"Research article" doi: 10.71767/jinev.2024.4013014

Using Multiple Regression in the Hurwicz Method to Determine the Cut Score of Tests as an Alternative to Angoff and Bookmark Methods[1]

Maryam Parsaeian [2], Ebrahim Khodaie[3*], Balal Izanloo [4], Keyvan Salehi[5], Sima Naghizadeh[6]

(Received: 2024.01.17 - Accepted: 2024.05.14)

1- The present article is taken from the thesis of Maryam Parsaeian, Ph. D. student in the field of Evaluating and measuring in University of Tehran.
2- Ph.D. student, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.
3- Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.
*- Corresponding Author: khodaie@ut.ac.ir
4- Assistant Professor, University of Kharazmi, Faculty of Psychology and Education, Tehran, Iran
5- Associate Professor, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.
6- Assistant Professor, National Organization of Educational Testing (NOET), Tehran, Iran,

Abstract

The aim of the current research is to determine the cutoff score of the Tolimo test for doctoral candidates applying for study opportunities abroad by applying multiple regression in the Hurwicz method, which is a decision-making approach. The statistical population of this research includes the candidates of one course of Tolimo exam, which were 461 people. The research method of this study is based on a quantitative approach, and in terms of its purpose, it is part of applied research, and in terms of analysis, it is part of secondary analysis studies, and the scores of all 461 candidates in each of the grammar, reading comprehension, listening, and writing sections are used to determine the cutoff score. Based on the results of the study and considering the small amount of classification error, if the total score is used, the average value of the Hurwicz index is the cutoff score of 481, and if the linear regression model is significant, using the results of the Hurwicz index in multiple regression, the cutoff score is 494. It can be used as a cutoff score. The proposed method significantly improves the accuracy of determining the cutoff score compared to Angoff and Bookmark methods. According to the findings obtained from cross-validation and error values, the cutoff score obtained from the combined strategy leads to more accurate results with zero classification error.

***Keywords*:** Cut score, Hurwicz method, multiple regression, Angoff method, bookmark method, cross-validation, Tolimo test

نشریه علمی آموزش و ارزشیابی سال ۱۷، شماره ۶۵، بهار ۱۴۰۳، ص.ص. ۱۹۲-۱۶۹

«مقاله پژوهشی» doi: 10.71767/jinev.2024.4013014

استفاده از رگرسیون چندگانه در روش هارویچ برای تعیین نمره برش آزمون‌ها به عنوان جایگزینی برای روش‌های آنگوف و بوک‌مارک[۱]

مریم پارساییان[۲]، ابراهیم خدایی[۳]*، بلال ایزانلو[۴]، کیوان صالحی[۵]، سیما نقی‌زاده[۶]

(دریافت: ۱۴۰۲/۱۰/۲۷ ـ پذیرش: ۱۴۰۳/۰۲/۲۵)

چکیده

هدف پژوهش حاضر این است که با به کارگیری رگرسیون چندگانه در روش هارویچ که یک رویکرد تصمیم‌گیری است نمره برش آزمون تولیمو برای داوطلبان دکتری متقاضی فرصت‌های مطالعاتی خارج از کشور را تعیین کند. جامعه آماری این پژوهش شامل داوطلبان یک دوره از آزمون تولیمو است که ۴۶۱ نفر بودند. روش پژوهش این مطالعه مبتنی بر رویکرد کمی است و از نظر هدف جزء پژوهش‌های کاربردی و از نظر تحلیلی جزء مطالعات تحلیل ثانویه محسوب می‌شود و از نمره‌های همه ۴۶۱ داوطلب به هر کدام از بخش‌های گرامر، درک مطلب، شنیداری و نوشتاری برای تعیین نمره برش استفاده می‌گردد. بر اساس نتایج مطالعه و با توجه به مقدار خطای طبقه‌بندی ناچیز، در صورت استفاده از نمره کل، میانگین مقادیر شاخص هارویچ به نمره برش ۴۸۱ و در صورت معنی‌دار بودن مدل رگرسیون خطی با به کارگیری نتایج شاخص هارویچ در رگرسیون چندگانه، از نمره برش ۴۹۴ به عنوان نمره برش می‌توان استفاده کرد. روش پیشنهادی به طور قابل توجهی دقت تعیین نمره برش را در مقایسه با روش‌های آنگوف و بوک‌مارک بهبود می‌بخشد. با توجه یافته‌های به دست آمده از اعتبارسنجی متقابل و مقادیر خطا، نمره برش به‌دست‌آمده از استراتژی ترکیبی منجر به نتایج دقیق‌تر و با خطای طبقه‌بندی صفر می‌شود.

واژه‌های کلیدی: نمره برش، روش هارویچ، رگرسیون چندگانه، روش آنگوف، روش بوک‌مارک، اعتبارسنجی متقابل، آزمون تولیمو

۱ - مقاله حاضر برگرفته از رساله دکتری مریم پارساییان دانشجوی دوره دکتری رشته سنجش و اندازه گیری دانشگاه تهران می‌باشد.

۲ – دانشجوی دکتری سنجش و اندازه‌گیری، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران.

۳ – دانشیار، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران.

* – نویسنده مسئول: khodaie@ut.ac.ir

۴ – استادیار ، دانشکده روانشناسی و علوم تربیتی ، دانشگاه خوارزمی، تهران، ایران.

۵ – دانشیار، دانشکدهٔ روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران.

۶ – استادیار، سازمان سنجش آموزش کشور، تهران، ایران.

مقدمه

نمره برش نشان‌دهنده یک نقطه تصمیم‌گیری است که در آن تصمیم برای پذیرش یا رد یک آزمودنی بر اساس مقایسه نمره به دست آمده فرد با نمره برش تعیین شده حاصل می‌شود. بنابراین نمره برش مقدار آستانه‌ای را نشان می‌دهد که پذیرش و عدم پذیرش در آزمون را مشخص می‌کند. بدین صورت که هر مقدار بیشتر از نمره برش، به معنای قبولی در آزمون و هر مقدار کمتر از نمره برش به معنای عدم قبولی در آزمون در نظر گرفته می‌شود. در تعیین نمره برش برای طبقه‌بندی آزمودنی‌ها، دو نوع خطا رخ می‌دهد. ۱) خطای مرتبط با نمره برش که خطای اندازه‌گیری[1] یا خطای استاندارد[2] نامیده می‌شود و ۲) خطای مرتبط با تصمیم‌گیری بر اساس نمره برش که خطای طبقه‌بندی[3] نامیده می‌شود (برک[4]، ۱۹۸۶). خطای طبقه‌بندی به معنای پذیرش دانش‌آموزان ناکارآمد و رد دانش‌آموزان کارآمد است (لین، روبرتس و خوانا[5]، ۲۰۲۰). اگر نمره برش خیلی بالا تعیین شود، دانش‌آموزانی که واقعاً شایسته قبولی هستند شکست خواهند خورد و برعکس اگر نمره برش خیلی پایین تعیین شوند، دانش‌آموزانی که واقعاً مستحق شکست هستند قبول خواهند شد. بالا یا پایین بردن نمره برش برای کاهش یک نوع خطا، لزوماً احتمال خطای نوع دیگر را افزایش می‌دهد. به عنوان مثال می‌توان تعداد دانش‌آموزانی را که قبول می‌شوند، اما واقعاً مستحق شکست هستند را با افزایش نمره برش کاهش داد که نتیجه انجام این کار افزایش تعداد دانش‌آموزانی می‌شود که شکست می‌خورند اما واقعاً شایسته قبولی هستند. تدوین آزمون خوب و شیوه‌های خوب برای تعیین نمره‌های برش می‌تواند تعداد خطاهای طبقه‌بندی را کاهش دهد، اما هیچ راهی برای کاهش خطاها به صفر وجود ندارد (گرابواسکای و وینر[6]، ۲۰۱۷).

گاهی اوقات تشخیص آسیب نسبی ناشی از این دو نوع خطا آسان است. به عنوان مثال در آزمون صدور گواهی‌نامه برای خلبانی بدیهی است که قبولی در آزمونی که مستحق رد شدن است، به وضوح مضرتر از قبول نشدن در آزمونی است که مستحق قبولی است. با این حال در بیشتر محیط‌های دانشگاهی، تعیین آسیب نسبی ناشی از این دو نوع خطا بسیار دشوارتر است. به عنوان مثال برخی افراد می‌گویند که استفاده از استانداردهای سختگیرانه‌ای که دانش‌آموزان حاشیه‌ای را رد می‌کند، صرفاً دانش‌آموزان را به دلیل ناکامی در سیستم آموزشی تنبیه می‌کند و ضرر شکست دانش‌آموزان بسیار بیشتر از فایده آن است. برخی دیگر می‌گویند که اعمال استانداردهای دقیق تنها راه بهبود مدارس است و قبولی دانش‌آموزان حاشیه‌ای که ممکن است فاقد مهارت‌های مهم باشند برای دانش‌آموزان و جامعه مضر است. حق با کدام

1- Measurement error
2- Standard Error
3- Classification error
4- Berk
5- Lane, Roberts, & Khanna
6- Grabovsky, & Wainer

گروه است؟ به طور کلی افرادی که در تعیین نمره‌های برش عملیاتی نقش دارند باید هر دو نوع خطا را در قضاوت خود در نظر بگیرند و تصمیم بگیرند که کدام نوع خطا را مضرتر بدانند. در واقع نمره برش باید نوع مضرتر خطا را کاهش دهد (زیکی و پری[1]، بی.تا.). توجه داشته باشید که اگر یکی از انواع خطاها آسیبی به همراه نداشته باشد، نیازی به تعیین نمره‌های برش وجود ندارد. برای مثال اگر قبولی دانش‌آموزانی که مستحق شکست هستند به هیچ وجه آسیبی به همراه نداشته باشد، بهترین استراتژی صرفاً قبولی همه دانش‌آموزان است.

