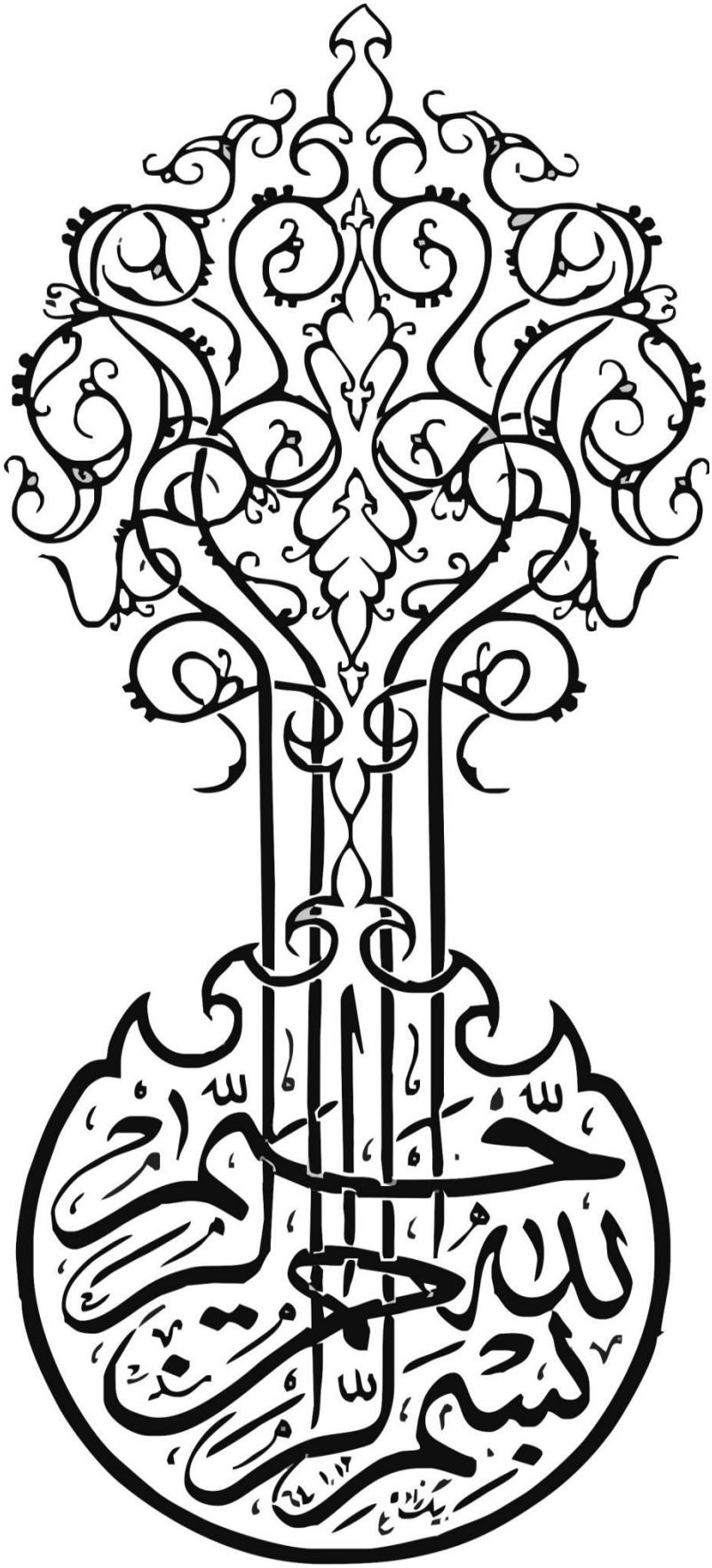

بسم الله الرحمن الرحيم

دانشکده: شیمی

گروه شیمی تجزیه

پایان‌نامه کارشناسی ارشد

پیش بینی فعالیت ضد ایدز مشتقات غیرنوکلئوزیدی تیوکربومات به روش جنگل‌های تصادفی

زهرا عجم

استاد راهنما:

دکتر منصور عرب چم‌جنگلی

استاد مشاور:

دکتر مجید سلامی

بهمن ماه ۱۳۹۳

پیوست۲

اگر شایسته تقدیم باشد:

تقدیم به وجود مقدس پدر و مادر عزیزم،

که استوار ایستادن تا من بایستم

و توانستند تا من بتوانم

و ماندن تا من در کنارشان تکیه زده

و استوار چون ایشان بمانم.

تقدیم به همسر مهربانم،

که وجودش بی شک بزرگترین افتخار زندگیم می باشد، او که اسوه صبر و تحمل بوده و مشکلات مسیر را برایم تسهیل نمود.

و تمامی عزیزانی که کلمه ای مرا آموختند.

ضمن تشکر و قدردانی از:

استاد راهنمای گرانقدرم:

جناب آقای دکتر منصور عرب چم جنگلی،

که آنچه هم اکنون پیش رو می بینم حاصل همفکری عالمانه، زحمات بی دریغ و دلسوزی های ایشان است،

استاد مشاور ارجمندم:

جناب آقای دکتر مجید سلامی،

همکاری های ایشان را قابل تقدیر و ستایش دانسته،

داوران عالیقدر:

جناب آقای دکتر قدمعلی باقریان دهمقی و جناب آقای دکتر ناصر کودرزی،

مساعدت و لطف ایشان را سپاسگزار بوده،

همسر عزیزم:

که مشکلات زندگی را برایم تسهیل نمود و پیمودن این راه بدون همراهی او ممکن نمی شد،

خانواده خوبم و خانواده محترم همسرم:

که همواره از حمایت و لطف آن ها بهره مند بوده ام،

از درگاه لایزال احدیت، توفیق روز افزون آن بزرگواران را مسئلت داشته

و همواره سعادت و بهروزیشان را آرزومندم.

تعهد نامه

اینجانب **زهرا عجم** دانشجوی دوره کارشناسی ارشد رشته شیمی تجزیه دانشکده شیمی دانشگاه شاهرود نویسنده پایان نامه **پیش بینی فعالیت ضد ایدز مشتقات غیرنوکلئوزیدی تیوکربومات به روش جنگل¬های تصادفی** تحت راهنمائی دکتر منصور عرب چم‌جنگلی 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

بازدارنده‌های غیرنوکلئوزیدی نسخه‌بردار معکوس، یک طبقه جدید از داروهای ضد رتروویروسی هستند که برای درمان عفونت ویروس نقص ایمنی اکتسابی(HIV) استفاده می‌شوند. این بازدارنده‌ها از فعالیت آنزیم نسخه‌بردار معکوس که تکثیر ژنوم HIV را درکنترل دارد، جلوگیری می‌کنند. در بخش اول پروژه، مدل QSAR برای پیش‌بینی فعالیت ضد HIV مشتقات تیوکربامات به عنوان بازدارنده‌های غیرنوکلئوزیدی با استفاده از توصیف‌گرهای سهم گروه و الگوریتم جنگل‌های تصادفی (RF) ساخته شد. سری داده‌ها به طور تصادفی به ۱۵۷ ترکیب آموزش و ۴۰ ترکیب آزمون تقسیم شد و پارامترهای مؤثر بر قدرت جنگل‌های تصادفی با استفاده از سری آموزش بهینه گردید. ارزیابی مدل به دست آمده، با استفاده از سری آزمون و رد مرحله‌ای تک‌تک (LOO) انجام گرفت. نتایج نشان داد که مجذور ضریب همبستگی برای پیش‌بینی خاصیت ضد HIV ترکیبات سری آزمون با استفاده از روش جنگل-های تصادفی برابر ۰/۸۷۴۸ است. در بخش دوم تحقیق ارتباط کمی ساختار – فعالیت بازدارنده‌های نکروپتوز (نکروستاتین-۵) مورد مطالعه قرار گرفتند (نکروپتوز نوعی مرگ سلول می‌باشد و نکروستاتین‌ها ترکیباتی با وزن مولکولی پایین هستند که به عنوان مهارکننده‌های نکروپتوز شناخته شده‌اند). داده‌های مورد بررسی شامل ۱۶۰ ترکیب می‌باشد که ۵۱ ترکیب فعال و بقیه غیرفعال هستند. ابتدا با استفاده از مدل جنگل‌های تصادفی مدلی طراحی شد تا بتواند این ترکیبات را به دو گروه فعال و غیرفعال دسته‌بندی کند. ارزیابی سری آزمون نشان می‌دهد مدل ارائه شده قادر است ۸۰٪ نمونه‌ها را به درستی طبقه‌بندی کند. در مرحله بعد پیش‌بینی خاصیت دارویی ترکیبات فعال با استفاده از شبکه عصبی مصنوعی و توصیف‌گرهای انتخاب شده توسط روش جنگل‌های تصادفی-رگرسیون مرحله‌ای (RF-SR) و روش رگرسیون مرحله ای (SR) انجام شد. مجذور ضریب همبستگی برای سری آزمون و روش رد مرحله‌ای تک‌تک به ترتیب برابر ۰/۹۶۷۷ و ۰/۸۲۹۵ می‌باشد.

کلید واژه: رابطه کمی ساختار-فعالیت- جنگل‌های تصادفی- شبکه عصبی مصنوعی- سهم گروه

نتایج حاصل از این پایان‌نامه در دو پوستر تحت عناوین:

1- Application of random forest regression in the modeling of the Anti-HIV activity of thiocarbamate derivatives, as potent HIV-1 reverse transcriptase inhibitors, using group contribution descriptors

2- Classification Study of necroptosis inhibitor, Necrostatin-5, based on random forest

در هفدهمین کنگره شیمی رفسنجان پذیرفته شد.

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل اول

مقدمه

۱-۱- رتروویروس

رتروویروس‌ها[1]، گروهی از ویروس‌ها هستند که ژنوم[2] آن‌ها ازRNA تشکیل شده‌است. بنابراین برای تکثیر خود به آنزیم نسخه‌بردار معکوس[3] وابسته هستند که ژنوم RNA آن‌ها را به DNA نسخه‌برداری کند تا بعد رتروویروس بتواند آن را به کمک آنزیم اینتگراز[4] وارد ژنوم میزبان کند و به این ترتیب امکان تکثیر ویروس ایجاد می‌گردد. کلمه رترو (معکوس) در این ویروس‌ها به این علت است که معمولاً نسخه‌برداری از DNA به RNA انجام می‌شود در حالیکه در این ویروس، عکس این عمل صورت گرفته و نسخه‌برداری از RNA انجام می‌گیرد. یکی از زیر خانواده‌های رتروویروس‌ها، لنتی‌ویرینه[5] است که شامل عواملی هستند که قادرند عفونت‌های مزمن با تخریب آهسته و پیش‌رونده عصبی ایجاد کنند. ویروس نقص ایمنی اکتسابی[6] (HIV) از دسته رتروویروس‌های لنتی‌ویرینه است.

HIV با آلوده کردن گروهی از سلول‌های دستگاه ایمنی به نام لنفوسیتT از نوع CD_4^+ بیماری ایدز را به وجود می‌آورد. این سلول‌ها زیر گروهی از گلبول‌های سفید هستند که به طور طبیعی پاسخ ایمنی به عفونت را تنظیم می‌کنند. پروتئین سازنده پوششHIV به آسانی وارد خون شده و مانند چسب سلول‌های TCD_4^+ را به هم می‌چسباند از طرف دیگر دستگاه ایمنی بدن به این سلول‌ها حمله کرده و باعث از بین رفتن TCD_4^+ می‌شوند و زمانی که میزان سلول‌های TCD_4^+ در فرد آلوده به حد معینی سقوط کند باعث ابتلا فرد به طیفی از عفونت‌ها می‌شود [۱].

1 - Retrovirus
2 - Genome
3 - Reverse Transcriptase
4 - Integrase
5 - Lentivirinae
6 - Human Immunodeficiency Virus

۱-۱-۱- درمان ضد رتروویروس HIV

درمان ضد رتروویروس، نوع اصلی درمان ایدز می‌باشد که این درمان شامل داروهائی است که از طریق مهار توانایی تکثیر ویروس، عمر افراد مبتلا را به طور قابل ملاحظه‌ای افزایش می‌دهند. این داروها سرعت نابودی CD_4^+ را کاهش می‌دهند. با توجه به اینکه این ویروس از سه آنزیم نسخه‌بردار معکوس، پروتئاز و اینتگراز برای تکثیر خود و آلوده کردن سلول‌ها استفاده می‌کند، بنابراین داروهای سنتز شده بیشتر این نوع آنزیم‌ها را هدف قرار می‌دهند.

۱-۱-۲-داروهای ضد رتروویروس

بازدارنده‌های ورودی[۱]

در سطح بیرونی ویروس HIV گلیکوپروتئین‌های gp41, gp120 وجود دارد که به وسیله آن‌ها ویروس به گیرنده‌های CD_4^+ موجود در سطح لنفوسیت‌های T سلول میزبان متصل شده و وارد سلول می‌گردد. بازدارنده‌های ورودی به گلیکوپروتئین‌های موجود در سطح ویروس می‌چسبند و از پیوستن ویروس به سلولهای CD_4^+ جلوگیری می‌کنند. بدین ترتیب با بلوکه کردن یکی از این پروتئین‌ها، تکثیر آهسته می‌شود. تاکنون یک نمونه از آن به نام فیوژن[۲] وارد بازار شده است که از طریق اتصال به پروتئین gp41، خاصیت بازدارندگی خود را اعمال می‌کند [۲].

بازدارنده‌های نوکلئوزیدی آنزیم نسخه‌بردار معکوس[۳]

مهارکننده‌های نوکلئوزیدی به زنجیره نوکلئوزیدی DNA که توسط آنزیم نسخه‌بردار معکوس درحال تولید است، می‌چسبد و مانع از اتصال نوکلئوزیدهای بیشتر به DNA می‌شوند به این ترتیب یک DNA ناقص تولید می‌شود [۳].

1 - Entry inhibitors
2 - Fusion
3 - Nucleoside Reverse transcriptase Inhibitors

مهارکننده‌های غیرنوکلئوزیدی نسخه‌بردار معکوس[1]

این نوع مهارکننده‌ها نیز فرآیند رونویسی را مختل می‌کنند. آن‌ها با چسباندن خود به آنزیم نسخه‌بردار معکوس و محدود کردن تحرک آنزیم، مانع از نسخه‌برداری می‌شوند [۳]. مهمترین ویژگی این نوع مهارکننده‌ها اختصاصی بودن آن‌هاست. نویراپین[2] نخستین گروه از این مهارکننده‌ها بود که در سال ۱۹۹۶ توسط سازمان غذا و دارو مورد تأیید قرار گرفت.

بازدارنده‌های نوع پروتئاز[3]

این داروها اولین بار در سال ۱۹۹۵ تأیید شدند. آنزیم پروتئاز در HIV به زنجیره طویل آنزیم‌ها و پروتئین‌ها، در سلول حمله کرده و آنها را به قطعات کوچکتری تقسیم می‌کند تا از این طریق پروتئین‌های لازم برای تولید یک ویروس بالغ تولید شود. مهارکننده پروتئاز به مکان‌هایی از آنزیم که بریدن[4] زنجیره پلی‌پپتیدی از آن مکان رخ می‌دهد، می‌چسبد و مانع از تولید پروتئین‌های لازم برای تولید ویروس جدید می‌شود. به عبارتی، مهارکننده پروتئاز قادر است فرآیند تبدیل ویروس نابالغ غیرعفونی را به ویروس بالغ عفونی، آهسته کند [۴].

۱-۱-۳- مشتقات تیوکربامات[5]

همانطور که گفته شد مهارکننده‌های غیر نوکلئوزیدی با هدف قرار دادن آنزیم نسخه‌بردار معکوس، در درمان عفونت HIV مؤثر هستند. مشتقات تیوکربامات به عنوان دسته جدیدی از مهارکننده های غیرنوکلئوزیدی شناخته شده‌اند. چهار گروه از این ترکیبات عبارتند از مشتقات:

---

[1] - Non Nucleoside Reverse transcriptase Inhibitors
[2] - Nevirapin
[3] -Protease inhibitors(PI)
[4] - Cut
[5]- Thiocarbamate(TC) derivatives

O- (۲- فتالامید و اتیل)- N- آریل- N- آسیل تیوکربامات[۱]

O- (۲- فنیل اتیل)- N- فنیل تیوکربامات[۲]

O- (۲- (۲-پیریدیل) اتیل)- N- فنیل تیوکربامات[۳]

O- (۲- فتالامید و اتیل)- N- آریل تیوکربامات[۴]

مطالعات داکینگ[۵] نشان داده‌است که این مهارکننده‌ها دارای صورت‌بندی پروانه‌ای[۶] و فاقد پیوند هیدروژنی درونی می‌باشند. این مزیت باعث شده مشتقات تیوکربامات از آزادی صورت‌بندی بیشتری برخوردار باشند که از این ویژگی می‌توان برای طراحی دارو با مقاومت کمتر استفاده کرد [۵].

از آن جایی که بیشتر داروهای شیمیایی توسط آزمون و خطا سنتز می‌شوند که بسیار وقت‌گیر و هزینه‌بر است، توسعه روش‌های نظری[۷] به عنوان یک ابزار کارآمد برای رفع این محدودیت‌ها مورد توجه قرار گرفته است. هدف پژوهش حاضر، یافتن رابطه کمی بین ساختار مشتقات تیوکربامات و فعالیت بازدارندگی آن‌ها است.

۱-۱-۴- مروری بر کارهای انجام شده

ریکیک[۸] و همکارانش در سال ۲۰۰۷ ارتباط کمی ساختار- فعالیت[۹] (QSAR) ۹۵ ترکیب را که به عنوان بازدارنده آنزیم نسخه‌بردار معکوس عمل می‌کنند، با استفاده از شبکه عصبی مصنوعی بررسی کردند و قابلیت پیش‌بینی مدل ساخته شده که شامل ۵ توصیف‌گر

1- O-(2-Phthalimidoethyl)-N-aryl-N-acylthiocarbamates
2- O-(2-Phenylethyl)-N-phenylthiocarbamates
3- O-[2-(2-pyridyl)ethyl]-N-phenylthiocarbamate
4- O-(2-Phthalimidoethyl)-N-arylthiocarbamates
5- Docking
6- Butterfly-like conformation
7- Theoretical approaches
8- Rakik
9- Quantitative Structure- Activity Relationship

بود، توسط رد مرحله‌ای تک‌تک[1]مورد ارزیابی قرار دادند. همچنین میزان مشارکت[2] توصیف‌گرهای وارد شده در مدل را نیز مورد تجزیه و تحلیل قرار دادند و نشان دادند که توصیف‌گرهایی که به خاصیت آب‌گریزی ترکیبات مربوط می‌شوند، دارای بیشترین اثر مشارکت هستند [۶].

در سال ۲۰۰۹ زارعی و عتباتی فعالیت ضد اچ- آی- وی ۲۳ ترکیب از مشتقات HEPT را بر اساس رگرسیون خطی چندگانه (MLR)[3] بررسی کردند. ضریب همبستگی مدل خطی شامل ۱۳ توصیف‌گر ۰/۹۹۹۹ به دست آوردند [۷].

به منظور دستیابی به ساختارهای مورد نیاز جهت بازداری آنزیم اینتگراز، مطالعه QSAR بر روی ۳۷ مشتق فتالیمید توسط دسالیو[4] انجام گرفت. نتایج این بررسی نشان داد که فعالیت ضد رتروویروسی به اندازه، شکل و قطبیت استخلاف‌ها بستگی دارد و روشن ساخت که با اصلاح اندازه استخلاف و میزان قطبیت آن می‌توان برهمکنش آنزیم- بازدارنده را بهبود بخشید [۸].

عرب چم جنگلی در سال ۲۰۰۹ با استفاده از روش غیرخطی بر مبنای شبکه عصبی، مدلی مناسب برای فعالیت ضدHIV مشتقات پیرول آریل‌سولفون‌ها ارائه داد به طوری که ضریب همبستگی برای سری آزمون ۰/۹۸۹بدست آمد [۹]. در همان سال مطالعه دیگری بر اساس شبکه عصبی توسط آلمریکو[5] برای مطالعه ضد ویروسی تعدادی از آنالوگ‌های اسایکلوویر[6] انجام شد [۱۰].

1 - Leave One Out Cross Validation
2 - Contribution
3 - Multiple Linear Regression
4 - Dessalew
5 - Almerico
6 - Acyclovir

در سال ۲۰۱۰ پاتل[1] 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 توصیف‌گر بود و با توجه به علامت ضرائب این توصیف‌گرها، نتیجه گرفتند که استخلاف‌های آب‌گریز و استخلاف‌های با شاخه بزرگتر، تأثیر مثبتی بر بازداری آنزیم نسخه‌بردار معکوس دارند [۱۱].

در سال ۲۰۰۸ سیکرو[2] و سزارینی[3] تعداد ۱۱۱ ترکیب از مشتقات ترکیب O- (۲-فتالامید و اتیل)- N- آریل تیوکربامات که دارای خاصیت بازدارندگی آنزیم نسخه‌بردار معکوس (RT) است، مورد بررسی قرار دادند. به منظور شناسایی ویژگی‌های مهم و تأثیرگذار این ترکیبات از استراتژی محاسباتی اتصال مولکولی[4] و سپس پیش‌بینی با روش COMFA[5] و COMSIA[6] به کار گرفته شده است. نتایج حاصل نشان می‌دهدکه وابستگی قابل قبولی بین ساختار این مشتقات و محل اتصال RT وجود دارد [۱۲].

محمد گودرزی و همکارانش در سال۲۰۱۲ فعالیت ضد ایدز گروهی از مشتقات ترکیب O- (۲- فتالامید و اتیل)- N- آریل تیوکربامات که جزو مهارکننده‌های نسخه بردار غیرنوکلئوزیدی است را مورد بررسی قرار دادند و از روش کلونی مورچگان[7] برای انتخاب متغیر استفاده کردند. برای ایجاد رابطه خطی بین توصیف‌گرهای انتخاب شده و فعالیت بازدارنده از روش حداقل مربعات جزئی[8](PLS) و رگرسیون خطی چندگانه استفاده کرده‌اند و رابطه

1 - Patel
2 - Elena Cichero
3 - Sara Cesarini
4- Docking
5- omparative Molecular Field Analysis
6- Comparative Molecular Similarity Index Analysis
7- Ant colony optimization (ACO)
8- Partial least squares

غیرخطی با استفاده از تابع پایه‌ای شعاعی[1] (RBF) و شبکه عصبی مصنوعی ایجاد شده است که تکنیک‌های PLS و RBF نتایج بهتری را نشان می‌دهد [۱۳].

۱-۲- مرگ سلول

مرگ سلول مرحله نهایی آسیب سلول است و عوامل مختلفی دارد که شایع‌ترین دلیل آن کمبود خون‌رسانی است. دو نوع مرگ سلول وجود دارد، نکروز[2] و آپوپتوز[3].

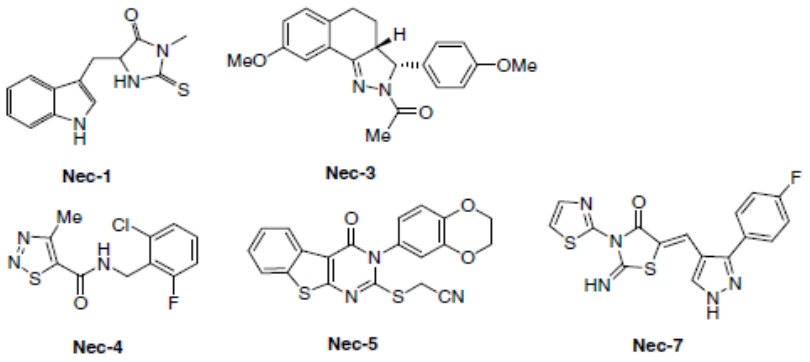
نکروز، نوعی مرگ سلولی است که به دلیل آسیب رسیدن به سلول ایجاد می‌شود. از جمله این آسیب‌ها، کمبود خون‌رسانی، تماس با سموم، ضربه و عفونت‌های مختلف است.

آپوپتوز مرگ سلولی برنامه‌ریزی شده است که در جاهای مختلف رخ می‌دهد مثلاً مرگ عده‌ای از سلول‌ها در دوران جنینی. آپوپتوز مرگ سلولی فعال است و به انرژی نیاز دارد و به شدت تحت کنترل بدن است. در واقع با آپوپتوز سلول‌هایی که آسیب DNA در آنها ترمیم نشده است برای جلوگیری از انتقال آسیب به نسل بعدی کشته می‌شوند. هرگونه اختلال در روند آپوپتوز، منجر به بیماری می‌شود که می‌تواند ناشی از کاهش مرگ سلولی باشد و باعث رشد سلول‌های سرطانی و یا اختلالات خودایمنی می‌گردد [۱۴].

اخیراً الکسی دگترو [4] و همکارانش یک نوع از نکروز را کشف کردند که مانند آپوپتوز کنترل شده می‌باشد ولی ویژگی‌های مورفولوژی نکروز را داراست که آن را نکروپتوز[5] نام نهادند. کشف نکروپتوز استراتژی‌های درمانی جدیدی ایجاد می‌کند. لذا شناخت و بهینه‌سازی مولکول‌های با وزن مولکولی پایین (به عنوان مثال نکروستاتین ها[6]) با قابلیت بازداری نکروپتوز در حذف این فرآیند در پاتو- فیزیولوژی بیماری برای توسعه‌ی درمانی قابل استفاده می‌باشد. با این هدف

1- Radial basis function
2 - Necrosis
3 -Apoptosis
4- Alexei Degterev
5- Necroptosis
6- Necrostatin

ساختار چندین نکروستاتین شناخته شده است. مکانیزم چگونگی عمل بازدارنده کاملاً واضح نیست و احتمالاً از طریق مهار برخی آنزیم‌های آبشار پیام‌دهی (کاسپازها) صورت می‌گیرد. نکروستاتین‌ها با وزن مولکولی پایین با اتصال به گیرنده‌های خاص موجود در پروتئین باعث غیر فعال شدن آن در پیام‌دهی می‌شوند. ساختار چندین نکروستاتین در شکل (۱-۱) نشان داده شده است [۱۵].

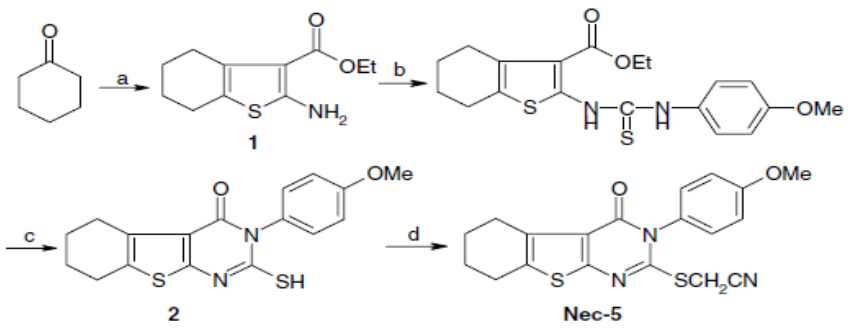


شکل(۱-۱)- ساختار نکروستاتین‌ها

۱-۲-۱- نکروستاتین-۵ (Nec-5)

نکروستاتین-۵ به عنوان یک مهارکننده‌ی قوی نکروپتوز شناخته شده است و مکانیسم اثر آن از طریق مهار غیرمستقیم RIP 1 کیناز[1] می‌باشد. به طور شیمیایی Nec-5، با عنوان ۳-*p*- متوکسی فنیل ۱-۵ و ۶- تترا متیلنو تینو [۲و۳-*d*] پیریمیدین-۴-یک-۲- مرکاپتو اتیل سیانید[2] نامیده می‌شود و روش سنتز آن به طور تجربی در شکل (۱-۲) نشان داده شده است. در بخش دوم تحقیق ارتباط کمی ساختار فعالیت مشتقات نکروستاتین-۵ مورد بررسی قرار گرفته است.

1- Respector intracting protein kinase 1
2-3-p-methoxyphenyl-5,6-tetramethylenothieno[2,3-d]pyrimidin-4-one-2-mercaptoethylcyanide



شکل (۱-۲)-روش سنتز Nec-5 [۴۰]

شرایط و واکنش‌گرها: (a) سیانواستات، S_8، Et_2NH، EtOH، ۱۲ساعت رفلاکس، ۷۲٪ بازده؛ (b) p –متوکسی فنیل ایزوتیوسیانات، EtOH، ۵-۶ ساعت رفلاکس، ۸۵٪ بازده؛ (c) HCl اتانولی، ۱۲-۲۴ ساعت رفلاکس، ۷۸٪ بازده؛ (d) KOH در EtOH ۷۰٪ سپس BrCH2CN، ۱-۲ ساعت، ۹۲٪ بازده.

۱-۲-۲- مروری بر کارهای انجام شده

عرب چم‌جنگلی و اشرفی در سال ۲۰۱۳ با استفاده از روش غیرخطی بر مبنای شبکه عصبی، مدلی مناسب برای فعالیت مهارکنندگی نکروپتوز مشتقات [۱و۲و۳]- تیادیازول و تیوفن [۱] ارائه دادند. قابلیت پیش‌بینی مدل ساخته شده که شامل ۷ توصیف‌گر بود، توسط رد مرحله‌ای تک‌تک مورد ارزیابی قرار گرفت که ضریب همبستگی محاسبه شده برابر ۰/۹۴۴۶ بود [۱۶].

۱-۳- ضرورت تحقیق

پیشرفت علوم در زمینه‌های مختلف روش‌های آزمایشگاهی کمک شایانی به بهبود نتایج حاصل از این روش‌ها کرده است. با این وجود، مشکلات و محدودیت‌هایی از جمله وجود عدم قطعیت نتایج آزمایشگاهی، هزینه‌های بالا، در دسترس نبودن امکانات باعث توسعه روش‌های نظری شده است. با گذشت زمان و ارزشمندتر شدن وقت و لزوم صحت و دقت در محاسبات، نرم‌افزارها و ابزارهای محاسباتی به یاری محققان رسیده‌اند. از جمله روش‌های

1- [1,2,3] thiadiazole and thiophene derivatives

نظری، رابطه کمی ساختار- فعالیت[1] (QSAR) است که در آن می‌توان از روش‌های مختلفی برای ایجاد روابط خطی و غیرخطی میان ساختار و خاصیت دارویی ترکیباتی که فعالیت دارویی آنها با استفاده از روش‌های تجربی اندازه‌گیری شده‌اند، استفاده نمود. در این پایان‌نامه از روش جنگل‌های تصادفی[2] همراه با توصیف‌گرهای سهم‌گروه[3] برای مدل‌سازی فعالیت داروئی ترکیبات ضدایدز استفاده شده است. در بخش دوم پژوهش شبکه عصبی مصنوعی برای مدل‌سازی فعالیت داروئی ترکیبات نکروستاتین-۵ به کار گرفته شده است.

1 - Quantitative Structure- Activity Relationship
2 - Random Forests
3 - Group Contribution

فصل دوم

کمومتریکس و کاربرد آن در مطالعات QSAR

۲-۱- کمومتریکس[1]

با اجرای هر آزمایش مجموعه‌ای پراکنده از داده‌های کمی یا کیفی به‌دست می‌آید که به کمک روش‌های تحلیلی و تفسیری مناسب همچون روش‌های منطق فکری و استدلالی، آماری و ریاضی می‌توان از آن‌ها در جهت رسیدن به نتایج سودمند بهره گرفت. مجموعه روش‌های ساده‌ای که به این منظور مورد استفاده قرار می‌گیرند، با عنوان کمومتریکس خوانده می‌شوند. کمومتریکس شاخه‌ای جدید از علم شیمی است که انجمن کمومتریکس، آن را این چنین تعریف کرده است: سازماندهی مطالب شیمی با روش‌های ریاضی، آماری، منطقی، علوم رایانه‌ای که به منظور طراحی عملیات آزمایشگاهی، انتخاب روش‌های بهینه اندازه‌گیری و دستیابی حداکثر به اطلاعات، به کمک تفسیر و تحلیل داده‌های شیمیایی، انجام می‌گیرد [۱۷]. بسیاری از شیمیدان‌ها اسوانت ولد[2] را به عنوان اولین کسی که روش کمومتریکس را معرفی کرده است، نام می‌برند و به او لقب پدر علم کمومتریکس را داده‌اند [۱۸].

برخی از کاربردهای کمومتریکس در شیمی تجزیه عبارتند از:

- شبیه‌سازی و جستجوی طیف‌های جرمی
- پیش‌بینی میزان جابجایی C NMR13
- شناسایی الگوها از داده‌های کروماتوگرافی
- پیش‌بینی ضریب پاسخ آشکارسازهای شعله - یونشی و هدایت گرمایی
- پیش‌بینی ضرایب انتخاب‌گری الکترودهای یون‌گزین
- پیش‌بینی خواص و فعالیت مولکول‌ها
- شناخت الگوها و طبقه‌بندی داده‌ها

[1] - Chemometrics
[2] - Svante wold

۲-۲- ارتباط کمی ساختار- فعالیت (QSAR)

ارتباط کمی ساختار- فعالیت (QSAR)، به نحوه ارتباط بین فعالیت بیولوژیکی و ساختار شیمیایی ترکیبات می‌پردازد و یکی از زمینه‌های مهم کمومتریکس می‌باشد. مولکول‌های مشابه با تغییر کوچکی در ساختارشان می‌توانند فعالیت بیولوژیکی کاملاً متفاوتی داشته باشند و در واقع QSAR نیز سعی در پیدا کردن رابطه‌ی هماهنگ میان فعالیت بیولوژیکی و ویژگی‌های مولکولی، به منظور کاربرد این قواعد برای ارزیابی فعالیت ترکیبات جدید دارد [۱۹].

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.

به طور کلی مطالعات QSAR شامل چهار بخش اصلی است; انتخاب سری مولکولی، بهینه‌سازی ساختار مولکول‌ها، محاسبه توصیف‌گرها و انتخاب بهترین آنها، مدل‌سازی و ارزیابی مدل.

۲-۲-۱- فراهم کردن سری داده‌ها

اولین مرحله در مطالعات QSAR، جمع‌آوری و انتخاب تعدادی مولکول است که خاصیت یا فعالیت مورد نظر آنها به طور تجربی در دسترس باشد (لازم به ذکر است که فعالیت یا خاصیت ترکیبات باید در شرایط آزمایشگاهی یکسان تعیین شده باشند). بدیهی است که هر چقدر اطلاعات تجربی قابل دسترس برای طراحی مدل بیشتر باشد بدون تردید مدل کارآیی بهتری خواهد داشت.

۲-۲-۲- بهینه سازی ساختار مولکول‌ها

برای بهینه‌سازی ساختار هندسی مولکول‌ها از روش‌های شیمی محاسباتی استفاده می‌شود. ایجاد توصیف‌گر های هندسی و هیبریدی بر مبنای ساختار و هندسه دقیق مولکولی استوار است و اگر ساختارها به شکل صورتبندی[1] با حداقل انرژی نباشند مقادیر غیرصحیحی برای این توصیف‌گرها ایجاد می‌شود [۹]. بنابراین باید ساختار هندسی ترکیب بهینه شود بدین منظور از نرم‌افزار Hyperchem جهت رسم و بهینه‌سازی ساختار مولکول‌ها استفاده می‌شود. به کمک این نرم‌افزار می‌توان طول پیوند، زاویه پیوندی و زوایای پیچشی را در مولکول تعیین کرد. داده‌های حاصل از این نرم‌افزار را می‌توان به عنوان ورودی به سایر نرم‌افزارها معرفی نمود.

۲-۲-۳- محاسبه توصیف‌گرها

توصیف‌گرها کمیت‌هایی هستند که به ساختار مولکول ارتباط دارند و برای هر مولکولی مقادیر خاصی را به خود اختصاص می‌دهند. در واقع هر توصیف‌گر بیان‌گر خصوصیت ویژه‌ای از مولکول است که ممکن است بر فعالیت مورد نظر مؤثر باشد و هر توصیف‌گر اطلاعات خاصی از مولکول را که بر کمیت مورد مدل‌سازی اثر می‌گذارد را در اختیار قرار می‌دهد.

توصیف‌گرها از نظر بیان چگونگی خصوصیات مولکول به دسته‌های متفاوتی تقسیم می‌شوند که معروف‌ترین آن‌ها عبارتند از:

- توصیف‌گرهای توپولوژیکی[2]
- توصیف‌گرهای هندسی[3]

1- Conformation
2- Topological Descriptors
3- Geometrical Descriptors

- توصیف‌گرهای الکترونی [1]

- توصیف‌گرهای فیزیکوشیمیایی [2]

- توصیف‌گرهای هم‌بستگی دو بعدی [3]

امروزه نرم‌افزارهای مختلفی طراحی شده‌اند که براساس محاسبات دقیق تئوری می‌توانند توصیف‌گرهای بیشتری را با دقت خوب و سرعت زیاد محاسبه کنند. در این پروژه برای ایجاد مدل QSAR مشتقات تیوکربامات از توصیف‌گرهای سهم گروه استفاده شده است که این توصیف‌گرها بسیار ساده‌اند و به راحتی از روی ساختار ترکیب و بدون نیاز به نرم‌افزار محاسبه می‌شوند. برای بررسی فعالیت مشتقات نکروستاتین-۵ از نرم‌افزار Dragon استفاده شده است که در۱۸ کلاس توصیف‌گرها را محاسبه می‌کند.

۲-۲-۴-توصیف‌گر های سهم گروه

استفاده از روش سهم گروه[4]، یکی از معتبرترین و در عین حال ساده‌ترین روش‌های پیش‌بینی خواص مواد از روی ساختار مولکولی است. روش سهم گروه‌ها به طور گسترده‌ای در پیش‌بینی خواص فیزیکی و ترمودینامیکی مواد به کار می‌رود [۲۰]. این روش به رابطه‌ی جمع پذیر گروه‌ها معروف است. مبنای ایجاد هر روش سهم گروه شامل دو مرحله است:

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 یک گروه مشخص در یک مولکول با تمام مولکول‌ها یکسان در نظر گرفته می‌شود.

1 - Electronic Descriptors
2 - Physicochemical Descriptors
3 - Autocorrelation Descriptors
4- Group contribution

توصیف‌گرهای سهم گروه در واقع نشان دهنده‌ی گروه‌های تشکیل دهنده یک ترکیب است و بدیهی است که هر چه تمایز بین گروه‌ها بیشتر شود، صحت روش سهم گروه‌ها افزایش می‌یابد تا جایی که سرانجام هر ترکیب فقط شامل گروه‌های مربوط به خودش باشد. توصیف‌گرهای سهم گروه می‌تواند شامل قطعات اصلی مولکول، محل استخلاف‌ها، نوع استخلاف‌ها و... باشد. در این پژوهش نیز از توصیف‌گرهای سهم گروه استفاده شده است که از روی ساختار ترکیبات و بدون نیاز به نرم‌افزار محاسبه شده‌اند.

