14(56), Summer 2024, 43-68
eISSN: 2470-6240 ISSN: 2252-004x
Article Type: Reserch Article
Received in revised form February 13, 2024
Accepted April 30, 2024
Received December 07, 2023
Published Online July 05, 2024

Educational Measurement
https://jem.atu.ac.ir/

ATU PRESS

Determining the Cut-off Score of Criterion-Referenced Tests Using Non-Parametric Estimation Methods of the Youden Index

Maryam Parsaeian — Ph.D. Candidate, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.E-mail: maryam.parsaeian@ut.ac.ir

Ebrahim Khodaie * — *Corresponding Author,*Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.Email : khodaie@ut.ac.ir

Balal Izanloo — Assistant Professor, Faculty of Psychology and Education, University of Kharazmi, Tehran, Iran.E-mail: b.ezanloo@gmail.com

Keyvan Salehi — Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. E-mail : keyvansalehi@ut.ac.ir

Sima NaghiZadeh — Assistant Professor, National Organization of Educational Testing (NOET), Tehran, Iran. E-mail : s_naghizadeh@yahoo.com

Abstract

The Youden index serves as a frequently utilized summary metric for Receiver Operator Characteristic (ROC) curves, which comprehensively measures the effectiveness of a criterion-referenced test while also unequivocally defining the cutoff score value for the test. The primary objective of this research is to compare and assess the efficacy of three empirical non-parametric estimation methods: kernel with Silverman's bandwidth method and kernel with maximum likelihood cross-validation bandwidth method in determining the value of the Youden index. This research also incorporates statistical measures such as bootstrap standard error (BSE), root mean square error (RMSE), square integrated error (ISE), and mean square integrated error (MISE) to evaluate the performance of these methods. The dataset used in this study consisted of scores from the different sections (grammar, listening, reading, and writing) of 461 participants who took the Tolimo language proficiency test, which was developed and administered by the National Organization for Educational Testing in Iran. The findings indicated that the kernel method with maximum likelihood cross-validation yielded a higher Youden index score. The following cutoff scores were achieved: 479 for the kernel methods and 465 for the empirical method. Based on the assessment of performance indices, kernel methods, particularly the one that incorporates the optimal bandwidth through maximum likelihood cross-validation, yielded highly reliable estimates. Thus, they provide dependable values for the Youden index and its corresponding cutoff score.

Keywords: Cut score; Youden index; Empirical nonparametric estimation; Kernel nonparametric estimation; Bootstrap standard error; ISE and MISE

Cite this Article: Parsaeian, M., Khodaie, E., Izanloo, B., Salehi, K. & NaghiZadeh, S. (2024). Determining the Cut-off Score of Criterion-Referenced Tests Using Non-Parametric Estimation Methods of the Youden Index Subject of Study: Tolimo Test. *Educational Measurement*, *14*(56), 43-68. https://doi.org/10.22054/jem.2024.77090.3508

© 2016 by Allameh Tabataba'i University Press
Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2024.77090.3508

1. Introduction

Youden's index is extensively employed in medical diagnostics and patient categorization. However, there is a conspicuous scarcity of research in the educational realm focused on ascertaining the optimal criterion-referenced assessment threshold. Non-parametric estimation techniques are frequently used for calculating the Youden index, necessitating an examination of three unique non-parametric estimation methods (including two bandwidth approaches) in this study to determine the Youden index for setting the cutoff score of the Tolimo test. The Tolimo test is a well-established and recognized English language test in Iran, designed and administered by the National Organization for Educational Testing in Iran to assess the language proficiency of students. Tolimo is an abbreviation that stands for "The test of Language by the Iranian Measurement Organization".

This study aspires to evaluate and compare three empirical non-parametric estimation techniques for determining the Youden index, an important metric derived from the Receiver Operating Characteristic (ROC) curve. The fundamental aim is to identify the most dependable method for calculating the cutoff score for the Tolimo test, as this determination will have a significant impact on the test's interpretation and have implications for decision-making and educational practices.

2. Literature Review

The literature review section of the manuscript focuses on the methodologies utilized in determining the cut score for criterion-referenced assessments. The article acknowledges the need for an efficient approach in determining the cut score and highlights the importance of the Youden index, which is a concise metric derived from the receiver operator characteristic (ROC) curve, in this process. The literature review lays the foundation for this investigation by highlighting the gap in existing research and the necessity of identifying a reliable method for setting cut scores in educational settings. Of particular importance is the focus on studies that employ nonparametric estimation methodologies.

According to the theoretical foundations and historical context of studies such as Fluss et al. (2005), Ewald (2006), Leeflang et al. (2008), and Hirschfeld and do Brasil (2014), the application of experimental optimization methods to estimate the Youden index often yields highly variable results due to the method's sensitivity to randomness in the

sample. Moreover, in situations where the data exhibit normal distribution or approximate normality, the application of the kernel density method with Silverman's bandwidth produces more consistent estimates of the Youden index compared to the experimental non-parametric estimation technique.

3. Methodology

The research methodology section of the article outlines the approach taken to compare and assess non-parametric estimation techniques. The focus of the investigation is on three specific methods: the kernel method with Silverman's bandwidth, the kernel method with maximum likelihood cross-validation bandwidth, and the empirical method. Various evaluation metrics, including bootstrap standard error (BSE), root mean square error (RMSE), square integrated error (ISE), and mean square integrated error (MISE), are utilized by the researchers to assess the efficacy of these methods. The significance of acquiring reliable estimates for the Youden index and the cut-off score in criterion-referenced assessments is highlighted in this section. It emphasizes the need for a robust methodology to ensure accurate outcomes. The research methodology aims to address these issues and provide insights into the best approach for determining the cut-off score.

4. Results

The findings revealed that the kernel method using Silverman's bandwidth produced a Youden index of 0.63 with a threshold of 479. The root mean square error for this approach was 0.715, with a bootstrap standard error of 0. As a result, the calculation of the Youden index through this approach showed a high level of consistency and precision. In contrast, the kernel method using MLCV produced a superior Youden index of 0.75 with the same threshold of 479. This suggests that the kernel method with MLCV offers a more accurate assessment of the Youden index compared to using Silverman's bandwidth, leading to a higher ability in differentiating between qualified and unqualified individuals. However, its bootstrap standard error was 0.039, indicating a small level of variability in the approximation. The empirical method generated a Youden index of 0.31 with a threshold of 465. The root mean square error for this

technique was 0.68, resulting in a similar bootstrap standard error of 0.017.

5. Conclusion

The findings of this study indicate that the choice of bandwidth determination method in relation to the kernel method plays a vital role in estimating the Youden index. When the data exhibit normality, selecting Silverman's bandwidth for the kernel method leads to consistent results; whereas, for non-normal data, it is recommended to use the maximum likelihood cross-validation bandwidth to compute and assess the Youden index in criterion-referenced assessments and establish the cut score. Nevertheless, further research is necessary to validate these findings and assess the effectiveness of these approaches across different domains. This can improve the accuracy of individual classification and reduce the incidence of inaccurate test outcomes. In addition, it is important to examine the distinct effects of setting the cut score on both individual and institutional outcomes, as this can help inform the decision-making process regarding the determination of the cut score. These efforts are significant in improving the assessment and decision-making framework within educational and institutional contexts.

Acknowledgments

The research from which this article is taken derives from the doctoral thesis in the measurement and evaluation field at Tehran University, titled "[Title of the Thesis]: Comparison of Benchmark Test Scores in Deep Learning Algorithms and Selected Studied Methods: Tolimo Test." This research was supported by Tehran University and the [country's] Education Assessment Organization. We are grateful for the Education Measurement Organization's cooperation in implementing this research, as well as the Education Measurement Quarterly for reviewing the article.

اندازه‌گیری تربیتی
https://jem.atu.ac.ir/

تعیین نمره برش آزمون‌های ملاک مرجع با استفاده از روش‌های برآورد ناپارامتری شاخص یودن موردمطالعه: آزمون تولیمو

مریم پارساییان — دانشجوی دکتری رشته سنجش و اندازه‌گیری، دانشگاه تهران، تهران، ایران. رایانامه: maryam.parsaeian@ut.ac.ir

ابراهیم خدایی * — نویسنده مسئول، دانشیار، دانشگاه تهران، تهران، ایران. رایانامه: khodaie@ut.ac.ir

بلال ایزانلو — استادیار گروه برنامه‌ریزی درسی، دانشگاه خوارزمی، تهران، ایران. رایانامه: b.ezanloo@gmail.com

کیوان صالحی — دانشیار، دانشگاه تهران، تهران، ایران. رایانامه: keyvansalehi@ut.ac.ir

سیما نقی‌زاده — استادیار، سازمان سنجش آموزش کشور، تهران، ایران. رایانامه: s_naghizadeh@yahoo.com

ISSN: 2252-004x
eISSN: 2470-6240
دوره ۱۴، شماره ۵۶، تابستان ۱۴۰۳، ۴۳-۶۸
نوع مقاله: مقاله پژوهشی
تاریخ دریافت: ۱۴۰۲/۰۹/۱۶؛ تاریخ بازنگری: ۱۴۰۲/۱۱/۲۴
تاریخ پذیرش: ۱۴۰۳/۰۲/۱۱؛ تاریخ انتشار: ۱۴۰۳/۰۴/۱۵

چکیده

شاخص یودن یک معیار خلاصه متداول برای منحنی ویژگی عملکرد (ROC) است که هم کارایی یک آزمون ملاک مرجع را می‌سنجد و هم مقدار نمره برش را برای آزمون مشخص می‌کند. این پژوهش با هدف مقایسه و ارزیابی سه روش برآورد ناپارامتری تجربی، هسته با روش پهنای باند سیلورمن و هسته با روش پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برای محاسبه مقدار شاخص یودن انجام شده است. در این پژوهش برای ارزیابی عملکرد از شاخص‌های خطای استاندارد بوت استرپ (BSE)، ریشه میانگین مربعات خطا (RMSE)، مربع خطای یکپارچه (ISE) و میانگین مربعات خطای یکپارچه (MISE) استفاده شده است. مجموعه داده‌های مورداستفاده در این مطالعه شامل نمرات بخش‌های مختلف (گرامر، شنیداری، درک مطلب و نوشتاری) ۴۶۱ شرکت‌کننده بود که در آزمون مهارت زبان تولیمو شرکت داشتند که توسط سازمان سنجش آموزش کشور تدوین و اجرا شد. نتایج نشان داد که روش هسته با اعتبارسنجی متقابل ماکسیمم درستنمایی دارای مقدار شاخص یودن بالاتری بود. نمرات برش به‌دست‌آمده برای روش‌های هسته ۴۷۹ و برای روش تجربی ۴۶۵ به دست آمد. با توجه به نتایج قابل‌قبول شاخص‌های ارزیابی، روش‌های هسته به‌ویژه با پهنای باند بهینه اعتبارسنجی متقابل ماکسیمم درستنمایی منجر به برآوردهای قابل‌اعتمادتری از شاخص یودن و نمره برش شد.

