
بسم الله الرحمن الرحیم

دانشکده شیمی

گروه شیمی تجزیه

کاربرد روش جنگل‌های تصادفی به عنوان ابزاری برای انتخاب متغیر و پیش‌بینی اندیس بازداری تعدادی از ترکیبات آلی به عنوان آلاینده‌ی محیط زیست

فرشته سادات عمادی جندقی

اساتید راهنما

دکتر ناصر گودرزی

دکتر داوود شاهسونی

استاد مشاور

دکتر منصور عرب چم‌جنگلی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

بهمن ماه ۱۳۹۲

تقدیم به فرشتگان مقدسم

پدر و مادر عزیزم که نمی‌توانم سیاهی را به موی سپیدشان بازگردانم اما دستان گرمشان را هزاران بار بوسه‌باران می‌کنم.

و خواهران خوبم که بی‌آلایش‌ترین دوستان زندگی‌ام هستند.

تشکر وقدردانی

سپاس خدای مهربانم را که در تمام مراحل زندگی‌ام، حضورش بسان نسیم ملایمی در روح خسته‌ام دمیده است.

و سپاس پدر و مادر عزیزم را به دلیل زحمات بی‌چشمداشتشان و اساتید گرانقدرم جناب آقای دکتر گودرزی، جناب آقای دکتر شاهسونی و جناب آقای دکتر عرب که بدون حضور و مساعدتشان، انجام کار ممکن نبود.

تعهد نامه

اینجانب **فرشته سادات عمادی جندقی** دانشجوی دوره کارشناسی ارشد رشته **شیمی‌تجزیه** دانشکده **شیمی** دانشگاه صنعتی شاهرود نویسنده پایان نامه **کاربرد روش جنگل‌های تصادفی به عنوان ابزاری برای انتخاب متغیر و پیش-بینی اندیس بازداری تعدادی از ترکیبات آلی به عنوان آلاینده‌ی محیط زیست** تحت راهنمائی **آقای دکتر ناصر گودرزی و آقای دکتر داوود شاهسونی** متعهد می شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه 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 می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

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

چکیده

ترکیبات گوگرددار و هیدروکربن‌های آروماتیک چندحلقه‌ای به عنوان دو گروه آلوده کننده‌ی پایدار شناخته می‌شوند. علاوه بر این، این ترکیبات سرطان‌زا هستند. تکنیک‌های تجزیه‌ای مختلفی مانند کروماتوگرافی گازی برای جداسازی و شناسایی این ترکیبات وجود دارد که اندیس‌های بازداری را با استانداردها مقایسه می‌کنند. بنابراین اندیس بازداری یک حل‌شونده، یک پارامتر مهم می‌باشد که می‌تواند برای شناسایی ترکیبات گوگرددار و هیدروکربن‌های آروماتیک چندحلقه‌ای به کار رود. اما از طرف دیگر، این روش‌ها زمان و هزینه‌ی زیادی صرف می‌کنند. از این رو، برای غلبه بر این مشکلات، می‌توان از روش‌های کمومتریکس برای پیش‌بینی اندیس بازداری این ترکیبات استفاده نمود.

در بخش اول این پایان‌نامه، روش جدید جنگل‌های تصادفی (RF) برای پیش‌بینی اندیس بازداری یک گروه از ترکیبات گوگرددار به کار گرفته شد. هم‌چنین عملکرد مدل RF با مدل‌های شبکه عصبی مصنوعی (ANN) و رگرسیون خطی چندگانه (MLR) مقایسه گردید. نتایج این مقایسه نشان داد که جنگل‌های تصادفی دارای برتری نسبی نسبت به شبکه عصبی مصنوعی و رگرسیون خطی چندگانه برای پیش‌بینی این ترکیبات می‌باشد.

در بخش دوم، مانند بخش اول، این مدل‌ها برای پیش‌بینی اندیس بازداری بعضی از ترکیبات آروماتیک چندحلقه‌ای به کار گرفته شدند. نتایج به دست آمده نشان می‌دهد که روش جنگل‌های تصادفی، اندیس بازداری این ترکیبات را بهتر از سایر روش‌ها پیش‌بینی نموده است.

کلمات کلیدی: ترکیبات گوگرددار، هیدروکربن‌های آروماتیک چندحلقه‌ای، اندیس بازداری، RF، ANN، MLR

مقالات مستخرج از پایان‌نامه

Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons

در Journal of chromatography A

QSPR study of retention time of some volatile organic compounds by using random forests (RF), multiple linear regression (MLR) and artificial neural network (ANN) methods

در شانزدهمین کنگره شیمی ایران در دانشگاه یزد _ شهریور۹۲

Application of Random Forests (RF) method to predict the retention index of some sulfur compounds

در شانزدهمین کنگره شیمی ایران در دانشگاه یزد _ شهریور۹۲

فهرست مطالب

فصل اول: مقدمه

۱-۱- اندیس بازداری ۲
۱-۲- معرفی ترکیبات گوگرددار ۲
۱-۲-۱- شناسایی و جداسازی ترکیبات گوگرددار ۳
۱-۲-۲- مروری بر کارهای انجام شده ۴
۱-۳- معرفی هیدروکربن‌های آروماتیک چندحلقه‌ای (PAHs) ۵
۱-۳-۱- شناسایی و جداسازی PAHs ۶
۱-۳-۲- مروری بر کارهای انجام شده ۶
۱-۴- ضرورت انجام تحقیق ۷
۱-۵- ساختار پایان‌نامه ۸

فصل دوم: کمومتریکس

۲-۱- فراهم کردن سری داده‌ها ۱۱
۲-۲- بهینه‌سازی ساختار هندسی ترکیبات ۱۱
۲-۳- محاسبه توصیف‌گرها ۱۲
۲-۴- انتخاب توصیف‌گرهای مهم ۱۳
۲-۴-۱- رگرسیون مرحله‌ای ۱۴
۲-۴-۲- راهبرد CDFS ۱۴
۲-۵- ساخت مدل ۱۵
۲-۶- رگرسیون خطی چندگانه ۱۵
۲-۷- شبکه عصبی مصنوعی ۱۶
۲-۷-۱- مراحل آموزش در شبکه‌های پیشخور ۱۷
۲-۷-۲- آموزش شبکه با تکنیک پس‌انتشار ۱۷
۲-۸- روش جنگل‌های تصادفی ۱۸
۲-۸-۱- روش درخت‌های رده‌بندی و رگرسیونی ۱۸
۲-۸-۱-۱- الگوریتم تشکیل درخت رگرسیونی ۲۰
۲-۸-۱-۲- اندازه‌ی درخت و هرس کردن ۲۱
۲-۸-۲- جنگل‌های تصادفی ۲۲
۲-۸-۲-۱- الگوریتم روش جنگل‌های تصادفی ۲۲
۲-۸-۲-۲- تعیین اهمیت توصیف‌گرها در روش جنگل‌های تصادفی ۲۴
۲-۸-۲-۳- مزایای روش جنگل‌های تصادفی ۲۵
۲-۹- ارزیابی مدل ۲۵
۲-۱۰- نرم‌افزارهای مورد استفاده ۲۸
۲-۱۰-۱- بسته نرم‌افزاری Hyperchem ۲۸
۲-۱۰-۲- بسته نرم‌افزاری Dragon ۲۸
۲-۱۰-۳- بسته نرم‌افزاری SPSS ۲۸

۲-۱۰-۴- نرم‌افزار MATLAB ۲۹
۲-۱۰-۵- بسته نرم‌افزاری R ۲۹

فصل سوم: مدل‌سازی اندیس بازداری تعدادی از ترکیبات آلی گوگرددار

۳-۱- سری داده‌ها ۳۲
۳-۲-بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌گرها ۳۸
۳-۳- انتخاب بهترین توصیف‌گرها ۳۹
۳-۳-۱- انتخاب بهترین توصیف‌گرها با روش رگرسیون مرحله‌ای ۳۹
۳-۳-۲- انتخاب بهترین توصیف‌گرها با راهبرد CDFS ۴۰
۳-۳-۳- انتخاب بهترین توصیف‌گرها با روش جنگل‌های تصادفی ۴۲
۳-۴- مدل‌سازی با رگرسیون خطی چندگانه ۴۴
۳-۵- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای(SR-ANN) ۴۵
۳-۵-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه ۴۶
۳-۵-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۴۷
۳-۶- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط راهبرد CDFS (CDFS-ANN) ۴۸
۳-۶-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه ۴۸
۳-۶-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۴۹
۳-۷- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی (RF-ANN) ۴۹
۳-۷-۱- بهینه‌سازی پارامترهای مؤثر ۵۰
۳-۷-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۵۱
۳-۸- مدل‌سازی جنگل‌های تصادفی ۵۱
۳-۸-۱- بهینه‌سازی مقادیر ntree و m ۵۲
۳-۸-۲- مدل‌سازی جنگل‌های تصادفی با استفاده از توصیف‌گرهای مهم ۵۳
۳-۹- ارزیابی مدل‌ها ۵۴
۳-۹-۱- ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون ۵۴
۳-۹-۲- ارزیابی مدل‌ها توسط اعتبارسنجی متقاطع با رد مرحله‌ای تک‌تک ۵۴
۳-۹-۳- ارزیابی مدل‌ها با استفاده از پارامترهای آماری ۶۳
۳-۹-۴- ارزیابی مدل‌ها با استفاده از آزمون Y-تصادفی ۶۴
۳-۱۰- بررسی ارتباط توصیف‌گرهای منتخب مدل برتر با اندیس بازداری ۶۴
۳-۱۱- نتیجه‌گیری ۶۷

فصل چهارم: مدل‌سازی جهت پیش‌بینی اندیس بازداری بعضی ترکیبات آروماتیک چندحلقه‌ای

۴-۱- سری داده‌ها ۷۰
۴-۲- بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌گرها ۷۰
۴-۳- انتخاب بهترین توصیف‌گرها ۷۰
۴-۳-۱- انتخاب بهترین توصیف‌گرها با روش رگرسیون مرحله‌ای ۷۹
۴-۳-۲- انتخاب بهترین توصیف‌گرها با راهبرد CDFS ۸۰

۴-۳-۳- انتخاب بهترین توصیف‌گرها با روش جنگل‌های تصادفی ۸۱
۴-۴- مدل‌سازی با رگرسیون خطی چندگانه ۸۳
۴-۵- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای(SR-ANN) ۸۴
۴-۵-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه ۸۴
۴-۵-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۸۵
۴-۶- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط راهبرد CDFS (CDFS-ANN) ۸۶
۴-۶-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه ۸۶
۴-۶-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۸۷
۴-۷- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی (RF-ANN) ۸۷
۴-۷-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه ۸۸
۴-۷-۲- ساختار شبکه عصبی مصنوعی بهینه شده ۸۹
۴-۸- مدل‌سازی جنگل‌های تصادفی ۸۹
۴-۸-۱- بهینه‌سازی مقادیر ntree و m ۹۰
۴-۸-۲- مدل‌سازی جنگل‌های تصادفی با استفاده از توصیف‌گرهای مهم ۹۱
۴-۹- ارزیابی مدل‌ها ۹۱
۴-۹-۱- ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون ۹۱
۴-۹-۲- ارزیابی مدل‌ها توسط اعتبارسنجی متقاطع با رد مرحله‌ای تک‌تک ۹۴
۴-۹-۳- ارزیابی مدل‌ها با استفاده از پارامترهای آماری ۹۴
۴-۹-۴- ارزیابی مدل‌ها با استفاده از آزمون Y-تصادفی ۹۴
۴-۱۰- بررسی ارتباط توصیف‌گرهای منتخب مدل برتر با اندیس بازداری ۱۰۲
۴-۱۱- نتیجه‌گیری ۱۰۴
آینده‌نگری ۱۰۵
فهرست منابع ۱۰۶

فهرست جداول

جدول (۳-۱)- نام و اندیس بازداری سری داده‌ها ۳۳
جدول (۳-۲)- توصیف‌گرهای انتخاب شده با روش رگرسیون مرحله‌ای ۴۰
جدول (۳-۳)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای ۴۰
جدول (۳-۴)- توصیف‌گرهای ۱۶ مدل به دست آمده از تقسیم‌بندی مختلف داده‌ها ۴۱
جدول (۳-۵)- توصیف‌گرهای انتخاب شده با راهبرد CDFS ۴۱
جدول (۳-۶)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط راهبرد CDFS ۴۱
جدول (۳-۷)- توصیف‌گرهای انتخاب شده با روش جنگل‌های تصادفی ۴۳
جدول (۳-۸)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی ۴۳
جدول (۳-۹)- پارامترهای آماری برای مدل‌های به دست آمده از روش MLR ۴۴
جدول (۳-۱۰)- ضرایب غیراستاندارد و استاندارد توصیف‌گرهای موجود در مدل خطی برتر ۴۵
جدول (۳-۱۱)- توابع و پارامترهای شبکه‌های بهینه (SR-ANN) به دست آمده ۴۷
جدول (۳-۱۲)- مشخصات شبکه بهینه (SR-ANN) ۴۷
جدول (۳-۱۳)- توابع و پارامترهای شبکه‌های بهینه (CDFS-ANN) به دست آمده ۴۸
جدول (۳-۱۴)- مشخصات شبکه بهینه (CDFS-ANN) ۴۹
جدول (۳-۱۵)- توابع و پارامترهای شبکه‌های بهینه (RF-ANN) به دست آمده ۵۰
جدول (۳-۱۶)- مشخصات شبکه بهینه (RF-ANN) ۵۱
جدول (۳-۱۷)- کمترین مقادیر RMSE همراه با m و ntree متناظر آنها ۵۲
جدول (۳-۱۸)- تغییرات خطای OOB با تعداد توصیف‌گر متفاوت ۵۳
جدول (۳-۱۹)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون ۵۵
جدول (۳-۲۰)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از روش رد مرحله‌ای تک‌تک ۵۷
جدول (۳-۲۱)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای داده‌های سری آزمون ۶۳
جدول (۳-۲۲)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای کل داده‌ها به روش رد تک‌تک ۶۳
جدول (۳-۲۳)- مقادیر R^2 برای داده‌های سری آزمون با استفاده از آزمون Y-تصادفی ۶۴
جدول (۳-۲۴)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر SCBO ۶۵
جدول (۳-۲۵)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر BELe6 ۶۶
جدول (۴-۱)- نام و اندیس بازداری سری داده‌ها ۷۱
جدول (۴-۲)- توصیف‌گرهای انتخاب شده با روش رگرسیون مرحله‌ای ۷۹
جدول (۴-۳)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای ۷۹
جدول (۴-۴)- توصیف‌گرهای ۱۶ مدل به دست آمده از تقسیم‌بندی مختلف داده‌ها ۸۰
جدول (۴-۵)- توصیف‌گرهای انتخاب شده با راهبرد CDFS ۸۰
جدول (۴-۶)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط راهبرد CDFS ۸۱
جدول (۴-۷)- توصیف‌گرهای انتخاب شده با روش جنگل‌های تصادفی ۸۲
جدول (۴-۸)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی ۸۲
جدول (۴-۹)- پارامترهای آماری برای مدل‌های به دست آمده از روش MLR ۸۳
جدول (۴-۱۰)- ضرایب غیراستاندارد و استاندارد توصیف‌گرهای موجود در مدل خطی برتر ۸۳
جدول (۴-۱۱)- توابع و پارامترهای شبکه‌های بهینه (SR-ANN) به دست آمده ۸۵

جدول (۴-۱۲)- مشخصات شبکه بهینه (SR-ANN) ۸۵
جدول (۴-۱۳)- توابع و پارامترهای شبکه‌های بهینه (CDFS-ANN) به دست آمده ۸۶
جدول (۴-۱۴)- مشخصات شبکه بهینه (CDFS-ANN) ۸۷
جدول (۴-۱۵)- توابع و پارامترهای شبکه‌های بهینه (RF-ANN) به دست آمده ۸۸
جدول (۴-۱۶)- مشخصات شبکه بهینه (RF-ANN) ۸۹
جدول (۴-۱۷)- کمترین مقادیر RMSE همراه با m و ntree متناظر آنها ۹۰
جدول (۴-۱۸)- تغییرات خطای OOB با تعداد توصیف‌گر متفاوت ۹۱
جدول (۴-۱۹)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون ۹۲
جدول (۴-۲۰)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از روش رد مرحله‌ای تک‌تک ۹۵
جدول (۴-۲۱)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای داده‌های سری آزمون ۱۰۱
جدول (۴-۲۲)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای کل داده‌ها به روش رد تک‌تک ۱۰۱
جدول (۴-۲۳)- مقادیر R^2 برای داده‌های سری آزمون با استفاده از آزمون Y-تصادفی ۱۰۱
جدول (۴-۲۴)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر G2 ۱۰۳
جدول (۴-۲۵)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر PiID ۱۰۴

فهرست اشکال

شکل (۱-۱)- شمای کلی پایان‌نامه ۸
شکل (۲-۱)- ساختمان یک نرون محاسباتی ۱۶
شکل (۲-۲)- نمونه‌ای از افراز صحیح (راست) و غیر صحیح (چپ) در فضای دومتغیره ۱۹
شکل (۲-۳)- روند سلسله مراتبی افراز شکل (۲-۲)-راست ۱۹
شکل (۲-۴)- نمودار درختی فضای افراز شده در شکل (۲-۳) ۱۹
شکل (۲-۵)- نمودار پراکندگی داده‌های دومتغیره ۲۰
شکل (۲-۶)- برخی از افرازهای ممکن در راستای توصیف‌گر x_1 ..... ۲۱
شکل (۲-۷)- برخی از افرازهای ممکن در راستای توصیف‌گر x_2 ..... ۲۱
شکل (۳-۱)- نمودار اهمیت نسبی توصیف‌گرها ۴۲
شکل (۳-۲)- نمودار تغییر خطای شبکه (SR-ANN) با تغییر پارامتر μ ۴۷
شکل (۳-۳)- نمودار تغییر خطای شبکه (CDFS-ANN) با تغییر پارامتر μ ۴۹
شکل (۳-۴)- نمودار تغییر خطای شبکه (RF-ANN) با تغییر پارامتر μ ۵۱
شکل (۳-۵)- نمودار تغییرات مجذور خطای OOB یا RMSE در مقادیر متفاوت m و ntree ۵۲
شکل (۳-۶)- نمودار مقادیر پیش‌بینی شده اندیس بازداری داده‌های سری آزمون توسط مدل‌های مختلف در مقابل مقادیر تجربی ۵۶
شکل (۳-۷)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای داده‌های سری آزمون با مدل‌های مختلف ۵۶
شکل (۳-۸)- نمودار مقادیر پیش‌بینی شده اندیس بازداری کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل‌های مختلف بر حسب مقادیر تجربی ۶۲
شکل (۳-۹)- نمودار مقادیر خطای مطلق بر حسب مقادیر تجربی برای کل داده‌ها با مدل‌های مختلف به روش رد مرحله‌ای تک‌تک ۶۲
شکل (۴-۱)- نمودار اهمیت نسبی توصیف‌گرها ۸۱
شکل (۴-۲)- نمودار تغییر خطای شبکه (SR-ANN) با تغییر پارامتر μ ۸۵
شکل (۴-۳)- نمودار تغییر خطای شبکه (CDFS-ANN) با تغییر پارامتر μ ۸۷
شکل (۴-۴)- نمودار تغییر خطای شبکه (RF-ANN) با تغییر پارامتر μ ۸۹
شکل (۴-۵)- نمودار تغییرات مجذور خطای OOB در مقادیر متفاوت m و ntree ۹۰
شکل (۴-۶)- نمودار مقادیر پیش‌بینی شده اندیس بازداری داده‌های سری آزمون توسط مدل‌های مختلف در مقابل مقادیر تجربی ۹۳
شکل (۴-۷)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای داده‌های سری آزمون با مدل‌های مختلف ۹۳
شکل (۴-۸)- نمودار مقادیر پیش‌بینی شده اندیس بازداری کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل‌های مختلف در مقابل مقادیر تجربی ۱۰۰
شکل (۴-۹)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای کل داده‌ها با مدل‌های مختلف به روش رد مرحله‌ای تک‌تک ۱۰۰

فصل اول

مقدمه

۱-۱- اندیس بازداری[1]

اندیس بازداری (RI) در سال ۱۹۵۸ برای اولین بار به عنوان 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 مورد نظر بین زمان بازداری آن دو باشد، امکان‌پذیر است. در این مورد پارافین‌های نرمال به عنوان استاندارد در نظر گرفته می‌شوند. طبق قرارداد مقدار این اندیس برای این پارافین‌ها صد برابر تعداد کربن‌های آنها در نظر گرفته می‌شود که مقداری ثابت بوده و مستقل از نوع فاز ساکن، درجه حرارت و سایر شرایط می‌باشد. در حالیکه 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 $\acute{t}_{Ra}$، $\acute{t}_{Rn}$ و $\acute{t}_{Rn+1}$ به ترتیب عبارتند از زمان‌های بازداری تصحیح شده ماده‌ی مورد تجزیه، پارافین نرمال حاوی n و n+1 کربن که بلافاصله قبل و بعد از ماده مورد تجزیه از ستون خارج می‌شوند[۱].

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

۱-۲- معرفی ترکیبات گوگرددار

ترکیبات گوگرددار، ترکیباتی هستند که در بسیاری از صنایع مثل غذا یا پتروشیمی نقش مهمی را ایفا می‌کنند.

مهم‌ترین منابع این ترکیبات، سوخت‌های فسیلی، ذغال سنگ و گاز طبیعی هستند. صنایع تولید کاغذ و نیشکر، مرکاپتان، سولفید و دی‌سولفید تولید می‌کنند و هم‌چنین پساب کارخانه‌های تبدیل ذغال سنگ به کک و تصفیه روغن حاوی تیول‌ها و سولفیدها است[۲].

1-Retention index

این ترکیبات یک منبع مهم آلودگی هوا هستند و هم‌چنین باعث تخریب محیط زیست می‌گردند؛ زیرا باعث از بین رفتن جنگل‌ها، اسیدی شدن آب دریا و دریاچه و خوردگی فلزات می‌شوند. در صنایع پتروشیمی، وجود مقدار اندک این ترکیبات، باعث سمی شدن محصولات یا آلودگی هوا هنگام سوختن این محصولات نفتی می‌گردد. حضور آنها حتی به مقدار کم در مواد غذایی و نوشیدنی، طعم و بوی بدی ایجاد می‌کند. در صورت وجود تیول‌های فرار در مواد غذایی فاسد، مصرف آنها می‌تواند منجر به مرگ انسان شود. هم‌چنین، وجود این ترکیبات در مکان‌هایی که اکسیژن کمی در فضا وجود دارد؛ خطر خفگی را افزایش می‌دهد[۲].

۱-۲-۱- شناسایی و جداسازی ترکیبات گوگرددار

آنالیز این ترکیبات در ماتریکس‌های مختلف، کار دشواری است، زیرا اکثرآنها با غلظت پایینی در بافت‌های پیچیده‌ای وجود دارند که سایر گونه‌ها با غلظتی چندین برابر، باعث ایجاد مزاحمت در اندازه-گیری این ترکیبات می‌شود. 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، اسپکتروفوتومتری، 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[۲]. اما استفاده از این تکنیک نیز 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 بافت‌های پیچیده، به یک دتکتور حساس و انتخاب پذیر نیاز است[۲].

به دلیل وجود این مشکلات، یک روش تئوری، جایگزین خوبی برای روش‌های تجربی است. مطالعه

ارتباط کمی ساختار-خاصیت (QSPR)[1]، یک روش توانمند برای پیش‌بینی اندیس بازداری است. در این روش از توصیف‌گرهای ساختاری که برهمکنش بین حل‌شونده و فاز ساکن ستون کروماتوگرافی را توصیف می‌کنند، برای ساخت یک مدل استفاده می‌شود.

۱-۲-۲- مروری بر کارهای انجام شده

در سال ۲۰۰۳، میلر[2] و برانو[3]، اندیس بازداری کواتس[4] برای ترکیبات گوگرددار را محاسبه نموده و آنها را با مقادیر به دست آمده از روش تجربی، مقایسه کردند. نتایج نشان داد که این روش تئوری، در پیش‌بینی اندیس بازداری این ترکیبات موفق عمل کرده است[۳]. در سال ۲۰۰۴، دیو[5] و همکارانش، مدل رگرسیون خطی چندگانه (MLR)[6] را برای پیش‌بینی زمان و اندیس بازداری این ترکیبات گوگرددار به کار بردند. نتایج به دست آمده توافق خوبی با مقادیر تجربی زمان و اندیس بازداری نشان می‌دهد[۴]. کن[7] و همکارانش در سال ۲۰۰۵، از پنج مدل شبکه‌ی عصبی مصنوعی (ANN)[8] برای پیش‌بینی زمان بازداری ترکیبات گوگردی چند حلقه‌ای آروماتیک استفاده کردند. برای ساخت مدل‌ها، تنها از توصیف‌گر های ساختاری ترکیبات استفاده شده بود. ارزیابی مدل‌ها نیز با استفاده از ارزیابی تقاطعی[9] انجام شده و بهترین مدل نیز که دارای دو توصیف‌گر است، گزارش شده است[۵].

1-Quantitative structure-property relationship
2-Miller
3-Bruno
4-kovats
5-Du
6-Multiple linear regression
7-Can
8-Artificial neural network
9-Cross validation

۱-۳- معرفی هیدروکربن‌های آروماتیک چندحلقه‌ای[1]

هیدروکربن‌های آروماتیک چندحلقه‌ای (PAHs)، گروهی از ترکیبات هستند که از دو یا چند حلقه‌ی آروماتیک تشکیل شده‌اند. این ترکیبات، خواص فیزیکی و شیمیایی مشابهی دارند و از مهم‌ترین آلوده کننده‌های محیطی هستند. هم‌چنین این ترکیبات سمی و سرطان‌زا بوده و با روش‌های مختلف از جمله تغذیه و تنفس وارد بدن انسان و جانوران شده و بر روی سلامت آنها اثر می‌گذارند[۶].

منابع تولید این هیدروکربن‌ها، صنایعی چون تولید آلومینیوم، ذغال، سیمان، بتن، آسفالت و پتروشیمی می‌باشد. احتراق انواع سوخت‌ها مانند گاز طبیعی، چوب، ذغال و ذغال سنگ نیز باعث تولید ترکیبات متنوعی از PAHs می‌شود. هم‌چنین این ترکیبات در اثر عواملی مثل دود سیگار، دود ناشی از کارخانه‌ها و موتورهای دیزلی و سوختن انواع سوخت‌های فسیلی در اتمسفر آزاد می‌گردند.

بسیاری از این ترکیبات اثرات مخربی روی ساختار و اعمال ارگانیسم‌های زنده دارند. آنها می‌توانند ذرات بسیار ریزی تشکیل دهند که به دلیل کوچک بودن می‌توانند زمان زیادی در فضا معلق باشند و مسافت‌های طولانی را طی کنند[۷]. هم‌چنین می‌توانند از طریق تنفس وارد بدن شده و صدمات زیادی به شش‌ها وارد نمایند. آب رودخانه‌ها و دریاچه‌های موجود در نزدیکی مکان‌هایی که زباله و زوائد کارخانه‌ها وجود دارد 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. استحمام با آب آلوده، مصرف شامپو و کرم‌هایی که حاوی برخی از این ترکیبات هستند و تماس با خاک آلوده، می‌تواند این هیدروکربن‌ها را جذب بدن نموده و باعث ایجاد تاول و سرطانی شدن سلول‌های پوستی گردد. [۸].

1-Polycyclic aromatic hydrocarbons

۱-۳-۱- شناسایی و جداسازی PAHs

از جمله تکنیک‌های مناسب برای آنالیز کمی و کیفی هیدروکربن‌های آروماتیک چندحلقه‌ای کروماتوگرافی مایع با عملکرد بالا[1] و کروماتوگرافی گازی با آشکارساز طیف‌سنج جرمی[2] که روش متداول‌تری است، می‌باشد. اما این روش نیز مشکلاتی را به همراه دارد که عبارتند از:

۱- هیدروکربن‌ها با وزن مولکولی زیاد تمایل زیادی به اتصال به فاز ساکن دارند و شسته نمی‌شوند.

۲- کنترل دما برای این سیستم مهم است، زیرا تغییرات دما باعث نوسان در زمان بازداری می‌شود.

۳- تهیه‌ی نمونه جهت آنالیز نیاز به مراحل زیادی دارد که مستلزم صرف وقت و هزینه‌ی زیاد است[۹].

به دلیل وجود این مشکلات می‌توان از روش‌های تئوری به عنوان روش جایگزین استفاده نمود. با استفاده از مطالعات QSPR می‌توان روابط خطی یا غیرخطی بین ویژگی‌های ساختاری و خواص ترکیبات برقرار کرده و خاصیت مورد نظر را برای ترکیبات پیش‌بینی کرد.

۱-۳-۲- مروری بر کارهای انجام شده

در سال ۱۹۸۵، آدلر[3] و همکاران، با استفاده از اندیس بازداری اندازه‌گیری شده با ستون ژل آبی[4] و حلال THF و محاسبه‌ی عدد واینر[5]، مقدار اندیس بازداری را برای PAHs پیش‌بینی کردند[۱۰]. در سال ۲۰۰۰، گروهی بر روی ۴۸ هیدروکربن آروماتیک چندحلقه‌ای مطالعات QSPR انجام دادند. آنها از مقادیر تجربی اندیس بازداری که با استفاده از کروماتوگرافی فاز معکوس اندازه‌گیری شده بود، برای

1-High performance liquid chromatography
2-Gas chromatography/mass spectrometry
3-Adler
4-Water gel column
5-Wiener number

مدل‌سازی این پارامتر با روش رگرسیون خطی چندگانه استفاده کردند. نتایج به دست آمده، توانایی مدل خطی برای پیش‌بینی اندیس بازداری را نشان می‌دهد[۱۱]. فریرا[۱] و ریبیرو[۲] در سال ۲۰۰۳ با استفاده از روش رگرسیون خطی اندیس بازداری ۶۷ مولکول آروماتیک چندحلقه‌ای را با صحت خوبی پیش‌بینی کردند[۱۲]. در سال ۲۰۰۶ مقدار اندیس بازداری به دست آمده بر روی دو فاز ساکن مختلف با روش شبکه عصبی مصنوعی مدل‌سازی شد. مدل‌های به دست آمده توافق خوبی با مقادیر تجربی نشان داده‌اند[۱۳]. در سال ۲۰۰۸ نیز، گروهی با استفاده از روش رگرسیون مرحله‌ای و مدل‌سازی MLR، اندیس بازداری را برای یک سری از هیدروکربن‌ها محاسبه نمودند. نتایج به دست آمده صحت خوبی را ارائه داده است[۱۴].

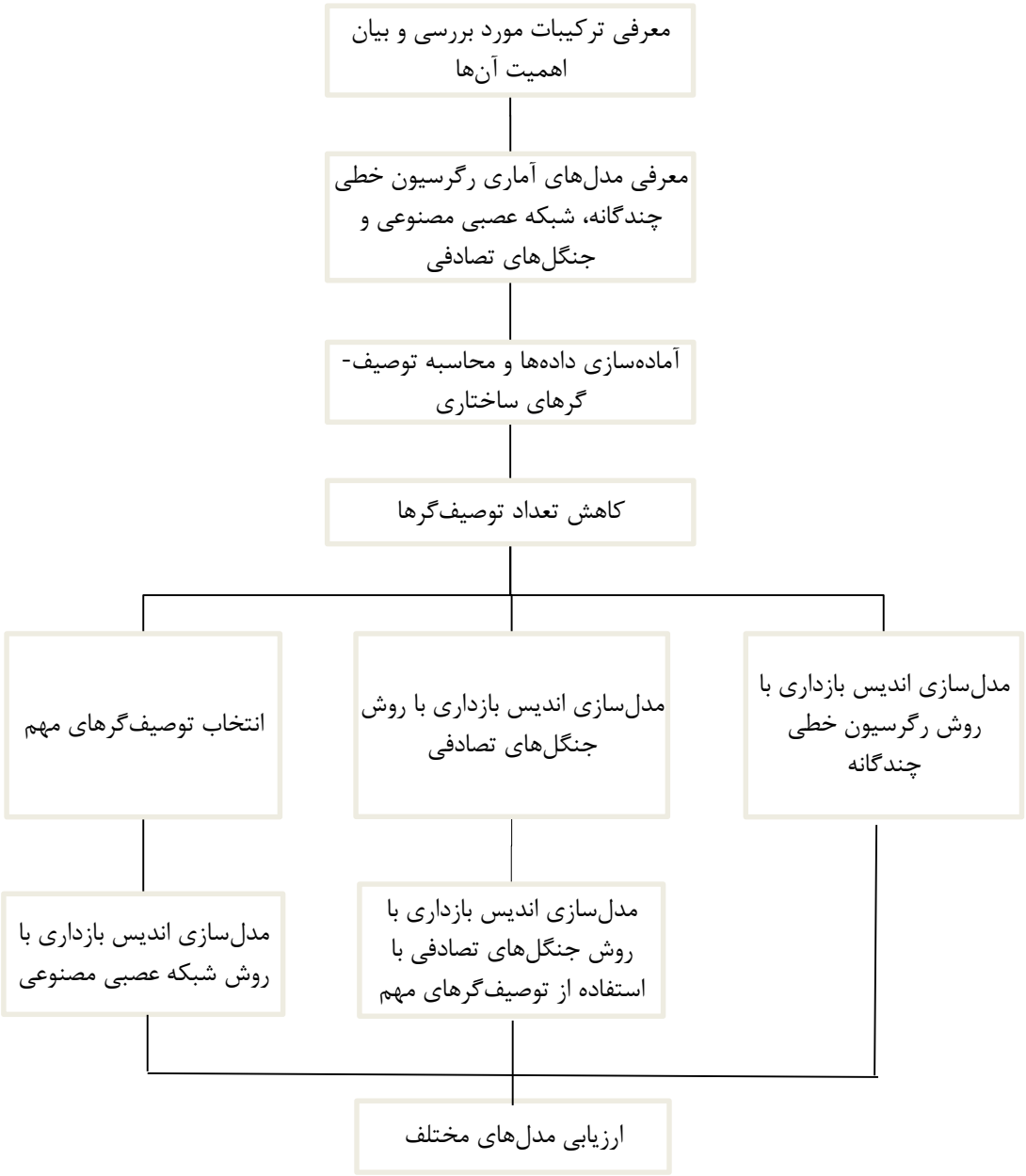
۱-۴- ضرورت انجام تحقیق

روش جنگل‌های تصادفی[۳] یکی از روش‌های پیش بینی خواص مختلف برای ترکیبات گوناگون است که بدون نیاز به یک روش انتخاب متغیر، مدل غیرخطی با صحت قابل قبول را ارائه می‌دهد. به همین دلیل در این پایان‌نامه از آن برای مدل‌سازی اندیس بازداری استفاده شده است. هم‌چنین به منظور ارزیابی عملکرد این روش، شبکه‌ی عصبی مصنوعی و رگرسیون خطی چندگانه برای مقایسه به کار برده شده‌اند.

1-Ferreira
2-Ribeiro
3-Random Forests

۱-۵- ساختار پایان‌نامه

هدف این پایان‌نامه، یافتن یک مدل مناسب برای پیش‌بینی اندیس بازداری ترکیبات مورد بررسی می‌باشد. شکل زیر، شمای کلی از مطالب ذکر شده در این پایان‌نامه را نشان می‌دهد. هر یک از این مراحل برای هر سری داده که شامل ترکیبات گوگرددار و ترکیبات آروماتیک چندحلقه‌ای است، انجام شده است.



شکل (۱-۱)- شمای کلی پایان‌نامه

فصل دوم
کمومتریکس

این فصل به معرفی مطالعات QSPR اختصاص داده شده است که یکی از کاربردهای روش کمومتریکس[1] می‌باشد. اصطلاح کمومتریکس برای اولین بار توسط اسوانت ولد[2] دانشمند جوان سوئدی مطرح گردید. طبق تعریف ICS[3]، کمومتریکس، کاربرد روش‌های ریاضی و آماری برای برقراری ارتباط بین سنجش‌های انجام شده روی یک سیستم و فرایند شیمیایی به منظور درک بهتر اطلاعات شیمیایی است[۱۵]. با توجه به اینکه گاهی اوقات شیمیدانان با موادی سروکار دارند که سمی و یا گران‌قیمت هستند، می توان به عنوان یک روش جایگزین از روش‌های ریاضی و آماری برای توصیف و توجیه نتایج آزمایش‌های مختلف استفاده کرد[۱۶]. به همین دلیل در سال‌های اخیر، شیمیدانان بیشترین استفاده از روش‌های کمومتریکس را داشته‌اند.

یکی از کاربردهای ویژه‌ی روش‌های کمومتریکس توسعه مطالعات QSPR و ارزیابی داده‌های تجزیه-ای از طریق آن است. تغییرات کوچک در ساختار مولکول‌های مشابه می‌تواند باعث ایجاد خواص شیمیایی کاملا متفاوتی گردد. نظریه QSPR بر این اساس است که خاصیت یک ترکیب به ویژگی‌های ساختاری آن بستگی دارد[۱۷] و هدف این روش یافتن رابطه‌ی هماهنگ میان خاصیت شیمیایی و ویژگی‌های مولکولی، به منظور کاربرد این قواعد برای ارزیابی خاصیت ترکیبات جدید می‌باشد.

مطالعات QSPR به طور گسترده در پیش‌بینی خواص فیزیکی و شیمیایی مختلفی مانند نقطه جوش و ذوب، فشاربخار، حلالیت، اندیس بازداری و غیره استفاده شده‌اند. این مطالعات شامل مراحل زیر است:

۱- فراهم کردن سری داده‌ها

۲- رسم و بهینه کردن ساختار هندسی ترکیبات

۳- محاسبه توصیف‌گرها

۴- انتخاب بهترین توصیف‌گرها

۵- مدل‌سازی

1-Chemometrics
2-Svante wold
3-International chemometrics society

۶- ارزیابی قدرت پیش‌بینی مدل[۱۸] .

