Journal of Statistical Sciences, Autumn and Winter, 2024
Vol. 18, No. 2, pp 415-434
DOI: 10.61186/jss.18.2.2

Survival Data Analysis Using Different Statistical Learning Methods

Madadi, M., Motarjem, K.
Department of Statistics, Tarbiat Modares University, Tehran, Iran.

Corresponding author: K. Motarjem, K.motarjem@modares.ac.ir

Received: 15/1/2023 **Revised:** 11/12/2024 **Accepted and Published Online:** 12/12/2024.

Introduction
The increasing volume and complexity of data in survival analysis necessitate using statistical learning methods. These methods offer the capability to estimate survival probabilities and evaluate the influence of various factors on patient survival outcomes. This paper aims to compare the performance of the Cox model, a standard model in survival analysis, with penalized methods such as Cox regression with ridge and lasso penalties, as well as statistical learning approaches like random survival forests and neural networks. The comparison will shed light on the effectiveness of these methods in different scenarios, enabling researchers to choose the most suitable approach for their specific research questions and data.

Material and Methods
Simulation studies indicate that under conditions with a linear relationship among variables, the performance of the above mentioned methods is nearly equivalent to that of the Cox model. However, approaches such as Cox lasso, random survival forests, and neural networks demonstrate superior performance in non-linear scenarios and with high dimensionality of variables. Furthermore, to evaluate the performance of these models, they were applied to data from patients with atheromatous, revealing that in the presence of a large number of explanatory variables, statistical learning approaches generally outperform classical survival analysis models.

Results and Discussion
The primary objective of this paper is to compare various statistical learning

methods in survival data analysis. Although classical approaches like the proportional hazards Cox model are widely popular and have a long history, recent statistics and data science advancements highlight the need for more flexible methods. Simulation results show that the choice between statistical learning methods, such as random survival forests and neural networks, versus classical methods, like the Cox model, depends on the specific characteristics inherent in the data. Additionally, random survival forests and neural networks are more adaptable and capable of elucidating complex relationships between independent variables and survival time. Notably, random survival forests and neural networks are particularly beneficial when interactions and non-linear relationships between explanatory variables and survival time exist. The practical example examined in this study also demonstrated that statistical learning methods can effectively handle many variables due to their inherent variable selection or regularization capabilities. A comparison of performance metrics revealed that Cox lasso and random survival forests performed better in analyzing atherosclerosis patients' survival data.

Conclusion
In conclusion, this study underscores the advantages of employing statistical learning methods in survival analysis, particularly in complex datasets with high dimensionality and non-linear relationships.

Keywords: Survival Analysis, Statistical Learning,Cox Model, Random Survival Forests, Neural Networks .

Mathematics Subject Classification (2010): 62N01, 68T05.

مجله علوم آماری، پاییز و زمستان ۱۴۰۳
جلد ۱۸، شماره ۲، ص ۴۱۵ -- ۴۳۴
DOI: 10.61186/jss.18.2.2
مقاله پژوهشی

تحلیل داده‌های بقا با روش‌های مختلف یادگیری آماری

مهرنوش مددی، کیومرث مترجم
گروه آمار، دانشگاه تربیت مدرس

نویسنده مسئول : کیومرث مترجم، k.motarjem@modares.ac.ir
تاریخ دریافت: ۱۴۰۲/۹/۲۵ تاریخ بازنگری: ۱۴۰۳/۸/۲۱ تاریخ پذیرش و انتشار: ۱۴۰۳/۸/۲۲

چکیده: با توجه به حجم و پیچیدگی داده‌های نوظهور در تحلیل بقا، بکارگیری روش‌های یادگیری آماری در این حوزه به امری اجتناب ناپذیر بدل شده است. این روش‌ها قادر به برآورد احتمال بقا و تأثیر عوامل مختلف بر بقا هستند. در این مقاله، عملکرد مدل کاکس به عنوان یک مدل رایج در تحلیل بقا با روش‌های مبتنی بر تاوان مانند کاکس ریج و لاسو و روش‌های مبتنی بر یادگیری آماری مانند جنگل بقای تصادفی و شبکه عصبی مورد مقایسه قرار گرفته است. نتایج شبیه‌سازی‌ها در این مطالعه نشان می‌دهد که در شرایط وجود رابطه خطی بین متغیرها، عملکرد مدل‌های مذکور تقریباً مشابه مدل کاکس است اما در حالات غیرخطی و بالا بودن بعد متغیرها، روش‌هایی مانند کاکس لاسو، جنگل بقای تصادفی و شبکه عصبی عملکرد بهتری دارند. درنهایت به منظور ارزیابی عملکرد این مدل‌ها در تحلیل داده‌های بیماران مبتلا به آترواسکلروز مورد استفاده قرار گرفتند و نتایج نشان داد که در مواجهه با داده‌هایی با تعداد متغیرهای تبیینی زیاد، رویکردهای یادگیری آماری به طور کلی عملکرد بهتری نسبت به مدل‌های کلاسیک تحلیل بقا دارند.
واژه‌های کلیدی: تحلیل بقا، یادگیری آماری، مدل کاکس، جنگل بقای تصادفی، شبکه عصبی
کد موضوع‌بندی ریاضی (۲۰۱۰): 62N01 ، 68T05.

۱ مقدمه

اهمیت تحلیل بقا در علوم پزشکی و پژوهش‌های مرتبط با آن، باعث شده پژوهشگران به دنبال روش‌های کارآمدی برای تحلیل داده‌های بقا باشند. از طرفی با پیشرفت فناوری و امکان دسترسی به داده‌های بیشتر، نیاز به روش‌های

دقیق‌تر در این حوزه هر روز بیشتر احساس می‌شود (اسمیت، ۲۰۲۰). بطور کلی روش‌های تحلیل بقا را می‌توان به دو دسته عمده طبقه‌بندی نمود که شامل روش‌های آماری مرسوم و روش‌های مبتنی بر یادگیری آماری است. هر دو رویکرد به دلیل برآورد توابع بقا و خطر دارای اهدف مشترکی هستند، با این حال روش‌های آماری مرسوم، بیشتر بر توصیف زمان تا وقوع رویداد و خواص آماری برآورد پارامترها و همچنین برآورد تابع بقا تمرکز دارند؛ اما در مقابل روش‌های یادگیری آماری عمدتاً برای پیش‌بینی وقوع رویداد در یک زمان معین با ترکیب روش‌های آماری مرسوم و تکنیک‌های مختلف یادگیری آماری متمرکز هستند (هارل، ۲۰۰۱). یکی از عمده‌ترین دلایل استفاده از روش‌های مبتنی بر یادگیری آماری در تحلیل بقا توانایی آنها در مدل‌بندی روابط پیچیده و غیرخطی، به ویژه در مواجهه با داده‌های بقای بعد بالا است درحالیکه روش‌های آماری مرسوم عمدتاً در برازش به داده‌هایی که روابط بین متغیرها خطی و بعد داده‌های نیز پایین باشد مورد استفاده قرار می‌گیرند. در دهه اخیر، روش‌های مبتنی بر یادگیری آماری مانند شبکه‌های عصبی، جنگل بقای تصادفی و روش‌های مبتنی بر تاوان مانند کاکس ریج و لاسو، در این زمینه بسیار مورد توجه قرارگرفته‌اند. تحلیل داده‌های بقا با استفاده از این روش‌ها موجب حصول دقت و صحت بیشتری در پیش‌بینی و ارزیابی عوامل موثر بر زمان بقا شده است (اسمیت و جونز، ۲۰۲۰). در این مقاله، عملکرد مدل کاکس به عنوان یک مدل پرکاربرد در تحلیل داده‌های بقا با روش‌های مبتنی بر تاوان مانند کاکس ریج و لاسو و همچنین روش‌های مبتنی بر یادگیری آماری مانند جنگل بقای تصادفی و شبکه عصبی مورد مقایسه قرارگرفته است. ادامه مقاله، به بررسی روش‌های مختلف تحلیل بقا با تأکید بر روش‌های یادگیری آماری و مقایسه آنها با مدل کاکس، پرداخته و بدین صورت سازماندهی شده است. در بخش ۲، به بررسی نظریه یادگیری آماری و روش‌های مبتنی بر این رویکرد یعنی رگرسیون کاکس لاسو و ریج پرداخته می‌شود. در بخش ۳، به معرفی جنگل بقای تصادفی به عنوان یکی از روش‌های یادگیری آماری برای تحلیل بقا پرداخته شده است. در بخش ۴، مدل بقای شبکه عصبی و نحوه کاربرد آن در تحلیل داده‌های بقا مورد بررسی قرارگرفته است. بخش ۵ به مطالعه شبیه‌سازی برای ارزیابی عملکرد مدل‌های معرفی شده می‌پردازد. در بخش ۶، کاربست روش‌های معرفی شده در تحلیل داده‌های بیماران مبتلا به آترواسکلروز ارائه می‌شود و در انتها نیز به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ نظریه یادگیری آماری

