
بسم الله الرحمن الرحيم
الحمد لله رب العالمين الرحمن الرحيم مالك يوم الدين
اهدنا الصراط المستقيم صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين

دانشکده شیمی

پایان‌نامه کارشناسی ارشد شیمی تجزیه

کاربرد روش‌های کمومتریکس جهت پیش‌بینی اندیس بازداری برخی ترکیبات آلی

نگارنده:

الهه سلامتی

اساتید راهنما:

دکتر ناصر گودرزی

دکتر داوود شاهسونی

استاد مشاور:

دکتر منصور عرب چم جنگلی

آذر ماه ۱۳۹۷

شماره: ۱۸۷۶ د/ش
تاریخ: ۹۷/۱۰/۳۰

باسمه‌تعالی

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم الهه سلامتی با شماره دانشجویی ۹۴۰۹۴۳۴ رشته شیمی گرایش تجزیه تحت عنوان کاربرد روش‌های کمومتریکس جهت پیش‌بینی اندیس بازداری برخی ترکیبات آلی که در تاریخ ۱۳۹۷/۰۹/۲۷ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه: خیلی خوب) ☑ **مردود** ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر ناصر گودرزی	دانشیار	
۲- استادراهنمای دوم	دکتر داوود شاهسونی	دانشیار	
۳- استاد مشاور	دکتر منصور عرب چم جنگلی	استاد	
۴- نماینده تحصیلات تکمیلی	دکتر علی کیوانلو	دانشیار	
۵- استاد ممتحن اول	دکتر فاطمه مصدرالامور	استادیار	
۶ـ استاد ممتحن دوم	دکتر قدمعلی باقریان	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر مهدی میرزایی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به

همسرم

به‌پاس تعبیر عظیم و انسانی‌اش از کلمه ایثار و از خودگذشتگی، به‌پاس عاطفه سرشار و گرمای امیدبخش وجودش که در این سردترین روزگاران بهترین پشتیبان است، به‌پاس قلب بزرگش که فریادرس است و سرگردانی و ترس در پناهش به شجاعت می‌گراید و به‌پاس محبت‌های بی‌دریغش که هرگز فروکش نمی‌کند.

مادرم

تو روشنایی‌بخش تاریکی جان هستی و ظلمت اندیشه را نور می‌بخشی. چگونه سپاس گویم مهربانی و لطف تو را که سرشار از عشق و یقین است. چگونه سپاس گویم تأثیر عشق آموزی تو را که چراغ روشن هدایت را بر کلبه محقر وجودم فروزان ساخته‌ای. آری در مقابل این همه عظمت و شکوه تو مرا نه توان سپاس است و نه کلام وصف

تشکر و قدردانی

تشکر و سپاس بی‌پایان مخصوص خداوندی است که بشر را آفریده و به او قدرت اندیشیدن داده و توانایی‌های بالقوه را در وجود انسان قرار داده و او را امر به تلاش و کوشش نموده و راهنمایانی را برای هدایت بشر فرستاده است. پس از ارادت خاضعانه به درگاه خداوند بر همتا لازم است از اساتید ارجمند جناب دکتر ناصر گودرزی و جناب دکتر داوود شاهسونی و جناب دکتر منصور عرب چم جنگلی به خاطر سعه‌صدر و رهنمودهای دلسوزانه شان که در تهیه این تحقیق مرا مورد لطف خود قرار داده‌اند و راهنمایی‌های لازم را نموده‌اند تشکر و قدردانی نموده و موفقیت همگان را از درگاه احدیت خواهانم، و همچنین از اساتید داور جناب آقای دکتر باقریان و سرکار خانم مصدرالامور به خاطر مطالعه پایان‌نامه و شرکت در جلسه دفاع قدردانی می‌نمایم

همسر عزیزم که مشکلات زندگی را برایم تسهیل نمود و پیمودن این راه بدون همراهی او ممکن نمی‌شد، سپاسگذارم

واز سرکار خانم مظفری که مرا در انجام این پروژه یاری نمودند کمال تشکر را دارم

از درگاه لایزال احدیت، توفیق روز افزون آن بزرگواران را مسئلت داشته و همواره سعادت و بهروزیشان را آرزومندم

تعهدنامه

اینجانب الهه سلامتی دانشجوی دوره کارشناسی ارشد رشته شیمی تجزیه دانشکده شیمی دانشگاه صنعتی شاهرود نویسنده پایان‌نامه کاربرد روش‌های کمومتریکس جهت پیش‌بینی اندیس بازداری برخی ترکیبات آلی. تحت راهنمائی دکتر ناصر گودرزی و دکتر داوود شاهسونی متعهد می‌شوم .

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج بانام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا چینی‌جاهای آن‌ها) استفاده‌شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده‌شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

*متن این صفحه نیز باید در ابتدای نسخه های چاپ شده پایان نامه وجود داشته باشد .

چکیده

به‌طورکلی میزان گوگرد موجود در نفت خام، بسته به نوع منبع آن بین ۲۵ ۰/۰ تـا بعضـاً حـدود ۷/۸۹ درصد وزنی متغیر است. گوگرد به دو صورت معدنی و آلی در نفت خـام وجـود دارد. عـلاوه بـر گوگرد عنصری بیش از ۲۰۰ ترکیب آلی گوگرددار در نفت شناسایی‌شده، که سولفیدها و تیول‌هـا یـا مرکاپتان‌ها و تیوفن‌ها را شامل می‌شود، تیوفن‌ها و دی بنزوتیوفن‌ها مولکول‌های بسـیار پیچیـده‌تـری دارند. این ترکیبات به عنوان آلوده‌کننده‌های پایدار شناخته می‌شـوند عـلاوه بـر ایـن، ایـن ترکیبـات سرطان‌زا هستند. روش‌های مختلفـی ماننـد کروماتوگرافی گـازی بـرای جداسـازی و شناسـایی ایـن ترکیبات پیشنهادشده است که اندیس بازداری را با استاندارها مقایسه کند. بنـابراین انـدیس بـازداری یک حل‌شونده، پارامتر مهم می‌باشد که مـی‌توانـد بـرای شناسـایی ترکیبـات هتروسـیکل آروماتیـک سولفوردار به کار رود. اما از طرف دیگر این روش‌ها پرهزینه و وقت‌گیر هستند. اما از این رو، برای غلبه بر این مشکلات، می‌توان از روش‌های کمومتریکس برای پیش بینـی انـدیس بـازداری ایـن ترکیبـات استفاده نمود. با استفاده از این روش‌های کمومتریکس می‌توان اندیس بـازداری ترکیبـات PASHs را پیش‌بینی نمود که با استفاده از روش‌هـای جنگـل‌هـای تصـادفی(RF) و مـارس (MARS) و شـبکه عصبی مصنوعی SR-ANN و RF-ANN اقدام به پیش‌بینی اندیس بازداری این ترکیبات شده است. نتایج حاصله نشان می‌دهد که استفاده از روش آماری ذکرشده توانایی خـوبی در پـیش‌بینـی انـدیس بازداری تجربی ایـن ترکیبـات دارنـد کـه R^2 حاصـل از روش RF و MARS و SR-ANN و RF-ANN به ترتیب ۰/۹۴۶ و ۰/۹۸۴ و ۰/۹۵۸ و ۰/۹۳۲می‌باشد.

کلمات کلیدی:کمومتریکس، جنگل تصادفی، MARS، شبکه عصبی مصنوعی اندیس بازداری، ترکیبات هتروسیکل آروماتیک گوگرددار

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل اول

مقدمه

۱-۱-اندیس بازداری[1]

اندیس بازداری (RI) در سال ۱۹۵۸ برای اولین بار به‌عنوان یک پارامتر برای شناسایی ترکیبات شیمیایی با استفاده از کروماتوگرافی معرفی شد. تعیین این اندیس برای هر ترکیب، از کروماتوگرام مربوط به مخلوط آن جسم و دو پارافین نرمال که زمان بازداری ترکیب موردنظر بین زمان بازداری آن دو باشد، امکان‌پذیر است. در این مورد پارافین‌های نرمال به‌عنوان استاندارد در نظر گرفته می‌شوند. طبق قرارداد مقدار این اندیس برای این پارافین‌ها صد برابر تعداد کربن‌های آن‌ها در نظر گرفته می شوند که مقداری ثابت بوده و مستقل از نوع فاز ساکن، درجه حرارت و سایر شرایط می‌باشد. درحالی‌که برای سایر ترکیبات با تغییر شرایط ستون، مقدار این اندیس تغییر می‌کند. اندیس بازداری ماده مورد تجزیه با استفاده از معادله (۱-۱) محاسبه می‌گردد که در این رابطه $\grave{t}_{Rn}$ ،$\grave{t}_{Ra}$،$\grave{t}_{Rn+1}$ به ترتیب عبارت‌اند از زمان‌های بازداری هیدروکربن سنگین‌تر بعد از مجهول، زمان بازداری ترکیب مجهول، زمان بازداری ترکیب سبک‌تر از مجهول و n تعداد کربن‌های نرمال می‌باشد[۱].

$$RI = n \times 100 + 100 \frac{log\grave{t}_{Ra} - log\ \grave{t}_{Rn}}{log\grave{t}_{Rn+1} - log\ \grave{t}_{Rn}} \qquad (۱\text{-}۱)$$

۱-۲-معرفی ترکیبات گوگرددار و شناسایی و جداسازی آن‌ها

پس از غذا سوخت مهم‌ترین منبع انرژی بشر است به‌طوری‌که ۸۵ درصد انرژی از سوخت‌های فسیلی، ۸ درصد از انرژی هسته‌ای و ۷ درصد از منابع دیگر که به‌طورکلی آب، برق و... است، تأمین می‌شود. زغال‌سنگ، نفت و گاز سه سوخت اصلی هستند که نفت ۴۰ درصد، زغال‌سنگ حدود ۲۴ درصد و گاز طبیعی حدود ۱۲ درصد از مصرف کل جهان را دربرمی‌گیرد[۲]. بر این اساس، نفت مهم‌ترین این سوخت‌هاست.

[1] Retention index

بسیاری از ترکیبات در نفت خام شامل عناصری غیر از هیدروژن و کربن نیز می‌باشند که گاهی درون ساختمان حلقه و یا در گروه‌های استخلافی متصل شده به ساختمان هیدروکربن مشاهده می‌شوند. ترکیبات آلی گوگرددار در نفت خام مولکول‌های هتروسیکل هستند، در مقابل عمده مولکول‌های اکسیژن‌دار به‌صورت کربوکسیلیک اسیدها هستند. اسیدهای آلیفاتیک سیرشده (اسیدهای نفتیک)، تیوفن آروماتیک‌ها و به مقدار ناچیزی از فنول‌ها و فوران‌ها نیز ممکن است موجود باشد. مولکول‌های بسیار بزرگ و سنگین رزین‌ها و آسفالت‌ها که مخلوطی از ساختمان آروماتیک و هتروسیکلیک هستند نیز وجود دارند.

به‌طورکلی پس از کربن و هیدروژن گوگرد سومین عنصر فراوان در این نوع سوخت‌هاست[۳].

جدول(۱-۱) - میزان عناصر موجود در یک نمونه نفت خام

عنصر تشکیل‌دهنده	درصد
کربن	۸۳-۸۷
گوگرد	۰-۳
اکسیژن	۰-۰/۵
هیدروژن	۱۱-۱۴
نیتروژن	۰/۱
عناصر فلزی	۰-۰/۲

استفاده از سوخت‌های فسیلی به‌عنوان منبع انرژی موجب آلودگی محیطی با ترکیباتی با منشأ غیر زیستی شده است. زیرا این سوخت‌های فسیلی محتوای گوگرد هستند که سوختن مستقیم آن‌ها موجب آزادسازی مقدار زیادی از اکسید گوگرد به اتمسفر می‌شود[۴]. اکسید گوگرد تولیدی مسئول

کیفیت بدهوا، باران اسیدی و غیرفعال شدن کاتالیست‌های شیمیایی و کاهش لایه اوزون است علاوه بر این اکسید گوگرد موجب به خطر افتادن سلامتی انسان می‌شود[۵]. به گونه‌ای که مشکلاتی از قبیل سرطان دستگاه تنفسی، بیماری‌های قلبی را به همراه دارد[۶]. آنالیز این ترکیبات در ماتریکس‌های مختلف، کار دشواری است، زیرا اکثر آن‌ها با غلظت پائینی در بافت‌های پیچیده وجود دارند که سایر گونه‌ها با غلظتی چند برابر، باعث ایجاد مزاحمت در اندازه‌گیری این ترکیبات می‌شود[۹]. همچنین، آنالیز این ترکیبات در هوا پیچیده است زیرا اکسنده‌های اتمسفری مقدار قابل اندازه‌گیری این ترکیبات را کاهش می‌دهند. چندین روش از جمله کروماتوگرافی گازی، اسپکتروفتومتری، پلاروگرافی، کولومتری، پتانسیومتری، و.....برای اندازه‌گیری این ترکیبات وجود دارد که از بین این روش‌ها کروماتوگرافی گازی با دتکتور انتخابی گوگرد، به سایر روش‌ها ترجیح داده می‌شود زیرا انتخاب پذیری و حد تشخیص بالایی دارد. اما استفاده از این تکنیک‌ها با مشکلات زیادی همراه است که در زیر به مهمترین آن‌ها اشاره شده است:

۱-استانداردهای بسیاری از این ترکیبات در دسترس نیست.

۲-امکان جذب و از بین رفتن گونه‌ها در سیستم کروماتوگرافی وجود دارد.

۳-در بافت پیچیده، به یک دتکتور حساس و انتخاب پذیر نیاز است.

به دلیل وجود این مشکلات، یک روش تئوری جایگزین خوبی برای روش‌های تجربی است. مطالعه ارتباط کمی ساختار-خاصیت(QSPR)، یک روش توانمند برای پیش بینی اندیس بازداری است.

۱-۳-معرفی ترکیبات هتروسیکل آروماتیک سولفوردار چند حلقه‌ای[1]

تیوفن[2] با فرمول شیمیایی C_4H_4S یک ترکیب شیمیایی است که شکل ظاهری آن، مایع بی‌رنگ است. تیوفن یک ترکیب هتروسیکل آروماتیک است که شامل چهار اتم کربن و یک اتم گوگرد می‌باشد. آنالوگ‌های تیوفن شامل فوران و پیرول می‌باشند که در آن‌ها اتم S جایگزین اتم‌های O و NHشده است. تیوفن به وسیلهٔ ویکتور میر[3] در سال ۱۸۸۳ به‌عنوان یک جز باقی‌مانده در بنزن کشف شد. بعضی از انواع مهم تیوفن‌ها شامل بنزو تیوفن و دی بنزوتیوفن می‌باشد که به ترتیب دارای یک و دو گروه بنزنی هستند. تیوفن در دمای محیط مایع بی‌رنگ است و بویی مشابه بنزن دارد. تیوفن به‌عنوان ترکیبی آروماتیک شناخته‌شده است ولی محاسبات تئوری نشان داده که میزان آروماتیک بودن آن از بنزن کمتر است. سهم جفت الکترون گوگرد در اوربیتال مشهود می‌باشد و به خاطر همین خواص آروماتیک تیوفن است که این ترکیب خواصی متفاوت نسبت به تیو اتر از خود نشان می‌دهد. به‌عنوان‌مثال این ترکیب توانایی آلکیل دار شدن توسط متیل یدید را ندارد. همچنین اتم گوگرد در این ترکیب واکنش ناپذیر است. واکنش‌پذیری بالای تیوفن در برابر سولفوناسیون اساس جداسازی بنزن از تیوفن را تشکیل می‌دهد. (ازآنجایی‌که تفاوت نقطه‌جوش بنزن و تیوفن در دمای محیط حدود ۴ درجه سانتی‌گراد می‌باشد، جداسازی آن‌ها از طریق تقطیر مشکل می‌باشد). بنابراین اضافه کردن اسیدسولفوریک به مخلوط بنزن و تیوفن در فرآیند سولفوناسیون تیوفن باعث تشکیل ترکیب محلول در آب تیوفن سولفونیک اسید می‌شود. در بعضی موارد حلقه بنزنی توسط تیوفن جایگزین می‌شود بی‌آنکه در خواص ترکیب تغییر قابل‌ملاحظه‌ای روی دهد. گوگردزدایی از این ترکیبات توسط رانی نیکل[4] باعث ایجاد مشتقاتی از بوتان با استخلاف در موقعیت های ۱ و ۴ می‌شود. پلیمرهای حاصل از پیوند شدن تیوفن در موقعیت‌های ۱ و ۵ پلی تیوفن نامیده می‌شوند. پلی تیوفن‌ها در اثر اکسایش

[1] Polycyclic aromatic sulfur heterocycles
[2] Thiophen
[3] Victor Meyer
[4] Raney Nickel

جزئی دارای خاصیت رسانایی می‌گردند. با در نظر گرفتن پایداری بالای تیوفن‌ها، این ترکیبات از بسیاری از ترکیبات حاوی گوگرد و وهیدروکربن (به‌خصوص هیدروکربن‌های غیراشباع مانند استیلن) به وجود می‌آیند. درروش سنتز کلاسیک این ترکیبات از واکنش ۱،۴-دی‌کتون با معرف‌های سولفور کننده مانند P_4S_{10} تهیه می‌شوند. تیوفن‌های خاص می‌توانند از طریق واکنش تراکمی استرها در حضور گوگرد عنصری تهیه شوند. تیوفن و مشتقات آن معمولاً به همراه مواد نفتی در غلظت‌های یک تا سه درصد یافت می‌شوند. انواع مایعات تیوفنی در صنعت معمولاً توسط فرآیند هیدرو دسولفوریزاسیون جداسازی می‌گردد. در این فرآیند مایع یا گاز از روی بستر کاتالیزوری شامل مولیبدونیم دی سولفید و تحت‌فشار هیدروژن عبور داده‌شده و ترکیبات گوگردی ازجمله تیوفن‌ها جداسازی می‌شوند. بسیاری از این ترکیبات اثرات مخربی روی ساختار و اعمال ارگانیسم‌های زنده دارند. آن‌ها می‌توانند از طریق تنفس وارد بدن شده و صدمات زیادی به شش‌ها وارد نمایند آب رودخانه ها و دریاچه‌ها موجود در نزدیکی مکان‌هایی که زباله و زوائد کارخانه‌ها وجود دارند نیز آلوده به این ترکیبات هستند. خوردن آب و غذای آلوده باعث وارد شدن آن‌ها به خون شده که برای کبد و کلیه مضر هستند و تماس با خاک آلوده، می‌تواند این ترکیبات را جذب بدن نموده و باعث ایجاد سرطان‌های پوستی گردد[۱۰].

۱-۴-ضرورت انجام تحقیق

در سال‌های اخیر پیشرفت‌های زیادی در جهت آسان‌تر شدن، دقیق‌تر شدن و سازگار بودن بیشتر روش‌های آزمایشگاهی با محیط زیست صورت گرفته است. با این حال این روش‌ها هنوز محدودیت هایی مانند عدم توانایی در اندازه‌گیری همه ترکیبات، هزینه بالا و در دسترس نبودن امکانات مورد نظر دارند. استفاده از روش‌های نظری می‌تواند در رفع این محدودیت‌ها مفید واقع شود. به‌کارگیری این روش‌ها می‌تواند علاوه بر پیش‌بینی خواص مورد نظر، به هدفمند و جهت‌دارتر ساختن آزمایش

های تجربی و همینطور توضیح پارامترهای موثر بر نتایج این آزمایش‌ها و مکانیسم‌های درگیر کمک کند.

در این تحقیق سعی شده با به‌کارگیری روش‌های QSPR مدلی برای پیش‌بینی اندیس بازداری ترکیبات هتروسیکل آروماتیک سولفوردار چند حلقه‌ای(PASHs) طراحی شود. هنگامی که رابطه معتبری بدست آمد، امکان پیش‌بینی این خواص برای ترکیباتی با ساختار مشابه وجود خواهد داشت. در تحقیق حاضر تلاش شده تا مقایسه‌ای بین پاسخ سری داده‌ها به مدل خطی و غیر خطی نیز صورت پذیرد.

۱-۵-مروری بر کارهای انجام‌شده

در سال ۲۰۰۳ میلر[۱] و برانو[۲]، اندیس بازداری کواتس[۳] برای ترکیبات گوگرددار را محاسبه نموده و آن‌ها را با مقادیر به‌دست‌آمده از روش تجربی، مقایسه کردند. نتایج نشان داد مقادیر پیش بینی شده با مقادیر بدست آمده با پیش‌بینی لگاریتمی مرسوم موافق بوده‌اند.[۱۱].

در سال ۲۰۰۴ دیو[۴] و همکارانش، مدل رگرسیون خطی چندگانه[۵] را برای پیش‌بینی زمان و اندیس بازداری ترکیبات گوگرددار به کاربردند. نتایج به‌دست‌آمده توافق خوبی با مقادیر تجربی زمان و اندیس بازداری نشان می‌دهد[۱۲].

در سال ۲۰۰۵ کن[۶] و گروهش مطالعات کمومتریکس[۷] بروی ۸۰ ترکیب PASHs انجام داده‌اند و از روش شبکه عصبی مصنوعی[۸] برای بهینه‌سازی تعداد توصیف‌کننده ورودی[۹] استفاده کردند،

[1] miller
[2] Bruno
[3] kovats
[4] Du
[5] Multiple linear regression(MLR)
[6] Can
[7] Chemometrics
[8] Artificial neural network
[9] Input descriptors

همچنین میزان مشارکت توصیف کننده‌های واردشده در مدل را نیز مورد تجزیه و تحلیل قرار دادند و توسط روش متقاطع[1] مدل را مورد ارزیابی قراردادند[۱۳]. مدل مورد ارزیابی مدل مناسبی برای پیش بینی اندیس بازداری ترکیبات گزارش شد.

در سال ۲۰۰۴ توسط گرکانی نژاد و همکارانش برای مجموعه‌ای از ۸۴۶ ترکیبات آلی مربوط به شیمی تجزیه با استفاده از توصیف‌کننده‌های استخراج شده از دراگون مدل‌سازی صورت گرفت. مدل‌سازی به روش‌های حداقل مربعات جزئی(PLS) [2] و مدل رگرسیون خطی چند گانه(MLR) انجام شد. انحراف معیار خطاهای پیش بینی شده (SEP) در چهار آزمایش با مجموعه های آموزشی و پیش بینی شده توزیع شده برآورد شد. SEP برای بهترین مدل‌ها حدود ۸۰ واحد شاخص نگهداری است[۱۴].

لیائو[3] و همکارانش در سال ۲۰۰۷، روش MLR را برای پیش‌بینی زمان‌بازداری ۶۹ ترکیب روغن ضروری موجود در گونه خاصی از گل‌ها به-کار بردند[۱۵]. ضریب همبستگی ۰/۹۴ در این روش برای سری داده‌ها گزارش شده است.

در سال ۲۰۰۵، جینگ هو و همکارانش اندیس بازداری را برای ترکیبات آروماتیک چند حلقه‌ای نیتروژن با استفاده از ارتباط کمی ساختار خاصیت(QSPR) [4] پیش بینی کردند[۱۶]. برای نشان دادن تاثیر توصیف‌کننده‌های ملکولی بر اندیس بازداری سه مدل خطی چند متغیره از سه گروه توصیف‌ـ کننده ملکولی ساخته شد. علاوه بر این، هر یک از توصیف کننده های مولکولی در این مدل

[1] Leave One Out Cross Validation
[2] Partial least square(PLS)
[3] Liao
[4] Quantitative structure property relationship

ها به خوبی درک رابطه بین ساختارهای مولکولی واندیس بازداری آن‌ها مورد بحث قرار گرفت. مدل های پیشنهادی به نتایج زیر رسیدند: با ضرایب تعیین R^2 برای مدل ها با یک، دو و سه توصیف کننده مولکولی به ترتیب ۰,۹۵۷۱، ۰,۹۷۷۶ و ۰,۹۴۸۶ بود.

در سال ۲۰۰۸ ریاحی و همکارانش روش GA-MLR را برای مدل‌سازی شاخص بازداری ۴۳ ترکیب آلی به‌کار گرفتند. R^2 گزارش شده برای سری آموزش متشکل از ۳۵ ترکیب، ۰/۹۷۷ و برای بقیه ترکیبات که به عنوان سری تست در نظر گرفته شده اند، ۰/۹۷۸ بود[۱۷].

ریاحی و همکارانش تحقیقاتی در سال ۲۰۰۸، روش ماشین بردار پشتیبان[1] را برای پیش‌بینی شاخص بازداری صد ترکیب آلی، با روش‌های خطی مورد مقایسه قرار دادند و آن را به عنوان روش بهتری نسبت به MLR گزارش کردند[۱۸].

در سال ۲۰۰۸ هیوییینگ[2] و همکارانش اندیس بازداری را برای ۱۱۴ ترکیب PASHs محاسبه نموده و آن‌ها را با مقادیر به‌دست‌آمده از روش تجربی مقایسه کردند و نتایج نشان داد روش تئوری در پیش‌بینی اندیس بازداری این ترکیبات موفق عمل کرده است[۱۹].

در سال ۲۰۰۵ ژنگ [3]و همکارانش اندیس بازداری را برای ۲۰۷ ساختار متیل‌آلکان پیش بینی و ۱۲۷ توصیف‌کننده توپولوژیکی محاسبه کرده‌اند. روش GAPLS، که یک روش انتخاب متغیر ترکیبی با الگوریتم های ژنتیک (GA)، حداقل مقادیر جزئی و جزئی (PLS) است، هفت توصیف-کننده توپولوژیکی از ۱۲۷ توصیف‌کننده توپولوژیکی با استفاده از روش GAPLS برای ساخت مدل QSPR با کیفیت رگرسیون بالا انتخاب شده است، ضریب تعیین (R^2) از ۰/۹۹۹۹، انحراف معیار ۲/۸۸. خطای مدل مشابه خطای آزمایشی است[۲۰]. اعتبارسنجی مدل با استفاده از تکنیک

[1] Support vector machine(SVM)
[2] Hui-Ying
[3] Xiang, Zheng

اعتبارسنجی مجدد بررسی می شود. نتیجه اعتبارسنجی نشان می دهد که مدل ساخته شده با کیفیت پیش بینی بالا قابل اعتماد و پایدار است.

در سال ۲۰۰۴ هونگبین[۱] و همکارانش ترکیبات گوگرد شامل مرکاپتان‌ها، سولفیدها و تیوفن‌ها با استفاده از کروماتوگرافی گازی با تشخیص انتشار اتمی شناسایی شدند. زمان و اندیس بازداری آن‌ها با توصیف‌کننده‌های مولکولی تولید شده از ساختار مولکولی آنها مرتبط بود. بهترین مدل رگرسیون خطی چند گانه پارامتر توانایی پیش بینی خوبی را نشان داد. توصیف‌ کننده‌هایی که در این مدل‌ها دخیل هستند، خواص هندسی، توپولوژیکی و الکترونیکی مولکولها را نشان می دهند که مربوط به تعامل بین حلال و فاز ثابت است. برای ۳۴ ترکیبات گوگرد (بنزوتیوفن و دیابنسوتیوفن)، دو مدل دیگر برای زمان بازداری و اندیس بازداری با خطاهای استاندارد به ترتیب ۰/۶۱ و ۱/۶۳ به دست آمد. چنین مدل هایی برای زمان بازداری و یا اندیس بازداری می‌تواند برای شناسایی قله‌های کروماتوگرافی ناشناخته با تطابق زمان بازداری آن‌ها یا اندیس بازداری با ترکیبات گوگرد از ساختار شناخته شده مولکولی زمانی که استانداردهای شیمیایی مربوطه در دسترس نیستند[۲۱].

در سال ۲۰۱۶ پاشا زانوسی و همکارانش[۲۲] با استفاده از سه روش استخراج به نام‌های ریز استخراج با فاز جامد در فضای فوقانی[۲] استخراج مایکروویو بدون حلال[۳] و تقطیر با آب[۴] مقدار اندیس بازداری را برای ترکیب روغن‌های فرّار پیش بینی کردند و مدل خود را توسط QSPR ارزیابی کردند و از روش رگرسیون خطی چند گانه به بالاترین ضریب همبستگی ($R^2 = 0.952$) و F=104.4[۵] و پایین ترین خطای استاندارد[۶] دست پیدا کردند.

[1] Du, Hongbin
[2] headspace solid-phasemicroextraction (HS-SPME)

[3] solvent-free microwave extraction (SFME)
[4] Hydrodistillation (HD)
[5] Statistic Term
[6] Standard error(SE)

در سال ۲۰۱۶ گودرزی و همکـاران از روش‌هـای جنگـل تصـادفی[1] و شـبکه عصـبی مصـنوعی و رگرسیون خطی چندگانه برای پیش‌بینی اندیس بازداری تعدادی از ترکیبات آلی گـوگرددار اسـتفاده کردند و نتایج نشان داد مدل جنگل تصادفی از مدل‌های دیگر برتر و ارجح است به دلیـل R^2 بـالا بـا مقدار ۰/۹۹۵ و خطای پیش بینی[2] پائین نسبت به سایر روش‌ها با مقدار ۱/۱۴۶ و F بـا مقـدار ۱۱۹۷ باعث شده این بهتر عمل کند[۲۳].

در سال ۲۰۱۴ گودرزی و دیگر همکارانش بر اساس روش جدید جنگل های تصادفی (RF) برای پیش بینی شاخص های نگهداری (RIs) برخی از ترکیبات هیدروکربن آروماتیک چند حلقه ای (PAH)، تحقیق رابطه QSRR کمی انجام دادند. RIs این ترکیبات با استفاده از توصیف های نظری تولید شده از ساختار مولکولی آنها محاسبه شد. اثرات پارامترهای مهم موثر بر توانایی قدرت پیش بینی RF مانند تعداد درختان (nt) و تعداد متغیرهای تصادفی انتخاب شده برای تقسیم هر گره (m) مورد بررسی قرار گرفت. بهینه سازی این پارامترها نشان داد که در نقطه m = 70، nt = 460، روش RF می تواند بهترین نتایج را بدست آورد. همچنین عملکرد مدل RF با شبکه عصبی مصنوعی (ANN) و تکنیک های رگرسیون خطی چندگانه (MLR) مقایسه شد. نتایج بدست آمده نشان دهنده برتری نسبی روش RF در سطح MLR و ANN است[۵۵].

[1] Random forests
[2] Relative Error Prediction

فصل دوم
مطالعات کمومتریکس و کاربرد آن در QSPR

۲-۱-کمومتریکس[1]

عبارت کمومتریکس توسط اسوانت ولد[2] و بروس کووالسکی[3] در اوایل ۱۹۷۰ تعریف شد. قبلا عباراتی مانند بیومتریکس[4] و اکونومتریکس[5] نیز در زمینه‌های علوم زیستی و همین‌طور اقتصاد معرفی شده-اند. پس از آن بود که انجمن بین‌المللی کمومتریکس تاسیس شد. از آن زمان تا کنون پیشرفت‌های زیادی در این زمان به‌وجودآمده و اکنون کمومتریکس در شاخه‌های مختلف شیمی، به ویژه شیمی تجزیه مورد استفاده قرار می‌گیرد. برخی از کاربردهای آن عبارتند از: کنترل فرآیندها، تجزیه و تحلیل و شناخت الگوها، پردازش علائم، بهینه کردن شرایط.

یکی از تعاریفی که از کمومتریکس وجود دارد، "سازماندهی شیمیایی با استفاده از روش‌های ریاضی و آمار برای (الف)-طراحی و انتخاب بهترین فرآیندهای محاسباتی و آزمایشگاهی و (ب)-فراهم نمودن بیشترین اطلاعات شیمیایی با بررسی داده‌های دردسترس" می‌باشد[۲۴].

هاوری[6] و هیرش[7] در اوایل سال ۱۹۸۰ پیشرفت‌های کمومتریکس را در مراحل مختلف رده بندی کردند. اولین مرحله قبل از ۱۹۷۰ است که در آن تعدادی از روش‌های ریاضی توسعه یافتند و در زمینه‌های مختلف ریاضی، علم رفتار و مهندسی به‌کارگرفته شدند. در این محدوده زمانی، شیمیدانان خود را به روش‌های آنالیز داده‌ها، یعنی محاسبات پارامترهای آماری مانند میانگین، انحراف استاندارد و حدود اطمینان محدود کرده بودند. هاوری و هیرش تحقیقات خود را به‌طور ویژه روی همبستگی تعداده زیادی از داده‌های شیمیایی و ویژگی‌های ملکولی مربوط، متمرکز کردند. این اقدامات اولیه اساس یکی از شاخه‌های مهم کمومتریکس یعنی ارتباط کمی ساختار-فعالیت (QSAR)[8] را فراهم کرد. دومین مرحله تکامل کمومتریکس در سال ۱۹۷۰ واقع شد، هنگامی که کمومتریکس توجه

[1] chemometrics
[2] Svant Wold
[3] Bruce R. Kowalaski
[4] biometrics
[5] Econometrics
[6] Howery
[7] Hirsch
[8] Quantitative Structure- Activity Relationship(QSAR)

شیمیدانان، به‌خصوص شیمیدانان تجزیه را که به این ترتیب روش‌های بیشتری برای آنالیز داده‌های دردسترس و همچنین روش‌های نوینی برای بدست آوردن اطلاعات پیش‌روی خود می‌دیدند، به خود جلب کرد. مرحله بعدی تکامل کمومتریکس که ابتدا توسط هاوری و هیرش و سپس توسط براون[1] پیش‌بینی شده بود از اوایل ۱۹۸۰ آغاز شد. هم‌اکنون کمومتریکس ابزار قوی و کارآمد برای شیمیدانان به‌شمار رفته و به عنوان یک زیر شاخه رشته شیمی در بسیاری از دانشگاه‌های آمریکا و اروپا و بعضی از دانشگاه‌های چین و سایر کشورها تدریس می‌شود. پیشرفت‌ها در علوم کامپیوتر فناوری اطلاعات آمار و ریاضی کاربردی عناصر جدیدی را در کمومتریکس معرفی کرده است[۲۵،۲۶]. اولین مقاله منتشر شده در زمینه کمومتریکس در ایران به سال ۱۹۹۸ برمی‌گردد[۲۷]. با رشد سریع کمومتریکس در جهان، استفادهٔ شیمی‌تجزیه‌دانان ایرانی از کمومتریکس در تحقیقاتشان افزایش یافت. تا سال ۲۰۰۵ حدود ۲۰۰ مقاله علمی توسط محققین کمومتریکس در ایران منتشر و بیش از ۸ گروه تحقیقاتی فعال در این زمینه مشغول فعالیت می‌باشند. انجمن کمومتریکس ایران نیز به عنوان یکی از شاخه‌های فعال انجمن شیمی ایران در سال ۲۰۰۰ تاسیس شد و همچنین سمینارهای داخلی و بین‌المللی در زمینهٔ کمومتریکس در ایران برگزار می‌گردد. زمینه‌هایی که تاکنون بیشتر در ایران مورد توجه قرار گرفته‌اند مطالعات(QSAR/QSPR[2])، کالیبراسیون چند متغیره[3] و هوش مصنوعی[4] هستند[۲۸].

