی، هزینه تصمیم بطور یکنواخت برای داده‌های آموزشی کاهش خواهد یافت. منظور از هزینه تصمیم در اینجا، هزینه رده بندی اشتباه درخت است. بهترین زیر درخت با توجه به فرمول زیر به دست آید:

$$\hat{R}(T_{k_0}) = min_k \hat{R}(T_k)$$

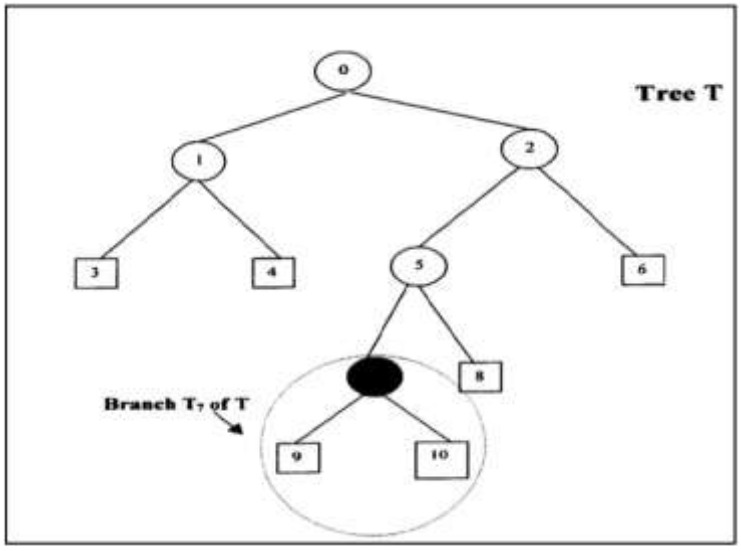
بنابراین موضوع مورد بحث در انتخاب درخت بهینه، ساختن برآوردی نااریب از هزینه رده بندی اشتباه است. دو روش برای این کار وجود دارد: ۱) اعتبار سنجی مدل؛ ۲) برآورد نمونه مستقل.

۱-۹- فرایند هرس کردن درخت

1. Optimal Tree Selection
2. Cross validation

درخت بهینه، درختی است که از لحاظ اندازه و برآورد رده بندی اشتباه بهینه باشد. بنابراین ابتدا باید برآورد رده بندی اشتباه بررسی شود، سپس با استفاده از تکنیک‌های مرسوم درخت را هرس کرد. برای هرس کردن ابتدا باید درخت را تا حد ممکن بزرگ کرد، تا درخت بیشه T_{MAX} ساخته شود. برای دستیابی به این درخت باید فرایند افراز را تا جایی ادامه داد که تمام گره‌های پایانی کوچک یا خالص و یا تنها شامل بردارهای با اندازه یکسان باشند.

نمودار ۱- نمودار درختی همراه با شاخه گره T_{max}



روش توافقی پذیرفته شده برای رشد یک درخت اولیه که بطور مناسبی بزرگ باشد، T_{max} تعیین یک عدد N_{min} و ادامه افراز تا وقتی که هر گره پایانی خالص شود یا $N_{(t)} < N_{min}$ و یا شامل فقط بردارهای اندازه یکسان باشند. بطور کلی $N_{min}=5$ تعیین می‌گردد. فرایند هرس با درخت بزرگ T_{max} شروع می‌شود. این فرایند با تولید یک

توالی از زیر درخت‌های هرس شده از Tmax انجام می‌شود. که این توالی قطع یا قطع زیرشاخه‌های درخت T_{max} تا رسیدن به گره ریشه‌ای اجرا می‌شود. برای آشنایی با فرایند هرس کردن درخت درک تعاریف زیر ضروری است.

۱- یک شاخه T_t از درخت T با گره ریشه ای t شامل گره t و تمام زیر گره‌های آن است.

۲- هرس یک شاخه T_t از یک درخت T شامل حذف تمام زیرگروه‌های گره t است که جدا کردن ۰ قطع کردن) تمام Tt به جز گره ریشه ای است. درختی که به این روش هرس شود با T_t-T نشان داده می‌شود.

