مجله شیمی کوانتومی و اسپکتروسکوپی
۱۳:٤۱ (۱٤۰۲) ۷۳-۱۰۹
Journal of Quantum Chemistry and Spectroscopy (JQCS)

مراحل، محاسبات و نتایج حاصل از مطالعات پیش‌بینی‌های نظری ارتباط کمی ساختار بازداری (QSRR) اسانس گیاه میخک زینتی

مجید محمدحسینی[1]*، مهدی نکوئی[1]
[1]گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران

تاریخ ثبت اولیه:۱۴۰۲/۰۱/۲۵، تاریخ دریافت نسخه اصلاح شده:۱۴۰۲/۰۴/۲۲ ، تاریخ پذیرش قطعی:۱۴۰۲/۰۴/۳۰

چکیده

در این مقاله، به تشریح مبسوط مدل‌های خطی توانمند در پیش‌بینی شاخص بازداری کواتس گروه وسیعی از ترکیبات طبیعی شناسایی شده در روغن اسانسی گیاه میخک زینتی به عنوان یکی از گیاهان دارویی پرداخته شده است. در این راستا، اساس کار مبتنی بر روابط کمی ساختار بازداری (QSRR) می‌باشد که در منابع علمی از اهمیت بسزایی جهت برقراری ارتباط منطقی و هدفمند بین شاخص کواتس به عنوان یک متغیر وابسته و گروهی از توصیف‌کننده‌های مولکولی به عنوان متغیرهای مستقل برخوردار است. در این راستا، پس از ترسیم ساختار ترکیبات مفروض در محیط نرم‌افزار هایپرکم و بهینه‌سازی آن‌ها، جهت استخراج توصیف‌کننده‌های مولکولی مربوطه از نرم‌افزار دراگون استفاده شد. در مرحله بعد، پس از حذف توصیف‌کننده‌های غیر مرتبط و اضافی، نهایتاً با روش‌های مرحله‌ای و روش انتخاب متغیر مبتنی بر الگوریتم ژنتیک گروهی از توصیف‌کننده‌های مهم و مؤثر شناسایی و ارتباط خطی آن‌ها با شاخص بازداری کواتس مورد بحث و بررسی قرار گرفت. نتایج حاصله حاکی از توانمندی بالای مدل‌های ارائه شده جهت پیش‌بینی شاخص کواتس گروه وسیعی از ترکیبات طبیعی دارد.

***واژه‌های کلیدی:** ارتباط کمی ساختار- بازداری، رگرسیون خطی چندگانه، توصیف‌کننده‌های مولکولی، الگوریتم ژنتیک، شاخص بازداری کواتس، میخک زینتی.*

۱. مقدمه

هم‌زمان با پیدایش انسان‌ها، استفاده از گیاهان داروئی نیز آغاز شد. برخی از ترکیبات شیمیایی موجود در گیاه دارای ساختار پیچیده‌ای هستند و سنتز آن‌ها در آزمایشگاه یا غیرممکن یا با صرف زمان و هزینه زیاد امکان‌پذیر است. در قرن ۱۸ و اوایل قرن ۱۹، محققان پیشرفت قابل‌توجهی در خالص‌سازی و شناسایی ترکیبات شیمیایی موجود در گیاهان داشته و موادی را به صورت فرآورده-

***عهده‌دار مکاتبات:** مجید محمدحسینی
نشانی: گروه شیمی، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران
تلفن:۰۲۳۳۲۳۹۴۲۷۸ **پست الکترونیک:**E-mail: majidmohammadhosseini@yahoo.com

های داروئی برای مصرف عرضه کردند. هم‌زمان با انقلاب صنعتی، علم شیمی پیشرفت چشمگیری داشت که باعث به‌وجودآمدن این تفکر در محیط‌های علمی شد که می‌توان از طریق سنتز ترکیبات شیمیایی به‌خصوص مواد داروئی مشکل دارو و درمان بیماری-ها را حل کرد [۱]. به همین دلیل تولید داروهای شیمیایی در قرن بیستم سرعت روزافزونی پیدا کرد و داروهای گیاهی به‌دست فراموشی سپرده شدند. پس از مواجه شدن با مشکلاتی نظیر آلودگی آب و هوا و خاک که توسط کارخانجات تولید مواد شیمیایی ایجاد شده بوده و عوارض جانبی داروهای شیمیایی که بعضاً پس از چند نسل ظاهر می‌شوند، به فکر استفاده از مواد طبیعی فن-آوری‌های غیر مخرب افتادند. به‌طوری‌که در کشورهای صنعتی مصرف داروهای گیاهی از مرز ۷ درصد گذشته است.

با توجه به کاربرد داروهای گیاهی، پژوهشکده‌ی صنایع شیمیایی سازمان پژوهش‌های علمی و صنعتی ایران نیز تحقیق‌هایی را در این زمینه به اجرا در آورده است. تاریخچه‌ی گیاهان داروئی به طور دقیق مشخص نیست و در طول تاریخ استفاده از گیاهان داروئی با خرافات و آداب خاصی همراه بوده است. مصری‌ها و چینی‌ها از اولین اقوامی هستند که از حدود ۲۷۰۰ سال پیش‌از میلاد مسیح از گیاهان به عنوان دارو استفاده می‌کردند. تئوفراست یکی از شاگردان ارسطو بنیانگذار مکتب درمان با گیاه است. دیوسکورید در قرن اول میلادی مجموعه‌ای را مشتمل بر خواص داروئی ۶۰۰ گیاه جمع‌آوری نمود که این اثر منشأ بسیاری از مطالعات در قرون بعد گردید [۱]. در قرون هشتم تا دهم میلادی، بوعلی‌سینا و محمد زکریای رازی سبب توسعه‌ی دانش درمان با گیاه شدند و در قرن سیزدهم، ابن بیطار خصوصیات بیش از ۱۴۰۰ گیاه را در کتابی گردآوری نموده و در قرن نوزدهم داروهای شیمیایی به سرعت جایگزین بسیاری داروهای گیاهی گردید. سپس، در اواخر قرن بیستم عوارض جانبی و مضر داروهای شیمیایی سبب رویکرد دوباره دانشمندان به گیاهان داروئی شد به‌طوری‌که این دوره را رنسانس گیاهان داروئی نامیدند. تا قرن نوزدهم، گیاهان داروئی به شکل بسیار ابتدایی مورد مصرف قرار می‌گرفتند تا این‌که استخراج مواد مؤثر گیاهی، از قرن نوزدهم آغاز گردید [۱].

نظر به کاربرد گیاهان داروئی در زمینه‌های مختلف، گیاهان اساساً در سه گروه اصلی گیاهان داروئی، گیاهان ادویه‌ای، گیاهان عطری طبقه‌بندی می‌شوند. اما به‌طورکلی، گیاه داروئی به گیاهی اطلاق می‌شود که در یک یا چند اندام خود حاوی مواد مؤثر بوده و کاشت، داشت و برداشت آن‌ها صرفاً به منظور استفاده از این مواد انجام گردد. بنابراین، با آن‌که اندام‌های برخی گیاهان نظیر برگ‌های گردو و کاکل ذرت و پوست میوه لوبیا حاوی مواد مؤثری هستند که کاربردهای داروئی نیز دارند ولی از این نظر که کاشت، داشت و برداشت این گیاهان تنها به‌منظور استفاده از این گیاهان تنها به‌منظور استفاده از مواد مؤثر موجود در آن‌ها انجام نمی‌گیرد، گیاه دارویی محسوب نمی‌شوند. نظر به تنوع و اهمیت حیاتی گیاهان داروئی، یکی از جامع‌ترین دائره‌المعارف‌های گیاهی توسط دکتر ولی‌الله مظفریان به زیور طبع آراسته شده است [۲].

یکی از مهم‌ترین مواد مؤثر در گیاهان داروئی را روغن‌های فرار یا روغن‌های اسانسی تشکیل می‌دهند. این مواد در قسمت‌های مختلف بسیاری از گیاهان داروئی وجود دارند. بسیاری از گیاهان داروئی به‌علت داشتن روغن‌های فرار به‌طور مستقیم در پزشکی مصرف می‌شوند، ولی در بیشتر موارد روغن‌های فرار را از مواد خام جدا نموده و به‌عنوان دارو به‌کار می‌برند. معمولاً، از اسانس‌ها

در تهیه‌ی عطرها و اسپری‌های خوشبوکننده و هم‌چنین به عنوان معطرکننده‌ی صابون و خمیردندان و شامپوهای طبی مورد استفاده قرار می‌گیرند. از اسانس‌ها در صنایع داروسازی نیز استفاده‌های زیادی به‌عمل می‌آید، زیرا بسیاری از این مواد دارای اثر ضد باکتری و ضدعفونی‌کننده هستند. از اثرهای داروئی اسانس‌ها، می‌توان به خواصی مانند ضد تورم، ضد دل‌درد، آرام بخش، ضد نفخ، اشتها آور و خلط‌آور اشاره نمود. در صنایع غذایی و کنسروسازی، روغن‌های اسانسی به‌دست آمده از گیاهان ادویه‌ای برای بهبود طعم مواد غذایی استفاده می‌شوند. هم‌چنین، روغن‌های اسانسی می‌توانند ضمن بهبود طعم برخی از داروها به‌عنوان محافظ هم مورد استفاده قرار گیرند.

هدف از تحقیق اخیر، مدل‌سازی‌های توانمند مبتنی بر روابط کمی ساختار-بازداری (QSRR) ۶۰ ترکیب طبیعی شناسایی شده در روغن اسانسی گیاه *Pittosporum undulatum* می‌باشد. این مدل‌سازی، با استفاده از روش‌های رگرسیون خطی چندگانه مرحله‌ای (SW-MLR) و الگوریتم ژنتیک-رگرسیون خطی چندگانه (GA-MLR) انجام شده است.

۲. روش‌های محاسباتی

۲-۱. سری داده‌ها

در این مطالعه، اندیس کواتس ۶۰ ترکیب از اسانس گیاه داروئی میخک زینتی (شکل ۱)، با استفاده از روش رگرسیون خطی چندگانه مورد بررسی قرار گرفته است. سری داده‌ی مربوطه، شامل مجموعه‌ای از ترکیبات مونوترپن‌ها، سزکوئی‌ترپن‌ها و سایر ترکیبات طبیعی، از تحقیق یک گروه از محققین پرتقالی به سال ۲۰۰۷ میلادی گرفته شده است [۳]. مراحل انجام این پژوهش، شامل انتخاب سری داده‌ها، تعیین‌توصیف‌کننده‌ها، رگرسیون خطی چندگانه و مدل‌سازی می‌باشد. نام، ساختار و مقادیر عددی اندیس‌های بازداری (RI) آن‌ها در جدول ۱، نشان داده شده است. برای مدل‌سازی، ابتدا اندیس بازداری ترکیبات، به‌عنوان پارامتر مورد بحث، انتخاب شد. در مرحله‌ی دوم، ساختار تمام گونه‌های مورد نظر رسم و بهینه گردید.

شکل ۱. تصویر گیاه میخک زینتی

۲-۲. محاسبه و انتخاب توصیف‌کننده‌ها

توصیف‌کننده‌ها، کدهای عددی هستند که بیان‌گر ویژگی‌های خاصی از هر مولکول بوده که با خواص فیزیکو-شیمیایی ماده - ارتباط دارند. با توجه به موضوع مورد بحث (بررسی ارتباط کمی ساختار- بازداری)، ابتدا ساختار مولکول‌ها به وسیله‌ی نرم‌افزار Hyperchem رسم و ساختار هندسی آن‌ها با احتساب اتم‌های هیدروژن و با استفاده از مدل هامیلتونی AM1 و الگوریتم Polak-Ribiere که از روش‌های نیمه‌تجربی می‌باشند، بهینه شد. سپس، ساختارهای رسم شده، توسط نرم‌افزار مذکور به‌عنوان ورودی به نرم‌افزار دیگری به نام Dragon، معرفی شد. محاسبه‌ی توصیف‌کننده‌ها، توسط این نرم‌افزار صورت گرفت. بالغ بر ۱۴۰۰ توصیف-کننده از هجده گروه مختلف برای هر ترکیب با این بسته‌ی نرم‌افزاری، قابل محاسبه است.

