دانشکده شیمی

رشته شیمی گرایش تجزیه

پایان نامه کارشناسی ارشد

مطالعه ارتباط کمی ساختار-فعالیت ترکیبات دارویی با استفاده از روش جنگل‌های تصادفی

نگارنده: عطیه عسلی

اساتید راهنما:

دکتر ناصر گودرزی

دکتر منصور عرب چم جنگلی

شهریور ۱۳۹۵

اگر شایسته تقدیم باشد

تقدیم به:

«زیباترین نقش های عالم هستی»

پدر بزرگوارم؛

مادر عزیز و دلسوزم؛

و خواهر مهربانم

و

همه کسانی که دوستشان دارم...

سپاسگزاری

سپاس بیکران ایزد منان را که در پرتو لایزالش توفیق آموختن میسر نمود تا منت پذیر آستان کبریایی اش باشم. امروز که به توفیق ایزد مهربان، راهی دیگر از زندگی را با موفقیت سپری کردم، پیشانی شکر بر سجده گاه عبودیت می سایم و بر خود واجب می دانم که از منت گذاران این راه قدردانی نمایم و با شهادت قلم چند سطری را به پاس زحمات بی دریغشان بنگارم. در همین راستا بر اساس روایت مشهور لم یشکر مخلوق لم یشکر الخالق، تحسنت سزاوار است نهایت سپاس قلبی خود را تقدیم حضور اساتید راهنمای گرامیم جناب آقای دکتر ناصر گودرزی و جناب آقای دکتر منصور عرب چم جنگلی گردانم که زحمات بی شائبه ای متحمل گشته اند و در تمامی این مدت با بردباری مرا راهنمایی فرموده اند و بی شک انجام مراحل مختلف این پایان نامه بدون حمایت و پشتیبانی ایشان امکان پذیر نبود. از داوران محترم جناب آقای دکتر قدمعلی باقریان و سرکار خانم دکتر مصدرالامور به خاطر نظرات ارزنده و اصلاحیات بجا و دلسوزانه شان ممنون و سپاسگزارم.

در این تلاش کوچک با تمام عشق و اشتیاق از پدر و مادر عزیزم تحسین آموزگاران زندگیم تشکر کرده و بوسه بر دستان مردانه پدر و چشمان دعاگوی مادرم می زنم، آنان که امروز من آرزوی دیروزشان بود و از خداوند منان می خواهم عمری بیفزاید تا گوشه ای از زحماتشان را جبران کنم، و دوست همیشگی ام خواهر عزیزم، آن که همواره حامی و مشوقم بوده و پیمودن روزهای سخت و آسان زندگی ام بدون دعای خیر و برکت وجودش غیر ممکن بود. سرو وجودشان، همیشه سرسبز و استوار باد.

تعهد نامه

اینجانب **عطیه عسلی** دانشجوی دوره کارشناسی ارشد رشته **شیمی تجزیه** دانشکده **شیمی** دانشگاه شاهرود نویسنده پایان نامه **مطالعه ارتباط کمی ساختار-فعالیت ترکیبات دارویی با روش جنگل تصادفی** تحت راهنمائی **دکتر ناصر گودرزی و دکتر منصور عرب چم جنگلی** 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.**

چکیده

در بخش اول این مطالعه، مدل‌های QSAR برای پیش‌بینی فعالیت ضـــدHIV مشـــتقات پیریدینون ساخته شدند. برای انتخاب توصیف‌کننده‌ها از روش سهم گروه و رگرسیون خطی مرحله‌ای استفاده شد. سپس توصیف‌کننده‌های مهم به عنوان ورودی برای ایجاد مدل‌های QSAR با استفاده از روش جنگل‌های تصادفی(RF) و شبکه عصبی مصنوعی (ANN) استفاده شد. سری داده‌ها برای الگوریتم جنگل تصادفی به دو سری آموزش (۴۳ترکیب) و سری آزمون (۱۰ترکیب) تقسیم شدند. پارامترهای موثر بر روش جنگل‌های تصـــادفی شـــامل تعداد مشـــاهدات باقی‌مانده در هر گره (N_{tree}) و تعداد توصیف‌کننده‌های انتخاب شده در هر مرحله افراز (Mtry) با استفاده از سری آموزش بهینه گردید. همچنین برای روش شبکه عصبی مصنوعی سری داده‌ها به سه دسته آموزش (۳۷ترکیب)، ارزیابی (۸ترکیب) و آزمون (۸ترکیب) تقسیم شدند. بعد از بهینه‌سازی پارامترهای موثر، برای سری ارزیابی، بهترین مدل انتخاب گردید. ارزیابی مدل‌های بدست آمده توسط پیش‌بینی فعالیت سری آزمون انجام گرفت و مجذور ضریب همبستگی برای مدل‌های RF و ANN به ترتیب ۰/۹۲۹ و ۰/۹۳۵به دست آمد.

در بخش دوم این تحقیق، روش‌های RF و ANN برای مدل‌سازی و پیش‌بینی صحیح فعالیت ضد HIV یکسری از مشتقات تیازول استفاده شد. داده‌ها به دو سری آموزش و آزمون به ترتیب شامل ۱۸ و ۶ ترکیب تقسیم شدند. برای انتخاب مهم‌ترین توصیف کننده‌ها روش رگرسیون مرحله‌ای مورد استفاده قرار گرفت. ارزیابی مدل‌های بدست آمده توسط سری آزمون انجام گرفت. مجذور ضریب همبستگی برای سری آزمون برای مدل‌های RF وANN به ترتیب برابر با ۰/۹۸۲ و ۰/۸۹۶ به دست آمد. نتایج به دست آمده، توانایی پیش‌بینی مناسب مدل‌های پیشنهادی را در ارزیابی فعالیت ضدHIV مشتقات فوق نشان می‌دهد.

کلمات کلیدی: QSAR، ضدHIV، شبکه عصبی مصنوعی، جنگل‌های تصادفی، سهم گروه، رگرسیون خطی مرحله‌ای

نتایج حاصل از این پایان‌نامه در پوستری تحت عنوان

“Application of random forest (RF) for the modeling of Anti-HIV activity of some dibenzyle pyridinon derivatives, as HIV-1 integrase inhibitor, using group contribution descriptors”

در پنجمین سمینار دوسالانه کمومتریکس ایران دانشگاه تهران در آذرماه ۱۳۹۴ ارائه گردید.

فهرست مطالب

فهرست اشکال

فهرست جداول

۱- فصل اول

مقدمه

۱-۱- ایدز

سـندرم نقص ایمنی اکتسـابی [۱](AIDS) یک بیماری عفونی کشنده است که اولین بار در سال ۱۹۸۱ شـناسایی شد. ویروس ایدز سلول‌های ایمنی موجود در خون (لنفوسیت‌ها) و سلول‌های ایمنی موجود در بافت‌ها مانند مغز اسـتخوان، طحال، کبد و گره‌های لنفاوی را درگیر می‌سـازد. این سلول‌ها در تولید پادتن برای مقابله با بیماری‌ها و سرطان‌ها نقش دارند. در مجموع باید گفت که ایدز یک نوع نقص ایمنی ثانویه است که در سیر عفونت با ویروس ایدز ایجاد می‌شود. ویروس HIV از خانواده رترو-ویروس‌ها[۲]و از زیر خانواده لنتی ویروس ها[۳] می‌باشـد. رتروویروس‌ها، گروهی از ویروس‌ها هسـتند که ژنوم آن‌ها از RNA تشـکیل شـده اسـت. بنابراین برای تکثیر خود به آنزیمی به نام آنزیم نسخه بردار معکوس[۴] وابسـته هسـتند که ژنوم RNA آن‌ها را به DNA نسـخه برداری کند تا بعد بتواند آن را به کمـک آنزیم اینتگراز[۵] وارد ژنوم میزبـان کند و به این ترتیب امکان تکثیر ویروس به وجود آید. کلمه رترو (معکوس) در نام این ویروس‌ها به همین خاطر اسـت چرا که معمولا نسـخه برداری از DNA به RNA انجام می‌شود اما در این دسته از ویروس‌ها عکس این عمل صورت می‌گیرد. رتروویروس‌ها باعث کاهش توانایی سیستم ایمنی بدن میزبان می‌شوند. علایمی که در بیماری ایدز مشاهده می‌شود مربوط به بیماری‌هایی است که در اثر نقص دستگاه ایمنی بدن تولید می‌شوند.

۱-۲- ساختمان ویروس HIV

یک ذره کاملا کروی با قطر ۱۲۰-۱۰۰ نانومتر اسـت که در میکروسکوپ الکترونی این ویروس به شـکل دوازده وجهی اسـت که با غشای دولایه لیپیدی پوشیده شده است. در سطح غشای لیپیدی

---

۱ - Acquired Immunodeficiency Syndrom

2 – Retrovirus

3 – Lentivirus

4- Reverse transcriptase

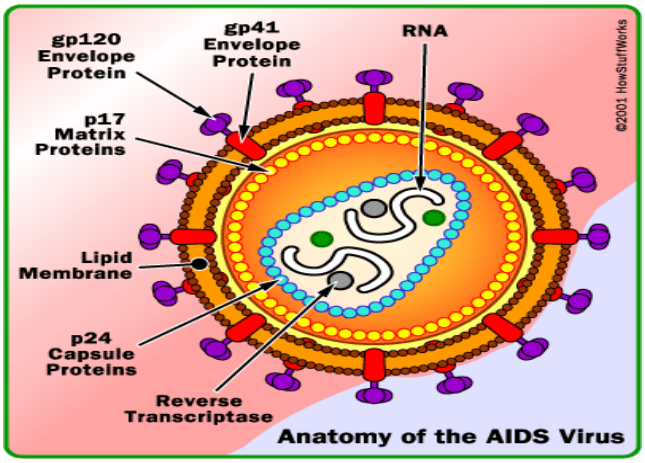
5- Integrase

دو نوع گلیکوپروتئین (gp120، gp 41) وجود دارد کـه ویروس از طریق این گلیکوپروتئین‌هـا قـادر خواهد بود به گیرنده های CD_4^+ موجود در سـطح لنفوسـیت T کمک کننده[1] بچسـبد. سـطح داخلی غشـای لیپیدی توسـط ماتریکس احاطه شـده و اطلاعات ژنتیکی مورد نیاز برای تکثیر ویروسی را در خود جای داده اسـت. این اطلاعات شامل دو کپی یکسان از ژنوم RNA ویروسی و سه آنزیم اینتگراز، پروتئاز و نسـخه بردار معکوس می‌باشـد [۱]. ویریون[2] (یک ذره ویروسی که توان آلوده کردن سلول را دارد) از سـطح سـلول های آلوده جوانه زده و به محیط بیرون آزاد می‌شـود و به همین علت است که ویروس HIV بطور مداوم پوشـش خود را عوض می کند و به هنگام جوانه‌زدن از غشای سلول میزبان انواع متفاوتی پروتئین و گلیکوپروتئین جدا می‌کند که ترکیب آن منحصـر به فرد می‌باشـد. در شکل ۱-۱، به طور واضح لایه‌ها و هسته‌ی ویروس قابل مشاهده است. دو قطعه RAN به صورت جدا از هم در مرکز ویروس مستقر هستند. این ویروس ها حاوی آنزیم ترانس کریپتاز معکوس هستند و به کمک همین آنزیم است که پس از ورود به داخل سلول می توانند از روی RNA ژنوم دو رشته ای DNA را سـنتز کنند. پس از این مرحله DNA سنتز شده، به هسته سلول میزبان رفته و توسط آنزیم‌های خود در DNA سـلول میزبان رخنه می‌کند و سیستم همانند سازی کننده میزبان را وادار به همانند سازی از ژنوم خود می‌کند. در مراحل بعدی اجزای ژنوم ویروس HIV به سرعت سنتز شده و با پدیده جوانه زدن از سـلول میزبان سـبب آسیب‌های جدی به غشای سلول میزبان شده و در نهایت سلول تخریب می‌شود. با توجه به این‌که سلول‌های میزبان ویروس، گلبول‌های سفید هستند، این اتفاق سبب کاهش

[1] - T helper lymphocyte

[2] -Virion

شدید این سلول‌ها شده و موجب تضعیف و ایجاد نقص در سیستم ایمنی بدن می‌شود.



شکل ۱-۱: ساختمان ویروس HIV

۱-۳- انواع HIV

در حال حاضر دو گونه از HIV وجود دارد:

- ✓ HIV-1
- ✓ HIV-2

نوع غالب در جهان HIV-1 می‌باشد و هنگامی که بطور کلی و بدون مشخص کردن نوع درباره HIV بحث می‌شود، منظور HIV-1 می‌باشد. راه‌های انتقال هر دو نوع ویروس HIV یکسان است هرچند که ویروس HIV-2 مشکل‌تر از از HIV-1 منتقل می‌شود و فاصله زمانی بین آلوده شدن با HIV-2 تا ایجاد بیماری طولانی تر است.

۱-۴- میزان پایداری ویروس HIV

ویروس HIV در اثر خشک شدن به سرعت غیرفعال می‌شود. در محیط کشت عاری از سلول و در درجه حرارت اتاق تا پانزده روز فعال بوده و در همان محیط در درجه حرارت ۳۷ درجه تا یازده روز زنده می‌ماند. در درجه حرارت ۶۰-۵۰ درجه سانتیگراد به مدت سی دقیقه عفونت زایی خود را از

دست می‌دهد. همچنین این ویروس در اثر استفاده از بعضی عوامل به مدت ده دقیقه در درجه حرارت اتاق به طور کامل غیر فعال می‌شـــود. این عوامل شـــامل هیپوکلریت ســـدیم ۰/۵ ٪ ، اتانول ۷۰ ٪، پارافرمالدهید ۰/۵ ٪ ، و قرار گرفتن در pH بیشـتر از ۱۰ می‌باشـد [۲]. از طرفی بدون درمان دارویی، بطور متوســط طی ۱۰ ســال آلودگی به ویروس HIV به ســمت ایدز پیش‌می‌رود که البته این مدت ۱۰سـال برای فردی اسـت که تغذیه مناسـبی دارد. اما فردی که در منطقه فقیر نشین است و بخوبی تغذیه نمی‌شود ممکن است بسیار سریعتر به سمت ایدز و نهایتا مرگ پیش برود.

۱-۵- سیر بیماری و علایم آن

ابتلا به بیماری ایدز در اغلب موارد با عوارض ســاده‌ای مانند اســهال، خارش پوســت و افزایش خفیف حرارت بدن شـروع می‌شـود. این عوارض پس از چند هفته خودبخود برطرف شـده و شـخص خیال می‌کند که به سرماخوردگی دچار بوده است. از آنجا که ساخته شدن پادتن‌های موثر حدود ۱۴ روز طول می‌کشد در این زمان جدال بین پادتن‌ها و ویروس‌ها در جریان است. پس از سپری شدن این مرحله، دورانی فرا می‌رسـد که حال شخص بیمار بهبود یافته اما در خون او پادتن ضد ایدز را می‌توان تشخیص داد و از خون وی ویروس ایدز را بدست آورد.

لنفوسـیت T_4 که آلوده به ویروس شـده اسـت شروع به تقسیم شدن می‌کند و همین امر باعث تورم غدد لنفاوی می‌شـــود. تورم غدد لنفاوی گاه ماه‌ها یا ســـال‌ها بعد از ابتلای اولیه پدیدار شـــده و می‌تواند مدت‌ها به همان حال باقی بماند و در بسیاری از موارد ناراحتی‌های جدی دیگری به آن اضافه می‌شـود. بیماران اکثرا به شدت لاغر می‌شوند چون دیواره روده آن‌ها مواد غذایی را بطور کامل جذب نمی‌کند. مرحله بعدی که در آن تصـویر نهایی بیماری ایدز به نمایش گذاشته می‌شود حاصل از درهم شـکسـته شدن کامل سیستم دفاعی بدن است. تعداد سلول‌های لنفوسیت T_4 به شدت کاهش یافته و در این حالت انواع بیماری‌های باکتریایی، ویروسی و انگلی قادر هستند که فرد مبتلا را از پای درآورند.

۱-۶- تشخیص بیماری

تشـخیص این بیماری از طریق آزمایش خون امکان‌پذیر اسـت. سـه نوع کلی آزمایش تشخیص HIV وجود دارد که عبارتند از:

۱. تست آنتی بادی HIV[۱] : این تست نشان می‌دهد که فرد به HIV آلوده است یا نه.
۲. تسـت آنتی ژن P24[۲] : این تسـت بطور اولیه برای غربالگری نمونه‌های خون استفاده می‌شود ولی در برخی مناطق آن را بعنوان تست تشخیص HIV به‌کار می‌برند. آنتی ژنP24، یک پروتئین است که جزئی از سـاختمان HIV بوده و در مراحل اولیه عفونت، به مقدار زیاد تولید می‌شود و بوسیله تسـت‌های تشـخیصـی می‌توان آن را در خون آشکار کرد. تست P24 می‌تواند آلودگی به HIV را قبل از تسـت آنتی بادی HIV آشـکار کند. بنابراین، تسـت آنتی ژنP24 در تشـخیص HIV در مراحل اولیه به کار می‌رود.
۳. تسـت میزان ویروس[۳] : این تسـت هنگامی استفاده می‌شود که شخص از آلوده بودن خود با HIV آگاه است و با این تست میزان ویروس در خون مشخص می‌شود.

۱-۷- درمان

در حال حاضـر داروها در معالجه عفونت ایدز موثر نیستند. برای مقابله با عفونت‌ها یا پیشگیری از آنها ممکن است آنتی بیوتیک تجویز شود. داروهای ضد HIV در سه دسته قرار می‌گیرند:

---

[۱] -HIV antibody test

[۲] -P24 Load test

[۳] - HIV Load test

۱) مهـار کننـده نوکلئوزیـدی ترانس کریپتـاز معکوس[۱]: زیـدووودین[۲]، لامیوودین[۳]، دیـدانوزین[۴]، زالسیتابین[۵] و استاوودین[۶] [۳].

۲) مهارکننده غیرنوکلئوزیدی ترانس کریپتاز معکوس[۷]: نویراپینف دلاوردین و افاویرنز [۴].

۳) مهار کننده پروتئاز: ایندیناویر، ساکویی ناویر، نلفیناویر و آمپرناویر.

این داروها ممکن اسـت پیشرفت بیماری را به تاخیر اندازند و تجویز آنها باید توسط پزشکان با تجربـه انجـام پـذیرد. بـا این داروها (مثل زیدووین) می توان انتقال عفونت به نوزادان مادران باردار را کاهش داد. تحقیقات زیادی در رابطه با داروهای جدیدتر و واکسـن آن از طریق مهندسـی ژنتیک در دست انجام هستند.

۱-۸- هدف‌های استفاده از داروهای ضدHIV

۱. سرکوب ویروس و کاهش روند تکثیر آن‌ها
۲. تحکیم و بهبود عملکرد سیستم ایمنی
۳. بهبود کیفیت زندگی بیماران با کاهش عوارض بیماری
۴. بهبود امید به زندگی و افزایش طول عمر بیماران
۵. کاهش احتمال سرایت

---

[۱] -Nucleoside Reverse Transcriptase Inhibitors (NRTIs)

[۲] -Zidovudine

[۳] -Lamivudine

[۴] -Didanosine

[۵] -Zalcitabine

[۶]-Stavudine

[۷] -Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs)

۱-۹- دلایـل اســتفـاده از روش‌هــای تئوری برای پیش‌بینی فعالیت داروها

امروزه پیشـــرفـت علم باعث بهبود روش‌های آزمایشـــگاهی گردیده اســـت. اما به علت وجود محدودیت‌هایی همچون عدم قطعیت نتایج آزمایشگاهی، هزینه بر بودن، وقت گیر بودن و در دسترس نبودن امکانات، توســـعه روش‌های نظری[۱] به عنوان یک ابزار کارآمد برای رفع این محدودیت‌ها مورد توجه قرار گرفته اســـت. به‌کارگیری این روش‌ها می‌تواند علاوه بر پیش‌گویی فعالیت بیولوژیکی مورد نظر، به هدفمندتر شـــدن انجام آزمایش‌های تجربی و همینطور توضـــیح پارامترهای موثر بر نتایج این آزمـایش‌هـا و مکـانیســـم‌هـای درگیر کمک کند. از جمله روش‌های نظری، رابطه کمی ســـاختار – فعالیت[۲](QSAR) اســـت که در آن می‌توان از روش‌های مختلفی برای ایجاد روابط خطی و غیرخطی میان ســـاختار و خاصـــیت دارویی ترکیباتی که فعالیت دارویی آن‌ها با اســـتفاده از روش‌های تجربی اندازه‌گیری شده‌اند، استفاده نمود.

۱-۱۰- مروری بر کارهای انجام شده

جلالی هروی و پرســتار در ســال ۲۰۰۰ در چارچوب مطالعات QSAR برای ایجاد مدل‌هایی که ارتباط دهنده‌ی فعالیت ضـد HIV با سـاختار ۱۰۷ مشـتق (۱-۲- (هیدروکسـی اتوکسی) متیل)-۶- (فنیل تیو) تیمین (HEPT) باشـد، مطالعاتی انجام دادند. در این کار مدل‌سـازی توسـط شبکه عصبی مصـــنوعی(ANN)[۳] نســـبت به روش رگرســـیون خطی چندگانه (MLR)[۴] نتایج بهتری را ارائه داد به

---

[۱] -Theoretical approaches

[۲] -Quantitative Structure- Activity Relationship

[۳] - Artificial Neural Network

[۴] - Multiple Linear Regression

طوری که میانگین مربع خطا (MSE)[1] برای سری تست با استفاده از روش خطی چندگانه ۰/۷۸۰و با استفاده از شبکه عصبی مصنوعی ۰/۳۷۲ به دست آمد [۵].

دوئلی[2] و همکارانش در سال۲۰۰۴ ارتباط کمی ساختار- فعالیت ۸۰ ترکیب از مشتقاتHEPT که بازدارنده آنزیم نسخه بردار معکوس در ویروس HIV می‌باشند، را با استفاده از شبکه عصبی مصنوعی مورد بررسی قرار دادند و نتایج این مدل را با روش‌های رگرسیون خطی چندگانه (MLR) و حداقل مربعات جزئی (PLS)[3] که توسط لوکو[4] و فرتی[5] ارائه شده بود، مقایسه کردند و از مقادیر مجذور ضریب همبستگی، قابلیت پیش‌بینی بهتر مدل ارائه شده توسط ANN را نتیجه گرفتند [۶].

عرب چم جنگلی در سال ۲۰۰۹ با استفاده از روش غیرخطی بر مبنای شبکه عصبی مصنوعی، مدلی مناسب برای فعالیت ضد HIV مشتقات پیرول آریل سولفون‌ها ارائه داد به طوری که ضریب همبستگی برای سری تست ۰/۹۸۹بود [۷].

در سال ۲۰۱۰ پاتل[6] و همکارانش مطالعه QSAR را بر روی ۴۶ مشتق دی‌هیدرو- آلکوکسی- بنزیل- اکسوپیریمیدین (DABOs) که به عنوان بازدارنده‌های غیرنوکلئوزیدی ویروس HIV عمل می‌کنند، انجام دادند. مدل نهایی به دست آمده شامل ۴ توصیف‌کننده بود و با توجه به علامت ضرائب این توصیف‌کننده‌ها، نتیجه گرفتند که استخلاف‌های آبگریز و استخلاف‌های با شاخه بزرگتر، تأثیر مثبتی بر بازداری آنزیم نسخه‌بردار معکوس دارند [۸].

---

[1] - Mean Square Error

[2] - Douali

[3] - Partial Least Square

[4] - Luco

[5] - Ferreti

[6] - Patel

عرب چم جنگلی و اشـرفی در سال ۲۰۱۳ با استفاده از روش غیرخطی بر مبنای شبکه عصبی مصنوعی، مدلی مناسب برای فعالیت مهارکنندگی نکروپتوز مشتقات [۱و۲و۳] تیادیازول و تیوفن ارائه دادند. قابلیت پیش‌بینی مدل ساخته شده که شامل ۷ توصیف‌کننده بود، توسط رد مرحله‌ای تک تک مورد ارزیابی قرار گرفت که ضریب همبستگی محاسبه شده برابر ۰/۹۴۴۶ بود [۹].

عرب چم جنگلی و عجم در سـال ۲۰۱۵ با اسـتفاده از توصـیف‌کننده‌های سـهم گروه و روش جنگل تصـادفی، مدلی مناسـب برای فعالیت بازدارنده‌های غیرنوکلئوزیدی مشـتقات تیوکربامات ارائه دادند. قابلیت پیش‌بینی مدل ساخته شده که شامل ۱۲ توصیف‌کننده بود توسط رد مرحله‌ای تک‌تک و با اسـتفاده از سـری آزمون مورد ارزیابی قرار گرفت و مجذور ضـریب همبستگی محاسبه شده برای پیش‌بینی خاصیت ضد HIV برابر ۰/۸۷۴۸ بود [۱۰].

۱-۱۱- ضرورت تحقیق

از آنجا که ایدز هم یکی از دلایل اصلی مرگ و میر در جهان است نیاز به درمان توسط مواد دارویی برای آن امری ضروری است. هرچند که جهش‌های[۱] مکرر ویروس HIV سبب افزایش مقاومت-دارویی[۲] و کاهش کارایی درمان توسط داروهای شیمیایی شده است، ولی نتایج تحقیقات نشان می‌دهد، اگر دو یا چند داروی ضد رتروویروسی همزمان مصرف شوند، مقاومت دارویی به میزان قابل توجهی کاهش می‌یابد. بنابراین درمان ترکیبی[۳] سبب گسترش تحقیقات جدید در جهت کشف داروهای گزینش پذیرتر و قویتر شده است و از آنجایی که بیشتر داروهای شیمیایی توسط آزمون و خطا سنتز می‌شوند بسیار وقت گیر و هزینه بر هستند [۱۱،۱۲].

در این پایان نامه سعی شده است با به‌کار گیری روش‌های QSAR مدل‌هایی از جمله روش جنگل‌های تصادفی[۴] و شبکه عصبی مصنوعی[۵] برای مدل‌سازی فعالیت دارویی و پیش‌بینی فعالیت بیولوژیکی داروهای ضد ایدز ارائه شود.

---

[۱] -Mutations

[۲] -Drug-Resistance

[۳] -Combination Therapy

[۴] -Rndom Forests

[۵] - Artificial Neural Networks

۲- فصل دوم

تئوری

۲-۱- کمومتریکس[۱]

کمومتریکس عبارت است از کاربرد روش‌های ریاضی و آمار به منظور برقراری ارتباط بین سنجش‌های انجام شده روی یک سیستم یا فرایند شیمیایی، طراحی عملیات آزمایشگاهی، دستیابی حداکثر به اطلاعات، انتخاب روش‌های بهینه اندازه‌گیری برای تفسیر و تحلیل بهتر اطلاعات شیمیایی [۱۳].

کمومتریکس شاخه ای نوپا در بین علوم می‌باشد. هرچند که نام آن از شیمی گرفته شده است ولی در بین شاخه‌های دیگر علوم هم دارای کاربردهای گسترده می‌باشد. کمومتریکس از آمار، جبرخطی و کامپیوتر برای آنالیز داده های شیمیایی استفاده می‌کند و هدف استخراج بیشترین اطلاعات از داده‌های شیمیایی می‌باشد. از نظر شیمی‌دانان فعالیت‌ها و خواص یک ترکیب ناشی از ویژگی‌های ساختاری آن است. هرگاه مطالعات به‌صورت ارتباط بین ساختارمولکولی و خواصی از نوع بیولوژیکی (مانند فعالیت دارویی) به‌کار گرفته شود، اشاره به ارتباط کمی ساختار-فعالیت (QSAR) خواهد داشت [۱۴].

در مطالعات QSAR سعی بر این است تا رابطه‌ی هماهنگ میان فعالیت‌های شیمیایی و فیزیکی با ویژگی‌های مولکولی پیدا شود، به گونه‌ای که بتوان این قواعد را برای ارزیابی فعالیت ترکیبات جدید به کار برد. در واقع نتایج این نوع مطالعات علاوه بر شفاف سازی نحوه‌ی ارتباط بین خواص مولکول‌ها و ویژگی‌های ساختاری آن‌ها، به پژوهشگران در پیش بینی رفتار مولکول‌های جدید براساس رفتار مولکول‌های مشابه کمک می‌کند [۱۵].

یکی از زمینه‌های مهم کاربرد کمومتریکس در مطالعاتی است که خواص مولکول‌ها را به

---

[۱] - Chemometrics

ویژگی‌های ساختاری آن‌ها نسبت می‌دهد. موارد خاصی از این تحقیقات و مطالعات شامل رابطه کمی ساختار- فعالیت (QSAR)[۱]، رابطه کمی ساختار- خصوصیت (QSPR)[۲]، رابطه کمی ساختار- سمیت[۳](QSTR) و رابطه کمی ساختار- بازداری[۴](QSRR) است. فرض اساسی در مطالعات QSAR این است که ساختار یک مولکول (یعنی خواص هندسی، الکترونی و چیدمان فضایی اتم‌ها) حاوی ویژگی‌هایی است که مسئول خواص فیزیکی، شیمیایی و زیستی ترکیبات شیمیایی می‌باشند. این ویژگی‌ها که به صورت کمی هم قابل نمایش هستند تحت عنوان توصیف کننده[۵] شناخته می‌شوند [۱۶].

برخی از کاربردهای کمومتریکس در شیمی تجزیه عبارتند از:

۱) شناسایی الگوها از داده‌های کروماتوگرافی
۲) پیش بینی خواص و فعالیت مولکول‌ها
۳) شناخت الگوها و طبقه بندی داده‌ها
۴) شبیه سازی و جستجوی طیف‌های جرمی
۵) پیش بینی میزان جا به جایی شیمیایی
۶) پیش بینی ضرایب گزینش‌پذیری الکترودهای یون‌گزین

۲-۲- ارتباط کمی ساختار- فعالیت (QSAR)

مدل‌سازی QSAR در حوزه سم‌شناسی متولد شد. در واقع تلاش‌ها برای کمی کردن روابط بین

---

[۱] -Quantitative Structure Activity Relationship

[۲] - Quantitative Structure Property Relationship

[۳] - Quantitative Structure Toxicity Relationship

[۴] - Quantitative Structure Retention Relationship

[۵] -Descriptor

ساختار شیمیایی و پتانسیل سمیت بحرانی قسمتی از دستاوردهای منتشر شده سم‌شناسی برای بیش از ۱۰۰ سال بود. در حال حاضر علم QSAR به شاخه‌های مختلف شیمی، بیوشیمی، شیمی پزشکی، داروسازی و شیمی زراعی راه یافته است و به ابزاری اساسی در طراحی دارو و آفت‌کش تبدیل شده و نقش رو به رشدی در علوم زیست محیطی دارد [۱۶].

مولکول‌های مشابه با تغییر کوچکی در ساختارشان می‌توانند فعالیت بیولوژیکی کاملا متفاوتی داشته باشند [۱۷].

بررسی بیولوژیکی فرآورده‌های طبیعی و برخی از مشتقات آن‌ها که خواص دارویی دارند، انگیزه‌ای جهت پیدا کردن وابستگی‌های احتمالی ساختار شیمیایی با فعالیت بیولوژیکی ایجاد نمود و مشخص شد که برخی از واحدهای ساختار مولکول‌هایی که از نظر بیولوژیکی فعال هستند، در مولکول‌های دیگری که دارای همان ویژگی بیولوژیکی می‌باشند، وجود دارد. این امر رهنمودی برای طراحی ساختار ترکیباتی با فعالیت بیشتر و در عین حال عملکردی اختصاصی‌تر بود.

در حال حاضر اصلاح و تغییر ساختار یک ترکیب الگو، عمده ترین روش و روند دستیابی به داروهای جدید می‌باشد. تغییرات ساختاری منجر به پیدایش ویژگی‌های جدید می‌شود که در نهایت موجب تغییر پخش دارو در سلول بافت، تغییر در کیفیت اثر متقابل با جایگاه‌های فعال آنزیم و گیرنده‌ها و در نهایت تغییر سرعت واکنش‌ها در جایگاه‌های مورد نظر می‌گردد [۱۸].

۲-۳- کاربردهای QSAR در طراحی دارو

QSAR به راحتی می‌تواند برای کمک به بهینه‌سازی سنتز داروها مورد استفاده قرار گیرد. تحقیقات برای پیشنهاد داروهای جدید در روش‌های داروسازی سنتی تا حد زیادی برپایه‌ی آزمون و خطا و استخراج موادی با خاصیت دارویی از طبیعت می‌باشد. پس از شناسایی اصول مولکولی حیات، دانشمندان به اطلاعاتی ارزشمند در مورد نحوه‌ی اثر مولکول‌های دارویی بر بدن دست یافتند. در این

تلاش‌ها مشخص شد که اثر دارویی هر ترکیب به ویژگی‌های ساختاری آن ترکیب وابسته است. هریک از این ویژگی‌های ساختاری باعث بروز یکی از آثار دارویی مربوط به آن می‌شود. در علم داروسازی مدرن با تکیه بر روش‌های QSAR، تولید دارو به یک روند تحقیقاتی کاملا جهت‌دار تبدیل شده است. برای انجام یک مطالعه QSAR ابتدا باید مقادیر فعالیت دارویی ترکیبات از طریق تجربی بدست آید و همچنین توصیف‌کننده‌های مناسب از طریق تجربی یا با محاسبه توسط کامپیوتر تولید شوند. مرحله بعد دستیابی به ارتباط ریاضی بین ویژگی‌های ساختاری و مقادیر فعالیت است.

۲-۴- مراحل QSAR

به طورکلی ایجاد یک مدل QSAR شامل ۵ مرحله اصلی زیر است:

۱) جمع آوری و پیش پردازش[۱] داده‌ها
۲) بهینه سازی ساختار مولکول‌ها
۳) محاسبه و غربال‌گری توصیف کننده‌ها و انتخاب متغیر
۴) مدلسازی
۵) ارزیابی مدل

۲-۵- جمع آوری و پیش پردازش داده‌ها

اولین مرحله مدل‌سازی جمع‌آوری و انتخاب یک سری مولکولی است که مقادیر تجربی فعالیت مورد نظر آن‌ها باید با صحت قابل قبول اندازه‌گیری شده باشد. فاکتور محدود کننده در توسعه مدل‌های QSAR در دسترس بودن داده‌های تجربی با کیفیت بالا است. در واقع باید اعتبار و صحت هر مدل QSAR به صحت تعیین تجربی فعالیت مورد نظر بستگی دارد. داده‌های تجربی می‌توانند در

[۱] -Preprocessing

آزمایشگاه تولید یا از مقادیر منتشر شده در مقالات استخراج شوند. لازم به ذکر است که فعالیت ترکیبات باید در شرایط آزمایشگاهی یکسان تعیین شده باشد. سری مولکولی ممکن است شامل یک سری متجانس از ترکیبات شیمیایی باشد یا طبقه‌ای از مواد شیمیایی با یک تنوع ساختاری موجه را در برگیرد. این تنوع می‌تواند منجر به عمومیت یافتن بیشتر مدل شود. برای اجتناب از تغییرپذیری اضافی درون داده‌ها به علت استفاده از منابع اطلاعاتی مختلف، داده‌ها باید فقط از یک منبع یا منابع مشابه استخراج شوند یا به عبارتی دیگر، پیوستگی داده‌ها باید رعایت شود. بعد از اینکه سری داده جمع آوری شد لازم است که روی آن‌ها پیش پردازش انجام شود. این مرحله وابسته به طبیعت ، مقدار و کیفیت داده‌ها است. پیش پردازش داده‌ها شامل حذف داده‌های نامربوط، نامطمئن و با نویز زیاد است که می‌تواند تاثیر منفی روی کل فرآیند مدل‌سازی داشته باشد. این روش‌ها بستگی به اهداف مدلسازی و دانش اولیه درباره خصوصیات داده‌ها دارد.