نظریه یادگیری آماری پایه مهمی برای رویکرد یادگیری ماشین و تحلیل بوسیله یادگیری از داده‌ها است. با بکارگیری نظریه یادگیری آماری، تلاش می‌شود تا الگوهای موجود در داده‌ها به طور خودکار شناسایی و استخراج شوند، به گونه‌ای که بتوان از آن‌ها برای پیش‌بینی و تصمیم‌سازی بهره برد. به بیان ساده در یادگیری آماری هدف پیش‌بینی یک متغیر تصادفی $\boldsymbol{Y}$ بر اساس متغیرهای کمکی $\boldsymbol{X} = \boldsymbol{x}$ با استفاده از نمونه تصادفی $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ است. در این رویکرد روش‌های مختلفی وجود دارند که هر کدام در جنبه‌های خاصی از الگوریتم‌ها و روش‌های مورد استفاده در داده‌کاوی، هوش مصنوعی و یادگیری ماشین کاربرد دارند. توسیع یادگیری آماری در تحلیل بقا بواسطه وجود سانسور به عنوان یک ویژگی اصلی داده‌ها، از چالش‌های پیش رو است و نیاز به تغییر در روش‌های سنتی

یادگیری آماری را طلب می‌نماید. اهمیت سانسور در داده‌های بقا از این منظر است که بر محاسبه تابع زیان تأثیر می‌گذارد زیرا زیان واقعی موارد سانسور شده را نمی‌توان محاسبه کرد. تابع زیان در نظریه یادگیری آماری معیاری است که برای اندازه‌گیری عملکرد یا کیفیت یک مدل پیش‌بینی استفاده می‌شود. این تابع برای ارزیابی اختلاف بین مقدار واقعی و مقدار پیش‌بینی شده توسط مدل استفاده می‌شود. هدف اصلی در انتخاب تابع زیان، کاهش این اختلاف و بهبود عملکرد مدل است.

روش‌های متعددی برای گنجاندن سانسور در فرایند یادگیری آماری معرفی شده‌اند که یکی از مهمترین آنها وزن دهی به روش احتمال معکوس[1] است، که توسط رابینز روتنیتزکی (۱۹۹۲) پیشنهاد شده است. علاوه بر سانسور، چالش دیگر، یافتن توابع زیان مناسب برای داده‌های بقا با توجه به ماهیت ذاتی این داده‌ها است. برخی از توابع زیان پیشنهادی که برای تحلیل داده‌های بقا مناسب هستند در هندرسون (۱۹۹۵) معرفی شده است. در ادامه این بخش با توجه به اهمیت توابع زیان در مدل‌سازی داده‌های بقا دو روش مهم مبتنی بر رویکرد منظم‌سازی در تحلیل بقا یعنی رگرسیون کاکس لاسو و ریج معرفی می‌شوند. رویکرد منظم سازی در نظریه یادگیری ماشین به روش‌هایی اشاره دارد که هدفشان جلوگیری از پدیده بیش‌برازش[2] در مدل‌های یادگیری ماشین است. بیش‌برازش به معنای تطبیق بیش از حد مدل به داده‌های آموزشی است، به گونه‌ای که مدل توانایی تعمیم به داده های جدید را از دست می‌دهد. این امر موجب می‌شود مدل بتواند بهتر به داده‌های جدید تعمیم یابد.

در مدل کاکس، پارامترهای رگرسیونی با بیشینه کردن تابع درستنمایی جزئی کاکس L_p یا لگاریتم آن ℓ_p به دست می‌آیند. با این حال، بیشتر روش‌های جدید در حوزه یادگیری آماری بر کمینه کردن تابع زیان تمرکز دارند. لذا به منظور استفاده عملی، بازفرموله کردن مسئله الزامی است. با استفاده از معادله

$$\arg\max_{\boldsymbol{x}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}}(-\boldsymbol{x})$$

برای هر $\boldsymbol{x}$ دلخواه، برآورد ماکسیمم درستنمایی را می‌توان با مینیمم‌سازی لگاریتم درستنمایی منفی (جزئی) بدست آورد. روش تاوان L_1، در مدل‌های رگرسیونی توسط تیبشیرانی (۱۹۹۷) برای مدل مخاطرات متناسب کاکس بسط داده شد. او روشی را پیشنهاد کرد که درستنمایی جزئی را با محدودیت کمتر بودن مجموع مقادیر قدرمطلق پارامترها از یک مقدار ثابت t کمینه نماید و بدین ترتیب تابع درستنمایی جزئی مدل مخاطرات متناسب با بردار ضرایب $\boldsymbol{\beta}$ را به‌صورت

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{x}_{(i)}^T \boldsymbol{\beta})}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta})} \qquad (1)$$

نوشت، که در آن $R(t)$ مجموعه‌ای از افراد در معرض خطر در نقطه زمانی t است. با نشان دادن لگاریتم درستنمایی

[1] Inverse Probability Weighting
[2] Overfitting

جزئی با $\ell_p(\boldsymbol{\beta})$، پارامتر $\boldsymbol{\beta}$ در رابطه (۱) را می‌توان از طریق معیار

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}), \quad s.t. \quad \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \leq t$$

و الگوریتم زیر برآورد کرد.

الگوریتم ۱. برآورد لاسو برای رگرسیون کاکس *(تیبشیرانی، ۱۹۹۷)*

گام ۱- t را ثابت و مقدار اولیه $\hat{\boldsymbol{\beta}} = 0$ در نظر گرفته شود.

گام ۲- $\boldsymbol{\eta}$، $\boldsymbol{u}$، $\boldsymbol{A}$ و $\boldsymbol{z}$ براساس مقدار فعلی $\hat{\boldsymbol{\beta}}$ محاسبه شوند.

گام ۳- $(\boldsymbol{z} - \boldsymbol{X\eta})^T \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{X\eta})$ به شرط $\sum_j |\boldsymbol{\beta}_j| \leq t$ مینیمم‌سازی شود.

گام ۴- مراحل ۲ و ۳ تا زمانی که $\hat{\boldsymbol{\beta}}$ تغییر نکند تکرار شود.

پارامتر تاوان[1] یا توسط کاربر انتخاب می‌شود یا توسط خود داده‌ها تعیین می‌شود. توجه شود که چون مدل مخاطرات متناسب کاکس فاقد عرض از مبدا است، لذا، مرحله ۳ نیز نیازی به عرض از مبدا ندارد. در مباحث مربوط به پیاده‌سازی، رویکرد رایج الگوریتمی است که توسط سایمن و همکاران (۲۰۱۱) معرفی شد که راهکاری سریع و کارآمد برای استفاده از مدل کاکس مبتنی بر تاوان است. این به معنای ماکسیمم‌سازی درستنمایی جزئی با ترکیب تاوان ریج و لاسو است؛ به طوری که با قرار دادن $\alpha = 1$ رگرسیون لاسو و با قرار دادن $\lambda = 1$ رگرسیون ریج حاصل می‌شود. این الگوریتم در واقع مشابه الگوریتم ارائه شده توسط تیبشیرانی (۱۹۹۷) عمل می‌کند. این الگوریتم در ابتدا برای رگرسیون خطی پیشنهاد شد و توسط محققین زیادی بهبود یافته است (لیو و همکاران، ۲۰۲۳). در این روش، هدف کلی کمینه کردن مقدار یک تابع هدف خطی با محدودیت‌هایی بر روی پارامترهای مدل است. تابع هدف کلی به صورت

$$M(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (w(\eta_i)) \left(z(\eta_i) - \boldsymbol{x_i}^T \boldsymbol{\beta}\right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\boldsymbol{\beta}_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \boldsymbol{\beta}_j{}^2\right) \qquad (۲)$$

تعریف می‌شود که در آن، n تعداد داده‌ها، $w(\eta)$ وزن داده‌ها، $\boldsymbol{x}_i$ بردار ویژگی‌های داده‌ی i–ام، $z(\eta)_i$ مقدار هدف برای داده i–ام، $\boldsymbol{\beta}$ بردار پارامترهای مدل و λ و α پارامترهای تنظیم‌کننده هستند. سپس با محاسبه مشتق (۲) نسبت به بردار پارامترهای مدل عبارت

$$\frac{\partial M(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{2}{n} \sum_{i=1}^{n} w(\eta_i) \left(z(\eta_i) - \boldsymbol{x_i}^T \boldsymbol{\beta}\right) \mathbf{x}_i + \lambda \left(\alpha \operatorname{sgn}(\boldsymbol{\beta}) + (1-\alpha)\boldsymbol{\beta_j}\right)$$