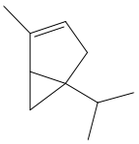
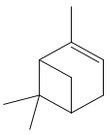
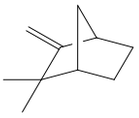
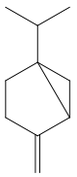
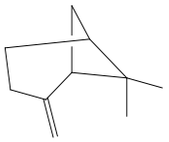
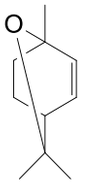
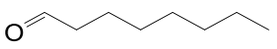
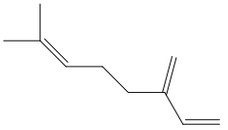
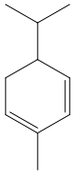
۲-۳. تجزیه و تحلیل آماری توصیف‌کننده‌ها

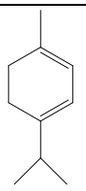
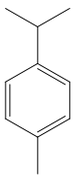
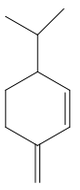
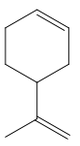
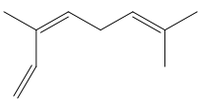
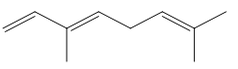
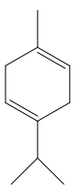
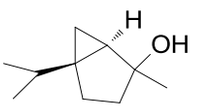
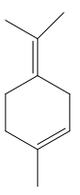
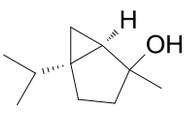
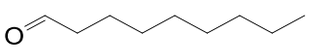
برای ایجاد مدلی که بیان‌گر ارتباط ساختاری ترکیبات مورد بررسی با اندیس بازداری آن‌ها باشد، از روش رگرسیون خطی چندگانه (MLR) استفاده شد. برای مشاهده و مرتب‌کردن توصیف‌کننده‌ها، خروجی نرم‌افزار Dragon را به نرم‌افزار Excel انتقال می-دهیم. طبیعی است که تعداد زیاد توصیف‌کننده‌ها باعث پیچیدگی محاسبات شده و هم‌چنین احتمال وجود فاکتورهایی دارای برهم-کنش باهم را افزایش می‌دهد. لذا، تعدادی از این توصیف‌کننده‌ها که تمام مقادیر آن‌ها صفر و یا بالای ٪۹۰ مقادیر یکسان بودند حذف شدند. نهایتاً، پس از مرحله‌ی حذف مقدماتی اولیه، توصیف‌کننده‌های باقیمانده در یک ماتریس داده‌ی $n \times m$ جای گرفتند. در این ماتریس، جملات n و m، به ترتیب بیان‌گر تعداد ترکیبات و تعداد توصیف‌کننده‌ها هستند. مقادیر عددی این دو جمله به ترتیب برابر ۶۰ و ۱۹۸ و شکل ماتریس مربوطه به صورت ۱۹۸ × ۶۰ می‌باشد. در پایان، بردار ستونی (y) برای محاسبه‌ی متغیر وابسته (اندیس کواتس) ساخته شد.

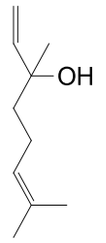
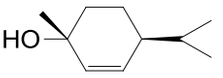
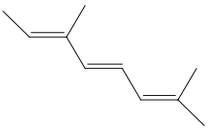
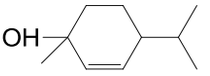
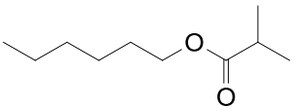
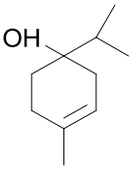
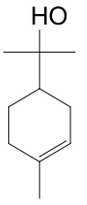
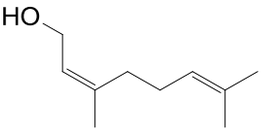
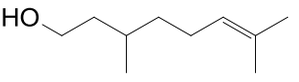
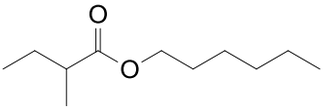
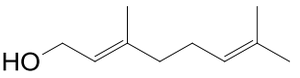
۲-٤. الگوی پراکنش آنالیز جزء اصلی (PCA) و تقسیم سری داده

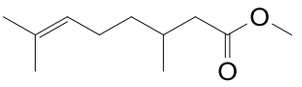
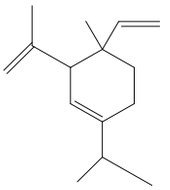
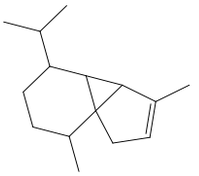
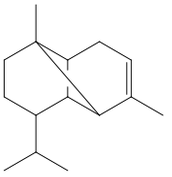
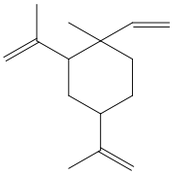
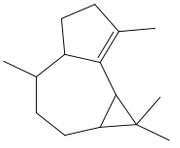
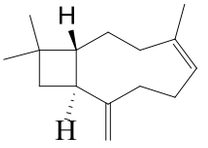
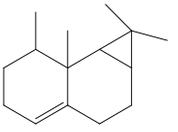
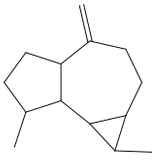
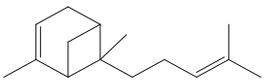
در حقیقت، PCA یک راهکار مطمئن در تجزیه و تحلیل و طبقه‌بندی اطلاعات محسوب می‌شود. PCA، می‌تواند چند جزء زمینه-ای اصلی را تعیین نموده و به تشریح واریانس مبسوط مشاهده شده در اکثر داده‌ها کمک شایانی نماید [۴،۵]. هدف اصلی در هر آنالیز مبتنی بر PCA، مشخص ساختن هر شیء در ماتریس ورودی (ردیف‌ها) بدون تجزیه‌ی هرگونه متغیر (ستون‌ها) می‌باشد. در این راستا، داده‌ها در یک زیر مجموعه‌ی بسیار کوچک‌تر از متغیرهای جدید و یا امتیازات مؤلفه‌های اصلی پیش‌بینی می‌شوند. این متغیرها یا فاکتورهای جدید، ارتباطات خطی مقادیر اولیه بوده و بیان‌گر واریانس مربوط به هر سری داده هستند که می‌توانند موارد اضافی را حذف نمایند. در حقیقت، اجزاء اصلی متوالی، در مرتبه‌ی کاهشی مقادیر ویژه مرتب می‌شوند [۶]. اجزاء اصلی یا PCها قائم یا مستقل بوده و به نحوی مقیاس‌بندی می‌شوند که واریانس‌های مربوطه با مقدار واحد برابر شوند. هم‌چنین، اجزاء اصلی به گونه‌ای آرایش می‌یابند که واریانس بیان شده با اولین PC بیشترین باشد و واریانس‌های مربوط به PC دوم، سوم و .... به ترتیب روند کاهشی داشته باشند. بدین‌ترتیب، آخرین واریانس کم‌ترین مقدار را خواهد داشت [۷]. در ادامه پس از اصلاح و کاهش توصیف-کننده‌ها، با اعمال دستور متنی آنالیز جزء اصلی (PCA) در محیط Matlab، شکل (۲) حاصل شد. در این شکل، هر علامت + نشان‌گر یک ترکیب و شماره‌ی مندرج در کنار آن، بیان کننده‌ی شماره‌ی آن ترکیب است.

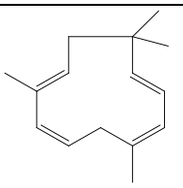
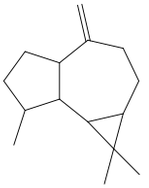
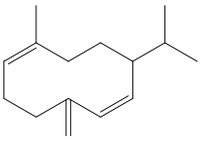
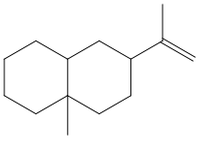
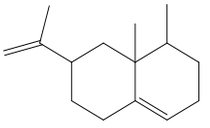
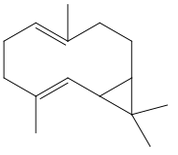
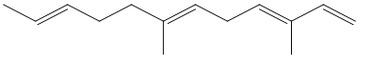
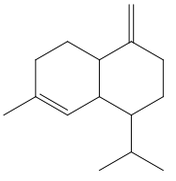
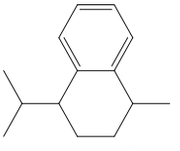
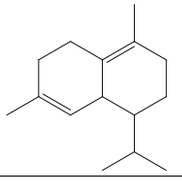
جدول ۱. نام، ساختار و اندیس کواتس ترکیبات اسانس گیاه میخک زینتی

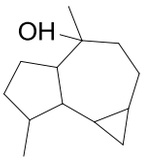
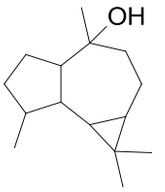
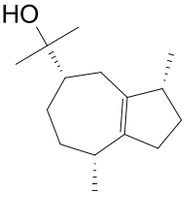
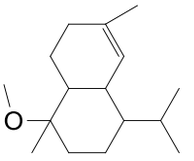
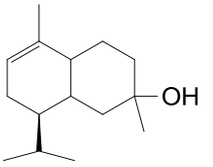
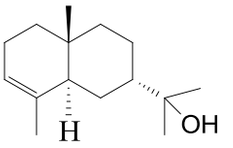
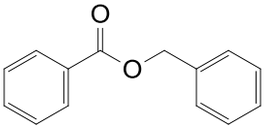
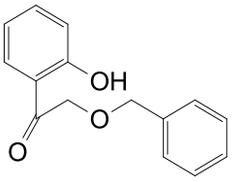
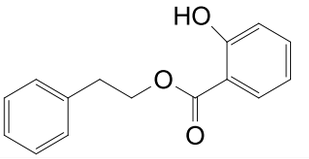
شماره	نام ترکیب	RI	ساختار ترکیب
۱	α-Thujene	924	
۲	α-Pinene	930	
۳	Camphene	938	
۴	Sabinene	958	
۵	β-Pinene	963	
۶	Dehydro-1,8-Cineole	973	
۷	*n*-Octanal	973	
۸	Myrcene	975	
۹	α-Phellandrene	995	

	1002	α-Terpinene	۱۰
	1003	*p*-Cymene	۱۱
	1005	β -Phellandrene	۱۲
	1009	Limonene	۱۳
	1017	*cis*-β-Ocimene	۱۴
	1027	*trans*-β-Ocimene	۱۵
	1035	γ-Terpinene	۱۶
	1037	*trans*-Sabinene hydrate	۱۷
	1064	Terpinolene	۱۸
	1066	*cis*-Sabinene hydrate	۱۹
	1073	*n*-Nonanal	۲۰

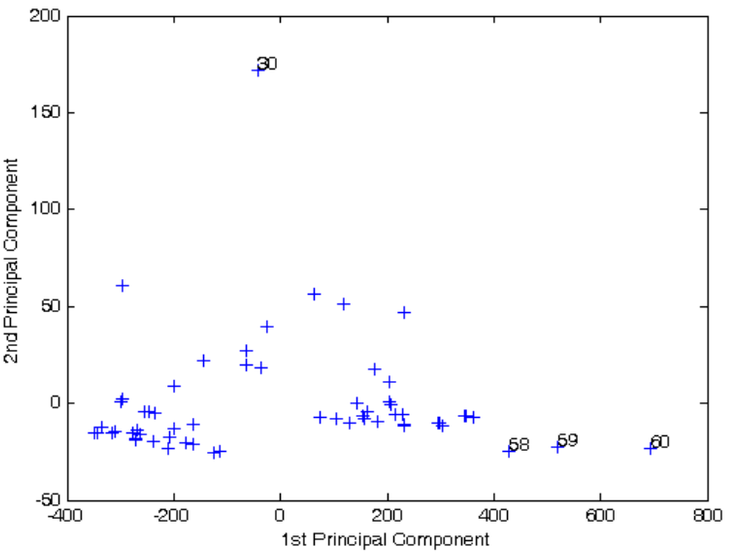
	1074	Linalool	۲۱
	1095	*trans-p*-Menth-2-en-l-ol	۲۲
	1110	*allo*-Ocimene	۲۳
	1110	*cis-p*-Menthen-l-ol	۲۴
	1127	Hexyl isobutyrate	۲۵
	1148	Terpinene-4-ol	۲۶
	1159	α-Terpineol	۲۷
	1206	Nerol	۲۸
	1208	Citronellol	۲۹
	1222	Hexyl-2-Methylbutyrate	۳۰
	1236	Geraniol	۳۱

	1245	Methyl citronellate	۳۲
	1332	δ-Elemene	۳۳
	1345	α-Cubebene	۳۴
	1375	α-Copaene	۳۵
	1388	β-Elemene	۳۶
	1400	α-Gurjunene	۳۷
	1414	β-Caryophyllene	۳۸
	1426	β-Gurjunene	۳۹
	1428	Aromadendrene	۴۰
	1434	*trans*-α-Bergamotene	۴۱

	1447	α-Humulene	۴۲
	1454	*allo*-Aromadendrene	۴۳
	1474	Germacrene-D	۴۴
	1476	β-Selinene	۴۵
	1477	Valencene	۴۶
	1487	Bicyclogermacrene	۴۷
	1500	α-*trans*-*trans*-Farnesene	۴۸
	1500	γ-Cadinene	۴۹
	1505	Calamenene	۵۰
	1505	δ-Cadinene	۵۱

	1566	Globulol	۵۲
	1569	Viridiflorol	۵۳
	1575	Guaiol	۵۴
	1616	T-Cadinol	۵۵
	1618	δ-Cadinol	۵۶
	1634	α-Eudesmol	۵۷
	1701	Benzyl benzoate	۵۸
	1790	Benzyl salicylate	۵۹
	1966	Phenylethyl salicylate	۶۰

بر اساس این شکل، قسمت عمده‌ی ترکیبات در خوشه‌ی اصلی جای گرفته ولی ترکیبات دارای شماره‌های ۳۰، ۵۸، ۵۹ و ۶۰ دارای فاصله‌ی محسوس با سایر ترکیات در سری داده‌ی مفروض هستند. این شکل، اساساً دارای دو پیام است: نخست اینکه احتمال دارد این ترکیبات دارای رفتاری متفاوت با ترکیبات دیگر باشند که این امر در برخی از موارد به حصول نتایج نامناسب و عدم مدل‌سازی مناسب منجر می‌شود. دوم اینکه ۴ ترکیب فوق را نمی‌توان در سری آزمون در نظر گرفت، چرا که در سری آزمون باید ترکیبات دارای رفتار و ماهیت مشابه یا نزدیک با ترکیبات موجود در خوشه‌ی اصلی قرار گیرند. نهایتاً، با در نظر گرفتن الگوی پراکنش نقطه-ای آنالیز مؤلفه‌ی اصلی، جهت تقسیم هوشمند سری داده به مجموعه‌های آموزش و آزمون ترکیباتی در سری آموزش جای گرفتند که نماینده‌ی مناسبی از کل ترکیبات در سری داده‌ی اصلی باشند. وظیفه‌ی خطیر مدل‌سازی، بر دوش ترکیبات سری آموزش قرار داشته و ترکیبات سری آزمون در این امر دخالت ندارند. در عوض، قدرت پیشگوئی مدل و توانمندی آن با اعمال به سری آزمون مورد بررسی قرار می‌گیرد. در مدل سازی مبتنی بر رگرسیون خطی چندگانه روی اسانس گیاه داروئی میخک زینتی، سری اطلاعات اولیه‌ی متشکل از ۶۰ مولکول را به مجموعه‌های آموزشی شامل ۵۱ ترکیب و آزمایشی شامل ۹ ترکیب تقسیم می‌کنیم. این تقسیم-بندی با نسبت تقریبی ۴ به ۱ صورت می‌پذیرد که در نهایت ٪۸۵ از سری اطلاعات در مجموعه‌ی آموزشی و ٪۱۵ از آن‌ها در مجموعه‌ی آزمایشی قرار می‌گیرند.