در مدل‌سازی QSAR سری داده‌ها عموما به دو دسته سری آموزشی[۱] و سری آزمون[۲] تقسیم می‌شود. عملیات مدل‌سازی بر روی سری آموزشی انجام می‌شود. از سری آزمون نیز جهت بررسی قدرت پیش‌بینی و اعتبار مدل استفاده می‌شود. سری آزمون به نحوی انتخاب می‌شود که نماینده جمعیت مولکول‌های سری آموزشی باشد.

۲-۶- بهینه سازی ساختار مولکول‌ها

کمومتریکس از روش‌های مفیدی که قادر به استخراج اطلاعات موجود در داده‌ها هستند استفاده می‌کند [۱۹]. بدین منظور از شیمی محاسباتی[۳] استفاده می‌شود. شیمی محاسباتی

---

[۱] -Training set

[۲] -Test set

[۳] -Computational chemistry

ســـاختارهای مولکولی را به صـــورت پارامترهای عددی معرفی و رفتار آن‌ها را با معادلات کوانتومی و فیزیک کلاسـیک شبیه سازی می نماید. این امر به دانشمندان امکان می‌دهد که بتوانند از این طریق به اطلاعات مولکول از جمله ســاختار هندســی، انرژی و خواص الکترونیکی و اثرات حلال دســت پیدا کنند. در سـال‌های اخیر اسـتفاده از روش‌های محاسـباتی در بین شیمیدانان رواج یافته است. در این روش‌ها می‌توان به راحتی محاســبات را انجام داد، بدون اینکه از اصـول اولیه و روش محاســبه آگاهی دقیق داشـت. روش‌های محاسـباتی در شـیمی به چندین دسـته تقسـیم می‌شـوند. روشی که در این تحقیق به کار گرفته شـده، روش AM1[۱] اسـت که جز روش‌های محاسباتی نیمه تجربی[۲] می‌باشد. در روش‌های نیمه تجربی، که در برنامه‌هایی نظیرHyperChem وارد شـده‌اند، محاسبات براساس مکانیک کوانتومی صـــورت می‌گیرد. در این روش‌هـا یـک ســـری از مقادیر تجربی نیز وارد شـــده‌اند و فقط الکترون‌های لایه ظرفیت در محاســـبه در نظر گرفته می‌شـــوند. در این تحقیق از MNDO[۳] یا روش اصلاح شده صرفنظر از همپوشانی‌های تفاضلی استفاده شده است.

روش AM1 همان روش MNDO اصـلاح شـده اسـت که روی مولکول‌های قطبی نتایج خوبی می‌دهد ولی پیوند هیدروژنی را نمی‌تواند در نظر بگیرد. در این روش هسته و لایه‌های داخلی به شکل یک هسته مرکزی[۴] در نظر گرفته می‌شود و محاسبات روی الکترون‌های ظرفیت انجام می‌گردد [۱۴و ۲۰]. AM1 معمولا گرمای تشـــکیل و انرژی اکتیواســـیون را بســـیار دقیق‌تر از MNDO پیش بینی می‌کند. از جمله ویژگی بارز روش نیمه تجربی این اســت که زمان محاســبات در این روش‌ها کوتاه‌تر است.

---

[۱] -Austin Methods

[۲] -Semi-emprical methods

[۳] -Modified neglected of differential overlap

[۴] -Core

۲-۷- محاسبه و غربالگری توصیف کننده ها

توصیف کننده‌ها مقادیر عددی هستند که ویژگی‌های مختلف ساختاری و الکترونی مولکول‌ها را به طور کمی نشـان می‌دهند. یعنی هر توصـیف‌کننده بیانگر خصـوصـیت ویژه‌ای از مولکول اسـت که ممکن است بر خاصیت مورد نظر تاثیر گذار باشد. توصیف کننده‌های مولکولی به دو دسته کلی تقسیم می‌شـوند: توصیف کننده‌های حاصل از اندازه‌گیری‌های تجربی (مانند قطبش پذیری و ممان دوقطبی) و توصـــیف کننده‌های مولکولی نظری (تئوری) که از ســـاختار مولکول مشـــتق شـــده و می‌توانند به دسته‌های بیشتری مطابق با انواع مختلفی از نمایش مولکولی[۱] یا بر اساس ابعاد توصیف کننده مولکولی تقسیم شوند.

محاســـبه توصـــیف کننده‌ها یکی از مهمترین مراحل مدل‌ســـازی و تعیین و انتخاب توصـــیف کننده‌ها می‌باشـد. توصـیف کننده‌های مولکولی نتیجه نهایی یک اسـتدلال و روش ریاضی هستند که اطلاعات شـیمیایی را به رمز تبدیل می‌کنند و آن‌ها را به صـورت یک نماد نشـان می‌دهند طوری که ارائه دهنده ویژگی‌های یک مولکول به صـورت یک عدد مفید باشند. در مدل‌سازی QSAR از توصیف کننده‌هایی باید اسـتفاده شـود که در ارتباط با ویژگی مورد نظر مولکول بوده و الگوریتم محاسـباتی آن‌ها تعریف واضـحی داشته باشد و همچنین تفسیر آن‌ها واضح باشد. همچنین باید تغییرات تدریجی در سـاختار مولکول به تغییرات تدریجی در مقادیر توصیف کننده منجر شده و محدود به طبقه بسیار کوچکی از ساختارهای مولکولی نباشد. به طور کلی توصیف کننده‌ها براساس چگونگی بیان خصوصیات مولکول به چند دسته کلی تقسیم می‌شوند که به تعدادی از آن‌ها در زیر اشاره می‌شود [۲۱]:

توصـیف کننده‌های ساختاری[۲]: این سری توصیف کننده‌ها عمومی ترین توصیف کننده‌های

---

[۱] -Molecular representation

[۲] -Constitutional descriptors

مورد اســتفاده می‌باشــند که ساختار ترکیب را بدون هیچگونه اطلاعاتی درباره هندسه مولکولی آن در اختیار قرار می‌دهند.

توصیف کننده‌های توپولوژیکی[۱]: جزء ساده‌ترین نوع توصیف کننده‌ها می‌باشند و به ساختار فضایی مولکولی ارتباطی نداشته و تنها به نوع اتم‌ها، نوع پیوندها و نحوه ارتباط اتم‌ها به یکدیگر وابسته اند این پارامترها را می‌توان بدون بهینه کردن ساختار مولکول محاسبه کرد.

توصیف کننده‌های هندسی[۲]: نسـبت به نوع توپولوژیکی از اهمیت بیشتری برخوردارند. یکی از خصـوصیات بارز این توصیف کننده‌ها، ارتباط آن‌ها با ساختار سه بعدی مولکول‌ها می‌باشد. بنابراین قبل از محاسبه این توصیف کننده‌ها باید پایدارترین ساختار فضایی مولکول تعیین شود.

توصـــیف کننده‌های الکترونی[۳]: بیانگر خواص الکترونی مولکول‌ها می‌باشـــند و مقادیر آن‌ها اطلاعات مفیدی راجع به برهمکنش‌های الکترونی بین مولکولی و درون مولکولی ارائه می‌دهند.

توصـــیف کننده‌های فیزیکو...شـــیمیایی[۴]: بیانگر بعضـــی از خواص فیزیکی و شـــیمیایی مولکول‌ها می‌باشــند و به ساختار مولکول وابستگی شدیدی دارند که از جمله می‌توان به نقطه ذوب و نقطه جوش اشاره نمود.

توصـــیف کننده‌های تئوری بدون نیاز به داده‌های تجربی به کمک نرم افزارهای مناســـب برای طیف وسـیعی از ترکیبات قابل محاسـبه هسـتند. بنابراین توصیف کننده‌های تئوری در این مطالعات بسـیار با اهمیت می‌باشــند. برای محاسبه توصیف کننده‌ها، ابتدا مولکول‌ها با استفاده از نرم افزارهایی

---

[۱] -Topological descriptors

[۲] -Geometrical descriptors

[۳] -Electronic descriptors

[۴] -Physico-Chemical descriptors

مانند Hyperchem، PC Model، Chemoffice رسم شده، ساختار مولکول‌ها به وسیله روش مکانیک کوانتومی[۱] و یا روش‌های نیمه تجربی[۲] بهینه می‌شوند. سپس خروجی این نرم افزارها که حاوی اطلاعات سه بعدی ساختارهای مولکولی مورد نظر است به عنوان ورودی نرم افزارهای مخصوص محاسبه توصیف کننده مورد استفاده قرار می‌گیرد.

در نوع دیگری از تقسیم بندی، توصیف کننده‌ها به دسته‌های صفربعدی، یک بعدی، دوبعدی و سه بعدی طبقه‌بندی می‌شوند. برای توضیحات بیشتر در مورد این نوع دسته بندی به مرجع [۲۲] مراجعه شود.

۲-۸- توصیف کننده‌های سهم گروه[۳]

امروزه در طراحی صنعتی، روش‌های ساده، کارآمد و در عین حال معتبر برای پیش‌بینی خواص مواد از روی ساختار مولکولی آن‌ها ضروری است. یکی از این روش‌ها، روش سهم گروه است که امروزه به‌طور گسترده‌ای در پیش‌بینی خواص فیزیکی و ترمودینامیکی مواد به کار می‌رود. این روش که به رابطه جمع‌پذیری گروه‌ها[۴] نیز معروف است، ارتباطی بین خاصیت یک ماده با ساختار شیمیایی آن برقرار می‌کند. مبنای بوجود آمدن روش سهم گروه‌ها این است که اگر چه هزاران ترکیب شیمیایی در دنیای علم و تکنولوژی وجود دارد اما تعداد گروه‌های ساختاری و عاملی سازنده‌ی همه‌ی این ترکیبات، بسیار کمتر از تعداد این ترکیبات می‌باشد. برای ایجاد توصیف کننده‌های سهم گروه دو مرحله وجود دارد: در مرحله اول خواص مواد شناخته شده به ساختار شیمیایی آن‌ها مرتبط می‌گردد تا گروه‌های اصلی در این مواد شناسایی شده و مقادیر جمع پذیر مولی آن‌ها مشخص شود. در مرحله دوم سهم

---

[۱] -Quantum mechanics

[۲] -Semi-empirical

[۳]- Group Contribution

[۴] -Group additivity relation

یک گروه مشخص در یک مولکول با تمام مولکول‌ها یکسان در نظر گرفته می‌شود. با این ایده، روش سهم گروه فرض می‌کند که خاصیت فیزیکی یا ترمودینامیکی هر ماده را می‌توان از مجموع سهم همه‌ی گروه‌های موجود در مولکول آن ماده در خاصیت مورد نظر، به دست آورد. به این ترتیب یک روش عملی برای پیش‌بینی خواص تعداد زیادی ماده بر حسب تعداد کمتری پارامتر به دست می‌آید که این پارامترها، سهم گروه‌های ساختاری را در خاصیت مورد نظر مشخص می‌کنند. اکثر این روش‌ها تجربی هستند اما اطلاعات نظری راجع به سهم گروه‌ها در خواص مواد می‌تواند به عنوان راهنما در توسعه‌ی روابط سهم گروه‌ها به کار گرفته شود. البته لازم به توضیح است که تعریف گروه‌های تشکیل دهنده‌ی هر ماده، یک مقوله اختیاری است [۲۳]. هرچه تمایز بین گروه‌ها بیشتر باشد صحت روش سهم گروه افزایش می‌یابد تا جایی که سرانجام هر ترکیب فقط شامل گروه‌های مربوط به خودش باشد. توصیف کننده‌های سهم گروه می‌تواند شامل قطعات اصلی مولکول، محل استخلاف‌ها، نوع استخلاف‌ها و... باشد. در این پژوهش نیز از توصیف کننده‌های سهم گروه استفاده شده است که از روی ساختار ترکیب و بدون نیاز به نرم افراز محاسبه شده‌اند.

۲-۹- بدست آوردن بهترین توصیف کننده‌ها برای ورود به مدل

یکی از مهمترین مراحل QSAR ، انتخاب توصیف کننده‌های مناسب می‌باشد. زیرا توصیف کننده‌های نامناسب کار برازش و مدل‌سازی را طولانی می‌کنند و تاثیری در بهبود نتایج نخواهند داشت. به این منظور باید در انتخاب توصیف کننده‌ها دقت لازم به عمل آید و توصیف کننده‌هایی که بیشترین و نزدیکترین ارتباط را به با پارامتر مورد نظر دارند، انتخاب شوند. از جمله روش‌های انتخاب

متغیر می‌توان به رگرسیون مرحله‌ای[۱]،سهم گروه[۲]، الگوریتم ژنتیک[۳] و روش جایگزینی[۴] اشاره کرد.

۲-۹-۱- انتخاب توصیف کننده به روش رگرسیون مرحله‌ای

یکی از روش‌های انتخاب توصیف کننده‌های مناسب روش رگرسیون مرحله‌ای است که با محاسبه ضرایب همبستگی[۵] مشخص می‌شوند. ضریب همبستگی میزان نزدیکی اطلاعات بدست آمده به خط رگرسیون را نشان داده که مقدار آن بین ۰ و ۱ تغییر می‌کند. "صفر" یعنی هیچ ارتباطی بین فعالیت و توصیف کننده وجود ندارد و "یک" یعنی این ارتباط کامل است. بنابراین توصیف کننده‌هایی با ضرایب همبستگی بزرگ در برازش منحنی برای ساختن مدل استفاده می‌شوند. در این روش ابتدا توصیف‌کننده‌ای که بیشترین همبستگی با متغیر وابسته (خاصیت مورد نظر) را دارد، وارد مدل شده و با ورود هر توصیف کننده جدید توصیف کننده‌های موجود در مدل بررسی شده و اگر هر کدام از آن‌ها سطح معناداری خود را از دست داده باشد قبل از ورود توصیف کننده جدید از مدل خارج می‌شود [۲۴].

۲-۹-۲- انتخاب توصیف کننده بر مبنای روش سهم گروه

انتخاب توصیف کننده‌های مناسب برای برقراری ارتباط بین ساختار مولکول و فعالیت، در ساخت مدل اهمیت بسیاری دارد. انتخاب توصیف کننده‌ها بر مبنای روش سهم گروه‌ها، یک روش بسیار ساده است که در آن ابتدا باید گروه‌های ساختاری مولکول‌ها را مشخص نمود. سپس گروه‌های ساختاری در هر ترکیب به عنوان توصیف کننده‌های آن ترکیب انتخاب می‌شود، در صورتی که هریک

---

۱ -Stepwise regression

2- Group Contribution

۳ -Genetic Algorithm

۴ -Replasement Method

۵ -Correlation coefficient

از این گروه‌های ساختاری در ترکیبی وجود نداشته باشد توصیف کننده مربوط به آن صفر در نظر گرفته می‌شود.

۲-۱۰- تقسیم‌بندی داده‌ها

قبل از ساختن مدل، سری داده‌ها معمولا به سه زیر مجموعه تقسیم می‌شوند. اولین زیر مجموعه سری آموزش[۱] می‌باشد که اکثریت داده‌ها را در برگرفته و از آن برای ساخت مدل غیرخطی استفاده می‌شود. دومین زیر مجموعه، سری ارزیابی[۲] است که در روند آموزش دخالت ندارد و مدل‌های به دست آمده از سری آموزش توسط آن ارزیابی شده و به این ترتیب توصیف‌کننده‌های بهینه برای ساخت مدل غیرخطی مشخص می‌شود. زیر مجموعه‌ی سوم سری تست[۳] است که در طول فرایند مدل‌سازی هیچ‌گونه دخالتی نداشته و از آن برای مقایسه مدل‌های مختلف استفاده می‌شود.

۲-۱۱- مدل‌سازی

مدل، در واقع یک رابطه ریاضی است که بیان کننده‌ی رابطه بین متغیر وابسته و مستقل می‌باشد و به کمک آن می‌توان با داشتن مقادیر متغیرهای مستقل، متغیر وابسته را ارزیابی کرد.

پس از انتخاب مناسب‌ترین توصیف کننده‌ها، با استفاده از روش‌های آماری مختلف به جستجوی مدل مناسبی پرداخته می‌شود که بتواند ارتباط بین توصیف کننده‌های انتخابی و پارامترهای مورد مدل‌سازی را به درستی بیان کند. برای مدل‌سازی از روش‌های گوناگون خطی و غیر خطی می‌توان استفاده کرد. در زیر به دسته بندی کلی این روش‌ها اشاره می‌شود:

---

[۱]-Training set

[۲]-Validation set

[۳]-Test set

روش‌های آماری چند متغیره (روش‌های خطی)

رگرسیون خطی چندگانه[۱] (MLR)

حداقل مربعات جزئی[۲] (PLS)

رگرسیون اجزای اصلی[۳] (PCR)

روش‌های غیر خطی

شبکه عصبی مصنوعی[۴] (ANN)

جنگل‌های تصادفی[۵] (RF)

در این پایان نامه از روش‌های غیرخطی برای مدل‌سازی استفاده شده است.

۲-۱۲- مقدمه ای بر 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 پردازش کننده 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[۶] توسط بخشی از ساختار خود به نام دندریت جمع‌آوری می‌کند. دندریت پس از دریافت اطلاعات، آن‌ها را به شکل سیگنال به هسته

---

[۱]-Multiple Linear Regression

[۲] -Partial Least Square

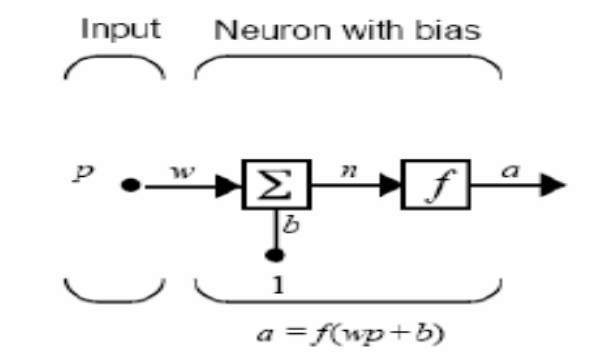
[۳] -Principle Component Regression

[۴] -Artificial Neural Networks

[۵] - Random Forests

[۶] -Neuron

سلول هدایت می‌کند. هسته پس از پردازش سیگنال‌های دریافتی، آن‌ها را ازطریق جزء دیگری به نام اکسون به نرون‌های دیگر انتقال می‌دهد. این کار توسط سیناپس‌ها که ارتباط دهنده نرون‌ها هستند، صورت گرفته و بدین ترتیب فعالیت‌های مغزی انجام می‌شود [۲۶و۲۵]. 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 [۲۷].



شکل ۲-۱: ساختمان یک نرون محاسباتی[۲۸]

۲-۱۲-۱- ساختار و عملکرد نرون مصنوعی

در شکل بالا کمیت‌های p و a به ترتیب ورودی و خروجی نرون می‌باشد. میزان تأثیر ورودی p روی خروجی a به وسیله پارامتر وزن (w) تعیین می‌شود. ورودی دیگر یک مقدار ثابت ۱ است که در جمله بایاس[1] ضرب شده و سپس با wp جمع می‌شود. این حاصل جمع، ورودی خالص n برای تابع محرک f را تشکیل می‌دهد (a= f(n)). در واقع b وw دو پارامتر تنظیم شونده در نرون‌ها می‌باشند و ایده اصلی شبکه عصبی این است که با تغییر مقادیرb وw، شبکه یک رفتار یا تصمیم را اتخاذ کند

[1]- Bias

[۲۹].

۲-۱۲-۲- تابع انتقال[1]

زمانی که ورودی‌ها با توجه به اهمیت آنها با یکدیگر جمع جبری شـوند توسـط تابع انتقال به نرون بعدی منتقل می‌شـوند. در واقع تابع انتقال یکی از اجزای شـبکه عصـبی می‌باشـد که برحسـب کاربرد می‌تواند توابع متفاوتی باشـد. دو تابع انتقال رایج در شبکه‌های عصبی که در این پژوهش برای بهینه‌سازی شبکه به کار گرفته شده است عبارتند از:

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

از این تابع انتقال در شـبکه‌های پس انتشـار اسـتفاده می‌شـود این تابع مقادیر ورودی را در محدوده منفی بی‌نهایت تا مثبت بی‌نهایت دریافت کرده و خروجی بین ۰ و۱ تولید می‌نماید.

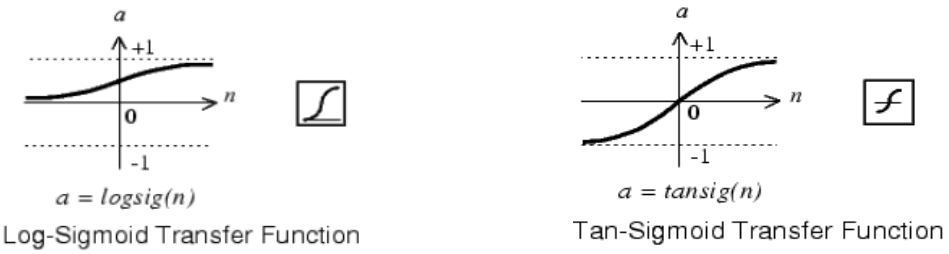
- تابع انتقال تانژانت سیگموئید[3] (tansig)

این تـابع انتقـال مقادیر ورودی را در محدوده منفی بی‌نهایت تا مثبت بی‌نهایت دریافت کرده و خروجی بین ۱+ و۱- تولید می‌کند.

[1] -Transfer function

[2] - Logarithm sigmoid transfer function

[3] - Hyperbolic tangent transfer function



شکل ۲-۲: توابع انتقال [۳۰]

۲-۱۲-۳- ساختارهای شبکه

بر اساس ساختار اتصال نرون‌ها، شبکه‌های عصبی مصنوعی را می‌توان در دو دسته‌ی کلی طبقه‌بندی کرد:

- شبکه‌های جلوسو[۱] که در آنها گراف حلقه ندارد.
- شبکه‌های برگشتی[۲] که در گراف آنها بدلیل وجود بازخورد[۳] در ساختار شبکه، حلقه به وجود می‌آید.

در معمول‌ترین خانواده‌ی شبکه‌های جلوسو که پرسپترون چندلایه[۴] نامیده می‌شوند، نرونها درلایه‌هایی قرار می‌گیرند و اتصال بین آنها یک طرفه است. اتصالات مختلف سبب رفتارهای متفاوت شبکه‌ها می‌شود. بطور کلی می‌توان گفت شبکه‌های جلوسو، استاتیک هستند، به این معنی که از ورودی داده شده تنها یک دسته مقدار خروجی تولید می‌کنند نه یک دنباله از مقادیر خروجی! شبکه‌های جلوسو بی‌حافظه هستند و پاسخ آنها به یک ورودی مستقل از وضعیت قبلی شبکه است.

۱- Feed-Forward Netwoks

۲- Recurrent Networks

۳- Feedback

۴- Multi Layer Perceptron

از طرف دیگر شبکه‌های برگشتی، سیستم‌های دینامیک هستند و زمانی که یک دنباله‌ی ورودی جدید به آنها داده شود، خروجی نرون‌ها محاسبه می‌شود.

در اثر وجود مسیرهای بازخورد، ورودی هر نرون اصلاح می‌شود و این باعث می‌شود که شبکه وارد یک وضعیت جدید شود. ساختارهای مختلف شبکه، به الگوریتم یادگیری مناسب خود نیاز دارند.

۲-۱۲-۴- آموزش شبکه‌های جلو سو با تکنیک پس‌انتشار[۱]

پس‌انتشار خطا یک روش متداول آموزش با ناظر برای شبکه‌های جلوسو است یعنی برای به دست آوردن ارتباط بین متغیرهای ورودی و خروجی در یادگیری به الگوی آموزشی نیاز است. به طور کلی آموزش به کمک تکنیک پس‌انتشار بر طبق مراحل زیر انجام می‌شود [۳۱]:

۱- انتشار ورودی‌ها از نرون‌های ورودی به سمت نرون‌های خروجی

۲-اختصاص ماتریس وزن‌های تصادفی به هریک از اتصالات

۳- مقایسه خروجی‌های شبکه با مقادیر واقعی (مقادیر هدف) و محاسبه‌ی خطای شبکه

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

۵- ارزیابی عملکرد شبکه با توجه به تابع کارآیی تعیین شده

مراحل فوق تا زمانی تکرار می‌شود که به حداکثر تکرار[۲] مجاز رسیده باشد یا مقدار تابع کارآیی از مقداری که تعیین شده کمتر باشد. شبکه عصبی انتخاب شده در این پژوهش یک شبکه جلوسو با

---

[۱] - Back propagation

[۲]- Epoch

الگوریتم آموزشی پس‌انتشار می‌باشد.

۲-۱۳- جنگل‌های تصادفی [۱] (RF)

یکی از روش‌های نوین مدل‌سازی و پیشرفته آماری، روش جنگل‌های تصادفی است که یکی از روش‌های یادگیری ماشین[۲] است. این روش در سال ۲۰۰۱ توسط بریمن[۳] متخصص آمار دانشگاه برکلی آمریکا و آدله کاتلر[۴] ارائه شد [۳۲] . قابلیت این روش، هنگامی که تعداد متغیرهای توضیحی[۵] زیاد است، بیشتر بروز می‌کند. بریمن روش درخت‌های رده بندی و رگرسیونی (CART)[۶] را به عنوان جایگزینی برای روش‌های مبتنی بر شبکه معرفی کرده است. که به ترتیب برای طبقه بندی داده‌های کیفی و پیش بینی داده‌های کمی به کار می‌روند. در این پروژه هدف پیش بینی داده‌های کمی است لذا تنها به بخش رگرسیونی روش CART با نام درخت تصمیم اشاره می‌شود.

۲-۱۳-۱- روش درخت رگرسیونی[۷] (تصمیم)

این روش بر پایه تقسیم کردن مجموعه داده‌ها به قسمت‌های کوچکتر است. در روش رگرسیون خطی، پیش بینی کننده‌های رگرسیونی، مدل‌هایی هستند که در آن‌ها یک مدل پیش بینی واحد روی فضای کل داده‌ها انجام می‌دهد. 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، تقسیم بندی

---

[۱] -Random Forests

[۲] -Machine learning

[۳] -Breiman

[۴] -Adele Cutler

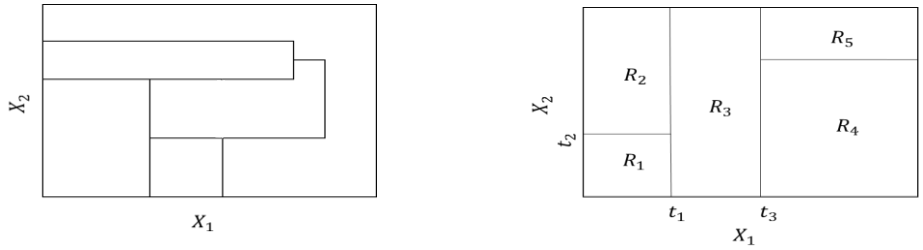
[5]- Explanatory variable

[۶] -Classification And Regression Tree

[۷]- Regression Tree

فضـای داده‌ها به بخش‌های کوچکتر اسـت تا بتوان رفتار متغیر پاسـخ را به طور موضـعی مدل سازی نمود. در روش درخت رگرسیونی هدف این است که مقادیر متغیرهای توضیحی (ابر مکعب در فضای m بعدی) در هر ناحیه، به گونه‌ای تقســـیم بندی شـــود که داده‌های واقع در هر بخش تا حد ممکن همگون باشند به طوری‌که بتوان آن‌ها را توسط ساده ترین مدل پیش بینی نمود.

به طور کلی، افراز[1] فضـای متغیرهای توضـیحی را می‌توان به هر شـکلی انجام داد. اما در روش درخت رگرســیونی، تنها افرازی قابل قبول اســت که تمامی نواحی ســاخته شــده به شــکل مربع یا مستطیل باشند. در شکل ۲-۳ نمونه‌ای از افراز قابل قبول و غیرقابل قبول در فضای دومتغیره با روش درخت رگرسیونی نشان داده شده است.

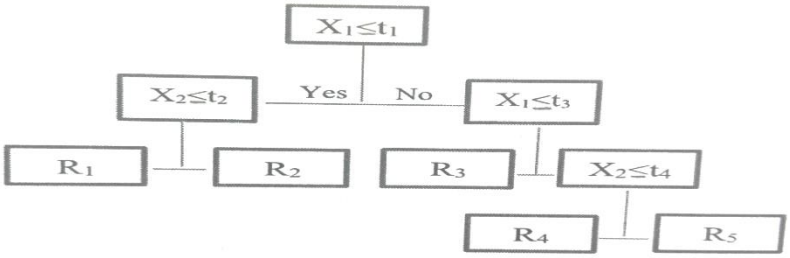


ب) الف)

شکل ۲-۳: شکل (الف) افراز قابل قبول و شکل (ب) افراز غیرقابل قبول

[1]- Partition

در شکل ۲-۴، نوع دیگری از نمایش این افراز نشان داده شده است. دلیل نامگذاری درخت رگرسیونی را می‌توان به نمایش درختی افراز فضـای متغیرهای ورودی نسـبت داد که در آن هر مشاهده از نقطه بالای نمودار درختی وارد شده و در یکی از نواحی افراز قرار می‌گیرد.



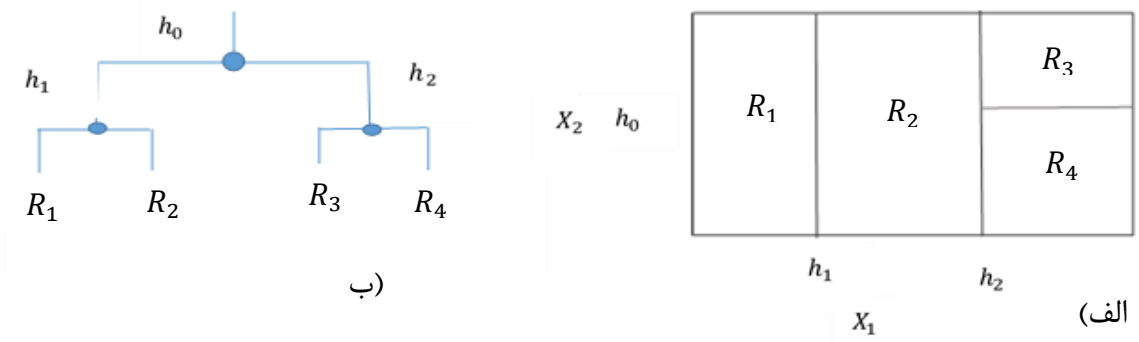
شکل ۲-۴: نمودار درختی فضای افراز شده در افراز قابل قبول شکل(۲-۳)

تقسـیم بندی فضـای مذکور توسط یک الگوریتم سلسله مراتبی انجام می‌شود و در آن به سوالات زیر پاسخ داده می‌شود:

۱) تقسـیم فضـا در راسـتای محور کدام یک از متغیرهای توضیحی و در چه مقداری از آن متغیر بایستی ایجاد شود؟

۲) معیار انتخاب راستای یاد شده چیست؟

۳) تقسیم فضا تا چه مرحله ای بایستی ادامه یابد؟



شکل ۲-۵ : الف) تقسیم بندی فضا، ب) نمودار درختی تقسیم بندی فضا [۳۳]

بـه مقـداری از هر متغیر که افراز در آن نقطه انجام می‌گیرد، گره[1] می‌گویند. مثلا در حالت دو متغیر توضیحی اگر تقسیم بندی فضا به صورت قسمت الف شکل ۲-۵ باشد آنگاه h_2 ,h_1 ,h. مجموعه گره‌ها را تشـکیل می‌دهد. اگر به ساختار درختی این تقسیم بندی (قسمت ب شکل۲-۳) توجه کنیم، در واقع گره محلی است که زیر شاخه‌های درخت بهم متصل می‌شوند.

لازم به ذکر اسـت که در روش درخت تصـمیم، درخت‌ها دوتایی[2] هسـتند به این معنی که هر زیر فضا (شاخه) فقط می‌تواند به دو مجموعه (شاخه) تقسیم شود.

اکنون با فرض اینکه فضـای متغیرهای X_1 و X_2 به چهار ناحیه R_1 ، R_2، R_3، R_4 تفکیک شده باشـد، به طور قطع هر ناحیه R_m: m=1,…,4 شـامل زیر مجموعه ای از داده‌های مشاهده شده به نام A_m اسـت، به طوری که $U_{m}^{s} = 1^{A_m} = S$. که در آن s فضای کل متغیرهای توضیحی می‌باشد. بنا به هدف دنبال شـده در روش درخت رگرسـیونی، می‌توان رویه پاسخ را در ناحیه R_m توسط یک مدل سـاده یعنی $\hat{y}_{R_m} = c_m$ تقریب نمود. بنابراین روش درخت رگرسیونی به صورت زیر بیان می‌شود.

$$\hat{y} = \sum_{m=1}^{S} c_m \, I_{A_m}(X_1, X_2) \qquad (۱-۲)$$

که در آن c_mها مقادیری ثابت می‌باشند و $I_{A_m}(X) = \begin{cases} 1, & X \in A_m \\ 0, & X \notin A_m \end{cases}$.

با توجه به روش کمترین توان‌های دوم خطا، به آسـانی ثابت می‌شـود که c_m، میانگین مقادیر متغیر پاسخ واقع در ناحیه R_m است. به عبارت دیگر

---

[1] -Node

[2] -Binary tree

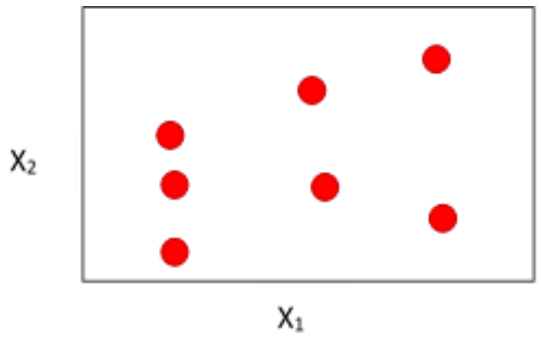
$$\hat{y}_{R_m} = c_m = \bar{y}_{A_m} = \frac{1}{n_m} \sum_{y_j \in A_m} y_j \qquad (۲-۲)$$

که در آن n_m تعداد مشاهدات مجموعه A_m است.

۲-۱۳-۲- الگوریتم تشکیل درخت رگرسیونی

تشکیل درخت رگرسیونی یا افراز فضای نمونه، طی سلسله مراتبی انجام می‌شود، که هر مرحله، براسـاس آن است که تفکیک فضا در راستای کدام متغیر و در چه مقدار از آن متغیر باید انجام گیرد. بدین منظور جزئیات الگوریتم را با یک مثال شرح می‌دهیم.

