

مدل های جدید مبتنی بر ساختار با استفاده از توصیف کننده های مخلوط برای پیش بینی دمای اشتعال مخلوط های آلی دو جزئی

زهره فرامرزی^۱، فاطمه عباسی تبار^{۲*}، وحید زارع شاه آبادی^۱

^۱ دانشکده معدن، نفت و انرژی، دانشگاه آزاد اسلامی واحد ماهشهر، خوزستان، ایران.

^۲ گروه شیمی، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران.

چکیده: به دلیل اهمیت مخلوطهای دو تایی و کاربرد وسیع آنها در صنعت، تخمین نقطه اشتعال آنها حائز اهمیت است. از این رو، در این مطالعه، مدل های رابطه کمی ساختار-خاصیت (QSPR) برای پیش بینی نقطه اشتعال پذیری مخلوطهای آلی دو تایی ایجاد شد. بزرگترین چالش در مطالعات QSPR مخلوط ها، محاسبه توصیف کننده عددی برای توصیف ویژگی یک مخلوط است. در این تحقیق، ۲۲ مجموعه از توصیف کننده ها حاصل از روابط ریاضی مختلف برای محاسبه توصیف کننده های مخلوط از توصیف کننده های مولکولی ترکیبات خالص به کار برده شد. الگوریتم کولونی مورچگان حافظه دار (Memorized-ACO) همراه با رگرسیون خطی چند متغیره (MLR) برای انتخاب بهترین زیر مجموعه توصیف کننده هایی که سهم قابل توجهی در خاصیت اشتعال پذیری دارند به کار برده شد. مدل مبتنی بر توصیفگر sqr-fmol-sum به عنوان بهترین مدل انتخاب شد. مقادیر آموزش R^2 ، آزمون R^2 این مدل به ترتیب ۰/۹۳ و ۰/۹۴ بود. میانگین خطای مطلق (MAE) برای مجموعه داده های آموزشی و آزمون به ترتیب ۳/۷۵ و ۲/۹۷ به دست آمد. یک مدل گروهی نیز از طریق یک راهکار ساده میانگین گیری و با استفاده از مدل هایی که توسط بهترین توصیف کننده ها ایجاد شده بودند، به دست آمد. این امر موجب بهبود کیفیت آماری مدل نهایی QSPR شد.

واژگان کلیدی: رابطه کمی ساختار-خاصیت، نقطه اشتعال، الگوریتم کولونی مورچگان حافظه دار، توصیف کننده مخلوط، رگرسیون خطی چند متغیره.

fa.abbasitabar@iau.ac.ir

است [۱، ۲]. داده های نقطه اشتعال برای مواد شیمیایی خالص را می توان از منابع مختلف مانند اس دی اس های علمی ترمو فیشر^۱، کتاب راهنمای شیمی لانترن [۳]، ایندکس مرک [۴] یا راهنمای جیبی نیوش برای خطرات شیمیایی بدست آورد. متأسفانه، به دلیل وابستگی نقطه اشتعال یک مخلوط به اجزای تشکیل دهنده آن، اطلاعات کمی از نقطه اشتعال مخلوط ها در دسترس است. علاوه بر این، در اکثر موارد، مخلوط ها رفتار غیر ایده آل نشان می دهند که موجب افزایش و یا کاهش شدید نقطه

۱- مقدمه

نقطه اشتعال^۱ پایین ترین درجه ی دمای است که در آن ماده، بخار کافی برای تشکیل مخلوط قابل احتراق با هوا تولید می کند. این خصوصیت یک مسئله کلیدی ایمنی در ارزیابی خطر فرآیندهای صنعتی است که در چارچوب های مختلف نظارتی برای استفاده، ذخیره سازی و حمل و نقل به مواد شیمیایی اختصاص یافته

^۱Thermo Fisher Scientific's SDSs

^۱Flammability Point

اشتعال نسبت به مقادیر نقطه اشتعال هر یک از اجزای تشکیل دهنده مخلوط می‌گردد [۵، ۶]. آزمون‌های محفظه باز و محفظه بسته برای تعیین نقطه اشتعال مخلوط‌ها به طور تجربی استفاده می‌شوند [۷]. با این حال، اندازه‌گیری نقطه اشتعال در مواردی مانند مخلوط‌های رادیواکتیو و سمی، دشوار و پرهزینه است [۸]. بنابراین، با توجه به اینکه داده‌های نقطه اشتعال برای مخلوط‌ها به راحتی در دسترس نیستند، پیش‌بینی دقیق آنها برای آنالیزهای ایمنی ضروری است. مدل‌های مختلفی برای تخمین نقطه اشتعال مخلوط‌ها گزارش شده است. این مدل‌ها شامل مدل‌های تجربی [۹، ۱۰]، مدل‌های مبتنی بر روش UNIFAC [۱۱-۱۳] و مدل‌های رابطه کمی خصوصیت-ساختار (QSPR) [۱۱، ۱۷-۱۴] است. در مدل تجربی نقطه اشتعال مخلوط‌ها با استفاده از داده‌های تجربی به عنوان توصیف‌کننده (متغیرهای مستقل) پیش‌بینی می‌شود. مدل‌های مبتنی بر روش UNIFAC نیز تا حدودی پیچیده هستند.

در سال‌های اخیر، تخمین نقطه اشتعال با استفاده از رابطه کمی خصوصیت-ساختار (QSPR) مورد توجه قرار گرفته است. مدلسازی QSPR توانایی پیش‌بینی خصوصیت بر اساس صرفاً توصیف‌کننده‌های ساختارهای مولکولی را دارد. مدل‌های پیش‌بینی‌کننده QSPR بار تجربی و هزینه‌های انتخاب عوامل مناسب برای فرایندهای صنعتی را به طور قابل توجهی کاهش می‌دهند. با این حال، تاکنون، روابط کمی خصوصیت-ساختار عمدتاً برای پیش‌بینی نقطه اشتعال ترکیبات خالص به کار برده شده است و فقط تعداد کمی از کارهای اخیر به تخمین خصوصیات مخلوط‌ها اختصاص یافته است [۸، ۱۸-۲۰]. چالش برانگیزترین مسئله در کاربرد مطالعات QSPR در مخلوط‌ها، محاسبه توصیف‌کننده‌های عددی است. برای پیش‌بینی خواص مخلوط‌ها، معمولاً قوانین اختلاط برای محاسبه توصیف‌کننده‌های مخلوط با استفاده از اجزاء تشکیل‌دهنده مخلوط به کار می‌رود. قوانین اختلاط مختلفی پیشنهاد شده است که برخی برای مخلوط‌های ایده‌آل و برخی دیگر برای مخلوط‌های غیر ایده‌آل با در نظر گرفتن برهم‌کنش بین اجزاء تشکیل‌دهنده مخلوط با استفاده از ضرایب فعالیت به کار

می‌رود [۲۱-۲۳]. جیو و همکاران مدل سازی QSPR را برای تعیین نقطه اشتعال مخلوط‌های دو جزئی آلی مورد بررسی قرار دادند. برای محاسبه توصیف‌کننده مخلوط، ابتدا، شاخص وضعیت الکتروتوپولوژیک (وضعیت E) اجزای سازنده مخلوط محاسبه شد. سپس، شاخص وضعیت الکتروتوپولوژیک هر جزء توسط کسر مولی مربوطه وزن داده شد و نهایتاً توصیف‌کننده مخلوط از جمع شاخص‌های وزن داده شده اجزاء تشکیل دهنده مخلوط بدست آمد [۱۸]. ثباتی و همکاران نیز [۸] مدل‌های QSPR جدید برای پیش‌بینی نقطه اشتعال مخلوط‌های آلی چند جزئی ارائه کردند. در این کار، آنها توصیف‌کننده‌های مخلوط را توسط قوانین مخلوط کی؟ قانون اختلاط الهام گرفته از فاصله اقلیدسی و با استفاده از ریشه سوم توصیف‌کننده مخلوط بدست آوردند. اخیراً، گادین و همکاران نیز با به کار بردن فرمول‌های ریاضی مختلف برای محاسبه توصیف‌کننده‌های مخلوط از اجزاء تشکیل‌دهنده، مدل‌های QSPR جدید برای پیش‌بینی نقطه اشتعال مخلوط‌های دوتایی ایجاد کردند [۲۴]. سولوو و همکاران [۲۵] مدل‌های QSPR را برای پیش‌بینی نقطه جوش نرمال آزیوتروپ‌های دوتایی با استفاده از رگرسیون خطی چندگانه گروهی و توصیف‌گرهای مولکولی زیرساختاری ایجاد کردند. توصیف‌گرهای یک مخلوط با ادغام توصیف‌گرهای هر یک از اجزای مولکولی آن تولید شدند. زارع شاه آبادی و همکاران [۲۰] نیز مدل‌های QSPR برای نقطه جوش نرمال آزیوتروپ‌های دوتایی به دست آوردند. توصیف‌گرهای مخلوط، بر اساس دو رویکرد مختلف بود: میانگین توصیف‌گرهای مولکولی مربوط به هر یک از اجزای مولکولی در مخلوط و وزن‌دهی توصیف‌گرهای مولکولی منفرد با کسرهای مولی آنها. در محاسبه حالت دوم، به مقادیر تجربی کسرهای مولی نیاز است.