شکل ۲. الگوی پراکنش نقطه‌ای ترکیبات سری داده‌ی اصلی میخک زینتی با استفاده از دستور آنالیز جزء اصلی(PCA)

۳. نتایج و بحث

۳-۱. تحلیل مدل‌های آماری و انتخاب مدل مناسب

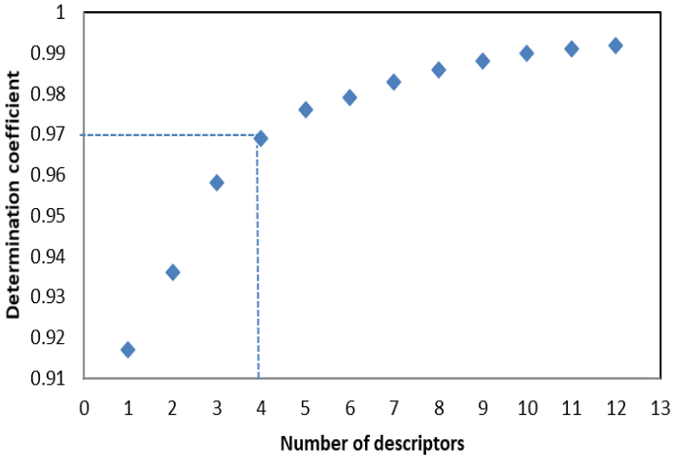
برای به‌دست‌آوردن مدلی مناسب، سری اطلاعات مجموعه‌ی آموزشی را به‌عنوان متغیر مستقل و مقادیر تجربی اندیس بازداری را به عنوان متغیر وابسته به نرم‌افزار آماری SPSS معرفی کرده و با استفاده از منوی آنالیز، گزینه‌ی رگرسیون خطی با روش مرحله‌ای مدل‌سازی صورت می‌پذیرد. نهایتاً، چندین مدل مختلف به‌طور جداگانه به‌دست می‌آید که با توجه به خصوصیات آماری آن‌ها، از جمله ضریب رگرسیون یا ضریب هم‌بستگی (R)، آماره‌ی F و خطای استاندارد (SE)، پس از رسم مقادیر R و R^2 و SE برحسب

تعداد توصیف‌کننده‌ها بهترین مدل که دارای بیش‌ترین مقدار R و F، و کم‌ترین مقدار خطای استاندارد و شامل توصیف‌کننده‌های تا حد امکان قابل توجیه باشد، به عنوان مدل نهایی برای ارتباط اندیس بازداری مولکول‌ها با ساختار آن‌ها انتخاب شد. بر اساس محاسبات نهائی، مجموعه‌ای شامل ۱۴ مدل رگرسیون در نرم‌افزار SPSS حاصل شد. نتایج عددی خلاصه‌ی مدل‌های ارائه شده در محیط SPSS، در جدول ۲ نمایش داده شده است.

در شکل ۳، روند تغییرات مقادیر ضرایب تعیین (R^2) بر حسب تعداد توصیف‌کننده‌ها نشان داده شده است. به عنوان یک معیار کلی، در روش‌های رگرسیون خطی چندگانه، با رسم روند تغییر ضریب اندازه‌گیری به‌عنوان تابعی از توصیف‌کننده و تعیین نقطه‌ی شکست نمودار، می‌توان به تعداد توصیف‌کننده‌ی بهینه پی برد. بدیهی است هرچه تعداد توصیف‌کننده‌های انتخابی کم‌تر باشد، مدل انتخابی ساده‌تر بوده و از اعتبار بیشتری برخوردار است.

جدول ۲. خلاصه‌ی ۱۲ مدل منتخب پس از محاسبات رگرسیون در محیط نرم‌افزار SPSS سری داده‌ی اسانس گیاه میخک زینتی

شماره‌ی مدل	ضریب هم-بستگی (r)	ضریب اندازه‌گیری (r^2)	ضریب اندازه‌گیری تنظیم شده ($Adj.r^2$)	خطای استاندارد تخمین
۱	0.958	0.917	0.916	75.96
۲	0.967	0.936	0.933	67.75
۳	0.979	0.958	0.955	55.28
۴	0.985	0.969	0.967	47.84
۵	0.988	0.976	0.973	43.1
۶	0.990	0.979	0.977	40.04
۷	0.991	0.983	0.980	37.3
۸	0.993	0.986	0.984	33.4
۹	0.994	0.988	0.986	31.1
۱۰	0.995	0.990	0.988	29.1
۱۱	0.996	0.991	0.989	27.7
۱۲	0.996	0.992	0.990	26.1



شکل ۳. منحنی تغییرات مقادیر ضرایب تعیین (R^2) برحسب تعداد توصیف‌کننده‌ها با استفاده از نرم‌افزار SPSS سری داده‌ی اسانس گیاه میخک زینتی

در این مرحله، مدل انتخابی را برای مجموعه‌ی آزمایشی نیز اعمال می‌کنیم. از آن‌جایی که توصیف‌کننده‌ها باید متغیرهایی مستقل باشند، لذا برای اطمینان از عدم هم‌بستگی بین آن‌ها، در نرم‌افزار SPSS، با استفاده از منوی آنالیز، گزینه‌ی هم‌بستگی دو متغیره را انتخاب می‌کنیم. چنانچه هم‌بستگی بین توصیف‌کننده‌ها بیش‌تر از ۰/۸۵ باشد، توصیف‌کننده‌های وابسته را از سری اطلاعات حذف کرده و مراحل مذکور را تا جایی که هم‌بستگی بین توصیف‌کننده‌ها کم‌تر از این مقدار باشد، مجدداً انجام می‌دهیم. مقادیر عددی توصیف‌کننده‌های وارد شده در مدل، در جدول ۳ آورده شده است.

جدول ۳. مقادیر عددی توصیف‌کننده‌های واردشده در مدل خطی نهائی سری داده‌ی اسانس گیاه میخک زینتی

شماره ترکیبات	توصیف کننده های نهایی انتخاب شده			
	LPRS	Mor29m	R7m+	MATS1e
Training set				
1	30.437	-0.208	0.001	0.059
2	30.023	-0.108	0	0.059
3	29.615	-0.213	0	0.059
4	30.437	-0.241	0.001	0.059
5	34.504	-0.198	0	0.048
6	29.343	-0.042	0.017	-0.004
7	33.542	-0.079	0.025	-0.064
8	31.601	-0.076	0.01	0
9	31.601	-0.08	0.009	0
10	31.601	0.009	0.01	0
11	31.601	-0.073	0.01	0
12	33.542	-0.064	0.029	-0.064
13	33.542	-0.044	0.021	-0.064
14	31.601	-0.019	0.009	0
15	51.62	-0.283	0.051	0.016
16	31.601	-0.077	0.009	0
17	35.153	-0.191	0.001	-0.066
18	34.733	-0.038	0.012	-0.004
19	36.453	-0.061	0.015	-0.072
20	33.542	0.008	0.019	-0.064

21	36.453	-0.079	0.014	-0.079
22	45.177	-0.095	0.02	0.017
23	36.011	-0.052	0.009	-0.072
24	36.32	-0.087	0.012	-0.072
25	38.959	0.054	0.019	-0.078
26	38.959	-0.024	0.018	-0.073
27	50.278	-0.179	0.024	0.019
28	38.959	0.054	0.019	-0.078
29	57.073	-0.173	0.033	0
30	55.962	-0.333	0.01	0.078
31	57.073	-0.12	0.029	0
32	55.604	-0.31	0.007	0.078
33	56.902	-0.325	0.013	0.04
34	55.604	-0.298	0.007	0.078
35	58.548	-0.095	0.033	0.04
36	57.938	-0.181	0.017	0
37	55.604	-0.425	0.009	0.078
38	58.08	-0.132	0.02	0
39	56.501	-0.071	0.02	0.04
40	57.423	-0.236	0.027	0.04
41	62.188	-0.037	0.021	-0.042
42	56.588	0.047	0.014	0.039
43	56.588	-0.147	0.014	0.04
44	60.845	-0.336	0.01	-0.049
45	60.845	-0.5	0.011	-0.049
46	61.991	-0.183	0.024	-0.054
47	61.818	-0.002	0.015	-0.054
48	62.168	-0.016	0.028	-0.054
49	65.768	0.12	0.032	0.001
50	71.272	0.132	0.046	-0.069

51	78.125	0.213	0.035	-0.06
Test set				
1	30.023	-0.077	0	0.059
2	31.601	-0.068	0.011	0
3	38.152	-0.099	0.025	-0.078
4	49.799	-0.017	0.017	0.017
5	55.742	-0.034	0.008	0.078
6	55.264	-0.265	0.006	0.078
7	56.677	-0.116	0.022	0.04
8	56.588	-0.07	0.014	0.04
9	61.818	-0.045	0.016	-0.054

۳-۲. حوزه‌ی کاربرد (AD)

در تحقیق اخیر، نمودار ویلیام یا نمودار روند تغییرات باقیمانده‌های استانداردشده بر حسب لورج یا قدرت نفوذ [۸] جهت نمایش حوزه‌ی کاربرد فرآیند مدل‌سازی استفاده و نتایج حاصله در شکل ۴، نمایش داده شده است. طبق تعریف، لورج یا قدرت نفوذ نشان‌گر فاصله‌ی یک ترکیب از مرکز ثقل محور X می‌باشد. طبق یک قرارداد آماری، لورج یک ترکیب در فضای متغیر اولیه به صورت رابطه‌ی ۱، تعریف می‌شود [۹].

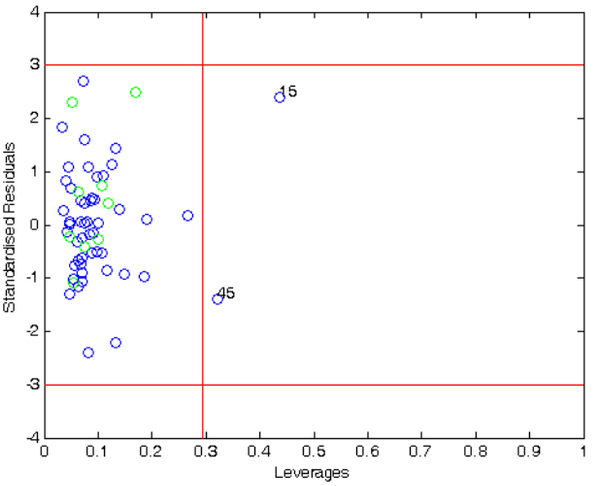
$$h_i = x_i^T \left(X^T X \right)^{-1} x_i \qquad (۱)$$

در این رابطه، x_i بردار توصیف‌کننده‌ی ترکیب مفروض و X ماتریس توصیف‌کننده‌ی حاصل از مقادیر توصیف‌کننده‌ی سری آموزش می‌باشد. مقدار لورج هشدار (h^*)، به‌صورت رابطه‌ی ۲ تعریف می‌شود [۱۰].

$$h^* = 3\,(p+1)/\,n \qquad (۲)$$

در رابطه‌ی (۲)، n تعداد ترکیبات در سری آموزش، p، تعداد متغیرهای پیشگوئی‌کننده می‌باشد. منظور از تعداد متغیرهای پیشگوئی-کننده، همان تعداد توصیف‌کننده‌های مولکولی منتخب در مدل‌سازی نهائی است. در فرآیندهای مدل‌سازی، ترکیبات دارای لورج بیشتر از لورج هشدار ($h_i > h^*$) می‌توانند در بسیاری از موارد به شدت عملکرد مدل‌سازی را تحت تأثیر قرار دهند. در عین حال، در کلیه‌ی موارد نمی‌توان آن را به عنوان یک ترکیب انحرافی در نظر گرفت چرا که ممکن است باقیمانده‌ی استاندارد شده‌ی آن جزئی و قابل اغماض باشد. در شرایطی که برای یک ترکیب معین هم شرط $h_i > h^*$ برقرار و هم باقیمانده‌ی استاندارد شده‌ی ترکیب مداری بزرگ و قابل توجه باشد، ترکیب مربوطه از حوزه‌ی کاربرد و در نتیجه سری داده‌ی اولیه حذف و مدل‌سازی با ترکیبات

باقیمانده انجام خواهد شد. تحت این شرایط، باید توجیهی قانع‌کننده پیرامون علت حذف این ترکیب با در نظرگرفتن ساختار آن ارائه نمود. بعلاوه، مقدار ۳ برابر باقیمانده‌ی استاندارد شده معمولاً تحت عنوان مقدار برش برای قبول پیشگوئی‌ها استفاده می‌شود. چرا که ثابت شده که نقاطی که در بازه‌ی ۳± برابر باقیمانده‌های استاندارد شده از مقدار میانگین قرار می‌گیرند، ۹۹٪ اطلاعات دارای پراکنش نرمال را در برمی‌گیرند [۱۱]. بدین‌ترتیب، ترکیب لورج و باقیمانده‌های استاندارد شده برای مشخص نمودن حوزه‌ی کاربرد مورد استفاده قرار گرفت.