۲-۱- فراهم کردن سری داده‌ها

برای دستیابی به مدل مناسب، اولین مرحله، انتخاب یک دسته از مولکول‌ها است که مقادیر تجربی یک خصوصیت فیزیکی یا شیمیایی آنها مانند زمان بازداری، فاکتور آبگریزی و ..... در دسترس باشد. در این خصوص لازم است شرایط اندازه‌گیری خاصیت مورد نظر برای همه‌ی ترکیبات یکسان باشد. یک مدل زمانی می‌تواند خاصیت مورد نظر را به درستی پیش‌بینی کند که مولکول‌های مورد استفاده، تقریبا مشابه بوده یا گروهی از مشتقات یک ترکیب باشند. یافتن سری داده‌ی مناسب از طریق جستجو در مقالات صورت می‌گیرد.

۲-۲- بهینه‌سازی ساختار هندسی ترکیبات

ایجاد توصیف‌گرهای هندسی و هیبریدی، بر مبنای ساختار و هندسه‌ی دقیق مولکولی استوار است. اگر ساختارها صورت بندی[1] با حداقل انرژی نداشته باشند، مقادیر غیرصحیحی برای این توصیف‌گرها ایجاد می‌شود[۱۹]. بهینه کردن ساختار ترکیبات توسط شیمی محاسباتی و نرم‌افزار Hyperchem انجام می‌شود. روش‌های مختلفی در شیمی محاسباتی برای بهینه‌سازی ساختار مولکول مورد استفاده قرار می‌گیرند. این روش‌ها شامل روش مکانیک مولکولی و روش مکانیک کوانتومی است. روش‌های مکانیک مولکولی بر مبنای روابط کلاسیک بنا شده است و به طور ساده ساختمان مولکولی به صورت گوی و فنر در نظر گرفته می‌شود. در این روش انرژی مولکول به صورت مجموعه‌ای از انرژی‌های کششی، خمشی و الکترواستاتیکی بیان می‌شود. سپس می‌توان طول پیوندها، زوایای پیوندی و صورت‌بندی را به گونه‌ای تغییر داد تا ساختاری که عبارت انرژی مکانیک مولکولی را مینیمم کند، پیدا شود. این روش بهینه‌سازی

1-Conformation

بیشتر برای ماکرومولکول‌ها به کار می‌رود[۲۰]. ولی در مکانیک کوانتومی از معادله شرودینگر[۱] برای محاسبه‌ی انرژی مولکول استفاده می‌شود و نسبت به روش مکانیک مولکولی دارای صحت بیشتری است؛ زیرا اثرات الکترونی اعمال شده روی مولکول را نیز در محاسبات وارد می‌کند[۲۱]. روش‌های محاسباتی بر پایه مکانیک کوانتومی شامل گروه‌های زیر می‌باشند[۲۲]:

۱- روش‌های آغازین[۲]

۲- روش‌های نیمه‌تجربی[۳]

۳- روش‌های تابع چگالی[۴]

روش‌های نیمه‌تجربی، روش‌هایی هستند که فقط الکترون‌های لایه ظرفیت را در محاسبات وارد می‌کنند. روش‌های نیمه‌تجربی مختلفی مانند چشم‌پوشی کامل از انتگرال‌های هم‌پوشانی دیفرانسیلی[۵] (CNDO)، چشم‌پوشی از هم‌پوشانی دو اتمی اصلاح شده (MNDO)[۶] و AM1[۷] وجود دارد[۲۳]. روش AM1 که در این پایان‌نامه از آن استفاده شده است، جزء دقیق‌ترین روش‌های نیمه‌تجربی است که همان روش MMDO اصلاح شده می‌باشد. در این روش هسته و لایه‌های داخلی به شکل یک هسته‌ی مرکزی در نظر گرفته شده و محاسبات روی الکترون‌های لایه ظرفیت انجام می‌گیرد[۲۱].

۲-۳- محاسبه توصیف‌گرها

توصیف‌گرهای مولکولی، مقادیری عددی هستند که ساختار یا شکل مولکول را توصیف نموده و ویژگی‌های ساختاری و الکترونی مولکول را به طور کمی نشان می‌دهند. برخی از ویژگی‌های یک توصیف‌گر مناسب عبارت است از[۲۴]:

۱- ساده بودن

1-Schrodinger equation
2-Ab initio methods
3-Semi-empirical methods
4-Density functional theory (DFT)
5-Complete neglect of differential overlap
6-Modified neglected of diatomic overlap
7-Austian method 1

۲- مستقل بودن

۳- توانایی تفسیر ساختار مولکول

۴- عدم هم‌بستگی با سایر توصیف‌گرها

۵- قابلیت تمایز بین ایزومرهای مختلف یک مولکول

۶- قابل کاربرد برای دامنه وسیعی از ساختارهای مولکولی

توصیف‌گرهای مولکولی به روش‌های مختلفی دسته‌بندی می‌شوند. ساده‌ترین طبقه‌بندی بر اساس ماهیت آنها (تجربی یا نظری) است. توصیف‌گرهای تجربی بر اساس اندازه‌گیری‌های تجربی به دست می‌آیند، در حالیکه توصیف‌گرهای نظری با استفاده از ساختار مولکول و بدون نیاز به داده‌های تجربی محاسبه می‌شوند که باعث صرفه‌جویی در وقت، هزینه، مواد و تجهیزات خواهد شد و برای هر مولکولی که هنوز سنتز نشده است نیز در دسترس می‌باشد[۱۹و ۲۵]. در این تحقیق از توصیف‌گرهای نظری استفاده شده است که محاسبه‌ی آن‌ها توسط نرم‌افزار DRAGON انجام می‌گیرد.

۲-۴- انتخاب توصیف‌گرهای مهم

انتخاب بهترین توصیف‌گرها که بیشترین وابستگی با خاصیت مورد نظر را دارند، از مهم‌ترین مراحل می‌باشد. وجود توصیف‌گرهای زیاد و هم‌چنین وجود هم‌بستگی میان آنها، محاسبات و مدل‌سازی را دشوار و پیچیده می‌کند. لذا باید توصیف‌گرهایی انتخاب شوند که با خاصیت مورد نظر بیشترین وابستگی را داشته و تغییرات آن را به خوبی توصیف کنند. روش‌های به کار برده شده برای انتخاب توصیف‌گر مناسب در این پایان‌نامه، رگرسیون گام‌به‌گام یا مرحله‌ای [1]و راهبرد تلفیق تقسیم داده‌ها و انتخاب ویژگی (CDFS)[2] است که توسط نرم‌افزار SPSS انجام می‌شود.

1-Stepwise
2-Combined data splitting–feature selection

۲-۴-۱- رگرسیون مرحله‌ای

این روش با محاسبه‌ی ضرایب هم‌بستگی که معیاری برای سنجش هم‌بستگی بین یک توصیف‌گر و متغیر پاسخ است، انجام می‌شود. مقدار این ضریب بین صفر و یک است که مقدار صفر برای یک توصیف‌گر بیانگر این است که هیچ ارتباط خطی بین توصیف‌گر و خاصیت وجود ندارد و مقدار یک، بیانگر ارتباط کامل بین توصیف‌گر و خاصیت است. در این روش ابتدا توصیف‌گری که بیشترین هم-بستگی با متغیر وابسته (خاصیت موردنظر) را دارد، وارد مدل شده و با ورود هر توصیف‌گر جدید، کلیه‌ی توصیف‌گرهای موجود در مدل بررسی شده و اگر هر کدام از آنها سطح معناداری خود را از دست داده باشند، قبل از ورود توصیف‌گر جدید، از مدل خارج می‌شوند[۲۶].

۲-۴-۲- راهبرد CDFS

اشکال بزرگی که در روش رگرسیون مرحله‌ای برای انتخاب توصیف‌گر وجود دارد، امکان انتخاب توصیف‌گرهایی است که وابستگی زیادی با متغیر پاسخ ندارند ولی در مدل ظاهر می‌شوند. این اتفاق مخصوصا زمانی که نسبت تعداد داده‌ها به تعداد توصیف‌گرها کم باشد، رخ خواهد داد. راهبرد CDFS یک روش انتخاب توصیف‌گرهای مهم است که با استفاده از آن می‌توان با تعداد توصیف‌گرهای کمتر، قدرت پیش‌بینی مدل را افزایش داد. مراحل این روش به صورت زیر است:

ابتدا تعدادی از مولکول‌ها به طور تصادفی به عنوان داده‌های آزمون انتخاب و کنار گذاشته می‌شوند. داده‌های باقی مانده چندین بار به طور تصادفی به دو سری آموزش و ارزیابی تقسیم می‌گردند. رگرسیون مرحله‌ای روی هر سری آموزش انجام شده و از بین مدل‌های به دست آمده، مدلی که بهترین عملکرد را برای پیش‌بینی سری ارزیابی مربوطه انجام دهد، به عنوان مدل برتر انتخاب می‌شود. بنابراین به ازای هر دسته‌بندی یک مدل بهینه به دست خواهد آمد. بررسی مدل‌های تولید شده نشان می‌دهد که بعضی از توصیف‌گرها در همه یا اکثر مدل‌ها تکرار شده‌اند که نشان‌دهنده‌ی اهمیت آنها در پیش‌بینی متغیر پاسخ می‌باشد. توصیف‌گرهایی که در بیش از ۵۰٪ مدل‌ها تکرار شده باشند، به عنوان توصیف‌گرهای

نهایی انتخاب می‌شوند. به دلیل اینکه توصیف‌گرها در این روش توسط داده‌های آموزش مختلفی به دست می‌آیند، نسبت به نوع مولکول‌ها حساس نیستند[۲۷ و ۲۸].

۲-۵- ساخت مدل

ساخت مدل به منظور برقراری ارتباط بین خاصیت مورد بررسی و توصیف‌گرها صورت می‌گیرد. یک مدل مناسب که هم‌بستگی قابل قبولی بین مقادیر تجربی خاصیت مورد نظر و مقدار به دست آمده به صورت تئوری نشان می‌دهد، به پیش‌بینی رفتار مولکول‌های جدید کمک می‌نماید. در مطالعات QSPR از روش‌های خطی مانند رگرسیون خطی چندگانه (MLR) و حداقل مربعات جزئی (PLS)[1] و روش‌های غیرخطی مانند شبکه عصبی مصنوعی (ANN)، ماشین بردار پشتیبان (SVM)[2] و جنگل‌های تصادفی (RF) استفاده می‌شود. در این پایان‌نامه، روش‌های MLR، ANN و RF به کار برده شده‌اند.

۲-۶- رگرسیون خطی چندگانه

در مدل رگرسیون خطی چندگانه، رابطه‌ای خطی بین یک متغیر وابسته (خاصیت مورد نظر) و متغیرهای مستقل (توصیف‌گرها) به صورت زیر تعریف می‌شود:

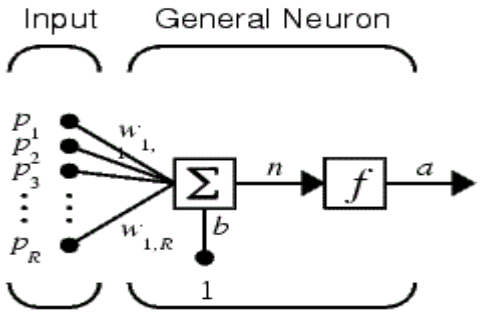
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (۱-۲)$$

که در آنها β_i ضرایب رگرسیون، Y متغیر وابسته و x_i متغیرهای مستقل هستند. مقدار ضریب متغیر مستقل در معادله، میزان اهمیت آن متغیر در پیش‌بینی مقادیر متغیر وابسته را نشان می‌دهد[۲۶].

1-Partial least square
2- Support vector machin

۲-۷- شبکه عصبی مصنوعی

شبکه عصبی مصنوعی یک الگوی پردازش اطلاعات است که از سیستم‌های عصبی بیولوژیکی مانند مغز الهام می‌گیرد. این الگو از تعداد زیادی عناصر پردازش‌کننده که به هم متصل هستند تشکیل شده است که به طور هماهنگ برای حل مسائل ویژه عمل می‌کنند و مانند مغز انسان با مثال، آموزش می‌بینند. در مغز انسان یک نرون سیگنال‌ها را از سایر نرون‌ها توسط بخشی از ساختار خود به نام دندریت جمع‌آوری می‌کند. دندریت پس از دریافت اطلاعات، آنها را به شکل سیگنال به هسته سلول هدایت می‌کند. هسته سلول پس از پردازش سیگنال‌های دریافتی، آنها را از طریق جزء دیگری به نام اکسون به نرون‌های دیگر انتقال می‌دهد. این کار توسط سیناپس‌ها که ارتباط‌دهنده‌ی نرون‌ها هستند، صورت گرفته و بدین ترتیب فعالیت‌های مغزی انجام می‌شود[۲۹و۳۰]. با توجه به ساختار نرون محاسباتی در شکل (۲-۱) و مقایسه آن با نرون طبیعی، میتوان ورودی‌ها را به دندریت، تابع محرک را به بدنه سلول، وزن‌ها را به شدت سیناپس‌ها و خروجی را به سیگنال گذرنده از اکسون تشبیه کرد[۳۱].



شکل (۲-۱)- ساختمان یک نرون محاسباتی

در شکل فوق، کمیت‌های p و a، به ترتیب ورودی و خروجی نرون می باشد. میزان تاثیر ورودی p روی خروجی a به وسیله پارامتر وزن (w) تعیین می‌شود. ورودی دیگر که مقدار ثابت ۱ است و بایاس[1] نامیده می‌شود، در وزن مربوطه یعنی b ضرب شده و با wp جمع می‌شود. این حاصل‌جمع، ورودی خالص n را برای تابع محرک f تشکیل می‌دهد. ورودی و خروجی یک نرون از طریق روابط زیر به دست می‌آید[۳۱]:

1-Bias

$$n = w_{1.1}p_1 + w_{1,2}p_2 + \cdots + w_{1,R}p_R + b \quad (۲-۲)$$

$$a = f(w_{1.1}p_1 + w_{1.2}p_2 + \cdots + w1_{1,R}p_R + b) \quad (۳-۲)$$

۲-۷-۱- مراحل آموزش در شبکه‌های پیشخور[1]

این مراحل شامل ایجاد شبکه، مقداردهی آغازین به وزن‌ها و آموزش شبکه می‌باشد[۳۲]. در نرم-افزار MATLAB تابع newff یک شبکه پیشخور ایجاد می‌کند. فرایند آموزش شبکه نیاز به یک سری مثال‌ها در خصوص رفتار مورد انتظار شبکه دارد که شامل ورودی شبکه و مقادیر هدف آن ورودی می‌باشد. در طول آموزش، وزن‌ها و بایاس‌ها طوری تنظیم می‌شوند تا تابع کارایی حداقل شود. تابع کارایی پیش‌فرض برای شبکه‌های پیشخور میانگین مربع خطا (MSE)[2] است.

۲-۷-۲- آموزش شبکه با تکنیک پس‌انتشار[3]

به طور کلی آموزش به کمک تکنیک پس‌انتشار طبق مراحل زیر انجام می‌شود:

۱- انتشار ورودی‌ها از نرون‌های ورودی

۲- دادن وزن تصادفی به هر یک از اتصالات

۳- محاسبه خطای شبکه از طریق مقایسه‌ی خروجی شبکه با مقادیر هدف

۴- پس‌انتشار خطا از نرون‌های خروجی به نرون‌های ورودی و اصلاح وزن‌ها

۵- ارزیابی عملکرد شبکه

مراحل فوق تا به حداقل رسیدن تابع کارایی یا رسیدن به حد مجاز تکرار انجام می‌شود[۳۲].

برای توضیحات بیشتر در مورد شبکه عصبی مصنوعی به مرجع [۳۳] مراجعه شود.

1-Feed forward
2-Mean square error
3-Back propagetion

۲-۸- روش جنگل‌های تصادفی

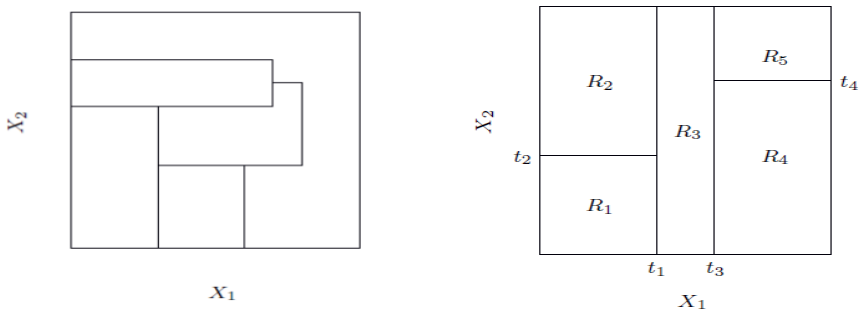
جنگل‌های تصادفی یکی از روش‌های یادگیری ماشین[1] است که در سال ۲۰۰۱ توسط بریمن[2] متخصص آمار دانشگاه برکلی آمریکا ارائه شد [۳۴]. از آنجا که ساختار جنگل‌های تصادفی را مجموعه‌ای از درخت‌های رده‌بندی یا رگرسیونی (CART)[3] مستقل تشکیل می‌دهد، ابتدا به توضیح این روش پرداخته می‌شود.

۲-۸-۱- روش درخت‌های رده‌بندی و رگرسیونی

این روش که توسط بریمن به عنوان جایگزینی برای روش‌های مبتنی بر شبکه معرفی شده است، هم برای رده‌بندی داده‌های کیفی و هم برای پیش‌بینی داده‌های کمی به کار می‌رود. از آنجایی که تمرکز این پایان‌نامه روی پیش‌بینی داده‌های کمی است، تنها به بخش رگرسیونی آن با نام درخت رگرسیونی پرداخته می‌شود. این روش بر پایه تقسیم کردن مجموعه داده به قسمت‌های کوچکتر است. در روش رگرسیون خطی، پیش‌بینی‌کننده‌های رگرسیونی، مدل‌هایی هستند که در آنها یک مدل پیش-بینی واحد روی فضای کل داده برقرار است. اما ممکن است به دلیل تفاوت رفتار متغیر پاسخ در نواحی مختلف، برقراری یک مدل واحد، کارایی لازم را نداشته باشد. لذا یک روش جایگزین، تقسیم‌بندی فضای داده به بخش‌های کوچکتر است تا بتوان رفتار متغیر پاسخ را به طور موضعی مدل‌سازی نمود. در روش درخت رگرسیونی هدف این است که از مجموعه کل داده‌ها، داده‌هایی که مقادیر متغیر پاسخ آنها به یکدیگر نزدیک است، در یک ناحیه قرار گیرند و سپس به هر ناحیه، یک رویه پاسخ[4] یا یک مدل رگرسیونی ساده نسبت داده شود. در این روش، افراز یا تقسیم فضای کل توصیف‌گرها باید با این شرط انجام شود که تمامی نواحی ساخته شده به شکل مربع یا مستطیل باشد. در حالت سه متغیره و بیشتر،

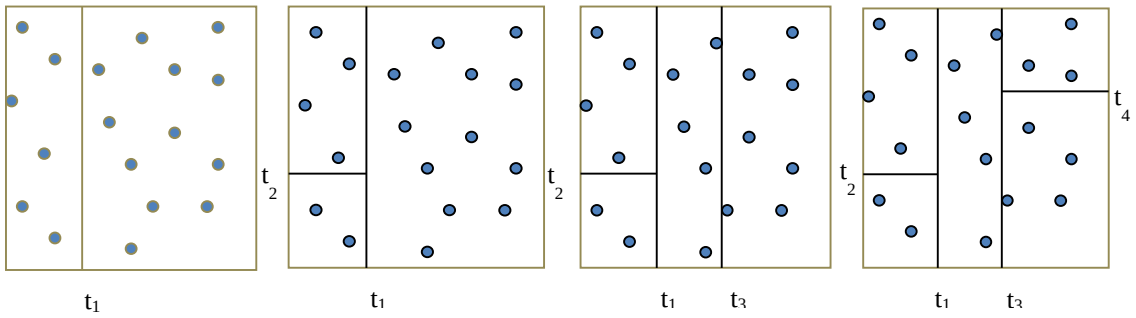
1-Machine learning
2-Breiman
3-Classification and regrression trees
4-Response surface

این قطعات به صورت مکعب مستطیل یا ابرمکعب خواهد بود. در شکل (۲-۲) افراز قابل‌قبول و غیرقابل-قبول در این روش نشان داده شده است که در آن، x_1 و x_2 متغیرهای مستقل (توصیف‌گرها) هستند.

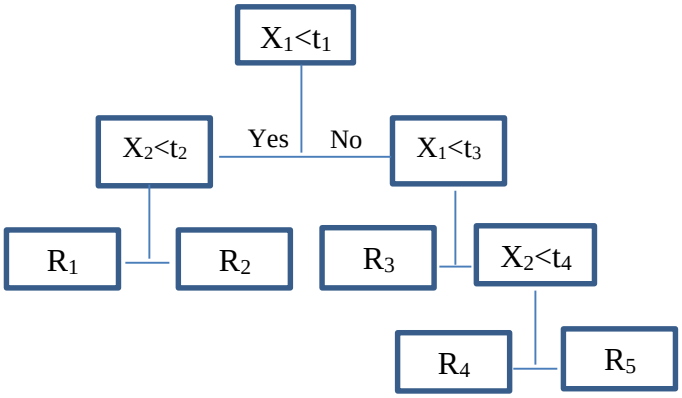


شکل (۲-۲)- نمونه‌ای از افراز صحیح (راست) و غیر صحیح (چپ) در فضای دومتغیره

دلیل نام‌گذاری درخت رگرسیونی را می‌توان به نمایش درختی افراز فضای توصیف‌گرهای ورودی نسبت داد. فرض کنید روند سلسله‌مراتبی افراز شکل ((۲-۲)-راست) به صورت شکل (۲-۳) باشد که می‌توان این روند را به صورت یک نمودار درختی نشان داد. (شکل (۲-۴)).



شکل (۲-۳)- روند سلسله مراتبی افراز شکل (۲-۲)-راست



شکل (۲-۴)- نمودار درختی فضای افراز شده در شکل (۲-۳)

در این نمودار درختی، هر مشاهده از نقطه بالای نمودار وارد شده و بسته به مقادیر توصیف‌گرهای خود، در یکی از نواحی افراز قرار می‌گیرد و مقدار $\hat{Y}$ مربوط به همان ناحیه را اختیار می‌کند.

بنا به هدف دنبال شده در روش درخت رگرسیونی، رویه پاسخ در هر ناحیه مانند R_m، توسط یک مدل ساده یعنی $\hat{Y}_{R_m} = C_m$ مدل‌سازی می‌شود که در آن بهترین تقریب برای c_m، میانگین مقادیر متغیر پاسخ در ناحیه R_m است. به عبارت دیگر:

$$\hat{y}_{R_m} = \bar{y}_{R_m} = \frac{1}{n_m}\sum_{j=1}^{n} y_j \,,\ y_j \in R_m \qquad (۴-۲)$$

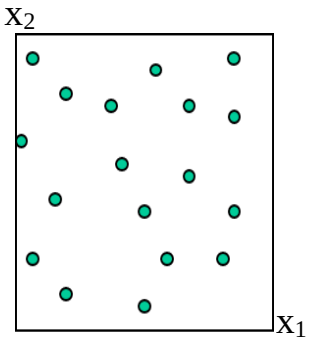
که در آن n_m تعداد مشاهدات مجموعه R_m است. بنابراین روش درخت رگرسیونی به صورت زیر بیان می‌شود:

$$\hat{y} = \sum_{m=1} c_m I_{R_m}(x_1, x_2) \qquad (۵-۲)$$

که در آن، $I_{R_m}(x) = \begin{cases} 1, & x \in R_m \\ 0, & x \notin R_m \end{cases}$ می‌باشد[۳۵].

۲-۸-۱-۱- الگوریتم تشکیل درخت رگرسیونی

برای ساده کردن موضوع، ابتدا یک مسئله با دو توصیف‌گر را در نظر می‌گیریم. فرض کنید نمودار پراکندگی داده‌ها به صورت شکل (۲-۵) باشد:

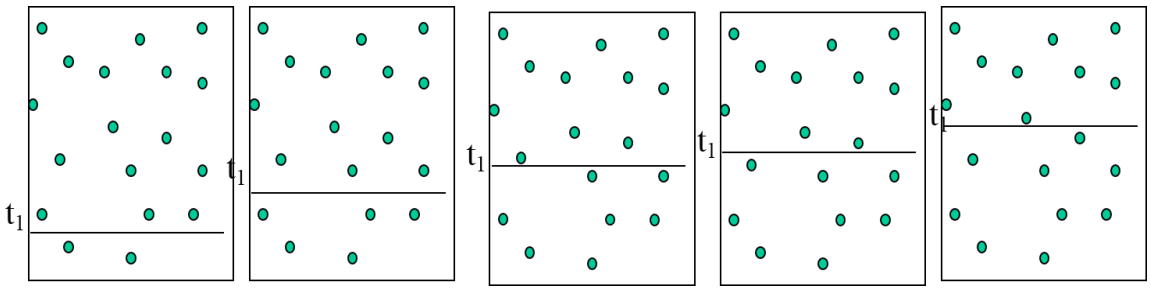


شکل (۲-۵)- نمودار پراکندگی داده‌های دومتغیره

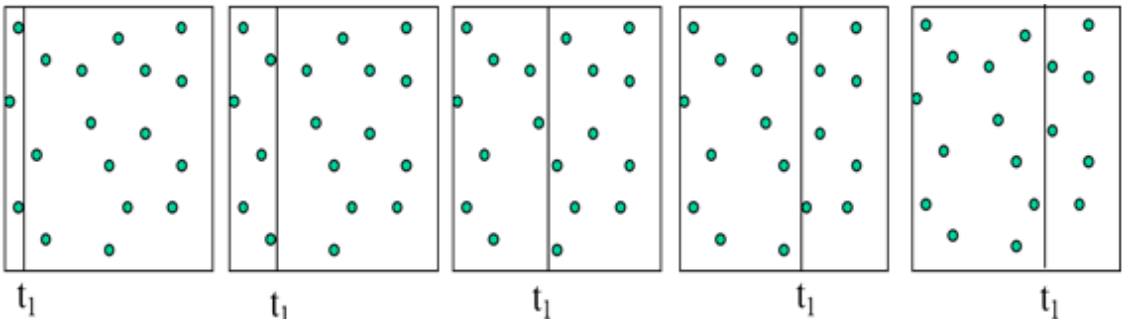
برای اولین افراز، انتخاب‌های مختلفی وجود دارد. شکل‌های (۲-۶) و (۲-۷)، برخی از افرازهای ممکن در راستای دو توصیف‌گر را نشان می‌دهند. برای انتخاب بهترین تقسیم‌بندی، حاصل‌جمع مجموع مربعات خطا (A) در دو قسمت ایجاد شده به صورت زیر محاسبه می‌شود:

$$A = \sum_{i=1}^{n}(y_{i1} - \bar{y}_{i1})^2 + \sum_{i=1}^{m}(y_{i2} - \bar{y}_{i2})^2 \qquad (۶-۲)$$

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 و $\bar{y}_i$، 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، فرایند فوق در هر یک از نواحی تولید شده تکرار می‌شود.



شکل (۲-۶)- برخی از افرازهای ممکن در راستای توصیف‌گر x_1



شکل (۲-۷)- برخی از افرازهای ممکن در راستای توصیف‌گر x_2

در حالت چندمتغیره باید تمام افرازهای ممکن در راستای هر توصیف‌گر در نظر گرفته شده و عبارت متناظر (۲-۶) محاسبه شود. به عبارت دیگر، روش CART در هر مرحله افراز فضای توصیف‌گرها باید هم‌زمان جستجو کند که کمترین مقدار عبارت (۲-۶) به ازای افراز در چه راستایی و از کدام نقطه (گره) حاصل می‌شود[۳۵].

۲-۸-۱-۲- اندازه‌ی 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 توصیف‌گرها 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[1] وجود دارد. برای انتخاب اندازه‌ی درخت نیاز به یک پارامتر در ساختار مدل است به طوری که این پارامتر مقدار بهینه‌ی اندازه 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 بر اساس 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 درخت در آن مرحله صرف نظر کرد. این میزان کاهش در مجموع توان‌های دوم خطا می‌تواند در اختیار کاربر باشد[۳۵].

۲-۸-۲- جنگل‌های تصادفی

اساس روش جنگل های تصادفی وابسته به ماهیت روش درخت رگرسیونی است و متشکل از ترکیب مجموعه‌ای از درخت‌های رگرسیونی می‌باشد. هر درخت، یک مدل تولید می‌کند و از برآیند یا ترکیب این مدل‌ها، مدل نهایی به دست می‌آید. بر خلاف روش درخت رگرسیونی، در این روش برای ساخت هر درخت، تنها بخشی از داده‌ها مورد استفاده قرار می‌گیرند؛ زیرا انتخاب داده‌های آموزش باجایگذاری[2] انجام می‌شود. هم‌چنین در جنگل‌های تصادفی برای افراز فضای توصیف‌گرها و ساخت هر گره، به جای استفاده از تمام توصیف‌گرها، از زیرمجموعه‌ای از توصیف‌گرها استفاده می‌شود.

۲-۸-۲-۱- الگوریتم روش جنگل‌های تصادفی

فرض کنید $i=1,...,N$ و (X_i,Y_i) مجموعه‌ای از داده‌های آموزش باشد که در آن$Xi=(x_{i1},.....x_{iM})$، برداری از M توصیف‌گر و Yi متغیر پاسخ متناظر آنها است. اگر تعداد کل درخت‌های مدل با ntree نشان داده شود، مراحل زیر بیان‌گر ساخت هر درخت است:

1-Over fitting
2-With replacement

الف-یک نمونه N تایی از داده‌های آموزش به روش باجایگذاری برای ساخت مدل انتخاب می‌شوند. حدود یک‌سوم از داده‌های اصلی که در ساخت درخت شرکت نمی‌کنند، داده‌های خارج از کیسه (OOB)[1] نام دارند که برای هر درخت متفاوت هستند و نقش داده‌های ارزیابی را برای آن درخت ایفا می‌کنند.

ب- با استفاده از داده‌های انتخاب شده در مرحله اول و انتخاب تصادفی m توصیف‌گرکه زیرمجموعه‌ای از تعداد کل توصیف‌گرها است، جستجو برای یافتن بهترین توصیف‌گر، فقط در بین m توصیف‌گر مذکور انجام شده و فضای توصیف‌گرها 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 هر یک از دو ناحیه تولید شده، با استفاده از همان نمونه آموزش انتخاب شده در مرحله (الف)، مجددا مرحله (ب) انجام می شود.

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

ح- مدل درخت رگرسیونی iام ساخته شده که فضای توصیف‌گرها را به r ناحیه ی $R_{i1},R_{i2},\ldots..R_{ir}$ تقسیم کرده است، به صورت زیر می باشد:

$$\hat{f}_i(x) = \sum_{j=1}^{r_i} \hat{c}_j I_{R_{ij}}(x) \qquad (۷-۲)$$

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 ntree درخت رگرسیونی وجود دارد، می توان گفت که تعداد ntree مدل به صورت معادله (۲-۷) خواهیم داشت. اگر برای هر داده، مدل خروجی درخت iام را به صورت $\hat{f}_i(x)$ نشان دهیم، برآورد متغیر پاسخ برای این داده، با میانگین‌گیری از مدل خروجی

1-Out of bag

تمام درخت‌ها یعنی از رابطه زیر به دست می آید[۳۴]:

$$\hat{y}(x) = \frac{1}{ntree}\sum_{i=1}^{ntree}\hat{f}_i(x) \qquad (۸-۲)$$

۲-۸-۲-۲- تعیین اهمیت توصیف‌گرها در روش جنگل‌های تصادفی

در ساختار روش RF امکانی وجود دارد که می‌توان توصیف‌گرهایی که نقش بیشتری در مدل دارند را شناسایی کرد. برای این کار از داده‌های OOB در هر درخت استفاده می‌شود. فرض کنید داده‌های OOB در درخت iام با OOB_i نمایش داده شود. از آنجا که این داده‌ها، حکم داده‌های ارزیابی را دارند، لذا می‌توان با استفاده از رابطه (۲-۸) مقدار برآورد متغیر پاسخ آنها را یافته و سرانجام خطای برآورد یعنی $EOOB_i$ را محاسبه نمود. برای تعیین اهمیت توصیف‌گر jام در درخت iام باید مقادیر این توصیف‌گر را در بلوک داده‌های OOB_i به صورت تصادفی جابه‌جا کرد تا داده‌های دیگری($\widetilde{OOB_i^j}$) به وجود آید که برای آن می‌توان میزان خطای برآورد یعنی $\widetilde{EOOB_i^j}$ را محاسبه نمود. واضح است که هر چقدر تفاوت مقدار این دو خطا بیشتر باشد، نشان دهنده‌ی آن است که توصیف‌گر jام، توصیف‌گر مؤثری بوده و برعکس اگر این تفاوت کم باشد، نشان از عدم تأثیرگذاری توصیف‌گر jام است[۳۴]. لذا می‌توان میزان اهمیت توصیف‌گر jام در درخت iام ($VI_i(x^j)$) را به صورت زیر تعریف نمود:

$$VI_i(x^j) = \widetilde{EOOB_i^j} - EOOB_i \qquad (۹-۲)$$

سرانجام، میزان اهمیت توصیف‌گر jام در مدل نهایی ($VI(x^j)$) به صورت زیر است:

$$VI(x^j) = \frac{1}{ntree}\sum_{i=1}^{ntree}\left(VI_i(x^j)\right) \qquad (۱۰-۲)$$

لازم به ذکر است که برای دقت بیشتر می‌توان جابه‌جا نمودن تصادفی مقادیر توصیف‌گر jام را به تعداد بیشتر از یک بار انجام داد و در هر بار خطای برآورد را محاسبه نمود و سرانجام مقدار $\widetilde{EOOB_i^j}$ را به صورت میانگین خطاهای حاصله در نظر گرفت[۳۴ و ۳۶].

۲-۸-۲-۳- مزایای روش جنگل‌های تصادفی

۱- این روش در مجموعه داده‌هایی که تعداد توصیف‌گر زیادی دارد از کارایی بالایی برخوردار است. زیرا در هر مرحله از مدل‌سازی تنها زیرمجموعه‌ای از توصیف‌گرها 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 نظرگرفتن تعدادی از داده‌ها به عنوان سری ارزیابی وجود ندارد.

۴- این روش توانایی انتخاب متغیر را دارا است و برخلاف سایر روش‌های یادگیری ماشین، نیازمند روشی دیگر برای انتخاب متغیر نیست[۳۴].

۲-۹- ارزیابی مدل

شاخص‌هایی که به وسیله آنها بتوان مدل برازش شده را ارزیابی نمود عبارتند از:

ضریب تعیین[1]: نشان‌دهنده‌ی نسبت تغییرات بیان شده توسط مدل انتخاب شده به تغییرات کل است.

رابطه ریاضی مربوط به ضریب تعیین به صورت زیر است:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (۱۱-۲)$$

که SSR[2]، SST[3] و SSE[4] طبق روابط زیر تعریف می شوند:

مجموع مربعات انحراف مقادیر پیش‌بینی شده‌ی متغیر وابسته از میانگین مقادیر آن:

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (۱۲-۲)$$

مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از میانگین آن:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (۱۳-۲)$$

1-Coefficient of determination
2-Sum square of regression
3-Sum square of total
4-Sum square of error

مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش‌بینی شده آن:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (۱۴-۲)$$

با توجه به روابط فوق می‌توان نوشت:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (۱۵-۲)$$

از آنجا که اگر به ازای تمام نقاط $y_i=\hat{y}_i$ باشد، مقدار این ضریب یک است، لذا نزدیک بودن ضریب تعیین به مقدار ۱ نشان دهنده مناسب بودن مدل انتخاب شده است.

مجموع مربع باقی مانده‌ی پیش‌بینی[۱]

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (۱۶-۲)$$

میانگین مربع خطا

$$MSE = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n} \qquad (۱۷-۲)$$

خطای استاندارد پیش‌بینی[۲]

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \qquad (۱۸-۲)$$

خطای مطلق میانگین[۳]

$$MAE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n} \qquad (۱۹-۲)$$

خطای نسبی پیش‌بینی[۴]

$$REP(\%) = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \qquad (۲۰-۲)$$

1-Predictive residual sum of square
2-Standard error of prediction
3-Mean absolute error
4-Relative error of prediction

میانگین خطای نسبی[1]

$$\mathrm{MRE} = \frac{\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}}{y_i}\right|}{n} \times 100 \qquad (۲۱-۲)$$

در روابط فوق، y_i و $\hat{y}_i$ به ترتیب نشان‌دهنده‌ی مقدار واقعی و مقدار پیش‌بینی شده‌ی متغیر وابسته در ترکیب iام است. $\bar{y}$ نیز میانگین مقادیر واقعی متغیر وابسته و n تعداد ترکیبات مورد بررسی است[۳۷].

تعدادی از شاخص‌های ارزیابی نیز برای مقایسه مدل‌های مختلف با تعداد متفاوت توصیف‌گر استفاده می‌شوند. این شاخص‌ها عبارتند از:

ضریب تعیین تعدیل یافته[2]

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p-1}.\frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - (1-\mathrm{R}^2)\frac{n-1}{n-p-1} \qquad (۲۲-۲)$$

آماره F : یا آزمون فیشر در واقع آزمون معنی‌دار بودن آماری در تحلیل رگرسیون ساده و چندمتغیره است و برابر با نسبت میانگین مربعات رگرسیون (MSR) به میانگین مربعات باقی مانده‌ها (MSE) است.

$$\mathrm{F} = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR}/p}{\mathrm{SSE}/n-p-1} \qquad (۲۳-۲)$$

FIT[3]: این پارامتر نیز به عنوان یک معیار برای مقایسه مدل‌های مختلف با تعداد متفاوت توصیف‌گر به کار می‌رود.

$$\mathrm{FIT} = \frac{\mathrm{R}^2(n-p-1)}{(n+p^2).(1-\mathrm{R}^2)} \qquad (۲۴-۲)$$

در این روابط n تعداد ترکیبات مربوط به مدل و p تعداد توصیف‌گرهای مدل است. هرچه مقدار این سه پارامتر بیشتر باشد، آن مدل، مناسب‌تر است[۳۷ و ۳۸].