۲-۲-۵- انتخاب توصیف‌گر های مهم

قبل از انتخاب توصیف‌گر، فرآیند کاهش متغیر صورت می‌گیرد زیرا زیاد بودن توصیفگرها و همبستگی بالای برخی از آن‌ها با یکدیگر باعث دشوار شدن فرآیند مدل‌سازی می‌شود. بنابراین متغیرها با همبستگی بالا حذف می‌شوند. پس از کاهش متغیرها، روش انتخاب متغیر به کار گرفته می‌شود که این مرحله شامل یافتن توصیف‌گرهایی است که حاوی اطلاعات مفید و در ارتباط با خاصیت مورد بررسی باشند. از جمله روش‌های انتخاب متغیر می‌توان به رگرسیون مرحله ای[۱]، الگوریتم ژنتیک[۲]، روش جایگزینی[۳] اشاره کرد. که در این پایان‌نامه از روش رگرسیون مرحله‌ای در نرم‌افزار SPSS برای انتخاب توصیف‌گر استفاده شده است. در روش رگرسیون مرحله‌ای که از طریق محاسبه ضریب همبستگی[۴] بین توصیف‌گر و فعالیت بیولوژیکی انجام می‌شود اگر توصیف‌گری دارای ضریب همبستگی برابر یک باشد یعنی می‌تواند فعالیت مورد نظر را به درستی توصیف کند و ضریب همبستگی صفر بدین معنی است که هیچ ارتباطی بین توصیف‌گر و فعالیت دارویی وجود ندارد. در این روش ابتدا توصیف‌گری که بیشترین همبستگی با متغیر وابسته (خاصیت مورد نظر) را دارد، وارد

1 - Stepwise
2 - Genetic Algorithm
3 - Replacement Method
4 - Correlation coefficient

مدل شده و با ورود هر توصیف‌گر جدید توصیف‌گرهای موجود در مدل بررسی شده و اگر هر کدام از آنها سطح معناداری خود را از دست داده باشند قبل از ورود توصیف‌گر جدید از مدل خارج می‌شوند [۲۱].

۲-۲-۶- ساخت مدل

مرحله اصلی در مطالعات QSAR ساختن مدل می‌باشد. پس از جمع‌آوری داده‌ها باید مدلی ساخته شود که بتواند خاصیت مورد نظر را پیش‌بینی کند. در واقع مدل بین خاصیت مورد بررسی و توصیف‌گرها ارتباط برقرار می‌کند. یک مدل مناسب که همبستگی قابل قبولی بین مقادیر تجربی فعالیت مورد نظر و مقدار بدست آمده به صورت تئوری نشان می‌دهد، به پیش‌بینی رفتار مولکول‌های جدید کمک می‌نماید. چندین روش متفاوت برای ساختن مدل QSAR وجود دارد که می‌توان به رگرسیون خطی چندگانه (MLR)، رگرسیون اجزای اصلی(PCR)[1]، حداقل مربعات جزئی(PLS)، شبکه عصبی مصنوعی (ANN) و جنگل‌های تصادفی[2](RF) اشاره کرد. در این پژوهش از دو روش شبکه عصبی مصنوعی و جنگل‌های تصادفی استفاده شده است.

۲-۳- مقدمه‌ای بر شبکه عصبی مصنوعی

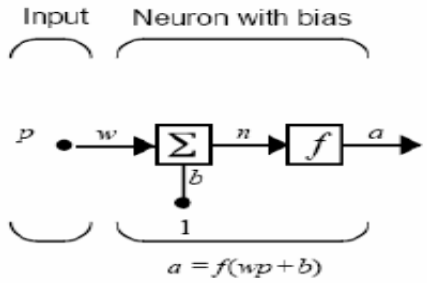
شبکه عصبی مصنوعی یک سامانه پردازشی داده‌ها است که از مغز انسان ایده گرفته و پردازش داده‌ها را به عهده پردازنده‌های کوچک و بسیار زیادی سپرده که به صورت شبکه‌ای به هم پیوسته و موازی با یکدیگر رفتار می‌کنند تا یک مسئله را حل نمایند. در این شبکه‌ها به کمک دانش برنامه‌نویسی، ساختار داده‌ای طراحی می‌شود که می‌تواند همانند نرون[3] عمل کند. به این ساختار داده نرون گفته می‌شود. با ایجاد شبکه‌ای بین این نرون‌ها و اعمال یک الگوریتم آموزشی به آن، شبکه را

1 - Principle Component Regression
2 - Random forestes
3 - Neuron

آموزش می‌دهند. در این شبکه عصبی نرون‌ها دارای دو حالت فعال (روشن یا یک) و غیرفعال (خاموش یا صفر) می‌باشند و هر یال (سیناپس یا ارتباط بین گره‌ها) دارای یک وزن[1] می‌باشد. یال‌های با وزن مثبت موجب تحریک یا فعال نمودن گره فعال بعدی می‌شوند و یال‌های با وزن منفی، گره متصل بعدی را غیرفعال یا مهار می‌کنند [۲۲] .

۲-۳-۱- ساختار و عملکرد نرون مصنوعی

شکل(۲-۱) الگویی از یک نرون محاسباتی با توجه به نحوه عملکرد نرون طبیعی ارائه می‌دهد. در شکل زیر 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[2] ضرب شده و سپس با wp جمع می‌شود. این حاصل جمع، ورودی خالص n برای تابع محرک f را تشکیل می‌دهد. 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 [۲۴].



شکل(۲-۱)- ساختمان یک نرون محاسباتی [۲۳]

۲-۳-۲- تابع انتقال[3]

زمانی که ورودی‌ها با توجه به اهمیت آنها با یکدیگر جمع جبری شوند توسط تابع انتقال به نرون بعدی منتقل می‌شوند. در واقع تابع انتقال یکی از اجزای شبکه عصبی می‌باشد که برحسب کاربرد

1- Weight
2- Bias
3 -Transfer function

می‌تواند توابع متفاوتی باشد. سه تابع انتقال رایج در شبکه‌های عصبی که در این پژوهش برای بهینه‌سازی شبکه به کار گرفته شده است عبارتند از:

- تابع انتقال لگاریتم سیگموئید[1](log sig)

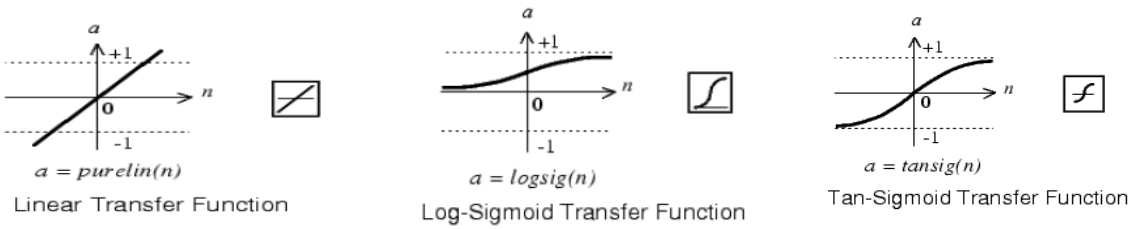
از این تابع انتقال در شبکه‌های پس انتشار استفاده می‌شود این تابع مقادیر ورودی را در محدوده منفی بی‌نهایت تا مثبت بی‌نهایت دریافت کرده و خروجی بین ۰ و۱ تولید می‌نماید.

- تابع انتقال تانژانت سیگموئید[2] (tansig)

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

- تابع انتقال خطی[3] (purelin)

نرون‌هایی که از این تابع انتقال استفاده می‌کنند معمولاً برای تقریب خطی در فیلترهای خطی به کار می‌روند این تابع همان مقدار ورودی را به عنوان خروجی برمی‌گرداند. شکل(۲-۲) مقایسه بین توابع انتقال را نشان می‌دهد.



شکل (۲-۲)- انواع توابع انتقال[۲۴]

1 - Logarithm sigmoid transfer function
2 - Hyperbolic tangent transfer function
3- Linear transfer function

۲-۳-۳- ساختارهای شبکه

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

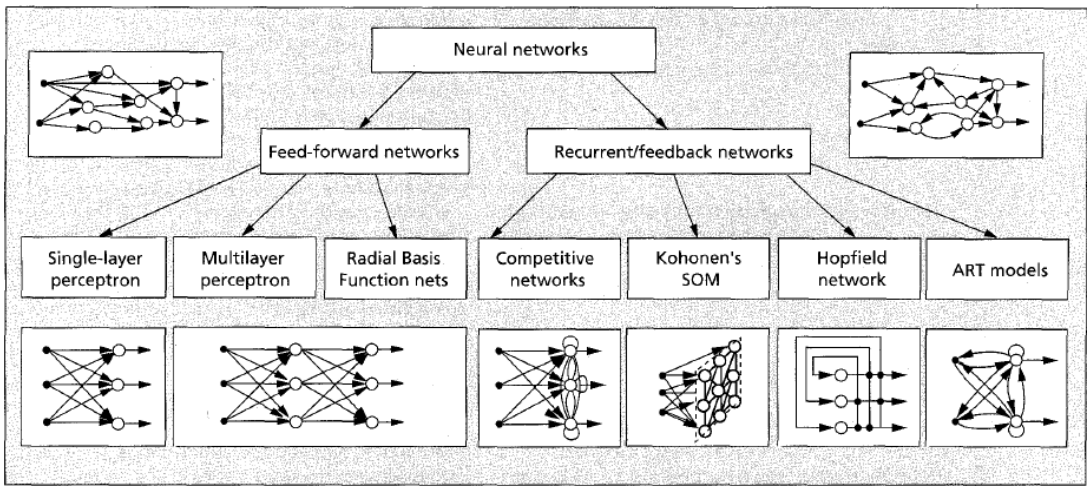
• شبکه‌های جلوسو[۱] که در آنها گراف حلقه ندارد.

• 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. در شکل(۲-۳) شبکه‌های اصلی هر دسته نشان داده شده 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. 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.

1- Feed-Forward Netwoks
2- Recurrent Networks
3- Feedback
4- Multi Layer Perceptron



شکل (۲-۳)- انواع شبکه‌های عصبی جلوسو و برگشتی[۲۵]

۲-۳-۴-آموزش شبکه‌های جلو سو با تکنیک پس‌انتشار[1]

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 [۲۶]:

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[2] 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 الگوریتم آموزشی پس‌انتشار می‌باشد.

1 - Back propagation
2- Epoch

۲-۴- جنگل‌های تصادفی

یکی از روش‌های نوین و پیشرفته آماری، روش جنگل‌های تصادفی است که جزو روش‌های یادگیری ماشین بوده و از این پس آن را به اختصار با نماد RF نمایش می‌دهیم. این روش در سال ۲۰۰۱ توسط لیو بریمن[1]، آماردان دانشگاه برکلی آمریکا ارائه شد [۲۷]. اساس کار مدل جنگل تصادفی ریشه در مدل درخت‌های رده‌بندی یا رگرسیون (CART)[2] دارد، لذا لازم است ابتدا روش CART توضیح داده شود.

۲-۴-۱- درخت رگرسیون و طبقه‌بندی CART

درخت رگرسیون و طبقه‌بندی یکی از روش‌های غیرپارامتری در داده‌کاوی[3] است که توسط بریمن و همکارانش در سال۱۹۸۴ ارائه شد. این روش هر دو قابلیت رگرسیون و رده‌بندی را داراست. به این معنی که وقتی متغیر پاسخ از نوع کیفی باشد قابلیت طبقه‌بندی و در صورتی که متغیر پاسخ از نوع کمی باشد (پیوسته) از قابلیت رگرسیون برخوردار است.

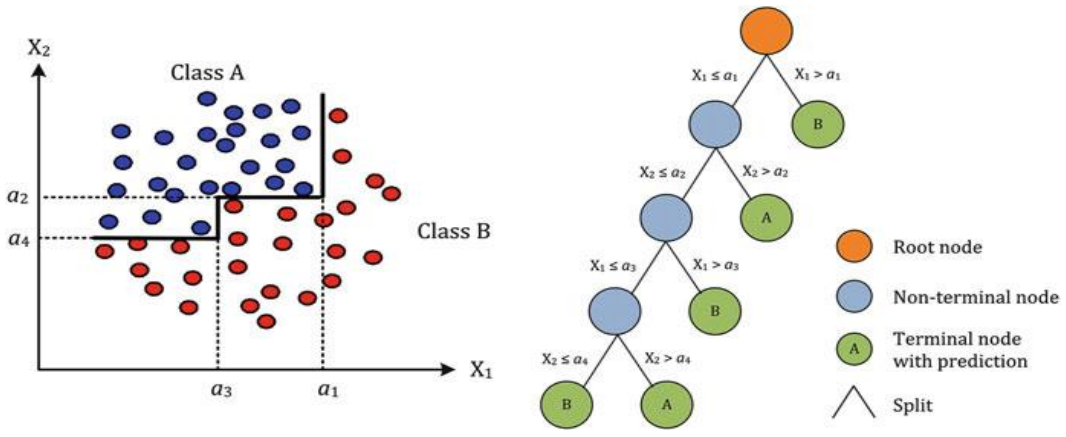
۲-۴-۲- درخت طبقه‌بندی [۲۸]

درخت طبقه‌بندی قادر به تولید قانون‌هایی براساس متغیرهای توضیحی، برای پیش‌بینی طبقه متغیر پاسخ می‌باشد. این قوانین را می‌توان برای سهولت درک آن به شکل ساختار درختی نمایش داد. فرآیند ساخت درخت طبقه‌بندی طی دو مرحله‌ی ساخت و هرس درخت انجام می‌شود که در ادامه شرح داده می‌شود.

1- Leo Breiman
2- Classification And Regression Tree
3- Data mining

- ساخت درخت

فضای به وجود آمده توسط متغیرها یک فضایN بعدی است. هدف در درخت طبقه‌بندی، افراز[1] این فضا به ابر مکعب مستطیل‌های مجاور هم است به گونه‌ای که مشاهدات واقع در هر ابرمکعب دارای بیشترین همگونی از لحاظ طبقه خاصی از متغیر پاسخ باشند. نحوه افراز فضای متغیرها، سلسله مراتبی[2] و به صورت دودویی[3] است یعنی در هر مرحله، این فضا به دو قسمت تقسیم می‌شود. فرض کنید فضای متغیرها در راستای متغیر X_1 و به ازای مقادیر a_1 و a_3 و در راستای متغیر X_2 و به ازای مقادیر a_2 و a_4 افراز شده باشد (a مقداری در دامنه متغیر X است) آنگاه این افراز را می‌توان به صورت شکل (۲-۴) نشان داد. هر یک از نواحی موجود در درخت تصمیم در اصطلاح یک گره[4] نامیده می‌شود. در واقع گره، مقداری از هر متغیر است که افراز در آن نقطه انجام می‌گیرد. ناحیه اول که شامل کل داده‌هاست را گره ریشه[5] (مادر) و سایر گره‌ها را گره‌های فرزند می‌نامند. به گره‌های فرزندی که خود نیز به دو گره جدید تقسیم شده‌اند گره داخلی یا غیرپایانی و به گره‌هایی که تقسیم نشده و پیش‌بینی انجام می‌دهند گره پایانی می‌گویند. شکل (۲-۴) این گره‌ها را نشان می‌دهد.



شکل (۲-۴)- نحوه افراز در درخت طبقه‌بندی[۲۸]

1- Partition
2- Hierarchy
3- Binary
4- Node
5-Root node

فرض کنید ناحیه اصلی شامل کل داده‌ها (R) و نواحی چپ و راست ایجاد شده به ترتیب R_1 و R_2 باشد. مجموعه متغیرهای توضیحی و متغیر پاسخ (با k طبقه) را به ترتیب با x_i و y_i نشان دهیم آنگاه فراوانی نسبی طبقه K واقع در ناحیه t که با P(k|t) نمایش داده می‌شود برابر است با [۲۹]:

$$P(k|t) = \frac{1}{N_t}\sum_{x_i \epsilon t} I\,(y_i = k) \qquad (۱-۲)$$

در این رابطه N_t تعداد مشاهدات درون ناحیه t و I به صورت زیر تعریف می‌شود:

$$I = \begin{cases} ۱\,, y_i = k \\ ۰\,, y_i \neq k \end{cases} \qquad (۲-۲)$$

آنگاه شاخص جینی[1] که معرف میزان ناخالصی ناحیه t می‌باشد، عبارت است از:

$$i(t) = \sum_{k \neq L} P(k|t)\, P(L|t) \qquad (۳-۲)$$

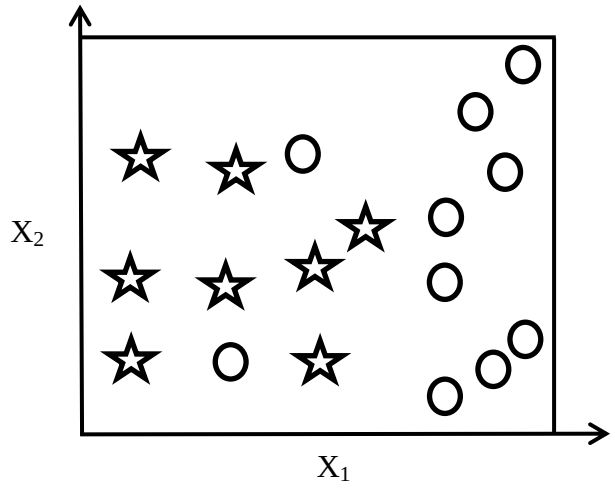
منظور از L طبقه‌ای دیگر از متغیر پاسخ می‌باشد. شاخص جینی را می‌توان برای هر یک از سه ناحیه R_2, R_1, R محاسبه نمود تغییرات تابع ناخالصی[2] ترکیبی از این سه مقدار است که با $\Delta i(t)$ نشان داده شده و به صورت رابطه (۲-۴) تعریف می‌شود.

$$\Delta i(t) = i(t_R) - E[i(t_c)] = i(t_R) - P_{R1} i(t_{R1}) - P_{R2} i(t_{R2}) \qquad (۴-۲)$$

که در آن t_c نواحی R_1 وR_2 ایجاد شده و P_{R1} و P_{R2} نسبتی از کل مشاهدات ناحیه R است که به ترتیب در R_1 وR_2 واقع شده‌اند. با تکرار فرآیند فوق برای تمامی متغیرها و به ازای مقادیر مختلف آنها مقدار $\Delta i(t)$ برای همه افرازهای ممکن محاسبه شده و در نهایت افرازی که مقدار بیشینه $\Delta i(t)$ را ایجاد کند، به عنوان افراز بهینه در نظر گرفته خواهد شد. نواحی جدید ایجاد شده نیز به همین صورت به نواحی کوچکتر تقسیم می‌شوند. برای فهم بهتر موضوع، مثالی از نحوه افراز فضای متغیرها برای حالتی که با دو متغیر سروکار داریم و متغیر پاسخ دو مقدار ۰ و ۱ را اختیار می‌کند ارائه می‌کنیم.

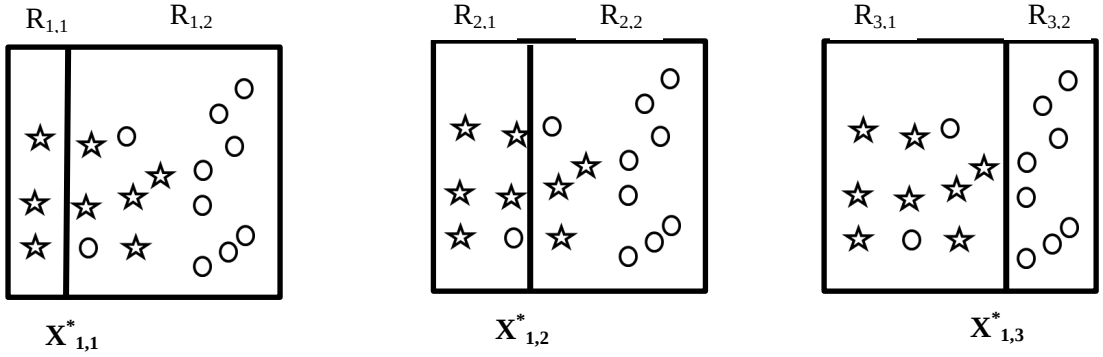
1 - Gini Index
2 - Impurity function

فرض کنید مجموعه داده‌ها به صورت شکل (۲-۵) توزیع شده باشند که در آن منظور از دایره و ستاره به ترتیب مشاهداتی هستند که متغیر پاسخ آنها مقدار صفر و یک را اختیار می‌کنند [۲۹].

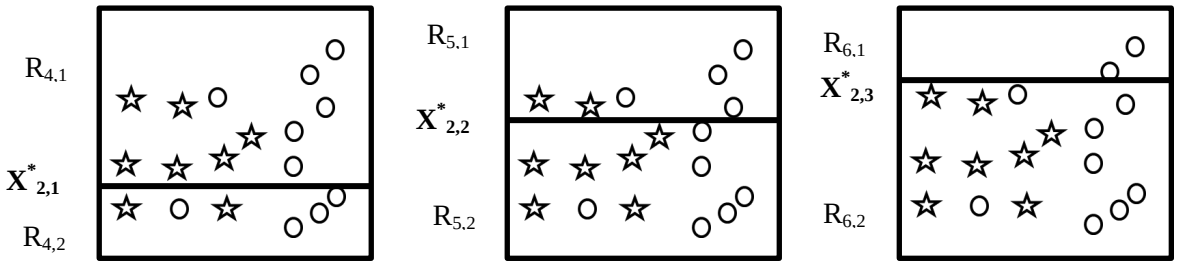


شکل (۲-۵)- فضای ایجاد شده توسط دو متغیر توضیحی x_1 , x_2

همان طور که بیان شد، هدف در روش درخت تصمیم، افراز این فضای دوبعدی به مستطیل‌های مجاور هم به گونه‌ای است که مشاهدات موجود در هر مستطیل بیشترین همگونی را داشته باشند، از آنجایی که برای هر یک از متغیرهای X_2, X_1، تعداد نقاط افراز زیادی می‌توان در نظر گرفت، لذا در اینجا جهت سهولت، فقط چند نقطه‌ای که به ظاهر برای انجام افراز در راستای هر یک از متغیرها محتمل‌تر هستند در نظر می‌گیریم (شکل۲-۶ و شکل ۲-۷).



شکل(۲-۶)- نقاط افراز محتمل در راستای متغیر X_1



شکل(۲-۷)- نقاط افراز محتمل در راستای متغیر X_2

$X^*_{j,k}$، متغیر jام به ازای مقدار k می‌باشد. حال باید معیار ناخالصی (۲-۴) را برای تک تک حالات فوق محاسبه کرده و از میان آن‌ها، راستا و نقطه‌ای که این معیار به ازای آن ماکزیمم می‌شود را انتخاب کنیم. در اینجا معیار ناخالصی مربوط به افراز ایجاد شده توسط $X^*_{j,k}$ را با $\Delta_{j,k}i(t)$ نشان می‌دهیم. از آنجایی که متغیر پاسخ دارای دو سطح است روابط (۲-۲) و (۲-۳) را به صورت زیر بازنویسی می‌کنیم.

$$I = \begin{cases} ۱ , y_k = ۱ \\ ۰ , y_k = ۰ \end{cases} \qquad\qquad i(t) = P\left(۱, t\right) P(۰, t)$$

به عنوان مثال، تابع ناخالصی مربوط به افراز ایجاد شده در راستای متغیر X_1 و مقدار $X^*_{1,1}$ به صورت زیر محاسبه می‌شود:

$$P\left(۱|R\right) = \frac{۸}{۱۸} \quad , \quad P\left(۰|R\right) = \frac{۱۰}{۱۸}$$

$$P\left(۱\middle|R_{1,1}\right) = \frac{۳}{۳} \quad , \quad P\left(۰\middle|R_{1,1}\right) = \frac{۰}{۳}$$

$$P\left(۱\middle|R_{1,2}\right) = \frac{۵}{۱۵} \quad , \quad P\left(۰\middle|R_{1,2}\right) = \frac{۱۰}{۱۵}$$

$$P\left(R_{1,1}\right) = \frac{۳}{۱۸} \quad , \quad P\left(R_{1,2}\right) = \frac{۱۵}{۱۸}$$

$$i\left(R_{1,1}\right) = P\left(۱\middle|R_{1,1}\right) P\left(۰\middle|R_{1,1}\right) = \frac{۳}{۳} \times ۰ = ۰$$

$$i(R_{1,2}) = P\left(۱\middle|R_{1,2}\right) P\left(۰\middle|R_{1,2}\right) = \frac{۵}{۱۵} \times \frac{۱۰}{۱۵} = \frac{۲}{۹}$$

$$i(R) = P\left(۱\middle|R\right) P\left(۰\middle|R\right) = \frac{۸}{۱۸} \times \frac{۱۰}{۱۸} = \frac{۲۰}{۸۱}$$

$$\Delta_{1,1} i(t) = \frac{۲۰}{۸۱} - \left(\frac{۳}{۱۸} \times ۰\right) - \left(\frac{۱۵}{۱۸} \times \frac{۲}{۹}\right) = \frac{۲۰}{۸۱} - \frac{۱۵}{۸۱} = ۰/۰۶۱$$

تابع ناخالصی برای بقیه نقاط نیز به همین صورت محاسبه می شود که نتایج به قرار زیر است:

$\Delta_{1,1} i(t) = ۰/۰۶۱$ $\Delta_{2,1} i(t) = ۰/۰۰۶$

$\Delta_{1,2} i(t) = ۰/۰۷۵$ $\Delta_{2,2} i(t) = ۰/۰۰۶$

$\Delta_{1,3} i(t) = \mathbf{۰/۱۵۸}$ $\Delta_{2,3} i(t) = ۰/۰۲$

با توجه به نتایج به دست آمده داریم:

$$\mathrm{Max}_{j=1,2\ ,\ K=1,2,3}\ \Delta_{j,k} i(t) = \Delta_{1,3} i(t)$$

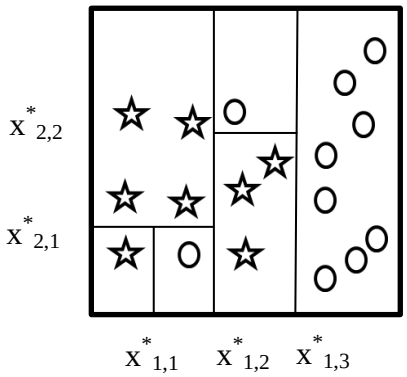
بنابراین اولین افراز در راستای متغیرX_1 و به ازای مقدار $X^*_{1,3}$ ایجاد می‌گردد. نواحی $R_{3,1}$ و $R_{3,2}$ نیز به همین صورت به نواحی کوچکتر افراز می‌شوند. فرآیند رشد درخت زمانی متوقف می‌شود که حداقل یکی از حالات زیر اتفاق بیافتد.

۱. در هر یک از نواحی جدید ایجاد شده، تنها یک مشاهده وجود داشته باشد.

۲. همه مشاهدات داخل یک ناحیه، دارای رده مشابهی از متغیر پاسخ باشند.

۳. یک پیش‌فرض توسط کاربر در نظر گرفته شود. مثلاً کاربر علاقمند باشد که در هر گره پایانی حداقل ۵ مشاهده وجود داشته باشد.

در نهایت پس از رشد کامل درخت، فضای متغیرها به صورت شکل (۲-۸) افراز می‌شود.



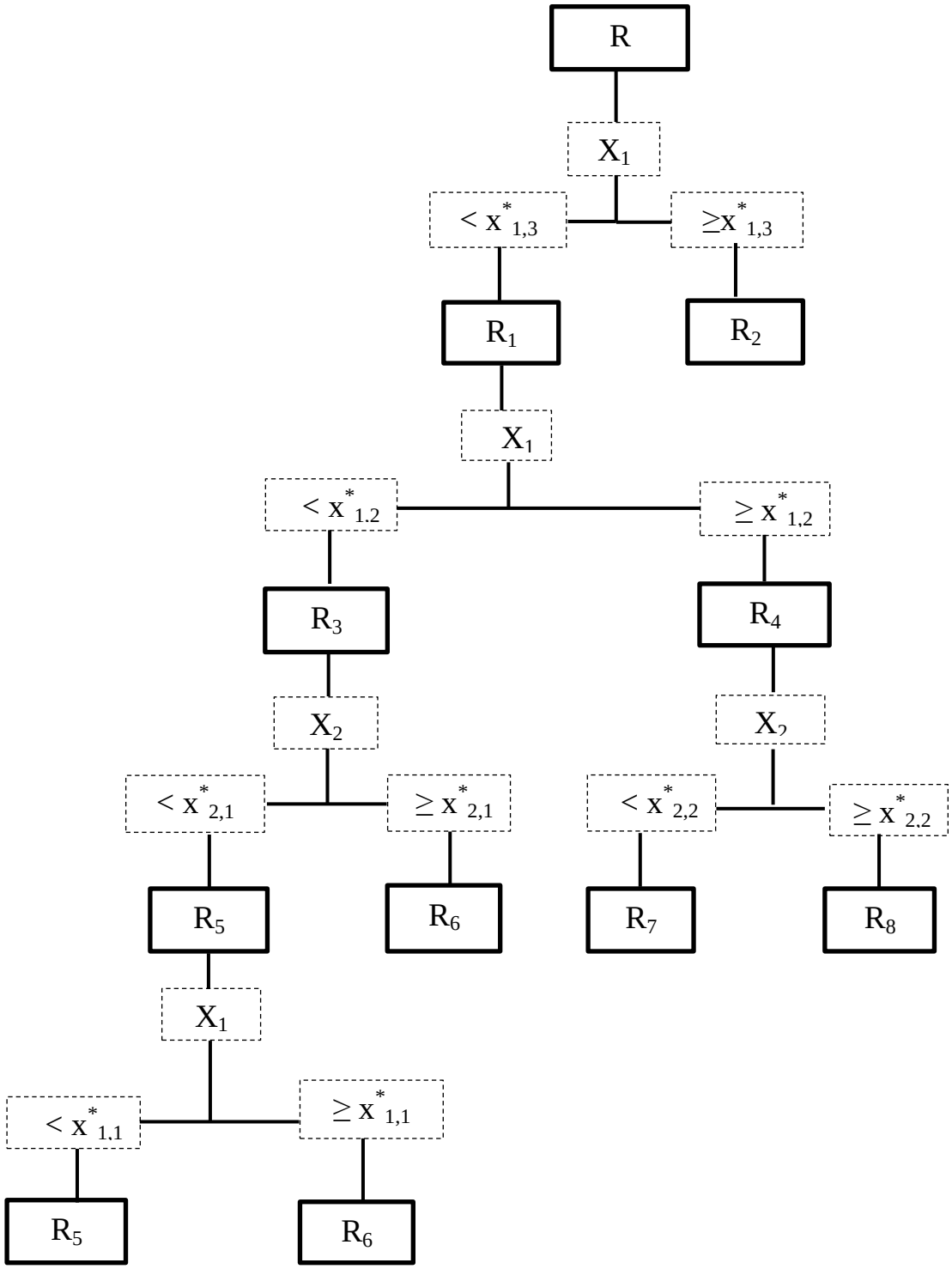
شکل(۲-۸)- افراز کامل فضای ایجاد شده توسط دو متغیر، در مدل درخت طبقه‌بندی

این افراز را می‌توان به صورت یک درخت (شکل۲-۹) نیز نشان داد که علت نام‌گذاری این روش به نام درخت تصمیم، همین ساختار درخت‌گونه آن است.

رشد بیش از حد درخت ممکن است موجب بروز پدیده بیش‌برازشی[1] شود. اصطلاح بیش برازشی زمانی برای یک مدل به کار می‌رود که اختلاف خطای طبقه‌بندی برای داده‌های مدل‌ساز[2] و داده‌های آزمون[3] بسیار زیاد باشد. برای حل این مشکل باید اجازه دهیم درخت به طور کامل رشد کند. پس از اینکه درخت به طور کامل رشد کرد نوبت به هرس کردن آن می‌رسد. فرآیند هرس کردن درخت به این صورت است که در هر گام یکی از شاخه‌های درخت که نبود آن باعث کمترین افزایش در خطای طبقه‌بندی می‌شود حذف می‌گردد.

در نهایت با داشتن یک مشاهده جدید و قرار دادن آن در درخت، این مشاهده با توجه به مختصاتش درون یکی از گره‌های پایانی قرار می‌گیرد. عمل طبقه‌بندی برای این مشاهده به این صورت است که طبقه‌ای که بیشترین فراوانی را در گره پایانی مربوطه داراست، به عنوان طبقه مشاهده مربوطه انتخاب می‌شود.

1 - Over fitting
2 - Train data
3 - Test data



شکل(۲-۹)- ساختار درختی مدل درخت طبقه‌بندی

۲-۴-۳- درخت رگرسیون

تشکیل درخت رگرسیون همانند درخت طبقه‌بندی با افراز فضای نمونه، طی مراحلی سلسله مراتبی انجام می‌شود که هر مرحله، براساس آن است که تفکیک فضا در راستای کدام متغیر و در چه مقداری از آن متغیر بایستی انجام گیرد. در این روش، فضای متغیرها به گونه‌ای تفکیک می‌شود که در هر ناحیه بتوان رویه پاسخ را توسط یک مدل ساده مانند یک صفحه تقریب نمود که همانند درخت طبقه‌بندی، افراز فضای متغیرها به شکل مربع یا مستطیل است.

فرض کنید $S = \{(X_{1,i}, X_{2,i})\ ;\ i = 1 \dots n\}$ مجموعه مشاهدات دو متغیر توصیفی X_1 و X_2 و همچنین $\{y_1, \dots \dots y_n\}$ مجموعه مقادیر متغیر پاسخ Y باشند. اکنون با فرض این که فضای متغیرهای X_1 و X_2 به پنج ناحیه $R_1, R_2, \dots, R_5$ تفکیک شده باشد، به طور قطع هر ناحیه‌ی R_m ; $m=1,\dots5$ شامل زیر مجموعه‌ای از داده‌های مشاهده شده به نام A_m است. در بنیان روش درخت رگرسیون می‌توان رویه پاسخ را در ناحیه R_m توسط یک مدل ساده یعنی $\hat{y}_{R_m} = C_m$ تقریب نمود. در نتیجه خواهیم داشت:

$$\hat{y} = \sum_{m=1}^{5} C_m I_{A_m}(x_1, x_2) \qquad (۵-۲)$$

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

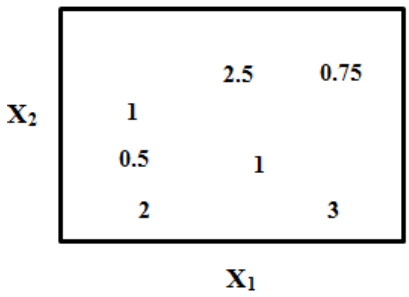
با توجه به روش کمترین توان‌های دوم خطا، به آسانی ثابت می‌شود که C_m، میانگین مقادیر متغیر پاسخ واقع در ناحیه R_m است و به عبارت دیگر:

$$\hat{y}_{R_m} = C_m = \bar{y}_{A_m} = \frac{1}{n_m} \sum_{y_j \epsilon A_m} yj \qquad (۶-۲)$$

که n_m تعداد مشاهدات مجموعه A_m است.

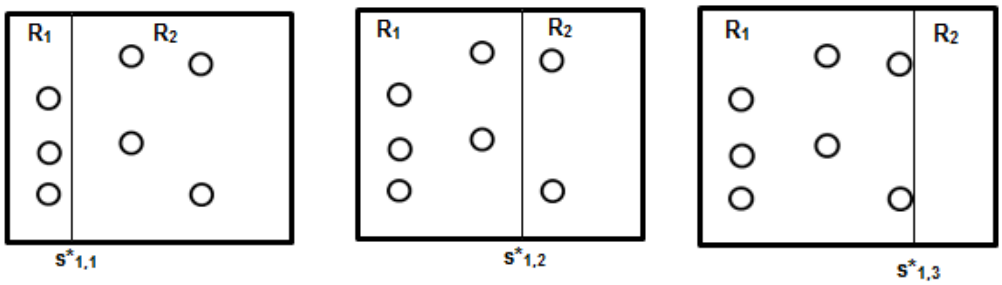
برای درک بهتر روش درخت رگرسیون این الگوریتم را با یک مثال ساده شرح می‌دهیم.

فرض کنید تعداد ۷ مشاهده با مقادیر پاسخ متفاوت برای دو متغیر X_1 وX_2 در اختیار باشد که نمودار پراکنش آنها در شکل (۲-۱۰) آمده است.

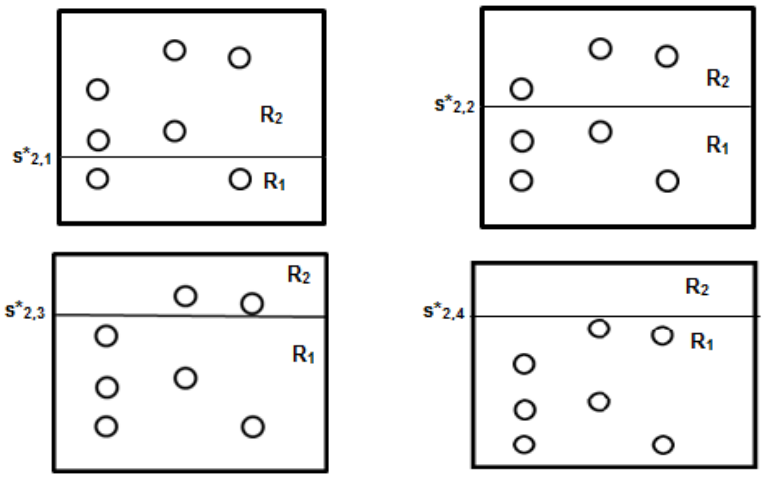


شکل(۲-۱۰)- نمودار پراکنش داده‌های دو متغیره

شکل های (۲-۱۱) و (۲-۱۲) کلیه افرازهای ممکن و مؤثر را به ترتیب در راستای محورهای X_1 و X_2 نشان می‌دهند و برای راحتی مقادیر پاسخ به شکل دایره نشان داده شده است.