هدف از این مطالعه، توسعه مدل‌های QSPR جدید و توانمند برای پیش‌بینی نقطه اشتعال مخلوط‌های آلی دو جزئی می‌باشد. پتانسیل فرمول‌های مختلف ریاضی برای توسعه چنین مدل‌هایی جهت پیش‌بینی نقطه اشتعال مورد ارزیابی قرار گرفت. در واقع، تمرکز بر این موضوع به این دلیل بود که تعریف توصیف‌کننده‌های مخلوط به عنوان یک مرحله مهم در مدل‌های QSPR برای مخلوط‌ها، به ویژه برای خواصی که

می‌توانند یک روند غیرخطی با کسر مولی اجزاء تشکیل‌دهنده را دنبال کنند، شناخته شده است.

۲- مواد و روش ها

۲-۱- داده های تجربی

در مطالعه حاضر، یک مجموعه داده بزرگ شامل ۴۰۰ داده تجربی برای مخلوط های آلی دوتایی از منابع منتشر شده استخراج شده است [۲۶-۳۷]. این مجموعه داده شامل ۳۳ ترکیب خالص است. از مخلوط این ترکیب ها با ترکیب درصد های متفاوت، ۴۰۰ مخلوط دوتایی با ویژگی های متفاوت ایجاد گردیده است. ترکیبات شیمیایی با گروه های عاملی مختلف شامل کتون ها، هیدروکربن ها، و الکل ها در این مجموعه داده یافت می شوند. مقادیر نقطه اشتعال در محدوده $27/5^{\circ}\text{C}$ تا 89°C و کسر های مولی در محدوده ۰/۰۰۵ تا ۰/۹۹۵ است. مجموعه داده ها به طور تصادفی به ترتیب به دو دسته آموزش و آزمون با اندازه های ۳۲۰ و ۸۰ تقسیم شدند.

۲-۲- توصیف کننده های مولکولی

برای محاسبه توصیف کننده ها، ساختار مولکولی ترکیبات تشکیل دهنده هر مخلوط بطور جداگانه در ChemDraw 2D ترسیم شد [۳۸]. برای ایجاد ساختار های سه بعدی از ماژول مربوطه در CS Chem3D Ultra استفاده شد. سپس ساختار های سه بعدی در نرم افزار عملیات مولکولی (MOE) ۱ با استفاده از روش AM1 تحت بهینه سازی هندسی قرار گرفتند. در مجموع ۳۴۷۵ توصیف کننده شامل ۲۵۱ توصیف کننده از MOE و ۳۲۲۴ توصیف کننده از نرم افزار دراگون برای توصیف ساختار مولکولی هر ترکیب خالص محاسبه شد [۳۹]. از بین توصیف کننده های محاسبه شده، توصیف کننده هایی که مقادیر صفر یا ثابت داشتند حذف شدند، زیرا این توصیف کننده ها تفاوت های ساختاری میان ترکیباتی که دارای مقادیر نقطه اشتعال متفاوت می باشند را رمز گذاری نمی کنند. سپس، ضریب همبستگی میان توصیف کننده های باقیمانده محاسبه و از بین

توصیف کننده هایی که ضریب همبستگی بالای ۰/۹ داشتند، توصیف کننده ای که بیشترین همبستگی با نقطه اشتعال مخلوط های دو تایی را داشت نگه داشته و بقیه حذف شدند. در نهایت، ۲۵۵ توصیف کننده مولکولی برای هر ترکیب باقی ماند که در قسمت های بعدی کار مورد استفاده قرار گرفت.

۲-۲-۲- توصیف کننده مخلوط

در مطالعات QSPR مخلوط ها، راهکار های متفاوتی برای محاسبه توصیف کننده های مخلوط با استفاده از توصیف کننده های مولکولی ترکیبات خالص به کار برده می شود [۴۰، ۴۱]. در این پژوهش، بیست و دو نوع توصیف کننده مخلوط، از توصیف کننده های مولکولی که از قسمت قبل نتیجه شدند، محاسبه شد. برای این کار، فرمول های مختلفی که بر مبنای معادلات ریاضی خطی و غیرخطی هستند بکار گرفته شدند (جدول ۱). لازم به ذکر است که از بین بیست و دو نوع توصیف کننده برای مخلوط ها، ۱۰ نوع توسط گروه تحقیقاتی ما ارایه شده است [۲۰] و پژوهش حاضر در ادامه کارهای تحقیقاتی قبلی و با هدف نشان دادن توانایی فرمول های ارایه شده در توصیف مخلوط ها جهت مدل سازی QSPR و پیش بینی خصوصیات مختلف آنها است. فهرستی از کلیه توصیف کننده های مخلوط، همراه با توضیحات مختصر و فرمول ریاضی، در جدول ۱ آورده شده است. این توصیف کننده ها را می توان به سه گروه طبقه بندی کرد: (۱) دسته ای که کسر مولی در فرمول آنها استفاده شده است، (۲) دسته ای که کسر مولی و اطلاعات دیگر را در نظر نمی گیرد و (۳) دسته ای که اطلاعات کسر مولی و انرژی پتانسیل را استفاده می کند. آخرین دسته برای اولین بار در گروه تحقیقاتی ما پیشنهاد شده است. انرژی های پتانسیل در نرم افزار MOE با استفاده از روش MNDO Hamiltonian محاسبه شدند. برای توضیح بیشتر، یک مخلوط دوتایی شامل دو جزء خالص A و B را در باید در نظر گرفت. برای درک و مقایسه بهتر انرژی پتانسیل مخلوط با ترکیبات خالص، انرژی پتانسیل سه سیستم شامل AA، BB و AB محاسبه شدند. ساختار مولکول های شرکت کننده در هر سیستم به صورت SIMLES به نرم افزار MOE وارد شد. سپس، بهینه سازی هندسی با استفاده از روش میدان

جدول ۱. انواع توصیف کننده های مخلوط، توصیف و فرمول های مربوط.

توصیف کننده	فرمول	توصیف
centroid	$\frac{D_A + D_B}{2}$	میانگین توصیف کننده های ترکیبات خالص
sqr-diff	$(D_A - D_B)^2$	مربع تفاوت توصیف کننده های ترکیبات خالص
abs-diff	$ D_A - D_B $	قدر مطلق تفاوت توصیف کننده های ترکیبات خالص
fmol-sum	$X_A D_A + X_B D_B$	مجموع توصیف کننده های وزن دار شده توسط کسرهای مولی
fmol-diff	$X_A D_A - X_B D_B$	تفاوت توصیف کننده های وزن دار شده توسط کسر مولی
sqr-fmol	$X_A^2 D_A + X_B^2 D_B$	مجموع توصیف کننده های وزن دار شده توسط مجذور کسرهای مولی
root-fmol	$\sqrt{X_A D_A} + \sqrt{X_B D_B}$	مجموع توصیف کننده های وزن دار شده توسط ریشه کسرهای مولی
sqr-fmol-sum	$(X_A D_A + X_B D_B)^2$	مجموع مربع توصیف کننده های وزن دار شده توسط کسرهای مولی
norm-cont	$\sqrt{(X_A D_A)^2 + (X_B D_B)^2}$	اندازه اقلیدوسی توصیف کننده های وزن داده شده با کسر مولی
mol-dev	$ (D_A - D_B) \times [1 - (X_A - X_B)]$	ترکیب انحرافی ۱: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق تفاوت کسر مولی ها
sqr- mol-dev	$ (D_A - D_B) \times [1 - (X_A^2 - X_B^2)]$	ترکیب انحرافی ۲: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق تفاوت توان دو کسر مولی ها
mol-dev-sqr	$ (D_A - D_B) \times [1 - (X_A - X_B) ^2]$	ترکیب انحرافی ۳: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق توان دو تفاوت کسر مولی ها
poten-sum	$C_A D_A + C_B D_B$	مجموع توصیف کننده های وزن دار شده توسط ضرایب پتانسیل
poten-diff	$C_A D_A - C_B D_B$	تفاوت توصیف کننده های وزن دار شده توسط ضرایب پتانسیل
sqr-poten	$C_A^2 D_A + C_B^2 D_B$	مجموع توصیف کننده های وزن دار شده توسط مربع ضرایب پتانسیل
root-poten	$\sqrt{C_A D_A} + \sqrt{C_B D_B}$	مجموع توصیف کننده های وزن دار شده توسط ریشه ضرایب پتانسیل
sqr-poten-sum	$\sqrt{C_A D_A} + \sqrt{C_B D_B}$	مربع مجموع توصیف کننده های وزن دار شده توسط ضرایب پتانسیل
poten-norm-cont	$\sqrt{(C_A D_A)^2 + (C_B D_B)^2}$	اندازه اقلیدوسی توصیف کننده های وزن داده شده با ضرایب پتانسیل
poten-dev	$ (D_A - D_B) \times [1 - (C_A - C_B)]$	ترکیب انحرافی ۴: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق تفاوت ضرایب پتانسیل
sqr-poten-dev	$ (D_A - D_B) \times [1 - (C_A^2 - C_B^2)]$	ترکیب انحرافی ۵: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق تفاوت توان دو ضرایب پتانسیل
poten-dev-sqr	$ (D_A - D_B) \times [1 - (C_A - C_B) ^2]$	ترکیب انحرافی ۶: حاصلضرب قدر مطلق تفاوت توصیف کننده های مولکولی و یک عبارت وابسته به قدر مطلق توان دو تفاوت ضرایب پتانسیل
fmol-poten-sum	$C_A X_A D_A + C_B X_B D_B$	مجموع توصیف کننده های وزن دار شده توسط کسر مولی و ضرایب پتانسیل

$$C_A = \frac{(E_{AB} - E_{AA})^2}{[(E_{AB} - E_{AA})^2 + (E_{AB} - E_{BB})^2]} \quad (1)$$

$$C_B = \frac{(E_{AB} - E_{BB})^2}{[(E_{AB} - E_{AA})^2 + (E_{AB} - E_{BB})^2]} \quad (2)$$

که در آن E_{AB} ، E_{AA} و E_{BB} به ترتیب به کل انرژی پتانسیل (kcal mol^{-1}) محاسبه شده برای سیستم های حاوی AA، BB و AB اشاره دارد. اگر چه با روش فوق، می توان تخمین های تقریبی از انرژی های پتانسیل را به دست آورد، ضرایب منتج شده، یعنی C_A و C_B ، سهم هر ترکیب خالص در پایداری مخلوط را نشان خواهد داد.