[1] Penalty Parameter

به دست می‌آید، که در آن، $\text{sgn}(\boldsymbol{\beta}_j)$ نشانگر علامت $\boldsymbol{\beta}_j$ است و تابع signum را نمایش می‌دهد. هدف به دست آوردن بهترین برآورد برای پارامترهای مدل بر اساس

$$\boldsymbol{\beta}j^* = \frac{S(n^{-1}\sum_{i=1}^{n} w(\eta)_i \boldsymbol{x}_{i,k}) \cdot \left[z(\eta)_i - \sum_{j \neq j^*} \boldsymbol{x}_{ij}\boldsymbol{\beta}_j\right], \lambda\alpha)}{n^{-1}\sum_{i=1}^{n} w(\eta)_i \boldsymbol{x}_{ij^*}^2 + \lambda(1-\alpha)}$$

است، که در آن $S(\boldsymbol{x}, \lambda) = \text{sgn}(\boldsymbol{x})(|\boldsymbol{x}| - \lambda)_+$ تابع Soft-Thresholding است. این معادله به صورت چرخه‌ای برای تمامی پارامترهای مدل انجام می‌شود تا با دست آوردن بهترین مقدار برای پارامترهای مدل، کمینه کردن تابع هدف حاصل شود (هیستی و همکاران، ۲۰۰۹). الگوریتم‌های زیادی برای ترکیب منظم‌سازی با مدل مخاطرات متناسب معرفی شده است. بهینه‌سازی می‌تواند از طریق ترکیبی از روش کاهش گرادیان و الگوریتم نیوتن رافسون یا الگوریتم LARS[1] انجام شود. سایمن و همکاران (۲۰۱۱) روش‌های گوناگونی را با توجه به مدت زمان اجرا برای حالت‌های مختلف داده مورد مقایسه قرار دادند. آنها نتیجه گرفتند که الگوریتم مبتنی بر کاهش مختصات چرخه‌ای به ویژه در مواجهه داده‌های بعد بالا، به طور قابل توجهی سریع‌تر از سایر روش‌ها عمل می‌کند.

۳ جنگل بقای تصادفی

روش‌های مبتنی بر درخت، یکی از محبوب‌ترین روش‌های یادگیری ماشین هستند. این روش‌ها با استفاده از یک ساختار درختی، داده‌های ورودی را به دسته‌های مختلف در طبقه‌بندی یا مقادیر پیوسته در رگرسیون تقسیم می‌کنند. با توسعه مفهوم درخت رگرسیون و تقسیم گره‌ها به صورت بازگشتی، یک مدل پیچیده‌تر به نام درختان رگرسیون[2] شکل می‌گیرد. در این مدل، تقسیم‌ها به صورت پیاپی انجام می‌شود و باعث تولید یک ساختار پیچیده‌تر می‌شود. اگرچه روش‌های متعددی برای تحلیل مبتنی بر درخت وجود دارد، اما یکی از رایج ترین الگوریتم‌ها توسط بریمن (۲۰۰۱) معرفی شده است که الگوریتم درختان طبقه‌بندی و رگرسیون(CART)[3] نامیده می‌شود. جنگل تصادفی[4]، یک روش جمعی است که برای غلبه بر مشکلات درختان رگرسیون و دستیابی به عملکرد پایدارتر استفاده می‌شود. کاربرد عمده جنگل‌های تصادفی بر مسائل رگرسیون و طبقه‌بندی متمرکز است اما جنگل‌های بقای تصادفی به مدل‌بندی داده‌های بقا سانسور شده می‌پردازند. برای ساخت یک درخت تکی رگرسیون برای داده‌های بقا، انتخاب یک معیار تقسیم مناسب، یعنی تابعی که باید بهینه شود، بسیار حائز اهمیت است. در همین ارتباط شیموکاوا و همکاران (۲۰۱۵) مقایسه‌ای از نه تابع تقسیم مختلف برای استفاده در درخت‌های بقا ارائه دادند. به عنوان مثال آنها، از تابع زیان لگاریتم درستنمایی به عنوان معیار تقسیم استفاده نمودند، که توسط دیویس و اندرسون (۱۹۸۹) پیشنهاد شده بود.

[1] Least Angle Regression
[2] Regression Trees
[3] Classification and Regression Trees
[4] Random Forest

این تقسیم، زیان را در میان تمام تقسیم‌های ممکن کمینه می‌کند.

فرض کنید N نمایانگر یک گره در درخت باشد و مجموعه نمونه‌ها برای آموزش به صورت $D = \{(t_i^*, \delta_i, x_i) \mid i = ۱, \ldots, n\}$ تعریف شود، که شامل زمان بقای مشاهده شده $T^* = \min(T, C)$، تابع نشانگر سانسور به صورت $\delta = ۱$ برای حالت $\{T \leq C\}$، و بردار متغیرهای مستقل X است. در تابع نشانگر سانسور، وقتی $\delta = ۱$ به معنای وقوع یک رویداد است، و وقتی $\delta = ۰$ به این معناست که مشاهده واقعی سانسور شده و فقط زمان سانسور قابل مشاهده است. در این صورت، خطر یک گره پایانی به صورت $h(x|N) = h_N$ تعریف می‌شود.

با روش ماکسیمم درستنمایی، برآورد h_N به صورت $\hat{h}_N = \frac{\sum_{i \in D_N} \delta_i}{\sum_{i \in D_N} x_i}$ به دست می‌آید همچنین، بر اساس شیموکاوا و همکاران (۲۰۱۵)، زیان لگاریتم درستنمایی برای یک گره N به صورت

$$-\log L_i(\hat{h}_N) = \sum_{i \in D_N} \delta_i - \sum_{i \in D_N} \delta_i \log L_i(\hat{h}_N)$$

محاسبه می‌شود. بنابراین تقسیم بهینه، باید این معیار را به حداقل برساند. در عمل استفاده از رگرسیون تک لایه یا درخت بقای تکی در عمل بسیار محدود است. به همین سبب در این بخش تمرکز اصلی بر روش ترکیبی جنگل‌های بقای تصادفی است که توسط ایشواران و همکاران (۲۰۰۸) پیشنهاد شده است. آن‌ها مدل جنگل تصادفی بریمن (۲۰۰۱) را به داده‌های بقای سانسور شده از راست تعمیم دادند و الگوریتمی به‌صورت زیر ارائه نمودند:

الگوریتم ۲. الگوریتم جنگل بقای تصادفی

گام ۱- تعداد M نمونه بوت استرپ از داده های اصلی استخراج شود.

گام ۲- برای هر نمونه بوت‌استرپ یک درخت بقا ساخته شود. در هر گره، p متغیر کاندید به طور تصادفی انتخاب شود. گره با استفاده از متغیر کاندیدی تقسیم شود که تفاوت بقا بین گره‌های فرزند را حداکثر می‌کند.

گام ۳- درخت به اندازه کامل رشد داده شود با این محدودیت که یک گره پایانی نباید کمتر از $d_۰ > ۰$ مرگ منحصر به فرد داشته باشد.

گام ۴- یک تابع خطر تجمعی *(CHF)*[1] برای هر درخت محاسبه شود و میانگین آنها در نظر گرفته شود تا CHF جمعی[2] به دست آید.

گام ۵- خطای پیشگویی برای *CHF* جمعی محاسبه شود.