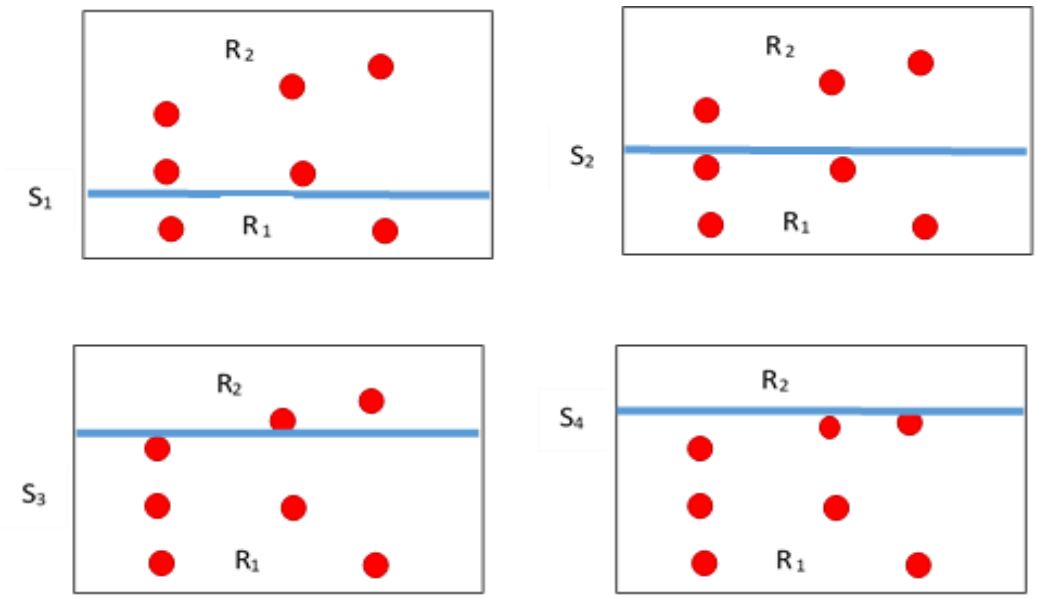
فرض کنید تعداد ۷ مشاهده برای دو متغیر X_1 و $X2$ در اختیار باشد که نمودار پراکنش[1] آن‌ها در شکل ۲-۶ آمده است.



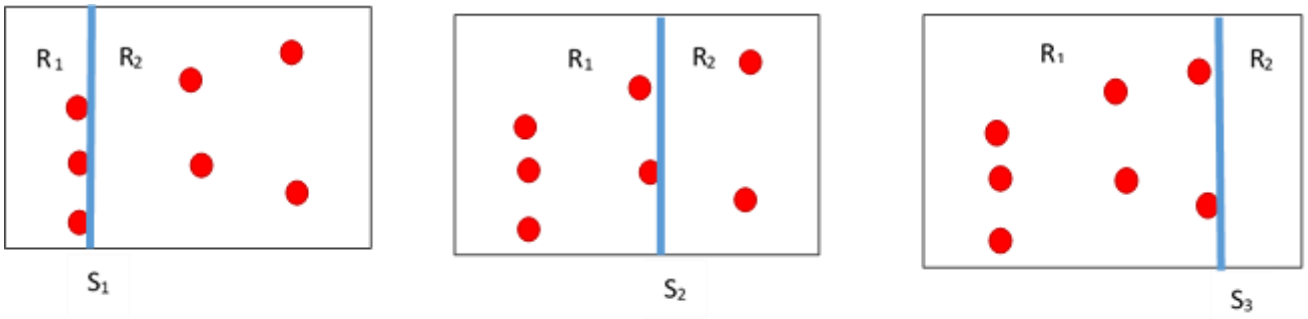
شکل ۲-۶: نمودار پراکنش داده‌های دو متغیره

شـکل ۲-۷ و شکل ۲-۸ کلیه افرازهای ممکن و موثر را به ترتیب در راستای محورهای X_1 و X_2 نشان می‌دهند.

[1]- Transmittal



شکل ۲-۷: چهار افراز ممکن در راستای متغیر x_2



شکل۲-۸: سه افراز ممکن در راستای متغیر x_1

برای انتخـاب بهترین افراز مرحلـه اول، بـه هریک از هفت حالت فوق، یک اندازه نســـبت داده می‌شود. این اندازه به صورت رابطه زیر تعریف می‌گردد.

$$M = \sum_{R_1}(Y_i - \overline{Y})^2 + \sum_{R_2}(Y_i - \overline{Y})^2 \qquad (۳-۲)$$

در این رابطه، Y_i متغیر پاسخ داده‌ها، $\overline{Y}$ میانگین مقادیر متغیر پاسخ در هر قسمت و M، برابر با مجموع توان‌های دوم خطا در نواحی R_1 و R_2 برای هریک از نمودارهای شکل۲-۸ و شکل ۲-۷ است. افرازی که دارای کمترین مقدار M باشد، به عنوان بهترین افراز در مرحله اول انتخاب می‌شود و بدین ترتیب بهترین جهت و بهترین نقطه برای تفکیک فضا در گام اول بدست خواهد آمد. همچنین می‌توان نوشت

$$M = n_1 Var_1 + n_1 Var_2 \qquad (۴-۲)$$

که در آن $n_1 Var_1$ و $n_2 Var_2$ به ترتیب واریانس مشاهدات واقع در نواحی R_1 و R_2 و همچنین n_1 و n_2 تعداد مشاهدات هریک از دو ناحیه افراز شده می‌باشد. بنابراین می‌توان M را مجموع واریانس‌های مشاهدات در دو ناحیه R_1 و R_2 دانست.

اکنون فضا به دو ناحیه تفکیک شده است. در مرحله دوم و مراحل بعدی فرآیند فوق در هر یک از نواحی تولید شده تکرار می‌شود تا در نهایت فضای متغیرهای توضیحی به صورتی شبیه به قسمت قابل قبول شکل ۲-۳ (الف)، افراز شود.

۲-۱۳-۳- اندازه درخت و هرس کردن[۱]

در این بخش به یک مطلب مهم درباره ساختار درخت رگرسیونی پرداخته می‌شود. تفکیک فضای متغیرها تا چه مرحله ای می‌تواند ادامه یابد یا به عبارت دیگر، هر درخت تا کجا می‌تواند رشد کند. اگر درخت به اندازه کافی رشد نکند ممکن است مدل مناسبی ارائه نشود و اگر هم بیش از حد

۱- Prune

رشـد کند، احتمال این که بیش‌برآوردی[1] رخ دهد، بسیار است. برای انتخاب اندازه درخت، نیاز به یک پارامتر در ساختار مدل است بطوری که این پارامتر مقدار بهینه اندازه درخت را بدست آورد.

در هر مرحله از رشـد درخت، مدل حاصـل از روش درخت رگرسـیونی دقیق‌تر شـده و مجموع توان‌های دوم خطا کاهش می یابد. راه توقف رشـد درخت براسـاس کاهش مجموع توان‌های دوم خطا اسـت، بدین ترتیب 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 مورد مجموع توان‌های دوم خطا می‌تواند در اختیار کاربر باشـــد [۳۴]. در پایان این بخش باید یادآور شـــد که روش درخت رگرسیونی با بیش‌برآوردی همراه است، که برای جلوگیری از این مشکل، درخت نیاز به هرس دارد.

۲-۱۳-۴- معرفی روش جنگل‌های تصادفی

اساس روش RF وابسته به ماهیت روش درخت رگرسیونی است و از آنجایی که ساختار درخت رگرسیونی بطور کامل معرفی گردید، اینک می‌توان به معرفی روش RF پرداخت.

در روش RF مجموعه‌ای از درخت‌های رگرسـیونی تشکیل می‌شوند و هر درخت مدلی را تولید می‌کنـد کـه مـدل نهـایی، برآینـد یا ترکیبی از همه این مدل‌ها اســـت. به عبارت دیگر هریک از این درخت‌ها سـهمی در مدل نهایی دارد. یکی از تفاوت‌های اساسی RF با درخت رگرسیونی آن است که در درخت رگرسـیونی برای افراز فضـای متغیرها در هر مرحله، از کلیه متغیرها اسـتفاده می‌گردد در حالی که در RF فقط از زیر مجموعه‌ای از متغیرها اسـتفاده می‌شود. دیگر تفاوت اساسی این دو روش داده‌هایی است که، در ساخت مدل شرکت دارند. به این معنا که در درخت رگرسیونی، همه داده‌ها در سـاخت مدل شـرکت دارند، اما در روش RF تنها بخشـی از داده‌ها در ساخت مدل استفاده می‌شوند. بطور کلی می‌توان گفت روش RF، ترکیبی از چندین درخت رگرســیونی اســت که در ســاخت آن

[1] -Over Estimation

چنـدین نمونه با جایگذاری از داده‌ها شـــرکت دارند و در هر درخت برای ســـاخت هر گره، تنها یک زیرمجموعه تصادفی از متغیرهای توضیحی شرکت می‌کنند. همچنین روش جنگل‌های تصادفی علاوه بر اینکه به عنوان یک روش مدل‌سازی به کار می‌رود به عنوان یک روش انتخاب متغیر نیز می‌تواند به کار برده شود.

۲-۱۳-۴-۱- الگوریتم روش جنگل‌های تصادفی

فرض کنید $i = 1 , \dots , N$ و (X_i , Y_i) مجموعه داده‌های مدل سـاز (آموزش) باشند که در آن، $X_i = (X_{1i,\dots,}X_{Mi})$ برداری از M متغیر توضـیحی و Y_i متغیر پاسـخ متناظر آن اسـت. اگر تعداد کل درخت‌های مدل با n_{tree} نشـــان داده شـــود، مراحل پنج‌گانه زیر بیانگر الگوریتم ســـاخت درخت iام $(i=1,\dots,n_{tree})$ است.

۱) یـک نمونـه N تـایی بـه روش با جایگذاری از مجموعه داده‌های آموزشـــی گرفته می‌شـــود. زیر مجموعه‌ای از داده‌های اصلی (حدود یک‌سوم از داده‌های آموزشی) که در این نمونه حضور ندارند را OOB[۱] نامیده که برای هر درخت نقش داده‌های آزمون را ایفا می‌کند. روش جداسازی داده‌های OOB روش نمونـه گیری خودگردان اســـت. با تکرار عملیات نمونه گیری، تعدادی مجموعه داده OOB از مجموعه آموزشـی بوجود می‌آید که می‌توان برای هر مجموعه داده آموزشی یک درخت تصمیم محاسبه کرد.

۲) به طور تصـادفی از بین M متغیر توضـیحی، m متغیر انتخاب می‌شـود $(m \ll M)$. برای تقسـیم فضای متغیرهای توضیحی به دو قسمت و براساس اصول درخت رگرسیونی، فقط از این m متغیر و نمونه N تایی انتخاب شـده در گام ۱ اسـتفاده می‌شود تا بهترین متغیر و بهترین نقطه افراز در اولین مرحله از مراحل بازگشـتی بدسـت آید. محدودیت این مرحله آن اسـت که تعداد مشاهدات

[۱] -Out of bag

موجود در هر یک از دو ناحیه باید بیشتر از n_r باشد و n_r پارامتری است که در اختیار کاربر بوده و بیانگر حداقل تعداد مشـاهدات موجود در هر ناحیه اسـت. معمولا در مدل رگرسـیونی، $m=\frac{M}{3}$ پیشنهاد می‌شود و در مدل رده‌بندی $m = \sqrt{M}$ در نظر گرفته می‌شود.

۳) برای هریک از دو ناحیه تولید شـده در گام ۲، مجددا بطور تصادفی از بین M متغیر توضیحی، m متغیر انتخاب می‌گردد و با اسـتفاده از همان نمونه N تایی انتخاب شـده در گام ۱، روش درخت رگرسـیون اعمال می‌گردد. این عمل منجر به افراز هر یک از نواحی موجود به دو قسمت می‌شود. لازم به ذکر است که افراز نواحی در صورتی انجام می‌شود که تعداد مشاهدات موجود در کل ناحیه بیشـتر از $2n_r$ باشد. یعنی اگر هریک از نواحی دارای تعداد مشاهداتی کمتر از $2n_r$ باشد نباید افراز دیگری روی هیچ یک از نواحی صورت گیرد.

۴) گام ۳ برای تمام نواحی افراز شده تا زمانی تکرار می‌شود که تعداد مشاهدات در تمامی این نواحی کمتر از $2n_r$ باشد.

۵) حال درخت رگرسـیونی iام طوری تشکیل شده است که فضای متغیرهای توضیحی به r_i ناحیه‌ی R_{1i} , R_{2i}, …, R_{r_i} تقسـیم گردیده اسـت و تعداد مشاهدات هر ناحیه کمتر از $2n_r$ می‌باشد، مدل درخت رگرسیونی به دست آمده، به صورت زیر است.

$$\hat{f}_i(X) = \sum_{j=1}^{r_i} \hat{c}_j I_{R_{ij}}(X) \qquad (۵-۲)$$

که در آن، $\hat{c}_j = \hat{y}_j$ و $I_{R_{ij}}(X) = \begin{cases} 1, X \in R_{ij} \\ 0, X \notin R_{ij} \end{cases}$.

از آن‌جایی که در روش RF تعداد n_{tree} درخت رگرسـیونی وجود دارد، می‌توان گفت که تعداد n_{tree} مدل به صـورت معادله (۲-۵) خواهیم داشـت. اگر برای مقدار مشـاهده شـده‌ی x مدل خروجی درخت iام را به صـورت $\hat{f}_i(X)$ نشان دهیم، آن‌گاه برآورد متغیر پاسخ در این نقطه، با میانگین‌گیری از مقادیر $\hat{f}_i(X)$ یعنی

$$\hat{y}(X) = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} \hat{f}_i(X) \qquad (۶-۲)$$

به دست می‌آید.

لازم به ذکر است که در روش RF، تعداد درخت‌ها (پارامتر n_{tree})، تعداد متغیرهای به تصادف انتخاب شده (پارامتر m) و تعداد حداقل مشاهدات در هر افراز (پارامتر n_r)، پارامترهایی هستند که قابل تغییر بوده و توسط کاربر تعیین می‌گردند.

۲-۱۳-۴-۲- تعیین اهمیت متغیرها در روش جنگل‌های تصادفی

در ساختار روشRF امکانی وجود دارد که می‌توان میزان اهمیت متغیرها (VI)[۱] را در مدل، تعیین نموده و متغیرهایی که دارای نقش بیشتری در هر درخت و در مدل نهایی هستند شناسایی شوند. همانطور که در الگوریتم روش RF اشاره شد، برای تشکیل هر درخت، یک نمونه با جایگذاری از داده‌های اصلی مورد استفاده قرار می‌گیرد. همچنین داده‌هایی که در این نمونه حضور ندارند را OOB نامیدیم که به نوعی نقش داده‌های آزمایشی را برای ارزیابی آن درخت ایفا می‌کند. فرض کنید خطای پیش‌بینی Y برای داده‌های OOB در درخت iام، با نماد $EOOB_i$ نشان داده شود. برای تعیین اهمیت متغیر jام، مقادیر این متغیر را بطور تصادفی n_{Perm} مرتبه (n_{Perm}، پارامتری است که در اختیار کاربر می‌باشد) جا به جا کرده و مجددا خطا، به ازای مجموعه جدید محاسبه می‌شود که مقدار این خطا با نماد $\widetilde{EOOB}_i^j$ نشان داده می‌شود. میزان اهمیت متغیر jام در مدل درخت iام با ($VI_i(x^j)$) نشان داده و به صورت رابطه (۲-۷) تعریف می‌شود:

$$VI_i(x^j) = \widetilde{EOOB}_i^j - EOOB_i \qquad (۷-۲)$$

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

[۱]-Variable Importance

$$VI(x^j) = \frac{1}{ntree} \sum_{i=1}^{ntree} (VI_i(x^j)) \qquad (۸-۲)$$

برای دقت بیشتر می‌توان جا به جا نمودن تصادفی مقادیر متغیر 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 [۳۵ و۳۲] . قابل ذکر است که اندازه اهمیت یک متغیر، به تنهایی قابل تفسیر نبوده و فقط برای رتبه بندی متغیرها براساس اهمیت آن‌ها در مدل به کار می‌رود.

۲-۱۳-۴-۳- مزیت‌های روش جنگل‌های تصادفی

۱) به دلیل شرکت کردن نمونه‌ای از داده‌ها در تشکیل مدل، این روش از لحاظ هزینه محاسبات مقرون به صرفه است. ضمن اینکه از دقت قابل قبولی نیز برخوردار می‌باشد.

۲) روش RF در مجموعه داده‌هایی که متغیرهای توضیحی زیادی وجود دارد، از کارایی بالایی برخوردار است. دلیل این ویژگی انتخاب تصادفی m متغیر از M متغیر کل در هر مرحله از رشد مدل می‌باشد.

۳) در این روش، با وجود این‌که هر درخت بطور کامل رشد می‌کند و هرس نمی‌گردد، با این حال مدل نهایی دچار بیش‌برآوردی نمی‌شود.

۴) این روش قابلیت استفاده در رده‌بندی داده‌های کیفی و پیش‌بینی داده‌های کمی را دارا می‌باشد.

۵) در صورت در اختیار داشتن تعداد مشاهدات کم، با توجه به نقش داده‌های OOB، می‌توان از آن به عنوان داده‌های آزمون استفاده کرد.

۶) این روش توانایی انتخاب متغیر را داراست و بر خلاف سایر روش‌های یادگیری ماشین، نیازمند روشی دیگر برای انتخاب متغیر نیست [۳۲] .

۲-۱۴- ارزیابی قدرت پیش‌بینی مدل

۲-۱۴-۱- استفاده از پارامترهای آماری

به منظور ارزیابی توانایی مدل‌ها در برآورد یا پیش بینی متغیر پاسخ مورد نظر، نیاز به معرفی بعضی معیار ها است تا بتوان نتایج مدل‌های مختلف را با یکدیگر مقایسه کرد. در واقع این معیارها با اندازه‌گیری خطای برآورد یا دقت برآورد یک مدل، نتایج حاصل را قابل ارزیابی و قیاس‌پذیر با سایر نتایج می‌کنند. محققین در پایگاه‌های مختلف داده از معیارهای متفاوتی استفاده می‌کنند که به تعدادی از آن‌ها اشاره می‌شود.

ضریب همبستگی[۱]: ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر، محاسبه آماره ضریب همبستگی آن‌ها است. ضریب همبستگی دو متغیر x و y با رابطه (۲-۹) تعریف می‌شود:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (۹-۲)$$

مقدار این آماره بین ۱ تا ۱- متغیر است. مقدار بزرگتر قدر مطلق آن نشانه این است که ارتباط خطی بیشتری میان متغیر وابسته و متغیرهای مستقل وجود دارد.

ضریب تعیین[۲]: ضریب تعیین یک معیار آماری است که بیانگر این است که چه درصدی از تغییرات متغیر پاسخ را می‌توان توسط مدل ساخته شده توجیه کرد. ضریب تعیین را معمولا با R^2 نشان می‌دهند و مقدار آن بین صفر و یک است. در واقع R^2 معیاریست که بیانگر دقت مدل مورد نظر در برآورد متغیر پاسخ می‌باشد. فرمول محاسباتی R^2 به صورت زیر است.

---

[۱]-Correlation Cofficient

[۲]- Determination coefficient

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (۱۰-۲)$$

کـه SSR[۱] در رابطـه (۲-۱۰) بیـانگر مجموع مربعات انحراف مقادیر پیش‌بینی شـــده‌ی متغیر وابسته از میانگین مقادیر آن است.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (۱۰-۲)$$

SST[۲] در رابطه (۲-۱۱) نشـانگر مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از میانگین مقادیر آن است.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (۱۱-۲)$$

SSE[۳] در رابطه (۲-۱۲) نیز مبین مجموع مربعات انحراف مقادیر واقعی متغیر وابســته از مقادیر پیش‌بینی شده برای آن است.

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

که در این روابط $\hat{y}_i$ مقدار پیش‌بینی شده‌ی متغیر وابسته، y_i مقدار واقعی متغیر وابسته و $\bar{y}$ در هر رابطه، میانگین مقادیر متغیر وابسته است.

---

۱ - Sum Square Regression

۲- Sum Square Total

۳ - Sum Square Error

بنابراین با توجه به روابط فوق می‌توان نوشت:

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

طبق رابطه (۲-۱۳) اگر تمام مشاهدات بر روی خط برازش شده قرار گرفته باشند، یعنی به ازای تمام نقاط $y_i = \hat{y}_i$ باشد، مقدار R^2 برابر یک می‌شود و هرگونه انحرافی از این حالت باعث می‌شود که مقدار R^2 از یک کوچکتر شود.

خطای استاندارد پیش‌بینی[۱] (SEP)

SEP به صورت رابطه (۲-۱۴) تعریف می‌شود:

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

مجموع مربع باقیمانده ها[۲] (PRESS): PRESS برابر مجموع مربعات تفاوت بین مقدار کمیت مشاهده شده(y_i) و مقدار تخمین زده شده ($\hat{y}_i$) است که مشابه رابطه (۲-۱۲) است.

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

میانگین مربع خطاها[۳] (MSE)

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

---

۱ - Standard Error of Prediction

۲ - Predictive Residual Sum of Squares

۳ - Mean Square Error

میانگین خطای مطلق[۱] (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i}{n} \quad (۱۷-۲)$$

میانگین خطای نسبی[۲] (MRE)

$$MRE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \times 100 \quad (۱۸-۲)$$

خطای نسبی پیش بینی[۳] (REP)

$$REP(\%) = \frac{100}{\overline{y}} \times \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (۱۹-۲)$$

۲-۱۴-۲- استفاده از نمودار برگشتی

در نمودار برگشتی مقادیر پیش بینی شده‌ی کمیت مورد نظر برحسب مقادیر تجربی رسم می‌شود و به کمک ضریب تعیین (R^2) به دست آمده از نمودار، پراکندگی نقاط اطراف خط برگشت تعیین می‌شود. محدوده‌ی تغییرات ضریب تعیین به دست آمده بین ۰ و ۱ است که هرچه قدر به ۱ نزدیک‌تر باشد، نتایج پیش‌بینی شده به مقادیر واقعی نزدیک ترند. اما اگر این مقدار برابر صفر باشد، بین داده‌ها هیچ گونه همبستگی وجود ندارد.

---

[۱] - Mean Absolute Error

[۲] - Mean Relative Error

[۳] - Relative Error of Prediction

۲-۱۴-۳- استفاده از نمودار خطای باقی‌مانده

منظور از عبارت خطای باقیمانده، اختلاف بین مقادیر پیش‌بینی شده و مقادیر تجربی است. اگر پراکندگی مقادیر در دو طرف خط صفر یکسان باشد، این امر نشان دهنده‌ی تصادفی بودن خطاست. ولی اگر عمده‌ی نقاط در این نمودار، در یک طرف خط صفر باشند، این بدان معناست که خطای جهت‌داری رخ داده است.

۲-۱۴-۴- استفاده از سری تست

اعتبار و اهمیت مدل‌های پیش‌بینی کننده وقتی مشخص می‌شود که بتوانند خاصیت مورد نظر را برای مولکول‌های جدیدی که در دسته‌ی آموزش موجود نیستند پیش‌بینی کنند. بدین منظور از ابتدای کار تعدادی از مولکول‌ها به عنوان سری تست خارجی کنار گذاشته می‌شود و در نهایت مدل منتخب بهینه شده، جهت پیش‌بینی خاصیت مورد نظر مولکول‌های موجود در سری تست که در مدل‌سازی استفاده نشده‌اند، به کار گرفته می‌شود.

۲-۱۴-۵- استفاده از روش رد مرحله‌ای تک‌تک

در این روش هم برای تعیین اعتبار مدل‌های پیش‌بینی کننده هر بار یک ترکیب به عنوان سری آزمون کنار گذاشته می‌شود و در شرایط بهینه به دست آمده برای سری ارزیابی یا سری آموزش با ترکیبات باقی‌مانده مدل‌سازی صورت می‌گیرد، سپس مدل بدست آمده برای پیش‌بینی فعالیت ترکیب کنار گذاشته شده به کار گرفته می‌شود و این فرآیند برای تمام اعضای سری داده‌ها تکرار می‌گردد.

۳- فصل سوم

مطالعه ارتباط کمی ساختار-فعالیت

مشتقات پیریدینون

۳-۱- مدل‌سازی فعالیت دارویی مشتقات پریدینون

در این فصل، روش جنگل تصادفی و شبکه عصبی مصنوعی به منظور مطالعه ارتباط کمی ساختار-فعالیت برای پیش‌بینی فعالیت ضدHIV مشتقات ۴-(۱و۵-دی بنزیل-۱و۲-دی هیدرو-۲-اکسو پیریدین-۳-ایل)-۲-هیدروکسی-۴-اکسوبوت-۲-انوئیک اسید[۱] به کار گرفته شده است. و توانایی مدل حاصله در پیش‌بینی فعالیت‌های ضد ایدز این ترکیبات مورد ارزیابی قرار گرفته است. بطور کلی این تحقیق شامل معرفی سری داده‌ها، بهینه سازی ساختار مولکول‌ها و محاسبه توصیف کننده‌های مولکولی با استفاده از روش سهم گروه و رگرسیون مرحله‌ای و همچنین مدل‌سازی توسط روش‌های جنگل‌های تصادفی و شبکه عصبی مصنوعی و ارزیابی مدل‌های برتر است.

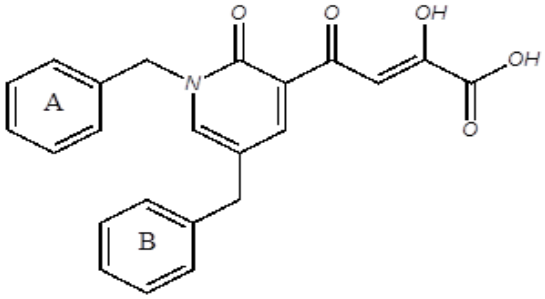
۳-۲- سری داده‌ها

سری داده‌ها از نتایج تجربی ۵۳ ترکیب فعال از مشتقات پیریدینون می‌باشد. اسکلت اصلی این ترکیبات در شکل ۳-۱ نشان داده شده است. جدول ۳-۱ استخلاف‌های مختلف تمام این ترکیبات را نشان می‌دهد. همچنین مقادیر pEC_{50} (منهای لگاریتم EC_{50}[۲] بر حسب مولار) این ترکیبات به عنوان معیاری کمی از فعالیت ضد HIV این ترکیبات در جدول ۳-۱ گزارش شده است که از مرجع [۳۶] استخراج گردیده است. EC_{50} غلظتی از دارو می‌باشد که تا ٪۵۰ تکثیر ویروس HIV-1 را مهار می‌کند.

---

[۱] - 4-(1,5-dibenzyl-1,2-dihydro-2-oxopyridin-3-yl)-2-hydroxy-4-oxobut-2-enoic acid

[۲] -Effective Concentration



شکل ۳-۱: اسکلت اصلی ترکیب مورد مطالعه

جدول ۳-۱: استخلاف‌های ترکیب مورد مطالعه

شماره ترکیب	حلقه آریلی A			حلقه آریلی B			pEC_{50}
	ارتو	متا	پارا	ارتو	متا	پارا	
۱	H	H	H	H	H	H	۵/۶۷
۲	F	H	H	H	H	H	۶/۰۴
۳	H	H	F	H	H	H	۶/۰۰
۴	F	H	F	H	H	H	۶/۲۲
۵	F	H	H	F	H	H	۶/۱۵
۶	H	H	OMe	H	H	H	۵/۷۴
۷	H	H	F	H	H	F	۶/۰۹
۸	H	H	Me	H	H	F	۵/۷۹
۹	H	Cl	F	H	H	H	۶/۰۹
۱۰	F	H	F	F	H	F	۶/۵۲
۱۱	H	F	H	H	H	H	۵/۶۹
۱۲	H	Cl	H	H	Cl	H	۵/۹۵
۱۳	H	Cl	F	H	Cl	H	۵/۹۲
۱۴	F	H	H	H	H	F	۶/۱۵
۱۵	H	Cl	F	H	Cl	F	۵/۶۱
۱۶	H	H	Me	H	Cl	H	۵/۶۰
۱۷	H	H	Me	F	H	H	۶/۱۵
۱۸	F	Cl	F	H	H	H	۶/۲۲
۱۹	H	H	F	H	Cl	H	۵/۸۵
۲۰	2,6-di-F	H	H	H	H	F	۵/۸۵
۲۱	H	H	F	F	H	H	۶/۰۹

ادامه جدول (۳-۱)

شماره ترکیب	حلقه آریلی A			حلقه آریلی B			pEC_{50}
	ارتو	متا	پارا	ارتو	متا	پارا	
۲۲	F	Cl	H	F	Cl	H	۵/۶۱
۲۳	H	H	Cl	H	Cl	H	۵/۶۰
۲۴	H	Cl	H	H	Cl	H	۵/۸۵
۲۵	F	Cl	H	H	Me	H	۵/۸۵
۲۶	F	H	F	H	F	H	۶/۲۲
۲۷	H	Me	H	H	Cl	F	۵/۶۵
۲۸	F	Cl	H	H	H	H	۶/۱۵
۲۹	F	5-Cl	H	H	H	H	۶/۰۴
۳۰	F	Cl	H	H	Cl	H	۵/۸۵
۳۱	F	H	F	H	Cl	F	۵/۸۵
۳۲	H	Cl	H	H	H	F	۵/۹۲
۳۳	H	Cl	H	F	H	H	۶/۱۵
۳۴	2,6-di-F	H	H	H	Cl	F	۵/۲۹
۳۵	2,6-di-F	H	H	H	H	H	۵/۸۲
۳۶	H	H	H	F	H	F	۶/۰۹
۳۷	F	H	H	H	Cl	F	۵/۶۵
۳۸	H	F	H	F	H	H	۵/۷۲
۳۹	F	H	H	H	Cl	H	۵/۷۲
۴۰	F	F	H	H	H	H	۶/۰۴
۴۱	H	Me	H	F	H	H	۶/۰۴
۴۲	F	F	H	H	H	F	۶/۱۵
۴۳	H	H	CN	F	H	H	۵/۹۵
۴۴	F	F	H	F	H	H	۶/۰۹
۴۵	H	F	H	H	H	F	۶/۰۰
۴۶	F	H	F	Cl	H	H	۶/۰۹
۴۷	H	Cl	H	F	Cl	H	۵/۸۲
۴۸	H	H	F	F	H	F	۶/۲۲
۴۹	H	Me	H	F	H	F	۶/۱۵
۵۰	H	H	H	Me	H	F	۵/۸۸
۵۱	H	H	F	Me	H	F	۶/۰۰
۵۲	F	H	H	F	H	F	۶/۳۰
۵۳	F	H	H	Me	H	F	۵/۸۸

۳-۳- نرم‌افرازهای مورد استفاده

دانش کمومتریکس بسته‌های نرم‌افزاری متنوعی را برای انجام تمام مراحل مدل‌سازی به کار گرفته است که در ادامه به اختصار بسته‌های نرم‌افزاری استفاده شده در این تحقیق معرفی می‌شوند.

۳-۳-۱- بسته نرم‌افزاری Hyperchem

از بسته نرم‌افزاری Hyperchem [۳۷] برای رسم شکل مولکول‌ها و بهینه‌سازی ساختار با استفاده از روش‌های کوانتومی و مکانیکی، استفاده می‌شود. به کمک این برنامه می‌توان طول پیوند، زاویه پیوندی و زوایای پیچشی را در مولکول تعیین کرد. داده‌های حاصل از این نرم‌افراز را می‌توان به عنوان ورودی به سایر نرم افرازها معرفی نمود. به عنوان مثال ساختار سه بعدی مولکول‌ها که با استفاده از روش مکانیک کوانتومی در Hyperchem بهینه شده است، را می‌توان جهت محاسبه توصیف کننده‌های مختلف به وسیله نرم‌افراز دراگون[۱] ، به کار برد. همچنین به کمک این نرم‌افراز می‌توان تعدادی از توصیف کننده‌ها از جمله حجم مولی و قطبش‌پذیری را محاسبه کرد.

۳-۳-۲- بسته نرم‌افزاری دراگون

نرم‌افراز Dragon5.5 امکان محاسبه ۳۲۲۴ توصیف کننده مختلف را فراهم می‌کند. جهت محاسبه توصیف کننده به کمک این نرم‌افراز لازم است ساختار هندسی بهینه مولکول مورد استفاده قرار گیرد. برای این منظور می‌توان ساختار بهینه مولکول‌ها را به صورت فایل‌هایی با فرمت mol، solf و hin به عنوان اطلاعات ورودی به کار برد.

---

[۱]- Dragon

۳-۳-۳- بسته نرم‌افزاری SPSS[۱]

SPSS [۳۸] که نخستین نسخه آن در سال ۱۹۷۰ توسط جمعی از فارغ التحصیلان دانشگاه استانفورد آمریکا ارائه شده، امکان تجزیه و تحلیل آماری داده‌ها را فراهم می‌آورد. برخی از قابلیت‌های این بسته نرم‌افزاری عبارتند از:

۱) تعیین تعداد فراوانی‌های هر یک از گروهها در یک متغیر

۲) محاسبه میانگین ساده برای داده‌ها

۳) نمایش اطلاعات به صورت متنوع در قالب نمودار و جدول

۴) انجام رگرسیون تک متغیره و چند متغیره

۳-۳-۴- نرم‌افراز MATLAB

نرم‌افراز MATLAB [۳۹] یکی از جامع‌ترین و کارآمد ترین نرم‌افزارهای علمی و محاسباتی است که طی چند سال گذشته تهیه و به بازار عرضه شد و در سال های اخیر با تدوین نسخه‌های جدیدتر و کامل‌تر روز به روز بر غنای آن افزوده می‌شود.

MATLAB[۲]به معنای آزمایشگاه ماتریس است. ورودی‌ها اساساً به صورت ماتریس در نظر گرفته می‌شوند و هیچ نیازی به مشخص کردن ابعاد ماتریس نمی‌باشد. در MATLAB حتی اعداد اسکالر، ماتریس‌های (x) به حساب می‌آیند و بردارها، حالت خاصی از ماتریس‌های سطری یا ستونی در نظر گرفته می‌شوند. از جمله کاربردهای جالب آن که مورد توجه شیمیدانان قرارگرفته است شبکه های عصبی مصنوعی، جنگل‌های تصادفی و ماشین بردار پشتیبان است که در این پژوهش از آن‌ها استفاده شده است. اطلاعات شیمیایی به عنوان ورودی و به صورت یک ماتریس به نرم افراز داده

---

[۱] - Statistical Package for the Social Science

[۲] - Matrix Laboratory

می‌شـود. در محیط این برنامه با اسـتفاده از آرایه‌ها و فرامان‌های موجود امکان مدل‌سـازی غیرخطی فعالیت بیولوژیکی با ساختار ترکیبات فراهم می‌گردد.

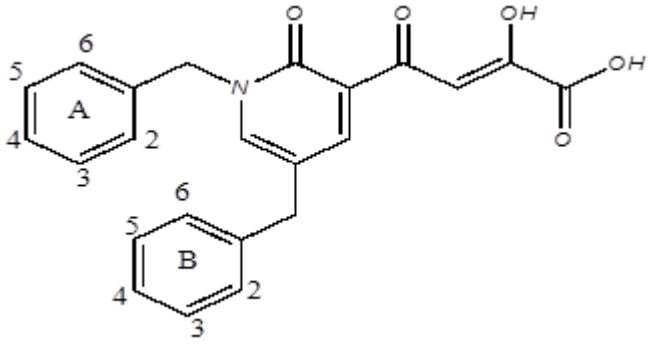
۳-۴- بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌کننده‌ها

زمانی که متغیرها را به روش‌هایی مانند رگرسـیون مرحله‌ای انتخاب می‌کنیم نیاز اسـت که ساختار مولکول‌ها را بهینه کنیم. ساختار مولکول‌ها با استفاده از نرم افراز Hyperchem رسم و با روش نیمـه تجربی AM1 برای رسـیدن جذر میانگین مربعات گرادیان انرژی به ۰/۰۰۱ کیلوکالری بر مول بهینه شـدند. سپس بوسیله نرم افراز دراگون، تعداد ۳۲۲۴ توصیف‌کننده با استفاده از این ساختارهای بهینه شده محاسبه شد.