1-Mean relative error
2-Adjusted determination coefficient
3-Fitness function

۲-۱۰- نرم‌افزارهای مورد استفاده

نرم‌افزارهای مورد استفاده در این تحقیق عبارتند از:

۲-۱۰-۱- بسته نرم‌افزاری Hyperchem

بسته نرم‌افزاری Hyperchem [۲۱] در رسم شکل و بهینه کردن ساختار ترکیبات به کار می‌رود. داده‌های حاصل از این نرم‌افزار را می‌توان به عنوان ورودی به سایر نرم‌افزارها معرفی کرد. هم‌چنین به کمک این نرم‌افزار می‌توان تعدادی از توصیف‌گرها مثل حجم مولی و قطبش‌پذیری را محاسبه نمود.

۲-۱۰-۲- بسته نرم‌افزاری Dragon

این نرم‌افزار توسط گروه تحقیقاتی کمومتریکس میلانو ارائه شده است[۳۹] و امکان محاسبه ۱۴۸۱ توصیف‌گر مختلف را فراهم می‌کند. برای محاسبه‌ی توصیف‌گر توسط این نرم‌افزار، ساختار بهینه‌ی مولکول مورد استفاده قرار می‌گیرد.

۲-۱۰-۳- بسته نرم‌افزاری SPSS[1]

این نرم‌افزار امکان تجزیه و تحلیل داده‌های آماری را فراهم می‌آورد. این بسته نرم‌افزاری، قابلیت محاسبه‌ی میانگین ساده برای داده‌ها، نمایش اطلاعات به صورت متنوع و انجام رگرسیون تک‌متغیره و چندمتغیره را دارا است[۴۰].

1-Statistical package for the social sciencae

۲-۱۰-۴- نرم‌افزار MATLAB[1]

این نرم‌افزار یکی از جامع‌ترین و کامل‌ترین نرم‌افزارهای علمی و محاسباتی است. ورودی‌ها در این نرم‌افزار به صورت ماتریس در نظر گرفته می‌شوند. از جمله کاربردهای جالب آن، شبکه عصبی مصنوعی است که در این پایان‌نامه از آن استفاده شده است. در محیط این نرم‌افزار با استفاده از آرایه‌ها و فرامین موجود امکان مدل‌سازی غیرخطی فراهم می‌گردد[۴۱].

۲-۱۰-۵- بسته نرم‌افزاری R

بسته نرم‌افزاری R، بسته‌ای منبع باز[2] است که در آن توسط یک زبان برنامه‌نویسی که بسیار شبیه S-plus (بسته نرم‌افزاری مشهور آماری) است، انجام محاسبات پیچیده آماری امکان‌پذیر می‌باشد. پروژه R توسط یک تیم بین‌المللی نگهداری شده و داوطلبانه توسعه می‌یابد. این زبان دارای راهنمای داخلی خوبی است و یادگیری آن بسیار ساده است. هم‌چنین توابع آماری پیش‌ساخته فراوانی دارد و بسته‌های زیادی به آن اضافه می‌شود. در این نرم‌افزار، بسته‌ای به نام Random Forests وجود دارد که در این پایان‌نامه از آن استفاده شده است[۴۲].

1-Matrix laboratory
2-Open source

فصل سوم

مدل‌سازی اندیس بازداری

تعدادی از ترکیبات آلی گوگرددار

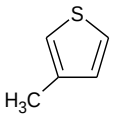
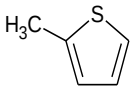
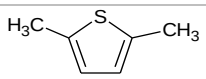
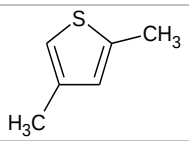
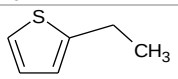
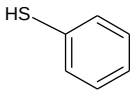
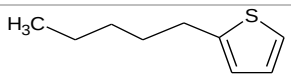
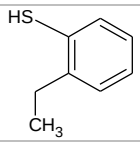
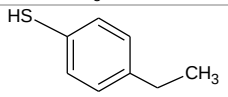
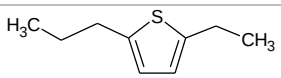
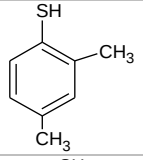
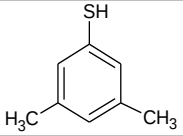
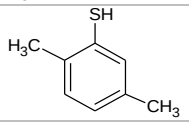
در این فصل، به منظور مطالعه ارتباط کمی ساختار-خاصیت گروهی از ترکیبات گوگرددار، از جنگل-های تصادفی، شبکه عصبی مصنوعی و رگرسیون خطی چندگانه به عنوان روش‌های مدل‌سازی برای پیش‌بینی اندیس بازداری این ترکیبات استفاده شده و توانایی مدل‌های به دست آمده از هر سه روش مورد ارزیابی قرار گرفت. به طور کلی بخش تجربی این تحقیق شامل معرفی سری داده‌ها، بهینه‌سازی ساختار مولکول‌ها و محاسبه توصیف‌گرهای مولکولی، انتخاب توصیف‌گرهای مولکولی مناسب با استفاده از روش‌های رگرسیون مرحله‌ای و راهبرد CDFS و هم‌چنین مدل‌سازی و ارزیابی مدل‌های برتر است.

۳-۱- سری داده‌ها

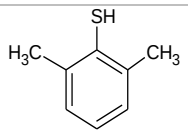
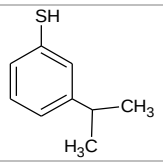
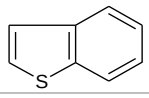
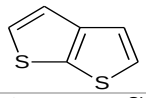
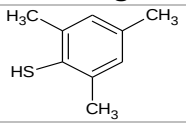
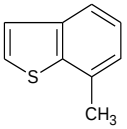
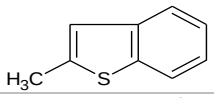
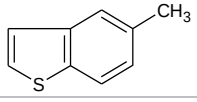
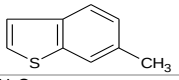
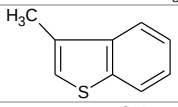
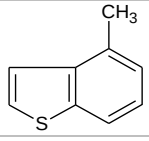
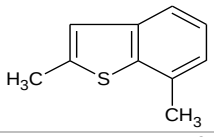
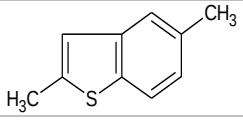
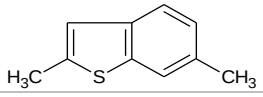
سری داده‌ها از نتایج تعیین اندیس بازداری ۷۲ ترکیب آلی گوگرددار با کروماتوگرافی گازی که فاز ساکن و آشکارساز مورد استفاده، متیل پلی سیلوکسان[1] و آشکارساز نشر اتمی[2] است، به دست آمد[۴]. نام، ساختار و اندیس بازداری (RI) تجربی این ترکیبات در جدول(۳-۱) نشان داده شده است. سری داده‌ها به طور تصادفی به سه دسته‌ی آموزش(۶۰٪)، ارزیابی (۲۰٪) و آزمون (۲۰٪) تقسیم شد که در جدول (۳-۱) دسته‌بندی شده‌اند. داده‌های سری آموزش به منظور ساخت مدل‌های مختلف، داده‌های سری ارزیابی برای بهینه‌سازی پارامترهای مؤثر بر قدرت پیش‌بینی مدل‌ها و داده‌های سری آزمون برای ارزیابی و مقایسه‌ی توانایی مدل‌های مختلف در پیش‌بینی اندیس بازداری ترکیبات به کار برده شدند.

1-Methyl polysiloxane
2-Atomic emission detector

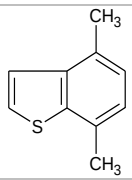
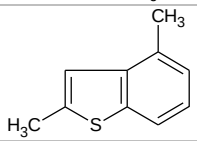
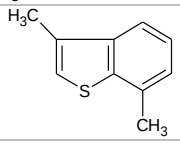
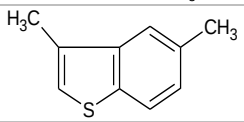
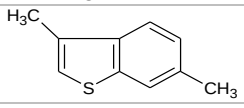
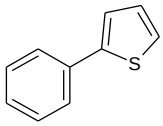
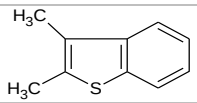
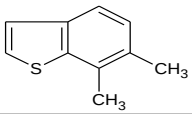
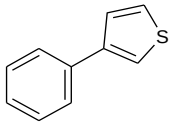
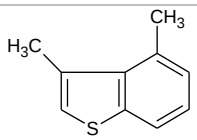
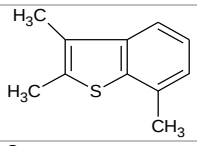
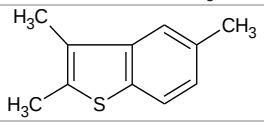
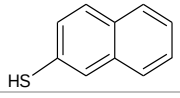
جدول (۳-۱)- نام و اندیس بازداری سری داده‌ها

شماره	نام	ساختار	RI	سری*
۱	Thiophene		۱۰۰/۰۰	tr
۲	3-Methylthiophene		۱۱۰/۶۶	tr
۳	2-Methylthiophene		۱۱۱/۱۸	v
۴	2,5-Dimethylthiophene		۱۲۲/۳۸	tr
۵	2,4-Dimethylthiophene		۱۲۲/۵۱	v
۶	2-Ethylthiophene		۱۲۳/۲۹	tr
۷	Benzene thiol		۱۵۰/۴۱	tr
۸	2-Pentylthiophene		۱۷۵/۳۲	tr
۹	2-Ethyl benzenethiol		۱۸۶/۸۳	Tr
۱۰	4-Ethyl thiophenol		۱۷۸/۵۱	Tr
۱۱	2-n-Butyl-5-ethyl thiophene		۱۸۹/۴۰	Tr
۱۲	2,4-Dimethyl benzenethiol		۱۸۹/۷۷	Tr
۱۳	3,5-Dimethyl benzenethiol		۱۸۹/۹۶	V
۱۴	2,5-Dimethyl benzenethiol		۱۹۰/۲۸	Te

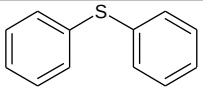
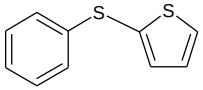
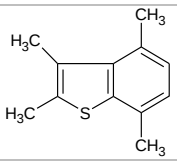
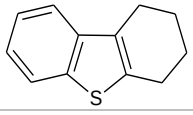
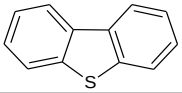
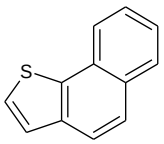
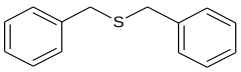
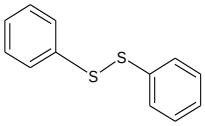
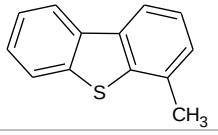
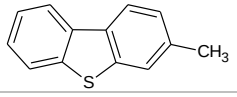
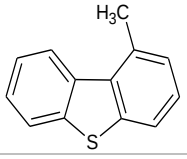
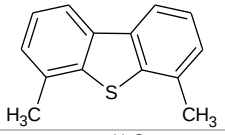
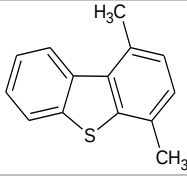
ادامه جدول (۳-۱)

شماره	نام	ساختار	RI	سری*
۱۵	2,6-Dimethyl benzenethiol		۱۹۱/۸۳	tr
۱۶	2-Isopropyl benzenethiol		۱۹۵/۷۶	tr
۱۷	Benzothiophene		۲۰۰/۰۰	v
۱۸	Thieno[2,3-b]thiophene		۲۰۴/۳۱	tr
۱۹	2,4,6-Trimethyl thiophenol		۲۰۹/۳۵	tr
۲۰	7-Methyl-benzo[b]thiophene		۲۱۶/۱۱	v
۲۱	2-Methyl-benzo[b]thiophene		۲۱۶/۸۳	tr
۲۲	5-Methyl-benzo[b]thiophene		۲۱۸/۲۵	te
۲۲	6-Methyl-benzo[b]thiophene		۲۱۹/۲۸	te
۲۴	3-Methyl-benzo[b]thiophene		۲۲۱/۲۲	te
۲۵	4-Methyl-benzo[b]thiophene		۲۲۱/۵۰	tr
۲۶	2,7-Dimethyl-benzo[b]thiophene		۲۳۰/۵۸	v
۲۷	2,5-Dimethyl-benzo[b]thiophene		۲۳۴/۱۵	tr
۲۸	2,6-Dimethyl-benzo[b]thiophene		۲۳۴/۹۴	tr

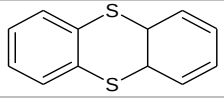
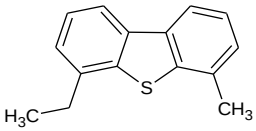
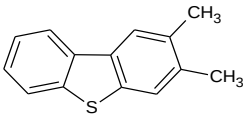
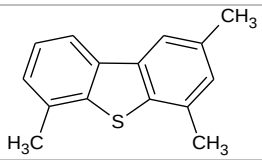
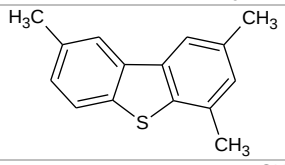
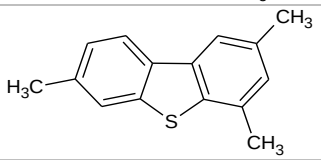
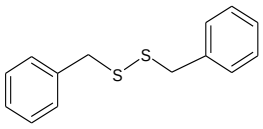
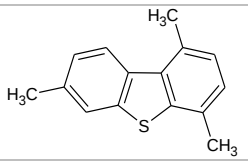
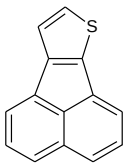
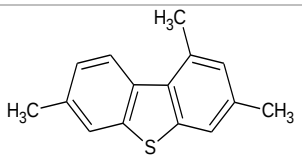
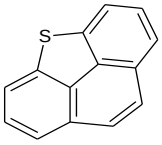
ادامه جدول (۳-۱)

شماره	نام	ساختار	RI	سری*
۲۹	4,7-Dimethyl-benzo[b]thiophene		۲۳۵/۳۹	Te
۳۰	2,4-Dimethyl-benzo[b]thiophene		۲۳۶/۰۷	tr
۳۱	3,7-Dimethyl-benzo[b]thiophene		۲۳۶/۱۵	te
۳۲	3,5-Dimethyl-benzo[b]thiophene		۲۳۸/۰۹	tr
۳۳	3,6-Dimethyl-benzo[b]thiophene		۲۳۸/۳۰	v
۳۴	2-Phenyl thiophene		۲۳۸/۵۶	tr
۳۵	2,3-Dimethyl-benzo[b]thiophene		۲۳۹/۵۴	te
۳۶	6,7-Dimethyl-benzo[b]thiophene		۲۴۰/۶۶	v
۳۷	3-Phenyl thiophene		۲۴۲/۳۷	tr
۳۸	3,4-Dimethyl-benzo[b]thiophene		۲۵۰/۱۴	tr
۳۹	2,3,7-Trimethyl-benzo[b]thiophene		۲۵۳/۹۱	tr
۴۰	2,3,5-Trimethyl-benzo[b]thiophene		۲۵۶/۱۷	v
۴۱	2-Naphthalenethiol		۲۶۳/۱۹	tr

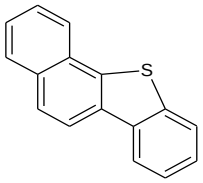
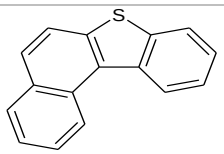
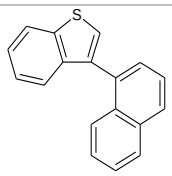
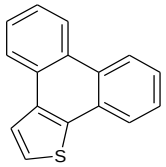
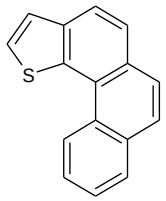
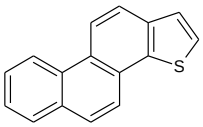
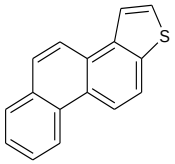
ادامه جدول (۳-۱)

شماره	نام	ساختار	RI	سری*
۴۲	Phenyl sulfide		۲۷۰/۷۵	tr
۴۳	2-Phenylthio thiophene		۲۷۱/۸۳	v
۴۴	2,3,4,7-Tetramethyl-benzo[b]thiophene		۲۸۰/۱۰	tr
۴۵	1,2,3,4-Tetrahydro dibenzothiophene		۲۹۳/۹۳	tr
۴۶	Dibenzothiophene		۳۰۰/۰۰	v
۴۷	Naphtho[1,2-b]thiophene		۳۰۲/۱۳	tr
۴۸	Benzyl sulfide		۳۰۷/۷۷	tr
۴۹	Phenyl disulfide		۳۱۳/۰۵	tr
۵۰	4-Methyl-dibenzothiophene		۳۱۳/۷۶	tr
۵۱	3-Methyl-dibenzothiophene		۳۱۷/۸۷	te
۵۲	1-Methyl-dibenzothiophene		۳۲۳/۳۸	te
۵۳	4,6-Dimethyl-dibenzothiophene		۳۲۷/۶۴	tr
۵۴	1,4-Dimethyl-dibenzothiophene		۳۳۷/۲۰	v

ادامه جدول (۳-۱)

شماره	نام	ساختار	RI	سری*
۵۵	Thianthrene		۳۳۹/۵۱	tr
۵۶	4-Ethyl-6-methyl-dibenzothiophene		۳۴۰/۴۳	te
۵۷	2,3-Dimethyl-dibenzothiophene		۳۴۲/۰۸	tr
۵۸	2,4,6-Trimethyl-dibenzothiophene		۳۴۳/۲۱	te
۵۹	2,4,8-Trimethyl-dibenzothiophene		۳۴۵/۰۵	te
۶۰	2,4,7-Trimethyl-dibenzothiophene		۳۴۶/۴۴	te
۶۱	Benzyl disulfide		۳۴۸/۹۴	tr
۶۲	1,4,7-Trimethyl-dibenzothiophene		۳۵۲/۹۰	tr
۶۳	Acenaphtho[1,2-b]thiophene		۳۵۳/۰۲	te
۶۴	1,3,7-Trimethyl-dibenzothiophene		۳۵۵/۸۶	tr
۶۵	Phenanthro[4,5-bcd]thiophene		۳۶۰/۹۳	tr

ادامه جدول (۳-۱)

شماره	نام	ساختار	RI	سری*
۶۶	Benzo(b)naphtho[2,1-d]thiophene		۴۰۰/۰۰	tr
۶۷	Benzo(b)naphtha(1,2-d)thiophene		۴۰۳/۶۹	v
۶۸	3-[1-Naphthalenyl]benzo[b] thiophene		۴۰۵/۸۹	tr
۶۹	Phenanthro(9,10-b) thiophene		۴۱۵/۲۸	tr
۷۰	Phenanthro(4,3-b) thiophene		۴۱۶/۴۷	v
۷۱	Phenanthro[1,2-b] thiophene		۴۱۷/۶۱	tr
۷۲	Phenanthro(2,1-b) thiophene		۴۲۶/۵۷	tr

*داده‌های سری آموزش : tr ,داده‌های سری ارزیابی : v ,داده‌های سری آزمون : te

۳-۲-بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌گرها

ابتدا ساختار مولکول‌ها با استفاده از نرم‌افزار Hyperchem رسم و با روش نیمه‌تجربی AM1 برای رسیدن جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول بهینه شد. سپس به وسیله‌ی نرم‌افزار Dragon، ۱۴۸۱ توصیف‌گر با استفاده از این ساختارهای بهینه شده محاسبه شد.

۳-۳- انتخاب بهترین توصیف‌گرها

با توجه به اینکه تعداد زیاد توصیف‌گر باعث پیچیدگی محاسبات شده و تعدادی از توصیف‌گرها حاوی اطلاعات یکسانی هستند، لذا باید تعداد توصیف‌گرها به نوعی کاهش یابد. به همین منظور، ابتدا توصیف-گرهایی که برای تمام مولکول‌ها مقادیر ثابت یا تقریبا ثابت داشتند، حذف شدند. سپس یکی از هر دو توصیف‌گری که هم‌بستگی بزرگتر از ۰/۹ داشتند نیز با استفاده از برنامه‌ی Remcol در نرم‌افزار MATLAB ، حذف گردیدند. در پایان این مرحله ۲۱۵ توصیف‌گر باقی ماند.

به دلیل این که روش جنگل‌های تصادفی قابلیت انجام مدل‌سازی با تعداد زیاد توصیف‌گر را دارا است، از کل ۲۱۵ توصیف‌گر به عنوان ورودی مدل استفاده شد (بخش ۳-۸). اما از آنجا که شبکه‌های عصبی با تعداد زیاد توصیف‌گر ورودی، نتیجه‌ی بیش برازشی به بار می‌آورند، لذا باید ابتدا توصیف‌گرهای مهم را از بین ۲۱۵ توصیف‌گر انتخاب و سپس مدل‌سازی شبکه‌های عصبی را انجام داد. در اینجا از روش‌های رگرسیون مرحله‌ای و راهبرد CDFS برای انتخاب بهترین توصیف‌گرها استفاده شده است که در زیر به توضیح آن می‌پردازیم. شایان ذکر است که روش جنگل‌های تصادفی به دلیل قابلیت خود در تعیین اهمیت توصیف‌گرها می‌تواند به عنوان یک روش انتخاب متغیر نیز استفاده شود.

۳-۳-۱- انتخاب بهترین توصیفگرها با روش رگرسیون مرحله‌ای

با به کارگیری نرم‌افزار SPSS و با استفاده از داده‌های سری آموزش و اجرای رگرسیون مرحله‌ای برای ۲۱۵ توصیف‌گر یاد شده به عنوان متغیرهای مستقل، و اندیس بازداری به عنوان متغیر وابسته، تعداد ۱۰ توصیف‌گر انتخاب شدند که نام و طبقه‌ی آنها در جدول (۳-۲) نشان داده شده است. هم‌چنین جدول (۳-۳) ماتریس هم‌بستگی بین این توصیف‌گرها را ارائه می‌دهد که نتایج این جدول نشان می‌دهد که بین این توصیف‌گرها هم‌بستگی معناداری وجود ندارد.

جدول (۳-۲)- توصیف‌گرهای انتخاب شده با روش رگرسیون مرحله‌ای

No	Symbol	Class	Meaning
1	SCBO	Constitutional	Sum of conventional band orders(H-depleted)
2	PSA	Properties	Fragment-based polar surface area
3	L3s	WHIM	3rd component size directional WHIM index/weighted by atomic electrotopological states
4	RDF020v	RDF	Radial distribution function-2.0/weighted by atomic van der waals volumes
5	RDF080p	RDF	Radial distribution function-8.0/weighted by atomic polarizabilities
6	ATS5p	2D autocorrelations	Broto-Moreau autocorrelation of a topological structure-lag5/weighted by atomic polarizabilities
7	Mor03p	3-D-MORSE	3-D-MORSE signal 03/Weighted by atomic polarizability
8	Mor24v	3-D-MORSE	3-D-MORSE signal 24/Weighted by atomic van der waals volumes
9	PJI3	Geometrical	3D Petijean shape index
10	RDF090u	RDF	Radial distribution function-9.0/unweighted

جدول (۳-۳)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای

	SCBO	PSA	L3s	RDF020v	RDF080p	ATS5p	Mor03p	Mor24v	PJI3	RDF090u
SCBO	۱									
PSA	-۰/۱۷۲	۱								
L3s	۰/۱۶۲	۰/۱۱۷	۱							
RDF020v	۰/۴۷۱	-۰/۰۸۵	۰/۴۰۴	۱						
RDF080p	۰/۷۰۹	-۰/۳۰۸	۰/۲۶۴	۰/۵۲۳	۱					
ATS5p	۰/۵۴۲	۰/۴۷۰	۰/۱۳۹	۰/۴۴۷	۰/۵۲۳	۱				
Mor03p	۰/۵۸۷	۰/۱۰۲	-۰/۲۱۱	-۰/۴۲۴	۰/۴۴۷	۰/۴۱۸	۱			
Mor24v	۰/۴۳۶	-۰/۳۳۷	-۰/۱۶۴	۰/۱۴۶	-۰/۴۲۴	۰/۳۸۷	-۰/۴۰۵	۱		
PJI3	۰/۱۷۳	-۰/۱۵۰	۰/۱۴۹	۰/۲۴۱	۰/۱۴۶	۰/۲۶۳	-۰/۲۰۴	۰/۲۴۸	۱	
RDF090u	۰/۵۲۱	۰/۳۸۰	۰/۳۲۷	۰/۵۸۲	۰/۲۴۱	۰/۴۶۷	-۰/۵۶۴	۰/۲۲۱	۰/۲۷۰	۱

۳-۳-۲- انتخاب بهترین توصیف‌گرها با راهبرد CDFS

برای انجام این روش ابتدا داده‌های سری آزمون (۲۰٪) کنار گذاشته شد و مجموعه‌ی متشکل از داده‌های سری آموزش و ارزیابی به تعداد ۱۶ بار به طور تصادفی به دو مجموعه آموزش و ارزیابی (۲۰٪ ارزیابی و ۶۰٪ آموزش) تقسیم گردیدند. رگرسیون مرحله‌ای برای هر دسته از سری‌های آموزش انجام شده و مدل‌های به دست آمده توسط داده‌های سری ارزیابی همان دسته مورد بررسی قرار گرفت تا بهترین مدل آن دسته‌بندی به دست آید. بنابراین در انتها ۱۶ مدل متفاوت برای سری داده‌ها ایجاد شد. توصیف‌گرهای انتخاب شده در این مدل‌ها در جدول (۳-۴) ذکر شده‌اند.

جدول (۳-۴)- توصیف‌گرهای ۱۶ مدل به دست آمده از تقسیم‌بندی مختلف داده‌ها

شماره مدل	توصیف‌گرها	تعداد توصیف‌گرها
۱	SCBO, Mor06m	۲
۲	SCBO, PSA, Mor06m	۳
۳	SCBO,PSA,E3s	۳
۴	SCBO, Mor06m, Mor32m	۳
۵	SCBO, Mor06m, Mor24v, E3s, PSA, RDF080p, Mor03p, Mor21m, HATS1m	۹
۶	SCBO, PSA, Mor06m	۳
۷	SCBO, PSA	۲
۸	SCBO, Mor06m, E3s	۳
۹	SCBO, RDF040e, E3s, Mor27p, E2s, Mor21m, H047, C028	۸
۱۰	SCBO, E3s, Mor06m, PSA	۴
۱۱	SCBO, PSA	۲
۱۲	SCBO, Mor06m, E3s	۳
۱۳	SCBO, Mor21m, Mor06m, Mor32m, E3s, L3s, Ds,	۷
۱۴	SCBO, PSA, E3s	۳
۱۵	SCBO, Mor06m	۲
۱۶	SCBO, PSA, L3s, RDF020v, Mor06m	۵

همانطور که مشاهده می‌شود در این مدل‌ها، ۴ توصیف‌گر معرفی شده در جدول (۳-۵)، بیش از ۵۰٪ تکرار شده‌اند که نشان‌دهنده‌ی اهمیت این توصیف‌گرها است. جدول (۳-۶) ماتریس هم‌بستگی بین این توصیف‌گرها را نشان می‌دهد که مشاهده می‌شود بین این توصیف‌گرها هم‌بستگی معناداری وجود ندارد.

جدول (۳-۵)- توصیف‌گرهای انتخاب شده با راهبرد CDFS

No	Symbol	Class	Meaning
1	SCBO	Constitutional	Sum of conventional band orders(H-depleted)
2	Mor06m	3-D-MORSE	3-D-MORSE signal 06/Weighted by atomic masses
3	PSA	Properties	Fragment-based polar surface area
4	E3s	WHIM	3rd component accessibility directional WHIM index/weighted by atomic electrotopological states

جدول (۳-۶)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط راهبرد CDFS

	SCBO	Mor06m	PSA	E3s
SCBO	۱			
Mor06m	۰/۶۹۰	۱		
PSA	-۰/۱۷۲	۰/۰۵۸	۱	
E3s	۰/۱۹۴	۰/۲۶۷	۰/۳۲۲	۱

۳-۳-۳- انتخاب بهترین توصیف‌گرها با روش جنگل‌های تصادفی

همان‌طور که در بخش ۲-۸-۲-۲ توضیح داده شد، روش جنگل‌های تصادفی علاوه بر این که به عنوان یک روش مدل‌سازی به کار می‌رود، می‌تواند به عنوان یک روش انتخاب متغیر نیز به کار برده شود. پس از ساخت مدل جنگل‌های تصادفی و بهینه کردن پارامترهای مؤثر بر توان پیش‌بینی مدل، الگوریتم جنگل‌های تصادفی با استفاده از نرم‌افزار R صد بار تکرار شده و در هر بار، اهمیت توصیف‌گرها محاسبه گردید. سرانجام میانگین مقادیر اهمیت هر توصیف‌گر به عنوان شاخص نهایی در نظر گرفته شده و هم‌چنین اهمیت نسبی هر توصیف‌گر نسبت به توصیف‌گری که بیشترین اهمیت را دارا است، تعیین شد. شکل (۳-۱) میزان اهمیت نسبی توصیف‌گرها را نمایش می‌دهد و همان‌طور که مشاهده می‌شود، بسیاری از توصیف‌گرها دارای اهمیت نسبی کمی بوده و تنها ۱۱ توصیف‌گر اهمیت نسبی قابل توجهی دارند که این ۱۱ توصیف‌گر همراه با مقدار اهمیت و نوع طبقه‌ی آن‌ها در جدول (۳-۷) نشان داده شده‌اند. هم‌چنین ماتریس هم‌بستگی بین این توصیف‌گرها در جدول (۳-۸) آمده است که نشان-دهنده‌ی عدم هم‌بستگی معنادار بین آنها می‌باشد.

شکل (۳-۱)- نمودار اهمیت نسبی توصیف‌گرها

جدول (۳-۷)- توصیف‌گرهای انتخاب شده با روش جنگل های‌تصادفی

Rank	Symbol	Class	Meaning	Importance	Relative Importance
1	SCBO	Constitutional	Sum of conventional band orders(H-depleted)	17.8838	1.00
2	BELe6	BCUT	Lowest eigenvalue n.6 of Burden matrix/weighted by atomic sanderson electronegativities	11.0211	0.61
3	Mor03p	3D-MoRSE	3D-MoRSE-signal 03/weighted by atomic polarizabilities	8.0560	0.45
4	MWC08	Molecular walk count	Molecular walk count of order 08	7.1639	0.40
5	Yindex	Topological	Balaban Y index	6.8560	0.38
6	BELe5	BCUT	Lowest eigenvalue n.5 of Burden matrix/weighted by atomic sanderson electronegativities	6.6732	0.37
7	IC5	Topological	Information content index (neighborhood symmetry of 5-order)	6.4170	0.35
8	Mor06m	3D-MoRSE	3D-MoRSE-signal 06/weighted by atomic masses	6.4163	0.35
9	DELS	Geometrical	Molecular electrotopological variation	5.9972	0.33
10	RDF070p	RDF	Radial distribution function-8.0/weighted by atomic polarizabilities	5.7654	0.32
11	H3m	GETAWAY	H autocorrelation of lag 3/weighted by atomic masses	5.6107	0.31

جدول (۳-۸)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی

	SCBO	BELe6	Mor03p	MWC08	Yindex	BELe5	IC5	Mor06m	DELS	RDF070p
SCBO	۱									
BELe6	۰/۴۱۳	۱								
Mor03p	-۰/۳۹۸	-۰/۲۸۷	۱							
MWC08	۰/۵۳۲	۰/۳۸۰	-۰/۲۹۷	۱						
Yindex	۰/۴۸۰	-۰/۳۷۳	۰/۴۱۲	-۰/۳۰۸	۱					
BELe5	۰/۴۹۵	۰/۴۰۲	-۰/۳۶۵	۰/۳۹۳	-۰/۳۸۵	۱				
IC5	۰/۴۷۵	۰/۲۶۶	-۰/۳۱۷	۰/۴۰۸	-۰/۳۹۰	۰/۳۶۷	۱			
Mor06m	۰/۲۵۰	۰/۲۳۹	-۰/۲۲۰	۰/۳۹۷	-۰/۲۴۷	۰/۲۷۴	۰/۲۴۲	۱		
DELS	۰/۲۶۶	۰/۲۶۴	-۰/۱۱۱	۰/۲۶۳	-۰/۱۴۳	۰/۲۴۹	۰/۱۷۵	۰/۱۰۸	۱	
RDF070p	۰/۴۰۳	۰/۳۸۵	-۰/۲۸۳	۰/۲۸۵	-۰/۴۰۵	۰/۳۹۳	۰/۳۰۹	۰/۱۵۴	۰/۱۵۱	۱

۳-۴- مدل‌سازی با رگرسیون خطی چندگانه

برای ساخت مدلی که بیان‌گر ارتباط خطی بین ساختار ترکیبات و اندیس بازداری آنها است، از روش رگرسیون خطی چندگانه (MLR) استفاده شد. با استفاده از داده‌های سری آموزش و توصیف-گرهای ارائه شده در جدول (۳-۲)، مدل‌های مختلف با تعداد توصیف‌گر متفاوت به دست آمد و سپس بهترین مدل که دارای بیشترین مقدار R^2_{adj} و FIT برای داده‌های سری ارزیابی است، به عنوان مدل نهایی انتخاب شد. مقادیر این پارامترها برای مدل‌های مختلف در جدول (۳-۹) ذکر شده است و روابط مربوط به آنها در بخش ۲-۹ ارائه گردیده است.

جدول (۳-۹)- پارامترهای آماری برای مدل‌های به دست آمده از روش MLR

مدل	توصیف‌گرها	FIT	R^2_{adj}
۱	SCBO, PSA	۵۷/۶۰	۰/۹۹۲
۲	SCBO, PSA, L3s	۶۰/۱۲	۰/۹۹۱
۳	SCBO, PSA, L3s, RDF020v	۶۰/۱۶	۰/۹۹۳
۴	**SCBO, PSA, L3s, RDF020v, RDF080p**	**۶۵/۰۷**	**۰/۹۹۴**
۵	SCBO, PSA,L3s, RDF020v, RDF080p, ATS5p	۱/۲۶	۰/۸۱۴
۶	SCBO, PSA,L3s, RDF020v, RDF080p, ATS5p, Mor03p	۲/۲۹	۰/۹۱۳
۷	SCBO, PSA, L3s, RDF020v, RDF080p, ATS5p, Mor03p, Mor24v	۲/۰۷	۰/۹۲۲
۸	SCBO, PSA, L3s, RDF020v, RDF080p, ATS5p, Mor03p, Mor24v, PJI3	۰/۲۸	۰/۵۷۷
۹	SCBO, PSA, L3s, RDF020v, RDF080p, ATS5p, Mor03p, Mor24v, PJI3, RDF09u	۰/۱۵	۰/۳۵۰

بدین ترتیب مدل ۴ به عنوان بهترین مدل برگزیده شد که معادله آن به صورت زیر است:

RI=-43.577+14.678 SCBO+1.307 PSA-52.059 L3s+44.193 RDF020v-6.473 RDF080p (۱-۳)

ولی با توجه به اینکه واحدهای اندازه‌گیری متغیرهای مستقل (ضرایب رگرسیون) یکسان نیستند، هرگز نمی‌توان از روی ضرایب رگرسیون غیراستاندارد به میزان اهمیت و تاثیر یک متغیر مستقل بر روی متغیر وابسته پی برد. برای حل این مشکل می‌توان از ضرایب استاندارد شده که طبق رابطه زیر محاسبه می‌شود استفاده نمود[۲۶ و ۴۳]:

$$\acute{\beta}_K = \left(\frac{S_k}{S_y}\right) \beta_k \qquad (۲-۳)$$

در این رابطه $\acute{\beta}_K$ ضریب گرسیون استاندارد شده توصیف‌گر kام، β_K ضریب رگرسیون غیراستاندارد همان توصیف‌گر، S_k انحراف استاندارد متغیر مستقل مورد نظر(توصیف‌گر kام) و S_y نیز انحراف استاندارد متغیر وابسته (اندیس بازداری) می‌باشد که از روابط زیر به دست می‌آیند:

$$S_k = \sqrt{\frac{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}{n-1}} \qquad (۳-۳)$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \qquad (۴-۳)$$

که x_{ik} مقدار توصیف‌گر kام برای ترکیب iام و $\bar{x}_k$ مقدار میانگین توصیف‌گر kام ، n تعداد ترکیبات، y_i مقدار متغیر وابسته برای ترکیب iام و $\bar{y}$ مقدار میانگین متغیر وابسته است.

بدین ترتیب در این تحقیق نیز بر اساس روابط فوق، ضرایب رگرسیون استاندارد شده برای مدل برتر محاسبه شد که در جدول زیر ارائه شده است:

جدول (۳-۱۰)- ضرایب غیراستاندارد و استاندارد توصیفگرهای موجود در مدل خطی برتر

توصیف‌گر	SCBO	PSA	L3s	RDF020v	RDF080p
ضریب غیر استاندارد	۱۴/۶۷۸	۱/۳۰۷	-۵۲/۰۵۹	۴۴/۱۹۳	-۶/۴۷۳
ضریب استاندارد	۰/۹۹۳	۰/۰۸۴	-۰/۰۰۰۱	۰/۰۵۷	-۰/۰۶۲

۳-۵- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای (SR-ANN)

یکی از راه‌های یافتن رابطه‌ی غیرخطی بین متغیرهای مستقل و متغیر وابسته، استفاده از شبکه عصبی مصنوعی برای مدل‌سازی می‌باشد. ابتدا مقادیر ۱۰ توصیف‌گر به عنوان توصیف‌گرهای ورودی و اندیس بازداری متناظر با آنها به عنوان متغیر هدف در نظر گرفته شده و سپس یک شبکه پیشخور با تکنیک پس‌انتشار ایجاد گردید. شایان ذکر است که پارامترهای مؤثر بر قدرت پیش‌بینی شبکه که شامل تعداد توصیف‌گرهای ورودی، تعداد گره در لایه پنهان، نوع تابع آموزش و انتقال یا محرک، تعداد دور

آموزش (epoch) و پارامتر ممنتوم (μ) برای دستیابی به بهترین نتیجه، باید بهینه شوند. تابع کارایی شبکه نیز میانگین مربع خطا (MSE) می‌باشد.