نیروی MMFF94x انجام و در آخر انرژی پتانسیل با استفاده از همان نرم افزار MOE محاسبه گردید [۴۲]. در مرحله بعد، انرژی پتانسیل سیستم AB به طور جداگانه با سیستم های AA و BB مقایسه می شود. اگر، به عنوان مثال، انرژی پتانسیل سیستم AB به سیستم AA نزدیک باشد، فرض می شود که A بیشترین سهم را در پایداری مخلوط AB دارد. بنابراین، باید در محاسبه توصیف کننده مخلوط، به A وزن بیشتری داده شود. ضرایب انرژی پتانسیل که سهم هر ترکیب خالص در پایداری مخلوط را نشان می دهد، با استفاده از معادلات ۱-۲ به دست آمدند:

۲-۳- ایجاد مدل و انتخاب توصیف کننده

در این مطالعه، رگرسیون خطی چندگانه (MLR) برای ارتباط نقاط اشتغال مخلوط های آلی دوتایی به ویژگی های فیزیکی و شیمیایی مخلوط ها به کار گرفته شد. توصیف کننده های مخلوط ها یا همان ویژگی های فیزیکی و شیمیایی بر اساس جدول ۱ و با استفاده از توصیف کننده های مولکولی محاسبه شد. علیرغم مزایای سادگی و تفسیر آسان، مشکل اصلی رگرسیون خطی چندگانه نیاز به یک ابزار انتخاب متغیر کارآمد برای انتخاب زیر مجموعه ای مناسب از متغیرها (توصیف کننده ها) است. در این مطالعه از الگوریتم بهینه سازی کولونی مورچه حافظه دار (Memorized-ACO)، برای انتخاب بهترین توصیف کننده ها و ایجاد مدل MLR استفاده شد. این الگوریتم از رفتار مورچه های واقعی، که می توانند کوتاه ترین مسیر را از یک منبع غذایی به لانه پیدا کنند، الهام گرفته شده است. الگوریتم ACO حافظه دار از حافظه خارجی مبتنی بر دانش تکرارهای قبلی ACO استفاده می کند. در ابتدا حافظه خالی است، اما با اجرای چندین بار الگوریتم ACO پر می شود. بعد از هر ACO، بهترین مورچه منتخب در حافظه ذخیره می شود و فرایند تا پر شدن کل حافظه ادامه می یابد. به روزرسانی فرومون توسط کل مورچه های منتخب جمع آوری شده در حافظه انجام می شود. سپس حافظه خالی شده و با انجام چندین الگوریتم ACO با استفاده از مسیرهای فرومون به روز شده دوباره پر می شود. این روند چندین مرتبه تکرار می شود. در پایان، حافظه شامل چندین راه حل برتر مساله پیش رو است. تعداد دفعات ظاهر شدن هر توصیف کننده در حافظه خارجی ملاک اهمیت آن است. سرانجام، پیش بینی توسط منتخب ترین مورچه و تفسیر با در نظر گرفتن اهمیت هر توصیف کننده انجام می شود [۴۳-۴۵]. قابلیت اطمینان مدل های نهایی ساخته شده با استفاده از اعتبار سنجی داخلی و خارجی ارزیابی گردید. مقادیر آماری R^2_{training} ، R^2_{test} و q^2 برای مقایسه مدلها استفاده شد. R^2_{training} و R^2_{test} به ترتیب مجذور ضریب همبستگی بین نقاط اشتغال پیش بینی شده و مقادیر تجربی برای مجموعه آموزش و آزمون می باشند. q^2 مجذور ضریب همبستگی

است که از طریق روش ارزیابی متقابل یکتایی (LOO-CV) بدست آمده است. اعتبار مدل های MLR بدست آمده توسط روش های ارزیابی تقاطع مونته کارلو (MCCV)، ارزیابی متقاطع ۵ برابری (5-CV)، و ارزیابی متقاطع دو حلقه ای (D-CV) نیز مورد بررسی قرار گرفت. روش ارزیابی متقاطع یکتایی در یک زمان، یک داده از داده ها را از سری آموزش حذف می کند، مدل را با داده های باقی مانده می سازد و این مدل برای پیش بینی خصوصیت داده بیرون گذاشته شده به کار می رود. این شیوه تکرار می شود تا اینکه خصوصیت مربوط به تمام داده ها (در اینجا، مخلوط های دوتایی) توسط این روش پیش بینی شود. ارزیابی متقاطع پنج برابری، مجموعه داده را به ۵ زیر مجموعه تقسیم کرده، هر بار چهار زیر مجموعه از ۵ زیر مجموعه برای تشکیل یک مجموعه آموزشی و یک مجموعه باقیمانده را به عنوان مجموعه آزمون در نظر گرفته و متوسط خطا در پنج ارزیابی محاسبه می گردد [۴۶]. در روش ارزیابی متقاطع مونته کارلو، داده ها مجموعه آموزش به طور تصادفی و با نسبت معین از پیش تعیین شده به دو دسته آموزش و آزمون تقسیم می شوند. مدل سازی برای دسته آموزش جدید انجام و مقادیر نقطه اشتغال دسته آزمون پیش بینی می شود و مقادیر مجذور ضریب همبستگی مقادیر پیش بینی شده و مقادیر تجربی دسته آزمون محاسبه و ثبت می شود. روش ارزیابی متقاطع پنج برابری و مونته کارلو ۱۰۰ مرتبه تکرار می شوند و از مقادیر محاسبه شده میانگین گرفته و گزارش می شوند [۴۷، ۴۸]. ارزیابی متقاطع دو حلقه روشی مناسب هم برای انتخاب مدل و هم ارزیابی مدل است [۴۹]. این روش شامل دو حلقه داخلی و خارجی است. در ابتدا، کل داده های آموزش به طور تصادفی به دو زیر مجموعه آموزش و آزمون تقسیم می شوند. مجموعه آزمون فقط برای ارزیابی مدل استفاده می شود. مجموعه آموزش در حلقه داخلی استفاده شده و به مجموعه داده های آموزش و ارزیابی به طور مکرر تقسیم می شود (برای مثال ۵۰ مرتبه). مدل توسط مجموعه داده آموزش ساخته شده ولی مجموعه داده آزمون برای تخمین خطای مدل استفاده می شود. کل الگوریتم D-CV، ۴۰۰ مرتبه تکرار شد.

*Leave-one-out Cross Validation

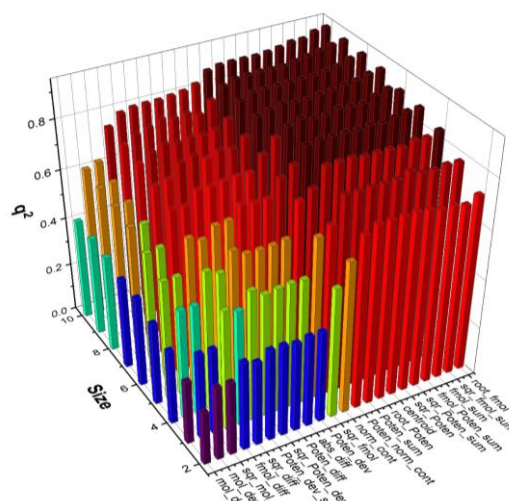
*Monte Carlo Cross Validation

*5-Fold Cross Validation

*Double Cross Validation

*Multiple Linear Regression

*Memorized_Ant Colony Optimization



شکل ۱. مقادیر q^2 مدل های QSPR ایجاد شده با اندازه های متفاوت بر اساس انواع مختلف توصیف کننده های مخلوط.

کننده مخلوط، بقیه توصیف کننده های مخلوط بکار گرفته شده در ایجاد مدل های QSPR پیش‌بینی‌کننده، توانمند هستند. ده توصیف‌کننده مخلوط با کارایی پایین عبارتند از شش توصیف‌کننده $mol-dev$ ، $sqr-mol-dev$ ، $sqr-dev$ ، $poten-dev$ ، $poten-dev-sqr$ ، $sqr-poten-dev$ (به دست آمده از فرمول انحراف) و چهار توصیف‌کننده $sqr-diff$ ، $abs-diff$ ، $fmol-diff$ و $poten-diff$ که براساس رابطه تفاضل ایجاد گردیده بودند. مقادیر آموزش R^2 و آزمون R^2 مدل‌های MLR بدست آمده با استفاده از این توصیف کننده ها به ترتیب در محدوده $0.70-0.72$ و $0.71-0.73$ قرار داشتند. محدوده q^2 برای این توصیف کننده ها $0.65-0.66$ بود. میانگین خطای مطلق (MAE) برای مجموعه آموزش $8/43$ و مجموعه آزمون $9/89$ °C بود که خطای زیادی را نشان می‌دهد. مدل های QSPR به دست آمده با استفاده از باقی توصیف‌کننده‌ها از کیفیت آماری خوبی برخوردار بودند. مقادیر آموزش R^2 ، آزمون R^2 و q^2 برای مدل‌های MLR با اندازه‌های مختلف به ترتیب در محدوده $0.70-0.90$ ، $0.63-0.88$ و $0.69-0.89$ بدست آمد که برآزش خوب مدل‌ها را نشان می‌دهد. میانگین مقادیر MAE مجموعه آموزش و آزمون نیز به ترتیب در محدوده $4/74-8/61$ و $10/09-5/35$ درجه سانتی گراد بود. بر اساس پارامترهای آماری مدل‌ها و همانطور که در شکل ۱ دیده می‌شود، ترتیب توصیف‌کننده‌های مخلوط با کارایی قابل قبول در ایجاد مدل‌های QSPR پیش‌بینی‌کننده به صورت زیر می‌باشد:

تمام محاسبات بر روی یک رایانه شخصی با Core™ i7 3/50 GHz Intel CPU و رم ۴ گیگابایتی اجرا شد. سیستم عامل میکروسافت ویندوز XP بود. تمام برنامه‌های لازم در متلب (MATLAB, MathWork) نوشته شده است.