۳-۵- انتخاب بهترین توصیف کننده‌ها

۳-۵-۱- معرفی توصیف کننده‌ها به روش سهم گروه (GC)

در این بخش برای انتخاب توصـیف کننده‌ها از روش سـهم گروه اسـتفاده شده است. برای این منظور، مولکول‌ها به اجزای مختلف شکسته شدند که در نهایت برای ۵۳ مولکول، گروه‌های ساختاری به صـورتی که در جدول ۳-۳ نشـان داده خواهد شـد، بدسـت آمد. توصیف کننده‌های موجود در این روش بیانگر نوع اسـتخلاف و جایگاه موقعیت‌های اسـتخلاف شـده در حلقه‌های آروماتیک A و B می‌باشد. برای نشان دادن نوع موقعیت‌هایی که دارای استخلاف هستند، این جایگاه‌ها همانطور که در شـکل ۳-۲ نشان داده شده در حلقه‌های آروماتیک A و B شماره گذاری گردید. برای مثال استخلاف در جایگاه شماره دو در حلقه آریلی A با نماد A_2 نمایش داده شد به همین شیوه A_3, A_4, A_5, A_6 و B_2, B_3, B_4, B_5, B_6 به‌دست آمدند.



شکل ۳-۲: نمایش نوع موقعیت‌ها (جایگاه‌های) استخلاف شده

همچنین برای تشخیص اینکه توصیف کننده دارای چه نوعی استخلافی است مطابق جدول ۳-۲ از اعداد استفاده شد. لازم به ذکر است که برای توصیف کننده‌های دارای استخلاف گیرنده الکترون علامت مثبت و برای توصیف کننده‌هایی با استخلاف دهنده الکترون علامت منفی قرار داد شد. علاوه بر این، مقدار اعداد نسبت داده شده به توصیف کننده‌ها براساس قدرت الکترون کشندگی است و به‌طور قراردادی نسبت داده شده است، بطور مثال برای فلوئور که قویترین کشنده الکترون است عدد ۱ و برای کلر که قدرت الکترون کشندگی آن کمتر از فلوئور است عدد ۰/۵ در نظر گرفته شد.

جدول ۳-۲: مقادیر نسبت داده شده به توصیف کننده‌ها

نوع توصیف کننده	مقدار توصیف کننده
F	۱
CN	۰/۷
Cl	۰/۵
H	۰
Me	-۰/۵
OMe	-۱

جدول ۳-۳: توصیف‌کننده‌های سهم گروه

Number	A_2	A_3	A_4	A_5	A_6	B_2	B_3	B_4	B_5	B_6
۱	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰
۲	۱	۰	۰	۰	۰	۰	۰	۰	۰	۰
۳	۰	۰	۱	۰	۰	۰	۰	۰	۰	۰
۴	۱	۰	۱	۰	۰	۰	۰	۰	۰	۰
۵	۱	۰	۰	۰	۰	۰	۰	۰	۰	۱
۶	۰	۰	−۱	۰	۰	۰	۰	۰	۰	۰
۷	۰	۰	۱	۰	۰	۰	۰	۱	۰	۰
۸	۰	۰	−۰/۵	۰	۰	۰	۰	۱	۰	۰
۹	۰	۰/۵	۱	۰	۰	۰	۰	۰	۰	۰
۱۰	۱	۰	۱	۰	۰	۰	۰	۱	۰	۱
۱۱	۰	۱	۰	۰	۰	۰	۰	۰	۰	۰
۱۲	۰	۰/۵	۰	۰	۰	۰	۰	۰	۰/۵	۰
۱۳	۰	۰/۵	۰	۰	۰	۰	۰	۱	۰/۵	۰
۱۴	۱	۰	۰	۰	۰	۰	۰	۱	۰	۰
۱۵	۰	۰/۵	۱	۰	۰	۰	۰	۱	۰/۵	۰
۱۶	۰	۰	−۰/۵	۰	۰	۰	۰	۰	۰/۵	۰
۱۷	۰	۰	−۰/۵	۰	۰	۰	۰	۰	۰	۱
۱۸	۱	۰/۵	۱	۰	۰	۰	۰	۰	۰	۰
۱۹	۰	۰	۱	۰	۰	۰	۰	۰	۰/۵	۰
۲۰	۱	۰	۰	۰	۱	۰	۰	۱	۰	۰
۲۱	۰	۰	۱	۰	۰	۰	۰	۰	۰	۱
۲۲	۱	۰/۵	۰	۰	۰	۰	۰	۰	۰/۵	۱
۲۳	۰	۰	۰/۵	۰	۰	۰	۰	۰	۰/۵	۰
۲۴	۰	۰/۵	۰	۰	۰	۰	۰	۰	۰/۵	۰
۲۵	۱	۰/۵	۰	۰	۰	۰	۰	۰	−۰/۵	۰
۲۶	۱	۰	۱	۰	۰	۰	۰	۰	۰	۱
۲۷	۰	−۰/۵	۰	۰	۰	۰	۰	۱	۰/۵	۰
۲۸	۱	۰/۵	۰	۰	۰	۰	۰	۰	۰	۰
۲۹	۱	۰	۰	۰/۵	۰	۰	۰	۰	۰	۰
۳۰	۱	۰/۵	۰	۰	۰	۰	۰	۰	۰/۵	۰
۳۱	۱	۰	۱	۰	۰	۰	۰	۱	۰/۵	۰

ادامه جدول(۳-۳)

Number	A_2	A_3	A_4	A_5	A_6	B_2	B_3	B_4	B_5	B_6
۳۲	۰	۰/۵	۰	۰	۰	۰	۰	۱	۰	۰
۳۳	۰	۰/۵	۰	۰	۰	۰	۰	۰	۰	۱
۳۴	۱	۰	۰	۰	۱	۰	۰	۱	۰/۵	۰
۳۵	۱	۰	۰	۰	۱	۰	۰	۰	۰	۰
۳۶	۰	۰	۰	۰	۰	۰	۰	۱	۰	۱
۳۷	۱	۰	۰	۰	۰	۰	۰	۱	۰/۵	۰
۳۸	۰	۱	۰	۰	۰	۰	۰	۰	۰	۱
۳۹	۱	۰	۰	۰	۰	۰	۰	۰	۰/۵	۰
۴۰	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰
۴۱	۰	-۰/۵	۰	۰	۰	۰	۰	۰	۰	۱
۴۲	۱	۱	۰	۰	۰	۰	۰	۱	۰	۰
۴۳	۰	۰	۰/۷	۰	۰	۰	۰	۰	۰	۱
۴۴	۱	۱	۰	۰	۰	۰	۰	۰	۰	۱
۴۵	۰	۱	۰	۰	۰	۰	۰	۱	۰	۰
۴۶	۱	۰	۱	۰	۰	۰	۰	۰	۰	۰/۵
۴۷	۰	۰/۵	۰	۰	۰	۰	۰	۰	۰/۵	۱
۴۸	۰	۰	۱	۰	۰	۰	۰	۱	۰	۱
۴۹	۰	-۰/۵	۰	۰	۰	۰	۰	۱	۰	۱
۵۰	۰	۰	۰	۰	۰	۰	۰	۱	۰	-۰/۵
۵۱	۰	۰	۱	۰	۰	۰	۰	۱	۰	-۰/۵
۵۲	۱	۰	۰	۰	۰	۰	۰	۱	۰	۱
۵۳	۱	۰	۰	۰	۰	۰	۰	۱	۰	۰/۵

همانطور که در جدول ۳-۳ مشاهده می‌شود همه توصیف کننده‌های ستون‌های B_2 و B_3 صفر هستند، لذا پس از حذف داده‌های ستون‌های B_2 و B_3 ، داده‌های ۸ ستون باقی‌مانده به عنوان ورودی برای روش جنگل تصادفی استفاده شد.

۳-۵-۲- توصیف کننده‌های حاصل از دراگون و انتخاب بهترین توصیف‌کننده‌ها با روش رگرسیون مرحله‌ای(SR)

با توجه به این‌که تعداد زیاد توصیف‌کننده باعث پیچیدگی محاسبات شده و تعدادی از توصیف‌کننده‌ها حاوی اطلاعات یکسانی هستند، تعداد توصیف‌کننده‌ها باید به نوعی کاهش یابد. به همین منظور در نرم افراز دراگون هنگام ذخیره سازی توصیف کننده‌ها، آن‌ها با همبستگی ۰/۹۶ ذخیره گردید که با این‌کار توصیف‌کننده‌هایی که برای تمام مولکول‌ها مقادیر ثابت یا تقریبا ثابت داشتند، حذف شدند و همچنین یکی از دو توصیف‌کننده‌ای که همبستگی بزرگتر از ۰/۹۶ داشت حذف گردید. در نهایت از ۳۲۲۴ توصیف‌کننده، ۶۰۰ توصیف کننده باقی ماند. با به کارگیری نرم افراز SPSS و با اجرای رگرسیون مرحله‌ای برای ۶۰۰ توصیف کننده‌های یاد شده به عنوان متغیرهای مستقل، و فعالیت به عنوان متغیر وابسته، تعداد ۹ توصیف کننده انتخاب شدند که نام و طبقه آن‌ها در جدول ۳-۴ نشان داده شده است. همچنین جدول ۳-۵ ماتریس همبستگی بین این توصیف کننده‌ها را ارائه می‌دهد، نتایج این جدول نشان می‌دهد که بین این توصیف کننده ها همبستگی معناداری وجود ندارد.

جدول ۳-۴: توصیف کننده‌های انتخاب شده توسط رگرسیون مرحله‌ای

No	Symbol	Class	Meaning
۱	B07NCl	2D binary fingerprints	Presence/absence of N-Cl at topological distance 07
۲	Mor25m	3D-MoRSE descriptors	3D-MoRSE- signal 25/Weighted by atomic masses
۳	MATS8m	2D autocorrelation	Moaran autocorrelation –lag8//Weighted by atomic masses
۴	Mor29v	3D-MoRSE descriptors	3D-MoRSE- signal 29/Weighted by atomic van der waals volumes
۵	RDF110u	RDF descriptors	Radial Distribution Function - 11.0/unweighted
۶	TF..Cl	Topological descriptors	Sum of topological distances between F..Cl
۷	G1p	WHIM descriptors	1st component symmetry directional WHIM index/ Weighted by atomic polarizabilities
۸	CIC5	Information indices	Complementary information content(neighborhood symmetry of 5-order)
۹	RDF095u	RDF descriptors	Radial Distribution Function -9.5/unweighted

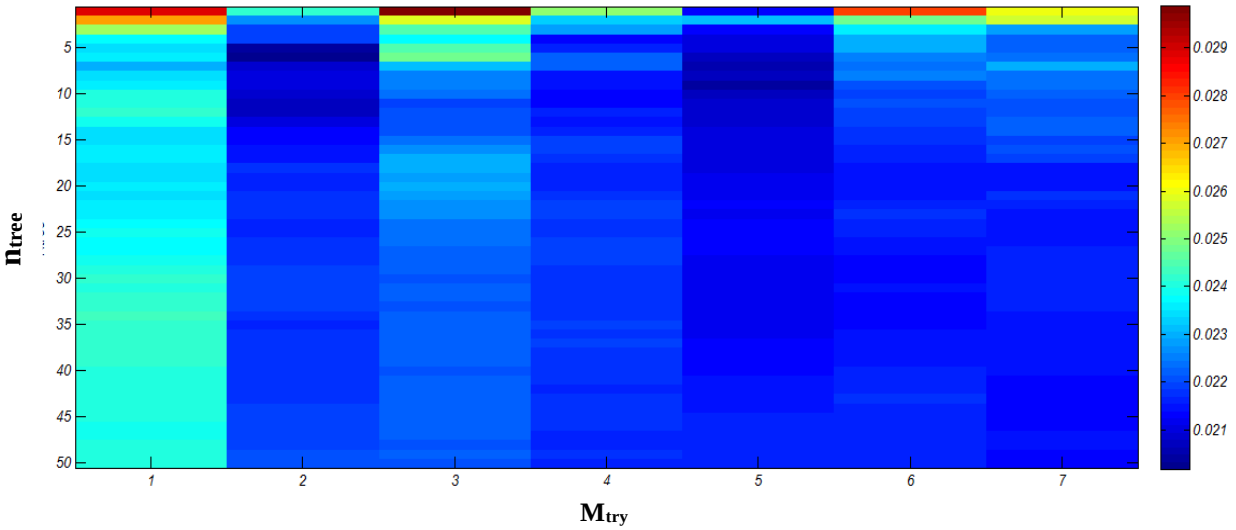
جدول ۳-۵: ماتریس همبستگی کل توصیف کننده‌های انتخاب شده توسط رگرسیون مرحله‌ای

	B07NCl	Mor25m	MATS8m	Mor29v	RDF110u	TF..Cl	G1p	ClC5	RDF095u
B07NCl	۱								
Mor25m	۰/۵۳۴	۱							
MATS8m	-۰/۳۹۹	-۰/۱۰۷	۱						
Mor29v	-۰/۱۸۳	۰/۰۶۷	۰/۱۹۹	۱					
RDF110u	۰/۰۴۵	۰/۰۵۲	۰/۰۴۱	-۰/۰۰۱	۱				
TF..Cl	۰/۶۰۱	۰/۵۳۳	-۰/۲۶۸	-۰/۰۳۷	-۰/۱۰۵	۱			
G1p	-۰/۰۵۸	-./۰۲۵	-۰/۰۳۱	۰/۱۶۳	-۰/۲۷۷	-۰/۰۹۹	۱		
ClC5	-۰/۳۸۵	-۰/۵۲۱	-۰/۱۷۴	-۰/۱۱۷	۰/۱۶۶	-۰/۴۱۲	-۰/۱۱۹	۱	
RDF095u	۰/۰۱۳	۰/۰۷۲	۰/۰۳۸	-۰/۰۷۰	۰/۲۳۰	-۰/۰۷۱	-۰/۰۹۳	-۰/۰۰۹	۱

۳-۶- مدل‌سازی جنگل‌های تصادفی

۳-۶-۱- مدل‌سازی جنگل‌های تصادفی با توصیف کننده‌های سهم گروه (GC-RF)

ابتدا کل داده‌ها به دو سری آموزش (۴۳ ترکیب) و سری آزمون (۱۰ ترکیب) تقسیم شدند. برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده گردید. همانطور که در بخش (۲-۱۲-۴-۱) گفته شد، $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند، این داده‌ها OOB هستند که برای بهینه سازی پارامترهای موثر در نظر گرفته شدند. برای دستیابی به بهترین مدل ابتدا سه پارامتر مهم و تاثیر گذار یعنی تعداد درختان (n_{tree})، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز (M_{try}) و تعداد مشاهدات باقی مانده در هر گره (اندازه گره) بهینه گردیدند. برای بهینه نمودن این پارامترها، تعداد درخت‌ها از ۱۰۰ تا ۵۰۰ با گام صد، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز (M_{try}) از ۲ تا ۸ با گام یک و تعداد مشاهدات باقی‌مانده در هر گره (Node Size) از ۱ تا ۱۰ با گام یک، بطور همزمان تغییر داده شد و در هر مرحله مقدار MSE مربوط به مجموعه OOB محاسبه گردید. نتایج بهینه سازی پارامترهای M_{try} و n_{tree} در شکل ۳-۳ نشان داده شده است همچنین جدول ۳-۶ چندین حالت که کمترین MSE را برای OOB دارد نشان می‌دهد.



شکل ۳-۳: بهینه سازی پارامترهای M_{try} و n_{tree} در Node size بهینه

همانطور که در شکل ۳-۳ مشاهده می‌شود نقاط آبی رنگ خطای کمتری را نشان می‌دهد.

جدول ۳-۶ : کمترین مقادیر MSE همراه با M_{try} و n_{tree} متناظر با آن‌ها

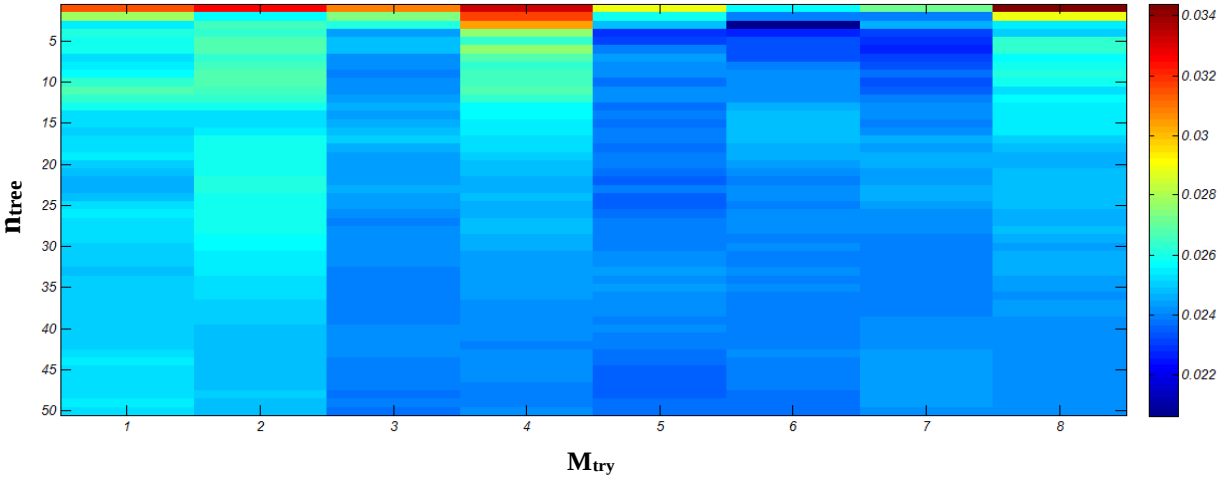
n_{tree}	M_{try}	Node Size	MSE OOB
۲۰۰	**۸**	**۱**	**۰/۰۰۲۴۱**
۱۰۰	۲	۲	۰/۰۰۳۵۵
۱۰۰	۸	۳	۰/۰۰۴۹۸
۱۰۰	۸	۴	۰/۰۰۶۷۰
۱۰۰	۸	۵	۰/۰۰۸۲۰
۱۰۰	۸	۶	۰/۰۰۹۶۵
۲۰۰	۸	۷	۰/۰۱۰۴۶
۱۰۰	۸	۸	۰/۰۱۱۸۱
۱۰۰	۸	۹	۰/۰۱۲۵۵
۱۰۰	۸	۱۰	۰./۰۱۳۴۷

با توجه به نتایج جدول، تعداد درختان =۲۰۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز =۸ و تعداد گره =۱ به عنوان مقادیر بهینه انتخاب شدند.

۳-۶-۲- مدل‌سازی جنگل‌های تصادفی با توصیف‌کننده‌های انتخاب شده به وسیله روش رگرسیون مرحله‌ای(SR)

ابتدا کل داده‌ها به دو سری آموزش (۴۳ ترکیب) و سری آزمون (۱۰ ترکیب) تقسیم شدند. برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده گردید، $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند، این داده‌ها OOB هستند که برای بهینه سازی پارامترهای موثر در نظر گرفته شدند. برای دستیابی به بهترین مدل ابتدا سه پارامتر مهم و تاثیر گذار یعنی تعداد درختان (n_{tree})، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز (M_{try}) و تعداد مشاهدات باقی مانده در هر گره (Node Size) برای داده‌های OOB بهینه گردیدند. برای بهینه نمودن این پارامترها، تعداد درخت

از ۱۰۰ تا ۵۰۰ با گام صـد، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز (M_{try}) از ۲ تا ۹ با گام یک و تعداد مشاهدات باقی‌مانده در هر گره (Node Size) از ۱ تا ۱۰ با گام یک، بطور همزمان تغییر داده شـــد و در هر مرحله مقدار MSE مربوط به مجموعه OOB محاســـبه گردید. نتایج بهینه سازی پارامترهای M_{try} و n_{tree} در شکل ۳-۴ نشان داده شده است که برای نقاط آبی رنگ مقدار mse کمتر است و جدول ۳-۷ چندین حالت که کمترین MSE را برای OOB دارد نشان می‌دهد.



شکل ۳-۴: بهینه سازی پارمترهای M_{try} و n_{tree} در Node size بهینه

جدول ۳-۷: کمترین مقادیر MSE همراه با M_{try} و n_{tree} متناظر با آن‌ها

n_{tree}	M_{try}	Node Size	MSE OOB
۳۰۰	**۴**	**۱**	**۰/۰۰۲۹۹**
۱۰۰	۵	۲	۰/۰۰۳۱۵
۱۰۰	۷	۳	۰/۰۰۳۵۴
۲۰۰	۹	۴	۰/۰۰۴۰۰
۳۰۰	۹	۵	۰/۰۰۵۰۹
۳۰۰	۹	۶	۰/۰۰۶۰۱
۵۰۰	۸	۷	۰/۰۰۷۰۶
۲۰۰	۹	۸	۰/۰۰۷۸۴
۴۰۰	۹	۹	۰/۰۰۹۱۲
۳۰۰	۸	۱۰	۰/۰۱۰۰۷

با توجه به نتایج به دست آمده، تعداد درختان =۳۰۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز =۴ و تعداد گره =۱ به عنوان مقادیر بهینه انتخاب شدند.

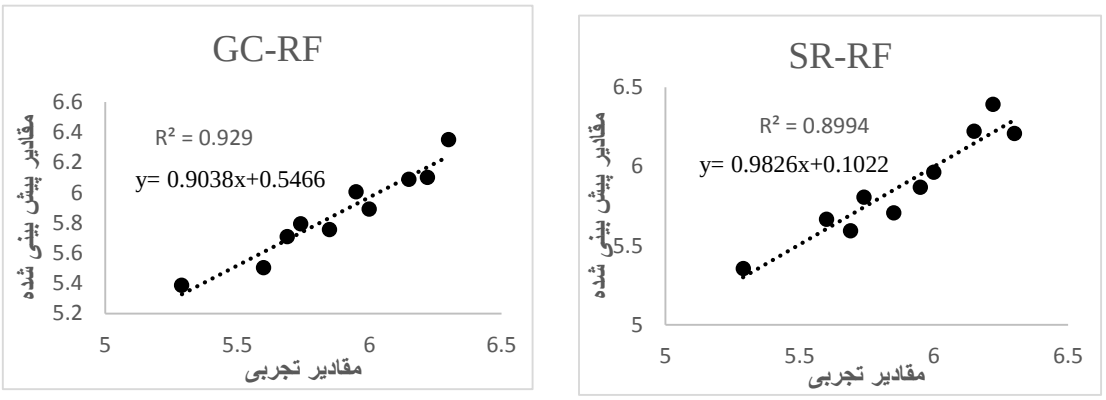
۳-۶-۳- ارزیابی مدل جنگل‌های تصادفی

۳-۶-۳-۱- ارزیابی مدل‌های RF با استفاده از داده‌های سری آزمون

اهمیت مدل‌های پیش بینی وقتی مشخص می‌گردد که مدل حاصله بتواند فعالیت مولکول‌هایی را که در مدل‌سازی به کار نرفته‌اند پیش بینی کند. بدین منظور مدل‌های منتخب، برای پیش‌بینی فعالیت داده‌های سری آزمون به کار برده شدند. جدول ۳-۸ نتایج پیش‌بینی حاصل از ارزیابی مدل‌های جنگل تصادفی (GC-RF و SR-RF) با استفاده از داده‌های سری آزمون و شکل ۳-۵ نیز نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد.

جدول ۳-۸: نتایج حاصل از ارزیابی مدل‌های جنگل تصادفی با استفاده از داده‌های سری آزمون

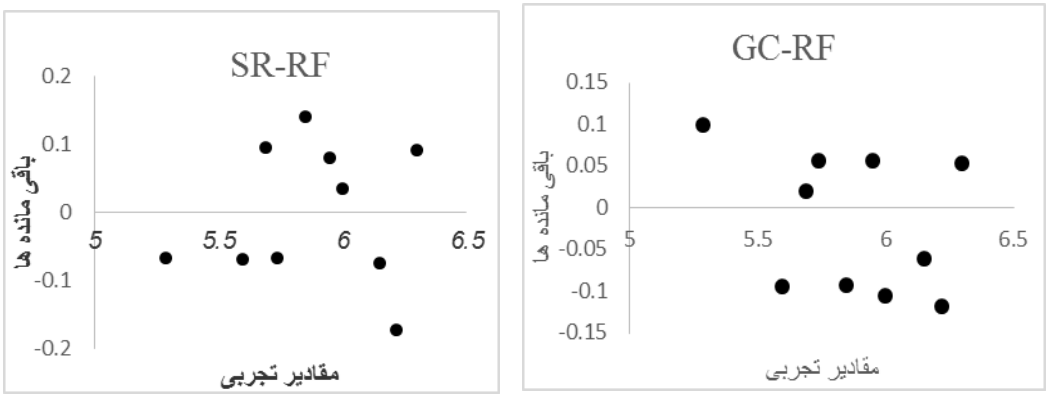
شماره ترکیب	مقدار تجربی (pEC50)	مقدار پیش‌بینی شده		درصد خطا	
		GC-RF	SR - RF	GC- RF	SR - RF
۶	۵/۷۴	۵/۸۰	۵/۸۱	۱/۰۴	۱/۲۱
۱۱	۵/۶۹	۵/۷۱	۵/ ۵۹	۰/۳۵	-۱/۷۵
۲۳	۵/۶۰	۵/۵۰	۵/۶۷	-۱/۷۸	۱/۲۵
۲۸	۶/۱۵	۶/۰۹	۶/۲۲	-۰/۹۷	۱/۱۳
۳۰	۵/۸۵	۵/۷۶	۵/۷۰	-۱/۵۳	-۲/۳۹
۳۴	۵/۲۹	۵/۳۹	۵/۳۶	۱/۸۹	۱/۳۲
۴۳	۵/۹۵	۶/۰۱	۵/۸۷	۱/۰۰	-۱/۳۴
۴۵	۶/۰۰	۵/۸۹	۵/۹۶	-۱/۸۳	-۰/۶۶
۴۸	۶/۲۲	۶/۱۰	۶/۳۹	-۱/۹۲	۲/۷۳
۵۲	۶/۳۰	۶/۳۵	۶/۲۱	۰/۷۹	-۱/۴۲



شکل ۳-۵: نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی توسط مدل‌هایGC-RF و SR-RF برای داده‌های سری آزمون

۳-۶-۳-۲- ارزیابی مدل‌های RF با استفاده از نمودار خطای باقی‌مانده

اختلاف مقادیر پیش‌بینی‌شـده و مقادیر تجربی، خطای باقیمانده نامیده می‌شود. توزیع متقارن داده‌ها حول محور افقی (خطای صـفر) حاکی از عدم وجود خطای سـیسـتماتیک است. نمودار خطای باقیمانده برحسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل ۳-۶ نشان داده شده است.



شکل ۳-۶: نمودار خطای باقیمانده برحسب مقادیر تجربی برای داده‌های سری آزمون با مدل‌های جنگل تصادفیGC-RF و SR-RF

۳-۶-۳-۳- ارزیابی مدل RF به وسیله روش رد مرحله‌ای تک‌تک

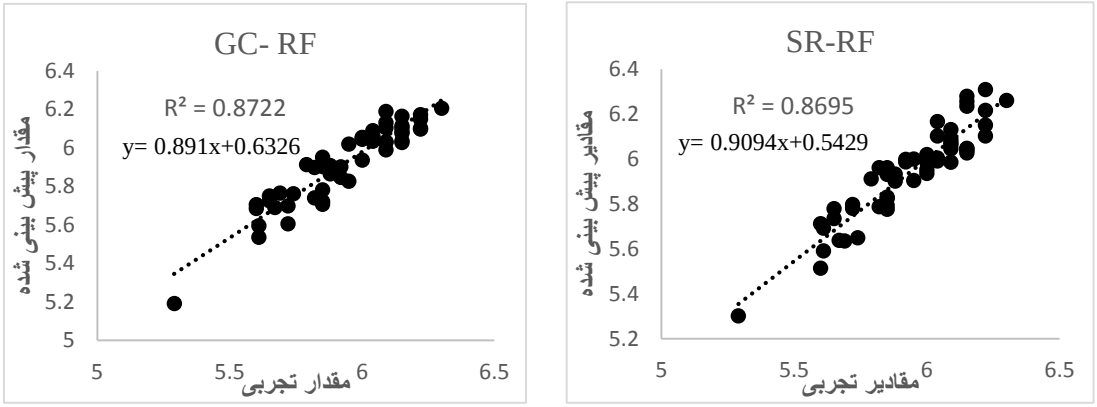
برای ارزیابی مدل توسط این روش هر بار یک ترکیب از۵۳ ترکیب به عنوان سری آزمون کنار گذاشته شد و در شرایط بهینه به دست آمده برای سری آموزش با ترکیبات باقی‌مانده مدل‌سازی صورت گرفت. سپس مدل بدست آمده برای پیش‌بینی pEC_{50} ترکیب کنار گذاشته شده به کار گرفته شد و این فرآیند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از این روش در جدول ۳-۹ ارائه شده است. نتایج به دست آمده در شکل ۳-۷ نشان می‌دهند که مدل جنگل‌های تصادفی توانایی خوبی برای پیش‌بینی فعالیت ضد ایدز مشتقات پیریدینون دارد. همچنین در شکل ۳-۸ که مقادیر باقی‌مانده‌ها بر حسب مقادیر تجربی pEC_{50} ترکیبات مورد بحث ترسیم شده است پراکندگی نقاط در دو طرف محور افقی متقارن است که نشان دهنده‌ی عدم وجود خطای سیستماتیک برای بیشتر ترکیبات است.

جدول ۳-۹ : نتایج حاصل از رد مرحله‌ای تک‌تک برای مدل‌های به دست آمده توسط روش جنگل‌های تصادفی

شماره ترکیب	مقدار تجربی (pEC50)	مقدار پیش‌بینی		درصد خطا	
		GC-RF	SR - RF	GC- RF	SR - RF
۱	۵/۶۷	۵/۶۹	۵/۶۴	۰/۳۵	-۰/۵۲
۲	۶/۰۴	۶/۰۷	۵/۹۹	۰/۴۹	-۰/۸۲
۳	۶/۰۰	۶/۰۵	۵/۹۸	۰/۸۳	-۰/۳۳
۴	۶/۲۲	۶/۱۰	۶/۱۰	-۱/۹۲	-۱/۹۲
۵	۶/۱۵	۶/۱۶	۶/۰۵	۰/۱۶	-۱/۶۲
۶	۵/۷۴	۵/۷۶	۵/۶۵	۰/۳۴	-۱/۵۶
۷	۶/۰۹	۶/۰۳	۶/۰۴	-۰/۹۸	-۰/۸۲
۸	۵/۷۹	۵/۹۱	۵/۹۱	۲/۱۱	۲/۰۷
۹	۶/۰۹	۶/۰۳	۶/۱۳	-۰/۹۸	۰/۶۵
۱۰	۶/۵۲	۶/۰۵	۶/۰۲	۰/۸۳	۰/۳۳
۱۱	۵/۶۹	۵/۷۷	۵/۶۳	۱/۴۰	-۱/۰۵
۱۲	۵/۹۵	۵/۸۳	۵/۹۰	-۲/۰۱	-۰/۸۴
۱۳	۵/۹۲	۵/۸۵	۶/۰۰	-۱/۱۸	۱/۳۵
۱۴	۶/۱۵	۶/۰۳	۶/۰۴	-۱/۹۵	-۱/۷۸
۱۵	۵/۶۱	۵/۶۰	۵/۶۹	-۰/۱۷	۱/۴۲
۱۶	۵/۶۰	۵/۷۱	۵/۷۱	۱/۹۶	۱/۹۶
۱۷	۶/۱۵	۶/۱۱	۶/۰۳	-۰/۶۵	-۱/۹۵
۱۸	۶/۲۲	۶/۱۷	۶/۲۲	-۰/۸۰	۰/۰۰
۱۹	۵/۸۵	۵/۷۲	۵/۷۹	-۲/۲۲	-۱/۰۲
۲۰	۵/۸۵	۵/۹۵	۵/۹۶	۱/۷۰	-۱/۳۶
۲۱	۶/۰۹	۶/۱۰	۶/۰۶	۰/۱۶	-۰/۴۹
۲۲	۵/۶۱	۵/۵۴	۵/۵۹	-۱/۲۴	-۰/۳۵
۲۳	۵/۶۰	۵/۶۹	۵/۵۱	۱/۶۰	-۱/۶۰
۲۴	۵/۸۵	۵/۹۰	۵/۸۳	۰/۸۵	-۰/۳۴
۲۵	۵/۸۵	۵/۹۲	۵/۹۶	۱/۱۹	۱/۸۸
۲۶	۶/۲۲	۶/۱۰	۶/۳۱	-۱/۹۲	۱/۴۴
۲۷	۵/۶۵	۵/۷۲	۵/۷۸	۱/۲۳	۲/۳۰
۲۸	۶/۱۵	۶/۱۱	۶/۲۸	-۰/۶۵	۲/۱۱
۲۹	۶/۰۴	۶/۰۳	۶/۱۰	-۰/۱۶	۰/۹۹
۳۰	۵/۸۵	۵/۷۸	۵/۸۲	-۱/۱۹	-۰/۵۱
۳۱	۵/۸۵	۵/۷۱	۵/۹۳	-۲/۳۹	۱/۳۶

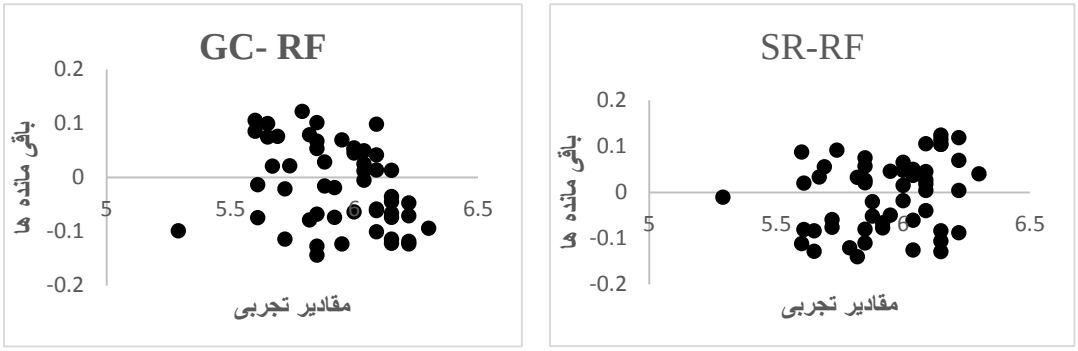
ادامه جدول(۳-۹)