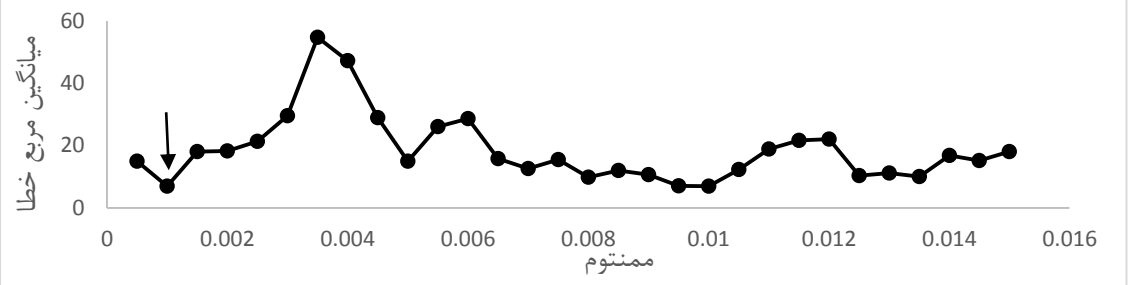
۳-۵-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

از آنجایی که در بیشتر موارد به نظر می‌رسد که یک لایه پنهان در ساختار شبکه عصبی مناسب باشد، در این تحقیق نیز یک لایه پنهان مورد استفاده قرار گرفت[۴۴]. هم‌چنین در لایه خروجی از تابع انتقال خطی (purelin) استفاده شد، اما برای یافتن مقدار بهینه سایر پارامترها، مقادیر مختلفی از آنها در ساختار شبکه قرار داده شد. برای این منظور، هر شبکه با تعداد ورودی از ۲ تا ۱۰ و با دو الگوریتم آموزشی لونبرگ-مارکوارت (trainlm)[1]و تنظیم بایزین (trainbr)[2]و دو تابع انتقال لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) و تعداد گره از ۲ تا ۱۰ و تعداد دور آموزش از ۲ تا ۴۰، آموزش داده شد. در اینجا تنها مقدار پارامتر μ، ثابت در نظر گرفته شد که در مرحله بعد بهینه گردید. در جدول (۳-۱۱) بهترین شبکه‌های به دست آمده در حین بهینه‌سازی پارامترهای مختلف بر اساس کمترین مقدار MSE نشان داده شده‌اند. با توجه به این جدول، شبکه عصبی ۳ لایه با تابع آموزشی لونبرگ-مارکوارت و تابع انتقال تانژانت سیگموئید، ۵ توصیف‌گر ورودی، ۶ نرون در لایه مخفی و تعداد دور آموزشی۱۲ کمترین MSE را نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی در نظر گرفته شد. برای یافتن مقدار بهینه پارامتر μ ، در شرایط بهینه به دست آمده از مرحله‌ی قبل، مقدار این پارامتر بین ۰/۰۰۰۵ تا ۰/۰۱۵ با گام‌های ۰/۰۰۰۵ تغییر داده شد و برای هر مورد مقدار MSE سری ارزیابی محاسبه گردید. شکل (۳-۲). کمترین مقدار MSE مربوط به μ= ۰/۰۰۱ می‌باشد که با مقدار پیش‌فرض آن در تابع آموزشی لونبرگ-مارکوارت برابر است.

1-Levenberg-marquart
2-Bayesian regularization

جدول (۳-۱۱)- توابع و پارامترهای شبکه‌های بهینه (SR-ANN) به دست آمده

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۸/۱۲	۲۲	۶	لگاریتم-سیگموئید	لونبرگ-مارکوارت	۵
۶/۹۸	**۱۲**	**۶**	**تانژانت-سیگموئید**	**لونبرگ-مارکوارت**	**۵**
۱۰/۵۶	۳۳	۹	تانژانت-سیگموئید	تنظیم بایزین	۷
۸/۸۵	۴۴	۴	لگاریتم-سیگموئید	تنظیم بایزین	۹



شکل (۳-۲)- نمودار تغییر خطای شبکه (SR-ANN) با تغییر پارامتر μ

۳-۵-۲- ساختار شبکه عصبی مصنوعی بهینه شده

در جدول (۳-۱۲)، مشخصات شبکه بهینه‌ی به دست آمده نشان داده شده است.

جدول (۳-۱۲)- مشخصات شبکه بهینه (SR-ANN)

تابع آموزش	لونبرگ-مارکوارت (Trainlm)
تابع انتقال لایه پنهان	تانژانت-سیگموئید (Tansig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۶
تعداد توصیف‌گر ورودی	۵
تعداد دور آموزشی	۱۲
پارامتر μ	۰/۰۰۱

۳-۶- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط راهبرد CDFS (CDFS-ANN)

در این بخش توصیف‌گرهای انتخاب شده با راهبرد CDFS به عنوان توصیف‌گرهای ورودی به یک شبکه عصبی با یک لایه پنهان و لایه خروجی purlin مورد استفاده قرار گرفت. برای یافتن مقادیر بهینه پارامترهای مؤثر بر کارایی شبکه، تعداد ورودی از ۲ تا ۴ و سایر پارامترها مانند مرحله قبل تغییر داده شد. میانگین مربع خطا نیز به عنوان معیار انتخاب شبکه بهینه در نظر گرفته شد.

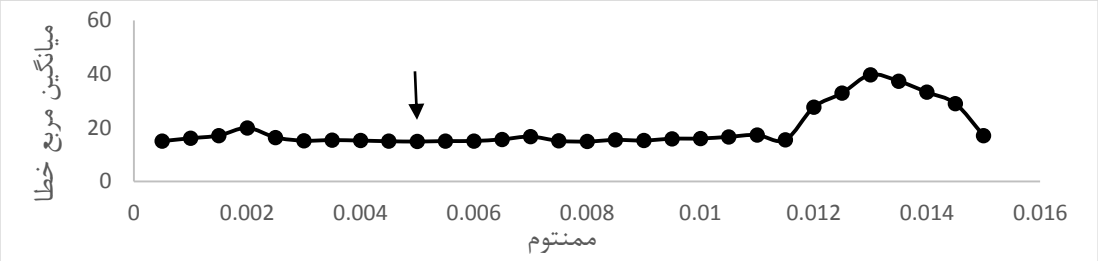
۳-۶-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

در این مرحله نیز جهت ایجاد شبکه بهینه تمامی پارامترها به طور هم‌زمان بهینه گردید و تنها مقدار ممنتوم ثابت در نظر گرفته شد که در مرحله بعد مقدار این پارامتر نیز بهینه گردید. با توجه به جدول (۳-۱۳) که بهترین شبکه‌های به دست آمده در حین بهینه‌سازی را نمایش می‌دهد، مشاهده می‌شود که شبکه عصبی با ۳ توصیف‌گر ورودی، ۹ نرون در لایه مخفی ، تعداد دور آموزشی ۱۲، تابع آموزشی تنظیم بایزین و تابع انتقال لگاریتم سیگموئید، کمترین MSE را نسبت به سایر شبکه‌ها نشان می دهد. بنابراین شبکه‌ای با این ساختار به عنوان شبکه‌ی بهینه برای مدل‌سازی در نظر گرفته شد.

جدول (۳-۱۳)- توابع و پارامترهای شبکه‌های بهینه (CDFS-ANN) به دست آمده

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۱۶/۱۰	۸	۷	لگاریتم-سیگموئید	لونبرگ-مارکوارت	۳
۱۴/۹۳	۲۳	۷	تانژانت-سیگموئید	لونبرگ-مارکوارت	۴
۱۸/۳۲	۷	۱۰	تانژانت-سیگموئید	تنظیم بایزین	۴
۱۴/۸۶	**۱۲**	**۹**	**لگاریتم-سیگموئید**	**تنظیم بایزین**	**۳**

برای بهینه‌سازی پارامتر μ، به ازای تنظیمات مدل انتخاب شده در مرحله‌ی قبل، مقدار این پارامتر از ۰/۰۰۰۵ تا ۰/۰۱۵ با گام‌های ۰/۰۰۰۵ تغییر داده شده و در هر مرحله تغییر، میانگین مربع خطای سری ارزیابی محاسبه گردید. طبق شکل (۳-۳)، مقدار بهینه μ برابر با مقدار پیش‌فرض آن در تابع آموزشی تنظیم بایزین یعنی ۰/۰۰۵ به دست آمد.



شکل (۳-۳)- نمودار تغییر خطای شبکه (CDFS-ANN) با تغییر پارامتر μ

۳-۶-۲- ساختار شبکه عصبی مصنوعی بهینه شده

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

جدول (۳-۱۴)- مشخصات شبکه بهینه (CDFS-ANN)

تابع آموزش	تنظیم بایزین (trainbr)
تابع انتقال لایه پنهان	تانژانت سیگموئید (tansig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۹
تعداد توصیف‌گر ورودی	۳
تعداد دور آموزشی	۱۲
پارامتر μ	۰/۰۰۵

۳-۷- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی (RF-ANN)

در این بخش نیز از شبکه عصبی با یک لایه پنهان و لایه خروجی purlin وتوصیف‌گرهای منتخب روش جنگل‌های تصادفی استفاده گردید. برای یافتن مقادیر بهینه پارامترهای مؤثر بر کارایی شبکه،

تعداد ورودی ها از ۲ تا ۱۱ و سایر پارامترها مانند مرحله قبل تغییر داده شده و تابع کارایی نیز میانگین مربع خطا در نظر گرفته شد.

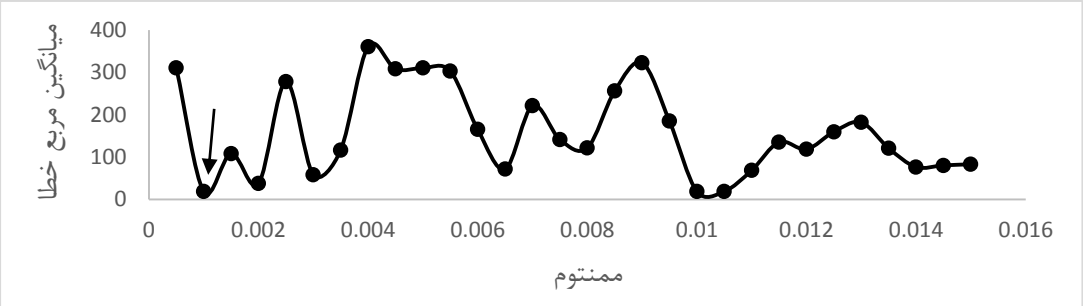
۳-۷-۱- بهینه‌سازی پارامترهای مؤثر

در این مرحله، تنظیمات مختلفی برای همه پارامترها به جز ممنتوم در نظر گرفته شد و با به دست آمدن تنظیمات مدل بهینه، بهترین مقدار μ نیز محاسبه گردید. جدول (۳-۱۵) بهترین شبکه‌های به دست آمده در حین بهینه‌سازی را نمایش می‌دهد که در آن شبکه عصبی با ۹ توصیف‌گر ورودی، ۳ نرون در لایه مخفی ، تعداد دور آموزشی ۳۷، تابع آموزشی لونبرگ-مارکوارت و تابع انتقال لگاریتم سیگموئید، کمترین MSE را نسبت به سایر شبکه‌ها دارا است و لذا شبکه‌ای با این ساختار می‌تواند به عنوان شبکه بهینه برای مدل‌سازی اندیس بازداری ترکیبات در نظر گرفته شود.

جدول (۳-۱۵)- توابع و پارامترهای شبکه‌های بهینه (RF-ANN) به دست آمده

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۱۸/۷۹	**۳۷**	**۳**	**لگاریتم-سیگموئید**	**لونبرگ-مارکوارت**	**۹**
۲۲/۹۷	۴۰	۵	تانژانت-سیگموئید	لونبرگ-مارکوارت	۲
۲۰/۴۵	۳۱	۱۰	تانژانت-سیگموئید	تنظیم بایزین	۹
۱۸/۸۳	۳۲	۷	لگاریتم-سیگموئید	تنظیم بایزین	۹

برای بهینه‌سازی مقدار پارامتر μ، شبکه‌ای با شرایط بهینه به دست آمده از مرحله‌ی قبل، طراحی شد. سپس مقدار این پارامتر از ۰/۰۰۰۵ تا ۰/۰۱۵ با گام‌های ۰/۰۰۰۵ تغییر داده شده و در هر تغییر میانگین مربع خطای سری ارزیابی محاسبه گردید. طبق شکل (۳-۴) مقداربهینه μ برابر با مقدار پیش‌فرض آن در تابع آموزشی لونبرگ-مارکوارت یعنی ۰/۰۰۱ به دست آمد.



شکل (۳-۴)- نمودار تغییر خطای شبکه (RF-ANN) با تغییر پارامتر μ

۳-۷-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده از مرحله قبل، مشخصات شبکه عصبی بهینه در جدول (۳-۱۶) نشان داده شده است.

جدول (۳-۱۶)- مشخصات شبکه بهینه (RF-ANN)

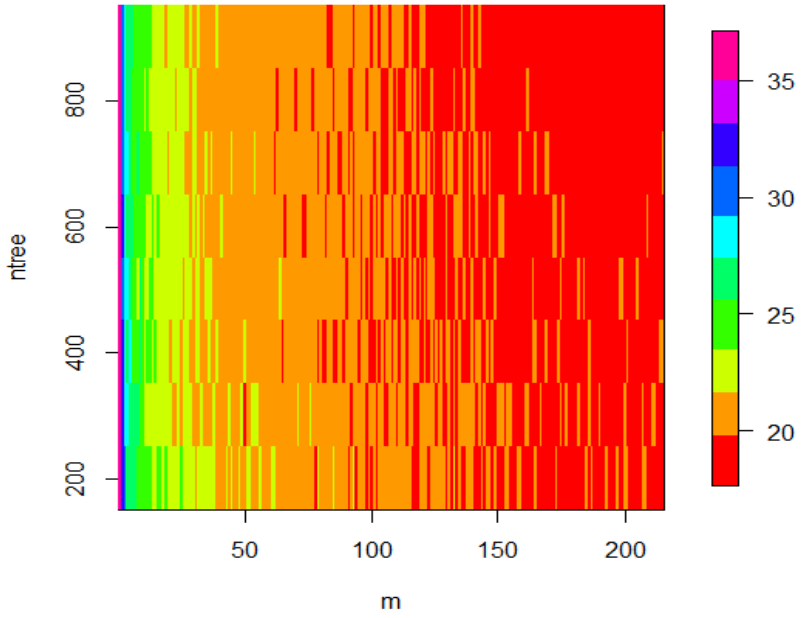
تابع آموزش	لونبرگ-مارکوارت (trainlm)
تابع انتقال لایه پنهان	لگاریتم سیگموئید (logsig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۳
تعداد توصیف‌گر ورودی	۹
تعداد دور آموزشی	۳۷
پارامتر μ	۰/۰۰۱

۳-۸- مدل‌سازی جنگل‌های تصادفی

برای ایجاد مدل جنگل‌های تصادفی از سری آموزش استفاده گردید. داده‌های سری OOB برای بهینه‌سازی پارامترهای مؤثر و داده‌های سری آزمون نیز برای ارزیابی نهایی مدل در نظر گرفته شدند. تعداد درختان (ntree) وتعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (m) ، برای دستیابی به بهترین مدل، بهینه گردیدند.

۳-۸-۱- بهینه‌سازی مقادیر ntree و m

برای بهینه نمودن تعداد درخت و تعداد توصیفگر مورد استفاده در هر مرحله افراز، ntree از ۲۰۰ تا ۹۰۰ با گام‌های ۱۰۰ و m از ۲ تا ۲۱۵ با گام‌های ۱ تغییر داده شده و در هر مرحله جذر خطای OOB یا RMSE[1] محاسبه گردید. نمودار ارائه شده در (۳-۵)، روند تغییر خطا را در مقادیر متفاوت ntree و m نشان می‌دهد. نقاطی با کمترین مقدار خطا که در قسمت قرمزرنگ شکل (۳-۵) وجود دارند، همراه با مقادیر ntree و m متناظر خود، در جدول(۳-۱۷) نشان داده شده‌اند.



شکل (۳-۵)- نمودار تغییرات مجذور خطای OOB یا RMSE در مقادیر متفاوت m و ntree

جدول (۳-۱۷)- کمترین مقادیر RMSE همراه با m و ntree متناظر آنها

RMSE (OOB)	**۱۷/۶۹**	۱۷/۷۳	۱۷/۸۶	۱۷/۸۷	۱۷/۹۹
m	**۷۳**	۱۵۵	۲۰۸	۲۰۱	۲۰۳
ntree	**۶۰۰**	۴۰۰	۴۰۰	۸۰۰	۴۰۰

از آنجا که هدف در این روش کاهش تعداد توصیف‌گرها است، m=73 و ntree=600 به عنوان مقادیر بهینه انتخاب شدند.

1-Root mean square error

۳-۸-۲- مدل‌سازی جنگل‌های تصادفی با استفاده از توصیف‌گرهای مهم

از آنجایی که در مطالعات QSPR، هر چه تعداد توصیف‌گرهای به کار رفته در ساخت مدل کمتر باشد، می‌توان آن را مدل بهتری دانست، لذا ایجاد مدل با روش جنگل‌های تصادفی فقط با توصیف‌گرهای مهم که در جدول (۳-۷) نشان داده شده‌اند، نیز مجددا انجام شد. به این صورت که ابتدا دو توصیف‌گر مهم به عنوان ورودی به کار برده شد و برای یافتن بهترین مدل، بهینه‌سازی انجام شد. به دلیل اینکه اهمیت پارامتر ntree زیاد نیست و تغییرات آن تاثیر اندکی بر قدرت پیش‌بینی دارد، تعداد درخت برابر با ۶۰۰ قرار داده شده و خطای OOB به ازای تغییر m محاسبه گردید. در مرحله بعد، سه توصیف‌گر وارد شده و این مراحل تکرار شد[۳۶]. جدول (۳-۱۸) بهترین نتیجه هر مرحله را نشان می‌دهد.

جدول (۳-۱۸)- تغییرات خطای OOB با تعداد توصیف‌گر متفاوت

MSE(OOB)	تعداد بهینه m	دامنه تغییر m	تعداد توصیف‌گر ورودی
۱۷۰/۹۳	**۲**	**۲**	**۲**
۱۷۴/۵۱	۳	۲-۳	۳
۱۷۵/۵۳	۳	۲-۴	۴
۲۲۱/۸۵	۲	۲-۵	۵
۲۲۸/۳۵	۲	۲-۶	۶
۲۳۰/۴۳	۲	۲-۷	۷
۲۶۱/۴۰	۲	۲-۸	۸
۲۳۵/۷۸	۲	۲-۹	۹
۳۰۸/۳۹	۲	۲-۱۰	۱۰
۲۵۲/۵۴	۴	۲-۱۱	۱۱

همان طور که مشاهده می‌شود، مدلی با ۲ توصیف‌گر و با تعداد درخت ۶۰۰ و m برابر با ۲ کمترین میزان خطا را نشان می‌دهد.

۳-۹- ارزیابی مدل‌ها

۳-۹-۱- ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون

امتیاز و اهمیت مدل‌های پیش‌بینی وقتی مشخص می‌گردد که خواص مولکول‌هایی که در مدل‌سازی به کار نرفته‌اند را پیش‌بینی کند. بدین منظور مدل‌های منتخب، برای پیش‌بینی اندیس بازداری ۱۴ مولکول داده‌ی سری آزمون، به کار برده شدند. جدول (۳-۱۹) نتایج پیش‌بینی و شکل (۳-۶) نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد. در شکل (۳-۷) نیز مقادیر باقی مانده‌ها بر حسب مقادیر تجربی RI ترکیبات مورد بررسی ترسیم شده است که توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است. نتیجه کلی از این ارزیابی این است که همه مدل‌ها توانسته‌اند با صحت خوبی اندیس بازداری را پیش‌بینی کنند اما روش‌های جنگل‌های تصادفی و رگرسیون خطی چندگانه عملکرد بهتری نشان داده‌اند.

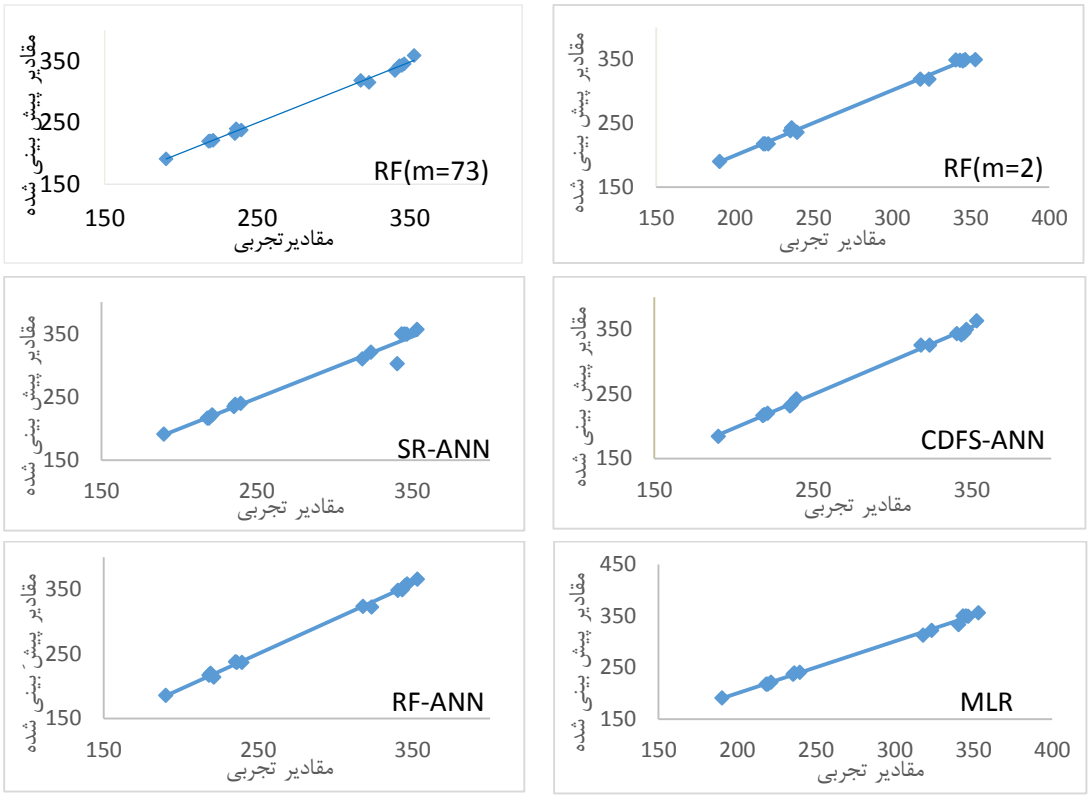
۳-۹-۲- ارزیابی مدل‌ها توسط اعتبارسنجی متقاطع با رد مرحله‌ای تک تک[1]

در این روش که به منظور بررسی تعمیم پذیری مدل‌ها انجام می‌شود، هر بار یک ترکیب از سری داده‌ها کنار گذاشته شده و مدل‌سازی توسط ۵ روش مدل‌سازی ذکر شده با ۷۱ ترکیب باقی مانده صورت گرفت. سپس مدل‌های به دست آمده برای پیش‌بینی اندیس بازداری ترکیب کنار گذاشته شده، به کار گرفته شد. نتایج حاصل از این روش در جدول (۳-۲۰) ارائه شده است. مقادیر پیش‌بینی شده در مقابل مقادیر تجربی نشان داده شده در شکل (۳-۸) و نمایش خطای مطلق در شکل (۳-۹) بر توانایی این مدل‌ها برای پیش‌بینی اندیس بازداری این ترکیبات و عدم وجود خطای سیستماتیک دلالت دارد.

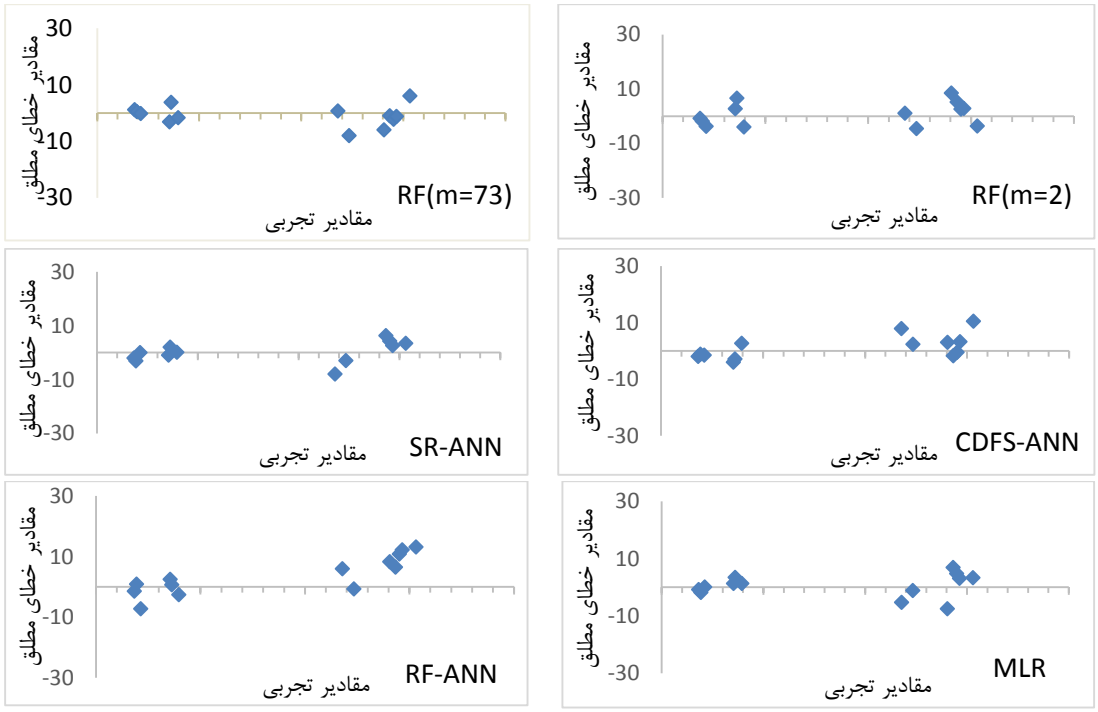
1-Leave one out cross validation

جدول (۳-۱۹)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده						درصد خطا					
		RF (m=73)	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=73)	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۱۴	۱۹۰/۲۸	۱۹۰/۸۱	۱۹۰/۰۹	۱۹۱/۰۳	۱۸۴/۳۱	۱۸۵/۶۷	۱۹۰/۶۶	۰/۲۸	-۰/۱۰	۰/۳۹	-۳/۱۴	-۲/۴۲	۰/۲۰
۲۲	۲۱۸/۲۵	۲۱۹/۵۰	۲۱۷/۵۳	۲۱۶/۲۷	۲۱۶/۲۶	۲۱۶/۹۰	۲۱۷/۵۱	۰/۵۷	-۰/۳۳	-۰/۹۱	-۰/۹۱	-۰/۶۲	-۰/۳۴
۲۳	۲۱۹/۲۸	۲۱۹/۸۴	۲۱۶/۵۳	۲۱۶/۲۷	۲۱۸/۰۹	۲۲۰/۲۷	۲۱۷/۵۱	۰/۲۶	-۱/۲۵	-۱/۳۷	-۰/۵۴	۰/۴۵	-۰/۸۱
۲۴	۲۲۱/۲۲	۲۲۱/۱۴	۲۱۷/۵۳	۲۲۱/۲۲	۲۱۹/۷۶	۲۱۳/۹۹	۲۲۱/۴۰	-۰/۰۴	-۱/۶۷	۰/۰۰	-۰/۶۶	۳/۲۷	۰/۰۸
۲۹	۲۳۵/۳۹	۲۳۲/۲۹	۲۳۸/۱۹	۲۳۴/۴۴	۲۳۱/۳۶	۲۳۷/۸۹	۲۳۶/۷۴	-۱/۳۲	۱/۱۹	-۰/۴۰	-۱/۷۱	۱/۰۶	۰/۵۷
۳۱	۲۳۶/۱۵	۲۴۰/۰۱	۲۴۲/۸۰	۲۳۸/۱۸	۲۳۳/۳۳	۲۳۶/۸۶	۲۳۹/۶۶	۱/۶۴	۲/۸۲	۰/۸۶	-۱/۱۹	۰/۳۰	۱/۴۹
۳۵	۲۳۹/۵۴	۲۳۷/۹۶	۲۳۵/۵۹	۲۳۹/۷۱	۲۴۲/۲۰	۲۳۷/۰۱	۲۴۰/۹۰	-۰/۶۶	-۱/۶۵	۰/۰۷	۱/۱۱	-۱/۰۶	۰/۵۷
۵۱	۳۱۷/۸۷	۳۱۸/۶۴	۳۱۹/۰۶	۳۰۹/۹۵	۳۲۵/۷۹	۳۲۳/۸۹	۳۱۲/۷۰	۰/۲۴	۰/۳۷	-۲/۴۹	۲/۴۹	۱/۸۹	-۱/۶۳
۵۲	۳۲۳/۳۸	۳۱۵/۴۱	۳۱۸/۸۸	۳۲۰/۴۹	۳۲۵/۷۱	۳۲۲/۷۲	۳۲۲/۲۷	-۲/۴۶	-۱/۳۹	-۰/۸۹	۰/۷۲	-۰/۲۰	-۰/۳۴
۵۶	۳۴۰/۴۳	۳۳۴/۵۸	۳۴۹/۰۴	۳۰۲/۳۲	۳۴۳/۴۳	۳۴۸/۷۷	۳۳۳/۰۳	-۱/۷۲	۲/۵۳	-۱۱/۱۹	۰/۸۸	۲/۴۵	-۲/۱۷
۵۸	۳۴۳/۲۱	۳۴۲/۲۶	۳۴۸/۵۳	۳۴۹/۵۴	۳۴۱/۴۷	۳۴۹/۷۹	۳۵۰/۱۹	-۰/۲۸	۱/۵۵	۱/۸۴	-۰/۵۱	۱/۹۲	۲/۰۳
۵۹	۳۴۵/۰۵	۳۴۲/۸۸	۳۴۷/۶۹	۳۴۹/۳۳	۳۴۴/۶۰	۳۵۵/۹۱	۳۴۹/۸۸	-۰/۶۳	۰/۷۶	۱/۲۴	-۰/۱۳	۳/۱۵	۱/۴۰
۶۰	۳۴۶/۴۴	۳۴۵/۲۶	۳۴۹/۳۸	۳۴۹/۱۵	۳۴۹/۶۷	۳۵۸/۷۲	۳۴۹/۶۱	-۰/۳۴	۰/۸۵	۰/۷۸	۰/۹۳	۳/۵۴	۰/۹۱
۶۳	۳۵۳/۰۲	۳۵۹/۰۸	۳۴۹/۴۸	۳۵۶/۵۶	۳۶۳/۴۹	۳۶۶/۱۹	۳۵۶/۴۴	۱/۷۲	-۱/۰۰	۱/۰۰	۲/۹۶	۳/۷۳	۰/۹۷



شکل (۳-۶)- نمودار مقادیر پیش‌بینی شده اندیس بازداری داده‌های سری آزمون توسط مدل‌های مختلف در مقابل مقادیر تجربی



شکل (۳-۷)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای داده‌های سری آزمون با مدل‌های مختلف

جدول (۳-۲۰)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از روش رد مرحله‌ای تک‌تک

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۱	۱۰۰/۰۰	۱۰۸/۳۳	۱۱۸/۲۷	۹۵/۸۵	۹۹/۷۷	۸۸/۳۴	۸/۳۳	۱۸/۲۷	-۴/۱۵	-۰/۲۳	-۱۱/۶۶
۲	۱۱۰/۶۶	۱۱۵/۰۶	۱۱۱/۱۱	۱۳۲/۹۵	۱۱۳/۳۷	۱۱۱/۳۸	۳/۹۸	۰/۴۱	۲۰/۱۵	۲/۴۵	۰/۶۵
۳	۱۱۱/۱۸	۱۱۴/۴۴	۱۱۰/۸۴	۱۳۳/۳۹	۱۱۱/۶۹	۱۱۰/۸۷	۲/۹۳	-۰/۳۰	۱۹/۹۷	۰/۴۶	-۰/۲۸
۴	۱۲۲/۳۸	۱۱۹/۳۸	۱۲۹/۱۰	۱۴۲/۳۶	۱۴۲/۳۰	۱۳۳/۳۳	-۲/۴۵	۵/۴۹	۱۶/۳۲	۱۶/۲۸	۸/۹۵
۵	۱۲۲/۵۱	۱۱۹/۹۸	۱۲۸/۹۳	۱۴۲/۴۱	۱۲۴/۳۱	۱۳۳/۲۳	-۲/۰۷	۵/۲۴	۱۶/۲۴	۱/۴۷	۸/۷۵
۶	۱۲۳/۲۹	۱۱۸/۷۶	۱۰۵/۸۷	۱۴۵/۹۰	۱۳۰/۲۸	۱۲۳/۶۱	-۳/۶۸	-۱۴/۱۳	۱۸/۳۴	۵/۶۷	۰/۲۶
۷	۱۵۰/۴۱	۱۲۰/۰۸	۱۴۹/۱۵	۱۶۰/۷۰	۱۴۲/۹۸	۱۵۲/۰۷	-۲۰/۱۷	-۰/۸۴	۶/۸۴	-۴/۹۴	۱/۱۰
۸	۱۷۵/۳۲	۱۹۰/۹۶	۱۷۰/۶۴	۱۸۰/۳۸	۱۶۶/۰۹	۱۷۴/۳۶	۸/۹۲	-۲/۶۷	۲/۸۹	-۵/۲۶	-۰/۳۹
۹	۱۸۶/۸۳	۱۹۰/۷۴	۱۷۹/۷۷	۱۸۶/۳۶	۱۸۰/۱۷	۱۸۳/۴۱	۲/۰۹	-۳/۷۸	-۰/۲۵	-۳/۵۶	-۱/۸۳
۱۰	۱۷۸/۵۱	۱۹۲/۴۳	۱۸۵/۹۹	۱۸۸/۷۲	۱۸۸/۸۳	۱۸۲/۴۶	۷/۸۰	۴/۱۹	۵/۷۲	۵/۷۸	۲/۲۲
۱۱	۱۸۹/۴۰	۱۹۷/۹۳	۱۹۹/۹۸	۱۹۵/۳۴	۱۸۷/۸۷	۱۸۶/۱۱	۴/۵۰	۵/۵۹	۳/۱۴	-۰/۸۱	-۱/۷۴
۱۲	۱۸۹/۷۷	۱۹۰/۱۱	۱۸۹/۹۸	۱۸۹/۵۱	۱۸۷/۴۵	۱۸۸/۶۰	۰/۱۸	۰/۱۱	-۰/۱۴	-۱/۲۲	-۰/۶۲
۱۳	۱۸۹/۹۶	۱۹۰/۶۹	۱۸۹/۶۸	۱۹۰/۶۰	۲۰۲/۶۹	۱۸۷/۶۰	۰/۳۸	-۰/۱۵	۰/۳۴	۶/۷۰	-۱/۲۴
۱۴	۱۹۰/۲۸	۱۹۰/۳۸	۱۸۹/۶۷	۱۸۷/۹۶	۱۸۶/۱۲	۱۸۹/۴۷	۰/۰۵	-۰/۳۲	-۱/۲۲	-۲/۱۸	-۰/۴۳
۱۵	۱۹۱/۸۳	۱۸۹/۶۲	۱۸۹/۸۱	۱۹۰/۵۱	۱۸۹/۳۱	۱۸۷/۹۶	-۱/۱۵	-۱/۰۵	-۰/۶۹	-۱/۳۱	-۲/۰۲

ادامه جدول (۳-۲۰)

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۱۶	۱۹۵/۷۶	۱۹۴/۴۹	۲۱۹/۳۴	۲۰۲/۲۷	۲۰۶/۱۵	۱۹۲/۴۰	-۰/۶۴	۱۲/۰۵	۳/۳۲	۵/۳۱	-۱/۷۲
۱۷	۲۰۰/۰۰	۱۹۶/۲۸	۱۹۸/۹۱	۲۰۹/۱۵	۲۰۹/۴۵	۱۹۹/۳۲	-۱/۸۶	-۰/۵۵	۴/۵۸	۴/۷۲	-۰/۳۴
۱۸	۲۰۴/۳۱	۱۸۷/۴۷	۱۶۳/۶۲	۲۱۳/۴۵	۱۶۰/۹۳	۱۸۸/۹۲	-۸/۲۴	-۱۹/۹۱	۴/۴۸	-۲۱/۲۳	-۷/۵۳
۱۹	۲۰۹/۳۵	۱۸۵/۹۶	۲۰۶/۱۶	۲۰۴/۷۷	۱۷۱/۷۳	۲۰۵/۳۶	-۱۱/۱۷	-۱/۵۲	-۲/۱۹	-۱۷/۹۷	-۱/۹۰
۲۰	۲۱۶/۱۱	۲۱۹/۲۹	۲۱۶/۴۰	۲۲۲/۹۱	۲۱۶/۲۴	۲۱۷/۲۹	۱/۴۷	۰/۱۳	۳/۱۵	۰/۰۶	۰/۵۵
۲۱	۲۱۶/۸۳	۲۱۹/۱۱	۲۱۸/۰۴	۲۲۵/۹۳	۲۱۹/۵۴	۲۱۹/۴۸	۱/۰۲	۰/۵۶	۴/۲۰	۱/۲۵	۱/۲۲
۲۲	۲۱۸/۲۵	۲۱۸/۹۲	۲۱۶/۰۴	۲۲۲/۸۵	۲۱۸/۲۶	۲۱۷/۲۰	۰/۳۰	-۱/۰۱	۲/۱۱	۰/۰۱	-۰/۴۸
۲۳	۲۱۹/۲۸	۲۱۸/۷۲	۲۱۵/۸۹	۲۲۴/۳۲	۲۲۱/۷۵	۲۱۷/۱۵	-۰/۲۵	-۱/۵۴	۲/۳۰	۱/۱۲	-۰/۹۷
۲۴	۲۲۱/۲۲	۲۱۸/۲۰	۲۱۹/۰۸	۲۲۵/۱۶	۲۵۱/۳۷	۲۲۰/۹۴	-۱/۳۶	-۰/۹۷	۱/۷۸	۱۳/۶۳	-۰/۱۳
۲۵	۲۲۱/۵۰	۲۱۸/۱۶	۲۱۶/۱۷	۲۲۴/۳۱	۲۲۰/۲۸	۲۱۷/۷۶	-۱/۵۱	-۲/۴۱	۱/۲۷	-۰/۵۵	-۱/۶۹
۲۶	۲۳۰/۵۸	۲۳۸/۳۹	۲۳۶/۵۱	۲۳۹/۶۵	۲۳۶/۴۲	۲۳۷/۷۳	۳/۸۹	۲/۵۷	۳/۹۳	۲/۵۳	۳/۰۱
۲۷	۲۳۴/۱۵	۲۴۷/۲۰	۲۳۶/۳۹	۲۳۹/۸۸	۲۰۶/۵۴	۲۳۷/۶۴	۵/۵۷	۰/۹۶	۲/۴۵	-۱۱/۷۹	۱/۴۹
۲۸	۲۳۴/۹۴	۲۳۶/۳۷	۲۱۸/۹۹	۲۴۱/۴۲	۲۴۵/۷۸	۲۳۰/۵۳	۰/۶۱	-۶/۷۹	۲/۷۶	۴/۶۲	-۱/۸۸
۲۹	۲۳۵/۳۹	۲۳۸/۹۰	۲۳۴/۸۴	۲۳۸/۱۰	۲۳۹/۱۵	۲۳۶/۳۴	۱/۴۹	-۰/۲۳	۱/۱۵	۱/۶۰	۰/۴۰
۳۰	۲۳۶/۰۷	۲۳۵/۵۳	۲۳۶/۶۲	۲۴۱/۴۸	۲۵۵/۰۲	۲۳۷/۹۹	-۰/۲۳	۰/۲۳	۲/۲۹	۸/۰۳	۰/۸۱