کلیدواژه‌ها: نمره برش، شاخص یودن، برآورد ناپارامتریک تجربی، برآورد ناپارامتریک هسته، خطای استاندارد بوت‌استرپ، ISE و MISE

استناد به این مقاله: پارساییان، مریم، خدایی، ابراهیم، ایزانلو، بلال، صالحی، کیوان، و نقی‌زاده، سیما. (۱۴۰۳). تعیین نمره برش آزمون‌های ملاک مرجع با استفاده از روش‌های برآورد ناپارامتری شاخص یودن، موردمطالعه: آزمون تولیمو. *فصلنامه اندازه‌گیری تربیتی*، ۱۴(۵۶)، ۴۳-۶۸. https://doi.org/10.22054/jem.2024.77090.3508

© ۲۰۱۶ دانشگاه علامه طباطبائی
ناشر: دانشگاه علامه طباطبائی

مقدمه

آزمون‌های ملاک مرجع اغلب نیاز به ایجاد یک مقدار برش برای طبقه‌بندی افراد به‌عنوان قبول یا مردود دارند. انتخاب این نمره برش به هزینه‌های مرتبط و پیامدهای طبقه‌بندی نادرست افراد بستگی دارد. علیرغم اینکه تفسیر یک آزمون با نتایج باینری سرراست و ساده است، تفسیر یک آزمون با نتایج پیوسته به‌صورت باینری به این سادگی نیست؛ زیرا در آزمونی با نتایج پیوسته (یا چندگانه)، هر مقدار آزمون را می‌توان به‌عنوان نمره برش در نظر گرفت. به‌عنوان‌مثال حداقل نمره قبولی در دروس کارشناسی، کارشناسی ارشد و دکتری در ایران به ترتیب برابر ۱۰، ۱۲ و ۱۴ است. حداقل نمره قبولی برای فرصت مطالعاتی دانشجویان دکتری در آزمون‌های $MSRT$، تولیمو و تافل به ترتیب برابر ۵۰، ۴۸۰ و ۶۰ است. درواقع با در نظر گرفتن این حداقل نمرات که همان نمرات برش هستند می‌توان دانشجویان را به دودسته قبول و رد طبقه‌بندی کرد (Thiele & Hirschfeld, 2020).

اکثر آزمون‌ها دارای مقیاس پیوسته‌ای از مقادیر هستند که یکی از این مقدارها را می‌توان به‌عنوان نقطه برش برای جداسازی افراد لایق و نالایق انتخاب کرد. بدیهی است نتایج نادرست آزمون‌ها مخصوصاً آزمون‌های سرنوشت‌ساز می‌تواند پیامدهای منفی زیادی هم برای داوطلبان و هم برای سازمان برگزارکننده داشته باشد (Dardick & Weiss, 2019). ازجمله پیامدهای داوطلبانی که به اشتباه رد شده‌اند (خطای منفی کاذب) می‌توان به از دست دادن فرصت شغلی، افسردگی، اضطراب و همچنین مشکلات مالی ازجمله ثبت‌نام برای آزمون مجدد و یا از دست دادن درآمد ناشی از پیشنهاد شغلی اشاره کرد. از سوی دیگر پذیرش داوطلبانی که صلاحیت قبولی در آزمون را ندارند (خطای مثبت کاذب) دارای عواقب منفی مانند کاهش کیفیت آموزش می‌شود چراکه نامزدهای فاقد صلاحیت که در آزمون پذیرفته می‌شوند، ممکن است نتوانند با دوره آموزشی هماهنگ شوند یا استانداردهای برنامه را برآورده کنند. این امر می‌تواند منجر به پایین آمدن استانداردهای تحصیلی و کاهش کیفیت آموزش شود که منجر به ایجاد فضای رقابتی کمتری می‌شود و موفقیت را برای نامزدهای واجد شرایط دشوارتر می‌کند. درنهایت پذیرش نامزدهای فاقد صلاحیت منجر به هدر رفتن منابع مالی و زمانی می‌شود. به‌عنوان‌مثال، نامزدهای فاقد صلاحیت ممکن است نیاز به گذراندن دوره‌های تقویتی یا آزمون مجدد داشته باشند که منجر به صرف هزینه و زمان می‌شود. به‌غیراز داوطلبان، نتایج اشتباه آزمون برای سازمان

برگزارکننده نیز پیامدهایی مانند از دست دادن اعتماد عمومی و هزینه‌های مالی ازجمله هزینه آزمون مجدد را در پی دارد. لذا لازم است نمره برش به‌درستی و با دقت تعیین شود.

شاخص یودن معیاری را برای انتخاب مقدار آستانه "بهینه" (c^*) ارائه می‌دهد که مقداری بین ۰ و ۱ می‌گیرد و به‌صورت زیر تعریف می‌شود:

$$J = \max_{c}\{Se(c) + Sp(c) - 1\}\ , \forall c \qquad (1)$$

طبق روش یودن، نقطه‌ای به‌عنوان نقطه برش انتخاب می‌شود که تابع یودن را ماکسیمم کند (Youden, 1950) که حداکثر فاصله یا اختلاف عمودی بین منحنی ROC و خط مورب یا شانس است (Schisterman et al., 2005). بنابراین طبق رابطه (۱) در روش یودن، نقطه برش بهینه c^* عبارت از نقطه c ی است که منجر به ماکسیمم مقدار تابع یودن به ازای همه مقادیر نقطه برش ممکن c موجود در مقادیر داده‌ها می‌شود.

مقدار شاخص یودن بین ۰ و ۱ است که جداسازی کامل توزیع مقادیر نشانگر برای جمعیت‌های قبول و مردود منجر به $J = 1$ می‌شود درحالی‌که همپوشانی کامل منجر به $J = 0$ می‌شود (Fluss et al., 2005). به‌عبارت‌دیگر مقدار شاخص یودن ۱ نشان‌دهنده یک آزمون کامل است، درحالی‌که مقدار شاخص یودن ۰ نشان‌دهنده یک آزمون بی‌فایده است که هیچ‌گونه اطلاعاتی در مورد داوطلبان نمی‌دهد؛ بنابراین درصورتی‌که مقدار شاخص یودن برابر صفر باشد می‌توان نتیجه گرفت که طبقه‌بندی‌کننده بهتر از شانس عمل نکرده است.

برای برآورد منحنی ROC از رویکردهای پارامتریک، نیمه‌پارامتریک و ناپارامتریک استفاده شده است که با توجه به رابطه شاخص یودن با منحنی ROC ، این رویکردها با شاخص یودن نیز مرتبط هستند. در برآورد پارامتری، به این فرض نیاز هست که داده‌ها از کدام خانواده چگالی تولید می‌شوند. اغلب این فرض بر اساس شواهد کم یا بدون شواهد است و اگر فرض موردنظر اشتباه باشد، نتیجه نادرستی حاصل می‌شود. ولی در برآورد چگالی ناپارامتریک نیازی به چنین فرضی نیست. منظور از اصطلاح ناپارامتریک این نیست که چنین مدل‌هایی کاملاً فاقد پارامتر هستند، بلکه تعداد و ماهیت پارامترها انعطاف‌پذیر هستند و از قبل ثابت نیستند. برآورد چگالی ناپارامتریک می‌تواند به‌عنوان یک تحلیل اولیه یا خود به‌عنوان یک ابزار تجزیه‌وتحلیل استفاده شود. ایده پشت برآورد چگالی ناپارامتریک قدیمی است، اما بسیاری از روش‌ها کاملاً محاسباتی هستند. ازآنجایی‌که قدرت محاسباتی

