به کارگیری مدل‌های یادگیری ماشین برای پیش‌بینی تشکیل امولسیون اسید و نفت در تست‌های استاتیک اسیدکاری با استفاده از بانک اطلاعات ترکیبی

۱- نویسنده مقاله: سپیده عطربرمحمدی

۲- نویسنده مقاله: حسین خیرالهی

۳- نویسنده مقاله: سید شهاب الدین آیت اللهی* (نویسنده عهده دار مقاله، shahab@sharif.edu)

۴- نویسنده مقاله: سید محمودرضا پیشوائی* (نویسنده عهده دار مقاله، pishvaie@sharif.edu)

دانشکده مهندسی شیمی و نفت، دانشگاه صنعتی شریف، تهران ایران

Department of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran

چکیده

در طول مدت بهره‌برداری از مخازن نفتی، معمولاً نواحی نزدیک به دیواره چاه به دلایل مختلفی همچون جابجایی ذرات سازندی، تورم رس و موارد دیگر در معرض آسیب‌های مختلف قرار گرفته و نرخ تولید/تزریق از آن‌ها به شدت کاهش می‌یابد. یکی از پرکاربردترین روش‌های انگیزش چاه برای رفع این آسیب‌های سازندی روش اسیدکاری است که در طی آن اسید و مواد شیمیایی (افزایه‌ها) به داخل سازند تزریق می‌شوند تا با انحلال سنگ در سازندهای کربناته تراوایی سازند را افزایش دهند. با این وجود، عدم بررسی آزمایشگاهی سازگاری سیالات تزریقی با سیالات سازندی در مرحله طراحی می‌تواند منجر به ایجاد آسیب‌های القائی همچون تشکیل امولسیون اسید در نفت شود. تست‌های آزمایشگاهی که به منظور بررسی سازگاری سیالات مذکور و انتخاب سیالات تزریقی مناسب صورت می‌گیرند، زمان‌بر، پرهزینه و نیز از لحاظ ایمنی خطرناک می‌باشند. به همین منظور در این پژوهش سعی شده‌است تا نتایج اولیه تست‌های امولسیونی با استفاده از مدل‌های مبتنی بر داده و در زمان کوتاه‌تر پیش‌بینی شوند. بنابراین داده‌های مؤثر بر نتایج این آزمایش‌ها، شامل متغیرهای نوع و غلظت اسید، افزایه‌های ضد امولسیون، ضد لخته، کاهنده کشش سطحی، کاهنده یون آهن و همچنین ویژگی‌های ۱۳ نوع نفت از مخازن مختلف مانند گرانروی، چگالی و غلظت یون فریک، جمع‌آوری و به عنوان ورودی‌های یک مجموعه داده ثبت شدند. سپس مدل‌های طبقه‌بندی نظارت‌شده شامل الگوریتم جنگل تصادفی، ماشین بردار پشتیبان، پرسپترون چندلایه و تقویت گرادیان شدید جهت پیش‌بینی خروجی تست‌های ضد امولسیون بر روی مجموعه داده جمع‌آوری‌شده اعمال شدند. با توجه به کمبود داده‌های آزمایشگاهی از تکنیک آماری بیش نمونه‌گیری مصنوعی به منظور تولید نمونه داده‌ی مصنوعی و و بهبود عملکرد مدل‌های هوش مصنوعی استفاده گردید. براساس نتایج بدست آمده، روش تقویت گرادیان شدید با پنج تخمین‌گر به ترتیب با مقادیر کوهن-کاپای ۰/۷۹ و ۰/۵۲۳ برای مجموعه داده‌های آموزش و تست بهترین عملکرد را داشته است.

کلیدواژه‌ها: اسیدکاری، امولسیون نفت و اسید، تست‌های استاتیک اسیدکاری، یادگیری ماشین، طبقه‌بندی، بیش نمونه‌گیری مصنوعی.

۱- مقدمه

در طول مراحل مختلف از عمر یک مخزن خصوصاً در حین اجرای عملیات‌های ازدیاد برداشت، ممکن است سازند اطراف چاه دچار آسیب‌هایی همچون تورم رس، جابجایی و مهاجرت ذرات ریز و نیز رسوب مواد معدنی در حفرات شود. این آسیب‌ها می‌توانند نرخ تولید/تزریق پذیری و به صورت کلی بازدهی مخزن را کاهش دهند. اسیدکاری یکی از مهم‌ترین و پرکاربردترین روش‌های انگیزش چاه است که در صنعت نفت و گاز برای برطرف کردن آسیب‌های سازندی و بالا بردن بهره‌وری مخازن نفتی اعمال می‌گردد [1]. در عملیات اسیدکاری سیالی که شامل اسید است به سازند اطراف چاه تزریق می‌شود تا با انحلال سنگ‌های کربناته و یا انحلال مواد معدنی رسوب کرده در حفرات سازندهای ماسه‌سنگی مسیر عبور سیالات را باز کند. با این کار تراوایی سازند به مقدار اولیه‌ی خود بازیابی شده و یا به میزان بیشتر از مقدار اولیه افزایش می‌یابد و بدین ترتیب نفت تولیدی و یا آب تزریقی آسان‌تر و بیشتر جریان می‌یابد. با توجه به نوع سنگ‌های سازندی معمولاً اسید مورد استفاده در این عملیات هیدروکلریک اسید برای سازندهای کربناته و اسید گِل (مخلوط هیدروکلریک اسید و هیدروفلوئوریک اسید) برای سازندهای ماسه‌سنگی می‌باشد. در همان سال‌های ابتدایی اعمال این روش‌ها و نیز با گذشت زمان و اجرای عملیات‌های اسیدکاری مختلف بر روی میادین نفتی در سراسر جهان، پژوهشگران نفتی متوجه شدند که در صورتیکه عملیات اسیدکاری به درستی طراحی و اجرا نگردد میتواند منجر به ایجاد مشکلات جدیدی همچون خوردگی لوله‌های چاه، تورم رس، تشکیل لجن اسیدی (اسلاج)، تشکیل امولسیون اسید – نفت و غیره در سازند شود. این مشکلات می‌توانند منجر به کاهش تراوایی مخزن و در مواردی منجر به بسته شدن تمامی مسیرهای منتهی به چاه گردند و بدین ترتیب چاه بطور کامل بسته شده و حجم عظیمی از منابع نفتی از دست می‌رود. به همین دلیل مواد شیمیایی تحت عنوان افزایه‌های اسیدکاری برای جلوگیری از بوجود آمدن چنین مشکلات ثانویه ساخته شدند و به همراه اسید به مخزن تزریق می‌گردند [2].