اگر ً T با هرس متوالی شاخه‌ها از T بدست آید، آنگاه ً T زیر درخت هرس شده از T نامیده می‌شود و ً T<T است. گره ریشه ای هر دو یکی است. بدون توجه به این که T_{max} چگونه و با ضابطه‌ای افرازی ساخته شود، فرایند هرس با درخت T_{max} شروع می‌شود. سپس این درخت به تدریج به سمت بالا هرس می‌شود و در هر مرحله هرس (Rt) کوچک خواهد شد. اگر فرض شود که درخت T_{MAX} دارای L گره پایانی باشد، آن گاه یک توالی از درخت‌های کوچک و کوچک‌تر بصورت زیر ساخته می‌شود:

$[T_{max}, T_1, T_2, \ldots t_1]$

بطوری که برای هر مقدار H که ۱< L > H حالت زیر وجود خواهد داشت. برای اجرای فرایند هرس یک درخت از روش کمینه هزینه- پیچیدگی استفاده می‌شود. ایده اصلی این تکنیک عبارت است از:

۱- برای هر زیر درخت $T<T_{min}$ پیچیدگی به عنوان T که تعداد گره‌های پایانی در درخت T است، تعریف می‌شود. A≥O عدد حقیقی است که پارامتر پیچیدگی بصورت زیر خواهد بود:

$$R_{\propto}(T) = R(T) + \propto |\tilde{T}|$$

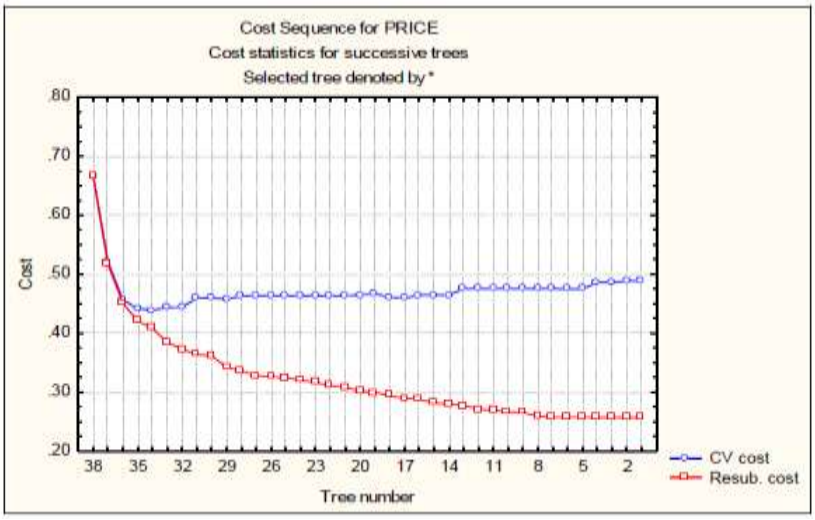
بنابراین $R_{\propto}(T)$ یک ترکیب خطی از هزینه و پیچیدگی درخت است. اگر $\propto$ به عنوان هزینه پیچیدگی درخت در نظر گرفته شود. (RT) با اضافه کردن یک هزینه

جبرانی برای پیچیدگی به هزینه رده بندی اشتباه، تشکیل می‌شود. اکنون برای هر مقدار، یک زیر درخت $T(a)< T_{max}$ را جستجو کرده که مقدار $R_{\propto}(T)$ را مینیموم کند. یعنی:

$$R_{\propto}(T(\alpha)) = minR_{\propto}(T)$$

۱-۱۰- انتخاب درخت بهینه[1]

هدف در انتخاب درخت بهینه تعریف شده با کارایی مورد انتظار روی یک مجموعه مستقل از داده‌ها یافتن درختی با دقت و اندازه مناسب است. یکی از روش‌های آماری برای انتخاب درخت بهینه از بین زیر درخت‌های حاصل از هرس کردن درخت، استفاده از خطای استاندارد و قاعده *ISE* است. این روش یک روش آماری برای اندازه‌گیری عدم قطعیت برآوردهای ($R^{ts}T$) و ($R^{cv}T$) به وسیله خطای استاندارد آن‌ها است.



نمودار ۲- منحنی رده بندی هزینه اشتباه برای داده‌های آموزشی و داده‌های آزمون

1. Optimal tree selection

۱-۱۱- مدل رگرسیون درختی

ساختار رگرسیون درختی همانند رده بندی درختی است که در آن فضای X بوسیله افرازهای دوتایی به ترتیب تفسیر می‌گردد تا گروه‌های پایانی بدست آیند که در هر گره پایانی، t مقدار پیش بینی شده پاسخ (yt) است.