بنابراین تصمیم‌گیری به عنوان یک فرآیند پیچیده ذهنی، یک برنامه حل مسئله است که هدف آن تعیین نتیجه مطلوب با توجه به جنبه‌های مختلف است. این فرآیند می‌تواند عقلانی یا غیرمنطقی باشد و از سوی دیگر می‌تواند از مفروضات ضمنی یا صریح استفاده کند که تحت‌تأثیر عوامل متعددی از جمله فیزیولوژیکی، بیولوژیکی، فرهنگی، اجتماعی و غیره باشد. مسائل پیچیده تصمیم‌گیری را می‌توان با استفاده از معادلات ریاضی، آمار چندگانه، ریاضیات، تئوری‌های اقتصادی و دستگاه‌های کامپیوتری حل کرد که به محاسبه و تخمین راه‌حل‌های مسائل تصمیم‌گیری به صورت خودکار کمک می‌کنند (طاهردوست و معدنچیان[2]، ۲۰۲۳).

لذا تعیین استاندارد جنبه جدایی‌ناپذیر هر سیستم ارزیابی است که طیفی از ذینفعان از جمله سیاست‌گذاران، تدوین‌کنندگان آزمون و متخصصان اندازه‌گیری را درگیر می‌کند تا اطمینان حاصل شود که نتایج آزمون معنادار و قابل دفاع خواهد بود چرا که به طور قابل توجهی بر تصمیم‌گیری در مورد پیشرفت آزمودنی‌ها به مرحله بعدی آموزش تأثیر می‌گذارد (لین، روبرتس و خوانا، ۲۰۲۰). در این مطالعه، ما رویکرد جدیدی برای تعیین نمرات برش آزمون با استفاده از رگرسیون چندگانه در روش هارویچ مورد بحث قرار داده و از نظر تجربی بررسی خواهیم کرد. این روش، عدم قطعیت و ذهنیت موجود در روش‌های سنتی را در نظر می‌گیرد و اعتبار و عادلانه بودن نتایج آزمون را بهبود می‌بخشد به طوری که می‌تواند نقش بسزایی در تعیین معیارهای قبولی آزمون‌های مهارت زبان مانند آزمون تولیمو در ایران داشته باشد. هدف ما از استفاده از این رویکرد، ارائه ارزیابی دقیق‌تر و مطمئن‌تر از عملکرد شرکت‌کنندگان است که می‌تواند به مربیان و برنامه‌ریزان آموزش عالی در تصمیم‌گیری آگاهانه در مورد فرصت‌های آموزشی دانشجویان کمک کند.

مبانی نظری و پیشینه پژوهش: تاکنون روش‌های زیادی برای تعیین نمره برش ایجاد شده‌اند ولی دو روشی که به طور گسترده به خصوص در حوزه آموزش مورد استفاده قرار گرفته‌اند روش‌های آنگوف[3] (۱۹۷۱) و بوک‌مارک[4](یا نشانک)(۱۹۹۶) هستند.

1- Zieky, & Perie
2- Taherdoost & Madanchian
3- Angoff
4- Bookmark

روش آنگوف یکی از طولانی‌ترین و پایدارترین روش‌های تعیین استاندارد در میان روش‌هایی است که توسط متخصصان در تجزیه و تحلیل محتویات ابزار آزمون برای تعیین استاندارد استفاده می‌شود که توسط آنگوف در سال ۱۹۷۱ پیشنهاد شد. روش آنگوف شامل تعریف و توصیف استاندارد عملکرد، برآورد احتمال یک آزمودنی فرضی یا گروهی از آزمودنی‌ها که به هر سؤال آزمون پاسخ درستی می‌دهند و یا ارزیابی در سطح استاندارد عملکرد با جمع برآوردهای سؤال برای هر داور و محاسبه میانگین نمرات حاصل شده داوران است که نتیجه این میانگین، مقدار نمره برش است (براندون[1]، ۲۰۰۴). به عبارت دیگر در روش آنگوف، اعضای پانل بایستی همه سؤال‌ها را ارزیابی کنند و نسبت آزمودنی‌ها با حداقل مهارت را که می‌توانند به درستی به هر سؤال پاسخ دهند را تخمین بزنند. مجموع میانگین رتبه[2] برای هر سؤال به عنوان یک نمره برش تعیین می‌شود (کیم[3] و همکاران، ۲۰۱۷). مشکلات اصلی این روش عبارتند از (الف) مفهوم‌سازی آزمون‌شونده مرزی و (ب) در مورد پاسخ صحیح به هر سؤال آزمون توسط آزمون‌شونده مرزی بایستی قضاوت احتمالی انجام شود (دیمیترو[4]، ۲۰۲۲).

در سال ۱۹۹۶ برای غلبه بر برخی محدودیت‌های روش آنگوف، لوئیس، میتزل و گرین[5] روش بوک‌مارک را معرفی کردند. روش بوک‌مارک مبتنی بر نظریه سؤال پاسخ است. یعنی چارچوبی که مهارت امتحان‌شوندگان و دشواری سؤال‌های آزمون را به طور همزمان مشخص می‌کند (لین[6]، ۲۰۰۶). در این روش، یک احتمال پاسخ[7] (RP) از پیش تعیین می‌شود و سپس سؤال‌های آزمون براساس دشواری برآورد شده در نظریه سؤال پاسخ به ترتیب صعودی در یک دفترچه سؤال[8](OIB) قرار داده می‌شوند و از اعضای پانل خواسته می‌شود که یک "نشانک" را در نقطه‌ای بین سخت‌ترین سؤالی که آزمون‌شونده مرزی در آزمون احتمالاً به درستی پاسخ می‌دهند و ساده‌ترین سؤالی که آزمون‌شونده مرزی احتمالاً به درستی پاسخ نمی‌دهند قرار دهند (زیکی و پری، بی.تا.). برای RP مقادیر مختلفی از جمله ۰/۵، ۰/۶۷ و ۰/۸ در نظر گرفته شده است (وانگ[9]، ۲۰۰۳) ولی اغلب مقدار RP=0/67 قرار داده می‌شود بدین‌معنی که ۶۷ درصد شانس وجود دارد که آزمودنی مرزی به یک سؤال، پاسخ درست دهد. نمره برش نقطه‌ای در مقیاس نظریه سؤال پاسخ است که با RP انتخابی یک پاسخ صحیح برای سؤال درست قبل از نشانک مطابقت دارد (در برخی موارد، نمره برش برابر با نقطه میانی بین سؤال نشانه‌گذاری شده و سؤال قبلی تعیین می‌شود).

1- Brandon
2- Rating
3- Kim
4- Dimitrov
5- Lewis, Mitzel, and Green
6- Lin
7- Response Propability
8- Ordered Item Booklet
9- Wang

بنابراین مشکلات اصلی در روش نشانک مربوط به مفهوم‌سازی آزمون‌شونده مرزی، انتخاب یک مقدار RP، قضاوت احتمال برای قرار دادن نشانک و تمرکز محدود بر دشواری سؤال است (دیمیترو، ۲۰۲۲).

مفهوم عملکرد مرزی در تعیین نمره‌های برش با استفاده از روش‌های آنگوف و بوک‌مارک بسیار مهم است. به عنوان مثال برای تعیین اینکه کدام دانش‌آموز مهارت دارد، باید دانش‌آموزان پایه را از افراد مسلط و دانش‌آموزان مسلط را از افراد پیشرفته تشخیص داد. بنابراین لازم است بین بهترین عملکرد دانش‌آموزی که هنوز در سطح پایه است و بدترین عملکرد دانش‌آموزی که هنوز در سطح تسلط است تمایز قائل شد. به طور مشابه لازم است بین بهترین عملکرد دانش‌آموزی که هنوز در سطح تسلط است و بدترین عملکرد دانش‌آموزی که هنوز در سطح پیشرفته است تمایز قائل شد. تمرکز بر دانش‌آموز متوسط در یک دسته، کمکی به ایجاد تمایزات لازم نمی‌کند. در واقع تمرکز باید در نقطه‌ای باشد که دانش‌آموز با بهترین عملکرد در یک رده از دانش‌آموز با بدترین عملکرد در رده بالاتر بعدی قابل تشخیص نباشد یا نقطه‌ای باشد که دانش‌آموز با بدترین عملکرد در یک دسته از دانش‌آموز با بهترین عملکرد در دسته پایین‌تر بعدی قابل تشخیص نباشد. در شرایط قبول و رد، قضاوت باید بر مرز بین دانش‌آموزی که بهترین عملکرد را دارد و هنوز شایستگی شکست را دارد و دانش‌آموزی با بدترین عملکرد که هنوز هم شایسته قبولی است، متمرکز شود (زیکی و پری، بی.تا.).