در چند دهه پیش به‌سرعت افزایش یافته است، روش‌های ناپارامتریک بیشتر موردتوجه قرار گرفته و پیشرفت‌های زیادی صورت گرفته است (Kile, 2010). Hsieh and Turnbull (1992) بدون مفروضات پارامتریک برای توزیع‌های اساسی، برآوردهای نقطه‌ای ناپارامتری را برای شاخص یودن بر اساس برآوردهای تجربی و هسته‌ای برای توزیع‌های زیربنایی مطالعه کردند. لذا در این مقاله نمره برش آزمون تولیمو با استفاده از شاخص یودن و بر اساس دو روش برآورد ناپارامتری تجربی و هسته به دست آورده شد و نتایج دو روش بر اساس دو شاخص خطای استاندارد بوت‌استرپ و ریشه میانگین مربعات خطا ارزیابی و مقایسه شدند. ازآنجایی‌که در روش هسته، مقدار پهنای باند تأثیر زیادی درنتیجه دارد لذا از دو روش تعیین پهنای باند در روش هسته استفاده شد و درنهایت نتایج با استفاده از معیار میانگین خطای مجذور یکپارچه[1] ($MISE$) و خطای مجذور یکپارچه[2] (ISE) مقایسه و تحلیل شدند. لازم به ذکر است برای دسترسی به نتایج دقیق‌تر برای تعیین نمره برش بهینه متقاضیان فرصت مطالعاتی تمام نمره‌های صحیح از ۴۷۰ تا ۵۰۰ در نظر گرفته شدند. روش‌های مختلف تخمین ناپارامتریک ازجمله روش هسته با پهنای باند سیلورمن، روش هسته با حداکثر احتمال پهنای باند اعتبار سنجی متقاطع و روش تجربی در تعیین نمره برش آزمون تولیمو با استفاده از روش مقایسه‌ای چگونه است و کدام روش تخمین‌های قابل‌اعتمادتری را ارائه می‌دهد؟

پیشینه پژوهش

منحنی ویژگی عملکرد[3] (ROC) یک روش گرافیکی محبوب است که از دهه ۱۹۷۰ باارزش‌ترین ابزار آماری برای توصیف و مقایسه دقت آزمون‌های تشخیصی برای تمایز بین دو جمعیت شناخته شده است (Zhou et al., 2009) به‌طوری‌که در بسیاری از زمینه‌های علمی از قبیل رادیولوژی (Metz, 1989)، روان‌پزشکی (Hsiao et al., 1989)، اپیدمیولوژی (Aoki et al., 1997) و سیستم‌های بازرسی ساخت (Somoza et al., 1990) استفاده شد. در اواخر دهه ۱۹۸۰، محققان شروع به استفاده از روش منحنی‌های ROC برای ارزیابی آزمون‌های تشخیصی پزشکی کردند (Hanley & McNeil, 1982). از منحنی ROC برای مشکلات زیست‌پزشکی به‌منظور بررسی اثربخشی نشانگرهای تشخیصی پیوسته در تشخیص افراد بیمار و سالم بسیار استفاده شده است (Shapiro, 1999; Greiner et al., 2000). اگر

1. Mean integrated squared error
2. Integrated square error
3. receiver operating characteristic

مقدار نشانگر آزمایش‌شده بیشتر از مقدار آستانه معین باشد، فرد بیمار (مثبت) ارزیابی می‌شود، در غیر این صورت آزمودنی سالم (منفی) تشخیص داده می‌شود. دقت هر مقدار آستانه معین را می‌توان با احتمال مثبت واقعی (حساسیت[1]) و احتمال منفی واقعی (وضوح[2]) اندازه‌گیری کرد (Fluss et al., 2005).

منحنی ROC نمودار حساسیت (Se(c)) را در مقابل (وضوح – ۱) (1-Sp(c)) روی تمام مقادیر آستانه ممکن (c) نشانگر رسم می‌کند. برای ارزیابی توانایی تمایز یک نشانگر، خلاصه کردن اطلاعات منحنی ROC در یک مقدار یا شاخص کلی رایج است. چندین شاخص ازجمله مساحت زیر منحنی ROC (AUC) و شاخص یودن در ادبیات یافت می‌شود که در کاربردهای مختلف مورداستفاده قرار گرفته شده‌اند (Shapiro, 1999; Aoki et al., 1997; Greiner et al., 2000; Grmec & Gašparovic, 2000).

با توجه به اهمیت تعیین نمره برش، تحقیقات زیادی بر روی تعیین تجربی نقطه‌های برش بهینه متمرکز شده است که در ادامه به برخی از تحقیق‌هایی که با روش یودن انجام شده است اشاره شده است.

Eckes (2017) به‌منظور تعیین نمرات برش در آزمون تعیین سطح زبان انگلیسی از ترکیب دو رویکرد منحنی ویژگی عملکرد و روش گروهی نمونه اولیه[3] استفاده کرد. روش گروهی نمونه اولیه، یک رویکرد اندازه‌گیری راش[4] را برای تجزیه‌وتحلیل مهارت آزمون‌شونده با مفهوم نمونه‌های اولیه برگرفته از تحقیقات در مورد قضاوت و طبقه‌بندی انسان ترکیب می‌کند. کارشناسان ابتدا زبان‌آموزانی را شناسایی می‌کنند که در هر یک از پنج سطح مهارت زبانی مشخص شده توسط چارچوب مرجع مشترک اروپایی[5] برای زبان‌ها مشخص شده است. به‌عبارت‌دیگر برای این مجموعه از نمونه‌های اولیه، برآوردهای مهارت زبان حاصل از مدل راش به‌عنوان ورودی برای تحلیل منحنی ویژگی عملکرد استفاده شد. درنهایت نمره‌های برش با استفاده از شاخص یودن به دست آمد که نرخ کلی طبقه‌بندی صحیح را به حداکثر می‌رساند و نرخ کلی طبقه‌بندی اشتباه را به حداقل می‌رساند. وی نتیجه گرفت که این روش امکان تنظیم نمرات برش را فراهم می‌کند که سطح بالایی از دقت

1. sensitivity
2. specificity
3. Prototype group method (PGM)
4. Rasch
5. Common european framework of reference (CEFR)

طبقه‌بندی را ازنظر مطابقت با طبقه‌بندی‌های متخصص نمونه‌های اولیه آزمون‌شونده نشان می‌دهد. علاوه بر این، نمرات برش مبتنی بر منحنی ویژگی عملکرد با دقت طبقه‌بندی بالاتری نسبت به نمره‌های برش حاصل از تجزیه‌وتحلیل رگرسیون لوجستیک از همان داده‌ها مرتبط بود.

Nakas و همکاران (2010) تعمیمی از شاخص یودن به نام J3 را برای مشکل طبقه‌بندی داده‌ها به سه کلاس برای انتخاب یک نقطه برش بهینه پیشنهاد کردند. آن‌ها هر دو روش پارامتری و ناپارامتریک را برای تخمین و آزمایش J3 برای داده‌های مربوط به بیماران با ویروس نقص ایمنی بررسی کردند که منجر به نتایج امیدوارکننده‌ای برای طبقه‌بندی دقیق بیماران در گروه‌های مختلف شد.

Luo and Xiong (2013) تعمیمی از شاخص یودن را برای طبقه‌بندی داده‌ها به سه گروه تشخیصی ترتیبی پیشنهاد کردند و از روش‌های پارامتریک و ناپارامتریک برای تخمین شاخص یودن بهینه و نقاط برش مربوطه و فواصل اطمینان متناظر استفاده کردند.

Carvalho and Branscum (2018) یک رویکرد ناپارامتریک بیزی برای تخمین شاخص یودن سه کلاسه و مقادیر برش بهینه متناظر آن برای ارزیابی اختلال شناختی در بیماران مبتلا به پارکینسون ارائه کردند. آن‌ها از روش مبتنی بر مخلوط‌های فرآیند دیریکله[1] برای تخمین شاخص یودن سه کلاسه و مقادیر برش بهینه استفاده کردند.

نتایج حاصل از این مطالعه اهمیت انتخاب یک رویکرد قابل‌اعتماد برای تعیین نمره برش را نشان می‌دهد که پیامدهای مهمی برای تجزیه‌وتحلیل آزمون و تصمیم‌گیری در محیط‌های آموزشی دارد. شاخص یودن یک معیار بسیار مورداستفاده در حوزه تشخیص پزشکی است، جایی که کارایی یک نشانگر زیستی (نشانگر غربالگری یا پیش‌بینی کننده) در طبقه‌بندی وضعیت بیماری موردبررسی قرار می‌گیرد؛ تحقیقات محدودی کاربرد این تکنیک را در زمینه‌های آموزشی بررسی کرده است. درنتیجه مقایسه و ارزیابی تکنیک‌های مختلف تخمین ناپارامتریک برای محاسبه شاخص یودن و سنجش قابلیت اطمینان آن‌ها در تقریب نمره برش در محیط‌های آموزشی می‌تواند در تعیین نمره برش آزمون‌های ملاک مرجع علی‌الخصوص از نوع سرنوشت‌ساز مؤثر باشد.

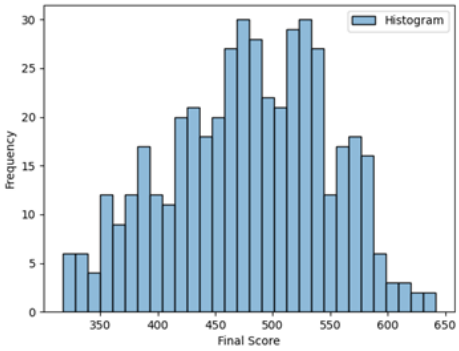
روش‌های برآورد ناپارامتری. اکثر روش‌ها برای برآورد نقاط برش بهینه با استفاده از شاخص یودن به‌عنوان متریک توسعه داده شده‌اند (Fluss et al., 2005; Leeflang et al., 2008).

1. Dirichlet process mixtures (DPM)

از بین روش‌های ناپارامتری می‌توان گفت روش‌های هیستوگرام و ناپارامتری تجربی قابل‌استفاده برای معیارهای دیگر غیر از شاخص یودن نیز هستند و روش هموارسازی مبتنی بر هسته مختص برآورد شاخص یودن است (Thiele & Hirschfeld, 2020).