در تمام مدل‌های رگرسیونی، داده‌ها شامل (X,Y) هستندکه X فضای اندازه گیری و متغیر مستقل نامیده می شودو Y متغیر پاسخ با مقادیر حقیقی است. تشکیل یک مدل رگرسیونی برای دو هدف است:

۱- پیش بینی متغیر پاسخ مطابق بردار اندازه متغیر مستقل

۲- فهمیدن رابطه بین متغیرهای اندازه گیری (مستقل) و متغیر پاسخ

برای ساخت مدل رگرسیون درختی باید مجموعه داده‌های آموزشی دارای N حالت بصورت $(X_1, Y_1)\ldots(X_n,y_n),\ldots,\ldots,\ldots$ باشند. بنابراین در رگرسیون درختی نیز سه جزء اساسی تعیین کننده درخت پیش بینی کننده عبارت‌اند از:

۱- روشی برای انتخاب یک افراز در هر گره میانی

۲- ضابطه ای برای تعیین گره‌های پایانی

۳- ضابطه ای برای تخصیص مقادیر y(t) به هر گره پایانی t (مقادیر پیش بینی شده متغیر پاسخ)

سؤال مطرح شده در مورد مدل رگرسیون درختی این است که چگونه می‌توان دقت پیش بینی حاصل از این مدل را بررسی کرد؟ برای پاسخ به این سؤال یک مجموعه داده آزمون بزرگی همانند مجموعه داده‌های آموزشی به حجم N2 در نظر گرفته می‌شود. دقت مدل رگرسیون درختی همانند رگرسیون معمولی بوسیله میانگین خطا قابل اندازه گیری است و بصورت زیر تعریف می‌شود:

$$\frac{1}{N_2}\sum_{n=1}^{N_2}|y_n' - d(x_n')|$$

۱-۱۲- افراز در رگرسیون درختی

بنابراین رگرسیون درختی همانند دیگر مدل‌های درختی بوسیله افرازهای مکرر که منجر به مکزیموم کردن کاهش در (TR) می‌گردند، ساخته می‌شود. اما در رده بندی درختی افرازی مطلوب است و انتخاب می‌گردد که برآورد رده بندی اشتباه را مینیموم کند. در بحث رگرسیون درختی می‌توان گفت که بهترین افراز کننده روی یک متغیر در گره t، افرازی است که بطور موفق مقادیر بزرگ متغیر پاسخ را از مقادیر کوچک آن تفکیک کند. پس بطور کلی در این مدل رگرسیونی یکی از (t_L) یا (t_R) از (ty) کمتر است و دیگری بزرگ‌تر.

۱-۱۳- برآورد R(t) در رگرسیون درختی

همانطور که در رده بندی درختی اشاره شد برای انتخاب یک درخت با اندازه بهینه از توالی زیر درخت‌های تولید شده برآورد (T_KR) ضروری است.

$\mathbf{T_1 > T_2 > \ldots . \{t_1\}}$

در رگرسیون درختی همانند مدل رده بندی برای برآورد (TR) از دو تکنیک اعتبارسنجی مدل و نمونه مستقل استفاده می‌شود. برای برآورد از روش نمونه مستقل، نمونه مورد نظر بطور تصادفی به دو قسمت نمونه آموزشی و نمونه آزمون تقسیم می‌شود. نمونه آموزشی برای برازش و ساخت درخت مورد استفاده قرار می‌گیرد. در عمل بیشتر از روش اعتبارسنجی برای برآورد استفاده می‌شود مگر این که حجم نمونه بسیار بزرگ باشد. نحوه اجرای این دو تکنیک در رگرسیون درختی همانند رده بندی درختی است.

۲- عملیاتی سازی مدل درختی

فرض کنید محققی قصد دارد تنوع مطالبات زنان را در مجموعه‌ای از کشورهای اسلامی با توجه به ساختارهای سیاسی، قومیتی و مذهبی مقایسه کند، استفاده از روش تحلیل درختی در این مطالعه مقایسه زوجی کشورهای مسلمان تحلیل مطالبات