علی‌رغم استفاده گسترده از روش‌های آنگوف و بوک‌مارک در عمل، دارای محدودیت‌هایی نیز هستند. به عنوان مثال، یکی از محدودیت‌های روش آنگوف این است که در روش آنگوف اصطلاح نامزد مرزی در ادبیات و در عمل به طور دقیق تعریف نشده است که می‌تواند منجر به متفاوت بودن و عدم اعتبار نتایج آزمون شود چرا که تصمیم‌گیری بر اساس قضاوت کارشناسان انجام می‌گیرد (لین، روبرتس و خوانا، ۲۰۲۰). از سوی دیگر از جمله محدودیت‌های روش بوک‌مارک انتخاب احتمال پاسخ، ناهماهنگی سؤال‌ها، حذف عوامل مهم غیر از دشواری سؤال‌ها و نیز مدل‌های نظریه سؤال پاسخ[1] است (لین، ۲۰۰۶). تشخیص مکان دقیق نشانک را نیز می‌توان به عنوان محدودیت دیگر روش بوک‌مارک بیان کرد چرا که ممکن است کارشناسان نظرات متفاوتی در مورد قرار دادن آن داشته باشند (زیکی و پری، بی.تا.). به غیر از این محدودیت‌ها که برای هر کدام از این روش‌ها وجود دارد می‌توان به محدودیت‌هایی مانند ۱) زمان‌بر و هزینه‌بر بودن روش‌های سنتی نیز اشاره کرد. زیرا یکی از راه‌های افزایش دقت در روش‌های سنتی در شناسایی افرادی که صلاحیت پذیرش را دارند این است که بایستی تعداد زیادی از کارشناسان واجد شرایط درگیر شوند تا اطمینان معقولی حاصل شود که رتبه‌بندی کارشناسان به اندازه کافی قابل اعتماد است و اگر این روند تکرار شود، نتایج قضاوت‌ها تفاوت زیادی نخواهد داشت (انجمن تحقیقات آموزشی آمریکا، ۲۰۱۸) که استخدام کارشناسان برای قضاوت در مورد سؤال‌های آزمون می‌تواند پرهزینه باشد. به علاوه در صورتی

1- Item Response Theory (IRT)

که سؤال‌های آزمون زیاد باشد فرآیند قضاوت در مورد سؤال‌های آزمون توسط کارشناسان زمان‌بر می‌شود که همین امر می‌تواند باعث عدم دقت آنها شود. ۲) از آنجایی که وضعیت واقعی یا نمره واقعی هر آزمون‌دهنده مشخص نیست و فقط نمره‌های مشاهده شده آنها موجود است و قضاوت کارشناسان موضوع در تعیین نمره برش و وضعیت آزمون‌دهنده‌ها تأثیر می‌گذارد لذا این خطوط تقسیم ممکن است بهترین انتخاب برای نمره برش به منظور به حداقل رساندن خطا نباشد (گرابواسکای و وینر، ۲۰۱۷). بنابراین روش‌های سنتی به شدت بر قضاوت متخصص متکی هستند که می‌تواند معرف ذهنیت و تنوع در پانل‌های مختلف متخصصان باشند. ۳) موضوع رانش[1] مفهومی در طول فرآیند تنظیم استاندارد است. بدین معنی که آیا مفاهیم حداقل شایستگی داوران در کل فرآیند تنظیم استاندارد یکسان باقی می‌ماند یا تصور داوران تحت‌تأثیر عواملی مانند قرار گرفتن در معرض سؤال‌های آزمون، بحث در میزگرد یا خستگی است؟ رانش یک مشکل بالقوه در هر روش قضاوتی است، به ویژه زمانی که نیازهای شناختی وظیفه تنظیم نمره برش بالا باشد (ریکر[2]، ۲۰۰۶). ۴) روش قرار دادن اعضای پانل در معرض نظارت کارشناسان، فرآیندهای مقایسه اجتماعی را به دنبال خواهد داشت که منجر به تجدیدنظر در نظرات اعضای گروه می‌شود و این تأثیر زمانی بیشتر می‌شود که نظر اجماع گروه به جای توزیع نظرات اعضاء به یکی از اعضاء معطوف شود. به عبارت دیگر کارشناسان ممکن است تعصبات یا ترجیحات خاص خود را داشته باشند که می‌تواند بر قضاوت آنها در مورد سؤال‌های آزمون تأثیر بگذارد (برک، ۱۹۸۶). ۵) استفاده از روش‌های سنتی برای تعیین نمرات برش در آزمون‌هایی مانند آزمون تولیمو که در آن سؤال‌های آزمون از یک بانک سؤال از پیش تعیین‌شده تولید می‌شوند می‌تواند توسط داورها به خطر بیفتد. زیرا آنها باید دارای تخصص موضوعی باشند و اگر به عنوان مدرس نیز خدمت کنند، ممکن است سؤال‌ها را در اختیار داشته باشند و آنها را بین دانشجویان خود توزیع کنند یا اینکه در بهترین وضعیت ممکن، نکات یا سرنخ‌هایی برای پاسخ‌های خاص به دانشجویان خود ارائه دهند که این می‌تواند منجر به برتری ناعادلانه برخی از آزمون‌شوندگان نسبت به دیگران شود و در نتیجه اعتبار کلی نتایج آزمون کاهش یابد.

برای رفع این محدودیت‌ها که نوعی مقابله با چالش‌های تصمیم‌گیری در شرایط عدم قطعیت محسوب می‌شوند، محققان و متخصصان روش‌ها و تکنیک‌های مختلفی را توسعه داده‌اند. هدف این روش‌ها بهبود کیفیت تصمیم‌گیری از طریق کاهش تأثیر عدم قطعیت بر فرآیند تصمیم‌گیری است. به عبارت دیگر تصمیم‌گیری زمانی چالش‌برانگیزتر می‌شود که اطلاعات موجود ناقص، مبهم یا مشمول تفاسیر متعدد باشند. در چنین شرایطی تصمیم‌گیرندگان باید بر قضاوت و تجربه خود تکیه کنند تا بهترین تصمیم ممکن را بگیرند. این نوع تصمیم‌گیری در شرایط عدم قطعیت در بسیاری از زمینه‌ها از جمله تجارت، اقتصاد، سیاست و مراقبت‌های بهداشتی رایج است. به عنوان مثال مدیران کسب و کار باید در مورد سرمایه‌گذاری،

1- Drift
2- Ricker

توسعه محصول و گسترش بازار در مواجهه با شرایط نامشخص اقتصادی و فشارهای رقابتی تصمیم بگیرند (پرز[1] و همکاران، ۲۰۱۵). بنابراین افراد تصمیم‌گیرنده بایستی روش‌های تصمیم‌گیری مناسبی را اتخاذ کنند که بتوانند به طور مؤثر عدم اطمینان را مدیریت کنند و بهترین تصمیم را اخذ نمایند.

در شرایط عدم قطعیت می‌توان از روش‌های مختلفی از جمله مدل‌سازی احتمالی[2]، درخت تصمیم[3]، شبیه‌سازی مونت کارلو[4] و تکنیک‌های تصمیم‌گیری چند معیاره[5] استفاده کرد. مدل‌سازی احتمالی شامل کمی کردن عدم قطعیت در تصمیم‌گیری است که در آن به نتایج مختلف بر اساس اطلاعات موجود، احتمال تخصیص داده می‌شود. درخت تصمیم، تصویری واضح از مساله تصمیم‌گیری در شرایط عدم قطعیت را ارائه می‌دهد و به تصمیم‌گیرندگان کمک می‌کند تا تصمیم بهینه را اخذ کنند. در واقع با استفاده از درخت تصمیم می‌توان بهترین مسیر اقدام را با در نظر گرفتن تمام نتایج ممکن و احتمالات مرتبط با آنها شناسایی نمود. درخت تصمیم به ویژه زمانی مفید است که چندین تصمیم و نتایج ممکن وجود داشته باشد و احتمالات هر نتیجه به خوبی تعریف نشده باشند (هان[6] و همکاران، ۲۰۱۱). شبیه‌سازی مونت کارلو تکنیکی است که در تصمیم‌گیری در شرایط عدم قطعیت استفاده می‌شود که شامل استفاده از مدل‌های آماری برای شبیه‌سازی نتایج احتمالی مختلف یک تصمیم است که یک روش محاسباتی است که از نمونه‌گیری تصادفی برای تولید تعداد زیادی سناریو استفاده می‌کند که هر کدام نشان‌دهنده نتیجه احتمالی یک تصمیم است. شبیه‌سازی مونت‌کارلو به ویژه زمانی مفید است که متغیرهای نامشخص زیادی در مسئله تصمیم‌گیری دخیل باشند (راس[7]، ۲۰۱۳). تصمیم‌گیری چند معیاره یکی از مسائل اصلی تصمیم‌گیری است که هدف آن تعیین بهترین گزینه با در نظر گرفتن بیش از یک معیار در فرآیند انتخاب است. روش‌های تصمیم‌گیری چندمعیاره دارای ابزارها و روش‌های متعددی است که می‌توانند در زمینه‌های مختلف به کار گرفته شوند (طاهردوست و معدنچیان، ۲۰۲۳). همانطور که تصمیم‌گیری با بزرگ شدن مسائل به طور فزاینده‌ای پیچیده می‌شود، روش‌های تصمیم‌گیری چندمعیاره به عنوان ابزار قدرتمندی برای انتخاب‌های آگاهانه ظاهر شده‌اند. در واقع یکی از کاربردهای مهم روش‌های تصمیم‌گیری چندمعیاره در تعیین نمره‌های برش است تا مشخص شود که یک فرد، سازمان یا محصول دارای معیارهای خاصی است یا خیر.

نتیجه انتخاب تصمیم‌گیرنده در شرایط عدم قطعیت به دو عامل بستگی دارد: کدام گزینه انتخاب خواهد شد و کدام سناریو (وضعیت طبیعی) در آینده رخ خواهد داد. پیامد هر تصمیمی نه تنها توسط خود تصمیم،

1- Pérez
2- Probabilistic modeling
3- Decision tree
4- Monte Carlo simulation
5- Multiple Criteria Decision Making (MCDM)
6- Han
7- Ross

بلکه توسط یک عامل خارجی تعیین می‌شود که خارج از کنترل تصمیم‌گیرنده است. جدول ۱ بیانگر تصمیم‌گیری در شرایط عدم قطعیت است که در آن سطرها بیانگر گزینه‌ها (یا سناریوها) و ستون‌ها بیانگر ویژگی‌های مربوط به هر گزینه می‌باشند. در جدول ۱، مقادیر a_{ij} بیانگر مقدار مرتبط با گزینه i و ویژگی j است. روش‌های ماکسیماکس[1]، مینیماکس[2] ، ساواج[3] ، لاپلاس[4] و هارویچ[5] همگی روش‌های تصمیم‌گیری چندمعیاره در شرایط عدم قطعیت هستند که به تصمیم‌گیرندگان کمک می‌کنند گزینه‌ها را ارزیابی کرده و در موقعیت‌هایی که ابهام، ریسک یا اطلاعات ناقص وجود دارد، تصمیم بگیرند. به عبارت دقیق‌تر هدف از استفاده از این‌گونه روش‌های تصمیم‌گیری چندمعیاره کمک به تصمیم‌گیرندگان برای تصمیم‌گیری آگاهانه و منطقی در موقعیت‌های نامشخص است.