۳- نتایج و بحث

در این پژوهش، پتانسیل ۲۲ نوع مختلف از توصیف‌کننده‌های مخلوط برای ایجاد مدل‌های QSPR جدید جهت پیش‌بینی نقطه اشتعال مخلوط های دو جزئی مورد ارزیابی قرار گرفت. این توصیف‌کننده‌ها براساس فرمول‌های ریاضی متفاوت و با استفاده از توصیف‌کننده‌های مولکولی که برای هر جزء بدست آمده بود، محاسبه شدند. جدول ۱ اطلاعات مورد نیاز این توصیف‌کننده‌ها را ارائه می‌دهد.

لازم به ذکر است که برای هر نوع توصیف‌کننده مخلوط، تعداد ۲۵۵ توصیف‌کننده که ویژگی‌های مختلفی مانند ویژگی‌های فیزیکی و شیمیایی یک مخلوط را توصیف می‌کنند، محاسبه شده است. برای ایجاد مدل MLR، لازم است، از میان این انبوه توصیف‌کننده‌ها، تعداد معدودی، اغلب زیر ۱۰ عدد، انتخاب گردند. انتخاب بهترین توصیف‌کننده‌ها از میان یک جمعیت انبوه یک مساله NP سخت می‌باشد. مساله NP سخت به مساله‌ای گفته می‌شود که زمان برای حل دقیق آن با افزایش اندازه مساله (در اینجا، تعداد توصیف‌کننده‌ها) به صورت نمایی بالا می‌رود. بنابراین در مواجهه با مسائلی از این دست، به جای حل دقیق، با استفاده از الگوریتم‌هایی همچون کولونی مورچگان و ژنتیک الگوریتم، راه حل تقریبی ارایه می‌گردد. در اینجا، برای هر نوع توصیف‌کننده، مدل‌های MLR با اندازه ۲ تا ۱۰ با استفاده از توصیف‌کننده‌هایی که الگوریتم Memorize انتخاب کرده است، ایجاد می‌گردد. ارزیابی آماری انواع توصیف‌کننده‌های مخلوط در ایجاد مدل‌های QSPR با اندازه های مختلف در جدول S1 پیوست آورده شده است. شکل ۱ مقادیر q^2 محاسبه شده برای مدل‌های MLR ساخته شده برای انواع توصیف‌کننده‌های مخلوط را نشان می‌دهد. نمودارهای مربوط به آموزش R^2 و آزمون R^2 در شکل‌های S1-S2 فایل پیوست آورده شده است. نتایج ارایه شده در شکل ۱ و جدول S1 فایل پیوست نشان می‌دهند که به جز ۱۰ توصیف

در ستون‌های ۹-۱۰ جدول ۲ آورده شده است، تایید می‌شوند. نتیجه ارزیابی D-CV در مقادیر ارزیابی R^2_{D-CV} و آزمون R^2_{D-CV} خلاصه می‌شوند و همانطور که در جدول ۲ مشهود است، هر دو به طور همزمان با افزایش اندازه مدل تا ۶ افزایش می‌یابند. لازم به یادآوری است که فقط مجموعه آموزش در فرآیند ارزیابی متقاطع دو برابری (ارزیابی D-CV) به کار برده می‌شود. برای بهترین مدل QSPR ایجاد شده بر اساس توصیف‌کننده $sqr-fmol-sum$ و با اندازه ۶، مقادیر آموزش R^2 ، آزمون R^2 و q^2 به ترتیب ۰/۹۱، ۰/۸۸ و ۰/۹۰ بدست آمد. مقادیر MAE برای مجموعه آموزش و آزمون نیز به ترتیب ۴/۵۱ و ۴/۴۴ محاسبه شد. پارامترهای آماری بیشتری برای تایید بهترین مدل QSPR بدست آمده در جدول ۳ آورده شده است.

جدول ۲. پارامترهای آماری مدل های QSPR بر اساس توصیف کننده $sqr-fmol-sum$ با اندازه‌های مختلف.

تعداد توصیف کننده	آموزش R^2	آموزش RMSE	q^2	$RMSE_{cv}$	$5CV_a$		$MCCV_b$		$R^2_{testDCV}$	$R^2_{DCVvalidation}$	آموزش R^2	آموزش RMSE	F	آموزش MAE	آزمون MAE
					آموزش R^2	آموزش RMSE	آموزش R^2	آموزش RMSE							
۲	۰/۷۱	۱۱/۶	۰/۷۰	۱۱/۸	۰/۶۹	۱۱/۸	۰/۷۱	۰/۷۱	۰/۷۱	۰/۷۱	۰/۷۱	۱۳/۳	۳۸۱/۸	۸/۱۲	۹/۳۱
۳	۰/۷۹	۹/۷	۰/۷۹	۹/۹	۰/۷۹	۹/۹	۰/۷۹	۰/۷۹	۰/۷۹	۰/۷۸	۰/۷۷	۷/۸	۴۰۴/۵	۷/۰۲	۶/۶۱
۴	۰/۸۴	۸/۶	۰/۸۳	۸/۸	۰/۸۳	۸/۸	۰/۸۳	۰/۸۴	۰/۸۳	۰/۸۳	۰/۷۰	۹/۱	۴۰۵/۶	۶/۵۷	۷/۳۵
۵	۰/۸۸	۷/۵	۰/۸۷	۷/۶	۰/۸۷	۷/۷	۰/۸۷	۰/۸۸	۰/۸۷	۰/۸۷	۰/۸۲	۷/۱	۴۵۱/۸	۵/۳۶	۵/۵۸
۶	۰/۹۱	۶/۴	۰/۹۰	۶/۶	۰/۹۰	۶/۷	۰/۹۰	۰/۹۱	۰/۹۰	۰/۹۰	۰/۸۸	۵/۹	۵۳۳/۹	۴/۵۱	۴/۴۴
۷	۰/۹۲	۶/۲	۰/۹۱	۶/۵	۰/۹۱	۶/۵	۰/۹۱	۰/۹۲	۰/۹۰	۰/۹۰	۰/۸۷	۵/۸	۴۸۰/۳	۴/۵۹	۴/۵۳
۸	۰/۹۲	۵/۹	۰/۹۲	۶/۲	۰/۹۲	۶/۲	۰/۹۲	۰/۹۲	۰/۹۱	۰/۹۱	۰/۸۹	۵/۴	۴۷۱/۸	۴/۱۳	۳/۹۳
۹	۰/۹۳	۵/۸	۰/۹۲	۶/۰	۰/۹۲	۶/۱	۰/۹۲	۰/۹۳	۰/۹۲	۰/۹۲	۰/۸۸	۵/۹	۴۲۹/۸	۴/۰۲	۴/۷۰
۱۰	۰/۹۳	۵/۶	۰/۹۲	۵/۹	۰/۹۲	۵/۹	۰/۹۲	۰/۹۳	۰/۹۲	۰/۹۲	۰/۸۸	۵/۷	۴۱۳/۹	۳/۹۸	۴/۵۵

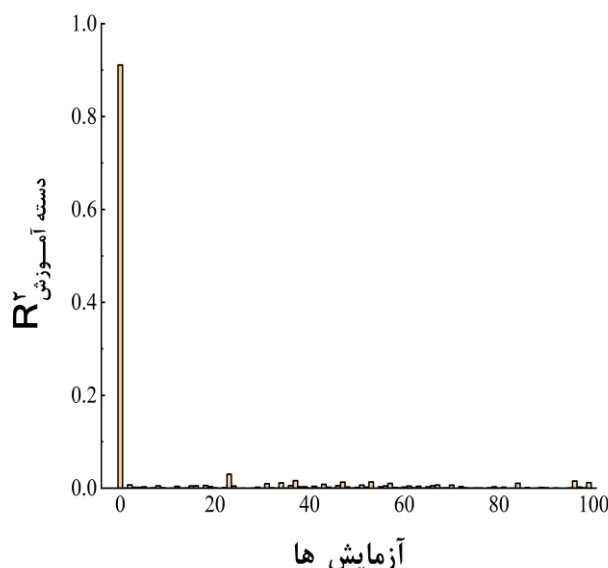
^a مقادیر گزارش شده میانگین ۱۰۰ تکرار 5-CV است.

^b مقادیر گزارش شده میانگین ۱۰۰ تکرار MCCV است.

جدول ۳. پارامترهای آماری برای بهترین مدل QSPR بر اساس توصیف کننده $sqr-fmol-sum$.