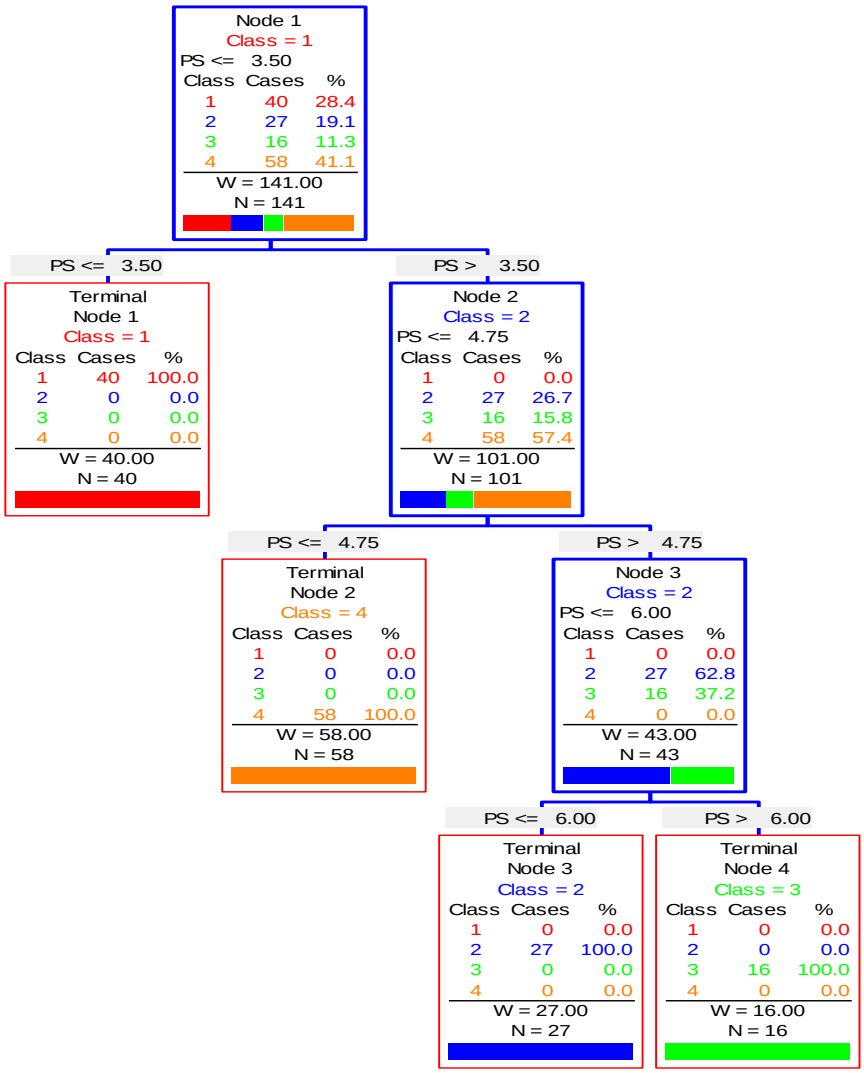
جنبش‌های زنان در آن‌ها را فراهم می‌کند که خود منجر به شناخت و پیش‌بینی نیازهای این گروه اجتماعی می‌شود.

با استفاده از روش تحلیل درختی استدلال‌های منطقی در تبیین دلایل تنوع مطالبات جنبش اجتماعی زنان بررسی و نسبت آن با نتایج به‌دست آمده مشخص می‌شود. در واقع، برحسب این‌که شکاف فعال در هر یک از کشورهای مورد بررسی در کدام بخش است، مبنایی برای تحلیل جنبش‌های اجتماعی زنان و مطالعه تنوع مطالبات آنان حاصل می‌شود.

در این بخش از تحقیق به‌منظور بررسی تأثیر سهم هر یک از متغیرهای پیش‌بینی شده در هر کشور از نمودار درختی استفاده شده است. این نمودار قادر است این تغییرات را بر حسب متغیرهای مهم و تأثیرگذار در هر کشور به‌تفکیک مشخص کند. همان‌طور که قبلاً مطرح شد این تحلیل کشورها را دوتا دوتا با هم مقایسه می‌کند.