جدول ۱: ماتریس تصمیم
Table 1
Decision matrix

ویژگی m Feature m	...	ویژگی ۲ Feature2	ویژگی ۱ Feature1	
a_{1m}	⋯	a_{12}	a_{11}	گزینه ۱ Option 1
a_{2m}	⋯	a_{22}	a_{21}	گزینه ۲ Option 2
⋮	⋯	⋮	⋮	⋮
a_{nm}	⋯	a_{n2}	a_{n1}	گزینه n Option n

روش ماکسیماکس به دنبال به حداکثر رساندن حداکثر بازده ممکن است. در این روش فرض بر این است که تصمیم‌گیرنده خوش‌بین است و مایل است برای دستیابی به بالاترین پاداش ممکن ریسک کند. روش ماکسیماکس گزینه‌ای را با بالاترین حداکثر نتیجه ممکن انتخاب می‌کند. به عبارت دیگر؛

$$m_j^* = \max_j\{m_j\} = \max_j \max_i\{a_{ij}\}$$

روش مینیماکس به دنبال به حداقل رساندن حداکثر ضرر ممکن است. در این روش فرض بر این است که تصمیم‌گیرنده بدبین است و تمایلی به ریسک ندارد و می‌خواهد از بدترین نتیجه ممکن اجتناب کند. معیار مینیماکس گزینه‌ای را با کمترین حداکثر ضرر ممکن انتخاب می‌کند. به عبارت دیگر؛

$$w_j^* = \min_j\{w_j\} = \min_j \max_i\{a_{ij}\}$$

1- Maximax
2- Minimax
3- Savage
4- Laplace
5- Hurwicz

معیار پشیمانی مینیماکس ساواج[1] و یا معیار ساواج به دنبال به حداقل رساندن حداکثر پشیمانی است. منظور از پشیمانی تفاوت بین سود بهترین گزینه و گزینه انتخاب شده است. معیار ساواج گزینه‌ای را انتخاب می‌کند که حداکثر پشیمانی را به حداقل می‌رساند. ساواج (۱۹۶۱) پیشنهاد کرد که ماتریس سود با یک جدول پشیمانی جدید که طبق فرمول (۱) محاسبه شده است جایگزین شود و به هر تصمیم بر اساس معادله (۲) یک شاخص اختصاص داده شود که نشان‌دهنده بدترین پشیمانی از گزینه j ام است:

$$r_{ij} = \max_j\{a_{ij}\} - a_{ij} \quad (1)$$

$$s_j = \max_i\{r_{ij}\} \quad (2)$$

که r_{ij} به معنای از دست دادن فرصت غیرمنفی است. در معیارساواج، توصیه می‌شود تصمیمی انتخاب شود که تا حد امکان دارای شاخص پایین باشد:

$$s_j^* = \min_j\{s_j\} = \min_j \max_i\{r_{ij}\}$$

روش لاپلاس یک معیار احتمالی برای تصمیم‌گیری در شرایط عدم قطعیت است که ارزش مورد انتظار هر گزینه را محاسبه می‌کند. معیار لاپلاس فرض می‌کند که همه حالت‌های طبیعت به یک اندازه محتمل هستند و وزن‌های برابری برای آنها قائل است. در واقع لاپلاس استدلال کرد که اگر کسی چیزی در مورد سناریوی واقعی نداند، این بدان معناست که همه حالت‌ها احتمال برابری دارند (رندر[2] و همکاران، ۲۰۱۶). معیار لاپلاس گزینه‌ای را انتخاب می‌کند که بالاترین میانگین بازدهی را در تمام حالت‌های طبیعت داشته باشد. به عبارت دیگر تصمیم‌گیرنده با مقدار مورد انتظار انتخاب خود روبرو می‌شود و باید گزینه‌ای را انتخاب کند که در معادلات (۳) و (۴) صدق کند:

$$l_j = \frac{1}{m}\sum_i a_{ij} \quad (3)$$

$$l_j^* = \max_j\{l_j\} \quad (4)$$

استفاده از ارزش‌های مورد انتظار، این رویکرد را از معیارهایی متمایز می‌کند که فقط از سودهای شدید استفاده می‌کنند. این ویژگی، رویه را شبیه به تصمیم‌گیری تحت ریسک می‌کند (پازک و روزمان[3]، ۲۰۰۹).

معیار هارویچ توسط لئونید هارویچ[4] (۱۹۵۱؛ ۱۹۵۲)، برنده جایزه نوبل در اقتصاد ایجاد شد؛ نوعی روش تصمیم‌گیری است که به یافتن نمره برش در یک سیستم امتیازدهی کمک می‌کند. هارویچ در مدل واقع‌گرایی که بیشتر به عنوان معیار هارویچ شناخته می‌شود استدلال کرد که تصمیم‌گیرنده باید گزینه‌ها را بر اساس میانگین وزنی سطوح امنیت و خوش‌بینی رتبه‌بندی کند.

1- Savage Minimax Regret Criterion
2- Render
3- Pažek, & Rozman
4- Leonid Hurwicz

شاخص هارویچ را می‌توان به عنوان میانگین وزنی بهترین و بدترین تحقق عدم قطعیت در نظر گرفت که به تصمیم‌گیرنده این امکان را می‌دهد تا با بیان یک "ضریب خوش‌بینی" که تأکید بر بهترین نتیجه و یا یک "ضریب بدبینی" که بیانگر بدترین نتیجه است به طور همزمان بهترین و بدترین نتایج ممکن را در نظر بگیرد (اون[1]، ۲۰۱۲).

معیار هارویچ شامل محاسبه میانگین وزنی بهترین نتیجه ممکن (حداکثر) و بدترین نتیجه ممکن (حداقل) برای هر گزینه در یک سناریوی تصمیم‌گیری است و وزنی که به هر نتیجه داده می‌شود با سطح خوش‌بینی یا بدبینی تصمیم‌گیرندگان تعیین می‌شود (هارویچ، ۱۹۵۲). به عبارت دقیق‌تر برای اعمال شاخص هارویچ لازم است مراحل زیر طی شود:

۱) بهترین نتیجه ممکن و بدترین نتیجه ممکن برای هر گزینه تعیین می‌شود.

۲) تعیین نگرش تصمیم‌گیرنده نسبت به ریسک که مقداری بین ۰ و ۱ را می‌گیرد که مقادیر نزدیک به ۱ نشان‌دهنده خوش‌بینی بیشتر (تصمیم‌گیرندگان خوش‌بین افراطی یا تصمیم‌گیرندگان ریسک‌پذیر) و مقادیر نزدیک به ۰ نشان‌دهنده بدبینی بیشتر (تصمیم‌گیرندگان محتاط) است و با α نشان داده می‌شود.

۳) شاخص هارویچ بر اساس رویکرد خوش‌بینانه از رابطه (۵) محاسبه می‌شود:

$$h_j = \alpha \max_i\{a_{ij}\} + (1-\alpha)\min_i\{a_{ij}\} \qquad (5)$$

۴) گزینه بهینه برابر با گزینه‌ای است که دارای مینیمم مقدار h_j باشد (گاس‌پارس- وای‌لوچ[2]، ۲۰۱۴).

روش پژوهش

این مطالعه با هدف تعیین نمره برش مناسب برای داوطلبان متقاضی فرصت‌های مطالعاتی از دانشگاه‌ها و رشته‌های مختلف انجام شده است که بایستی حد نصابی را در آزمون زبان کسب کنند. لذا داده‌های مورد استفاده در این پژوهش شامل نمرات ۴۶۱ داوطلب شرکت‌کننده در یک دوره آزمون زبان تولیمو است که توسط سازمان سنجش آموزش کشور طراحی و اجرا شده است. در آزمون تولیمو چهار مهارت گرامر، درک مطلب، درک شفاهی و درک نوشتاری مورد ارزیابی قرار می‌گیرند. سطوح عملکرد داوطلبان بر اساس قضاوت و نظر کارشناسان به عالی، خوب، متوسط و ضعیف طبقه‌بندی شده است. محدوده نمرات آزمون تولیمو از ۳۱۰ تا ۶۷۷ تعریف شده است و نمره قبولی تا قبل از سال ۱۴۰۰ برابر ۴۸۰ و از ۱۴۰۰ به بعد برابر ۵۰۰ تعیین شده است. از آنجایی‌که اعتبار روش‌های سنتی تعیین نمره برش مانند آنگوف و بوک‌مارک با افزایش تعداد اعضای پانل و تکرار فرآیند ارزیابی افزایش می‌یابد لذا این مطالعه با استفاده از رویکردی جایگزین برای روش‌های سنتی، بینشی در تعیین یک نمره برش مناسب برای داوطلبان متقاضی

1- Ion
2- Gaspars-Wieloch

فرصت‌های مطالعاتی بر اساس نمره کل و نمره‌های بخش‌های مختلف آزمون با به کارگیری نتایج مربوط به معیار هارویچ در رگرسیون چندگانه را فراهم می‌کند. به منظور اصلاح روش هارویچ، پارامتر α بر اساس اعضای پانل بسیار خوش‌بین و بسیار بدبین تعریف می‌شوند و سپس معیار هارویچ برای α بهینه و یا میانگین مقادیر معیار هارویچ برای مقادیر مختلف α محاسبه می‌شود. برای دستیابی به هدف موردنظر، دو نوع استراتژی به کار برده می‌شود: استراتژی اول تنها بر اساس نمرات کل انجام می‌شود که به نام استراتژی خالص[1] نامیده شده و استراتژی دوم که به نام استراتژی ترکیبی[2] خوانده شده بر اساس نمرات مؤلفه‌های مختلف آزمون در یک رویکرد خوش‌بینانه برای به حداقل رساندن مقدار شاخص هارویچ انجام می‌شود.