در واقع CHF تجمعی، تابعی است که از میانگین‌گیری CHF های چندین مدل یا درخت حاصل می‌شود غالباً از آزمون لگاریتم رتبه برای مقایسه تفاوت بین دو گروه بر اساس تابع بقا استفاده می‌شود همچنین این آزمون می‌تواند به عنوان یک ابزار برای تقسیم‌بندی بر اساس زمان‌های بقا نیز به کار گرفته شود و در عین حال می‌تواند به عنوان

[1] Cumulative Hazard Function
[2] Ensemble CHF

ابزاری برای به حداکثر رساندن تفاوت‌های بقای بین گره بکار گرفته شود (لبلانک و کراولی، ۱۹۹۳). به عنوان یک راه حل فرض کنید $t_1 < \cdots < t_k$ زمان‌های رویداد متمایز در گره والد g باشند. علاوه بر این $d_{i,h}$ و $n_{i,h}$ به ترتیب تعداد مرگ‌ها و تعداد بیماران در معرض خطر در زمان t_i در گره‌های فرزند احتمالی $h = 1$ و $h = 2$ باشند. برای یک تقسیم در مقدار t برای پیش‌بینی کننده x_j، آزمون لگاریتم رتبه به صورت

$$LR(x_j, t) = \frac{\sum_{k=1}^{K}(d_{k,1} - n_{k,1}d_k/n_k)}{\sqrt{\sum_{k=1}^{K}\frac{n_{k,1}}{n_k}.(1 - \frac{n_{k,1}}{n_k})(\frac{n_k - d_k}{n_k - 1})d_k}} \quad (3)$$

تعریف می‌شود که در آن اندیس K تعداد زمان‌های وقوع رویداد متفاوت در گره والد است. لازم به ذکر است که بیشینه ساختن (۳) منجر به بهترین تقسیم می‌شود. در این روش، برای هر متغیر، یک تقسیم تصادفی انتخاب می‌شود و سپس متغیری که بیشترین آماره آزمون لگاریتم رتبه را دارد، برای تقسیم نهایی انتخاب می‌گردد. سوال بعدی پیش‌بینی در گره‌های پایانی است، یعنی چگونه می‌توان خطر تجمعی برای گره پایانی یک درخت و مجموعه CHF را برآورد نمود (مرحله ۴ الگوریتم). برای یک گره پایانی h و زمان‌های رویداد متمایز آن یعنی $t_{1,h} < t_{2,h} < \cdots < t_{k(h),h}$ برآورد CHF به صورت $\hat{\Lambda}_h(t) = \sum_{t_{k,h} \le t} \frac{d_{k,h}}{n_{k,h}}$ است، که در آن $d_{k,h}$ و $n_{k,h}$ به ترتیب بیانگر تعداد مرگ و میر و تعداد افراد در معرض خطر در زمان $t_{k,h}$ هستند. هر مشاهده جدید با متغیرهای x^* در یک گره پایانی منحصر به فرد h قرار می‌گیرد. بنابراین، پیش‌بینی بر اساس یک درخت منفرد به صورت $\Lambda(t|x^*) = \hat{\Lambda}_h(t)$ است که در آن $\Lambda_m{}^*(t|x)$ به عنوان CHF برای یک درخت واحد بر اساس یک نمونه بوت استرپ تعریف می‌شود. سپس، تابع مشخصه تجمعی بوت استرپ برای مشاهده i به صورت $\Lambda_e{}^*(t|x_i) = \frac{1}{M}\sum_{m=1}^{M}\Lambda_m{}^*(t|x_i)$ برآورد می‌شود. روش استاندارد جنگل‌های بقای تصادفی توسط اوتکین و همکاران (۲۰۱۹) به نسخه موزون توسعه داده شده است. آنها برای دستیابی به عملکرد بهتر، از میانگین وزنی به جای میانگین حسابی در برآورد تابع بقای جنگل استفاده نمودند. آخرین مرحله از الگوریتم پیاده‌سازی جنگل بقای تصادفی، محاسبه خطای پیش‌بینی است که با استفاده از شاخص هارل[1] احتمال هماهنگی[2] را برآورد می‌کند یک انتخاب مناسب است (هارل و همکاران، ۱۹۸۲). این شاخص که به اختصار C-index نوشته می‌شود برای ارزیابی عملکرد پیش‌بینی در مدل‌های بقا استفاده می‌شود.

۴ مدل‌های بقای شبکه عصبی

شبکه‌های عصبی عمدتاً در حوزه‌هایی که روابط بین متغیرها غیرخطی باشد مورد استفاده قرار می‌گیرند که یکی از مهمترین این حوزه‌ها تحلیل بقا است. تحلیل داده‌های بقا با استفاده از شبکه عصبی به دلیل وجود داده‌های سانسور شده بسیار پیچیده‌تر از آن چیزی است که در ابتدا تصور می‌شود. اولین تلاش‌ها برای غلبه بر این مشکل توسط

[1] Harrell's C-index
[2] Concordance Probability

فاراگی و سیمون (۱۹۹۵) مورد بررسی قرار گرفت. هدف آن‌ها مدل‌سازی داده‌های بقای سانسور شده با یک شبکه عصبی پیش‌خور بسیار ساده و رابطه ورودی–خروجی به عنوان مبنایی برای مدل مخاطرات متناسب غیرخطی بود. رویکرد آنها استفاده از یک معماری ساده شامل فقط یک گره خروجی بود. خروجی یک شبکه عصبی تک‌لایه پنهان با H گره پنهان و یک بردار ورودی از متغیرهای کمکی داده شده به‌صورت $x_i = (x_{i\circ}, \ldots, x_{ip})^T$ است. وزن W_{ph} مرتبط با ورودی p و گره پنهان $h(h = ۱, \ldots, H)$ است. حاصل ضرب مقادیر ورودی و وزن‌ها به تابع فعال‌ساز f داده می‌شود. به طور کلی، یک شبکه عصبی دارای یک گره پنهان خاص (گره ۰) است که به گره‌های خروجی متصل است اما به گره‌های ورودی متصل نیست که نقش آن مشابه عرض از مبدا در رگرسیون خطی است. با این حال، در مدل مخاطرات متناسب بقا نیازی به آن نیست زیرا اثر آن در تابع خطر پایه گنجانده می‌شود. بردار وزن‌های مرتبط با ورودی‌های گره پنهان h به‌صورت $W_h = (W_{\circ h}, \ldots, W_{ph}^T)$ نشان داده می‌شود. ورودی گره پنهان h برای حالت i یک تصویر خطی $(w^T{}_h x_i)$ است. خروجی گره پنهان h همان $f(w^T{}_h x_i)$ است که در آن f تابع فعال‌سازی است. توابع فعال‌ساز متنوعی در شبکه عصبی مورد استفاده قرار می‌گیرد که از متداول‌ترین آنها می‌توان به تابع لجستیک، تابع خطی و تابع مماس هذلولی اشاره کرد (دستراس، ۲۰۲۳).

با در نظر گرفتن وزن‌های $W_{\circ h}$ مرتبط با ورودی واحد $\mathbf{x}_{i\circ}$ به عنوان «بایاس»، خروجی شبکه به صورت مجموع وزنی از خروجی گره‌های پنهان با وزن‌های $(\alpha_\circ, \ldots, \alpha_H)$ و گره‌های پنهان H برای یک بردار ورودی داده شده $\mathbf{x}_i$ به صورت

$$g(\mathbf{x}_i, \theta) = \alpha_\circ + \sum_{h=۱}^{H} \alpha_h f(\mathbf{w}_h^T \mathbf{x}_i) = \alpha_\circ + \sum_{h=۱}^{H} \frac{\alpha_h}{۱ + \exp(-\mathbf{w}_h^T \mathbf{x}_i)}$$

است، که در آن تعداد کل پارامترها برابر $m = (H + ۱) + (P + ۱)H$ و با

$$\theta = (w_{\circ ۱}, w_{۱۱}, \ldots, w_{p۱}, w_{۱۲}, \ldots, w_{p۲}, \ldots, w_{\circ H}, w_{۱H}, \ldots, w_{pH}, \alpha_\circ, \ldots, \alpha_H)^T$$

نشان داده شده و $\mathbf{g}(x, \theta) = [\mathbf{g}(x_۱, \theta), \ldots, \mathbf{g}(x_n, \theta)]^T$ بردار خروجی است. در حقیقت نقطه قوت کار فاراگی و سیمون (۱۹۹۵) این بود که توانستند معماری شبکه‌های عصبی را با مدل کاکس ترکیب نمایند تا از اطلاعات زمان بقا به انضمام مشاهدات سانسور شده برای مدل‌سازی روابط غیرخطی استفاده کنند. از آن زمان، علاقه فزاینده‌ای به ترکیب معماری شبکه‌های عصبی در تحلیل بقا به وجود آمده است. با توجه به آنکه تابع خطر به زمان t و بردار متغیرهای کمکی $\mathbf{x}_i$ وابسته است و به صورت $h(t, \mathbf{x}_i) = h_\circ(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i)$ است. بردار پارامترهای مدل با ماکسیمم‌سازی تابع درستنمایی جزئی کاکس (کاکس، ۱۹۷۵) ، یعنی $L_c(\boldsymbol{\beta}) = \prod_{i=۱}^{k} \frac{\exp(\mathbf{x}_{(i)}^T \boldsymbol{\beta})}{\sum_{j \in R(t_{(i)})} \exp(\mathbf{x}_j^T \boldsymbol{\beta})}$، با روش نیوتن–رافسون و بر اساس مشاهدات بدون سانسور برآورد می‌شود. با جایگزینی تابع خطی $\mathbf{x}_{(i)}^T \boldsymbol{\beta}$ در مدل کاکس با خروجی‌های $g(\mathbf{x}_i, \theta)$ یک شبکه عصبی، مدل مخاطرات متناسب به صورت $h(t, \mathbf{x}_i) =$