توصیف کننده ها	توصیف	beta	t	p-value	VIF
intercept		۱۵/۳۷	۴۲/۶۸	$10^{-133} \times 1/6$	
PEOE_PC+	Total positive partial charge	۲۳/۳۱	۴۳/۶۰	$10^{-135} \times 5/1$	۲/۲
vsurf_D8	Properties of molecular interaction fields with a hydrophobic DRY probe	۲۰/۳۸	۲۴/۳۰	$10^{-74} \times 4/8$	۵/۴
Jhetm	Balaban-type index from mass weighted distance matrix	-۸/۱۲	-۱۶/۵۰	$10^{-44} \times 1/8$	۱/۹
MAXDP	maximal electrotopological positive variation	۷/۰۲	۱۶/۲۰	$10^{-43} \times 2/7$	۱/۴
EEig08d	Eigenvalue 08 from edge adj. matrix weighted by dipole moments	-۸/۰۷	-۱۳/۳۴	$10^{-32} \times 2/0$	۲/۸
H6m	H autocorrelation of lag6 / weighted by atomic masses	-۱۸/۹۶	-۲۶/۸۴	$10^{-83} \times 3/4$	۳/۸

که امکان ارزیابی عدم قطعیت در پیش‌بینی خصوصیت یک ترکیب جدید را فراهم می‌کند. شایان ذکر است که یک مدل QSPR بسته به نوع ترکیبات شیمیایی استفاده شده در ایجاد مدل و همچنین مقادیر توصیف‌کننده‌های در نظر گرفته شده، دامنه کاربرد خاص خود را دارد. از این رو، تمام مدل‌های MLR ایجاد شده دارای دامنه کاربرد متفاوتی هستند زیرا بر اساس انواع متفاوتی از توصیف‌کننده‌های مخلوط به دست آمده‌اند. از میان راهکارهای متفاوتی که برای تعریف دامنه کاربرد ارایه شده است، در اینجا نمودار ویلیامز به کار برده شد [۴۷، ۵۱]. این نمودار با استفاده از مقیاس دوری و درجه نفوذ مخلوط‌های دوتایی، داده‌های پرت را شناسایی می‌کند. باقیمانده‌های استاندارد شده را در مقابل درجه نفوذ^۲ نمایش می‌دهد. باقیمانده‌های استاندارد، تفاوت بین مقادیر مشاهده شده و مقادیر پیش‌بینی شده مقیاس‌بندی شده و درجه نفوذ (h) تأثیر آن ترکیب بر مدل را نشان می‌دهد. اگر مقادیر باقیمانده استاندارد خارج از محدوده ± 3 قرار گیرند، داده به عنوان داده‌ی پرت شناسایی می‌شود. همچنین نقاطی که درجه نفوذ آن‌ها بیشتر از مقدار بحرانی باشند، از لحاظ ساختاری بر مدل تأثیر زیادی دارند. مقدار بحرانی برای درجه نفوذ به صورت زیر تعریف می‌شود:



شکل ۲. آزمون بهم ریختگی Y به کار برده شده برای مدل QSAR ایجاد شده بر اساس root-fmol. نوار اول مقدار آموزش R^2 را برای مدل بر اساس داده‌های واقعی نشان می‌دهد.

همانطور که مشاهده می‌شود، تمام مقادیر t معنی دار بودند. فاکتور تورم واریانس (VIF) برای کلیه توصیف‌کننده‌های انتخابی در این مدل کمتر از ۱۰ بود که نشان دهنده عدم وجود همبستگی چندگانه در بین توصیف‌کننده‌ها است [۵۰]. همچنین، به منظور نشان دادن عدم وابستگی دوتایی بین توصیف‌کننده‌ها، ماتریس همبستگی در جدول ۴ آورده شده است. داده‌ها جدول ۴ به وضوح نشان می‌دهند که هیچ همبستگی معناداری میان توصیف‌کننده‌ها وجود ندارد.

برای امکان‌سنجی همبستگی شانس در مدل QSPR انتخابی، آزمون به هم ریختگی Y استفاده شد. این آزمون الگوریتم تکراری دارد و معمولاً ۱۰۰ مرتبه اجرا می‌شود. در هر مرتبه، مقادیر نقطه اشتغال به طور تصادفی به مولکول‌های متفاوت اختصاص داده می‌شود و سپس، یک مدل رگرسیون خطی با توصیف‌کننده‌های انتخابی (جدول ۳) با استفاده از نقاط اشتغال بهم ریخته، ساخته شده و ضریب همبستگی محاسبه می‌شود. در صورت شناسی بودن مدل حاصله، پارامتر آموزش R^2 برای مدل‌های ساخته شده در این آزمون نیز مقادیر بالایی را از خود نشان می‌دهند. همانطور که در شکل ۲ نشان داده شده است بیشترین میزان همبستگی مربوط به داده‌های واقعی بوده است. همبستگی پایین مربوط به ۱۰۰ بار بهم ریختگی نشان دهنده‌ی این است که ارتباط شانس بین مقادیر تجربی نقطه اشتغال و توصیف‌کننده‌های محاسباتی وجود ندارد و اثباتی است بر این حقیقت که مدل اصلی بصورت شانس توسعه نیافته و اهمیت مدل را در پیش‌بینی نقطه اشتغال مخلوط‌های دو جزئی تأیید می‌کند. QSPR تعریف شد. دامنه کاربرد یک منطقه فضایی تئوری است جدول ۴. ماتریس همبستگی میان توصیف‌کننده‌های ظاهر شده در بهترین مدل.

H6m	EEig08d	MAXDP	Jhetm	vsurf_D8	PEOE_PC+
					PEOE_PC+
				۱	۰/۰۴
			۱	۰/۲۴	۰/۰۲
		۱	۰/۱۰	۰/۰۲	۰/۰۶
	۱	۰/۰۳	۰/۰۶	۰/۳۷	۰/۰۳
۱	۰/۰۱	۰/۰۰	۰/۰۴	۰/۳۱	۰/۳۵

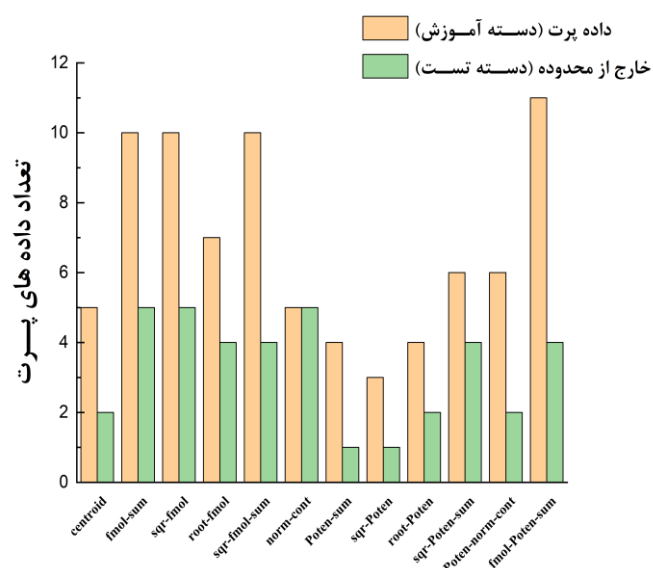
$$h^* = (3 \times (p + 1)) / n \quad (3)$$

که p تعداد پارامترهای مدل و n تعداد مخلوط های دوتایی در دسته آموزش می باشد. برای هر توصیف کننده مخلوط، بهترین اندازه مدل یا تعداد بهینه توصیف کننده ظاهر شده در مدل با توجه به ماکسیمم مقدار q^2 انتخاب گردید.

با توجه به اینکه در این مقاله از چندین فرمول ریاضی مختلف برای محاسبه توصیف کننده های مخلوط استفاده شده، برای بررسی بیشتر میزان کاربردی بودن آنها، دامنه کاربردی برای هر نوع توصیف کننده مخلوط با در نظر گرفتن بهترین اندازه مدل محاسبه و داده ها پرت یا دورافتاده شناسایی شدند. براساس تعداد داده های پرت شناسایی شده در دسته آموزش و تعداد داده های خارج محدوده در دسته آزمون، عملکرد توصیف کننده های استفاده شده به صورت زیر می باشند (شکل ۳):

$\text{sqr-Poten} < \text{Poten-sum} < \text{root-Poten} < \text{centroid} < \text{Poten-norm-cont} < \text{norm-cont} < \text{sqr-Poten-sum} < \text{root-fmol} < \text{sqr-fmol-sum} < \text{fmol-sum} < \text{sqr-fmol} < \text{fmol-Poten-sum}$

همانطور که از ترتیب بدست آمده مشهود است، بهترین توصیف کننده از حیث داشتن کمترین داده پرت و خارج از محدوده و یا داشتن بزرگترین دامنه کاربردی، sqr-Poten و سپس Poten-sum می باشد. در ادامه نیز یکی دیگر از توصیف کننده هایی که بر مبنای پتانسیل های انرژی محاسبه

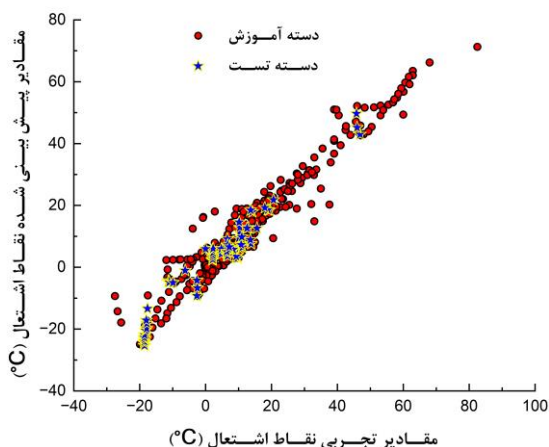


شکل ۳. مقایسه تعداد داده های پرت مدل های QSPR متفاوت ایجاد شده بر اساس انواع توصیف کننده های مخلوط.