ادامه جدول (۳-۲۰)

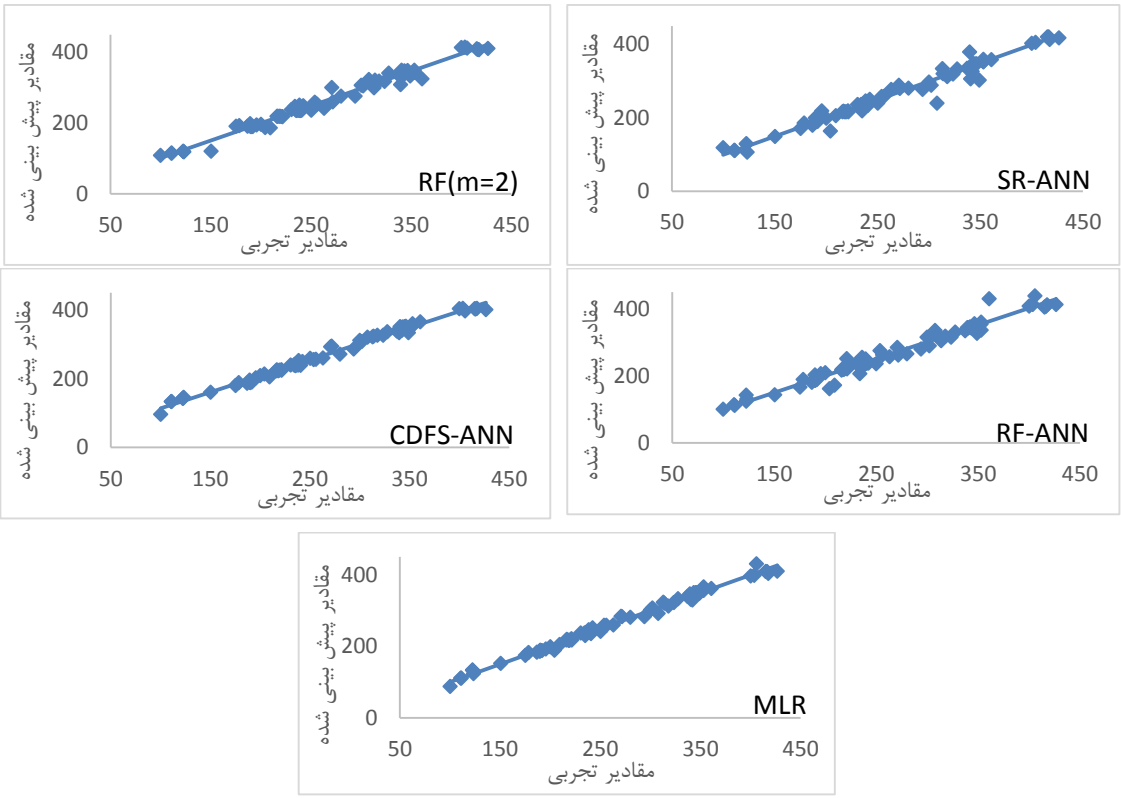
شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۳۱	۲۳۶/۱۵	۲۴۳/۵۶	۲۳۸/۰۲	۲۳۹/۱۶	۲۳۵/۳۰	۲۳۹/۲۱	۳/۱۴	۰/۷۹	۱/۲۸	-۰/۳۶	۱/۲۹
۳۲	۲۳۸/۰۹	۲۴۲/۱۸	۲۳۸/۰۱	۲۳۹/۱۰	۲۳۱/۶۸	۲۳۹/۲۱	۱/۷۲	-۰/۰۳	۰/۴۲	-۲/۶۹	۰/۴۷
۳۳	۲۳۸/۳۰	۲۳۴/۹۴	۲۳۷/۷۹	۲۴۰/۶۴	۲۳۵/۴۳	۲۳۹/۰۰	-۱/۴۱	-۰/۲۱	۰/۹۸	-۱/۲۱	۰/۲۹
۳۴	۲۳۸/۵۶	۲۵۱/۰۷	۲۴۶/۱۴	۲۵۲/۴۱	۲۳۶/۸۳	۲۴۶/۴۰	۵/۲۵	۳/۱۸	۵/۸۱	-۰/۷۳	۳/۲۸
۳۵	۲۳۹/۵۴	۲۳۵/۱۶	۲۳۹/۲۶	۲۴۱/۹۸	۲۴۹/۷۳	۲۴۰/۳۲	-۱/۸۳	-۰/۱۲	۱/۰۲	۴/۲۵	۰/۳۲
۳۶	۲۴۰/۶۶	۲۳۵/۸۰	۲۳۴/۱۹	۲۳۸/۸۲	۲۳۶/۷۲	۲۳۵/۴۳	-۲/۰۲	-۲/۶۹	-۰/۷۶	-۱/۶۴	-۲/۱۷
۳۷	۲۴۲/۳۷	۲۴۸/۷۸	۲۵۰/۳۳	۲۴۹/۴۶	۲۳۷/۹۸	۲۵۱/۴۷	۲/۶۵	۳/۲۸	۲/۹۲	-۱/۸۱	۳/۷۵
۳۸	۲۵۰/۱۴	۲۳۶/۴۲	۲۳۹/۷۲	۲۵۹/۴۳	۲۳۵/۹۴	۲۴۱/۵۲	-۵/۴۸	-۴/۱۶	۳/۷۲	-۵/۶۸	-۳/۴۵
۳۹	۲۵۳/۹۱	۲۵۸/۴۳	۲۵۸/۲۳	۲۵۵/۷۹	۲۷۵/۱۹	۲۵۸/۷۸	۱/۷۸	۱/۷۰	۰/۷۴	۸/۳۸	۱/۹۲
۴۰	۲۵۶/۱۷	۲۵۳/۴۵	۲۵۸/۲۴	۲۵۶/۰۵	۲۶۲/۷۷	۲۵۸/۵۰	-۱/۰۶	۰/۸۱	-۰/۰۴	۲/۵۸	۰/۹۱
۴۱	۲۶۳/۱۹	۲۴۱/۵۹	۲۷۷/۳۰	۲۵۹/۸۵	۲۵۶/۸۹	۲۵۹/۵۸	-۸/۲۱	۵/۳۶	-۱/۲۷	-۲/۳۹	-۱/۳۷
۴۲	۲۷۰/۷۵	۳۰۰/۲۴	۲۸۸/۹۷	۲۹۲/۳۰	۲۸۴/۲۹	۲۸۴/۰۲	۱۰/۸۹	۶/۷۳	۷/۹۶	۵/۰۰	۴/۹۰
۴۳	۲۷۱/۸۳	۲۵۹/۶۹	۲۷۹/۸۰	۲۹۴/۱۹	۲۶۱/۸۲	۲۸۲/۶۲	-۴/۴۷	۲/۹۳	۸/۲۲	-۳/۶۸	۳/۹۷
۴۴	۱۸۰/۱۰	۲۷۵/۸۹	۲۸۰/۵۶	۲۷۱/۱۸	۲۶۵/۴۵	۲۸۱/۲۷	-۱/۵۰	۰/۱۶	-۳/۱۸	-۵/۲۳	۰/۴۲
۴۵	۲۹۳/۹۳	۲۷۵/۷۹	۲۷۷/۲۴	۲۸۶/۱۴	۲۷۹/۶۵	۲۸۳/۳۷	-۶/۱۷	-۵/۶۸	-۲/۶۵	-۴/۸۶	-۳/۵۹

ادامه جدول (۳-۲۰)

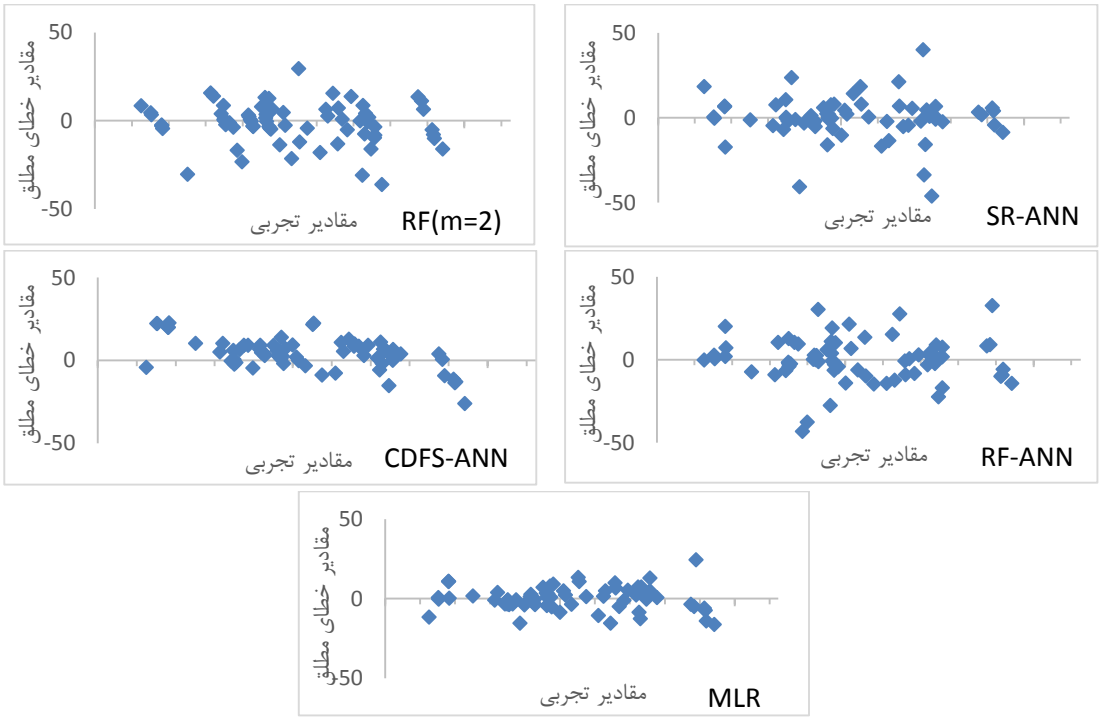
شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۴۶	۳۰۰/۰۰	۳۰۶/۴۴	۲۹۷/۷۸	۳۱۰/۹۲	۳۱۵/۱۸	۳۰۱/۳۹	۲/۱۵	-۰/۷۴	۳/۶۴	۵/۰۶	۰/۴۶
۴۷	۳۰۲/۱۳	۳۰۴/۶۸	۲۸۸/۵۳	۳۰۷/۵۹	۳۸۹/۷۱	۳۰۷/۰۱	۰/۸۴	-۴/۵۰	۱/۸۱	-۴/۱۱	۱/۶۱
۴۸	۳۰۷/۷۷	۳۲۳/۳۲	۲۳۹/۸۵	۳۲۰/۵۳	۳۳۵/۲۶	۲۹۲/۴۱	۵/۰۵	-۲۲/۰۷	۴/۱۵	۸/۹۳	-۴/۹۹
۴۹	۳۱۳/۰۵	۲۹۹/۹۴	۳۳۴/۲۹	۳۲۱/۵۹	۳۱۲/۱۳	۳۲۳/۰۳	-۴/۱۹	۶/۷۸	۲/۷۳	-۰/۲۹	۳/۱۹
۵۰	۳۱۳/۷۶	۳۲۰/۹۴	۳۲۰/۶۰	۳۲۳/۵۷	۳۰۴/۶۱	۳۲۰/۵۳	۲/۲۹	۲/۱۸	۳/۱۳	-۲/۹۲	۲/۱۶
۵۱	۳۱۷/۸۷	۳۱۸/۶۳	۳۱۲/۵۷	۳۲۶/۴۴	۳۱۸/۳۷	۳۱۲/۸۳	۰/۲۴	-۱/۶۷	۲/۷۰	۰/۱۶	-۱/۵۸
۵۲	۳۲۳/۳۸	۳۱۸/۱۱	۳۱۸/۹۳	۳۲۶/۲۸	۳۱۵/۰۴	۳۲۲/۲۰	-۱/۶۳	-۱/۳۸	۰/۹۰	-۲/۵۸	-۰/۳۷
۵۳	۳۲۷/۶۴	۳۴۱/۳۶	۳۳۳/۰۶	۳۳۶/۷۸	۳۳۰/۵۴	۳۳۲/۸۸	۴/۱۹	۱/۶۵	۲/۷۹	۰/۸۸	۱/۶۰
۵۴	۳۳۷/۲۰	۳۳۷/۰۲	۳۳۵/۱۴	۳۳۸/۵۴	۳۳۴/۰۹	۳۳۹/۴۲	-۰/۰۵	-۰/۶۱	۰/۴۰	-۰/۹۲	۰/۶۶
۵۵	۳۳۹/۵۱	۳۰۸/۵۵	۳۷۹/۵۶	۳۳۳/۸۱	۳۴۳/۴۶	۳۴۶/۸۰	-۹/۱۲	۱۱/۸۰	-۱/۶۸	۱/۱۶	۲/۱۵
۵۶	۳۴۰/۴۳	۳۴۹/۰۰	۳۰۶/۶۱	۳۵۱/۴۹	۳۴۲/۵۲	۳۳۱/۷۴	۲/۵۲	-۹/۹۳	۳/۲۵	۰/۶۱	-۲/۵۵
۵۷	۳۴۲/۰۸	۳۳۴/۴۴	۳۲۶/۲۴	۳۴۰/۲۳	۳۴۶/۰۱	۳۲۹/۴۰	-۲/۲۳	-۴/۶۳	-۰/۵۴	۱/۱۵	-۳/۷۱
۵۸	۳۴۳/۲۱	۳۴۷/۳۴	۳۴۷/۸۳	۳۵۰/۰۲	۳۴۳/۱۹	۳۵۰/۵۵	۱/۲۰	۱/۳۵	۱/۹۹	-۰/۰۱	۲/۱۴
۵۹	۳۴۵/۰۵	۳۴۴/۳۰	۳۴۷/۴۸	۳۵۰/۸۹	۳۴۲/۶۹	۳۵۰/۰۶	-۰/۲۲	۰/۷۰	۱/۶۹	-۰/۶۸	۱/۴۵
۶۰	۳۴۶/۴۴	۳۴۸/۳۸	۳۴۷/۲۴	۳۵۲/۱۱	۳۵۵/۱۶	۳۴۹/۸۰	۰/۵۶	۰/۲۳	۱/۶۴	۲/۵۲	۰/۹۷

ادامه جدول (۳-۲۰)

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۶۱	۳۴۸/۹۴	۳۳۲/۸۰	۳۰۲/.۶۸	۳۳۳/۶۳	۳۲۶/۴۲	۳۴۸/۵۱	-۴/۶۳	-۱۳/۲۶	-۴/۳۹	-۶/۴۵	-۰/۱۲
۶۲	۳۵۲/۹۰	۳۴۳/۱۱	۳۵۳/۴۷	۳۵۲/۸۸	۳۵۴/۵۸	۳۵۷/۷۸	-۲/۷۷	۰/۱۶	-۰/۰۱	۰/۴۸	۱/۳۸
۶۳	۳۵۳/۰۲	۳۴۹/۵۰	۳۵۹/۶۴	۳۵۹/۵۴	۳۳۶/۰۳	۳۶۵/۹۱	-۱/۰۰	۱/۸۸	۱/۸۵	-۴/۸۱	۳/۶۵
۶۴	۳۵۳/۰۲	۳۴۴/۶۷	۳۵۲/۱۳	۳۵۴/۶۶	۳۶۰/۴۸	۳۵۶/۳۲	-۲/۳۶	-۰/۲۵	۰/۴۶	۲/۱۱	۰/۹۴
۶۵	۳۶۰/۹۳	۳۲۴/۶۵	۳۵۸/۴۲	۳۶۴/۸۹	۴۳۰/۱۴	۳۶۱/۵۴	-۱۰/۰۵	۰/۷۰	۱/۱۰	۱۹/۱۸	۰/۱۷
۶۶	۴۰۰/۰۰	۴۱۳/۴۱	۴۰۳/۱۴	۴۰۳/۹۱	۴۰۸/۱۵	۳۹۶/۳۹	۳/۳۵	۰/۷۹	۰/۹۸	۲/۰۴	-۰/۹۰
۶۷	۴۰۳/۶۹	۴۱۴/۸۷	۴۰۵/۴۱	۴۰۴/۲۳	۴۱۲/۶۴	۳۹۸/۷۳	۲/۷۷	۰/۴۳	۰/۱۳	۲/۲۲	-۱/۲۳
۶۸	۴۰۵/۸۹	۴۱۲/۲۳	۴۵۹/۷۶	۳۹۶/۵۷	۴۳۸/۵۲	۴۳۰/۳۴	۱/۵۶	۱۷/۹۱	-۲/۳۰	۸/۰۴	۶/۰۲
۶۹	۴۱۵/۲۸	۴۱۰/۰۱	۴۲۰/۹۷	۴۰۳/۸۸	۴۰۵/۲۵	۴۰۹/۲۱	-۱/۲۷	۱/۳۷	-۲/۷۴	-۲/۴۲	-۱/۴۶
۷۰	۴۱۶/۴۷	۴۰۸/۵۱	۴۲۰/۴۶	۴۰۲/۹۸	۴۰۷/۱۰	۴۰۹/۰۳	-۱/۹۱	۰/۹۶	-۳/۲۴	-۰۲/۲۵	-۱/۷۹
۷۱	۴۱۷/۶۱	۴۰۷/۵۳	۴۱۳/۲۲	۴۰۴/۷۴	۴۱۱/۸۴	۴۰۳/۵۱	-۲/۴۱	-۱/۰۵	-۳/۰۸	-۱/۳۸	-۳/۳۸
۷۲	۴۲۶/۵۷	۴۱۰/۵۲	۴۱۷/۸۲	۴۰۰/۴۷	۴۱۲/۲۷	۴۱۰/۲۲	-۳/۷۶	-۲/۰۵	-۶/۱۲	-۳/۳۵	-۳/۸۳



شکل (۳-۸)- نمودار مقادیر پیش‌بینی شده اندیس بازداری کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل‌های مختلف بر حسب مقادیر تجربی



شکل (۳-۹)- نمودار مقادیر خطای مطلق بر حسب مقادیر تجربی برای کل داده‌ها با مدل‌های مختلف به روش رد مرحله‌ای تک‌تک

۳-۹-۳- ارزیابی مدل‌ها با استفاده از پارامترهای آماری

ده پارامتر آماری که توضیح آنها در بخش ۲-۹ ارائه شده است، جهت ارزیابی توانایی پیش‌بینی مدل‌های به کار گرفته شده، برای داده‌های سری آزمون محاسبه و با یکدیگر مقایسه گردید. مقدار این پارامترها که برتری روش‌های جنگل‌های تصادفی و رگرسیون خطی چندگانه را نشان می‌دهد، در جدول (۳-۲۱) ارائه شده است. جدول (۳-۲۲) نیز مقدار این پارامترها برای کل داده‌ها را توسط مدل‌های مختلف نشان می‌دهد.

جدول (۳-۲۱)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای داده‌های سری آزمون

	RF(m=73)	RF(m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
R^2	۰/۹۹۷	۰/۹۹۵	۰/۹۶۷	۰/۹۹۴	۰/۹۸۶	۰/۹۹۶
R^2_{adj}	-	۰/۹۹۴	۰/۹۴۷	۰/۹۹۳	۰/۹۵۵	۰/۹۹۷
SEP	۳/۴۹۷	۴/۱۱۶	۱۰/۷۵۷	۴/۴۳۱	۶/۹۷۶	۳/۷۴۵
REP	۱/۲۴۶	۱/۴۶۶	۳/۸۳۳	۱/۵۷۷	۲/۴۸۵	۱/۳۳۴
MAE	۲/۵۶۵	۳/۴۶۴	۵/۳۳۴	۳/۵۱۹	۵/۵۵۹	۲/۹۵۵
MRE	۰/۸۶۷	۱/۲۱۵	۱/۶۷۶	۱/۲۷۸	۱/۸۶۲	۰/۹۶۵
MSE	۱۲/۲۳۱	۱۶/۹۳۹	۱۱۵/۷۳۶	۱۹/۶۳۳	۴۸/۶۵۷	۱۴/۰۲۶
PRESS	۱۷۱/۲۳۳	۲۳۷/۱۴۴	۱۶۲۰	۲۷۴/۸۵۷	۶۸۱/۱۹۷	۱۹۶/۳۶۵
F	-	۱۱۹۷	۴۷/۶۹۸	۶۶۲/۴۰۰	۳۸/۴۶۷	۴۱۰/۶۷۳
FIT	-	۱۲۶/۹۲۱	۶/۰۶۰	۷۷/۸۵۰	۳/۰۱۷	۵۱/۴۹۲

جدول (۳-۲۲)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای کل داده‌ها به روش رد تک‌تک

	RF(m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
R^2	۰/۹۸۱	۰/۹۶۳	۰/۹۸۵	۰/۹۶۵	۰/۹۹۳
R^2_{adj}	۰/۹۸۰	۰/۹۶۰	۰/۹۸۵	۰/۹۶۰	۰/۹۹۲
SEP	۱۱/۴۷۳	۱۵/۹۷۳	۹/۹۷۵	۱۵/۳۶۱	۷/۲۲۳
REP	۴/۳۴۹	۶/۰۵۵	۳/۷۸۱	۵/۸۲۳	۲/۷۴۰
MAE	۸/۳۲۰	۹/۱۱۳	۷/۷۳۶	۱۰/۱۸۰	۵/۳۷۶
MRE	۳/۳۳۳	۳/۵۸۶	۳/۵۹۳	۴/۰۲۷	۲/۱۲۶
MSE	۱۳۱/۶۲۰	۲۵۵/۱۴۵	۹۹/۴۹۹	۲۳۵/۹۶۲	۵۲/۲۵۵
PRESS	۹۴۷۶	۱۸۳۷۱	۷۱۶۴	۱۶۹۸۹	۳۷۶۲
F	۱۷۵۱	۳۶۵	۱۳۹۶	۲۰۹	۱۷۲۰
FIT	۴۶/۰۵۳	۱۷/۴۷۵	۵۶/۶۰۲	۱۱/۲۸۶	۸۷/۹۶۶

۳-۹-۴- ارزیابی مدل‌ها با استفاده از آزمون Y-تصادفی[1]

این تکنیک با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام می‌شود. در این آزمون، مقادیر تجربی متغیر وابسته در محدوده همان مقادیر به طور تصادفی تغییر داده شده و مدل‌های جدید با تمام روش‌های ذکر شده، ایجاد شدند. نتایج حاصل از ۱۰ بار اجرای این آزمون در جدول (۳-۲۳) نشان داده شده است. تفاوت قابل ملاحظه بین مقادیر ضریب تعیین مدل‌های اصلی و مدل‌های جدید، بیانگر عدم وجود ارتباط تصادفی در مدل‌ها است.

جدول (۳-۲۳)- مقادیر R^2 برای داده‌های سری آزمون با استفاده از آزمون Y-تصادفی

تکرار / روش	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
RF(m=73)	۰/۱۵۹	۰/۳۸۷	۰/۰۶۱	۰/۰۱۲	۰/۰۱۱	۰/۱۰۲	۰/۰۱۰	۰/۰۸۷	۰/۰۶۳	۰/۲۸۴
RF(m=2)	۰/۰۵۱	۰/۱۸۰	۰/۰۴۴	۰/۰۱۷	۰/۰۵۶	۰/۱۰۹	۰/۰۲۸	۰/۰۰۸	۰/۰۷۷	۰/۳۱۴
SR-ANN	۰/۰۵۲	۰/۱۸۷	۰/۰۷۹	۰/۱۳۳	۰/۳۵۵	۰/۱۸۹	۰/۱۵۹	۰/۲۳۵	۰/۳۴۵	۰/۳۴۸
CDFS-ANN	۰/۴۱۶	۰/۰۵۹	۰/۴۶۶	۰/۱۱۷	۰/۴۷۹	۰/۲۴۰	۰/۳۷۱	۰/۴۳۲	۰/۴۸۵	۰/۴۸۲
RF-ANN	۰/۲۸۱	۰/۱۱۶	۰/۴۶۶	۰/۳۹۰	۰/۰۸۷	۰/۲۴۵	۰/۲۰۰	۰/۳۶۱	۰/۲۲۰	۰/۲۷۰

۳-۱۰- بررسی ارتباط توصیف‌گرهای منتخب مدل برتر با اندیس بازداری

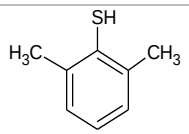
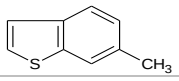
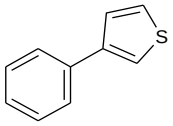
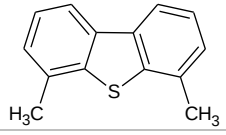
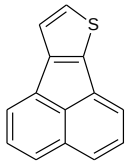
مدل برتر برای پیش‌بینی اندیس بازداری، روش جنگل‌های تصادفی است. دو توصیف‌گر که در ساخت این مدل به کار برده شدند، SCBO و BELe6 می‌باشند.

توصیفگر SCBO از جمله توصیف‌گرهای زیرساختاری[2] است و این گروه از توصیف‌گرها، ساده‌ترین و پر استفاده‌ترین توصیف‌گرها هستند که ساختار مولکولی یک ترکیب را بدون هیچ اطلاعاتی در رابطه با هندسه مولکول‌ها بررسی می‌کنند. معمول‌ترین توصیف‌گرهای زیرساختاری شامل تعداد اتم، تعداد پیوندهای چندگانه و تعداد حلقه‌های آروماتیک می‌باشند. SCBO با مجموع پیوندهای دوگانه متناسب است. با افزایش تعداد پیوند دوگانه در مولکول، مقدار این توصیف‌گر افزایش می‌یابد. هم‌چنین با افزایش

1-Y-randomaization test
2-Constitutional

تعداد پیوند دوگانه در مولکول، میزان قطبیت ترکیب افزایش یافته و در نتیجه تمایل مولکول به فاز ساکن قطبی افزایش و اندیس بازداری نیز افزایش می‌یابد. جدول (۳-۲۴) چگونگی ارتباط این توصیف‌گر با تعداد پیوند دوگانه و اندیس بازداری را برای تعدادی از مولکول‌های سری داده‌ها نشان می‌دهد.

جدول (۳-۲۴)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر SCBO

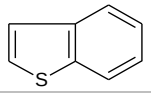
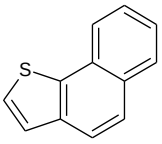
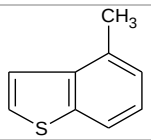
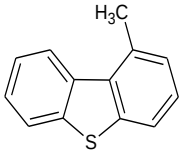
شماره ترکیب	ساختار	تعداد پیوند دوگانه	مقدار SCBO	اندیس بازداری
۱۵		۳	۱۲	۱۹۱/۸۳
۲۳		۴	۱۵	۲۱۹/۲۸
۳۷		۵	۱۷	۲۴۲/۳۷
۵۳		۶	۲۳	۳۲۷/۶۴
۶۳		۷	۲۵	۳۵۳/۰۲

توصیف‌گر دوم BELe6 از گروه توصیف‌گرهای BCUT[1] می‌باشد. این توصیف‌گرها مقادیر ویژه‌ای از یک ماتریس ارتباطی تغییر شکل یافته به نام ماتریس بوردن[2] هستند. ماتریس بوردن یک گراف مولکولی تهی از هیدروژن را ارائه می‌کند که این گراف مولکولی با علامت G=G(V,E) تعریف می‌شود که V نشانگر رئوس گراف است که بیانگر تعداد اتم‌های موجود در ساختار مولکول است و E نشانگر یال‌های گراف است که متناظر با پیوندهای متصل کننده‌ی این رئوس (اتم‌ها) به یکدیگر است. تعداد یال متصل به یک رأس را درجه‌ی آن رأس می‌گویند. گرافی که بدون در نظر گرفتن اتم هیدروژن به دست آید، گراف مولکولی تهی از هیدروژن است. در ماتریس بوردن، عناصر قطری در ارتباط با اعداد اتمی،

1-Burden-CAS-university of texas eigenvalue
2-Burden matrix

الکترونگاتیویته، قطبش‌پذیری و غیره و عناصر غیرقطری در ارتباط با مرتبه پیوند دو اتم متصل به یکدیگر هستند. عناصر غیرقطری این ماتریس در صورتی که $d_{ij}=k$ باشد، یک و در غیر این صورت صفر هستند. K به عنوان lag تعریف می‌شود که فاصله توپولوژیکی بین دو اتم است و می‌تواند مقداری از یک تا هشت داشته باشد. مقادیر ویژه بالا و پایین این ماتریس منجر به ایجاد توصیف‌گرهای متفاوتی می‌گردد. BEL پایین‌ترین مقدار ویژه منفی از ماتریس بوردن است و حاوی اطلاعاتی در مورد برهمکنش‌های قطبی و قطبش‌پذیری مولکول‌ها می‌باشد و مقدار صفر یا یک دارد. یک بودن مقدار این توصیف‌گر برای یک مولکول به معنای قطبیت بیشتر آن ترکیب است و در نتیجه‌ی این قطبیت، تمایل به فاز ساکن قطبی زیاد و اندیس بازداری نیز افزایش می‌یابد[۳۸ و ۴۵]. جدول (۳-۲۵) چگونگی تاثیر مقدار این توصیف‌گر بر اندیس بازداری را برای تعدادی از مولکول‌ها نشان می‌دهد. با اضافه شدن حلقه و تجمع بار منفی، قطبیت و اندیس بازداری نیز افزایش می‌یابد.

جدول (۳-۲۵)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر BELe6

شماره ترکیب	ساختار	مقدار BELe6	اندیس بازداری
۱۷		۰	۲۰۰/۰۰
۴۷		۱	۳۰۲/۱۳
۲۵		۰	۲۲۱/۵۰
۵۲		۱	۳۲۳/۳۸

۳-۱۱- نتیجه‌گیری

در این تحقیق از پنج مدل مختلف برای پیش‌بینی اندیس بازداری ترکیبات گوگرددار استفاده شد. از بین سه شبکه عصبی مصنوعی به کار برده شده ، شبکه با توصیف‌گرهای انتخاب شده با راهبرد CDFS عملکرد بهتری نسبت به دو روش دیگر داشته است. اما در حالت کلی، روش‌های جنگل‌های تصادفی و رگرسیون خطی چندگانه نسبت به روش شبکه عصبی مصنوعی نتایج بهتری را ارائه داده‌اند. بر اساس این مدل‌ها و تنها با دانستن ساختار ترکیبات و محاسبه‌ی توصیف‌گرهای مربوطه، می‌توان اندیس بازداری ترکیبات مشابه و یا همین طبقه از ترکیبات را که هنوز سنتز نشده‌اند، پیش‌بینی نمود.

فصل چهارم
مدل‌سازی جهت پیش‌بینی اندیس بازداری بعضی ترکیبات آروماتیک چندحلقه‌ای

هدف در این فصل، یافتن یک مدل برای پیش‌بینی اندیس بازداری ترکیبات آروماتیک چند حلقه‌ای است. بخش تجربی شامل معرفی سری داده‌ها، بهینه‌سازی ساختار مولکول‌ها و محاسبه توصیف‌گرهای مولکولی، انتخاب توصیف‌گرهای مولکولی مناسب با روش‌های رگرسیون مرحله‌ای و راهبرد CDFS، مدل‌سازی توسط روش‌های رگرسیون خطی چندگانه، شبکه عصبی مصنوعی، جنگل‌های تصادفی و ارزیابی مدل‌های برتر می‌باشد.

۴-۱- سری‌داده‌ها

سری داده‌ها از نتایج تعیین اندیس بازداری ۸۳ ترکیب آروماتیک با کروماتوگرافی گازی همراه با فاز ساکن قطبی و آشکارساز نشر اتمی به دست آمد[۴۶]. نام، ساختار و اندیس بازداری (RI) تجربی این ترکیبات در جدول (۴-۱) نشان داده شده است. هم‌چنین در این جدول مشخص شده است که هر داده متعلق به کدام سری (آموزش، ارزیابی و یا آزمون) می‌باشد.

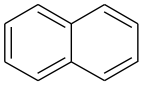
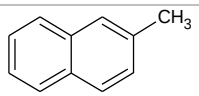
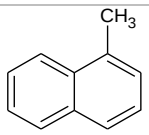
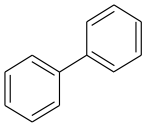
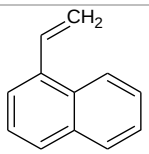
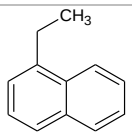
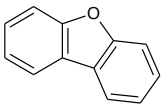
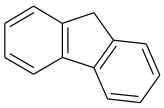
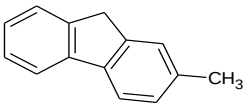
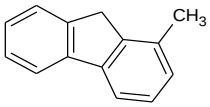
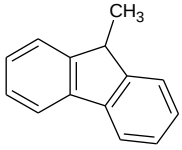
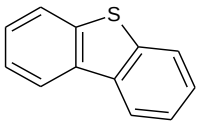
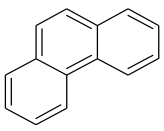
۴-۲- بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌گرها

در این مرحله، ساختار ترکیبات به وسیله نرم‌افزار Hyperchem رسم و برای بهینه کردن ساختار آنها، روش AM1 به کار برده شد. سپس ۱۴۸۱ توصیف‌گر توسط نرم‌افزار Dragon با استفاده از این ساختارهای بهینه شده، محاسبه گردید.

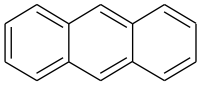
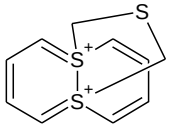
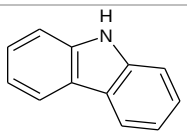
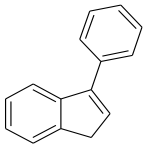
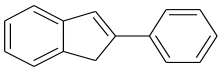
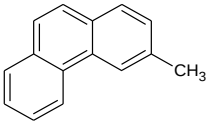
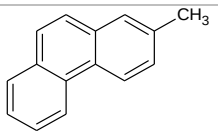
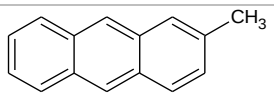
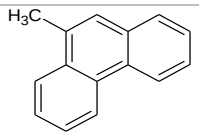
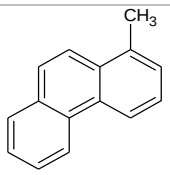
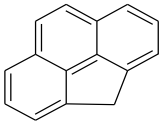
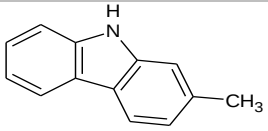
۴-۳- انتخاب بهترین توصیف‌گرها

برای جلوگیری از پیچیدگی محاسبات، ابتدا توصیف‌گرهایی که برای تمام مولکول‌ها مقادیر ثابت یا تقریبا ثابت داشتند. هم‌چنین یکی از هر دو توصیف‌گر با هم‌بستگی بزرگتر از ۰/۹ نیز با استفاده از نرم‌افزار MATLAB حذف گردیدند به طوری که در پایان این مرحله، ۱۹۳ توصیف‌گر باقی ماند.