در نمودار شماره ۳ طبقه یک نمایانگر وضعیت جنبش زنان در ترکیه، طبقه دو در مصر، طبقه سه عربستان سعودی و طبقه چهار وضعیت جنبش زنان در مالزی را بر اساس سهم مهم‌ترین و تأثیرگذارترین متغیرها در تفکیک کشورهای مورد بررسی نشان می‌دهد که ارقام مندرج در نمودار نشان می‌دهد متغیر ساختار سیاسی در هر چهار کشور مهم‌ترین عامل تفکیک جنبش زنان محسوب می‌شود. از میان کشورهای مورد بررسی عربستان و مصر به‌دلیل بالا بودن میانگین اقتدارگرایی در ساختار سیاسی در مقایسه با کشور مالزی و ترکیه مشابهت بیشتری با هم دارند. از آن‌جا که متغیر ساختار سیاسی به‌عنوان یک تبیین‌کننده قوی در معادله وارد شده است، سهم متغیرهای دیگر

نظیر ساختار قومیتی، موضوعات مبارزه، گرایش‌های زنان به مشارکت در جنبش و سازمان جنبش زنان را در تفکیک کشورها کنار زده است.



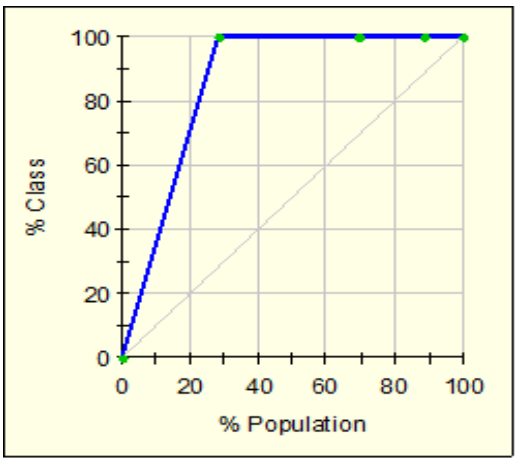
نمودار ۳- نمودار درختی

جدول ۱- درصد تبیین متغیرهای پیش بین تنوع مطالبات زنان بر حسب کشور

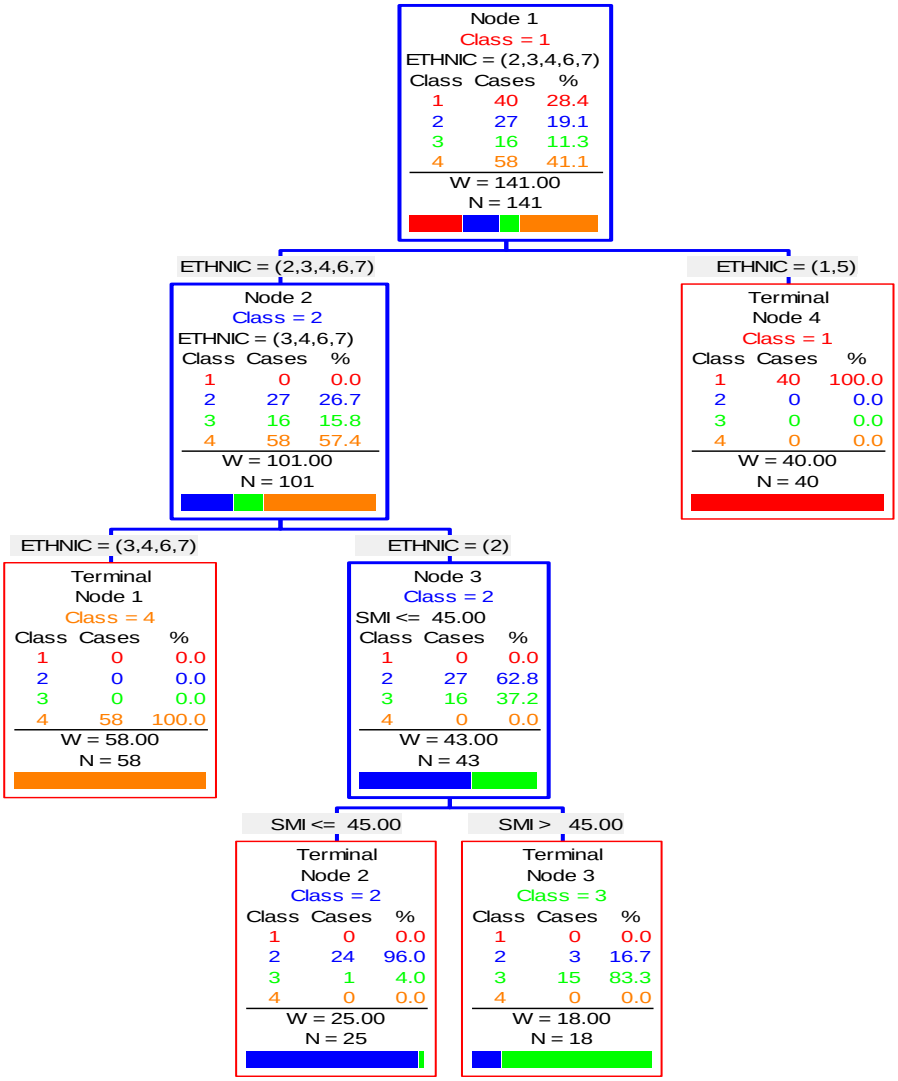
متغیرهای تبیین‌گر	درصد تبیین
ساختار سیاسی	۱۰۰
ساختار قومیتی	٦٦/٦
موضوعات مبارزه جنبش	٤٤/٤
عملکرد سازمان جنبش	۱۸/۱
گرایش اعضاء به مشارکت در جنبش	۹/۷