شکل (۲-۱۱)- سه افراز ممکن در راستای متغیر X_1



شکل (۲-۱۲)- چهار افراز ممکن در راستای متغیر X_2

هر یک از مشاهدات در نواحی R_1 و R_2 دارای یک پاسخ (y_i) هستند، میانگین پاسخ در هر ناحیه نیز $\bar{y}$ می‌باشد. در نتیجه برای انتخاب اولین افراز، مقدار M که برابر با مجموع توان‌های دوم خطا در نواحی R_1 و R_2 است را در هر یک از شکل‌های بالا با استفاده از رابطه (۲-۷) بدست آورده و افرازی که دارای کمترین مقدار M باشد به عنوان بهترین افراز در مرحله اول انتخاب می‌شود.

$$M = \sum_{R1}(y_i - \bar{y})^2 + \sum_{R2}(y_i - \bar{y})^2 \qquad (۲-۷)$$

مقادیر M برای افرازهای ایجاد شده به صورت زیر خواهد بود:

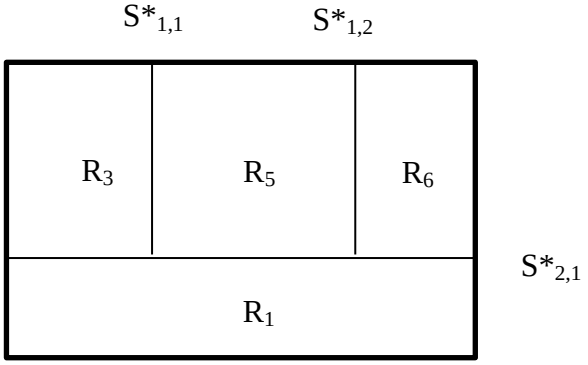
$M_{1,1}$ =۱/۱۶ + ۳/۶۷ = ۴/۸۳ $M_{1,2}$ =۲/۷ + ۲/۵۳ = ۵/۲۳

$M_{1,3}$ =۵/۵۵ + ۰ = ۵/۵۵

$M_{2,1}$ =۰/۵ + ۲/۴۵ = ۲/۹۵ $M_{2,2}$ =۳/۶۸ + ۱/۷۹ = ۵/۴۸

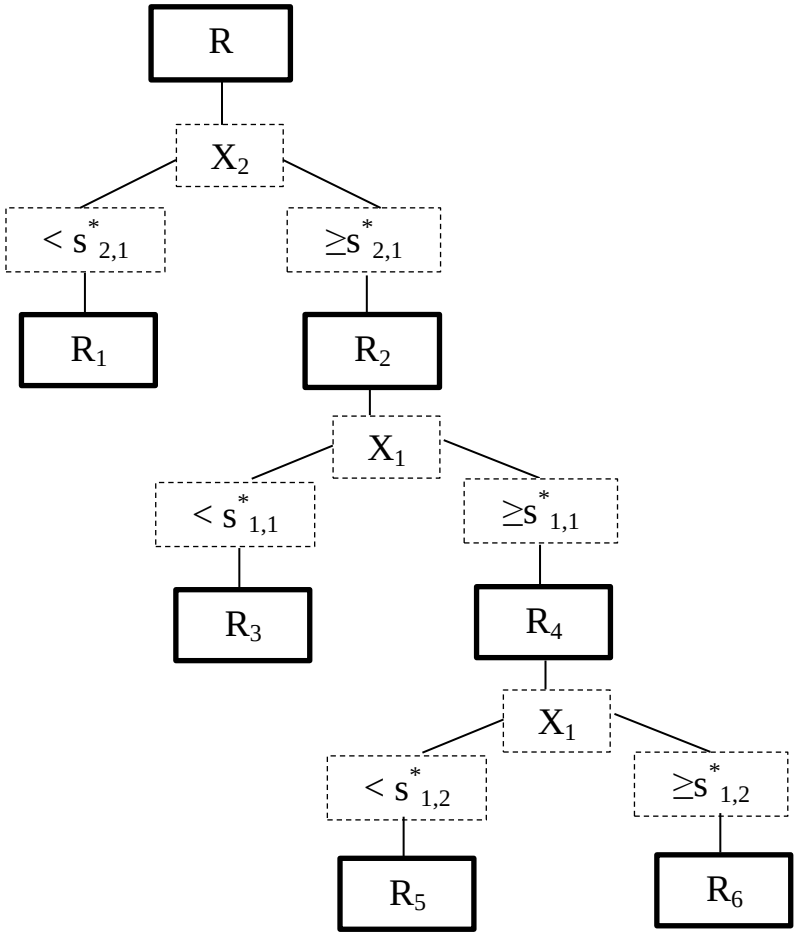
$M_{2,3}$ =۴ + ۱/۵۳ = ۵/۵۳ $M_{2,4}$ =۵/۵۵ + ۰ = ۵/۵۵

همانطور که مشاهده می‌کنید افراز در راستای متغیر X_2 و به ازای مقدار $S^*_{2,1}$ دارای کمترین مقدار M می‌باشد. اکنون با انتخاب اولین افراز فضای متغیرها به دو ناحیه تفکیک شده است. در مراحل بعدی فرآیند فوق در هر یک از نواحی تولید شده تکرار می‌شود تا در نهایت فضای متغیرهای توضیحی به صورت زیر تقسیم می‌گردد شکل (۲-۱۳).



شکل(۲-۱۳)- افراز فضای متغیرها

منظور از $S_{j,k}$ در واقع مقداری از هر متغیر است که افراز در آن نقطه صورت می‌گیرد و به آن گره گفته می‌شود. این افراز را می‌توان به صورت یک درخت در شکل (۲-۱۴) نشان داد.



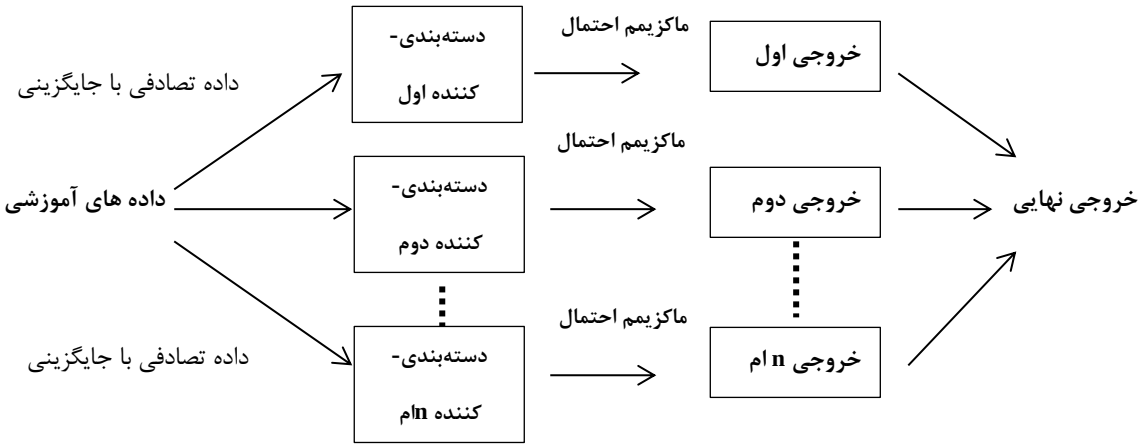
شکل(۲-۱۴)- ساختار درختی مدل درخت رگرسیون

۲-۵- جنگل‌های تصادفی(RF)

۲-۵-۱- تکنیک Bagging

جنگل تصادفی در واقع مجموعه‌ای از دسته‌بندی کننده‌های[1] مبتنی بر الگوریتم درخت تصمیم‌گیری و به طور خاص درخت تصمیم‌گیری CART محسوب می‌شود [۳۰]. عملکرد موازی دسته‌بندی کننده‌های درون بدنه جنگل و انتخاب ویژگی و داده آموزشی در هر یک از درخت‌ها به طور تصادفی، این دسته‌بندی کننده‌ها را بسیار قدرتمند ساخته است. طبق یک قاعده تقریباً عمومی، برآیند چندین دسته‌بندی کننده‌ی ضعیف و ساده می‌تواند دسته‌بندی کننده‌ای بسیار قوی و پیچیده باشد [۳۱]. با این رویکرد روش‌های بسیاری برای طراحی چنین دسته‌بندی کننده‌هایی پیشنهاد شده است. تکنیک Bagging یکی از متداول‌ترین نوع ساختارهای تجمعی به شمار می‌رود.

تکنیک Bagging ساختاری را می‌سازد که در آن دسته‌بندی کننده‌ها در کنار هم قرار می‌گیرند. ورودی در آن واحد به تمامی دسته‌بندی کننده‌ها داده می‌شود و هر دسته‌بندی کننده داده را مستقل از سایرین طبقه‌بندی می‌کند در نتیجه یک ساختار موازی به وجود می‌آید [۳۱]. این ساختار در شکل (۲-۱۵) نشان داده شده است.



شکل(۲-۱۵)- تکنیک مبتنی بر Bagging

1- Classifiers

۲-۵-۲- الگوریتم جنگل‌های تصادفی

جنگل تصادفی یک دسته‌بندی‌کننده تجمعی قدرتمند برگرفته از ساختار Bagging و شامل دسته‌بندی‌کننده‌های ضعیف درخت تصمیم‌گیری است. در الگوریتم‌های مبتنی بر تکنیک Bagging هر دسته‌بندی‌کننده مجموعه داده آموزشی مربوط به خود را استفاده می‌کند. بنابراین در جنگل تصادفی نیز هر یک از درختان حاضر در جنگل به طور مستقل و کاملاً تصادفی نمونه‌های آموزشی خود را از بین نمونه‌های موجود انتخاب می‌کنند. از طرفی هیچ مجموعه یا ترکیب ثابتی از ویژگی‌ها در کلاس‌بندی مورد استفاده قرار نمی‌گیرد. بلکه در تک‌تک گره‌های درخت تعدادی از ویژگی‌ها بصورت تصادفی انتخاب شده و یک ترکیب تصادفی از این متغیرها، در طبقه‌بندی داده‌ها استفاده خواهد شد. این موضوع در مورد تمامی درختان حاضر در جنگل صدق می‌کند [۳۲]. از آنجا که نمونه برداری داده‌ها با جایگزینی صورت می‌گیرد، ممکن است تعدادی از داده‌های آموزشی یک درخت تکراری باشند و در مقابل تعدادی از داده‌ها (به طور نسبی حدود $\frac{1}{3}$ مجموع داده‌ها) در آموزش درخت شرکت داده نشوند. دلیل عمده قدرت جنگل تصادفی همان انتخاب تصادفی داده‌ها و ویژگی‌هاست[۳۲]. با تولید ترکیبات تصادفی متعدد از ویژگی‌ها توسط درختان حاضر در جنگل و انتخاب ترکیبات تقریباً بهینه برای طبقه‌بندی امید این می‌رود که نهایتاً ویژگی‌های اصلی در مجموعه داده‌ها شناسایی شده و میزان تأثیر ویژگی‌های نامربوط در نتیجه کلاس‌بندی کاهش یابد.

این شیوه استفاده از بردار ویژگی و داده آموزشی، سبب می‌شود درختان جنگل، آموزش متفاوت، تصمیم‌گیری‌های متفاوت و در نتیجه کارآیی متفاوت داشته باشند.

فرض کنید (x_i, y_i) ; $i=1,\ldots,N$ مجموعه‌ی داده‌های آموزشی باشد که در آن X_i برداری از M متغیر و y_i متغیر پاسخ متناظر آن است. اگر تعداد کل درخت های مدل با n_{tree} نشان داده شود مراحل زیر بیانگر الگوریتم ساخت هر درخت است.

۱- با شروع آموزش جنگل، یک کپی از تمام داده‌های آموزشی به تمام درخت‌ها ارسال می‌شود سپس هر درخت حدود $\frac{2}{3}$ از داده‌های دریافتی را به صورت تصادفی برای آموزش انتخاب می‌کند و $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند که داده‌های خارج از کیسه[1] نام دارند و نقش داده‌های ارزیابی آن درخت را ایفا می‌کنند.

۲- برای تصمیم‌گیری درباره داده‌هایی که به یک گره می‌رسند مجموعه‌ای ازm متغیر به صورت تصادفی از بین M متغیر انتخاب می‌شود (m<<M) معمولاً در مدل رگرسیون $m = \frac{M}{3}$ و در مدل رده‌بندی $m = \sqrt{M}$ در نظر گرفته می‌شود. برای تقسیم فضای متغیرها به دو قسمت و بر اساس اصول درخت رگرسیونی فقط از این m متغیر و نمونهN تایی انتخاب شده در گام اول استفاده می‌شود تا بهترین متغیر و بهترین نقطه افراز در اولین مرحله بدست آید. محدودیت آن است که تعداد مشاهدات موجود در هر یک از دو ناحیه باید بیشتر از $2n_r$ باشد و $2n_r$ پارامتری است که در اختیار کاربر بوده و بیانگر حداقل تعداد مشاهدات موجود در هر ناحیه است و مقدار پیش‌فرض این پارامتر برای مسائل رگرسیون برابر ۵ و برای مسائل طبقه‌بندی برابر ۱ است.

۳- برای هر یک از دو ناحیه‌ی تولید شده در گام۲، مجدداً به طور تصادفی از بین M متغیر توضیحی، m متغیر انتخاب می‌گردد و با استفاده از همان نمونه Nتایی انتخاب شده در گام۱، روش درخت رگرسیون اعمال می‌گردد. این عمل منجر به افراز هر یک از نواحی موجود به دو قسمت می‌شود.

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

1- Out Of Bag

۵- حال درخت رگرسیونی iام طوری تشکیل شده است که فضای متغیرهای توضیحی به r_i ناحیه‌ی $R_{i1}, R_{i2}, \ldots R_{i\,ri}$ تقسیم گردیده است و تعداد مشاهدات هر ناحیه کمتر از $2n_r$ می‌باشد مدل درخت رگرسیون به دست آمده به صورت زیر است.

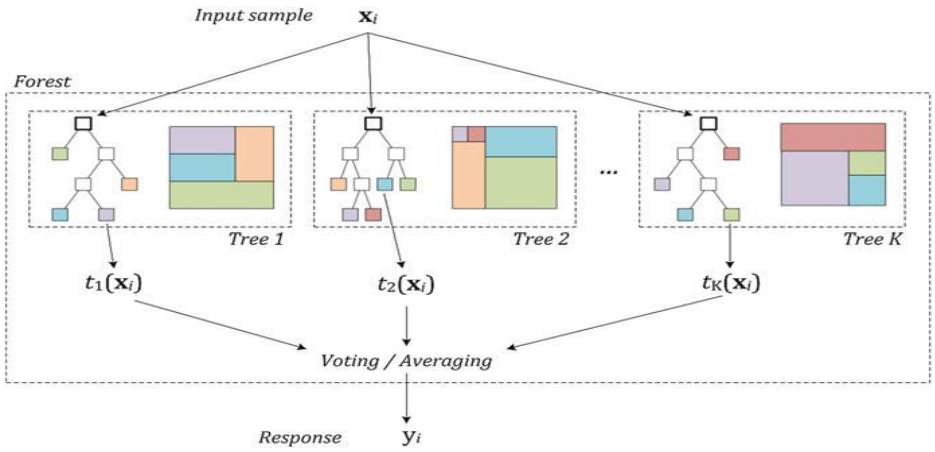
$$\widehat{f}_i(X) = \sum_{j=1}^{r_i} \hat{C}_j I_{R_{ij}}(X) \qquad (۸-۲)$$

که در آن $\hat{C}_j = \bar{y}_j$ و $I_{Rij}(X) = \begin{cases} 1 , & X \in R_{ij} \\ 0 , & X \notin R_{ij} \end{cases}$ می‌باشد

از آنجایی که در روشRF تعداد n_{tree} درخت رگرسیونی وجود دارد، می‌توان گفت که تعداد n_{tree} مدل به صورت معادله (۲-۸) خواهیم داشت. هنگامی که داده‌های جدید $T_{test} = \{X_k\}$ به جنگل ارائه می‌شود هر داده x_k همزمان در تمامی درخت‌ها پخش شده و با گذر از گره‌های میانی به گره نهایی خواهد رسید و میزان متغیر پاسخ برای این داده در مسائل رگرسیونی با میان‌گیری از مدل خروجی تمام درخت‌ها یعنی از رابطه (۲-۹) بدست می‌آید.

$$\hat{y}(x) = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} \hat{f}_i(x) \qquad (۹-۲)$$

در مسائل طبقه‌بندی احتمال تعلق به طبقه b در هر یک از درخت‌ها مقادیر متفاوتی را اختیار خواهد کرد که طبقه‌بندی نهائی داده با مکانیزم رأی‌گیری از نتیجه کلیه درخت‌ها صورت می‌گیرد. این فرآیند در شکل(۲-۱۶) نشان داده شده است[۳۳].



شکل(۲-۱۶)- مجموعه‌ای از درختان تصمیم‌گیری در یک مدل گروهی

۲-۵-۳- تنظیم پارامترهای آموزشی جنگل

پارامترهای دخیل در میزان کارآیی و دقت الگوریتم RF عمدتاً عبارتند از تعداد درختان تصمیم‌گیری، تعداد متغیرهای مورد استفاده در هر گره و ماکزیمم تعداد مجاز مشاهدات باقی مانده در هر گره. انتخاب مقادیر نامناسب برای تعداد متغیرهای طبقه‌بندی کننده می‌تواند به کاهش دقت منجر شود زیرا با افزایش متغیرها احتمال استفاده از متغیرهای تکراری و در نتیجه تصمیم‌گیری‌های تکراری افزایش خواهد یافت. با وجود این تاکنون هیچ تحقیق معتبری روی تنظیم پارامترهای جنگل انجام نشده و تعیین آنها مطابق روابط پیش‌فرض صورت می‌گیرد [۳۳ و۳۲].

متداول‌ترین مقداری که برای تعداد متغیرها در مسائل طبقه‌بندی در نظر گرفته می‌شود برابر مجذور تعداد کل ویژگی‌ها است و در مسائل رگرسیون $\frac{1}{3}$ تعداد کل متغیرها لحاظ می‌شود. ماکزیمم تعداد مجاز مشاهده باقیمانده در هر گره در مسائل طبقه‌بندی به طور پیش فرض ۱ در نظر گرفته می‌شود یعنی هیچ دو داده‌ای در یک گره باقی نمی‌مانند و تصمیم‌گیری به سطح دیگری منتقل می‌شود و در مسائل رگرسیونی به طور پیش‌فرض این مقدار برابر با ۵ است.

۲-۵-۴- تعیین اهمیت متغیرها در روش جنگل‌های تصادفی

در ساختار روش RF امکانی وجود دارد که میزان اهمیت متغیرها را در مدل، تعیین نموده و متغیرهایی که دارای نقش بیشتری در هر درخت و در مدل نهایی هستند شناسایی می‌شود. همان‌طور که اشاره شد، داده‌هایی که در ساخت درخت شرکت نمی‌کردند، داده‌های خارج از کیسه (OOB) نامیدیم که به نوعی نقش داده‌های آزمایشی را برای ارزیابی آن درخت ایفا می‌کنند، فرض کنید خطای پیش‌بینی Y برای داده‌های OOB برای درخت i ام با نماد EOOBi نشان داده شود. برای تعیین اهمیت متغیر j ام (xj)، مقادیر این متغیر به طور تصادفی، چندین مرتبه (تعداد مراتب، پارامتری است که در اختیار کاربر می‌باشد) جابه‌جا کرده و مجدداً خطای پیش‌بینی را برای مشاهدات OOB

محاسبه می‌شود که مقدار این خطا با نماد $\widetilde{\mathrm{EOOB}}_i$ نشان داده می‌شود. میزان اهمیت متغیر jام حول درخت iام را با $\mathrm{VI_i}(x^j)$ نشان داده و به صورت زیر تعریف می‌گردد.

$$\mathrm{VI_i}\left(x^j\right) = \widetilde{EOOB}_i^j - EOOB_i \qquad (۱۰-۲)$$

اگر x_j یک متغیر وابسته به پاسخ و مهم باشد میزان $\widetilde{EOOB}$ افزایش خواهد یافت و صحت پیش بینی برای مشاهدات OOB کاهش می‌یابد. سرانجام میزان اهمیت متغیر jام در مدل نهاییRF با میان‌گیری از اهمیت محاسبه شده توسط تک‌تک درختان به صورت زیر محاسبه می‌شود.

$$VI(x^J) = \frac{1}{ntree}\sum_{i=1}^{ntree}(VI_i(x^J)) \qquad (۱۱-۲)$$

۲-۶- ارزیابی مدل

منظور از ارزیابی مدل، ارائه شاخص‌های کمی است، به وسیله این شاخص‌ها صحت نتایج ارائه شده توسط مدل مورد سنجش قرار می‌گیرند. این شاخص‌ها عبارتند از:

ضریب همبستگی[1]: ساده‌ترین راه برای بررسی میزان همبستگی دو یا چند متغیر، محاسبه آماره‌ی ضریب همبستگی آنهاست. ضریب همبستگی دو متغیرx,y با رابطه (۲- ۱۲) تعریف می شود.

$$R = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \qquad (۱۲-۲)$$

ضریب تعیین[2]: ضریب تعیین به عنوان یک شاخص برای بیان دقت خط رگرسیون برآورد شده، به کار می‌رود و نشان‌دهنده نسبت تغییرات متغیر وابسته توسط متغیر مستقل است. رابطه ریاضی مربوط به ضریب تعیین به صورت زیر است:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (۱۳-۲)$$

1 - Correlation coefficient
2 - Determination coefficient

که SSR[1] طبق رابطه (۲-۱۴) بیانگر مجموع مربعات انحراف مقادیر پیش‌بینی شده‌ی متغیر وابسته از میانگین مقادیر آن است.

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

SST[2] طبق رابطه (۲-۱۵) نشانگر مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از میانگین مقادیر آن است.

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

SSE[3] نیز مبین مجموع مربعات انحراف مقادیر واقعی متغیر وابسته از مقادیر پیش‌بینی شده برای آن است.

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

بنابراین با توجه به روابط فوق می‌توان نوشت:

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

طبق رابطه (۲-۱۷) 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.

ضریب تعیین تصحیح شده[4]: ضریب تعیین تصحیح شده یکی از پارامترهایی است که در ارزیابی و مقایسه مدل‌های مختلف با تعداد متفاوت متغیر استفاده می‌شود و از آنجایی که R^2 (ضریب تعیین)،

1 - Sum Square Regression
2- Sum Square Total
3 - Sum Square Error
4 - Adjusted determination coefficient

تعداد پارامترهای موجود در مدل را به حساب نمی‌آورد، بدین منظور از ضریب تعیین تصحیح شده طبق رابطه زیر برای انتخاب مدل برتر استفاده می‌شود:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1} \cdot \frac{SSE}{SST} = 1 - (1-R^2)\frac{n-1}{n-p-1} \qquad (۱۸-۲)$$

که در این رابطه p تعداد متغیرهای مستقل و n تعداد ترکیبات مورد بررسی می‌باشد.

مجموع مربع باقیمانده‌ها[1] (PRESS): برابر مجموع مربعات تفاوت بین مقدار کمیت مشاهده شده (y_i) و مقدار تخمین زده شده $(\hat{y}_i)$ است:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (۱۹-۲)$$

خطای استاندارد پیش‌بینی[2] (SEP)

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \qquad (۲۰-۲)$$

خطای مطلق میانگین[3] (MAE)

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}|}{n} \qquad (۲۱-۲)$$

خطای نسبی پیش بینی[4] (REP)

$$REP(\%) = \frac{100}{\overline{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \qquad (۲۲-۲)$$

میانگین مربع خطاها[5] (MSE)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} \qquad (۲۳-۲)$$

1 - Predictive Residual Sum of Squares
2 - Standard Error of Prediction
3 - Mean Absolute Error
4 - Relative Error of Prediction
5 - Mean Square Error

میانگین خطای نسبی[1] (MRE)

$$MRE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}}{y_i} \right|}{n} \times 100 \qquad (۲۴-۲)$$

در مدل‌سازی فعالیت ترکیبات دارویی، بعضی از مشتقات که به طور تجربی سنتز شده‌اند، فاقد فعالیت دارویی (غیرفعال) می‌باشند. لذا لازم است مدلی ارائه شود که قادر باشد ترکیبات را به دو دسته فعال و غیرفعال تقسیم‌بندی نماید. در ارزیابی جنگل‌های تصادفی به کار گرفته‌شده در طبقه‌بندی ترکیبات، به دو دسته فعال و غیرفعال، اگر N_{TP}، تعداد ترکیبات فعال که مدل آن‌ها را درست دسته‌بندی کرده است، N_{TN} تعداد ترکیبات غیرفعالی که مدل آن‌ها را درست دسته‌بندی کرده است، N_{FP} تعداد ترکیباتی که مدل به غلط آن ها را فعال شناخته است و N_{FN} تعداد ترکیباتی که مدل به غلط آن‌ها را غیرفعال شناخته است، باشند با استفاده از این کمیت‌ها، می‌توان شاخص‌های کمی ارزیابی را مطابق روابط زیر محاسبه کرد [۳۴].

نسبت صحیح[2] (FC) : درصدی از کل نمونه‌ها که به درستی دسته‌بندی شده‌اند.

$$FC = \frac{N_{TN}+N_{TP}}{N_{total}} \times 100 \qquad (۲۴-۲)$$

نسبت هشدار[3] (FAR) : درصدی از ترکیبات غیرفعال است که نادرست دسته‌بندی شده‌اند.

$$FAR = \frac{N_{FP}}{N_{TN}+N_{TP}} \times 100 \qquad (۲۵-۲)$$

احتمال آشکارسازی[4](POD): درصدی از ترکیبات است که به درستی دسته‌بندی شده‌اند.

$$POD = \frac{N_{TP}}{N_{FN}+N_{TP}} \times 100 \qquad (۲۶-۲)$$

بر این اساس بهترین مدل آن است که بیشترین نسبت صحیح و احتمال آشکارسازی و کم‌ترین نسبت هشدار را دارا باشد.

1 - Mean Relative Error
2 - Fraction Correct
3 - False Alarm Rate
4- Detection Of Probability

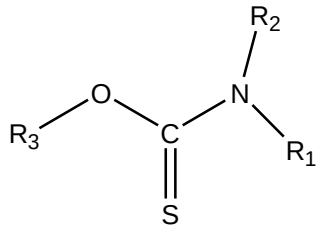
فصل سوم

مطالعه ارتباط کمی ساختار- فعالیت مشتقات تیوکربامات به عنوان دسته‌ی جدیدی از بازدارنده‌های غیر نوکلئوزیدی HIV با استفاده از روش جنگل-های تصادفی و توصیف‌گرهای سهم گروه

در راستای مطالعه ارتباط کمی ساختار فعالیت ضد ایدز گروهی از مشتقات تیوکربامات، تعداد ۱۹۷ ترکیب از آن‌ها انتخاب و مدل‌سازی با استفاده از روش جنگل‌های تصادفی انجام گرفت. در این پژوهش توصیف‌گرها با استفاده از روش سهم گروه محاسبه شده‌اند که در این بخش به معرفی سری داده‌ها، محاسبه توصیف‌گرهای سهم گروه، مدل‌سازی جنگل‌های تصادفی و ارزیابی مدل خواهیم پرداخت.

۳-۱- سری داده‌ها

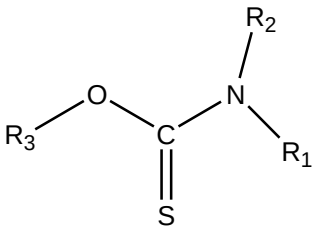
سری داده‌ها از نتایج تجربی ۱۹۷ ترکیب از مشتقات تیوکربامات می‌باشد. اسکلت اصلی این ترکیبات در شکل (۳-۱) نشان داده شده است. جدول (۳-۱) استخلاف‌های مختلف تمامی این ترکیبات را نشان می‌دهد. مقادیر pEC50 این ترکیبات (EC_{50}[1] بر حسب مولار می‌باشد) که از مراجع [۳۸-۳۵] استخراج گردیده‌اند در جدول (۳-۱) آمده است. مقادیر EC_{50} بایستی در شرایط یکسان اندازه‌گیری شود و EC_{50} غلظتی از دارو می‌باشد که تا ۵۰٪ سلول MT-4 را از HIV محافظت می‌کند[۳۷].

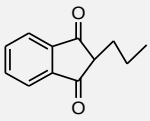
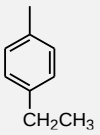
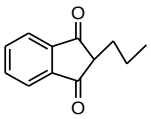
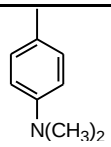
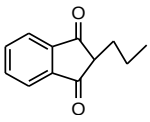
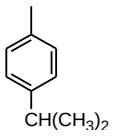
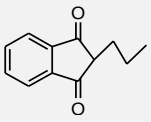
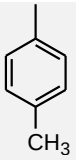
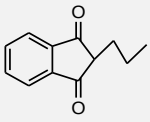
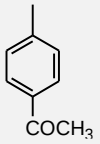
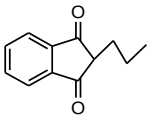
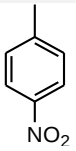
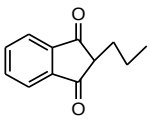
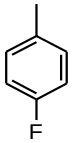
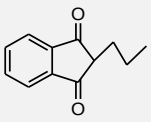
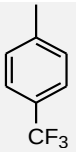
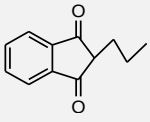
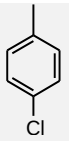
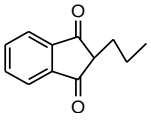
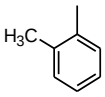
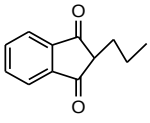
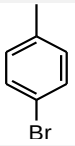
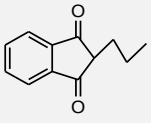
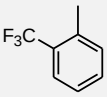
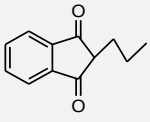
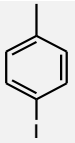
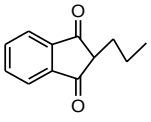
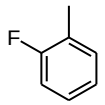
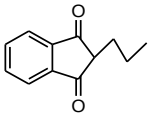
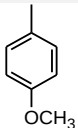
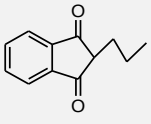
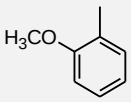
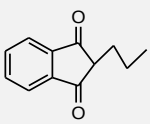
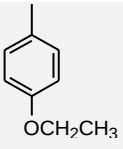
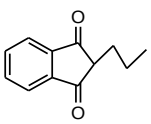
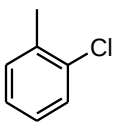


شکل (۳-۱)- اسکلت اصلی ترکیبات مورد مطالعه

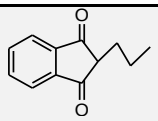
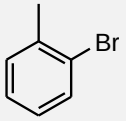
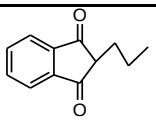
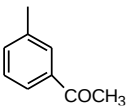
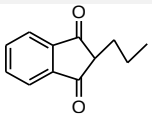
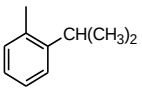
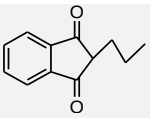
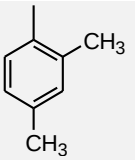
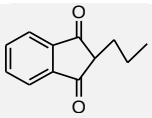
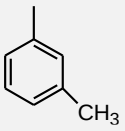
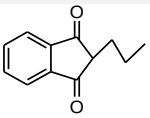
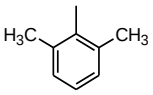
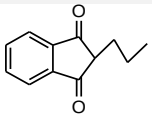
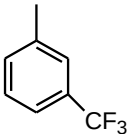
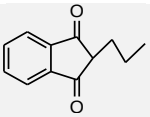
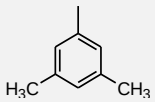
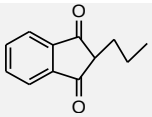
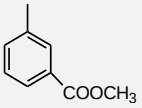
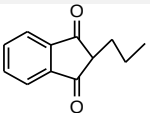
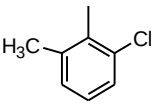
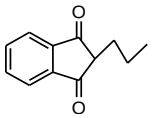
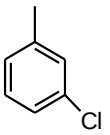
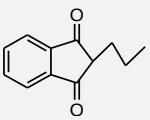
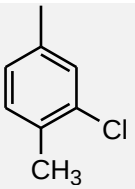
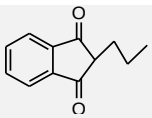
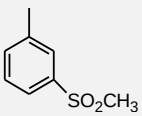
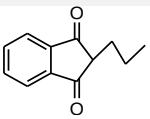
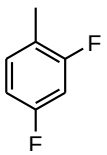
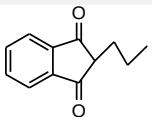
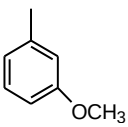
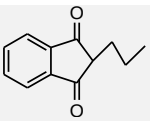
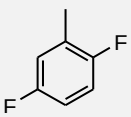
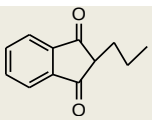
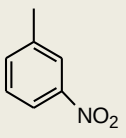
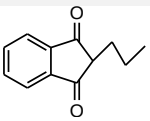
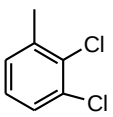
1- Effective Concentration

جدول (۳-۱)- ترکیبات مورد مطالعه

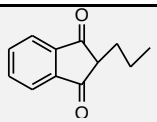
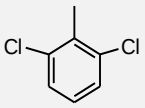
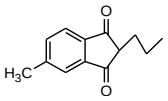
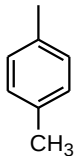
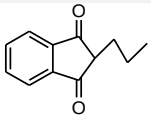
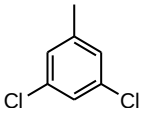
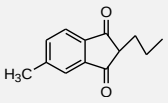
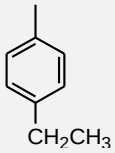
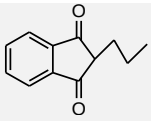
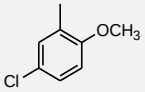
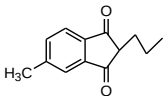
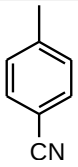
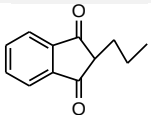
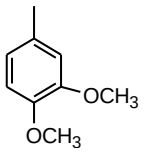
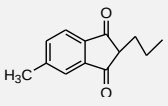
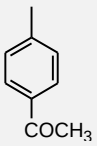
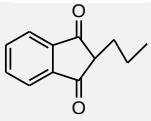
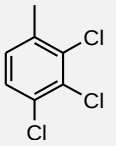
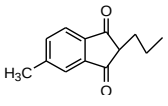
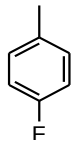
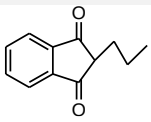
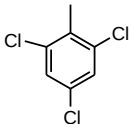
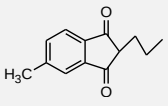
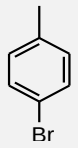
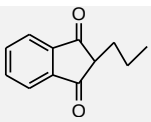
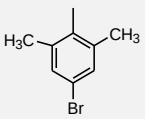
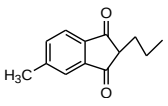
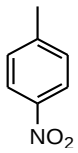
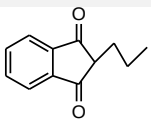
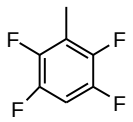
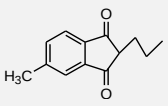
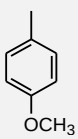
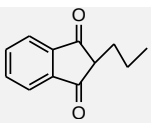
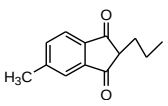
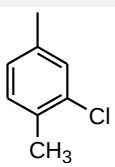


No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۱			H	۷/۱۰	۱۰			H	۶/۲۲
۲			H	۷/۴۰	۱۱			H	۷/۷۰
۳			H	۶/۸۵	۱۲			H	۷/۴۰
۴			H	۷/۰۰	۱۳			H	۷/۴۰
۵			H	۷/۴۰	۱۴			H	۵/۴۳
۶			H	۷/۵۲	۱۵			H	۴/۱۹
۷			H	۷/۷۰	۱۶			H	۵/۵۲
۸			H	۷/۵۲	۱۷			H	۵/۲۵
۹			H	۵/۱۵	۱۸			H	۴/۷۷

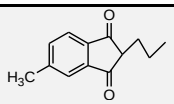
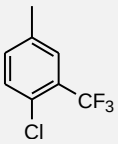
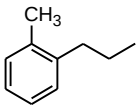
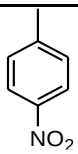
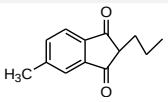
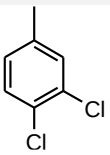
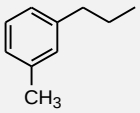
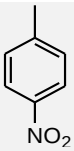
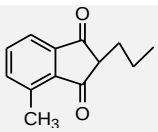
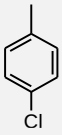
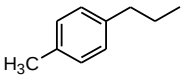
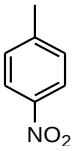
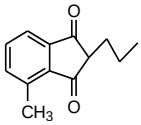
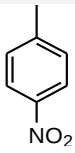
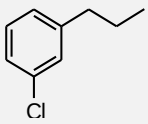
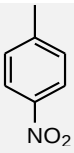
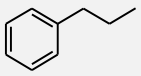
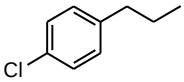
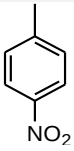
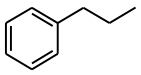
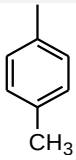
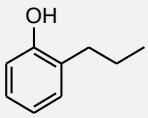
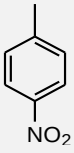
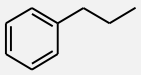
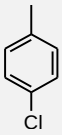
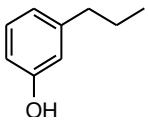
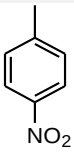
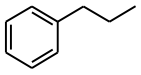
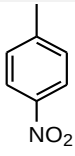
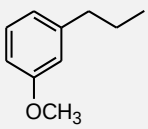
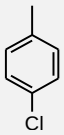
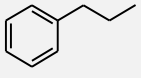
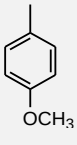
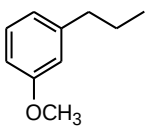
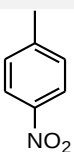
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۱۹			H	۴/۸۵	۲۸			H	۵/۳۶
۲۰			H	۴/۳۶	۲۲۹			H	۵/۲۲
۲۱			H	۵/۷۵	۳۰			H	۴/۱۲
۲۲			H	۴/۸۷	۳۱			H	۴/۹۲
۲۳			H	۴/۶۸	۳۲			H	۴/۶۴
۲۴			H	۵/۷۰	۳۳			H	۶/۱۰
۲۵			H	۵/۱۵	۳۴			H	۶/۳۰
۲۶			H	۵/۳۹	۳۵			H	۵/۱۰
۲۷			H	۶/۲۲	۳۶			H	۴/۵۲

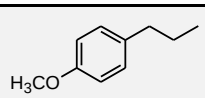
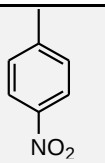
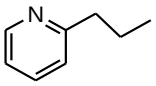
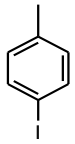
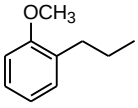
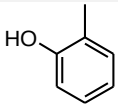
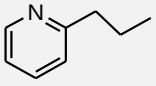
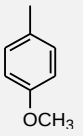
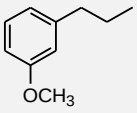
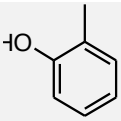
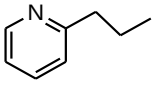
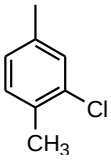
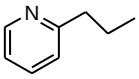
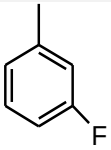
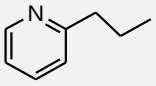
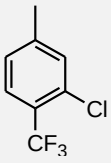
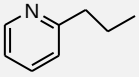
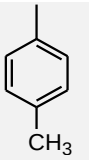
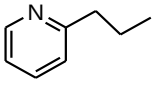
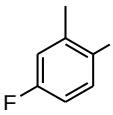
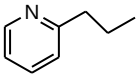
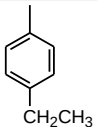
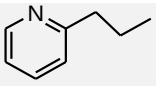
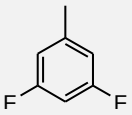
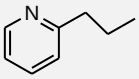
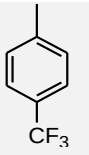
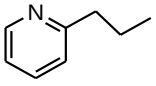
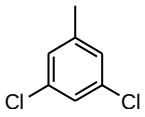
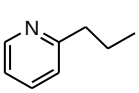
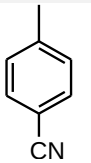
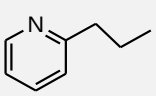
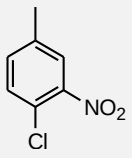
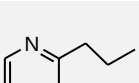
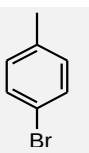
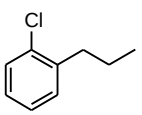
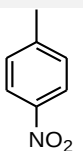
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۳۷			H	۴/۶۸	۴۶			H	۷/۷۰
۳۸			H	۵/۴۰	۴۷			H	۷/۷۰
۳۹			H	۴/۳۲	۴۸			H	۸
۴۰			H	۵/۴۵	۴۹			H	۷/۱۵
۴۱			H	۴/۹۶	۵۰			H	۷/۵۲
۴۲			H	۵/۰۰	۵۱			H	۸/۱۰
۴۳			H	۵/۰۵	۵۲			H	۸/۰۰
۴۴			H	۵/۰۰	۵۳			H	۷/۵۲
۴۵			H	۵/۹۲	۵۴			H	۶/۴۰