$h_۰(t)\exp(g(\mathbf{x}_i,\theta))$ بازنویسی می‌شود و بر اساس آن تابع درستنمایی جزئی نیز به صورت

$$L_c(\theta)=\prod_{i=۱}^{k}\frac{\exp\left\{\sum_{h=۱}^{H}\alpha_h/(۱+\exp(-\mathbf{w}_h^T\mathbf{x}_i))\right\}}{\sum_{j\in R(t_{(i)})}\exp\left\{\sum_{h=۱}^{H}\alpha_h/(۱+\exp(-\mathbf{w}_h^T\mathbf{x}_i))\right\}}$$

قابل محاسبه است. مدل شبکه عصبی در حالت مخاطرات متناسب، فاقد مقدار عرض از مبدا یعنی $\alpha_۰$ است. همچنین، هنگام جایگزینی $x_{(i)}^T\beta$ با خروجی $g(x_i,\theta)$ شبکه، تابع فعال‌سازکه معمولاً به خروجی شبکه اعمال می‌شود حذف شده، بطوریکه خروجی $x_{(i)}^T\beta$ روی کل مجموعه اعداد حقیقی است و به بازه $(۰,۱)$ محدود نمی‌شود. در نهایت با استفاده از روش نیوتن–رافسون می‌توان لگاریتم درستنمایی جزئی را به حداکثر رساند و پارامترهای (وزن‌های) بهینه شبکه عصبی را به دست آورد (وانگ و همکاران، ۲۰۲۳). برآوردهای ماکسیمم درستنمایی پارامترهای شبکه عصبی باروش نیوتن–رافسون و استفاده از توابع احتمال و ماکسیمم احتمال برای برآورد پارامترهای شبکه، باعث شده تا بتوان از آزمون‌ها و روش‌های آماری برای ارزیابی شبکه استفاده کرد.

۵ مطالعه شبیه‌سازی

در این بخش یک مطالعه شبیه‌سازی با هدف مقایسه عملکرد مدل‌های معرفی شده در دو سناریوی مختلف انجام شده است. با توجه به پیچیدگی شبیه‌سازی داده‌های بقا، روش‌هایی که بر مبنای آنها می‌توان داده‌های بقا را شبیه‌سازی کرد، عمدتاً زمان‌بر هستند. برای سهولت، در این مطالعه از روش مترجم (۱۴۰۰) بر مبنای تابع توزیع خطر تجمعی استفاده شده است. بر این اساس در سناریوی اول یک مجموعه ۱۰۰۰ تایی داده بقا با ۳۰ متغیر پیش‌بین تولید شده است که در آن متغیرهای تبیینی شامل سه دسته $(X_۱$ تا $X_{۱۰})$، $(X_{۱۱}$ تا $X_{۲۰})$ و $(X_{۲۱}$ تا $X_{۳۰})$ می‌باشند. دسته اول از توزیع برنولی با پارامترهای $p=۰/۰۵, ۰/۱, ..., ۰/۴۵, ۰/۵۰$، دسته دوم از توزیع پواسن با با پارامترهای $\mu=۰/۵, ۱, ..., ۴/۵, ۵$ و دسته سوم از توزیع نمایی با پارامترهای $\theta=۰/۳, ۰/۶, ..., ۳$ تولید شده‌اند. تابع خطر پایه از توزیع وایبول با پارامترهای شکل و مقیاس به ترتیب برابر با ۰/۵ و ۲ و در تابع توزیع خطر تجمعی بردار متغیرهای تبیینی X با توان دوم در عبارت $\exp(\beta' X^۲)$ به عنوان رابطه غیر خطی در نظر گرفته شده‌اند. در این شبیه‌سازی سانسور راست ۱۰ درصدی ناآگاهی بخش بر داده‌ها اعمال شده است یعنی به‌صورت کاملا تصادفی به ۱۰ درصد از داده‌ها وضعیت صفر یعنی سانسور شده نسبت داده شده است این کار با استفاده از شبیه‌سازی داده از یک توزیع برنولی با پارامتر $p=۰/۹$ انجام شده است. پس از پیاده سازی مدل‌ها برای ارزیابی عملکرد از شاخص‌های C-index و B-score استفاده شد. بر این اساس نتایج شبیه‌سازی در جدول ۱ نشان می‌دهد که جنگل بقای تصادفی با مقدار C-index برابر با ۰/۸۶۶۸ و شاخص B-score برابر با ۰/۱۲۸۱ عملکرد مطلوب‌تری نسبت به سایر روش‌ها دارد. به طور کلی، مدل کاکس لاسو و شبکه عصبی نیز در این بخش عملکردی نزدیک به هم داشته‌اند. در سناریوی دوم ۱۰۰۰ داده بقا با تابع خطر پایه نمایی با پارامتر ۱ و بردار متغیرهای تبیینی X به‌صورت خطی در عبارت $\exp(\beta' X)$ تابع

جدول ۱. جدول مقایسه عملکرد مدل‌ها برای سناریوی اول شبیه‌سازی

مدل	C-index	B-score
کاکس	۰٫۷۱۰۳	۰٫۲۵۱۷
کاکس ریج	۰٫۷۵۱۴	۰٫۲۲۱۳
کاکس لاسو	۰٫۸۹۱۸	۰٫۱۶۲۱
جنگل بقای تصادفی	۰٫۹۶۶۸	۰٫۱۲۸۱
شبکه عصبی	۰٫۸۷۳۶	۰٫۱۷۰۲

توزیع خطر تجمعی لحاظ شده و در آن تنها سه متغیر مستقل به عنوان متغیرهای پیش‌بین در نظر گرفته شده است. متغیر پیش‌بینی کننده اول (X_1)، از توزیع نرمال با میانگین ۰ و واریانس ۱، متغیر پیش‌بینی کننده دوم (X_2) از توزیع یکنواخت در بازه $[0, 1]$ و متغیر پیش‌بینی کننده سوم (X_3) از توزیع نمایی با پارامتر ۰٫۵ تولید شد. نتایج این قسمت از شبیه‌سازی در جدول ۳ نشان می‌دهد که در قسمت دوم از شبیه‌سازی تفاوت قابل توجهی در شاخص‌های محاسبه شده بین مدل‌ها وجود ندارد و همه مدل‌ها عملکردی تقریباً مشابه دارند. این نتیجه نشان می‌دهد در شرایطی

جدول ۲. جدول مقایسه عملکرد مدل‌ها برای سناریوی دوم شبیه‌سازی

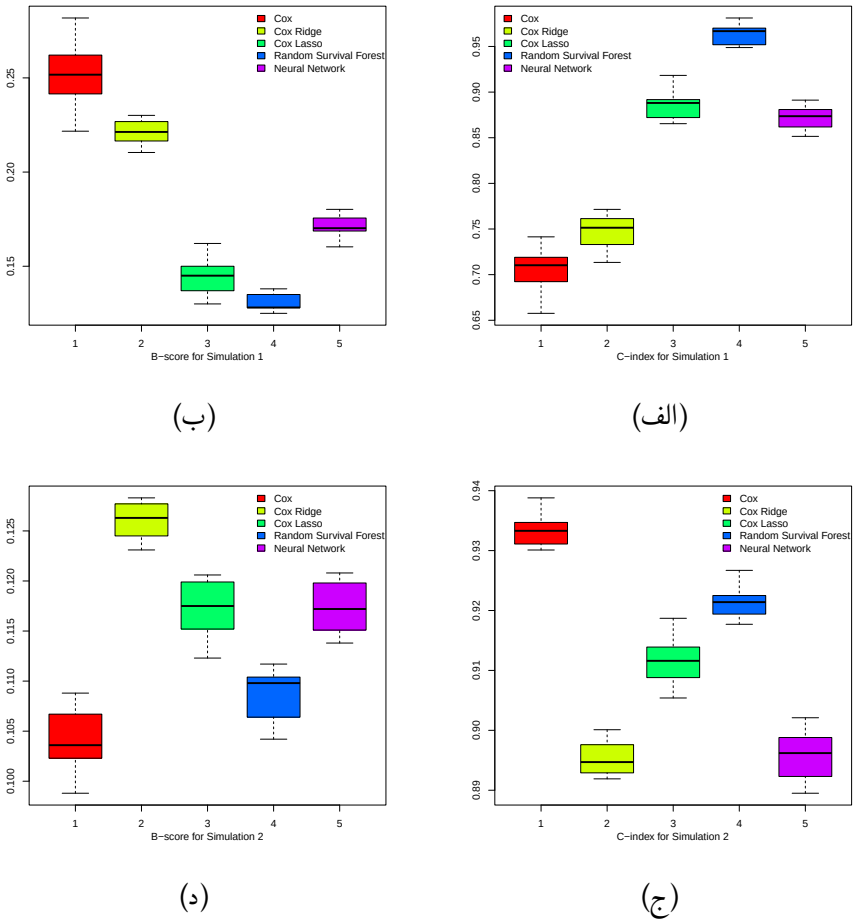
مدل	C-index	B-score
کاکس	۰٫۹۳۳۳	۰٫۱۰۳۶
کاکس ریج	۰٫۸۹۴۷	۰٫۱۲۶۳
کاکس لاسو	۰٫۹۱۱۶	۰٫۱۱۷۵
جنگل بقای تصادفی	۰٫۹۲۱۴	۰٫۱۰۹۸
شبکه عصبی	۰٫۸۹۶۲	۰٫۱۱۷۲