۲-۲-مراحل مطالعه ارتباط کمی ساختار – خاصیت (QSPR)

در مطالعات QSPR برای بیان ارتباط ساختار ملکول با خاصیت مورد نظر باید مراحل زیر به ترتیب طی شود .

۱-فراهم کردن سری داده‌ها

[1] Brown
[2] Quantitative Structure- Property Relationship(QSPR)
[3] Multivariate Calibration(MVC)
[4] Artificial Intelligence(AI)

۲-رسم و بهینه کردن ساختار هندسی ترکیبات

۳-محاسبهٔ توصیف کننده‌ها

۴-انتخاب بهترین توصیف کننده‌ها

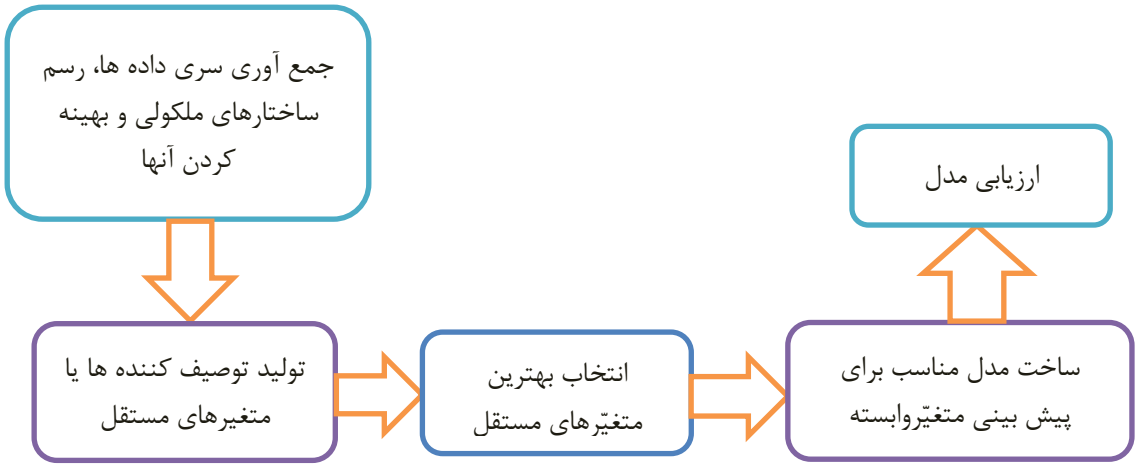
۵-مدل‌سازی

۶-ارزیابی قدرت پیش‌بینی مدل[۲۹].

۲-۳-ارتباط کمی ساختار-خاصیت[1]

QSPR در لغت به معنی برقراری ارتباط کمی بین ساختار و خواص مولکول می‌باشد. درواقع با استفاده ازآنچه قبلاً به‌صورت تجربی انجام‌شده، این ارتباط برقرار می‌شود و سپس از آن برای پیش‌بینی فعالیت یا خواص ترکیبات جدید استفاده می‌گردد. زمانی که نمونه‌های استاندارد ازلحاظ فعالیت در دسترس نباشند، آزمایش‌ها وقت‌گیر، پیچیده، پرهزینه و یا با خطر همراه باشند. روش QSPR روش مناسبی برای حل مشکل خواهد بود. درروش QSPR فرض بر آن است که بین خواص فیزیکی-شیمیایی و ساختار مولکولی ترکیبات، یک رابطه خطی یا غیرخطی وجود دارد و در همین راستا سعی بر آن دارد که با برقراری روابط ساده ریاضی، فعالیت/خاصیت مجموعه‌ای از ترکیبات را پیش‌بینی نماید. بدین منظور، ابتدا بایستی توصیف‌کننده‌های مولکولی مناسب که بتواند معرف ترکیبات مختلف باشند، انتخاب‌شوند و سپس با استفاده از الگوریتم‌هایی نظیر رگرسیون خطی چندگانه و یا شبکه‌های عصبی مصنوعی، مدل QSPR مربوطه ساخته شود. مهم‌ترین مسئله در توسعهٔ مدل‌های QSPR انتخاب توصیف‌کننده‌هایی است که بستگی بسیار زیادی به خاصیت/فعالیت موردنظر داشته باشد و معمولاً این انتخاب باید از میان تعداد زیادی توصیف‌کننده‌های مولکولی صورت پذیرد.

[1] Quantitative structure-property relationship



شکل ۲-۱-شمای کلی مراحل به‌کار گرفته شده در یک فرآیند QSPR

۲-۳-۱- فراهم کردن سری داده‌ها

برای دست‌یابی به مدل مناسب، اولین مرحله انتخاب یک دسته از مولکول‌هاست که مقادیر تجربی یک خصوصیت فیزیکی یا شیمیایی آن‌ها مانند زمان بازداری، نقطه‌جوش و...در دسترس باشد. در این خصوص لازم است شرایط اندازه‌گیری خاصیت موردنظر برای همه ترکیبات یکسان باشد. بدیهی است که هرچقدر اطلاعات تجربی قابل‌دسترس برای طراحی مدل بیشتر باشد بدون تردید مدل کارایی بهتری خواهد داشت. فراهم کردن سری داده‌های مناسب از طریق جستجو در مقالات صورت می‌گیرد.

۲-۳-۲- بهینه‌سازی ساختار هندسی ترکیبات

کمومتریکس از روش‌های مفیدی که قادر به استخراج اطلاعات موجود در داده هستند، استفاده می‌کند. به این منظور از شیمی محاسباتی استفاده می‌شود. شیمی محاسباتی ساختارهای ملکولی را به صورت پارامترهای عددی معرفی و رفتار آن‌ها را با معادلات کوانتومی و فیزیک کلاسیک شبیه سازی می‌نماید. این امر به دانشمندان امکان می‌دهد که بتوانند از این طریق به اطلاعات ملکول از جمله ساختار هندسی، انرژی‌ها و خواص الکترونیکی و اثرات حلال دست پیدا کنند. در سال‌های اخیر

روش‌های محاسباتی در بین شیمیدانان رواج یافته است. در این روش‌ها می‌توان به راحتی محاسبات را انجام داد، بدون اینکه از اصول اولیه و روش محاسبه آگاهی دقیق داشت. روش‌های محاسباتی در شیمی به چندین دسته تقسیم می‌شوند. روشی که در این تحقیق به‌کار گرفته شده، روش AM1[1] است که جز روش‌های محاسباتی نیمه‌تجربی[2] می‌باشد. در روش‌های نیمه‌تجربی، که در برنامه‌هایی نظیر Hyperchem وارد شده‌اند، محاسبات براساس مکانیک کوانتومی صورت می‌گیرد. Hyperchem یک نرم‌افزار قدرتمند مدل‌سازی مولکولی است که تمام ابزارهای موردنیاز برای ترسیم ساختارهای دوبعدی و سه‌بعدی مولکولی را در محیطی ساده و انعطاف‌پذیر در خود گردآوری کرده است که برای بهینه‌سازی ساختار هندسی مولکول‌ها از این نرم‌افزار استفاده می‌شود که توسط روش نیمه تجربی AM1 با استفاده از الگوریتم Polak-Rabiere تا ریشه میانگین مربع گرادیان ۰/۰۰۱انجام شد. ایجاد توصیف‌کننده‌های هندسی و هیبریدی، برمبنای ساختار و هندسه دقیق مولکولی استوار است و اگر ساختارها به شکل صورت‌بندی[3] با حداقل انرژی نباشند مقادیر غیر صحیحی برای این توصیف کننده‌ها ایجاد می‌شود[۳۰]. به کمک این نرم‌افزار می‌توان طول پیوند، زاویهٔ پیوندی و زوایای پیچشی را در مولکول تعیین کرد. داده‌های حاصل از این نرم‌افزار را می‌توان به‌عنوان ورودی به سایر نرم‌افزارها معرفی نمود.

۲-۳-۳- محاسبهٔ توصیف‌کننده‌ها

توصیف کننده‌ها کمیت‌هایی هستند که به ساختار مولکول ارتباط دارند و برای هر مولکولی مقادیر خاصی را به خود اختصاص می‌دهند. درواقع هر توصیف کننده بیان‌گر خصوصیت ویژه‌ای از مولکول است که ممکن است بر فعالیت و خاصیت موردنظر مؤثر باشد و هر توصیف کننده اطلاعات خاصی از مولکول را که بر کمیت مورد مدل‌سازی اثر می‌گذارد را در اختیار می‌دهد. محاسبهٔ توصیف کننده‌ها توسط نرم‌افزار Dragon انجام می‌شود. از این نرم‌افزار جهت استخراج توصیف‌کننده‌های

[1] Austin methods
[2] Semi-empirical methods
[3] Conformation

مولکولی استفاده می‌شود، Dragon قادر به محاسبهٔ بیش از ۳۲۲۴ توصیف‌کننده می‌باشد که بـه ۲۲ دسته اصلی تقسیم‌بندی می‌شوند که در جدول (۲-۱) نام این ۲۲ دسته اصلی آورده شده است. علاوه بر محاسبهٔ ساده‌تـرین نـوع اتـم‌هـا و گـروه‌هـای عـاملی و شـمارش اجـزا مـی‌تـوان تعـداد زیـادی از توصیف‌کننده‌های توپولوژیکی و هندسی را نیز توسط این نرم‌افزار محاسبه نمود.

جدول (۲-۱)- انواع توصیف کننده‌های محاسبه‌شده توسط نرم‌افزار Dragon

۱- توصیف‌کننده‌های زیر ساختاری[۱]	۲- توصیف‌کننده‌های توپولوژیکی[۲]
۳- شمارنده‌های مولکولی مورس[۳]	۴- توصیف‌کننده‌های BCUT[۴]
۵- شاخص‌های بار[۵]	۶- خود ارتباطی‌های دوبعدی[۶]
۷- توصیف‌کننده‌های بار[۷]	۸- شاخص‌های آروماتیسیته[۸]
۹- پروفایل‌های مولکولی راندیک[۹]	۱۰- توصیف‌کننده‌های هندسی[۱۰]
۱۱- توصیف‌کننده‌های عملکرد توزیع شعاعی[۱۱]	۱۲- توصیف‌کننده سه‌بعدی پراش الکترونی[۱۲]
۱۳- توصیف‌کننده‌های ملکولی مقیاس وزنی[۱۳]	۱۴- توصیف‌کننده‌های GETAWAY[۱۴]
۱۵- گروه‌های عاملی[۱۵]	۱۶- اجزای میان اتمی
۱۷- توصیف‌کننده‌های تجربی[۱۶]	۱۸- خصوصیات مولکولی
۱۹-شاخص‌های اتصال[۱۷]	۲۰- اثرانگشت فرکانسی[۱۸]
۲۱-شاخص‌های اطلاعاتی[۱۹]	۲۲-شاخص‌های مجاورت[۲۰]

[1] Substructural Descriptors
[2] Topological Descriptors
[3] Morse Molecular Counts
[4] BCUT Descriptors
[5] Charge Indices
[6] 2D Autocorrelations
[7] Charge Descriptors
[8] Aromatic Indices
[9] Rancid Molecular Profiles
[10] Geometrical Descriptors
[11] Radial Distribution Function Descriptors(RDF)
[12] 3D Molecule Representation of Structure Backed on Electron Diffraction Descriptors
[13] Weighted Holistic Invariant Molecular Descriptors(WHIM)
[14] Geometry, Topology and Atom Weights Assembly
[15] Functional Groups
[16] Empirical Descriptors
[17] Connectivity indices
[18] fingerprint Frequency
[19] Information indices
[20] Adge adjacency indices

۲-۳-۴- انتخاب توصیف‌کننده‌های مهم

یکی از مهم‌ترین مراحل QSPR، انتخاب توصیف‌کننده‌های مناسب می‌باشد زیرا توصیف‌کننده‌های نامناسب کار برازش و مدل‌سازی را طولانی می‌کنند و تأثیری در بهبود نتایج نخواهند داشت. به این منظور باید در انتخاب توصیف‌کننده‌ها دقت لازم به عمل آید و توصیف‌کننده‌هایی که بیشترین و نزدیک‌ترین ارتباط را به پارامتر موردنظر دارند، انتخاب شوند (توصیف‌کننده‌هایی که بیشترین وابستگی را با خاصیت موردنظر دارند). ازجمله روش‌های انتخاب توصیف‌کننده[۱] می‌توان به رگرسیون گام‌به‌گام[۲]، سهم گروه[۳]، الگوریتم ژنتیک[۴] و روش جایگزینی[۵] اشاره کرد.که در این پایان‌نامه از روش رگرسیون گام‌به‌گام در نرم‌افزارSPSS[۶] برای انتخاب بهترین توصیف‌کننده‌ها استفاده‌شده است.

۲-۳-۴-۱- رگرسیون گام‌به‌گام

این روش با محاسبهٔ ضرایب همبستگی که معیاری برای سنجش همبستگی بین یک توصیف‌کننده و متغیر پاسخ است، انجام می‌شود. مقدار این ضریب بین -۱ و ۱ است. در این روش ابتدا توصیف‌کننده‌ای که بیشترین همبستگی با متغیر وابسته(خاصیت موردنظر) را دارد، وارد مدل شده و با هر توصیف‌کنندهٔ جدید، کلیه‌ی توصیف‌کننده‌های موجود در مدل بررسی شده و اگر هر کدام از آن‌ها سطح معنادار خود را از دست داده باشند، قبل از ورود توصیف‌کننده جدید، از مدل خارج می‌شوند[۳۱].

[1] descriptor
[2] Step by Step regression (stepwise)
[3] Group contribution
[4] Genetic algorithm (GA)
[5] Replacement method
[6] Statistical Package for the social science (SPSS)

۲-۳-۵- ساخت مدل

مدل درواقع یک رابطهٔ ریاضی است که بیان‌کنندهٔ رابطهٔ بین متغیر وابسته و متغیرهای مستقل می‌باشد و به کمک آن می‌توان با داشتن مقادیر متغیرهای مستقل، متغیر وابسته را ارزیابی کرد. پس از انتخاب مناسب‌ترین توصیف‌کننده‌ها با استفاده از روش‌های آماری مختلف به جستجوی مدل مناسبی پرداخته می‌شود که بتواند ارتباط بین توصیف‌کننده‌های انتخابی و پارامترهای مورد مدل‌سازی را به‌درستی بیان کند. برای مدل‌سازی از روش‌های گوناگون خطی و غیرخطی می‌توان استفاده کرد. روش‌های آماری خطی شامل: رگرسیون خطی چندگانه(MLR)، حداقل مربعات جزئی[۱]، رگرسیون اجزا اصلی و جنگل‌های تصادفی و روش‌های آماری غیرخطی شامل: شبکه عصبی مصنوعی (ANN) و رگرسیون اسپلاین تطبیقی چند متغیره[۲] (MARS) می‌باشد.

در این پایان‌نامه از روش خطی RF و روش‌های غیرخطی ANN و MARS برای ساخت مدل استفاده‌شده است.

۲-۴- مقدمه‌ای بر شبکه عصبی مصنوعی

شبکه عصبی مصنوعی یک الگوی پردازش اطلاعات است که از سیستم‌های عصبی بیولوژیکی مانند مغز الهام می‌گیرد. عنصر کلیدی این ایده، ایجاد ساختارهایی جدید برای سامانهٔ پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق‌العاده به‌هم‌پیوسته بانام نرون[۳] تشکیل‌شده که برای حل یک مسئله با هم هماهنگ عمل می‌کنند و توسط سیناپس[۴]ها (ارتباطات الکترومغناطیسی) اطلاعات را منتقل می‌کنند. در این شبکه‌ها اگر یک سلول آسیب ببیند بقیه

[1] Partial least square (PLS)
[2] Multivariate Adaptive Regression Spline (MARS)
[3] neuron
[4] synapse

سلول‌ها می‌توانند نبود آن را جبران کرده و نیز در بازسازی آن سهیم باشند. این شبکه‌ها قادر به یادگیری‌اند، یادگیری در این سیستم‌ها به‌صورت تطبیقی صورت می‌گیرد، یعنی با استفاده از مثال‌ها وزن سیناپس‌ها به‌گونه‌ای تغییر می‌کند که در صورت دادن ورودی‌های جدید، سیستم پاسخ درستی تولید کند. به‌طورکلی یک شبکه عصبی مصنوعی از سه لایهٔ ورودی، خروجی و پردازش تشکیل می‌شود. هر لایه شامل گروهی از سلول‌های عصبی (نرون) است که عموماً با کلیهٔ نرون های لایهٔ دیگر در ارتباط هستند، مگر اینکه کاربرد ارتباط بین نرون‌ها را محدود کند ولی نرون‌های هر لایه با سایر نرون‌های همان لایه، ارتباطی ندارد. نرون کوچک‌ترین واحد پردازشگر اطلاعات است که اساس عملکرد شبکه‌های عصبی را تشکیل می‌دهد. یک شبکه عصبی مجموعه‌ای از نرون‌هاست که با قرار گرفتن در لایه‌های مختلف، معماری خاصی را بر مبنای ارتباطات بین نرون‌ها در لایه‌های مختلف تشکیل می‌دهند. نرون می‌تواند یک تابع ریاضی غیرخطی باشد، درنتیجهٔ یک شبکهٔ عصبی که از اجتماع این نرون‌ها تشکیل می‌شود، نیز می‌تواند یک سامانهٔ کاملاً پیچیده و غیرخطی باشد. در شبکه عصبی هر نرون به‌طور مستقل عمل می‌کند و رفتار کلی شبکه، برآیند رفتار نرون‌های متعدد است. به‌عبارت‌دیگر، نرون‌ها در یک‌روند همکاری، یکدیگر را تصحیح می‌کنند.

۲-۴-۱-عملکرد نرون مصنوعی

یک نرون با برچسب j که ورودی $P_j(t)$ را از نرون‌های ماقبل خود دریافت می‌کند از اجزا زیر تشکیل می‌شود:

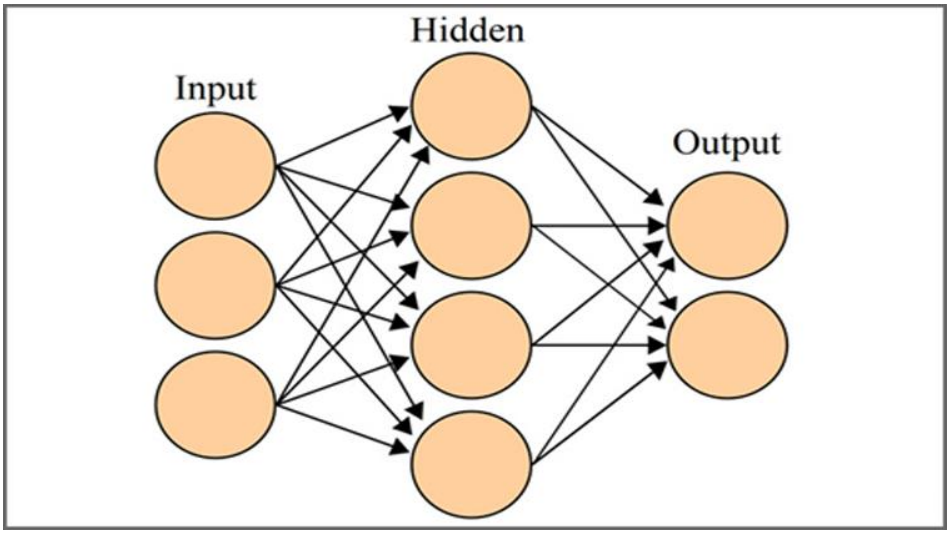
- یک فعال‌ساز که به یک پارامتر گسسته زمان وابسته است.
- احتمالاً یک آستانهٔ θ_j، ثابت است مگر آنکه توسط تابع یادگیرنده تغییر کند.
- یک تابع فعال‌سازی f که فعال‌ساز جدید را در زمان داده‌شده t+1 از θ_j،$a_j(t)$، ورودی خالص P_j حساب می‌کند و رابطه (۲-۱) را بدست می‌آورد:

$$a_j(t+1) = f\left(a_j(t)\right), (P_j(t), \theta_j) \quad (۱-۲)$$

- و یک تابع خروجی که خروجی را از فعال‌ساز حساب می‌کند.

$$O_j(t) = fout\left(a_j(t)\right) \quad (۲-۲)$$

تابع خروجی اغلب تابع همانی است. نرون ورودی نرون ماقبل ندارد اما برای کل شبکه نقش رابط ورودی را ایفا می‌کند. به‌طور مشابه نرون خروجی تالی ندارد و لذا به‌عنوان رابط خروجی کل شبکه ایفای نقش می‌کند.



شکل ۲-۲-شبکه عصبی مصنوعی

شکل(۲-۲) یک شبکه عصبی مصنوعی می‌باشد که شبیه شبکه بزرگ نرون‌ها در مغز است. در اینجا هر رأس دایره نمایشگر یک نرون مصنوعی است و پیکان نمایشگر یک اتصال از خروجی یک نرون مصنوعی به ورودی نرون دیگر است

۲-۴-۲-تابع انتقال[1]

زمانی که ورودی‌ها با توجه به اهمیت آن‌ها با یکدیگر جمع جبری شوند توسط تابع انتقال به نرون بعدی منتقل می‌شوند. درواقع تابع انتقال یکی از اجزای شبکه عصبی می‌باشد که در این پایان‌نامه برای بهینه‌سازی شبکه به کار گرفته‌شده عبارت‌اند از:

➢ تابع انتقال لگاریتم سیگموئید[2] (log sig)

از این تابع انتقال در شبکه‌های پس انتشار[3] استفاده می‌شود. این تابع انتقال مقادیر ورودی را در محدوده ∞- و∞+ دریافت کرده و خروجی بین ۰ و ۱ تولید می‌نماید.

➢ تابع انتقال تانژانت سیگموئید[4] (tan sig)

این تابع انتقال مقادیر ورودی را در محدوده -∞ و∞+ دریافت کرده و خروجی بین -۱ و+۱ تولید می‌نماید.

شکل ۲-۳- توابع انتقال

[1] Transfer function
[2] Logarithm sigmoid transfer function
[3] Back propagation
[4] Hyperbolic tangent transfer function

۲-۴-۳ -ساختارهای شبکه

بر اساس ساختار اتصال نرون‌ها، شبکه عصبی مصنوعی را می‌توان در دودستهٔ کلی طبقه‌بندی کرد:

- شبکه‌های پیش‌خور[1]
- شبکه‌های برگشتی[2]

شبکه‌های پیش‌خور: شبکه‌هایی هستند که مسیر پاسخ در آن‌ها همواره روبه‌جلو پردازش می‌شود و به نرون‌های لایه قبل بازنمی‌گردد. در این نوع شبکه‌ها به سیگنال[3] اجازه می‌دهد که در مسیر یک‌طرفه عبور کند یعنی از ورودی تا خروجی، بنابراین بازخوردی وجود ندارد بدین معنی که خروجی هر لایه تأثیری بر همان لایه و همچنین لایه‌های قبلی ندارد.

شبکه‌های برگشتی: که در گراف آن‌ها به دلیل وجود بازخورد در ساختار شبکه، حلقه به وجود می‌آید. در معمول‌ترین خانواده‌ی شبکه پیش‌خور که پرسپترون چندلایه نامیده می‌شوند، نرون‌ها در لایه‌هایی قرار می‌گیرند و اتصال بین آن‌ها یک‌طرفه است. اتصالات مختلف سبب رفتارهای متفاوت شبکه‌ها می‌شود. به‌طورکلی می‌توان گفت شبکه‌های پیش‌خور استاتیک هستند، به این معنی که از ورودی داده‌شده تنها یک دسته مقدار خروجی تولید می‌کنند نه یک دنباله از مقدار خروجی. این شبکه‌ها بی حافظه هستند و پاسخ آن‌ها به یک ورودی مستقل از وضعیت قبلی شبکه است، از طرف دیگر شبکه‌های برگشتی، سیستم‌های دینامیک هستند و زمانی که یک دنبالهٔ ورودی جدید به آن‌ها داده شود، خروجی نرون‌ها محاسبه می‌شود.

[1] Feed forward
[2] Recurrent networks
[3] Signal

۲-۴-۴-آموزش شبکه‌های پیش‌خور با تکنیک پس انتشار[1]

شبکه پس انتشار نوعی شبکه عصبی چندلایه با تابع انتقال غیرخطی می‌باشد. از بـردار ورودی و هدف در راستای آموزش این نوع شبکه برای تقریب زدن یک تابع، یافتن رابطه بین ورودی و خروجی و دسته‌بندی ورودی‌ها استفاده می‌شود. یک شبکه BP با دارا بـودن بایـاس[2]، یـک‌لایـه سـیگموئید و یک‌لایه خروجی خطی، توانایی تخمین هر تابعی با تعداد نقاط ناپیوستگی محدود را داراست[۳۲]. BP یک الگوریتم استاندارد با کاهش شیب می‌باشد که در آن وزن‌های شبکه در جهت خلاف شـیب تـابع کارایی[3] حرکت می‌کنند. لغت پس انتشار به رفتار شبکه BP در محاسبهٔ شیب در شبکه‌های غیرخطی چندلایه اشاره دارد. پس انتشار خطا یک روش متداول آموزش با ناظر برای شبکه‌های پیش‌خور اسـت یعنی برای به دست آوردن ارتباط بین متغیرهای ورودی و خروجی در یادگیری به الگوی آموزشی نیاز است؛ بنابراین باید داده‌های آموزشی به شبکه ارائه شود. دو روش مختلف برای ارائهٔ الگوهای آموزشی و پیاده‌سازی الگوریتم وجود دارد:

- آموزش گام‌به‌گام[4]
- آموزش دسته‌ای[5]

در آموزش گام‌به‌گام وزن‌ها و بایاس‌ها بعد از اعمال هر ورودی به‌روز می‌شـوند درحـالی‌کـه روش دسته‌ای پس از اعمال تمام ورودی‌ها (اعضای مجموعه آموزشی) عملیات بـه‌روزرسـانی وزن‌هـا انجـام می‌شود. بدین‌صورت که شیب‌های محاسبه‌شده برای هر ورودی باهم جمع می‌شوند تا درنهایت وزن‌ها و بایاس‌ها از طریق آن به‌روز شود[۳۳].

به‌طورکلی آموزش به کمک تکنیک پس انتشار بر طبق مراحل زیر انجام می‌شود[۳۴]:

[1] Back propagation
[2] bias
[3] Performance function
[4] Incremental training
[5] Batch training

۱-انتشار ورودی‌ها از نرون‌های ورودی به سمت نرون‌های خروجی

۲-اختصاص ماتریس وزن‌های تصادفی به هر یک از اتصالات

۳-مقایسهٔ خروجی‌های شبکه با مقادیر واقعی (مقادیر هدف) و محاسبهٔ خطای شبکه

۴-پس انتشار خطا از نرون‌های خروجی به سمت نرون‌های ورودی و اصلاح وزن‌ها

۵-ارزیابی عملکرد شبکه با توجه به تابع کارایی تعیین‌شده

مراحل فوق تا زمانی تکرار می‌شود که به حداکثر تکرار[1] مجاز رسیده باشد یا مقدار تابع کارایی از مقداری که تعیین‌شده کمتر باشد. به دلیل فرآیندی که در طی این تکنیک رخ می‌دهد، این روش انتشار به عقب نامیده می‌شود زیرا ابتدا وزن‌های ارتباطی بین لایهٔ خروجی و لایهٔ پنهان و لایهٔ ورودی اصلاح می‌گردند.

۲-۵- جنگل‌های تصادفی[2]

جنگل تصادفی یک روش مدرن از روش‌های درخت-مبنا[3] هستند که شامل انبوهی از درخت‌های رده‌بندی و رگرسیونی[4] می‌باشد[۳۵]. علاوه بر عملکرد بسیار بالای جنگل‌های تصادفی در پیشگویی متغیر پاسخ، این توانایی را نیز دارند که میزان اهمیت متغیرهای توضیحی را در پیشگویی متغیر پاسخ تعیین کنند. یک جنگل تصادفی مجموعه‌ای از درخت‌های هرس نشده است که هر درخت با الگوریتم

[1] Epoch
[2] Random forests
[3] Tree-based
[4] Classification and Regression trees

افرازهای بازگشتی[1] [۳۶] بدست می‌آید. الگوریتم ساخت یک جنگل تصادفی با T درخت از یک مجموعه داده با n مشاهده و p متغیر بدین‌صورت است:

i) یک نمونه تصادفی با جایگذاری n تایی از مشاهدات انتخاب می‌شود.

ii) برای نمونه انتخاب‌شده یک درخت کلاس‌بندی با استفاده از الگوریتم افرازهای بازگشتی، رشد می‌کند. در هر گره افراز بر اساس یک نمونه تصادفی m تایی از p متغیر پیشگو انجام می‌شود.

iii) الگوریتم افرازهای بازگشتی آن‌قدر ادامه می‌یابد تا درخت به بزرگ‌ترین اندازه خود برسد بدون آن‌که درخت هرس شود.

(iv) مراحل (i) تا (iii) T بار تکرار می‌شود تا یک جنگل تصادفی ساخته شود[۳۷]. انتخاب‌های رایج برای T، ۱۰۰۰ درخت و برای m، $\sqrt{p}$ و log P هستند[۳۸]. یک جنگل تصادفی آن‌قدر بزرگ است که تفسیر آن کار بسیار دشواری است، لذا نیازمند خلاصه کردن اطلاعات آن با استفاده از شاخص‌های کمی هستیم. یکی از این شاخص‌ها اهمیت متغیر[2] (VI) است، VI شاخصی برای رتبه‌بندی متغیرها برحسب اهمیت آن‌ها در اثرگذاری رویه پاسخ[3] است. معروف‌ترین شاخص‌های اهمیت شاخص اهمیت جینی[4] و شاخص اهمیت جایگشتی[5] می‌باشد.

شاخص اهمیت جایگشتی: برای محاسبهٔ این شاخص، الگوریتم جنگل‌های تصادفی از تمام مشاهدات نمونه برای ساخت درخت استفاده نمی‌کند بلکه یک نمونه تصادفی با جایگذاری به حجم n_1 (معمولا برابر $2n/3$) از مشاهدات انتخاب می‌شود. به مشاهدات انتخاب‌شده نمونه آموزشی[6] (LS) به بقیهٔ آن‌ها نمونه خارج کیسه[7] گفته می‌شود. درخت‌ها با مشاهدات LS ساخته می‌شوند و از OOB

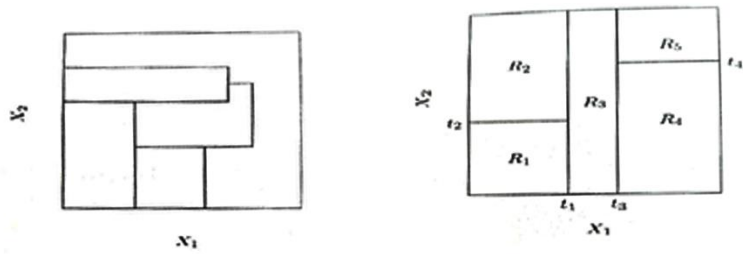
[1] Recurrent partitioning
[2] Variable importance
[3] Response surface
[4] Gini importance index
[5] Permataine importance mdex
[6] Learning sample
[7] Out of bag

برای اندازی گیری ناخالصی درخت استفاده می‌شود. در هر درخت ابتدا اندازه ناخالصی روی مشاهدات OOB محاسبه می‌شود، سپس مقادیر متغیر X_i مشاهدات OOB به‌طور تصادفی جابه‌جا[1] می‌شوند و اندازه ناخالصی درخت روی مقادیر جابه‌جاشده محاسبه می‌شود. اندازه اهمیت متغیر X_i در هر درخت، اختلاف بین این دو اندازه ناخالصی است و میانگین این مقادیر شاخص اهمیت جایگشتی است. انگیزه این روش این است که اگر X_i متغیر مهم می‌باشد جابه‌جا شدن مقادیر آن به‌طور تصادفی منجر به افزایش ناخالصی درخت می‌شود درحالی‌که اگر متغیر تأثیرگذاری نباشد، تغییری ایجاد نمی‌شود. کلیه تحلیل‌ها با استفاده از نرم‌افزار Matlab انجام شد.