شکل ٤. نمودار ویلیام برای ترکیبات سری‌های آموزش و آزمون سری داده‌ی اسانس گیاه میخک زینتی

بررسی شکل ۴، خاطر نشان می‌سازد که دو ترکیب (ترکیبات شماره‌ی ۱۵ و ۴۵) در سری آموزش دارای مقادیر لورج بیشتر از مقدار هشدار (h*) هستند. براساس رابطه‌ی ۲، مقدار لورج هشدار برابر با h*= 0.2941 است. بنابراین، می‌توانند جزو ترکیبات انحرافی باشند. خوشبختانه، در دو مورد اخیر مقادیر اطلاعات عددی پیشگوئی شده مناسب و قابل قبول هستند. لذا، این دو ترکیب را لورج خوب خوانده و از حوزه‌ی کاربرد و سری داده‌ی اولیه حذف نمی‌کنیم. در یک جمع‌بندی، می‌توان گفت که نمودار حوزه‌ی کاربرد (نمودار ویلیام یا نمودار AD) هم تأئیدکننده‌ی مناسب بودن مدل ساخته‌شده و هم بیان‌گر تقسیم مناسب سری داده‌ی اولیه به مجموعه‌های آموزش و آزمون است.

۳-۳. فاکتور تورم تغییر (VIF)

هم‌بستگی خطی چندگانه‌ی بین توصیف‌کننده‌های منتخب، با محاسبه‌ی پارامتری موسوم به فاکتور تورم تغییر (VIF) تعیین شد که در قالب رابطه‌ی ۳ قابل بیان است.

$$VIF = \frac{1}{1-r^2} \qquad (۳)$$

در رابطه‌ی اخیر، r ضریب هم‌بستگی رگرسیون چندگانه‌ی بین متغیرها در مدل نامیده می‌شود. طبق قرارداد، اگر VIF برابر یک باشد، هیچ‌گونه هم‌بستگی داخلی بین متغیرها وجود ندارد و در صورتی‌که بازه‌ی تغییرات VIF بین ۱ تا ۵ باشد، مدل مربوطه قابل قبول بوده و اگر مقادیر عددی محاسبه‌شده برای VIF بزرگ‌تر از ۱۰ باشد، مدل مربوطه ناپایدار بوده و یک بازنگری و تصحیح اساسی جهت توسعه‌ی آن ضروری خواهد بود [۱۲]. مقادیر VIF در جدول ۴ گزارش شده است.

۳-٤. اثر متوسط (MF_j) توصیف‌کننده‌های منتخب

جهت بررسی اهمیت نسبی و تأثیر هر توصیف‌کننده در مدل، مقدار اثر متوسط (MF_j) برای هر توصیف‌کننده محاسبه شد. محاسبه‌ی مقادیر عددی این پارامتر مهم آماری با استناد به رابطه‌ی ۴، صورت گرفت [۱۳].

$$MF_j = \frac{\beta_j \sum_{i=1}^{n} d_{ij}}{\sum_{i}^{m} \beta_j \sum_{i}^{n} d_{ij}} \qquad (۴)$$

در این رابطه، MF_j اثر متوسط توصیف‌کننده‌ی مفروض jام، β_j ضریب توصیف‌کننده‌ی j، d_{ij} مقدار توصیف‌کننده‌های هدف برای هر مولکول و m شماره‌ی توصیف‌کننده در مدل ارائه شده می‌باشند. در حقیقت، مقدار MF_j نشان‌گر اهمیت نسبی یک توصیف‌کننده در مقایسه با سایر توصیف‌کننده‌ها در مدل می‌باشد. در مجموعه‌ی توصیف‌کننده‌های انتخاب‌شده‌ی بهینه با روش‌های انتخاب متغیر مختلف، توصیف‌کننده‌ی LPRS دارای بیشترین مقدار MF_j، بیشترین تأثیر را نسبت به سایر متغیرها در مدل ارائه شده خواهد داشت. علامت آن، نشان‌گر جهت تغییر در مقادیر عددی متغیر وابسته به عنوان نتیجه‌ای از کاهش یا افزایش در مقادیر توصیف‌کننده است. مقادیر عددی محاسبه شده‌ی اثر متوسط (MF_j) در جدول ۴ آورده شده‌اند. بر اساس این جدول، مقدار MF_j نسبت داده شده به توصیف‌کننده‌ی به مراتب بیشتر از سایر توصیف‌کننده‌ها بوده بنابراین انتظار می‌رود این توصیف‌کننده دارای بیشترین تأثیر باشد. بنابراین، معادله‌ی به‌دست آمده برای داده‌های اندیس بازداری این سری از ترکیبات به‌صورت معادله‌ی ریاضی ۵ قابل بیان می‌باشد:

RI= 406.340(±24.9) + 21.199(±0.6) **LPRS** + 305.364(±58.2) **MATS1e** -5515.943(±816.7) **Mor29m** - 644.946(±156.6) **R7m+** (۵)

اعداد داخل پرانتز، بیان‌گر خطای استاندارد (انحراف استاندارد) ضرایب و عرض از مبدأ واردشده در مدل خطی بر اساس رگرسیون خطی چندگانه هستند. طبق جدول ۴، هرکدام از توصیف‌کننده‌های وارد شده در مدل، یک خصوصیت ساختاری از مولکول را بیان می‌کنند که نشان‌دهنده‌ی تأثیر بسزای ساختار مولکول‌های مورد بررسی بر اندیس کواتس آن‌ها بوده و از طرفی ضریب هم‌بستگی میان توصیف‌کننده‌های وارد شده در مدل منتخب، نشان‌دهنده‌ی عدم هم‌بستگی قابل‌توجه میان این توصیف‌کننده‌ها می‌باشد.

علامت مثبت نسبت داده شده به هر توصیف‌کننده که در جدول ۴ آورده شده است، دلالت بر تقویت مدل به وسیله‌ی آن‌ها داشته در حالی‌که علامت منفی بیان‌گر رابطه‌ی معکوس بین متغیر وابسته (اندیس کواتس) و متغیرهای مستقل و توصیف‌کننده‌های مولکولی دارد. بر این اساس، توصیف‌کننده‌های LPRS و MATS1e به‌ترتیب از گروه‌های توپولوژیکال و خود ارتباطی دوبعدی،

دارای تأثیر سازنده و افزایشی بر مدل خطی ارائه شده هستند. یعنی، با افزایش مقادیر این دو توصیف‌کننده مقدار اندیس کواتس به عنوان یک متغیر وابسته نیز افزایش می‌یابد. در عین‌حال، توصیف‌گرهای Mor29m و R7m+ دارای علامت منفی هستند.

جدول ٤. مشخصات مدل منتخب سری داده‌ی اسانس گیاه میخک زینتی به‌روش SW-MLR

No.	Symbol	Descriptor description	Descriptor group	Coefficient	Mean effect	VIF
1	Intercept		-	405.8	-	-
2	LPRS	log of product of row sums (PRS)	Topological	21.2	1.1413	1.6181
3	MATS1e	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity	2D autocorrelations	302.6	-0.0406	1.4854
4	Mor29m	signal 29 / weighted by mass	3D-MoRSE	-5445.8	-0.1065	1.8875
5	R7m+	R maximal autocorrelation of lag 7 / weighted by mass	GETAWAY	-641.7	0.0058	1.4133

این توصیف‌گرها به‌ترتیب به خانواده‌های 3D-MoRSE و GETAWAY تعلق دارند. نکته‌ی قابل تأمل دیگر، آن است که قدر مطلق مقدار اثر متوسط (MF_j) (رابطه‌ی ۴)، توصیف‌کننده‌ی LPRS در مقایسه با سایر توصیف‌کننده‌ها بیشتر است. این بدان معنی است که این توصیف‌کننده در قیاس با سایر توصیف‌کننده‌ها دارای بیشترین تأثیر در مدل خطی ارائه شده است. مورد دیگر قابل توجه آن است که مقادیر عددی فاکتور تورم تغییر (VIF)، براساس رابطه‌ی ۳، در گستره‌ی ۱/۴ تا ۱/۹ قرار داشته و این به‌مفهوم قابل قبول‌بودن و پایداربودن مدل ارائه‌شده می‌باشد.هم‌چنین، جدول ۵ ماتریس هم‌بستگی توصیف‌کننده‌های وارد شده در مدل نهایی برای سری داده‌ها را نمایش می‌دهد. مقادیر عددی مندرج در این جدول، بیان‌گر ضرایب هم‌بستگی بین هر جفت توصیف‌کننده‌ی انتخاب شده در مدل خطی نهائی هستند. بر اساس جدول ۵، بیشترین مقدار ضریب هم‌بستگی در این ماتریس برابر با 0.541 و بین توصیف‌کننده‌های MATS1e و Mor29m می‌باشد. بازه‌ی عددی بسیار کم توصیف‌کننده‌های انتخابی دلالت بر اعتبار مدل ارائه شده و شاهدی بر رفتار مستقل این متغیرها می‌باشد.

جدول ۵. ماتریس هم‌بستگی ٤ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی اسانس گیاه میخک زینتی به‌روش SW-MLR

	LPRS	MATS1e	Mor29m	R7m+
LPRS	1	-0.066	0.541	0.052
MATS1e	-0.066	1	0.346	-0.5
Mor29m	0.541	0.346	1	-0.316
R7m+	0.052	-0.5	-0.316	1

در ادامه هنگامی که از عدم وجود هم‌بستگی بین توصیف‌کننده‌ها اطمینان حاصل شد، با استفاده از معادله‌ی ریاضی ۵ مربوط به مدل ارائه شده، ضرایب رگرسیون موجود در جدول ۴ و نیز مقادیر وارد شده‌ی توصیف‌کننده‌ها در مدل، اندیس‌های بازداری ترکیبات را توسط نرم‌افزار قدرتمند MATLAB محاسبه و پیش‌بینی می‌کنیم. نتایج حاصل‌شده در این محاسبه در جدول ۶ نشان داده شده است.

جدول ٦. مقادیر تجربی و پیش‌بینی‌شده‌ی اندیس‌های بازداری (RI)، اسانس گیاه میخک زینتی توسط روش رگرسیون خطی چندگانه‌ی مرحله‌ای

No.	Compound name	K.I. (Lit.)[a]	K.I. (Cal.)[b]	Dif. [c]	R.E.[d] %
Training set					
1	α-Thujene	924	944.5	20.5	2.2
2	α-Pinene	930	971.8	41.8	4.5
3	Camphene	938	931.1	-6.9	-0.7
4	Sabinene	958	934.4	-23.6	-2.5
5	Dehydro-1,8-Cineole	973	1046.4	73.4	7.5
6	*n*-Octanal	973	924.4	-48.6	-5
7	Myrcene	975	996.7	21.7	2.2
8	α-Phellandrene	995	997.9	2.9	0.3
9	α-Terpinene	1002	1002.2	0.2	0
10	*p*-Cymene	1003	1023.8	20.8	2.1
11	Limonene	1009	998.8	-10.2	-1
12	*cis*-β-Ocimene	1017	979.2	-37.8	-3.7
13	*trans*-β-Ocimene	1027	1029.4	2.4	0.2
14	γ-Terpinene	1035	1020.8	-14.2	-1.4
15	*trans*-Sabinene hydrate	1037	1122.6	85.6	8.3
16	Terpinolene	1064	1003.1	-60.9	-5.7
17	*cis*-Sabinene hydrate	1066	1130.3	64.3	6
18	*n*-Nonanal	1073	1067.4	-5.6	-0.5
19	Linalool	1074	1097.3	23.3	2.2
20	*allo*-Ocimene	1110	1056.3	-53.7	-4.8
21	*cis*-*p*-Menthen-l-ol	1110	1128.7	18.7	1.7
22	Hexyl isobutyrate	1127	1213.8	86.8	7.7

23	Terpinene-4-ol	1148	1150.7	2.7	0.2
24	α-Terpineol	1159	1130	-29	-2.5
25	Nerol	1206	1194.2	-11.8	-1
26	Citronellol	1208	1172.7	-35.3	-2.9
27	Hexyl 2-Methylbutyrate	1222	1272.9	50.9	4.2
28	Geraniol	1236	1194.2	-41.8	-3.4
29	δ-Elemene	1332	1381.4	49.4	3.7
30	α-Cubebene	1345	1385.5	40.5	3
31	β-Elemene	1388	1419.6	31.6	2.3
32	α-Gurjunene	1400	1401.5	1.5	0.1
33	β-Caryophyllene	1414	1415.9	1.9	0.1
34	Aromadendrene	1428	1405.2	-22.8	-1.6
35	*trans*-α-Bergamotene	1434	1410.7	-23.3	-1.6
36	α-Humulene	1447	1485.5	38.5	2.7
37	*allo*-Aromadendrene	1454	1355.4	-98.6	-6.8
38	Germacrene-D	1474	1487	13	0.9
39	Valencene	1477	1446.3	-30.7	-2.1
40	Bicyclogermacrene	1487	1376.9	-110.1	-7.4
41	α-*trans*-*trans*-Farnesene	1500	1624.6	124.6	8.3
42	Calamenene	1505	1517.9	12.9	0.9
43	δ-Cadinene	1505	1458	-47	-3.1
44	Globulol	1566	1570	4	0.3
45	Viridiflorol	1569	1514.4	-54.6	-3.5
46	Guaiol	1575	1567.1	-7.9	-0.5
47	δ-Cadinol	1618	1668.3	50.3	3.1
48	α-Eudesmol	1634	1599.7	-34.3	-2.1
49	Benzyl benzoate	1701	1660.1	-40.9	-2.4
50	Benzyl salicylate	1790	1748.3	-41.7	-2.3
51	Phenylethyl salicylate	1966	1973.2	7.2	0.4