که تعداد متغیرهای تبیینی کم و بر اساس یک رابطه خطی در مدل تعریف شوند مدل کاکس عملکردی مشابه با سایر مدل‌ها خواهد داشت. اما در صورت وجود روابط غیرخطی بین متغیرهای تبیینی و زمان بقا استفاده از آن با دقت کمتری همراه خواهد بود و طبیعتاً استفاده از مدل‌هایی مانند شبکه‌های عصبی و جنگل بقای تصادفی بیشتر مورد توجه خواهد بود.

در هر دو سناریوی شبیه‌سازی به منظور اعتبارسنجی داخلی مدل‌ها، از روش بوت‌استرپ استفاده شد که در همه مدل‌ها بر اساس $B = ۱۰۰$ نمونه بوت‌استرپ تصادفی با اندازه‌ی مشابه داده اصلی، تنظیم شده است. سپس نتایج برای ارزیابی عملکرد در مجموعه داده بوت‌استرپ و همچنین مجموعه داده کامل استفاده شده است. این روش موجب می‌شود برآوردی دقیق‌تر از عملکرد مدل‌ها حاصل شود و حداکثر محدوده عملکرد مورد انتظار به دست آید. نتایج به دست آمده برای هر دو بخش شبیه‌سازی در شکل ۱ نشان داده شده است.

۶ تحلیل داده‌های بیماری آترواسکلروز

آترواسکلروز یک بیماری مزمن است که عمدتاً به دلیل تجمع پلاکت‌ها و چربی‌ها در دیواره رگ‌ها رخ می‌دهد. این تجمعات می‌توانند به تدریج باعث تشکیل پلاک‌های آتروماتوز شوند که از موادی مانند کلسترول، تری‌گلیسرید و



شکل ۱. نمودارهای الف و ب مقادیر C−index و B−score برای سناریوی اول شبیه‌سازی و نمودارهای ج و د مقادیر C−index و B−score برای قسمت دوم شبیه‌سازی

کلسیم تشکیل شده‌اند. زمانی که پلاکت‌ها و چربی‌ها در دیواره رگ‌ها تجمع می‌کنند، این پلاک‌ها به طور تدریجی رشد کرده و سخت می‌شوند. این فرآیند به عنوان آتروماتوز[1] شناخته می‌شود و می‌تواند باعث باریک شدن شریان‌ها و کاهش جریان خون به اعضاء و اندام‌های بدن شود. در نتیجه، اکسیژن و مواد مغذی کافی به سلول‌ها نمی‌رسد و ممکن است باعث آسیب و نارسایی اعضای بدن شود. داده‌های این مطالعه شامل ۳۸۷۳ نمونه از زمان بقای بیماران مبتلا به آترواسکلروز بر حسب روز با سانسور راست حدود ۸۸ درصدی و ۲۷ مولفه بالینی به عنوان متغیرهای تبیینی

[1] Atheromatous

است که پس از تمیزسازی اولیه مورد تحلیل قرار گرفت.

رگرسیون کاکس مبتنی بر تاوان: در اولین قدم مدل مخاطرات متناسب کاکس بر روی مجموعه داده‌های بیماران آترواسکلروز پیاده‌سازی شده است سپس از روش‌های کاکس لاسو و ریج در برازش به داده‌ها استفاده است. تفاوت اصلی بین این دو روش در نوع تاوان استفاده شده است. در روش کاکس لاسو، از تاوان L_1 استفاده می‌شود که به عنوان تاوان لاسو نیز شناخته می‌شود و سبب می‌شود تا برخی از ضرایب برابر با صفر شوند و متغیرهای مهم و کلیدی ترجیحاً با ضرایب غیر صفر ماندگار شوند. با این روش، متغیرهای غیرضروری حذف می‌شوند و مدلی ساده‌تر و قابل تفسیرتری به دست آید. در مقابل، در روش کاکس ریج، از تاوان L_2 استفاده می‌شود این امر موجب کاهش اندازه ضرایب شده و اهمیت برابری به متغیرهای مختلف داده می‌شود، اما به طور کلی ضرایب غیر صفر ماندگار هستند. به بیان دیگر تفاوت اساسی بین مدل‌های کاکس لاسو و ریج در تعداد متغیرهای باقیمانده در مدل است. کاکس لاسو تمایل به حذف متغیرهای غیرضروری دارد و مدل ساده‌تری را ترجیح می‌دهد، در حالی که کاکس ریج

جدول ۳. نتایج حاصل از برازش مدل کاکس بر داده‌های آترواسکلروز

نام متغیر	ضریب	خطر نسبی
سن	۰٫۰۳۰۳	۱٫۰۳۰۷
آسیب مغزی	۰٫۲۷۳۸	۱٫۳۱۵۰
انسداد شریان قلب	۰٫۳۵۴۸	۱٫۴۲۵۹
آنوریسم آئورت شکمی	۰٫۷۰۳۴	۲٫۰۲۰۶
عوامل محیطی	۰٫۲۷۲۸	۱٫۳۱۳۵
پاک‌کنندگی کراتینین	۰٫۰۰۲۱	۱٫۰۰۲۱
ضخامت مدیا اینتیما	۰٫۴۱۹۶	۱٫۵۲۱۴
آلبومین	۰٫۳۳۶۸	۱٫۴۰۰۵
مقدار مصرف سیگار	۰٫۰۰۵۱	۱٫۰۰۵۱

تمایل به حفظ متغیرها را دارد و تغییرات کوچکتری در ضرایب ایجاد می‌کند توجه شود که در مدل کاکس لاسو تعداد متغیرها ممکن است کاهش یابد و تنها متغیرهای مهم باقی بمانند، در حالی که در مدل کاکس ریج تعداد متغیرها بیشتری اما با تاثیر کمتر در مدل باقی می‌مانند. نتایج برازش مدل‌های کاکس، کاکس ریج و کاکس لاسو در جداول ۳ تا ۴ آمده است.

جدول ۴. نتایج برازش مدل کاکس لاسو بر داده‌های آترواسکلروز

نام متغیر	ضریب	خطر نسبی
سن	۰٫۰۲۷۴	۱٫۰۲۷۸
انسداد شریان قلب	۰٫۳۵۲۸	۱٫۴۲۳۰
آنوریسم آئورت شکمی	۰٫۷۶۱۰	۲٫۱۴۰۴
پاک‌کنندگی کراتینین	۰٫۰۰۲۰	۱٫۰۰۲۰
ضخامت مدیا اینتیما	۰٫۴۶۴۷	۱٫۵۹۱۵
آلبومین	۰٫۳۵۳۴	۱٫۴۲۳۹

مدل کاکس لاسو متغیرهای سن، انسداد شریان قلب، آنوریسم آئورت شکمی، ضخامت مدیا اینتیما، پاک‌کنندگی کراتینین و آلبومین به عنوان متغیرهای تاثیرگذار در مدل حفظ نموده است. مدل کاکس ریج با تمایل به حفظ متغیرهای

جدول ۵. نتایج برازش مدل کاکس ریج بر داده‌های آترواسکلروز

نام متغیر	ضریب	خطر نسبی
جنسیت	−۰٫۱۴۷۶	۰٫۸۶۲۸
سن	۰٫۰۳۰۳	۱٫۰۳۰۸
آسیب مغزی	۰٫۲۷۳۸	۱٫۳۱۵۰
انسداد شریان قلب	۰٫۳۵۴۸	۱٫۴۲۵۹
آنوریسم آئورت شکمی	۰٫۷۰۳۴	۲٫۰۲۰۶
عوامل محیطی	۰٫۲۷۲۸	۱٫۳۱۳۶
پاک‌کنندگی کراتینین	۰٫۰۰۲۱	۱٫۰۰۲۱
ضخامت مدیا اینتیما	۰٫۴۱۹۶	۱٫۵۲۱۳
آلبومین	۰٫۳۳۶۸	۱٫۴۰۰۴
مقدار مصرف سیگار	۰٫۰۰۵۱	۱٫۰۰۵۱

تبیینی بیشتر در مدل عمل می‌کند در اینجا، مدل کاکس ریج متغیرهای جنسیت، سن، آسیب مغزی، انسداد شریان قلب، آنوریسم آئورت شکمی، عوامل محیطی، پاک‌کنندگی کراتینین، ضخامت مدیا اینتیما، آلبومین و مقدار مصرف سیگار را در مدل حفظ نموده است. به منظور سهولت در ارائه نتایج در مدل‌های فوق، متغیرهایی مستقل غیرمعنی‌دار با مقدار ضریب کمتر از ۰٫۰۰۱ حذف شده‌اند.