۲-۵-۱- روش درخت رگرسیونی[2] (تصمیم)

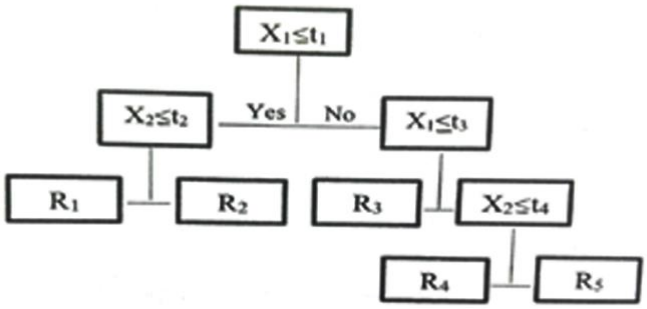
در روش رگرسیون خطی یک مدل واحد برای کل فضای داده‌ها در نظر گرفته می‌شود. اما ممکن است به دلیل تفاوت رفتار متغیر پاسخ در نواحی مختلف فضای داده‌ها، برقراری یک مدل واحد، کارایی لازم را نداشته باشد. لذا یک روش جایگزین، تقسیم‌بندی فضای داده‌ها به بخش‌های کوچک‌تر است تا بتوان رفتار متغیر پاسخ را به‌طور موضعی مدل‌سازی نمود. درروش درخت رگرسیونی هدف این است که مقادیر متغیرهای توضیحی در هر ناحیه از ابر مکعب p بعدی، به‌گونه‌ای تقسیم‌بندی می‌شوند که داده‌های واقع در هر ناحیه تا حد ممکن همگون باشند به‌طوری‌که بتوان آن‌ها را توسط ساده‌ترین مدل پیش‌بینی نمود. به‌طوری که افراز فضای متغیرهای توضیحی را می‌توان به هر شکلی انجام داد. اما درروش درخت رگرسیونی، تنها افرازی قابل‌قبول است که تمامی نواحی ساخته‌شده به شکل مربع یا مستطیل باشند. در شکل (۲-۴) نمونه‌ای از افراز قابل‌قبول و غیرقابل‌قبول در فضای دومتغیره با روش درخت رگرسیونی نشان داده‌شده است.

[1] Permute
[2] Regression tree



شکل۲-۴-شکل (الف) افزار قابل‌قبول شکل (ب) افزار غیرقابل‌قبول

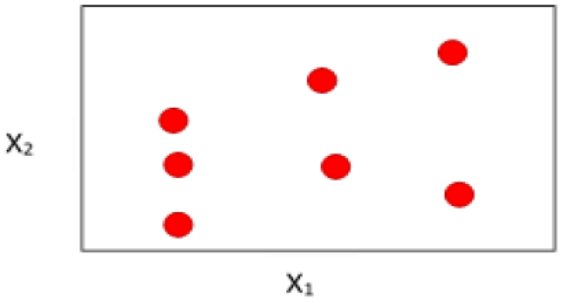
در شکل (۲-۵)، نوع دیگری از نمایش این افراز نشان داده‌شده است. دلیل نام‌گذاری درخت رگرسیونی را می‌توان به نمایش درختی افراز فضای متغیرهای ورودی نسبت داد که در آن هر مشاهده از نقطه بالای نمودار درختی واردشده و دریکی از نواحی افراز قرار می‌گیرد.



شکل۲-۵-نمودار درختی فضای افراز شده در افراز قابل‌قبول

۲-۵-۲-الگوریتم تشکیل درخت رگرسیونی

برای ساده کردن موضوع، ابتدا یک مسئله با دو توصیف‌کننده را در نظر می‌گیریم. فرض کنید نمودار پراکندگی داده‌ها به صورت شکل(۲-۶) باشد:

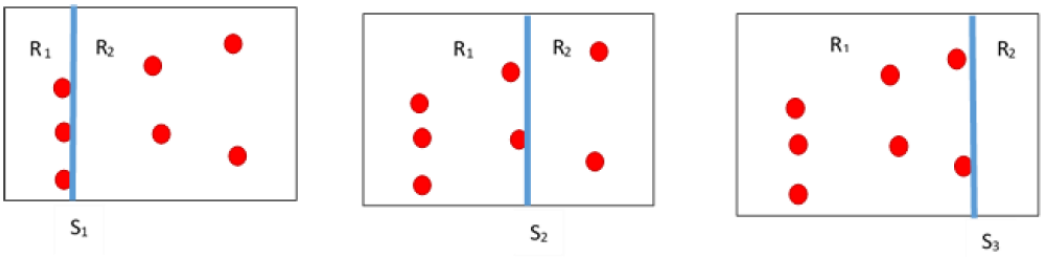


شکل(۲-۶)-نمودار پراکندگی داده‌های دومتغیره

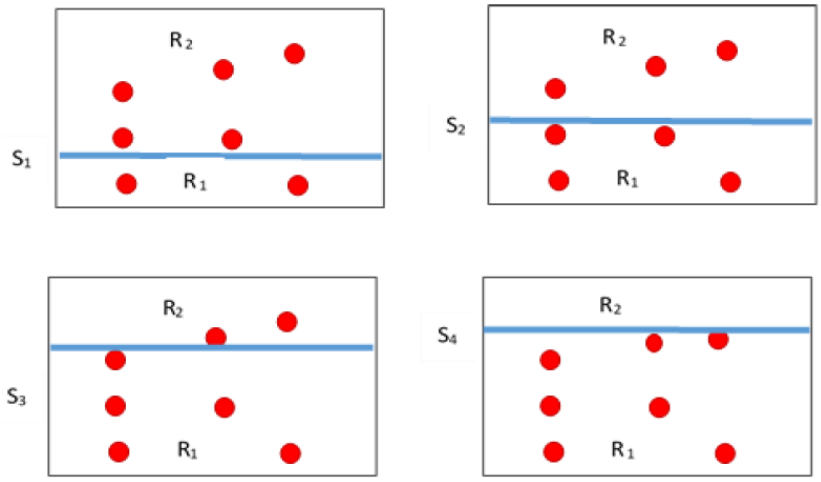
برای اولین افراز، انتخاب‌های مختلفی وجود دارد. شکل‌های (۲-۷) و (۲-۸)، برخی از افرازهای ممکن در راستای دو توصیف‌کننده را نشان می‌دهند. برای انتخاب بهترین تقسیم‌بندی، حاصل جمع مجموع مربعات خطا (A) در دو قسمت ایجاد شده به صورت زیر محاسبه می‌شود:

$$A = \sum_{i=1}^{n}(y_{i1} - \bar{y}_{i1})^2 + \sum_{i=1}^{m}(y_{i2} - \bar{y}_{i2})^2 \qquad (۲-۳)$$

در این رابطه n تعداد داده‌های قرار گرفته در قسمت اول و m تعداد داده‌های قرار گرفته در قسمت دوم، y_{i1} و y_{i2} متغیر پاسخ داده‌های قرار گرفته در قسمت اول و دوم $\bar{y}_i$، میانگین مقادیر متغیر پاسخ در هر قسمت است. افراز بهینه افرازی است که در آن، مقدار عبارت فوق کمینه شود. پس از انتخاب بهترین افراز و به دست آمدن دو ناحیه مجزا، فرآیند فوق در هریک از نواحی تولید شده تکرار می‌شود.



شکل(۲-۷)-سه افراز ممکن در راستای متغیر x_1



شکل(۲-۸)- چهار افراز ممکن در راستای متغیر x_2

در حالت چند متغیره باید تمام افرازهای ممکن در راستای هر توصیف‌کننده در نظر گرفته شده و عبارت متناظر (۲-۳) محاسبه شود. به عبارت دیگر، روش CART[1] در هر مرحله افراز فضای توصیف-توصیف‌کننده‌ها باید هم‌زمان جستجو کند که کمترین مقدار عبارت (۲-۳) به ازای افراز در چه راستایی و از کدام نقطه (گره) حاصل می‌شود[۳۹].

۲-۵-۳-اندازه درخت و هرس کردن

یک مطلب مهم در ساخت درخت رگرسیونی این است که هر درخت تا کجا می‌تواند رشد کند. اگر درخت به اندازه کافی رشد نکند و یا به عبارت دیگر تفکیک فضای توصیف‌کننده‌ها در مراحل اولیه متوقف شود، ممکن است مدل مناسبی ارائه نشود و اگر هم درخت بیش از حد رشد کند، احتمال رخ دادن بیش برازشی[2] وجود دارد. برای انتخاب اندازهٔ درخت نیاز به یک پارامتر در ساختار مدل است

[1] Classification and regression trees
[2] Over fititing

به‌طوری که این پارامتر مقدار بهینهٔ اندازهٔ درخت را به دست آورد. در هر مرحله از رشد درخت، مدل حاصل از روش درخت رگرسیونی دقیق‌تر شده و مجموع توان‌های دوم خطا کاهش می‌یابد. یک راه توقف برای رشد درخت می‌تواند براساس میزان کاهش مجموع توان‌های دوم خطا باشد. بدین ترتیب که اگر در یک مرحله از رشد درخت، مجموع توان‌های دوم خطا دچار کاهش چندانی نشود، می‌توان از رشد در خت در آن مرحله صرف نظر کرد. این میزان کاهش در مجموع توان‌های دوم خطا می‌تواند در اختیار کاربر باشد[۳۹].

۲-۵-۴- الگوریتم جنگل تصادفی

اساس روش جنگل‌های تصادفی وابسته به ماهیت درخت رگرسیونی است و متشکل از ترکیب مجموعه‌ای از درخت‌های رگرسیونی می‌باشد[۴۰]. هر درخت یک مدل تولید می‌کند و از برآیند یا ترکیب این مدل‌ها، مدل نهایی به دست می آید. برخلاف روش درخت رگرسیونی، در این روش برای ساخت هر درخت، تنها بخشی از داده‌ها مورد استفاده قرار می‌گیرند، زیرا انتخاب داده‌های آموزش با جایگذاری[1] انجام می‌شود. همچنین در جنگل‌های تصادفی برای افراز فضای توصیف‌کننده‌ها و ساخت هر گره، به جای استفاده از تمام توصیف‌کننده‌ها، از زیر مجموعه‌ای از توصیف‌کننده‌ها استفاده می‌شود. فرض کنید $i = 1, \dots \dots, N$ و (X_i, Y_i) مجموعه‌ای از داده‌های آموزش باشد که در آن $X_i = (x_{i1}, \dots \dots, x_{iM})$، برداری از M توصیف‌کننده و Y_i متغیر پاسخ متناظر آن‌هاست. اگر تعداد کل درخت‌های مدل با n_{tree} نشان داده شود، مراحل زیر بیان‌گر ساخت هر درخت است:

الف-یک نمونه N تایی از داده‌های آموزش به روش جایگذاری برای ساخت مدل انتخاب می‌شوند. حدود یک سوم از داده‌های اصلی که در ساخت درخت شرکت نمی‌کنند، داده‌های خارج از کیسه

[1] With replacement

(OOB) نام دارند که برای هر درخت متفاوت هستند و نقش داده‌های ارزیابی را برای آن درخت ایفا می‌کنند.

ب-داده‌های انتخاب شده در مرحله اول و انتخاب تصادفی m توصیف‌کننده که زیر مجموعه‌ای از تعداد کل توصیف‌کننده‌ها است، جستجو برای یافتن بهترین توصیف‌کننده، فقط در بین m توصیف-کننده مذکور انجام شده و فضای توصیف‌کننده‌ها به دو قسمت تقسیم می‌شود. در این مرحله پیشنهاد بریمن این است که در مدل رگرسیونی می‌توان این تعداد را برابر با $M/3$، $M/6$ و یا $2M/3$ در نظر گرفت.

ج-برای هریک از دو ناحیه تولید شده، با استفاده از همان نمونه آموزش انتخاب شده در مرحله (الف)، مجددا مرحله (ب) انجام می‌شود.

چ-مرحله (ج) برای تمامی نواحی افزار شده تا زمانی تکرار می‌شود که تعداد مشاهدات در تمامی نواحی کمتر از $2n_r$ باشد. پارامتر n_r در اختیار کاربر بوده و بیانگر حداقل تعداد مشاهدات موجود در هر ناحیه است. مقدار پیش فرض این پارامتر برای مسائل رگرسیون برابر با ۵ است.

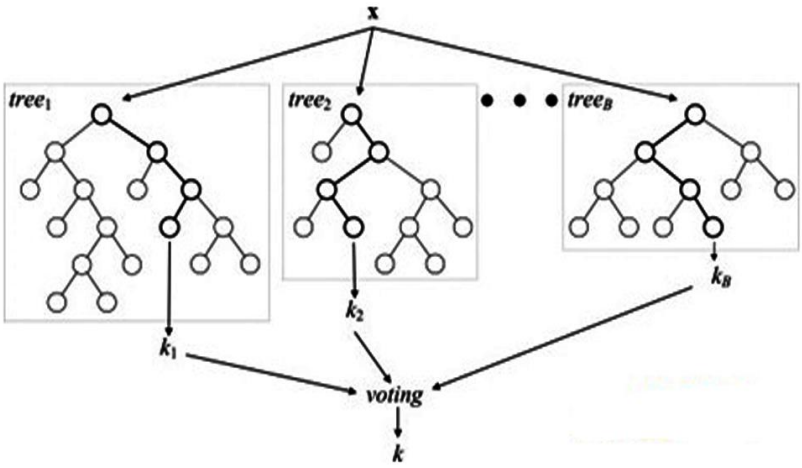
ح-مدل درخت رگرسیونی iام ساخته شده که فضای توصیف‌کننده‌ها را به ناحیه‌ی $R_{i1}, R_{i2}, \dots \dots R_{ir}$ تقسیم کرده است، به صورت زیر می‌باشد:

$$\hat{f}_i(x) = \sum_{j=1}^{r_i} \hat{C}_j I_{R_{ij}}(x) \qquad (۴-۲)$$

که در آن، $\hat{C}_j = \bar{y}_j$ و $I_{R_{ij}} = \begin{cases} 1, x \in R_{ij} \\ 0, x \notin R_{ij} \end{cases}$ می‌باشد.

از آنجایی که در روش جنگل‌های تصادفی تعداد n_{tree} درخت رگرسیونی وجود دارد، می‌توان گفت که تعداد n_{tree} مدل به صورت معادله(۲-۴) خواهیم داشت. اگر برای هر داده، مدل خروجی درخت i ام را به صورت $\hat{f}_i(X)$ نشان دهیم، برآورد متغیر پاسخ برای این داده، با میانگین‌گیری از مدل خروجی تمام درخت‌ها یعنی از رابطه زیر بدست می‌آید[۴۱]:

$$\hat{y}(x) = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} \hat{f}_i(x) \qquad (۵-۲)$$



شکل۲-۹- نمای کلی الگوریتم

شکل ۲-۹ نمایی کلی از الگوریتم جنگل تصادفی است که یک متد یادگیری تجمعی برای رده‌بندی و رگرسیون است. مدل‌های پایه، درختان رگرسیونی CART هستند و خروجی K، بر اساس رأی گیری خروجی‌های B عدد درخت تعیین می‌شود. در مورد رده‌بندی، رأی نهایی بر مبنای مُد خروجی‌ها و در مورد رگرسیون بر اساس میانگین گیری تعیین می‌شود.

بکار گیری این استراتژی باعث عملکرد مناسب و کارا در مقایسه با دیگر الگوریتم‌ها همچون آنالیزهای جداسازی، بردارهای پشتیبان و شبکه عصبی مصنوعی می‌شود. همچنین این تکنیک به

تنظیم دو پارامتر (۱) تعداد متغیرهای قرارگرفته در زیرمجموعه متغیرها به تصادف انتخاب‌شده (۲) تعداد درختان مورداستفاده، نیاز دارد و علاوه بر آن، حساسیت شدیدی نسبت به مقادیر این پارامترها ندارد.

۲-۵-۵- تعیین اهمیت متغیرها درروش جنگل‌های تصادفی

در ساختار روش RF امکانی وجود دارد که می‌توان میزان اهمیت متغیرها[1] را در مدل، تعیین نموده و متغیرهایی که دارای نقش بیشتری در هر درخت و در مدل نهایی دارند شناسایی کرد. همان‌طور که در الگوریتم روش RF اشاره شد، برای تشکیل هر درخت، یک نمونه با جایگذاری از داده‌های اصلی مورداستفاده قرار می‌گیرد. همچنین داده‌هایی که در این نمونه حضور ندارند را OOB نامیدیم که به‌نوعی نقش داده‌های آزمایشی را برای ارزیابی آن درخت ایفا می‌کند. فرض کنید خطای پیش‌بینی Y برای داده‌های OOB در درخت iام، با نماد $EOOB_i$ نشان داده شود .برای تعیین اهمیت متغیر jام، مقادیر این متغیر را به‌طور تصادفی n_{perm} مرتبه(n_{perm}، پارامتری است که در اختیار کاربر می‌باشد) جابه‌جا کرده و مجدداً خطا، به ازای مجموعهٔ جدید محاسبه می‌شود که مقدار این خطا با نماد $\widetilde{EOOB_i^j}$ نشان داده می‌شود. میزان اهمیت متغیر jام در مدل درخت iام با $(VI_i(x^j))$ و به‌صورت زیر تعریف می‌شود:

$$VI_i(x^j) = \widetilde{EOOB_i^j} - EOOB_i \qquad (۶-۲)$$

سرانجام، میزان اهمیت متغیر jام در مدل نهایی $(VI_i(x^j))$ به‌صورت زیر است:

$$VI_i(x^j) = \frac{1}{ntree}\sum_{i=1}^{ntree}(VI_i(x^j)) \qquad (۷-۲)$$

[1] Descriptor importance

برای دقت بیشتر می‌توان جابه‌جا نمودن تصادفی مقادیر متغیر jام را به تعداد بیشتر از یک‌بار انجام داد و در هر بار خطای برآورد را محاسبه نمود و سر انجام مقدار $\widetilde{EOOB_i^j}$ را به‌صورت میانگین خطاهای حاصله در نظر گرفت [۴۲،۴۳]. قابل‌ذکر است که اندازه اهمیت یک متغیر، به‌تنهایی قابل تفسیر نبوده و فقط برای رتبه‌بندی متغیرها بر اساس اهمیت آن‌ها در مدل به کار می‌رود.

۲-۵-۶- مزیت روش جنگل تصادفی

۱-در این روش در مجموعه داده‌هایی که تعداد توصیف‌کننده زیادی دارد از کارایی بالایی برخوردار است. زیرا در هر مرحله از مدل‌سازی تنها زیرمجموعه‌ای از توصیف‌کننده‌ها انتخاب می‌شود.

۲-در این روش بیش برازشی رخ نمی‌دهد.

۳-در صورت کم بودن تعداد داده‌ها، به دلیل وجود داده‌های OOB، نیاز به در نظر گرفتن تعدادی از داده‌ها به‌عنوان سری ارزیابی وجود ندارد.

۴-این روش توانایی انتخاب متغیر را داراست و برخلاف سایر روش‌های یادگیری ماشین، نیازمند روشی دیگر برای انتخاب متغیر نیست[۴۴].

۲-۶-رگرسیون اسپلاین تطبیقی چند متغیره(MARS)

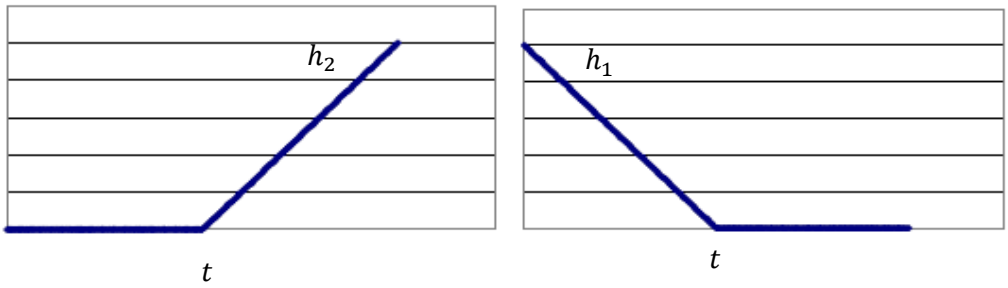
الگوی رگرسیون اسپلاین تطبیقی چند متغیره توسط فریدمن[1] در سال ۱۹۹۱ معرفی شد که تعمیمی از رگرسیون خطی گام به گام یا شکل اصلاح شدهٔ درخت رگرسیون است [۴۵]. در این روش توابع زیر موسوم به توابع مبنا، مدل رگرسیونی را مشخص می‌کنند:

[1] Friedman

$$h_1(x) = (x-t)_+ = \begin{cases} x-t & x > t \\ 0 & x \le t \end{cases} \qquad (۸-۲)$$

$$h_2(x) = (t-x)_+ = \begin{cases} t-x & x < t \\ 0 & x \ge t \end{cases} \qquad (۹-۲)$$

که در آن ثابت t گره نامیده شده و نماد "+" به معنای بخش مثبت است. گره t در عمل یکی از مشاهدات متغیر x است. توابع $h_1(x)$ و $h_2(x)$ را چوب هاکی و یا زوج منعکس یافته[1] نیز می‌نامند. شکل (۲-۱۰) نمودار این توابع را نشان می‌دهد.



شکل (۲-۱۰)-نمودار توابع مبنای $\boldsymbol{h_1}$ و $\boldsymbol{h_2}$

مجموعه C (رابطهٔ (۲-۱۰)) را تمامی زوج توابع منعکس یافته در نظر می‌گیریم که $\{x_1, x_2, x_3 \dots \dots \dots, x_n\}$ مقادیر مشاهده‌ای x هستند.

$$C = \{(x-t)_+, (t-x)_+\} \quad \forall\, t \in \{x_1, x_2, x_3 \dots, x_n\} \qquad (۱۰-۲)$$

ساختار کلی الگوی رگرسیون یک متغیرهٔ مارس برای برآورد Y مطابق رابطهٔ (۲-۱۱) است که h_i یکی از توابع مجموعه C است.

$$Y = f_{2m}(x) = b_0 + \sum_{i=1}^{2m} b_i\ h_i(x) \qquad (۱۱-۲)$$

[1] Reflected pair

ضرائب $\{ b_0, b_1, \dots \dots \dots, b_{2m}\}$ با کمینه‌سازی مجموع مربعات خطا برآورد می‌شوند. روش MARS سه مرحله دارد ۱) آرایش (ایجاد الگوی اولیه به روش پیشرو) ۲)پیرایش (اصلاح الگوی اولیّه) و ۳) گزینش (انتخاب الگوی بهینه)[۴۵].

الگوهای مختلفی در مراحل آرایش و پیرایش براساس توابع پایه بر داده‌ها برازش ومجموع مربعات خطا[1] اندازه‌گیری می‌شود. الگوی نهایی در مرحله گزینش و با کمترین مقدار معیار اعتبار متقابل تعمیم یافته[2]، $(GCV_j = SSE_j \div (1 - \frac{vm_j}{n}))$ (J=1,2,…,2n-2 با، انتخاب می‌شود. m_j تعداد توابع پایه در الگوی j، SSE_j مجموع مربعات خطای الگوی j، v پارامتر هموارسازی و n حجم نمونه است.

۲-۶-۱-MARS یک متغیره

در ابتدا مدل مارس را فقط با یک متغیر ورودی X بیان کرده، سپس در بخش‌های بعد آن را به ورودی‌های چندگانه تعمیم می‌دهیم. دراین حالت، داده‌های مشاهده شده را به صورت $((X_1,Y_1), \dots . (X_n,Y_n))$ که (X_i,Y_i) معرف i امین مشاهده است، در نظر می‌گیریم.

➢ مرحله پیش‌رو

در ابتدا مدل در نظر گرفته شده فقط شامل عرض از مبدا است، یعنی

$$\widehat{f_1}(\mathrm{x}) = \widehat{\beta}_{.}, \qquad (۱۲-۲)$$

که در آن $\bar{y} = \widehat{\beta}_{.}$ این مدل، مدل ۱ نامیده می‌شود. سوال این است که کدام‌یک از زوج توابع پایه مجموعه C باید وارد مدل شوند، تمام مدل‌های ممکن که با هریک از زوج توابع پایه ساخته می‌شوند (به اندازه K زوج توابع پایه تعداد 2K مدل داریم.) را در نظر می‌گیریم و سپس برای ساختن مدل ۲،

[1] Sum of square error (SSE)
[2] Generalized cross validation

از بین این مدل‌ها، مدلی را که دارای کمترین مقدار مجموع توان‌های دوم خطا نسبت به سایر مدل‌ها است، انتخاب می‌کنیم.

$$\hat{f}_{21}(X) = \hat{\beta}_{\cdot} + \hat{\beta}_1(X - x_1)_+ + \hat{\beta}_2(x_1 - X)_+ \qquad (۱۳-۲)$$

$$\hat{f}_{22}(X) = \hat{\beta}_{\cdot} + \hat{\beta}_3(X - x_2)_+ + \hat{\beta}_4(x_2 - X)_+ \qquad (۱۴-۲)$$

$$\hat{f}_{2n}(X) = \hat{\beta}_{\cdot} + \hat{\beta}_{2n-1}(X - x_n)_+ + \hat{\beta}_{2n}(x_n - X)_+ \qquad (۱۵-۲)$$

فرض کنید مدل $\hat{f}_{22}$ دارای کمترین مقدار SSE است لذا دومین مدل عبارت است از :

$$\hat{f}_2(X) = \hat{\beta}_{\cdot} + \hat{\beta}_3(X - x_2)_+ + \hat{\beta}_4(x_2 - X)_+ \qquad (۱۶-۲)$$

که به صورت مختصرتر می‌توان آن را به صورت زیر نمایش داد و آن را مدل ۲ نامید.

$$\hat{f}_2(X) = \hat{\beta}_{\cdot} + \hat{\beta}_3 h_3(X)_+ + \hat{\beta}_4 h_4(X)_+. \qquad (۱۷-۲)$$

اکنون مدل ۳ را با افزودن سایر توابع پایه به مدل ۲ که افزودن آن توابع، بزرگترین کاهش را در مقدار مجموع توان دوم‌های خطا دارد، ساخته می‌شود. بنابراین مدل ۳ از بین مدل‌های زیر انتخاب خواهد شد.

$$\hat{f}_{31}(X) = \hat{\beta}_{\cdot} + \hat{\beta}_1 * h_3(X) + \hat{\beta}_2 * h_4(X) + \hat{\beta}_3 * h_1(X) + \hat{\beta}_4 * h_2(X). \qquad (۱۸-۲)$$

$$\hat{f}_{31}(X) = \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots + \hat{\beta}_3 * h_5(X) + \hat{\beta}_4 * h_6(X), \qquad (۱۹-۲)$$

$$\hat{f}_{3(n-1)}(X) = \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots + \hat{\beta}_3 * h_{2n-1}(X) + \hat{\beta}_4 * h_{2n}(X) \qquad (۲۰-۲)$$

همانطور که گفته شد مدلی از بین مدل‌های فوق انتخاب می‌شود که بیشترین کاهش در معادله (۲-۲۱) داشته باشد. شایان ذکر است که در هر بار ساختن مدل جدید ضرایب β موجود در مدل

جدید توسط مینیمم کردن مجموع توان‌های دوم مجددا برآورد می‌شوند. این فرآیند تا زمانی که k تابع پایه به مدل اضافه شود، ادامه می‌یابد. توجه شود $k \leq 2n$ یا توسط کاربر مشخص می‌شود.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \qquad (۲۱-۲)$$

➢ مرحله پس‌رو

در پایان مرحله پیش‌رو، مدل تولید شده مدلی است که اگر چه برای داده‌های مدل‌ساز ممکن است مناسب باشد اما موجب بیش‌برازش داده‌های آزمون است. برای رفع این مشکل، روش مارس وارد مرحلهٔ دوم یعنی "حذف توابع پایه از مدل" می‌شود. بنابراین اکنون به حذف ممکن هریک از توابع مدل f_k که دارای یک عرض از مبدأ و دو (k-1)۲ تابع پایه است، توجه می‌شود. توابه پایه‌ای از مدل حذف خواهند شد که حذف آن‌ها کمترین افزایش را در مقدار مجموع توان‌های دوم خطا داشته باشد. این فرآیند تا زمانی که همهٔ توابع از مدل، به جز عرض از مبدأ حذف شوند، ادامه می‌یابد.

وقتی که فرآیند حذف توابع پایه به‌طور کامل انجام شد، تعداد 2k-2 مدل ساخته می‌شود (در هر مرحله یک مدل) که هر یک از آن‌ها، نامزد برای مدل نهایی هستند. برای هرکدام از این مدل‌ها، معیار اعتبار سنجی متقابل تعمیم یافته (GCV)، محاسبه می‌شود. اگر GCV_l مقدار شاخص GCV برای l–امین مدل در فرآیند پس‌رو به ازای l=1,….,2K-2 باشد داریم[۴۳].

$$GCV_l = \frac{SSE_l}{1-(\frac{(vm_l+1)}{n})} \qquad (۲۲-۲)$$

که در آن m_l و SSE_l به ترتیب تعداد توابع پایه و مجموع توان‌های دوم خطا در مدل lام و v جریمهٔ هر تابع پایه است. در واقع می‌توان گفت که v به عنوان یک پارامتر هموارسازی که کاربر تعریف می‌کند، عمل می‌نماید. در حقیقت v مبادلهٔ بین مدل‌های پیچیده و ساده را کنترل می‌نماید.

مدلی که در بین 2k-2 مدل دارای کمترین مقدار GCV به عنوان مدل نهایی و برآورد مارس انتخاب می‌شود.

۲-۶-۲-MARS دو متغیره

اگر بیش از یک متغیر پیش بین وجود داشته باشد، مارس این توانایی را دارد که جز به توابع پایه، حاصلضرب آن‌ها یعنی اثر متقابل‌شان را در مدل لحاظ کند[۴۶].

$$h_m(X) = h_i(X)(X_j - t)_+, h_{m+1}(X) = h_i(X)(t - X_j)_+ \qquad (۲۳-۲)$$

که در آن $\{(X_j - t)_+, (t - X_j)_+\}$ جفت‌های منعکس شده از مجموعه C به ازای J=1,……,P می‌باشند و $1 \le i \le m-1$ $h_i(X)$ توابع پایه‌ای هستند که از قبل در مدل وجود داشته‌اند. انتخاب جفت توابع پایه $\{ h_{m+1}(X), h_m(X),\}$ از طریق جستجوی حاصلضرب‌هایی از توابع پایه $h_i(X)$ از مدل و جفت‌های منعکس شده در مجموعه C انجام می‌شود، به طوری‌که $h_i(X)$ و جفت‌های منعکس شده هیچ‌یک در ساختارشان متغیر یکسان X_j را نداشته و وقتی به مدل اضافه می‌شوند، بیشترین کاهش را در مجموعه توان‌های دوم خطا داشته باشند[۵۶].

۲-۶-۳-گام‌های ساخت MARS

مارس روش رگرسیون غیر خطی و ناپارامتریک است که پاسخ‌های غیر خطی را بین ورودی‌ها و خروجی‌های یک سیستم به‌وسیلهٔ مجموعه‌ای از قطعه‌های خطی تکه‌ای (کثیرالجمله‌های چند قطعه-ای) با گرادیان‌های متفاوت مدل‌سازی می‌کند. فرضی ثابت در رابطه تابعی اساسی بین متغیرهای ورودی و خروجی لازم نیست. نقاط انتهایی این قطعه‌ها گره نامیده می‌شوند. گره انتهای یک ناحیه از داده‌ها و ابتدای ناحیه‌ای دیگر از داده‌ها را مشخص می‌کند. منحنی‌های تکه‌ای منتج شناخته شده به‌

عنوان توابع پایه‌ای انعطاف پذیری بیشتری به مدل می‌دهند و انحناها، آستانه‌ها و دیگر انحراف‌های حاصل از توابع خطی را در نظر می‌گیرند. مارس توابع پایه‌ای را با جستجو به روش مرحله‌ای ایجاد می‌کند. الگوریتم رگرسیون انطباقی برای انتخاب موقعیت گره به کار می‌رود. مدل‌های مارس به روش‌های دو مرحله‌ای ایجاد می‌شوند، مرحله مقدم توابع را جمع می‌بندد و گره‌های احتمالی را برای بهبود عملکرد می‌یابد که به مدلی با برازش کامل می‌انجامد. مرحله دوم دربرگیرنده حذف کمترین جمله‌های حقیقی است. مدل‌سازی مارس فرآیند پدیدآمده از داده‌هاست برای برازش مدل ابتدا روش انتخابی مقدم روی داده‌های آموزشی انجام می‌شود و سپس مدل تنها با جفت پایه ایجاد می‌شود که بزرگترین کاهش را در خطای آموزشی به وجود می‌آورد که با توجه به مدل جاری با توابع پایه M جفت بعدی به مدل اضافه می‌شود. برای کاهش تعداد جمله‌ها از رشتهٔ حذفی مؤخر پیروی می‌شود. هدف این روش پیدا کردن مدل نزدیک به حد مطلوب با حذف متغیرهای غیر اصلی است، مسیر این روش به این صورت است که از مدل با حذف توابع پایه همراه با کمترین سهم نسبت به مدل می‌زند تا اینکه بهترین مدل را بیابد، بنابراین توابع پایه حفظ شده در مدل بهینهٔ نهایی از مجموعه تمام توابع پایه انتخاب می‌شود که در مرحلهٔ انتخابی مقدم مورد استفاده قرار می‌گیرد. زیر مجموعه‌های مدل با استفاده از روش ارزیابی متقابل تعمیم یافته (GCV)که به لحاظ محاسباتی کم هزینه است محاسبه می‌شوند. GCV نه تنها تعداد توابع پایه‌ای مدل بلکه تعداد گره‌ها را نیز برآورد می‌کند[۴۷].

۲-۶-۴-مزیت‌های MARS

مارس به لحاظ محاسباتی در پیدا کردن مدل بهینه کارآمدتر است چرا که در اصل با برازش مدل‌های قابل انعطافی می‌سازد و مدل را با تقسیم به شیب‌های جداگانه در فاصله‌های مشخص متغیرهای ورودی انتخاب می‌کند. متغیرها و محل‌های گره برای هر متغیر از طریق روش جستجو

سریع اما متمرکز تعیین می‌شود همچنین در مرحله پس‌رو و پیش‌رو تضمین می‌کند که مدل بهینه می‌تواند تشخیص داده شود.