قدیمی‌ترین و پرکاربردترین برآوردگر چگالی احتمال ناپارامتری، هیستوگرام است که به اوایل قرن نوزدهم برمی‌گردد. با توجه به مبدأ x_0 و a بلوک به عرض h ، بلوک‌های هیستوگرام به‌صورت بازه‌های $[x_0 + mh\,, x_0 + (m+1)h]$ به ازای اعداد صحیح مثبت و منفی m تعریف می‌شوند (Silverman, 2018). به‌عبارت‌دیگر در این روش، با تقسیم داده‌ها به بازه‌های یکنواخت و شمارش تعداد مشاهدات در هر بازه، برآوردی از چگالی احتمال به دست آورده می‌شود. بنابراین روش هیستوگرام تنها یک برآورد تقریبی از چگالی احتمال است و وابسته به تعداد و انتخاب بازه‌ها است. نمودار مربوط به داده‌های این مقاله با ۳۰ بازه در شکل ۱ نشان داده شده است.

شکل ۱. برآورد هیستوگرام نمره کل آزمون تولیمو با ۳۰ بازه



به‌منظور شرح روش‌های تجربی و هسته، فرض کنید که نتایج یک آزمون تشخیصی (نشانگر) $x_1, \dots, x_m$ و $y_1, \dots, y_n$ از دو نمونه تصادفی در جمعیت‌های قبول (P) و مردود (R) با توابع توزیع تجمعی F_P و G_R در دسترس هستند. برای هر آستانه معین c می‌توان حساسیت و وضوح نمره برش را برحسب توابع توزیع تجمعی به‌صورت زیر تعریف کرد:

$$Se(c) = 1 - G_R(c) \qquad , \qquad Sp(c) = F_P(c)$$

بنابراین شاخص یودن عبارت می‌شود از:

$$J = Se(c) + Sp(c) - 1 = F_P(c) - G_R(c) \qquad (2)$$

مقدار c که منجر به مقدار ماکسیمم J می‌شود به‌عنوان آستانه بهینه c^* در نظر گرفته می‌شود. برآورد Jبا برآورد G_R و F_P (یعنی $(\hat{G}_R, \hat{F}_P)$) و جایگزینی این برآوردها در معادله (۲) یعنی $\hat{J} = \max_c\{\hat{F}_P(c) - \hat{G}_R(c)\}$ حاصل می‌شود. در این مقاله برای برآورد G_R و F_P و تعیین مقدار شاخص یودن و تعیین نمره برش آزمون تولیمو از روش‌های برآورد ناپارامتری تجربی (EMP) و هسته (KDE) استفاده می‌شود.

– روش تجربی (EMP). ساده‌ترین روش برای بهینه‌سازی یک متریک، روش ناپارامتری تجربی است که نقطه برشی را برمی‌گزیند که مقدار متریک بهینه را در نمونه به دست می‌دهد بدین‌صورت که در تمام نقاط برش جستجو می‌کند و درنهایت نقطه برشی را انتخاب می‌کند که منجر به مقدار متریک بهینه در نمونه می‌شود؛ بنابراین تعیین نمره برش با استفاده از روش برآورد تجربی معادل انتخاب یک نمره برش بر اساس تجزیه‌وتحلیل منحنی ROC (هموار نشده) است (Thiele & Hirschfeld, 2020).

شاخص یودن در روش EMP مستقیماً از داده‌های مشاهده‌شده بدون هیچ فرض توزیعی محاسبه می‌شود. در روش تجربی از تابع توزیع تجربی به‌منظور برآورد تابع توزیع تجمعی احتمالات پیش‌بینی‌شده از یک مدل طبقه‌بندی باینری و بر اساس داده‌های مشاهده‌شده استفاده می‌شود. به‌عبارت‌دیگر در روش تجربی، تابع توزیع تجمعی نشانگر با تابع توزیع تجمعی تجربی نمونه برآورد می‌شود (Fluss et al., 2005). تابع توزیع تجمعی تجربی به‌صورت زیر محاسبه می‌شوند:

$$\hat{F}_P(c) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le c) \qquad , \qquad \hat{G}_R(c) = \frac{1}{m}\sum_{i=1}^{m} I(x_i \le c)$$

که در آن؛

$$I(u \le c) = \begin{cases} 1 & u \le c \\ 0 & u > c \end{cases}$$

پس از محاسبه تابع توزیع تجمعی تجربی هر یک گروه‌های قبول و مردود، J توسط رابطه (۲) به دست می‌آید و درنهایت مقدار c که مقدار $\hat{J}$ را بیشینه می‌کند به‌عنوان c^* تعیین می‌شود (Ruopp et al., 2008).

– روش هسته (KDE). روش ناپارامتری هسته، برای اولین بار توسط Parzen (1962) معرفی شد که برای برآورد تابع چگالی استفاده می‌شود که یکی از روش‌های متداول برای

هموارسازی است. این روش نسبت به روش تابع توزیع تجربی دارای دقت بسیار بیشتری است (Silverman, 2018). یکی از مزایای این روش این است که تخمین هموارى از تابع توزیع تجمعی را ارائه می‌دهد. (Fluss et al., 2005). فرض کنید متغیرهای تصادفی $X_n,\ldots,X_2,X_1$ مستقل و هم‌توزیع با تابع چگالی نامعلوم f باشند. یک برآوردگر چگالی هسته[1] تک متغیره (KDE) به‌صورت رابطه (۳) تعریف می‌شود.

$$\hat{f}_n(t)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-X_i}{h}\right),\quad t\in R\ ,\ h>0 \qquad (3)$$

که در آن $K(.)$ تابع هسته نامیده می‌شود که به‌طورکلی یک تابع هموار و متقارن مانند مثلثی، مستطیلی، گاوسی، اپاچنیکوف[2] و ... است و $h>0$ پهنای باند یا ضریب هموارسازی نامیده می‌شود که میزان هموارسازی را کنترل می‌کند و برای عملکرد برآوردگر بسیار مهم است، t یک مقدار آستانه[3] و n، حجم نمونه است. بنابراین هدف از برآورد چگالی هسته، برآورد چگالی احتمال $f(t)$ با استفاده از نمونه تصادفی $X_n,\ldots,X_2,X_1$ است (Park & Marron, 1990). انتخاب پهنای باند چالش‌برانگیزترین مرحله در ساخت برآوردگر چگالی با استفاده از روش هسته است. زیرا اگر حجم نمونه بزرگ باشد و داده‌ها به‌طور فشرده[4] توزیع شده باشند، بهتر است مقدار بزرگی برای h انتخاب شود تا چگالی هسته بیشترین احتمال را به هر نقطه بدهد. اما وقتی حجم نمونه کوچک است و داده‌ها پراکنده[5] هستند، بهتر است مقدار کوچکی برای h انتخاب شود تا چگالی هسته احتمالات بیشتری را در مقادیر همسایه یک نقطه پخش کند. به‌عبارت‌دیگر پهنای کم هسته به این معناست که میانگین‌گیری در هر نقطه بر روی تعداد کمی از مشاهدات انجام می‌شود و لذا این برآورد، توجه زیادی به نمونه‌ها داشته و به نمونه اجازه تغییرپذیری نمی‌دهد. چنین برآوردی، کم‌برآورد[6] نامیده می‌شود. بزرگ‌تر شدن پهنای هسته، منجر به برآورد هموارتری می‌شود به‌طوری‌که اگر به میزان کافی، پهنای باند بزرگ شود و درواقع مقدار مناسبی برای پهنای باند انتخاب شود ساختار اساسی چگالی داده‌ها قابل‌شناسایی است (Wand & Jones, 1994). از سوی دیگر کارایی مجانبی برآوردگر چگالی برای توابع مختلف هسته توسط

1. Kernel Density Estimator
2. Epanechnikov
3. threshold value
4. tightly
5. sparse
6. underestimation

Silverman (1986) بررسی شده است (Silverman, 2018). او نشان داد که با فرض انتخاب بهینه برای پارامتر هموارسازی، بازده مجانبی توابع مختلف هسته نزدیک به یک است و تفاوت چندانی بین آن‌ها وجود ندارد. بنابراین در عمل انتخاب نوع تابع هسته تأثیر چندانی ندارد و انتخاب بهینه پارامتر هموارسازی اهمیت بسیار بیشتری در دقت برآورد چگالی خواهد داشت. ازاین‌رو، در این مقاله از تابع هسته گاوسی که دارای خواص مطلوبی ازجمله تقارن است و نسبت به دیگر توابع هسته متداول‌تر است استفاده شده است (Trosset, 2009). با انتخاب تابع هسته گاوسی در رابطه (۳)، $K(t) = \Phi(t)$ می‌شود که نشانگر یک توزیع نرمال استاندارد $N(0,1)$ است که بایستی برای هر دو گروه قبول و مردود با حجم نمونه n و m نوشته شود. طبق رابطه (۳) و با فرض تابع هسته گاوسی، مشاهده می‌شود که پارامتر h در مخرج توابع توزیع بالا نقش انحراف استاندارد توزیع نرمال را دارد. یعنی h نقطه مهمی به‌عنوان عامل مقیاس دارد که پراکندگی تابع هسته را مشخص می‌کند. درنهایت بعد از محاسبه توابع هسته مربوط به گروه‌های قبول و مردود طبق رابطه (۳)، مقدار c که مقدار $\hat{f}$ موجود در رابطه (۲) را بیشینه می‌کند به‌عنوان c^* تعیین می‌شود (Zou et al., 1998).