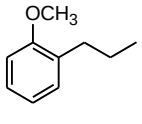
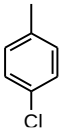
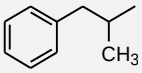
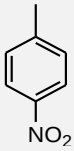
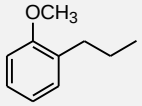
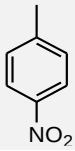
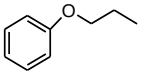
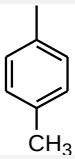
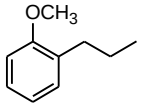
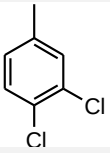
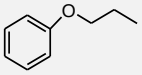
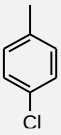
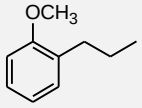
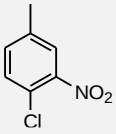
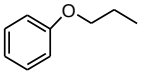
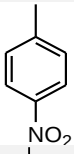
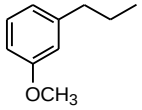
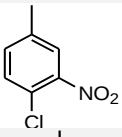
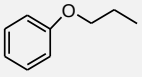
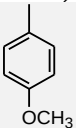
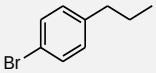
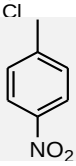
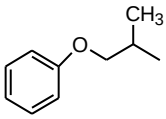
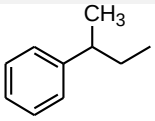
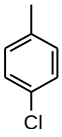
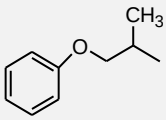
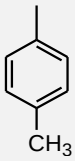
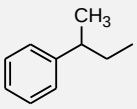
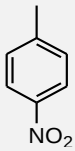
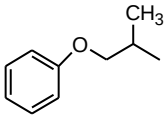
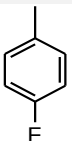
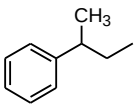
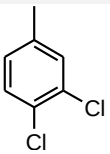
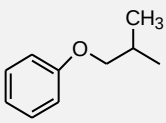
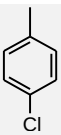
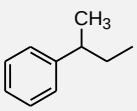
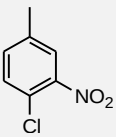
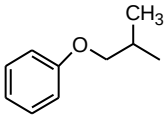
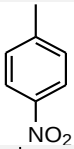
دامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۵۵			H	۶/۳۰	۶۴			H	۶/۱۵
۵۶			H	۶/۷۰	۶۵			H	۵/۷۷
۵۷			H	۷/۳۰	۶۶			H	۶/۵۲
۵۸			H	۷/۱۵	۶۷			H	۵/۸۲
۵۹			H	۵/۲۲	۶۸			H	۴/۸۹
۶۰			H	۵/۳۷	۶۹			H	۵/۰۰
۶۱			H	۶/۳۰	۷۰			H	۵/۴۳
۶۲			H	۶/۰۵	۷۱			H	۷/۰۹
۶۳			H	۴/۷۷	۷۲			H	۶/۷۰

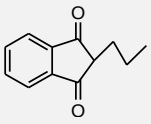
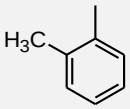
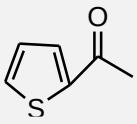
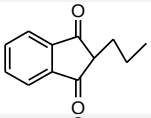
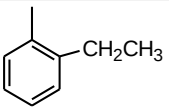
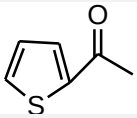
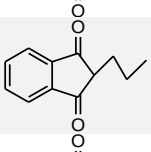
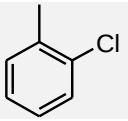
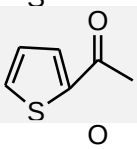
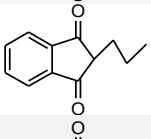
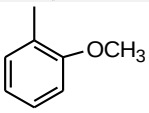
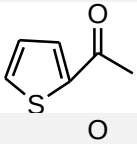
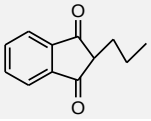
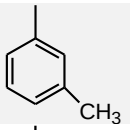
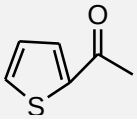
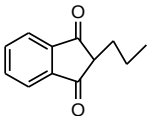
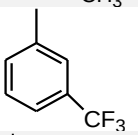
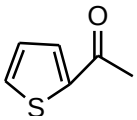
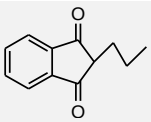
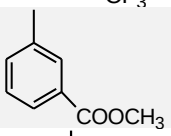
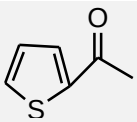
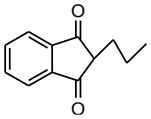
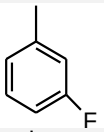
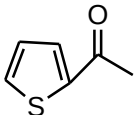
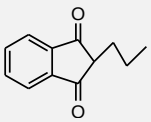
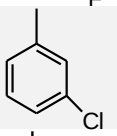
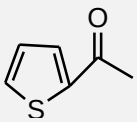
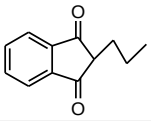
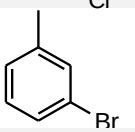
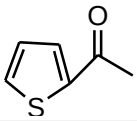
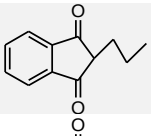
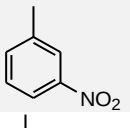
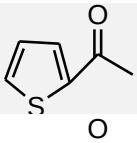
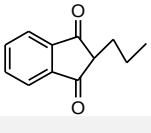
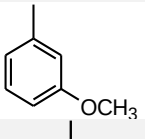
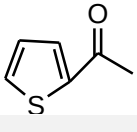
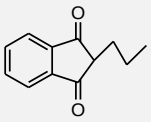
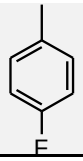
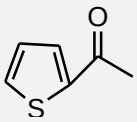
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۷۳			H	۵/۸۵	۸۲			H	۶/۳۰
۷۴			H	۶/۵۲	۸۳			H	۵/۲۵
۷۵			H	۶/۰۰	۸۴			H	۶/۱۵
۷۶			H	۵/۹۵	۸۵			H	۵/۲۶
۷۷			H	۶/۳۰	۸۶			H	۵/۰۵
۷۸			H	۵/۰۰	۸۷			H	۵/۰۵
۷۹			H	۴/۸۹	۸۸			H	۵/۵۳
۸۰			H	۶/۷۰	۸۹			H	۷/۰۰
۸۱			H	۶/۷۰	۹۰			H	۶/۴۰

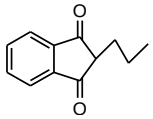
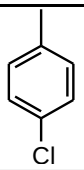
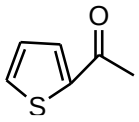
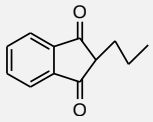
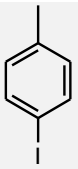
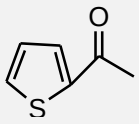
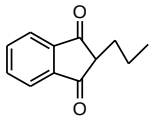
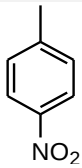
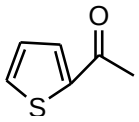
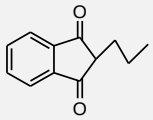
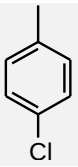
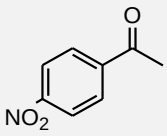
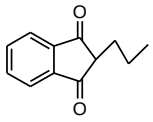
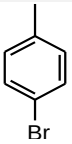
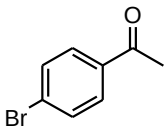
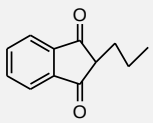
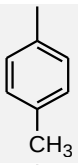
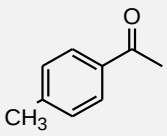
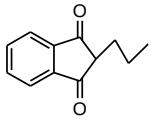
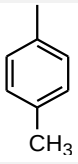
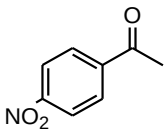
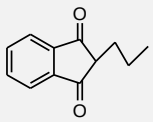
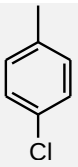
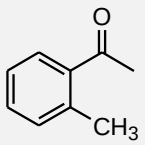
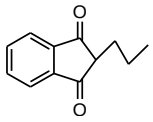
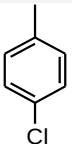
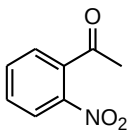
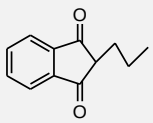
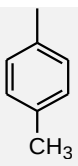
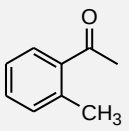
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}	No	R_3	R_2	R_1	pEC_{50}
۹۱			H	۷/۱۵	**۱۰۲**			H	۵/۸۸
۹۲			H	۷/۰۰	**۱۰۳**			H	۵/۰۰
۹۳			H	۶/۵۲	**۱۰۴**			H	۵/۵۳
۹۴			H	۶/۰۴	**۱۰۵**			H	۵/۰۷
۹۵			H	۵/۷۰	**۱۰۶**			H	۴/۷۷
۹۶			H	۵/۱۰	**۱۰۷**			H	۵/۶۹
۹۷			H	۶/۷۰	**۱۰۸**			H	۵/۹۲
۹۸			H	۶/۲۲	**۱۰۹**			H	**۴/۹۶**
۹۹			H	۶/۵۲	**۱۱۰**			H	۵/۸۰
۱۰۰			H	۶/۳۰	**۱۱۱**			H	**۵/۵۸**
۱۰۱			H	۵/۰۴	**۱۱۲**			H	۵/۳۹

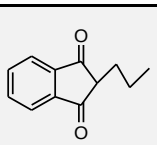
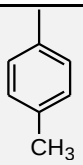
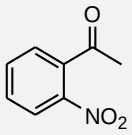
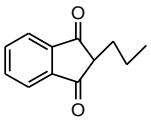
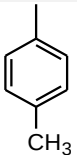
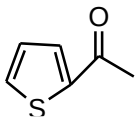
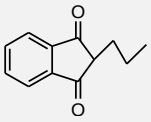
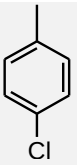
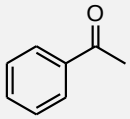
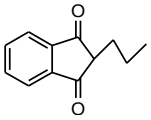
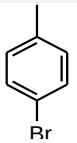
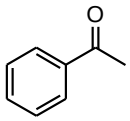
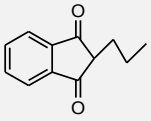
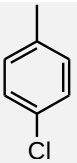
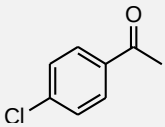
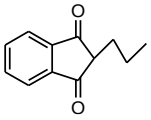
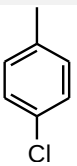
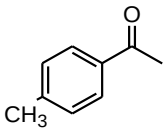
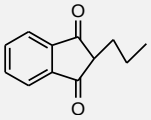
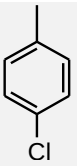
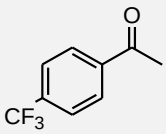
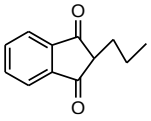
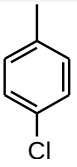
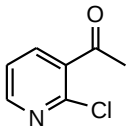
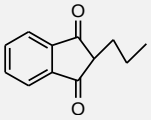
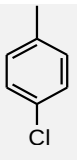
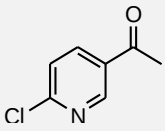
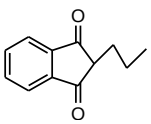
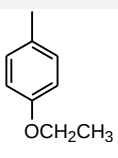
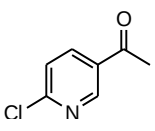
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۱۳				۴/۹۶
۱۱۴				۴/۹۶
۱۱۵				۴/۹۶
۱۱۶				۴/۹۶
۱۱۷				۵/۰۰
۱۱۸				۴/۹۶
۱۱۹				۵/۲۲
۱۲۰				۴/۹۶
۱۲۱				۵/۲۲
۱۲۲				۵/۹۲
۱۲۳				۶/۴۲
۱۲۴				۵/۴۶
۱۲۵				۷/۰۰

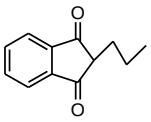
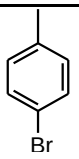
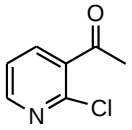
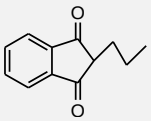
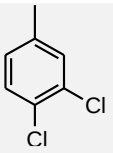
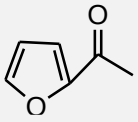
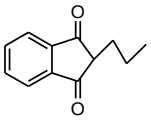
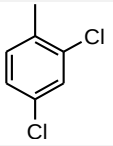
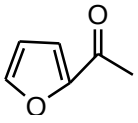
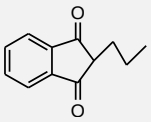
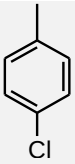
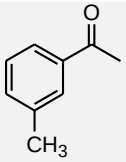
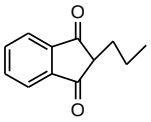
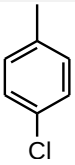
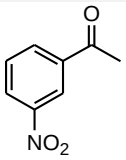
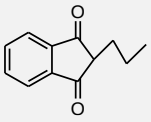
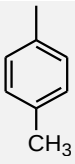
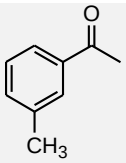
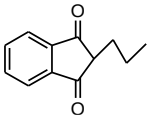
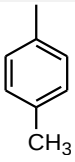
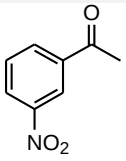
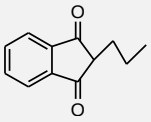
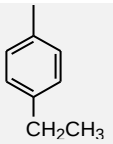
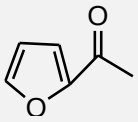
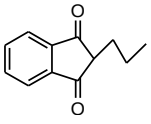
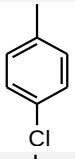
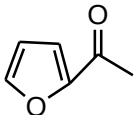
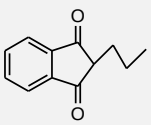
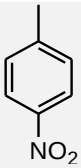
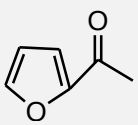
ادامه جدول (۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۲۶				۸/۱۰
۱۲۷				۸/۰۰
۱۲۸				۸/۰۰
۱۲۹				۷/۲۲
۱۳۰				۷/۴۶
۱۳۱				۷/۷۰
۱۳۲				۷/۵۲
۱۳۳				۷/۴۰
۱۳۴				۷/۱۵
۱۳۵				۶/۷۰

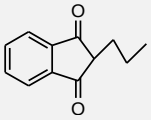
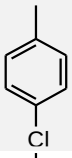
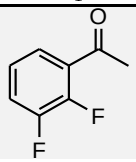
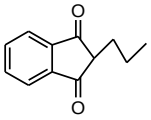
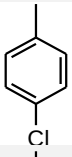
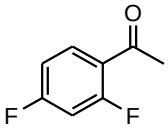
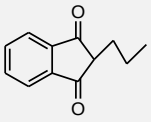
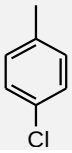
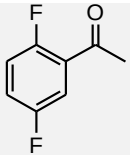
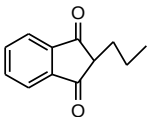
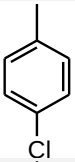
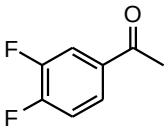
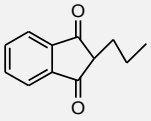
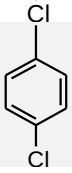
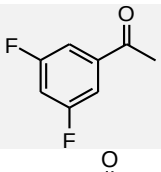
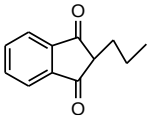
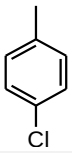
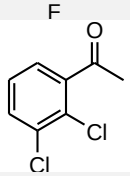
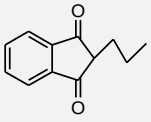
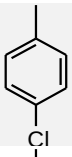
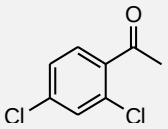
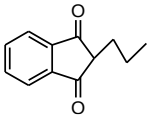
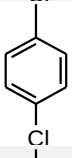
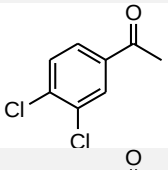
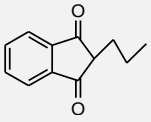
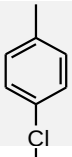
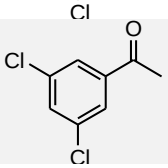
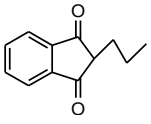
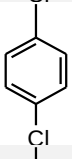
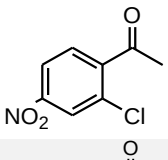
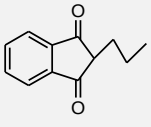
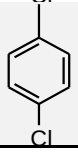
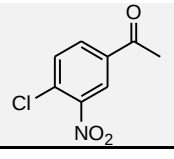
ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۳۶				۶/۴۰
۱۳۷				۶/۳۹
۱۳۸				۷/۰۰
۱۳۹				۷/۴۶
۱۴۰				۷/۶۰
۱۴۱				۷/۳۰
۱۴۲				۸/۳۰
۱۴۳				۸/۳۰
۱۴۴				۸/۲۲
۱۴۵				۶/۳۰

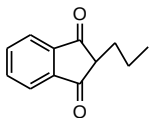
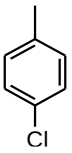
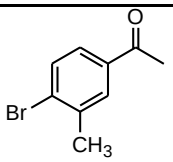
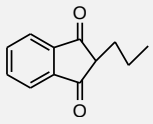
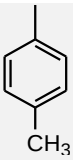
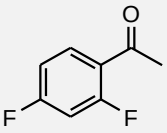
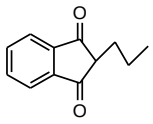
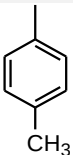
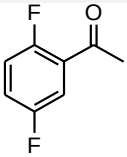
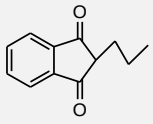
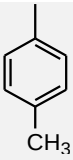
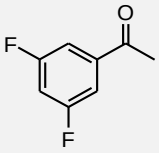
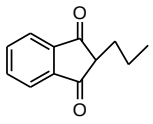
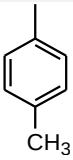
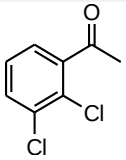
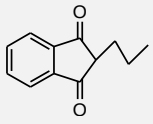
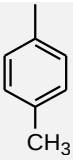
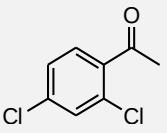
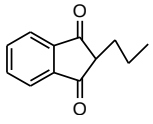
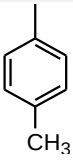
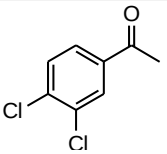
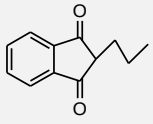
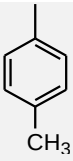
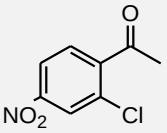
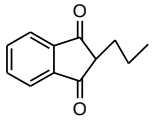
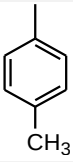
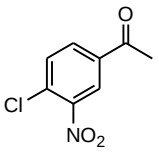
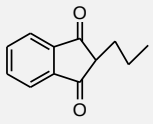
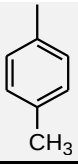
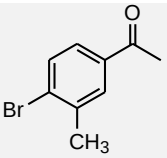
ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۴۶				۶/۷۰
۱۴۷				۶/۱۶
۱۴۸				۵/۵۹
۱۴۹				۸/۰۰
۱۵۰				۸/۸۲
۱۵۱				۷/۴۰
۱۵۲				۷/۷۰
۱۵۳				۷/۵۲
۱۵۴				۸/۱۵
۱۵۵				۸/۱۰

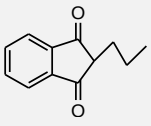
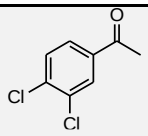
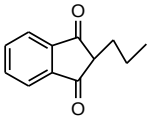
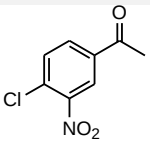
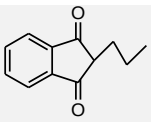
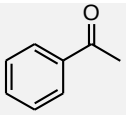
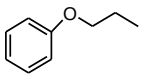
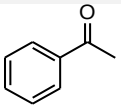
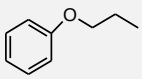
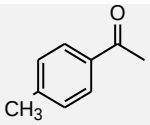
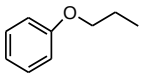
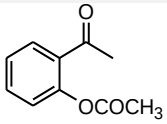
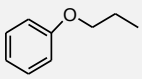
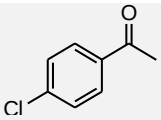
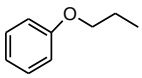
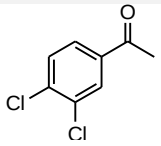
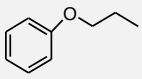
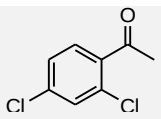
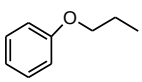
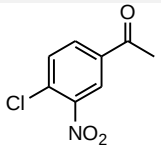
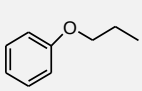
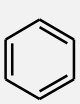
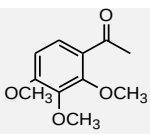
ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۵۶				۸/۱۰
۱۵۷				۸/۰۰
۱۵۸				۷/۷۰
۱۵۹				۸/۱۰
۱۶۰				۸/۱۵
۱۶۱				۷/۰۰
۱۶۲				۷/۵۲
۱۶۳				۸/۰۵
۱۶۴				۸/۱۰
۱۶۵				۸/۱۰
۱۶۶				۷/۷۰

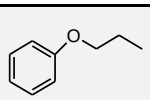
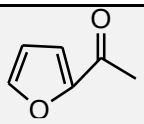
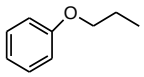
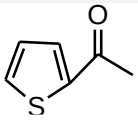
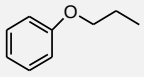
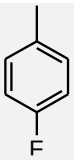
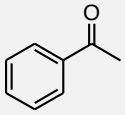
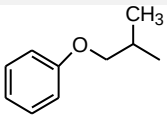
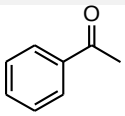
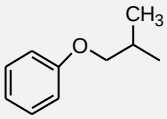
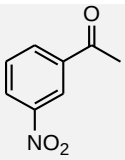
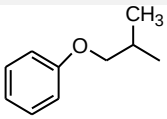
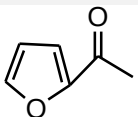
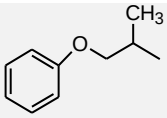
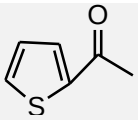
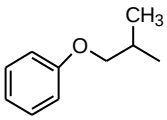
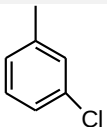
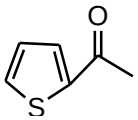
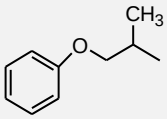
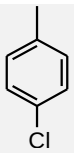
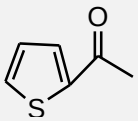
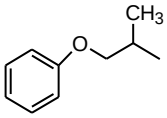
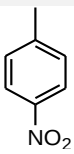
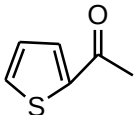
ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۶۷				۸/۱۵
۱۶۸				۷/۵۲
۱۶۹				۷/۴۰
۱۷۰				۷/۳۰
۱۷۱				۶/۷۰
۱۷۲				۶/۵۲
۱۷۳				۷/۳۰
۱۷۴				۷/۰۵
۱۷۵				۷/۲۲
۱۷۶				۷/۳۰

ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۷۷				۶/۳۰
۱۷۸				۶/۴۶
۱۷۹				۶/۴۰
۱۸۰				۵/۰۹
۱۸۱				۵/۰۰
۱۸۲				۵/۰۲
۱۸۳				۴/۹۸
۱۸۴				۴/۹۳
۱۸۵				۵/۰۵
۱۸۶				۵/۱۲
۱۸۷				۵/۰۱

ادامه جدول(۳-۱)

No	R_3	R_2	R_1	pEC_{50}
۱۸۸				۵/۰۷
۱۸۹				۵/۰۶
۱۹۰				۵/۴۰
۱۹۱				۵/۸۸
۱۹۲				۵/۴۴
۱۹۳				۵/۸۸
۱۹۴				۵/۷۰
۱۹۵				۴/۹۶
۱۹۶				۴/۹۶
۱۹۷				۵/۲۲

۳-۲- نرم‌افزار مورد استفاده

متلب[1] یک زبان سطح بالا و با محیطی جذاب می‌باشد، که در ابتدا براساس زبان برنامه نویسی C توسعه داده شد. واژه متلب، هم به معنی محیط محاسبات رقمی و هم به معنی خود زبان برنامه‌نویسی مربوطه است که از ترکیب دو واژه MATrix (ماتریس) و LABoratory (آزمایشگاه) ایجاد شده‌است. این نام حاکی از رویکرد ماتریس‌محور، برنامه است، که در آن حتی اعداد منفرد هم به عنوان **ماتریس** در نظر گرفته می‌شوند. متلب امکان آسان عملیات ماتریسی، محاسباتی و توابعی، استفاده از الگوریتم‌های مختلف و همچنین امکان ارتباط آسان با زبان‌های مختلف برنامه‌نویسی را به کاربر می‌دهد. از محیط برنامه متلب، برای نوشتن الگوریتم جنگل‌های تصادفی جهت انجام مدل‌سازی مطالعات کمی ساختار-فعالیت استفاده گردیده است.

۳-۳- محاسبه و استخراج توصیف‌گرهای مولکولی

برای محاسبه توصیف‌گرهایی که در عین سادگی و قابل تفسیر بودن، بتوانند کل ویژگی‌های ترکیبات مورد بررسی را نشان دهند و به‌علاوه به‌راحتی قابل محاسبه باشند، چهار دسته توصیف‌گر با در نظر گرفتن ساختار تمام ۱۹۷ ترکیب تعریف گردید. این توصیف‌گرها عبارتند از:

دسته اول: توصیف‌گرهای قطعات مولکولی

دسته دوم: توصیف‌گرهای جایگاه (محل) استخلاف

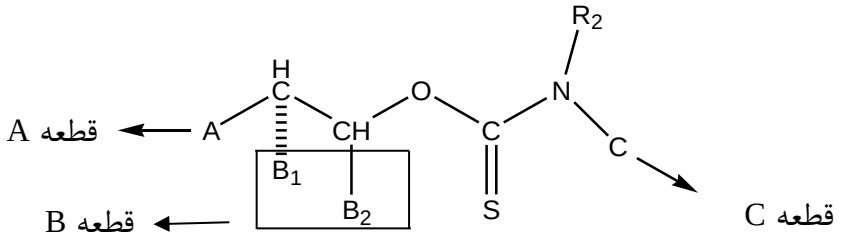
دسته سوم: توصیف‌گرهای ماهیت استخلاف

دسته چهارم: توصیف‌گرهای مربوط به پیوند هیدروژنی و تعداد حلقه‌های بنزن

1- MATLAB

۳-۳-۱-توصیف‌گرهای قطعات مولکولی

توصیف‌گرهای این دسته نمایانگر قطعات اصلی مولکول می‌باشند و با در نظر گرفتن ماهیت استخلاف‌های R_1 و R_3 در شکل (۳-۱) در سه گروه مختلف A, B, C مطابق شکل (۳-۲) تعریف شدند. این دسته از توصیف‌گرها در مجموع مشتمل بر ۱۲ توصیف‌گر (۶ توصیف‌گر از نوع A، ۲ توصیف‌گر از نوع B و ۴ توصیف‌گر از نوع C) می‌باشند.

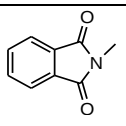
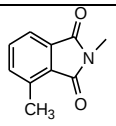
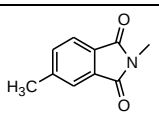
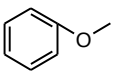


شکل(۳-۲)-نمایش توصیف‌گرهای قطعه مولکولی

- توصیف‌گرهای قطعه A

در ساختار ۱۹۷ مولکول، در استخلاف R_3، شش نوع قطعه اصلی می‌تواند قرار بگیرد که این قطعات در جدول (۳-۲) نمایش داده شده‌اند. بر اساس این قطعات شش نوع توصیف‌گر A_1 تا A_6 تعریف گردید.

جدول(۳-۲)-توصیف‌گرهای حاصل از Fragment(A)

قطعه مولکولی در موقعیت (A)						
علامت توصیف‌گر	A_1	A_2	A_3	A_4	A_5	A_6

طبق جدول (۳-۲) توصیف‌گرهای A_1 تا A_6 حضور ۶ قطعه مختلف را در مولکول نمایش می‌دهند. برای به دست آوردن مقادیر این توصیف‌گرها این گونه عمل گردید: اگر هر کدام از

قطعات A در مولکول حضور داشت مقدار توصیف‌گر مربوط به آن قطعه برابر یک خواهد بود و عدم حضور آن باعث می‌شود که مقدار توصیف‌گر برابر صفر باشد.

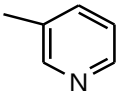
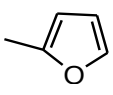
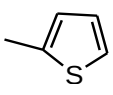
- توصیف‌گرهای قطعه B

با توجه به اینکه در برخی از مولکول‌های مورد بررسی در دو موقعیت B_1 و B_2 گروه متیل وجود دارد، برای در نظر گرفتن ویژگی این بخش از مولکول و بررسی تأثیر آن در فعالیت دارویی ترکیبات، دو توصیف‌گر تعریف شد که با B_1 و B_2 در شکل (۳-۲) نشان داده شده است. این دو توصیف‌گر حضور و یا عدم حضور متیل را نشان می‌دهند، یعنی اگر متیل در موقعیت B_1 یا B_2 حضور داشت مقدار این توصیف‌گر برابر ۱ در نظر گرفته‌شد و در غیر این صورت مقدار صفر به آن اختصاص داده شده است.

- توصیف‌گرهای قطعه C

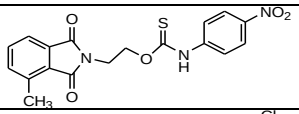
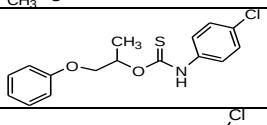
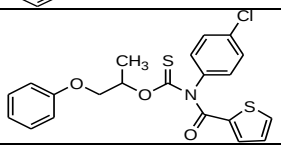
در ساختار ۱۹۷ مولکول، در استخلاف R_1، چهار نوع قطعه اصلی وجود دارد که این قطعات در جدول (۳-۳) نمایش داده شده‌اند. بر اساس این قطعات چهار نوع توصیف‌گر C_1 تا C_4 تعریف گردید و این توصیف‌گرها حضور ۴ قطعه مختلف را در R_1 نشان می‌دهند. برای به دست آوردن مقادیر این توصیف‌گرها این گونه عمل گردید که اگر هر کدام از قطعات C در مولکول حضور داشت مقدار توصیف‌گر مربوط به آن قطعه برابر یک خواهد بود و عدم حضور آن باعث می‌شود که مقدار توصیف‌گر برابر صفر باشد.

جدول (۳-۳)-توصیف‌گرهای حاصل از Fragment (C)

قطعه مولکولی در موقعیت (C)				
علامت توصیف‌گر	C_1	C_2	C_3	C_4

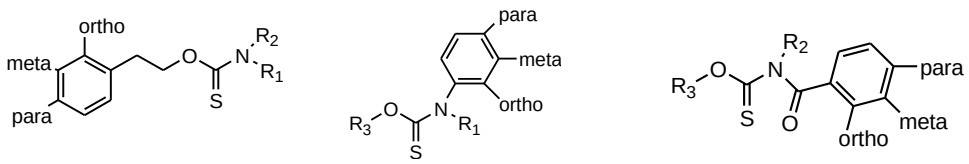
جدول (۳-۴) مقادیر دوازده توصیف‌گر قطعه مولکولی محاسبه شده برای چند ترکیب را نشان می‌دهد.

جدول (۳-۴)- توصیف‌گرهای قطعه مولکولی برای چند ترکیب

		توصیف گر											
No	ترکیب	A_1	A_2	A_3	A_4	A_5	A_6	B_1	B_2	C_1	C_2	C_3	C_4
۵۸		0	1	0	0	0	0	0	0	0	0	0	1
۱۱۰		0	0	0	0	0	1	0	1	0	0	0	1
۱۹۶		0	0	0	0	0	1	0	1	0	0	1	0

۳-۳-۲- توصیف‌گرهای جایگاه (محل) استخلاف

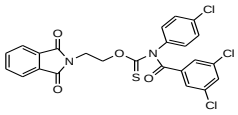
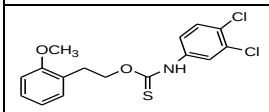
توصیف‌گرهای موجود در این دسته بیانگر تعداد و نوع موقعیت‌های استخلاف شده در حلقه‌های آروماتیک گروه‌های R_1، R_2، R_3 می‌باشند. این دسته از استخلاف‌ها با علائم $R_{1,n}$، $R_{2,n}$، $R_{3,n}$ نشان داده شدندکه n بیان کننده‌ی نوع موقعیت استخلاف شده در حلقه آروماتیک موجود در گروه‌های R_1، R_2، R_3 بوده و می‌تواند ارتو ، متا و پارا باشد. مثلا توصیف‌گر $R_{1,para}$ نشان دهنده تعداد استخلاف‌هایی است که در موقعیت پارا گروه استخلافی R_1 قرار گرفته است. در پایان ۹ توصیف‌گر از این نوع تعریف گردید. این توصیف‌گرها در شکل (۳-۳) نمایش داده شده‌اند.



شکل (۳-۳)- نمایش توصیف‌گرهای جایگاه (محل) استخلاف

برای درک بهتر چند مثال از این توصیف‌گر در جدول (۳-۵) گردآوری شده است.

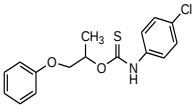
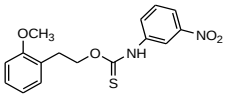
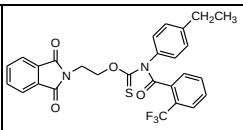
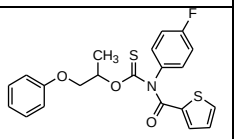
جدول (۳-۵)- توصیف‌گرهای جایگاه (محل) استخلاف برای چند ترکیب

	توصیف گر								
ترکیب	$R_{1,ortho}$	$R_{1,meta}$	$R_{1,para}$	$R_{2,ortho}$	$R_{2,meta}$	$R_{2,para}$	$R_{3,ortho}$	$R_{3,meta}$	$R_{3,para}$
	0	2	0	0	0	1	0	0	0
	0	0	0	0	1	1	1	0	0

۳-۳-۳- توصیف‌گرهای ماهیت استخلاف

حضور استخلاف‌های مختلف می‌تواند روی خاصیت دارویی ترکیب تأثیر داشته باشد، بدین منظور برای در نظر گرفتن ویژگی این بخش از مولکول و بررسی تأثیر آن در فعالیت دارویی، توصیف‌گرهای ماهیت استخلاف برای ترکیبات مورد نظر تعریف شد. برای استخلاف‌های روی حلقه آروماتیک R_1 هشت توصیف‌گر تعریف شد که این توصیف‌گرها با $R_{1\_Substituent}$ نمایش داده می‌شوند. استخلاف‌هایی که روی حلقه آروماتیک R_2 قرار می‌گیرند با $R_{2\_Substituent}$ نمایش داده می‌شوند و شامل ۱۷ توصیف‌گر می باشند و استخلاف‌هایی که روی حلقه R_3 قرار می‌گیرد شامل ۵ توصیف‌گر است که با $R_{2\_Substituent}$ نمایش می‌دهیم. در مجموع ۳۰ توصیف‌گر برای نمایش ماهیت استخلاف‌های روی حلقه‌های آروماتیک تعریف شد. زمانی که این توصیف‌گرها در مولکول حضور داشته باشند عدد ۱ و اگر حضور نداشته باشند مقدار صفر را به آن اختصاص می‌دهیم. برای مثال توصیف‌گر R_{2_Cl} استخلاف کلر روی حلقه R_2 را نشان می‌دهد پس اگر این استخلاف روی حلقه R_2 در هر ترکیبی حضور داشت مقدار این توصیف‌گر برای این ترکیب برابر یک خواهد بود. این توصیف‌گرها برای چند ترکیب در جدول (۳-۶) نمایش داده شده‌اند.

جدول (۳-۶)-توصیف‌گرهای ماهیت استخلاف

	ترکیب			
توصیف گر				
R_{2_Cl}	1	0	0	0
R_{2_Br}	0	0	0	0
R_{2_NO2}	0	1	0	0
R_{2_OCH3}	0	0	0	0
R_{2_OH}	0	0	0	0
R_{2_F}	0	0	0	1
R_{2_I}	0	0	0	0
R_{2_CH3}	0	0	0	0
R_{2_CF3}	0	0	0	0
R_{2_CH2CH3}	0	0	1	0
R_{2_COCH3}	0	0	0	0
$R_{2\_OCH2CH3}$	0	0	0	0
$R_{2\_CH(CH3)2}$	0	0	0	0
R_{2_COOCH3}	0	0	0	0
R_{2_SO2CH3}	0	0	0	0
R_{2_CN}	0	0	0	0
$R_{2\_N(CH3)2}$	0	0	0	0
R_{1_Cl}	0	0	0	0
R_{1_Br}	0	0	0	0
R_{1_NO2}	0	0	0	0
R_{1_OCH3}	0	0	0	0
R_{1_F}	0	0	0	0
R_{1_CH3}	0	0	0	0
R_{1_CF3}	0	0	1	0
R_{1_COOCH3}	0	0	0	0
R_{3_Cl}	0	0	0	0
R_{3_Br}	0	0	0	0
R_{3_OCH3}	0	1	0	0
R_{3_OH}	0	0	0	0
R_{3_CH3}	0	0	0	0

۳-۳-۴- توصیف‌گرهای مربوط به پیوند هیدروژنی و تعداد حلقه‌های بنزن

دسته چهارم توصیف‌گرهای nO، nH، nHACC، nHDon، N#BNZ، PCL، به ترتیب از راست به چپ مربوط به تعداد اتم‌های اکسیژن و هیدروژن، تعداد پیوندهای پذیرنده و دهنده هیدروژن، تعداد حلقه بنزن و تعداد استخلاف کلر در حلقه پیریدین، می‌باشند. در واقع nHACC با شمارش تعداد اتم‌های (N، O، F) و nHDon با شمارش گروه‌های (NH و OH) در ترکیب محاسبه می‌شود.

در پایان ۵۷ توصیف‌گر برای ترکیبات مورد مطالعه به دست آمد، این توصیف‌گرها با یکدیگر همبستگی ندارند و از همه‌ی آن ها به عنوان ورودی در مدل جنگل‌های تصادفی استفاده شد.