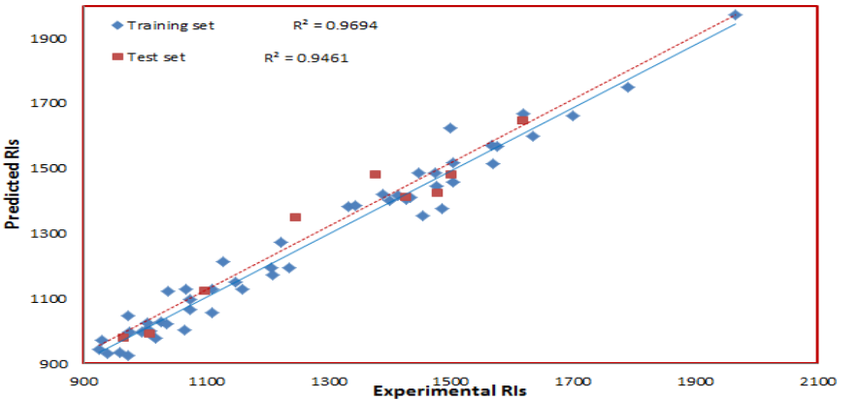
Test set					
1	β-Pinene	963	981.2	18.2	1.9
2	β-Phellandrene	1005	994.8	-10.2	-1
3	*trans-p*-Menth-2-en-l-ol	1095	1124.2	29.2	2.7
4	Methyl citronellate	1245	1352.1	107.1	8.6
5	α-Copaene	1375	1483.2	108.2	7.9
6	β-Gurjunene	1426	1413.6	-12.4	-0.9
7	β-Selinene	1476	1425.3	-50.7	-3.4
8	γ-Cadinene	1500	1481.6	-18.4	-1.2
9	T-Cadinol	1616	1649.7	33.7	2.1

A : Retention index in literature: اندیس کواتس در منابع علمی B: Retention index calulated: اندیس کواتس محاسبه شده C: Difference: تفاوت D: Relative Error: خطای نسبی

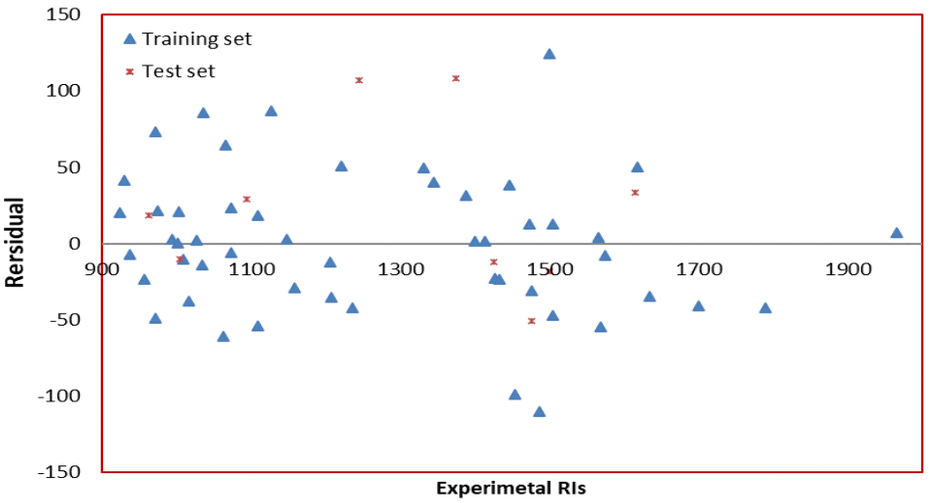
۳-۵. نمودارهای پیش‌بینی و باقیمانده

یکی از مواردی که می‌توان با استناد به آن از معتبر بودن مدل انتخابی اطمینان‌خاطر پیدا کرد، به‌دست‌آوردن ضریب اندازه‌گیری (R^2) بین مقادیر تجربی یک کمیت و مقادیر نظیر پیش‌گوئی شده برای آن می‌باشد. براساس یک قاعده‌ی کلی، هرقدر مقدار ضریب اندازه‌گیری در نمودار پیشگوئی به یک نزدیک‌تر باشد، مدل ساخته‌شده معتبرتر خواهد بود. جهت ارزیابی شمائی اعتبار مدل‌های انتخاب‌شده، مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری را برحسب مقادیر تجربی مربوطه (شکل ۵) رسم می‌کنیم. براساس شکل ۵، یک هم‌گرائی مطلوب بین مقادیر تجربی و مقادیر پیشگوئی‌شده به‌چشم می‌خورد. در این شکل، مقدار ضریب اندازه‌گیری برای مجموعه‌های آموزش و آزمون به‌ترتیب معادل ۰/۹۶۹۴ و ۰/۹۴۶۱ است که دلالت بر اعتبار مدل و توانمندی بالای آن جهت پیش‌گوئی مقادیر اندیس کواتس گستره‌ی وسیعی از ترکیبات دارد.

مورد دیگری که می‌توان برای اعتبار مدل به آن استناد کرد، رسم مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی شده‌ی اندیس‌های بازداری برحسب مقادیر تجربی آن می‌باشد. این نمودار، موسوم به نمودار باقی‌مانده در شکل ۶ آورده شده است.



شکل ۵. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن



شکل ٦. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن

در نمودار باقی‌مانده، می‌توان پراکندگی طبیعی نسبتاً یکسان نقاط را حول مقادیر صفر، ناشی از عدم وجود خطای معین در روش و اعتبار بالای مدل دانست.

۳-٦. اعتبارسنجی تقاطعی مدل خطی ارائه شده

یکی از رایج‌ترین روش‌های دیگر که پیش‌تر بدان پرداخته شد، اعتبارسنجی تقاطعی (CV) نام دارد. در این روش، در هر مرحله یک یا یک گروه کوچک از داده‌ها کنار گذاشته می‌شود. سپس برای داده‌هایی که باقی‌مانده، مدلی محاسبه شده و پاسخ از روی مدل محاسبه شده برای یک داده یا گروهی از داده‌ها که کنار گذاشته شده است پیش‌بینی می‌شود. در تکنیک حذف تکی (*LOO*)، در مراحل متوالی تنها یک ترکیب و مقادیر عددی آن از سری آموزش نهائی حذف شده و مدل‌سازی با بقیه انجام می‌شود. در مرحله‌ی بعد، مدل برای ترکیب کنار گذاشته شده، اعمال می‌شود. از طرف دیگر، در تکنیک حذف گروهی (*LGO*)، در هر مرحله داده‌های مربوط به چند مولکول (پنج ملکول) منحصراً از مجموعه‌ی آموزشی، کنار گذاشته می‌شود. در ادامه، با توجه به داده‌های باقی‌مانده مدل خطی ساخته می‌شود. سپس، این مدل را برای گروه کنار گذاشته شده اعمال می‌کنیم که مبنای پیش‌بینی اندیس‌های بازداری برای آن گروه می‌باشد. در نهایت، با استفاده از یک معادله‌ی ریاضی، تحت مدل ورود اجباری (Enter) و اعمال ضرائب توصیف‌کننده‌های وارد شده در مدل، اندیس‌های بازداری ترکیبات توسط نرم‌افزار قدرتمند MATLAB هم در شرایط حذف تک مولکول و در شرایط حذف گروهی مولکول‌ها محاسبه و پیش‌بینی شد. نتایج حاصل از انجام روش اعتبارسنجی تقاطعی با روش‌های حذف تکی (*LOO*) و حذف گروهی (*LGO*)، در جدول ۷ نشان داده شده است. طبق قرارداد، در مراجع علمی در مطالعات ارتباط کمی ساختار- فعالیت یا ارتباط کمی ساختار- ویژگی شامل اعتبارسنجی تقاطعی، ضریب اندازه‌گیری یا مجذور ضریب هم‌بستگی بین مقادیر تجربی و مفادیر پیش‌گوئی‌شده، با علامت Q^2 نمایش داده می‌شود. بر اساس مقالات مرجع و پراستناد در پایگاه‌های علمی، مقادیر بالا و قابل توجه Q^2، بیانگر استحکام [۱۴]، پایداری ، صحت و قدرت پیشگوئی بالای مدل خطی ارائه‌شده دارند [۱۵-

۱۸]. براساس نتایج بدست آمده، مقدار Q^2 برای روش‌های حذف تکی (*LOO*) و حذف گروهی (*LGO*)، به‌ترتیب برابر ۰/۹۵۸ و ۰/۹۵۵ است. این مقادیر، نشان‌گر اعتبار تقاطعی مدل ارائه‌شده هستند.

۳-۷. نتایج آماری مدل ارائه‌شده اسانس گیاه میخک زینتی به‌روش SW-MLR

در این بخش، مقادیر تجربی و پیش‌بینی‌شده‌ی اندیس‌های بازداری را به‌طور جداگانه در یک فایل متنی قرار داده و با استفاده از یک برنامه‌ی جامع در نرم‌افزار MATLAB به‌همراه نتایج حاصل از نرم‌افزار SPSS، پارامترهای آماری شاخص مربوطه را محاسبه می‌کنیم. نتایج حاصل از این محاسبه، در جدول ۷ درج شده است.

جدول ۷. پارامترهای آماری جامع به‌دست آمده برای مجموعه‌های آموزش و آزمون اسانس گیاه *Pittosporum undulatum* به‌روش SW-MLR

شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2 الف)	0.969
	سری آزمون		0.946
۲	سری آموزش	Adjusted R^2 ب)	0.967
۲	سری آموزش	REP (%) ج)	3.6
	سری آزمون		4.3
۳	سری آموزش	RMSEP د)	45.4
	سری آزمون		56.4
۴	سری آموزش	SEP ه)	45.8
	سری آزمون		59.9
۵	سری آموزش	RSEP و)	3.5
	سری آزمون		4.3
۶	سری آموزش	MAE ز)	82.8
	سری آزمون		218.9
۷	سری آموزش	F ح)	362.8
	سری آزمون		16.3

الف) ضریب اندازه‌گیری، ب) ضریب اندازه‌گیری تنظیم‌شده، ج) درصد خطای نسبی پیشگوئی، د) جذر میانگین مجذور خطای پیشگوئی، ه) خطای استاندارد پیشگوئی، و) خطای استاندارد نسبی پیشگوئی، ز) خطای مطلق میانگین،ح) آماره‌ی F

در جدول ۷، مقادیر عددی ضریب اندازه‌گیری تنظیم شده (Adjusted R^2) و آماره‌ی F به‌توسط روابط ۶ و ۷ قابل تعیین هستند.

$$R^2_{Adj.}=1-(1-R^2)\left(\frac{n-1}{n-p-1}\right) \qquad (۶)$$

$$F=\frac{MSR}{MSE} \qquad (۷)$$

در رابطه‌ی (۶)، R^2 ،n و p به‌ترتیب بیان‌گر ضریب اندازه‌گیری، تعداد ترکیبات در سری آموزش و تعداد توصیف‌کننده‌ها هستند. در منابع علمی، $R^2_{Adj.}$ به‌روش مشابهی با R^2 (ضریب اندازه‌گیری) با در نظر گرفتن تعداد درجات آزادی بیان می‌شود. پارامتر $R^2_{Adj.}$

با تقسیم‌کردن باقیمانده‌ی مجموع مربعات و مجموع مربعات کل بر درجات آزادی مربوطه تنظیم می‌شود. مقدار $R^2_{Adj.}$ در صورتی‌که متغیر (توصیف‌کننده‌ی) اضافه‌شده به معادله، واریانس توضیح داده نشده را کاهش ندهد کم خواهد شد [۱۹]. بنابراین، $R^2_{Adj.}$ معیاری مناسب برای مقایسه‌ی مدل‌های دارای تعداد متفاوتی از متغیرهای پیشگوئی‌کننده (توصیف‌کننده‌ها) می‌باشد.

در علم آمار، آماره‌ی F اساساً برای ارزیابی دقت مدل‌های پیشنهادی به‌طور وسیع مورد استفاده قرار می‌گیرد. در رابطه‌ی تعیین آماره‌ی F (۷)، MSR و MSE به ترتیب دلالت بر میانگین مربعات رگرسیون و میانگین مربعات خطا دارند [۲۰]. این دو پارامتر به‌توسط روابط ۸ و ۹ قابل تعیین هستند.

$$MSR=\frac{SSR}{m} \qquad (۸)$$

$$MSE=\frac{SSE}{n-m-1} \qquad (۹)$$

در معادلات ۸ و ۹، SSR و SSE نشان‌گر مجموع مربعات رگرسیون و مجموع مربعات خطا هستند. m و n هم به ترتیب بیانگر تعداد توصیف‌کننده‌ها (۴) و تعداد ترکیبات در سری آموزش (۵۱) می‌باشند.

بعلاوه، پارامتر SSE، موسوم به مجموع مربعات خطاها بر اساس رابطه‌ی ۱۰، قابل بیان می‌باشد.

$$SSE=\sum_{i=1}^{n}(y_{exp}-y_{pred})^2 \qquad (۱۰)$$

تأمل روی رابطه‌ی ۱۰ نشان می‌دهد که رابطه‌ی SSE دقیقاً با رابطه‌ی PRESS برابر می‌باشد. مجموع مربعات خطا، برطبق رابطه‌ی ۱۱، قابل محاسبه است.

$$SSR=\sum_{i=1}^{n}(y_{pred}-\bar{y}_{pred})^2 \qquad (۱۱)$$

اساساً بین SSR و مجموع مربعات کل (SST)، رابطه‌ی ۱۲ برقرار است.

$$SST=SSR+SSE \qquad (۱۲)$$

رابطه‌ی مجموع مربعات کل با مقادیر تجربی و میانگین مقادیر تجربی، عبارت است از:

$$SST=\sum_{i=1}^{n}(y_{exp}-\bar{y}_{exp})^2 \qquad (۱۳)$$

در روابط ۱۰تا ۱۳، y_{exp}، y_{pred}، $\bar{y}_{pred}$ و $\bar{y}_{exp}$ به‌ترتیب دلالت بر مقدار تجربی کمیت، مقدار پیشگوئی‌شده، میانگین مقادیر پیشگوئی‌شده و میانگین مقادیر تجربی دارند.