۲-۷-ارزیابی قدرت پیش بینی مدل

۲-۷-۱-استفاده از پارامترهای آماری

برای اطمینان از اینکه مدل به‌دست‌آمده، مدل مناسبی است که توانایی پیشگویی نمونه‌های مختلفی از یک جمعیت را داراست، باید مدل را ارزیابی کرد. این ارزیابی از طریق شاخص‌های کمی است که به‌وسیلهٔ آن‌ها صحّت نتایج ارائه‌شده توسط مدل موردسنجش قرار می‌گیرند. برخی از این شاخص‌های کمی عبارت‌اند از:

ضریب همبستگی[1]:ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر، محاسبهٔ آماری ضریب همبستگی آن‌هاست. ضریب همبستگی دو متغیر x,y با رابطهٔ زیر تعریف می‌شود. مقدار این آماره بین -۱ و ۱ است مقدار بزرگ‌تر آن نشان‌دهندهٔ این است که ارتباط خطی بیشتری میان متغیر وابسته و متغیرهای مستقل وجود دارد.

$$R = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}} \qquad (۲۴-۲)$$

ضریب تعیین[2]:شاخصی برای بیان دقت خط رگرسیون برآورد شده است و مقدار آن بین صفر و یک تغییر می‌کند و مبین واریانس مشترک بین متغیر وابسته و مستقل است یعنی نسبتی از واریانس متغیر وابسته که توسط متغیرهای مستقل به‌حساب آمده معین می‌شود و توسط رابطهٔ زیر تعریف می‌شود.

$$R^2 = \frac{SSR}{SST} \qquad (۲۵-۲)$$

[1] Correlation coefficient
[2] Coefficient of determination

که SSR در این رابطه بیانگر مجموع مربعات انحراف مقادیر پیش‌بینی‌شدهٔ متغیر وابسته از میانگین مقادیر آن است.

$$SSR = \sum_{i=1}^{n}(\hat{y}i - \overline{y})^2 \qquad (۲۶-۲)$$

SST [۱] بیانگر مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از میانگین مقادیر آن است.

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad (۲۷-۲)$$

که در این روابط $\hat{y}i$ مقدار پیش‌بینی‌شدهٔ متغیر وابسته، y_iمقدار واقعی متغیر وابسته و $\bar{y}$ در هر رابطه، میانگین مقادیر متغیر وابسته است. بنابراین با توجه به روابط فوق می‌توان نوشت:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}i)^2}{\sum_{i=1}^{n}(yi - \overline{y})^2} \qquad (۲۸-۲)$$

طبق این رابطه اگر تمام مشاهدات بر روی خط برازش شده قرارگرفته باشند، یعنی به ازای تمام نقاط $y_i = \hat{y}i$ باشد، مقدار R^2 برابر یک می‌شود و هرگونه انحرافی از این حالت باعث می‌شود که مقدار R^2 از یک کوچک‌تر شود.

خطای استاندارد پیش‌بینی (SEP)[۲]:

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{n}} \qquad (۲۹-۲)$$

مجموع مربعات باقی‌ماندهٔ پیش‌بینی(PRESS)[۳]:

$$\mathrm{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}i)^2 \qquad (۳۰-۲)$$

[1] Sum square of total
[2] Standard error of prediction
[3] Predictive residual

میانگین مربع خطاها (MSE):[1]

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}\mathrm{i})^2}{n} \qquad (۳۱-۲)$$

میانگین خطای مطلق (MAE):[2]

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}\mathrm{i}|}{n} \qquad (۳۲-۲)$$

میانگین خطای نسبی (MRE):[3]

$$\mathrm{MRE} = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}\mathrm{i}|}{y_i}\right|}{n} \qquad (۳۳-۲)$$

خطای نسبی پیش‌بینی (REP):[4]

$$REP\% = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}\mathrm{i})^2}{n}} \qquad (۳۴-۲)$$

ضریب تعیین تعدیل‌یافته:[5]

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p-1} \times \frac{SSE}{SST} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad (۳۵-۲)$$

آمار F:[6] یا آزمون فیشر درواقع آزمون معنی‌دار بودن آماری در تحلیل رگرسیون ساده و چند متغیره است و برابر با نسبت میانگین رگرسیون (MSR)[7] به میانگین مربعات باقی‌مانده (MSE) است.

[1] Mean Square Error
[2] Mean Square Regression
[3] Mean relative error
[4] Relative error of Prediction Coefficient
[5] Adjusted determination coeffiction
[6] Statistic term
[7] Statistic term

$$F = \frac{MSR}{MSE} = \frac{SSR/P}{SSE/n-p-1} \quad (۳۶-۲)$$

FIT[1]: این پارامتر نیز به‌عنوان یک معیار برای مقایسه مدل‌های مختلف با تعداد متفاوت توصیف‌کننده به کار می‌رود.

$$FIT = \frac{R^2(n-p-1)}{(n+p^2)\times(1-R^2)} \quad (۳۷-۲)$$

در این روابط n تعداد ترکیبات مربوط به مدل و p تعداد توصیف‌کننده‌های مدل است، هرچه مقدارF و FIT بیشتر باشد(FIT به صورت مقایسه‌ای است و هر مدلی که FIT بالاتری داشته باشد مدل مناسب تری است) بهتر است.[۴۷،۴۸].

۲-۷-۲- استفاده از نمودار خطای باقی‌مانده

منظور از عبارت خطای باقی‌مانده، اختلاف بین مقادیر پیش‌بینی‌شده و مقادیر تجربی است. اگر پراکندگی مقادیر در دو طرف خط صفر یکسان باشد، این امر نشان‌دهندهٔ تصادفی بودن خطاست، ولی اگر عمدهٔ نقاط در این نمودار، در یک‌طرف خط صفر باشند، این بدان معناست که خطای جهت‌داری رخ‌داده است.

۲-۷-۳- استفاده از روش ارزیابی متقاطع

در این روش هم برای تعیین اعتبار مدل‌های پیش‌بینی‌کننده هر بار یک ترکیب از سری داده‌ها کنار گذاشته می‌شود و در شرایط بهینه به‌دست‌آمده برای سری ارزیابی یا سری آموزش با ترکیبات باقی‌مانده مدل‌سازی صورت می‌گیرد، سپس مدل به‌دست‌آمده برای پیش‌بینی فعالیت/خاصیت[2]

[1] Fitness function
[2] Activity/property

ترکیب کنار گذاشته‌شده به کار گرفته می‌شود و این فرآیند برای تمام اعضای سری داده‌ها تکرار می‌گردد.

۲-۷-۴- آزمون Y-تصادفی[1]

این تکنیک برای مطالعهٔ همبستگی تصادفی مدل طراحی‌شده است. وقتی در مدلی، همبستگی متغیرهای مستقل با متغیرهای پاسخ به‌صورت تصادفی باشد، دارای همبستگی تصادفی هستند. در این آزمون، مقادیر پاسخ در محدودهٔ پاسخ‌ها به‌صورت تصادفی تغییر داده می‌شود. سپس همبستگی متغیرهای مستقل با متغیرهای پاسخ با استفاده از یکی از شاخص‌های آماری موردبررسی قرار می‌گیرد تفاوت قابل ملاحظه بین مقادیر ضریب تعیین مدل‌های اصلی و مدل‌های جدید بیانگر عدم وجود ارتباط تصادفی در مدل‌ها است. این فرآیند را چندین بار باید تکرار کرد [۴۸].

[1] Y-randomization test

فصل سوم

مطالعهٔ ارتباط کمی ساختار-خاصیت برروی تعدادی از ترکیبات هتروسیکل آروماتیک سولفوردار چند حلقه‌ای با استفاده از ANN و RF و MAR

۳-۱-مدل‌سازی جهت پیش‌بینی‌اندیس بازداری ترکیبات سولفور آروماتیک چند حلقه‌ای سولفوردار

در این فصل به‌منظور مطالعه ارتباط کمی ساختار-خاصیت بروی ترکیبات سولفور چند حلقه‌ای، از جنگل‌های تصادفی، شبکه عصبی مصنوعی و رگرسیون اسپلاین تطبیقی چند متغیره به عنوان روش‌های مدل‌سازی برای پیش‌بینی اندیس بازداری این ترکیبات استفاده‌شده و توانایی مدل‌های به‌دست‌آمده از هر سه روش مورد ارزیابی قرار گرفته است. به‌طورکلی بخش تجربی این تحقیق شامل معرفی سری داده‌ها، بهینه‌سازی ساختار مولکول‌ها، محاسبهٔ توصیف‌کننده‌های مولکولی، انتخاب توصیف‌کننده‌های مولکولی مناسب با استفاده از روش رگرسیون گام‌به‌گام و همچنین مدل‌سازی و ارزیابی مدل‌های برتر است.

۳-۲-نرم افزارهای مورد استفاده

در این تحقیق از بسته نرم‌افزاری متنوعی در مراحل مدل‌سازی استفاده شده است که هریک به اختصار توضیح داده می‌شوند.

۳-۲-۱-نرم‌افزار Hyperchem

این نرم‌افزار جهت رسم و بهینه‌سازی اولیه ساختار ترکیبات شیمیایی مورد استفاده قرار می‌گیرد. در این نرم‌افزار، ساختار ترکیبات می‌تواند با چهار روش آغازین[1]، نیمه تجربی[2]، مکانیک مولکولی[3] و تابع دانسیته[4] بهینه‌سازی شود. در این پروژه، بهینه‌سازی تا زمانی ادامه یافت که گرادیان انرژی به ۰/۰۰۱کیلوکالری بر مول رسید. به کمک این برنامه می‌توان طول پیوند، زاویه پیوندی و زوایای پیچشی را در مولکول تعیین کرد. داده‌های حاصل از این نرم‌افزار را می‌توان به‌عنوان ورودی به سایر

[1] Ab initio methods
[2] Semi-empirical methods
[3] Molecular mechanicse
[4] Density functional theory(DFT)

نرم‌افزارها معرفی نمود. همچنین به کمک این نرم‌افزار می‌توان تعدادی از توصیف‌کننده‌ها ازجمله حجم مولی و قطبش‌پذیری را محاسبه کرد.

۳-۲-۲- نرم‌افزار Dragon

از این نرم‌افزار جهت استخراج توصیف‌کننده‌های مولکولی استفاده می‌شود. نرم‌افزار Dragon توسط گروه QSPR و کمومتریکس دانشگاه میلانو طراحی شده است. این نرم‌افزار قادر به محاسبه ۳۲۲۴ توصیف‌کننده می‌باشد. علاوه بر محاسبه ساده‌ترین نوع اتم‌ها و گروه‌های عاملی و شمارش اجزا می‌توان تعداد زیادی از توصیف‌کننده‌های توپولوژیکی و هندسی را نیز توسط این نرم‌افزار محاسبه نمود. برای اجرای این نرم‌افزار به فایل‌های ساختار مولکولی ایجاد شده به‌وسیله نرم‌افزارهای مدل‌سازی مولکولی، مانندHyperchem نیاز می‌باشد. اکثر فایل‌های مولکولی با فرمت‌های معمول در این نرم‌افزار قابل قبول هستند. برای استفاده بهتر از محاسبات این نرم‌افزار باید بهینه‌سازی سه بعدی ساختارها با هیدروژن‌هایشان به‌کار رود. لازم به ذکر است که Dragon به‌عنوان یک نرم‌افزار QSPR/QSAR طراحی نشده است یعنی فقط توصیف‌کننده‌های مولکولی را نمایش می‌دهد و هیچ گونه آنالیز QSPR/QSAR را انجام نمی‌دهد. با این حال Dragon یک فایل خروجی کامل را که به سادگی توسط هر نرم‌افزار آنالیز همبستگی قابل کاربرد است فراهم می‌سازد تا به‌کمک آن بتوان مطالعات QSPR/QSAR را انجام داد و همبستگی‌های خاصی را می‌توان توسط آن حذف کرد.

۳-۲-۳-نرم‌افزارSPSS

SPSS یک نرم‌افزار آماری است که توانایی تحلیل کلی اطلاعات را دارا بوده و عمومی‌ترین نرم‌افزار آماری می‌باشد. اولین نسخه آن در سال ۱۹۷۰ توسط جمعی از فارغ التحصیلان دانشگاه استانفورد آمریکا ارائه شده است. پس از ۲۸ جولای ۲۰۰۹ که شرکت سازنده این نرم‌افزار توسط

IBM خریداری شد این نرم‌افزار با نامPASW[1] منتشر گردید. SPSS ازجمله نرم‌افزارهایی است که برای تحلیل‌های آماری در علوم اجتماعی، به‌صورت بسیار گسترده‌ای استفاده می‌شود. افزون بر تحلیل‌های آماری، مدیریت داده‌ها و مستندسازی داده‌ها نیز از ویژگی‌های این نرم‌افزار هستند. ازجمله توانایی‌های نرم‌افزار SPSS عبارتند از:

*تهیه خلاصه‌های آماری مانند گراف‌ها، جداول، آماره‌ها

*محاسبه انواع توابع ریاضی مانند قدر مطلق، تابع علامت، لگاریتم، توابع مثلثاتی

*تهیه انواع جداول سفارشی مانند جداول فراوانی، فراوانی تجمعی، درصد فراوانی

*انواع توزیع‌های آماری شامل توزیع‌های گسسته و پیوسته

*تهیه انواع طرح‌های آماری

*انجام آنالیز واریانس یک طرفه، دو طرفه، چند طرفه و آنالیز کوواریانس

*محاسبه میانگین ساده برای داده‌ها

*تکنیک‌های تجزیه و تحلیل سری‌های زمانی

*محاسبه انواع آماره‌های توصیفی

*ایجاد داده‌های تصادفی و پیوسته

*انجام رگرسیون تک متغیره و چند متغیره

۳-۲-۴- نرم‌افزار MATLAB[2]

نرم‌افزار MATLAB برنامه‌ای کامپیوتری است که برای کسانی که با محاسبات عددی سر و کار دارند تهیه شده است. هدف اولیه این نرم‌افزار قادر ساختن مهندسین و دانشمندان به حل مسائل، شامل عملیات ماتریسی بدون نیاز به نوشتن برنامه در زبان‌های فورترن و ..... بود. کاربردهای نوعی آن شامل ریاضیات، الگوریتم محاسباتی، شبیه سازی، مدل‌سازی و آنالیز اطلاعات می‌باشد. یکی از

[1] Predictive Analytics Software
[2] Matrix Laboratory

کاربردهای مهم آن‌که مورد استقبال شیمیدانان قرارگرفته امکان استفاده از آن برای آنالیز آماری و مدل‌سازی داده‌های شیمیایی است. مدل‌سازی به‌روش شبکه‌ی عصبی مصنوعی ازجمله کاربردهای مهم دیگر آن است که در تحقیق حاضر مورد استفاده قرارگرفته است. اطلاعات شیمیایی ترکیبات به‌صورت ماتریس به‌عنوان ورودی به نرم‌افزار ارائه می‌شوند. سپس در محیط MATLAB با ورژن ۲۰۱۵ و با استفاده از آرایه‌ها و فرامین موجود امکان مدل‌سازی غیرخطی فعالیّت شیمیایی ترکیبات امکان‌پذیرمی‌گردد. سایر روش‌های کمومتریکس شامل آنالیز (رگرسیون) اجزای اصلی، روش کمترین مربعات جزئی، الگوریتم ژنتیک و غیره نیز توسط این نرم‌افزار قابل اجرا است.

برای رسم نمودار سه بعدی در این تحقیق از نرم‌افزار sigmaplot با ورژن ۱۲ استفاده شده است.

۳-۲-۵- نرم‌افزار R

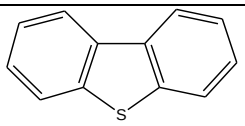
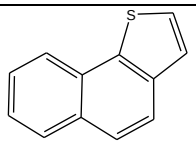
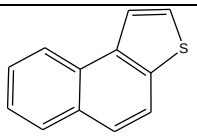
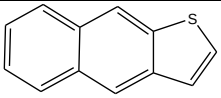
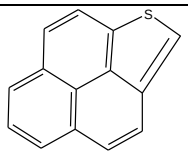
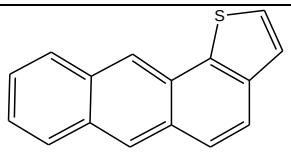
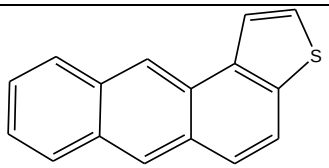
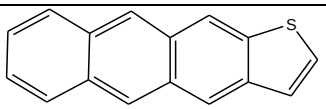
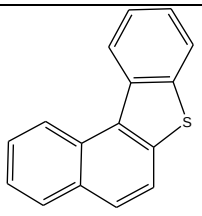
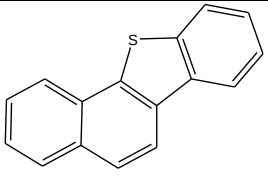
R یک زبان برنامه‌نویسی و محیط نرم‌افزاری برای محاسبات آماری و علم داده‌هاست و اکثر زمینه‌های آمار کاربردی مانند تحلیل سری‌های زمانی، رگرسیون خطی و غیر خطی، آمون فرض‌های کلاسیک، کدگذاری، خوشه بندی و ....را پوشش می‌دهد. R یک نرم افزار متن باز[1] و رایگان است. R نرم‌افزار قدرتمندی برای ایجاد اشکال گرافیکی و نمودارهاست[۴۹].

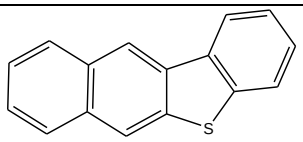
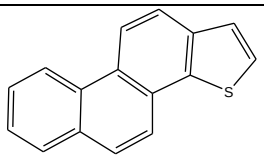
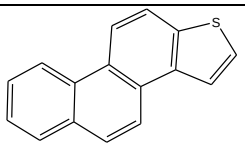
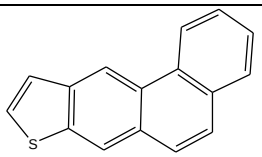
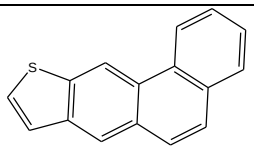
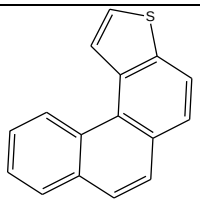
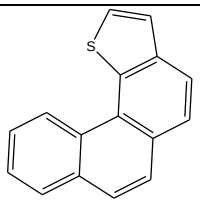
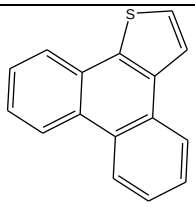
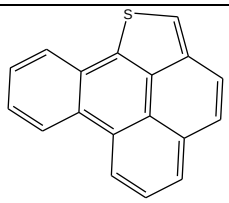
۳-۳- سری داده‌ها

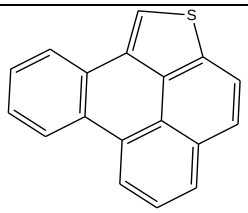
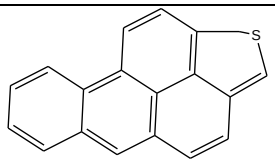
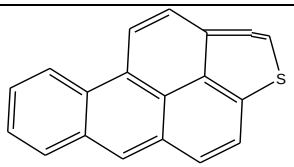
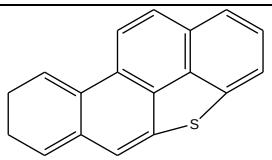
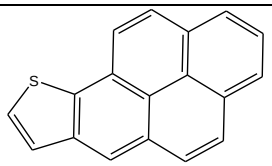
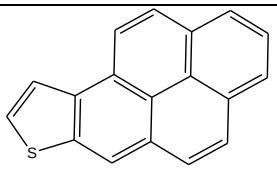
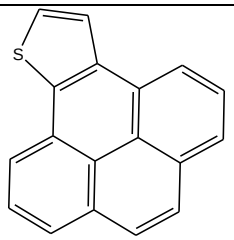
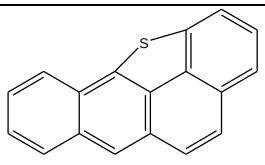
سری داده‌ها شامل نتایج اندیس بازداری ۱۵۲ ترکیب آروماتیک سولفور چند حلقه‌ای می‌باشد که توسط والتر ویلسون [2]و همکارانش گزارش‌شده است[۵۰،۵۱]. نام، ساختار و اندیس بازداری (RI) تجربی این ترکیبات در جدول (۳-۱) نشان داده‌شده است. شایان ذکر است که اندیس بازداری به عنوان خاصیت ترکیبات ذکر شده، متغیر وابسته محسوب می‌شود.

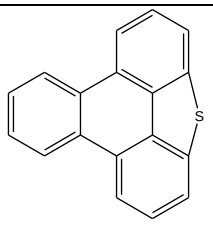
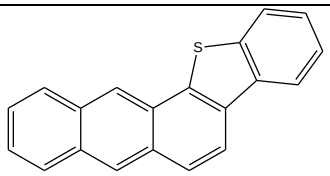
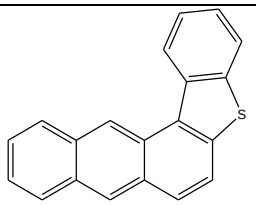
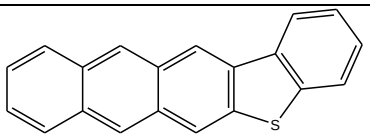
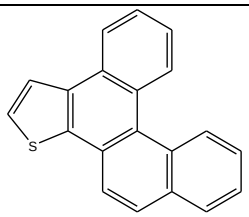
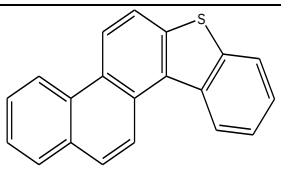
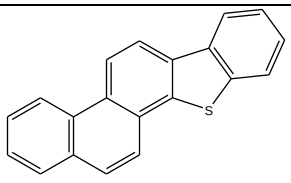
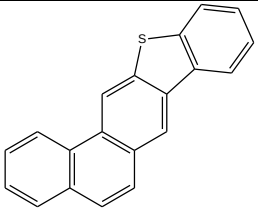
[1] Open source
[2] Walter B . wilson

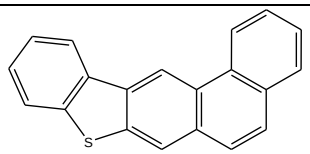
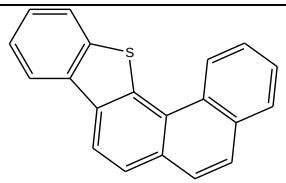
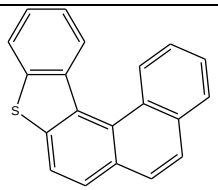
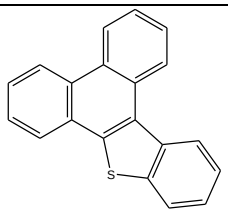
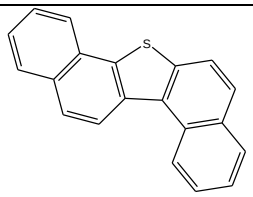
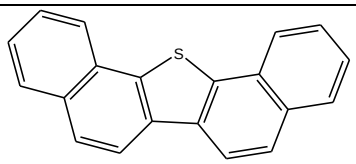
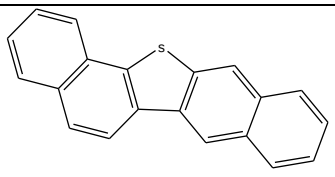
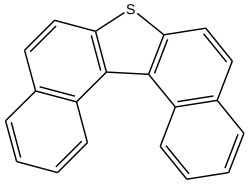
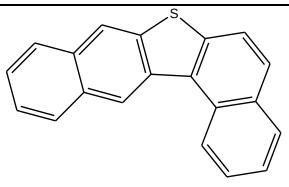
جدول(۳-۱)-نام و اندیس بازداری سری داده‌ها

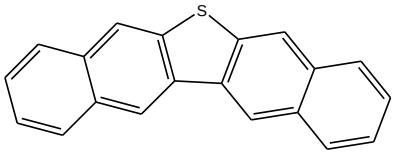
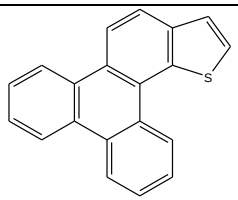
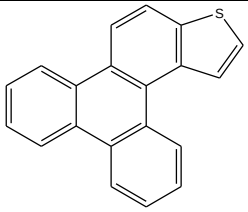
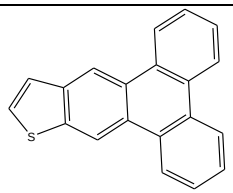
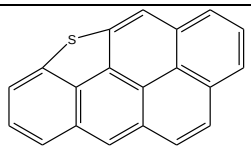
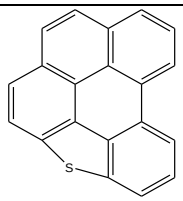
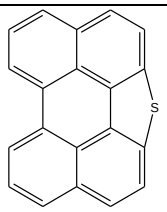
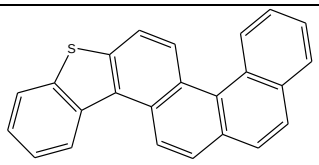
شماره	نام	ساختار	RI
۱	Dibenzothiophene		۳/۲۸
۲	Naphtho[1,2-b]thiophene		۲/۸۴
۳	Naphtho[2,1-b]thiophene		۲/۷۸
۴	Naphtho[2,3-b]thiophene		۲/۹۱
۵	Phenaleno[1,9-bc]thiophene		۳/۴۷
۶	Anthra[1,2-b]thiophene		۳/۹۲
۷	Anthra[2,1-b]thiophene		۳/۷۶
۸	Anthra[2,3-b]thiophene		۳/۹۰
۹	Benzo[b]54ikiped[1,2-b]thiophene		۴/۰۰
۱۰	Benzo[b]54ikiped[2,1-b]thiophene		۴/۱۳

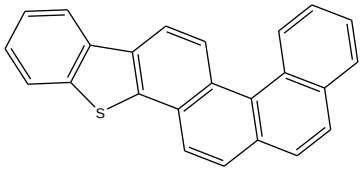
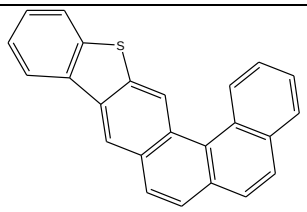
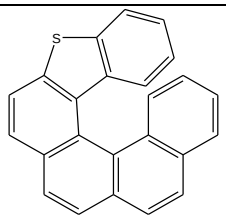
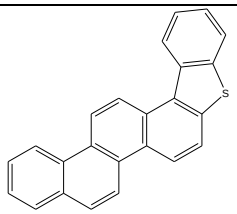
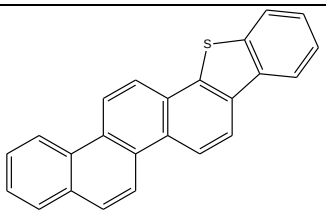
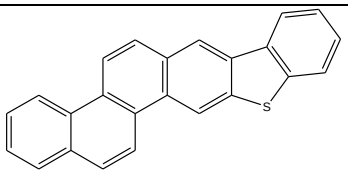
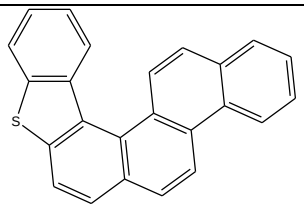
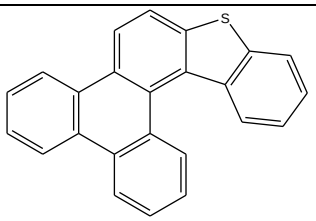
شماره	نام	ساختار	RI
۱۱	Benzo[b]55ikiped[2,1-b]thiophene		۴/۰۵
۱۲	Phenanthro[1,2-b]thiophene		۳/۹۰
۱۳	Phenanthro[2,1-b]thiophene		۳/۷۰
۱۴	Phenanthro[2, 3-b]thiophene		۳/۷۰
۱۵	Phenanthro[3,2-b]thiophene		۳/۶۶
۱۶	Phenanthro[3,4-b]thiophene		۳/۷۳
۱۷	Phenanthro[4,3-b]thiophene		۳/۸۶
۱۸	Phenanthro[9,10-b]thiophene		۳/۸۸
۱۹	Benzo[1,2]phenaleno[3,4-bc]thiophene		۴/۶۳

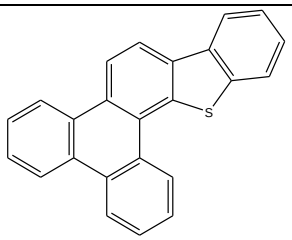
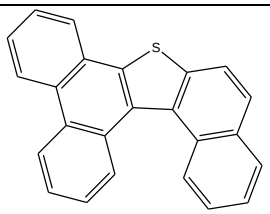
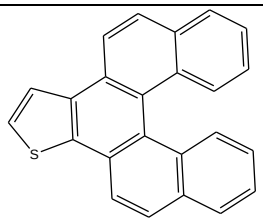
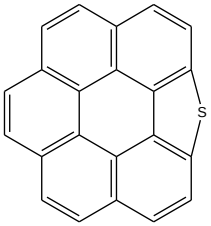
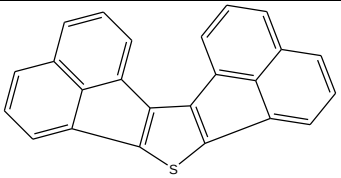
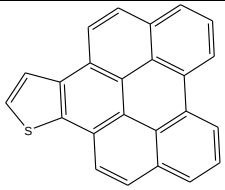
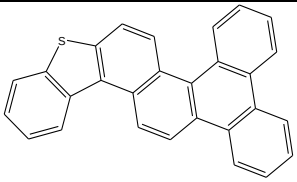
شماره	نام	ساختار	RI
۲۰	Benzo[1,2]phenaleno[4,3-bc]thiophene		۵/۰۲
۲۱	Benzo[4,5]phenaleno[1,9-bc]thiophene		۴/۴۷
۲۲	Benzo[4,5]phenaleno[9,1-bc]thiophene		۴/۴۷
۲۳	Chryseno[4,5-bcd]thiophene		۴/۷۶
۲۴	Pyreno[1,2-b]thiophene		۴/۵۹
۲۵	Pyreno[2,1-b]thiophene		۴/۴۳
۲۶	Pyreno[4,5-b]thiophene		۴/۵۴
۲۷	Tetrapheno[1,12-bcd]thiophene		۴/۹۷

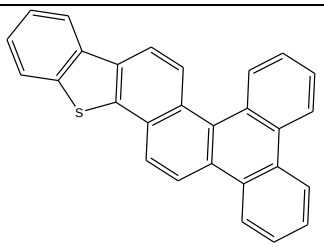
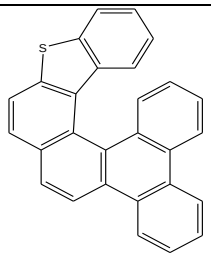
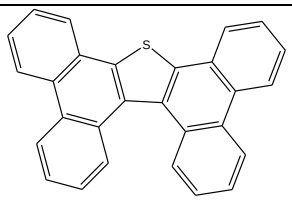
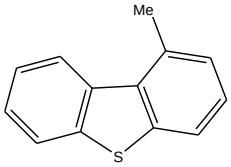
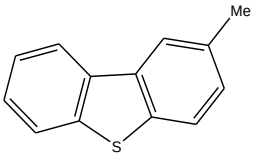
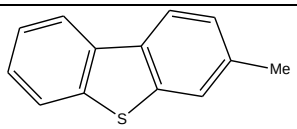
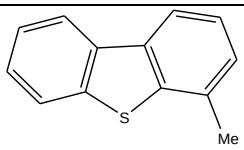
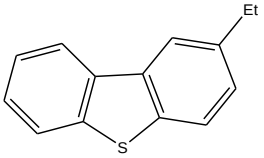
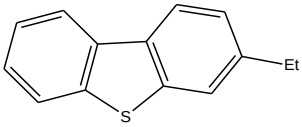
شماره	نام	ساختار	RI
۲۸	Triphenyleno[4,5-bcd]thiophene		۴/۶۰
۲۹	Anthra[1,2-b]benzo[d]thiophene		۵/۳۷
۳۰	Anthra[2,1-b]benzo[d]thiophene		۴/۹۹
۳۱	Anthra[2,3-b]benzo[d]thiophene		۵/۰۶
۳۲	Benzo[3,4]phenathro[1,2-b]thiophene		۴/۸۹
۳۳	Benzo[b]phenathro[1,2-b]thiophene		۵/۲۲
۳۴	Benzo[b]phenathro[2,1-b]thiophene		۵/۳۶
۳۵	Benzo[b]phenathro[2,3-b]thiophene		۴/۵۱