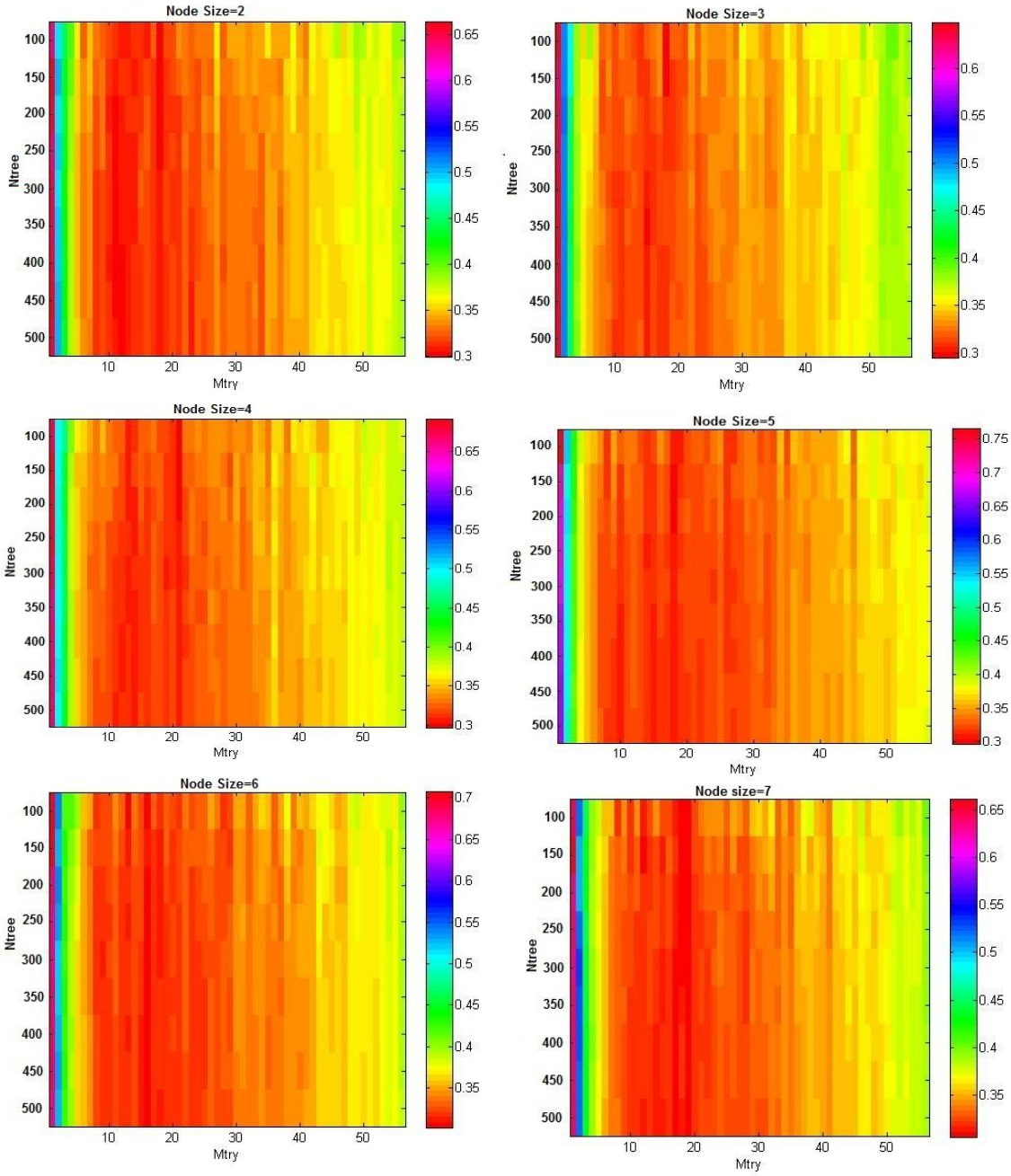
۳-۴- مدل‌سازی جنگل‌های تصادفی

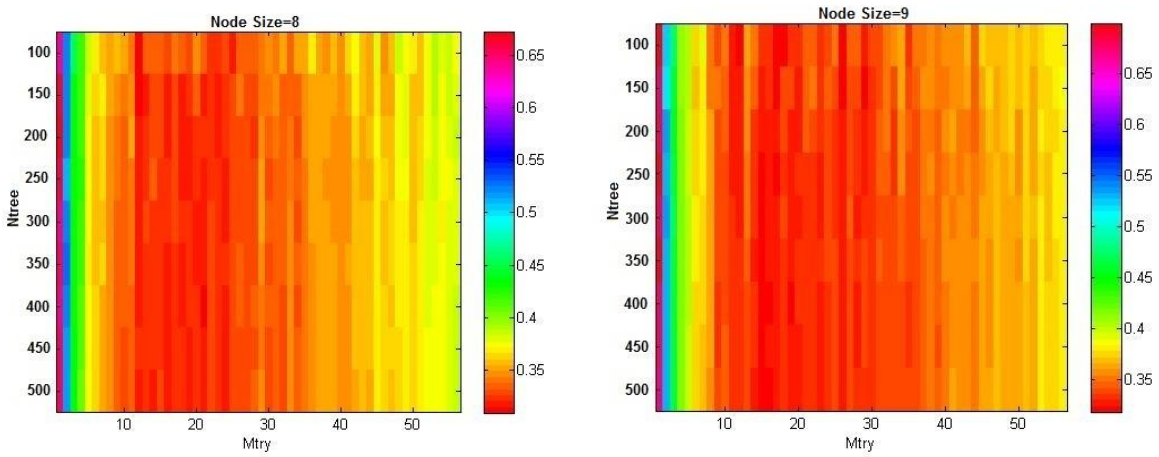
ابتدا سری داده‌ها به طور تصادفی به دو مجموعه آموزش (۱۵۷ ترکیب) و سری آزمون (۴۰ ترکیب) تقسیم شدند. برای ایجاد مدل جنگل‌های تصادفی از داده‌های سری آموزش استفاده گردید. همان‌طور که در بخش (۲-۵-۲) گفته شد، $\frac{1}{3}$ داده‌ها در ساخت درخت شرکت ندارند، این داده‌ها OOB هستند که برای بهینه‌سازی پارامترهای مؤثر در نظرگرفته شدند. برای دستیابی به بهترین مدل ابتدا سه پارامتر مهم و تأثیرگذار یعنی تعداد درختان (n_{tree})، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (Mtry) و تعداد مشاهدات باقی‌مانده در هر گره (Node Size) بهینه گردیدند.

۳-۴-۱- بهینه‌سازی مقادیر n_{tree} , Mtry, Node size

برای بهینه نمودن این پارامترها ، تعداد درخت از ۱۰۰ تا ۵۰۰ با گام ۵۰، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (Mtry) از ۲ تا ۵۷ با گام یک و تعداد مشاهدات

باقی‌مانده در هر گره (Node size) از ۱ تا ۱۰ با گام یک همزمان تغییر داده شد و در هر مرحله مقدار MSE مربوط به مجموعه OOB محاسبه گردید. نتایج بهینه‌سازی در شکل (۳-۴) آمده است و نقاط قرمز رنگ خطای کمتری را نشان می‌دهند. همچنین جدول (۳-۷) مقادیر n_{tree} ،Mtry و node size همراه با MSE متناظر، برای چندین حالت که کمترین MSE برای OOB ایجاد شده، نشان می‌دهد.


Node Size=2
Node Size=3
Node Size=4
Node Size=5
Node Size=6
Node size=7
Ntree
Mtry



ادامه شکل(۳-۴)- بهینه‌سازی پارامترهای ntree, Mtry, Node Size

جدول (۳-۷)- کمترین مقادیر MSE همراه با Mtry و n_{tree} متناظر آنها

N_{tree}	Mtry	Node size	MSE OOB
۳۰۰	**۱۶**	**۲**	**۰/۲۹۵۰**
۱۰۰	۱۹	۳	۰/۲۹۶۵
۱۰۰	۲۲	۴	۰/۲۹۶۹
۱۵۰	۱۹	۵	۰/۲۹۷۱
۴۰۰	۱۷	۶	۰/۳۰۲۶
۲۵۰	۲۰	۷	۰/۳۰۵۹
۱۵۰	۱۳	۸	۰/۳۰۱۰
۱۰۰	۱۸	۹	۰/۳۱۸۸
۳۵۰	۲۸	۱۰	۰/۳۲۳۲

n_{tree}=300 و Mtry=16 و Node size=2 به عنوان مقادیر بهینه انتخاب شدند.

۳-۵- ارزیابی مدل جنگل‌های تصادفی

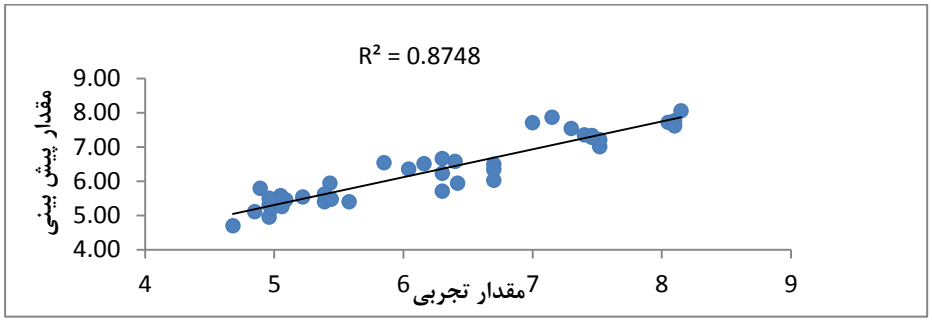
۳-۵-۱- ارزیابی با استفاده از سری آزمون

در این مرحله الگوریتم جنگل‌های تصادفی با استفاده از مقادیر بهینه تعداد درخت، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز و تعداد مشاهدات باقی‌مانده در هر گره اجرا شد سپس مقادیر pEC50 مربوط به ۴۰ ترکیب سری آزمون که در مدل‌سازی از آن‌ها استفاده نشده بود، پیش‌بینی گردید. جدول (۳-۸) نتایج این پیش‌بینی را نشان می‌دهد. هم‌چنین در

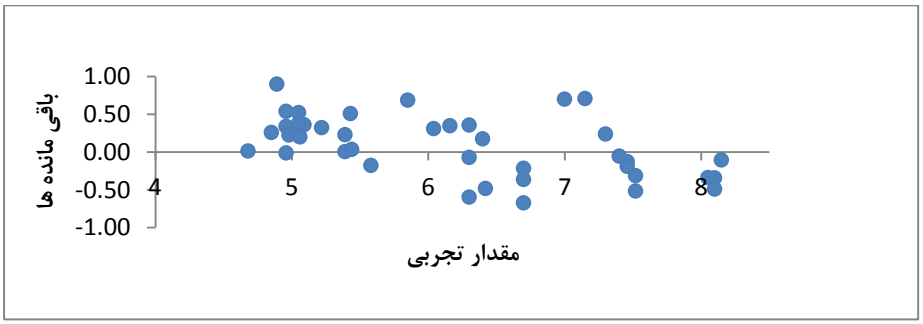
شکل (۳-۴) مقدار پیش‌بینی توسط مدل جنگل‌های تصادفی بر حسب مقدار تجربی برای داده‌های سری آزمون ترسیم شده است و شکل (۳-۵) نیز نمودار باقی‌مانده‌ها را نشان می‌دهد. نتیجه کلی از این ارزیابی این است که مدل حاصله می‌تواند با صحت قابل قبولی فعالیت مشتقات تیوکربامات را پیش‌بینی کند و وجود تقارن در نمودار باقی‌مانده‌ها، برای بیشتر ترکیبات، عدم وجود خطای سیستماتیک در مدل را نشان می‌دهد.

جدول (۳- ۸) – نتایج حاصل از ارزیابی مدلRF با استفاده از سری آزمون

pEC_{50}							
شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا	شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا
۵	۷/۴	۷/۳۴	-۰/۸۱	۱۵۹	۸/۱۵	۸/۰۵	-۱/۲۲
۸	۷/۵۲	۷/۰۰	-۶/۹۱	۱۶۲	۸/۰۵	۷/۷۱	-۴/۲۲
۱۸	۴/۸۵	۵/۱۱	۵/۳۶	۱۶۴	۸/۱۰	۷/۶۱	-۶/۰۳
۲۵	۵/۳۹	۵/۴۰	۰/۱۸	۱۷۵	۷/۳۰	۷/۵۴	۳/۲۸
۳۳	۶/۳۰	۵/۷۰	-۹/۵۲	۵۸	۵/۲۲	۵/۵۴	۶/۱۳
۳۶	۴/۶۸	۴/۷۰	۰/۴۲	۷۸	۴/۸۹	۵/۷۹	۱۸/۴۰
۴۲	۵/۰۵	۵/۵۷	۱۰/۲۹	۶۹	۵/۴۳	۵/۹۴	۹/۴۰
۵۲	۷/۵۲	۷/۲۱	-۴/۱۲	۷۲	۵/۸۵	۶/۵۴	۱۱/۷۸
۵۳	۶/۴	۶/۵۸	۲/۸۱	۷۹	۶/۷۰	۶/۰۳	-۱۰/۰۷
۵۵	۶/۷۰	۶/۴۹	-۳/۱۳	۹۳	۶/۰۴	۶/۳۵	۵/۱۳
۱۱۳	۴/۹۶	۴/۹۵	-۰/۱۹	۹۶	۶/۷۰	۶/۳۴	-۵/۳۷
۱۱۷	۴/۹۶	۵/۳۰	۶/۸۵	۹۹	۶/۳۰	۶/۲۳	-۱/۱۱
۱۲۲	۶/۴۲	۵/۹۴	-۷/۴۷	۱۱۰	۵/۵۸	۵/۴۰	-۳/۲۲
۱۲۵	۸/۱۰	۷/۷۶	-۴/۲۰	۱۱۱	۵/۳۹	۵/۶۲	۴/۲۶
۱۲۹	۷/۴۶	۷/۳۳	-۱/۷۴	۱۷۹	۵/۰۹	۵/۴۵	۷/۰۷
۱۳۳	۷/۱۵	۷/۸۶	۹/۹۳	۱۸۲	۴/۹۸	۵/۲۱	۴/۶۱
۱۳۷	۷/۰۰	۷/۷۰	۹/۹۹	۱۸۴	۵/۰۵	۵/۴۳	۷/۵۱
۱۳۸	۷/۴۶	۷/۲۷	-۲/۵۴	۱۸۸	۵/۰۶	۵/۲۶	۳/۹۴
۱۴۴	۶/۳	۶/۶۶	۵/۶۹	۱۹۱	۵/۴۴	۵/۴۷	۰/۵۵
۱۴۶	۶/۱۶	۶/۵۱	۵/۶۸	۱۹۵	۴/۹۶	۵/۵۰	۱۰/۸۸



شکل (۳-۵)- نمودار مقادیر پیش‌بینی pEC_{50} در مقابل مقادیر تجربی داده‌های آزمون



شکل (۳-۶)- نمودار باقی‌مانده‌ها بر حسب مقدار تجربی pEC_{50} برای سری آزمون

۳-۵-۲- ارزیابی مدل 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 (۳-۴) با ۱۹۶ ترکیب باقی‌مانده مدل‌سازی صورت گرفت سپس مدل به دست آمده برای پیش‌بینی pEC50 ترکیب کنار گذاشته شده به کار گرفته شد و این فرآیند برای تک‌تک ترکیبات تکرار گردید. نتایج حاصل از این روش در جدول (۳-۹) ارائه شده است. نتایج به دست آمده و ضریب همبستگی مشاهده شده در شکل (۳-۶) نشان می‌دهد که مدل توانایی خوبی برای پیش‌بینی فعالیت ضد ایدز مشتقات تیوکربامات دارد. همچنین در شکل (۳-۷) مقادیر باقی‌مانده‌ها بر حسب مقدار تجربی pEC50 ترکیبات مورد بحث ترسیم شده است که تقارن پراکندگی نقاط در دو طرف محور افقی عدم وجود خطای سیستماتیک را برای بیشتر ترکیبات نشان می‌دهد.

جدول (۳- ۹)- نتایج حاصل از ارزیابی مدل RF با استفاده از رد مرحله‌ای تک‌تک برای کل داده‌ها

pEC_{50}							
شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا	شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا
۱	۷/۱۰	۷/۱۳	۰/۴۴	۲۷	۶/۲۲	۵/۴۹	-۱۱/۷۶
۲	۷/۴۰	۷/۰۱	-۵/۲۴	۲۸	۵/۳۶	۵/۳۷	۰/۱۳
۳	۶/۸۵	۷/۴۳	۸/۴۲	۲۹	۵/۲۲	۵/۶۳	۷/۷۹
۴	۷/۰۰	۷/۳۶	۵/۱۸	۳۰	۴/۱۲	۵/۰۸	۲۳/۴۰
۵	۷/۴۰	۷/۳۸	-۰/۲۹	۳۱	۴/۹۲	۵/۳۶	۸/۹۳
۶	۷/۵۲	۷/۴۱	-۱/۴۷	۳۲	۴/۶۴	۴/۸۶	۴/۸۴
۷	۷/۷۰	۷/۴۴	-۳/۳۲	۳۳	۶/۱۰	۶/۶۶	۹/۱۵
۸	۷/۵۲	۷/۱۸	-۴/۴۹	۳۴	۶/۳۰	۵/۶۷	۱۰/۰۷
۹	۵/۱۵	۷/۰۰	۳۵/۸۶	۳۵	۵/۱۰	۵/۱۸	۱/۶۰
۱۰	۶/۲۲	۷/۱۲	-۱۴/۴۰	۳۶	۴/۵۲	۴/۸۰	۶/۲۰
۱۱	۷/۷۰	۷/۱۸	-۶/۸۱	۳۷	۴/۶۸	۴/۶۸	۰/۱۰
۱۲	۷/۴۰	۷/۴۳	۰/۳۵	۳۸	۵/۴۰	۵/۳۶	-۰/۶۶
۱۳	۷/۴۰	۷/۱۲	-۳/۷۹	۳۹	۴/۳۲	۵/۰۷	۱۷/۳۶
۱۴	۵/۴۳	۴/۸۱	-۱۱/۳۵	۴۰	۵/۴۵	۶/۵۰	۱۹/۲۷
۱۵	۴/۱۹	۵/۲۲	۲۴/۶۳	۴۱	۴/۹۶	۵/۴۶	۱۰/۰۵
۱۶	۵/۵۲	۴/۹۰	-۱۱/۲۱	۴۲	۵/۰۰	۵/۴۴	۸/۷۸
۱۷	۵/۲۵	۴/۹۲	-۶/۲۵	۴۳	۵/۰۵	۵/۵۵	۹/۸۲
۱۸	۴/۷۷	۴/۸۰	۰/۶۶	۴۴	۵/۰۰	۵/۰۵	۱/۰۸
۱۹	۴/۸۵	۵/۰۶	۴/۲۴	۴۵	۵/۹۲	۵/۶۵	-۴/۶۲
۲۰	۴/۳۶	۵/۱۰	۱۷/۰۵	۴۶	۷/۷۰	۷/۲۸	-۵/۵۱
۲۱	۵/۷۵	۵/۲۶	-۸/۵۵	۴۷	۷/۷۰	۷/۴۲	-۳/۶۶
۲۲	۴/۸۷	۵/۳۷	۱۰/۳۳	۴۸	۸/۰۰	۷/۶۰	-۴/۹۷
۲۳	۴/۶۸	۵/۳۷	۱۴/۸۱	۴۹	۷/۱۵	۷/۳۰	۲/۱۲
۲۴	۵/۷۰	۵/۳۶	-۵/۹۵	۵۰	۷/۵۲	۷/۵۶	۰/۵۶
۲۵	۵/۱۵	۵/۲۶	۲/۱۲	۵۱	۸/۱۰	۷/۵۷	-۶/۵۳
۲۶	۵/۳۹	۵/۳۸	-۰/۱۷				

ادامه جدول (۳- ۹)- نتایج حاصل از ارزیابی مدل RF با استفاده از رد مرحله‌ای تک‌تک برای کل داده‌ها

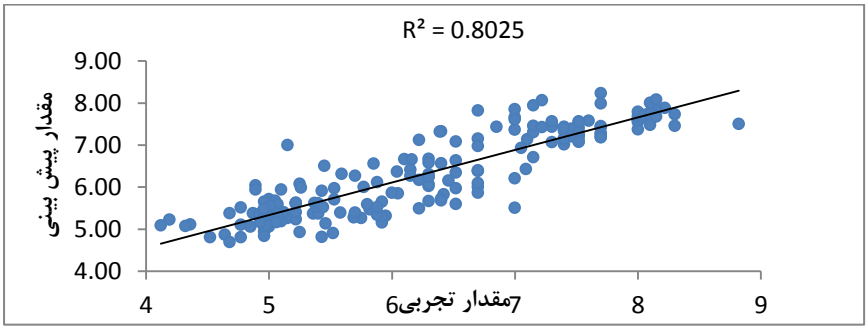
pEC_{50}							
شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا	شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا
۵۲	۸/۰۰	۷/۵۴	-۵/۷۴	۷۷	۶/۳۰	۶/۰۶	-۳/۷۵
۵۳	۷/۵۲	۷/۲۴	-۳/۷۲	۷۸	۵/۰۰	۵/۷۱	۱۴/۲۲
۵۴	۶/۴۰	۶/۵۶	۲/۴۹	۷۹	۴/۸۹	۵/۹۴	۲۱/۳۹
۵۵	۶/۳۰	۶/۶۷	۵/۸۹	۸۰	۶/۷۰	۵/۸۶	-۱۲/۵۵
۵۶	۶/۷۰	۶/۳۸	-۴/۸۴	۸۱	۶/۷۰	۶/۰۹	-۹/۰۸
۵۷	۷/۳۰	۷/۴۶	۲/۲۴	۸۲	۶/۳۰	۶/۲۴	-۰/۸۹
۵۸	۷/۱۵	۷/۴۵	۴/۲۴	۸۳	۵/۲۵	۶/۰۷	۱۵/۶۵
۵۹	۵/۲۲	۵/۶۱	۷/۵۳	۸۴	۶/۱۵	۶/۲۷	۱/۹۱
۶۰	۵/۳۷	۵/۶۲	۴/۵۹	۸۵	۵/۲۶	۵/۹۸	۱۳/۷۴
۶۱	۶/۳۰	۶/۳۰	۰/۰۰	۸۶	۵/۰۵	۵/۲۴	۳/۶۷
۶۲	۶/۰۵	۵/۸۴	-۳/۳۹	۸۷	۵/۰۵	۵/۶۷	۱۲/۳۵
۶۳	۴/۷۷	۵/۵۱	۱۵/۵۱	۸۸	۵/۵۳	۵/۹۷	۷/۸۸
۶۴	۶/۱۵	۶/۴۱	۴/۲۶	۸۹	۷/۰۰	۵/۵۰	-۲۱/۳۷
۶۵	۵/۷۷	۶/۰۰	۳/۹۱	۹۰	۶/۴۰	۵/۶۸	-۱۱/۲۷
۶۶	۶/۵۲	۵/۹۷	-۸/۴۲	۹۱	۷/۱۵	۶/۷۱	-۶/۲۲
۶۷	۵/۸۲	۵/۴۶	-۶/۱۳	۹۲	۷/۰۰	۶/۲۰	-۱۱/۴۰
۶۸	۴/۸۹	۶/۰۴	۲۳/۵۱	۹۳	۶/۵۲	۶/۶۳	۱/۷۲
۶۹	۵/۰۰	۵/۶۳	۱۲/۶۱	۹۴	۶/۰۴	۶/۳۶	۵/۳۱
۷۰	۵/۴۳	۵/۹۱	۸/۷۷	۹۵	۵/۷۰	۶/۲۶	۹/۸۵
۷۱	۷/۰۹	۶/۴۳	-۹/۳۸	۹۶	۵/۱۰	۵/۹۳	۱۶/۳۶
۷۲	۶/۷۰	۶/۰۰	-۱۰/۴۵	۹۷	۶/۷	۶/۳۹	-۴/۶۱
۷۳	۵/۸۵	۶/۵۵	۱۲/۰۰	۹۸	۶/۲۲	۶/۱۶	-۰/۹۳
۷۴	۶/۵۲	۵/۶۰	-۱۴/۱۵	۹۹	۶/۵۲	۶/۳۴	-۲/۸۰
۷۵	۶/۰۰	۵/۸۶	-۲/۳۶	۱۰۰	۶/۳۰	۶/۲۴	-۰/۸۹
۷۶	۵/۹۵	۵/۳۱	-۱۰/۷۰	۱۰۱	۵/۰۴	۵/۵۵	۱۰/۱۱

ادامه جدول (۳- ۹)- نتایج حاصل از ارزیابی مدل RF با استفاده از رد مرحله‌ای تک‌تک برای کل داده‌ها

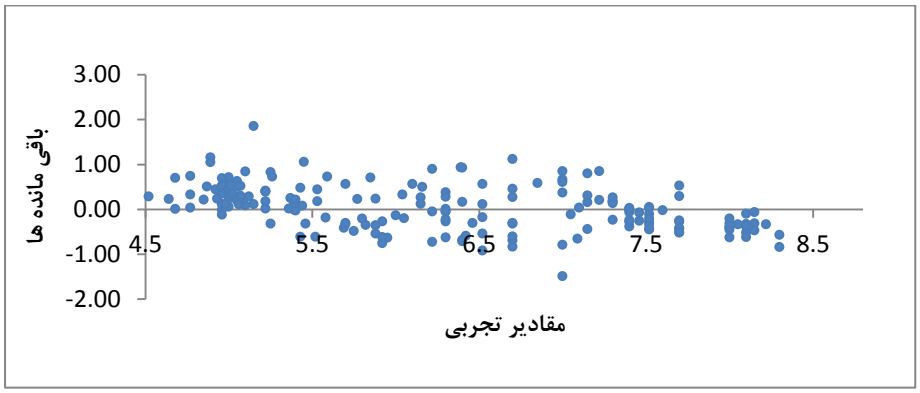
pEC_{50}							
شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا	شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا
۱۰۲	۵/۸۸	۶/۱۱	۳/۹۵	۱۲۷	۸/۰۰	۷/۳۷	-۷/۹۴
۱۰۳	۵/۰۰	۵/۴۶	۹/۱۲	۱۲۸	۸/۰۰	۷/۵۹	-۵/۱۰
۱۰۴	۵/۵۳	۵/۷۰	۳/۱۱	۱۲۹	۷/۲۲	۸/۰۶	۱۱/۶۷
۱۰۵	۵/۰۷	۵/۵۹	۱۰/۱۷	۱۳۰	۷/۴۶	۷/۳۸	-۱/۰۵
۱۰۶	۴/۷۷	۵/۱۰	۶/۸۳	۱۳۱	۷/۷۰	۷/۲۶	-۵/۶۵
۱۰۷	۵/۶۹	۵/۲۷	-۷/۳۶	۱۳۲	۷/۵۲	۷/۳۱	-۲/۸۲
۱۰۸	۵/۹۲	۵/۳۰	-۱۰/۵۳	۱۳۳	۷/۴۰	۷/۳۳	-۰/۹۸
۱۰۹	۴/۹۶	۵/۶۵	۱۳/۹۱	۱۳۴	۷/۱۵	۷/۹۴	۱۱/۰۸
۱۱۰	۵/۸۰	۵/۵۸	-۳/۷۱	۱۳۵	۶/۷۰	۷/۱۵	۶/۷۰
۱۱۱	۵/۵۸	۵/۳۹	-۳/۴۴	۱۳۶	۶/۴۰	۷/۳۲	۱۴/۳۸
۱۱۲	۵/۳۹	۵/۵۲	۲/۴۳	۱۳۷	۶/۳۹	۷/۳۲	۱۴/۵۱
۱۱۳	۴/۹۶	۵/۰۷	۲/۲۳	۱۳۸	۷/۰۰	۷/۶۰	۸/۶۳
۱۱۴	۴/۹۶	۴/۹۶	۰/۰۰	۱۳۹	۷/۴۶	۷/۲۰	-۳/۴۷
۱۱۵	۴/۹۶	۴/۸۴	-۲/۴۹	۱۴۰	۷/۶۰	۷/۶۰	۰/۰۰
۱۱۶	۴/۹۶	۵/۰۷	۲/۱۴	۱۴۱	۷/۳۰	۷/۴۸	۲/۴۷
۱۱۷	۵/۰۰	۵/۳۹	۷/۸۳	۱۴۲	۸/۳۰	۷/۴۵	-۱۰/۲۴
۱۱۸	۴/۹۶	۵/۴۶	۱۰/۰۳	۱۴۳	۸/۳۰	۷/۷۲	-۶/۹۳
۱۱۹	۵/۲۲	۵/۲۲	۰/۰۰	۱۴۴	۸/۲۲	۷/۸۸	-۴/۱۴
۱۲۰	۴/۹۶	۵/۳۰	۶/۹۴	۱۴۵	۶/۳۰	۶/۵۸	۴/۳۷
۱۲۱	۵/۲۲	۵/۴۰	۳/۳۷	۱۴۶	۶/۷۰	۷/۸۲	۱۶/۶۶
۱۲۲	۵/۹۲	۵/۱۶	-۱۲/۹۱	۱۴۷	۶/۱۶	۶/۶۵	۷/۹۸
۱۲۳	۶/۴۲	۵/۸۲	-۹/۲۸	۱۴۸	۵/۵۹	۶/۳۱	۱۲/۹۳
۱۲۴	۵/۴۶	۵/۱۳	-۶/۰۱	۱۴۹	۸/۰۰	۷/۶۸	-۳/۹۷
۱۲۵	۷/۰۰	۷/۶۶	۹/۴۰	۱۵۰	۸/۸۲	۷/۵۰	-۱۴/۹۹
۱۲۶	۸/۱۰	۷/۷۱	-۴/۸۰	۱۵۱	۷/۴۰	۷/۴۰	۰/۰۰

ادامه جدول (۳- ۹)- نتایج حاصل از ارزیابی مدل RF با استفاده از رد مرحله‌ای تک‌تک برای کل داده‌ها

pEC_{50}							
شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا	شماره ترکیب	مقدار تجربی	مقدار پیش بینی	درصد خطا
۱۵۲	۷/۷۰	۷/۱۸	-۶/۷۲	۱۷۵	۷/۲۲	۷/۴۲	۲/۷۹
۱۵۳	۷/۵۲	۷/۰۷	-۶/۰۰	۱۷۶	۷/۳۰	۷/۵۶	۳/۵۳
۱۵۴	۸/۱۵	۷/۸۳	-۳/۹۱	۱۷۷	۶/۳۰	۶/۰۲	-۴/۴۰
۱۵۵	۸/۱۰	۷/۷۷	-۴/۰۹	۱۷۸	۶/۴۶	۶/۱۵	-۴/۸۴
۱۵۶	۸/۱۰	۷/۷۴	-۴/۴۴	۱۷۹	۶/۴۰	۵/۷۱	-۱۰/۷۴
۱۵۷	۸/۰۰	۷/۷۹	-۲/۶۳	۱۸۰	۵/۰۹	۵/۳۰	۴/۲۱
۱۵۸	۷/۷۰	۷/۹۹	۳/۷۰	۱۸۱	۵/۰۰	۵/۲۳	۴/۵۴
۱۵۹	۸/۱۰	۸/۰۰	-۱/۲۷	۱۸۲	۵/۰۲	۵/۳۷	۶/۹۳
۱۶۰	۸/۱۵	۸/۰۸	-۰/۸۸	۱۸۳	۴/۹۸	۵/۰۹	۲/۲۲
۱۶۱	۷/۰۰	۷/۸۵	۰۹۱۲	۱۸۴	۴/۹۳	۵/۱۶	۴/۶۸
۱۶۲	۷/۵۲	۷/۵۵	۰/۳۵	۱۸۵	۵/۰۵	۵/۲۷	۴/۴۲
۱۶۳	۸/۰۵	۷/۷۱	-۴/۲۰	۱۸۶	۵/۱۲	۵/۴۰	۵/۴۶
۱۶۴	۸/۱۰	۷/۷۳	-۴/۵۵	۱۸۷	۵/۰۱	۵/۵۴	۱۰/۶۲
۱۶۵	۸/۱۰	۷/۴۸	-۷/۷۱	۱۸۸	۵/۰۷	۵/۳۷	۵/۸۲
۱۶۶	۷/۷۰	۸/۲۳	۶/۸۲	۱۸۹	۵/۰۶	۵/۱۶	۲/۰۲
۱۶۷	۸/۱۵	۷/۶۷	-۵/۸۵	۱۹۰	۵/۴۰	۵/۶۱	۳/۸۴
۱۶۸	۷/۵۲	۷/۰۸	-۵/۸۲	۱۹۱	۵/۸۸	۵/۳۳	-۹/۳۰
۱۶۹	۷/۴۰	۷/۱۵	-۳/۳۲	۱۹۲	۵/۴۴	۵/۵۲	۱/۳۹
۱۷۰	۷/۳۰	۷/۴۳	۱/۸۱	۱۹۳	۵/۸۸	۵/۵۲	-۶/۱۹
۱۷۱	۶/۷۰	۶/۹۷	۴/۰۱	۱۹۴	۵/۷۰	۵/۳۹	-۵/۴۳
۱۷۲	۶/۵۲	۷/۰۸	۸/۵۷	۱۹۵	۴/۹۶	۵/۴۴	۹/۶۳
۱۷۳	۷/۳۰	۷/۰۶	-۳/۲۸	۱۹۶	۴/۹۶	۵/۴۹	۱۰/۶۱
۱۷۴	۷/۰۵	۶/۹۳	-۱/۷۲	۱۹۷	۵/۲۲	۵/۶۱	۷/۴۵



شکل (۳-۷) – نمودار مقادیر پیش بینی شده بر حسب مقادیر تجربی pEC_{50} به روش رد مرحله‌ای تک‌تک برای کل داده‌ها



شکل (۳-۸)- نمودار باقی‌مانده‌ها بر حسب مقدار تجربی pEC_{50} برای کل داده‌ها

۳-۵-۳- ارزیابی مدل RF با استفاده از پارامترهای آماری

مطابق جدول (۳-۱۰) چند پارامتر آماری، جهت ارزیابی مدل RF ساخته شده به کار گرفته شد. این پارامترها طبق روابط توضیح داده شده در بخش (۲-۶) محاسبه شده اند. نتایج نشان می‌دهد که مدل به دست آمده توانایی رضایت‌بخشی برای پیش‌بینی فعالیت ضد ایدز مشتقات تیوکربامات دارد.

جدول (۳-۱۰) - پارامترهای آماری برای مدل RF

پارامتر	سری آزمون(N=40)	کل داده ها(N=197)
MAE	۰/۳۴۵۳	۰/۴۱۳۵
MSE	۰/۱۸۷۳	۰/۲۴۲۲
PRESS	۷/۳۷۰	۵۱/۲۰
SEP	۰/۴۲۹۳	۰/۵۰۹۴
R^2	۰/۸۴۴۸	۰/۸۰۲۵
REP (%)	۶/۹۲۱	۸/۱۵۳
MRE(%)	۵/۷۰۸	۶/۹۰۵

۳-۵-۴- ارزیابی مدل ارائه شده با استفاده از آزمون Y- تصادفی[1]

این تکنیک ارزیابی مدل با هدف بررسی هر گونه ارتباط تصادفی بین داده‌ها انجام شد. با توجه به اینکه مقادیر pEC_{50} ترکیبات مورد مطالعه در بازه ۹/۰۰-۴/۰۰ قرار دارد، با استفاده از نرم افزار متلب مقادیر تصادفی از متغیر وابسته در این بازه تولید گردید. مدل QSAR جدید با استفاده از ماتریس متغیرهای مستقل اصلی و مقادیر تصادفی از متغیر وابسته توسعه یافت. اگر مدل اصلی هیچ‌گونه ارتباط تصادفی نداشته باشد، تفاوت قابل توجهی بین مقدار ضریب تعیین مدل اصلی و مدل QSAR که با پاسخ‌های تصادفی توسعه یافته، وجود دارد. نتایج حاصل از چندین بار اجرای آزمون Y- تصادفی در جدول (۳-۱۱) نشان داده شده است. مقادیر کوچک ضریب تعیین (R^2) بیان‌گر عدم وجود ارتباط شانسی یا وابستگی ساختاری به سری آموزش در مدل توسعه یافته توسط RF می‌باشد.

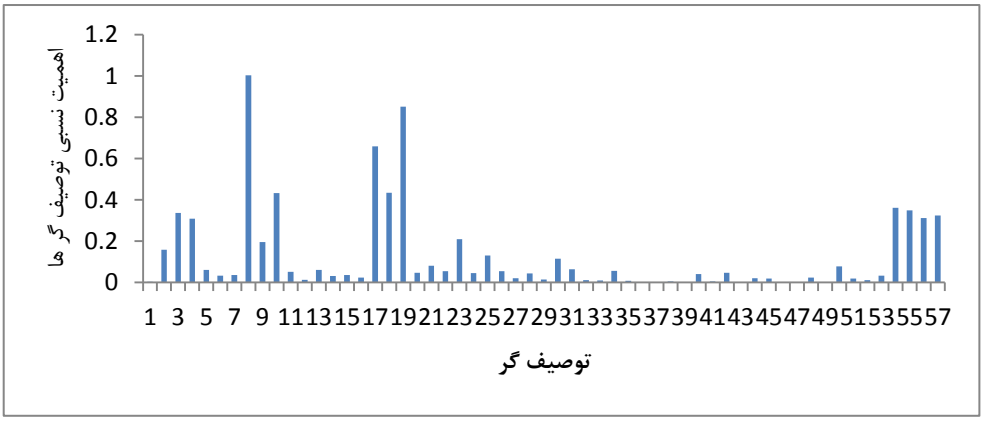
جدول (۳-۱۱)– مقادیر R^2 برای سری آزمون با استفاده از آزمونY- تصادفی

تکرار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
R^2	۰/۰۰۱	۰/۰۰۲	۰/۰۲۵	۰/۰۴۵	۰/۰۰۶۵	۰/۰۲۱	۰/۱۸۷	۰/۰۱۶	۰/۰۰۰	۰/۱۲۳

1 - Y- randomization test

۳-۶- انتخاب توصیف‌گرهای مهم با روش جنگل‌های تصادفی

همان‌طور که در بخش (۲-۵-۳) توضیح داده شد، روش جنگل‌های تصادفی علاوه بر اینکه به عنوان یک روش مدل‌سازی به کار می‌رود، می‌تواند به عنوان یک روش انتخاب متغیر نیز به کار برده شود. پس از ساخت مدل جنگل‌های تصادفی و بهینه کردن پارامترهای مؤثر، الگوریتم جنگل‌های تصادفی با استفاده از مقادیر بهینه صد بار تکرار شد و در هر بار میزان کاهش MSE به عنوان مقدار اهمیت توصیف‌گرها محاسبه گردید. سرانجام میانگین مقادیر اهمیت هر توصیف‌گر به عنوان شاخص نهایی در نظر گرفته شد و سپس اهمیت نسبی هر توصیف‌گر نسبت به توصیف‌گری که بیشترین اهمیت را دارا است، تعیین شد. شکل (۳-۸) میزان اهمیت نسبی توصیف‌گرها را نمایش می‌دهد و همان‌طور که مشاهده می‌شود، بسیاری از توصیف‌گرها دارای اهمیت نسبی کمی بوده و تنها ۶ توصیف‌گر اهمیت نسبی قابل توجهی دارند که در جدول (۳-۱۲) این توصیف‌گرها همراه با مقدار اهمیت نسبی و ضریب اثر آن‌ها گردآوری شده است.