۳-۸. آزمون Y تصادفی یا آزمون هم‌بستگی شانسی

آزمون Y تصادفی، به عنوان یک معیار مناسب جهت تأیید عدم وجود هم‌بستگی شانسی بین متغیرهای واردشده در مدل و استحکام مدل در مدل‌سازی‌های QSAR، QSPR، QSRR و.... در نظر گرفته می‌شود [۲۱]. در این آزمون، بردار متغیر وابسته (اندیس کواتس) به طور تصادفی در هم آمیخته‌شده یا در هم ریخته‌شده و مدل‌های جدید رابطه‌ی کمی ساختار-فعالیت یا رابطه‌ی کمی

ساختار-ویژگی، با استفاده از ماتریس متغیرهای مستقل اولیه (توصیف‌کننده‌های منتخب) ساخته می‌شود. این در هم‌ریختگی، هرگونه رابطه‌ی منطقی بین توصیف‌کننده‌ها و متغیر وابسته را از بین می‌برد. بنابراین، انتظار می‌رود که با استفاده از این رویکرد پس از چندین با تکرار، مدل‌های ساخته شده دارای مقادیر کم و قابل اغماض R^2 و Q^2 باشند. در جدول ۸، برخی از پارامترهای آماری حاصل از انجام آزمون Y تصادفی پس از ۱۰ بار تکرار آورده شده‌اند.

جدول ۸. پارامترهای آماری شاخص حاصل از انجام آزمون Y تصادفی

PRESS (الف)	S_{PRESS} (ب)	SST (ج)	R^2CV (د)	PRESS/SST (ه)	REP (و)	RMSEP (ز)
3447433.0582	273.7594	1240222.9990	0.0894	2.7797	272132.0672	259.9937
4051128.4856	296.7625	512329.1367	0.0015	7.9073	319786.3311	281.8402
4300086.3960	305.7452	470887.4879	0.0238	9.1319	339438.4692	290.3712
3164133.5830	262.2699	624770.4901	0.0935	5.0645	249769.0886	249.0820
4198907.2720	302.1268	302585.7967	0.0507	13.8767	331451.6327	286.9347
3969905.2911	293.7725	1089984.7350	0.0215	3.6422	313374.7676	279.0005
4573501.1133	315.3156	624436.3183	0.0294	7.3242	361021.1689	299.4603
4171664.6610	301.1451	829740.2613	0.0009	5.0277	329301.1666	286.0024
4076430.0950	297.6878	395942.7187	0.0111	10.2955	321783.5792	282.7189
4667150.0829	318.5275	352433.5929	0.1546	13.2426	368413.5931	302.5107

الف) مجموع مربعات باقی‌مانده‌ی پیش‌بینی ب) عدم قطعیت مجموع مربعات باقی‌مانده‌ی پیش‌بینی ج) مجموع مربعات کل د) مجذور ضریب هم‌بستگی اعتبار تقاطعی ه) مجموع مربعات باقی‌مانده‌ی پیش‌بینی تقسیم بر مجموع مربعات کل و) خطای نسبی پیش‌بینی ز) جذرمیانگین مجذور خطای پیش‌بینی

۳-۹. نتایج محاسبات و مدل‌های خطی حاصل از سری داده میخک زینتی به‌روش GA-MLR

پس از انجام محاسبات با استفاده از الگوریتم توانمند GA-MLR با تغییر متوالی در پارامترهای مربوطه در دستورهای متنی در محیط نرم‌افزار متلب، مجموعه‌های شامل ۳ تا ۷ توصیف‌کننده‌ی مولکولی حاصل شد. در جداول متوالی ۹ تا ۱۶، مشخصات و جزئیات کامل این مدل‌ها آورده شده‌اند.

جدول ۹. فهرست توصیف‌کننده‌های منتخب در مدل‌های خطی سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR

Model number	Descriptor Name(s)	R^2		Q^2	
		Training set	Test set	Q^2LOO	Q^2LGO
1	-	–	–	–	–
2	-	–	–	–	–
3	LPRS; Mor18v; R3v	0.9338	0.9775	0.921	0.912
4	LPRS; MATS2e; GATS8e; Mor24e	0.9397	0.9761	0.901	0.882
5	LPRS; MATS2e; GATS8e;Mor11e; R4p+	0.9513	0.9678	0.922	0.910
6	LPRS; MATS2e;SP12;RDF060m; Mor32u;Mor22m	0.946	0.9865	0.920	0.898
7	LPRS;MATS4p; Mor18v; Mor12e; R7m+;R3v; R1e+	0.962	0.9739	0.920	0.912

جدول ۱۰. مشخصات مدل‌های منتخب سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR

No.	Symbol	Descriptor description	group Descriptor	Coefficient	Mean effect	VIF
Linear Model 1 (3 molecular descriptors)						
1	Intercept	-	-	388.7963	-	-
2	**LPRS**	log of product row sums (PRS)	topological	17.6513	**0.934**	1.446
3	Mor18v	3D-MoRSE - signal 18 / weighted by atomic van der Waals volumes	3D-MoRSE	- 143.4047	0.036	1.249
4	R3v	R autocorrelation of lag 3 / weighted by atomic van der Waals volumes	GETAWAY	47.0196	0.031	1.475
Linear Model 2 (4 molecular descriptors)						
1	Intercept	-	-	430.6207	-	-
2	**LPRS**	log of product row sums (PRS)	topological	19.0013	**1.056**	1.058
3	MATS2e	Moran autocorrelation - lag 2 / weighted by atomic Sanderson electronegativitie	2D autocorrelations	-386.9065	-0.037	1.088
4	GATS8e	Geary autocorrelation - lag 8 / weighted by atomic Sanderson electronegativities	2D autocorrelations	- 20.9688	-0.018	1.540
5	Mor24e	3D-MoRSE - signal 24 / weighted by atomic Sanderson electronegativities	3D-MoRSE	- 14.0612	**0.000**	1.550
Linear Model 3 (5 molecular descriptors)						
1	Intercept	-		473.1039	-	-
2	**LPRS**	log of product row sums (PRS)	topological	19.2497	**1.127**	1.076
3	MATS2e	Moran autocorrelation - lag 2 / weighted by atomic Sanderson electronegativitie	2D autocorrelations	- 394.9530	-0.040	1.082
4	GATS8e	Geary autocorrelation - lag 8 / weighted by atomic Sanderson electronegativities	2D autocorrelations	-25.1673	-0.023	1.305
5	Mor11e	3D-MoRSE - signal 11 / weighted by atomic Sanderson electronegativities	3D-MoRSE	- 48.3640	-0.059	1.024
6	R4p+	R maximal autocorrelation of lag 4 / weighted by atomic polarizabilities	GETAWAY	- 34.2877	-0.004	1.371
Linear Model 4 (6 molecular descriptors)						
1	Intercept	-	-	395.0259	-	-
2	**LPRS**	log of product row sums (PRS)	topological	22.0531	**1.175**	5.238
3	MATS2e	Moran autocorrelation - lag 2 / weighted by atomic Sanderson electronegativitie	2D autocorrelations	- 434.8939	-0.040	1.461
4	SP12	shape profile no. 12	Randic molecular profiles	- 14.7261	-0.057	2.372
5	RDF060m	Radial Distribution Function - 6.0 / weighted by atomic masses	RDF	- 24.0879	-0.034	3.211
6	Mor32u	3D-MoRSE - signal 32 / unweighted	3D-MoRSE	4.7085	-0.001	1.846
7	Mor22m	3D-MoRSE - signal 22 / weighted by atomic masses	3D-MoRSE	- 274.0886	-0.043	1.500
Linear Model 5 (7 molecular descriptors)						
1	Intercept	-	-	622.846	-	-
2	**LPRS**	log of product row sums (PRS)	topological	17.901	**1.291**	2.580
3	MATS4p	Moran autocorrelation - lag 4 / weighted by atomic polarizabilities	2D autocorrelations	- 81.381	0.014	1.469
4	Mor18v	3D-MoRSE - signal 18 / weighted by atomic van der Waals volumes	3D-MoRSE	- 80.698	0.028	1.918
5	Mor12e	3D-MoRSE - signal 12 / weighted by atomic Sanderson electronegativities	3D-MoRSE	19.045	-0.036	2.195
6	R7m+	R maximal autocorrelation of lag 7 / weighted by atomic masses	GETAWAY	-3248.092	-0.010	1.315
7	R3v	R autocorrelation of lag 3 / weighted by atomic van der Waals volumes	GETAWAY	- 11.856	-0.011	2.031
8	R1e+	R maximal autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities	GETAWAY	- 548.037	-0.277	1.454

جدول ۱۱. فهرست مدل‌های خطی نهائی حاصل در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR

Model number	Linear Model Form
1	**RI** = 388.7963(±47.7) + 17.6513(±0.9) × **LPRS** - 143.4047(±43.6) × **Mor18v** + 47.0196(±87.0)× **R3v**
2	**RI** = 430.6207(±34.4) + 19.0013(±0.7) × **LPRS** -386.9065(±123.7) × **MATS2e** - 20.9688 (±10.4)× **GATS8e** - 14.0612(±40.7) × **Mor24e**
3	**RI** = 473.1039(±43.0) + 19.2497(±0.7) × **LPRS** - 394.9530(±112.1) × **MATS2e** - 25.1673(±8.7) × **GATS8e** - 48.3640 (±14.8) × **Mor11e** - 34.2877(±325.2) × **R4p+**
4	**RI** = 395.0259(±43.8) + 22.0531(±1.5) × **LPRS** - 434.8939(±138.6) × **MATS2e** - 14.7261(±6.6)× **SP12** - 24.0879(±14.1) × **RDF060m** + 4.7085 (±60.5) × **Mor32u** - 274.0886(±91.4) × **Mor22m**
5	**RI** = 622.846 (±94.6) + 17.901(±0.9) × **LPRS** - 81.381(±111.2) × **MATS4p** - 80.698(±42.8) × **Mor18v** + 19.045(±22.6) × **Mor12e** -3248.092(±637.8) × **R7m+** - 11.856(±80.9) × **R3v** - 548.037(±213.3) **R1e+**

جدول ۱۲. پارامترهای آماری جامع به دست آمده برای مجموعه‌های آموزش و آزمون در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR در مدل شامل ۳ توصیف‌کننده

شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2	0.9338
	سری آزمون		0.9775
۲	سری آموزش	Adjusted R^2	0.934
۲	سری آموزش	REP (%)	5.3
	سری آزمون		2.7
۳	سری آموزش	RMSEP	66.7
	سری آزمون		35.6
٤	سری آموزش	SEP	67.4
	سری آزمون		37.7
٥	سری آموزش	RSEP	5.2
	سری آزمون		2.7
٦	سری آموزش	MAE	99.9
	سری آزمون		178.9
۷	سری آموزش	F	220.9
	سری آزمون		59.9

جدول ۱۳. پارامترهای آماری جامع به دست آمده برای مجموعه‌های آموزش و آزمون در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR در مدل شامل ٤ توصیف‌کننده

شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2	0.9338
	سری آزمون		0.9775
۲	سری آموزش	Adjusted R^2	0.934
۲	سری آموزش	REP (%)	5.0
	سری آزمون		2.7
۳	سری آموزش	RMSEP	63.7
	سری آزمون		35.3
٤	سری آموزش	SEP	64.4
	سری آزمون		37.5
٥	سری آموزش	RSEP	4.9
	سری آزمون		2.7
٦	سری آموزش	MAE	100.8
	سری آزمون		180.6
۷	سری آموزش	F	179.1
	سری آزمون		34.9

جدول ١٤. پارامترهای آماری جامع به دست آمده برای مجموعه‌های آموزش و آزمون در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR در مدل شامل ٥ توصیف‌کننده

شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2	0.9397
	سری آزمون		0.9761
۲	سری آموزش	Adjusted R^2	0.946
۲	سری آموزش	REP (%)	4.5
	سری آزمون		3.2
۳	سری آموزش	RMSEP	57.3
	سری آزمون		41.9
٤	سری آموزش	SEP	57.8
	سری آزمون		44.5
٥	سری آموزش	RSEP	4.4
	سری آزمون		3.2
٦	سری آموزش	MAE	96.5
	سری آزمون		207.7
۷	سری آموزش	F	175.7
	سری آزمون		14.4

جدول ۱۵. پارامترهای آماری جامع به دست آمده برای مجموعه‌های آموزش و آزمون در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR در مدل شامل ٦ توصیف‌کننده