دو روش تعیین پهنای باند در ادامه بیان شده است:

در برآورد چگالی هسته، انتخاب پارامتر پهنای باند بسیار مهم است زیرا همواری و دقت چگالی برآورد شده را تعیین می‌کند. روش سرانگشتی سیلورمن متداول‌ترین روش تعیین پهنای باند است که زمانی که تابع هسته و توزیع واقعی نرمال فرض شود بر اساس میانگین مجانبی خطای مربع یکپارچه $(AMISE)$ عمل می‌کند (Węglarczyk, 2018). این روش با فرض وجود یک تابع هسته در هر نقطه مشاهداتی و تجمیع آن‌ها به همراه مقدار مناسب پهنای باند، تابع چگالی احتمال را به روش ناپارامتری محاسبه می‌کند (Adamowski, 1987). پهنای باند بهینه در این روش بر اساس انحراف استاندارد نمونه و حجم نمونه برآورد می‌شود. طبق این روش پهنای باند با فرض هسته گاوسی برای نمونه‌ای به حجم n از رابطه (۴) به دست می‌آید.

$$h = 1/06\, s\, n^{-0/2} \qquad (4)$$

رابطه (۴) نسبت به داده پرت حساس است چراکه داده پرت منجر به بزرگ شدن مقدار s می‌شود و درنتیجه برآورد بزرگ‌تری از پهنای باند در برآورد تابع چگالی حاصل می‌شود. لذا از دامنه میان چارکی داده‌ها استفاده کرده و رابطه زیر به‌عنوان یک برآوردگر نیرومند برای تعیین h بهینه با استفاده از روش سرانگشتی سیلورمن حاصل می‌شود:

$$h_{opt} = 1/06n^{-0.2} \times min\left\{s, \frac{iqr}{1.34}\right\} \qquad (5)$$

که در آن s انحراف معیار نمونه و iqr دامنه بین چارکی است که از تفاضل چارک سوم از چارک اول نمونه داده‌ها به دست می‌آید (Thiele & Hirschfeld, 2020; Zucchini et al., 2003).

رویکرد دیگر برای تعیین پهنای باند، روش اعتبارسنجی متقابل ماکسیمم درستنمایی[1] است که توسط Habbema و همکاران (1974) و Duin (1976) پیشنهاد شد. طبق این روش، برای انتخاب h بایستی شبه درستنمایی[2] $\prod_{i=1}^{n} \hat{f}_h(X_i)$ ماکسیمم شود. در این روش تابع هسته بر روی زیرمجموعه‌ای از X_j بر اساس رویکرد اعتبارسنجی متقابل leave-one-out[3] برآورد می‌شود. بدین‌صورت که به تعداد حجم نمونه، داده‌ها تقسیم می‌شوند و هر بار به‌طور سیستماتیک یک مشاهده از یک مجموعه داده برای ارزیابی و آزمون کنار گذاشته می‌شود. به‌عنوان‌مثال مجموعه داده‌ای به حجم n طوری تقسیم می‌شود که هر بار تنها یک داده برای ارزیابی و آزمون نگه داشته می‌شود و از سایر داده‌ها برای آموزش مدل استفاده می‌شود و این کار n بار تکرار می‌شود. تابع هدف که $MLCV$ را به حداکثر می‌رساند به‌صورت زیر بیان می‌شود:

$$MLCV(h) = \frac{1}{n}\sum_{i=1}^{n} log\left[\sum_{j} K\left(\frac{X_j - x_i}{h}\right)\right] - log\left[(n-1)h\right]$$

که در آن، $h_{mlcv} = \underset{h>0}{\operatorname{argmax}}\, MLCV(h)$.

خوانندگان به‌منظور بررسی عمیق روش‌های انتخاب پهنای باند می‌توانند به Hall (1982)، Park and Marron (1990)، Van Es (1991)، Jones و همکاران (1996)، Loader (1999)، Heidenreich و همکاران (2013) و Barbeito and Cao (۲۰۲۰) مراجعه کنند.

– معیارهای ارزیابی پهنای باند: همه روش‌های تعیین باند بر اساس حداقل خطا عمل می‌کنند بدین‌معنی که پهنای باندی برآورد می‌شود که دارای حداقل خطا باشد. در این مقاله فرض می‌شود که پارامتر هموارسازی در تمام مقادیر x ثابت است بنابراین منطقی است که از یک معیار خطای کلی برای انتخاب پهنای باند استفاده شود (Kile, 2010). بدین‌منظور از معیار میانگین خطای مجذور یکپارچه ($MISE$) استفاده شد که به‌صورت رابطه (۶) تعریف می‌شود:

1. Maximum likelihood cross-validation
2. pseudo-likelihood
3. Leave-one-out cross validation ($LOOCV$)

$$MISE(h) = E\left[\int \left(\hat{f}_h(x) - f(x)\right)^2 dx\right] \quad (6)$$

هدف ما یافتن h ی است که حتی برای نمونه‌های خیلی کوچک، $MISE$ را کمینه کند. پهنای باندی که منجر به حداقل $MISE(h)$ می‌شود با h_{MISE} نشان داده می‌شود و به‌عنوان پهنای باند $MISE$ داده‌ها در نظر گرفته می‌شود. برای یافتن h بهینه از رابطه (۶) برای برآوردگر هسته لازم است رابطه مجانبی برای $MISE$ پیدا شود که سرراست‌تر باشد که بدین‌منظور از بسط تیلور استفاده می‌شود. لذا رابطه (۶) به‌صورت زیر نوشته می‌شود:

$$\begin{aligned} MISE(h) &= \int E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] dx \\ &= \int \left[E\left(\hat{f}_h(x) - f(x)\right)\right]^2 dx + \int Var\left(\hat{f}_h(x)\right) dx \\ &= \int \left[Bias\left(\hat{f}_h(x)\right)\right]^2 + \int Var\left(\hat{f}_h(x)\right) dx \end{aligned}$$

که در آن؛

$$\begin{aligned} Bias\left(\hat{f}_h(x)\right) &= E\left(\hat{f}_h(x)\right) - f(x) \\ &= E[K_h(x - X)] - f(x) \\ &= \int K_h(x - X) f(x) dx - f(x) \end{aligned}$$

و طبق رابطه $Var(X) = E(X^2) - [E(X)]^2$ خواهیم داشت:

$$Var\left(\hat{f}_h(x)\right) = \frac{1}{nh^2} Var\left(K\left(\frac{x - X}{h}\right)\right)$$

$$= \frac{1}{n}\int \frac{1}{h^2} K^2\left(\frac{x - X}{h}\right) f(x) dx$$
$$- \frac{1}{n}\left[\frac{1}{h}\int K\left(\frac{x - X}{h}\right) f(x) dx\right]^2$$

معیار میانی بین $MISE$ و MSE ، خطای مجذور یکپارچه (ISE) است که یک اندازه‌گیری ناهمخوانی[1]است که برای برآورد مقدار پارامتر هموارسازی استفاده می‌شود و به‌صورت زیر تعریف می‌شود:

$$ISE(h) = \int_{-\infty}^{+\infty} \left(\hat{f}_h(x) - f(x)\right)^2 dx$$

1. discrepancy

– داده‌ها: آزمون تولیمو. آزمون تولیمو[1] یکی از آزمون‌های استاندارد و معتبر تعیین سطح دانش زبان انگلیسی دانشجویان داخل ایران است که توسط سازمان سنجش طراحی و برگزار می‌شود. دانشجویان برای تحصیل در مقاطع کارشناسی‌ارشد و دکترا در رشته‌های مختلف دانشگاهی در ایران و یا برای اخذ بورس تحصیلی و نیز افرادی که در جستجوی کار هستند به‌منظور استخدام در برخی مؤسسات و ارگان‌های دولتی باید در یکی از آزمون‌های معتبر زبان مانند آزمون تولیمو، آزمون تافل، آزمون $MCHE$ و یا هر آزمون دیگری نمره قابل قبولی را کسب کنند که این نمره موردقبول در هر یک از این آزمون‌ها را هر یک از مؤسسات و دانشگاه‌ها بر اساس انتظارات خود تعیین می‌کنند. به‌عنوان نمونه در حال حاضر برای فرصت مطالعاتی دانشجویان دکتری، حداقل نمره قبولی موردنیاز برای آزمون تولیمو ۴۸۰ تعیین شده است.

در آزمون تولیمو هر چهار مهارت گرامر، شنیداری، درک مطلب و نوشتاری مورد آزمون قرار می‌گیرد. بدین‌صورت که در جلسه آزمون تولیمو دو دفترچه در اختیار داوطلبان قرار داده می‌شود که بخش نوشتاری در یک دفترچه و شامل یک سؤال تشریحی و سه بخش دیگر در دفترچه دیگر و به‌صورت ۱۰۵ سؤال در اختیار داوطلبان قرار می‌گیرد که هر بخش شامل ۳۵ سؤال چهارگزینه‌ای است. بخش تشریحی آزمون تولیمو به بخش نوشتاری آزمون آیلتس و آزمون جی‌آرای (GRE) شباهت دارد و نشان‌دهنده مهارت داوطلبان در نوشتن متن و مقاله به زبان انگلیسی است. ازآنجایی‌که در این آزمون نمره منفی وجود ندارد لذا معمولاً داوطلبان به همه سؤال‌های چهارگزینه‌ای پاسخ می‌دهند. نمره کل داوطلبان در بازه ۲۰۰ تا ۶۷۷ قرار می‌گیرد.