شکل(۳-۹)- نمودار اهمیت نسبی توصیف‌گرهای سهم گروه

جدول(۳-۱۲)-توصیف گرهای برتر

No	علامت	مفهوم توصیفگر	اهمیت نسبی	ضریب اثر
۱	A_1	حضور یا عدم‌حضور فتالامید در استخلاف R_3	۱/۰۰	+۰/۷۹۴
۲	$R_{2,para}$	حضور استخلاف در موقعیت پارا حلقه R_2	۰/۸۵۰	+۰/۸۴۶
۳	$R_{2,ortho}$	حضور استخلاف در موقعیت اورتو حلقه R_2	۰/۶۵۸	-۰/۹۶۴
۴	$R_{2,meta}$	حضور استخلاف در موقعیت متا حلقه R_2	۰/۴۳۳	-۰/۵۷۰
۵	A_6	حضور یا عدم‌حضور متوکسی بنزن در استخلاف R_3	۰/۴۳۲	-۰/۵۳۷
۶	nHACC	تعداد اتم های گیرنده پیوند هیدروژنی(N,O,F)	۰/۳۶۱	+۰/۰۵۳

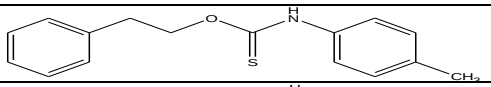
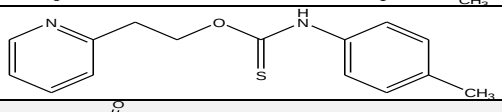
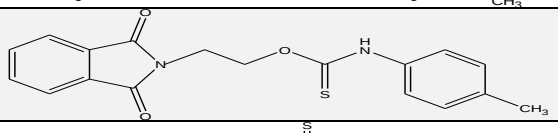
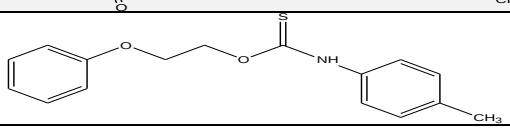
۳-۷- بررسی ارتباط بین توصیف‌گرهای منتخب و خاصیت مورد نظر

در این بخش به بررسی ارتباط بین توصیف‌گرهای منتخب که در بخش (۳-۶) محاسبه شدند، می‌پردازیم و سعی می‌کنیم تا حد امکان توضیحی برای ارتباط بین این توصیف‌گرها و فعالیت مورد نظر بیابیم.

الف-توصیف‌گر A_1

ترکیباتی که حاوی گروه‌های فتالامید هستند دارای خاصیت ضد التهابی، ضد سرطان و مهارکنندگی آنزیم اینتگراز می‌باشند. به علت حضور گروه‌های CO در فتالامید قادر به تشکیل پیوندهای هیدوژنی با سایت‌های فعال آمینواسیدها می‌باشد. این توصیف‌گر دارای ضریب اثر مثبت در مدل است و نشان می‌دهد مشتقات تیوکربامات که دارای گروه فتالامید هستند نسبت به سایر مشتقات pEC_{50} بیشتری دارند، در نتیجه از خاصیت دارویی بیش‌تری برخوردارند. چند مثال از اثر این توصیف‌گر در جدول (۳-۱۳) ارائه شده است.

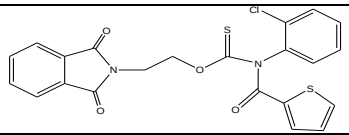
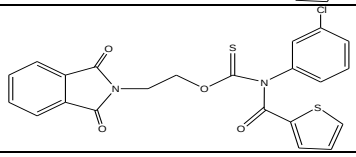
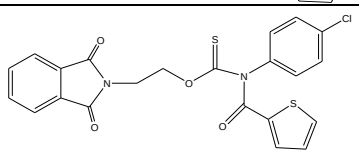
جدول(۳-۱۳)- مثال‌هایی از اثرگروه phthalimide بر فعالیت دارویی

No	ترکیب	PEC_{50}
۱		۵/۳۷
۲		۶/۳۰
۳		**۷/۷۰**
۴		۵

ب) توصیف‌گر R_{2-n}

توصیف‌گرهای R_{2-n}، موقعیت‌های ارتو، متا و پارا روی حلقه بنزن متصل به N را نشان می‌دهد. $R_{2,ortho}$ و $R_{2,meta}$ دارای ضریب منفی و $R_{2,para}$ دارای ضریب مثبت در مدل می‌باشد. مقدار pEC_{50} برای ترکیباتی که دارای استخلاف در موقعیت پارا حلقه R_2 هستند زیاد است، بنابراین حضور استخلاف در موقعیت پارا حلقه R_2 باعث افزایش خاصیت دارویی می‌شود. همچنین ترکیبات با استخلاف در موقعیت ارتو و متا دارای pEC_{50} کمتری هستند و خاصیت دارویی کمتری دارند. بررسی ترکیبات مورد مطالعه نیز این موضوع را نشان می‌دهد جدول (۳-۱۴).

جدول(۳-۱۴)-بررسی اثر موقعیت استخلاف بر فعالیت دارویی

No	ترکیب	PEC_{50}
۱		۴/۹۶
۲		۵/۲۲
۳		۸/۱۰

ج) توصیف‌گر nHACC[1]

این توصیف‌گر تعداد سایت‌های پذیرنده پیوند هیدروژنی را نشان می‌دهد. در واقع پیوندهای پذیرنده هیدروژنی در یک دارو بستگی به تعداد اتم‌های N و O و F دارد. nHACC در واقع توانایی یک دارو در ایجاد پیوند هیدروژنی را نشان می‌دهد که این توانایی یک ویژگی مهم برای تعامل بین آنزیم / مهارکننده و یا یک گیرنده/ لیگاند است. در یک مولکول با تعداد قابل توجهی پیوندهای دهنده و پذیرنده هیدروژنی ، انرژی قابل توجهی برای شکستن این پیوندها در آب مورد نیاز است تا عبور از لایه لیپیدی تسهیل شود. بنابراین مولکول‌هایی که قطبی هستند عبور سخت‌تری از غشای سلولی دارند و فعالیت کمتری خواهند داشت [۳۱]. دراینجا با افزایش تعداد nHACC مقدار pEC_{50}افزایش یافته است و این توصیف‌گر دارای ضریب مثبت است.

۳-۸- نتیجه‌گیری نهایی

با توجه به این که تاکنون مطالعات استفاده از توصیف‌گرهای سهم گروه برای پیش‌بینی خاصیت دارویی انجام نشده است و این اولین گزارش در مورد بکارگیری جنگل‌های تصادفی با توصیف‌گرهای سهم گروه در مورد ترکیبات ضد ایدز می‌باشد، این مدل می‌تواند به محققان در طراحی ترکیبات جدید از این مشتقات با استخلاف‌های مختلف کمک زیادی نماید. توصیف‌گرهای سهم گروه از روی ساختار ترکیب به دست می‌آیند بنابراین ساده و قابل تفسیر می‌باشند و محور اصلی این پژوهش نیز ارائه توصیف‌گرهای ساده، در عین حال مرتبط با خاصیت مورد نظر است. با توجه به نتایج به دست آمده می‌توان گفت که فعالیت بازدارندگی این ترکیبات وابستگی قابل قبولی به ساختمان دو بعدی ترکیب دارد.

1 - number of acceptor atoms for H-bonds

همان‌طور که در بخش (۱-۴) گفته شد دو مطالعه در مورد ۱۱۱ ترکیب از مشتقات تیوکربامات در سال ۲۰۰۸ و۲۰۱۲ انجام شده است [۱۲و۱۳]. این مشتقات در مجموعه داده-های این پژوهش نیز قرار دارند، بنابراین برخی از پارامترهای آماری مربوط به دو پژوهش اخیر در جدول (۳-۱۵) از مقالات مربوطه آن‌ها جمع‌آوری شده است و با نتایج حاصل از مدل RF ارائه شده در این پژوهش مورد مقایسه قرار گرفته است. با توجه به نتایج حاصل از این جدول می‌توان گفت مدل جنگل‌های تصادفی ارائه شده نسبت به سایر روش‌های مدل‌سازی که تاکنون برای مشتقات تیوکربامات انجام شده، برتری دارد. در حالیکه تعداد ترکیبات متنوع‌تری در این مدل وارد شده است.

جدول (۳-۱۵)-مقایسه نتایج حاصل از مدل‌سازی مشتقات تیوکربامات با روش های مختلف

پارامتر	سری آزمون (n=40)	سری آزمون (n=17)	سری آزمون (n=17)	سری آزمون (n=17)
	RF method	ACO/PLS method [13]	ACO/RBFNN method [13]	COMFA method [12]
MAE(%)	۵/۷۱	۱۶/۱۴	۱۴/۲۴	۱۵/۷۶
MSE	۰/۱۸۷	۰/۵۳	۰/۳۷	۰/۵۲
PRESS	۷/۳۷۰	۴/۷۶	۲/۳۲	۶/۹۶
RSEP(%)	۶/۹۲۱	۹/۴۷	۶/۶۰	۹/۲۴
R^2	۰/۸۴۴	۰/۶۴	۰/۸۰۷	۰/۷۰

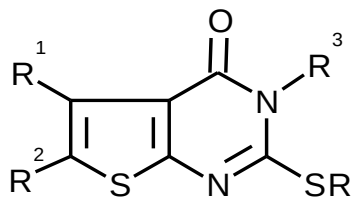
فصل چهارم

ارتباط کمی ساختار- فعالیت مشتقات نکروستاتین-۵ به عنوان بازدارنده-های نکروپتوز با استفاده از شبکه عصبی مصنوعی

در این پروژه به منظور مطالعه ارتباط کمی ساختار-فعالیت گروهی از مشتقات نکروستاتین- ۵، ابتدا از الگوریتم جنگل‌های تصادفی برای طبقه‌بندی ترکیبات به دو گروه فعال و غیرفعال استفاده گردید و در مرحله‌ی بعد یک مدل شبکه‌ی عصبی مصنوعی طراحی و بهینه گردید که بتواند فعالیت دارویی (pEC_{50}) ۵۱ ترکیب فعال از مشتقات نکروستاتین-۵ را به درستی پیش‌بینی نماید.

۴-۱- سری داده‌ها

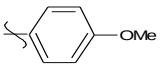
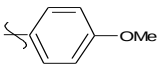
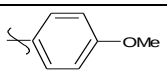
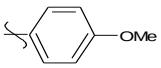
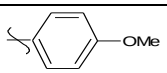
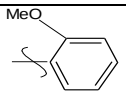
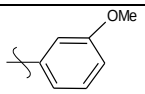
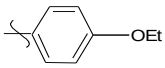
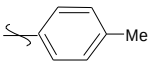
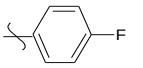
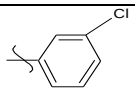
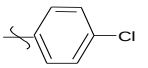
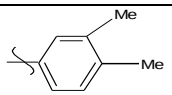
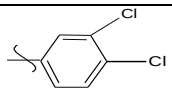
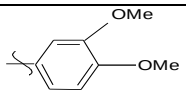
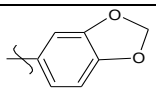
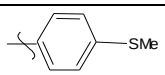
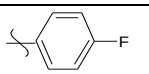
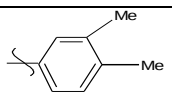
ترکیبات مورد استفاده در این پروژه، داده‌های ۱۶۰ ترکیب از مشتقات نکروستاتین-۵ (Nec-5) می‌باشند که از مرجع شماره‌ی [۴۰] گردآوری شده است. مقادیر منهای لگاریتم داده‌های EC_{50} بر حسب مولار ($-\log EC_{50}\times 10^{-6}$) به عنوان متغیر وابسته، در این مدل‌سازی مورد استفاده قرار گرفته است. از لحاظ شیمیایی Nec-5، با عنوان ۳-*p* متوکسی فنیل ۱-۵ و ۶- تترا متیلنو تینو [۲و۳-*d*] پیریمیدین-۴-یک-۲- مرکاپتو اتیل سیانید[1] نامیده می‌شود و ساختار پایه در شکل (۴-۱) آمده است. در جدول (۴-۱) نیز جزئیات استخلاف‌ها، فعال یا غیرفعال بودن هر ترکیب و مقادیر عددی pEC_{50} برحسب مولار متناظر هر ترکیب فعال آورده شده است. با توجه به نتایج تجربی از ۱۶۰ مشتق Nec-5، ۵۱ مورد آن دارای فعالیت مهارکنندگی بوده و بقیه غیرفعال هستند. برای ترکیبات فعال مقادیر pEC_{50} بر حسب مولار در ناحیه تقریبی ۷/۰۰-۴/۰۰ قرار دارند.



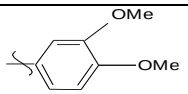
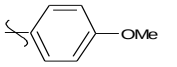
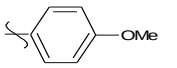
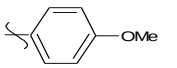
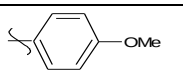
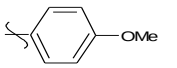
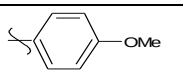
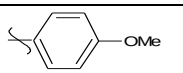
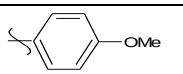
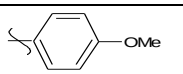
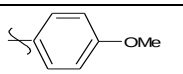
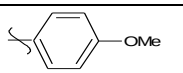
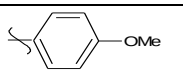
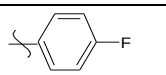
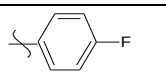
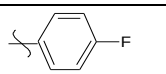
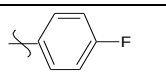
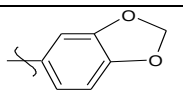
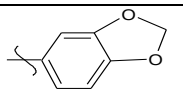
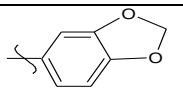
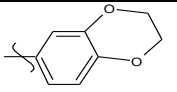
شکل (۴-۱)-اسکلت مولکولی ترکیبات مورد بررسی

1 - 3-p-methoxyphenyl-5,6-tetramethylenothieno[2,3-d]pyrimidin-4-one-2-mercaptoethylcyanide

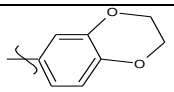
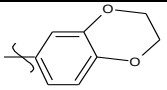
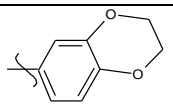
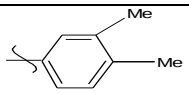
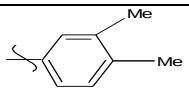
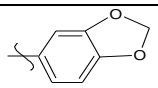
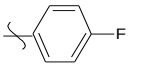
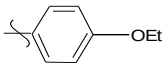
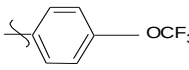
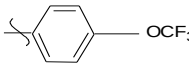
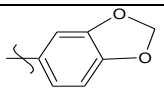
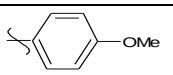
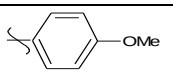
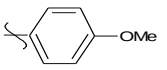
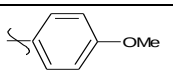
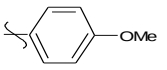
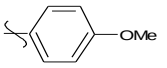
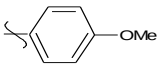
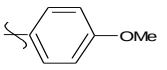
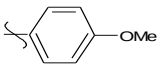
جدول(۴-۱)-جزئیات ساختار مولکولی مشتقات نکروستاتین ۵

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۱	CH_2CN	$-(CH_2)_4-$			۶/۶۲
۲	Me	$-(CH_2)_4-$			۶/۶۲
۳	$CH_2C\equiv CH$	$-(CH_2)_4-$			۵/۲۲
۴	CH_2CH_2CN	$-(CH_2)_4-$			۵/۲۸
۵	CH_2Cl	$-(CH_2)_4-$			۵/۶۵
۶	CH_2CN	$-(CH_2)_4-$			۴/۷۵
۷	CH_2CN	$-(CH_2)_4-$			۵/۸۳
۸	CH_2CN	$-(CH_2)_4-$			۶/۲۶
۹	CH_2CN	$-(CH_2)_4-$			۵/۷۴
۱۰	CH_2CN	$-(CH_2)_4-$			۶/۶۲
۱۱	CH_2CN	$-(CH_2)_4-$			۵/۲۹
۱۲	CH_2CN	$-(CH_2)_4-$			۵/۰۷
۱۳	CH_2CN	$-(CH_2)_4-$			۴/۷۸
۱۴	CH_2CN	$-(CH_2)_4-$			۵/۲۴
۱۵	CH_2CN	$-(CH_2)_4-$			۵/۲۴
۱۶	CH_2CN	$-(CH_2)_4-$			۶/۰۵
۱۷	CH_2CN	$-(CH_2)_4-$			۵/۶۳
۱۸	Me	$-(CH_2)_4-$			۵/۸۰
۱۹	Me	$-(CH_2)_4-$			۴/۷۸

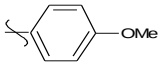
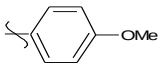
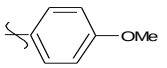
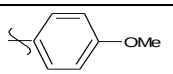
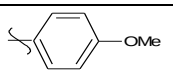
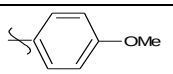
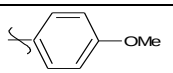
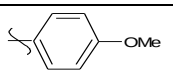
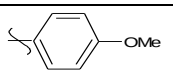
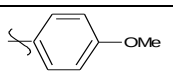
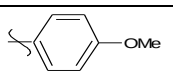
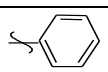
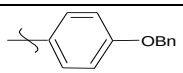
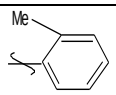
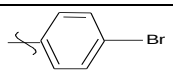
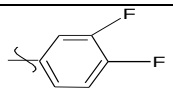
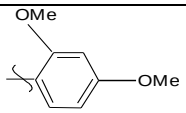
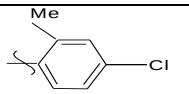
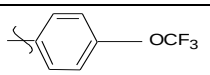
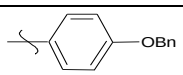
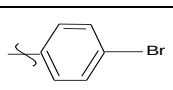
ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۲۰	Me	—$(CH_2)_4$—			۴/۷۸
۲۱	CH_2CN	Me	Me		۶/۳۵
۲۲	CH_2CN	Me	Et		۵/۲۸
۲۳	CH_2CN	Et	Me		۵/۹۷
۲۴	CH_2CN	—$(CH_2)_3$—			۶/۳۵
۲۵	CH_2CN	—$(CH_2)_5$—			۶/۰۲
۲۶	CH_2CN	-CH=CHCH=CH-			۶/۷۴
۲۷	Me	—$(CH_2)_3$—			۶/۶۲
۲۸	Me	-CH=CHCH=CH-			۶/۶۲
۲۹	$CH_2C\equiv CH$	—$(CH_2)_3$—			۵/۶۰
۳۰	CH_2CH_2OH	—$(CH_2)_3$—			۵/۱۵
۳۱	$CH_2C\equiv CH$	Me	Me		۵/۳۱
۳۲	CH_2CH_2OH	Me	Me		۵/۱۴
۳۳	CH_2CN	Me	Me		۵/۳۵
۳۴	CH_2CN	—$(CH_2)_3$—			۶/۰۵
۳۵	Me	Me	Me		۵/۵۸
۳۶	Me	—$(CH_2)_3$—			۵/۵۲
۳۷	CH_2CN	Me	Me		۵/۷۸
۳۸	CH_2CN	Et	Me		۵/۷۲
۳۹	CH_2CN	—$(CH_2)_3$—			۵/۹۳
۴۰	CH_2CN	—$(CH_2)_3$—			۶/۶۰

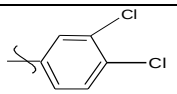
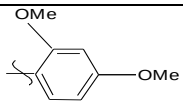
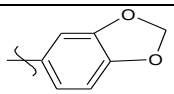
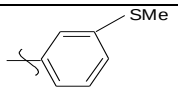
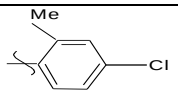
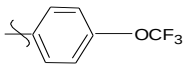
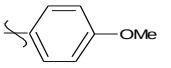
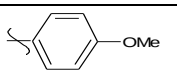
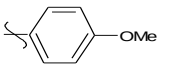
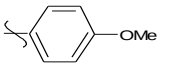
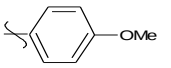
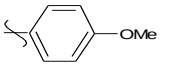
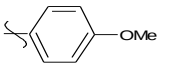
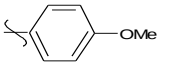
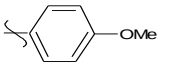
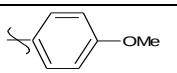
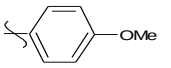
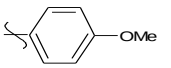
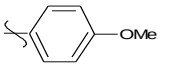
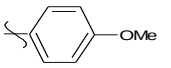
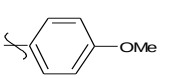
ادامه جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۴۱	CH_2CN	$-(CH_2)_4-$			۶/۶۶
۴۲	CH_2CN	-CH=CHCH=CH-			۶/۸۲
۴۳	CH_2CN	Me	Me		۶/۶۰
۴۴	CH_2CN	Me	Me		۵/۵۵
۴۵	CH_2CN	$-CH_2CH_2CH(CH_3)CH_2-$			۴/۷۸
۴۶	CH_2CN	-CH=CHCH=CH-			۶/۵۷
۴۷	Me	-CH=CHCH=CH-			۵/۵۱
۴۸	CH_2CN	Me	Me		۵/۱۱
۴۹	CH_2CN	Me	Me		۵/۴۳
۵۰	CH_2CN	$-(CH_2)_3-$			۵/۴۳
۵۱	Me	Me	Me		۵/۹۵
۵۲	Et	$-(CH_2)_4-$			غیرفعال
۵۳	*n*-Pr	$-(CH_2)_4-$			غیرفعال
۵۴	*n*-Bu	$-(CH_2)_4-$			غیرفعال
۵۵	*n*-Pent	$-(CH_2)_4-$			غیرفعال
۵۶	*n*-Hex	$-(CH_2)_4-$			غیرفعال
۵۷	$CH_2CH=CH_2$	$-(CH_2)_4-$			غیرفعال
۵۸	$CH_2C_6H_5$	$-(CH_2)_4-$			غیرفعال
۵۹	$CH_2(C_6H_4Me$-4)	$-(CH_2)_4-$			غیرفعال
۶۰	$CH_2(C_6H_4OMe$-4)	$-(CH_2)_4-$			غیرفعال

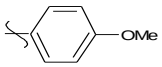
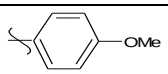
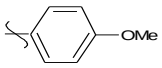
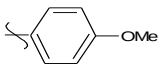
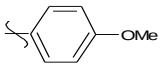
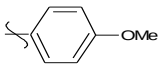
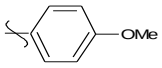
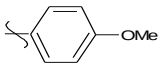
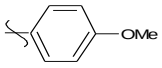
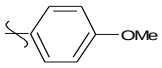
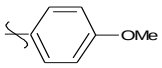
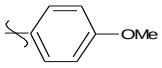
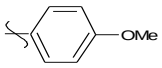
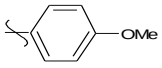
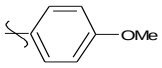
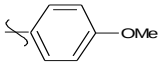
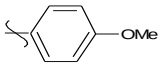
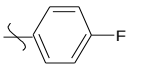
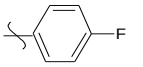
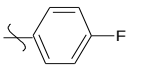
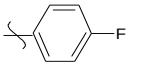
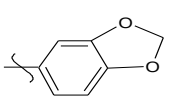
ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۶۱	$CH_2(C_6H_4NO_2$-4)	—$(CH_2)_4$—			غیرفعال
۶۲	CH_2COMe	—$(CH_2)_4$—			غیرفعال
۶۳	CH_2COOMe	—$(CH_2)_4$—			غیرفعال
۶۴	CH_2CONH_2	—$(CH_2)_4$—			غیرفعال
۶۵	COMe	—$(CH_2)_4$—			غیرفعال
۶۶	COC_3H_7-*n*	—$(CH_2)_4$—			غیرفعال
۶۷	COC_6H_5	—$(CH_2)_4$—			غیرفعال
۶۸	CH_2NO_2	—$(CH_2)_4$—			غیرفعال
۶۹	$CH_2C(O)NH(C_6H_4CF_3$-2)	—$(CH_2)_4$—			غیرفعال
۷۰	$CH_2CH(OH)CH_3$	—$(CH_2)_4$—			غیرفعال
۷۱	CH_2COOH	—$(CH_2)_4$—			غیرفعال
۷۲	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۳	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۴	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۵	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۶	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۷	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۸	CH_2CN	—$(CH_2)_4$—			غیرفعال
۷۹	CH_2CN	—$(CH_2)_4$—			غیرفعال
۸۰	Me	—$(CH_2)_4$—			غیرفعال
۸۱	Me	—$(CH_2)_4$—			غیرفعال

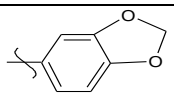
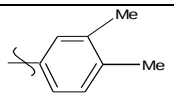
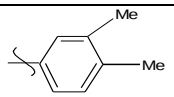
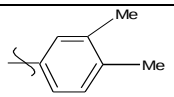
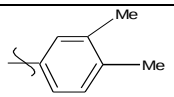
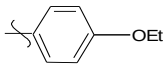
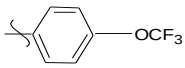
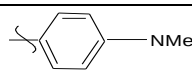
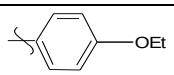
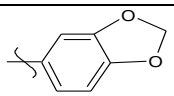
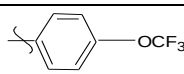
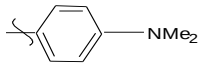
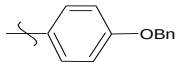
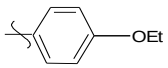
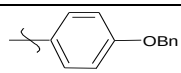
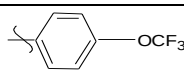
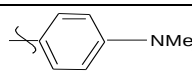
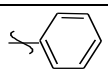
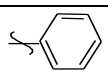
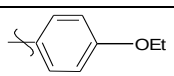
ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۸۲	Me	—$(CH_2)_4$—			غیرفعال
۸۳	Me	—$(CH_2)_4$—			غیرفعال
۸۴	Me	—$(CH_2)_4$—			غیرفعال
۸۵	Me	—$(CH_2)_4$—			غیرفعال
۸۶	Me	—$(CH_2)_4$—			غیرفعال
۸۷	Me	—$(CH_2)_4$—			غیرفعال
۸۸	CH_2CN	H	H		غیرفعال
۸۹	CH_2CN	H	Me		غیرفعال
۹۰	CH_2CN	H	Et		غیرفعال
۹۱	CH_2CN	Me	H		غیرفعال
۹۲	CH_2CN	Me	*n*-Pr		غیرفعال
۹۳	CH_2CN	Me	*i*-Pr		غیرفعال
۹۴	CH_2CN	Me	$C_{14}H_{29}$		غیرفعال
۹۵	CH_2CN	$-CH_2CH_2CHCH_3CH_2-$			غیرفعال
۹۶	CH_2CN	$-CH_2CH_2NEtCH_2-$			غیرفعال
۹۷	CH_2CN	$-CH_2CH_2N(i\text{-Pr})CH_2-$			غیرفعال
۹۸	Et	—$(CH_2)_3$—			غیرفعال
۹۹	n-Pr	—$(CH_2)_3$—			غیرفعال
۱۰۰	n-Bu	—$(CH_2)_3$—			غیرفعال
۱۰۱	n-Pent	—$(CH_2)_3$—			غیرفعال
۱۰۲	$CH_2CH=CH_2$	—$(CH_2)_3$—			غیرفعال

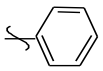
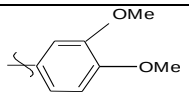
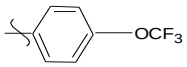
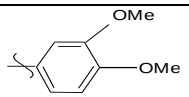
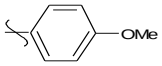
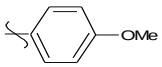
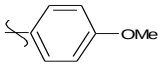
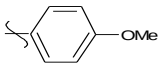
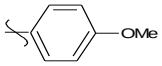
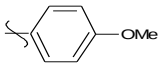
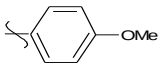
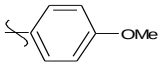
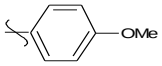
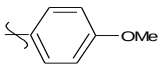
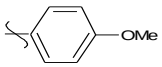
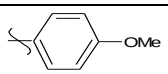
ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۱۰۳	$CH_2C_6H_5$	$-(CH_2)_3-$			غیرفعال
۱۰۴	$CH_2(C_6H_4NO_2$-4)	$-(CH_2)_3-$			غیرفعال
۱۰۵	CH_2COMe	$-(CH_2)_3-$			غیرفعال
۱۰۶	CH_2NO_2	$-(CH_2)_3-$			غیرفعال
۱۰۷	$CH_2C\equiv CH$	H	H		غیرفعال
۱۰۸	$CH_2C\equiv CH$	Me	H		غیرفعال
۱۰۹	$CH_2C\equiv CH$	Me	Et		غیرفعال
۱۱۰	$CH_2C\equiv CH$	Me	*n*-Pr		غیرفعال
۱۱۱	$CH_2C\equiv CH$	$-(CH_2)_5-$			غیرفعال
۱۱۲	Et	Me	Me		غیرفعال
۱۱۳	Et	Me	Et		غیرفعال
۱۱۴	Et	$-(CH_2)_5-$			غیرفعال
۱۱۵	CH_2CH_2CN	Me	Me		غیرفعال
۱۱۶	CH_2CH_2OH	Me	Et		غیرفعال
۱۱۷	CH_2CH_2OH	Me	*n*-Pr		غیرفعال
۱۱۸	Et	Me	*n*-Pr		غیرفعال
۱۱۹	$CH_2C(O)OH$	Me	Me		غیرفعال
۱۲۰	CH_2CN	Me	Et		غیرفعال
۱۲۱	CH_2CN	$-(CH_2)_5-$			غیرفعال
۱۲۲	CH_3	Me	Et		غیرفعال
۱۲۳	CH_3	$-(CH_2)_5-$			غیرفعال
۱۲۴	CH_3	Et	Me		غیرفعال

ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۱۲۵	CH_3	$—(CH_2)_3—$			غیرفعال
۱۲۶	CH_2CN	$—(CH_2)_3—$			غیرفعال
۱۲۷	CH_3	Me	Me		غیرفعال
۱۲۸	CH_3	$-CH_2CH_2CH(CH_3)CH_2-$			غیرفعال
۱۲۹	CH_3	$—(CH_2)_3—$			غیرفعال
۱۳۰	CH_2CN	-CH=CHCH=CH-			غیرفعال
۱۳۱	CH_2CN	-CH=CHCH=CH-			غیرفعال
۱۳۲	CH_2CN	-CH=CHCH=CH-			غیرفعال
۱۳۳	CH_3	-CH=CHCH=CH-			غیرفعال
۱۳۴	CH_3	-CH=CHCH=CH-			غیرفعال
۱۳۵	CH_3	-CH=CHCH=CH-			غیرفعال
۱۳۶	CH_2CN	Me	Me		غیرفعال
۱۳۷	CH_2CN	Me	Me		غیرفعال
۱۳۸	CH_3	Me	Me		غیرفعال
۱۳۹	CH_3	Me	Me		غیرفعال
۱۴۰	CH_3	Me	Me		غیرفعال
۱۴۱	CH_3	Me	Me		غیرفعال
۱۴۲	CH_2CN	H	H		غیرفعال
۱۴۳	CH_2CN	$—(CH_2)_3—$			غیرفعال
۱۴۴	CH_2CN	$—(CH_2)_3—$			غیرفعال

ادامه‌ی جدول (۴-۱)

No	R	R^1	R^2	R^3	$P(EC_{50}\times10^{-6})(M)$
۱۴۵	CH_2CN	—$(CH_2)_5$—			غیرفعال
۱۴۶	CH_2CN	$-CH_2CH_2CH(CH_3)CH_2-$			غیرفعال
۱۴۷	CH_2CN	—$(CH_2)_5$—			غیرفعال
۱۴۸	CH_2CN	—$(CH_2)_3$—			غیرفعال
۱۴۹	CH_3	H	H		غیرفعال
۱۵۰	CH_3	Me	H		غیرفعال
۱۵۱	CH_3	Me	Me		غیرفعال
۱۵۲	CH_3	Et	Me		غیرفعال
۱۵۳	CH_3	Me	Et		غیرفعال
۱۵۴	CH_3	Me	*n*-Pr		غیرفعال
۱۵۵	CH_3	Me	*i*-Pr		غیرفعال
۱۵۶	CH_3	Me	$C_{14}H_{29}$		غیرفعال
۱۵۷	CH_3	—$(CH_2)_5$—			غیرفعال
۱۵۸	CH_3	$-CH_2CH_2CHCH_3CH_2-$			غیرفعال
۱۵۹	CH_3	$-CH_2CH_2NEtCH_2-$			غیرفعال
۱۶۰	CH_3	$-CH_2CH_2N(i\text{-Pr})CH_2-$			غیرفعال

۴-۲- نرم‌افزارهای مورد استفاده

۴-۲-۱- بسته نرم‌افزاری Hyperchem

از بسته نرم‌افزاری Hyperchem [۴۱] برای رسم شکل مولکول‌ها و بهینه‌سازی ساختار با استفاده از روش‌های کوانتومی و مکانیکی، استفاده می‌شود. به کمک این نرم‌افزار، ساختار ترکیبات به طور تقریبی رسم شد، در ادامه با اضافه کردن اتم‌های هیدروژن و انتخاب گزینه‌ی بهینه‌سازی، شکل هندسی مولکول بهینه شد.

۴-۲-۲- بسته نرم‌افزاری Dragon

نرم‌افزار Dragon که توسط گروه تحقیقاتی کمومتریکس میلانو ارائه شده [۴۲]، امکان محاسبه ۱۴۸۱ توصیف‌گر مختلف را فراهم می‌کند. جهت محاسبه توصیف‌گر به کمک این نرم‌افزار 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، hin، ... به عنوان اطلاعات ورودی به کار برد.

۴-۲-۳- بسته نرم‌افزاری SPSS[1]

این نرم‌افزار به منظور تجزیه و تحلیل داده‌ها تدوین شده است و یکی از قدرتمندترین ابزارها برای تحلیل آماری داده‌ها به شمار می‌آید. یادگیری آسان، سادگی استفاده و قدرت بالا در انجام محاسبات پیچیده، SPSS را به یکی از رایج‌ترین بسته‌های نرم‌افزاری مبدل نموده است. در این تحقیق از این نرم‌افزار برای بدست آوردن بهترین مدل استفاده شده است.

1- Statistical Package for the Social Science

۴-۲-۴- نرم افزار MATLAB

از محیط نرم‌افزار متلب برای استفاده از الگوریتم طبقه‌بندی جنگل‌های تصادفی و شبکه عصبی مصنوعی بهره گرفته شده است.

۴-۳- رسم و بهینه‌سازی ساختار مولکول‌ها

ساختار دوبعدی مشتقات Nec-5 در نرم افزار Hyperchem 8 ترسیم شدند و ساختار نهایی به روش نیمه‌تجربی و با استفاده از روش AM1 بهینه گردید. بهینه‌سازی تا زمانی ادامه یافت که جذر میانگین مربعات گرادیان[1] انرژی به ۰/۰۰۱ کیلوکالری بر مول برسد.

۴-۴- طبقه‌بندی ترکیبات نکروستاتین-۵

همانطور که بیان شد، بعضی از ترکیبات که به طور تجربی سنتز شده‌اند، فاقد فعالیت مهارکنندگی می‌باشند. لذا در مرحله اول تحقیق، هدف ارائه مدلی است که قادر به تقسیم‌بندی ترکیبات به دو گروه فعال و غیرفعال باشد. این کار بدان جهت انجام گرفت که اگر ترکیبی از این خانواده وارد مدل شود، مدل نوع فعالیت آن را پیش‌بینی نماید، سپس رابطه‌ی کمی ساختار – فعالیت pEC_{50} ترکیبات فعال، بررسی شود. برای ایجاد مدل طبقه‌بندی از برنامه جنگل‌های تصادفی در متلب استفاده شد. درانتخاب توصیف‌گر و در مدل‌سازی RF، ترکیبات فعال با کد 1 ($pEC_{50}=1$) و ترکیبات غیرفعال با کد 0 ($pEC_{50}=0$) در نظر گرفته شدند که در ادامه شرح داده می‌شود.

1 - Root Mean Square Gradient

۴-۴-۱- محاسبه توصیف‌گرها

ساختارهای رسم شده برای ۱۶۰ ترکیب موجود به عنوان ورودی به نرم افزار Dragon داده شد و ۱۴۸۱ توصیف‌گر برای این ساختارهای بهینه محاسبه گردید. با توجه به اینکه تعداد زیاد توصیف‌گر باعث پیچیدگی محاسبات می‌شود، لذا باید تعداد توصیف‌گرها به نوعی کاهش یابد. به همین منظور، ابتدا توصیف‌گرهایی که برای تمام مولکول‌ها مقادیر ثابت داشتند، حذف شدند. سپس یکی از هر دو توصیف‌گری که همبستگی بزرگتر از ۰/۹ داشتند، آنکه همبستگی کمتری با متغیر وابسته داشت، با استفاده از برنامه‌ی Remcol در نرم‌افزار متلب حذف گردید. در پایان این مرحله ۴۵۵ توصیف‌گر باقی ماند.

۴-۴-۲- مدل‌سازی و بهینه‌سازی پارامترهای مؤثر بر قدرت جنگل‌های تصادفی

برای استفاده از قابلیت انتخاب توصیف‌گر و انجام مدل‌سازی توسط جنگل‌های تصادفی، ابتدا باید پارامترهای مؤثر بهینه شوند. برای انجام این کار ابتدا ترکیبات به دو سری آموزش(۱۲۰ترکیب) و سری آزمون (۴۰ترکیب) تقسیم شدند. مقادیر مربوط به ۴۵۵ توصیف‌گر به عنوان متغیر مستقل و مقادیر pEC_{50} با کد صفر و یک به عنوان متغیر وابسته برای برنامه تعریف شدند. داده‌های سری OOB نیز برای بهینه‌سازی تعداد درختان (n_{tree})، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (Mtry) در نظر گرفته شدند و مقدار مشاهدات باقی‌مانده در هر گره (Node size) ثابت و برابر یک قرار داده شد. برای بهینه نمودن، تعداد درخت از ۱۰۰ تا ۵۰۰ با گام ۱۰۰، تعداد توصیف‌گرهای انتخاب شده در هر مرحله افراز (Mtry) از ۲ تا ۴۵۵ با گام یک تغییر داده شد و در هر مرحله مقدار خطای طبقه‌بندی مربوط به مجموعه OOB محاسبه گردید. جدول (۴-۲) مقادیر Mtry، n_{tree} همراه با خطای طبقه‌بندی متناظر آن‌ها را در چندین حالت که کمترین خطا برای OOB ایجاد شده، نشان می‌دهد.

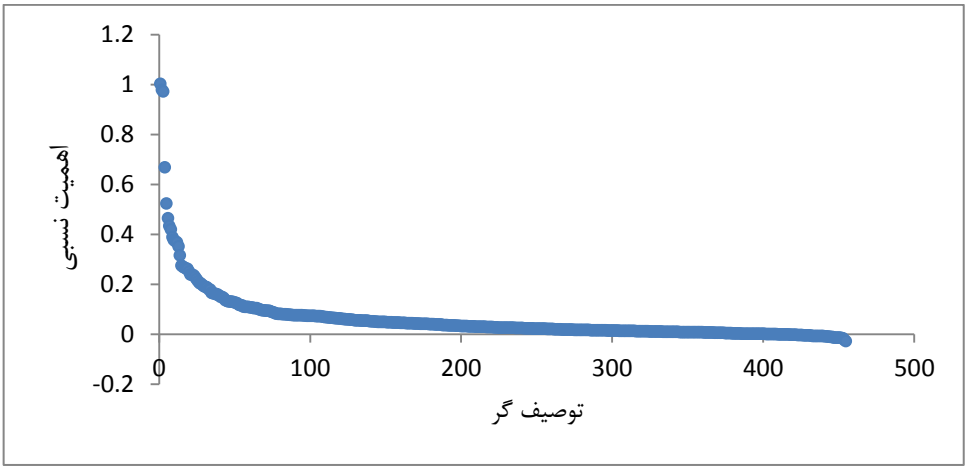
جدول (۴-۲)- نتایج بهینه‌سازی پارامترهای موثر

ntree	Mtry	Node size	Error OOB
۱۰۰	۱۰۸	۱	۰/۲۰۸
۲۰۰	۷۵	۱	۰/۱۷۵
۳۰۰	**۲۱**	**۱**	**۰/۱۵۸**
۴۰۰	۲۲۴	۱	۰/۱۷۵
۵۰۰	۳۴۱	۱	۰/۱۸۳۳

n_{tree}=300 و Mtry=21 به عنوان مقدار بهینه انتخاب شد.