جدول شماره ۱ نشان می‌دهد که در هر چهار کشور مورد بررسی نقش ساختارهای سیاسی بسیار زیاد است. پس از آن ٦٦/٦٪ سهم ساختارهای قومیتی است. اهمیت موضوعات مبارزه جنبش زنان در کشورهای مورد مطالعه ٤٤/٤٪، عملکرد سازمان جنبش ۱۸/۱٪ و گرایش اعضاء به مشارکت زنان ۹/۷٪ به‌دست آمده است.

نمودار ٤- نیکویی برازش مدل



نمودار شماره ٤ نشان می‌دهد که دقت نمودار درختی در سنجش تأثیر متغیرهای وارد شده بسیار بالا است. هر چه سطح نمودار زیر منحنی بالاتر باشد نشان دهنده نیکویی برازش مدل است.



نمودار ۵- نمودار درختی

نمودار درختی شماره ۵ تأثیر ساختارهای قومیتی و موضوعات مبارزه زنان را در هر یک از کشورهای مورد بررسی به‌تفکیک نشان می‌دهد. بر اساس ارقام مندرج در این نمودار ساختار قومیتی بیشترین اثرگذاری را در جنبش زنان مالزی و ترکیه گذاشته است. بنابراین این دو کشور بر اساس ساختار قومیتی قابل تفکیک هستند و موضوعات مبارزه زنان در این دو کشور تفاوت معنی‌داری ندارند. در حالی‌که در کشور عربستان و مصر علاوه بر تأثیرگذاری ساختار قومیتی، موضوعات مبارزه زنان نیز تبیین‌کننده مناسبی برای تفکیک این دو کشور محسوب می‌شود. نمودار درختی همچنین نشان داده است که در کشور عربستان میانگین موضوعات جنبش به‌شکل معناداری بیش از کشور مصر است.

جدول ۲- درصد تبیین متغیرهای پیش‌بین تنوع مطالبات زنان بر حسب کشور

متغیرهای تبیین‌گر	درصد تبیین
ساختار قومیتی	۱۰۰
موضوعات مبارزه جنبش	٦٦/٦
عملکرد سازمان جنبش	۲۵/٦
گرایش اعضاء به مشارکت در جنبش	١٤/٦

ارقام مندرج در جدول شماره ۲ نشان می‌دهد که با حذف تأثیر ساختار سیاسی که تبیین‌گر بسیار مهم در تفکیک کشورها به‌لحاظ جنبش زنان محسوب می‌شود، ساختار قومیتی در مقام اول اهمیت قرار می‌گیرد، اما در این وضعیت نقش موضوعات مبارزه زنان در کشورهای مختلف اهمیت خود را نشان می‌دهد که ارقام مندرج در جدول نشان می‌دهد در شرایطی که ساختار قومیتی ۱۰۰٪ تمایز میان کشورها را تبیین کند موضوعات مبارزه ٦٦/٦٪، عملکرد سازمان جنبش ۲۵/٦٪ و گرایش اعضاء به

مشارکت در جنبش ۱٤/٦٪ سهم دارند. در هر دو حالت متغیر گرایش‌های اعضاء به مشارکت زنان کمترین سهم را در تفکیک جنبش در کشورهای مختلف نشان داد.

بحث و نتیجه‌گیری

در این قسمت با توجه به بررسی ساختار درختی و مقایسه آن با دیگر انواع تحلیل‌های رگرسیونی، لازم است مزایا و معایب این روش‌ها مورد بحث و بررسی قرار گیرد.