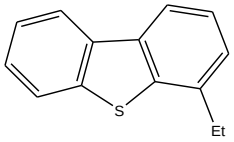
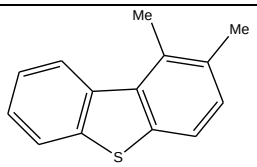
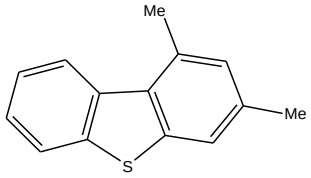
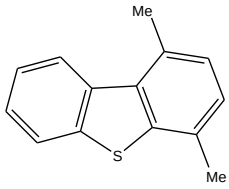
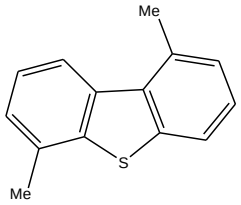
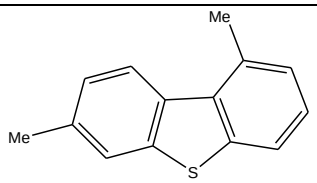
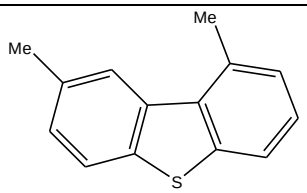
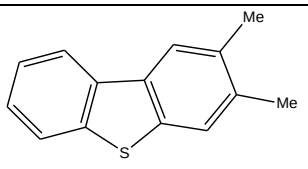
شماره	نام	ساختار	RI
۳۶	Benzo[b]phenathro[3,2-b]thiophene		۴/۹۱
۳۷	Benzo[b]phenathro[3,4-b]thiophene		۵/۳۲
۳۸	Benzo[b]phenathro[4,3-b]thiophene		۴/۹۱
۳۹	Benzo[b]phenathro[9,10-b]thiophene		۵/۳۰
۴۰	Dinaphtho[1,2:b-1,2-d]thiophene		۵/۴۰
۴۱	Dinaphtho[1,2:b-2,1-d]thiophene		۵/۶۶
۴۲	Dinaphtho[1,2:b-2,3-d]thiophene		۵/۴۷
۴۳	Dinaphtho[2,1:b-1,2-d]thiophene		۴/۹۷
۴۴	Dinaphtho[2,1:b-2,3-d]thiophene		۴/۹۹

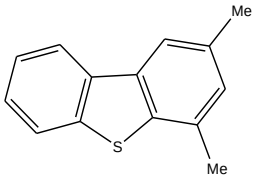
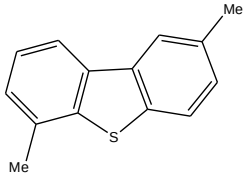
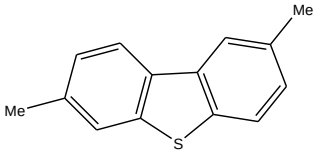
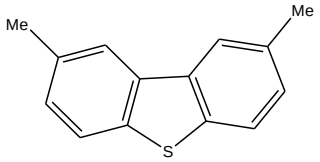
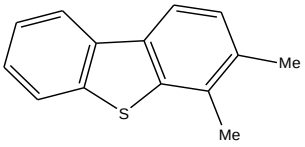
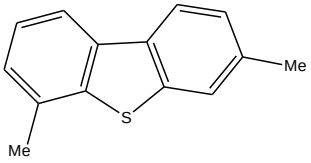
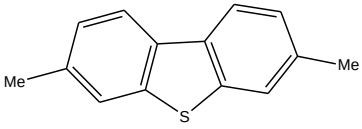
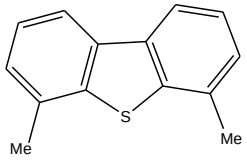
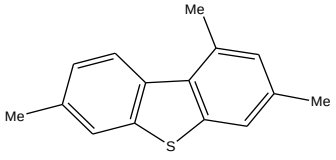
شماره	نام	ساختار	RI
۴۵	Dinaphtho[2,3-b-2,3-d]thiophene		۵/۱۹
۴۶	Triphenyleno[1,2-b]thiophene		۴/۷۶
۴۷	Triphenyleno[2,1-b]thiophene		۴/۶۴
۴۸	Triphenyleno[2,3-b]thiophene		۴/۵۱
۴۹	Benzo[10,11]chryseno[4,5-bcd]thiophene		۵/۸۳
۵۰	Benzo[4,5]triphenyleno[1,12-bcd]thiophene		۵/۵۷
۵۱	Peryleno[1,12-bcd]thiophene		۵/۰۵
۵۲	Benzo[b]benzo[5,6]phenathro[1,2-d]thiophene		۶/۰۵

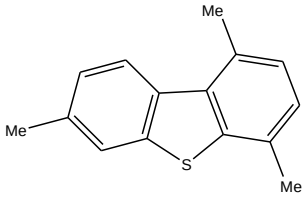
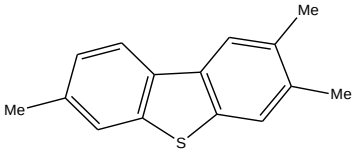
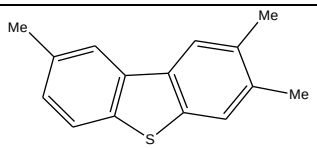
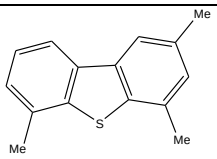
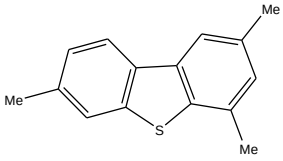
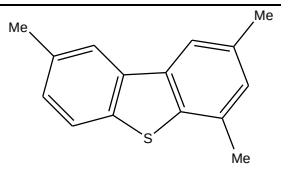
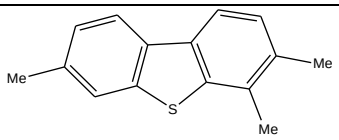
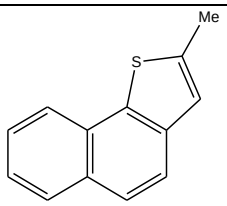
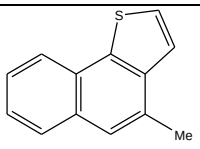
شماره	نام	ساختار	RI
۵۳	Benzo[b]benzo[5,6]phenathro[2,1-d]thiophene		۶/۳۳
۵۴	Benzo[b]benzo[5,6]phenathro[2,3-d]thiophene		۵/۸۲
۵۵	Benzo[b]benzo[5,6]phenathro[4,3-d]thiophene		۵/۱۴
۵۶	Benzo[b]chryseno[1,2-d]thiophene		۶/۱۶
۵۷	Benzo[b]chryseno[2,1-d]thiophene		۵/۷۰
۵۸	Benzo[b]chryseno[2,3-d]thiophene		۵/۰۰
۵۹	Benzo[b]chryseno[4,3-d]thiophene		۵/۸۹
۶۰	Benzo[b]triphenyleno[1,2-d]thiophene		۵/۷۸

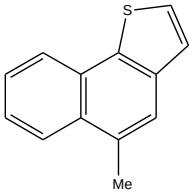
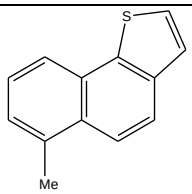
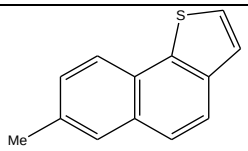
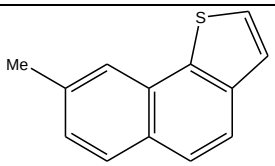
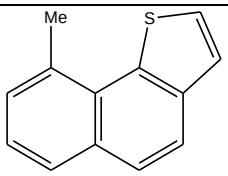
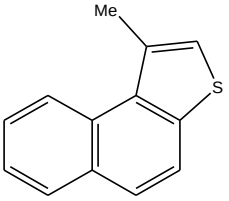
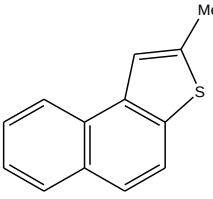
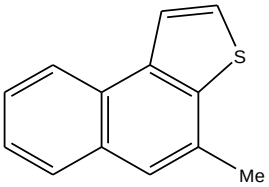
شماره	نام	ساختار	RI
۶۱	Benzo[b]triphenyleno[2,1-d]thiophene		۶/۲۷
۶۲	Naphtho[2,1-d]phenanthro[9,10-d]thiophene		۶/۱۹
۶۳	Naphtho[2,1:3,4-d]phenanthro[1,2-b]thiophene		۵/۵۵
۶۴	Benzo[6,7]peryleno[1,12-bcd]thiophene		۶/۴۱
۶۵	Diacenaphtho[1,2-b:1,2-d]thiophene		۶/۲۵
۶۶	N[218:345]PY12T		۵/۴۳
۶۷	Benzo[b]benzo[5,6]chryseno[1,2-d]thiophene		۶/۷۰

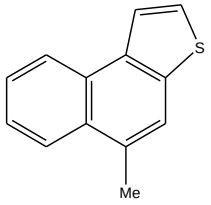
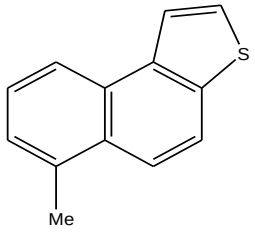
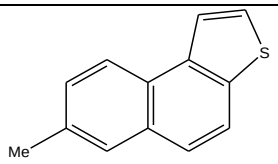
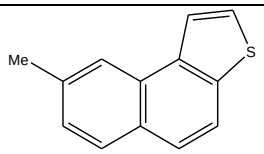
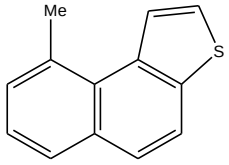
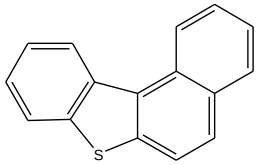
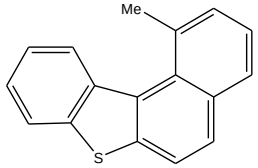
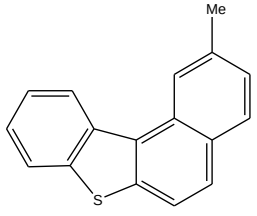
شماره	نام	ساختار	RI
۶۸	Benzo[b]benzo[5,6]chryseno[2,1-d]thiophene		۷/۰۳
۶۹	Benzo[b]benzo[5,6]chryseno[3,4-d]thiophene		۶/۴۳
۷۰	Diphenanthro[9,10-b:9,10-d]thiophene		۷/۳۷
۷۱	1-Methyl Dibenzothiophene		۳/۵۴
۷۲	2-Methyl Dibenzothiophene		۳/۵۷
۷۳	3-Methyl Dibenzothiophene		۳/۶۴
۷۴	4-Methyl Dibenzothiophene		۳/۶۸
۷۵	2-Ethyl Dibenzothiophene		۳/۹۸
۷۶	3-Ethyl Dibenzothiophene		۴/۰۰

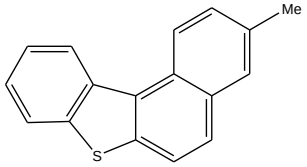
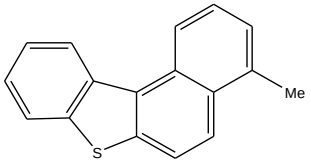
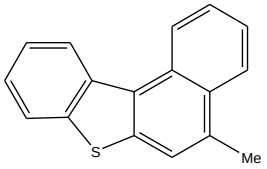
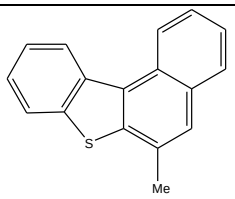
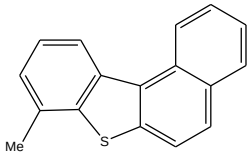
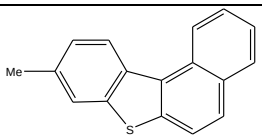
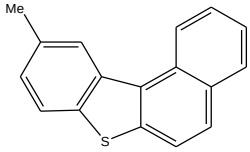
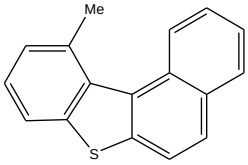
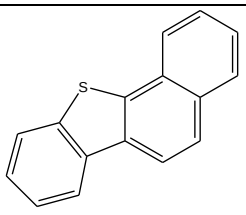
شماره	نام	ساختار	RI
۷۷	4-Ethyl Dibenzothiophene		۴/۰۸
۷۸	1,2-Dimethyl dibenzothiophene		۴/۰۱
۷۹	1,3-Dimethyl dibenzothiophene		۴/۱۱
۸۰	1,4-Dimethyl dibenzothiophene		۴/۱۶
۸۱	1,6-Dimethyl dibenzothiophene		۴/۱۶
۸۲	1,7-Dimethyl dibenzothiophene		۴/۱۰
۸۳	1,8-Methyl Dibenzothiophene		۴/۰۰
۸۴	2,3-Dimethyl dibenzothiophene		۴/۰۱

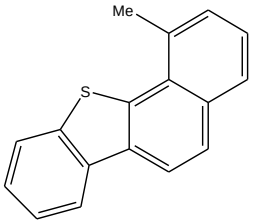
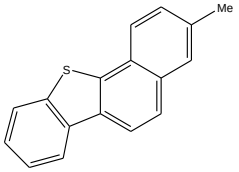
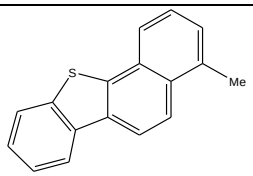
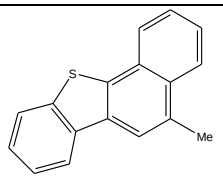
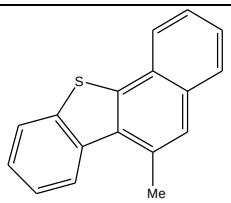
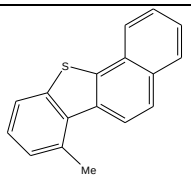
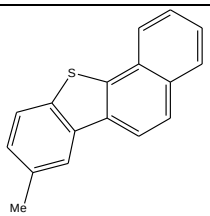
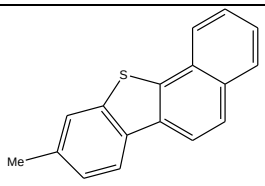
شماره	نام	ساختار	RI
۸۵	2,4-Dimethyl dibenzothiophene		۴/۱۷
۸۶	2,6-Dimethyl dibenzothiophene		۴/۱۷
۸۷	2,7-Dimethyl dibenzothiophene		۴/۱۲
۸۸	2,8-Dimethyl dibenzothiophene		۴/۰۵
۸۹	3,4-Methyl Dibenzothiophene		۴/۱۰
۹۰	3,6-Dimethyl dibenzothiophene		۴/۲۴
۹۱	3,7-Dimethyl dibenzothiophene		۴/۱۹
۹۲	4,6-Dimethyl dibenzothiophene		۴/۲۵
۹۳	1,3,7-Trimethyl dibenzothiophene		۴/۶۶

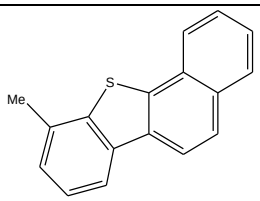
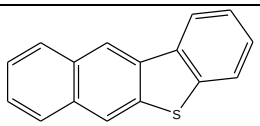
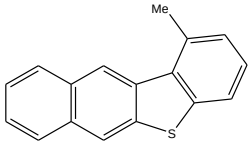
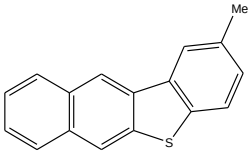
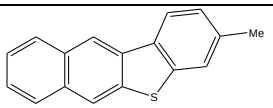
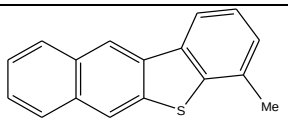
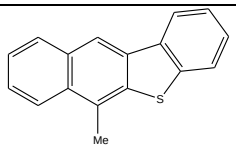
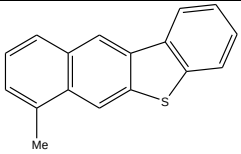
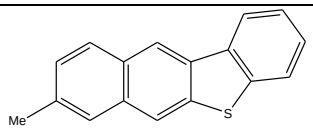
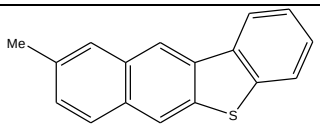
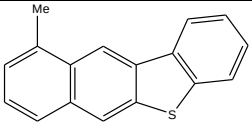
شماره	نام	ساختار	RI
۹۴	1,4,7-Trimethyl dibenzothiophene		۴/۷۱
۹۵	2,3,7-Trimethyl dibenzothiophene		۴/۵۶
۹۶	2,3,8-Trimethyl dibenzothiophene		۴/۴۹
۹۷	2,4,6-Trimethyl dibenzothiophene		۴/۷۵
۹۸	2,4,7-Trimethyl dibenzothiophene		۴/۷۳
۹۹	2,4,8-Trimethyl dibenzothiophene		۴/۶۵
۱۰۰	3,4,7-Trimethyl dibenzothiophene		۴/۶۶
۱۰۱	2-Methyl naphtha[1,2-b]thiophene		۳/۵۹
۱۰۲	4-Methyl naphtha[1,2-b]thiophene		۳/۴۸

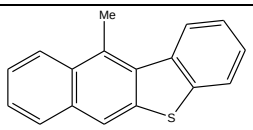
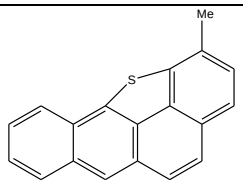
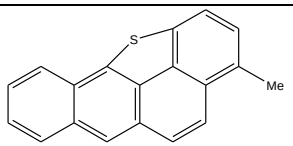
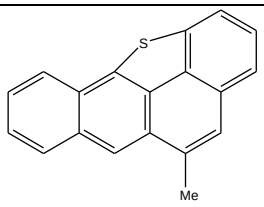
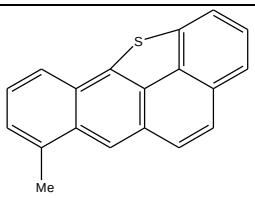
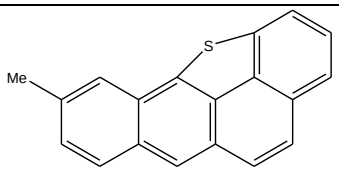
شماره	نام	ساختار	RI
۱۰۳	5-Methyl naphtha[1,2-b]thiophene		۳/۴۷
۱۰۴	6-Methyl naphtha[1,2-b]thiophene		۳/۴۵
۱۰۵	7-Methyl naphtha[1,2-b]thiophene		۳/۵۱
۱۰۶	8-Methyl naphtha[1,2-b]thiophene		۳/۴۲
۱۰۷	9-Methyl naphtha[1,2-b]thiophene		۴/۰۲
۱۰۸	1-Methyl naphtha[2,1-b]thiophene		۳/۳۰
۱۰۹	2-Methyl naphtha[2,1-b]thiophene		۳/۳۹
۱۱۰	4-Methyl naphtha[2,1-b]thiophene		۳/۳۸

شماره	نام	ساختار	RI
۱۱۱	5-Methyl naphtha[2,1-b]thiophene		۳/۳۳
۱۱۲	6-Methyl naphtha[2,1-b]thiophene		۳/۲۹
۱۱۳	7-Methyl naphtha[2,1-b]thiophene		۳/۳۴
۱۱۴	8-Methyl naphtha[2,1-b]thiophene		۳/۲۷
۱۱۵	9-Methyl naphtha[2,1-b]thiophene		۳/۳۱
۱۱۶	Benzo[b]naphtha[1,2-d]thiophene		۴/۰۰
۱۱۷	1-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۲۳
۱۱۸	2-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۴۴

شماره	نام	ساختار	RI
۱۱۹	3-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۵۴
۱۱۲۰	4-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۵۲
۱۲۱	5-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۵۶
۱۲۲	6-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۶۸
۱۲۳	8-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۷۰
۱۲۴	9-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۶۴
۱۲۵	10-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۴۹
۱۲۶	11-Methyl Benzo[b]naphtha[1,2-d]thiophene		۴/۳۷
۱۲۷	Benzo[b]naphtha[2,1-d]thiophene		۴/۱۳

شماره	نام	ساختار	RI
۱۲۸	1-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۷
۱۲۹	3-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۹
۱۳۰	4-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۵
۱۳۱	5-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۶۸
۱۳۲	6-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۴
۱۳۳	7-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۴
۱۳۴	8-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۷۵
۱۳۵	9-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۸۲

شماره	نام	ساختار	RI
۱۳۶	10-Methyl Benzo[b]naphtha[2,1-d]thiophene		۴/۸۴
۱۳۷	Benzo[b]naphtha[2,3-d]thiophene		۴/۰۵
۱۳۸	1-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۴۸
۱۳۹	2-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۷۵
۱۴۰	3-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۶۳
۱۴۱	4-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۷۱
۱۴۲	6-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۵۸
۱۴۳	7-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۴۹
۱۴۴	8-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۶۲
۱۴۵	9-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۵۹
۱۴۶	10-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۴۴

شماره	نام	ساختار	RI
۱۴۷	11-Methyl Benzo[b]naphtha[2,3-d]thiophene		۴/۵۱
۱۴۸	1-Methyl Tetrapheno[1,12-bcd]thiophene		۵/۷۹
۱۴۹	3-Methyl Tetrapheno[1,12-bcd]thiophene		۵/۶۷
۱۵۰	5-Methyl Tetrapheno[1,12-bcd]thiophene		۵/۶۰
۱۵۱	7-Methyl Tetrapheno[1,12-bcd]thiophene		۵/۵۸
۱۵۲	9-Methyl Tetrapheno[1,12-bcd]thiophene		۵/۲۸

۳-۴-بهینه‌سازی ساختار هندسی مولکول‌ها و محاسبهٔ توصیف‌کننده‌ها

ابتدا ساختار مولکول‌ها با استفاده از نرم‌افزار Hyperchem رسم و با روش نیمه تجربی AM1 برای رسیدن جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول بهینه شد. سپس به‌وسیلهٔ نرم‌افزار Dragon، ۳۲۲۴ توصیف‌کننده با استفاده از این ساختارهای بهینه‌شده محاسبه شد.

و روش رگرسیون گام‌به‌گام برای انتخاب بهترین توصیف‌کننده‌ها استفاده شد که به آن اشاره می‌شود.

۳-۴-۱-انتخاب بهترین توصیف‌کننده‌ها به روش رگرسیون گام‌به‌گام

با توجه به اینکه تعداد زیاد توصیف‌کننده باعث پیچیدگی محاسبات شده و تعدادی از توصیف‌کننده‌ها حاوی اطلاعات یکسانی هستند، تعداد توصیف‌کننده‌ها باید به نوعی کاهش یابد. به همین منظور با استفاده از correlation توصیف کننده‌هایی که همبستگی بالای ۰/۹ داشتند حذف شدند و توصیف-کننده‌هایی که برای تمام ملکول‌ها مقادیر ثابت یا تقریبا ثابت داشتند، حذف شدند و در نهایت از ۳۲۲۴ توصیف کننده، ۱۹۲ توصیف‌کننده باقی ماند. با به کارگیری نرم‌افزار SPSS و با اجرای رگرسیون مرحله‌ای برای ۱۹۲ توصیف‌کننده یاد شده به عنوان متغیرهای مستقل، و اندیس بازداری به عنوان متغیر وابسته، تعداد ۱۰ توصیف‌کننده انتخاب شدند که نام و طبقه‌بندی آن‌ها در جدول(۳-۲) نشان داده شده است. همچنین جدول(۳-۳) ماتریس همبستگی بین این توصیف‌کننده‌ها را ارائه می-دهد، نتایج این جدول نشان می‌دهد که بین این توصیف‌کننده‌ها همبستگی معناداری وجود ندارد.

جدول(۳-۲)-توصیف‌کننده‌های انتخاب‌شده با روش رگرسیون گام‌به‌گام

NO.	Symbol	Class	Meaning
۱	X3v	Connectivity indices	Valance connectivity index chi-3
۲	PCWTe	Charge descriptors	Partial charge weighted topological electronic descriptor
۳	F04CS	2D-frequency fingerprints	Frequency of c-s at topological distance 04
۴	RDF115e	RDF	Radial Distribution Function -11.5/ weighted by aromatic Sanderson electronegativities
۵	Mor19m	3D-MoRSE descriptors	3D-MoRSE-Signal 19/ weighted by aromatic masses
۶	PW3	Topological descriptors	Path/walk3-Randic shape index
۷	H6m	GETAWAY descriptors	H autocorrolation of lag 6/ weighted by aromatic masses
۸	R6u	GETAWAY descriptors	R autocorrolation of lag 6/unweighted
۹	R2m	GETAWAY descriptors	R autocorrolation of lag 2/ weighted by aromatic masses
۱۰	Mor29m	3D-MoRSE descriptors	3D-MoRSE-Signal 29/ weighted by aromatic masses

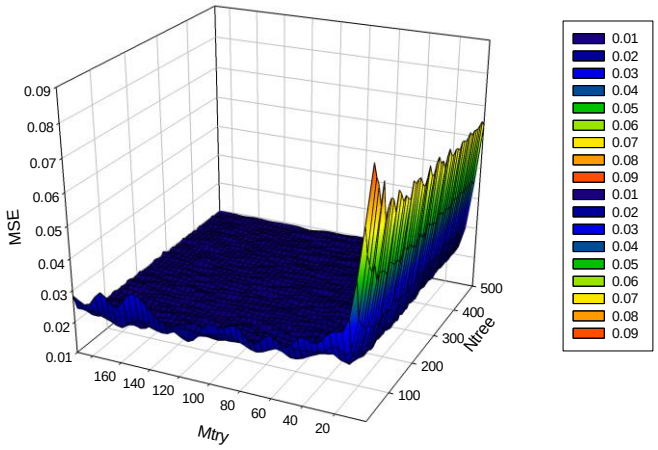
جدول(۳-۳)-ماتریس همبستگی توصیف‌کننده‌های انتخاب‌شده توسط روش رگرسیون گام‌به‌گام

	x3v	PCWTe	F04CS	Mor19m	PW3	H6m	RDF115e	R6u	R2m	Mor29m
x3v	۱									
PCWTe	۰/۵۷۵	۱								
F04CS	۰/۴۹۹	۰/۳۷۴	۱							
Mor19m	۰/۳۸۹	۰/۵۷	۰/۳۳۵	۱						
PW3	۰/۷۰۶	۰/۰۹۹	۰/۴۴۷	۰/۵۶۹	۱					
H6m	-۰/۵۴۲	۰/۴۲۹	۰/۱۸۴	۰/۰۳۷	۰/۲۱۶	۱				
RDF115e	۰/۶۲۱	۰/۴۷۰	۰/۲۳۷	۰/۰۸۳	۰/۲۲۶	۰/۷۵۲	۱			
R6u	۰/۲۲۹	-۰/۰۱۷	-۰/۱۳۴	-۰/۳۳۴	-۰/۱۳۲	-۰/۱۸۴	-۰/۱۴۰	۱		
R2m	-۰/۳۳۵	-۰/۵۴۱	-۰/۵۰۱	-۰/۲۶۱	-۰/۱۹۷	-۰/۳۰۳	-۰/۲۷۵	۰/۲۶۴	۱	
Mor29m	-۰/۱۰۳	-۰/۰۲۵	۰/۰۱۳	۰/۴۱	-۰/۱۶۹	-۰/۱۷۴	۰/۲۳۵	-۰/۰۱	-۰/۲۷۶	۱

۳-۵- مدل‌سازی و بهینه‌سازی پارامتر های مؤثر بر روش جنگل تصادفی

برای مدل‌سازی توسط جنگل‌های تصادفی، ابتدا باید پارامترهای مؤثر بهینه شوند. برای انجام این کار ابتدا ترکیبات به دو سری آموزش (۱۲۲ ترکیب) و سری آزمون (۳۰ ترکیب) تقسیم شدند. مقادیر مربوط به ۱۹۲ توصیف کننده به عنوان متغیر مستقل و مقادیر اندیس بازداری به عنوان متغیر وابسته منسوب شدند. داده‌های سری OOB نیز برای بهینه‌سازی تعداد درختان (ntree)، تعداد توصیف‌کننده‌های انتخاب‌شده در هر مرحله افراز (Mtry) و تعداد مشاهدات باقی‌مانده در هر گره (Node Size) در نظر گرفته شدند. برای بهینه نمودن، تعداد درخت از ۱۰ تا ۵۰۰ با گام ده، تعداد توصیف‌کننده‌های انتخاب‌شده در هر مرحله افراز (Mtry) از ۱ تا ۱۹۲ با گام ده و تعداد مشاهدات باقی‌مانده در هر گره از ۱ تا ۵ تغییر داده شد و در هر مرحله مقدار خطای مربوط به مجموعه OOB محاسبه گردید. نتایج بهینه‌سازی پارامترهای تعداد درختان و تعداد توصیف‌کننده‌های وارد شده در

هرمرحله افزار در گره بهینه در شکل (۳-۱) آمده است. شکل نشان می‌دهد که با بالا رفتن mtry خطا کمتر شده تا جایی که مقدار خطا ثابت می‌شود که در قسمت آبی رنگ قابل مشهود است و با پائین آمدن mtry خطا بیشتر شده به طوری که بع ماکسیمم مقدار خود می‌رسد. همچنین برای متغیرهای وارد شده جدول (۳-۴) چندین حالت که کمترین MSE را برای OOB دارد، نشان می-دهد.



شکل(۳-۱)-نمودار تغییرات خطای OOB در مقادیر متفاوت Mtry و Ntree

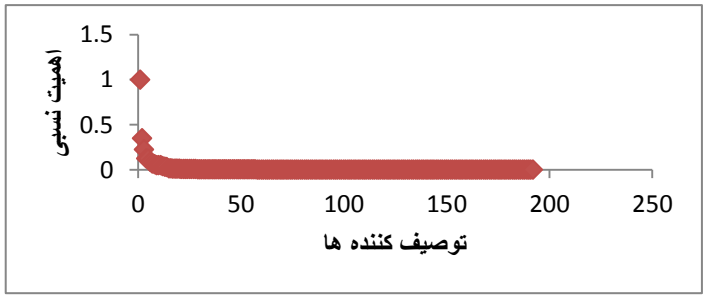
جدول (۳-۴)-کمترین مقادیر MSE همراه با Mtry وntree متناظر با آن‌ها

ntree	Mtry	Node Size	MSE Of OOB
۱۷۰	۱۱۱	۱	۰/۰۱۶۳
۱۷۰	**۱۱۱**	**۲**	**۰/۰۱۶۲**
۱۵۰	۹۱	۳	۰/۰۱۶۹
۱۱۰	۱۷۱	۴	۰/۰۱۷۵
۱۹۰	۱۴۱	۵	۰/۰۱۸۷

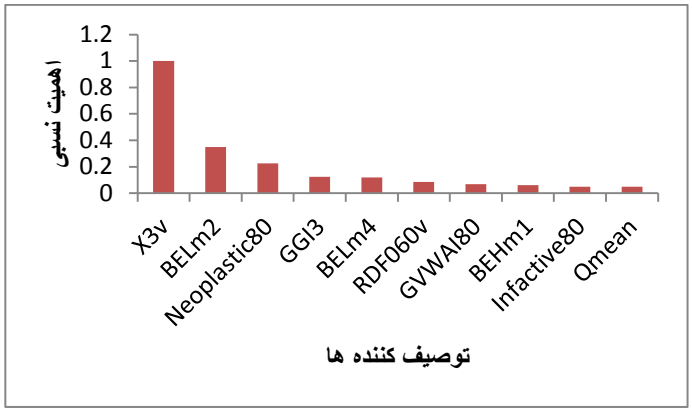
با توجه به جدول(۳-۴) و بر اساس مدل‌سازی جنگل‌های تصادفی با ۱۹۲ توصیف‌کننـده ورودی، شرایط بهینه N_{tree}=۱۷۰ و M_{try}=۱۱۱ و Nodesize=۲ به‌دست آمد و همانطور که می‌شـود در این حالت مقدار MSE مینیمم است. این مدل با توجه به ارزیابی‌های پیش‌رو می‌تواند بـا مـدل‌هـای برتر در رقابت باشد.

۳-۶- تعیین اهمیت نسبی توصیف‌کننده‌ها با روش جنگل تصادفی

همان‌طور که در بخش قبل توضیح داده شد، روش جنگل تصادفی علاوه بر اینکه به‌عنوان یک روش مدل‌سازی به کار می‌رود، می‌تواند به‌عنوان یک روش انتخاب متغیر نیز به‌کاربرده شود. پس از ساخت مدل جنگل‌های تصادفی و بهینه کردن پارامترهای مؤثر بر جنگل‌های تصادفی، با وارد کردن ۱۹۲ توصیف‌کننده به عنوان توصیف‌کننده ورودی در جنگل تصادفی طبق شرایط بهینه منتخب که در بخش ۳-۵ بدست آمد اهمیت نسبی توصیف‌کننده‌های ورودی محاسبه گردید. این توصیف‌کننده‌های ورودی براساس اهمیت چیده شدند. با توجه به شکل (۳-۲) از توصیف‌کننده ۱۰ به بعد مقدار اهمیت نسبی ثابت می‌شود به همین خاطر پراهمیت‌ترین آن‌ها از توصیف‌کننده اول تا دهم انتخاب می‌شود و به همین منظور ۱۰ توصیف‌کننده برتر به عنوان ورودی شبکه عصبی انتخاب می‌شوند. شکل (۳-۳) میزان اهمیت نسبی ۱۰ توصیف‌کننده را نمایش می‌دهد. همان‌طور که مشاهده می‌شود بسیاری از توصیف‌کننده‌ها دارای اهمیت نسبی بالایی هستند و توصیف‌کننده X3v دارای بالاترین اهمیت نسبی نسبت به سایر توصیف‌کننده‌ها است. ۱۰ توصیف‌کننده مورد نظر همراه با مقدار اهمیت و نوع طبقهٔ آن‌ها در جدول (۳-۵) نشان داده‌شده‌اند. همچنین ماتریس همبستگی این توصیف‌کننده-ها در جدول(۳-۶) آمده است که نشان‌دهنده عدم همبستگی معنی‌دار بین آن‌ها می‌باشد.