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

پس از ساخت مدل جنگل‌های تصادفی و بهینه‌کردن پارامترهای مؤثر، الگوریتم جنگل‌های تصادفی با استفاده از مقادیر بهینه صد بار تکرار شد و در هر بار میزان کاهش در دقت طبقه‌بندی به عنوان مقدار اهمیت توصیف‌گرها محاسبه گردید. سرانجام میانگین مقادیر اهمیت هر توصیف‌گر به عنوان شاخص نهایی در نظر گرفته شد و سپس اهمیت نسبی هر توصیف‌گر نسبت به توصیف‌گری که بیشترین اهمیت را دارا است، تعیین شد. شکل (۴-۲) میزان اهمیت نسبی توصیف‌گرها را نمایش می‌دهد و همان‌طور که مشاهده می‌شود، بسیاری از توصیف‌گرها دارای اهمیت نسبی کمی بوده و تنها ۹ توصیف‌گر اهمیت نسبی قابل توجهی دارند که در جدول (۴-۳) این توصیف‌گرها همراه با مقدار اهمیت نسبی آن‌ها گردآوری شده است. هم‌چنین جدول (۴-۴) عدم همبستگی توصیف‌گرها را نشان می‌دهد.



شکل(۴-۲) - اهمیت نسبی توصیف‌گرها

جدول (۴-۳)- توصیف‌گرهای انتخاب شده با روش جنگل‌های تصادفی

شماره	توصیف‌گر	طبقه	معنی	اهمیت نسبی
۱	RDF100m	RDF	Radial Distribution Function - 100 / weighted by mass	۱
۲	HATS8v	GETAWAY	leverage-weighted autocorrelation of lag 8 / weighted by van der Waals volume	۰/۹۸
۳	Mor09e	3D-MoRSE	signal 09 / weighted by Sanderson electronegativity	۰/۹۷
۴	Mor08u	3D-MoRSE	signal 08 / unweighted	۰/۶۶
۵	RPCG	Charge	relative positive charge	۰/۵۲
۶	HATS0m	GETAWAY	leverage-weighted autocorrelation of lag 0 / weighted by mass	۰/۴۶
۷	HATS8m	GETAWAY	leverage-weighted autocorrelation of lag 8 / weighted by mass	۰/۴۳
۸	R7p	GETAWAY	R autocorrelation of lag 7 / weighted by polarizability	۰/۴۲
۹	Mor09v	3D-MoRSE	signal 09 / weighted by van der Waals volume	۰/۳۸

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

	RDF100m	HATS8v	Mor09e	Mor08u	RPCG	HATS0m	HATS8m	R7P	Mor09v
RDF100m	۱								
HATS8v	۰/۵۳۰	۱							
Mor09e	-۰/۰۷۸	-۰/۲۸۳	۱						
Mor08u	۰/۲۶۵	-۰/۵۹۵	۰/۵۰۸	۱					
RPCG	-۰/۶۵۰	۰/۶۸۵	۰/۰۶۵	-۰/۳۹۹	۱				
HATS0m	-۰/۲۹۹	۰/۵۴۶	۰/۱۳۴	-۰/۴۰۱	۰/۳۷۴	۱			
HATS8m	-۰/۳۱۰	۰/۸۱۸	۰/۲۸۸	-۰/۴۶۷	۰/۴۶۷	۰/۷۵۵	۱		
R7P	-۰/۳۹۳	۰/۸۰۴	۰/۲۰۲	-۰/۴۷۵	۰/۴۶۵	۰/۵۷۸	۰/۷۵۷	۱	
Mor09v	-۰/۲۱۸	۰/۰۳۴	۰/۷۵۹	۰/۱۸۷	۰/۲۴۵	۰/۰۲۲	۰/۰۵۷	-۰/۰۸۸	

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

از آنجایی که در مطالعات QSAR، هر چه تعداد توصیف‌گرهای به کار رفته در ساخت مدل کمتر باشد، می‌توان آن را مدل بهتری دانست لذا ایجاد مدل با روش جنگل‌های تصادفی فقط با توصیف‌گرهای مهم که در جدول (۴-۳) آمده است، نیز مجددا انجام شد. به این صورت که ابتدا مقادیر مربوط به ۹ توصیف‌گر به عنوان ورودی به کار برده شد و بهینه‌سازی پارامتر n_{tree} و Mtry انجام شد که n_{tree}=100 و Mtry=2 به عنوان مقدار بهینه به دست آمد.

۴-۴-۵-ارزیابی RF برای طبقه‌بندی ترکیبات نکروستاتین-۵ با استفاده از سری آزمون

بعد از انتخاب مهم‌ترین توصیف‌گرها با استفاده از سری آموزش، مدل RF با ۹ توصیف‌گر برتر ساخته شد و قدرت پیش‌بینی RF، بوسیله ۴۰ داده‌ی سری آزمون که در آموزش استفاده نشده‌اند، مورد بررسی قرار گرفت. شاخص‌های ارزیابی مدل طبق روابط بخش (۴-۲) محاسبه شد. نتایج حاصله در جدول (۴-۵) نشان داده شده است. نتایج بیان‌گر این است که RF توانمندی قابل قبولی برای طبقه‌بندی مشتقات نکروستاتین-۵ دارد.

جدول (۴-۵)- نتایج حاصل از ارزیابی مدل با استفاده از سری آزمون

شاخص های ارزیابی	FC(نسبت صحیح)	FAR(نسبت هشدار)	POD(احتمال آشکارسازی)
سری آزمون	٪۸۰	٪۱۲	٪۷۳

۴-۵- مجموعه داده‌های مولکول‌های فعال

با توجه به اینکه ۵۱ ترکیب از مشتقات نکروستاتین-۵ در طبقه مولکول‌های فعال قرار دارند، در این بخش رابطه ساختار-فعالیت این مولکول‌ها توسط شبکه عصبی مورد بررسی قرار گرفت. ساختار و قدرت بازدارندگی (EC_{50}) این مشتقات در جدول (۴-۱) نشان داده شده است.

۴-۵-۱-محاسبه‌ی توصیف‌گرهای مولکولی

برای محاسبه توصیف‌گرها، ساختار بهینه‌شده‌ی ۵۱ مولکول فعال به نرم‌افزار Dragon منتقل شد و تمام ۱۴۸۱ توصیف‌گر برای هر مولکول فعال محاسبه گردید. برای انتخاب بهترین توصیف‌گرها از دو روش استفاده شده است که در ادامه شرح داده می‌شود.

۴-۵-۲-انتخاب توصیف‌گرهای مهم با استفاده از روش جنگل‌های تصادفی و رگرسیون مرحله‌ای(RF-SR)

انتخاب توصیف‌گرهای مهم که داده‌های EC_{50} را به ساختار مولکولی ارتباط دهند، مهم‌ترین مرحله در مطالعات QSAR می‌باشد. جهت انتخاب توصیف‌گرهای مناسب که ارتباط بیشتری با داده‌های EC_{50} داشته باشند و بتوانند به عنوان ورودی‌های شبکه عصبی مصنوعی مورد استفاده قرار گرفته شوند، مراحل زیر به ترتیب در کاهش تعداد متغیر و استخراج توصیف‌گرهای مؤثر بکار گرفته شد:

الف) ابتدا ترکیبات به دو سری آموزش (۴۱ترکیب) و سری آزمون (۱۰ترکیب) تقسیم شدند. برنامه جنگل‌های تصادفی با سری آموزش (۴۱ترکیب) بهینه گردید، نتیجه بهینه‌سازی در

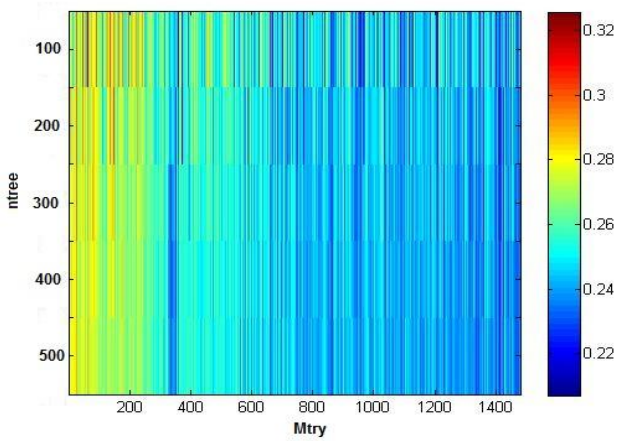
شکل (۴-۳) نمایش داده شده است و نقاط آبی خطای کمتری را نشان می‌دهند. با توجه به شکل مقادیر بهینه n_{tree}=500 و Mtry=375 محاسبه شد.

ب) از قابلیت محاسبه اهمیت متغیر در برنامه RF استفاده شد و برنامه با مقادیر بهینه صد بار تکرار شد و اهمیت هر متغیر محاسبه گردید و در پایان اهمیت نسبی متغیرها محاسبه شد.

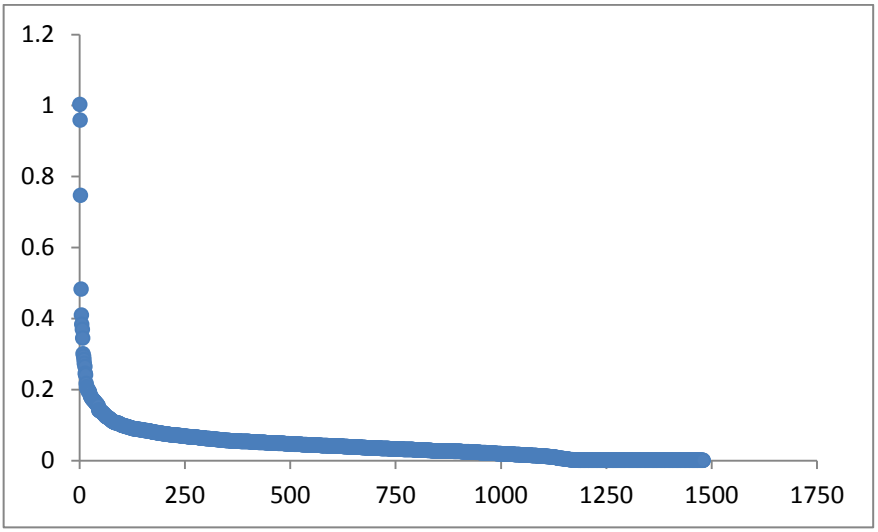
ج) نمودار اهمیت نسبی متغیرها ترسیم شد شکل (۴-۴) و متغیرهایی که اهمیت آنها صفر بود حذف گردید. هدف از انجام این کار کاهش و حذف متغیرهایی است که برای تمام مولکول‌ها یکسان است و یا اهمیت کم‌تری دارند. در این مرحله ۷۱۵ توصیف‌کننده باقی ماند.

د) برای انتخاب توصیف‌گرهای مناسب و مدل‌سازی، روش رگرسیون خطی مرحله‌ای با استفاده از ۷۱۵ توصیف‌گر باقی‌مانده، در نرم‌افزار SPSS انجام گردید. نتایج حاصله نشان داد که از این تعداد زیاد توصیف‌گر، فقط تعداد ۱۲ توصیف‌گر ارتباط معنی‌داری با داده‌های EC_{50} ترکیبات مورد بررسی داشته‌اند.

جدول (۴-۶) توصیف‌گرهای انتخاب شده و طبقه مربوط به آنها را نشان می‌دهد. همچنین ماتریس ضریب همبستگی این توصیف‌گرها که در جدول (۴-۷) ارائه شده، عدم همبستگی قابل ملاحظه بین توصیف‌گرها را نشان می‌دهد.



شکل(۴-۳)- بهینه سازی پارامترهای ntree و Mtry



شکل(۴-۴)-اهمیت نسبی کل توصیف‌گرها

جدول (۴-۶) - کل توصیف‌گرهای انتخاب شده توسط روش RF-SR

تعداد	نماد	طبقه	معانی
۱	H7u	GETAWAY	H autocorrelation of lag 7 / unweighted
۲	R8u	GETAWAY	R autocorrelation of lag 8 / unweighted
۳	Mor11u	3D-MoRSE	signal 11 / unweighted
۴	MLOGP	Molecular	Moriguchi octanol-water partition coeff. (logP)
۵	Uindex	topological	Balaban U index
۶	GO..S	Geometry	sum of geometrical distances between O..S
۷	BEHm6	BCUT	Highest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
۸	RDF140e	RDF	Radial Distribution Function - 140 / weighted by Sanderson electronegativity
۹	HVcpx	topological	graph vertex complexity index
۱۰	RDF075v	RDF	Radial Distribution Function - 075 / weighted by van der Waals volume
۱۱	GATS7e	2D autocorrelations	Geary autocorrelation of lag 7 weighted by Sanderson electronegativity
۱۲	H4e	GETAWAY	H autocorrelation of lag 4 / weighted by Sanderson electronegativity

جدول (۴- ۷) - ماتریس همبستگی کل توصیف‌گرهای انتخاب شده توسط روش RF-SR

	H7u	R8u	Mor11u	MLOGP	Uindex	GO..S	BEHm6	RDF140e	HVcpx	RDF075v	GATS7e	H4e
H7u	۱											
R8u	/۲۳۵ ۰	۱										
Mor11u	/۰۲۲ -۰	۰/۱۵۵	۱									
MLOGP	/۲۲۴ ۰	۰/۰۲۸ -	۰/۱۴۸	۱								
Uindex	/۲۱۸ ۰	۰/۰۸۱ -	-۰/۰۲۴	-۰/۴۲۹	۱							
GO..S	/۱۸۲ -۰	۰/۲۷۶ -	-۰/۲۰۷	-۰/۷۴۲	۰/۲۰۲	۱						
BEHm6	/۳۲۶ ۰	۰/۰۱۹ -	۰/۰۴۳	-۰/۱۶۳	۰/۴۵۲	۰/۲۲۷	۱					
RDF140e	/۰۲۰ ۰	۰/۴۴۸ -	-۰/۱۳۸	-۰/۱۹۴	۰/۵۰۸	۰/۱۶۹	۰/۳۱۲	۱				
HVcpx	/۱۹۲ ۰	۰/۵۵۳ -	-۰/۰۱۹	-۰/۲۳۵	۰/۶۵۳	۰/۳۹۳	۰/۵۹۵	۰/۶۴۲	۱			
RDF075v	/۱۱۴ ۰	۰/۱۸۷	-۰/۰۶۸	-۰/۵۴۷	۰/۷۳۰	۰/۳۴۲	۰/۴۴۶	۰/۲۹۴	۰/۳۵۰	۱		
GATS7e	/۳۸۸ ۰	۰/۱۷۲	-۰/۰۹۵	۰/۲۶۱	۰/۰۶۶	۰/۲۴۳ -	۰/۱۴۹	۰/۰۹۳	۰/۰۶۲	۰/۲۷۶	۱	
H4e	/۱۸۷ ۰	۰/۲۲۶ -	-۰/۲۷۲	-۰/۰۵۲	۰/۵۳۹	۰/۱۵۳	۰/۲۲۵	۰/۴۶۱	۰/۵۸۶	۰/۴۴۴	۰/۵۸۴	**۱**

۴-۵-۳- انتخاب بهترین توصیف‌گرها با استفاده از روش رگرسیون مرحله‌ای(SR)

با به کارگیری نرم‌افزار SPSS و اجرای رگرسیون مرحله‌ای برای ۱۴۸۱ توصیف‌گر یاد شده به عنوان متغیرهای مستقل، و فعالیت بازدارندگی به عنوان متغیر وابسته، تعداد ۱۹ توصیف‌گر انتخاب شدند که نام و طبقه‌ی آن‌ها در جدول (۴-۸) نشان داده شده است. همچنین ماتریس ضریب همبستگی این توصیف‌گرها که در جدول (۴-۹) ارائه شده، عدم همبستگی قابل ملاحظه بین توصیف‌گرها را نشان می‌دهد.

جدول (۴-۸) - کل توصیف‌گرهای انتخاب شده توسط روش رگرسیون مرحله‌ای

تعداد	نماد	طبقه	معانی
۱	H7u	GETAWAY	H autocorrelation of lag 7 / unweighted
۲	R8u	GETAWAY	R autocorrelation of lag 8 / unweighted
۳	n#CH	Functional group	number of terminal C(sp)
۴	Mor11e	3D-MoRSE	signal 11 / weighted by Sanderson electronegativity
۵	Mor23m	3D-MoRSE	signal 23 / weighted by mass
۶	RDF130m	RDF	Radial Distribution Function - 130 / weighted by mass
۷	BELm5	BCUT	lowest eigenvalue n. 5 of Burden matrix / weighted by atomic masses
۸	G2m	WHIM	2nd component symmetry directional WHIM index / weighted by mass
۹	BEHm6	BCUT	highest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
۱۰	BEHm7	BCUT	highest eigenvalue n. 7 of Burden matrix / weighted by atomic masses
۱۱	Mor04p	3D-MoRSE	signal 04 / weighted by polarizability
۱۲	PJI2	Topological	2D Petitjean shape index
۱۳	GATS7e	2D autocorrelations	Geary autocorrelation of lag 7 weighted by Sanderson electronegativity
۱۴	BEHe2	BCUT	highest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities
۱۵	G1m	WHIM	1st component symmetry directional WHIM index / weighted by mass
۱۶	G1p	WHIM	1st component symmetry directional WHIM index / weighted by polarizability
۱۷	piPC04	Walk and path counts	molecular multiple path count of order 4
۱۸	R2v_A	GETAWAY	R maximal autocorrelation of lag 2 / weighted by van der Waals volume
۱۹	GS..S	Geometrical	sum of geometrical distances between S..S

جدول (۴- ۹) - ماتریس هم‌بستگی کل توصیف‌گرهای انتخاب شده توسط روش SR

	H7u	R8u	nCH	Mor11e	Mor23m	RDF130m	BELm5	G2m	BEHm6	BEHm7	Mor04p	Pji2	GATS7e	BEHe2	G1m	G1p	pipC04	R2V_A	GS..S
H7u	۱																		
R8u	۰/۲۲۵	۱																	
n#CH	-۰/۱۲۱	-۰/۰۱۴	۱																
Mor11e	-۰/۰۶۳	۰/۱۲۱	۰/۰۵۹	۱															
Mor23m	۰/۰۱۷	۰/۳۹۳	-۰/۰۵۷	-۰/۱۶۷	۱														
RDF130m	۰/۰۹۷	-۰/۳۰۰	-۰/۰۷۱	-۰/۲۴۸	۰/۱۵۴	۱													
BELm5	۰/۰۰۷	-۰/۰۶۴	۰/۰۲۰۲	-۰/۲۴۹	۰/۱۳۴	۰/۲۵۶	۱												
G2m	-۰/۳۹۰	-۰/۰۳۹	-۰/۰۹۸	۰/۱۳۳	-۰/۰۳۱	-۰/۰۷۱	-۰/۴۳۹	۱											
BEHm6	۰/۲۲۶	-۰/۰۱۹	۰/۰۱۱	۰/۰۱۶	۰/۱۲۲	۰/۳۱۵	۰/۱۸۴	-۰/۴۰۶	۱										
BEHm7	۰/۲۴۰	۰/۰۴۷	۰/۰۹۳	-۰/۱۲۸	۰/۲۰۸	۰/۲۲۳	۰/۳۷۱	-۰/۵۵۰	۰/۷۹۶	۱									
Mor04p	۰/۲۸۸	۰/۵۰۶	-۰/۰۰۲	-۰/۲۸۱	-۰/۳۵۰	۰/۴۳۹	۰/۱۹۴	-۰/۲۳۳	۰/۲۹۷	۰/۳۸۵	۱								
PJI2	-۰/۲۴۶	۰/۳۷۷	-۰/۰۷۰	-۰/۲۱۹	-۰/۱۸۵	۰/۲۰۱	۰/۱۲۹	۰/۱۵۴	-۰/۰۳۹	-۰/۱۱۶	۰/۲۱۰	۱							
GATS7e	۰/۳۸۸	۰/۱۷۲	-۰/۰۸۵	-۰/۱۱۵	۰/۰۹۷	۰/۲۲۲	۰/۲۲۹	-۰/۳۱۸	۰/۱۴۹	۰/۲۵۱	۰/۳۴۵	-۰/۱۹۴	۱						
BEHe2	۰/۰۰۳	۰/۱۳۹	-۰/۱۲۹	-۰/۱۲۵	۰/۱۶۲	۰/۱۲۴	۰/۱۵۱	-۰/۲۰۸	۰/۲۹۲	۰/۳۰۳	-۰/۰۲۵	۰/۰۵۳	۰/۲۸۲	۱					
G1m	-۰/۳۰۳	۰/۱۰۲	-۰/۰۹۶	۰/۰۸۵	۰/۱۲۱	-۰/۲۹۷	-۰/۳۵۹	۰/۳۷۹	-۰/۳۴۳	-۰/۳۹۲	-۰/۳۳۰	۰/۱۱۳	-۰/۵۲۵	-۰/۲۳۳	۱				
G1p	-۰/۱۳۶	۰/۳۴۶	-۰/۰۹۱	۰/۳۴۶	-۰/۰۰۵	-۰/۴۲۳	-۰/۵۱۰	۰/۳۹۳	-۰/۳۷۰	-۰/۴۱۶	-۰/۴۶۸	۰/۲۷۹	-۰/۱۷۰	-۰/۲۲۰	۰/۴۰۷	۱			
piPC04	۰/۰۲۹	-۰/۳۰۱	-۰/۱۳۱	-۰/۰۲۲	۰/۰۱۴	۰/۳۸۷	۰/۰۴۱	-۰/۰۹۸	۰/۳۹۶	۰/۳۶۵	۰/۲۶۶	۰/۰۷۸	۰/۲۳۰	۰/۸۴۷	-۰/۲۲۶	-۰/۲۱۲	۱		
R2v_A	-۰/۱۹۴	-۰/۲۹۷	-۰/۰۷۷	۰/۰۴۶	۰/۳۰۱	-۰/۱۸۰	-۰/۶۲۹	۰/۴۸۰	-۰/۱۴۴	-۰/۲۲۳	-۰/۴۵۹	-۰/۰۸۶	-۰/۱۴۱	۰/۱۰۸	۰/۳۵۹	۰/۴۶۹	۰/۱۰۰	۱	
GS..S	-۰/۰۴۲	-۰/۱۴۸	-۰/۰۳۵	۰/۴۰۰	۰/۰۰۵	۰/۰۷۳	-۰/۰۲۱	-۰/۰۸۲	۰/۲۲۴	۰/۱۳۵	۰/۲۰۷	۰/۱۲۶	۰/۰۱۴	۰/۰۶۰	-۰/۰۷۸	-۰/۰۷۵	-۰/۱۵	۰/۱۱۰	۱

۴-۵-۴- مدل‌سازی به روش شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش RF-SR

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 ۱۲ توصیف‌گر انتخاب شده با روش RF-SR به عنوان متغیرهای ورودی و فعالیت بازدارندگی 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 و مقدار mu) بهینه‌سازی شدند.

۴-۵-۴-۱- بهینه‌سازی پارامترهای شبکه

در فرآیند بهینه‌سازی پارامترهای شبکه، سری داده‌ها به طور تصادفی به مجموعه آزمون (۱۰ ترکیب) و مجموعه آموزش (۴۱ 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[۴۳] 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 ۱۲ 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 آموزش داده شد. در تمامی این شبکه‌ها تعداد دورهای آموزش ثابت و برابر ۲۰ در نظر گرفته شد. همچنین از تابع انتقال خطی (purelin) در لایه خروجی استفاده گردید. در روند بهینه‌سازی پارامترهای فوق، به حداقل

رساندن میانگین مربع خطاهای حاصل از ارزیابی تقاطعی سری آموزش به عنوان معیار انتخاب شد. نتایج بهینه‌سازی این پارامترها در جدول (۴-۱۰)، گردآوری شده است.

جدول(۴-۱۰)- توابع و پارامترهای شبکه های بهینه RF-SR-ANN

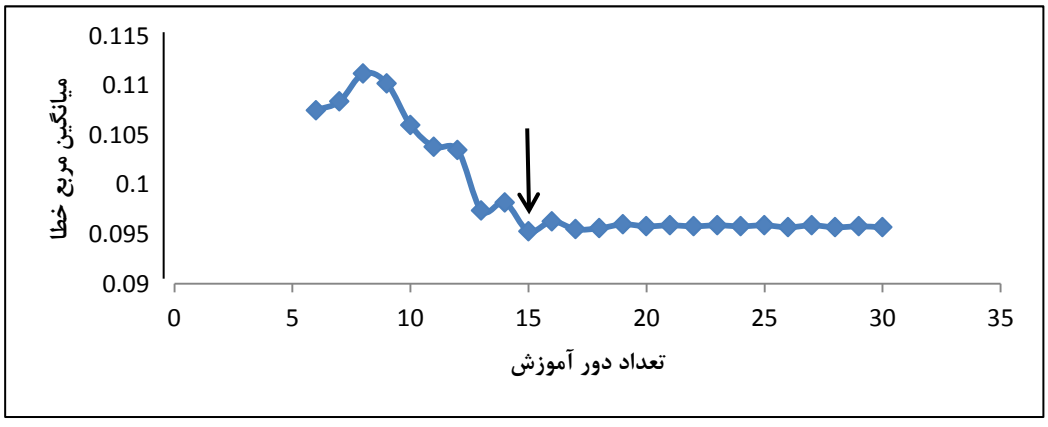
MSE	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف گر
۰/۱۴۵۵	۲	لگاریتم-سیگموئید	لونبرگ-مارکوات	۱۲
۰/۱۹۸۲	۱۰	تانژانت-سیگموئید	لونبرگ-مارکوات	۱۲
۰/۱۰۲	۳	لگاریتم-سیگموئید	تنظیم بایزین	۹
۰/۰۹۵۸	**۲**	**تانژانت-سیگموئید**	**تنظیم بایزین**	**۹**

با توجه به نتایج به دست آمده، الگوریتم آموزشی تنظیم بایزین و تابع انتقال تانژانت سیگموئید، ۹ توصیف‌گر ورودی و ۲ نرون در لایه مخفی کم‌ترین MSE را نشان می‌دهد. بنابراین این شبکه برای مدل‌سازی در نظر گرفته شدکه جدول (۴-۱۱) لیست این توصیف‌گرها به همراه طبقه آنها را نشان می‌دهد. در مرحله بعد تعداد دورهای آموزش بهینه شد برای این منظور شبکه بهینه به‌دست آمده از مرحله قبل در تعداد دورهای متفاوت (۲ تا ۳۰) آموزش داده شد و میانگین مربع خطاها برای سری آموزش ثبت گردید. شکل (۴-۴) نشان می‌دهد که با افزایش تعداد دورهای آموزش میانگین مربع خطا (MSE) کاهش می‌یابد تا جایی که در ۱۵ دور تکرار به حداقل خطا می‌رسد و سپس با افزایش بیشتر تعداد دورهای یادگیری، خطاها افزایش می‌یابد بنابراین ۱۵ بار تکرار به عنوان تعداد تکرارهای بهینه برای یادگیری شبکه انتخاب شد. جهت بهینه کردن مقدار mu، ساختار شبکه با ۹ متغیر ورودی، ۲ گره در لایه پنهان و الگوریتم آموزشی تنظیم بایزین در نظر گرفته شد. سپس مقدار mu از ۰/۰۴ تا ۰/۰۵ با گام‌های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری

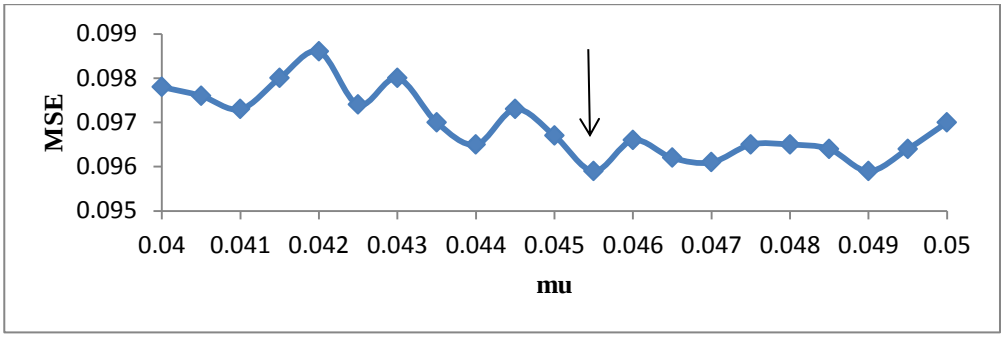
آموزش به روش ارزیابی تقاطعی محاسبه گردید. در نهایت مقدار میانگین مربع خطا بر حسب mu رسم شد. طبق شکل (۴- ۵) مقدار بهینه ۰/۰۴۵۵ به‌دست آمد.

جدول (۴-۱۱) – توصیف‌گرهای انتخاب شده توسط شبکه عصبی

No	Symbol	Class	Meaning
۱	H7u	GETAWAY	H autocorrelation of lag 7 / unweighted
۲	R8u	GETAWAY	R autocorrelation of lag 8 / unweighted
۳	Mor11u	3D-MoRSE	signal 11 / unweighted
۴	MLOGP	Molecular	Moriguchi octanol-water partition coeff. (logP)
۵	Uindex	Topological	Balaban U index
۶	GO..S	Geometry	sum of geometrical distances between O..S
۷	BEHm6	BCUT	Highest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
۸	RDF140e	RDF	Radial Distribution Function - 140 / weighted by Sanderson electronegativity
۹	HVcpx	Topological	graph vertex complexity index



شکل (۴- ۵)– نمودار میانگین مربع خطای حاصل از ارزیابی تقاطعی سری آموزش در تعداد دورهای آموزش متفاوت



شکل (۴- ۶) – نمودار میانگین مربع خطای حاصل از ارزیابی تقاطعی سری آموزش بر حسب پارامتر mu

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

نتایج به دست آمده در بهینه‌سازی شبکه، در جدول (۴-۱۲) نشان داده شده است.

جدول (۴- ۱۲) – مشخصات شبکه عصبی مصنوعی بهینه(RF-SR-ANN)

تابع آموزش	train br
تابع انتقال لایه پنهان	tansig
تابع انتقال لایه خروجی	pure line
تعداد نرون لایه پنهان	۲
تعداد متغیرهای ورودی	۹
تعداد دورهای آموزش	۱۵
پارامتر mu	۰/۰۴۸۵

۴-۵-۵- مدل‌سازی به روش شبکه عصبی مصنوعی با توصیف‌گرهای انتخاب شده توسط روش SR

در این بخش توصیف‌گرهای انتخاب شده با روش رگرسیون مرحله‌ای به عنوان توصیف‌گرهای ورودی به یک شبکه عصبی با یک لایه پنهان و لایه خروجی purlin مورد استفاده قرار گرفت. برای یافتن مقادیر بهینه پارامترهای مؤثر بر کارایی شبکه، تعداد ورودی از ۲ تا ۱۹ و سایر

پارامترها همانند قبل تغییر داده شد. میانگین مربع خطا نیز به عنوان معیار شبکه بهینه در نظر گرفته شد.

۴-۵-۵-۱- بهینه‌سازی پارامترهای مؤثر بر شبکه

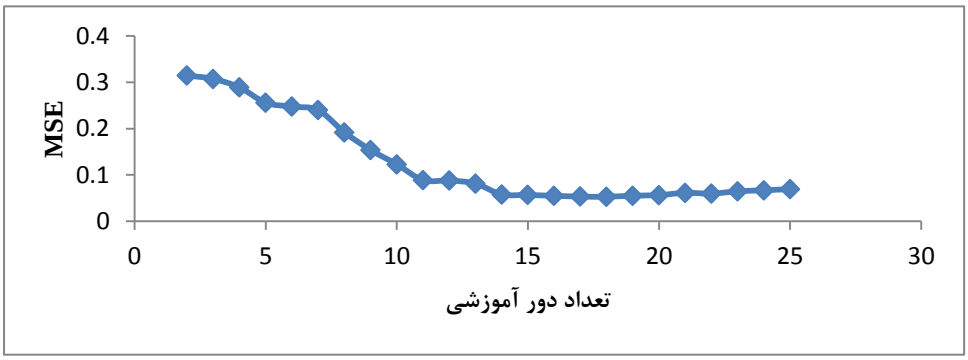
برای این منظور هر شبکه با تعداد ورودی ۲ تا ۱۹ و با دو الگوریتم آموزشی لونبرگ-مارکوات (trainlm) و تنظیم بایزین (trainbr) و دو تابع انتقال لگاریتم سیگموئید و تانژانت سیگموئید و تعداد گره از ۲ تا ۱۰ آموزش داده شد. در تمامی این شبکه‌ها تعداد دورهای آموزش ثابت و برابر ۲۰ در نظر گرفته شد. در روند بهینه‌سازی پارامترهای فوق، به حداقل رساندن میانگین مربع خطاهای حاصل از ارزیابی تقاطعی سری آموزش به عنوان معیار انتخاب شد. نتایج بهینه‌سازی این پارامترها در جدول (۴-۱۳)، گردآوری شده است.

جدول(۴-۱۳)- توابع و پارامترهای شبکه های بهینه SR-ANN

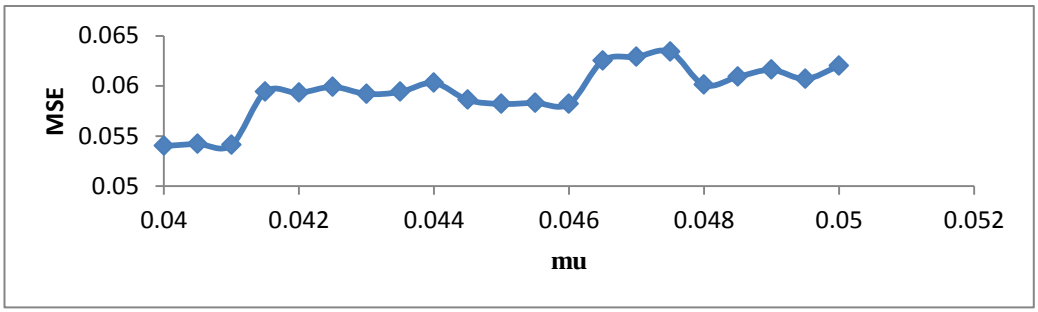
MSE	تعداد گره	تابع انتقال	تابع آموزش	تعداد توصیف گر
۰/۰۸۰۳	۱۰	لگاریتم-سیگموئید	لونبرگ-مارکوات	۱۶
۰/۰۵۸۶	۶	تانژانت-سیگموئید	لونبرگ-مارکوات	۱۸
۰/۱۲۶۷	۱۰	لگاریتم-سیگموئید	تنظیم بایزین	۷
۰/۰۵۶۶	**۶**	**تانژانت-سیگموئید**	**تنظیم بایزین**	**۱۸**

با توجه به نتایج به دست آمده، الگوریتم آموزشی تنظیم بایزین و تابع انتقال تانژانت سیگموئید، ۱۸ توصیف‌گر ورودی و ۶ نرون در لایه مخفی کم‌ترین MSE را نشان می‌دهد. در مرحله بعد تعداد دورهای آموزش بهینه شد برای این منظور شبکه بهینه به‌دست آمده از مرحله قبل در تعداد دورهای متفاوت (۲ تا ۲۵) آموزش داده شد و میانگین مربع خطاها برای سری آموزش ثبت گردید. شکل (۴-۶) نشان می‌دهد که با افزایش تعداد دورهای آموزش میانگین

مربع خطا (MSE) کاهش می‌یابد تا جایی که در ۱۸ دور تکرار به حداقل خطا می‌رسد و سپس با افزایش بیشتر تعداد دورهای یادگیری، خطاها افزایش می‌یابد بنابراین ۱۸ بار تکرار به عنوان تعداد تکرارهای بهینه برای یادگیری شبکه انتخاب شد. جهت بهینه کردن مقدار mu، ساختار شبکه با ۱۸ متغیر ورودی، ۶ گره در لایه پنهان و الگوریتم آموزشی تنظیم بایزین در نظر گرفته شد. سپس مقدار mu از ۰/۰۴ تا ۰/۰۵ با گام‌های ۰/۰۰۰۵ تغییر داده شد و آنگاه برای هر مورد مقدار میانگین مربع خطای سری آموزش به روش ارزیابی تقاطعی محاسبه گردید. در نهایت مقدار میانگین مربع خطا بر حسب mu رسم شد. طبق شکل (۴- ۷) مقدار بهینه ۰/۰۴ به‌دست آمد.



شکل (۴- ۷)– نمودار میانگین مربع خطای حاصل از ارزیابی تقاطعی سری آموزش در تعداد دورهای آموزش متفاوت



شکل (۴- ۸) – نمودار میانگین مربع خطای حاصل از ارزیابی تقاطعی سری آموزش بر حسب پارامتر mu

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

نتایج به دست آمده در بهینه‌سازی شبکه، در جدول (۴-۱۴) نشان داده شده است.

جدول (۴- ۱۴) – مشخصات شبکه عصبی مصنوعی بهینه(SR-ANN)

تابع آموزش	train br
تابع انتقال لایه پنهان	tansig
تابع انتقال لایه خروجی	pure line
تعداد نرون لایه پنهان	۶
تعداد متغیرهای ورودی	۱۸
تعداد دورهای آموزش	۱۸
پارامتر mu	۰/۰۴۰۰

۴-۵-۶-ارزیابی مدل‌ها با استفاده از سری آزمون

بعد از انتخاب شبکه عصبی بهینه ، قدرت پیش‌بینی شبکه، به وسیله ۱۰ داده سری آزمون که در آموزش استفاده نشده‌اند توسط هر دو مدل، مورد بررسی قرار گرفت. نتایج حاصله در جدول (۴-۱۵) و همچنین شکل (۴-۸) نشان داده شده است. ضرایب تعیین بیان‌گر این است که مدل‌های به دست آمده در پیش‌بینی فعالیت دارویی ترکیبات مورد بررسی، بسیار توانمند هستند و خطاهای به دست آمده ناچیز می‌باشد اما روش RF-SR-ANN نتایج بهتری را نشان می‌دهد.

۴-۵-۷- ارزیابی شبکه به روش رد مرحله‌ای تک‌تک

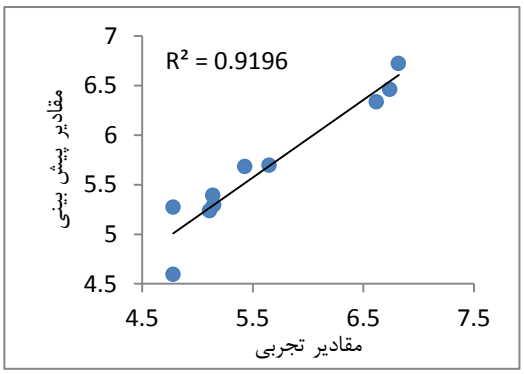
نتایج حاصل از ارزیابی شبکه با این روش در جدول (۴-۱۶) نشان داده شده است.