شماره ترکیب	مقدار تجربی (pEC50)	مقدار پیش‌بینی		درصد خطا	
		GC-RF	SR - RF	GC- RF	SR - RF
۳۲	۵/۹۲	۵/۹۰	۵/۹۹	-۰/۳۳	۱/۱۸
۳۳	۶/۱۵	۶/۰۸	۶/۲۳	-۱/۱۳	۱/۳۰
۳۴	۵/۲۹	۵/۱۹	۵/۳۰	-۱/۸۹	۰/۱۸
۳۵	۵/۸۲	۵/۹۰	۵/۹۶	۱/۳۷	۲/۴۰
۳۶	۶/۰۹	۶/۱۳	۶/۰۹	۰/۶۵	۰/۰۰
۳۷	۵/۶۵	۵/۷۵	۵/۷۳	۱/۷۶	۱/۴۱
۳۸	۵/۷۲	۵/۶۱	۵/۸۰	-۱/۹۲	۱/۳۹
۳۹	۵/۷۲	۵/۷۰	۵/۷۸	-۰/۳۴	۱/۰۴
۴۰	۶/۰۴	۶/۰۵	۶/۱۷	۰/۱۶	۲/۱۵
۴۱	۶/۰۴	۶/۰۹	۶/۰۰	۰/۸۲	-۰/۶۶
۴۲	۶/۱۵	۶/۰۴	۶/۲۶	-۱/۷۸	۱/۷۸
۴۳	۵/۹۵	۶/۰۱	۶/۰۰	۱/۱۷	۰/۸۴
۴۴	۶/۰۹	۵/۹۹	۵/۹۸	-۱/۶۴	-۱/۸۰
۴۵	۶	۵/۹۴	۵/۹۵	-۱/۰۰	-۰/۸۳
۴۶	۶/۰۹	۶/۱۹	۶/۰۷	۱/۶۴	-۰/۳۲
۴۷	۵/۸۲	۵/۷۴	۵/۷۹	-۱/۳۷	-۰/۵۱
۴۸	۶/۲۲	۶/۱۵	۶/۱۵	-۱/۱۲	-۱/۱۲
۴۹	۶/۱۵	۶/۰۸	۶/۰۵	-۱/۱۳	-۱/۶۲
۵۰	۵/۸۸	۵/۸۶	۵/۹۰	-۰/۳۴	۰/۳۴
۵۱	۶	۶/۰۵	۵/۹۳	۰/۸۳	-۱/۱۶
۵۲	۶/۳۰	۶/۲۱	۶/۲۶	-۱/۴۲	-۰/۶۳
۵۳	۵/۸۸	۵/۹۱	۵/۹۳	۰/۵۱	۰/۸۵



شکل ۳-۷: نمودار مقادیر پیش‌بینی شده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک

توسط مدل‌های جنگل‌های تصادفی (GC-RF و SR-RF) بر حسب مقادیر تجربی

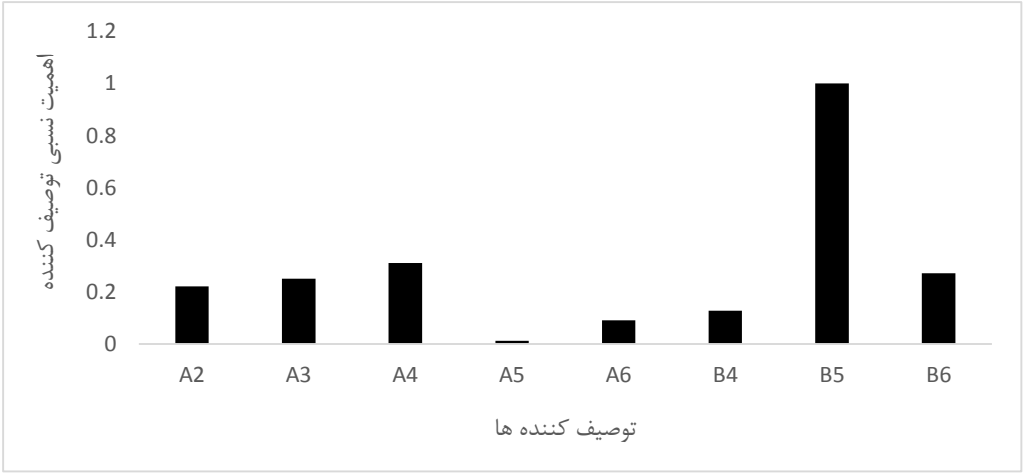


شکل ۳-۸: نمودار مقادیر خطای باقی‌مانده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل‌های

جنگل‌های تصادفی (GC-RF و SR-RF) بر حسب مقادیر تجربی

۳-۶-۴- انتخاب توصیف کننده‌های مهم با روش جنگل‌های تصادفی

روش جنگل‌های تصـادفی علاوه بر اینکه به عنوان یک روش مدل‌سـازی به کار می‌رود به عنوان یک روش انتخاب متغیر نیز به کار برده می‌شـود. پس از سـاخت مدل جنگل‌های تصـادفی و بهینه کردن پارامترهای موثر، الگوریتم جنگل‌های تصادفی با استفاده از مقادیر بهینه صد بار تکرار شد و در هر بار میزان کاهش MSE به عنوان مقدار اهمیت توصـیف‌کننده‌ها محاسبه گردید. سرانجام میانگین مقادیر اهمیت هر توصـیف‌کننده به عنوان شـاخص نهایی در نظر گرفته شـد و سـپس اهمیت نسـبی هر توصیف‌کننده نسبت به توصیف‌کننده‌ای که بیشترین اهمیت را داراست، تعیین شد. شکل ۳-۹ میزان اهمیت نسـبی توصـیف‌کننده‌ها را نمایش می‌دهد و همانطور که مشـاهده می‌شـود بعضـی از توصیف کننده‌ها دارای اهمیت نسـبی کمی بوده و تنها ۴ توصـیف کننده اهمیت نسبی قابل توجهی دارند که در جدول ۳-۱۰ این توصیف‌کننده‌ها همراه با مقدار اهمیت نسبی و ضریب اثر آن‌ها گردآوری شده اند.



شکل ۳-۹: نمودار اهمیت نسبی توصیف کننده‌های سهم گروه

جدول ۳-۱۰: توصیف کننده‌های برتر با استفاده از روش جنگل‌های تصادفی

No	علامت	مفهوم توصیف کننده	اهمیت نسبی
۱	A_4	حضور یا عدم حضور استخلاف در موقعیت پارا حلقهA	۰/۳۱۱۳
۲	A_6	حضور یا عدم حضور استخلاف در موقعیت ارتو حلقهA	۰/۰۹۰۶
۳	B_5	حضور یا عدم حضور استخلاف در موقعیت متا حلقه B	۱/۰۰۰۰
۴	B_6	حضور یا عدم حضور استخلاف در موقعیت ارتو حلقه B	۰/۲۷۱۱

۳-۷- مدل‌سازی شبکه عصبی مصنوعی

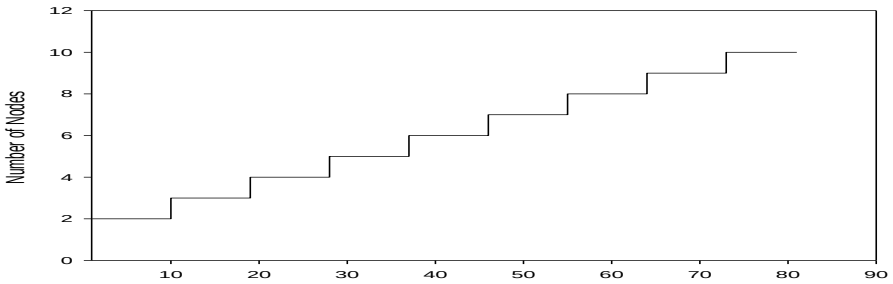
۳-۷-۱- مدل‌سازی شبکه عصبی مصنوعی با توصیف‌کننده‌های سهم گروه (GC-ANN)

یکی از راه‌های یافتن رابطه غیرخطی بین متغیرهای مستقل و متغیر وابسته، استفاده از شبکه عصبی مصنوعی برای مدل‌سازی می‌باشد. شبکه عصبی انتخاب شده در این مطالعه یک شبکه پیشرو با الگوریتم آموزشی پس انتشار می باشد که الگوریتم آن با برنامه متلب نوشته شده است در فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به سه بخش سری آزمون (۸ ترکیب) و سری ارزیابی (۸ ترکیب) و همچنین سری آموزش (۳۷ ترکیب) تقسیم شدند. ۸ توصیف کننده‌ای که از روش سهم گروه به دست آمدند به عنوان ورودی به شبکه عصبی مصنوعی داده شدند. خروجی شبکه نیز فعالیت متناظر با ورودی‌ها را نشان می‌دهد. برای آموزش شبکه، دو الگوریتم آموزشی لونبرگ-مارکوات (lm) و تنظیم بایزین (br)، و دوتابع انتقال تانژانت سیگموئید (tan sig) و لگاریتم سیگموئید (log sig) مورد استفاده قرار گرفتند که از ترکیب این دو الگوریتم و دو تابع انتقال، چهار نوع شبکه عصبی مختلف طراحی گردید. در هر یک از این شبکه‌ها باید تعداد ورودی‌ها و تعداد نرون‌های لایه پنهان (گره) و تعداد دور آموزشی بهینه گردند که مقدار بهینه هریک از این پارامترها زمانی است که میانگین مربع خطا (MSE) برای سری ارزیابی دارای کمترین مقدار باشد. برای بهینه‌سازی تعداد ورودی‌های

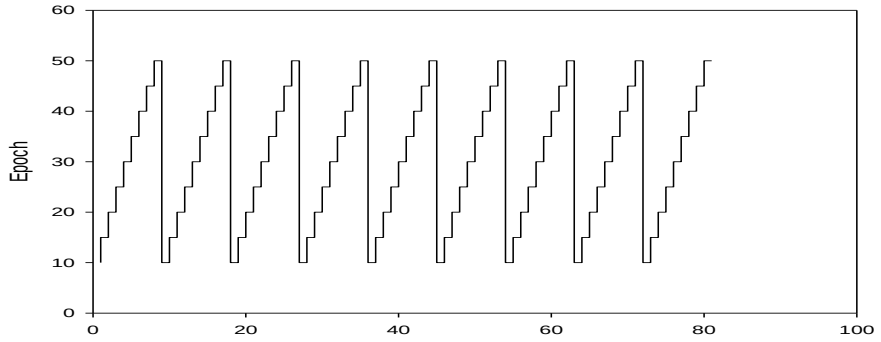
شبکه، تعداد نرون‌های لایه‌ی پنهان و تعداد دور آموزشی از روش بهینه‌سازی همزمان استفاده گردید. برای این منظور هر شبکه با تعداد ورودی ۸ و تعداد گره از ۲ تا ۱۰ و تعداد دور آموزشی از ۱۰ تا ۵۰ با گام ۵ به طور همزمان آموزش داده شـــد. در روند بهینه‌ســـازی فوق، به حداقل رســـاندن میانگین خطاهای سری ارزیابی به عنوان معیار انتخاب شد، این کار موجب می‌شود بتوانیم نقش تمام پارامترها را به طور همزمان و یکجا بررسی نماییم. بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE بدســت آمده، به صــورت نموداری برحســب یک بردار مرجع فرضی در شکل ۳-۱۰ تا شکل ۳-۱۳ آمده است.

بـا توجـه به نتایج حاصـــله برای توصـــیف‌کننده‌ها و تعداد مختلف گره‌ها و دورهای آموزش و همچنین توابع متفاوت آموزش و انتقال، بهترین شبکه‌های بدست آمده براساس کمترین مقدار MSE در جدول ۳-۱۱ خلاصه شده است.

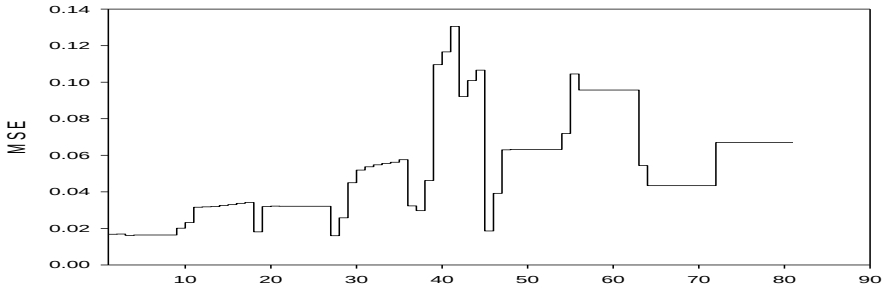
با توجه به جدول ۳-۱۱ شـبکه عصـبی با ۸ توصـیف‌کننده در لایه ورودی، سـه نرون در لایه مخفی و تعداد دور آموزشـــی ۴۵ با تابع آموزش لونبرگ_مارکوات و تابع انتقال تانژانت ســـیگموئیدی کمترین MSE را نسـبت به سـایر شـبکه‌ها نشـان می‌دهد. بنابراین شـبکه‌ای با این ساختار به عنوان بهترین مدل ANN برای مدل‌سازی داده‌ها در نظر گرفته شد.



الف)

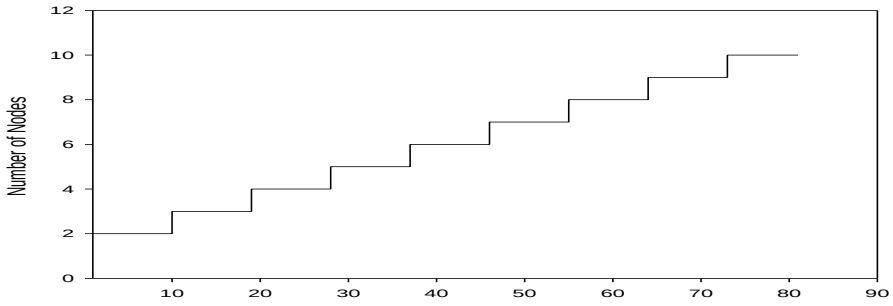


ب)

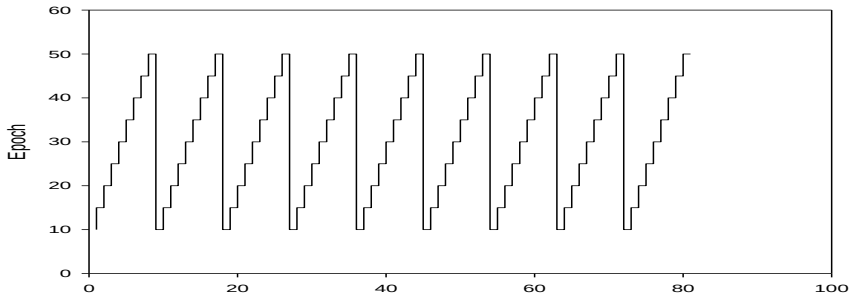


بردار مرجع ج)

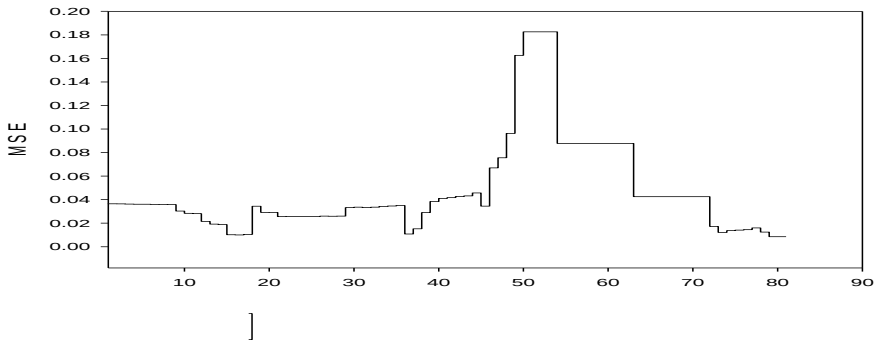
شکل ۳-۱۰: نمودارهای الف) تعداد گره‌ها، ب) تعداد دورهای آموزش و ج) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال لگاریتم سیگموئید



الف)



ب)

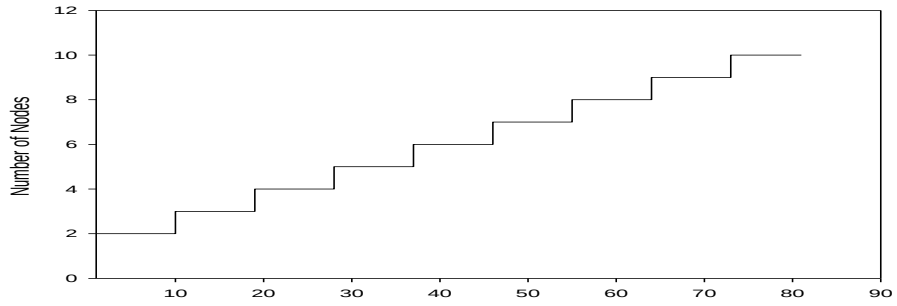


]

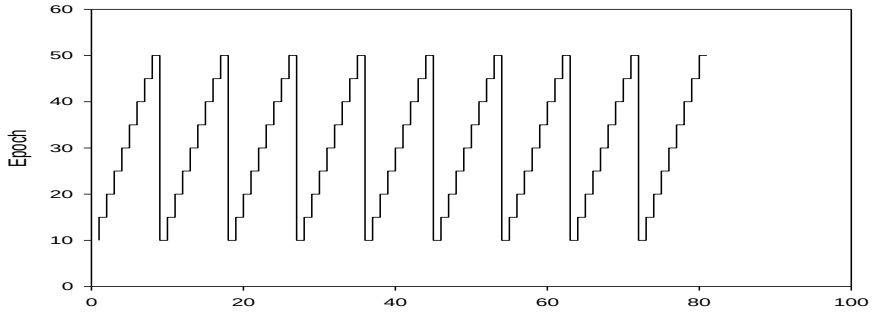
بردار مرجع

ج)

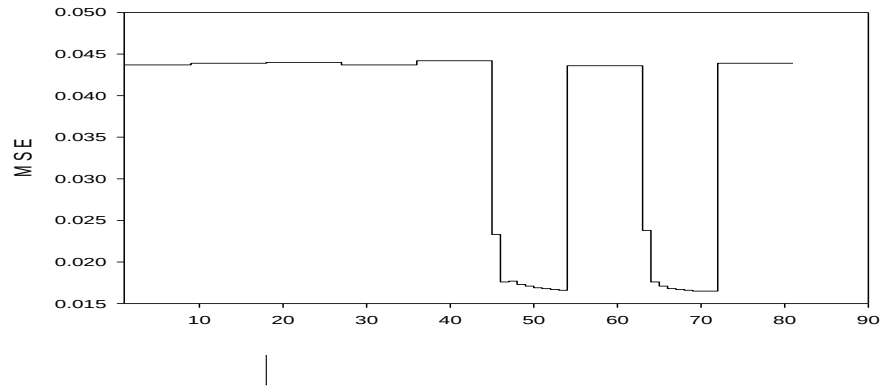
شکل ۳-۱۱: نمودارهای الف) تعداد گره‌ها، ب) تعداد دورهای آموزش و ج) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال تانژانت سیگموئید



الف)



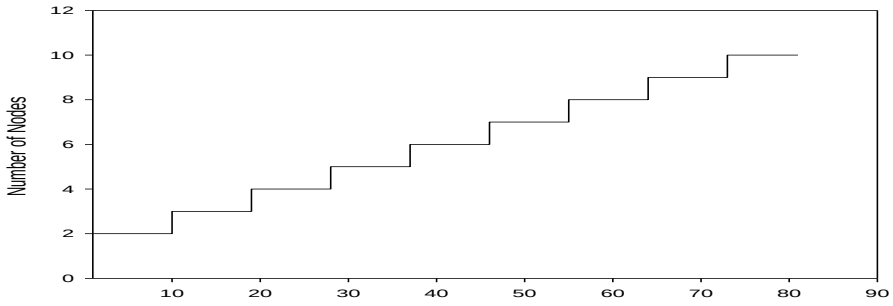
ب)



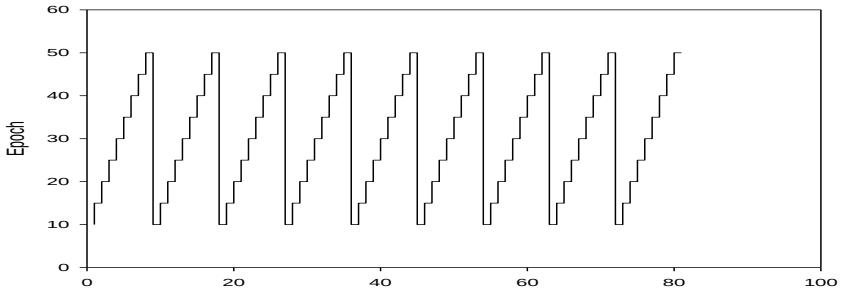
بردار مرجع

ج)

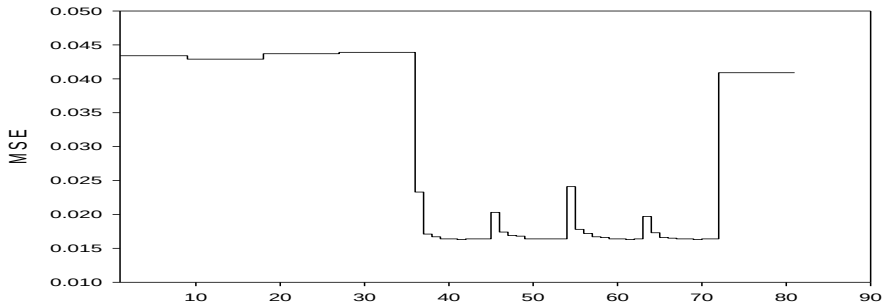
شکل ۳-۱۲: نمودارهای الف) تعداد گره‌ها، ب) تعداد دورهای آموزش و ج) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال لگاریتم سیگموئید



الف)



ب)



ج)

شکل ۳-۱۳: نمودارهای الف) تعداد گره‌ها، ب) تعداد دورهای آموزش و ج) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال تانژانت سیگموئید

جدول ۳-۱۱: توابع و پارامترهای شبکه‌های بهینه GC-ANN

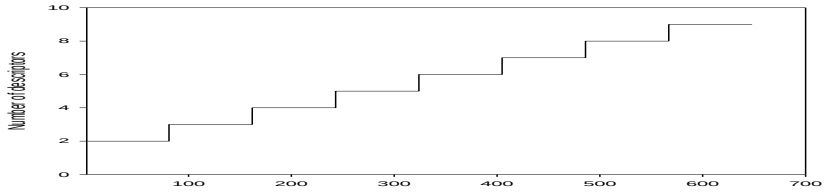
MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۰۱۶۰	۱۰	۵	لگاریتم-سیگموئید	لونبرگ- مارکوات	۸
۰/۰۱۰۰	**۴۵**	**۳**	**تانژانت-سیگموئید**	**لونبرگ- مارکوات**	**۸**
۰/۰۱۶۵	۴۰	۹	لگاریتم-سیگموئید	تنظیم بایزین	۸
۰/۰۱۶۳	۳۵	۶	تانژانت-سیگموئید	تنظیم بایزین	۸

۳-۷-۲- مدل‌سازی شبکه عصبی مصنوعی با توصیف کننده‌های انتخاب شده به وسیله روش رگرسیون مرحله‌ای (SR -ANN)

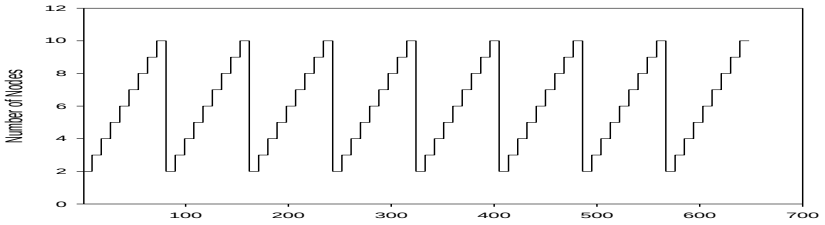
ابتدا مقادیر ۹ توصیف کننده انتخاب شده با روش SR (رگرسیون مرحله‌ای) به عنوان متغیرهای ورودی و فعالیت متناظر آن‌ها به عنوان متغیر هدف در نظر گرفته شد تا پاسخ شبکه با آن‌ها سنجیده شود. برای بدست آوردن بهترین مدل و کمترین خطا، پارامترهای موثر (تابع انتقال، تابع آموزش، تعداد متغیرهای ورودی شبکه، تعداد گره‌ها در لایه مخفی و تعداد دورهای آموزشی) به‌طور همزمان بهینه شدند. تابع کارایی شبکه نیز میانگین مربع خطا (MSE) بود.

در فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به سه بخش سری آزمون (۸ ترکیب) و سری ارزیابی (۸ ترکیب) و همچنین سری آموزش (۳۷ ترکیب) تقسیم شدند. سری ارزیابی برای محاسبه گرادیان و به روز کردن اوزان و بایاس‌های شبکه استفاده شد. از آنجایی که در بیشتر موارد به نظر می‌رسد که یک لایه پنهان در ساختار شبکه عصبی مناسب باشد، در این تحقیق نیز از یک لایه پنهان استفاده شده است [۴۰]. برای بهینه‌سازی تعداد متغیرهای ورودی شبکه، تعداد نرون‌های لایه‌ی پنهان و تعداد دور آموزشی از روش بهینه‌سازی همزمان استفاده گردید. برای یافتن مقدار بهینه پارامترهای موثر، مقادیر مختلفی از آن‌ها در ساختار شبکه قرار داده شد. برای این منظور هر شبکه با تعداد ورودی ۲ تا ۹ و با دو الگوریتم آموزشی لونبرگ-مارکوات (train lm) و تنظیم بایزین (train br) و دو تابع انتقال لگاریتم سیگموئید و تانژانت سیگموئید و تعداد گره از ۲ تا ۱۰ و تعداد دورهای

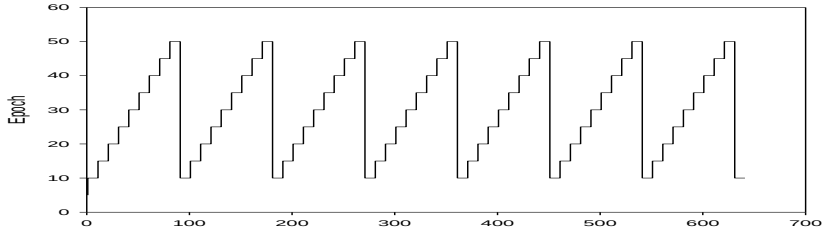
آموزش از ۱۰ تا ۵۰ با گام ۵، به طور همزمان آموزش داده شـد. در روند بهینه‌سازی فوق، به حداقل رساندن میانگین خطاهای سری ارزیابی به عنوان معیار انتخاب شد. بخشی از روند تغییرات پارامترهای شـبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE بدست آمده، به صورت نموداری برحسـب یک بردار مرجع فرضـی در شـکل ۳-۱۴ تا شـکل ۳-۱۷ آمده اسـت. نتیجه بهینه‌سازی این پارامترها در جدول ۳-۱۲ گردآوری شده است.



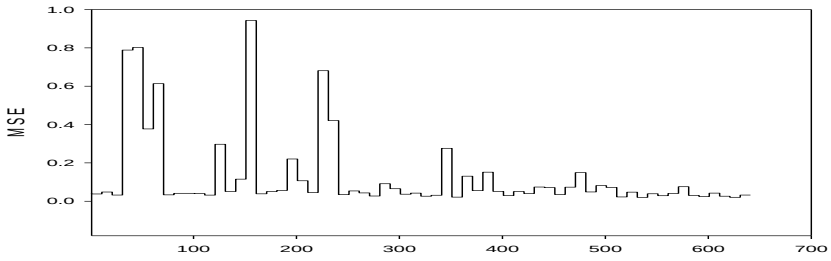
الف)



ب)



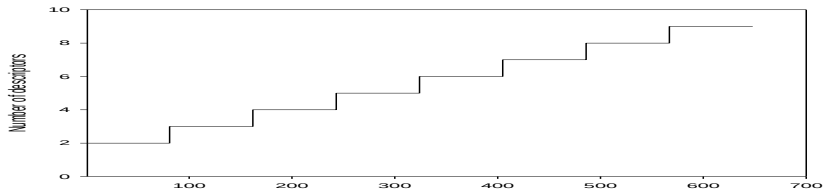
ج)



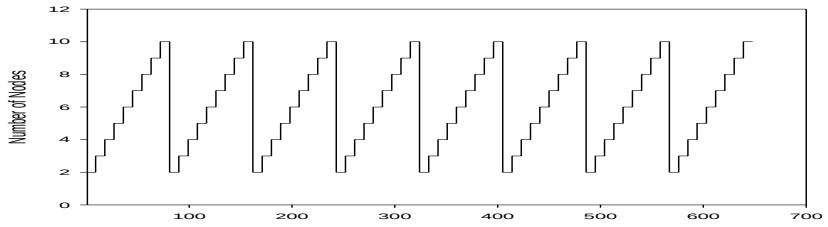
د)

بردار مرجع

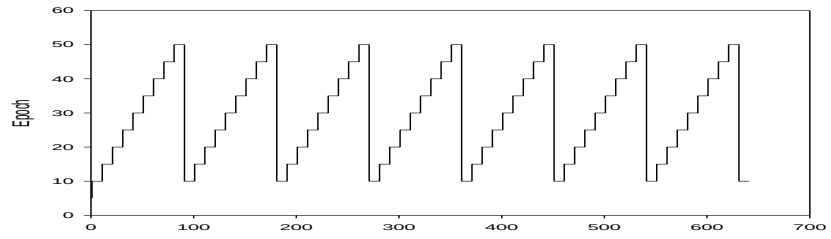
شکل ۳-۱۴: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال لگاریتم سیگموئید



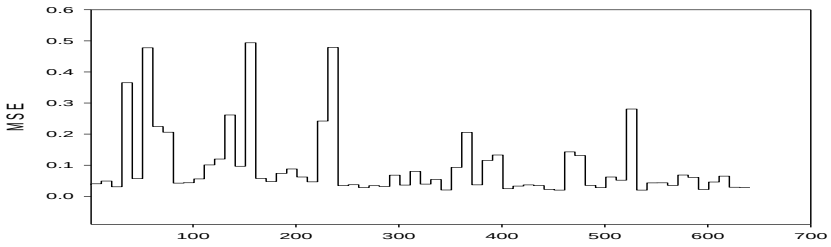
الف)



ب)



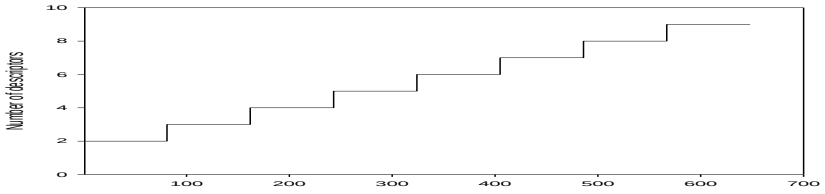
ج)



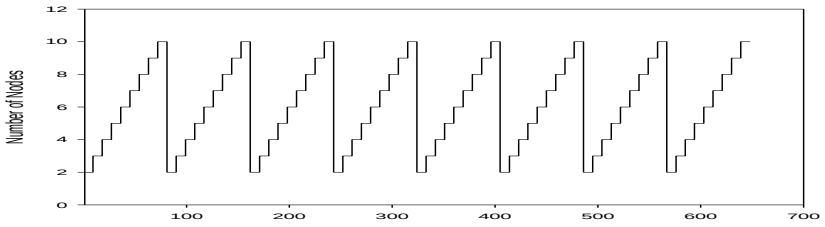
د)

بردار مرجع

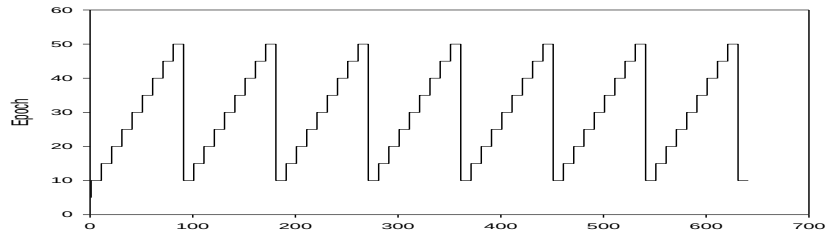
شکل ۳-۱۵: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال تانژانت سیگموئید



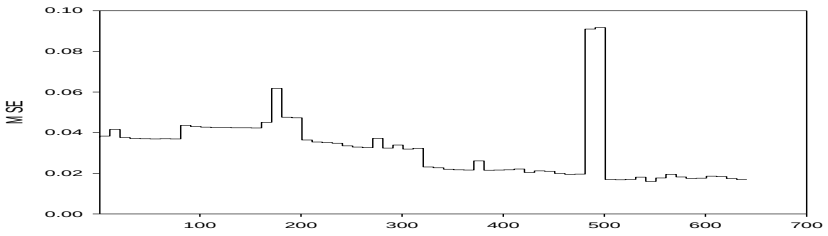
الف)



ب)



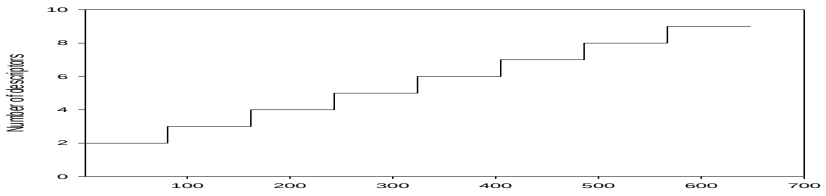
ج)



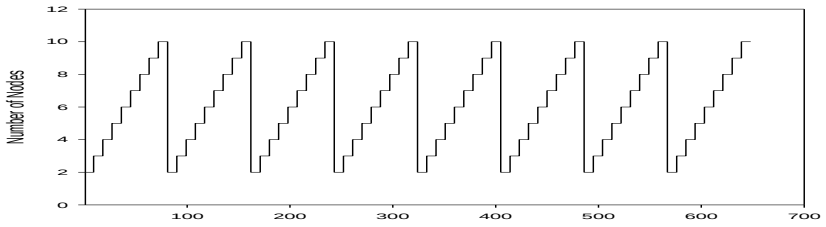
د)

بردار مرجع

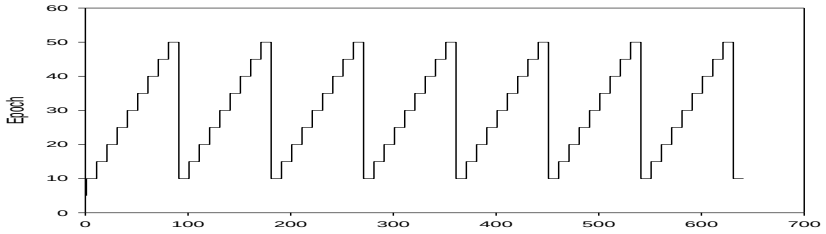
شکل ۳-۱۶: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال لگاریتم سیگموئید



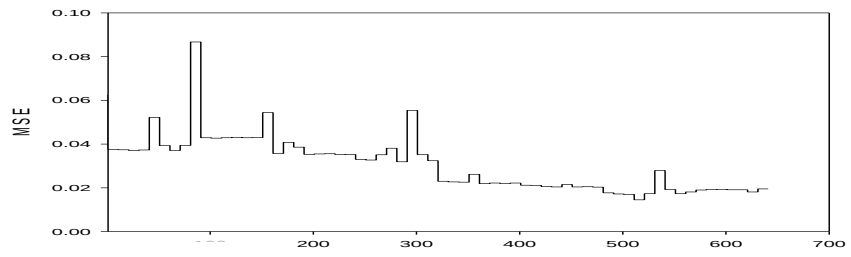
الف)



ب)



ج)



د)

بردار مرجع

شکل ۳-۱۷: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال تانژانت سیگموئید

جدول ۳-۱۲: توابع و پارامترهای شبکه‌های بهینه SR-ANN

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۰۱۶۱	۱۰	۶	لگاریتم-سیگموئید	لونبرگ- مارکوات	۹
۰/۰۰۹۶	**۱۵**	**۷**	**تانژانت-سیگموئید**	**لونبرگ- مارکوات**	**۸**
۰/۰۱۵۶	۱۵	۶	لگاریتم-سیگموئید	تنظیم بایزین	۸
۰/۰۱۴۵	۴۵	۵	تانژانت-سیگموئید	تنظیم بایزین	۸

با توجه به نتایج بدست آمده، الگوریتم آموزشی لونبرگ –مارکوات و تابع انتقال تانژانت سیگموئید، ۸ توصیف کننده ورودی و ۷ نرون در لایه مخفی با تعداد دور آموزشی ۱۵ کمترین MSE را نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی در نظر گرفته شد.