شدند، قرار دارد. این ترتیب، قدرت کاربردی فرمول های ارائه شده بر مبنای پتانسیل های انرژی برای محاسبه توصیف کننده های مخلوط را نشان می دهد. لازم به ذکر است که این فرمول ها اولین بار در گروه تحقیقاتی ما ارائه شده اند. شکل ۴ بهبود پارامترهای آماری مختلف را برای بهترین مدل QSPR که براساس sqr-fmol-sum بدست آمد، را نشان می دهد. میزان بهبود در آماره ها، در این شکل به وضوح دیده می شود. مقادیر آموزش R^2 ، آزمون R^2 و q^2 قبل از حذف داده های پرت به ترتیب ۰/۸۸، ۰/۹۱ و ۰/۹۰ و بعد از حذف داده های پرت به ترتیب به ۰/۹۳، ۰/۹۴ و ۰/۹۳ بهبود پیدا کردند. شکل ۵ مقادیر پیش بینی شده برحسب مقادیر تجربی نقاط اشتعال با بهترین مدل و بعد از حذف داده های پرت را نشان می دهد. فهرست کاملی از داده های پرت و داده های



شکل ۴. مقایسه آماره های مختلف برای مدل QSPR به دست آمده بر اساس توصیف کننده sqr-fmol-sum .



شکل ۵. نقاط اشتعال تجربی در مقابل مقادیر پیش بینی شده توسط مدل QSPR ایجاد شده بر اساس sqr-fmol-sum بعد از حذف داده های پرت.

خارج از محدوده کاربردی شناسایی شده و همچنین نقاط اشتعال‌پذیری تجربی و پیش‌بینی شده برای آنها در جدول ۵ آورده شده است. همانطور که قابل مشاهده است مخلوط دوتایی بعضی از ترکیبات مانند متانول/ ایزو-کتان با ترکیب درصدی متفاوت به عنوان داده پرت شناسایی شده‌اند. قابل تامل اینکه در بعضی از ترکیب درصدها، نقطه اشتعال‌پذیری این مخلوط دوتایی به خوبی پیش‌بینی شده است، به طوریکه از ۲۷ ترکیب درصد مختلف مخلوط متانول/ ایزو-کتان، ۱۲ ترکیب درصد در دسته آزمون ظاهر شده‌اند و از این ۱۲ ترکیب درصد، ۷ ترکیب درصد به عنوان داده خارج از محدوده کاربردی شناسایی شده‌اند. از میان ترکیب درصدیهای مختلف مخلوط متانول/ ایزو-کتان، آنهایی که میزان کسر مولی متانول کمتر از ۰/۳ داشتند به خوبی پیش‌بینی شده بودند. در مورد مخلوط استون/ ان-دکان می‌توان گفت که برای تقریباً تمام ترکیب درصدها، از کسر مولی پایین استون تا کسر مولی بالای آن، پیش‌بینی مدل برای این مخلوط قابل قبول نبوده است. در مجموع، با بررسی ساختار مولکولی اجزاء مخلوط‌هایی که به عنوان داده پرت یا داده خارج از محدوده شناسایی شده‌اند، نمی‌توان نتیجه‌گیری کرد که محدوده کاربردی مدل چگونه خواهد بود.

همانطور که عنوان شد، از بین ۲۲ نوع روش محاسبه توصیف‌کننده‌های مخلوط براساس توصیف‌کننده‌های مولکولی محاسبه شده برای اجزاء سازنده مخلوط، ۱۲ نوع از آنها منجر به توصیف‌کننده‌هایی شدند که براساس آنها مدل‌های QSPR با کیفیت خوب ایجاد شدند. از میان این ۱۲ نوع، بهترین عملکرد مربوط به sqr-fmol-sum بود، درحالی‌که بقیه توصیف‌کننده‌ها نیز عملکرد خوبی داشتند. داده‌ها ارایه شده در شکل ۱ و جدول S1 و اشکال S1-S2 این حقیقت را تایید می‌کند. در هر حال، در دست داشتن چندین مدل QSPR با قابلیت پیش‌بینی‌کنندگی خوب، استفاده از پیش‌بینی گروهی را ترغیب می‌کند [۵۲]. در مقایسه با پیش‌بینی یک مدل واحد، پیش‌بینی‌های انجام شده توسط گروهی از مدل‌ها از صحت بالاتری برخوردار خواهد بود که به دلیل ادغام پیش‌بینی چندین مدل مستقل برای رسیدن به پیش‌بینی نهایی است. این امر سبب می‌شود که خطاهای تصادفی که در نتایج ظاهر شده‌اند همدیگر را خنثی و حذف کنند. تمام ۱۲ مدل QSPR با قدرت پیش‌بینی بالا در ایجاد یک مدل

گروهی از طریق یک راهکار میانگین‌گیری ساده لحاظ شدند. پارامترهای آماری آموزش R^2 ، آزمون Q^2 و R^2 مربوط به مقادیر پیش‌بینی شده از مدل گروهی به ترتیب برابر با ۰/۹۴، ۰/۹۳ و ۰/۹۳ بودند. در مقایسه با مقادیر پیش‌بینی شده از مدل تکی بر مبنای sqr-fmol-sum نتایج بهبود پیدا کرده‌اند. لازم به ذکر است که در استفاده از مدل گروهی، برای هیچکدام از انواع توصیف‌کننده‌ها داده‌های پرت حذف نشده بودند. مهمترین علت آن امر، تنوع در داده‌های پرت شناسایی شده بود و برای ایجاد مدل گروهی لازم بود که تمام داده‌های پرت شناسایی شده برای هر توصیف‌کننده، از تمام دسته داده‌های مورد استفاده برای بقیه توصیف‌کننده‌ها حذف گردد. در نتیجه، تعداد زیادی از داده‌ها از دسته داده اولیه حذف خواهند شد که باعث محدود شدن کاربرد و گستردگی مدل‌های ایجاد شده می‌گردد. نکته دیگر اینکه، استفاده از راهکار پیش‌بینی گروهی به بهبود نتایج کمک می‌کند و این شک وجود داشت که با استفاده از این راهکار می‌توان بر مشکل اشاره شده فائق آمد. متأسفانه بررسی نتایج در جدول ۵ نشان می‌دهد که در اغلب موارد این راهکار بهبودی در کیفیت پیش‌بینی برای داده‌های پرت و داده‌های خارج از محدوده ایجاد نمی‌کند.

۳-۱- جزئیات بهترین مدل

بر اساس نتایج، مدل QSPR به دست آمده از sqr-fmol-sum به عنوان بهترین مدل انتخاب گردید. پس از حذف داده‌های پرت، مدل بدست آمده توسط معادله (۵) نشان داده شده است.

$$\begin{aligned} \text{FP} = & 13.72 + 20.07 \times (\pm 0.41) \times \\ & \text{PEOE_PC} + 15.87 \times (\pm 0.70) \times \\ & \text{vsurf_D8} - 7.05 \times (\pm 0.44) \times \text{Jhetm} + \\ & 5.85 \times (\pm 0.40) \times \text{MAXDP} - 7.32 \times \\ & (\pm 0.55) \times \text{EEig08d} - 7.86 \times (\pm 0.39) \times \\ & \text{H6m} \end{aligned} \quad (5)$$

که در معادله (۵)، FP نقطه اشتعال‌پذیری بر حسب درجه سانتی‌گراد است. لازم به ذکر است که ضرایب داده شده در معادله ۵ متفاوت از ضرایب موجود در جدول ۳ است، زیرا قبل از توسعه معادله ۵، تمام داده‌های پرت شناسایی شده حذف شدند. بزرگی ضرایب مربوط به هر توصیف‌کننده در این مدل نشانگر اهمیت

جدول ۵. داده‌های پرت و خارج از محدوده کاربردی به همراه نقاط اشتعال‌پذیری پیش‌بینی شده توسط مدل QSPR بر اساس توصیف کننده sqr-fmol-sum و تخمین گروهی.