شکل (۴-۹) هم مقادیر پیش‌بینی شده بر حسب مقادیر واقعی برای کل داده‌ها را نشان می‌دهد. نتایج بیان‌گر این است که مدل‌های ارائه شده خطای پیش‌بینی کم و قدرت تعمیم بالایی دارد. روش RF-SR-ANN نتایج بهتری را نشان می‌دهد لذا با این مدل می‌توان داده‌های

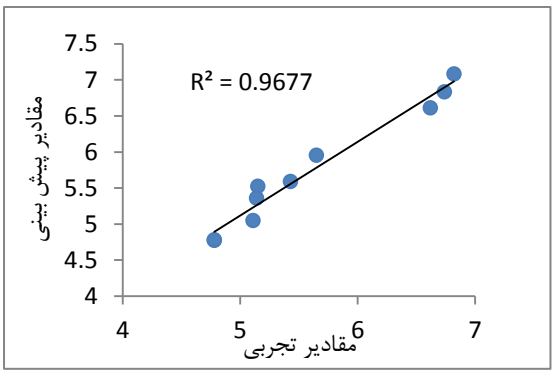
EC_{50} را برای هر ترکیب جدید با خطای کم و قابل قبول پیش‌بینی کرد. باقی‌مانده مقادیر محاسبه شده از مقادیر تجربی بر حسب مقادیر تجربی برای مولکول‌ها در شکل(۴-۱۰) رسم شده است. توزیع نسبتاً یکنواخت و اتفاقی باقی‌مانده‌ها در دو طرف خط صفر نشان دهنده‌ی این است که در مدل حاصل از شبکه‌ی عصبی مصنوعی، خطای معین وجود ندارد.

جدول (۴- ۱۵) – نتایج حاصل از ارزیابی مدل ANN با استفاده از سری آزمون

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		مدل (RF-SR-ANN)	مدل (SR-ANN)	مدل (RF-SR-ANN)	مدل (SR-ANN)
۲	۶/۶۲	۶/۶۱	۶/۳۳	-۰/۱۵۳	-۴/۲۹
۵	۵/۶۵	۵/۹۵	۵/۶۹	۵/۳۳	۰/۷۹
۱۳	۴/۷۸	۴/۷۷	۵/۲۷	-۰/۰۱۷	۱۰/۲۷
۱۹	۴/۷۸	۴/۷۷	۴/۵۹	-۰/۰۳۲	-۳/۹۱
۲۶	۶/۷۴	۶/۸۳	۶/۴۶	۱/۳۶	-۴/۱۶
۳۰	۵/۱۵	۵/۵۲	۵/۲۹	۷/۱۵	۲/۸۰
۳۲	۵/۱۴	۵/۳۴	۵/۳۹	۴/۲۳	۴/۸۸
۴۲	۶/۸۲	۷/۰۸	۶/۷۲	۳/۴۸	-۱/۴۳
۴۸	۵/۱۱	۵/۰۴	۵/۲۳	-۱/۷۲	۲/۴۶
۵۰	۵/۴۳	۵/۵۷	۵/۶۸	۲/۷۳	۴/۶۵



مدل(SR-ANN)



مدل(RF-SR-ANN)

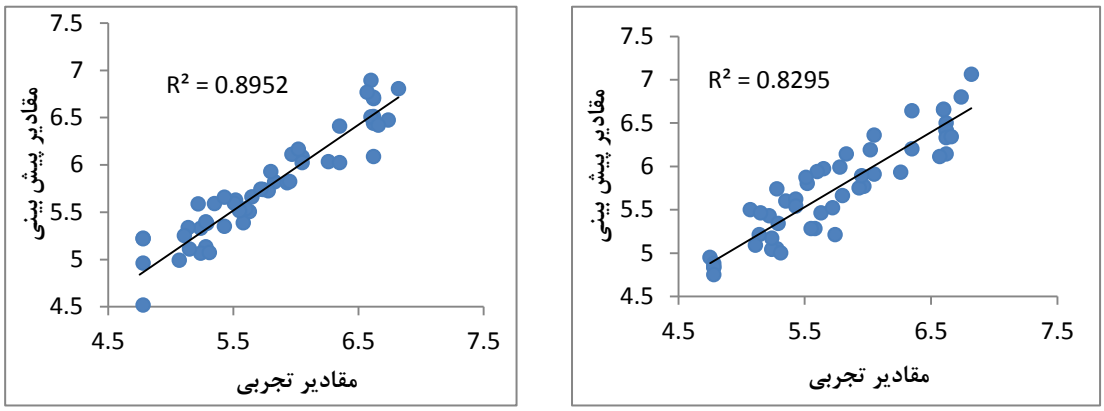
شکل (۴-۹)- ترسیمی از مقادیر پیش‌بینی شده بر حسب مقدار واقعی در سری آزمون

جدول (۴-۱۶)- نتایج حاصل از ارزیابی مدل‌ها با استفاده از روش رد تک‌تک

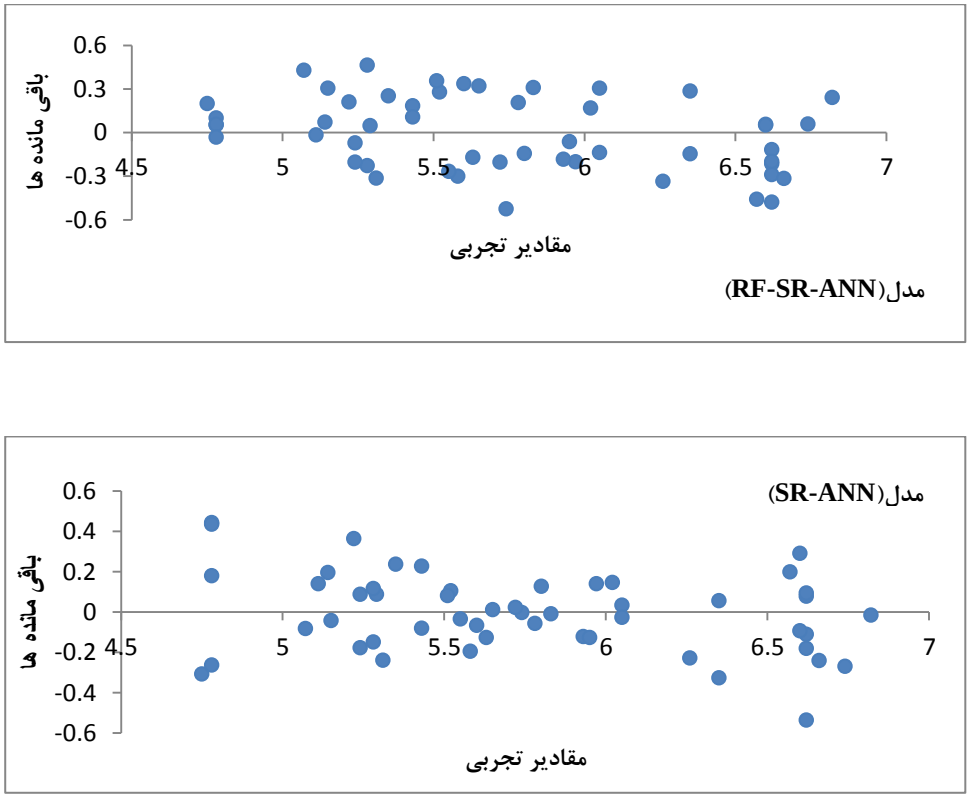
شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		مدل (RF-SR-ANN)	مدل (SR-ANN)	مدل (RF-SR-ANN)	مدل (SR-ANN)
۱	۶/۶۲	۶/۴۱	۶/۵۱	-۳/۲۴	-۱/۶۸
۲	۶/۶۲	۶/۵۰	۶/۴۴	-۱/۷۷	-۲/۷۳
۳	۵/۲۲	۵/۴۳	۵/۵۸	۴/۰۲	۶/۹۷
۴	۵/۲۸	۵/۰۵	۵/۱۳	-۴/۳۲	-۲/۸۲
۵	۵/۶۵	۵/۹۷	۵/۶۶	۵/۶۹	۰/۲۰
۶	۴/۷۵	۴/۹۵	۴/۴۴	۴/۲۰	-۶/۴۸
۷	۵/۸۳	۶/۱۴	۵/۸۲	۵/۳۰	-۰/۱۶
۸	۶/۲۶	۵/۹۳	۶/۰۳	-۵/۳۴	-۳/۶۳
۹	۵/۷۴	۵/۲۱	۵/۷۴	-۹/۱۶	۰/۰۰
۱۰	۶/۶۲	۶/۱۴	۶/۰۸	-۷/۲۱	-۸/۰۸
۱۱	۵/۲۹	۵/۳۴	۵/۳۸	۰/۹۰	۱/۶۷
۱۲	۵/۰۷	۵/۵۰	۴/۹۸	۸/۴۴	-۱/۶۰
۱۳	۴/۷۸	۴/۸۳	۵/۲۲	۱/۰۷	۹/۲۶
۱۴	۵/۲۴	۵/۰۴	۵/۳۲	-۳/۸۸	۱/۶۶
۱۵	۵/۲۴	۵/۱۷	۵/۰۶	-۱/۳۴	-۳/۳۸
۱۶	۶/۰۵	۶/۳۶	۶/۰۸	۵/۰۵	۰/۵۵
۱۷	۵/۶۳	۵/۴۶	۵/۵۰	-۳/۰۴	-۲/۲۵
۱۸	۵/۸۰	۵/۶۶	۵/۹۳	-۲/۴۷	۲/۲۱
۱۹	۴/۷۸	۴/۸۸	۴/۵۱	۲/۱۰	-۵/۴۸
۲۰	۴/۷۸	۴/۸۴	۴/۹۶	۱/۱۷	۳/۷۵
۲۱	۶/۳۵	۶/۲۰	۶/۰۲	-۲/۲۹	-۵/۱۲
۲۲	۵/۲۸	۵/۷۴	۵/۳۹	۸/۸۰	۲/۲۱
۲۳	۵/۹۷	۵/۷۷	۶/۱۱	-۳/۳۷	۲/۳۴
۲۴	۶/۳۵	۶/۶۴	۶/۴۰	۴/۵۰	۰/۸۹
۲۵	۶/۰۲	۶/۱۹	۶/۱۶	۲/۷۹	۲/۴۲

ادامه جدول(۴-۱۶)

شماره ترکیب	مقدار تجربی	مقدار پیش‌بینی		درصد خطا	
		مدل (RF-SR-ANN)	مدل (SR-ANN)	مدل (RF-SR-ANN)	مدل (SR-ANN)
۲۶	۶/۷۴	۶/۸۰	۶/۴۷	۰/۸۶	-۳/۹۹
۲۷	۶/۶۲	۶/۴۲	۶/۷۰	-۳/۰۱	۱/۱۸
۲۸	۶/۶۲	۶/۳۳	۶/۷۱	-۴/۳۶	۱/۴۱
۲۹	۵/۶۰	۵/۹۴	۵/۵۳	۶/۰۰	-۱/۱۹
۳۰	۵/۱۵	۵/۴۶	۵/۱۱	۵/۹۴	-۰/۸۱
۳۱	۵/۳۱	۵/۰۰	۵/۰۷	-۵/۸۸	-۴/۵۰
۳۲	۵/۱۴	۵/۲۱	۵/۳۳	۱/۳۸	۳/۸۰
۳۳	۵/۳۵	۵/۶۰	۵/۵۸	۴/۷۳	۴/۴۴
۳۴	۶/۰۵	۵/۹۱	۶/۰۲	-۲/۲۵	-۰/۴۴
۳۵	۵/۵۸	۵/۲۸	۵/۳۸	-۵/۳۶	-۳/۴۸
۳۶	۵/۵۲	۵/۸۰	۵/۶۲	۵/۰۵	۱/۹۱
۳۷	۵/۷۸	۵/۹۹	۵/۷۲	۳/۵۶	-۰/۹۸
۳۸	۵/۷۲	۵/۵۲	۵/۷۴	-۳/۵۴	۰/۳۹
۳۹	۵/۹۳	۵/۷۵	۵/۸۰	-۳/۱۰	-۲/۰۵
۴۰	۶/۶۰	۶/۶۶	۶/۵۰	۰/۸۴	-۱/۴۰
۴۱	۶/۶۶	۶/۳۴	۶/۴۲	-۴/۷۴	-۳/۶۱
۴۲	۶/۸۲	۷/۰۶	۶/۸۰	۳/۵۳	-۰/۲۳
۴۳	۶/۶	۶/۶۵	۶/۸۹	۰/۷۷	۴/۴۰
۴۴	۵/۵۵	۵/۲۸	۵/۵۱	-۴/۷۹	-۰/۶۲
۴۵	۴/۷۸	۴/۷۵	۵/۲۱	-۰/۶۶	۹/۱۲
۴۶	۶/۵۷	۶/۱۱	۶/۷۷	-۶/۹۷	۳/۰۱
۴۷	۵/۵۱	۵/۸۷	۵/۵۹	۶/۴۸	۱/۴۹
۴۸	۵/۱۱	۵/۰۹	۵/۲۵	-۰/۳۰	۲/۷۳
۴۹	۵/۴۳	۵/۶۲	۵/۳۵	۳/۴۱	-۱/۴۸
۵۰	۵/۴۳	۵/۵۴	۵/۶۵	۱/۹۷	۴/۱۹
۵۱	۵/۹۵	۵/۸۹	۵/۸۲	-۱/۰۵	-۲/۱۲



شکل (۴-۱۰)- مقادیر پیش‌بینی شده با شبکه عصبی مصنوعی بر حسب مقدار واقعی برای کل داده‌ها



شکل (۴-۱۱)- باقی‌مانده‌ی مقادیر پیش‌بینی شده‌ی PEC_{50} توسط شبکه عصبی مصنوعی از مقادیر واقعی بر حسب مقادیر تجربی برای کل ترکیبات

۴-۵-۸- ارزیابی مدل ANN با استفاده از پارامترهای آماری

برای ارزیابی کارآیی شبکه عصبی مصنوعی ارائه شده برای پیش‌بینی مهارکنندگی ترکیبات نکروستاتین-۵ پارامترهای آماری محاسبه شد و نتایج حاصل در جدول (۴-۱۷) آمده است. این نتایج حاکی از قدرت پیش‌بینی بالای مدل غیرخطی شبکه‌ی عصبی می‌باشد. با توجه به نتایج به‌دست آمده می‌توان گفت مدل SR-ANN عملکرد بهتری را نشان می‌دهد اما تعداد توصیف‌گر بیشتری در آن به‌کار گرفته شده است.

جدول (۴-۱۷)- پارامترهای آماری محاسبه شده

پارامتر	سری آزمون(N=10)		کل داده ها(N=51)	
	مدل (RF-SR-ANN) ۹ توصیفگر	مدل (SR-ANN) ۱۸ توصیفگر	مدل (RF-SR-ANN) ۹ توصیفگر	مدل (SR-ANN) ۱۸ توصیفگر
MAE	۰/۱۴۳۹	۰/۲۱۶۰	۰/۲۱۷۶	۰/۱۵۷۰
MSE	۰/۰۳۵۴	۰/۰۶۱۱	۰/۰۶۳۴	۰/۰۳۹۱
PRESS	۰/۳۵۴	۰/۶۱۰	۳/۲۳۵	۱/۹۷۶
SEP	۰/۱۸۱۸	۰/۲۴۷۱	۰/۲۵۱۹	۰/۱۹۶۹
R^2	۰/۹۶۷۷	۰/۹۱۹۶	۰/۸۲۹۵	۰/۸۹۵۰
REP (%)	۳/۳۴۶	۴/۳۹۵	۴/۳۷۸	۳/۴۲۲
MRE	۲/۵۷۷	۳/۹۶۷	۳/۷۸۰	۲/۷۹۷

۴-۵-۹- ارزیابی مدل ANN با استفاده از آزمون Y- تصادفی[1]

برای حصول اطمینان از تصادفی نبودن نتایج، یک مجموعه اعداد تصادفی در دامنه‌ی متغیر هدف تعریف کرده، مدل QSAR جدید با استفاده از ماتریس متغیرهای مستقل اصلی و مقادیر تصادفی از pEC_{50} توسعه یافت. نتایج حاصل از چندین مجموعه‌ی تصادفی ایجاد شده، در جدول (۴-۱۸) نشان داده شده است. ضریب هم‌بستگی پایین، نشان‌دهنده‌ی عدم هم‌بستگی تصادفی در مدل‌های شبکه عصبی مصنوعی ایجاد شده می‌باشد.

1 - Y- randomization test

جدول (۴-۱۸)– مقادیر R^2 برای سری آزمون با استفاده از آزمون Y- تصادفی

تکرار	۱	۲	۳	۴	۵	۶	۷	۸
R^2 (RF_SR_ANN)	۰/۲۵	۰/۰۳	۰/۰۶	۰/۱۷	۰/۰۸	۰/۰۰	۰/۰۲	۰/۱۸
R^2 (SR_ANN)	۰/۰۸۹	۰/۱۳۶	۰/۳۶۵	۰/۰۰۵	۰/۲۳۷	۰/۰۲۲	۰/۰۳۴	۰/۳۳۹

۴-۶-بررسی ارتباط توصیف‌گرهای منتخب با فعالیت بازدارندگی نکروپتوز

مهم‌تر از قدرت پیش‌بینی میزان فعالیت ترکیبات جدید، رسالت دیگر مدل‌های QSAR ایجاد ارتباط بین فعالیت مورد نظر و پارامترهای ساختاری یا توصیف‌کننده‌هاست که بتوان به کمک این ارتباط به بررسی تئوری و در نهایت عملی ترکیباتی جدید با خواص بهتر پرداخت. جدول (۴-۱۹) مقادیر اثر متوسط[1] توصیف‌کننده‌های مدل برتر یعنی RF-SR-ANN را نشان می‌دهد. اثر متوسط یک متغیر مستقل با استفاده از فرمول (۴-۱) به دست می آید:

$$T=\frac{\beta_x \times \sum X_n}{\sum pEC_{50}} \qquad (۱-۴)$$

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

جدول (۴-۱۹)-اثر متوسط توصیف گر های به دست آمده توسط روش رگرسیون مرحله ای

Variable	HVcpx	RDF140e	BEHm6	GO..S	Uindex	MLOGP	Mor11u	R8u	H7u
Mean effect	۰/۴۳۸	-۰/۳۲۱	-۰/۵۶۹	۰/۲۶۴	-۰/۷۳۸	-۰/۹۷۰	۰/۲۹۸	-۰/۵۳۴	-۰/۳۱۶

1- Mean effect

الف- توصیف‌گرهای گروه GETAWAY[1] [۴۴]

این توصیف‌گرها با توجه به مختصات فضایی اتم‌ها در یک مولکول به راحتی قابل محاسبه هستند. این توصیف‌گرها بر اساس ماتریس قدرت نفوذ [2] تعریف می‌شوند و بیان‌کننده‌ی ویژگی‌های هندسی و توپولوژیک مولکول‌ها هستند. توصیف‌گرهای GETAWAY به دو زیر مجموعه‌ی H و R تقسیم می‌شوند.

ماتریس قدرت نفوذ یا ماتریس تأثیر مولکول[3] (MIM) با رابطه‌ی زیر بیان می‌شود.

$$H= M .(M^T.M)^{-1}.M^T \qquad (۲-۴)$$

که M ماتریس مختصات اتمی، T به معنای ماتریس ترانهاده[4] و H ماتریس قدرت نفوذ است. H یک ماتریس متقارن A×A است که A تعداد اتم‌هاست. عناصر قطری ماتریس H لوریج‌ها نام دارند که هر یک بیان‌گر اثر یک اتم در ایجاد شکل کلی یک مولکول است.

ماتریس تاثیر- فاصله به صورت زیر تعریف می‌شوند.

$$[R]_{ij} = \left[\frac{\sqrt{h_{ii}.h_{jj}}}{r_{ij}} \right]_{ij} \quad i \neq j \qquad \text{رابطه‌ی (۴-۳)}$$

که h_{ii} و h_{jj} لوریج‌های دو اتم i و j و r_{ij} فاصله‌ی آن دو است. عناصر قطری این ماتریس صفر است.

دو توصیف گر ی H7u و R8u از این کلاس در مدل‌سازی به کار گرفته شده است. u در این توصیف‌کننده ها نشان‌دهنده‌ی آن است که آن‌ها با ویژگی خاصی از اتم‌ها محاسبه نشده‌اند.

1 - Geometry,Topology, and Atom-Weights AssemblY
2 - Leverage matrix
3 - Molecular Influence Matrix
4- Transpose

ضریب اثر این توصیف‌کننده‌ها منفی می‌باشد یعنی با افزایش مقدار آن‌ها میزان pEC_{50} کاهش می‌یابد. چند مثال از اثر این توصیف‌کننده در جدول (۴-۲۰) ارائه شده است.

ب-توصیف‌گرهای 3D – MORSE[1] [۴۵]

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

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (۴-۴)$$

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 وj ، N تعداد کل اتم‌ها را نشان می‌دهد. این روش باعث می‌شود که ساختار سه‌بعدی مولکول به یک کد ثابت تبدیل شود. این توصیف‌گرها آرایش سه بعدی اتم‌ها را بدون اینکه به اندازه مولکول ارتباط داشته باشند، بیان می‌کنند بنابراین برای تعداد زیادی مولکول با تفاوت‌های ساختاری زیاد قابل کاربرد است[۴۷]. هم‌چنین قادرند ارتباط بین ساختار سه‌بعدی ترکیبات آلی و خصوصیات فیزیکی، شیمیایی و بیولوژیکی فراهم سازند. از این گروه توصیف‌کننده‌ها، توصیف‌گر سه‌بعدی mor11u در مدل دیده می‌شود که دارای ضریب مثبت است.

ج) توصیف‌گرهای Molecular

از این طبقه، توصیف‌گر MLOGP انتخاب شده است. این توصیف‌گر ضریب تقسیم دارو در دو فاز غیرقابل امتزاج آب/ اکتانول را نشان می‌دهد و اطلاعاتی در مورد آب‌گریزی

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

مولکول در اختیار قرار می دهد. هر چه آب‌گریزی یک مولکول بیشتر باشد به طور وسیع‌تری در بدن پخش شده و از گزینش‌پذیری کمتری برخوردار است. ضریب این توصیف‌کننده منفی است و هر چه مقدار آن بیشتر باشد pEC_{50} کمتر خواهد بود و خاصیت دارو کاهش پیدا می‌کند. انتخاب این توصیف‌گر نشان می‌دهد که فعالیت دارو به آب‌گریزی مولکول بستگی دارد.

د) توصیف‌گرهای Topoligical [۴۵]

این توصیف‌گرها بر اساس نمایش گراف مولکول می‌باشند. در این گراف‌ها 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 توصیف‌گرها 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 توصیف‌گرها به سادگی از روی ساختمان دوبعدی مولکول‌ها امکان‌پذیر می‌باشد. از این گروه، توصیف‌گرهای Uindex[۱] و HVcpx[۲] انتخاب شده است.

Uindex شاخص بالابان نامیده می‌شود که بر اساس ماتریس فاصله در گراف مولکولی و از فرمول زیر محاسبه می‌شود.

$$U=\frac{M}{\mu+1}\sum\ (D_i.D_j)^{-1/2} \qquad (۴-۴)$$

M: تعداد پیوندها، $\boldsymbol{\mu}$: تعداد cyclomatic در گراف مولکول و برابر با حداقل تعداد پیوندی است که قبل از چند حلقه‌ای شدن گراف حذف می‌شود ، D_i نیز از ماتریس فاصله (D) محاسبه می‌شود که برابر مجموع عناصر ردیف i در ماتریس فاصله است.

1 - balaban u index

2 - graph vertex complexity

شاخص بالابان با افزایش اندازه مولکول، تعداد شاخه‌ها و درجه اشباع‌شدگی افزایش یافته و با زیاد شدن تعداد حلقه‌ها کاهش می‌یابد. این توصیف‌گر دارای ضریب منفی در مدل بوده و افزایش آن باعث کم شدن pEC_{50} می‌شود.

HVcpx نیز طبق رابطه (۴-۵) از ماتریس فاصله قابل محاسبه است:

$$H_v= 1/A*\sum_{i=1}^{A} v_i \qquad (۵-۴)$$

که در این رابطه A برابر تعداد اتم‌هاست و v_i هم پیچیدگی رئوس می باشد. این توصیف‌گر دارای ضریب مثبت در مدل می‌باشد.

ت-توصیف‌گرهای BCUT[1]

این توصیف‌گرها از دسته توصیف‌گرهای دو بعدی هستند و به عنوان مقدار ویژه یک ماتریس ارتباطی اصلاح شده[2] (ماتریس بوردن) می‌باشند. ماتریس بوردن یک گراف تهی از هیدروژن را ارائه می‌دهد که در آن B_{ii}، یعنی عناصر قطری در ارتباط با خاصیت اتمی مانند الکترونگاتیویته، حجم واندروالس، جرم اتمی و عناصر غیر قطری در ارتباط با مرتبه پیوند دو اتم پیوند داده شده، هستند [۴۵و۴۶]. توصیف گر BEHm6[3] بالاترین مقدار ویژه از ماتریس بوردن که با جرم اتمی وزن‌دار شده است و دارای ضریب مثبت است.

ه-توصیف‌گرهای RDF[4]

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:

1 - Burden-CAS- university of texas eigenvalues
2 - Modified connectivity matrix
3 - Highest eigenvalue n. 6 of Burden matrix / weighted by atomic masses
4 - Radial Distribution Function

$$g(r) = f \sum_{i=1}^{N-1} \sum_{j>i}^{N} A_i A_j e^{\beta(r-r_{ij})^2} \qquad (۷-۳)$$

که در این رابطه N تعداد اتم‌ها، A_i و A_j خاصیت اتمی (الکترونگاتیویته، جرم اتمی، قطبش‌پذیری اتمی و...)، r_{ij} شعاع بین اتم i و j، f فاکتور مقیاس، β فاکتور دما که باعث حرکت و جابجایی اتم‌ها است، می‌باشند[۳۷]. اهمیت این توصیف‌گرها به دلیل اختلاف در توزیع اتم‌ها در مولکول‌ها و اثر این توزیع بر روی فعالیت مهارکنندگی ترکیبات می‌باشد.

۴-۷- بررسی میزان مشارکت توصیف‌گرهای منتخب در شبکه عصبی

میزان مشارکت توصیف‌گرهای منتخب به صورت زیر تعیین شد:

۱-توصیف‌گر مورد نظر به همراه اوزان مربوطه‌اش از شبکه بهینه شده حذف گردید.

۲- با استفاده از بقیه توصیف‌گرها مقدار متغیر وابسته (pEC_{50}) برای هر ترکیب سری آموزش به روش ارزیابی تقاطعی پیش‌بینی شد.

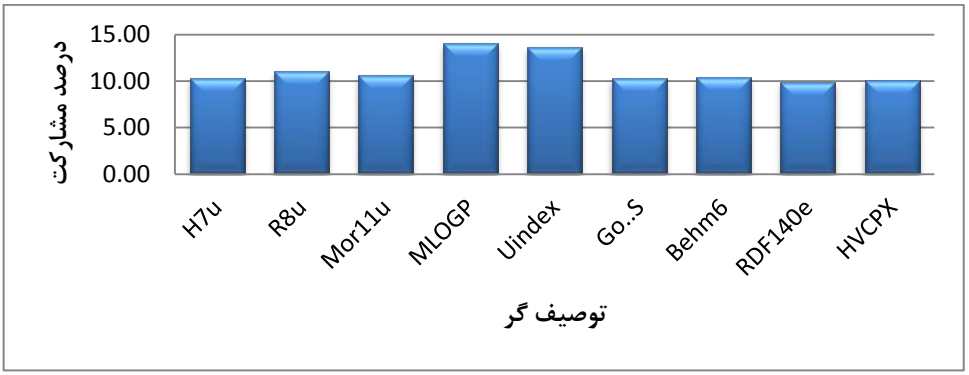
۳- میانگین خطای مطلق حاصل از ارزیابی تقاطعی ترکیبات سری آموزش محاسبه گردید.

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

۵- سرانجام درصد مشارکت هر توصیف‌گر توسط رابطه (۳-۱۲) برآورد شد.

$$c_i = 100 \frac{\Delta m_i}{\sum_{i=1}^{N} \Delta m_i} \qquad (۱۲-۳)$$

تعداد توصیف‌گرهای مدل و Nام، iمشارکت توصیف‌گر حذف شده درصد c_i در این رابطه ام را i میانگین خطای مطلق حاصل از ارزیابی تقاطعی سری آموزش در غیاب توصیفگر Δm_i نشان می‌دهد. که بر این اساس درصد مشارکت توصیف‌گرهای منتخب در ترکیبات مورد بررسی به شکل زیر می‌باشد.



شکل(۴-۱۲)- مشارکت توصیف‌گرها در شبکه عصبی بهینه RF-SR-ANN

دو توصیف‌گر MLOGP و Uindex دارای بیشترین اثر مشارکت هستند. Uindex در برگیرنده اطلاعات دو بعدی مولکول است. بنابراین می توان گفت که فعالیت بازداری این ترکیبات وابستگی قابل ملاحظه‌ای به توپولوژی و آب‌گریزی ترکیب بازدارنده دارد.

۴-۸- نتیجه‌گیری نهایی

نتایج نشان می‌دهند که مدل‌های غیرخطی توسعه یافته می‌توانند به عنوان روش‌های موفق برای مدل‌سازی و پیشگویی فعالیت بازدارندگی ترکیبات مورد مطالعه باشند و این مدل‌ها می‌توانند به محققان در طراحی داروهایی با فعالیت بازدارندگی قویتر کمک کنند. همچنین می‌توان از روش RF-SR، برای کاهش تعداد توصیف‌گر در مطالعات QSAR استفاده کرد.

آینده‌نگری

- روش جنگل‌های تصادفی را می‌توان برای پیش‌بینی فعالیت سایر 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.

- محققان می‌توانند از نتایج حاصل از این پروژه در طراحی و آزمایش ترکیبات جدید با ویژگی بازدارندگی قوی‌تر برای نکروپتوز استفاده کنند. این تحقیقات را می‌توان با پژوهش روی خواص دارویی ترکیبات جدید ادامه داد تا نهایتاً منجر به ساخت داروهای جدیدتر با اثربخشی بهتر گردد.

منابع

[1]- http://www.iranhealers.com

[2]- http://www.ray.tums.ac.ir/information/hiv.htm.

[3]www.nature.com/nature/journal/v410/n6831/fig_tab/410995a0_F3.htm

[4]- www.nature.com/nature/journal/v410/n6831/fig_tab/410995a0_F4.html.

[5]- Cesarini S., Spallarossa A., Ranise A., et al., (2008), *"Thiocarbamates as non-nucleoside HIV-1 reverse transcriptase inhibitors. Part 1: Parallel synthesis, molecular modelling and structure–activity relationship studies on O-[2-(hetero)arylethyl]-N-phenylthiocarbamates"*, Bioorganic & Medicinal Chemistry, 16(7): p 4160-4172

[6]-Zahouily M., Rakik J., Lazar M., Banlaoui M.A., Rayadh A., Komiha N., (2007), *"Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by artificial neural networks: HEPT derivatives"*, ARKIVOC, pp 245-256

[7]- Zarei K., Atabati M., (2009), *"QSAR study of anti-HIV activities against HIV-1 and some of their mutant strain for a group of HEPT derivatives"*, Chinese Chemical Society., 56, pp 206-213.

[8]- Dessalew N., (2009), *"Investigation of the structural requirement for inhibiting HIV integrase: QSAR study"*, Acta Pharmaceutica., 59, pp 31-43.

[9]- 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"*, Chemical Biology & Drug Design., 73, pp 456-465.

[10]- Almerico A.M, Tutone M., Lauria A., (2009),"*A QSAR study investigating the potential anti-HIV-1effect of some Acyclovir and Ganciclovir analogs"*, ARKIVOC, pp 85-94.

[11]-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

[12]- Cesarini S., Spallarossa A., Ranise A., Schenone S., et al., (2008), "*Parallel one-pot synthesis and structure–activity relationship study of symmetric formimidoester disulfides as a novel class of potent non-nucleoside HIV-1 reverse transcriptase inhibitors"*, Bioorganic & Medicinal Chemistry,. 16(12): p. 6353-6363

[13]- Goodarzi M., M.P. Freitas, and Y.V. Heyden, (2011) *" Linear and nonlinear quantitative structure–activity relationship modeling of the HIV-1 reverse transcriptase inhibiting activities of thiocarbamate"s*. Analytica Chimica Acta,. 705(1–2): p. 166-173.

[14]- هاشمی محمد، قوامی سعید، کرمی تهرانی فاطمه، (۱۳۸۲) "آپوپتوسیس، مرگ برنامه‌ریزی شده سلول"، ویژه‌نامه طبیب شرق، سال پنجم، شماره۱، ص ۷۱-۷۷

[15] - Weihong Zheng, Alexei Degterev, Emily Hsu ,Junying Yuan, Chengye Yuan, (2008), "Structure–activity relationship study of a novel necroptosis inhibitor, necrostatin-7", Bioorganic & Medicinal Chemistry Letters, 18, pp 4932-4935

[16]- Chamjangali, M.A. and M. Ashrafi, (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): p. 392-400.

[17]- Edited by D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, *" Handbook of Chemometrics and Qualimetrics: Part A"* Volume 20, Part A, Pages 1-867 (1998)

[18]- Wold s.,(1995), *"chemometrics;what do we mean with it and what do we want from it",*Chemometrics and Intelligent Laboratory System,30,pp 503-523

[19]-Jurs P. C. (2005) *"Assessing the reliability of a QSAR model's predictions",* J.Mol Graph Model, 23(6), pp 50

[20]- Constantinou, L. and R. Gani,(1994) New group contribution method for estimating properties of pure compounds. AIChE Journal,. 40(10): p. 1697-1710.

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

[22]- Mcculloch W.S., Pitts W.H.,(1943).*"A logical calculus of the ideas immanent in*

*neural nets",*Bulletin of Mathematical Biophysics,5,pp 115-133.

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

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

[25]- Anil.k.Jane,Jianchang Mao,K.M.Mohiuddin,(1996),” *Artificial neural networks: a tutorial”*

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

[27]- Breiman L., (2001), “Random forests” , Machine Learning, vol. 45, pp 5-32

[28]- Aldrich, C. and L. Auret,(2013), “*Tree-Based Methods, in Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods”*., Springer London. p. 183-220.

[29]- مهدوی تالارپشتی، ع،(۱۳۹۱) پایان نامه کارشناسی ارشد، *"مدیریت ریزش مشتری با استفاده از ابزارهای پیشرفته آماری و داده کاوی"، دانشگاه صنعتی شاهرود*

[30]- Breiman L., (1998), *“Arcing classifiers*”, The Annals of Statistics, vol. 24, pp: 801-849.

[31]-Cutler.A.,Zhao.G.,(2001),“*PERT–Perfect,”Random TreeEnsembles”*,Computing science and statistics,vol. 33, pp.490-497

[32]- Shi T. and S. Horvath,*(* 2006),*” Unsupervised Learning With Random Forest Predictors*”. Journal of Computational and Graphical Statistics, 15(1): p. 118-138

[33]- Aldrich, C. and L. Auret,(2013), ” *Tree-Based Methods, in Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*”. Springer London. p. 183-220

[34]- Hajmeer M., Basheer I., (2003), “*Comparison of logistic regression and neural network-based classifiers for bacterial growth*”, *Food Microbiol*, 20, pp 43-45

[35]- Ranise A., Spallarossa A., Cesarini S, Bondavalli F.,(2005),”*Structure-Based Design, Parallel Synthesis, Structure−Activity Relationship, and Molecular Modeling Studies of Thiocarbamates, New Potent Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitor Isosteres of Phenethylthiazolylthiourea Derivatives”*. Journal of Medicinal Chemistry, 48(11): p. 3858-3873.

[36]-Spallarossa A., Cesarini S., Ranise A., et al.,(2009),” *Novel modifications in the series of O-(2-phthalimidoethyl)-N-substituted thiocarbamates and their ring-opened congeners as non-nucleoside HIV-1 reverse transcriptase inhibitors”,* European Journal of Medicinal Chemistry, 44(4): p. 1650-1663

[37]- 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

[38]-Cesarini S., Spallarossa A., Ranise A., et al.(2008), " *Thiocarbamates as non-nucleoside HIV-1 reverse transcriptase inhibitors. Part 2: Parallel synthesis, molecular modelling and structure–activity relationship studies on analogues of O-(2-phenylethyl)-N-phenylthiocarbamate*. Bioorganic& MedicinalChemistry. 16(7):p.4173-4

[39]- Pillar C., and D. Sahm,(2012), "*Resistance Trends and Susceptibility Profiles in the US Among Prevalent Clinical Pathogens: Lessons from Surveillance, in Antibiotic Discovery and Development*", T.J. Dougherty and M.J. Pucci, Editors, Springer US. p. 753-792

[40]- Ke Wang, Jinfeng Li, Alexei Degterev, Emily Hsu, Junying Yuan and Chengye Yuan, (2007) "*Structure–activity relationship analysis of a novel necroptosis inhibitor, Necrostatin-5*", 17, pp 1455-1465

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

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

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

[44] - Consonni V, Todeschini R., Pavan M, (2002), "*Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors*", J. Chem. Inf. Comput. Sci, 42, pp 682-692.

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

[46]- Marjanovic M., Kralj M., Supek F., Frkanec L., Piantunida I., Smuc T., Tusek-Bozic L., (2007), *"Antitumor potential of crown ether: structure-activity relationship, cell cycle disturbances, and cell death studies of a series of Ionophores",* J. Med. Chem., 50, pp 1007-1018

[47]- Schuur J.H., Selzer P., Gasteiger J., (1996), *"The coding of the three dimensional structure of molecules by molecular transform and its application to structure- spectra correlation and studies of biological activity"*, J. Chem. Inf. Comput. Sci , 36, pp 334-344.

Abstract

Non-nucleoside reverse transcriptase inhibitors (NNRTIs) are a class of anti retroviral drugs used to treat HIV infection. NNRRTIs inhibit activity of reverse transcriptase (RT), an enzyme that controls the replications of the genetic material of HIV. In the first section, Quantitative structure – activity relationship (QSAR) models were constructed for predicting the Anti-HIV activity of a set of Thiocarbamate Derivatives with group contribution descriptors and Random forest method. Data set was randomly divided into training set and test set including 157 and 40 compounds, respectively. The validation study of the RF models was performed using test set and leave-one-out technique. The results obtained for prediction of anti HIV activity of the test set and leave-one-out method by RF models showed squared correlation coefficients of 0.8748 and 0.8025, respectively.

In the second section, some derivatives of necroptosis inhibitors (Necrostatin-5) were studied. In this study, A data set including 160 compounds were used to design predictive models. A new method Random Forest was used for classification necrostatin-5 compounds in to active and inactive groups. For a series of the 51 active componds, The most effective descriptors were selected using Random Forest and Rtepwise Regression(RF-SR) method. The selected descriptors were correlated with the bioactivities of the molecules using the well known Artificial Neural Networks(ANN). The prediction ability of the proposed models was evaluated by test set and leave-one-out method. The squared correlation coefficients obtained for test set and leave-one-out method by ANN models were 0.9677 and 0.8295, respectively.

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

University of Shahrood

Faculty of Chemistry

Prediction of anti HIV activities of thiocarbamates as Non-nucleoside derivative using Random Forest

Zahra Ajam

Superviser:

Dr. Mansour Arab Chamjangali

Advisor:

Dr. Majid Salami

February 2015

۱۳۸