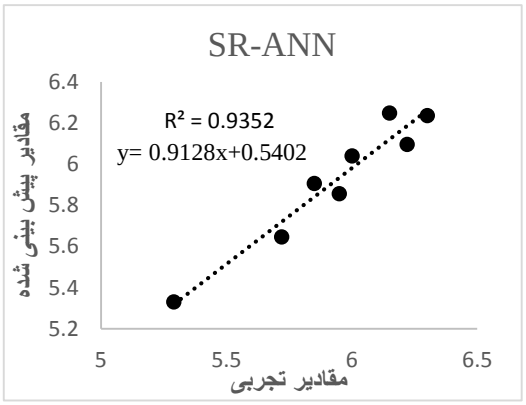
۳-۷-۳- ارزیابی مدل‌های شبکه عصبی مصنوعی

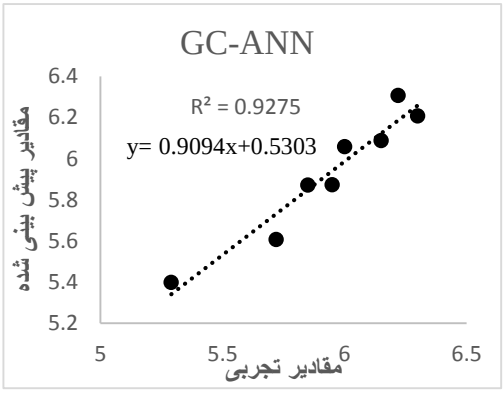
۳-۷-۳-۱- ارزیابی مدل‌های ANN با استفاده از داده‌های سری آزمون

جدول ۳-۱۳ نتایج پیش‌بینی حاصل از ارزیابی مدل‌های شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون و شکل ۳-۱۸ نیز نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد.

جدول ۳-۱۳: نتایج حاصل از ارزیابی مدل‌های شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		GC-ANN	SR -ANN	GC-ANN	SR -ANN
۱۰	۶/۰۰	۶/۰۶	۵/۸۷	۱/۰۰	-۲/۱۶
۱۲	۵/۹۵	۵/۸۷	۶/۰۲	-۱/۳۴	۱/۱۷
۱۸	۶/۲۲	۶/۳۱	۶/۱۹	۱/۴۴	-۰/۴۸
۲۰	۵/۸۵	۵/۸۷	۵/۹۳	۰/۳۴	۱/۳۶
۳۳	۶/۱۵	۶/۰۹	۶/۱۱	-۰/۹۷	-۰/۶۵
۳۴	۵/۲۹	۵/۴۰	۵/۳۵	۲/۰۷	۱/۱۳
۳۹	۵/۷۲	۵/۶۱	۵/۸۲	-۱/۹۲	۱/۷۴
۵۲	۶/۳۰	۶/۲۱	۶/۳۹	-۱/۴۲	۱/۴۲

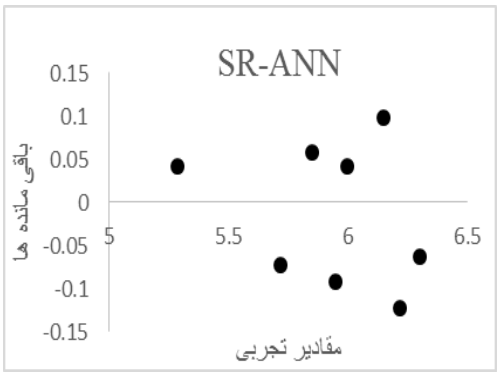


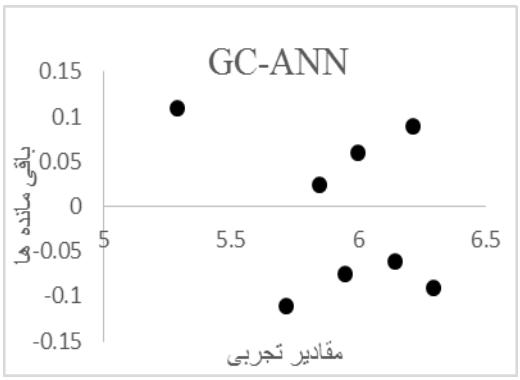


شکل ۳-۱۸: نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون

۳-۷-۳-۲- ارزیابی مدل‌های ANN با استفاده از نمودار خطای باقی‌مانده

شـکل ۳-۱۹ نمودار خطای باقیمانده برحسـب مقادیر تجربی، برای مدل‌های ذکر شده را نشان می‌دهد. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است.





شکل ۳-۱۹: نمودار خطای باقیمانده برحسب مقادیر تجربی برای داده‌های سری آزمون

۳-۷-۳-۳- ارزیابی مدل‌های ANN توسط روش رد مرحله‌ای تک‌تک

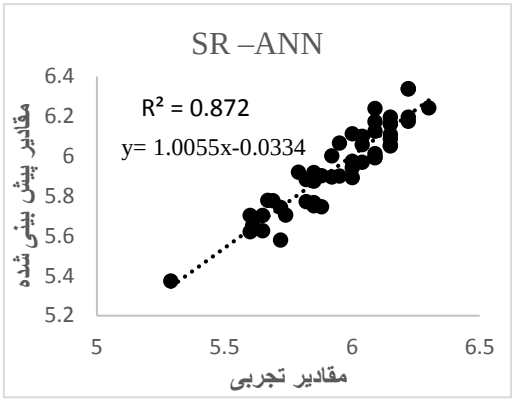
برای ارزیابی مدل‌های پیشنهادی از روش رد مرحله‌ای تک‌تک نیز استفاده گردید. نتایج حاصل از این روش در جدول ۳-۱۴ ارائه شده است. نتایج به دست آمده در شکل ۳-۲۰ نشان می‌دهند که مدل شبکه عصبی مصنوعی توانایی خوبی برای پیش‌بینی فعالیت ضد ایدز مشتقات پیریدینون دارد. همچنین شکل ۳-۲۱ که مقادیر باقی‌مانده‌ها بر حسب مقادیر تجربی pEC_{50} ترکیبات مورد بحث ترسیم شده است دارای تقارن پراکندگی نقاط در دو طرف محور افقی است که نشان دهنده‌ی عدم وجود خطای سیستماتیک برای بیشتر ترکیبات است.

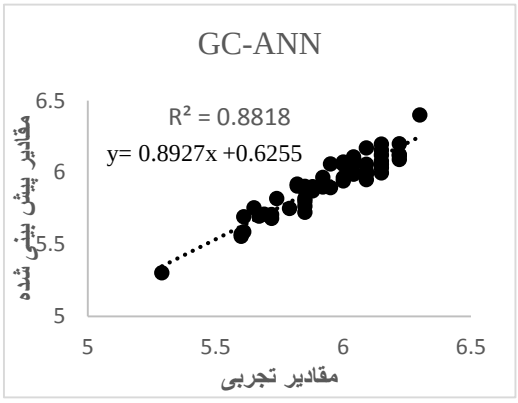
جدول ۳-۱۴: نتایج حاصل از ارزیابی مدل‌های (GC-ANN و SR-ANN) به دست آمده توسط روش رد مرحله‌ای تک‌تک

شماره ترکیب	مقدار تجربی (pEC50)	مقدار پیش‌بینی		درصد خطا	
		GC-ANN	SR -ANN	GC-ANN	SR -ANN
۱	۵/۶۷	۵/۷۰	۵/۵۴	۰/۵۲	-۲/۲۹
۲	۶/۰۴	۶/۱۱	۵/۹۴	۱/۱۵	-۱/۶۵
۳	۶/۰۰	۶/۰۷	۵/۹۴	۱/۱۶	-۱/۰۰
۴	۶/۲۲	۶/۱۳	۶/۲۲	-۱/۴۴	۰/۰۰
۵	۶/۱۵	۶/۰۰	۶/۲۴	-۲/۴۳	-۱/۴۶
۶	۵/۷۴	۵/۸۲	۵/۶۷	۱/۳۹	-۱/۲۱
۷	۶/۰۹	۶/۰۱	۵/۹۹	-۱/۳۱	-۱/۶۴
۸	۵/۷۹	۵/۷۵	۵/۷۰	-۰/۶۹	۱/۵۵
۹	۶/۰۹	۵/۹۹	۶/۰۶	-۱/۶۴	-۰/۴۹
۱۰	۶/۰۰	۶/۰۶	۵/۹۵	۱/۰۰	-۰/۸۳
۱۱	۵/۶۹	۵/۷۱	۵/۷۲	۰/۳۵	۰/۵۲
۱۲	۵/۹۵	۵/۹۰	۵/۸۷	-۰/۸۴	-۱/۳۴
۱۳	۵/۹۲	۵/۹۰	۶/۰۱	-۰/۳۳	۱/۵۲
۱۴	۶/۱۵	۶/۱۶	۶/۰۹	۰/۱۶	-۰/۹۷
۱۵	۵/۶۱	۵/۶۹	۵/۵۳	۱/۴۲	-۱/۴۲
۱۶	۵/۶۰	۵/۵۷	۵/۵۷	-۰/۵۳	-۰/۵۳
۱۷	۶/۱۵	۶/۲۰	۶/۲۰	۰/۸۱	-۱/۸۱
۱۸	۶/۲۲	۶/۰۹	۶/۲۷	-۲/۰۹	۰/۸۰
۱۹	۵/۸۵	۵/۷۷	۵/۹۷	-۱/۳۶	۲/۰۵

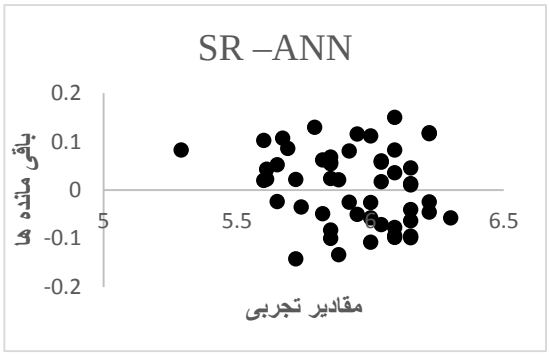
ادامه جدول (۳-۱۴)

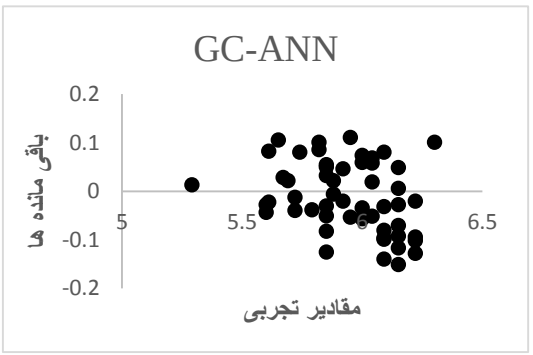
شماره ترکیب	مقدار تجربی (pEC50)	مقدار پیش‌بینی		درصد خطا	
		GC-ANN	SR -ANN	GC-ANN	SR -ANN
۲۰	۵/۸۵	۵/۸۲	۵/۹۴	-۰/۵۱	۱/۵۳
۲۱	۶/۰۹	۵/۹۹	۶/۰۰	-۱/۶۴	-۱/۴۷
۲۲	۵/۶۱	۵/۵۹	۵/۶۹	-۰/۳۵	۱/۴۲
۲۳	۵/۶۰	۵/۵۶	۵/۶۸	-۰/۷۱	۱/۴۲
۲۴	۵/۸۵	۵/۸۰	۵/۷۳	-۰/۸۵	-۲/۰۵
۲۵	۵/۸۵	۵/۹۰	۵/۷۲	۰/۸۵	-۲/۲۲
۲۶	۶/۲۲	۶/۱۲	۶/۳۳	-۱/۶۰	۱/۷۸
۲۷	۵/۶۵	۵/۹۰	۵/۵۳	۰/۸۵	-۲/۱۲
۲۸	۶/۱۵	۶/۰۳	۶/۱۷	-۱/۹۵	۰/۳۲
۲۹	۶/۰۴	۶/۱۰	۶/۱۳	۰/۹۹	۱/۴۹
۳۰	۵/۸۵	۵/۷۲	۵/۸۰	-۲/۲۲	-۰/۸۵
۳۱	۵/۸۵	۵/۸۸	۵/۷۱	۰/۵۱	-۲/۳۹
۳۲	۵/۹۲	۵/۹۷	۶/۰۳	۰/۸۴	۱/۸۵
۳۳	۶/۱۵	۶/۰۸	۶/۱۰	-۱/۱۳	-۰/۸۱
۳۴	۵/۲۹	۵/۳۰	۵/۲۷	۰/۱۸	-۰/۳۷
۳۵	۵/۸۲	۵/۹۱	۵/۹۰	۱/۵۴	۱/۳۷
۳۶	۶/۰۹	۶/۰۶	۶/۱۵	-۰/۴۹	۰/۹۸
۳۷	۵/۶۵	۵/۷۶	۵/۷۳	۱/۹۴	۱/۴۱
۳۸	۵/۷۲	۵/۶۸	۵/۸۰	-۰/۶۹	۱/۳۹
۳۹	۵/۷۲	۵/۷۱	۵/۷۹	-۰/۱۷	۱/۲۲
۴۰	۶/۰۴	۶/۰۶	۶/۱۳	۰/۳۳	۱/۴۹
۴۱	۶/۰۴	۵/۹۹	۶/۰۹	-۰/۸۲	۰/۸۲
۴۲	۶/۱۵	۶/۰۶	۶/۰۵	-۱/۴۶	-۱/۶۲
۴۳	۵/۹۵	۶/۰۶	۶/۰۱	۱/۸۴	۱/۰۰
۴۴	۶/۰۹	۵/۹۵	۶/۰۷	-۲/۲۹	-۰/۳۲
۴۵	۶	۵/۹۷	۶/۰۷	-۰/۵۰	-۱/۱۶
۴۶	۶/۰۹	۶/۱۷	۶/۱۵	۱/۳۱	۰/۹۸
۴۷	۵/۸۲	۵/۹۲	۵/۷۷	۱/۷۱	-۰/۸۵
۴۸	۶/۲۲	۶/۲۰	۶/۳۰	-۰/۳۲	۱/۲۸
۴۹	۶/۱۵	۶/۱۲	۶/۰۷	-۰/۴۸	-۱/۳۰
۵۰	۵/۸۸	۵/۸۷	۵/۹۹	-۰/۱۷	۱/۸۷
۵۱	۶	۵/۹۴	۵/۹۷	-۱/۰۰	-۰/۵۰
۵۲	۶/۳۰	۶/۴۰	۶/۱۷	۱/۵۸	-۲/۰۶
۵۳	۵/۸۸	۵/۹۰	۵/۹۳	۰/۳۴	۰/۸۵





شکل ۳-۲۰: نمودار مقادیر پیش‌بینی شده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل شبکه عصبی مصنوعی بر حسب مقادیر تجربی





شکل ۳-۲۱: نمودار مقادیر خطای باقی‌مانده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک توسط مدل شبکه عصبی مصنوعی بر حسب مقادیر تجربی

۳-۷-۴- مقایسه مدل‌های ارائه شده با استفاده از پارامترهای آماری

همـان‌طور کـه قبلاً بیـان شـــد، پارامترهای آماری مقادیر تعیین‌کننده‌ای برای ارزیابی مدل‌ها هستند. مطابق جدول۳-۱۵ از این مقادیر برای ارزیابی و مقایسه مدل‌های ساخته شده استفاده گردید. این پارامترها برای سـری ارزیابی، سری آزمون و کل داده‌ها به روش رد مرحله‌ای تک تک و مطابق با روابط توضیح داده شده در بخش ۲-۱۵-۱ محاسبه شده اند.

جدول۳-۱۵: پارامترهای آماری محاسبه شده برای ارزیابی و مقایسه مدل‌های SR-ANN ، GC-ANN ، SR-RF و GC-RF

		SR-ANN	GC-ANN	SR-RF	GC-RF
R	سری ارزیابی	۰/۹۸۳	۰/۹۷۳	—	—
	سری تست	۰/۹۶۷	۰/۹۶۳	۰/۹۴۸	۰/۹۶۳
	کل داده‌ها	۰/۹۳۴	۰/۹۳۹	۰/۹۳۲	۰/۹۳۴
R^2	سری ارزیابی	۰/۹۴۱	۰/۹۴۱	—	—
	سری تست	۰/۹۳۵	۰/۹۲۷	۰/۸۹۲	۰/۹۲۴
	کل داده‌ها	۰/۸۷۲	۰/۸۸۲	۰/۸۶۹	۰/۸۷۲
PRESS	سری ارزیابی	۰/۰۲۰	۰/۰۲۰	—	—
	سری تست	۰/۰۴۰	۰/۰۵۴	۰/۰۹۴	۰/۰۶۶
	کل داده‌ها	۰/۳۵۹	۰/۲۷۷	۰/۳۱۴	۰/۳۱۴
MSE	سری ارزیابی	۰/۰۰۲	۰/۰۰۳	—	—
	سری تست	۰/۰۰۵	۰/۰۰۷	۰/۰۱۷	۰/۰۰۷
	کل داده‌ها	۰/۰۰۷	۰/۰۰۵	۰/۰۰۶	۰/۰۰۶
MAE	سری ارزیابی	۰/۰۰۶	۰/۰۴۷	—	—
	سری تست	۰/۰۰۸	۰/۰۷۷	۰/۰۸۹	۰/۰۷۶
	کل داده‌ها	۰/۰۰۱	۰/۰۶۲	۰/۰۶۷	۰/۰۶۸
MRE	سری ارزیابی	۰/۰۶۳	۰/۷۹۸	—	—
	سری تست	۰/۵۰۰	۱/۱۳۵	۱/۵۱۰	۱/۲۹۱
	کل داده‌ها	۰/۰۳۸	۱/۰۴۱	۱/۱۳۱	۱/۱۴۲
SEP	سری ارزیابی	۰/۰۵۰	۰/۰۵۰	—	—
	سری تست	۰/۰۷۱	۰/۰۸۲	۰/۰۹۷	۰/۸۱۱
	کل داده‌ها	۰/۰۸۱	۰/۰۷۲	۰/۰۷۷	۰/۰۷۶۷
REP(%)	سری ارزیابی	۰/۸۴۹	۰/۸۵۲	—	—
	سری تست	۱/۱۹۰	۱/۳۸۶	۱/۶۴۸	۱/۳۸۶
	کل داده‌ها	۱/۳۴۹	۱/۲۲۰	۱/۲۹۵	۱/۲۹۹

۳-۸- بررسی ارتباط بین توصیف کننده‌های منتخب و فعالیت مورد نظر

در این قسمت با توجه به توصیف کننده‌های وارد شده در مدل‌های RF و ANN یک بررسی اجمالی روی اثرات این متغیرها بر فعالیت دارویی ترکیبات مورد مطالعه صورت خواهد گرفت. بهترین مدل انتخاب شده شامل ۸ توصیف‌کننده است که هر کدام بیانگر خصوصیت دو بعدی، هندسی و سه بعدی ترکیبات مورد بررسی می‌باشند. جدول ۳-۱۶ مقادیر اثر متوسط[۱] توصیف‌کننده‌های مدل برتر یعنی SR-ANN را نشان می‌دهد. اثر متوسط یک متغیر مستقل با استفاده از فرمول (۳-۱) به دست می آید:

(۳-۱)

$$T=\frac{\beta_x \sum X_n}{\sum pEC_{50}}$$

که در آن β_x ضریب متغیر مستقل X در مدل، X_n مقدار متغیر مستقل X مورد نظر برای ترکیب nام می‌باشد.

جدول ۳-۱۶: اثر متوسط توصیف‌کننده‌های به دست آمده توسط روش رگرسیون مرحله ای

Variable	B07NCl	Mor25m	MATS8m	Mor29v	RDF110u	TF..Cl	G1p	ClC5
Mean effect	-۸/۵۷۶	۶/۶۹۰	۳/۲۲۲	-۳/۵۵۰	-۳/۹۳۶	-۴/۵۶۹	-۳/۸۹۵	-۲/۹۳۲

الف) توصیف کننده‌های 2D binary fingerprints

این توصیف کننده حضور یا عدم حضور N-Cl در فاصله توپولوژیکی ۷ را نشان می‌دهد. اثر انگشت باینری اندازه مولکول‌ها را به‌طور غیر مستقیم بررسی می‌کند. در حالت تجربی مولکول‌های بزرگتر، دارای ویژگی‌های متنوع‌تری هستند. در حالت تئوری فاصله توپولوژیکی مولکول‌های بزرگ ثابت است و اثر انگشت باینری حضور یا عدم حضور ویژگی‌های موثر در عملکرد مولکول را نسبت به حالتی‌که

[۱]- Mean effect

ویژگی‌های واقعی بررسی شده‌اند مورد بررسی قرار می‌دهد. توصیف کننده‌های اثر انگشتی اندازه مولکول را بررسی می‌کنند و همه این ویژگی‌های بررسی شده برای مولکول باید با اندازه مولکول در سطح اولیه همبسته باشد. که از میان توصیف کننده‌های اثر انگشت B07NCl در مدل برتر دیده می‌شود [۴۱].

ب) توصیف کننده های 3D MORSE

توصیف کننده های 3D - Morse (نمایش سه بعدی ساختار مولکول براساس تفرق الکترون) از طریق معادله تبدیلی که در پراش الکترون استفاده می‌شود، محاسبه می‌گردند:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (۲-۳)$$

I شدت الکترون پراکنده شده، A_i و A_j خاصیت اتمی اتم i و j ، s زاویه پراکندگی، r_{ij} فاصله بین اتم‌های i و j ، N تعداد کل اتم‌ها را نشان‌می‌دهد. این روش باعث می‌شود که ساختار سه بعدی مولکول به یک کد ثابت تبدیل شود. این توصیف کننده ها قادرند ارتباط بین ساختار سه بعدی ترکیبات آلی و خصوصیات فیزیکی، شیمیایی و بیولوژیکی را فراهم سازند [۱۹]. برخی از این توصیف کننده های سه بعدی مانند Mor25m[۱] و Mor29v[۲] در مدل برتر دیده می‌شوند، که به ترتیب توسط جرم اتمی و حجم واندروالس وزن‌دار شده اند و این نشان می‌دهد که حجم کلی و جرم اتمی استخلاف، فاکتوری تعیین کننده برای فعالیت دارویی آن ترکیب می‌باشد.

---

[۱] - 3D –MORSE-signal 25 weighted by atomic masses

[۲] - 3D –MORSE-signal 29 weighted by atomic van der waals volumes

پ) توصیف کننده های Moran Autocorrelation

توصیف کننده های گروه Moran Autocorrelation توصیف کننده های دوبعدی هستند که از طریق رابطه زیر محاسبه می‌شوند:

$$I(d) = \frac{\frac{1}{2}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}\cdot(w_i - \overline{w})(w_j - \overline{w})}{\frac{1}{A}\cdot\sum_{i=1}^{A}(w_i - \overline{w})^2} \qquad (۳-۳)$$

که w یک ویژگی اتم، $\overline{w}$ میانگین مقدار آن روی مولکول، A تعداد اتم‌ها و d فاصله‌ی توپولوژیکی است و δ_{ij} نیز - که به تابع کرونکر[1] معروف است - در حالتی که $d_{ij} = d$ باشد یک است، در غیر اینصورت صفر است. مقدار این توصیف کننده که از جنس فاصله است، از ۱+ تا ۱- متغیر است [۱۹ و ۴۲]. MATS8m[2] از این گروه توصیف کننده هاست که در مدل برتر ظاهر شده است که با جرم اتمی وزن‌دار شده در مدل انتخاب شده و نشان می‌دهد که وزن کلی استخلاف، فاکتوری تعیین کننده برای فعالیت دارویی آن ترکیب می‌باشد. علامت مثبت این توصیف کننده بیانگر این است که که با افزایش مقادیر این توصیف کننده pEC50 نیز افزایش می‌یابد. جدول ۳-۱۷ چگونگی ارتباط این توصیف کننده با فعالیت دارویی را برای تعدادی از مولکول‌های سری داده‌ها نشان می‌دهد.

جدول ۳-۱۷: نمایش ارتباط فعالیت با مقدار توصیف کننده MATS8m برای بعضی از ترکیبات مورد مطالعه

شماره ترکیب	مقدار MATS8m	pEC_{50}
۱۱	-۰/۲۳۸	۵/۶۹
۲۴	-۰/۱۵۰	۵/۸۵
۳۶	۰/۰۲۰	۶/۰۹
۴۹	۰/۰۳۹	۶/۱۵

[1]- Kronecker

[2] - Moran autocorrelation-lag8 weighted by atomic masses

ت) توصیف کننده‌های گروه RDF[1]

می‌توان تابع توزیع شعاعی یک ترکیب شامل N اتم را به احتمال یافتن یک اتم در یک حجم کروی با شعاع r تعریف کرد و فرم کلی این تابع به صورت زیر است:

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{\beta(r-r_{ij})^2} \qquad (۴-۳)$$

که در این رابطه N تعداد اتم‌ها، A_i و A_j خاصیت اتمی (الکترونگاتیویته، جرم اتمی، قطبش پذیری اتمی و...)، r_{ij} شعاع بین اتم i وj، f فاکتور مقیاس، β فاکتور دما که باعث حرکت و جا به جایی اتم‌ها است، می‌باشند[۳۷]. از میان این توصیف کننده‌ها RDF110u در مدل برتر دیده می‌شود.

ث) توصیف کننده‌های Topoligical [۱۹]

این توصیف کننده بر اساس نمایش گراف مولکول می باشند. در این گراف ها هر نقطه نشان‌دهنده یک اتم بوده و خطوط بین نقاط نیز نشان‌دهنده‌ی پیوند شیمیایی بین اتم‌ها می‌باشد. معمولاً در گراف‌های مولکولی اتم هیدروژن را نشان نمی‌دهند. این توصیف کننده اطلاعاتی راجع به ساختمان، اندازه، شکل، تقارن، شاخه‌دار شدن، نحوه‌ی اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول در اختیار ما قرار می‌دهند. محاسبه‌ی این توصیف کننده به سادگی از روی ساختمان دوبعدی مولکول‌ها امکان‌پذیر می‌باشد. از این گروه، توصیف کننده‌های CIC5[2] و TF..Cl[3] انتخاب شده است.

ج) توصیف کننده‌های WHIM[4]

این شاخص از مختصات کارتزین ساختار سه‌بعدی مولکول، با استفاده از صورتبندی با حداقل انرژی

---

[1] - Radial Distribution Function

[2] - Complementary information content

[3] - Sum of topological distances between F..Cl

[4] -Weighted Holistic Invariant Molecular descriptors

محاسبه می‌شود و شامل اطلاعاتی درباره‌ی اندازه، شکل، تقارن و توزیع اتمی ساختار سه‌بعدی مولکول می‌باشد. این توصیف کننده از رابطه (۳-۵) به دست می‌آید:

$$S_{jk} = \frac{\sum_{i=1}^{A} w_i (q_{ij} - \overline{q_j})\,(q_{ik} - \overline{q_k})}{\sum_{i=1}^{A} w_i} \qquad (۳-۵)$$

که S_{jk} کوواریانس وزن‌دار شده بین کئوردینه j ام و k ام، A تعداد اتم‌ها، w_i وزن i امین اتم، q_{ij} و q_{ik} به ترتیب j امین و k امین کئوردینه‌های اتم i ام، و $\overline{q}$ مقدار میانگین متناظر است. شش طرح وزن‌دار شدن پیشنهاد شده است که عبارتند از:

۱ - حالت بدون وزن (u) ۲- وزن‌دار شده با جرم اتمی ۳- وزن‌دار شده با حجم واندروالس ۴- وزن‌دار شده با الکترونگاتیویته ساندرسون ۵- وزن‌دار شده با حالت الکتروتوپولوژیکی کی‌یر- هال[۱] ۶- وزن‌دار شده با قطبش‌پذیری[۱۹ و ۴۳]. از این دسته توصیف کننده G1p توسط مدل انتخاب گردید که به تأثیر صورتبندی دارو اشاره دارد.

۳-۹- بررسی میزان مشارکت توصیف‌کننده‌های منتخب در شبکه عصبی

میزان مشارکت توصیف‌کننده‌های منتخب توسط شبکه عصبی (SR-ANN) به صورت زیر تعیین شد:

۱- توصیف‌کننده مورد نظر به همراه اوزان مربوطه اش از شبکه بهینه شده حذف گردید.

۲- با استفاده از بقیه توصیف‌کننده‌ها مقدار متغیر وابسته (pEC_{50}) برای هر ترکیب سری ارزیابی پیش‌بینی شد.

۳- میانگین خطای مطلق ترکیبات سری ارزیابی محاسبه گردید.

۴- مراحل ۱تا ۳ برای دیگر توصیف‌کننده‌های منتخب نیز تکرار شد.

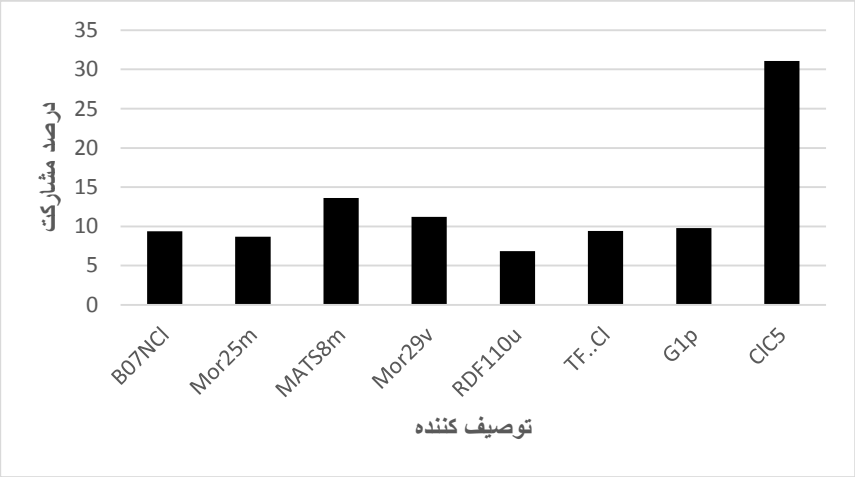
۵- سرانجام درصد مشارکت هر توصیف‌کننده توسط رابطه (۳-۶) برآورد شد [۴۴ و ۴۵].

---

[۱] - Kier-hall

$$c_i = 100 \frac{\Delta m_i}{\sum_{i=1}^{N} \Delta m_i} \qquad (۶-۳)$$

در این رابطه c_i درصد مشارکت توصیف‌کننده حذف شده iام، N تعداد توصیف‌کننده‌های مدل و Δm_i میانگین خطای مطلق سری ارزیابی در غیاب توصیف‌کننده‌iام را نشان می‌دهد. بر این اساس درصد مشارکت توصیف‌کننده‌های منتخب در ترکیبات مورد بررسی به صورت شکل ۳-۲۲ به دست آمد.



شکل ۳-۲۲: مشارکت توصیف‌کننده‌ها در شبکه عصبی بهینهSR-ANN

بر اساس نمودار فوق توصیف‌کننده‌های ClC5، MATS8m و Mor29v دارای بیشترین اثر مشارکت می‌باشند و از میان این سه، توصیف‌کننده ClC5 دارای بیشترین اثر مشارکت می‌باشد. این توصیف‌کننده در برگیرنده اطلاعات دو بعدی مولکول است. بنابراین می توان گفت که فعالیت ضد ایدز این ترکیبات وابستگی قابل ملاحظه‌ای به توپولوژی ترکیب ضد ایدز دارد. توصیف‌کننده دیگری که درصد مشارکت بالایی دارد، توصیف‌کننده MATS8m از دسته توصیف‌کننده دو بعدی است که وزن کلی استخلاف را در میزان فعالیت ضد ایدز بیان می‌کند. همچنین توصیف‌کننده Mor29v در برگیرنده اطلاعات سه بعدی مولکول است که این اطلاعات به صورتبندی دارو ارتباط دارد پس می‌توان گفت که فعالیت بازداری این ترکیبات وابستگی قابل ملاحظه‌ای به تغییرات صورتبندی دارو دارد.

۳-۱۰- نتیجه‌گیری نهایی

با توجه به این که به ندرت مطالعات استفاده از توصیف کننده‌های سهم گروه برای پیش‌بینی خاصیت دارویی استفاده شده است، در این گزارش از روش‌های جنگل‌های تصادفی و شبکه عصبی مصنوعی با استفاده از توصیف کننده‌های سهم گروه برای مدل‌سازی استفاده گردید. این مدل‌ها می‌توانند به محققان در طراحی ترکیبات جدید از این مشتقات با استخلاف‌های مختلف کمک زیادی نمایند. توصیف کننده‌های سهم گروه از روی ساختار ترکیب به دست می‌آیند بنابراین ساده و قابل تفسیر می‌باشند. محور اصلی این پژوهش نیز ارائه توصیف کننده‌های ساده، در عین حال مرتبط با فعالیت مورد نظر است. با توجه به نتایج به دست آمده می‌توان گفت که فعالیت ضد ایدز این ترکیبات وابستگی قابل قبولی به ساختمان دو بعدی ترکیب دارد. در نتیجه ترکیبی می‌تواند فعالیت ضد ایدز بهتری داشته باشد که مقدار سه توصیف کننده مهم ClC5، MATS8m و Mor29v برای آن زیاد باشد.

۴- فصل چهارم

مطالعه ارتباط کمی ساختار-فعالیت

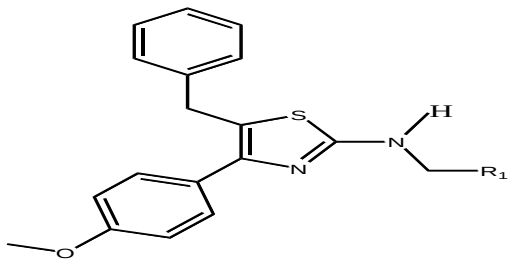
مشتقات تیازول

۴-۱- مدل‌سازی فعالیت دارویی مشتقات تیازول

هدف در این فصل، یافتن یک مدل برای پیش‌بینی فعالیت ضد HIV مشتقاتی از ترکیبات تیازول است. بخش تجربی شامل معرفی سری داده‌ها، بهینه سازی ساختار مولکول‌ها و محاسبه توصیف کننده‌های مولکولی با استفاده از روش رگرسیون مرحله‌ای و همچنین مدل‌سازی توسط روش‌های جنگل‌های تصادفی و شبکه عصبی مصنوعی و ارزیابی مدل‌های برتر است.