نقاط اشتعال بذیر									
ترکیب ۱	ترکیب ۲	کسر مولی ۱	کسر مولی ۲	مقادیر تجربی	پیش بینی شده		RE%	% گروهی	
					sqr-fmol-sum	مدل گروهی			
داده های پرت									
۱	فنول	استوفنون	۰/۵۰	۰/۵۰	۸۹	۱۰۹/۱	۹۲/۹	۲۲/۶	۴/۳
۲	ایزو-آمیل الکل	ایزو-آمیل استات	۰/۳۰	۰/۷۰	۳۸/۳	۴۶/۱	۴۴/۶	۲۰/۴	۱۶/۶
۳	۱-اکتانول	ان-آمیل استات	۰/۱۰	۰/۹۰	۴۱	۳۶/۰	۴۰/۱	-۱۲/۳	-۲/۳
۴	استون	ان-دکان	۰/۰۱	۱/۰۰	۴۱/۹	۱۹/۸	۱۲/۲	-۵۲/۸	-۷۰/۸
۵	ان-دکان	ان-دودکان	۰/۷۸	۰/۲۲	۵۶	۳۴/۳	۴۳/۹	-۳۸/۷	-۲۱/۵
۶	۱-اکتانول	ان-آمیل استات	۰/۵۰	۰/۵۰	۵۰	۴۰/۳	۴۹/۲	-۱۹/۳	-۱/۷
۷	اتانول	ان-تترا دکان	۰/۰۱	۰/۹۹	۳۹/۵	۲۳/۵	۲۴/۳	-۴۰/۶	-۳۸/۴
۸	هگزان-۱-ال	سیکلو هگزانون	۰/۱۰	۰/۹۰	۴۶/۱	۵۴/۳	۵۲/۰	۱۷/۸	۱۲/۷
۹	متیل استات	متیل آکریلات	۰/۱۰	۰/۹۰	-۲/۹	-۳/۸	-۵/۰	۳۱/۳	۷۲/۶
۱۰	استون	ان-دکان	۰/۱۰	۰/۹۰	-۱۱/۹	۱۳/۰	۹/۴	-۲۰۹/۳	-۱۷۹/۳
۱۱	ان-دکان	ان-دودکان	۰/۱۲	۰/۸۹	۷۶	۷۷/۶	۵۵/۲	۲/۰	-۲۷/۳
۱۲	استون	ان-دکان	۰/۰۸	۰/۹۲	-۱۱/۶	۱۴/۴	۱۰/۰	-۲۲۴/۱	-۱۸۶/۰
۱۳	۱-اکتانول	ان-آمیل استات	۰/۲۵	۰/۷۵	۴۴	۳۵/۳	۴۲/۷	-۱۹/۷	-۳/۰
۱۴	اتانول	ان-تترا دکان	۰/۰۲	۰/۹۸	۲۹/۴	۲۳/۶	۲۴/۲	-۱۹/۷	-۱۷/۸
۱۵	استون	ان-دکان	۰/۰۷	۰/۹۳	-۶/۶	۱۵/۱	۱۰/۳	-۳۲۸/۶	-۲۵۵/۴
۱۶	اتانول	ان-تترا دکان	۰/۱۰	۰/۹۰	۲۰	۲۴/۷	۲۳/۱	۲۳/۵	۱۵/۳
۱۷	اتانول	ان-هگزان	۰/۶۸	۰/۳۲	-۲۴/۲	-۲/۴	-۸/۸	-۹۰/۲	-۶۳/۷
۱۸	استون	ان-دکان	۰/۲۰	۰/۸۰	-۱۴/۸	۶/۵	۶/۸	-۱۴۳/۶	-۱۴۵/۷
۱۹	اتانول	ان-تترا دکان	۰/۱۵	۰/۸۵	۱۷/۶	۲۵/۲	۲۲/۴	۴۲/۹	۲۷/۲
۲۰	ایزو-آمیل الکل	ایزو-آمیل استات	۰/۵۰	۰/۵۰	۳۸/۷	۵۰/۵	۴۳/۲	۳۰/۴	۱۱/۵
خارج از دامنه کاربردی									
۱	استون	ان-دکان	۰/۴	۰/۶	-۱۷/۵	-۴/۹	۲۱/۵	-۷۱/۹	-۲۲۲/۸
۲	استون	ان-دکان	۰/۳	۰/۷	-۱۷/۵	۰/۵	۹۲/۹	-۱۰۲/۷	-۶۳۰/۶
۳	اتانول	ان-تترا دکان	۰/۲	۰/۸	۱۴/۲	۲۵/۵	-۱۲/۶	۷۹/۳	-۱۸۸/۷
۴	ایزو-آمیل الکل	ایزو-آمیل استات	۰/۴	۰/۶	۳۸/۶	۴۸/۹	۲۱/۲	۲۶/۸	-۴۵/۱
۵	متانول	ایزو-اکتان	۰/۹۳	۰/۰۷	-۱۱/۹	۲/۹	۱۱/۸	-۱۲۴/۰	-۱۹۹/۴
۶	متانول	ایزو-اکتان	۰/۹	۰/۱	-۱۱/۶	۲/۶	۶/۹	-۱۲۲/۵	-۱۵۹/۵
۷	متانول	ایزو-اکتان	۰/۸	۰/۲	-۱۱/۷	۱/۸	۴۰/۱	-۱۱۵/۲	-۴۴۲/۴
۸	متانول	ایزو-اکتان	۰/۷	۰/۳	-۱۱/۶	۰/۹	۲۶/۹	-۱۰۷/۹	-۳۳۲/۱
۹	متانول	ایزو-اکتان	۰/۶	۰/۴	-۱۱/۷	۰/۰	۱۸/۹	-۱۰۰/۳	-۲۶۱/۲
۱۰	متانول	ایزو-اکتان	۰/۴	۰/۶	-۱۱/۶	-۱/۸	-۷/۷	-۸۴/۶	-۳۳/۹
۱۱	متانول	ایزو-اکتان	۰/۳	۰/۷	-۱۱/۷	-۲/۷	۴/۷	-۷۶/۶	-۱۴۰/۴

عبارت دیگر، این توصیف‌کننده بیانگر چگالی بار مثبت روی اتم‌های مولکول سازنده مخلوط است. افزایش این توصیف‌کننده در یک مولکول یا مولکول‌های تشکیل‌دهنده یک مخلوط نشان دهنده حضور نواحی با بار مثبت قوی‌تر است، مانند زمانی که در مولکول گروه‌های الکترون‌کشنده بیشتر و قوی‌تری وجود دارند.

نسبی آن توصیف‌کننده در مدل و میزان تاثیر آن بر نقطه اشتعال‌پذیری مخلوط‌ها می‌باشد. همچنین علامت ضرایب، نشان دهنده مثبت و یا منفی بودن این اثر است. تعریف توصیف کننده‌های موثر در معادله (۵) در جدول ۳ آمده است. $\text{PEOE}_{\text{PC}+}$ شاخص بار مثبت جزئی را نشان می‌دهد. به

داشتن ضریب مثبت بزرگ (۲۰/۰۷) نشان دهنده این حقیقت است که هرچه مقدار این توصیف کننده بیشتر باشد، نقطه اشتعال بالاتر می رود یا به عبارت دیگر مخلوط هایی با بار مثبت جزئی بیشتر، پایداری حرارتی بیشتری دارند و دیرتر مشتعل می شوند. به بیان دیگر، داشتن بارهای مثبت جزئی بیشتر برهم کنش بین مولکول های اجزای مخلوط را افزایش می دهد. vsurf_D8 یکی از توصیف کننده های سطحی است که به قطبیت و توزیع بار سطحی مولکول مربوط است. مقدار بالای این توصیف کننده نشان دهنده سطح قطبی بیشتر و توزیع بار سطحی بالاتر است که سبب افزایش میزان برهم کنش میان مولکول ها شود. ضریب مثبت (۱۵/۸۷) این توصیف کننده، سبب افزایش میزان برهم کنش ها میان مولکول ها، فراریت کمتر و در نتیجه باعث می شود مولکول ها، فراریت کمتر و در نتیجه باعث می شود که مخلوط دیرتر مشتعل گردد. Jhetm از نوع بالابان^۱ یک توصیف گر توپولوژیکی است که از گراف مولکولی مشتق شده و به طور خاص وزن های جرمی اتم ها را در نظر می گیرد. این شاخص، نسخه توسعه یافته ای از نمایه کلاسیک بالابان است که در آن جرم نسبی اتم های موجود در مولکول (با در نظر گرفتن کربن به عنوان استاندارد مرجع) لحاظ می شود. این توصیف گر، شاخص هایی مانند انشعاب مولکولی، حلقوی بودن و توزیع جرم اتمی را در ساختار مولکول رمزگذاری می کند. مقدار بالاتر Jhetm یعنی ساختار مولکول پیچیده تر و متنوع تر است مثلاً به معنای شاخه دار یا دارای حلقه های متنوع در مولکول است. ضریب منفی (۷/۰۵-) نشان می دهد که افزایش مقدار این توصیف کننده باعث کاهش مقدار FP می شود. MAXDP بیشینه اختلاف پتانسیل دو نقطه از مولکول را نشان داده و شاخصی از قطبیت کلی مولکول است. ضریب مثبت آن نشان می دهد که افزایش قطبیت سبب افزایش برهم کنش های بین مولکولی و در نتیجه افزایش FP مخلوط می شود. EEig08d یک توصیف کننده ساختاری-الکترونیکی است که از ریاضیات گراف مولکولی به دست می آید. در محاسبه این توصیف کننده، ابتدا برای هر مولکول یک ماتریس مجاورت^۲ ایجاد می شود و سپس این ماتریس براساس مقادیر ممان دوقطبی هر پیوند وزن گذاری می شود. مقدار ویژه هشتم این

ماتریس به عنوان EEig08d گزارش می شود. بنابراین می توان گفت که این توصیف کننده تا حدودی ساختار مولکولی از نظر شاخه دار بودن و تا حدودی میزان ممان دوقطبی مولکول را نشان می دهد. در اینجا، ضریب منفی این توصیف کننده نشان می دهد که افزایش این توصیف کننده، به معنی افزایش شاخه دار شدن و لذا کاهش میزان برهم کنش و کاهش FP خواهد بود. H6m یکی از توصیف کننده های GETAWAY است که از عناصر ماتریس لوریج^۳ استخراج می شود که با استفاده از مختصات اتمی متمرکز شده محاسبه می گردد. مقدار بالای H6m نشان دهنده ساختار پیچیده تر و متصل تر است. هرچه مقدار این توصیف کننده بیشتر باشد، FP کاهش می یابد.