جنگل بقای تصادفی: در پیاده‌سازی ساخت جنگل بقا یک پارامتر تنظیم[1] (CP) که پیچیدگی مدل را اندازه‌گیری می‌کند، متغیرهای کم اهمیت را حذف می‌کند. در نتیجه، دنباله‌ای تو در تو از زیر درخت‌ها تولید می‌شود. اگر یک تقسیم نتواند کاهش کل ناسازگاری را به ضریب CP برساند، حذف می‌شود این روش، نه تنها یک راه برای جلوگیری از بیش‌برازش است، بلکه در زمان محاسباتی نیز صرفه‌جویی می‌کند. اگر پارامتر در یک مقدار بزرگتر تنظیم شود، درخت حاصل کوچکتر خواهد بود، به عبارت دیگر، کمترین تقسیم‌بندی‌ها بررسی خواهند شد. مقادیر CP برای انتخاب بهینه‌ترین مدل جنگل بقای تصادفی به کمک روش اعتبارسنجی متقابل در نظر گرفته می‌شوند. در این روش، مدل با مقادیر مختلف CP آموزش داده می‌شود و سپس عملکرد مدل بر روی داده‌های اعتبارسنجی ارزیابی می‌شود. در این مثال مقدار بهینه CP برابر ۰٫۰۴ بهترین عملکرد را بر روی داده‌های اعتبارسنجی ارائه می‌دهد، به عنوان مقدار نهایی به دست آمده است. در نهایت بر اساس ترکیب ۱۰۰ درخت ۶ متغیر سن، هموسیستئین، ضخامت مدیا اینتیما، کلسترول لیپوپروتئین با چگالی بالا، کلسترول کل و فشار خون دیاستولیک به عنوان متغیرهای پر اهمیت تشخیص داده شد.

شبکه‌های عصبی: در این بخش، داده‌ها با استفاده از رویکرد مبتنی بر شبکه‌های عصبی مورد بررسی و تحلیل قرار گرفته است. برای برازش شبکه از تابع ELMCox در نرم افزار R استفاده شده است، که ترکیبی از

[1] Control Parameter

یادگیری ماشین در قالب شبکه عصبی و مدل رگرسیون کاکس است، در نهایت این شبکه به وزن‌های منجر می‌شود که برای پیش‌بینی خطرها در مجموعه داده استفاده می‌شوند. در این الگوریتم از روش تخمین فاصله کاکس[1] در

جدول ۶. وزن‌های مربوط به متغیرهای موثر در شبکه عصبی

نام متغیر	وزن
سن	۰٫۲۵۱۳
جنسیت	۰٫۱۸۰۸
مقدار مصرف سیگار	۰٫۱۴۳۹
آسیب مغزی	۰٫۱۱۹۸
کلسترول کل	۰٫۱۰۲۱
دیابت	۰٫۰۹۵۴
مصرف الکل	۰٫۰۸۲۳
عوامل محیطی	۰٫۰۷۹۸
آلبومین	۰٫۰۵۶۸
تنگی شریان کاروتید	۰٫۰۳۱۱
هوموسیستئین	۰٫۰۲۹۲

برآورد پارامترهای مدل استفاده می‌شود که در آن مقادیر ویژگی‌ها و وزن‌های شبکه با استفاده از تابع هدف و الگوریتم بهینه‌سازی بهینه می‌شوند. هدف این بهینه‌سازی، کمینه کردن خطا و افزایش دقت پیش‌بینی‌ها است. در اینجا شبکه مورد نظر دارای ساختاری به‌صورت ۲۷ − ۱۰ − ۱ است که در آن ۷۰ درصد داده‌ها برای آموزش شبکه استفاده شد و شبکه با ۳۰ درصد دیگر از داده‌ها تست شد. این اطلاعان نشان می‌دهد که شبکه مورد نظر دارای سه لایه است که شامل لایه ورودی با ۲۷ متغیر تبیینی، لایه میانی (پنهان) با ۱۰ واحد و یک واحد در لایه خروجی است. هر واحد در لایه میانی با واحدهای لایه قبلی متصل شده است و هر واحد در لایه خروجی با واحدهای لایه قبلی وزن دارد. در این مثال آموزش شبکه پس از ۱۰۰ بار متوقف شده است با استفاده از این شبکه، با ارائه ویژگی‌های ورودی، می‌توان پیش‌بینی برای زمان رویداد را انجام داد. مقدار وزن‌های شبکه عصبی در جدول ۶ نشان داده شده است. این جدول با در نظر گرفتن آستانه ۰٫۰۱ برای وزن‌های ارائه شده است. در نهایت با توجه به وزن‌های بدست آمده متغیرهای سن، جنس، مصرف سیگار، آسیب مغزی، کلسترول کل، دیابت، مصرف الکل، عوامل محیطی، تنگی شریان کاروتید و هوموسیستئین به عنوان متغیرهای مهم شناسایی شدند.

ارزیابی عملکرد مدل‌ها: برای ارزیابی کلی عملکرد مدل‌ها، از شاخص‌های C−index و B−score استفاده شده است. به منظور اعتبارسنجی داخلی مدل‌ها، از روش بوت‌استرپ استفاده شد که در همه مدل‌ها بر اساس $B = ۲۰۰$ نمونه بوت‌استرپ تصادفی با اندازه‌ی مشابه داده اصلی، تنظیم شده است. به منظور تعیین عوامل خطر موثر بر بقای بیماران آترواسکلروز است برخی متغیرهای مهم که در اکثر مدل‌های برازش یافته مختلف حضور دارند جدول ۷ این متغیرهای مهم و تعداد کل حضور آنها در مدل‌های مختلف را نمایش می‌دهد. برای محاسبه ترتیب اهمیت متغیرها اگر متغیر در مدل حضور داشته است به آن امتیاز ۱ و در غیر اینصورت امتیاز ۰ تعلق گرفته است

[1] Cox Distance Estimation

در نهایت جمع امتیازات در ستون آخر نشان می‌دهد که متغیر مورد نظر در چند مدل حضور داشته است. نتایج نشان می‌دهد که عامل سن با جمع کل امتیاز ۵ در همه مدل‌ها حضور داشته و به عنوان موثرترین عامل در بیماری تصلب شرایین شناخته می‌شود.