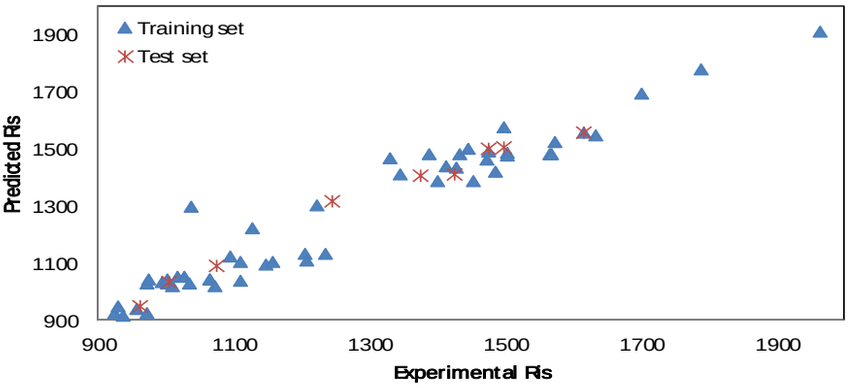
شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2	0.9513
	سری آزمون		0.9678
۲	سری آموزش	Adjusted R^2	0.939
۲	سری آموزش	REP (%)	4.8
	سری آزمون		2.4
۳	سری آموزش	RMSEP	60.3
	سری آزمون		31.0
٤	سری آموزش	SEP	60.9
	سری آزمون		32.9
٥	سری آموزش	RSEP	4.7
	سری آزمون		2.4
٦	سری آموزش	MAE	98.7
	سری آزمون		157.6
۷	سری آموزش	F	128.6
	سری آزمون		16.8

جدول ١٦. پارامترهای آماری جامع به دست آمده برای مجموعه‌های آموزش و آزمون در سری داده‌ی ترکیبات اسانس میخک زینتی به‌روش GA-MLR در مدل شامل ۷ توصیف‌کننده

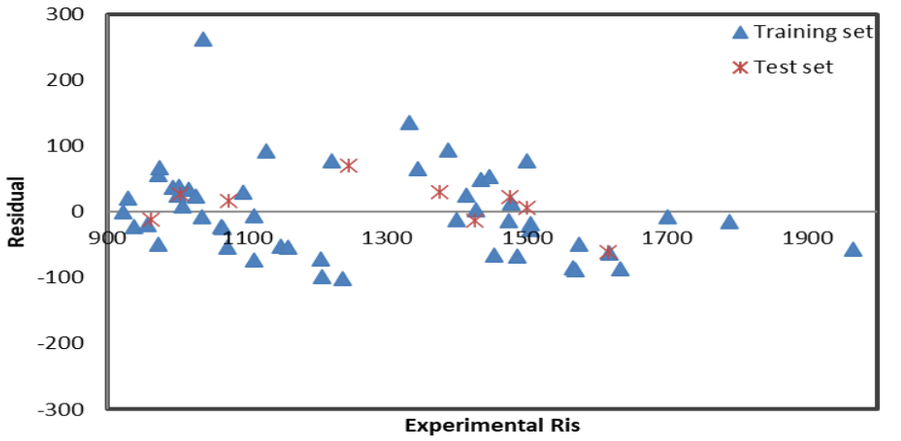
شماره	سری منتخب	پارامتر آماری	مقدار عددی
۱	سری آموزش	R^2	0.962
	سری آزمون		0.9739
۲	سری آموزش	Adjusted R^2	0.956
۲	سری آموزش	REP (%)	4.0
	سری آزمون		2.9
۳	سری آموزش	RMSEP	50.6
	سری آزمون		37.3
٤	سری آموزش	SEP	51.1
	سری آزمون		39.5
٥	سری آموزش	RSEP	3.9
	سری آزمون		2.8
٦	سری آموزش	MAE	87.2
	سری آزمون		188.2
۷	سری آموزش	F	155.6
	سری آزمون		4.8

جهت درک عمیق‌تر و مقایسه‌ی کامل‌تر مدل‌های ساخته‌شده، منحنی‌های پیش‌گوئی و باقی‌مانده برای مدل‌های خطی شامل ۳، ۴، ۵، ۶ و ۷ توصیف‌کننده‌ی مولکولی به‌ترتیب در شکل‌های ۱۰، ۱۱، ۱۲، ۱۳، ۱۴، ۱۵، ۱۶، ۱۷، ۱۸ و ۱۹ آورده شده است. نهایتاً، در جداول ۱۸، ۱۹، ۲۰، ۲۱ و ۲۲، ماتریس هم‌بستگی ۳، ۴، ۵، ۶ و ۷ توصیف‌کننده‌ی واردشده در مدل‌های خطی به‌روش GA-MLR

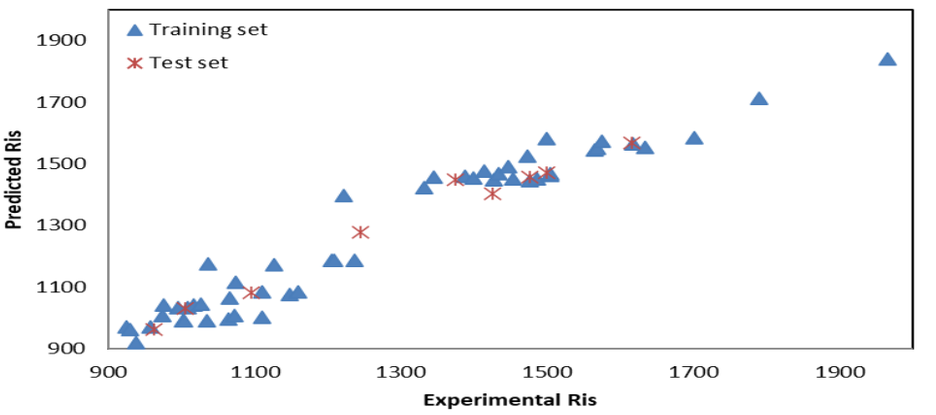
نمایش داده شده است. براساس این جداول، مقادیر ناچیز ضرائب هم‌بستگی بین هر جفت توصیف‌کننده دلالت بر رفتار مستقل آن‌ها و مناسب بودن مدل‌های ارائه‌شده دارد.



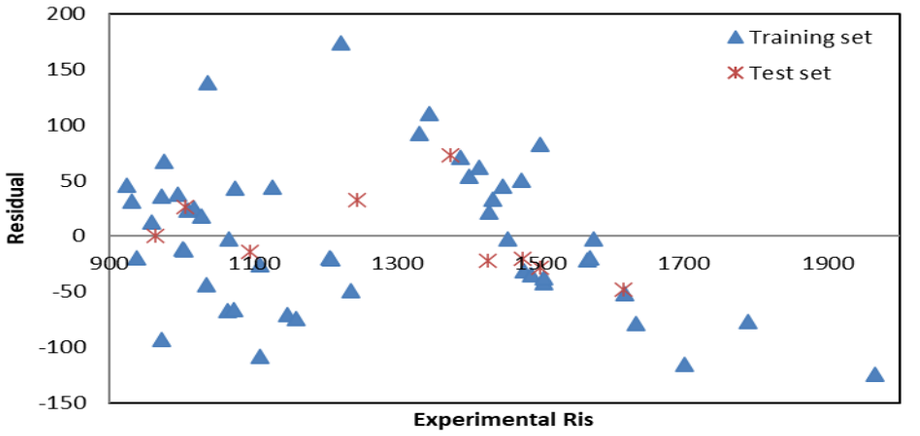
شکل ۱۰. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ۳ توصیف‌کننده



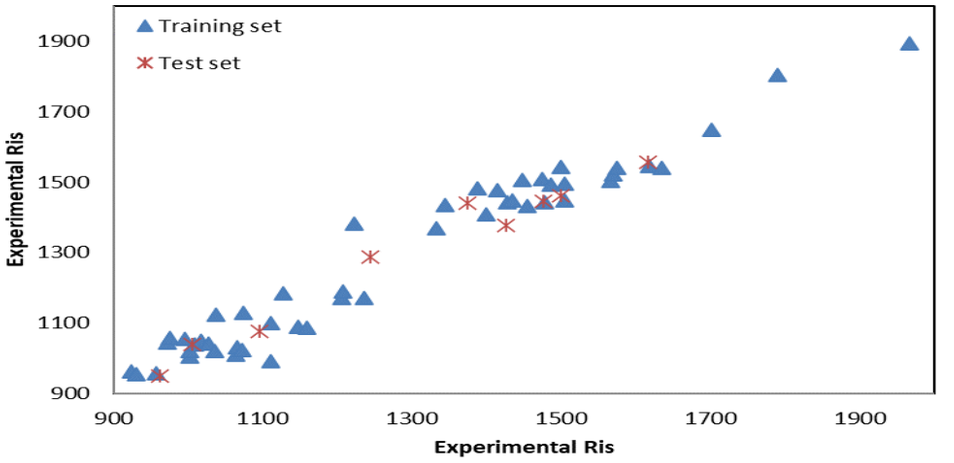
شکل ۱۱. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ۳ توصیف‌کننده



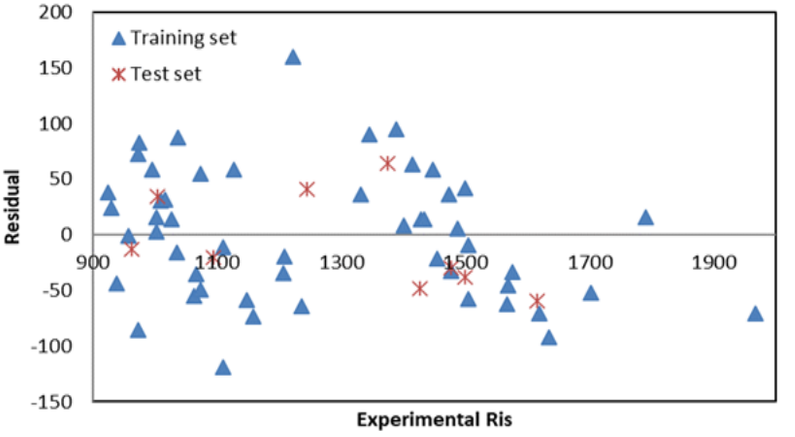
شکل ۱۲. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٤ توصیف‌کننده



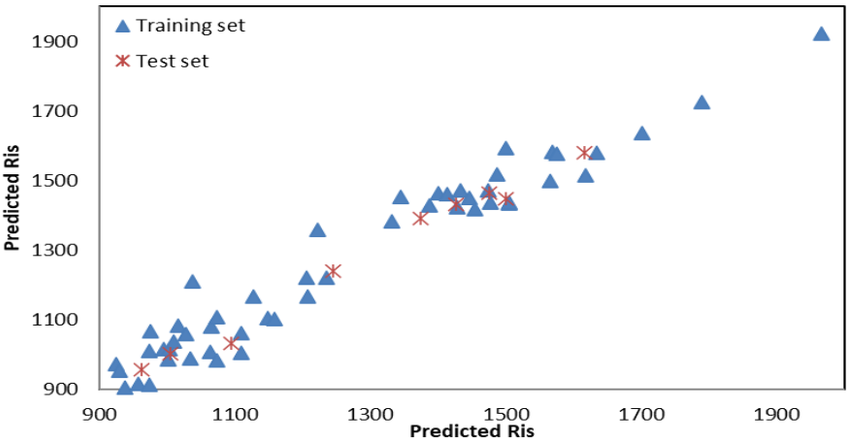
شکل ۱۳. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٤ توصیف‌کننده



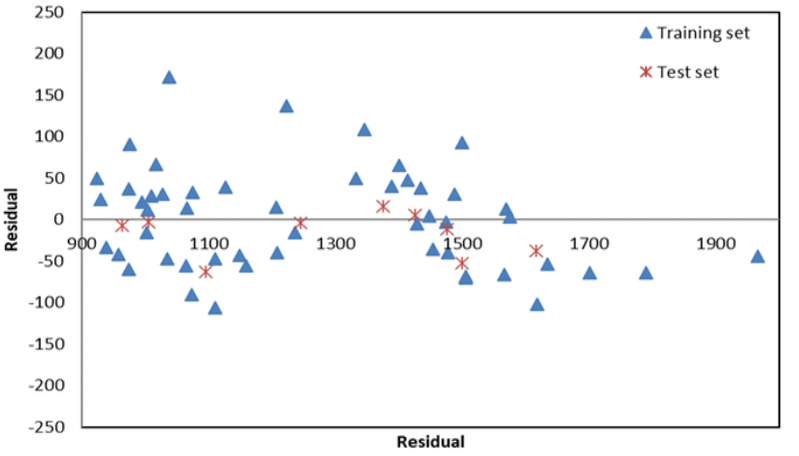
شکل ۱٤. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٥ توصیف‌کننده



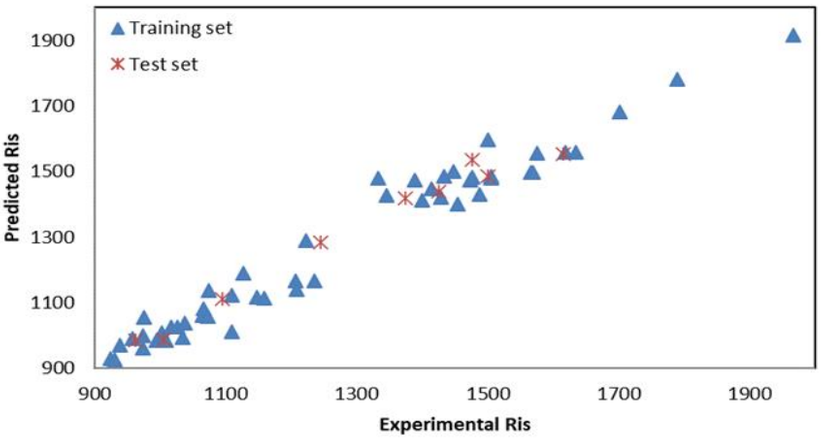
شکل ۱٥. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٥ توصیف‌کننده