لازم به ذکر است با توجه به عدم دسترسی به نحوه ساخت نمره کل آزمون تولیمو، در این پژوهش برای تعیین میزان تأثیر هر یک از بخش‌های آزمون در نمره کل از رگرسیون چندگانه برای روش‌های مطرح شده در استراتژی ترکیبی استفاده گردید.

در نهایت، یک مطالعه اعتبارسنجی متقابل[3] برای انتخاب راهکارهایی انجام می‌شود که نمره برش حاصل شده آنها بین ۴۸۰ تا ۵۰۰ باشد. داده‌ها در اعتبارسنجی متقابل به طور تصادفی به پنج زیرمجموعه مساوی و دو به دو ناسازگار تقسیم می‌شوند. فرآیند آموزش و آزمون پنج بار اجرا می‌شود بدین‌صورت که در هر اجرا، یک بخش به عنوان آزمون و سایر بخش‌ها به عنوان آموزش در نظر گرفته می‌شوند. میانگین نتایج پنج آزمون محاسبه شده و استراتژی تشکیل‌دهنده با بالاترین دقت انتخاب می‌شود. بعد از تأیید اعتبار مدل‌ها، بهترین مدل تعیین نمره برش از طریق محاسبه خطای استاندارد و خطای طبقه‌بندی برگزیده می‌شود. به منظور محاسبه خطاها داده‌ها به دو گروه آموزشی و آزمون تقسیم می‌شوند و نمره برش حاصل از داده‌های آزمون در داده‌های آموزشی بررسی می‌شود و در نهایت با فرض اینکه P و N به ترتیب بیانگر تعداد قبول‌شده‌ها و تعداد رد شده‌ها باشند خطای طبقه‌بندی (CE) که بیانگر نسبت داده‌هایی است که به غلط پیش‌بینی شده‌اند از رابطه (۶) محاسبه می‌شود.

$$CE = \frac{FP + FN}{P + N} \qquad (6)$$

که در آن FP بیانگر افرادی است که به غلط قبول شده‌اند یعنی افرادی هستند که صلاحیت قبول شدن نداشتند ولی الگوریتم آنها را به اشتباه در دسته قبول شده‌ها قرار داده است و FN بیانگر افرادی است که به غلط رد شده‌اند یعنی افرادی هستند که صلاحیت قبول شدن داشتند ولی الگوریتم آنها را به اشتباه در دسته رد شده‌ها قرار داده است.

1- Pure strategy
2- Combined strategy
3- K-fold cross-validation

خطای استاندارد (SE) از طریق جذر میانگین مربعات خطای داده‌های آزمون با استفاده از رابطه (۷) به دست آورده می‌شود.

$$SE = \sqrt{MSE} = \sqrt{\frac{p}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

که در آن n و p به ترتیب بیانگر تعداد داده‌های آزمون و تعداد متغیرهای مستقل هستند.

تجزیه و تحلیل داده‌ها در پایتون[1] که یک زبان برنامه نویسی رایگان و محبوب برای محاسبات علمی است انجام شد. چندین بسته برای اجرا و دستیابی به نتایج از جمله pandas ، numpy، scikit-learn استفاده گردید. pandas یکی از کتابخانه‌های مهم پایتون است که تعریف و فراخوانی داده‌ها از فایل اکسل را ممکن می‌سازد و همچنین شامل ابزارهای آماری است (مک‌کینی[2]، ۲۰۱۰). به عبارت ساده‌تر pandas یک کتابخانه نرم‌افزاری است که برای دستکاری و تجزیه و تحلیل داده‌ها در پایتون نوشته شده است. numpy یکی از مهم‌ترین کتابخانه‌های محاسباتی علمی برای جبر خطی و یادگیری ماشین است (ترفی[3]، ۲۰۲۱). Scikit-learn دارای الگوریتم‌های طبقه‌بندی مختلف از جمله رگرسیون است (وانگ[4] و همکاران، ۲۰۲۱).

۱) رگرسیون چندگانه

تحلیل رگرسیون چندگانه ابزار قدرتمندی برای بررسی رابطه بین متغیر وابسته و چندین متغیر مستقل است. هدف از رگرسیون چندگانه این است که مشخص شود چه مقدار واریانس در متغیر وابسته توسط متغیرهای مستقل قابل توضیح است. به عبارت دیگر، این تحلیل به تعیین اینکه کدام متغیرهای مستقل تأثیر قابل توجهی بر متغیر وابسته دارند کمک می‌کند.

به منظور انجام هر عمل آماری لازم است مفروضه‌هایی برقرار باشند تا بتوان به نتیجه حاصل شده از آنها اطمینان حاصل کرد (فیلد[5]، ۲۰۱۳). رگرسیون خطی چندگانه یک روش آماری است که برای نتایج دقیق نیازمند چندین پیش‌فرض است. این مفروضات شامل عدم وجود داده‌های پرت، استقلال مشاهدات، نرمال بودن خطا، همگنی واریانس خطا، عدم وجود هم‌خطی بین متغیرهای مستقل هستند. در این پژوهش به منظور شناسایی داده‌های پرت از فاصله ماهالانوبیس[6] استفاده می‌شود. بدین صورت که مقادیر به دست آمده برای فاصله ماهالانوبیس با توزیع خی‌دو با درجه آزادی سه مقایسه می‌شود (درجه آزادی برابر با تعداد

1- Python
2- Mckinney
3- Torfi
4- Wang
5- Field
6- Mahalanobis Distance

متغیرهای مستقل منهای یک می‌باشد). در صورتی که فاصله ماهالانوبیس موردی بیشتر از مقدار آماره خی‌دو باشد داده پرت محسوب می‌شود و باید از مجموعه داده‌ها حذف می‌شود (بلک و بابین[1]، ۲۰۱۹). همچنین برای بررسی نرمال بودن متغیر ملاک، آماره کولموگروف اسمیرنف به کار گرفته می‌شود. به علاوه برای بررسی استقلال مشاهدات از یکدیگر از آماره دوربین واتسون[2] استفاده می‌شود. مقدار این آماره بین ۰ تا ۴ تغییر می‌کند. اگر همبستگی بین باقیمانده‌های متوالی وجود نداشته باشد، مقدار آماره دوربین ـ واتسون باید نزدیک به ۲ شود و اگر مقدار این آماره به صفر نزدیک باشد، نشان‌دهنده وجود همبستگی مثبت بین باقیمانده‌های متوالی است. اگر مقدار این آماره به ۴ نزدیک باشد، نشان‌دهنده وجود همبستگی منفی بین مشاهدات متوالی است. به عنوان یک قاعده کلی، اگر مقدار مشاهده شده دوربین ـ واتسون بین ۱/۵ تا ۲/۵ باشد بیانگر استقلال مشاهدات است (مونت‌گومری[3] و همکاران، 2021). هم‌خطی چندگانه پدیده‌ای در تحلیل رگرسیون چندگانه است که در آن متغیرهای پیش‌بین وابستگی شدیدی را نشان می‌دهند که در رگرسیون چندگانه ایجاد مشکل می‌کند. به منظور بررسی وجود یا عدم وجود هم‌خطی چندگانه از عامل تورم واریانس[4] (VIF) استفاده می‌شود. به طورکلی حد قابل قبول مقدار VIF کمتر از پنج است و اگر آماره آزمون VIF به یک نزدیک باشد نشان‌دهنده عدم وجود هم‌خطی است (جیمز[5] و همکاران، ۲۰۱۳). بعد از برقراری این مفروضات لازم است اهمیت کلی مدل و نیز اهمیت هر یک از متغیرهای مستقل بررسی شود که به ترتیب با آماره‌های F و t آزمون می‌شوند (بلک و بابین، ۲۰۱۹).

در نهایت به منظور بررسی مناسب بودن مدل رگرسیون خطی چندگانه برای داده‌ها از شاخص ضریب تعیین (R^2) استفاده می‌شود که مقداری بین ۰ و ۱ را می‌پذیرد. هر چه مقدار R^2 به یک نزدیک‌تر باشد به معنی این است که مدل برازش شده، مدل مناسبی برای پیش‌بینی متغیر وابسته است و نسبت قابل‌توجهی از تغییرپذیری در متغیر وابسته را می‌توان با متغیرهای مستقل موجود توضیح داد.

۲) روش‌های استفاده از معیار هارویچ

در روش اول یا استراتژی خالص فرض بر این است که تصمیم‌گیرنده یک و تنها یک ویژگی را انتخاب کرده و به طور کامل اجرا می‌کند. ویژگی موجود در این روش، نمره کل شرکت‌کنندگان در یک دوره از آزمون تولیمو است. استراتژی خالص بعد از محاسبه بیشترین نمره کل (s_{max}) و کمترین مقدار نمره کل (s_{min}) و جایگذاری این مقادیر در رابطه (۵) به ازای α های مختلف به دو صورت اجرا می‌شود: ۱) نوع اول که همان روش کلاسیک هارویچ است مینیمم مقدار معیار هارویچ به عنوان نمره برش انتخاب می‌شود ۲) میانگین مقادیر معیار هارویچ به ازای α های مختلف به عنوان نمره برش در نظر گرفته می‌شود.