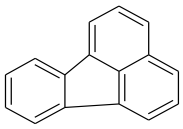
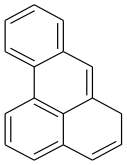
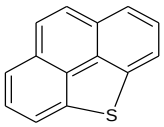
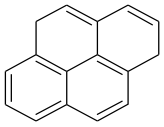
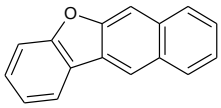
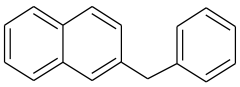
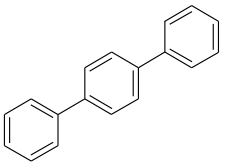
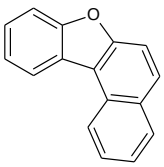
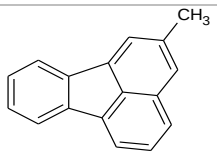
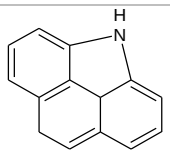
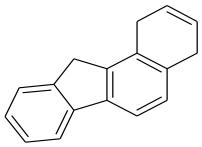
جدول (۴-۱)- نام و اندیس بازداری سری داده‌ها

شماره	نام	ساختار	RI	سری*
۱	Naphthalene		۲۰۰/۰۰	v
۲	2-methyl naphthalene		۲۲۱/۱۰	v
۳	1-methyl naphthalene		۲۲۴/۱۲	tr
۴	Biphenyl		۲۳۶/۵۶	tr
۵	Acenahthylene		۲۴۸/۲۷	tr
۶	Acenaphthene		۲۵۳/۵۶	te
۷	Di benzo furan		۲۵۹/۶۳	tr
۸	Flourene		۲۷۰/۳۹	v
۹	2-methyl flourene		۲۸۸/۲۸	tr
۱۰	1-methyl flourene		۲۸۹/۲۴	v
۱۱	9-Methyl flourene		۲۹۲/۲۸	tr
۱۲	Dibenzothiophene		۲۹۵/۹۵	v
۱۳	Phenanthrene		۳۰۰/۰۰	tr

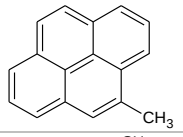
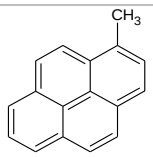
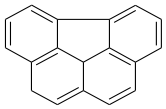
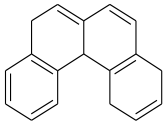
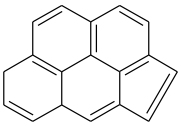
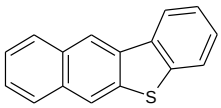
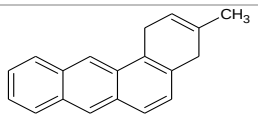
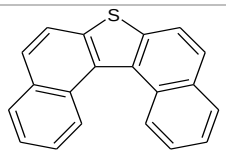
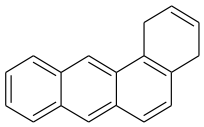
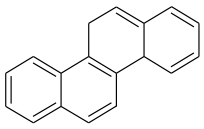
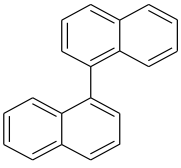
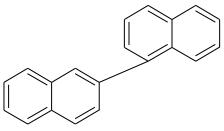
ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۱۴	Anthracene		۳۰۱/۵۳	tr
۱۵	Naphtha[2,3-b]thiophene		۳۰۴/۲۸	v
۱۶	Carbazole		۳۰۹/۸۰	tr
۱۷	1H-indene,1-phenyl		۳۱۱/۱۲	tr
۱۸	1H-Indene,2-phenyl		۳۱۴/۱۵	te
۱۹	2-Methyl phenantherene		۳۱۸/۸۳	v
۲۰	3-Methyl phenanthrene		۳۱۹/۶۷	te
۲۱	2-Methyl anthracene		۳۲۱/۱۲	v
۲۲	9-Methyl phenantherene		۳۲۲/۷۴	tr
۲۳	1-Methyl phenanthrene		۳۲۳/۵۴	te
۲۴	Cyclopentha[def]phenantheren		۳۲۵/۷۳	tr
۲۲۵	2-Methyl carazol		۳۲۶/۹۴	tr

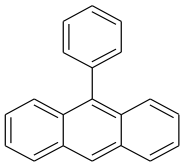
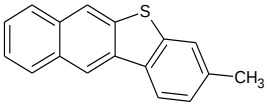
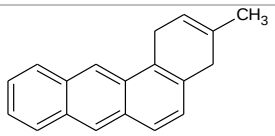
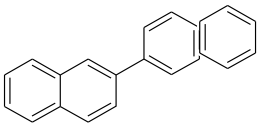
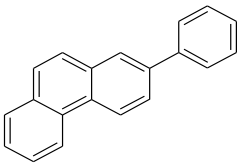
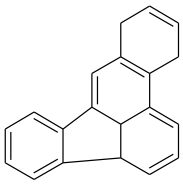
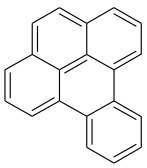
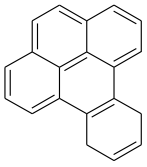
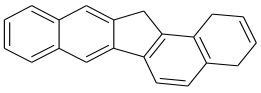
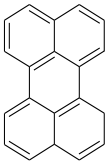
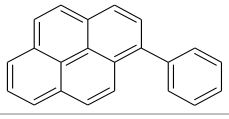
ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۲۶	Fluoranthene		۳۴۴/۹۶	tr
۲۷	Acephenanthrylene		۳۴۸/۱۴	te
۲۸	Phenanthro[4-5bcd]thiophene		۳۴۹/۴۲	tr
۲۹	Pyrene		۳۵۲/۴۱	tr
۳۰	Benzo[b]naphtha[2,3-d]furan		۳۵۳/۴۷	v
۳۱	Phenyl Methyl naphthalene		۳۵۶/۰۴	tr
۳۲	p-trephenyl		۳۵۷/۶۵	tr
۳۳	benzo[b]naphtha[1,2-d]furan		۳۵۷/۸۶	tr
۳۴	2-Methyl fluoranthan		۳۶۲/۰۹	tr
۳۵	4H-Benzo[def]carbazol		۳۶۲/۸۲	v
۳۶	Benzo[a]flourene		۳۶۶/۲۸	tr

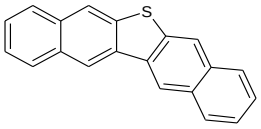
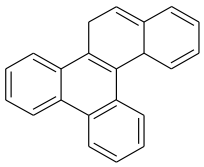
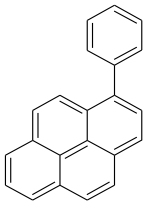
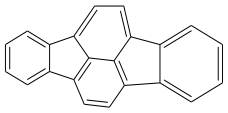
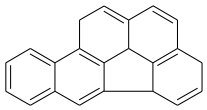
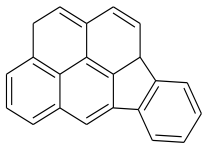
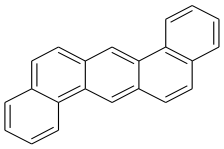
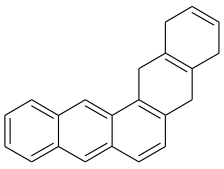
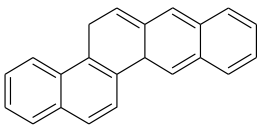
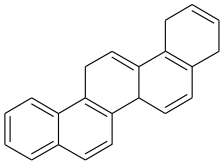
ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۳۷	4-Methyl pyrene		۳۶۸/۷۵	tr
۳۸	1-Methyl pyren		۳۷۳/۵۵	tr
۳۹	Benzo[ghi]flouranthene		۳۹۰/۱۸	te
۴۰	Benzo[c]phenanthrene		۳۹۰/۵۷	te
۴۱	Cyclopenta[cd]pyrene		۳۹۱/۱۰	te
۴۲	Benzo[b]naphtha[2,3-d]thiophene		۳۹۱/۷۳	tr
۴۳	Methylbenzo[a]anthracene		۲۹۲/۷۰	tr
۴۴	Dinaphtha[3,4-d]thiophene		۳۹۴/۷۶	tr
۴۵	Benzo[a]Anthracene		۳۹۸/۳۶	te
۴۶	Chrycene		۴۰۰/۰۰	tr
۴۷	1,1-binaphthalene		۴۰۲/۷۷	tr
۴۸	1,2-binaphthalene		۴۰۴/۲۸	v

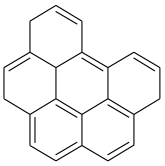
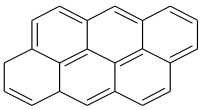
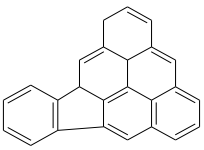
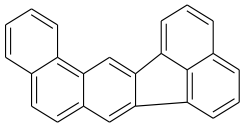
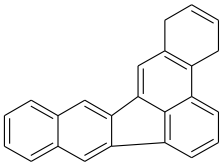
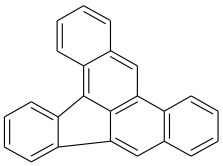
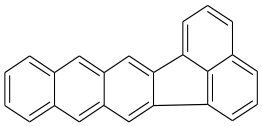
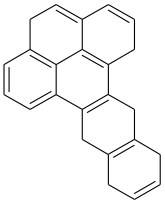
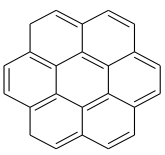
ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۴۹	9-phenyl anthracene		۴۰۶/۴۷	tr
۵۰	Methylbenzo[b]naphtha[2,3-d]thiophene		۴۰۹/۵۴	te
۵۱	Methylbenzo[a]anthracene		۴۱۰/۲۷	te
۵۲	2,2Binaphthalene		۴۲۱/۵۵	tr
۵۳	2-phenylphenanthrene		۴۲۴/۱۹	tr
۵۴	Benzo[b]fluorescence		۴۴۳/۹۳	tr
۵۵	Benzo[c]acephenanthrylene		۴۴۸/۰۷	tr
۵۶	Benzo[c]pyrene		۴۵۲/۹۳	te
۵۷	Dibenzo[a,h]flourane		۴۵۶/۶۰	v
۵۸	Perylene		۴۵۷/۹۸	tr
۵۹	Phenylpyrene		۴۵۹/۰۱	tr

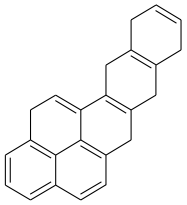
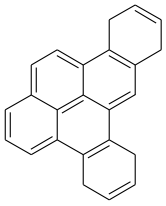
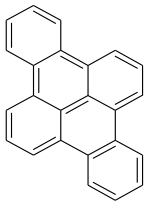
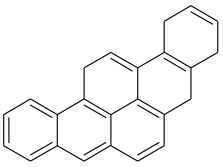
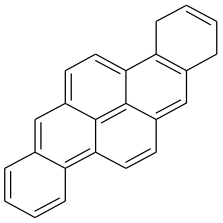
ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۶۰	Dinaphthothiophene		۴۸۱/۹۶	tr
۶۱	Benzo[b]trihenylene		۴۸۳/۲۷	te
۶۲	1-phenyl pyrene		۴۸۶/۳۴	tr
۶۳	Indeno[1,2,3-cd]flpurancene		۴۸۸/۳۸	v
۶۴	Indeno[7,1,2,3-cdef]chrycene		۴۹۱/۵۳	tr
۶۵	Indeno[1,2,3-cd]pyrene		۴۹۴/۹۱	tr
۶۶	Dibenzo[ah]anthracene		۴۹۵/۸۹	tr
۶۷	Pentaphene		۴۹۶/۶۷	tr
۶۸	Benzo[b]chrycene		۴۹۸/۸۶	tr
۶۹	Picene		۵۰۰/۰۰	tr

ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۷۰	Benzo[ghi]perylene		۵۰۱/۶۴	tr
۷۱	Benzo[def,mno]chrysene		۵۰۵/۹۷	tr
۷۲	Dibenzo[be]fluoranthene		۵۳۳/۲۲	te
۷۳	Naphtha[1,2-k]fluoranthene		۵۳۶/۹۲	te
۷۴	Dibenzo[b,k]fluranthene		۵۳۹/۵۹	te
۷۵	Dibenzo[a,e]flouranthene		۵۴۰/۷۱	tr
۷۶	Naphtha[2,3,k]fluoranthene		۵۴۳/۰۶	v
۷۷	Naphtha[2,3,e]pyrene		۵۴۷/۷۱	te
۷۸	Coronene		۵۵۰/۴۳	tr

ادامه جدول (۴-۱)

شماره	نام	ساختار	RI	سری*
۷۹	Naphtha[2,3-a]pyrene		۵۵۱/۱۰	v
۸۰	Dibenzo[a,e]pyrene		۵۵۱/۶۵	v
۸۱	Dibenzo[a.l]pyrene		۵۵۲/۸۲	tr
۸۲	Dibenzo[a,i]pyrene		۵۵۶/۳۲	v
۸۳	Dibenzo[a,h]pyrene		۵۵۹/۹۰	tr

* داده‌های سری آموزش:tr , داده‌های سری ارزیابی:v , داده‌های سری آزمون:te

برای مدل‌سازی به روش شبکه عصبی مصنوعی از روش‌های رگرسیون مرحله‌ای و راهبرد CDFS برای انتخاب مهم‌ترین توصیف‌گرها استفاده شد. شایان ذکر است که روش جنگل‌های تصادفی به دلیل قابلیت خود در تعیین اهمیت توصیف‌گرها می‌تواند به عنوان یک روش انتخاب متغیر نیز به کار برده شود.

۴-۳-۱- انتخاب بهترین توصیف‌گرها با روش رگرسیون مرحله‌ای

مانند بخش ۳-۳-۱، روش رگرسیون مرحله‌ای انجام شد. ۱۰ توصیف‌گر در این روش انتخاب شد که نام و طبقه‌ی آنها در جدول (۴-۲) و ماتریس هم‌بستگی بین آنها در جدول (۴-۳) ارائه شده است. نتایج جدول (۴-۳) نشان می‌دهد که بین این توصیف‌گرها هم‌بستگی معناداری وجود ندارد.

جدول (۴-۲)- توصیف‌گرهای انتخاب شده با روش رگرسیون مرحله‌ای

No	Symbol	Class	Meaning
۱	G2	Geometrical	Gravitational index G2(bond-restricted)
۲	Mor32v	3D-MoRSE	3D-MoRSE-signal 32/weighted by atomic van der waaals volumses
۳	Hy	Empirical	Hydrophilic factor
۴	ATS5e	2D autocorrelation	Broto-Morea autocorrelation of a topological structure –lag 5/weighted by atomic sanderson electronagtivities
۵	RWW	Topological	Reciprocal hyper-detour index
۶	RDF045p	RDF	Radial distribution function -4.5/weighted by atomic polarization
۷	Mor11u	3D-MoRSE	3D-MoRSE-signal 11/unweighted
۸	GATS1e	2D autocorrelation	Geary autocorrelation-lag1/weighted by atomic Sanderson electronegativities
۹	GATS5e	2D autocorrelation	Geary autocorrelation-lag5/weighted by atomic Sanderson electronegativities
۱۰	RDF100u	RDF	Radial distribution function -10.0/unweighted

جدول (۴-۳)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای

	G2	Mor32v	Hy	ATS5e	RWW	RDF045p	Mor11u	GATS1e	GATS5e	RDF100u
G2	۱									
Mor32v	۰/۵۳۰	۱								
Hy	-۰/۲۷۳	-۰/۲۱۲	۱							
ATS5e	۰/۴۲۲	-۰/۰۷۷	۰/۱۷۸	۱						
RWW	-۰/۲۵۱	-۰/۲۹۶	-۰/۰۹۰	-۰/۲۳۳	۱					
RDF045p	۰/۴۶۲	۰/۱۲۴	۰/۰۱۴	۰/۱۹۶	-۰/۱۷۳	۱				
Mor11u	۰/۵۲۳	۰/۴۸۱	-۰/۱۶۳	۰/۲۴۷	-۰/۰۳۱	۰/۲۴۲	۱			
GATS1e	۰/۳۵۵	۰/۴۲۶	-۰/۰۴۳	-۰/۰۴۸	-۰/۱۴۲	۰/۴۰۶	۰/۰۳۳	۱		
GATS5e	-۰/۲۳۱	-۰/۲۴۹	۰/۳۴۰	۰/۲۶۰	-۰/۱۶۱	-۰/۰۴۸	۰/۲۴۷	-۰/۰۶۸	۱	
RDF100u	۰/۵۸۶	۰/۲۶۹	-۰/۱۵۲	۰/۰۹۵	-۰/۰۹۱	۰/۳۰۳	۰/۳۶۳	۰/۰۷۳	-۰/۲۹۰	۱

۴-۳-۲- انتخاب بهترین توصیف‌گرها با راهبرد CDFS

برای انجام این روش طبق بخش ۳-۳-۲ عمل و در انتها ۱۶ مدل متفاوت برای سری داده‌ها ایجاد شد. توصیف‌گرهای انتخاب شده در این مدل‌ها در جدول (۴-۴) ذکر شده است که طبق این جدول، ۴ توصیف‌گر در این مدل‌ها بیش از ۵۰٪ تکرار شده‌اند که نشان دهنده‌ی اهمیت این توصیف‌گرها است. این توصیف‌گرها در جدول (۴-۵) معرفی شده‌اند و در جدول (۴-۶) ماتریس هم‌بستگی بین آنها نشان داده شده است. نتایج این جدول نشان می‌دهد که بین این توصیف‌گرها هم‌بستگی معناداری وجود ندارد.

جدول (۴-۴)- توصیف‌گرهای ۱۶ مدل به دست آمده از تقسیم‌بندی مختلف داده‌ها

شماره مدل	توصیف‌گرها	تعداد توصیف-گرها
۱	G2,Hy,Ms,RDF045p,Mor11u	۵
۲	G2, C024, RDF045m, RWW, Mor11u, Hy	۶
۳	G2, Mor32v, Hy, Ms, PiID, MATS4e, Mor20u	۷
۴	G2, GATS6v, RWW, GATS1e, RDF060p, GATS3e, MOR06e	۷
۵	G2, ATS6v, Hy, Mor16u, ATS4e, RWW	۶
۶	G2, Mor16u, Mor17m, GATS1e, GATS3e, RDF040m, RDF045m, RWW	۸
۷	G2, Mor11u, GATS1e, RWW, MAXDP, RDF045p, RDF040v	۷
۸	G2, Mor25u, Mor16u, Hy, ATS5e	۵
۹	G2, ATS6v, Hy, MAXDP, Mor16u	۵
۱۰	G2, C024, Hy, RDF045m, Mor11u	۵
۱۱	G2, Ms, Hy, RDF045p, Mor11u, RWW	۶
۱۲	G2, Mor32v, Hy, GATS3e, Mor11u, PCR, Mor08u, SPI, JGT	۹
۱۳	G2, Mor32v, Hy, MAXDP, RWW, RDF1150e, MATS1v, E1e, Mor06e, R1v	۱۰
۱۴	G2, Mor32v, Hy, RWW, ATS4e, RDF045m, Mor11u	۷
۱۵	G2, Hy, Ms, RDF045p,Mor11u, PCR, Mor15m	۷
۱۶	G2, ATS6v, Hy, ATS4e, RWW, RDF060p	۶

جدول (۴-۵)- توصیف‌گرهای انتخاب شده با راهبرد CDFS

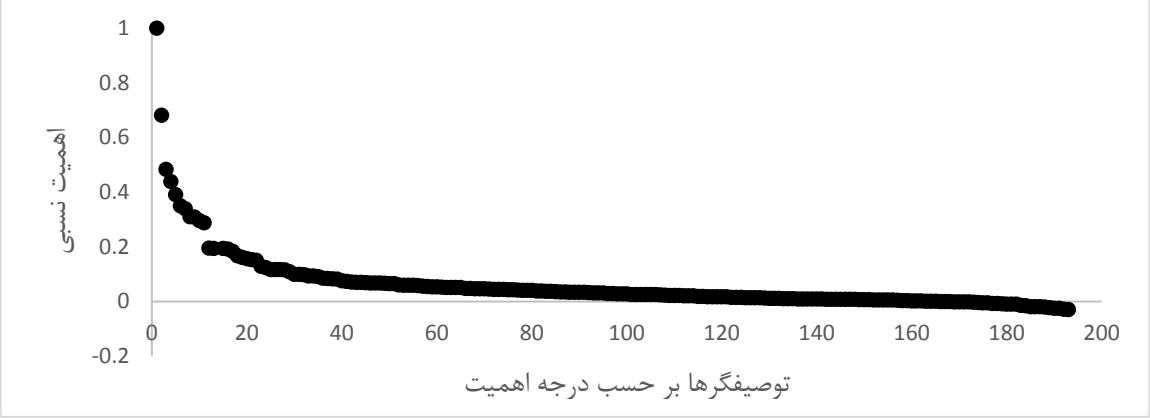
No	Symbol	Class	Meaning
1	G2	Geometrical	Gravitational index G2(bond-restricted)
2	Hy	Empirical	Hydrophilic factor
3	RWW	Topological	Reciprocal hyper-detour index
4	Mor11u	3D-MoRSE	3D-MoRSE-signal 11/unweighted

جدول (۴-۶)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط راهبرد CDFS

	G2	Hy	RWW	Mor11u
G2	۱			
Hy	-۰/۲۷۳	۱		
RWW	-۰/۲۵۱	-۰/۰۹	۱	
Mor11u	۰/۵۲۳	-۰/۱۶۳	-۰/۰۳۱	۱

۴-۳-۳- انتخاب بهترین توصیف‌گرها با روش جنگل‌های تصادفی

پس از ساخت مدل جنگل‌های تصادفی و بهینه کردن پارامترهای مؤثر بر قدرت پیش‌بینی مدل، مانند بخش (۳-۳-۳)، میزان اهمیت نسبی هر توصیف‌گر در قدرت پیش‌بینی مدل محاسبه گردید شکل (۴-۱) اهمیت نسبی این توصیف‌گرها و جدول (۴-۷) مهم‌ترین توصیف‌گرها همراه با مقدار اهمیت و طبقه‌ی آنها را نشان می‌دهد. هم‌چنین جدول (۴-۸) ماتریس هم‌بستگی بین این توصیف‌گرها را نشان می‌دهد که بیانگر عدم وجود هم‌بستگی معنادار بین این توصیف‌گرها است.



شکل (۴-۱)- نمودار اهمیت نسبی توصیف‌گرها

جدول (۴-۷)- توصیف‌گرهای انتخاب شده با روش جنگل‌های تصادفی

Rank	Symbol	Class	Meaning	Importance	Relative Importance
1	G2	Geometrical	Gravitational index G2 (bond restricted)	20.98	1
2	piID	Topological	Conventional bond-order ID number	13.81	0.66
3	SPI	Topological	Superpendentic index	8.74	0.42
4	Hy	Empirical	Hydrophilic factor	8.58	0.41
5	BELe8	BCUT	Lowest eigenvalue n.8 of Burden matrix/weighted by atomic sanderson electronegativities	8.24	0.39
6	BELe7	BCUT	Lowest eigenvalue n.7 of Burden matrix/weighted by atomic sanderson electronegativities	6.58	0.35
7	BELp8	BCUT	Lowest eigenvalue n.8 of Burden matrix/weighted by atomic polarizabilities	6.37	0.34
8	GGI5	Galves topol.charg indices	Topological charg index of order 5	5.84	0.31
9	BELe6	BCUT	Lowest eigenvalue n.6 of Burden matrix/weighted by atomic sanderson electronegativities	5.81	0.30
10	PCR	Topological	Ratio of multiple path counts to path counts	5.56	0.29
11	BEHm1	BCUT	Highest eigenvalue n.1 of Burden metrix/weighted by atomic masses	5.40	0.28

جدول (۴-۸)- ماتریس هم‌بستگی توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی

	G2	PiID	SPI	Hy	BELe8	BELe7	BELp8	GGI5	BELe6	PCR	BEHm1
G2	۱										
piID	۰/۶۹۹	۱									
SPI	۰/۱۰۰	-۰/۰۸۷	۱								
Hy	-۰/۲۷۳	-۰/۱۷۱	-۰/۰۵۰	۱							
BELe8	۰/۶۳۸	۰/۵۵۱	-۰/۰۴۳	-۰/۳۷۲	۱						
BELe7	۰/۶۱۵	۰/۵۵۸	۰/۰۱۵	-۰/۳۲۷	۰/۶۰۸	۱					
BELp8	۰/۶۱۹	۰/۵۰۳	۰/۰۱۷	-۰/۳۴۴	۰/۶۶۹	۰/۶۵۸	۱				
GGI5	۰/۵۸۱	۰/۷۶۹	۰/۰۵۱	-۰/۲۵۱	۰/۵۱۹	۰/۷۸۸	۰/۷۸۹	۱			
BELe6	۰/۶۱۶	۰/۵۸۵	۰/۰۳۳	-۰/۲۶۶	۰/۵۷۹	۰/۶۰۱	۰/۶۰۰	۰/۵۲۱	۱		
PCR	۰/۷۵۴	۰/۵۵۷	-۰/۰۸۵	-۰/۲۳۱	۰/۶۵۷	۰/۶۴۵	۰/۶۱۹	۰/۷۲۶	۰/۶۶۸	۱	
BEHm1	۰/۶۰۹	۰/۴۵۶	-۰/۰۹۹	-۰/۰۷۴	۰/۴۲۹	۰/۴۴۱	۰/۴۱۲	۰/۵۳۷	۰/۴۷۲	۰/۴۶۰	۱

۴-۴- مدل‌سازی با رگرسیون خطی چندگانه

برای به دست آوردن بهترین مدل خطی، R^2_{adj} و FIT برای مدل‌های مختلف با تعداد متفاوت توصیف‌گر که با استفاده از داده‌های سری آموزش و توصیف‌گرهای ارائه شده در جدول (۴-۲) به دست آمده بود، محاسبه و سپس مدلی که دارای بیشترین مقدار این پارامترها برای داده‌های سری ارزیابی بود، به عنوان مدل نهایی انتخاب شد. مقادیر این پارامترها برای مدل‌های مختلف در جدول (۴-۹) و روابط مربوط به آنها در بخش ۲-۹ ارائه شده است.

جدول (۴-۹)- پارامترهای آماری برای مدل‌های به دست آمده از روش MLR

مدل	توصیف‌گرها	FIT	R^2_{adj}
۱	G2, Mor32v	۶۴/۷۸	۰/۹۸۸
۲	G2, Mor32v, Hy	۸۲/۵۵	۰/۹۹۳
۳	G2, Mor32v, Hy, ATS5e	۷۸/۱۶	۰/۹۹۴
۴	G2, Mor32v, Hy, ATS5e, RWW	۶۱/۵۳	۰/۹۹۴
۵	**G2, Mor32v, Hy, ATS5e, RWW, RDF045p**	**۲۰۵/۳۲**	**۰/۹۹۹**
۶	G2, Mor32v, Hy, ATS5e, RWW, RDF045p, Mor11u	۱۳۲/۷۳	۰/۹۹۸
۷	G2, Mor32v, Hy, ATS5e, RWW, RDF045p, Mor11u, GATS1e	۳۷/۱۲	۰/۹۹۵
۸	G2, Mor32v, Hy, ATS5e, RWW, RDF045p, Mor11u, GATS1e, GATS5e	۲۳/۹۲	۰/۹۹۳
۹	G2, Mor32v, Hy, ATS5e, RWW, RDF045p, Mor11u, GATS1e, GATS5e, RDF100u	۱۲/۴۱	۰/۹۸۹

بدین ترتیب مدل ۵ به عنوان بهترین مدل برگزیده شد که معادله آن به صورت زیر است:

RI=784.517+37.782 G2+12.749 Mor32v+63.581 Hy-790.117 ATS5e-3.316 RWW-1.592 RDF045p (۱-۴)

بر اساس روابط (۳-۲) تا (۳-۴)، ضرایب رگرسیون استاندارد شده برای مدل برتر محاسبه شد که در جدول زیر ارائه شده است:

جدول (۴-۱۰)- ضرایب غیراستاندارد و استاندارد توصیف‌گرهای موجود در مدل خطی برتر

توصیف‌گر	G2	Mor32v	Hy	ATS5e	RWW	RDF045p
ضریب غیر استاندارد	۳۷/۷۸۲	۱۲/۴۷۹	۶۳/۵۸۱	-۷۹۰/۱۱۷	-۳/۳۱۶	-۱/۵۹۲
ضریب استاندارد	۱/۰۲۴	۰/۰۲۸	۰/۰۷۶	۰/۰۸۷	۰/۰۲۶	۰/۰۲۶

۴-۵- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای (SR-ANN)

برای یافتن رابطه غیرخطی بین متغیرهای مستقل و متغیر وابسته و ایجاد یک شبکه عصبی مصنوعی، ابتدا مقادیر ۱۰ توصیف‌گر انتخاب شده با روش رگرسیون مرحله‌ای به عنوان توصیف‌گر ورودی و اندیس بازداری متناظر با آنها به عنوان متغیر هدف در نظر گرفته شدند و یک شبکه پیشخور با تکنیک پس-انتشار ایجاد شد. پارامترهای مؤثر بر قدرت پیش‌بینی شبکه شامل تعداد توصیف‌گرهای ورودی، تعداد گره در لایه پنهان، نوع تابع آموزش و انتقال، تعداد دور آموزش (epoch) و پارامتر μ بهینه شدند. تابع کارایی شبکه نیز میانگین مربع خطای استاندارد (MSE) در نظر گرفته شد.

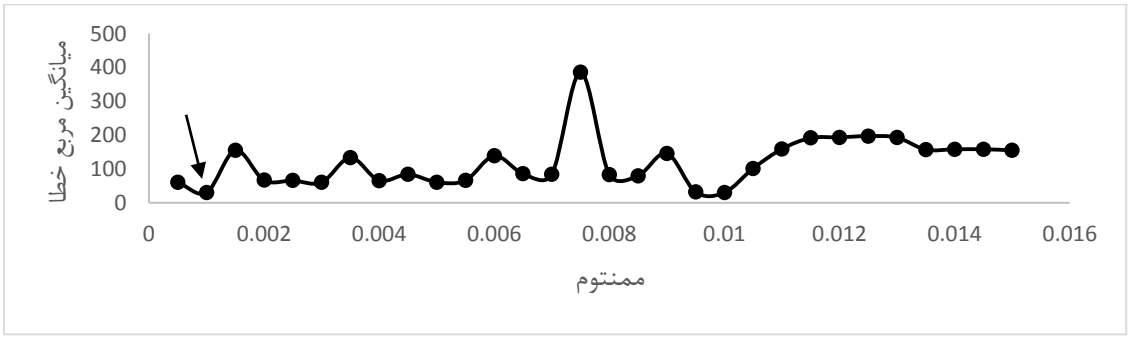
۴-۵-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

از آنجایی که در بیشتر موارد به نظر می‌رسد که یک لایه پنهان در ساختار شبکه عصبی مناسب باشد، در این تحقیق نیز یک لایه پنهان مورد استفاده قرار گرفت[۴۴] و هم‌چنین در لایه خروجی از تابع انتقال خطی (purelin) استفاده شد، اما مقدار سایر پارامترها بهینه شد. برای این کار، هر شبکه با تعداد ورودی از ۲ تا ۱۰ و با دو الگوریتم آموزشی لونبرگ-مارکوارت (trainlm) و تنظیم بایزین (trainbr) و دو تابع انتقال لگاریتم سیگموئید (logsig) و تانژانت سیگموئید (tansig) و تعداد گره از ۲ تا ۱۰ و تعداد دور آموزش از ۲ تا ۴۰، آموزش داده شد. تنها مقدار پارامتر ممنتوم ثابت در نظر گرفته شد که در مرحله بعد این پارامتر نیز بهینه گردید. در جدول (۴-۱۱) بهترین شبکه‌های به دست آمده در حین بهینه‌سازی پارامترهای مختلف بر اساس کمترین مقدار MSE نشان داده شده‌اند. با توجه به این جدول، شبکه عصبی ۳ لایه با تابع آموزشی لونبرگ مارکوارت و تابع انتقال لگاریتم سیگموئیدی، ۴ توصیف‌گر ورودی، ۵ نرون در لایه مخفی و تعداد دور آموزشی ۱۰ کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی اندیس بازداری ترکیبات در نظر گرفته شد.

جدول (۴-۱۱)- توابع و پارامترهای شبکه‌های بهینه (SR-ANN) به دست آمده

MSE	تعداد دورهای آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۳۰/۰۹	**۱۰**	**۵**	**لگاریتم-سیگموئید**	**لونبرگ-مارکوارت**	**۴**
۳۱/۱۷	۸	۴	تانزانت-سیگموئید	لونبرگ-مارکوارت	۵
۴۹/۱۱	۱۴	۵	تانزانت-سیگموئید	تنظیم بایزین	۶
۵۲/۸۷	۱۵	۸	لگاریتم-سیگموئید	تنظیم بایزین	۶

به منظور بهینه کردن مقدار پارامتر μ ، در شایط بهینه به دست آمده از مرحله‌ی قبل، مقدار این پارامتر، بین ۰/۰۰۰۵ تا ۰/۰۱۵ با گام های ۰/۰۰۰۵ تغییر داده شد و برای هر مورد مقدار MSE سری ارزیابی محاسبه گردید. در این تحقیق مقدار خطا مطابق با شکل(۴-۲) در μ=۰/۰۰۱ که همان مقدار پیش‌فرض در تابع آموزشی لونبرگ مارکوارت می‌باشد، به حداقل رسید.



شکل (۴-۲)- نمودار تغییر خطای شبکه (SR-ANN) با تغییر پارامتر μ

۴-۵-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده، مشخصات شبکه بهینه در جدول (۴-۱۲) نشان داده شده است.

جدول (۴-۱۲)- مشخصات شبکه بهینه (SR-ANN)

تابع آموزش	لونبرگ-مارکوارت (trainlm)
تابع انتقال لایه پنهان	لگاریتم سیگموئید (logsig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۵
تعداد توصیف‌گر ورودی	۴
تعداد دور آموزشی	۱۰
پارامتر μ	۰/۰۰۱

۴-۶- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط راهبرد CDFS (CDFS-ANN)

در این بخش توصیف‌گرهای منتخب راهبرد CDFS در یک شبکه عصبی با یک لایه پنهان و لایه خروجی purlin استفاده گردید. برای یافتن مقدار بهینه پارامترهای مؤثر بر کارایی شبکه، تعداد ورودی ها از ۲ تا ۴ و سایر پارامترها مانند مرحله قبل تغییر داده شد و هم‌چنین تابع کارایی، میانگین مربع خطا در نظر گرفته شد.

۴-۶-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

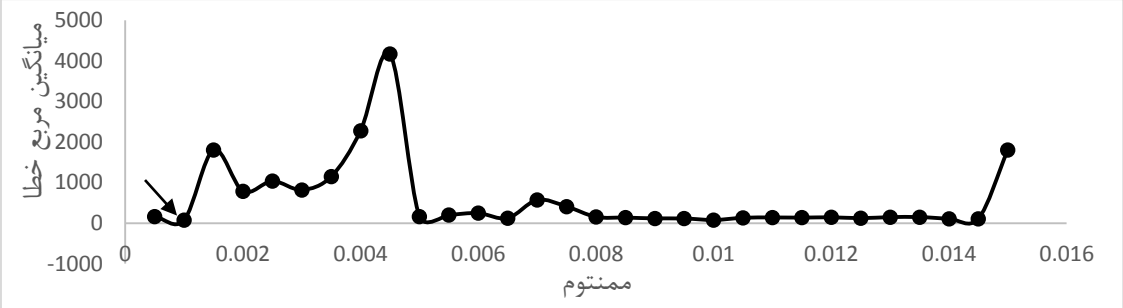
در این مرحله تمامی پارامترها به غیر از ممنتوم به طور هم‌زمان بهینه گردید و در مرحله بعد مقدار ممنتوم نیز برای یافتن مقدار بهینه آن تغییر داده شد. با توجه به جدول (۴-۱۳) که بهترین شبکه‌های به دست آمده در حین بهینه‌سازی را نمایش می‌دهد، مشاهده می‌شود که شبکه عصبی با ۳ توصیف‌گر ورودی، ۸ نرون در لایه مخفی، تعداد دور آموزشی ۳۹، تابع آموزشی لونبرگ-مارکوارت و تابع انتقال لگاریتم سیگموئید، کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین شبکه‌ای با این ساختار به عنوان بهترین شبکه برای مدل‌سازی اندیس بازداری ترکیبات در نظر گرفته شد.

جدول (۴-۱۳)- توابع و پارامترهای شبکه‌های بهینه (CDFS-ANN) به دست آمده

MSE	تعداد دورهای آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۷۹/۲۰	**۳۹**	**۸**	**لگاریتم-سیگموئید**	**لونبرگ-مارکوارت**	**۳**
۹۷/۴۹	۳۶	۵	تانژانت-سیگموئید	لونبرگ-مارکوارت	۴
۱۱۷/۱۲	۴۰	۴	لگاریتم-سیگموئید	تنظیم بایزین	۴
۱۱۸/۲۲	۷	۹	تانژانت-سیگموئید	تنظیم بایزین	۴

برای بهینه کردن پارامتر ممنتوم، به ازای تنظیمات مرحله‌ی قبل، μ از ۰/۰۰۰۵ تا ۰/۰۱۵ با گام‌های ۰/۰۰۰۵ تغییر داده شد و در هر تغییر میانگین مربع خطای سری ارزیابی محاسبه گردید. با توجه به

نتایج شکل (۴-۳)، مقدار بهینه μ برابر با مقدار پیش‌فرض آن در تابع آموزشی لونبرگ-مارکوارت یعنی ۰/۰۰۱ به دست آمد.



شکل (۴-۳)- نمودار تغییر خطای شبکه (CDFS-ANN) با تغییر پارامتر μ

۴-۶-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده از مرحله‌ی قبل، مشخصات شبکه عصبی بهینه در جدول (۴-۱۴) نشان داده شده است.

جدول (۴-۱۴)- مشخصات شبکه بهینه (CDFS-ANN)

تابع آموزش	لونبرگ-مارکوارت (trainlm)
تابع انتقال لایه پنهان	تانژانت سیگموئید (logsig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۸
تعداد توصیف‌گر ورودی	۳
تعداد دور آموزشی	۳۹
پارامتر μ	۰/۰۰۱

۴-۷- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی (RF-ANN)

در این بخش نیز از یک شبکه عصبی با یک لایه پنهان و لایه خروجی purlin و توصیف‌گرهای با اهمیت در روش جنگل‌های تصادفی (جدول (۴-۷)) استفاده شد. برای یافتن مقادیر بهینه پارامترهای

مؤثر بر کارایی شبکه، تعداد ورودی ها از ۲ تا ۱۱ و سایر پارامترها مانند مرحله قبل تغییر داده شده و تابع کارایی نیز میانگین مربع خطا در نظر گرفته شد.

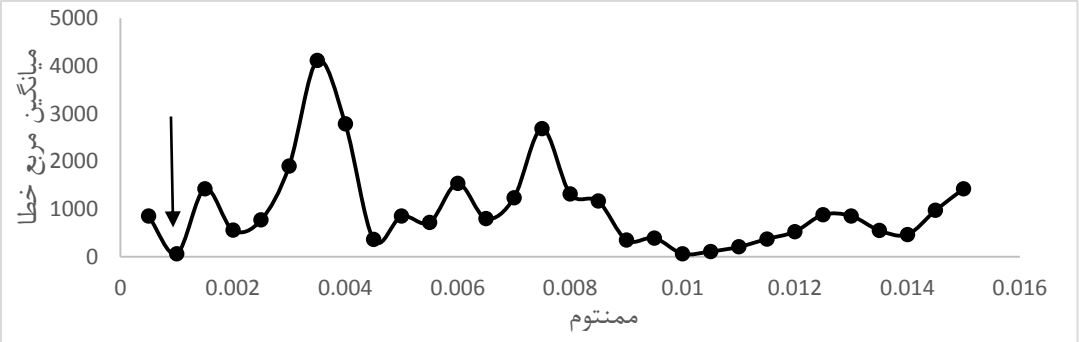
۴-۷-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

در این مرحله نیز تمامی پارامترها به طور هم‌زمان بهینه و تنها مقدار ممنتوم ثابت قرار داده شد که در مرحله بعد مقدار آن نیز بهینه گردید. جدول (۴-۱۵) بهترین شبکه‌های به دست آمده حاصل از بهینه‌سازی را نمایش می‌دهد. نتایج این جدول نشان می‌دهد که شبکه عصبی با ۱۱ توصیف‌گر ورودی، ۷ نرون در لایه مخفی، تعداد دور آموزشی ۴۰، تابع آموزشی لونبرگ-مارکوارت و تابع انتقال تانژانت سیگموئید، کمترین MSE را نسبت به سایر شبکه‌ها نشان می‌دهد. بنابراین شبکه‌ای با این ساختار به عنوان بهترین شبکه برای مدل‌سازی اندیس بازداری ترکیبات در نظر گرفته شد.