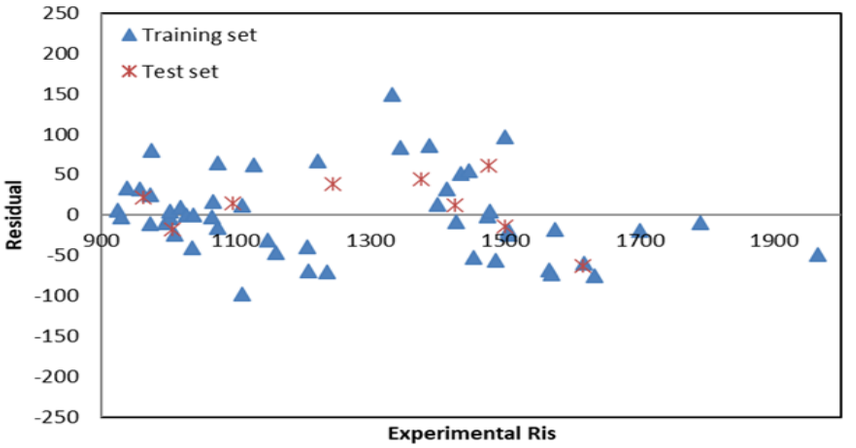
شکل ١٦. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٦ توصیف‌کننده



شکل ۱۷. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ٦ توصیف‌کننده



شکل ۱۸. مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ۷ توصیف‌کننده



شکل ۱۹. مقادیر باقی‌مانده‌ی حاصل از اختلاف مقادیر پیش‌بینی‌شده‌ی اندیس‌های بازداری ترکیبات اسانس گیاه میخک زینتی برحسب مقادیر تجربی آن به‌روش GA-MLR در مدل شامل ۷ توصیف‌کننده

جدول ۱۸. ماتریس هم‌بستگی ۳ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی ترکیبات اسانس گیاه میخک زینتی به‌روش GA-MLR

	LPRS	Mor18v	R3v
LPRS	1	-0.216	0.439
Mor18v	-0.216	1	0.256
R3v	0439	0.256	1

جدول ۱۹. ماتریس هم‌بستگی ٤ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی ترکیبات اسانس گیاه میخک زینتی به‌روش GA-MLR

	LPRS	MATS2e	GATS8e	Mor24e
LPRS	1	0.216	0.035	-0.014
MATS2e	0.216	1	-0.055	0.103
GATS8e	0.035	-0.055	1	0.577
Mor24e	-0.014	0.103	0.577	1

جدول ۲۰. ماتریس هم‌بستگی ٥ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی ترکیبات اسانس گیاه میخک زینتی به‌روش GA-MLR

	LPRS	MATS2e	GATS8e	Mor11e	R4p+
LPRS	1	0.216	0.035	-0.101	0.124
MATS2e	0.216	1	-0.055	-0.020	0.194
GATS8e	0.035	-0.055	1	0.061	-0.474
Mor11e	-0.101	-0.020	0.061	1	-0.123
R4p+	0.124	0.194	-0.474	-0.123	1

جدول ۲۱. ماتریس هم‌بستگی ٦ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی ترکیبات اسانس گیاه میخک زینتی به‌روش GA-MLR

	LPRS	MATS2e	SP12	RDF060m	Mor32u	Mor22m
LPRS	1	0.216	0.301	0.794	-0.469	0.273
MATS2e	0.126	1	-0.398	0.305	-0.258	0.234
SP12	0.301	-0.398	1	0.184	0.230	-0.365
RDF060m	0.794	0.305	0.184	1	-0.245	0.205
Mor32u	-0.469	-0.258	0.230	-0.245	1	-0.193
Mor22m	0.273	0.234	-0.365	0.205	-0.193	1

جدول ۲۲. ماتریس هم‌بستگی ۷ توصیف‌کننده‌ی وارد شده در مدل منتخب سری داده‌ی ترکیبات اسانس گیاه میخک زینتی به‌روش GA-MLR

	LPRS	MATS4p	Mor18v	Mor12e	R7m+	R3v	R1e+
LPRS	1	0.236	-0.216	-0.451	0.054	0.439	-0.518
MATS4p	0.263	1	-0.125	-0.142	-0.145	-0.256	-0.226
Mor18v	-0.216	-0.125	1	-0.366	0.238	0.256	0.026
Mor12e	-0.451	-0.142	-0.366	1	0.143	-0.506	0.334
R7m+	0.054	-0145	0.238	0.143	1	0.029	-0.083
R3v	0.439	-0.256	0.256	-0.506	0.029	1	-0.312
R1e+	-0.518	-0.226	-0.026	0.334	-0.083	-0.312	1

٤. نتیجه‌گیری

در این مقاله، به صورت جامع به مدلسازی‌های خطی انجام شده روی ترکیبات طبیعی تشکیل‌دهنده روغن اسانسی استحصال شده از گیاه پرداخته شده است. این مدل‌سازی‌ها، با استفاده از روش‌های SW-MLR و GA-MLR انجام پذیرفته‌اند. جهت نیل به روابط خطی هدفمند و مستحکم پس از انجام مراحلی مانند محاسبه و انتخاب توصیف‌کننده‌ها و تجزیه و تحلیل آماری آن‌ها، الگوی پراکنش آنالیز جزء اصلی (PCA) و تقسیم سری داده به مجموعه‌های آموزشی و آزمون به صورت منطقی انجام شد. نتایج حاصل از انجام این نوع آنالیز مشخص نمود که ترکیبات طبیعی با شماره‌های ۳۰، ۵۸، ۵۹ و ۶۰ در مجموعه‌ی داده‌های اصلی با فاصله محسوس نسبت به خوشه‌ی دربرگیرنده‌ی سایر ترکیبات طبیعی شناسایی شده قرار گرفته‌اند. رگرسیون خطی چندگانه مرحله‌ای منجر به مجموعه‌ای شامل ۱۴ مدل رگرسیون در نرم‌افزار SPSS شد. همچنین، ترکیبات شماره‌ی ۱۵ و ۴۵ در سری آموزش دارای مقادیر لورج بیشتر از مقدار هشدار (h^*) و می‌توانند جزء ترکیبات انحرافی باشند. خوشبختانه، در دو مورد اخیر مقادیر اطلاعات عددی پیشگوئی شده مناسب و قابل قبول هستند. لذا، این دو ترکیب لورج خوب خوانده شده و از حوزه‌ی کاربرد و سری داده‌ی اولیه حذف نمی‌شوند. نکته حائز اهمیت اینکه در کلیه مدل‌های خطی ارائه شده، هم‌گرائی‌های مطلوب بین مقادیر تجربی و مقادیر پیشگوئی شده حاصل شدند.

از طرف دیگر، در نمودارهای باقیمانده‌ی ترسیم شده مربوط به تفاضل مقادیر حقیقی و مقادیر پیشگوئی شده به توسط مدل‌های خطی مربوطه، پراکندگی طبیعی نسبتاً یکسان نقاط را حول مقادیر صفر را می‌توان ناشی از عدم وجود خطای معین در روش و اعتبار بالای مدل دانست. بعلاوه، پس از انجام روش اعتبارسنجی تقاطعی (CV)، مقادیر عددی محاسبه شده‌ی ضرایب اندازه‌گیری (یا مجذور ضریب هم‌بستگی بین مقادیر تجربی و مفادیر پیش‌گوئی‌شده) نشانگر پایداری، صحت و قدرت پیشگوئی بالای مدل‌های خطی مربوطه هستند.

در فاز دوم این مطالعه‌ی نظری با استفاده از روش انتخاب متغیر توانمند الگوریتم ژنتیک مبتنی بر نظریه‌ی تکامل طبیعی داروین همراه با رگرسیون خطی چندگانه (GA-MLR)، با تغییر متوالی در پارامترهای مربوطه در دستورهای متنی در محیط نرم‌افزار متلب، مجموعه‌های شامل ۳ تا ۷ توصیف‌کننده‌ی مولکولی حاصل شدند که هر یک به تنهایی از استحکام و قدرت پیشگویی قابل ملاحظه

ای جهت پیش‌بینی مقادیر عددی شاخص‌های بازداری کواتس ترکیبات طبیعی برخوردارند. به عنوان پیشنهاداتی تکمیلی، برای انجام مدل‌سازی‌های اینچنینی در آینده می‌توان از روش‌های غیرخطی مانند شبکه‌های عصبی مصنوعی و همچنین از الگوریتم‌های دیگر مانند الگوریتم کرم شب‌تاب، الگوریتم مورچه و یا جنگل تصادفی نیز می‌توان بهره‌برداری و توانمندی مدل‌های مختلف را با در نظر گرفتن شاخص‌های آماری مرتبط با هر نوع مدل‌سازی تعیین و با یکدیگر مقایسه نمود.

۵. مراجع

[1] Zargari, A. (1991). Medicinal Plants. Tehran University Publication, Tehran.

[2] Mozaffarian, V. (1996). A Dictionary of Iranian Plant Names. Farhang Moaser Press, Iran.

[3] Ferreira, N. J., de Sousa, I. G. M., Luis, T. C., Currais, A. J. M., Figueiredo, A. C., Costa, M. M., Lima, A. S. B., Santos, P. A. G., Barroso, J. G., Pedro, L. G. & Scheffer, J. J. C. (2007). *Pittosporum undulatum* Vent. grown in Portugal: Secretory structures, seasonal variation and enantiomeric composition of its essential oil. *Flavour and Fragrance Journal*, *22*, 1-9.

[4] Wold, S., Esbensen, K. & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*, 37-52.

[5] Heberger, K. (1999). Evaluation of polarity indicators and stationary phases by principal component analysis in gas-liquid chromatography *Chemometrics and Intelligent Laboratory Systems*, *47*, 41-49.

[6] Heberger, K., Milczewska, K. & Voelkel, A. (2005). Principal component analysis of polymer-solvent and filler-solvent interactions by inverse gas chromatography. *Colloids and Surfaces a-Physicochemical and Engineering Aspects*, *260*, 29-37.

[7] Heberger, K. & Gorgeny, M. (1999). Principal component analysis of Kovats indices for carbonyl compounds in capillary gas chromatography. *Journal of Chromatography A*, *845*, 21-31.

[8] OECD (2007). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. Organisation for Economic Co-Operation and Development, Paris.

[9] Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W. D., Veith, G. & Yang, C. H. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. *Atla-Alternatives to Laboratory Animals*, *33*, 155-173.

[10] Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M. & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives*, *111*, 1361-1375.

[11] Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Atla-Alternatives to Laboratory Animals*, *33*, 445-459.

[12] Agrawal, V. K. & Khadikar, P. V. (2001). QSAR prediction of toxicity of nitrobenzenes. *Bioorganic & Medicinal Chemistry*, *9*, 3035-3040.

[13] Pourbasheer, E., Riahi, S., Ganjali, M. R. & Norouzi, P. (2010). Quantitative structure-activity relationship (QSAR) study of interleukin-1 receptor associated kinase 4 (IRAK-4) inhibitor activity by the genetic algorithm and multiple linear regression (GA-MLR) method. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *25*, 844-853.

[14] Beheshti, A., Pourbasheer, E., Nekoei, M. & Vahdani, S. (2012). QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm–multiple linear regressions. *Journal of Saudi Chemical Society*, *20*, 282-290.
[15] Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science*, *26*, 694-701.
[16] Noorizadeh, H. & Farmany, A. (2012). Quantitative structure-retention relationship for retention behavior of organic pollutants in textile wastewaters and landfill leachate in LC-APCI-MS. *Environmental Science and Pollution Research*, *19*, 1252-1259.
[17] Noorizadeh, H., Noorizadeh, M. & Farmany, A. (2012). Advanced QSRR models of toxicological screening of basic drugs in whole blood by UPLC-TOF-MS. *Medicinal Chemistry Research*, *21*, 4357-4368.
[18] Jalali-Heravi, M. & Asadollahi-Baboli, M. (2009). Quantitative structure-activity relationship study of serotonin (5-HT7) receptor inhibitors using modified ant colony algorithm and adaptive neuro-fuzzy interference system (ANFIS). *European Journal of Medicinal Chemistry*, *44*, 1463-1470.
[19] Agirre-Basurko, E., Ibarra-Berastegi, G. & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. *Environmental Modelling & Software*, *21*, 430-446.
[20] Larose, D. T. (2006). Data Mining Methods and Models. John Wiley & Sons, Inc Publication, Hoboken, New Jersey.
[21] Tropsha, A., Gramatica, P. & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, *22*, 69-77.

Steps, calculations and results of studies of theoretical predictions of quantitative structure retention relationship (QSRR) of *Pittosporum undulatum* essential oil

Majid Mohammadhosseini[1]*, Mehdi Nekoei[1]

[1]*Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran*

Submited: 14 April 2023, Revised: 13 July 2023, Accepted: 21 July 2023

Abstract

In this article, a detailed description of the linear models capable of predicting the inhibition indices of a large group of natural compounds identified in the essential oil of *Pittosporum undulatum*, as one of the medicinal plants, has been discussed. In this regard, the work is based on quantitative structure retention relationship (QSRR), which is of prime importance in scientific resources to establish a logical and meaningful relationship between the Kovats index as a dependent variable and a group of molecular descriptors as independent variables. In this regard, after drawing the structure of the natural compounds using the Hypercam software and optimizing their molecular structures, Dragon software was used to extract the relevant molecular descriptors. In the next step, after removing irrelevant and redundant descriptors, a group of important and effective descriptors were identified and their linear relationship with the Kovats retention index was discussed and investigated using stepwise multiple linear regression method as well as another variable selection method based upon genetic algorithm feature selection approach. The obtained results indicate the high capability of the presented models to predict the Kovats index of a wide group of natural compounds.

Keywords: *Quantitative structure retent relationship (QSRR), multiple linear regression, molecular descriptors, genetic algorithm, Kovats retention index, Pittosporum undulatum.*