1- Black & Babin
2- Durbin Watson Statistic
3- Montgomery
4- Variance inflation factor
5- James

در روش دوم یا استراتژی ترکیبی که نگرش آگاهانه‌تری در فرآیند تصمیم‌گیری وجود دارد به تصمیم‌گیرنده این امکان داده می‌شود تا ترکیب وزنی از چندین ویژگی در دسترس را انتخاب و اجرا کند. ویژگی‌های موجود در این روش، نمره‌های بخش‌های گرامر، شنیداری، درک مطلب و نوشتاری آزمون تولیمو هستند و از رگرسیون خطی چندگانه برای مدل‌سازی این متغیرها با نمره کل استفاده شده است. استراتژی ترکیبی به چهار صورت اجرا می‌شود که در دو روش اول آن بعد از محاسبه مینیمم و ماکسیمم نمره‌های هر یک از بخش‌های گرامر، شنیداری، درک‌مطلب و نوشتاری و جایگذاری مقادیر مربوطه به ازای α های مختلف در رابطه (۵) بدین صورت عمل می‌کنیم: ۱) مینیمم مقدار معیار هارویچ در هر بخش به عنوان سطح تسلط آن بخش در نظر گرفته می‌شود که با جایگذاری این مقادیر در مدل برازش شده رگرسیون خطی، نمره برش مربوط به نمره کل حاصل می‌شود. ۲) میانگین مقادیر معیار هارویچ به ازای α های مختلف برای هر بخش آزمون در مدل رگرسیون خطی قرار داده می‌شود. برای اجرای روش‌های سوم و چهارم، بایستی مقادیر داده‌های هر یک از بخش‌های آزمون در مدل برازش شده رگرسیون خطی جایگذاری شود بنابراین به تعداد داده‌ها مقدار نمره کل پیش‌بینی شده $(\hat{y})$ به دست می‌آید. سپس مینیمم و ماکسیمم مقدار پیش‌بینی نمره کل محاسبه و با جایگذاری مقادیر مربوطه به ازای α های مختلف در رابطه (۵) بدین صورت عمل می‌کنیم: ۳) مینیمم مقدار معیار هارویچ به عنوان نمره برش تعیین می‌شود. ۴) میانگین نمره‌هایی به عنوان نمره برش انتخاب می‌شوند که از مقدار شاخص هارویچ (رابطه ۵) بزرگتر یا مساوی باشند یعنی در ستون مقادیر $\hat{y}$ ، نمره‌هایی در نظر گرفته می‌شوند که از مقدار معیار هارویچ بزرگتر یا مساوی باشند که میانگین این نمره‌ها برابر نمره برش می‌شود.

یافته‌ها

نمونه مورد بررسی شامل ۴۶۱ داوطلب شرکت‌کننده در دوره ۱۸۹ آزمون تولیمو هستند که در شهریور ۱۴۰۱ برگزار شده است. ویژگی‌ها شامل نمره‌های بخش‌های مختلف به انضمام نمره کل داوطلبان است. ویژگی‌های توصیفی داده‌ها شامل دامنه نمره‌ها، میانگین و انحراف استاندارد و چارک‌های هر یک از بخش‌های آزمون در جدول ۲ آورده شده است. طبق جدول ۲ مشاهده می‌شود که میانگین نمره کل داوطلبان شرکت‌کننده در این آزمون تولیمو برابر ۴۷۸ است و تنها ۲۵ درصد داوطلبان، نمره‌ای کمتر از ۴۳۲ ، ۵۰ درصد داوطلبان نمره‌ای کمتر از ۴۸۲ و ۷۵ درصد داوطلبان نمره‌ای کمتر از ۵۲۸ کسب کرده‌اند. بنابراین تنها نمره ۲۵ درصد داوطلبان بیشتر از ۵۲۸ شده است. قاعدتاً برای اینکه داوطلبان بیشتری امکان پذیرش را پیدا کنند لازم است نمره برش حداکثر برابر چارک دوم نمره‌ها یعنی ۴۸۲ باشد.

جدول ۲ : آمار توصیفی نمره‌های بخش‌های مختلف آزمون تولیمو دوره ۱۸۹

Table 2

Descriptive statistics of the scores of different sections of the 189th Tolimo exam

نمره کل Total score	نوشتاری Writing	درک مطلب Reading	شنیداری Listening	گرامر Grammar	
461	461	461	461	461	حجم جامعه population volume
317	100	4	2	3	مینیمم نمره‌ها Minimum scores
642	600	32	35	34	ماکسیمم نمره‌ها Maximum scores
478.0933	188.6117	18.50109	20.67245	18.58351	میانگین نمره‌ها Average scores
69.10718	115.3229	5.688518	8.923533	6.429219	انحراف استاندارد standard deviation
422	100	14	13	14	چارک اول first quarter
482	150	19	21	18	چارک دوم second quarter
528	250	23	28	24	چارک سوم third quarter

از آنجایی که برای تعیین نمره برش به شیوه‌های سنتی مانند آنگوف و بوک‌مارک لازم است سؤال‌های آزمون به چند داور برای تجزیه و تحلیل داده شود و از طرفی روش هارویچ به ازای مقادیر مختلف $\alpha \in (0,1)$، تصمیم‌گیرنده افراطی بدبین تا افراطی خوش‌بین را در برمی‌گیرد لذا داورها به سه صورت سخت‌گیر، متوسط‌گیر و آسان‌گیر در نظر گرفته می‌شوند. بدین‌صورت که یک داور سخت‌گیر، احتمال پایینی برای دادن پاسخ صحیح به یک سؤال برای آزمودنی مرزی در نظر می‌گیرد. یک داور آسان‌گیر، احتمال پاسخ صحیح به یک سؤال را برای آزمودنی مرزی بالا در نظر می‌گیرد و داور متوسط‌گیر، احتمال متوسطی را برای این منظور تعیین می‌کند. لذا بازه مربوطه به سه دسته تقسیم و میانگین هر دسته (نقطه وسط هر بازه) به عنوان ضریب هارویچ داورها در نظر گرفته می‌شود که نتایج حاصل در جدول ۳ آورده شده است.

جدول ۳ : دامنه مقادیر داورها از نظر نوع عملکرد

Table 3

The range of Referees' values according to the type of performance

ضریب α Coefficient α	دامنه مقادیر عملکرد Range of performance values	نوع داور Referee type
0.17	(0, 0.33)	داور سخت‌گیر Strict Referee
0.5	(0.34 , 0.67)	داور متوسط‌گیر Average Referee
0.84	(0.68 , 1)	داور آسان‌گیر Permissive Referee

در بخش زیر، پیاده‌سازی و مقایسه نتایج تجربی برای دو نوع استراتژی خالص و ترکیبی بر اساس روش پژوهش ارائه می‌شود و همچنین روش‌های ارزیابی نمره‌های برش به دست آمده نیز بیان می‌شود.

الف) استراتژی خالص

مقدار شاخص هارویچ در استراتژی خالص بر اساس مینیمم و ماکسیمم مقدار نمره کل به ازای مقادیر α در جدول ۴ آورده شده است.

جدول ۴ : مقدار شاخص هارویچ بر اساس نمره کل در استراتژی خالص

Table 4

Hurwicz index value based on total score in pure strategy

	$\alpha = 0.17$	$\alpha = 0.5$	$\alpha = 0.84$	میانگین Average
نمره برش Cut score	372.25	479.5	590	480.58

نتایج نشان می‌دهد که نمره برش بهینه بر اساس رویکرد خوش‌بینانه برابر ۳۷۲/۲۵ که به ازای ۰/۱۷ $= \alpha$ تعیین می‌شود. بنابراین در صورتی که فقط از یک تصمیم‌گیرنده برای تعیین نمره برش استفاده شود، تصمیم‌گیرنده بدبین با توجه به فرض توانایی پایین شرکت‌کنندگان در پاسخگویی صحیح به سؤالات، نمره برش ۳۷۲ را برای پذیرش مناسب می‌داند.

علاوه بر این، با بررسی تأثیر استفاده از هر سه نوع داور (سخت‌گیر، متوسط‌گیر و آسان‌گیر) بر نمره برش، مقدار میانگین شاخص هارویچ که تقریباً برابر ۴۸۱ است به عنوان نمره برش استفاده می‌شود. نتیجه حاصل بیانگر این مطلب است که استفاده از چندین داور برای تعیین نمره برش بر اساس تنها ویژگی نمره کل مطابق با رویکردهای سنتی، ممکن است منجر به نتایج قابل قبول‌تری شود.

ب) استراتژی ترکیبی

به منظور ارائه نتایج حاصل از استراتژی ترکیبی، مفروضه‌های رگرسیون چندگانه بررسی و با استفاده از روش حداقل مربعات، ضرایب رگرسیون چندگانه به صورت گرامر (۴/۱۴۲)، شنیداری (۳/۵۲۳)، درک‌مطلب (۴/۱۴۱) ، نوشتاری (۰/۰۰۳) با مقدار ثابت ۲۵۱/۰۴۸ برآورد شدند که به غیر از نمره ویژگی نوشتاری سایر ویژگی‌ها از نظر آماری در سطح ۰/۰۵ معنادار شدند ($p - value < 0/05$). بدین‌معنی که ضرایب گرامر، شنیداری، درک‌مطلب پیش‌بینی‌کننده معناداری برای عملکرد آزمون هستند به طوری که دانشجویانی که در این زمینه‌ها دارای مهارت بالایی هستند می‌توانند نمره آزمون بالایی را کسب کنند. از طرف دیگر مقدار ضریب تعیین در رگرسیون چندگانه برابر ۰/۹۸۸ و مقدار MSE برابر ۰/۰۰۰۵۶ به دست آمده است که دلیلی بر دقت مدل رگرسیونی در برازش پارامترها هستند. با توجه به مقدار R^2 این اطمینان حاصل می‌شود که پیش‌بینی‌کننده‌های انتخاب شده یعنی گرامر، شنیداری و درک‌مطلب سهم قابل توجهی در عملکرد آزمون دارند و بنابراین استفاده از این متغیرها در ایجاد نمره برش مجاز می‌باشد و منجر به نتایج قابل‌قبولی می‌شوند.

نتایج مربوط به مقادیر شاخص هارویچ برای هر یک از بخش‌های آزمون بر اساس استراتژی ترکیبی به ازای مقادیر α در جدول ۵ آورده شده است. طبق نتایج مشاهده شده در جدول ۵ بدیهی است که در روش اول استراتژی ترکیبی، مقدار αی بهینه براساس رویکرد خوش‌بینانه برابر ۰/۱۷ است که بر این اساس، نمره برش بهینه با جایگزینی این مقادیر در مدل رگرسیون خطی برابر ۳۴۸/۳۸۷۵۳ به دست می‌آید و برای روش دوم کافی است میانگین مقادیر αهای مختلف در هر بخش آزمون در مدل رگرسیون خطی جایگذاری شود که حاصل نمره برش بهینه تقریباً برابر ۴۶۹ به دست می‌آید.