۴-۲- سری داده‌ها

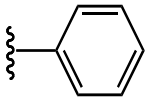
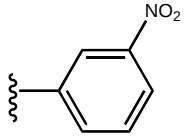
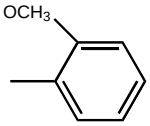
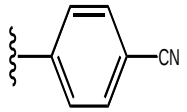
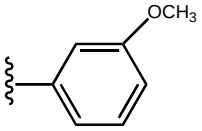
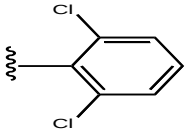
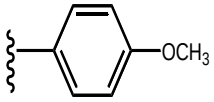
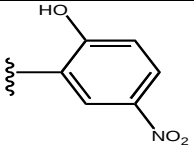
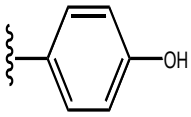
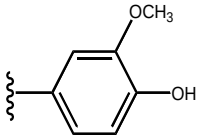
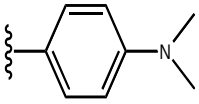
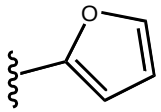
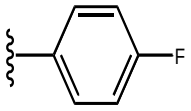
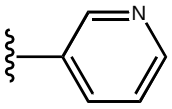
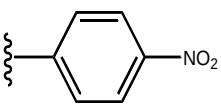
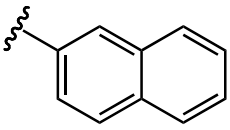
ترکیبات مورد استفاده در این تحقیق، داده‌های تجربی ۲۴ ترکیب از مشتقات تیازول می باشند که از مرجع [۴۶] گردآوری شده است. مقادیر منهای لگاریتم داده های IC_{50} (IC_{50})[1] کمترین دوزی از دارو که آنزیم ۵۰٪ بازداری داشته باشد) بر حسب مولار ($-\log IC_{50}\times 10^{-6}$) به عنوان متغیر وابسته، در این مدل‌سازی مورد استفاده قرار گرفته است. دو ساختار اصلی برای این ترکیبات در شکل ۴-۱ و شکل ۴-۲ نشان داده شده است. جدول ۴-۱ و جدول ۴-۲ ساختار استخلاف‌های آن و مقادیر تجربی pIC_{50} بر حسب مولار که در شرایط یکسان اندازه گیری شده است را نشان می‌دهد.

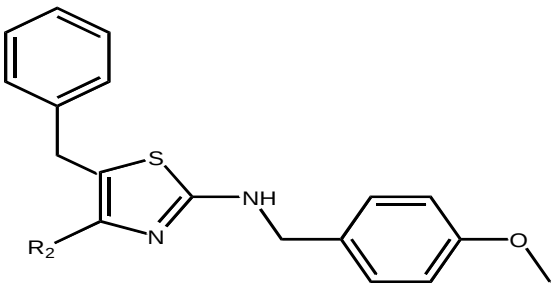


شکل ۴-۱: اسکلت اصلی شماره یک

[1]- describes the dose that causes haif-maximal inhibition by an antagonist

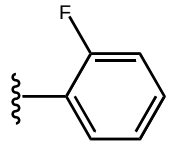
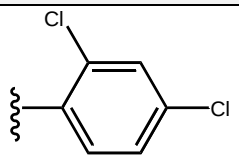
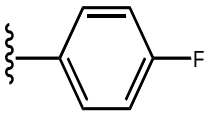
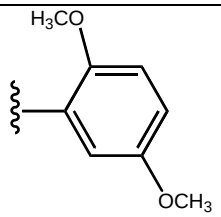
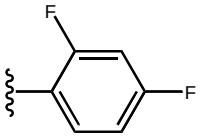
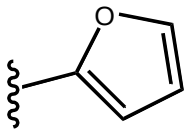
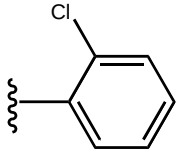
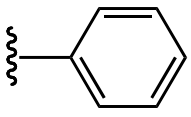
جدول۴-۱ : ساختار و مقادیر تجربی pIC_{50} مشتقات

شماره	R_1	pIC_{50}	شماره	R_1	pIC_{50}
۱		۶/۱۹	۹		۵/۸۳
۲		۵/۱۶	۱۰		۶/۳۵
۳		۶/۰۹	۱۱		۵/۱۴
۴		۶/۵۱	۱۲		۵/۲۸
۵		۶/۱۹	۱۳		۶/۴۸
۶		۶/۰۶	۱۴		۶/۰۰
۷		۶/۱۸	۱۵		۵/۳۳
۸		۶/۱۰	۱۶		۶/۰۲



شکل ۴-۲: اسکلت اصلی شماره دو

جدول ۴-۲: ساختار و مقادیر تجربی pIC_{50} مشتقات تیازول

شماره	R_2	pIC_{50}	شماره	R_2	pIC_{50}
۱۷		۷/۲۱	۲۱		۶/۸۵
۱۸		۵/۹۲	۲۲		۵/۲۱
۱۹		۷/۱۹	۲۳		۶/۱۵
۲۰		۷/۳۴	۲۴		۶/۷۴

۴-۳- بهینه‌سازی ساختمان هندسی مولکول‌ها و محاسبه توصیف‌کننده‌ها

در این مرحله، ساختار ترکیبات به وسیله نرم افراز Hyperchem رسم و برای بهینه کردن ساختار آن‌ها، روش AM_1 به کار برده شد. سپس بوسیله نرم افراز Dragon، ۳۲۲۴ توصیف‌کننده با استفاده از این ساختارهای بهینه شده محاسبه شد.

۴-۴- انتخاب بهترین توصیف‌کننده‌ها با روش رگرسیون مرحله‌ای(SR)

برای انتخاب توصیف کننده‌هایی که به طور معناداری فعالیت دارویی را به ساختار مولکول ارتباط دهند، مراحل زیر انجام گرفت:

الف- توصیف کننده‌هایی که دارای مقادیر یکسان برای همه ترکیبات بودند، از فهرست توصیف کننده‌هاحذف شدند.

ب- همبستگی میان توصیف کننده‌ها محاسبه گردید و از میان جفت توصیف کننده‌هایی که ضریب همبستگی بزرگتر از ۰/۹ داشتند، آنکه که همبستگی کمتری با متغیر وابسته داشت، حذف شد.

ج- روش رگرسیون خطی چندگانه مرحله‌ای بر اساس انتخاب رو به جلو و حذف رو به عقب جهت غربال کردن بهترین توصیف کننده‌های مورد نیاز با استفاده از نرم افراز.SPSS 18 انجام شد که در این مرحله ۱۰ توصیف کننده انتخاب شدند که بیشترین همبستگی با فعالیت ضد HIV ترکیبات مورد مطالعه را دارند. جدول ۴-۳ کل توصیف کننده‌های انتخاب شده و طبقه مربوط به آنها را نشان می‌دهد. همچنین ماتریس ضریب همبستگی این توصیف کننده‌ها که در جدول ۴-۴ ارائه شده، عدم همبستگی قابل ملاحظه بین توصیف کننده‌ها را نشان می‌دهد.

جدول۴-۳: توصیف کننده‌های انتخاب شده توسط رگرسیون مرحله‌ای

No	Symbol	Class	Meaning
۱	P2u	WHIM descriptors	2nd component shape directional WHIM index/unweighted
۲	F07SCl	2D frequency fingerprints	Frequency of S - Cl at topological distance 07
۳	MATS5m	2D autocorrelation	Moaran autocorrelation –lag5/Weighted by atomic masses
۴	GATS6p	2D autocorrelation	Geary autocorrelation –lag 6/ Weighted by atomic polarizabilities
۵	SIC3	Information indices	Structural information content (neighborhood symmetry of 3-order)
۶	Mor30v	3D-MoRSE descriptors	3D-MoRSE- signal 30/Weighted by atomic van der waals volumes
۷	RDF055v	RDF descriptors	Radial Distribution Function -5.5/Weighted by atomic van der waals volumes
۸	G2p	WHIM descriptors	2st component symmetry directional WHIM index/ Weighted by atomic polarizabilities
۹	EEig08d	edge adjacency indices	Eigenvalue 08 from edge adj matrix weighted by dipol moments
۱۰	qpmax	Charge descriptors	Maximum positive charge

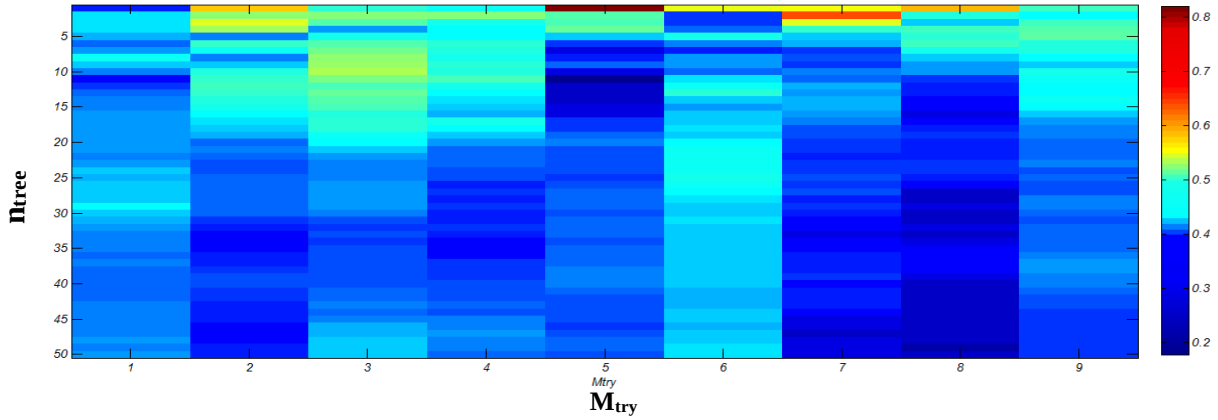
جدول ۴-۴: ماتریس همبستگی کل توصیف کننده‌های انتخاب شده توسط رگرسیون مرحله‌ای

	P2u	F07SC	MATS5m	GATS6p	SlC3	Mor30v	RDF055v	G2p	EEig08d	Qpmax
P2u	۱									
F07SCl	۰/۲۵۰	۱								
MATS5m	-۰/۴۱۱	۰/۴۴۱	۱							
GATS6p	-۰/۰۵۳	۰/۱۸۲	۰/۲۵۱	۱						
SlC3	۰/۱۶۶	۰/۲۳۷	۰/۱۸۳	-۰/۲۹۹	۱					
Mor30v	۰/۲۵۲	-۰/۱۶۸	۰/۰۰۹	-۰/۰۶۹	-۰/۰۶۹	۱				
RDF055v	-۰/۳۴۳	۰/۲۹۶	۰/۱۹۳	۰/۴۱۰	-۰/۲۰۹	-۰/۰۷۳	۱			
G2p	۰/۱۰۳	۰/۰۶۴	۰/۲۱۸	-۰/۲۷۸	-۰/۴۳۳	۰/۰۵۶	-۰/۱۸۰	۱		
EEig08d	-۰/۲۲۴	۰/۱۴۷	-۰/۰۴۹	-۰/۱۸۱	۰/۲۲۵	۰/۳۰۸	۰/۴۱۵	۰/۰۶۶	۱	
Qpmax	۰/۰۷۷	-۰/۰۸۷	-۰/۱۶۱	-۰/۳۴۹	۰/۳۴۸	-۰/۰۱۲	-۰/۱۸۱	۰/۱۴۲	۰/۵۷۴	۱

۴-۵- مدل‌سازی و بهینه‌سازی پارامترهای موثر بر روش جنگل‌های تصادفی

برای استفاده از قابلیت انتخاب توصیف‌کننده و انجام مدل‌سازی توسط جنگل‌های تصادفی، ابتدا باید پارامترهای مؤثر بهینه شوند. برای انجام این کار ابتدا ترکیبات به دو سری آموزش(۱۸ترکیب) و سری آزمون (۶ترکیب) تقسیم شدند. مقادیر مربوط به ۱۰ توصیف‌کننده به عنوان متغیر مستقل و مقادیر pIC_{50} به عنوان متغیر وابسته در نظر گرفته شدند. داده‌های سری OOB نیز برای بهینه‌سازی تعداد درختان (n_{tree})، تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز (M_{try}) و تعداد مشاهدات باقی‌مانده در هر گره (Node size) در نظر گرفته

شـدند. برای بهینه نمودن، تعداد درخت از ۱۰۰ تا ۵۰۰ با گام صد، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (M_{try}) از ۲ تا ۱۰ با گام یک و مقدار مشاهدات باقی‌مانده در هر گره (Node size) از ۱ تا ۱۰ تغییر داده شد و در هر مرحله مقدار خطای مربوط به مجموعه OOB محاسبه گردید. نتایج بهینه ســازی پارامترهای M_{try} و n_{tree} در Node size بهینه در شــکل ۴-۳ آمده اســت همانطور که مشاهده می‌شود نقاط آبی رنگ دارای خطای کمتری است. همچنین نتایج بهینه سازی در جدول ۴-۵ برای چندین حالت که کمترین MSE را دارد، آورده شده است.



شکل ۴-۳: بهینه سازی پارمترهای M_{try} و n_{tree} در Node size بهینه

جدول۴-۵: کمترین مقادیر MSE همراه با M_{try} و N_{tree} متناظر با آن‌ها

n_{tree}	M_{try}	Node Size	MSE OOB
۲۰۰	**۸**	**۱**	**۰/۰۳۹۸**
۱۰۰	۶	۲	۰/۰۴۳۴
۱۰۰	۱۰	۳	۰/۰۴۹۷
۲۰۰	۸	۴	۰/۰۵۲۸
۵۰۰	۱۰	۵	۰/۰۵۹۴
۲۰۰	۷	۶	۰/۰۵۸۸
۱۰۰	۱۰	۷	۰/۰۶۴۷
۱۰۰	۱۰	۸	۰/۰۷۳۵
۵۰۰	۱۰	۹	۰/۰۹۷۸
۳۰۰	۱۰	۱۰	۰/۱۲۷۱

با توجه به نتایج حاصل از جدول، تعداد درختان =۲۰۰ و تعداد توصیف کننده‌های انتخاب شده در هر مرحله افراز =۸ و تعداد گره =۱ به عنوان مقادیر بهینه انتخاب شدند.

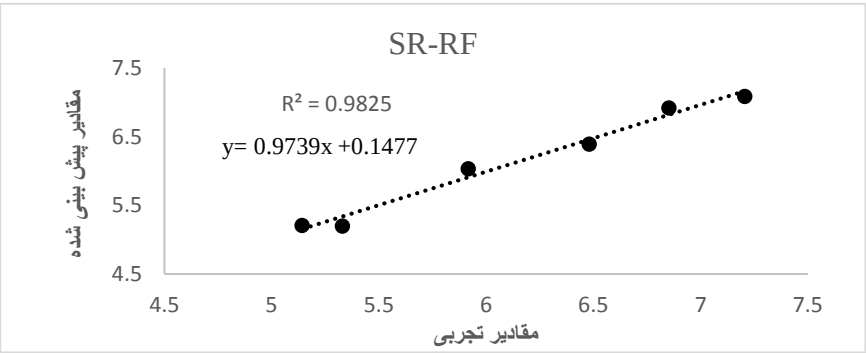
۴-۶- ارزیابی مدل جنگل‌های تصادفی

۴-۶-۱- ارزیابی مدل RF با استفاده از داده‌های سری آزمون

اهمیت مدل‌های بهینه وقتی مشخص می‌گردد که خواص مولکول‌هایی که در مدل‌سازی به کار نرفته‌اند را پیش بینی کند. بدین منظور مدل‌های منتخب، برای پیش‌بینی فعالیت داده‌های سری آزمون به کار برده شدند. جدول ۴-۶ نتایج پیش‌بینی حاصل از ارزیابی مدل جنگل‌های تصادفی با استفاده از داده‌های سری آزمون و شکل ۴-۴ نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد.

جدول ۴-۶: نتایج حاصل از ارزیابی مدل جنگل تصادفی به دست آمده با استفاده از داده‌های سری آزمون

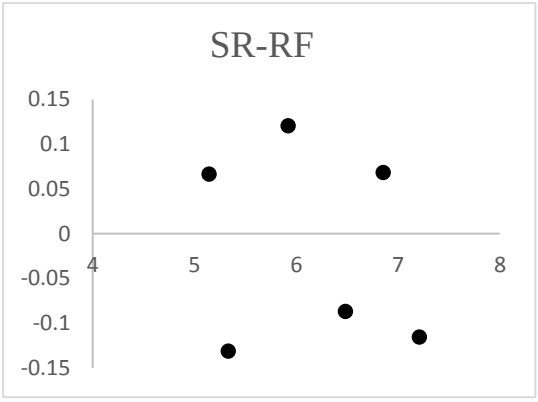
شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی	درصد خطا
۱۱	۵/۱۴	۵/۲۱	**۱/۳۶**
۱۳	۶/۴۸	۶/۳۹	**-۱/۳۸**
۱۵	۵/۳۳	۵/۲۰	**-۲/۴۳**
۱۷	۷/۲۱	۷/۰۹	**-۱/۶۶**
۱۸	۵/۹۲	۶/۰۴	**۲/۰۲**
۲۱	۶/۸۵	۶/۹۲	**۱/۰۲**



شکل ۴-۴: نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی به وسیله داده‌های سری آزمون مدل جنگل‌های تصادفی

۴-۶-۲- ارزیابی مدل RF با استفاده از نمودار خطای باقی‌مانده

اختلاف مقادیر پیش‌بینی‌شده و مقادیر تجربی، خطای باقیمانده نامیده می‌شود. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است. نمودار خطای باقیمانده برحسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل ۴-۵ نشان داده شده است.



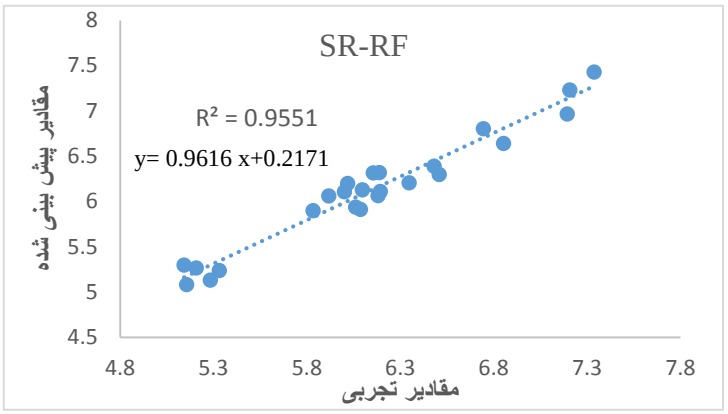
شکل ۴-۵: نمودار خطای باقیمانده برحسب مقادیر تجربی به وسیله داده‌های سری آزمون برای مدل جنگل‌های تصادفی

۴-۶-۳- ارزیابی مدل RF توسط روش رد مرحله‌ای تک‌تک

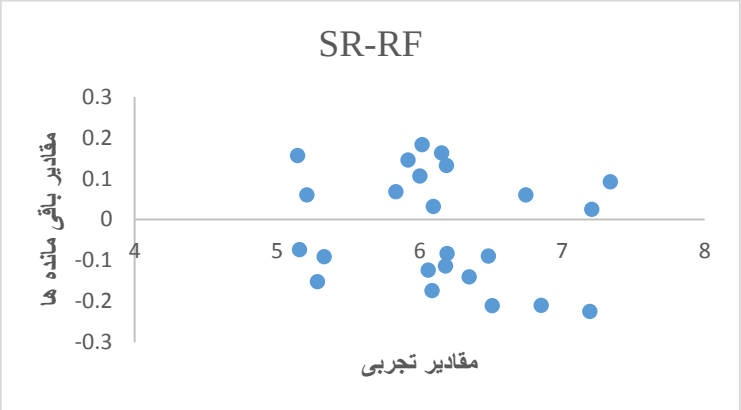
در این روش هر بار یک ترکیب به عنوان سری آزمون کنار گذاشته شد و در شرایط بهینه به دست آمده از بخش ۴-۵ با ترکیبات باقی‌مانده مدل‌سازی صورت گرفت. سپس مدل بدست آمده برای پیش‌بینی pIC50 ترکیب کنار گذاشته شده به کار گرفته شد و این فرآیند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از این روش در جدول ۴-۷ نشان شده است. ضریب‌های تعیین مشاهده شده در شکل ۴-۶ نشان می‌دهند که مدل‌ها توانایی خوبی برای پیش‌بینی فعالیت ضد ایدز مشتقات تیازول دارد. همچنین نتایج به دست آمده شکل ۴-۷ که در آن‌ها مقادیر باقی‌مانده‌ها بر حسب مقادیر تجربی pIC50 ترکیبات مورد بحث ترسیم شده است دارای تقارن پراکندگی نقاط در دو طرف محور افقی است که نشان دهنده‌ی عدم وجود خطای سیستماتیک برای بیشتر ترکیبات است.

جدول ۴-۷: نتایج حاصل از رد مرحله‌ای تک‌تک برای مدل جنگل‌های تصادفی

شماره ترکیب	مقدار تجربی (pIC_{50})	مقدار پیش‌بینی	درصد خطا
۱	۶/۱۹	۶/۳۲	**۲/۱۰**
۲	۵/۱۶	۵/۰۸	**-۱/۵۵**
۳	۶/۰۹	۵/۹۱	**-۲/۹۵**
۴	۶/۵۱	۶/۳۰	**-۳/۲۲**
۵	۶/۱۹	۶/۱۱	**-۱/۲۹**
۶	۶/۰۶	۵/۹۴	**-۲/۹۸**
۷	۶/۱۸	۶/۰۷	**-۱/۷۷**
۸	۶/۱۰	۶/۱۳	**۰./۴۹**
۹	۵/۸۳	۵/۹۰	**۱/۲۰**
۱۰	۶/۳۵	۶/۲۱	**-۲/۲۰**
۱۱	۵/۱۴	۵/۳۰	**۳/۱۱**
۱۲	۵/۲۸	۵/۱۳	**-۲/۸۴**
۱۳	۶/۴۸	۶/۳۹	**-۱/۳۸**
۱۴	۶/۰۰	۶/۱۱	**۱/۸۳**
۱۵	۵/۳۳	۵/۲۴	**-۱/۶۸**
۱۶	۶/۰۲	۶/۲۰	**۲/۹۹**
۱۷	۷/۲۱	۷/۲۳	**۰/۲۷**
۱۸	۵/۹۲	۶/۰۶	**۲/۳۶**
۱۹	۷/۱۹	۶/۹۷	**-۳/۰۵**
۲۰	۷/۳۴	۷/۴۳	**۱/۲۲**
۲۱	۶/۸۵	۶/۶۴	**-۳/۰۶**
۲۲	۵/۲۱	۵/۲۷	**۱/۱۵**
۲۳	۶/۱۵	۶/۳۲	**۲/۷۶**
۲۴	۶/۷۴	۶/۸۰	**۰/۸۹**



شکل ۴-۶: نمودار مقادیر پیش‌بینی شده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک بر حسب مقادیر تجربی برای مدل جنگل‌های تصادفی



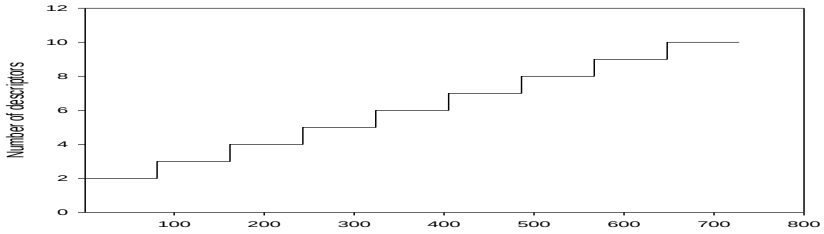
شکل ۴-۷: نمودار مقادیر خطای باقی‌مانده برای کل داده‌ها به روش رد مرحله‌ای تک‌تک برحسب مقادیر تجربی برای مدل جنگل‌های تصادفی

۴-۷- مدل‌سازی و بهینه‌سازی پارامترهای موثر بر شبکه عصبی مصنوعی

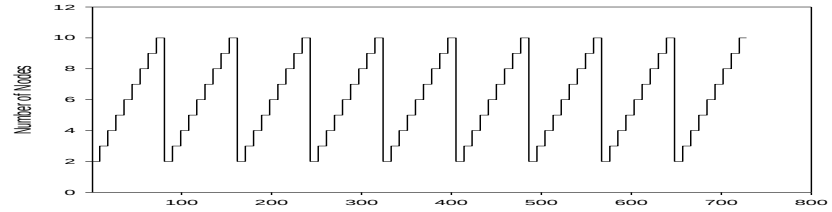
شبکه عصبی انتخاب شده در این مطالعه یک شبکه پیشخور با الگوریتم آموزشی پس انتشار می‌باشد. الگوریتم آن با برنامه متلب نوشته شده است. ابتدا مقادیر ۱۰ توصیف‌گر انتخاب شده با روش رگرسیون مرحله‌ای به عنوان متغیرهای ورودی و فعالیت ضدHIV متناظر آن‌ها به عنوان متغیر هدف در نظر گرفته شد تا پاسخ شبکه با آن‌ها سنجیده شود. برای به دست آوردن بهترین مدل و کم‌ترین خطا، پارامترهای مؤثر (تابع انتقال، تابع آموزش، تعداد متغیرهای ورودی شبکه، تعداد گره‌ها در لایه‌ی مخفی و تعداد دورهای آموزش) به‌طور همزمان بهینه‌سازی شدند.

در فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به به مجموعه آزمون (۶ ترکیب) و مجموعه آموزش (۱۸ ترکیب) تقسیم شدند. سری آموزش برای محاسبه گرادیان و به روز کردن اوزان و بایاس‌های شبکه با روش ارزیابی تقاطعی استفاده شد. از آنجایی که در بیشتر موارد به نظر می‌رسد یک لایه پنهان مناسب باشد [۴۰] در این تحقیق نیز از یک لایه پنهان استفاده شده است. برای یافتن مقدار بهینه پارامترهای مؤثر، مقادیر مختلفی از آنها در ساختار شبکه قرار داده شد. برای این منظور هر شبکه با تعداد ورودی ۲ تا ۱۰ و با دو الگوریتم آموزشی لونبرگ-مارکوات (trainlm) و تنظیم بایزین (trainbr) و دو تابع انتقال لگاریتم سیگموئید و تانژانت سیگموئید و تعداد گره از ۲ تا ۱۰ و تعداد دورهای آموزش از ۱۰ تا ۵۰ با گام ۵، به طور همزمان آموزش داده شد.

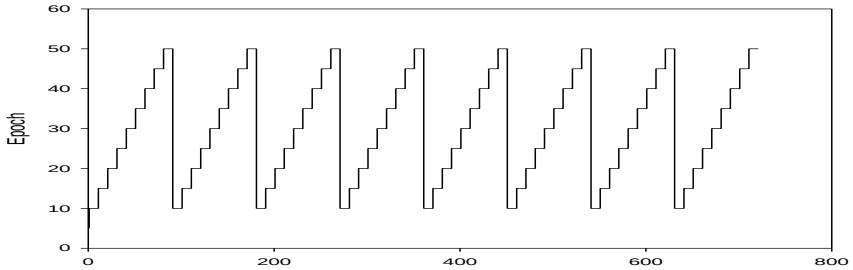
در روند بهینه‌سازی پارامترهای فوق، به حداقل رساندن میانگین مربع خطاهای حاصل از ارزیابی تقاطعی سری آموزش به عنوان معیار انتخاب شد. بخشی از روند تغییرات پارامترهای شبکه در حین بهینه‌سازی همزمان پارامترها به همراه مقادیر MSE بدست آمده، به صورت نموداری برحسب یک بردار مرجع فرضی در شکل ۴-۸ تا شکل ۴-۱۱ آمده است. نتایج بهینه‌سازی این پارامترها در جدول۴-۸، گردآوری شده است.



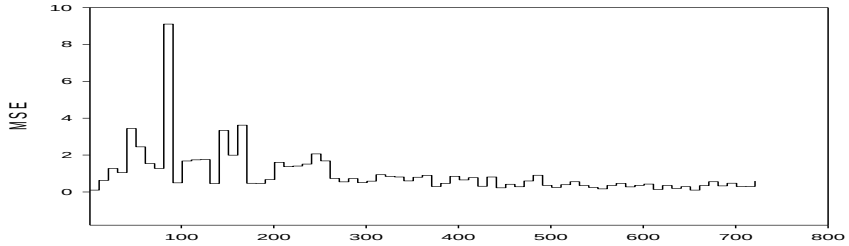
الف)



ب)



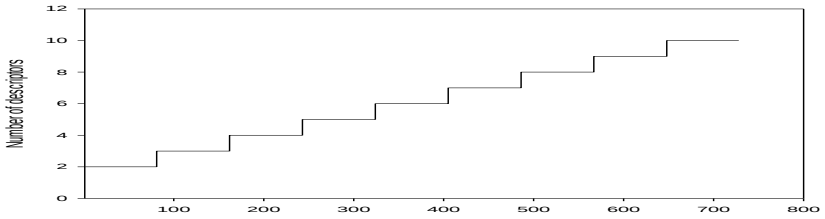
ج)



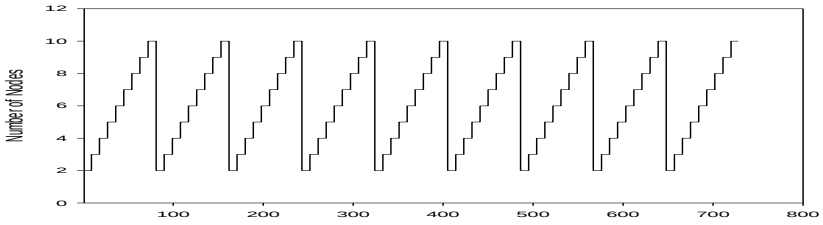
د)

بردار مرجع

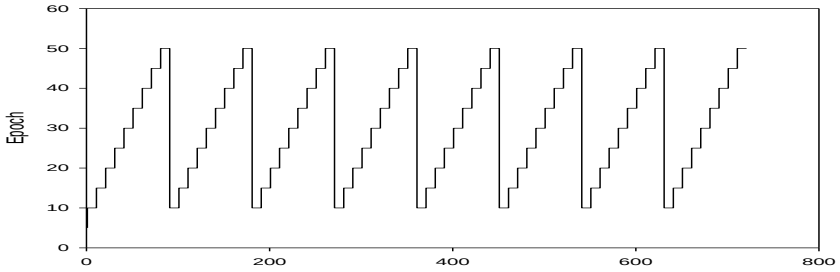
شکل ۴-۸: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال لگاریتم سیگموئید



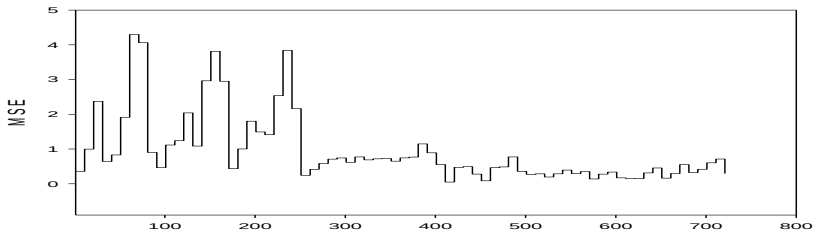
الف)



ب)



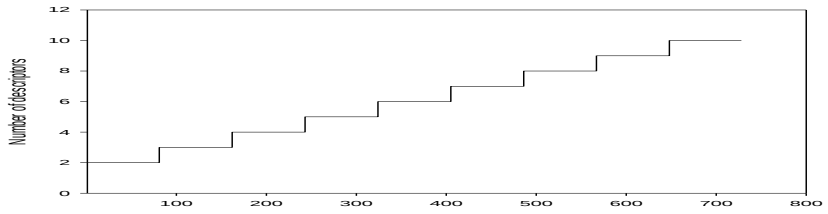
ج)



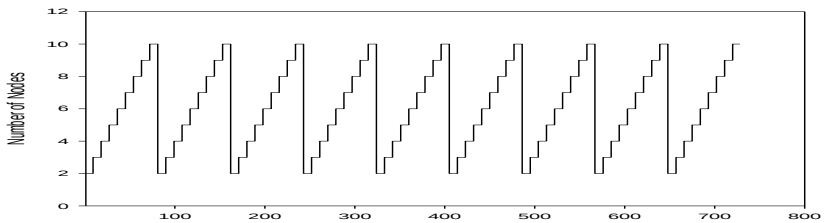
د)

بردار مرجع

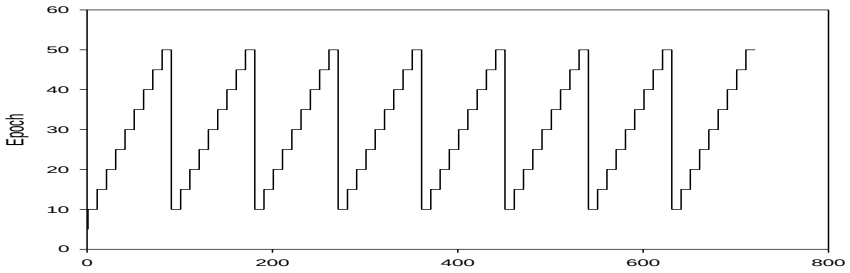
شکل ۴-۹: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش لونبرگ-مارکوات به همراه تابع انتقال تانژانت سیگموئید



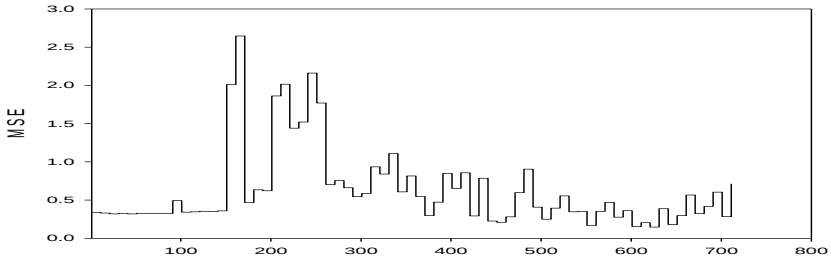
الف)



ب)

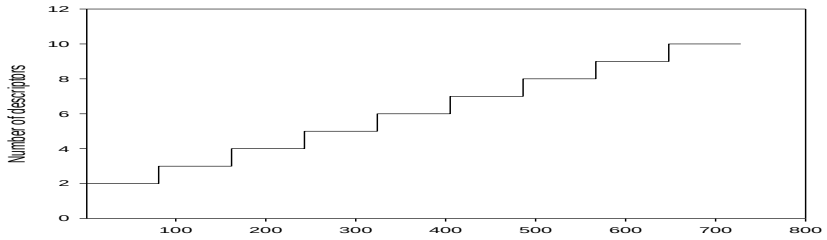


ج)

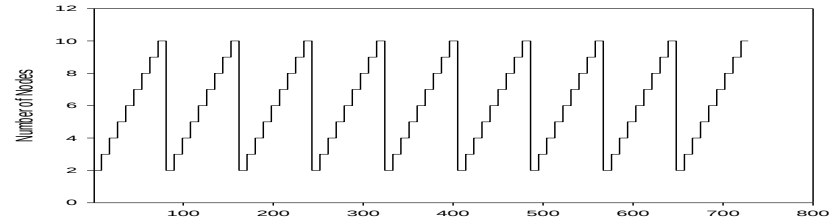


بردار مرجع د)

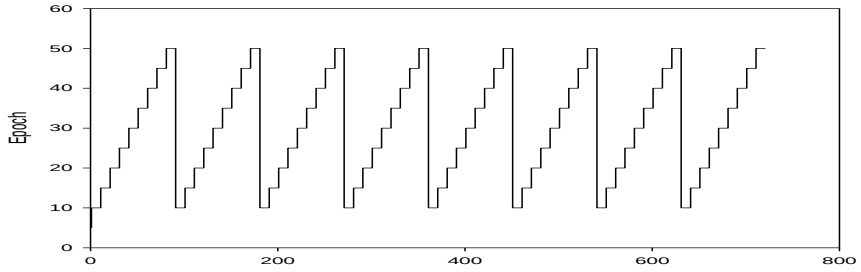
شکل ۴-۱۰: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال لگاریتم سیگموئید



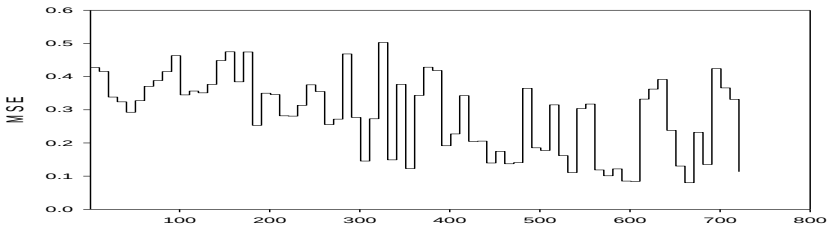
الف)



ب)



ج)



د)

بردار مرجع

شکل ۴-۱۱: نمودارهای الف)تعداد توصیف‌کننده، ب) تعداد گره‌ها، ج) تعداد دورهای آموزش و د) مقادیر MSE برای تابع آموزش تنظیم بایزین به همراه تابع انتقال تانزانت سیگموئید

جدول۴-۸: توابع و پارامترهای شبکه‌های بهینه SR-ANN

MSE	تعداد دور آموزش	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف کننده
۰/۰۹۹۵	۱۰	۲	لگاریتم-سیگموئید	لونبرگ- مارکوات	۱۰
۰/۰۴۶۴	**۴۰**	**۲**	**تانژانت-سیگموئید**	**لونبرگ- مارکوات**	**۷**
۰/۰۸۳۶	۴۰	۴	لگاریتم-سیگموئید	تنظیم بایزین	۹
۰/۰۶۹۰	۱۵	۱۰	تانژانت-سیگموئید	تنظیم بایزین	۱۰

با توجه به نتایج بدست آمده، الگوریتم آموزشی لونبرگ –مارکوات و تابع انتقال تانژانت سیگموئید، ۷ توصیف کننده ورودی و ۲ نرون در لایه مخفی با تعداد دور آموزشی ۴۰ کمترین MSE را نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی در نظر گرفته شد.