شکل(۳-۲)-نمودار اهمیت نسبی ۱۹۲ توصیف‌کننده



شکل(۳-۳)-نمودار میزان اهمیت نسبی توصیف‌کننده‌ها

جدول(۳-۵)-توصیف‌کننده‌های انتخاب شده توسط جنگل تصادفی

NO.	Symbol	class	Meaning	Importance	Relative Importance
۱	X3v	Connectivity indices	Valance connectivity index chi-3	۲۴/۶۷	۱
۲	BELm2	Burden eigen values	Lowest eigenvalue n.2 of Burden matrix/weighted by atomic masses	۸/۶۲	۰/۳٤۹
۳	Neoplastic80	Molecular properties	Ghose-viswanadhan-wendoloski antineoplastic-like- index at 80%	۵/۵۶۰	۰/۲۲۵
۴	GGI3	Topological charge indices	Topological charge index of order 3	۳/۰۴۷	۰/۱۲۳
۵	BELm4	Burden eigen values	highest eigenvalue n.1 of Burden matrix/weighted by atomic masses	۲/۹۵۴	۰/۱۱۹
۶	RDF060v	RDF	Radial Distribution Function -6.0/ weighted by aromatic van der waals volumes	۲/۱۳	۰/۰۸٦٤
۷	GVWAI80	Molecular properties	Ghose-viswanadhan-wendoloski drug-like- index at 80%	۱/۶۷۱	۰/۰٦۷۷
۸	BEHm1	Burden eigen values	highest eigenvalue n.1 of Burden matrix/weighted by atomic masses	۱/۵۱۸	۰/۰٦۱۵
۹	infective80	Molecular properties	Ghose-viswanadhan-wendoloski antiin fective-like index at 80%	۱/۲۳۳	۰/۰٤۹۹
۱۰	Qmean	Charge descriptors	Mean absolute charge	۱/۲۲۴	۰/۰٤۹٦

جدول(۳-۶)-ماتریکس همبستگی توصیف‌کننده‌ها

	X3v	BELm2	Neoplastic80	GGI3	BELm4	RDF060v	GVWAI80	BEHm1	infective80	Qmean
X3v	۱									
BELm2	۰/۷۷۵	۱								
Neoplastic80	-۰/۵۹۱	-۰/۷٦٤	۱							
GGI3	۰/۸۳٤	۰/۷٦٦	-۰/۷۰۷	۱						
BELm4	۰/۸٤۸	۰/٦٤۳	-۰/۵۱۳	۰/۷۸۲	۱					
RDF060v	۰/۸۱۸	۰/۷۳۱	-۰/۵٦۰	۰/۷٦۲	۰/۷٦۵	۱				
GVWAI80	-۰/۷٤۳	-۰/۵٦۸	۰/۲۹٤	-۰/۵۵۸	-۰/٦۵۲	-۰/٦۲۷	۱			
BEHm1	۰/۷٤۵	۰/۸۵۳	-۰/۷٦۷	۰/۸۰۸	۰/۵۳٤	۰/۷٤۸	-۰/٤٤۹	۱		
infective80	-۰/٦۷۵	-۰/٦۸۲	۰/۵۹۷	-۰/٦۵۸	-۰/٦٤۰	-۰/٦۱٦	۰/٤۹۲	-۰/٦۲۳	۱	
Qmean	-۰/۷٦۳	-۰/۸٤۳	۰/۸۱۸	-۰/۷۹۳	-۰/٦۵۹	-۰/٦٦۰	۰/٤۵۰	-۰/۸۲٤	۰/۷۰۲	۱

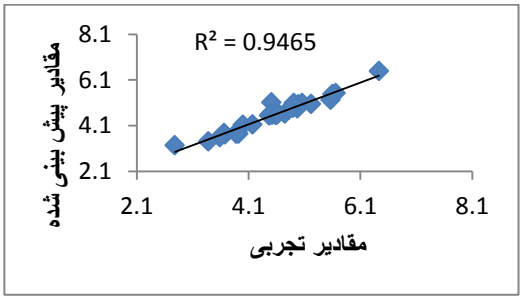
۳-۵-۱- ارزیابی مدل جنگل‌های تصادفی

۳-۵-۱-۱- ارزیابی مدل با استفاده از داده‌های سری آزمون

اهمیت مدل پیش‌بینی وقتی مشخص می‌گردد که مدل حاصله بتواند خاصیت مولکول‌هایی را که در مدل‌سازی به کار نرفته‌اند پیش‌بینی کند. بدین منظور مدل منتخب، برای پیش‌بینی اندیس بازداری داده‌های سری آزمون به کاربرده‌شد. جدول (۳-۷) نتایج پیش‌بینی حاصل از ارزیابی مدل‌های جنگل تصادفی (RF) با استفاده از داده‌های سری آزمون را نشان می‌دهد. نتایج بیانگر این است که مدل ارائه شده خطای پیش‌بینی کمی دارد. شکل (۳-۴) نمودار مقادیر پیش‌بینی‌شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد. نتایج بیانگر این است که مدل ارائه شده خطای پیش‌بینی کمی دارند. ضریب تعیین بالای مدل با ۱۹۲ توصیف‌کننده ورودی، برتری قابل ملاحظه‌ی مدل را در پیش‌بینی اندیس بازداری مربوط به داده‌های سری آزمون نشان می‌دهد.

جدول (۳-۷) نتایج حاصل از ارزیابی مدل جنگل تصادفی با استفاده از داده‌های آزمون

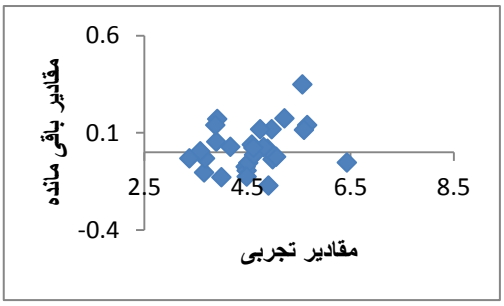
شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی‌شده	درصد خطا
۳	۲/۷۸	۳/۲۳	۱۶/۴۱
٦	۳/۹۲	۳/۷۴	-۴/۳۳
۹	۴	۴/۱۲	۳/۲۲
۱۲	۳/۹	۳/۸۴	-۱/۳۹
۱۵	۳/۶۶	۳/۷۶	۲/۸۵
۱۸	۳/۸۸	۳/۷۴	-۳/۵۷
۲۱	۴/۴۷	۴/۵۴	۱/۶۵
۲٤	۴/۵۹	۴/۵۵	-۰/۸۴
۲۷	۴/۹۷	۴/۸۵	-۲/۳۸
۳۰	۴/۹۹	۵/۰۲	۰/۷۵
۳۱	۵/۰۶	۵/۰۸	۰/۴۵
۳۲	۴/۸۹	۴/۸۷	-۰/۳۴
۳۳	۵/۲۲	۵/۰۴	-۳/۳۱
۳۵	۴/۵۱	۵/۱۱	۱۳/۴۵
۳٦	۴/۹۱	۵/۰۸	۳/۴۹
٤۱	۵/۶۶	۵/۵۲	-۲/۴۶
۵۰	۵/۵۷	۵/۲۲	-۶/۲۴
٦۹	۶/۴۳	۶/۴۸	۰/۸۲
۷٤	۳/۶۸	۳/۷۱	۰/۸۶
۸۵	۴/۱۷	۴/۱۴	-۰/۶۴
۹٦	۴/۴۹	۴/۵۸	۲/۱۰
۹۷	۴/۷۵	۴/۶۳	-۲/۴۶
۱۰۱	۳/۵۹	۳/۵۸	-۰/۱۲
۱۱۰	۳/۳۸	۳/۴۱	۰/۹۲
۱۲۰	۴/۵۲	۴/۵۷	۱/۲۰
۱۳۰	۴/۸۲	۴/۷۹	-۰/۵۴
۱٤۰	۴/۶۳	۴/۶۴	۰/۲۹
۱٤۳	۴/۴۹	۴/۶۱	۲/۷۸
۱٤٤	۴/۶۲	۴/۵۹	-۰/۵۰
۱۵۰	۵/۶	۵/۴۸	-۲/۰۵



شکل (۳-۴)-نمودارتغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی توسط مدل RF برای داده‌های سری آزمون

۳-۵-۱-۲- ارزیابی مدل با استفاده از نمودار خطای باقی‌مانده

اختلاف مقادیر پیش‌بینی‌شده و مقادیر تجربی، باقی‌مانده نامیده می‌شود. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است. نمودار خطای باقی‌مانده برحسب مقادیر تجربی، برای مدل‌های ذکرشده در شکل (۳-۵) نشان داده‌شده است.



شکل (۳-۵)- نمودار خطای باقی‌مانده برحسب مقادیر تجربی با استفاده از داده‌های سری آزمون توسط روش جنگل تصادفی

۳-۶-مدل‌سازی شبکه عصبی مصنوعی با توصیف‌کننده‌های انتخاب‌شده توسط روش رگرسیون گام‌به‌گام (SR-ANN)

ابتدا ۱۰ توصیف‌کننده انتخاب‌شده باروش رگرسیون گام‌به‌گام به‌عنوان متغیرهای ورودی و خاصیت متناظر آن‌ها به‌عنوان متغیر وابسته در نظر گرفته شد تا پاسخ شبکه با آن‌ها سنجیده شود. برای به دست آوردن بهترین مدل با کمترین خطا، پارامترهای مؤثر تابع انتقال، تابع آموزش، تعداد متغیرهای ورودی شبکه، تعداد گره‌ها در لایهٔ پنهان و تعداد دورهای آموزشی به‌طور هم‌زمان بهینه شدند. تابع کارایی شبکه نیز خطای مطلق میانگین (MAE) در نظر گرفته شد.

در فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به سه بخش سری آزمون (۳۰ ترکیب) و سری ارزیابی (۳۰ ترکیب) و سری آموزش (۹۲ ترکیب) تقسیم شدند. سری ارزیابی برای محاسبهٔ گرادیان و به‌روز کردن اوزان و بایاس‌های شبکه استفاده شد. برای این منظور هر شبکه با ورودی از ۲

تا ۱۰ با گام یک به عنوان توصیف‌کننده ورودی و با دو الگوریتم آموزشی لونبرگ-مارکوات[1] و تنظیم بایزن[2] و دو تابع انتقال لگاریتم سیگموئید و تانژانت سیگموئید و تعداد گره از ۲ تا ۱۰ با گام یک و تعداد دورهای آموزش از ۵ تا ۵۰ با گام ۵، به‌طور هم‌زمان آموزش داده شد. درروند بهینه‌سازی فوق، حداقل میانگین مربعات سری ارزیابی به‌عنوان معیار انتخاب شد. نتیجه بهینه‌سازی این پارامترها در جدول (۳-۸) گردآوری‌شده است.

جدول (۳-۸)-توابع و پارامترهای بهینه شبکهٔ عصبی مصنوعی

R	تعداد دوره آموزش	MSE	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌کننده
۰/۹۸۴	۱۰	۰/۰۲۳۱	۷	لگاریتم-سیگموئید	لونبرگ-مارکوات	۶
۰/۹۸۶	**۱۵**	**۰/۰۲۱۹**	**۸**	**تانژانت-سیگموئید**	**لونبرگ-مارکوات**	**۶**
۰/۹۸۷	۱۵	۰/۰۲۵	۴	لگاریتم-سیگموئید	تنظیم بایزن	۷
۰/۹۸۶	۲۵	۰/۰۲۲۶	۴	تانژانت-سیگموئید	تنظیم بایزن	۶

با توجه به نتایج به‌دست‌آمده، الگوریتم آموزشی لونبرگ-مارکوات و تابع انتقال تانژانت سیگموئید، ۶ توصیف‌کنندهٔ ورودی و ۸ گره در لایهٔ پنهان با تعداد دور آموزشی ۱۵ کمترین MSE را نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی در نظر گرفته شد.

۳-۶-۱-ساختار شبکه عصبی بهینه شده

در جدول (۳-۹) مشخصات شبکه بهینه به دست آمده نشان داده شده است.

جدول(۳-۹)-مشخصات شبکه بهینه (SR-ANN)

تابع آموزش	لونبرگ-مارکوات
تابع انتقال لایه پنهان	تانژانت سیگموئید
تابع انتقال لایه خروجی	Purelin
تعداد گره	۸
تعداد توصیف‌کننده ورودی	۶
تعداد دور آموزشی	۱۵
پارامتر MSE	۰/۲۱۹

[1] Levenberg-Marquart
[2] bayesinregularization

۳-۶-۲- ارزیابی مدل شبکه عصبی مصنوعیSR-ANN

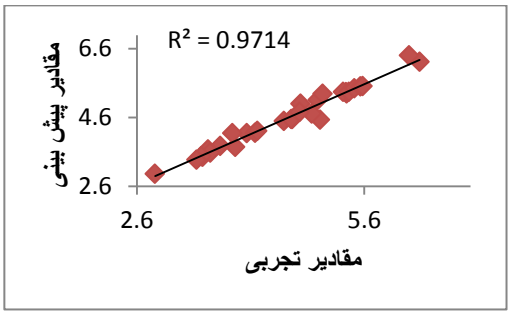
۳-۶-۲-۱-ارزیابی مدل با استفاده از داده‌های سری ارزیابی

بدین منظور، شبکه عصبی با ۹۲ مولکول سری آموزش، آموزش داده شد. معیار بهینه‌سازی شبکه، کمترین خطای مطلق میانگین برای سری ارزیابی است. نتایج حاصل از شبکه عصبی مصنوعی در جدول (۳-۱۰) و ترسیمی از مقادیر پیش‌بینی‌شده برحسب مقادیر تجربی برای سری ارزیابی در شکل (۳-۶) آورده شده است. درصد خطای مشاهده شده و ضریب تعیین به دست آمده نشان از برتری مدل به دست آمده دارد.

جدول (۳-۱۰)-نتایج حاصل از مدل شبکه عصبی مصنوعی برای داده‌های سری ارزیابی(SR-ANN)

درصد خطا	مقدار پیش‌بینی‌شده	مقدار واقعی	شماره ترکیب
۳/۹۱	۲/۹۵	۲/۸۴	۲
-۰/۵۸	۳/۴۴	۳/۴۷	۵
-۴/۰۶	۳/۷۴	۳/۹	۸
۲/۲۵	۴/۱۴	۴/۰۵	۱۱
۱/۸۴	۳/۷۶	۳/۷	۱۴
۷/۱۶	۴/۱۳	۳/۸۶	۱۷
-۹/۸۱	۴/۵۲	۵/۰۲	۲۰
۵/۰۲	۴/۹۹	۴/۷۶	۲۳
-۰/۹۷	۴/۴۹	۴/۵۴	۲۶
-۱/۴۲	۵/۲۹	۵/۳۷	۲۹
-۰/۴۶	۵/۳۳	۵/۳۶	۳۴
۰/۳۶	۵/۳۳	۵/۳۲	۳۷
-۴/۱۹	۴/۷۰	۴/۹۱	۳۸
-۱/۱۲	۵/۳۳	۵/۴۰	۴۰
-۰/۴۳	۵/۴۴	۵/۴۷	۴۲
۲/۳۳	۵/۱۰	۴/۹۹	۴۴
۴/۷۱	۵/۲۸	۵/۰۵	۵۱
-۱/۸۰	۶/۲۱	۶/۳۳	۵۳
۳/۴۹	۶/۴۰	۶/۱۹	۶۲
-۰/۹۸	۵/۴۹	۵/۵۵	۶۳
۳/۵۹	۳/۶۶	۳/۵۴	۷۱
-۰/۰۳	۳/۵	۳/۵۷	۷۲

۸۰	۴/۱۶	۴/۱۴	-۰/۳۸
۹۱	۴/۱۹	۴/۲۰	۰/۲۹
۹۹	۴/۶۵	۴/۵۴	-½
۱۰۴	۳/۴۵	۳/۴۶	۰/۴۶
۱۰۹	۳/۳۹	۳/۳۶	-۰/۷۳
۱۲۴	۴/۶۴	۴/۵۴	-۱/۹۸
۱۳۰	۴/۷۵	۴/۷۵	۰/۰۷۰
۱۵۱	۵/۵۸	۵/۵۱	-۱/۲۹



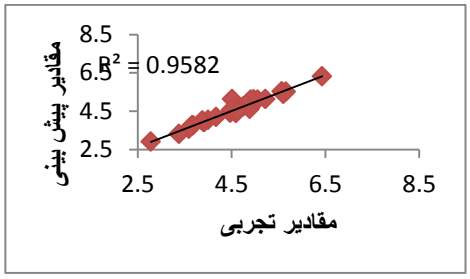
شکل (۳-۶)-نمودار مقادیر پیش‌بینی شده برحسب مقادیر تجربی برای داده‌های سری ارزیابی توسط روش ANN-SR

۳-۶-۲-۲- ارزیابی مدل با استفاده از داده‌های سری آزمون

بعد از انتخاب شبکه عصبی بهینه با استفاده از سری ارزیابی، قدرت پیش‌بینی شبکه به وسیلهٔ ۳۰ داده که در سری آزمون قرار دارند مورد بررسی قرار گرفت. نتایج حاصل در جدول (۳-۱۱) نشان داده‌شده است. همچنین شکل (۳-۷) مقادیر پیش‌بینی‌شده برحسب مقادیر تجربی برای سری آزمون نشان داده است. با توجه به میانگین خطای بدست آمده که که مقدار آن معادل ۰/۸۴۰ است و مقدار ضریب تعیین در SR-ANN برای داده‌های سری ارزیابی نسبت به داده‌های سری آزمون بیشتر بوده که انتظار نیز همین است همچنین میانگین خطای ارزیابی ۰/۱۰۳ است و ضریب تعیین بالاتری دارد.

جدول (۳-۱۱)-نتایج حاصل از مدل شبکه عصبی مصنوعی برای داده‌های سری آزمون(SR-ANN)

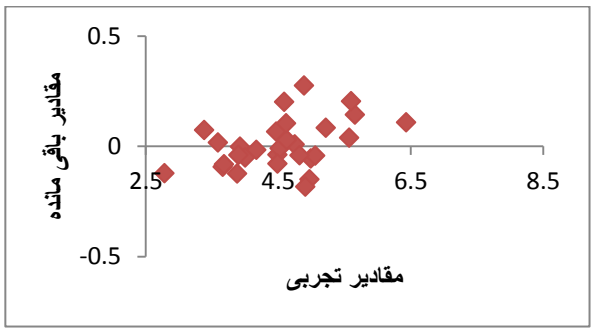
شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطا
۳	۲/۷۸	۲/۹۰	۴/۴۱
٦	۳/۹۲	۳/۹۲	۰/۰۴۰
۹	۴	۴/۰۵۱	۱/۲۸۰
۱۲	۳/۹	۳/۹۳	۱/۰۷
۱۵	۳/۶۶	۳/۷۵	۲/۵۳
۱۸	۳/۸۸	۴/۰۰	۳/۱۹
۲۱	۴/۴۷	۴/۴۰	-۱/۴۸
۲٤	۴/۵۹	۴/۳۸	-۴/۳۷
۲۷	۴/۹۷	۵/۱۲	۳/۰۲
۳۰	۴/۹۹	۵/۰۴	۱/۱۱
۳۱	۵/۰۶	۵/۱۰	۰/۸۵
۳۲	۴/۸۹	۴/۶۱	-۵/۶۲
۳۳	۵/۲۲	۵/۱۳	-۱/۶۰
۳۵	۴/۵۱	۵/۱۲	۱۳/۷۲
۳٦	۴/۹۱	۵/۰۹	۳/۷۴
٤۱	۵/۶۶	۵/۵۱	-۲/۵۲
۵۰	۵/۵۷	۵/۵۳	-۰/۶۸
٦۹	۶/۴۳	۶/۳۲	-۱/۶۸
۷٤	۳/۶۸	۳/۷۶	۲/۲۳
۸۵	۴/۱۷	۴/۱۸	۰/۴۰۷
۹٦	۴/۴۹	۴/۵۳	۰/۸۶
۹۷	۴/۷۵	۴/۷۴	-۰/۱۶
۱۰۱	۳/۵۹	۳/۵۷	-۰/۴۵
۱۱۰	۳/۳۸	۳/۳۰	-۲/۱۸
۱۲۰	۴/۵۲	۴/۵۲	۰/۲۰
۱۳۰	۴/۸۲	۴/۸۶	۰/۸۴
۱٤۰	۴/۶۳	۴/۶۰	-۰/۶۱
۱٤۳	۴/۴۹	۴/۵۶	۱/۷۶
۱٤٤	۴/۶۲	۴/۵۱	-۲/۲۵
۱۵۰	۵/۶	۵/۳۹	-۳/۶۳



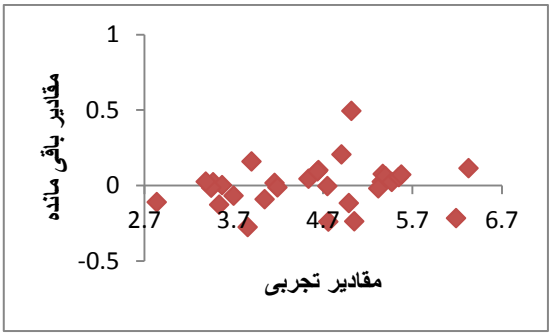
شکل (۳-۷)-نمودار مقادیر پیش‌بینی شدهٔ اندیس بازداری برحسب مقادیر تجربی برای داده‌های سری آزمون توسط روش SR-ANN

۳-۶-۲-۳- ارزیابی مدل SR-ANN توسط نمودار خطای باقی‌مانده

شکل‌های (۳-۸) و (۳-۹) نمودارهای خطای باقی مانده برحسب مقادیر تجربی سری‌های آزمون و ارزیابی برای مدل SR-ANN را نشان می‌دهد. توزیع متقارن داده ها حول محور افقی حاکی از عـدم وجود خطای سیستماتیک است.



شکل (۳-۸)-نمودار خطای باقی‌مانده برحسب مقادیر تجربی برای داده‌های آزمون(SR-ANN)



شکل (۳-۹)- نمودار خطای باقی‌مانده برحسب مقادیر تجربی برای داده‌های سری ارزیابی(SR-ANN)

۳-۶-۲-۴- ارزیابی مدل SR-ANN توسط روش ارزیابی متقاطع

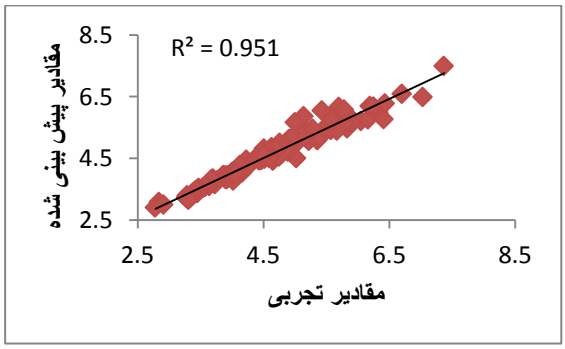
در این روش هر بار یک ترکیب از ۱۵۲ترکیب به‌عنوان داده‌های آزمون کنار گذاشته شد و در شرایط بهینه به‌دست‌آمده برای سری آموزش با ترکیبات باقی‌مانده مدل‌سازی صورت گرفت. سپس مدل به‌دست‌آمده برای پیش‌بینی اندیس بازداری ترکیب کنار گذاشته‌شده به کار گرفته شد و این فرآیند برای تک‌تک ترکیبات تکرار گردید. نتایج حاصل از این روش در جدول (۳-۱۲) ارائه‌شده است. این نتایج تعمیم پذیری بالای مدل طراحی شده را بیان می‌کند. همچنین ضریب همبستگی مشاهده شده بین مقادیر پیش بینی شده‌ی اندیس‌بازداری و مقادیر تجربی در شکل (۳-۱۰)، بر نزدیکی مقادیر پیش‌بینی شده توسط مدل نسبت به مقادیر تجربی آن‌ها دلالت دارد. درصد خطاهای مشاهده شده برای ترکیبات سولفور چندحلقه‌ای زیر ۱۵ درصد می‌باشد که نشان می‌دهد مدل مورد نظر صحت خوب و قابل قبولی را داراست، همچنین در شکل (۳-۱۱) مقادیر باقی‌مانده برحسب مقدار تجربی اندیس بازداری ترکیبات موردبحث ترسیم شده است که تقارن پراکندگی نقاط در دو طرف محور افقی عدم وجود خطای سیستماتیک را برای بیشتر ترکیبات نشان می‌دهد. طبق تکنیک ارزیابی متقاطع این مدل می‌تواند به عنوان یک مدل مناسب معرفی شود.

جدول (۳-۱۲)-نتایج حاصل از ارزیابی مدل SR-ANN با استفاده از ارزیابی متقاطع

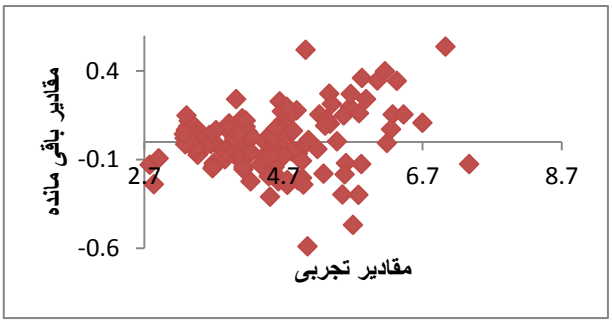
NO.	مقدارتجربی	مقدارپیش‌بینی	درصدخطا	NO.	مقدارتجربی	مقدارپیش‌بینی	درصدخطا
۱	۳/۲۸	۳/۲۶	-۰/۵۸	۸۱	۴/۱۶	۴/۰۴	-۲/۸۷
۲	۲/۸۴	۳/۰۸	۸/۴۶	۸۲	¼	۴/۲۱	۲/۷۸
۳	۲/۷۸	۲/۹۱	۴/۶۵	۸۳	۴	۳/۹۸	-۰/۴۰
۴	۲/۹۱	۳/۰۰	۳/۱۹	۸۴	۴/۰۱	۴/۰۶	۱/۲۷
۵	۳/۴۷	۳/۵۴	۲/۱۱	۸۵	۴/۱۷	۴/۲۲	۱/۳۵
۶	۳/۹۲	۳/۸۱	-۲/۶۲	۸۶	۴/۱۷	۴/۱۵	-۰/۲۵
۷	۳/۷۶	۳/۷۸	۰/۶۲	۸۷	۴/۱۲	۴/۱۸	۱/۶۵
۸	۳/۹	۳/۸۳	-۱/۷۱	۸۸	۴/۰۵	۴/۰۶	۰/۴۸
۹	۴	۴/۰۰	۰	۸۹	¼	۴/۱۹	۲/۲۱
۱۰	۴/۱۳	۴/۲۸	۳/۷۶	۹۰	۴/۲۴	۴/۲۹	۱/۳۹
۱۱	۴/۰۵	۴/۰۰	-۱/۰۷	۹۱	۴/۱۹	۴/۱۳	-۱/۲۸
۱۲	۳/۹	۳/۹٤	۱/۱۲	۹۲	۴/۲۵	۴/۲۲	-۰/۵۹

۱۳	۳/۷	۳/۸۳	۳/۷۱	۹۳	۴/۶۶	۴/۶۴	-۰/۲۲
۱۴	۳/۷	۳/۸۲	۳/۲۶	۹۴	۴/۷۱	۴/۶۲	-۱/۸۴
۱۵	۳/۶۶	۳/۶۷	۰/۳۳	۹۵	۴/۵۶	۴/۵۵	-۰/۱۲
۱۶	۳/۷۳	۳/۶۶	-۱/۷۴	۹۶	۴/۴۹	۴/۵۴	۱/۲۵
۱۷	۳/۸۶	۳/۹۵	۲/۵۷	۹۷	۴/۷۵	۴/۶۵	-۱/۹۴
۱۸	۳/۸۸	۳/۹۳	۱/۵۲	۹۸	۴/۷۳	۴/۶۶	-۱/۵۱
۱۹	۴/۶۳	۴/۸۵	۴/۸۲	۹۹	۴/۶۵	۴/۴۲	-۴/۸۸
۲۰	۵/۰۲	۴/۵۰	-۱۰/۳۲	۱۰۰	۴/۶۶	۴/۷۱	۱/۰۶
۲۱	۴/۴۷	۴/۴۷	۰/۰۳۵	۱۰۱	۳/۵۹	۳/۶۱	۰/۶۲
۲۲	۴/۴۷	۴/۴۴	۰/۵۴۱	۱۰۲	۳/۴۸	۳/۴۵	-۰/۷۵
۲۳	۴/۷۶	۵/۰۰	۵/۱۰	۱۰۳	۳/۴۷	۳/۴۹	۰/۸۱
۲۴	۴/۵۹	۴/۵۵	-۰/۶۶	۱۰۴	۳/۴۵	۳/۴۸	۱/۱۵
۲۵	۴/۴۳	۴/۴۲	-۰/۰۰۷	۱۰۵	۳/۴۵	۳/۳۷	-۲/۲۲
۲۶	۴/۵۴	۴/۶۲	۱/۸۲	۱۰۶	۳/۴۲	۳/۴۶	۱/۲۲
۲۷	۴/۹۷	۵/۱۷	۴/۰۹	۱۰۷	۴/۰۲	۳/۷۷	-۵/۹۹
۲۸	۴/۶	۴/۷۹	۴/۱۳	۱۰۸	۳/۳	۳/۲۳	-۲/۰۰
۲۹	۵/۳۷	۵/۲۶	-۱/۹۰	۱۰۹	۳/۳۹	۳/۳۲	-۲/۰۶
۳۰	۴/۹۹	۵/۰۱	۰/۵۰	۱۱۰	۳/۳۸	۳/۳۸	۰/۰۶۰
۳۱	۵/۰۶	۵/۰۵	-۰/۱۹	۱۱۱	۳/۳۳	۳/۲۴	-۲/۵۹
۳۲	۴/۸۹	۴/۷۱	-۳/۶۴	۱۱۲	۳/۲۹	۳/۳۰	۰/۴۲۰
۳۳	۵/۲۲	۵/۰۶	-۲/۹۶	۱۱۳	۳/۳۴	۳/۲۲	-۳/۵۴
۳۴	۵/۳۶	۵/۰۸	-۵/۰۴	۱۱۴	۳/۲۷	۳/۲۲	-۱/۳۴
۳۵	۴/۵۱	۴/۶۹	۴/۱۷	۱۱۵	۳/۳۱	۳/۱۶	-۴/۵۰
۳۶	۴/۹۱	۵/۱۲	۴/۴۵	۱۱۶	۴	۳/۹۸	-۰/۴۹
۳۷	۵/۳۲	۵/۱۹	-۲/۳۶	۱۱۷	۴/۲۳	۴/۴۵	۵/۳۱
۳۸	۴/۹۱	۴/۹۷	۱/۳۷	۱۱۸	۴/۴۴	۴/۵۶	۲/۸۶
۳۹	۵/۳	۵/۲۰	-۱/۷۲	۱۱۹	۴/۵۴	۴/۵۷	۰/۶۸
۴۰	۵/۴	۵/۱۸	-۳/۹۳	۱۲۰	۴/۵۲	۴/۴۸	-۰/۷۱
۴۱	۵/۶۶	۵/۴۶	-۳/۴۵	۱۲۱	۴/۵۶	۴/۵۴	-۰/۳۰۰
۴۲	۵/۴۷	۵/۴۷	-۰/۰۲۸	۱۲۲	۴/۶۸	۴/۵۱	-۳/۶۵
۴۳	۴/۹۷	۵/۰۷	۲/۱۸	۱۲۳	۴/۷	۴/۵۸	-۲/۵۱
۴۴	۴/۹۹	۵/۲۳	۴/۸۵	۱۲۴	۴/۶۴	۴/۵۶	-۱/۶۰
۴۵	۵/۱۹	۵/۲۳	۰/۸۰	۱۲۵	۴/۴۹	۴/۵۴	۱/۲۱
۴۶	۴/۷۶	۴/۹۷	۴/۵۳	۱۲۶	۴/۳۷	۴/۴۹	۲/۹۶
۴۷	۴/۶۴	۴/۶۲	-۰/۳۹	۱۲۷	۴/۱۳	۴/۲۷	۳/۴۵
۴۸	۴/۵۱	۴/۸۲	۶/۹۰	۱۲۸	۴/۷۷	۴/۷۳	-۰/۷۹
۴۹	۵/۸۳	۵/۴۶	-۶/۱۷	۱۲۹	۴/۷۹	۴/۶۴	-۳/۰۴
۵۰	۵/۵۷	۵/۴۲	-۲/۶۵	۱۳۰	۴/۷۵	۴/۶۷	-۱/۶۶
۵۱	۵/۰۵	۵/۶۴	۱۱/۶۹	۱۳۱	۴/۶۸	۴/۷۲	۱/۰۳
۵۲	۶/۰۵	۵/۷۰	-۵/۷۰	۱۳۲	۴/۷۴	۴/۷۱	-۰/۵۱

۵۳	۶/۳۳	۵/۹۸	-۵/۴۱	۱۳۳	۴/۷۴	۴/۷۷	۰/۷۲
۵۴	۵/۸۲	۵/۹۴	۲/۱۵	۱۳۴	۴/۷۵	۴/۷۶	۰/۲۵
۵۵	۵/۱۴	۵/۸۶	۱۴/۱۵	۱۳۵	۴/۸۲	۴/۸۷	۱/۱۲
۵۶	۶/۱۶	۵/۷۶	-۶/۴۵	۱۳۶	۴/۸۴	۴/۷۷	-۱/۲۸
۵۷	۵/۷	۶/۱۷	۸/۲۴	۱۳۷	۴/۰۵	۴/۰۰	-۱/۰۷
۵۸	۵	۵/۶۷	۱۳/۴۴	۱۳۸	۴/۴۸	۴/۵۸	۲/۳۹
۵۹	۵/۸۹	۵/۶۴	-۴/۰۸	۱۳۹	۴/۷۵	۴/۵۵	-۴/۱۴
۶۰	۵/۷۸	۶/۰۷	۵/۱۸	۱۴۰	۴/۶۳	۴/۶۹	۱/۴۴
۶۱	۶/۲۷	۶/۱۱	-۲/۴۶	۱۴۱	۴/۷۱	۴/۷۸	۱/۶۴
۶۲	۶/۱۹	۶/۱۹	۰/۱۰۵	۱۴۲	۴/۵۸	۴/۷۲	۳/۰۷
۶۳	۵/۵۵	۵/۸۴	۵/۳۶	۱۴۳	۴/۴۹	۴/۶۸	۴/۳۴
۶۴	۶/۴۱	۵/۷۷	-۹/۹۶	۱۴۴	۴/۶۲	۴/۵۳	-۱/۷۵
۶۵	۶/۲۵	۶/۱۸	-۱/۱۱	۱۴۵	۴/۵۹	۴/۵۱	-۱/۵۹
۶۶	۵/۴۳	۶/۰۵	۱۱/۵۱	۱۴۶	۴/۴۴	۴/۵۳	۲/۰۶
۶۷	۶/۷	۶/۵۹	-۱/۶۰	۱۴۷	۴/۵۱	۴/۵۹	۱/۸۶
۶۸	۷/۰۳	۶/۴۹	-۷/۶۴	۱۴۸	۵/۷۹	۵/۶۲	-۲/۷۸
۶۹	۶/۴۳	۶/۲۷	-۲/۴۰	۱۴۹	۵/۶۷	۵/۳۹	-۴/۷۶
۷۰	۷/۳۷	۷/۴۹	۱/۷۱	۱۵۰	۵/۶	۵/۷۲	۲/۱۶
۷۱	۳/۵۴	۳/۵۵	۰/۳۸	۱۵۱	۵/۵۸	۵/۷۶	۳/۳۱
۷۲	۳/۵۷	۳/۵۷	۰/۰۴۲	۱۵۲	۵/۲۸	۵/۴۶	۳/۴۱
۷۳	۳/۶۴	۳/۶۰	-۱/۰۶				
۷۴	۳/۶۸	۳/۸۲	۴/۰۶				
۷۵	۳/۹۳	۳/۹۴	۰/۴۸				
۷۶	۴	۴/۰۳	۰/۸۶				
۷۷	۴/۰۸	۴/۰۰	-۱/۸۴				
۷۸	۴/۰۱	۴/۰۰	-۰/۱۱				
۷۹	۴/۱۱	۳/۹۷	-۳/۲۲				
۸۰	۴/۱۶	۴/۰۷	-۲/۰۲				



شکل(۳-۱۰)-نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای ارزیابی مدل SR-ANN به روش ارزیابی متقاطع برای کل داده‌ها



شکل(۳-۱۱)-نمودار مقادیر باقی‌مانده بر حسب مقادیر تجربی برای ارزیابی مدل SR-ANN به روش ارزیابی متقاطع برای کل داده‌ها

۳-۷- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌کننده‌های انتخاب‌شده توسط روش RF (RF-ANN)

در این بخش، از شبکه عصبی با یک لایه پنهان و لایه خروجی و توصیف‌کننده‌های منتخب روش جنگل‌های تصادفی استفاده گردید. برای یافتن مقادیر بهینه پارامترهای موثر بر کارایی شبکه، تعداد ورودی‌ها از ۲ تا ۱۰ و با دو الگوریتم آموزشی لونبرگ-مارکوات (train lm) و تنظیم بایزن (trainbr) انتقال لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) و تعداد گره از ۲ تا ۱۰ و تعداد دور آموزش از ۵ تا ۵۰ با گام ۵ داده شد. در این مرحله، تنظیمات مختلفی برای همه پارامترها در نظر گرفته شد و خطای مطلق میانگین به عنوان معیار انتخاب شد و با به‌دست آمدن تنظیمات مدل بهینه کمترین مقدار میانگین مربعات خطا محاسبه گردید. جدول (۳-۱۳) بهترین شبکه‌های به دست آمده در حین بهینه سازی را نمایش می‌دهد، که در آن شبکه عصبی با ۶ توصیف‌کننده ورودی، ۷ نرون در لایه پنهان، تعداد دورهای آموزش ۱۰، تابع آموزشی لونبرگ-مارکوات و تابع انتقال تانژانت سیگموئید، کمترین MSE را نسبت به سایر شبکه‌ها داراست و لذا شبکه‌ای با این ساختار می‌تواند به عنوان شبکه بهینه برای مدل‌سازی اندیس بازداری ترکیبات در نظر گرفته شود.