مزایای استفاده از تحلیل درختی در تحقیقات اجتماعی

روش‌های مبتنی بر مدل‌های برخلاف روش‌های کلاسیک نیازمند فرض‌های کمتری هستند و شامل طیف وسیعی از داده‌ها هستند. به همین دلیل در دهه‌های اخیر این روش‌ها در برابر مدل‌های کلاسیک تحلیل ممیزی خطی و رگرسیون خطی مقبولیت عام یافته‌اند. بخصوص که این مدل‌ها برای حجم بالای داده‌ها برازش می‌شوند. همچنین مشکل ناشی از داده‌های گم شده در این مدل‌ها وجود ندارد.

انعطاف‌پذیری مدل‌های درختی نسبت به مدل‌های کلاسیک یکی از دلایل مقبولیت آن‌ها است. در واقع برای هر نوع از داده‌ها با تعدیل ضوابط افراز می‌توان مدل درختی برازش کرد. رده بندی درختی برای داده‌های اسمی، رگرسیون درختی برای داده‌های پیوسته و درخت بقا برای داده‌های بقا با داده‌های سانسور شده. بیان نتایج حاصل از مدل بر اساس ضوابط افراز منطقی به صورت یک درخت تصمیم و کشف ساختارهای پیچیده داده‌ها که منجر به تفسیر ساده و قابل درکی از آن‌ها می‌شود، یکی از نکات جالب توجه برای محققین در زمینه‌های گوناگون است.

در آنالیز بقا هدف عمده، یافتن رابطه بین مجموعه متغیرهای کمکی و زمان وقوع رخداد مورد نظر است. مدل خطرات متناسب کاکس روش مناسبی برای این امر است، اما مدل‌های درختی بقا علاوه بر دستیابی به این هدف، منجر به تعیین زیرگروه‌هایی با وضعیت مشابه می‌گردند. این امر برای محققین اجتماعی که تمایل به تحقیقات تطبیقی دارند، بسیار مهم‌تر است.

روش‌های مبتنی بر مدل‌های درختی برای داده‌های بقا بهتر از مدل‌های کلاسیک فاکتورهای پیش آگهی را تعیین می‌کنند. با استفاده از مدل‌های درختی می‌توان زیرگروه‌های همگنی از نمونه‌هایی به دست آورد که شرایط علّی گوناگونی دارند. نمودار درختی امکان تحلیل و پیش بینی وضعیت زنان در کشورهای مختلف اسلامی را فراهم می‌کند که از یک سو مطالبات متفاوتی دارند و از سوی دیگر شرایط ساختاری متفاوتی را تجربه می‌کنند.

محدودیت‌های مدل‌های درختی

۱- ترکیب متغیرها،

۲- یکی از معایب ساختار درختی استفاده از ساختار استاندارد افراز است که همه افرازها تنها با استفاده از یک متغیر صورت می‌پذیرند. در حالتی که ساختار درختی به ترکیب متغیرها بستگی داشته باشد، برنامه استاندارد درختی عملکرد بسیار ضعیفی خواهد داشت. در چنین حالتی مجموعه‌های مجاز از افزارهای توسعه یافته وجود دارد که شامل تمام ترکیبات خطی متغیرها است. در واقع، یک الگوریتم برای جستجو از میان افرازها و تلاش برای یافتن افرازی که نیکویی ضابطه افراز را افزایش دهد، ایجاد می‌شود. چنین حالتی که در افرازها از ترکیبات خطی متغیرها استفاده می‌شود، قابل قیاس با تحلیل ممیزی است و البته عملکرد آن خیلی بهتر از تحلیل ممیزی است.

منابع

- طالبان، محمدرضا، (۱۳۸۸). درآمدی روش شناسانه بر تحلیل بولی فوران از انقلاب ایران، ***فصلنامه علوم اجتماعی***، شماره ٤٢، ٤٣.
- کوثری، مسعود، (۱۳۸٦). تحلیل مقایسه‌ای کیفی در علوم اجتماعی، ***نامه علوم اجتماعی***، شماره ۳۱.