جدول ۵ : مقادیر شاخص هارویچ برای هر یک از بخش‌های آزمون در استراتژی ترکیبی

Table 5

Hurwicz index values for each part of the test in the combined strategy

	$\alpha = 0.17$	$\alpha = 0.5$	$\alpha = 0.84$	میانگین Average
گرامر Grammar	8.27	18.5	29.04	18.6
شنیداری Listening	7.61	18.5	29.72	18.61
درک‌مطلب Reading	8.76	18	27.52	18.09

در روش سوم، مقدار برآورد متغیر وابسته $(\hat{y})$ بر اساس مدل رگرسیون خطی به ازای همه داده‌ها به دست آورده می‌شود و از مینیمم و ماکسیمم مقدار پیش‌بینی نمره کل در رویکرد خوش‌بینانه شاخص هارویچ استفاده می‌شود و مینیمم مقدار معیار هارویچ به ازای αهای مختلف به عنوان نمره برش انتخاب می‌شود و در روش چهارم میانگین نمره‌هایی به عنوان نمره برش انتخاب می‌شوند که بزرگتر یا مساوی مقدار شاخص هارویچ باشند. مقدار مینیمم و ماکسیمم برآورد نمره کل به ترتیب برابر ۳۳۲/۴۳۶ و ۶۳۵/۵۶۸ است. مقدار شاخص هارویچ و نمره برش معادل به ازای هر سه مقدار α در جدول ۶ آورده شده است.

جدول۶ : مقادیر شاخص هارویچ و نمرهٔ برش در استراتژی ترکیبی

Table 6

Hurwicz index values and cutoff score in combined strategy

	$\alpha = 0.17$	$\alpha = 0.5$	$\alpha = 0.84$
شاخص هارویچ Hurwicz index	383.968	484.002	587.67
نمره برش Cut score	494	536.40	609.1875

طبق نتایج نوشته شده در جدول ۶، نمره برش براساس روش سوم تقریباً برابر ۳۸۴ به دست می‌آید و بر اساس روش چهارم، مقدار نمره برش برابر ۴۹۴ حاصل می‌شود که در این دو روش، مقدار ۰/۱۷ = α است که نشانگر بدبینی داور است.

ج) روش‌های ارزیابی

به منظور ارزیابی نمره‌های برش به دست آمده از استراتژی‌های خالص و ترکیبی، خطای استاندارد و خطای طبقه‌بندی داده‌های آزمون محاسبه شدند که نتایج آنها در جدول ۷ آورده شده است.

جدول ۷ : ارزیابی استراتژی‌های استفاده شده

Table 7

Evaluation of used strategies

	خطای استاندارد standard error	خطای طبقه‌بندی Classification error
دومین استراتژی خالص Second pure strategy	1.1694	0.03597
چهارمین استراتژی ترکیبی fourth Combined strategy	1.1694	0.0

همچنین برای بررسی اعتبار مدل‌هایی که نمره برش آنها بین ۴۸۰ تا ۵۰۰ شده از اعتبارسنجی متقابل استفاده شد. بدین صورت که داده‌ها به طور تصادفی به دو بخش آموزشی و آزمون تقسیم شدند به طوری که فرآیند آموزش و آزمون، پنج بار اجرا شد و در هر اجرا یک بخش به عنوان آزمون و سایر بخش‌ها برای آموزش به کار گرفته شدند و در نهایت میانگین αها و میانگین نمره برش حاصل از این اجراها به عنوان مقادیر بهینه در نظر گرفته شدند که در جدول ۸ آورده شده‌اند.

جدول ۸ : نتایج حاصل از اعتبارسنجی متقابل

Table 8

Results of cross-validation

	$\bar{\alpha}$	نمره برش Cut score
دومین استراتژی خالص Second pure strategy	0.17	481
چهارمین استراتژی ترکیبی fourth ombined strategy	0.17	494

$\bar{\alpha}$ = میانگین α های حاصل از پنج اجرا

بحث و نتیجه‌گیری

تعیین نمره‌های برش در آزمون یک وظیفه ضروری برای مربیان و کارفرمایان است که برای ارزیابی دانش و مهارت افراد استفاده می‌کنند. نمره برش حداقل نمره مورد نیاز برای یک فرد به عنوان دستیابی به سطح خاصی از عملکرد یا مهارت در یک آزمون مشخص است. این نمره در تصمیم‌گیری در مورد محل قرارگیری، تشخیص و واجد شرایط بودن افراد برای برنامه‌های مختلف ضروری است. نمره برش همچنین به شناسایی افرادی که ممکن است برای بهبود عملکرد خود نیاز به حمایت یا مداخله بیشتری داشته باشند کمک می‌کند. علاوه بر این، آنها معیاری برای مقایسه عملکرد افراد یا گروه‌های مختلف مردم ارائه می‌دهند. به عنوان مثال، عملکرد دانش‌آموز در یک آزمون ممکن است بر اساس نمرات برش به یکی از چندین دسته از قبیل پایه، ماهر یا پیشرفته طبقه‌بندی شود. تعیین نمره‌های برش در آزمون‌های پرکاربرد در زمینه‌های آموزشی مستلزم مشارکت سیاست‌گذاران، مربیان، متخصصان سنجش و دیگران در یک فرآیند قضاوتی چندمرحله‌ای است (زیکی و پری، بی.تا.). بنابراین نمره برش نشان‌دهنده مقدار آستانه‌ای است که پذیرش و عدم پذیرش در آزمون را مشخص می‌کند. بدین صورت که هر مقدار بیشتر از نمره برش، به معنای قبولی در آزمون و هر مقدار کمتر از نمره برش به معنای عدم قبولی در آزمون در نظر گرفته می‌شود. لذا تعیین دقیق نمره برش این اطمینان را به وجود می‌آورد که تصمیمات اتخاذ شده بر اساس نتایج آزمون منصفانه و عینی هستند.

مقدار R^2 در رگرسیون چندگانه به عنوان یک آمار خلاصه ارزشمند عمل می‌کند که می‌تواند در تصمیم گیری در حوزه آزمون آموزشی کمک کند. در این مطالعه خاص مقدار ۰/۹۸۸ برای R^2 حاصل شد که نشان‌دهنده مناسب بودن مدل رگرسیون خطی چندگانه برای مجموعه داده است. در نتیجه مربیان و کارشناسان ارزیابی می‌توانند قضاوت آگاهانه در مورد گنجاندن متغیرها در ارزیابی‌های خود و همچنین روش‌های مناسب اندازه‌گیری انجام دهند که به نوبه خود، ایجاد نمرات برش و سایر معیارهای ارزیابی را امکان‌پذیر می‌کند که با ارزش بالای این آماره در مطالعه فعلی این پیش نیازها برآورده می‌شود. در نتیجه با توجه به معنادار بودن مدل ($p - value < 0/05$) و بالا بودن مقدار ضرایب گرامر (۴/۱۴۲)، شنیداری (۳/۵۲۳)، درک‌مطلب (۴/۱۴۱) دریافتیم که دانش و درک قوی از اصول گرامری، مهارت‌های شنیداری قوی، مهارت‌های خواندن قوی می‌توانند تأثیر مثبتی بر نمره آزمون داشته باشد.

از آنجایی که روش‌های تعیین استاندارد باید به گونه‌ای طراحی شوند که اثرات مقایسه اجتماعی بر تصمیمات اتخاذ شده را به حداقل برسانند و تأثیرات برخی از اطلاعات را به حداکثر برسانند (برک، ۱۹۸۶). لذا ایده اصلی نویسندگان از استفاده از معیار هارویچ به جای روش‌های سنتی آنگوف و بوک‌مارک، افزایش اعتبار آزمون و عدم سوء استفاده از سؤال‌های آزمون بود. در این مقاله، چگونگی استفاده از روش هارویچ برای تعیین مناسب‌ترین نمره برش آزمون تولیمو بررسی شد و به نتایج قابل‌قبولی دست یافتیم.

شاخص هارویچ یک رویکرد ساده و انعطاف‌پذیر برای تصمیم‌گیری در سناریوهایی که شامل نتایج متعدد و سطوح مختلف خوش بینی یا بدبینی هستند، ارائه می‌دهد که به عنوان ابزاری برای تعیین نمره برش برای اقدامات یا انتخاب‌های خاص عمل می‌کند. روش هارویچ با استفاده از وزن‌های دلخواه برای بهترین و بدترین سناریو، از یک رویکرد میانگین وزنی استفاده می‌کند که بهترین حالت و بدترین سناریو را برای رسیدن به یک تصمیم ترکیب می‌کند که می‌تواند منجر به ذهنیت و سوگیری در فرآیند تصمیم‌گیری شود.

پس از تعیین نمره برش، اعتبارسنجی مدل بررسی و خطای استاندارد و خطای طبقه‌بندی به دست آورده می‌شوند. در این مرحله می‌توان اثربخشی نمره برش در تحقق هدف ارزیابی را بررسی نمود که شامل تجزیه و تحلیل عملکرد افرادی باشد که در آزمون شرکت کرده‌اند چه آنهایی که نمره حد نصاب تعیین شده را کسب کرده و چه آنهایی که کسب نکرده‌اند. همچنین لازم است تأثیر بالقوه نمره برش بر جمعیت مورد ارزیابی در نظر گرفته شود. بنابراین ارزیابی تصمیم برای اطمینان از منصفانه، معتبر و قابل اعتماد بودن آن یکی از گام‌های اصلی در تصمیم‌گیری است.