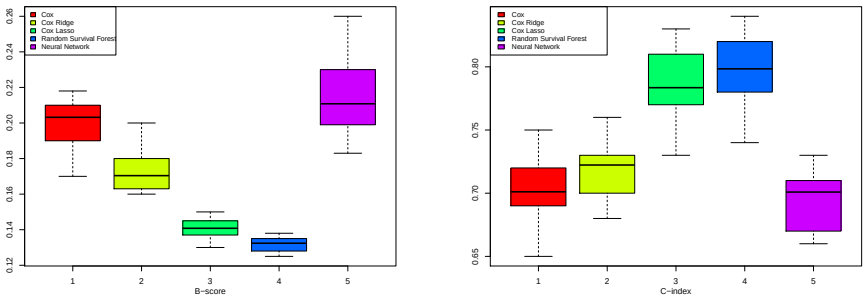
جدول ۷. مجموع امتیازات هر متغیر در مدل‌های مختلف

نام متغیر/مدل	کاکس	لاسو	ریج	شبکه عصبی	جنگل بقای تصادفی	جمع کل
سن	۱	۱	۱	۱	۱	۵
آلبومین	۱	۰	۱	۱	۱	۴
ضخامت مدیا اینتیما	۱	۱	۱	۰	۱	۴
آسیب مغزی	۱	۰	۱	۱	۰	۳
انسداد شریان قلب	۱	۱	۱	۰	۰	۳
پاک‌کنندگی کراتینین	۱	۱	۱	۰	۰	۳
عوامل محیطی	۱	۰	۱	۱	۰	۳
مقدار مصرف سیگار	۱	۱	۱	۰	۰	۳

جدول ۸. عملکرد مدل‌های مختلف بر اساس دو شاخص C-index و B-score

مدل	C-index	B-score
کاکس	۰/۷۰۱۱	۰/۲۰۳۲
کاکس ریج	۰/۷۲۲۳	۰/۱۷۰۴
کاکس لاسو	۰/۷۸۳۵	۰/۱۴۰۸
جنگل تصادفی	۰/۷۹۸۴	۰/۱۳۲۴
شبکه عصبی	۰/۷۰۰۹	۰/۲۱۰۸

در جدول ۸، نحوه عملکرد مدل‌های مختلف بر اساس دو شاخص C-index و B-score نشان داده شده است. این مقادیر بر اساس میانگین اندازه‌گیری‌های عملکرد مدل‌های بر اساس نمونه‌های بوت‌استرپ در هر مدل حاصل شده است. روش بوت استرپ همچنین امکان بررسی واریانس اندازه‌های عملکرد را از طریق ۲۰۰ نمونه تصادفی فراهم می‌کند. این امر با استفاده از نمودارهای جعبه‌ای نشان داده شده در شکل ۲ به ترتیب برای معیارهای C-index و B-score به تصویر کشیده شده است. بر اساس نتایج جدول ۸ و مقایسه مقادیر شاخص C-index می‌توان نتیجه گرفت که مدل جنگل بقای تصادفی با مقدار ۰/۷۹۸۴، بالاترین مقدار شاخص را داراست و به عنوان مدلی با بهترین عملکرد معرفی می‌شود. همچنین، مدل کاکس لاسو با مقدار ۰/۷۸۳۵ نیز نزدیک به مقدار جنگل بقای تصادفی قرار دارد و نشان می‌دهد که عملکرد خوبی دارد در رتبه‌های بعدی، مدل‌های کاکس و کاکس ریج با مقادیر ۰/۷۰۱۱ و ۰/۷۲۲۳ قرار دارند. اما مدل شبکه عصبی با مقدار ۰/۷۰۰۹، پایین‌ترین شاخص را داراست و عملکرد ضعیف‌تری نسبت به سایر مدل‌ها دارد. در مورد شاخص B-score، مدل جنگل بقای تصادفی با مقدار ۰/۱۳۲۴، کمترین مقدار B-score را داراست که نشان‌دهنده دقت بالای آن در پیش‌بینی است. همچنین، مدل کاکس لاسو با مقدار ۰/۱۴۰۸ در رتبه دوم قرار دارد و دقت نسبتاً بالایی دارد. مدل‌های کاکس و کاکس ریج با مقادیر ۰/۲۰۳۲ و ۰/۱۷۰۴ به ترتیب از عملکرد متوسطی برخوردار بوده‌اند. اما مدل شبکه عصبی با مقدار ۰/۲۱۰۸ بالاترین مقدار B-score را داراست و دارای عملکرد ضعیفی در میان مدل‌های مورد مطالعه است. به طور کلی، مدل جنگل بقای



شکل ۲. نمودار الف مقادیر بوت استرپ شده C−index و نمودار ب مقادیر بوت استرپ شده B−score

تصادفی با داشتن مقادیر مطلوبتر در شاخص‌های ارزیابی، به عنوان مدلی با عملکرد بهتر معرفی می‌شود. در عین حال، مدل کاکس لاسو نیز با نتایج نزدیک به مدل جنگل بقای تصادفی یک گزینه قوی دیگر است.

بحث و نتیجه‌گیری

نتایج شبیه‌سازی‌های انجام شده در این مقاله نشان داد که انتخاب بین روش‌های یادگیری آماری مانند جنگل بقای تصادفی و شبکه‌های عصبی در مقابل روش‌های کلاسیک مانند مدل کاکس در تحلیل بقا، به ویژگی‌های خاص نهفته در داده‌ها وابسته است. از سوی دیگر، مدل‌های جنگل بقای تصادفی و شبکه‌های عصبی انعطاف‌پذیرتر هستند و می‌توانند روابط پیچیده‌تر بین متغیرهای مستقل و زمان بقا را تبیین نمایند. لازم به ذکر است استفاده از مدل‌های جنگل بقای تصادفی و شبکه عصبی به خصوص زمانی مفید است که تعاملات و روابط غیرخطی بین متغیرهای تبیینی با زمان بقا وجود داشته باشد. مثال کاربردی مورد بررسی در این مطالعه نیز نشان داد که روش‌های یادگیری آماری قادر به مدیریت تعداد زیاد متغیر به دلیل تنظیم یا انتخاب داخلی متغیرها هستند. مقایسه معیارهای عملکرد مدل‌ها نشان داد که مدل‌های کاکس لاسو و جنگل‌های بقای تصادفی عملکرد بهتری در تحلیل داده‌های زمان بقای بیماران آترواسکلروز داشته‌اند.

تقدیر و تشکر

نویسندگان مقاله از اعضای محترم هیئت تحریریه و همچنین پیشنهادها و نظرات ارزشمند داوران و ویراستار محترم که موجب ارتقاء سطح آن گردید کمال تشکر و قدردانی را دارند.

مراجع

مترجم، ک. (۱۴۰۰)، معرفی یک مدل بقای فضایی با اثرات تصادفی چوله گاوسی و کاربرد آن در تحلیل داده‌های بیماری کووید-۱۹، مجله علوم آماری، ۱۵، ۵۶۷-۵۹۰

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**, 5-32.

Cox, R. D. (1975), Partial Likelihood, *Biometrika*, **62**, 269–276.

Davis, R. B. and Anderson, J. R. (1989), Exponential Survival Trees, *Statistics in Medicine*, **8**, 947–961.

Destras, O., Le Beux, S., De Magalhães, F. G., and Nicolescu, G. (2023), Survey on Activation Functions for Optical Neural Networks, *ACM Computing Surveys*, **56**(2), 1-30.

Faraggi, D. and Simon R. (1995), A Neural Network Model for Survival Data, *Statistics in Medicine*, **31**, 73-82.

Harrell, F. E., Cali, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982), Evaluating the Yield of Medical Tests, *Jama*, **247**, 2543-2546.

Harrell, F. E. (2001), *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer, New York.

Hastie, T. and Tibshirani, R. (1990), Generalized Additive Models, *Statistical Sciences*, **3**, 297-318.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, **2**, 1-758.

Henderson, R. (1995), Problems and Prediction in Survival Data Analysis, *Statistics in Medicine*, **14**, 161-184.

Ishwaran, H., Kogalur, U. B., Blackstone, E.H. and Lauer, M. S. (2008), Random Survival Forests, *The Annals of Applied Statistics*, **2**, 841–60.

LeBlanc, M. and Crowley J. (1993), Survival Trees by Goodness of Split, *Journal of the American Statistical Association*, **88**, 457–467.

Liu, L., Yang, F., Fan, Y., Kao, C., Wang, F., Yu, L., and Yu, Z. (2023), An Improved Training Algorithm Based on Ensemble Penalized Cox Regression for Predicting Absolute Cancer Risk, *China CDC Weekly*, **5**(9), 206.

Robins, J. M. and Rotnitzky, A. (1992), *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*, Springer, 297-331.

Shimokawa, A., Kawasaki, Y. and Miyaoka, E. (2015), Comparison of Splitting Methods on Survival Tree, *The International Journal of Biostatistics*, **1**, 175-188.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011), Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, *Journal of Statistical Software*, **39**, 1-13.

Smith, J. (2020),The impact of Technology on Data Analysis, *Journal of Data Science*, **8**(2), 123-135.

Smith, J., and Jones, A. (2020), Application of Statistical Learning Techniques in Survival Analysis, *Journal of Biostatistics*, **10**(3), 123-135.

Tibshirani, R. (1997), The Lasso Method for Variable Selection in the Cox Model, *Statistics in Medicine*, **16**, 385–395.

Utkin, L. V., Konstantinov, A. V., Chukanov, V. S., Kots, M. V., Ryabinin, M. A., and Meldo, A. A. (2019), A Weighted Random Survival Forest, *Knowledge-Based Systems*, **177**, 136-144.

Wang, D. and Gao, Z., (2023), Distributed Finite-Time Optimization Algorithms with a Modified Newton–Raphson Method, *Neurocomputing*, **536**, 73-79.