در این پژوهش، از نمره کل داوطلبان یک دوره از آزمون تولیمو استفاده شده است که در دوره مزبور، ۴۶۱ داوطلب از رشته‌ها و دانشگاه‌های مختلف شرکت کرده‌اند. دامنه نمره‌های کسب‌شده در این دوره از آزمون تولیمو از ۳۱۷ تا ۶۴۲ بوده است و جمعاً ۱۵۳ نمره به دست آورده شده است که با رسم جدول فراوانی مشخص شد مد داده‌ها نمره ۴۷۰ است که ۹ نفر این نمره را کسب کرده‌اند. میانگین و میانه نمره‌های کسب‌شده به ترتیب برابر ۴۷۸ و ۴۸۲ است که بیانگر این مطلب است که عملکرد معمولی داوطلبان در حدود محدوده متوسط نمرات است. از سوی دیگر انحراف استاندارد نمره‌ها برابر ۶۹ است که می‌توان این‌گونه تفسیر کرد که تنوع متوسطی در نمرات وجود دارد و برخی از داوطلبان به‌طور

1. The test of language by the Iranian measurement organization

قابل‌توجهی بالاتر از آن نمره یا کمتر از مقدار میانگین (یعنی ۴۷۸) را گرفته‌اند. مقادیر آماره‌های توصیفی داده‌های استفاده شده در این مطالعه در جدول ۱ آورده شده است.

جدول ۱. آماره‌های توصیفی داده‌ها

میانگین	میانه	مد	چارک اول	چارک سوم	انحراف استاندارد
٤٧٨	٤٨٢	٤٧٠	٤٣٢	٥٢٩	٦٩

روش

داده‌ها با استفاده از نرم‌افزار اکسل و محاسبات با استفاده از نرم‌افزار پایتون و با به‌کارگیری کتابخانه‌های *pandas*، *numpy*، *sklearn*، *seaborn* و *matplotlib* انجام شدند.

داده‌های این مقاله در ۳۱ فایل اکسل بدین‌صورت ذخیره شدند که از دستور *if* اکسل برای تعیین وضعیت پذیرش یا رد هر داوطلب بر اساس نمره نهایی آن‌ها در رابطه با هر یک از اعداد طبیعی در بازه ۴۷۰ تا ۵۰۰ مثلاً ۴۷۰ استفاده شد. داوطلبانی که نمرات نهایی بالاتر از ۴۷۰ داشتند پذیرفته درحالی‌که آن‌هایی که نمرات نهایی کمتر از ۴۷۰ داشتند رد شدند. این نتایج به ترتیب با ۱ و ۰ نمایش داده شدند؛ بنابراین هر فایل شامل دو ستون نمره نهایی و برچسب قبول و رد برحسب نمره برش انتخابی شد که برای هر فایل به‌طور مجزا مقدار شاخص یودن و به‌منظور برآورد پایایی و میزان تغییرپذیری روش‌ها، خطای استاندارد بوت استرپ و ریشه میانگین مربعات خطا برای هر سه روش تجربی، روش هسته با روش پهنای باند سیلورمن و روش هسته با تعیین پهنای باند به روش اعتبارسنجی متقابل ماکسیمم درستنمایی به دست آورده شد. خطاهای استاندارد بوت استرپ برآوردی از تغییرپذیری برای مقادیر شاخص یودن را ارائه می‌دهد. به‌منظور محاسبه خطای استاندارد بوت استرپ از مجموعه داده خود، k نمونه مکرر و با جایگذاری گرفته و برای هر نمونه، خطای استاندارد از طریق رابطه $\frac{s}{\sqrt{n}}$ محاسبه شد. میانگین k خطای استاندارد برابر با خطای استاندارد بوت‌استرپ در نظر گرفته شد. ازآنجایی‌که مبنای اساسی ریشه میانگین مربعات خطا در اختلاف بین مقدار پیش‌بینی‌شده حاصل از مدل یا برآوردگر آماری و مقدار واقعی است لذا ستون برچسب به‌عنوان متغیر وابسته واقعی در نظر گرفته شد. در مورد پیش‌بینی متغیر وابسته، روش زیر را اجرا کردیم: در مرحله اول، مدل مربوطه را برای هر یک از گروه‌های قبول و مردود برازش دادیم. با توجه به اینکه انتظار می‌رود گروه پذیرفته‌شدگان نمره بالاتری نسبت

به گروه مردودی‌ها کسب کرده باشند، به این افراد مقدار ۱ در متغیر وابسته پیش‌بینی‌شده و در غیر این صورت مقدار ۰ اختصاص داده شد. در ادامه مراحل اجرای هر یک از روش‌ها (EMP و KDE) که برای هر یک از ۳۱ فایل داده به‌طور جداگانه اجرا شد شرح داده می‌شود.

برای برآورد شاخص یودن با استفاده از روش EMP، تابع توزیع تجمعی تجربی برای هر دو گروه مردود (۰) و قبول (۱)، نرخ مثبت کاذب (FPR) و نرخ مثبت واقعی (TPR) محاسبه شد. مقادیر ایجادشده در دو ستون با عنوان‌های FPR و TPR در یک فایل اکسل ذخیره شدند. در مرحله بعد ماکسیمم قدر مطلق تفاضل این دو ستون به دست آورده شد. به‌عنوان‌مثال ۲۰۳ امین مقدار دارای ماکسیمم $|FPR - TPR|$ شد. درنهایت با صعودی کردن ستون مربوط به نمره کل و پیدا کردن ۲۰۳ امین عدد که در فایل ما ۴۷۰ بود، نمره برش برای فایل مربوطه تعیین شد.

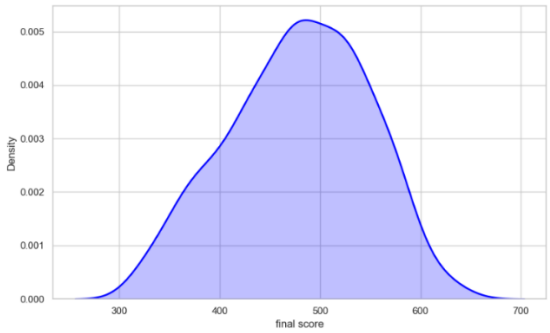
برای برآورد شاخص یودن با استفاده از روش هسته و با به‌کارگیری روش سرانگشتی سیلورمن ابتدا پیش‌فرض نرمال بودن داده‌ها بررسی شد. درواقع توزیع و نقاط پرت در داده‌ها به‌منظور برآورد شاخص یودن با استفاده از رویکرد ناپارامتریک هسته موردبررسی قرار گرفت. برای شناسایی نقاط پرت بالقوه در مجموعه داده‌ها از روش دامنه بین چارکی (IQR) استفاده شد. در این مقاله داده‌ای به‌عنوان نقطه پرت در نظر گرفته شد که خارج از محدوده مقادیر بین $Q_1 - 1.5IQR$ و $Q_3 + 1.5IQR$ قرار بگیرد. از شکل ۲ مشخص است که توزیع داده‌ها تقریباً نرمال است. درنهایت مشخص شد که داده‌های مربوط به نمره کل حاوی داده پرتی نیستند. لذا از روش سرانگشتی سیلورمن برای تعیین پهنای باند روش ناپارامتری هسته استفاده شد.

برای هر نمره برش بین ۴۷۰ تا ۵۰۰، مقدار هسته بر اساس تابع هسته گاوسی و با به‌کارگیری روش پهنای باند مرتبط برای هر دو گروه قبول و مردود برآورد شد و درنهایت ماکسیمم تفاضل مقادیر حاصل به‌عنوان مقدار شاخص یودن در نظر گرفته شد.

لازم به ذکر است که رویکرد $MLCV$ با استفاده از تابع $GridSearchCV$ موجود در کتابخانه $sklearn$ پیاده‌سازی شد. $GridSearchCV$ فرآیند تنظیم فراپارامتر را به‌منظور تعیین مقادیر بهینه برای یک مدل معین انجام می‌دهد. ازآنجایی‌که عملکرد مدل هسته به‌طور

قابل‌توجهی به مقدار پهنای باند بستگی دارد لذا با استفاده از تابع $GridSearchCV$، ۱۰۰۰ مقدار در بازه ۱۰ به توان اعداد بین ۱- و ۱ امتحان و مقدار بهینه پهنای باند بهینه انتخاب شد.

شکل ۲. نمودار تابع چگالی نمره کل آزمون تولیمو



یافته‌ها

داده استفاده شده در این پژوهش شامل نمره کل ۴۶۱ داوطلب آزمون تولیمو است که در این آزمون شرکت کرده‌اند. به‌منظور اجرای پژوهش، ستونی با مقادیر واقعی قبول و رد بر اساس نمره‌های ۴۷۰ تا ۵۰۰ تشکیل شد که دارای مقادیر ۰ (مردود) و ۱ (قبول) بودند. درمجموع ۳۱ فایل عددی حاصل شد که هر فایل داده برای هر روش استفاده شد و درنهایت بهترین مقدار برای هر روش بر اساس ماکسیمم مقدار شاخص یودن انتخاب گردید. نتایج برآورد و خطاهای استاندارد بوت‌استرپ (BSE) و ریشه میانگین مربعات خطا $(RMSE)$ برای این ۳۱ فایل داده که برگرفته از خروجی ۹۳ کد است در جدول ۱ آورده شده است.