ضریب هارویچ بر اساس مدل خوش‌بینی هارویچ، طیفی از مقادیر را برای داورهای سخت‌گیر، متوسط‌گیر و آسان‌گیر ارائه می‌دهد. در این مطالعه به منظور دستیابی به نتایج بهینه از دو استراتژی خالص و ترکیبی استفاده شد. برای تعیین نمره برش در استراتژی خالص تنها از نمره کل استفاده شد و در استراتژی ترکیبی از نمره‌های مربوط به بخش‌های مختلف آزمون استفاده گردید.

در روش اول یعنی استراتژی خالص هارویچ، نمره برش بهینه برای داوطلبان آزمون تولیمو برابر ۳۷۲ و در روش دوم بر اساس میانگین هر سه نوع ضریب α تقریباً برابر ۴۸۱ به دست آمد بدین‌معنی که در صورت استفاده از این رویکرد، تنها داوطلبانی صلاحیت ورود به دور بالاتر را خوهند داشت که حداقل نمره ۴۸۱ را کسب کرده باشند. همچنین در استراتژی ترکیبی، نمره برش بهینه برای هر چهار روش به ترتیب برابر ۳۴۹، ۴۶۹، ۳۸۴ و ۴۹۴ به دست آمد.

به علاوه از خطای استاندارد و خطای طبقه‌بندی داده‌های آزمون برای ارزیابی نمره‌های برش به دست آمده از استراتژی‌های خالص و ترکیبی، استفاده شد (به جدول شماره ۷ نگاه کنید).

با توجه به نتایج اعتبارسنجی متقابل (جدول شماره ۸) و مقادیر خطاهای به دست آمده (جدول شماره ۷) نتیجه گرفته شد که نمره برش حاصل از استراتژی ترکیبی منجر به نتایج دقیق‌تری می‌شود به طوری که خطای طبقه‌بندی این استراتژی برابر صفر شده است. ولی از سوی دیگر با توجه به اعتبار استراتژی خالص و نیز خطاهای ناچیز آن از روش مربوط به استراتژی خالص هم می‌توان برای تعیین حد تسلط در هر یک از بخش‌های آزمون استفاده کرد.

از آنجایی که نتایج ارزیابی و اعتبار مدل‌ها رضایت‌بخش بودند و نیز هدف ما تعیین نمره برشی بین ۴۸۰ و ۵۰۰ بود لذا اینگونه می‌توان گفت که در صورت استفاده از نمره کل، میانگین مقادیر شاخص هارویچ به ازای ضرایب خوش‌بینی مختلف منجر به نمره برش ۴۸۱ می‌شود و در صورت معنی‌دار بودن مدل رگرسیون خطی و بالا بودن مقدار R^2 با به کارگیری نتایج شاخص هارویچ در رگرسیون چندگانه و با در نظر گرفتن ضریب خوش‌بینی بالا ($\alpha = 0/17$) از نمره برش ۴۹۴ به عنوان حد نصاب قبولی می‌توان استفاده کرد.

تاکنون از روش‌های تصمیم‌گیری چندمعیاره از جمله هارویچ برای تعیین نمره برش در حوزه آموزش استفاده نشده است و نتایج به دست آمده حاکی از آن است که روش هارویچ ابزار مفیدی برای تعیین نمره برش برای آزمونی مانند آزمون تولیمو را فراهم می‌کند. بر اساس یافته‌های این پژوهش می‌توان نتیجه گرفت که روش هارویچ و رویکرد رگرسیون چندگانه می‌توانند به‌عنوان روشی جدید و قابل اعتماد برای تعیین نمرات برش در آزمون‌های آموزشی مورد استفاده قرار گیرند به طوری که به عنوان جایگزینی برای روش‌های سنتی تعیین نمره برش از جمله آنگوف و بوک‌مارک معرفی شوند. چرا که با استفاده از معیار هارویچ لزومی به تجزیه و تحلیل سؤال‌ها توسط چندین داور نیست و از لحاظ هزینه و زمان بسیار به صرفه‌تر است. تنها ذهنیتی که در روش هارویچ وجود دارد تعیین ضریب خوش‌بینی است. ضریب خوش‌بینی (α) یک پارامتر ذهنی است که بسته به ترجیحات تصمیم‌گیرنده می‌تواند متفاوت باشد. بدین‌صورت که معیار هارویچ به تصمیم‌گیرندگان این امکان را می‌دهد تا سطح خوش‌بینی یا بدبینی را بسته به ترجیحات ریسک خود در بازه بین ۰ و ۱ تنظیم کنند.

در واقع با استفاده از این روش می‌توان خطرات و پاداش‌های ناشی از نمره‌های برش مختلف را سنجید و در نهایت به تصمیمی رسید که منصفانه و سازگار باشد و بدین‌ترتیب خطرات احتمالی فاش شدن سؤال‌های آزمون از بین می‌رود و اعتبار آزمون افزایش می‌یابد. بنابراین استفاده از این رویکردها برای بهبود دقت نمرات برش در سایر ارزشیابی‌های آموزشی نیز می‌تواند مفید واقع شود و سازمان‌های ارزیابی آموزش می‌توانند از روش هارویچ و رویکرد رگرسیون چندگانه برای ایجاد نمره‌های برش منصفانه و دقیق برای آزمون‌ها و یا روش هارویچ اصلاح شده استفاده کنند. تحقیقات بیشتر می‌تواند کاربرد این رویکرد را در سایر انواع آزمون‌های آموزشی بررسی کند تا تعمیم‌پذیری و کاربرد آن در زمینه‌های مختلف مشخص شود.

در این مطالعه در مورد اهمیت داشتن یک نمره برش کاملاً تعریف شده به منظور اطمینان از انصاف و ثبات در آزمون بحث شد. همچنین روش‌های تصمیم‌گیری تحت عدم قطعیت و به طور ویژه روش هارویچ که یک تکنیک تصمیم‌گیری برای ایجاد تعادل ریسک و پاداش است معرفی شد. در این پژوهش

روش هارویچ برای تعیین نمره برش آزمون تولیمو به کار گرفته شد و نتیجه حاصل با سناریوی فعلی که در آن حدنصاب قبولی ۴۸۰ یا ۵۰۰ است مقایسه شد.

تقدیر و تشکر

بدینوسیله از سازمان سنجش آموزش کشور به خاطر همکاری در اجرای پژوهش حاضر که برگرفته از رساله دکتری با عنوان " مقایسه نمرات برش آزمون‌های ملاک مرجع در الگوریتم‌های یادگیری عمیق و روش‌های منتخب مورد مطالعه : آزمون تولیمو" و همچنین از نشریه علمی – پژوهشی آموزش و ارزشیابی دانشگاه آزاد اسلامی واحد تبریز برای قبول داوری مقاله سپاسگزاری می‌شود.

منابع **References**

Association, A. E. R. (2018). Standards for educational and psychological testing: *American Educational Research Association*.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational research,56(1): 137-172*

Black, W., & Babin, B. J. (2019). Multivariate data analysis: Its approach, evolution, and impact. *In The Great Facilitator: Reflections on the Contributions of Joseph F. Hair, Jr. to Marketing and Business Research*, pp.121-130: Springer.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1):59-88.

Dimitrov, D. M. (2022). The Response Vector for Mastery Method of Standard Setting. *Educational and Psychological Measurement*,82(4): 719-746.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*: sage.

Gaspars-Wieloch, H. (2014). Modifications of the Hurwicz's decision rule. *Central European Journal of Operations Research,22,* 779-794.

Grabovsky, I., & Wainer, H.(2017). The cut-score operating function: A new tool to aid in standard setting. *Journal of Educational and Behavioral Statistics*,42(3): 251-263.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hurwicz, L. (1952). *A criterion for decision making under uncertainty*. Retrieved from

ION, P. V. (2012). Methods and techniques underlying the decision - effective management tools. Paper presented at the *proceedings of the international scientific conference eco-trend*, târgu jiu ,romania .

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112): Springer.

Kim, B., Kim, J., & Yi, G. (2017). Analysis of clustering evaluation considering features of item response data using data mining technique for setting cut-off scores. *Symmetry*, 9(5): 62.

Lane, A. S., Roberts, C., & Khanna, P. (2020). Do We Know Who the Person With the Borderline Score is, in Standard-Setting and Decision-Making. *Health Professions Education*,6(4): 617-625.

Lin, J. (2006). The bookmark procedure for setting cut-scores and finalizing performance standards: Strengths and weaknesses. *Alberta journal of educational research*, 52(1).

McKinney, W. (2010). Data structures for statistical computing in python. *Paper presented at the Proceedings of the ۹th Python in Science Conference*.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis: *John Wiley & Sons*.

Pažek, K., & Rozman, Č. (2009). Decision making under conditions of uncertainty in agriculture: a case study of oil crops. *Poljoprivreda*, 15(1): 45-50.

Pérez, D. E., Hernández, J. G., García, M. J., & Hernández, G. J. (2015). Hurwicz method modified and the amplitude model (TAM). *Delener et al.(ed) GBATA ۲۰۱۵reading book. GBATA, Peniche*, 559-566.

Render, B., & Stair Jr, R. M. (2016). Quantitative Analysis for Management, 12e: *Pearson Education India*.

Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta journal of educational research*, 52(1).

Ross, S. M. (2013). Simulation: *Academic Press*.

Taherdoost, H., & Madanchian, M.(2023).Multi-Criteria Decision Making (MCDM) Methods and Concepts. *Encyclopedia*, 3(1): 77-87.

Torfi, A. (202). *Practical Linear Algebra for Machine Learning*: Instill AI.

Wang, J., Liu, J., DiStefano, C., Pan, G., Gao, R., & Tang, J. (2021). Utilizing deep learning and oversampling methods to identify children's emotional and behavioral risk. *Journal of Psychoeducational Assessment*, 39(2): 227-241.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, 40(3): 231-253.

Zieky, M., & Perie, M. A Primer on Setting Cut Scores on Tests of Educational Achievement.Retrieved from report

Zahed babelan, A., & karimianpour, G. (2020). The Relationship between Academic Optimism and Buoyance, the Mediator Role of Academic Self-efficacy, Educational and Scholastic studies, 9(1): 149-170. [In Persian]