۴- نتیجه گیری

نکات در این پژوهش، با هدف پیش بینی دقیق نقاط اشتعال پذیری مخلوط های دوتایی، مدل های QSPR قابل اعتماد ارائه و بررسی شدند. برای این منظور، یک مجموعه داده بزرگ حاوی ۴۰۰ مخلوط دوتایی متشکل از ۳۳ ترکیب خالص استفاده شد. با بهره گیری از فرمول های ریاضی گوناگون و ویژگی های مولکولی مولکول های تشکیل دهنده هر مخلوط، ۲۲ نوع توصیف کننده برای مخلوط ها محاسبه گردید. به کمک الگوریتم کولونی مورچگان حافظه دار، چندین مدل QSPR با اندازه های مختلف برای هر نوع توصیف کننده مخلوط توسعه یافت. مدل ایجاد شده بر اساس توصیف کننده sqr-fmol-sum عملکرد بهتری نسبت به سایر مدل ها داشت و از کیفیت آماری خوبی برخوردار بود. یک مدل گروهی نیز با استفاده از راهکار میانگین گیری ساده و با در نظر گرفتن بهترین مدل های به دست آمده بر اساس بهترین توصیف کننده ها ایجاد شد. استفاده از یک مدل گروهی، صحت پیش بینی را به طور قابل توجهی بهبود بخشید. همچنین تحلیل مدل QSPR بدست آمده بر اساس sqr-fmol-sum نشان داد که تمام توصیف کننده هایی که در مدل های QSPR ظاهر شدند را می توان با خواص مولکولی مربوط دانست.

مراجع

^۱Balaban

^۲Adjacency matrix

^۳Leverage matrix

14. B. Nazari, M. H. Keshavarz, A. Hassanzadeh. *Process Saf. Environ. Prot.* 132, 134 (2019).
15. F. A. Carroll, C. Y. Lin, F. H. Quina. *Industrial & Engineering Chemistry Research* 50, 4796 (2011).
16. X. Liu, Z. Liu. *Journal of Chemical & Engineering Data* 55, 2943 (2010).
17. G. Fayet, P. Rotureau, V. Prana, C. Adamo. In 14. International Symposium on Loss Prevention and Safety Promotion in the Process Industry, 31, 925 (2013).
18. L. Jiao, X. Zhang, Y. Qin, X. Wang, H. Li. *Chemometr. Intell. Lab. Syst.* 156, 211 (2016).
19. H. J. Liaw, Y. Y. Chiu. *J. Hazard. Mater.* 101, 83 (2003).
20. Z. Faramarzi, F. Abbasitabar, V. Zare-Shahabadi, H. J. Jahromi. *J. Mol. Liq.* 296, 111854 (2019).
21. W. A. Affens, G. W. McLaren. *J. Chem. Eng. Data* 17, 482 (1972).
22. T. Gaudin, P. Rotureau, G. Fayet. *Fire Saf. J.* 74, 61 (2015).
23. M. Vidal, W. Rogers, M. Mannan. *Process Saf. Environ. Prot.* 84, 1 (2006).
24. T. Gaudin, P. Rotureau, G. Fayet. *Industrial & Engineering Chemistry Research* 54, 6596 (2015).
25. V. P. Solov'ev, I. Oprisiu, G. Marcou, A. Varnek. *Industrial & Engineering Chemistry Research* 50, 14162 (2011).
26. D. Wu, R. Finkelman. In *Abstracts of Papers of the American Chemical Society*, 175, 12 (1978).
1. L. O. Chemicals. Globally harmonized system of classification and labelling of chemicals (GHS); 10th revised ed (United Nations, 2023).
2. D. A. Crowl, J. F. Louvar. *Chemical process safety: fundamentals with applications*; 4th ed (Pearson Education, 2019).
3. J. Shah, M. R. Jan, A. U. Haq. *Tenside, Surfactants, Detergents* 51, 240 (2014).
4. M. J. O'Neil, P. E. Heckleman, C. B. Koch, K. J. Roman. White House Station, NJ, USA 945, 160 (2006).
5. H. J. Liaw, S. C. Lin. *J. Hazard. Mater.* 140, 155 (2007).
6. H. J. Liaw, T. P. Lee, J. S. Tsai, W. H. Hsiao, M. H. Chen, T. T. Hsu. *J. Loss Prev. Process Ind.* 16, 173 (2003).
7. X. Y. Guo, S. Liang, Q. H. Tian. Removal of heavy metal ions from aqueous solutions by adsorption using modified orange peel as adsorbent. In *Advanced Materials Research*, 2011; Vol. 236-238, pp 237.
8. E. Torabian, M. A. Sobati. *Thermochim. Acta* 672, 162 (2019).
9. L. Catoire, S. Paulmier, V. Naudet. *Process Saf. Prog.* 25, 33 (2006).
10. Y. Pan, J. Cheng, X. Song, G. Li, L. Ding, J. Jiang. *J. Loss Prev. Process Ind.* 34, 56 (2015).
11. H. J. Liaw, T. P. Tsai. *Fluid Phase Equilib.* 345, 45 (2013).
12. H. J. Liaw, H. Y. Chen. *Industrial & Engineering Chemistry Research* 52, 7579 (2013).
13. H. J. Liaw, C. L. Tang, J. S. Lai. *Comb. Flame.* 138, 308 (2004).

42. S. Khorsand Ahmadi, M. Mahmoodian Moghadam, P. Mokaberi, M. Reza Saberi, J. Chamani. 33, 1880 (2015).
43. M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond. Anal. Chim. Acta 646, 39 (2009).
44. F. Abbasitabar, V. Zare-Shahabadi. SAR QSAR Environ. Res. 23, 1 (2012).
45. V. Zare-shahabadi, F. Abbasitabar. J. Comput. Chem. 31, 2354 (2010).
46. M. Jalali-Heravi, H. Ebrahimi-Najafabadi, A. Khodabandehloo. QSAR Comb. Sci. 28, 1432 (2009).
47. F. Abbasitabar, V. Zare-Shahabadi. Chemosphere 172, 249 (2017).
48. V. Zare-Shahabadi. Med. Chem. Res. 25, 2787 (2016).
49. D. Baumann, K. Baumann. J. Cheminform. 6, 47 (2014).
50. D. Massart, B. Vandeginste, L. Buydens, S. de Jong. Handb. Chemom. Qualimetr., 379 (1997).
51. K. Roy, S. Kar, P. J. C. Ambure, I. L. Systems. 145, 22 (2015).
52. V. Zare-Shahabadi, F. Abbasitabar, M. Akhond, M. Shamsipour. J. Braz. Chem. Soc. 24, 1561 (2013).
27. H. J. Liaw, W. H. Lu, V. Gerbaud, C. C. Chen. J. Hazard. Mater. 153, 1165 (2008).
28. H. J. Liaw, V. Gerbaud, Y. H. Li. Fluid Phase Equilib. 300, 70 (2011).
29. H. J. Liaw, V. Gerbaud, H. T. Wu. J. Chem. Eng. Data 55, 3451 (2010).
30. H. J. Liaw, V. Gerbaud, C. Y. Chiu. J. Chem. Eng. Data 55, 134 (2010).
31. M. Mitu, E. Brandes, D. Pawel. Rev. Chim. 57, 770 (2006).
32. M. Noorollahy, A. Z. Moghadam, A. A. Ghasrodashti. Chem. Eng. Res. Des. 88, 81 (2010).
33. D.-M. Ha, S. Lee. Korean J. Chem. Eng. 28, 1161 (2011).
34. A. Z. Moghaddam, A. Rafiei, T. Khalili. Fluid Phase Equilib. 316, 117 (2012).
35. A. Donaldson, S. Al-Sharafi. Eng. J. Qatar Univ. 7, 27 (1994).
36. M. Hristova. Cent. Eur. J. Chem. 11, 57 (2013).
37. M. Hristova, D. Damgaliev. Cent. Eur. J. Chem. 11, 388 (2013).
38. Ultra, ChemDraw; Cambridge soft corporation, Cambridge, USA: 2001.
39. Koala-Software for kohonen artificial neural networks; Milano chemometrics and QSAR research group: 2007.
40. E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min. Mol. Inform. 31, 202 (2012).
41. T. Gaudin, P. Rotureau, G. Fayet. Ind. Eng. Chem. Res. 54, 6596 (2015).



New structure-based models using mixture descriptors to predict the flammability temperature of binary organic mixtures

Zohreh Faramarzi¹, Fatemeh Abbasitabar^{2,*}, Vahid Zare-Shahabadi¹

¹ Institute of Mining, Oil and Energy, Mahs.C., Islamic Azad University, Mahshahr, Iran.

² Department of Chemistry, Marv. C., Islamic Azad University, Marvdasht, Iran.

Abstract: Due to the importance of binary mixtures and their wide application in industry, it is important to estimate their flammability temperature. Therefore, in this study, quantitative structure-property relationship(QSPR) models were developed to predict the flammability point of binary organic mixtures. The biggest challenge in QSPR studies of mixtures is the calculation of a numerical descriptor to describe the property of a mixture. In this study, a set of twenty-two formulas was used to calculate mixture descriptors from molecular descriptors of pure compounds. Memorized Ant Colony(Memorized-ACO) Algorithm along with multivariate linear regression(MLR) was used to select the best subset of descriptors that have a significant contribution to the flammability property. The model based on sqr-fmol-sum descriptor was selected as the best model. R^2_{training} and R^2_{test} of this model were 0.93 and 0.94, respectively. The mean absolute error(MAE) for the training and test datasets was 3.75 and 2.97, respectively. A group model was also obtained through a simple averaging strategy using the models generated by the best descriptors. This improved the statistical quality of the final QSPR model.

Keywords: QSPR, Flammability point, Memorized-Ant Colony optimization, Mixture descriptors, Multiple linear regression.