جدول ۲. برآورد J و c^* برای آزمون تولیمو

	$KDE-Silverman$	$KDE-MLCV$	EMP
J	۰/۶۳	۰/۷۵	۰/۳۱
$RMSE$	۰/۷۱۵	۰/۷۲	۰/۶۸
BSE	۰/۰۰	۰/۰۳۹	۰/۰۱۷
c^*	٤٧٩	٤٧٩	٤٦٥

بر اساس نتایج نوشته‌شده در جدول ۲ مشاهده می‌شود که مقدار شاخص یودن برای روش هسته با پهنای باند $MLCV$ از دو روش دیگر بیشتر است ولی نمره برش به‌دست‌آمده

در هر دو روش هسته ۴۷۹ به دست آمده است که تنها یک نمره با میانگین نمره‌ها تفاوت دارد و در روش تجربی برابر ۴۶۵ شد که از مد نمره‌ها یعنی ۴۷۰ تفاوت کمتری دارد. شاخص یودن یک معیار عملکردی است که معمولاً در آزمایش‌های تشخیصی ازجمله آزمون‌های ملاک مرجع مورداستفاده قرار می‌گیرد که نشان‌دهنده توانایی یک آزمون برای طبقه‌بندی صحیح افراد به دو گروه (قبول و مردود) است. مقدار شاخص یودن بالاتر نشان‌دهنده توانایی تشخیص بهتر افراد قبول و مردود است بنابراین بر اساس شاخص یودن، روش هسته با پهنای باند $MLCV$ روش بهتری برای طبقه‌بندی داوطلبان به دو گروه قبول و رد است. از سوی دیگر مقدار ریشه میانگین مربعات خطا و خطای استاندارد بوت استرپ روش هسته با پهنای باند $MLCV$ از دو روش دیگر کمی بیشتر است. خطای استاندارد بوت‌استرپ که برآوردی از تغییرپذیری نمونه‌برداری هر روش است در روش هسته با پهنای باند سیلورمن صفر شده است بدین معنی روش سیلورمن منجر به برآوردهای پایدارتر و دقیق‌تری شده است. لذا شاخص‌های ارزیابی هسته ISE و $MISE$ برای این دو روش محاسبه شد که نتایج به همراه پهنای باند بهینه برآورد شده این دو روش در جدول ۳ آورده شده است.

جدول ۳. ارزیابی عملکرد روش‌های هسته با نمره برش ٤٧٩ و پهنای باند بهینه

شاخص‌های ارزیابی	$KDE - Silverman$	$KDE - MLCV$
h_{opt}	۲۱/٤٦	٦
ISE	۰/۲٦	۰/۱۸
$MISE$	۰/۱۳	۰/۱۹

h_{opt} =پهنای باند بهینه

ISE اختلاف کلی بین شاخص یودن برآورد شده و مقدار واقعی را اندازه‌گیری می‌کند بنابراین مقدار کمتر ISE بیانگر برازش بهتر با شاخص واقعی یودن است. $MISE$ یک اندازه‌گیری متوسط از اختلاف در کل محدوده مقادیر ممکن را ارائه می‌دهد به‌طوری‌که $MISE$ کمتر نشان‌دهنده این است که عملکرد کلی مدل بهتر است. از سوی دیگر پهنای باند کمتر می‌تواند جزئیات محلی بیشتری را ثبت کند، درحالی‌که پهنای باند بیشتر ممکن است داده‌ها را بیش‌ازحد هموار کند و حتی باعث محو شدن ویژگی‌های مهم شود. با توجه به این توضیحات به نظر می‌رسد که روش هسته با $MLCV$ در مقایسه با روش هسته با پهنای باند

سیلورمن عملکرد بهتری در برآورد شاخص یودن دارد چراکه دارای پهنای باند کمتر، ISE کمتر و $MISE$ کمی بالاتر است به‌علاوه بر اساس جدول ۲ ، مقدار شاخص یودن که معیار مهمی برای ارزیابی عملکرد طبقه‌بندی است در روش هسته با $MLCV$ بیشتر است.

بحث و نتیجه‌گیری

تعیین دقیق نمره برش بهینه در آزمون‌های ملاک مرجع برای به حداکثر رساندن اثربخشی آن‌ها ضروری است. شاخص یودن یکی از شاخص‌هایی است که هم کارایی یک نشانگر تشخیصی را می‌سنجد و هم مقدار نمره برش را برای نشانگر انتخاب می‌کند. ازآنجایی‌که یک روش واحدی برای تعیین دقیق نمره برش وجود ندارد لذا برآورد دقیق شاخص یودن چالش‌برانگیز است. لذا در این مطالعه، مقدار شاخص یودن با استفاده از سه روش ناپارامتری تجربی، برآورد چگالی هسته با پهنای باند سیلورمن و برآورد چگالی هسته با پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برآورد شد. همچنین از شاخص‌های خطای استاندارد بوت‌استرپ (BSE)، ریشه میانگین مربعات خطا ($RMSE$)، مربع خطای یکپارچه (ISE) و میانگین مربعات خطای یکپارچه ($MISE$) برای ارزیابی عملکرد استفاده شد.

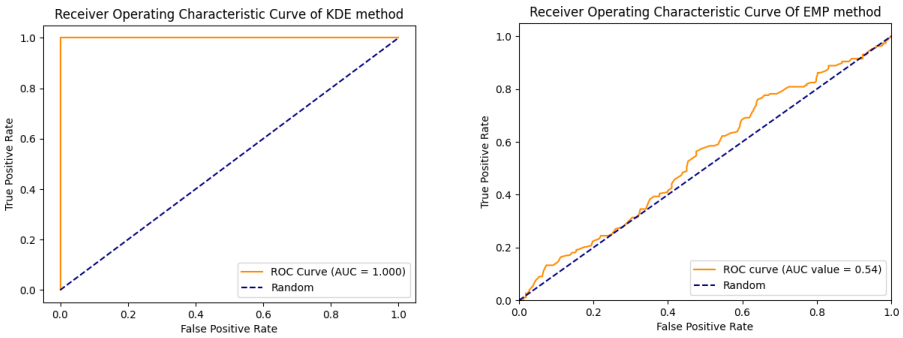
در روش هسته بایستی سه مورد تابع هسته، پهنای باند و حجم نمونه مشخص شود. با توجه به تحقیقات انجام شده توسط Silverman (1976) مشخص شد که تابع هسته نقش چندانی در برآورد چگالی ندارد و مقدار دقیق پهنای باند درنتیجه بسیار مهم است. به‌طوری‌که مقدار کم و مقدار زیاد h می‌تواند منجر به اطلاعات نادرستی شود؛ بنابراین برآورد مقدار بهینه h برای ایجاد معنی‌دارترین و دقیق‌ترین چگالی بسیار مهم است. روش‌های مختلفی برای تعیین پهنای باند وجود دارد ازجمله می‌توان به روش جایگذاری سیلورمن، اعتبارسنجی متقابل، اعتبارسنجی متقابل ماکسیمم درستنمایی ($MLCV$) و اعتبارسنجی متقابل کمترین توان‌های دوم[1] ($LSCV$) اشاره کرد (Silverman, 2018).

طبق مبانی نظری و پیشینه پژوهش‌های مختلف ازجمله Fluss و همکاران (2005)، Ewald (2006)، Leeflang و همکاران (2008)، Hirschfeld and do Brasil (2014) نشان دادند که هنگامی‌که از روش بهینه‌سازی تجربی برای برآورد شاخص یودن استفاده می‌شود به نتایج بسیار متغیری منجر می‌شود زیرا این روش به تصادفی بودن در نمونه حساس است. به‌علاوه درصورتی‌که داده‌ها نرمال و یا نزدیک به نرمال باشند روش چگالی هسته با پهنای باند

1. Least squares cross validation

سیلورمن منجر به برآوردهای پایدارتری از شاخص یودن در مقایسه با روش برآورد ناپارامتری تجربی می‌شود. برتری روش هسته نسبت به روش برآورد تجربی را می‌توان با رسم منحنی ROC و محاسبه سطح زیر نمودار (AUC) نیز نشان داد که در شکل ۳ آورده شده است.

شکل ۳. منحنی‌های ویژگی عملکرد و مقدار مساحت زیر منحنی نمره برش‌های بهینه به دست آمده از روش‌های هسته و تجربی



به‌طورکلی روش تعیین پهنای باند نقش مهمی در پایایی روش هسته دارد که بهترین کار، استفاده از چند روش و ارزیابی آن‌هاست. پایایی روش‌ها با استفاده از خطای استاندارد بوت‌استرپ ارزیابی شد.

نتایج نشان داد که روش هسته با پهنای باند سیلورمن منجر مقدار شاخص یودن ۰/۶۳ با نمره برش ۴۷۹ را به دست آورد. ریشه میانگین مربعات خطای این روش ۰/۷۱۵ با خطای استاندارد بوت‌استرپ آن ۰ به دست آمد؛ بنابراین برآورد شاخص یودن با استفاده از این روش نسبتاً پایدار و دقیق بود. از سوی دیگر، روش هسته با $MLCV$ دارای مقدار شاخص یودن بالاتر یعنی ۰/۷۵ با همان نمره برش ۴۷۹ را تولید کرد. یعنی روش هسته با $MLCV$ برآورد دقیق‌تری از شاخص یودن را در مقایسه با روش هسته با پهنای باند سیلورمن می‌دهد به معنای این است که قدرت تمایز بین افراد لایق و نالایق در این روش بیشتر است. ولی خطای استاندارد بوت استرپ آن برابر ۰/۰۳۹ است که نشان‌دهنده درجه کمی از تغییرپذیری در برآورد است. روش تجربی منجر به مقدار شاخص یودن ۰/۳۱ با نمره برش ۴۶۵ شد. ریشه

میانگین مربعات خطای این روش ۰/۶۸ بود و خطای استاندارد بوت‌استرپ آن برابر ۰/۰۱۷ شد.

این یافته‌ها نشان می‌دهند که انتخاب روش تعیین پهنای باند مربوط به روش هسته می‌تواند بر برآورد شاخص یودن تأثیر بسزایی داشته باشد. با توجه به نتایج حاصل‌شده در این مقاله و پیشینه روش‌ها می‌توان این‌گونه نتیجه گرفت که درصورتی‌که داده‌ها نرمال باشند روش هسته با پهنای باند سیلورمن می‌تواند منجر به نتایج پایداری شود و در غیر این صورت روش هسته با پهنای باند اعتبارسنجی متقابل ماکسیمم درستنمایی برای محاسبه و ارزیابی شاخص یودن در آزمون‌های ملاک مرجع و تعیین نمره برش ارجحیت دارد.