جدول (۳-۱۳)-توابع و پارامترهای بهینه شبکهٔ عصبی مصنوعی

R	MSE	نعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۹۸۰	۰/۰۴۰۶	۵	۶	لگاریتم-سیگموئید	لونبرگ-مارکوات	۵
۰/۹۷۶	**۰/۰۳۴۹**	**۱۰**	**۷**	**تانژانت-سیگموئید**	**لونبرگ-مارکوات**	**۶**
۰/۹۷۶	۰/۰۴۰۵	۲۰	۹	تانژانت-سیگموئید	تنظیم بایزن	۵
۰/۹۷۷	۰/۰۳۷۲	۳۰	۹	لگاریتم-سیگموئید	تنظیم بایزن	۵

۳-۷-۱-ساختار شبکه عصبی بهینه شده

با توجه به نتایج به دست آمده در مرحله قبل، مشخصات شبکه عصبی مصنوعی در جدول (۳-۱۴) نشان داده شده است.

جدول(۳-۱۴)-مشخصات شبکه بهینه (RF-ANN)

تابع آموزش	لونبرگ-مارکوات
تابع انتقال لایه پنهان	تانژانت سیگموئید
تابع انتقال لایه خروجی	Purelin
تعداد گره	۷
تعداد توصیف‌کننده ورودی	۶
تعداد دور آموزشی	۱۰
پارامتر MSE	۰/۳۴۹

۳-۷-۲- ارزیابی مدل شبکه عصبی مصنوعی با توصیف‌کننده‌های RF

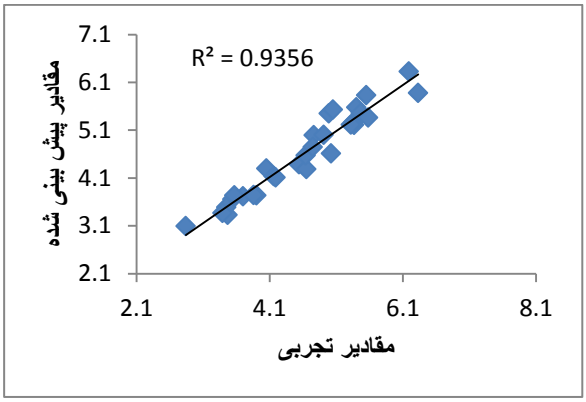
۳-۷-۲-۱-ارزیابی مدل با استفاده از داده‌های سری ارزیابی

جدول (۳-۱۵) نتایج پیش‌بینی حاصل از ارزیابی مدل RF-ANN در پیش بینی اندیس بازداری با استفاده از داده‌های سری ارزیابی و شکل (۳-۱۲) نیز نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری ارزیابی نشان می‌دهد. براساس مشاهدات و ارزیابی‌های انجام

شده می‌توان استناد کرد که روش SR-ANN برای داده‌های سری ارزیابی دارای ضریب تعیین بالاتر و میانگین درصد خطای کمتری به نسبت RF-ANN است و بنابراین مدل برتر است.

جدول (۳-۱۵)-نتایج حاصل از مدل شبکه عصبی مصنوعی با استفاده از داده‌های سری ارزیابی(RF-ANN)

شماره ترکیب	مقدار واقعی	مقدار پیش‌بینی‌شده	درصد خطا
۲	۲/۸۴	۳/۰۹۰	۸/۸۱۶
۵	۳/۴۷	۳/۳۲۲	-۴/۲۴۵
۸	۳/۹	۳/۷۳۳	-۴/۲۷۶
۱۱	۴/۰۵	۴/۲۹۶	۶/۰۸۹
۱۴	۳/۷	۳/۷۱۰	۰/۲۹۱
۱۷	۳/۸۶	۳/۷۴۲	-۳/۰۵۳
۲۰	۵/۰۲	۴/۶۱۱	-۸/۱۴۴
۲۳	۴/۷۶	۴/۹۸۸	۴/۷۹۸
۲۶	۴/۵۴	۴/۳۸۱	-۳/۴۸۵
۲۹	۵/۳۷	۵/۲۰۵	-۳/۰۵۹
۳۴	۵/۳۶	۵/۲۵۷	-۱/۹۰۴
۳۷	۵/۳۲	۵/۲۱۱	-۲/۰۳۰
۳۸	۴/۹۱	۵/۰۰۰	۱/۸۴۶
۴۰	۵/۴۰	۵/۵۷۵	۳/۲۵۶
۴۲	۵/۴۷	۵/۴۵۲	-۰/۳۱۵
۴۴	۴/۹۹	۵/۴۴۴	۹/۱۰۲
۵۱	۵/۰۵	۵/۵۲۸	۹/۴۸۴
۵۳	۶/۳۳	۵/۸۷۳	-۷/۲۱۰
۶۲	۶/۱۹	۶/۳۲۱	۲/۱۲۳
۶۳	۵/۵۵	۵/۸۲۸	۵/۰۱۱
۷۱	۳/۵۴	۳/۶۵۱	۳/۱۵۸
۷۲	۳/۵۷	۳/۷۳۳	۴/۵۹۳
۸۰	۴/۱۶	۴/۱۳۳	-۰/۶۲۸
۹۱	۴/۱۹	۴/۱۰۹	-۱/۹۲۶
۹۹	۴/۶۵	۴/۲۸۱	-۷/۹۱۴
۱۰۴	۳/۴۵	۳/۴۸۷	۱/۰۷۲
۱۰۹	۳/۳۹	۳/۳۶۳	-۰/۷۹۲
۱۲۴	۴/۶۴	۴/۵۷۱	-۱/۴۸۵
۱۳۰	۴/۷۵	۴/۷۴۳	-۰/۱۲۷
۱۵۱	۵/۵۸	۵/۳۵۷	-۳/۹۹۰



شکل (۳-۱۲)-نمودار مقادیر پیش‌بینی شده برحسب مقادیر تجربی برای داده‌های ارزیابی توسط روش RF-ANN

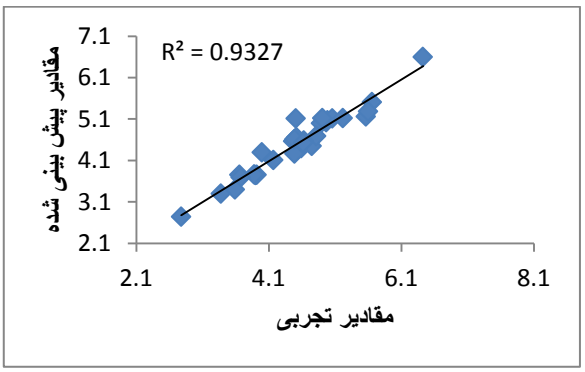
۳-۷-۲-۲-ارزیابی مدل با استفاده از داده‌های سری آزمون

اعتبار و اهمیت مدل با پیش‌بینی اندیس بازداری داده‌هایی که در مدل‌سازی حضور نداشته‌اند مشخص می‌شود. بدین منظور مدل‌های انتخاب شده براساس توصیف‌کننده‌های انتخاب شده توسط RF جهت پیش‌بینی اندیس بازداری ۳۰ ترکیب سری آزمون که در فرآیند مدل‌سازی استفاده نشده-اند، به کار گرفته شد و با استفاده از شبکه عصبی بهینه شده مقادیر اندیس بازداری این ترکیبات پیش‌بینی شد. جدول (۳-۱۶) نتایج حاصل از ارزیابی مدل شبکه عصبی مصنوعی در پیش بینی اندیس بازداری داده‌های سری آزمون را نشان می‌دهد. خطای کم در پیش‌بینی اندیس بازداری ترکیبات مورد بررسی، قدرت پیش بینی مدل را نشان می‌دهد. شکل(۳-۱۳) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد. برتری مدل SR-ANN نسبت به مدل RF-ANN مشهود است.

جدول (۳-۱۶)-نتایج حاصل از ارزیابی مدل شبکه عصبی مصنوعی بااستفاده از داده‌های سری آزمون(RF-ANN)

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی شده	درصد خطا
۳	۲/۷۸	۲/۷۴۳	-۱/۳۲۴
٦	۳/۹۲	۳/۷۵۵	-۴/۱۸۹
۹	۴	۴/۲۹۲	۷/۳۰۰
۱۲	۳/۹	۳/۷۴۷	-۳/۹۰۸

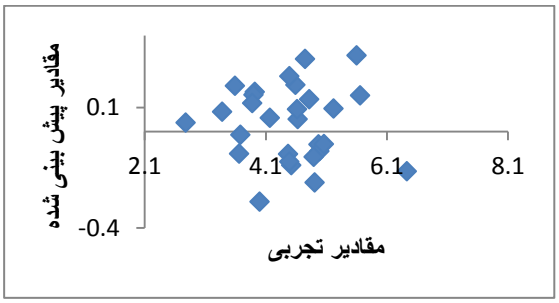
۱۵	۳/۶۶	۳/۷۵۲	۲/۵۲۹
۱۸	۳/۸۸	۳/۷۶۱	-۳/۰۴۲
۲۱	۴/۴۷	۴/۵۶۳	۲/۰۹۹
۲٤	۴/۵۹	۴/۳۹۶	-۴/۲۱۵
۲۷	۴/۹۷	۵/۰۲۳	۱/۰۸۵
۳۰	۴/۹۹	۵/۰۷۰	۱/۶۱۸
۳۱	۵/۰۶	۵/۱۱۱	۱/۰۲۵
۳۲	۴/۸۹	۴/۹۹۵	۲/۱۵۲
۳۳	۵/۲۲	۵/۱۲۵	-۱/۸۱۹
۳۵	۴/۵۱	۵/۱۱۳	۱۳/۳۸۰
۳٦	۴/۹۱	۵/۱۲۲	۴/۳۲۱
٤۱	۵/۶۶	۵/۵۱۰	-۲/۶۴۸
۵۰	۵/۵۷	۵/۱۶۰	-۷/۳۵۳
٦۹	۶/۴۳	۶/۵۹۵	۲/۵۶۷
۷٤	۳/۶۸	۳/۶۹۲	۰/۳۵۱
۸۵	۴/۱۷	۴/۱۱۲	-۱/۳۷۵
۹٦	۴/۴۹	۴/۲۶۰	-۵/۱۰۴
۹۷	۴/۷۵	۴/۴۴۸	-۶/۳۴۱
۱۰۱	۳/۵۹	۳/۳۹۹	-۵/۳۰۳
۱۱۰	۳/۳۸	۳/۲۹۷	-۲/۴۴۲
۱۲۰	۴/۵۲	۴/۶۵۹	۳/۰۹۴
۱۳۰	۴/۸۲	۴/۶۸۵	-۲/۷۸۷
۱٤۰	۴/۶۳	۴/۵۷۹	-۱/۱۰۱
۱٤۳	۴/۴۹	۴/۶۱۵	۲/۷۹۵
۱٤٤	۴/۶۲	۴/۵۲۷	-۲/۰۰۷
۱۵۰	۵/۶	۵/۲۸۴	-۵/۶۴۰



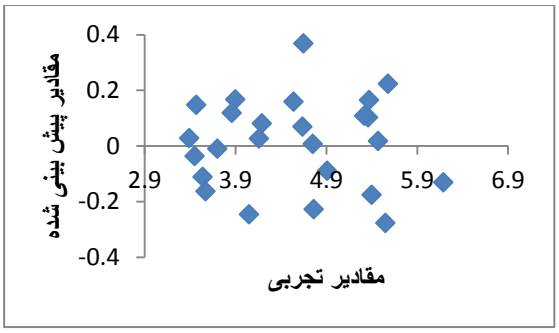
شکل (۳-۱۳)-نمودار مقادیر پیش‌بینی شده‌ی اندیس بازداری برحسب مقادیر تجربی با استفاده از سری داده‌های آزمون توسط روش RF-ANN

۳-۷-۲-۳- ارزیابی مدل توسط نمودار خطای باقی‌مانده

شکل‌های (۳-۱۴) و (۳-۱۵) نمودارهای خطای باقی‌مانده بر حسب مقادیر تجربی، برای سری‌های آزمون و ارزیابی را در مدل RF-ANN نشان می‌دهد. توزیع متقارن داده‌ها حول محور افقی حاکی از عدم وجود خطای سیستماتیک است. مشاهده می‌شود.



شکل (۳-۱۴)_نمودار خطای باقی‌مانده برحسب مقادیر تجربی برای داده‌های سری آزمون (RF-ANN)



شکل (۳-۱۵)_نمودار خطای باقی‌مانده برحسب مقادیر تجربی برای داده‌های سری ارزیابی (RF-ANN)

۳-۷-۲-۴- ارزیابی مدل RF-ANN توسط روش ارزیابی متقاطع

از دیگر روش‌هایی که برای ارزیابی مدل ارائه شده توسط ANN به کار رفته است، تکنیک ارزیابی متقاطع می‌باشد که نتایج حاصل از این روش در جدول (۳-۱۷) و شکل (۳-۱۶) ارائه شده است. این نتایج تعمیم پذیری بالای مدل طراحی شده را بیان می‌کند، اما در مقایسه با ارزیابی متقاطع به‌کار

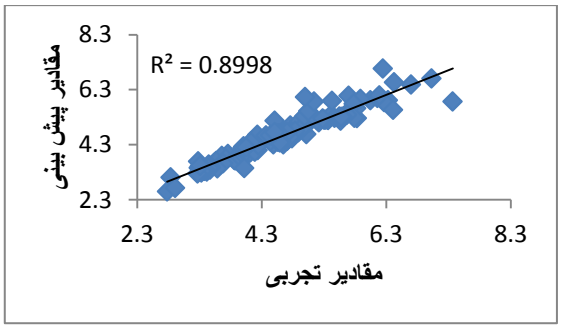
گرفته شده در SR-ANN درصد خطا بیشتر و ضریب تعیین پایین‌تری را داراست. در شکل (۳-۱۷) مقادیر باقی‌مانده بر حسب مقادیر تجربی مورد بررسی ترسیم شده است که تمرکز نقاط به سمت محور صفر، نشان‌دهنده عدم وجود خطای سیستماتیک در مدل است.

جدول (۳-۱۷)-نتایج حاصل از ارزیابی مدل RF-ANN با استفاده از ارزیابی متقاطع

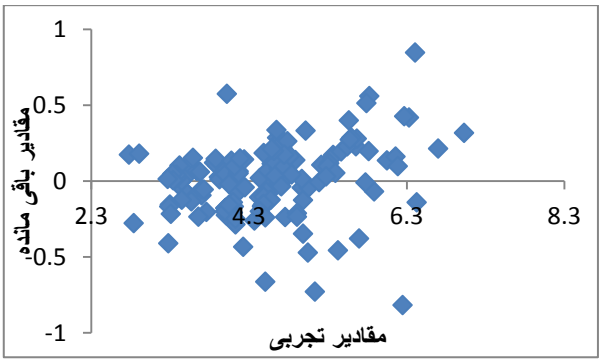
NO.	مقدارتجربی	مقدارپیش‌بینی	درصدخطا	NO.	مقدارتجربی	مقدارپیش‌بینی	درصدخطا
۱	۳/۲۸	۳/۶۹۰	۱۲/۵۱۰	۸۱	۴/۱۶	۴/۲۳۵	۱/۸۰۵
۲	۲/۸۴	۳/۱۱۷	۹/۷۶۲	۸۲	¼	۴/۲۴۰	۳/۴۳۵
۳	۲/۷۸	۲/۶۰۵	-۶/۲۷۹	۸۳	۴	۴/۱۷۹	۴/۴۸۲
۴	۲/۹۱	۲/۷۳۰	-۶/۱۷۰	۸۴	۴/۰۱	۴/۱۸۵	۴/۳۸۰
۵	۳/۴۷	۳/۳۸۸	-۲/۳۴۸	۸۵	۴/۱۷	۴/۱۰۵	-۱/۵۳۶
۶	۳/۹۲	۳/۸۹۲	-۰/۶۹۷	۸۶	۴/۱۷	۴/۱۱۸	-۱/۲۴۳
۷	۳/۷۶	۳/۹۶۳	۵/۴۱۹	۸۷	۴/۱۲	۴/۰۲۲	-۲/۳۶۴
۸	۳/۹	۳/۷۹۶	-۲/۶۴۱	۸۸	۴/۰۵	۴/۰۷۰	۰/۵۰۱
۹	۴	۴/۱۹۸	۴/۹۶۵	۸۹	¼	۴/۳۰۷	۵/۰۶۵
۱۰	۴/۱۳	۴/۴۱۵	۶/۹۰۷	۹۰	۴/۲۴	۴/۰۹۶	-۳/۳۷۵
۱۱	۴/۰۵	۴/۲۰۷	۳/۸۹۳	۹۱	۴/۱۹	۴/۰۴۱	-۳/۵۵۰
۱۲	۳/۹	۳/۷۹۲	-۲/۷۴۳	۹۲	۴/۲۵	۴/۲۹۱	۰/۹۷۵
۱۳	۳/۷	۳/۷۹۳	۲/۵۱۹	۹۳	۴/۶۶	۴/۳۷۳	-۶/۱۳۸
۱۴	۳/۷	۳/۷۵۳	۱/۴۴۳	۹۴	۴/۷۱	۴/۴۳۲	-۵/۸۹۳
۱۵	۳/۶۶	۳/۸۹۴	۶/۴۰۴	۹۵	۴/۵۶	۴/۴۰۵	-۳/۳۹۵
۱۶	۳/۷۳	۳/۸۷۶	۱/۵۱۵	۹۶	۴/۴۹	۴/۳۰۵	-۴/۱۰۹
۱۷	۳/۸۶	۳/۷۳۷	-۳/۱۶۲	۹۷	۴/۷۵	۴/۶۱۷	-۲/۷۹۲
۱۸	۳/۸۸	۳/۷۳۲	-۳/۷۹۵	۹۸	۴/۷۳	۴/۵۱۳	-۴/۵۷۶
۱۹	۴/۶۳	۴/۶۶۸	۰/۸۴۱	۹۹	۴/۶۵	۴/۳۱۳	-۷/۲۲۵
۲۰	۵/۰۲	۴/۶۸۷	-۶/۶۲۵	۱۰۰	۴/۶۶	۴/۴۴۶	-۴/۵۷۸
۲۱	۴/۴۷	۴/۶۳۲	۳/۶۳۷	۱۰۱	۳/۵۹	۳/۴۳۵	-۴/۲۸۹
۲۲	۴/۴۷	۴/۶۳۷	۳/۷۵۶	۱۰۲	۳/۴۸	۳/۴۶۲	-۰/۵۱۰
۲۳	۴/۷۶	۴/۹۹۷	۴/۹۸۱	۱۰۳	۳/۴۷	۳/۵۲۵	۱/۵۹۶
۲۴	۴/۵۹	۴/۴۷۴	-۲/۵۲۳	۱۰۴	۳/۴۵	۳/۵۷۲	۳/۵۶۰
۲۵	۴/۴۳	۴/۵۳۶	۲/۳۹۷	۱۰۵	۳/۴۵	۳/۴۴۹	-۰/۰۲۷۵
۲۶	۴/۵۴	۴/۵۳۷	-۰/۰۶۱۴	۱۰۶	۳/۴۲	۳/۳۱۷	-۳/۰۰۹
۲۷	۴/۹۷	۵/۰۱۴	۰/۸۸۹	۱۰۷	۴/۰۲	۳/۴۴۴	-۱۴/۳۲۳
۲۸	۴/۶	۴/۵۴۳	-۱/۲۳۸	۱۰۸	۳/۳	۳/۴۵۴	۴/۶۷۶
۲۹	۵/۳۷	۵/۱۹۸	-۳/۱۸۶	۱۰۹	۳/۳۹	۳/۳۶۸	-۰/۶۴۷
۳۰	۴/۹۹	۵/۱۱۵	۲/۵۲۴	۱۱۰	۳/۳۸	۳/۳۳۸	-۱/۲۲۴
۳۱	۵/۰۶	۵/۱۰۱	۰/۸۱۲	۱۱۱	۳/۳۳	۳/۲۸۳	-۱/۳۹۷
۳۲	۴/۸۹	۴/۷۵۱	-۲/۸۲۷	۱۱۲	۳/۲۹	۳/۴۵۷	۵/۰۸۳

۳۳	۵/۲۲	۵/۱۱۳	-۲/۰۳۸	۱۱۳	۳/۳۴	۳/۳۳۵	-۰/۱۳۱
۳۴	۵/۳۶	۵/۲۰۲	-۲/۹۳۹	۱۱۴	۳/۲۷	۳/۲۵۶	-۰/۴۲۳
۳۵	۴/۵۱	۵/۱۷۲	۱۴/۶۹۷	۱۱۵	۳/۳۱	۳/۵۲۶	۶/۵۵۰
۳۶	۴/۹۱	۵/۱۲۲	۴/۳۱۹	۱۱۶	۴	۴/۱۹۹	۴/۹۷۷
۳۷	۵/۳۲	۵/۲۰۱	-۲/۲۱۸	۱۱۷	۴/۲۳	۴/۶۶۴	۱۰/۲۷۲
۳۸	۴/۹۱	۵/۱۴۵	۴/۷۸۹	۱۱۸	۴/۴۴	۴/۴۲۰	-۰/۴۳۷
۳۹	۵/۳	۵/۲۰۱	-۱/۸۵۶	۱۱۹	۴/۵۴	۴/۴۷۰	-۱/۵۳۵
۴۰	۵/۴	۵/۳۴۶	-۰/۹۹۴	۱۲۰	۴/۵۲	۴/۵۹۱	۱/۵۸۶
۴۱	۵/۶۶	۵/۴۳۱	-۴/۰۴۴	۱۲۱	۴/۵۶	۴/۶۸۹	۲/۸۳۴
۴۲	۵/۴۷	۵/۲۷۴	-۳/۵۷۰	۱۲۲	۴/۶۸	۴/۵۹۹	-۱/۷۱۷
۴۳	۴/۹۷	۴/۹۵۶	-۰/۲۶۸	۱۲۳	۴/۷	۴/۷۲۸	۰/۶۰۵
۴۴	۴/۹۹	۵/۳۳۶	۶/۹۴۷	۱۲۴	۴/۶۴	۴/۵۱۹	-۲/۵۹۰
۴۵	۵/۱۹	۵/۱۹۶	۰/۱۲۹	۱۲۵	۴/۴۹	۴/۵۹۴	۲/۳۱۸
۴۶	۴/۷۶	۴/۶۸۵	-۱/۵۵۴	۱۲۶	۴/۳۷	۴/۶۳۰	۵/۹۶۳
۴۷	۴/۶۴	۴/۶۱۴	-۰/۵۴۷	۱۲۷	۴/۱۳	۴/۳۷۷	۵/۹۹۹
۴۸	۴/۵۱	۴/۶۴۲	۲/۹۴۸	۱۲۸	۴/۷۷	۴/۷۴۶	-۰/۵
۴۹	۵/۸۳	۵/۲۶۹	-۹/۶۱۰	۱۲۹	۴/۷۹	۴/۵۲۶	-۵/۵۰۳
۵۰	۵/۵۷	۵/۱۷۰	-۷/۱۶۶	۱۳۰	۴/۷۵	۴/۷۲۳	-۰/۵۵۴
۵۱	۵/۰۵	۵/۵۲۲	۹/۳۴۸	۱۳۱	۴/۶۸	۴/۶۹۰	۰/۲۳۴
۵۲	۶/۰۵	۵/۹۱۴	-۲/۲۳۹	۱۳۲	۴/۷۴	۴/۶۶۰	-۱/۶۸۳
۵۳	۶/۳۳	۵/۹۱۰	-۶/۶۲۱	۱۳۳	۴/۷۴	۴/۷۲۵	۰/۳۰۱
۵۴	۵/۸۲	۵/۶۲۳	-۳/۳۸۱	۱۳۴	۴/۷۵	۴/۶۵۴	-۲/۰۲۰
۵۵	۵/۱۴	۵/۸۶۸	۱۴/۱۷۴	۱۳۵	۴/۸۲	۴/۶۵۴	-۳/۴۳۷
۵۶	۶/۱۶	۵/۹۹۹	-۲/۶۰۶	۱۳۶	۴/۸۴	۴/۷۹۵	-۰/۹۱۵
۵۷	۵/۷	۶/۰۷۹	۶/۶۶۲	۱۳۷	۴/۰۵	۴/۲۰۷	۳/۸۹۳
۵۸	۵	۶/۰۳۴	۲۰/۶۹۹	۱۳۸	۴/۴۸	۴/۶۳۴	۳/۴۴۷
۵۹	۵/۸۹	۵/۹۵۸	۱/۱۶۷	۱۳۹	۴/۷۵	۴/۵۷۶	-۳/۶۴۷
۶۰	۵/۷۸	۵/۷۸۹	۰/۱۵۶	۱۴۰	۴/۶۳	۴/۵۹۷	-۰/۷۰۶
۶۱	۶/۲۷	۵/۸۴۳	-۶/۸۰۳	۱۴۱	۴/۷۱	۴/۷۴۴	۰/۷۳۲
۶۲	۶/۱۹	۶/۰۹۰	-۱/۶۰۰	۱۴۲	۴/۵۸	۴/۷۰۳	۲/۶۸۹
۶۳	۵/۵۵	۵/۳۲۰	-۴/۱۳۹	۱۴۳	۴/۴۹	۴/۶۴۲	۳/۳۹۰
۶۴	۶/۴۱	۵/۵۶۱	-۱۳/۲۳۱	۱۴۴	۴/۶۲	۴/۴۱۹	-۴/۳۳۸
۶۵	۶/۲۵	۷/۰۶۷	۱۳/۰۷۲	۱۴۵	۴/۵۹	۴/۳۷۶	-۴/۶۴۳
۶۶	۵/۴۳	۵/۸۸۶	۸/۴۱۰	۱۴۶	۴/۴۴	۴/۶۳۹	۴/۴۹۵
۶۷	۶/۷	۶/۴۸۶	-۳/۱۹۲	۱۴۷	۴/۵۱	۴/۷۴۹	۵/۳۰۸
۶۸	۷/۰۳	۶/۷۱۱	-۴/۵۲۳	۱۴۸	۵/۷۹	۵/۲۷۸	-۸/۸۹۴
۶۹	۶/۴۳	۶/۵۷۰	۲/۱۷۷	۱۴۹	۵/۶۷	۵/۳۹۰	-۴/۹۲۳
۷۰	۷/۳۷	۵/۸۶۷	-۲۰/۳۸۱	۱۵۰	۵/۶	۵/۳۱۴	-۵/۰۹۹
۷۱	۳/۵۴	۳/۶۳۴	۲/۶۷۹	۱۵۱	۵/۵۸	۵/۳۰۵	-۴/۹۱۲
۷۲	۳/۵۷	۳/۷۰۴	۳/۷۶۷	۱۵۲	۵/۲۸	۵/۲۴۸	-۰/۵۸۷

۷۳	۳/۶۴	۳/۵۷۳	-۱/۸۲۷				
۷۴	۳/۶۸	۳/۶۲۰	-۱/۶۲۷				
۷۵	۳/۹۳	۳/۹۱۹	-۰/۲۷۳				
۷۶	۴	۳/۹۵۳	-۱/۱۵۰				
۷۷	۴/۰۸	۳/۹۴۶	-۳/۲۷۴				
۷۸	۴/۰۱	۴/۲۳۴	۵/۶۰۷				
۷۹	۴/۱۱	۴/۲۷۸	۴/۰۹۱				
۸۰	۴/۱۶	۴/۱۰۹	-۱/۲۰۸				



شکل(۳-۱۶)-نمودار مقادیر پیش‌بینی شده بر حسب مقادیر تجربی برای ارزیابی مدل RF-ANN به روش ارزیابی متقاطع برای کل داده‌ها



شکل(۳-۱۷)-نمودار مقادیر باقی‌مانده بر حسب مقادیر تجربی برای ارزیابی مدل RF-ANN به روش ارزیابی متقاطع برای کل داده‌ها

۳-۸- مقایسهٔ مدل‌های ارائه شده با استفاده از پارامترهای آماری

همانطور که در مباحث قبل بیان شد، پارامترهای آماری مقادیر تعیین کننده‌ای برای ارزیابی مدل‌ها هستند. مطابق جدول (۳-۱۹) هفت پارامتر آماری عنوان شده جهت ارزیابی توانایی پیشگویی مدل‌های ساخته شده به روش‌های SR-ANN و RF-ANN و جنگل تصادفی به کاربرده شد. باتوجه به

نتایج موجود در جدول (۳-۱۹)، ضرایب تعیین بزرگتر و خطای استاندارد کوچکتر سری آزمون و ارزایابی برای SR-ANN، برتری روش شبکه عصبی مصنوعی در پیش‌بینی اندیس بازداری نشان می دهد. این پارامترها طبق روابط توضیح داده‌شده در بخش۲-۳-۶ محاسبه‌شده‌اند.

جدول (۳-۱۹)-محاسبهٔ پارامترهای آماری برای داده‌های آزمون و ارزیابی

پارامترآماری		R2	PRESS	MSE	SEP	MAE	REP%	MRE	F	FIT
SR-ANN	سری ارزیابی	۰/۹۷۱	۰/۶۵۷	۰/۰۲۱۹	۰/۱۴۸	۰/۱۰۴	۳/۲۱۸	۲/۲۶۷	۶۵۱/۶۸۰	۷/۴۰
	سری تست	۰/۹۵۸	۰/۷۲۴	۰/۰۲۴۱	۰/۱۵۵	۰/۱۰۳۴	۳/۴۲۳	۲/۲۸۵		
RF-ANN	سری ارزیابی	۰/۹۳۵	۱/٤۹۷	۰/۰٤۹	۰/۲۲۳	۰/۱۷٦	٤/۸٥٦	۳/۸۰۷	۲۲٤/۸۷٥	٤/٤۰
	سری تست	۰/۹۳۲	۱/۲۱۸	۰/۰٤۰٦	۰/۲۰۱	۰/۱٦۰	٤/٤٤۱	۳/٤۹۷		
RF	سری تست	۰/۹٤٦	۰/۹۳۹	۰/۰۳۱۳	۰/۱۷۷	۰/۱۱۷	۳/۸۹۹	۲/۷٥۱	۳٦۱/۳۲۲	۷/۰۲

۳-۹- ارزیابی مدل با استفاده از Y-تصادفی

این تکنیک ارزیابی مدل با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام می‌شود. در این آزمون مقادیر تصادفی از متغیر وابسته تولید گردید. اگر مدل اصلی هیچگونه ارتباط تصادفی نداشته باشند، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSPR با پاسخ‌های تصادفی توسعه یافته وجود دارد. نتایج حاصل از چندین بار اجرای آزمون Y-تصادفی برای سری آزمون و ارزیابی در روش SR-ANN و نیز آزمون -Yتصادفی برای داده های سری ارزیابی و آزمون در روش RF-ANN به ترتیب در جدول‌های (۳-۲۰) و (۳-۲۱) آورده شده است. مقادیر کوچک ضریب تعیین (R^2) بیانگر عدم وجود ارتباط یا وابستگی ساختاری در مدل توسعه یافته توسط شبکه می‌باشد، همانطور که مشاهده می‌شود در روش SR-ANN مقادیر R^2 کوچکتر است و تاییدی بر بهتر بودن این روش نسبت به RF-ANN می‌باشد.