جدول (۴-۱۵)- توابع و پارامترهای شبکه‌های بهینه (RF-ANN) به دست آمده

MSE	تعداد دورهای آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف‌گر
۷۵/۶۵	۱۲	۳	لگاریتم-سیگموئید	لونبرگ-مارکوارت	۱۱
۶۴/۱۵	**۴۰**	**۷**	**تانژانت-سیگموئید**	**لونبرگ-مارکوارت**	**۱۱**
۶۵/۳۸	۳۹	۷	تانژانت-سیگموئید	تنظیم بایزین	۱۱
۶۵/۷۳	۳۵	۵	لگاریتم-سیگموئید	تنظیم بایزین	۱۱

برای بهینه کردن پارامتر ممنتوم، شبکه بهینه به دست آمده از مرحله‌ی قبل، طراحی و مقدار μ از ۰/۰۰۰۵ تا ۰/۰۱۵ با گام های ۰/۰۰۰۵ تغییر داده شد و در هر تغییر میانگین مربع خطای سری ارزیابی محاسبه گردید. طبق شکل (۴-۴)، مقدار بهینه μ برابر با مقدار پیش‌فرض آن در تابع آموزشی لونبرگ-مارکوارت یعنی ۰/۰۰۱ به دست آمد.



شکل (۴-۴)- نمودار تغییر خطای شبکه (RF-ANN) با تغییر پارامتر μ

۴-۷-۲- ساختار شبکه عصبی مصنوعی بهینه شده

با توجه به نتایج به دست آمده از مرحله‌ی قبل، مشخصات شبکه عصبی بهینه در جدول (۴-۱۶) نشان داده شده است.

جدول (۴-۱۶)- مشخصات شبکه بهینه (RF-ANN)

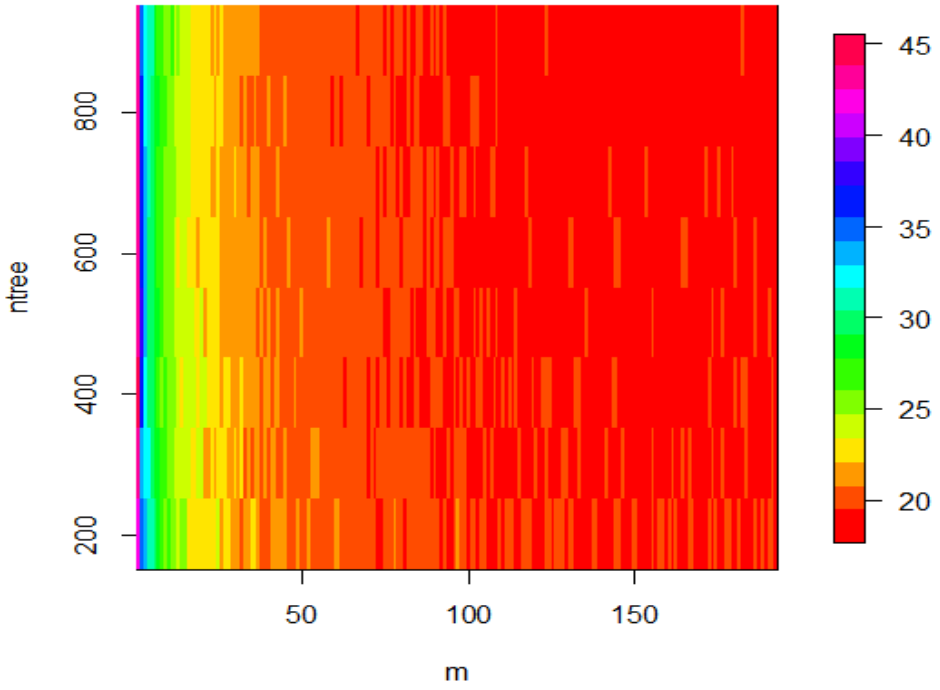
تابع آموزش	لونبرگ-مارکوارت (trainlm)
تابع انتقال لایه پنهان	تانژانت سیگموئید (tansig)
تابع انتقال لایه خروجی	purelin
تابع کارایی	MSE
تعداد گره	۷
تعداد توصیف‌گر ورودی	۱۱
تعداد دور آموزشی	۴۰
پارامتر μ	۰/۰۰۱

۴-۸- مدل‌سازی جنگل‌های تصادفی

برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده شد. در اینجا نیز تعداد درختان (ntree) وتعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (m)، برای دستیابی به بهترین مدل بهینه گردید.

۴-۸-۱- بهینه‌سازی مقادیر ntree و m

برای بهینه نمودن تعداد درخت و تعداد توصیفگر مورد استفاده در هر مرحله افراز، ntree از ۲۰۰ تا ۹۰۰ با گام‌های ۱۰۰ و m از ۲ تا ۱۹۳ با گام‌های ۱ تغییر داده شده و در هر مرحله مجذور خطای OOB یا RMSE محاسبه گردید. نمودار ارائه شده شکل (۴-۵)، روند تغییر خطا در مقادیر متفاوت ntree و m را نشان می‌دهد. نقاط زیادی با کمترین مقدار خطا در قسمت قرمز رنگ شکل (۴-۵) که وجود دارند که همراه با مقادیر ntree و m متناظر خود، در جدول(۴-۱۷) نشان داده شده‌اند.



شکل (۴-۵)- نمودار تغییرات مجذور خطای OOB در مقادیر متفاوت m و ntree

جدول (۴-۱۷)- کمترین مقادیر RMSE همراه با m و ntree متناظر آنها

RMSE(OOB)	۱۷/۷۴	۱۷/۷۷	۱۷/۸۳	**۱۷/۸۷**	۱۷/۹۲	۱۷/۹۲	۱۷/۹۹
m	۱۵۲	۱۴۴	۱۲۳	**۷۵**	۱۸۰	۱۴۸	۱۴۸
ntree	۲۰۰	۲۰۰	۲۰۰	**۵۰۰**	۶۰۰	۴۰۰	۳۰۰

مانند بخش ۳-۸-۱، چون هدف در این روش کاهش تعداد توصیف‌گرها است، m=75 و ntree=500 به عنوان مقادیر بهینه انتخاب شدند.

۴-۸-۲- مدل‌سازی جنگل‌های تصادفی با استفاده از توصیف‌گرهای مهم

در این مرحله، روش جنگل‌های تصادفی فقط با توصیف‌گرهای مهم (جدول (۴-۷)) مجددا انجام شد. به این صورت که ابتدا از دو توصیف‌گر به عنوان ورودی استفاده گردید و برای یافتن بهترین مدل، تعداد درخت برابر با ۵۰۰ قرار داده شد. سپس تعداد توصیف‌گرهای به کار رفته در هر مرحله از افراز تغییر داده شده و خطای OOB محاسبه گردید تا مقدار بهینه m به دست بیاید. در مرحله بعد، سه توصیف‌گر وارد شده و این مراحل تکرار شد[۳۶]. جدول (۴-۱۸) نتایج این کار را نشان می‌دهد و مشاهده می‌شود که تعداد ورودی ۲ با تعداد درخت ۵۰۰ و تعداد m برابر با ۲ کمترین میزان خطا را نشان می‌دهد.

جدول (۴-۱۸)- تغییرات خطای OOB با تعداد توصیف‌گر متفاوت

تعداد توصیف‌گر ورودی	دامنه تغییر m	تعداد بهینه m	MSE(OOB)
۲	**۲**	**۲**	**۲۴۶/۳۵**
۳	۲-۳	۳	۲۶۷/۰۶
۴	۲-۴	۳	۲۵۷/۰۸
۵	۲-۵	۵	۲۷۱/۵۳
۶	۲-۶	۴	۲۶۸/۰۸
۷	۲-۷	۴	۲۷۵/۶۶
۸	۲-۸	۳	۳۰۶/۲۷
۹	۲-۹	۶	۳۱۱/۰۳
۱۰	۲-۱۰	۵	۲۵۹/۶۹
۱۱	۲-۱۱	۶	۲۷۶/۱۳

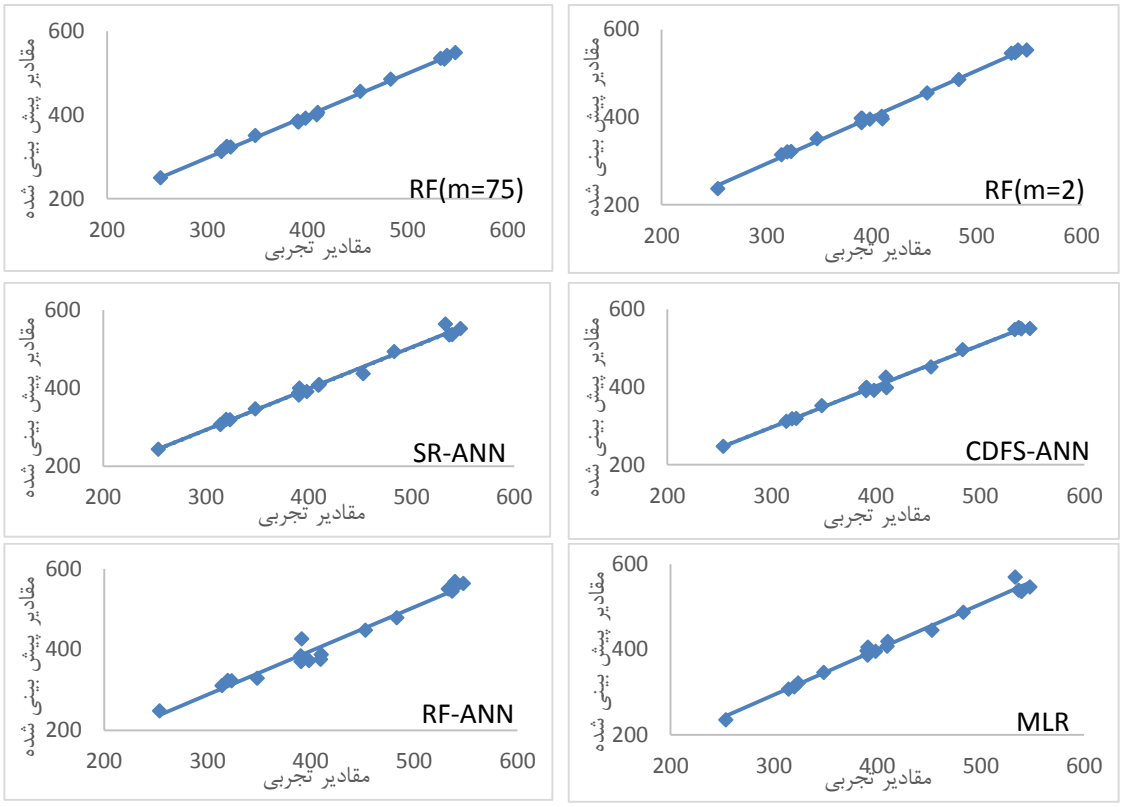
۴-۹- ارزیابی مدل‌ها

۴-۹-۱- ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون

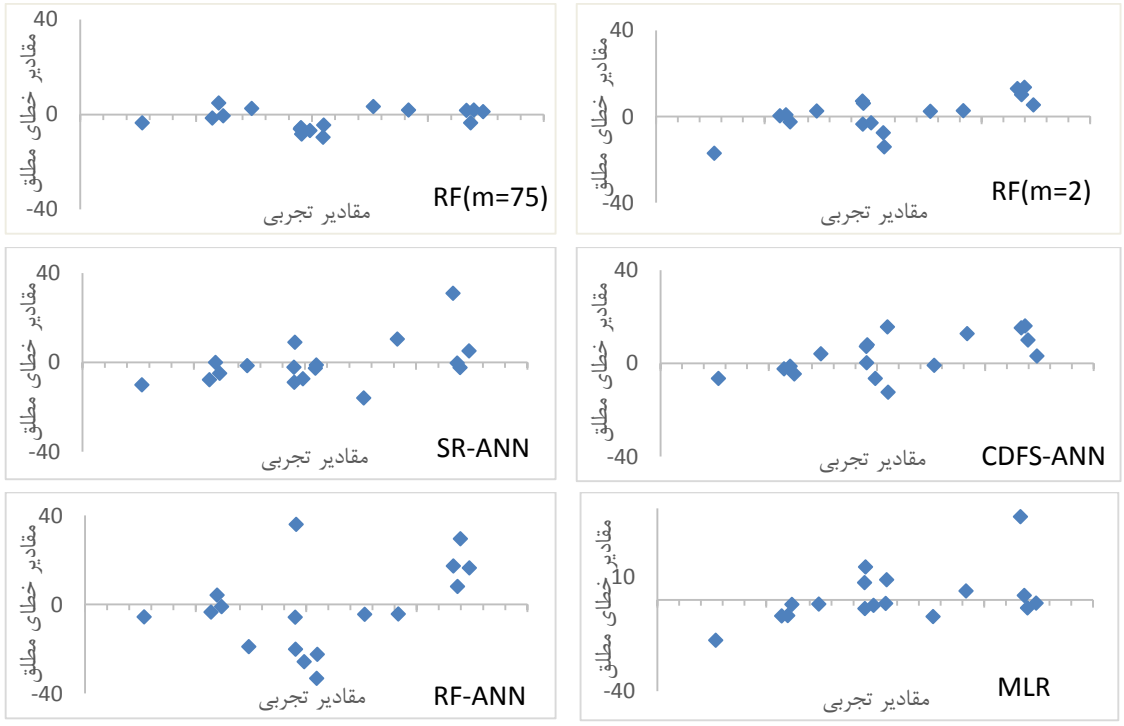
جدول (۴-۱۹) نتایج پیش‌بینی و شکل (۴-۶) مقدار پیش‌بینی شده توسط مدل‌های به کار برده شده در این تحقیق بر حسب مقدار تجربی برای داده‌های سری آزمون را ارائه می‌دهد که نشان می‌دهد روش جنگل‌های تصادفی بهتر از سایر روش‌ها عمل نموده است. هم‌چنین در شکل (۴-۷) مقادیر باقی مانده‌ها بر حسب مقادیر تجربی RI ترکیبات مورد بررسی ترسیم شده است که توزیع متقارن داده‌ها حول محور افقی (خطای صفر) عدم وجود خطای سیستماتیک را نشان می‌دهد.

جدول (۴-۱۹)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده						درصد خطا					
		RF (m=75)	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=75)	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۶	۲۵۳/۵۶	۲۵۰/۰۶	۲۳۶/۴۷	۲۴۳/۵۱	۲۴۷/۱۷	۲۴۷/۹۸	۲۳۵/۸۵	-۱/۳۸	-۶/۷۴	-۳/۹۶	-۲/۵۲	-۲/۲۰	-۶/۹۸
۱۸	۳۱۴/۱۵	۳۱۲/۵۵	۳۱۴/۴۴	۳۰۶/۳۲	۳۱۱/۷۷	۳۱۰/۸۰	۳۰۷/۲۰	-۰/۵۱	۰/۰۹	-۲/۴۹	-۰/۷۶	-۱/۰۷	۲/۲۱
۲۰	۳۱۹/۶۷	۳۲۴/۴۳	۳۲۰/۴۹	۳۱۹/۶۷	۳۱۸/۴۷	۳۲۳/۹۰	۳۱۲/۸۴	۱/۴۹	۰/۲۶	۰/۰۰	-۰/۳۸	۱/۳۲	-۲/۱۴
۲۳	۳۲۳/۵۴	۳۲۳/۰۰	۳۲۱/۰۹	۳۱۸/۵۴	۳۱۹/۰۲	۳۲۲/۵۸	۳۲۱/۶۳	-۰/۱۷	-۰/۷۶	-۱/۵۴	-۱/۴۰	-۰/۳۰	-۰/۵۹
۲۷	۳۴۸/۱۴	۳۵۰/۵۸	۳۵۰/۷۲	۳۴۶/۶۹	۳۵۲/۲۶	۳۲۹/۱۶	۳۴۶/۲۵	۰/۷۰	۰/۷۴	-۰/۴۲	۱/۱۸	-۵/۴۵	-۰/۵۴
۳۹	۳۹۰/۱۸	۳۸۴/۰۱	۳۹۷/۳۵	۳۸۷/۹۳	۳۹۷/۴۶	۳۸۴/۵۲	۳۹۷/۷۸	-۱/۵۸	۱/۸۴	-۰/۵۸	۱/۸۷	-۱/۴۵	۱/۹۵
۴۰	۳۹۰/۵۷	۳۸۵/۰۴	۳۸۶/۹۸	۳۸۱/۵۶	۳۹۰/۹۲	۳۷۰/۴۶	۳۸۶/۷۷	-۱/۴۲	-۰/۹۲	-۲/۳۱	۰/۰۹	-۵/۱۵	-۰/۹۷
۴۲	۳۹۱/۱۰	۳۸۲/۸۰	۳۹۷/۱۶	۴۰۰/۱۱	۳۹۹/۱۳	۴۲۷/۱۱	۴۰۵/۴۷	-۲/۱۲	۱/۵۵	۲/۳۰	۲/۰۵	۹/۲۰	۳/۶۷
۴۵	۳۹۸/۳۶	۳۹۱/۶۲	۳۹۵/۴۴	۳۹۰/۹۶	۳۹۱/۸۵	۳۷۲/۶۱	۳۹۵/۹۷	-۱/۶۹	-۰/۷۳	-۱/۸۶	-۱/۶۳	-۶/۴۶	-۰/۶۰
۵۰	۴۰۹/۵۴	۳۹۹/۹۶	۴۰۱/۹۲	۴۰۶/۹۴	۴۲۵/۲۹	۳۷۶/۲۹	۴۰۸/۰۵	-۲/۳۴	-۱/۸۶	-۰/۶۳	۳/۸۵	-۸/۱۲	-۰/۳۶
۵۱	۴۱۰/۲۷	۴۰۵/۷۸	۳۹۶/۰۹	۴۰۹/۱۳	۳۹۷/۸۸	۳۸۷/۸۲	۴۱۹/۰۳	-۱/۰۹	-۳/۴۶	-۰/۲۸	-۳/۰۲	-۵/۴۷	۲/۱۳
۵۶	۴۵۲/۹۳	۴۵۶/۲۸	۴۵۵/۳۴	۴۳۶/۹۸	۴۵۲/۱۲	۴۴۸/۵۵	۴۴۵/۶۴	۰/۷۴	۰/۵۳	-۳/۵۲	-۰/۱۸	-۰/۹۷	-۱/۶۱
۶۱	۴۸۳/۲۷	۴۸۵/۱۱	۴۸۵/۹۶	۴۹۳/۶۶	۴۹۶/۰۸	۴۷۸/۹۹	۴۸۷/۱۴	۰/۳۸	۰/۵۶	۲/۱۵	۲/۶۵	-۰/۸۹	۰/۸۰
۷۲	۵۳۳/۲۲	۵۳۴/۹۷	۵۴۶/۲۲	۵۶۴/۲۱	۵۴۸/۵۳	۵۵۰/۵۵	۵۶۹/۵۷	۰/۳۳	۲/۴۴	۵/۸۱	۲/۸۷	۳/۲۵	۶/۸۲
۷۳	۵۳۶/۹۲	۵۳۳/۴۱	۵۴۷/۰۰	۵۳۶/۵۳	۵۵۳/۰۰	۵۴۵/۰۹	۵۳۸/۸۶	-۰/۶۵	۱/۸۸	-۰/۰۷	۲/۹۹	۱/۵۲	۰/۳۶
۷۴	۵۳۹/۵۹	۵۴۱/۴۸	۵۵۳/۱۵	۵۳۷/۲۴	۵۴۹/۶۹	۵۶۹/۱۰	۵۳۶/۰۷	۰/۳۵	۲/۵۱	-۰/۴۴	۱/۸۷	۵/۴۷	-۰/۶۵
۷۷	۵۴۷/۷۱	۵۴۸/۸۱	۵۵۳/۰۶	۵۵۲/۸۲	۵۵۰/۹۴	۵۶۴/۱۳	۵۴۶/۲۷	۰/۲۰	۰/۹۸	۰/۹۳	۰/۵۹	۳/۰۰	-۰/۲۶



شکل (۴-۶)- نمودار مقادیر پیش‌بینی شده اندیس بازداری داده‌های سری آزمون توسط مدل‌های مختلف در مقابل مقادیر تجربی



شکل (۴-۷)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای داده‌های سری آزمون با مدل‌های مختلف

۴-۹-۲- ارزیابی مدل‌ها توسط اعتبارسنجی متقاطع با رد مرحله‌ای تک تک

در این روش هر بار یک ترکیب از سری داده‌ها کنار گذاشته شده و مدل‌سازی با ۸۲ ترکیب باقی مانده با تمامی مدل‌های ذکر شده، صورت گرفت. سپس مدل‌های به دست آمده برای پیش‌بینی اندیس بازداری ترکیب کنار گذاشته شده به کار گرفته شد که این فرایند برای تمام ترکیبات تکرار گردید. نتایج حاصل از این روش در جدول (۴-۲۰) ارائه شده است. ضرایب تعیین به دست آمده از شکل (۴-۸) و خطاهای مطلق متناظر نشان داده شده در شکل (۴-۹) بر توانایی این مدل‌ها برای پیش‌بینی اندیس بازداری این ترکیبات و عدم وجود خطای سیستماتیک دلالت دارد.

۴-۹-۳- ارزیابی مدل‌ها با استفاده از پارامترهای آماری

ده پارامتر آماری که توضیح آنها در بخش ۲-۹ ارائه شده است، جهت ارزیابی توانایی پیش‌بینی مدل‌های به کار گرفته شده، برای داده‌های سری آزمون محاسبه شد. مقدار این پارامترها که برتری مدل جنگل‌های تصادفی را نسبت به سایر روش‌ها نشان می‌دهد، در جدول (۴-۲۱) و مقدار این پارامترها برای کل داده‌ها توسط مدل‌های مختلف در جدول (۴-۲۲) نشان داده شده است.

۴-۹-۴- ارزیابی مدل‌ها با استفاده از آزمون Y-تصادفی

در این آزمون، مقادیر تجربی متغیر وابسته در محدوده همان مقادیر به طور تصادفی تغییر داده شدند و مدل‌های جدید با تمام روش‌های ذکر شده با استفاده از این مقادیر تصادفی ایجاد و ضریب تعیین داده‌های آزمون محاسبه شد. نتایج حاصل از ۱۰ بار اجرای این آزمون در جدول (۴-۲۳) نشان داده شده است. تفاوت قابل ملاحظه بین مقادیر ضریب تعیین مدل‌های اصلی و مدل‌های جدید، بیانگر عدم وجود ارتباط تصادفی در مدل‌ها است.

جدول (۴-۲۰)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از روش رد مرحله‌ای تک تک

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۱	۲۰۰/۰۰	۲۱۲/۷۶	۲۳۳/۲۰	۱۶۸/۱۰	۲۱۳/۸۰	۲۱۴/۹۱	۶/۳۸	۱۶/۶۰	-۱۵/۶۵	۶/۹۰	۷/۴۵
۲	۲۲۱/۱۰	۲۳۰/۱۲	۲۳۰/۴۴	۲۲۸/۰۸	۲۲۴/۱۴	۲۳۰/۶۸	۴/۰۸	۴/۲۲	۳/۱۶	۱/۳۸	۴/۳۳
۳	۲۲۴/۱۲	۲۲۸/۹۴	۲۲۷/۶۸	۲۲۵/۹۵	۲۲۱/۸۶	۲۳۰/۰۴	۲/۱۵	۱/۵۹	۰/۸۲	-۱/۰۱	۲/۶۴
۴	۲۳۶/۵۶	۲۳۶/۰۳	۲۴۷/۹۷	۲۵۶/۴۹	۲۶۷/۱۵	۲۳۰/۸۹	-۰/۲۲	۴/۸۲	۸/۴۲	۱۲/۹۳	-۲/۴۰
۵	۲۴۸/۲۷	۲۴۳/۷۸	۲۲۰/۶۴	۲۴۹/۶۳	۲۵۶/۹۴	۲۴۰/۲۴	-۱/۸۱	-۱۱/۱۳	۰/۵۵	۳/۴۹	-۳/۲۳
۶	۲۵۳/۵۶	۲۳۷/۳۶	۲۴۵/۶۲	۲۴۲/۷۹	۲۴۷/۴۶	۲۳۵/۲۸	-۶/۳۹	-۳/۱۳	-۴/۲۵	-۲/۰۴	-۷/۲۱
۷	۲۵۹/۶۳	۲۹۳/۲۷	۲۷۷/۱۳	۲۷۷/۶۱	۲۶۴/۴۶	۲۷۹/۹۷	۱۲/۹۶	۶/۷۴	۶/۹۲	۱/۸۶	۷/۸۳
۸	۲۷۰/۳۹	۲۹۶/۷۱	۲۷۳/۳۰	۲۶۳/۸۶	۲۷۵/۳۲	۲۷۵/۲۹	۹/۷۳	۱/۰۸	-۲/۴۲	۱/۸۲	۱/۸۱
۹	۲۸۸/۲۸	۲۹۹/۵۶	۲۹۵/۱۳	۲۹۶/۳۶	۲۸۶/۳۱	۲۹۲/۱۹	۳/۹۱	۲/۳۸	۲/۸۰	-۰/۶۸	۱/۳۶
۱۰	۲۸۹/۲۴	۲۹۸/۲۳	۲۹۱/۹۳	۲۹۶/۳۱	۲۹۱/۵۹	۲۸۹/۴۴	۳/۱۱	۰/۹۳	۲/۴۵	۰/۸۱	۰/۰۷
۱۱	۲۹۲/۲۸	۲۹۶/۳۴	۲۸۹/۱۵	۲۹۴/۲۵	۲۸۸/۳۸	۲۸۷/۴۵	۱/۳۹	-۱/۰۷	۰/۶۷	-۱/۳۳	-۱/۶۵
۱۲	۲۹۵/۹۵	۳۰۵/۸۸	۳۰۰/۳۸	۲۹۳/۲۸	۳۰۱/۱۷	۲۹۰/۲۸	۳/۳۶	۱/۵۰	-۰/۹۰	۱/۷۶	-۱/۹۲
۱۳	۳۰۰/۰۰	۳۰۲/۳۱	۳۰۱/۲۲	۲۹۱/۷۳	۲۹۵/۱۴	۲۸۹/۹۳	۰/۷۷	۰/۴۱	-۲/۷۶	-۱/۶۲	-۳/۳۶
۱۴	۳۰۱/۵۳	۳۰۱/۸۷	۲۹۴/۸۷	۲۹۱/۴۱	۲۵۷/۵۴	۳۰۵/۸۹	۰/۱۱	-۲/۲۱	-۳/۳۵	-۱۴/۵۹	۱/۴۵
۱۵	۳۰۴/۲۸	۲۸۴/۷۸	۲۶۳/۹۴	۲۷۷/۳۵	۲۰۵/۷۳	۳۰۰/۶۰	-۶/۹۸	-۱۳/۲۶	-۸/۸۵	-۳۲/۳۹	-۱/۲۱
۱۶	۳۰۹/۸۰	۲۷۶/۶۵	۳۲۷/۹۵	۳۵۶/۹۷	۳۰۹/۵۸	۳۱۸/۶۲	-۱۰/۷۰	۵/۸۶	۱۵/۲۳	-۰/۰۷	۲/۸۵
۱۷	۳۱۱/۱۲	۳۱۵/۷۶	۳۰۸/۳۲	۳۱۴/۰۹	۳۱۲/۷۲	۲۹۷/۲۱	۱/۴۹	-۰/۹۰	۰/۹۶	۰/۵۱	-۴/۴۷

ادامه جدول (۴-۲۰)

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۱۸	۳۱۴/۱۵	۳۱۳/۵۱	۳۰۳/۶۸	۳۱۱/۷۴	۳۱۲/۳۷	۳۰۵/۶۷	-۰/۲۰	-۳./۳۳	-۰/۷۷	-۰/۵۷	-۲/۷۰
۱۹	۳۱۸/۸۳	۳۲۱/۲۷	۳۰۹/۱۱	۳۱۷/۵۶	۳۲۲/۶۲	۳۱۳/۰۵	۰/۷۷	-۳/۰۵	-۰/۴۰	۱/۱۹	-۱/۸۱
۲۰	۳۱۹/۶۷	۳۲۰/۶۹	۳۱۱/۴۷	۳۱۶/۷۱	۳۱۹/۵۸	۳۱۳/۳۰	۰/۳۲	-۲/۵۶	-۰/۹۲	-۰/۰۳	-۱/۹۹
۲۱	۳۲۱/۱۲	۳۲۰/۱۴	۳۱۶/۳۰	۳۱۷/۱۵	۳۴۸/۶۹	۳۲۲/۸۴	-۰/۳۰	-۱/۵۰	-۱/۲۴	۸/۵۸	۰/۵۴
۲۲	۳۲۲/۷۴	۳۲۰/۲۰	۳۱۲/۸۵	۳۱۶/۷۱	۳۱۹/۵۹	۳۲۴/۵۲	-۰/۷۹	-۳/۰۶	-۱/۸۷	-۰/۹۸	-۰/۳۷
۲۳	۳۲۳/۵۴	۳۲۱/۱۶	۳۱۹/۷۴	۳۱۶/۶۹	۳۱۹/۵۰	۳۲۱/۱۷	-۰/۷۴	-۱/۱۸	-۲/۱۲	-۱/۲۵	-۰/۷۳
۲۴	۳۲۵/۷۳	۳۴۳/۸۶	۳۱۸/۸۱	۳۲۸/۰۹	۳۱۹/۳۲	۳۲۴/۷۷	۵/۵۷	-۲/۱۲	۰/۷۲	-۱/۹۷	-۰/۲۹
۲۵	۳۲۶/۹۴	۲۹۵/۷۷	۳۲۲/۳۷	۳۶۸/۰۹	۳۱۴/۲۶	۳۲۳/۷۹	-۹/۵۳	-۱/۴۰	-۱۸/۰۰	-۳/۸۸	-۰/۹۶
۲۶	۳۴۴/۹۶	۳۵۱/۵۳	۳۳۶/۰۰	۳۵۲/۴۲	۲۵۶/۱۶	۳۴۵/۶۴	۱/۹۰	-۲/۶۰	۲/۱۶	-۲۴/۷۳	۰/۲۰
۲۷	۳۴۸/۱۴	۳۵۰/۸۲	۳۴۱/۷۴	۳۵۲/۱۲	۳۴۴/۴۴	۳۴۴/۹۶	۰/۷۷	-۱/۸۴	۱/۱۴	-۱/۰۶	-۰/۹۱
۲۸	۳۴۹/۴۲	۳۵۷/۳۶	۳۴۵/۸۲	۳۳۹/۵۲	۳۱۷/۲۲	۳۴۸/۷۴	۲/۲۷	-۱/۰۳	-۲/۸۴	-۹/۲۲	-۰/۱۹
۲۹	۳۵۲/۴۱	۳۵۵/۰۶	۳۴۹/۰۷	۳۵۳/۶۴	۳۴۵/۶۱	۳۵۴/۷۷	۰/۷۵	-۰/۹۵	۰/۳۵	-۱/۹۳	۰/۶۷
۳۰	۳۵۳/۴۷	۳۶۱/۹۵	۳۴۸/۳۱	۳۷۳/۱۶	۴۱۹/۸۱	۳۶۲/۶۱	۲/۴۰	-۱/۴۶	۵/۵۷	۱۸/۷۷	۲/۵۹
۳۱	۳۵۶/۰۴	۳۵۴/۶۴	۳۵۰/۶۶	۳۵۰/۸۵	۳۵۶/۸۲	۳۳۷/۸۵	-۰/۳۹	-۱/۵۱	-۱/۴۶	۰/۲۲	-۵/۱۱
۳۲	۳۵۷/۶۵	۳۶۰/۱۱	۳۶۹/۹۴	۳۵۱/۰۵	۴۰۰/۸۹	۳۶۱/۷۰	۰/۶۹	۳/۴۴	-۱/۸۴	۱۲/۰۹	۱/۱۳
۳۳	۳۵۷/۸۶	۳۵۹/۸۰	۳۷۲/۸۱	۳۷۲/۰۵	۴۵۶/۲۹	۳۵۸/۴۵	۰/۵۴	۴/۱۸	۳/۹۶	۲۷/۵۰	۰/۱۷
۳۴	۳۵۷/۸۶	۳۵۶/۹۷	۳۵۸/۰۸	۳۶۷/۸۷	۳۷۰/۰۰	۳۵۶/۳۶	-۰/۲۵	۰/۰۶	۲/۸۰	۳/۳۹	-۰/۴۲

ادامه جدول (۴-۲۰)

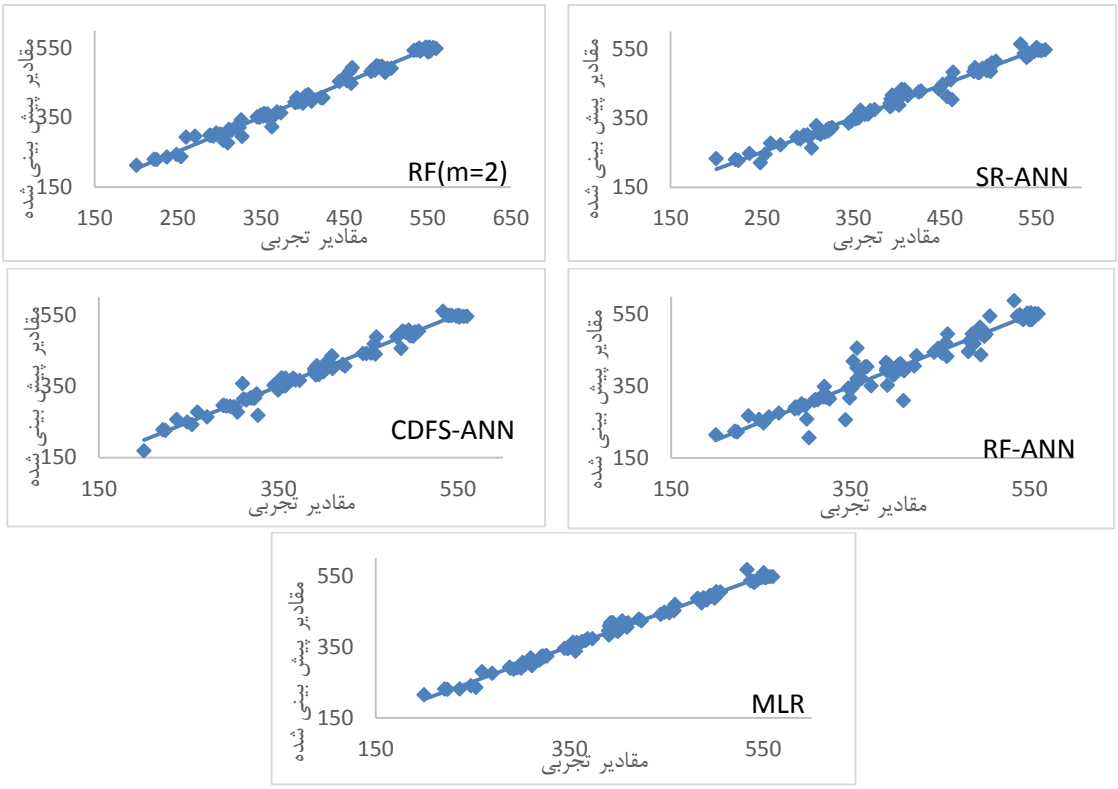
شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۳۵	۳۶۲/۸۲	۳۲۲/۹۵	۳۶۰/۹۰	۳۶۴/۷۸	۳۷۱/۵۸	۳۶۴/۳۳	-۱۰/۹۹	-۰/۵۳	۰/۵۴	۲/۴۱	۰/۴۲
۳۶	۳۶۶/۲۸	۳۵۷/۱۲	۳۶۱/۸۴	۳۷۲/۵۰	۴۰۳/۶۲	۳۶۷/۸۵	-۲/۵۰	-۱/۲۱	۱/۷۰	۱۰/۲۰	۰/۴۳
۳۷	۳۶۸/۷۵	۳۶۶/۴۶	۳۷۱/۶۲	۳۶۷/۲۱	۴۰۴/۴۴	۳۷۲/۷۸	-۰/۶۲	۰/۷۸	-۰/۴۲	۹/۶۸	۱/۰۹
۳۸	.۳۷۳/۵۵	۳۶۴/۰۵	۳۷۴/۵۱	۳۶۵/۹۴	۳۵۰/۶۴	۳۷۲/۵۳	-۲/۵۴	۰/۲۶	-۲/۰۴	-۶/۱۳	-۰/۲۷
۳۹	۳۹۱/۷۳	۳۹۶/۹۱	۳۸۶/۳۵	۳۸۱/۱۰	۳۵۲/۹۴	۳۸۶/۰۸	۱/۳۲	-۱/۳۷	-۲/۷۱	-۹/۹۰	-۱/۴۴
۴۰	۳۹۰/۱۸	۳۹۴/۰۶	۳۹۲/۲۷	۳۹۸/۸۲	۴۱۵/۸۳	۳۹۵/۴۹	۰/۹۹	۰/۵۴	۲/۲۱	۶/۵۷	۱/۳۶
۴۱	۳۹۰/۵۷	۳۹۴/۰۵	۳۸۳/۱۳	۳۹۱/۶۴	۳۹۳/۵۲	۳۸۴/۰۸	۰/۸۹	-۱/۹۰	۰/۲۷	۰/۷۶	-۱/۶۶
۴۲	۳۹۱/۱۰	۳۹۳/۴۲	۴۰۵/۳۷	۳۹۹/۹۳	۳۹۹/۰۰	۴۰۸/۸۴	۰/۵۹	۳/۶۲	۲/۲۶	۲/۰۲	۴/۵۴
۴۳	۳۹۲/۷۰	۴۰۷/۴۱	۴۱۵/۶۴	۴۰۷/۷۸	۴۱۰/۳۳	۴۱۸/۴۲	۳/۷۴	۵/۸۴	۳/۸۴	۴/۵۰	۶/۵۵
۴۴	۳۹۴/۷۶	۳۹۵/۳۱	۳۹۵/۶۱	۳۸۲/۰۶	۴۰۴/۰۰	۴۱۷/۷۴	۰/۱۴	۰/۲۱	-۳/۲۲	۲/۳۴	۵/۸۳
۴۵	۳۹۸/۳۶	۳۹۵/۲۰	۴۰۲/۴۴	۳۹۱/۵۰	۳۸۱/۵۴	۳۹۵/۹۶	-۰/۷۹	۱/۰۲	-۱/۷۲	-۴/۲۲	-۰/۶۰
۴۶	۴۰۰/۰۰	۳۹۱/۸۵	۳۸۷/۷۸	۳۹۰/۴۲	۳۸۳/۸۱	۳۹۳/۲۲	-۲/۰۴	-۳/۰۶	-۲/۴۰	-۴/۰۵	-۱/۷۰
۴۷	۴۰۲/۷۷	۴۱۲/۲۷	۴۳۲/۸۱	۴۱۴/۲۳	۴۰۴/۹۹	۴۱۰/۸۶	۲/۳۶	۷/۴۶	۲/۸۴	۰/۵۵	۲/۰۱
۴۸	۴۰۴/۲۸	۴۱۲/۹۳	۴۳۰/۰۳	۴۱۴/۴۶	۴۰۹/۴۰	۴۲۳/۱۹	۲/۱۴	۶/۳۷	۲/۵۲	۱/۲۷	۴/۶۸
۴۹	۴۰۶/۴۷	۴۱۶/۹۲	۴۳۱/۶۲	۴۱۶/۳۴	۴۱۲/۴۲	۴۱۵/۳۰	۲/۶۰	۶/۱۹	۲/۴۳	۱/۴۶	۲/۱۷
۵۰	۴۰۹/۵۶	۴۰۶/۳۹	۴۱۵/۲۴	۴۳۵/۱۸	۳۰۹/۹۳	۴۰۵/۷۵	-۰/۷۷	۱/۳۹	۶/۲۶	-۳۸/۷۴	-۰/۹۲
۵۱	۴۱۰/۲۷	۳۹۶/۸۰	۴۱۷/۴۶	۳۹۹/۰۱	۳۹۲/۷۸	۴۱۷/۸۶	-۳/۲۸	۱/۷۵	-۲/۷۵	-۴/۲۶	۱/۸۵