بااین‌حال تحقیقات بیشتری برای تأیید این نتایج و بررسی عملکرد این روش‌ها در زمینه‌های مختلف موردنیاز است. این تحقیقات می‌توانند به بهبود فرآیند طبقه‌بندی افراد و کاهش خطاهای منفی و مثبت کاذب در نتایج آزمون‌ها کمک کنند. همچنین مفید است اثرات مختلف تعیین نمره برش بر عواقب فردی و سازمانی نیز موردتوجه قرار گیرد تا تصمیم‌گیری‌های بهتری در تعیین نمره برش گرفته شود. این اقدامات می‌توانند به بهبود سیستم‌های ارزیابی و تصمیم‌گیری در محیط‌های آموزشی و سازمانی کمک کنند.

تعارض منافع

هیچ تعارض منافعی وجود ندارد.

سپاسگزاری

مقاله حاضر برگرفته از رساله دکتری رشته سنجش و اندازه‌گیری دانشگاه تهران با عنوان «مقایسه نمرات برش آزمون‌های ملاک مرجع در الگوریتم‌های یادگیری عمیق و روش‌های منتخب موردمطالعه: آزمون تولیمو» با حمایت دانشگاه تهران و سازمان سنجش آموزش کشور است. بدین‌وسیله از سازمان سنجش آموزش کشور به خاطر همکاری در اجرای پژوهش حاضر و از فصلنامه اندازه‌گیری تربیتی برای داوری مقاله سپاسگزاری می‌شود.

References

Adamowski, K. (1987). *Nonparametric Techniques For Analysis Of Hydrological Events.* Paper Presented At The Water For The Future: Hydrology In Perspective (Proceedings Of The Rome Symposium).

Aoki, K., Misumi, J., Kimura, T., Zhao, W., & Xie, T. (1997). *Evaluation Of Cutoff Levels For Screening Of Gastric Cancer Using Serum Pepsinogens And*

Distributions Of Levels Of Serum Pepsinogen I, Ii And Of Pg I/Pg Ii Ratios In A Gastric Cancer Case-Control Study. *Journal Of Epidemiology*, *7*(3), 143-151.

Barbeito, I., & Cao, R. (2020). *Nonparametric Curve Estimation And Bootstrap Bandwidth Selection*. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(3), E1488.

Carvalho, V. I. D., & Branscum, A. J. (2018). *Bayesian Nonparametric Inference For The Three-Class Youden Index And Its Associated Optimal Cutoff Points. Statistical Methods In Medical Research*, *27*(3), 689-700.

Dardick, W. R., & Weiss, B. A. (2019). *An Investigation Of Chi-Square And Entropy Based Methods Of Item-Fit Using Item Level Contamination In Item Response Theory. Journal Of Modern Applied Statistical Methods*, *18*.

Duin. (1976). *On The Choice Of Smoothing Parameters For Parzen Estimators Of Probability Density Functions*. *Ieee Transactions On Computers*, *100*(11), 1175-1179.

Eckes, T. (2017). *Setting Cut Scores On An Efl Placement Test Using The Prototype Group Method: A Receiver Operating Characteristic (Roc) Analysis.* Language Testing, *34*(3), 383-411.

Ewald, B. (2006). *Post Hoc Choice Of Cut Points Introduced Bias To Diagnostic Research.* Journal Of Clinical Epidemiology, *59*(8), 798-801.

Fluss, R., Faraggi, D., & Reiser, B. (2005). *Estimation Of The Youden Index And Its Associated Cutoff Point*. Biometrical Journal, *47*(4), 458-472. Doi:10.1002/Bimj.200410135

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). *Principles And Practical Application Of The Receiver-Operating Characteristic Analysis For Diagnostic Tests*. Preventive Veterinary Medicine, *45*(1-2), 23-41.

Grmec, Š., & Gašparovic, V. (2000). *Comparison Of Apache Ii, Mees And Glasgow Coma Scale In Patients With Nontraumatic Coma For Prediction Of Mortality.* Critical Care, *5*(1), 1-5.

Habbema, J., Hermans, J., & Van Den Broek, K. (1974). *A Stepwise Discriminant Analysis Program Using Density Estimation.*

Hall, P. (1982). *Cross-Validation In Density Estimation.* Biometrika, *69*(2), 383-390.

Hanley, J. A., & Mcneil, B. J. (1982). *The Meaning And Use Of The Area Under A Receiver Operating Characteristic (Roc) Curve*. Radiology, *143*(1), 29-36.

Heidenreich, N.-B., Schindler, A., & Sperlich, S. (2013). *Bandwidth Selection For Kernel Density Estimation: A Review Of Fully Automatic Selectors.* Asta Advances In Statistical Analysis, *97*, 403-433.

Hirschfeld, G., & Do Brasil, P. E. A. A. (2014). *A Simulation Study Into The Performance Of "Optimal" Diagnostic Thresholds In The Population:"Large" Effect Sizes Are Not Enough.* Journal Of Clinical Epidemiology, *67*(4), 449-453.

Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). *Diagnosing Diagnoses: Receiver Operating Characteristic Methods And Psychiatry.* Archives Of General Psychiatry, *46*(7), 664-667.

Hsieh, F., & Turnbull, B. (1992). *Nonparametric Methods For Evaluating Diagnostic Tests.* Retrieved From

Jones, M. C., Marron, J. S., & Sheather, S. J. (1996). *A Brief Survey Of Bandwidth Selection For Density Estimation*. Journal Of The American Statistical Association, *91*(433), 401-407.

Kile, H. (2010). *Bandwidth Selection In Kernel Density Estimation.* (Master). University Of Science And Technology, Norwegian.

Leeflang, M. M., Moons, K. G., Reitsma, J. B., & Zwinderman, A. H. (2008). *Bias In Sensitivity And Specificity Caused By Data-Driven Selection Of Optimal Cutoff Values: Mechanisms, Magnitude, And Solutions.* Clinical Chemistry, *54*(4), 729-737.

Loader, C. R. (1999). *Bandwidth Selection: Classical Or Plug-In?* The Annals Of Statistics, *27*(2), 415-438.

Luo, J., & Xiong, C. (2013). *Youden Index And Associated Cut-Points For Three Ordinal Diagnostic Groups*. Communications In Statistics-Simulation And Computation, *42*(6), 1213-1234.

Metz, C. E. (1989). *Some Practical Issues Of Experimental Design And Data Analysis In Radiological Roc Studies.* Investigative Radiology, *24*(3), 234-245.

Nakas, C. T., Alonzo, T. A., & Yiannoutsos, C. T. (2010). *Accuracy And Cut-Off Point Selection In Three-Class Classification Problems Using A Generalization Of The Youden Index.* Statistics In Medicine, *29*(28), 2946-2955.

Park, B. U., & Marron, J. S. (1990). *Comparison Of Data-Driven Bandwidth Selectors.* Journal Of The American Statistical Association, *85*(409), 66-72.

Parzen, E. (1962). *On Estimation Of A Probability Density Function And Mode.* The Annals Of Mathematical Statistics, *33*(3), 1065-1076.

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). *Youden Index And Optimal Cut-Point Estimated From Observations Affected By A Lower Limit Of Detection.* Biometrical Journal: Journal Of Mathematical Methods In Biosciences, *50*(3), 419-430.

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). *Optimal Cut-Point And Its Corresponding Youden Index To Discriminate Individuals Using Pooled Blood Samples.* Epidemiology, 73-81.

Shapiro, D. E. (1999). *The Interpretation Of Diagnostic Tests.* Statistical Methods In Medical Research, *8*(2), 113-134.

Silverman, B. W. (2018). Density Estimation For Statistics And Data Analysis: Routledge.

Somoza, E., Mossman, D., & Mcfeeters, L. (1990). *The Info-Roc Technique: A Method For Comparing And Optimizing Inspection Systems.* Review Of Progress In Quantitative Nondestructive Evaluation, 601-608.

Thiele, C., & Hirschfeld, G. (2020). *Cutpointr: Improved Estimation And Validation Of Optimal Cutpoints In R.* Arxiv Preprint Arxiv:*2002.09209*.

Trosset, M. W. (2009). *An Introduction To Statistical Inference And Its Applications With R:* Crc Press.

Van Es, B. (1991). *Likelihood Cross-Validation Bandwidth Selection For Nonparametric Kernel Density Estimators †*. Journal Of Nonparametric Statistics, *1*(1-2), 83-110. Doi:10.1080/10485259108832513

Wand, M. P., & Jones, M. C. (1994). *Kernel Smoothing:* Crc Press.

Węglarczyk, S. (2018). *Kernel Density Estimation And Its Application.* Paper Presented At The Itm Web Of Conferences.

Youden, W. J. (1950). *Index For Rating Diagnostic Tests.* Cancer, 3(1), 32-35.

Zhou, X.-H., Mcclish, D. K., & Obuchowski, N. A. (2009). *Statistical Methods In Diagnostic Medicine*: John Wiley & Sons.

Zou, K. H., Tempany, C. M., Fielding, J. R., & Silverman, S. G. (1998). *Original Smooth Receiver Operating Characteristic Curve Estimation From Continuous Data: Statistical Methods For Analyzing The Predictive Value Of Spiral Ct Of Ureteral Stones. Academic Radiology, 5*(10), 680-687.

Zucchini, W., Berzel, A., & Nenadic, O. (2003). *Applied Smoothing Techniques. Part I: Kernel Density Estimation, 15*, 1-20.