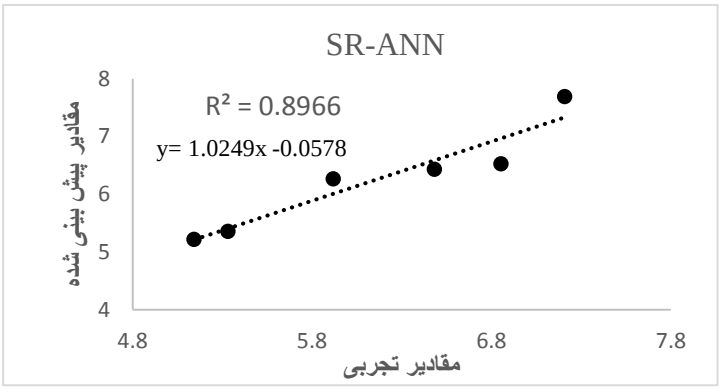
۴-۸- ارزیابی مدل شبکه عصبی مصنوعی

۴-۸-۱- ارزیابی مدل ANN با استفاده از داده‌های سری آزمون

جدول۴-۹ نتایج پیش‌بینی حاصل از ارزیابی مدل‌های شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون و شکل ۴-۱۲ نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی را برای داده‌های سری آزمون نشان می‌دهد.

جدول۴-۹: نتایج حاصل از ارزیابی مدل شبکه عصبی مصنوعی با استفاده از داده‌های سری آزمون

شماره ترکیب	مقدار تجربی (pIC_{50})	مقدار پیش‌بینی	درصد خطا
۱۱	۵/۱۴	۵/۲۲	**۱/۵۵**
۱۳	۶/۴۸	۶/۴۴	**-۰/۶۱**
۱۵	۵/۳۳	۵/۳۶	**۰/۵۶**
۱۷	۷/۲۱	۷/۶۹	**۶/۶۵**
۱۸	۵/۹۲	۶/۲۷	**۵/۹۱**
۲۱	۶/۸۵	۶/۵۳	**-۴/۶۷**



شکل ۴-۱۲: نمودار تغییرات مقادیر پیش‌بینی شده در مقابل مقادیر تجربی به وسیله داده‌های سری آزمون برای مدل شبکه عصبی مصنوعی

۴-۸-۲- ارزیابی مدل ANN با استفاده از نمودار خطای باقی‌مانده

اختلاف مقادیر پیش‌بینی‌شده و مقادیر تجربی، خطای باقیمانده نامیده می‌شود. توزیع متقارن داده‌ها حول محور افقی (خطای صفر) حاکی از عدم وجود خطای سیستماتیک است. نمودار خطای باقیمانده برحسب مقادیر تجربی، برای مدل‌های ذکر شده در شکل ۴-۱۳ نشان داده شده است.

شکل ۴-۱۳: نمودار خطای باقیمانده برحسب مقادیر تجربی برای داده‌های سری آزمون برای مدل شبکه عصبی مصنوعی

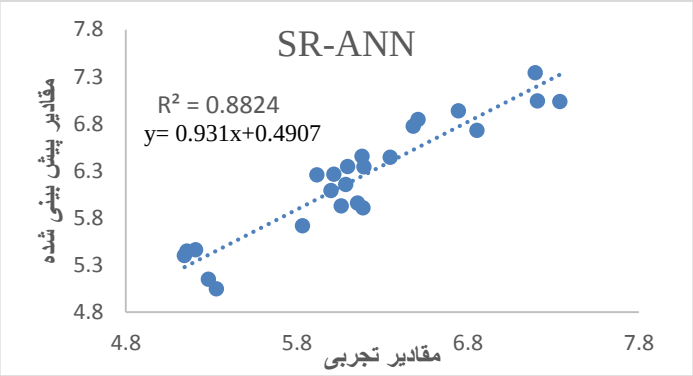
۴-۸-۳- ارزیابی مدل ANN توسط روش رد مرحله‌ای تک‌تک

در این روش هر بار یک ترکیب به عنوان سری آزمون کنار گذاشته شد و در شرایط بهینه به دست آمده از بخش ۷-۴ با ترکیبات باقی‌مانده مدل‌سازی صورت گرفت. سپس مدل بدست آمده برای

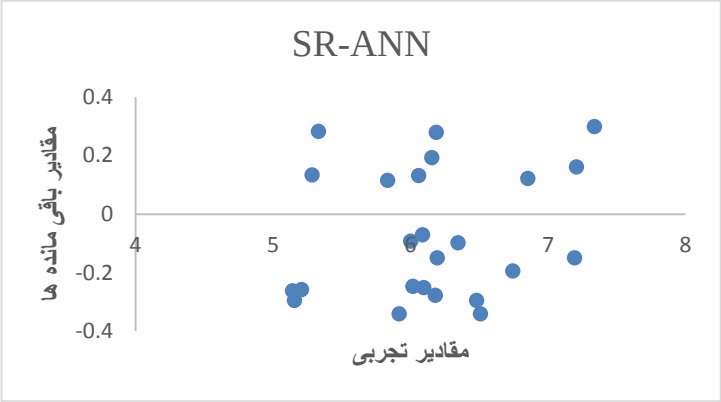
پیش‌بینی pIC_{50} ترکیب کنار گذاشته شده به کار گرفته شد و این فرآیند برای تمام اعضای سری داده‌ها تکرار شد. نتایج حاصل از این روش در جدول ۴-۱۰ نشان شده است. ضریب تعیین مشاهده در شکل ۴-۱۴ نشان می‌دهد که مدل توانایی خوبی برای پیش‌بینی فعالیت ضد ایدز مشتقات تیازول دارد. همچنین نتایج به دست آمده شکل ۴-۱۵ که در آن مقادیر باقی‌مانده‌ها بر حسب مقادیر تجربی pIC_{50} ترکیبات مورد بحث ترسیم شده است دارای تقارن پراکندگی نقاط در دو طرف محور افقی است که نشان دهنده‌ی عدم وجود خطای سیستماتیک برای بیشتر ترکیبات است.

جدول۴-۱۰: نتایج حاصل از رد مرحله‌ای تک‌تک برای مدل شبکه عصبی

شماره ترکیب	مقدار تجربی(pIC_{50})	مقدار پیش‌بینی	درصد خطا
۱	۶/۱۹	۵/۹۱	**-۴/۵۲**
۲	۵/۱۶	۵/۴۵	**۵/۶۲**
۳	۶/۰۹	۶/۱۶	**۱/۱۴**
۴	۶/۵۱	۶/۸۵	**۵/۲۲**
۵	۶/۱۹	۶/۳۴	**۲/۴۲**
۶	۶/۰۶	۵/۹۳	**-۲/۱۴**
۷	۶/۱۸	۶/۴۶	**۴/۵۳**
۸	۶/۱۰	۶/۳۵	**۴/۰۹**
۹	۵/۸۳	۵/۷۲	**-۱/۸۸**
۱۰	۶/۳۵	۶/۴۴	**۱/۴۱**
۱۱	۵/۱۴	۵/۴۰	**۵/۰۵**
۱۲	۵/۲۸	۵/۱۵	**-۲/۴۶**
۱۳	۶/۴۸	۶/۷۸	**۴/۶۲**
۱۴	۶/۰۰	۶/۰۹	**۱/۵۰**
۱۵	۵/۳۳	۵/۰۵	**-۵/۲۵**
۱۶	۶/۰۲	۶/۲۶	**۳/۹۸**
۱۷	۷/۲۱	۷/۰۵	**-۲/۲۱**
۱۸	۵/۹۲	۶/۲۶	**۵/۷۴**
۱۹	۷/۱۹	۷/۳۴	**۲/۰۸**
۲۰	۷/۳۴	۷/۰۴	**-۴/۰۸**
۲۱	۶/۸۵	۶/۷۳	**-۱/۷۵**
۲۲	۵/۲۱	۵/۴۶	**۴/۷۹**
۲۳	۶/۱۵	۵/۹۶	**-۳/۰۸**
۲۴	۶/۷۴	۶/۹۴	**۲/۹۶**



شکل ۴-۱۴: نمودار مقادیر پیش‌بینی شده برای کل داده‌ها به وسیله روش رد مرحله‌ای تک‌تک بر حسب مقادیر تجربی برای مدل شبکه عصبی مصنوعی



شکل ۴-۱۵: نمودار مقادیر خطای باقی‌مانده برای کل داده‌ها به وسیله روش رد مرحله‌ای تک‌تک برحسب مقادیر تجربی مدل شبکه عصبی مصنوعی

۴-۸-۴- مقایسه مدل‌ها با استفاده از پارامترهای آماری

همان‌طور که قبلاً بیان شد، پارامترهای آماری مقادیر تعیین‌کننده‌ای برای ارزیابی مدل‌ها هستند. مطابق جدول ۴-۱۱ در این مرحله از این مقادیر برای ارزیابی مدل‌های ساخته شده استفاده گردید. با توجه به نتایج حاصل از این جدول می‌توان گفت که مدل جنگل‌های تصادفی ارائه شده نسبت به شبکه عصبی مصنوعی برای مشتقات تیازول دارای برتری است.

جدول ۴-۱۱: پارامترهای آماری برای مدل‌های طراحی شده برتر توسط RF وANN

پارامتر	سری تست (N=10)		کل داده ها (N=24)	
	RF	ANN	RF	ANN
R	۰/۹۹۱	۰/۹۴۷	۰/۹۷۷	۰/۹۳۹
R^2	۰/۹۸۲	۰/۸۶۲	۰/۹۵۴	۰/۸۶۸۲
PRESS	۰/۰۶۲	۰/۴۷۶	۰/۴۲۶	۱/۲۲۱۳
MSE	۰/۰۱۰	۰/۰۷۹۴	۰/۰۱۸	۰/۰۵۱
MAE	۰/۰۹۸	۰/۲۱۹۰	۰/۱۲۱	۰/۲۱۰
MRE	۱/۶۲۳	۳/۳۶۲	۱/۹۸۲	۳/۴۷۴
SEP	۰/۲۴۸	۰/۲۸۲	۰/۱۳۳۱	۰/۲۲۶
REP(%)	۴/۰۴۶	۴/۵۰۸	۲/۱۷۴	۳/۶۳۱

۴-۹- بررسی ارتباط بین توصیف کننده‌های منتخب و خاصیت موردنظر

در این بخش به بررسی ارتباط بین توصیف‌گرهای منتخب در مدل‌سازی جنگل‌های تصادفی که مدل برتر است و در بخش ۴-۷ محاسبه شدند، می‌پردازیم و سعی می‌کنیم تا حد امکان توضیحی برای ارتباط بین این توصیف‌گرها و فعالیت مورد نظر بیابیم. جدول ۴-۱۲ مقادیر اثر متوسط[۱] (که روش محاسبه آن در بخش (۳-۸) بیان شد) توصیف‌کننده‌های مدل برتر یعنی SR-RF را نشان می‌دهد.

جدول ۴-۱۲: اثر متوسط توصیف‌کننده‌های به دست آمده توسط روش رگرسیون مرحله ای

Variable	P2u	F07SCl	MATS5m	GATS6p	SlC3	Mor30v	RDF055v	G2p
Mean effect	-۳/۵۸۰	-۷/۹۹۹	۱۵/۱۶۹	-۱۰/۹۴۰	-۷/۰۴۳	۶/۵۴۳	-۵/۳۴۹	-۴/۹۰۶

الف) توصیف کننده‌های WHIM[۲]

همان‌طور که در بخش ج-۸-۳ اشاره شد، این شاخص از مختصات کارتزین ساختار سه‌بعدی مولکول، با استفاده از صورتبندی با حداقل انرژی محاسبه می‌شود و شامل اطلاعاتی درباره‌ی اندازه، شکل، تقارن و توزیع اتمی ساختار سه‌بعدی مولکول می‌باشد. از این نوع توصیفگرها G2p, P2u توسط مدل انتخاب گردید که به تأثیر صورتبندی دارو اشاره دارد.

ب) توصیف کننده‌های 2D frequency fingerprints

همان‌طور که در بخش الف-۸-۳ در مورد توصیف کننده‌های اثر انگشت اشاره شد، این توصیف کننده‌ها اندازه مولکول‌ها را بررسی می‌کند. توصیف کننده اثر انگشت فرکانس که به بررسی مجموع

۱- Mean effect

۲ -Weighted Holistic Invariant Molecular descriptors

فرکانس دو اتم در فاصله هندسی آن اتم‌ها می‌پردازد، اندازه اتم‌ها را به‌طور مستقیم بررسی می‌کند. که از میان توصیف کننده‌های اثر انگشت F07SCl در مدل برتر دیده می‌شود که فرکانس S-Cl را در فاصله هندسی ۷ نشان می‌دهد [۴۷].

پ) توصیف کننده MATS5m

همان‌طور که در بخش پ-۸-۳ اشاره شد، توصیف کننده‌های گروه Moran Autocorrelation توصیف کننده‌های دوبعدی هستند MATS5m[۱] از این گروه توصیف کننده‌ها ست که در مدل برتر ظاهر شده است .

ت) توصیف کننده GATS6p

جزء (2D-autocorrelation) توصیف کننده‌های دوبعدی خودارتباطی است که از گراف مولکولی و از طریق محاسبه مجموع اوزان اتم‌های انتهایی کل مسیرها با طول مسیر مورد نظر (lag)، به دست می‌آیند. نوعی از این توصیف کننده‌ها گروه Geary Autocorrelation است که ضریب گری نام دارد و بدین صورت محاسبه می‌شود:

$$C(d) = \frac{\frac{1}{2\Delta}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}.(w_i - w_j)^2}{\frac{1}{A-1}.\sum_{i=1}^{A}(w_i - \overline{w})^2} \qquad (۱-۴)$$

که w یک ویژگی اتم، $\overline{w}$ میانگین مقدار آن ویژگی روی مولکول، A تعداد اتم‌ها و d فاصله‌ی توپولوژیکی است و δ_{ij} نیز - که به تابع کرونکر[۲] معروف است - در حالتی که $d_{ij} = d$ باشد یک است، و در غیر اینصورت صفر است و Δ هم مجموعه‌ی δ هاست. مقدار این توصیف‌کننده که از جنس فاصله

---

[۱] - Moran autocorrelation-lag15 weighted by atomic polarizabilities

[۲] - Kronecker

اســـت، از صـــفر تا بی‌نهایت متغیر اســـت[۱۹] از میان این توصـــیف کننده‌ها GATS6p و نقش قطبش‌پذیری اتمی ترکیبات را در جهت‌گیری به سوی مولکول هدف نشان می دهد. علامت منفی این توصـیف کننده بیانگر این است که مقادیر بالاتر از این توصیف کننده برای فعالیت ضد HIV مطلوب نیســـت. جـدول ۴-۱۳ چگونگی ارتبـاط این توصـــیف کننـده بـا فعالیت دارویی را برای تعدادی از مولکول‌های سری داده‌ها نشان می‌دهد.

جدول ۴-۱۳: نمایش ارتباط فعالیت با مقدار GATS6p

شماره ترکیب	مقدار GATS6p	pIC_{50}
۱۱	۱/۳۱۰	**۵/۱۴**
۱۵	۱/۱۶۶	**۵/۳۳**
۱۸	۰/۸۶۸	**۵/۹۲**

ث)توصیف کننده Topoligical

همان‌طور که در بخش ث-۸-۳ بیان شد این توصیف کننده بر اساس نمایش گراف مولکول می باشند. در این گراف‌ها هر نقطه نشان‌دهنده یک اتم بوده و خطوط بین نقاط نیز نشان‌دهنده‌ی پیوند شیمیایی بین اتم‌ها می‌باشـد. معمولاً در گراف‌های مولکولی اتم هیدروژن را نشان نمی‌دهند. این توصیف کننده اطلاعاتی راجع به سـاختمان، اندازه، شـکل، تقارن، شاخه‌دار شدن، نحوه‌ی اتصال اتم‌ها و نوع اتم‌های موجود در یک مولکول در اختیار ما قرار می‌دهند. محاسـبه‌ی این توصـیف کننده به سـادگی از روی سـاختمان دوبعدی مولکول‌ها امکان‌پذیر می‌باشد. . از میان این توصیف کننده‌ها SIC3[۱] در مدل برتر دیده می‌شود.

[۱] Structural information content

د) توصیف کننده‌های 3D – MORSE[1]

همان‌طور که در بخش ب-۸-۳ بیان شد توصیف کننده‌های 3D - Morse ساختار سه بعدی مولکول را براساس تفرق الکترون نشان می‌دهند. از بین این توصیف کننده‌ها Mor30v[2] در مدل برتر دیده می‌شود.

ه-توصیف کننده‌های RDF[3]

همان‌طور که در بخش ت-۸-۳ بیان شد می‌توان تابع توزیع شعاعی یک ترکیب شامل N اتم را به احتمال یافتن یک اتم در یک حجم کروی با شعاع r تعریف کرد [۴۸]. اهمیت این توصیف کننده‌ها به دلیل اختلاف در توزیع اتم‌ها در مولکول‌ها و اثر این توزیع بر روی فعالیت مهارکنندگی ترکیبات می‌باشد. از میان این توصیف کننده‌ها RDF055v در مدل برتر دیده می‌شود.

---

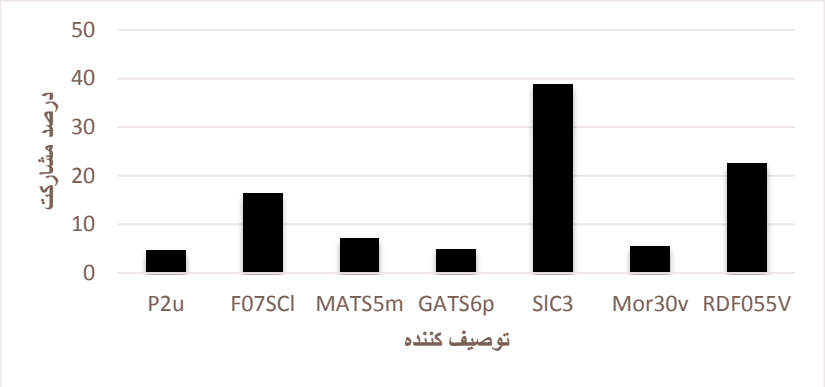
[1] - 3D- Molecular Representation of Structure based on Electron diffraction

[2] - 3D –MORSE-signal 30 volumes

[3] - Radial Distribution Function

۴-۱۰- بررسی میزان مشارکت توصیف کننده‌های منتخب در شبکه عصبی

بر اساس روشی که در بخش (۳-۹) توضیح داده شد، میزان مشارکت توصیف‌کننده‌های منتخب در پیش‌بینی فعالیت دارویی موردنظر برآورد شد که نتایج آن در شکل ۴-۱۶ ارائه شده است.



شکل ۴-۱۶: مشارکت توصیف کننده‌ها در شبکه عصبی بهینه SR-ANN

بر اساس نمودار فوق توصیف‌کننده‌های SCl3، RDF055v و F07SCl دارای بیشترین اثر مشارکت می‌باشند. طبق شکل فوق توصیف‌کننده SCl3 دارای بیشترین اثر مشارکت می‌باشد. این توصیف‌کننده در برگیرنده اطلاعات دو بعدی مولکول است. توصیف‌کننده دیگری که درصد مشارکت بالایی دارد، توصیف‌کننده RDF055v که اهمیت این توصیف کننده به دلیل اختلاف در توزیع اتم ها در مولکول‌ها و اثر این توزیع بر روی فعالیت مهارکنندگی ترکیبات می‌باشد. همچنین توصیف‌کننده F07SCl بیان کننده اندازه مولکول به‌طور مستقیم است و اثر انگشت فرکانس را در فاصله هندسی بیان می‌کند.

۴-۱۱- نتیجه‌گیری نهایی

در این تحقیق نتایج نشـــان می‌دهند که مدل‌های غیرخطی توســـعه یافته می‌توانند به عنوان روش‌های موفق برای مدل‌سازی و پیش‌بینی فعالیت ضد ایدز ترکیبات مورد مطالعه به کار روند و این مـدل‌ها می‌توانند به محققان در طراحی داروهایی با فعالیت ضـــد ایدز قویتر کمک کنند. بنابراین بر اسـاس این مدل‌ها می‌توان فعالیت ضـدHIV را برای همین طبقه از ترکیبات که هنوز سـنتز نشده‌اند پیش بینی کرد. در واقع این روش به طراحان دارو این امکان را می‌دهد که قبل از اینکه بخواهند یک ترکیب دارویی را سنتز کنند، ابتدا با استفاده از مدل معتبر خاصیت دارویی آن را پیش‌بینی کنند و در صـورتی که دارای فعالیت دارویی مناسـبی باشـند، ترکیب سنتز شود و بدین ترتیب با صرف هزینه و وقت کمتر در مسیر تهیه داروهایی با اثر بخشی بهتر گام بردارند.

پیشنهادات

- ✓ روش جنگل‌های تصـــادفی را می‌توان برای پیش‌بینی فعالیت ســـایر ترکیبات دارویی استفاده نمود.
- ✓ پژوهشگران می‌توانند با الگوگیری از روش سهم گروه مورد استفاده در این تحقیق، به بررسی روابط ساختار - فعالیت در گروه‌های دیگر مولکولی بپردازند.
- ✓ می‌توان از روش غیر خطی دیگر مانند SVM[1] استفاده کرد.
- ✓ این تحقیقات را می‌توان با پژوهش روی خواص دارویی ترکیبات جدید ادامه داد تا نهایتاً منجر به ساخت داروهای جدیدتر با اثربخشی بهتر گردد.

[1] - Supported Vector Machine

فهرست منابع

[1]-http: //www.soton.ac.uk/~ceb/teaching/2005/206-8.htm.

[2] -موسـوی جاهد،" ویروس شـناسـی پزشکی"، انتشارات دانشگاه علوم پزشکی و پیراپزشکی شهید بهشتی، ۱۳۷۸.

[3]-Debyser Z., Pauwels R., Andries K., Desmyter J., Kukla M., Janssen P. A., & An antiviral target on reverse transcriptase of human "De Clercq E. (1991), immunodeficiency virus type 1 revealed by tetrahydroimidazo-[4, 5, 1-jk][1, 4] , Proceedings of the National "benzodiazepin-2 (1H)-one and-thione derivatives Academy of Sciences, 88(4), 1451-1455.

[4]- Pillai A. D., Rani S., Rathod P. D., Xavier F. P., Vasu K. K., Padh H., & QSAR studies on some thiophene analogs as anti-"Sudarsanam V. (2005), inflammatory agents: enhancement of activity by electronic parameters and its , Bioorganic & medicinal chemistry, 13(4), "utilization for chemical lead optimization 1275-1283.

[5]- Jalali-Heravi M.; Parastar F., (2000),"Use of artificial neural network in a QSAR study of anti-HIV activity for a large group of HEPT derivatives", J. Chem. Inf. Comput. Sci., 40, pp 147-154.

[6]- Douali L., Villemin D., Cherqaoui D.; (2004), "Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by neural networks: TIBO derivatives", Int. Mol. Sci., 5, pp 48-55.

[7] -Arab Chamjangali M., (2009),"Modeling of cytotoxicity data (CC_{50}) of anti-HIV 1-[5-chlorophenyl) sulfonyl]-1H-pyrrole derivatives using calculated molecular descriptors and Levenberg- Marquardt artificial neural network", J.Chem. Bio. Drug. Des., 73, pp 456-465.

[8]-Patel A., Gupta L., Karthikeyan C., Trivedi P.(2010), "QSAR studies on dihydro-alkoxy-benzyl-oxopyrimidines(DABOs) derivatives, as new series of potent, broad-spectrum non-nucleoside reverse transcriptase inhibitors" J. Current Pharmaceutical Research , 1, pp 19-25

[9]- Chamjangali M. A., & Ashrafi M., (2013)," QSAR study of necroptosis inhibitory activities (EC50) of [1, 2, 3] thiadiazole and thiophene derivatives using Bayesian regularized artificial neural network and calculated descriptors", Medicinal Chemistry Research, 22(1), 392-400.

[10] عجم ز، (۱۳۹۲)، پایان نامه کارشناسی ارشد،" پیش بینی فعالیت ضد ایدز مشتقات غیر نوکلئوزیدی تیوکربومات به روش جنگل‌های تصادفی"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[11]- Arab Chamjangali M., (2009),"Modeling of cytotoxicity data (CC_{50}) of anti-HIV 1-[5-chlorophenyl) sulfonyl]-1H-pyrrole derivatives using calculated molecular descriptors and Levenberg- Marquardt artificial neural network", J.Chem. Bio. Drug. Des., 73, pp 456-465.

[12]- Jalali-Heravi M., Asadollahi-Baboli, (2008), "QSAR analysis of platelet-derived growth inhibitors using GA-ANN and shuffling cross validation", J. QSAR. Comb. Sci., 6, pp 750-757

[13]- http://www.wordiq.com/definition/chemometrics

[14] - عرب چم جنگلی م.، (۱۳۸۶) " پیش‌بینی فعالیت دارویی ضد ایدز (سیتوتوکسیتی) مشتقات ۵-فنیل-۱-فنیل آمینو۱-H- ایمیدازول به وسیله شبکه عصبی مصنوعی"، دانشگاه صنعتی شاهرود، گزارش طرح پژوهشی.

[15]- Arab Chamjangali M., Beglari M., Bagherian G., (2007), "Prediction of cytotoxicity data (CC_{50}) of anti-HIV 5-phenyl-1-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg-Marquardt algorithm", J. Mol. Graph. Model, 26, pp 360–367.

[16]- بهرام م، خیامیان ت، عبدالهی ح، فاطمی م.ح، قاسمی ج، کمپانی زارع م، ناصری ع، همتی نژاد ب، (۱۳۹۴)، " مقدمه‌ای بر کمومتریکس"، چاپ اول، نشر پژوهشی نوآوران شریف،۲۳۶، ص۱۷۴-۱۷۲.

[17]- Jurs P. C. (2005) "Assessing the reliability of a QSAR model's predictions", J.Mol Graph Model, 23(6), pp 503.

[18]-شعیری ف، (۱۳۸۵)، پایان نامه کارشناسی ارشد،"پیشگویی فعالیت پیرازولوکوئینولین‌ها با استفاده از شبکه عصبی موجک و کاربرد روشRAFA در اندازه‌گیری همزمان آمپی سیلین و آموکسی سیلین "، دانشگاه علوم پایه دامغان.

[19]- Todeschini R. and Consonni V. (2000), "Handbook of Molecular Descriptors", Jonh Wiley, New York, pp98.

[20]- Daniel W. W., & Wayne W. D. (1995). "Biostatistics: a foundation for

John WILEY.analysis in the health sciences".

[21]- بهرام م، خیامیان ت، عبدالهی ح، فاطمی م.ح، قاسمی ج، کمپانی زارع م، ناصری ع، همتی نژاد ب، (۱۳۹۴)، "

مقدمه‌ای بر کمومتریکس"، چاپ اول، نشر پژوهشی نوآوران شریف،۲۳۶، ص۱۷۷-۱۷۶.

[22]- محمدرضایی م، (۱۳۸۹)، پایان نامه کارشناسی ارشد، "مطالعه کمی ساختار- فعالیت برخی از ترکیبات سولفونانیلید به عنوان گروهی جدید از داروهای ضد سرطان و ضد HIV برخی از ترکیبات"، دانشکده شیمی، دانشگاه صنعتی شاهرود.

[23] - کشتکار ه، (۱۳۹۳)، پایان نامه کارشناسی ارشد، " مطالعه کمی ساختار- خاصیت (QSPR) دانسیته مایعات

یونی به صورت تابعی از دما و فشار با استفاده از روش‌های غیر خطی "، دانشکده شیمی، دانشگاه صنعتی شاهرود.

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

[25] – htpp://www.iranhiv.com/treatment.

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

[27] -منهاج م، (۱۳۸۷)، " مبانی شبکه‌های عصبی (هوش محاسباتی)"، جلد اول، چاپ پنجم، مرکز نشر دانشگاه صنعتی امیرکبیر، تهران.

[28] -بن کوپین،داورپناه،میرزایی ر، (۱۳۸۵)،" هوش مصنوعی"، انتشارات سخن گستر ، مشهد.

[29] -منهاج م، سیفی پور ن، (۱۳۷۷)،"کاربرد هوش محاسباتی در کنترل"، مرکز نشر دانشگاه صنعتی امیر کبیر.

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

[31] -فردوسی م.ع.،(۱۳۸۹)، پایان نامه کارشناسی ارشد، "پیش‌بینی ثابت‌های هنری بعضی از ترکیبات آلی با استفاده از روش‌های خطی و غیرخطی QSPR"، دانشکده شیمی، ، دانشگاه صنعتی شاهرود.

[32]- Breiman L, (2001), “ Random forests”, Mach Learn, 45,pp 5-32

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

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

[35]- Genuer R, Poggi J M, Tuleau-Malot Ch, (2010), “ V ariable selection using random forests”, Pattern Recogn Lett, 31, pp 2225-2236.

[36]- Seo, B. I., Uchil, V. R., Okello, M., Mishra, S., Ma, X. H., Nishonov, M., ... & Nair, V. (2011). Discovery of a potent HIV integrase inhibitor that leads to a prodrug with significant anti-HIV activity. ACS medicinal chemistry letters, 2(12), 877-881.

[37]- HyperChem7.0 Toronto, Canada: HyperCube Inc, http:// www.hyper.com

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

[39]- MATLAB 2009, the Math Work, Inc., Natick, MA, USA.

[40]- Zupan J., Gasteiger J., (1993), "Neural networks for chemists an introduction", VCH Publishers, Weinheim, Germany.

[41]- Ewing T., Baber J. C., & Feher M. (2006)," Novel 2D fingerprints for ligand-based virtual screening", Journal of chemical information and modeling, 46(6), 2423-2431.

[42]- P. Broto, G. Moreau, C. Vandycke, (1984), "Molecular structures: pp 66– Perception, autocorrelation descriptor and SAR studies", Eur. J. Med. Chem., 19, 70.

[43]- Todeschini R., Gromatica P., (1997), "3D- modeling and prediction by WHIM descriptors. Part5. Theory development and chemical meaning of WHIM descriptors", J. Quant. Struct., Acta., Relat., 16, pp 113-119.

[44]- Douali L., Villemin D., Cherqaoui D.; (2003), "Neural networks: Accurate nonlinear QSAR model for HEPT derivatives", J. Chem. Inf. Comput. Sci., 43, pp 1200-1207.

[45]- chen Z., Zhang Y., Fu W., (2010),"QSAR study of carboxylic acid derivatives as HIV-1integrase inhibitors", Eur. J. Med. Chem., 45, pp 3970-3980.

[46]- Xu, Z., Ba, M., Zhou, H., Cao, Y., Tang, C., Yang, Y., ... & Zhu, L. (2014). 2, 4, 5-Trisubstituted thiazole derivatives: a novel and potent class of non-nucleoside inhibitors of wild type and mutant HIV-1 reverse transcriptase. European journal of medicinal chemistry, 85, 27-42.

[47]- Crisan L., Iliescu S., & Funar-Timofei S.,(2016),"Structure-flammability relationship study of phosphoester dimers by MLR and PLS", Polímeros, (AHEAD), 0-0.

[48]- Spallarossa A., Cesarini S., Ranise A., et al.,(2009), Parallel synthesis," molecular modelling and further structure–activity relationship studies of new acylthiocarbamates as potent non-nucleoside HIV-1 reverse transcriptase inhibitors". European Journal of Medicinal Chemistry, 44(5): p. 2190-2201.

Abstract

In the first section of this study, quantitative structure – activity relationship (QSAR) models were constructed for predicting the anti-HIV activity of some pyridinone derivatives. The group contribution and the stepwise multiple linear regression methods were used for variable selection. Then the best descriptors were applied as input for QSAR modeling using random forest (RF) and artificial neural network (ANN) techniques. In modelling methods based on random forest the data set was divided into training set and test set including 43 and 10 compounds, respectively. Effective parameters of RF method including N_{tree} and Mtry were optimized using the training set. For ANN modeling the data set was divided into training set, validation set and test set including 37, 8 and 8 compounds, respectively. ANN model was optimized too and the best model was selected. The validation study of the RF and ANN models was performed by prediction of the activity of the test set compounds and the correlation coefficients for RF and ANN model are 0.929 and 0.935 respectively.

In the second section, RF and ANN models were constructed and the ability of prediction of anti-HIV activities for some of thiazole derivatives were investigated. The data set was divided into training and test set containing 18 and 6 chemical, respectively. The stepwaise regression method was used to select the most important descriptors. The validation study of the RF and ANN models was performed using test set compounds. The correlation coefficients obtained for test set were 0.982 and 0.896 for RF and ANN respectively. The results obtained showed proper prediction power of the proposed models.

Keyword: QSAR, Anti-HIV, Artificial neural network, Random forest, group contribution

Shahrood University of Technology

Faculty of Chemistry

MSc thesis in Analytical Chemistry

Quantitative structure-activity relationship study of the activity of some pharmaceutical compounds using Random forest

By: Atiyeh Asali

Supervisers:

Dr. Nasser Goudarzi

Dr. Mansour Arab Chamjangali

September 2016