ادامه جدول (۴-۲۰)

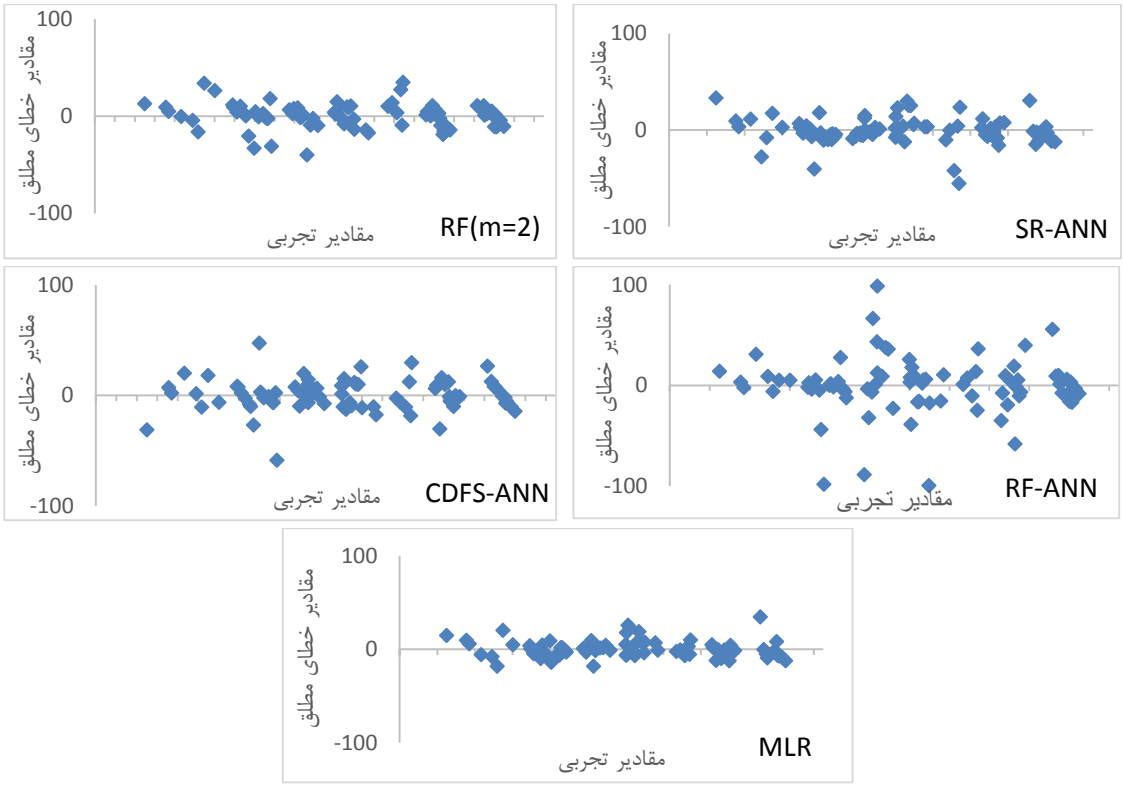
شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۵۲	۴۲۱/۵۵	۴۰۷/۱۶	۴۲۵/۱۱	۴۱۱/۱۲	۴۰۵/۷۵	۴۲۸/۳۱	-۳/۴۱	۰/۸۴	-۲/۴۷	-۳/۷۵	۱/۶۰
۵۳	۴۲۴/۱۹	۴۰۶/۸۸	۴۲۷/۷۴	۴۰۶/۵۸	۴۳۴/۵۸	۴۲۳/۱۸	-۴/۰۸	۰/۸۴	-۴/۱۵	۲/۴۵	-۰/۲۴
۵۴	۴۴۳/۹۳	۴۵۳/۹۷	۴۳۳/۹۴	۴۴۱/۲۱	۴۴۴/۶۷	۴۴۱/۶۷	۲/۲۶	-۲/۲۵	-۰/۶۱	۰/۱۷	-۰/۵۱
۵۵	۴۴۸/۰۷	۴۶۱/۷۰	۴۴۷/۸۸	۴۴۱/۶۴	۴۵۵/۲۹	۴۴۷/۳۵	۳/۰۴	-۰/۰۴	-۱/۴۴	۱/۶۱	-۰/۱۶
۵۶	۴۵۲/۹۳	۴۵۶/۵۲	۴۱۱/۰۴	۴۴۲/۲۴	۴۴۲/۳۲	۴۴۶/۲۵	۰/۷۹	-۹/۲۵	-۲/۳۶	-۲/۳۴	-۱/۴۸
۵۷	۴۵۷/۹۸	۴۴۸/۵۰	۴۰۲/۸۵	۴۳۹/۴۲	۴۳۳/۱۲	۴۵۲/۵۰	-۲/۰۷	-۱۲/۰۴	-۴/۰۵	-۵/۴۳	-۱/۲۰
۵۸	۴۵۹/۰۱	۴۹۳/۸۷	۴۸۲/۶۲	۴۸۸/۴۹	۴۹۵/۱۶	۴۶۹/۰۴	۷/۶۰	۵/۱۴	۶/۴۲	۷/۸۸	۲/۱۸
۵۹	۴۵۶/۶۰	۴۸۳/۶۸	۴۶۰/۵۶	۴۶۸/۸۳	۴۷۰/۰۶	۴۵۹/۷۵	۵/۹۳	۰/۸۷	۲/۶۸	۲/۹۵	۰/۶۹
۶۰	۴۸۱/۹۶	۴۸۳/۱۷	۴۸۴/۵۲	۴۸۷/۹۸	۴۴۶/۷۰	۴۸۶/۸۳	۰/۲۵	۰/۵۳	۱/۲۵	-۷/۳۱	۱/۰۱
۶۱	۴۸۶/۳۴	۴۸۶/۵۰	۴۸۱/۵۳	۴۵۵/۸۰	۴۹۵/۸۷	۴۷۴/۳۵	۰/۰۳	-۰/۹۹	-۶/۲۸	۱/۹۶	-۲/۴۷
۶۲	۴۸۳/۲۷	۴۸۷/۷۴	۴۹۵/۱۱	۴۹۲/۰۷	۴۷۵/۶۵	۴۸۴/۹۸	۰/۹۳	۲/۴۵	۱/۸۲	-۱/۵۸	۰/۳۵
۶۳	۴۸۸/۳۸	۴۹۹/۱۶	۴۸۱/۹۶	۵۰۴/۳۸	۴۶۸/۹۳	۴۸۸/۱۲	۲/۲۱	-۱/۳۱	۳/۲۸	-۳/۹۸	-۰/۰۵
۶۴	۴۹۱/۵۳	۴۹۸/۰۷	۴۹۳/۴۶	۵۰۱/۳۴	۴۹۵/۵۳	۴۸۱/۷۴	۱/۳۳	۰/۳۹	۲/۰۰	۰/۸۱	-۱/۹۹
۶۵	۴۹۴/۹۱	۴۹۷/۸۹	۴۹۲/۹۹	۵۰۶/۸۹	۵۱۳/۷۹	۴۹۴/۳۵	۰/۶۰	-۰/۳۹	۲/۴۲	۳/۸۱	-۰/۱۱
۶۶	۴۹۵/۸۹	۴۹۳/۶۴	۴۸۷/۷۳	۴۹۴/۷۰	۴۳۷/۴۲	۴۹۲/۹۶	-۰/۴۵	-۱/۷۱	-۰/۲۴	-۱۱/۷۹	-۰/۵۹
۶۷	۴۹۶/۶۷	۴۸۶/۹۰	۵۰۰/۲۴	۴۹۱/۲۳	۴۹۵/۲۳	۴۹۴/۱۰	-۱/۹۷	۰/۷۲	-۱/۱۰	-۰/۲۹	-۰/۵۲
۶۸	۴۹۸/۸۶	۴۸۰/۱۱	۴۹۰/۷۶	۴۹۲/۱۶	۵۰۴/۰۳	۴۹۲/۸۱	-۳/۷۶	-۱/۶۲	-۱/۳۴	۱/۰۴	-۱/۲۱

ادامه جدول (۴-۲۰)

شماره ترکیب	مقدار تجربی	مقدار پیش بینی شده					درصد خطا				
		RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR	RF (m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
۶۹	۵۰۰/۰۰	۴۸۴/۵۰	۴۸۴/۱۵	۴۸۹/۷۸	۴۸۹/۳۰	۴۸۷/۵۵	-۳/۰۱	-۳/۱۷	-۲/۰۴	-۲/۱۴	-۲/۴۹
۷۰	۵۰۱/۶۴	۴۹۱/۷۳	۵۰۹/۲۱	۵۰۱/۱۷	۴۹۴/۹۴	۵۰۵/۷۰	-۱/۹۸	۱/۵۱	-۰/۰۹	-۱/۳۴	۰/۸۱
۷۱	۵۰۵/۹۷	۴۹۱/۷۶	۵۱۳/۸۲	۵۰۴/۶۳	۵۴۵/۶۸	۵۰۴/۵۶	-۲/۸۱	۱/۵۵	-۰/۲۷	۷/۸۵	-۰/۲۸
۷۲	۵۳۳/۲۲	۵۴۴/۱۳	۵۶۳/۸۳	۵۵۹/۶۱	۵۸۸/۷۷	۵۶۷/۸۹	۲/۰۵	۵/۷۴	۴/۹۵	۱۰/۴۲	۶/۵۰
۷۳	۵۳۶/۹۲	۵۴۴/۴۶	۵۳۵/۸۱	۵۴۹/۲۱	۵۴۵/۸۷	۵۳۶/۵۸	۱/۴۱	-۰/۲۱	۲/۲۹	۱/۶۷	-۰/۰۶
۷۴	۵۳۹/۵۹	۵۵۰/۳۷	۵۲۴/۹۶	۵۴۷/۲۴	۵۴۹/۲۰	۵۳۴/۵۱	۲/۰۰	-۲/۷۱	۱/۴۲	۱/۷۸	-۰/۹۴
۷۵	۵۴۰/۷۱	۵۴۱/۷۰	۵۳۸/۵۶	۵۴۸/۱۳	۵۴۱/۹۰	۵۳۱/۳۲	۰/۱۸	-۰/۴۰	۱/۳۷	۰/۲۲	-۱/۷۴
۷۶	۵۴۳/۰۶	۵۴۵/۴۹	۵۳۲/۶۰	۵۴۸/۰۲	۵۳۵/۳۵	۵۳۷/۰۸	۰/۴۵	-۱/۹۳	۰/۹۱	-۱/۴۲	-۱/۱۰
۷۷	۵۴۷/۷۱	۵۵۲/۷۳	۵۴۱/۸۵	۵۴۶/۷۰	۵۵۳/۶۴	۵۴۵/۶۰	۰/۹۲	-۱/۰۷	-۰/۱۹	۱/۰۸	-۰/۳۹
۷۸	۵۵۰/۴۳	۵۳۹/۵۸	۵۵۳/۸۸	۵۴۸/۲۱	۵۳۴/۴۴	۵۵۸/۵۵	-۱/۹۷	۰/۶۳	-۰/۴۰	-۲/۹۱	۱/۴۸
۷۹	۵۵۱/۶۵	۵۵۲/۹۰	۵۴۸/۸۳	۵۴۶/۲۲	۵۵۴/۶۹	۵۴۵/۰۹	۰/۲۳	-۰/۵۱	-۰/۹۸	۰/۵۵	-۱/۱۹
۸۰	۵۵۲/۸۲	۵۴۱/۹۰	۵۴۵/۰۱	۵۴۵/۶۶	۵۳۵/۷۰	۵۴۵/۷۴	-۱/۹۷	-۱/۴۱	-۱/۳۰	-۳/۱۰	-۱/۲۸
۸۱	۵۵۱/۰۰	۵۴۹/۸۷	۵۴۷/۷۷	۵۴۳/۶۷	۵۴۴/۶۷	۵۴۷/۱۴	-۰/۲۱	-۰/۵۹	-۱/۳۳	-۱/۱۵	-۰/۷۰
۸۲	۵۵۶/۳۲	۵۵۲/۰۷	۵۴۴/۹۱	۵۴۵/۷۴	۵۵۲/۹۰	۵۴۷/۶۵	-۰/۷۶	-۲/۰۵	-۱/۹۰	-۰/۶۱	-۱/۵۶
۸۳	۵۵۹/۹۰	۵۴۹/۳۳	۵۴۷/۶۹	۵۴۵/۴۰	۵۵۱/۶۱	۵۴۷/۵۸	-۱/۸۹	-۲/۱۸	-۲/۵۹	-۱/۴۸	-۲/۲۰



شکل (۴-۸)- نمودار مقادیر پیش‌بینی شده اندیس بازداری کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل‌های مختلف در مقابل مقادیر تجربی



شکل (۴-۹)- نمودار مقادیر خطای مطلق در مقابل مقادیر تجربی برای کل داده‌ها با مدل‌های مختلف به روش رد مرحله‌ای تک‌تک

جدول (۴-۲۱)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای داده‌های سری آزمون

	RF(m=75)	RF(m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
R^2	۰/۹۹۷	۰/۹۹۱	۰/۹۸۶	۰/۹۸۹	۰/۹۵۴	۰/۹۸۳
R^2_{adj}	-	/۹۹۰	۰/۹۸۲	۰/۹۸۶	۰/۸۵۱	۰/۹۷۳
SEP	۴/۶۸۷	۸/۲۹۷	۱۰/۲۱۶	۹/۱۳۰	۱۸/۷۴۳	۱۱/۳۴۲
REP	۱/۱۳۱	۲/۰۰۳	۲/۴۶۶	۲/۲۰۴	۴/۵۲۴	۲/۷۳۸
MAE	۳/۹۴۶	۶/۵۸۰	۷/۱۱۳	۷/۴۸۶	۱۵/۰۸۴	۷/۵۳۶
MRE	۱/۰۰۸	۱/۶۳۷	۱/۷۲۴	۱/۷۵۹	۳/۶۰۵	۱/۹۲۱
MSE	۲۱/۹۶۴	۶۸/۸۴۱	۱۰۴/۳۶۳	۸۳/۳۵۶	۳۵۱/۳۰۹	۱۲۸/۶۳۸
PRESS	۳۷۳	۱۱۷۰	۱۷۷۴	۱۴۱۷	۵۹۷۰	۱۲۸۷
F	-	۸۷۱/۲۴۹	۲۴۴/۵۸۶	۴۴۳/۰۶۴	۱۱/۹۱۱	۱۱۱/۶۴۵
FIT	-	۷۲/۴۴۰	۲۵/۹۴۰	۴۴/۷۸۲	۰/۷۴۲	۱۰/۸۸۲

جدول (۴-۲۲)- پارامترهای آماری به دست آمده توسط مدل‌های مختلف برای کل داده‌ها به روش رد تک‌تک

	RF(m=2)	SR-ANN	CDFS-ANN	RF-ANN	MLR
R^2	۰/۹۸۳	۰/۹۷۸	۰/۹۷۸	۰/۹۰۸	۰/۹۹۱
R^2_{adj}	۰/۹۸۲	۰/۹۷۶	۰/۹۷۷	۰/۸۹۴	۰/۹۹۰
SEP	۱۲/۵۳۲	۱۴/۲۳۴	۱۴/۲۳۳	۲۹/۰۱	۹/۲۱۸
REP	۳/۱۴۶	۳/۵۷۳	۳/۵۷۳	۷/۲۸	۲/۳۱۴
MAE	۹/۰۰۳	۹/۸۸۱	۱۰/۳۳۳	۱۸/۰۵	۶/۷۶۴
MRE	۲/۴۵۹	۲/۶۹۴	۲/۸۲۴	۴/۷۶	۱/۸۴۱
MSE	۱۵۷/۰۶۳	۲۰۲/۶۲۰	۲۰۲/۵۹۲	۸۴۱/۵۵	۸۴/۹۶۹
PRESS	۱۳۰۳۶	۱۶۸۱۷	۱۶۸۱۵	۶۹۸۴۹	۷۰۵۲
F	۲۳۱۵	۶۸۹	۱۲۲۹	۷۶	۱۳۵۲
FIT	۵۲/۷۴۵	۳۱/۵۴۰	۳۷/۹۹۳	۳/۴۴۳	۶۸/۲۵۸

جدول (۴-۲۳)- مقادیر R^2 برای داده‌های سری آزمون با استفاده از آزمون Y-تصادفی

تکرار / روش	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
RF(m=73)	۰/۱۶۱	۰/۰۰۱	۰/۰۴۰	۰/۱۵۸	۰/۰۰۱	۰/۲۲۰	۰/۱۳۷	۰/۰۴۹	۰/۱۲۸	۰/۰۳۴
RF(m=2)	۰/۰۰۳	۰/۰۹۶	۰/۰۶۰	۰/۰۳۵	۰/۱۵۴	۰/۱۱۸	۰/۰۷۲	۰/۰۹۱	۰/۰۰۷	۰/۰۰۱
SR-ANN	۰/۰۳۶	۰/۴۱۹	۰/۳۱۱	۰/۲۸۷	۰/۰۲۹	۰/۴۶۲	۰/۲۱۰	۰/۱۳۹	۰/۱۰۹	۰/۰۱۴
CDFS-ANN	۰/۰۴۵	۰/۱۴۹	۰/۴۱۳	۰/۲۷۰	۰/۰۷۷	۰/۳۶۳	۰/۵۰۶	۰/۰۴۴	۰/۰۵۴	۰/۱۲۴
RF-ANN	۰/۱۹۴	۰/۰۵۶	۰/۴۸۶	۰/۱۱۵	۰/۱۵۸	۰/۰۱۶	۰/۱۸۲	۰/۱۱۹	۰/۱۶۹	۰/۲۱۴

۴-۱۰- بررسی ارتباط توصیف‌گرهای منتخب مدل برتر با اندیس بازداری

دو توصیف‌گر G2 و PiID در مدل‌سازی جنگل‌های تصادفی که مدل برتر می‌باشد، به کار برده شدند. توصیف‌گر G2 از دسته توصیف‌گرهای هندسی[1] است. توصیف‌گرهای هندسی از ساختار سه‌بعدی مولکول مشتق می‌شوند و لازمه‌ی محاسبه‌ی این توصیف‌گرها، بهینه‌سازی کامل ساختار ترکیبات می‌باشد. این توصیف‌گرها قدرت تمایز بالایی برای ساختارهای مشابه مولکول‌ها و صورت‌بندی‌های مختلف مولکول دارند و ویژگی‌های مولکول را به طور دقیق مشخص می‌کنند. این قدرت به این دلیل است که نمایش هندسی مولکول شامل موقعیت‌های نسبی اتم‌ها در فضای سه‌بعدی است. معمولا برای نمایش هندسی مولکول، اتم‌ها به صورت گره‌های سخت با شعاع وان‌دروالسی در نظر گرفته می‌شوند که در محل پیوندها با یکدیگر هم‌پوشانی دارند. سپس در فضای سه‌بعدی با استفاده از مختصات سه‌بعدی مراکز این کره‌ها و شعاع وان‌دروالسی آنها، این توصیف‌گرها محاسبه می‌شوند که شامل حجم وان‌دروالسی مولکول، مساحت سطح مولکول، فاکتور شکل و گشتاور اینرسی مولکول می‌باشد. G2 شاخص جاذبه گرانشی[2] است که تنها به تعداد اتم‌های متصل با پیوند محدود می‌باشد و توزیع جرم در یک مولکول را نشان می‌دهد و طبق معادله (۴-۲) تعریف می‌شود:

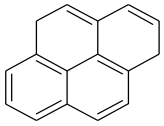
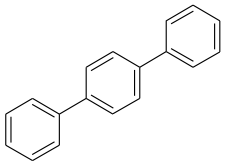
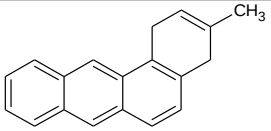
$$G2 = \left(\sum_{b=1}^{B} \frac{m_i m_j}{r_{ij}}\right)_b \qquad (۴-۲)$$

که m_i و m_j جرم اتمی اتم‌های موردنظر؛ r مسافت بین اتمی متناظر و B تعداد پیوندهای مولکول است. طبق این رابطه، G2 با زوج اتم‌های پیوند شده مشخص می‌شود. هرچه جرم مولکول بیشتر شود، به دلیل افزایش بیشتر تعداد صورت کسر نسبت به مخرج آن و افزایش تعداد جملات، G2 نیز بیشتر می‌شود. هم‌چنین با افزایش جرم و به دلیل افزایش تعداد پیوند دوگانه، قطبش‌پذیری افزایش یافته و

1-Geometrical
2-Gravitational index

برهم‌کنش مولکول با فاز ساکن بیشتر و در نتیجه اندیس بازداری افزایش می‌یابد. جدول (۴-۲۴) چگونگی تغییر اندیس بازداری با مقدار این توصیف‌گر را برای تعدادی از مولکول‌ها نشان می‌دهد.

جدول (۴-۲۴)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر G2

اندیس بازداری	مقدار G2	تعداد اتم	ساختار	شماره ترکیب
۳۵۲/۴۱	۱۰/۲۹	۲۸		۲۹
۳۵۷/۶۵	۱۱/۱۴	۳۲		۳۲
۴۱۰/۲۷	۱۲/۰۰	۳۵		۵۱

توصیف‌گر دوم، PiID است و از گروه توصیف‌گرهای توپولوژیکی[1] می‌باشد. توصیف‌گرهای توپولوژیکی از گراف مولکولی تهی از هیدروژن محاسبه می‌شوند. توصیف‌گرهای توپولوژیکی نشان دهنده نوع اتم، تعداد پیوند و چگونگی ارتباط آنها با یکدیگر بوده و به سادگی از روی ساختار دو بعدی مولکول قابل محاسبه‌اند و چون به طور مستقیم از ساختار مولکول محاسبه می‌شوند و مستقل از صورت‌بندی مولکول هستند، به عنوان توصیف‌گرهای عمومی مطرح هستند. این توصیف‌گرها می‌توانند به یک یا چند مشخصه‌ی ساختاری مولکول مثل اندازه، شکل، تقارن، شاخه‌دار بودن، حلقوی بودن و غیره حساس باشند و هم‌چنین می‌توانند اطلاعات شیمیایی را با توجه به نوع اتم و چندگانگی پیوند کدگذاری کنند. piID به صورت مرتبه پیوند قراردادی نام گذاری شده است و از رابطه (۴-۳) محاسبه می شود:

$$\mathrm{PiID} = \sum_{k=0}^{L} \mathrm{piPCK} \qquad (۴-۳)$$

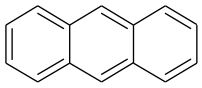
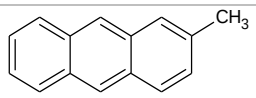
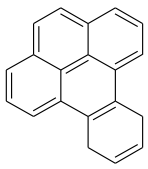
و piPCK نیز از رابطه‌ی زیر محاسبه می گردد:

$$\mathrm{piPCK} = \sum k_{pij} w_{ij} \qquad (۴-۴)$$

1-Topological

که در این رابطه، k_{pij} مسافت بین دو رأس مورد نظر و w_{ij}، وزن مسیر است که مرتبه پیوند مقدار آن را مشخص می‌کند. با افزایش تعداد اتم در مولکول و مرتبه پیوندها، مقدار این توصیف‌گر افزایش می‌یابد و به دلیل افزایش جرم ناشی از آن و افزایش قطبش‌پذیری، اندیس بازداری نیز افزایش می‌یابد [۳۸]. تاثیر مقدار این توصیف‌گر بر اندیس بازداری برای تعدادی از مولکول‌ها در جدول (۴-۲۵) نشان داده شده است.

جدول (۴-۲۵)- نمایش ارتباط اندیس بازداری با مقدار توصیف‌گر PiID

شماره ترکیب	ساختار	تعداد اتم	تعداد پیوند دوگانه	مقدار PiID	اندیس بازداری
۱۴		۲۴	۷	۸۱۵۳	۳۰۱/۵۳
۲۱		۲۷	۷	۹۳۱۲	۳۲۱/۱۲
۵۶		۳۴	۹	۱۰۱۱۶۳	۴۵۲/۹۳

۴-۱۱- نتیجه‌گیری

در این تحقیق از روش‌های مختلفی برای مدل‌سازی اندیس بازداری گروهی از هیدروکربن‌های آروماتیک چند حلقه‌ای با استفاده از توصیف‌گرهای ساختاری استفاده شد. از سه شبکه عصبی مصنوعی مورد استفاده، شبکه با توصیف‌گرهای منتخب راهبرد CDFS نتایج بهتری را نسبت به دو روش دیگر ارائه نموده و مدل جنگل‌های تصادفی نسبت به تمام روش‌ها عملکرد بهتری را نشان داده است. این مدل می‌تواند در پیش‌بینی اندیس بازداری گروهی از این ترکیبات که هنوز سنتز و یا اندیس بازداری آنها اندازه‌گیری نشده است، به کار رود.

آینده‌نگری

- روش جنگل‌های تصادفی را می‌توان برای پیش‌بینی اندیس بازداری یا سایر خواص ترکیبات دیگر استفاده نمود.
- علاوه بر روش جنگل‌های تصادفی و شبکه عصبی مصنوعی می‌توان از سایر روش‌های مدل‌سازی غیرخطی مانند SVM و LS-SVM[1] برای پیش‌بینی اندیس بازداری این ترکیبات استفاده نمود.
- علاوه بر روش‌های رگرسیون مرحله‌ای و راهبرد CDFS که در این تحقیق به عنوان روش‌های انتخاب توصیف‌گر استفاده شده است می‌توان از سایر روش‌های انتخاب متغیر مانند اجتماع مورچگان[2] و طرح‌ریزی پیاپی[3] استفاده کرد.

1-Least square supported vector machin
2-Ant colony optimization
3-Successive projection algorithm (SPA)

فهرست منابع

[۱]-راجر، ام اسمیت، (۱۳۸۳)، *"کروماتوگرافی گاز و مایع در شیمی تجزیه"*، ارباب زوار م، سرافراز یزدی ع، انتشارات دانشگاه فردوسی مشهد، ۴۸۶، ص ۱۸۵-۱۸۴.

[2]- Wardencki W, (1998)," Problems with the determination of environmental sulphur compounds by gas chromatography", *J. Chromatogr.A*, 793, pp 1-19.

[3]-Miller K. E, Bruno T J, (2003)," Isothermal Kova´ts retention indices of sulfur compounds on a poly (5% diphenyl–95% dimethylsiloxane) stationary phase", *J. Chromatogr. A*, 1007, pp 117-125.

[4]- Du H, Ring z, Briker Y, Arboled P,(2004), "Prediction of gas chromatographic retention times and indices of sulfur compounds in light cycle oil", *Catal. Today*, 98, pp 214-225.

[5]-Can H, Dimoglo 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.

[6]-Liu L, Hashi Y, Liu M, Wet Y, Lin J,(2007), "Determination of Particle-associated polycyclic aromatic hydrocarbons in urban air of Beijing by GC\MS", *Anal.Sci*, 23, pp 667-671.

[7]-U.S. Department of Health and Human Servers, (2000),"PAH prepared by the Wisconsin department of health and family series division of public health", POH 4606.

[8]-Canadian soil qualityguidlines, (2008),"Carcinogenic and other PAHs", PN1401-15BN978-1-896997-79-7-PDF.

[9]-Engkvist O, Borowski O, Lindh A, Colmsjo A, (1996),"On the relation between retention index and interaction between the solute and the column in gas-liquid chromatography" *J.Chem.Inf Compu Sci*, 36, pp 1153-1161.

[10]-Adler N, Babic D, Trinajstic, (1985),"On the calculation of the HPLC parameters for PAHs", *Frescnius Z Anal Chem*, 332, pp 426-429.

[11]-Nowski M, Peta M, Janeczec J, (2000),"Composition and source of polycyclic aromatic compound in deposited dust from selected sites around the upper Silesia ,Poland,", *Geol. Q. 2004*, 48, pp 169-180.

[12]-Ribeiro F, Ferreire M, (2003)."QSPR model of boiling point, octanol-water partition coefficient and retention index of PAHs", *J. Mol. Struct*, 663, pp 109-126.

[13]-Skrbic B, Onjia A, (2006),"Prediction of the Lee retention index of PAHs by ANN", *J. Chromatogr A*, 1108, pp 279-284.

[14]-Jw X, Yuqs J, (2008),"Quantitative structure-chromatographic RI for PAHs", *J. Chromatogr A*, 1158, pp 273-305.

[15]- Wold S. (1995), "chemometrics; what do we mean with it and what do we want from it", *J Chemolab*, 30, pp 109-115.

[16]- Manssnat D. L., Vandeginste B. G., Deming S. N. and Kaufman L. (1998) "chemometrics, A Text Bo", Elsevier, Amesterdom.

[17]- Hilbert, D. B. (1993), "Genetic algorithm in chemistr". *Chemometr. Intel. Lab. Syst*, 19, pp 277-293.

[18]- Anderson, G.G. Kaufmann, P. (2000),"Development of a generalized neural network", *Chemometr. Intel. Lab. Syst*. 50 pp. 101-105.

[19]- Arab Chamjangali M, (2009),"Modeling of cytotoxicity data of anti-HIV 1-[(5-chlorophenyl) sulfpnyl]-1H-pyrole derivatives using calculated molecular descriptors and Levenberg- Marquadt artificial neural network", *J.Chem. Bio. Drug. Des*, 73, pp 456-465.

[۲۰]- لواین ای. ان(۱۳۸۷)، "*شیمی کوانتومی*"، اسلامپور غ، مقاری ع؛ نجفی ب، جلد سوم چاپ اول، انتشارات فاطمی.

[21]- HyperChem7.0 Toronto, Canada: HyperCube Inc, htpp:\\www.hyper.com.

[22]- T. Clark, (1998), "*A Handbook Computational Chemistry"*, John wiley & Sons, New York.

[۲۳]- عرب چم جنگلی م، (۱۳۸۶) "پیش بینی فعالیت دارویی ضد ایدز مشتقات ۵-فنیل –۱-فنیل آمینو ۱-ایمیدازول به وسیله شبکه عصبی مصنوعی "، دانشگاه صنعتی شاهرود، گزارش طرح پژوهشی.

.

[24]- Karekson, M. (2000) "*Molecular descriptors in QSAR/QSPR*". Wiley-VCH, New York.

[25]- Habibi A, Danandeh M, (2007), "prediction of acidity constant for substituted acetic acids in water using artificial neural networks", *Indian J. Chem*, 46B, pp 478-487.

[۲۶]-فرشادفر ع، (۱۳۸۰)، "*اصول و روش های پیشرفته آماری(تجزیه رگرسیون)*"، چاپ دوم، انتشارات طاق بستان.

[27]-Hemmateenejad B, Javadnia K, Elyasi M, (2007),"Quantitative structure-retention relationship for Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy", *Anal. Chim. Acta*, 592, pp 72-81.

[28]- Hemmateenejad B, Javadnia K, Elyasi M, Miri R, (2008),"linear and nonlinear quantitative structure-property relationship models for solubility of some anthraquinone, anthrone and xanthone derivatives in supercritical carbon dioxide", *Anal, Chim. Acta*, 610, pp 25-34.

[29]-htpp://www.iranhiv.com/treatment.

[۳۰]-هاگان ت. ه، دیموث ه، بیل م، "*طراحی شبکه های مصنوعی*"، کیا م، انتشارات کیان رایانه سبز

.

[۳۱]-منهاج م، (۱۳۸۷)،"*مبانی شبکه های عصبی (هوش محاسباتی)*"، جلد اول، چاپ پنجم، مرکز نشر دانشگاه صنعتی امیرکبیر، تهران

.

[32]-http://foran.takdownload.ir/threads/25029-؟-شبکه –های-عصبی-مصنوعی-چیست.

[۳۳]- اشرفی م، (۱۳۸۹)، پایان نامه کارشناسی ارشد، "مطالعه ارتباط کمی ساختار-فعالیت مشتقات تیوکربامات‌ها به عنوان دسته جدیدی از بازدارنده‌های غیر نوکلئوزیدی HIV "، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[34]-Breiman L, (2001),"Random forests", Mach. Learn, 45, pp 5-32.

[35]-Hastie T, Tibshirani R, Friedman J, (2009),"The Elements of Statistical Learning Data Mining Inferences, and Prediction", 2nd Ed, Springer, New York.

[36]- Genuer R, Poggi J M, Tuleau-Malot Ch, (2010), "Variable selection using random forests", *Pattern Recogn Lett*, 31, pp 2225-2236

[37]-Mercader A.G, Pomilio A.B, (2010), "QSAR study of flavonoids and biflavonoids as influenza HINI virus neuraminidase inhibitors", *Eur. J. Med. Chem*, 45, pp 1724-1730.

[38]-Todeschini R, Consonni V, (2000), "*Handbook of molecular descriptors*", Wiley-VCH, Weinheim, Germany.

[39]-Todeschini R, Milano Chemometrics and QSPR Group.http://www.disat.unimib.it.

[40]-SPSS for windows Statistical package for IBM PC, SPSS Inc,http://www.spss.com

[41]-MATLAB 7.8, the Math Work, Inc, Natick, MA, USA.

[42]-http:\\en.wikipedia.org/wiki/R/R (programming language).

[43]-Garkani-Nejad Z, Karlovits M, Demuth W, Stimpfl T,Vycudilik W, Jalali-Heravi M, Varmuza K, (2004),"Prediction of gas chromatographic retention time indices of a diverse set of toxicologically relevant compounds", *J. chrom. A*, 1028, pp 287-295.

[44]-Zupan J, Gasteiger J, (1993),"*Neural Networks for chemists*", VCH publishers, New York, Chapter 2.

[45]- Goodarzi M, Funar-Timofei S, Heyden Y.V, (2013), "Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models", *Trends. Anal. Chem*, 42, pp 49-53

[46]- Wang Z, Li K, Lambert P, Yan Ch, (2007) ," Identification, characterization and quantitation of pyrogenic polycyclic aromatic hydrocarbons and other organic compounds in tire fire products", *J. Chromatogr. A*, 1139, pp 14-26.

Abstract

Organo-sulfur compounds and poly cyclic aromatic hydrocarbons (PAHs) are known to be two groups of persistent pollutants. In addition, there is evidence of the carcinogenic potential of these compounds. There are different analytical techniques such as gas chromatography for separation and identification of these compounds by comparison of retention indices with standards. So, the retention index of a solute is an important parameter can be used for identification of organo-sulfur and PAHs compounds, but on the other hand, these methods are time consuming and very expensive. Therefore, to overcome on these problems, chemometrics methods can be used for calculation of retention indices of these compounds.

In the first section of this thesis, a new method of Random Forests (RF) is used for predicting the retention indices of a group of sulfur compounds. Also the performance of the random forests model was compared with the artificial neural network (ANN) and multiple linear regressions (MLR). The determination coefficients obtained for test set by RF, SR-ANN, CDFS-ANN, RF-ANN and MLR are 0.997, 0.967, 0.994, 0.986 and 0.996 respectively. The results obtained showed that RF model has relative superiority than ANN and MLR to predict the retention indices of these compounds.

In the second section, as same as first part, these methods were used to predict the retention indices of some PAH compounds. The determination coefficients obtained for test set by RF, SR-ANN, CDFS-ANN, RF-ANN and MLR are 0.997, 0.986, 0.989, 0.954 and 0.983 respectively. Obtained result showed that RF is better than the other techniques for prediction of RIs of PAHs.

Keywords: organo-sulfur compounds, polycyclic aromatic hydrocarbons, retention index, RF, ANN, MLR.

Shahrood University of Technology

Faculty of chemistry

Application of Random Forests (RF) method as a tool for variable selection and modeling of retention index of some organic compounds as environmental pollutant

Fereshte sadat emadi jandaghi

Supervisors

Dr. Nasser goudarzi

Dr. davood shahsavani

Advisor

Dr. mansour arab chamjangali

February 2014