- محمدی، نعیما. (۱۳۹۳). بررسی موردی مطالبات زنان در کشورهای ترکیه، مصر، عربستان سعودی و مالزی. ***فصلنامه مطالعات سیاسی جهان اسلام***. دوره ٤، شماره ۱۲.
- محمدی، نعیما. (۱۳۹۲). بررسی جامعه شناختی جنبش‌های اجتماعی زنان در کشورهای عضو کنفرانس اسلامی. ***رساله دوره دکتری***، رشته جامعه شناسی سیاسی. دانشگاه تربیت مدرس.

- Banerjee M, Biswas D, Sakr W and Wood D. (2000). ***Recursive partitioning for prognostic grouping of patients with clinically localized prostate carcinoma***. American Cancer Society, 89: 404-411.
- Breiman L, Friedman J.H., OlShen R. A. and Ston C. J. (1984). ***Classification and Regression Trees***. California, A Division of Wadsworth Inc.
- Callaghan F. (2008). ***Classification trees for survival data with competing risk***. [dissertation], Doctor of philosophy, University of Pitters burgh,
- Cappelli C, D'Elia A. (2006). ***A tree-based method for selection of variables in models for ordinal data***, Quaderni di Statistica, 8: 125-135.
- Dean L.S. (2007). ***A Method for detecting optimal splits over time in survival analysis using***. tree-structured models. [dissertation], Doctor of philosophy, University of Pitters.burgh.
- Garzotto M, Beer T, Hodson R, Peters L, Hsieh Y, Barrera E,Klien T,Mori M. (2005). ***Improved detection of prostate cancer using classification and regression tree analysis***, Clinical Oncology; 23 (19): 4322-4329.
- Gimotty P, Elder D, Fraker D, Botbyl J,Seller K, Elenitsas R, Ming M E, Schuchter, L, Spitz F R,Czerniecki B J,Guerry D. (2007***). Identification***

of high-risk patients among those diagnosed with thin cutaneous melanomas, Clinical Oncology; 25(9): 1129-1134.

- Hess K R, Abbruzzese M C, Lenzi R, Raber M N, Abbruzzese J L. (1999). ***Classification and regression tree analysis 0f 1000 consecutive patients with unknown primary carcinoma***. Clinical Cancer Research; 5: 3403-3410.
- Hothorn T, Hornik K, Zeileis A. (2008). Unbiased recursive partitioning: A Conditional inference framework***. Journal of Computational and Graphical Statistics***, 15(3): 651-674.
- Hothorn T, Hornik K, Zeileis A. (2006). ***Unbiased Recursive Partitioning: A Conditional Inference Framework***. Computational and Graphical statistics; 15(3):651-674.
- Kantardzic M (2003). ***Data Mining: Concept, Model, Method and Algorithms***. Wiley, Inter science.
- Lamborn K. Chang S and Prados M. (2004). ***Prognostic factors for survival of patients with glioblastoma:*** Recursive partitioning analysis. Nerro_Oncology; 6: 227-235.
- Liu J, Letaief K, Coa Z et al. (2004). ***A Reduced- complexity maximum-likelihood method for Multiuser detection***, IEEE Transaction on Communication, 25(2):289-295.
- Molinaro A, Dudoit S, Van der Laan M. (2004). ***Tree-based multivariate regression and density estimation with right-censored data. Multivariate Analysis***, 90: 154-177.
- Oberc M. (1993). ***Tree-structured methods for the proportional hazards model***. [dissertation], Doctor of philosophy, University of Toronto,

- Schittgen R. (1999). Regression trees for survival Data-an approach to select Discontinuous split points by ***rank statistics Biometrical Journal***; 41: 943-954.
- Segal M. (1988). ***Regression trees for Censored data***. Biometrics,44: 35-48.
- Takashi O, Cook E.F, NakaMura T,Saito J, Ikawa F, Fukui T. (2006). ***Risk stratification for in-hospital mortality in spontaneous intra cerebral haemorrhage***: A Classification and Regression Tree analysis. QJ Med,99: 743-750.
- Wang J. (2002). ***Tree-structured classification for multivariate binary responses***. [Dissertation], Doctor of philosophy, North Carolina State University.
- Wernecke K, Possinger K, Kalb G and Stein J. Validating classification
- WU Y, Genton M, Stefanski L. (2009). ***A Comparison of Node-Splitting Rules in Recursive Partitioning Analysis of Multivariate Quantitative Structure Activity Data.*** Statistics in Biopharmaceutical Research, 1(2): 119-130.