جدول (۳-۲۰)-مقادیر R^2 برای داده‌های سری ارزیابی و آزمون بعد از ۱۰ بار اجرا Y-تصادفی در روش (SR-ANN)

تکرار	سری آزمون	سری ارزیابی
۱	۰/۰۰۶۴	۰/۰۳۸۳
۲	۰/۰۰۵۵	۰/۰۷۴۳
۳	۰/۰۶۸۳	۰/۰۰۱۶
۴	۰/۰۰۷۷	۰/۰۱۹۱
۵	۰/۰۱۱۷	۰/۰۰۰۰۲
۶	۰/۰۰۰۷	۰/۰۲۸۱
۷	۰/۰۰۲۵	۰/۰۰۰۲
۸	۰/۰۰۵۳	۰/۰۳۹۹
۹	۰/۰۰۱۶	۰/۰۵۸۴
۱۰	۰/۰۱۵۵	۰/۰۳۱

جدول (۳-۲۱)-مقادیر R^2 برای داده‌های سری ارزیابی و آزمون بعد از ۱۰ بار اجرا Y-تصادفی در روش (RF-ANN)

تکرار	سری آزمون	سری ارزیابی
۱	۰/۰۰۱۴	۰/۰۳۳۰
۲	۰/۰۰۵۶	۰/۰۰۰۰۶
۳	۰/۱۲۲	۰/۰۲۲۱
۴	۰/۰۷۲۰	۰/۰۰۶۲
۵	۰/۰۳۲۰	۰/۱۳۸
۶	۰/۰۰۶۶	۰/۰۰۰۸
۷	۰/۰۴۲۱	۰/۰۵۲۹
۸	۰/۰۰۱۲۷	۰/۰۲۷۸
۹	۰/۰۰۰۳۲	۰/۰۲۲۹
۱۰	۰/۰۲۰۱	۰/۰۶۰۹

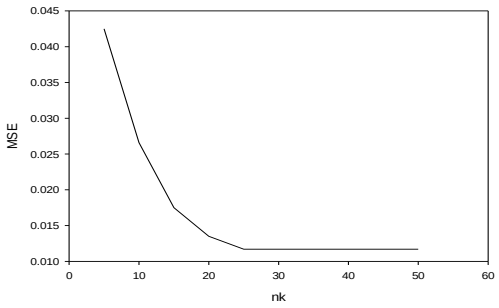
MARS-۱۰-۳

برای مدل‌سازی توسط مارس ابتدا باید پارامترهای موثر بهینه شوند، برای انجام این‌کار ابتدا کل داده‌ها به دو سری آزمون(۳۰ ترکیب) و سری آموزش (۱۲۲ ترکیب) تقسیم بندی شده‌اند و مقادیر مربوط به ۱۹۲ توصیف‌کننده به عنوان متغیر مستقل و مقادیر اندیس بازداری به عنوان متغیر وابسته منسوب شدند. برای برازش مدل مارس، همانطور که ذکر شد، پارامتری کلیدی این روش یعنی تعداد توابع پایه در گام پیش‌رو یعنی nk باید بهینه شود. با در نظر گرفتن nk از ۵ تا ۵۰ نتایج جدول (۳-۱۸) به دست آمده است. با توجه به جدول (۳-۱۸)، با تغییر nk مقادیر MSE و R-Square همگام با nk تغییر می‌کنند، در ابتدا این تغییر قابل محسوس است اما با افزایش بیشتر nk تغییری در آن‌ها

احساس نمی‌شود تا جایی که ثابت می‌شوند (با توجه به شکل (۳-۱۸)). به همین ترتیب مدل بهینه انتخاب شده مدلی است که دارای بیشترین مقدار ضریب تعیین(R-square) و کمترین مقدار اعتبار سنجی(GCV) با تنظیمات nk=۲۵ حاصل می‌شود. این مدل براساس اینکه یک مدل بهینه در مجموعه داده‌های مدل‌ساز می باشد، مشاهده می‌شود بالاترین R^2 را برای داده‌های آموزش ایجاد کرده و به عنوان مدل برتر شناخته می‌شود، اما برای داده‌های سری آزمون با مقدار ضریب تعیین ۰/۹۱۴ و با توجه به شکل (۳-۱۹) که مقدار پیش‌بینی شده با مقدار واقعی تطابق زیادی ندارند روش MARS برازش خوبی برای داده‌های آزمون ایجاد نکرده است.

جدول (۳-۱۸)-مقادیر R-square، GCV، RSS، MSE برای دادهای آزمون و آموزش براساس مقادیر مختلف nk

Training set				Testset	
nk	R-squre	MSE	GCV	R-squre	MSE
۵	۰/۹۴۰	۰/۰۴۵۲	۰/۰۵۰۹	۰/۹۳۰	۰/۰۵۴۱
۱۰	۰/۹۶۴	۰/۰۲۶۶	۰/۰۳۴۵	۰/۹۴۴	۰/۰۴۲۵
۱۵	۰/۹۷۶	۰/۰۱۷۵	۰/۰۲۶۶	۰/۹۱۸	۰/۰۵۶۹
۲۰	۰/۹۸۲	۰/۰۱۳۵	۰/۰۲۲۳	۰/۹۰۶	۰/۰۶۵۱
۲۵	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷
۳۰	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷
۳۵	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷
۴۰	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷
۴۵	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷
۵۰	۰/۹۸۴	۰/۰۱۱۷	۰/۰۲۱۹	۰/۹۱۴	۰/۰۵۸۷



شکل (۳-۱۸)- نمودار میانگین خطای برحسب توابع پایه (nk)

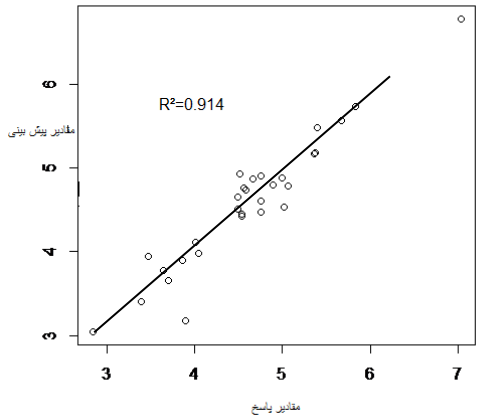
مدل نهایی مارس بعد از طی مراحل پیش‌رو و پس‌رو به صورت زیر است:

$$Y=4.78+0.53h(X3V-3.93)-0.82h(PJI3-0.79)-0.045h(4.87-LBW)-0.157h(LBW-4.87)+0.28h(0.822-AROM)-13.94h(AROM-0.822)-0.24h(4.169-RDF040V)-0.076h(RDF040V-4.16)-0.12h(1.554-RDF115e)-0.58h(RDF11e-1.55)-2.66h(0.029-Mor19m)-0.46h(Mor19m-0.029)+3.3h(0.154-Mor11p)-0.14h(15.06-PCWTe)-0.11h(4-F04CS)+0.14h(F04CS-4)$$

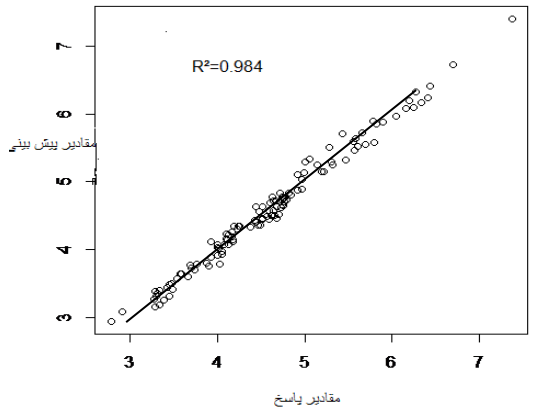
که در آن

$$h(u) = u_+ = \begin{cases} u & \text{اگر } u > 0 \\ 0 & \text{در غیر اینصورت} \end{cases}$$

عملکرد مدل MARS را می‌توان در شکل‌های (۳-۱۹) و (۳-۲۰) به وسیله نمودارهای مقادیر پیش بینی‌شده در مقابل مقادیر پاسخ دید. که می‌توان گفت مدل MARS برازش بهتری را به داده‌ها انجام داده است



شکل(۳-۱۹)-نمودار مقادیر پاسخ در مقابل مقادیر برازش داده شده توسط روش مارس برای داده‌های سری آزمون



شکل(۳-۲۰)-نمودار مقادیر پاسخ در مقابل مقادیر برازش داده شده توسط روش مارس برای داده‌های سری آموزش

۳-۱۱- نتیجه گیری مدل‌سازی با مجموع توصیف‌کننده‌های انتخاب‌شده توسط رگرسیون گام‌به‌گام

ارزیابی عملکرد مدل‌های مختلف به کمک فاکتور R^2 امکان پذیر است. این فاکتور زمانی می‌تواند ارزش رجحانی مدل‌ها را تعیین نماید که بین مقادیر مشاهده شده و مقادیر پیش‌بینی شده محاسبه شوند. توجه به مقادیر R^2 مزیت نسبی مدل‌ها را نشان می‌دهد. بیشترین ضریب تعیین معنی‌دار بهترین و کارآمدترین روش را معرفی می‌نماید بنابراین، ارزش رجحانی روش‌ها به ترتیب MARS و SR-ANN و RF وRF-ANN با میزان ضرائب تعیین ۰/۹۸۴، ۰/۹۷۱، ۰/۹۴۴، ۰/۹۳۸ انتخاب می‌گردد. شکل‌های (۳-۳) و (۳-۵) و (۳-۶) و (۳-۱۱) و (۳-۱۲) و(۳-۱۸) و (۳-۱۹) ارتباط بین مقادیر مشاهده شده و مقادیر پیش‌بینی در مدل‌های چهارگانه تحقیق براساس ضریب تعیین را نشان می‌دهد. در مقایسه کارایی این مدل‌ها هرچه نتایج پیش‌بینی بیشتر و منطبق بر نتایج مشاهده شده باشد مدل از کارایی مطلوب‌تری برخوردار است و همچنین مقدار MSE بدست آمده برای این مدل ها به این شرح است. روش‌ها به ترتیب MARS و SR-ANN و RF وRF-ANN با میزان میانگین مربعات خطا ۰/۰۱۱۷، ۰/۰۲۴۱،۰/۰۳۱۳، ۰/۰۴۰۶ انتخاب می‌گردد.

۳-۱۲- بررسی ارتباط بین توصیف‌کننده‌های منتخب و خاصیت مورد نظر

در این بخش بعضی از توصیف‌کننده‌های وارد شده در مدل MARS که به عنوان توصیف‌کننده ورودی در SR-ANN نیز محسوب می‌شود آورده شده است، که به بررسی اجمالی روی اثرات این متغیرها بر خاصیت ترکیبات مورد مطالعه پرداخته شده است. هرکدام از این توصیف‌کننده‌ها بیانگر خصوصیت دو بعدی، هندسی و سه بعدی ترکیبات مورد بررسی می‌باشند.

۳-۱۲-۱-توصیف‌کننده‌های2D frequency finger prints

این توصیف‌کننده‌ها اندازه مولکول هارا بررسی می‌کند. توصیف‌کننده اثرانگشت فرکانس که به بررسی مجموع فرکانس دو اتم در فاصله هندسی آن اتم‌ها می پردازد، اندازه اتم‌ها را به‌طور مستقیم بررسی می‌کند، که از میان توصیف‌کننده‌های اثرانگشت F04CS که فرکانس S-C را در فاصله ای هندسی نشان میدهد[۵۲].

۳-۱۲-۲-توصیف کننده‌های Topological

این توصیف‌کننده بر اساس نمایش گراف مولکول می‌باشد. در این گراف‌ها هر نقطه نشان دهندهٔ یک اتم بوده و خطوط بین نقاط نیز نشان دهندهٔ پیوند شیمیایی بین اتم‌ها می‌باشد. معمولا در گراف های مولکولی اتم هیدروژن را نشان نمی‌دهد. این توصیف‌کننده اطلاعاتی راجع به ساختمان، اندازه، شکل، تقارن، شاخه دار شدن، نحوهٔ اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول در اختیار ما قرار می‌دهند. محاسبهٔ این توصیف‌کننده به سادگی از روی ساختمان دو بعدی مولکول امکان پذیر می‌باشد. از میان این توصیف‌کننده‌ها PW3 است[۵۳].

۳-۱۲-۳- توصیف کننده‌های 3D MORSE

این توصیف‌کننده ساختار ۳ بعدی مولکول را بر اساس تفرق الکترون نشان می‌دهند. که توصیف‌کننده‌ها Mor19mو Mor29m را شامل می‌شود[۵۳].

۳-۱۲-۴- توصیف‌کننده‌های RDF[1]

این توصیف‌کننده می‌تواند تابع توزیع شعاعی یک ترکیب شامل N اتم را به احتمال یافتن یک اتم دریک حجم کروی با شعاع r ارتباط دهد. اهمیت این توصیف‌کننده‌ها به دلیل اختلاف در توزیع اتم‌ها در مولکول ها می‌باشد. از میان این توصیف‌کننده‌های RDF115e، دیده شده است[۵۳].

۳-۱۲-۵- توصیف‌کننده‌های Charge

با توجه به تئوری کلاسیک، برهم کنش‌های بین مواد از نظر ماهیّت، الکتروستاتیکی یا کووالانسی است. تغییرات بار الکتریکی در ملکول به عنوان نیروی پیش برندهٔ برهم‌کنش‌های الکتروستاتیک است. از جملهٔ این توصیف‌کننده PCWTe، شاخص دو قطبی محلی می‌باشد. علامت منفی آن در مدل، اهمیت برهم‌کنش‌های الکتروستاتیک بین بازدارنده و ملکول هدف را نشان می‌دهد و به عنوان یک شاخص برای اندازه‌گیری میزان این برهم‌کنش‌های الکتروستاتیک به کار می‌رود[۵۴].

۳-۱۲-۶- توصیف‌کننده‌های GATAWAY[2]

این توصیف‌کننده‌ها با توجه به مختصات فضایی اتم‌ها در یک ملکول به راحتی قابل محاسبه هستند. این توصیف‌کننده‌ها براساس ماتریس قدرت نفوذ[3] تعریف می‌شوند و بیان‌کنندهٔ ویژگی‌های هندسی و توپولوژیک ملکول‌ها هستند. توصیف‌کننده‌های GATAWAY به دو زیر مجموعهٔ H و R تقسیم می‌شوند.

ماتریس نفوذ یا ماتریس تاثیر ملکول (MIM) [4] با رابطهٔ زیر بیان می‌شود.

[1] Radial distribution function
[2] Geometry, Topology, and Atom-weights Assembly
[3] Leverage matrix
[4] Molecular influence matrix

$$H = M.(M^T.M)^{-1}.M^T$$ رابطه(۳-۱)

که M ماتریس مختصات اتمی، T به معنای ماتریس ترانهاده[1] و H ماتریس قدرت نفوذ است. H یـک ماتریس متقارن A*A است که A تعداد اتم‌هاست. عناصر قطری ماتریس H لوریج‌ها نام دارند که هر یک بیانگر ماتریس تاثیر-فاصله به صورت زیر تعریف می شود

$$[R]_{ij} = [\frac{\sqrt{h_{ii}.h_{jj}}}{r_{ij}}]_{ij}\ i \neq j$$ رابطه(۳-۲)

h_{jj} و h_{ii} لوریج‌های دو اتم i و j و r_{ij} فاصله آن دو است. عناصر قطری این ماتریس صفر است. سه توصیف‌کننده R6u و R2m و H6m از این کلاس در مدل‌سازی به کار گرفته شده است. u در این توصیف‌کننده‌ها نشان‌دهندهٔ آن است که آن‌ها با ویژگی خاصی از اتم‌ها محاسبه نشده‌اند[۵۴].

۳-۱۲-۷-توصیف‌کننده‌های connectivity indices

X3v توصیف‌کننده‌ای از گروه اندیس‌های ارتباط مولکولی است. این اندیس‌ها دسته‌ای از توصیف‌کننده‌های توپولوژیکی هستند که اطلاعاتی در مورد ساختمان، اندازه و میزان شاخه‌ای شدن مولکول، نحوه‌ی اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول را در اختیار قرار می‌دهند. جهت محاسبه‌ی این توصیف‌کننده‌ها ابتدا باید ساختار مولکول را با استفاده از تئوری گراف شیمیایی رسم نمود.

این اندیس‌ها اولین بار به‌وسیله راندیک[2] در سال ۱۹۷۵ ارائه شدند و سپس در سال ۱۹۷۶ توسط هال[3] اصلاح گردیدند. این توصیف‌کننده‌ها با رابطه‌ی (۳-۳) قابل محاسبه هستند:

[1] Transpose
[1]- Randic
[2]- Hall

$$ {}^{m}\chi^{v} = \sum_{i=1}^{N_s} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v} \right)^{1/2} \qquad (۳-۳) $$

که در این رابطه δ_k^v، درجه رأس ظرفیت[۳] اتم kام در گراف مولکولی است و به شکل رابطه‌ی زیر تعریف می‌شود:

$$ \delta_k^v = \frac{\left(Z_K^V - h_k \right)}{\left(Z - Z_k^v - 1 \right)} \qquad (۴-۳) $$

که در آن Z عدد اتمی اتم k، Z_k^v تعداد الکترون‌های لایه‌ی ظرفیت اتم k و h_k تعداد اتم‌های هیدروژن متصل به اتم غیر هیدروژنی k است.

[۳]- Valence vertex degree

۳-۱۳-نتیجه گیری:

در کار ارائه شده مطالعه ارتباط کمی ساختار-خاصیت جهت پیش بینی اندیس بازداری دسته وسیعی از ترکیبات آلی انجام شد. در ابتدا توصیف کننده هایی ایجاد و مناسب ترین آنها که بیشترین ارتباط را با اندیس بازداری داشتند به کمک روش رگرسیون گام‌به‌گام انتخاب گردیدند و به روش‌هایی از جمله جنگل تصادفی و شبکه عصبی مصنوعی و مارس مدل مناسب ایجاد شد. مدل ایجاد شده دارای توصیف‌کننده می باشد که این توصیف کننده ها از دسته های گوناگون از جمله توپولوژیکی، هندسی، ساختاری و.... می باشند. با استفاده از این توصیف کننده ها مدلی ایجاد شد که قدرت پیش بینی اندیس بازداری ترکیبات مشابه را داشت. بعضی از این مدل‌ها با استفاده از روش‌های مختلف از جمله ارزیابی متقاطع مورد ارزیابی قرار گرفت و همچنین پارامترهای آماری مختلف نشان داد که مدل های ارائه شده می تواند جهت پیش بینی اندیس بازداری سایر ترکیبات قرارگیرد. در ضمن از این روش می توان جهت مدل سازی و پیش بینی سایر خواص فیزیکوشیمیایی استفاده نمود.

آینده نگری:

- در این تحقیق نتایج نشان می‌دهد که مدل‌های خطی و غیر خطی توسعه یافته می‌توانند به‌عنوان روش‌های موفق برای مدل‌سازی و پیش‌بینی خاصیت مورد نظر (اندیس بازداری) ترکیبات مورد مطالعه به کار روند.
- روش جنگل‌های تصادفی را می‌توان برای پیش‌بینی اندیس بازداری یا سایر خواص ترکیبات دیگر که هنوز سنتز نشده‌اند استفاده نمود
- علاوه بر روش رگرسیون گام‌به‌گام که در این تحقیق به‌عنوان روش انتخاب توصیف‌کننده استفاده شده است می‌توان از سایر روش‌های انتخاب متغیر مانند اجتماع مورچگان و راهبرد CDFS و طرح‌ریزی پیاپی استفاده کرد.
- می‌توان از روش‌های خطی دیگر مانند حداقل مربعات جزئی و آنالیز اجزای اصلی را جایگزین روش‌های جنگل‌های تصادفی کرد و نتایج به‌دست آمده را باهم مقایسه کرد.

منابع:

[1]-راجر، ام اسمیت، (۱۳۸۳)،" کروماتوگرافی گاز و مایع در شیمی تجزیه"، ارباب زوارم، سرفراز یزدی ع ، انتشارات دانشگاه فردوسی مشهد،۴۸۶، ص ۱۸۵-۱۸۴

[2]-Gupta N., Roychoudhury Pk., Deb Jk. Biotechnology of desulfurization on diesel; Prospects and challenges. Applied Environmental Microbiology 2005; 66: 356-66

[3]-Baldi F., Pepi M ., Fara F. Growth of Rhodosporidium. Toruloides strain DBVPG 6662 on Dibenzothiophene Crystals and Orimulsion. Applied Environmental Microbiology 2003; 69(8): 4689-96.

[4]-Monticello Dj. Biodesulfurization and the upgrading of petroleum distiiates. Current opinion Microbiology 2000;11:540-6.

[5]-Schemidt M., Siebert W., Bangall Kw. The chemistry of Sulphur., Selenium., tellurium and polonium pergamon Texin Inorganic chemistry. 115, New York: Pergamon Press, oxford; 1973

[6]-Prayuen Yongp. Coal biodesulfurization Process. Science thechnology 2012; 24: 493-507

[7]-Kilbane JJ. Microbial bi Octalyst developments toupgrade fossil fuel. Energy Biotechnology 2006;17: 305-14.

[8]-Mei H., Mei1 B.w. and Yent.F.," A new method for obtaining ultra-low sulfur diesel fuel via ultra-sound as-sisted oxidative desulfurization" Fuel 82,PP. 405-414,2003.

[9]-Lin L., Hong L., Janhua Q., Jin Juan X. Progress in the technology for desulfurization of crude Oil. China petroleum processing Petrochemical technologhy 2010;12(4):1-6.

[10]-WWW.fa.wikipedia.org/wiki/تیوفن

[11]-Miller K.E,Bruno TJ,(2003)," Iso thermal Kovats retention indieces of sulfur compounds on poly(5% diphenyl-95% dimethyl siloxane) Stationary phase", J Chromatogr. A , 1007,PP 117-125

[12]-DUH, Ring Z, Briker Y, Arboled P,(2004),"Prediction of gas chromatographic retention times and indices of sulfur compound inlight cycle Oil" Catal. Todey, 98, PP 214-225

[13]-Can H, Diamoglo A, Kovalishyn V, (2005), "Application of artificial neural networks for the prediction of sulfur polycyclic aromatic compounds retention indices "J.Mol.Struct.723. PP 183-188

[14]-Garkani-Nejad, Z., et al. "Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds." *Journal of Chromatography A* 1028.2 (2004): 287-295.

[15]-Liao L., Mei H., Li J., Li Z., (2008), "Estimation and prediction on retention times of components from essential oil of paulownia tomentosa flowers by molecular electronegativity-distanse vector (MEDV)", J Mol Struct, 850, pp 1-8

[16]-Hu, Rong-Jing, et al. "QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors." *Talanta* 68.1 (2005): 31-39.

[17]-Riahi S., Ganjali M.R., Pourbasheer S., Norouzi P., (2008), "QSPR Study of GC Retention Indices of Oganic Compounds by Multiple Linear Regression with a Genetic Algorithm", Chromatographia, 67, pp 917-922

[18]-Riahi S., Gaanjali M., Pourbasheer S., Norouzi P.,(2009), "Investigation of different linear and nonlinear chemometric methods for modeling of retention index of organic components: Concerns to support vector machine" Journal of Hazardous Materials, 166, pp 853-859

[19]-Xu, Hui-Yuing et all, "Quantitative structure-chromatographic sulfur heterocyclies" Journal chromatography A, 1198(2008):202-207.

[20]-Xiang, Zheng, Yizeng Liang, and Qiannan Hu. "Relativity study of the topological index of methylalkane structures and chromatographic retention index." *Se pu= Chinese journal of chromatography* 23.2 (2005): 117-122.

[21]-Du, Hongbin, et al. "Prediction of gas chromatographic retention times and indices of sulfur compounds in light cycle oil." *Catalysis today* 98.1-2 (2004): 217-225.

[22]-Zanousi mohammad bagher, Mehdi nekoi and majid mohammad Hosseini,(2016),"composition of the essential oils and volatile fractions of Artemisia absinthium by three different extraction methods: Hydrodistillation, solvent-free microwave extraction and headspace solid-phase micro extraction combined with a novel QSRR Evaluation", Journal of essential oil bearing plants, 19:7(2016),1561-1581.

[23]-Goudarzi, Nasser, et al."Quantitative structure-property relationships of retention indices of some sulfure organic compounds using random forest technique as a variable selection and modeling method". Journal of sparation science 39.19(2016): 3835-3842

[24]-Otto M., (2007), "Chemometrics", WILEY-VCH Verlag GmbH & Co. KgaA, Weinheim, pp 1-11

[25]-Howery D.G., Hirsch R.F., (1983), "Chemometrics in the chemistry curriculum", J. Chem. Ed. 60, pp 656-659

[26]-Wold S., "Chemometrics: what do we mean with it, and what do we want from it?" And Brown, S. D., "Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics", papers on chemometrics:Philosoiphy, History and Direction of In CINC'94.

[27]-Moosavi-Movahedi A.A., Gharanfoli M., Nazeri K., Shamsipur M., Chamani J., Hemmateenejad B., Alavi M., Shokrollahi A., Habib-Rezaei M., Sorrenson C., Sheibani N., (2005), Colloids Suurf., 43, pp 150-157

[28]-Hemmateenejad B., (2005) "Chemometrics in Iran" Chemometrics and intelligent Laboratory System, 81, pp 202-208

[29]-Anderson, G.G. Kaufmann, P.(2000),"Development of a generalized neural network",chemometr. Intel. Lab. Syst.50 PP.101-105

[30]-Arab cham Jangali M, (2009), Modelling of Cytotoxicity data of anti- Hiv 1-[(5-chloropheny) sulfpnyl]-1 H-pyrole derrivatives using calculated molecular descriptors and Levenbery-Marquadt Artificial neural network", chem. Bio. Des, 73, PP 456-465.

[۳۱]-فرشاد فرع-(۱۳۸۰)،" اصول و روش های پیشرفته آماری در تجزیه رگرسیون" چاپ دوم، انتشارات طاق بستان.

[32]-Mohammad Asradel, “Prediction of combustion Dynamics in An Experimental Turbulent Swir stabilized combustor with secondary Fuel Injection”,university of Tehran,2015.

[۳۳]-کیا.م،(۱۳۸۷)،"شبکه های عصبی در MATLAB چاپ دوم،انتشارات کیان رایانه سبز

[۳۴]-فردوسی.م،ع، (۱۳۸۹)، پایان نامه کارشناسی ارشد،" پیش بینی ثابت های هنری بعضی از ترکیبات آلی با استفاده از روش های خطی و غیر خطی QSPR "، دانشگاه صنعتی شاهرود

[35]-Braiman L, Random Forests, Machine Learn 2001; 45:P 15-32

[36]-Hastie T TRFJ, The elements of statistical learning data mining inference and prediction,In Springer series in statistics New York, Springer 2001; Xvi:P.533

[37]-Braiman L, Classification and regression trees CA,WadsWorth Intenational Groups 1984.

[۳۸]-جندقی. ف (۱۳۹۲)، پایان نامه کارشناسی ارشد، "کاربرد روش جنگل‌های تصادفی به عنوان ابزاری برای انتخاب متغیر و پیش بینی اندیس بازداری تعدادی از ترکیبات آلی به عنوان آلاینده‌ی محیط زیست"، دانشگاه صنعتی شاهرود

[39]-Hastie T, Tibshirani R, Friedman J, (2009),” The Elements of statistical learning Data Mining Inferences, and Prediction”, 2^{nd} Ed, Springer, New York

[40]-Braiman L, Random Forests, Machine Learn 2001;45:P.5-32.

[41]-Genuer R, Poggi L M , Tueleau-Malotch,2010," Variable selection using random forests", Pattern Recoyn Lett, 31, PP.2225-2236.

[42]-Friedman JH (1991) Multivariate adaptive regression splines, J. Annals of statistics, 19:1-141

[۴۳]-تکتم ولی‌زاده، (۱۳۹۱)، پایان‌نامه کارشناسی ارشد، "اسپلاین‌های تطبیقی و غیر تطبیقی در مدل‌های رگرسیونی نیمه پارامتری"، دانشگاه صنعتی شاهرود

[۴۴]-زینب افتخاری، (۱۳۹۰)، پایان‌نامه کارشناسی ارشد، "مطالعه QSPR جهت پیشگویی حلالیّت حشره کش‌ها در آب با استفاده از روش MARS-ANFIS و اندازه‌گیری همزمان سرب و کادمیم به روش ولتامتری عاری سازی پالس تفاضلی با استفاده از روش حداقل مربعات جزئی(PLS)"، دانشگاه دامغان.

[45]-Mercader A. G, Pomilio A.B,(2010),"QSAR Study Of Flavonoids and bioflavonoids as in fluenza HINI virus neuramaini dase inhibitors", Eur.J. Med.Chem,45, PP 1724-1730

[46]-Todeschini R, Consonni V, (2000),"Hundbook of molecular descriptors",Wiley-VCH,Wein heim,Germany

[47]-http//WWW.downloadsoftware.ir

[48]-Walter B. Wilson., Lane C Sander.," Retention behavior of alkyl-sustituded Polycyclic aromatic sulfur heterocycles inreversed-phase liquid chromatography" J. Chromatography A, 1461(2016),120-130

[49]-Walter B. Wilson., Lane C Sander.," Retention behavior of isomeric polycyclic aromatic sulfur heterocyles in reversed-phase liquid chromatography" J. Chromatography A, 1461(2016)107-119

[50]-Crisan L., Iliescu S., (2016), "Structure-flammability relationship study of phosphoester dimers by MLR and PLS", Polimeros, (AHEAD), 0-0.

[51]-Cosonni V., Todeschini R., Pavan M, (2000), "Structure/Response Correlations and similarity/Diversity Analysis by GETAWAY Descriptors. Theory of the novel 3D Molecular Descriptors", J. Chem. Info. Comput. Sci, 42, PP 682-692

[52]-Karelson M., Lobanov V.S., Katritzky A.R. (1996), "Quantum-chemical descriptors in QSAR/QSPR Studies", J. Chem Rev., 96, PP 1027-1043. Charge

[۵۳]-اشرفی. م. (۱۳۸۹)، پایان‌نامه کارشناسی ارشد "مطالعه ارتباط کمی ساختار-فعالیت مشتقات کربامات به عنوان دسته‌ای جدید از بازدارنده‌های غیر نوکلئوزیدی HIV"، دانشگاه صنعتی شاهرود

[۵۴]-عجم. ز. (۱۳۹۳)، پایان‌نامه کارشناسی ارشد "پیش‌بینی فعالیت ضد ایدز مشتقات غیر نوکلئوزیدی تیوکربامات به روش RF "، دانشگاه صنعتی شاهرود

[55]-Goudarzi, N., et al. “Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons.” *Journal of Chromatography A* 1333 (2014): 25-31.

[56]- Kao, Szu-Pyng, Yao-Chung Chen, and Fang-Shii Ning. "A MARS-based method for estimating regional 2-D ionospheric VTEC and receiver differential code bias." *Advances in Space Research* 53.2 (2014): 190-200.

Abstract

Generally, the amount of sulfur in crude oil, depending on its source, varies from 0.25 to about 7.89 of Weight percent. Generally sulfur is observed both organic and inorganic forms in crude oil. In addition to elemental sulfur more than 200 sulfur organic compounds are identified in the oil, which contains sulfides, Thiols or mercaptans, thiophene, which theophany, and di-benzoate molecules are much more complex. These compounds are known as persistent contaminants. In addition, these compounds are carcinogenic. Various methods, such as gas chromatography, have been proposed to segregate and identify these compounds, which inhibition index compares with standards. Therefore, the inhibition index of a soluble is an important parameter that can be used to identify sulfur heterocyclic aromatic compounds. But on the other hand, these methods are costly and time-consuming. By applying chemo metrics methods, prediction of inhibitory index of compounds is plausible (PASHs) and by using random forest methods (RF) and MARS methods and the artificial neural network SR-ANN and RF-ANN we predicted the inhibitory effect of these compounds. The results indicate that the use of the mentioned statistical method has a good potential to predict the empirical inhibitory index of these compounds, which R^2 results from RF and MARS and SR-ANN and RF-ANN methods are 0/944and 0/984, 0/971and 0/932 respectively.

Keywords: Chemometrics, Random forest, MARS, Artificial Neural Network, Inhibition Indices (index), Aromatic heterocyclic sulfur compounds

University of Shahrood

Faculty of chemistry

M.Sc Thesis in analytical chemistry

Application of Chemometircs methods to Predict the Retention index of some organic Compounds

Elahe Salamati

Supervisors:

Dr. Naser goodarzi

Dr. Davood Shahsavani

Advisor:

Dr. Mansour ArabCham jangali

December 2018