یکی از مشکلات عمده در عملیات اسیدکاری تشکیل امولسیون اسید – نفت است که به دلیل پایه آبی بودن اسید و پایه روغنی بودن نفت تشکیل می‌شود [3]. پس از تزریق اسید به مخزن، نفت و اسید در تماس با یکدیگر قرار گرفته و منجر به تشکیل امولسیون اسید در نفت می‌شوند. با تشکیل این امولسیون، سیال اسید توسط نفت محصور شده و نمی‌تواند با سطح سنگ در تماس قرار گرفته و سنگ را در خود حل کند. این پدیده باعث کاهش بازدهی عملیات اسیدکاری می‌گردد. همچنین امولسیون اسید در نفت موجب افزایش ویسکوزیته‌ی سیال (حدود چهار برابر) می‌شود و بدین ترتیب حرکت نفت در داخل سازند را دشوار و بهره‌دهی چاه را بشدت کاهش می‌دهد [3]. در برخی موارد امولسیون به حدی است که بعد از اسیدکاری امکان بهره‌برداری از چاه وجود ندارد. تحقیقات قبلی نشان داده است که تشکیل امولسیون در نفت بدلیل حضور ترکیبات قطبی، رزین‌ها و ترکیبات هتروسیکلیک همچون اسیدها، بازها، فنلیک‌ها، آسفالتین‌ها و ترکیبات پیچیده با وزن مولکولی بالا بوده که همانند یک سورفکتانت عمل می‌کنند. ترکیبات فوق الذکر، قطرات آب را (با اندازه ۱ تا ۲۰ میکرومتر) در فاز نفت به دام می‌اندازند. ترکیبات آروماتیکی فرار، آروماتیک‌های مونو و آروماتیک‌های چندحلقه‌ای در نفت خام همچون بنزن و اتیل بنزن، این آسفالتین‌ها و رزین‌ها را حل می‌کنند؛ بنابراین نفت خامی که مقدار زیادی از این ترکیبات فرار را دارد، تمایل کمتری به تشکیل امولسیون هنگام تماس با آب دارد. افزون بر این، تحقیقات نشان داده است که پارامترهای مختلفی همچون نوع و غلظت اسید، دما، مدت زمان مجاورت اسید و نفت، نسبت حجمی اسید به نفت، غلظت یون آهن و نیز پارامترهای نفت همچون ویسکوزیته، ترکیب سارا و چگالی نفت در میزان امولسیون تشکیل شده در سازند مؤثر می‌باشند [3]–[5].

در صنعت نفت و گاز برای جلوگیری از بوجود آمدن امولسیون اسید-نفت در طی عملیات میدانی اسیدکاری از افزایه‌ی ضد امولسیون استفاده می‌شود. این افزایه که شامل سورفکتانت با ساختار انشعابی و نیز حلال‌های دوگانه است از ایجاد امولسیون اسید در نفت جلوگیری می‌کند. غلظت مناسب این افزایه جهت تزریق به مخزن بستگی به ویژگی‌های مختلف نفت و نوع و غلظت اسید تزریقی در عملیات اسیدکاری دارد و تعیین آن قبل از اجرای عملیات میدانی توسط تست‌های آزمایشگاهی صورت می‌پذیرد. این تست‌های آزمایشگاهی که برای تعیین غلظت مناسب افزایه‌های تزریقی انجام می‌شوند به تست‌های استاتیک اسیدزنی معروف هستند و انجام هر یک از آن‌ها هزینه‌بر، زمان‌بر و برای پرسنل آزمایشگاه خطرناک است [7]–[5], [3]. همچنین تعداد تست‌های که باید برای تعیین تکلیف سازگاری سیالات در آزمایشگاه اسیدزنی انجام شوند بسیار زیاد خواهد بود، زیرا شرایط بهینه برای هر نمونه نفتی، نوع و غلظت مناسب اسید در حضور افزایه‌های مختلف همچون افزایه‌ی ضد امولسیون تعیین تکلیف گردد. شرط لازم در دستورالعمل‌های عملیاتی بر این تاکید دارد که در پایان انجام تست، امولسیون اسید-نفت پس از گذشت ۳۰ دقیقه از زمان تشکیل آن باید از بین برود [4]. این تست‌ها در آزمایشگاه‌های شرکت‌های بزرگ سرویس[1] در سراسر جهان از سالیان گذشته انجام شده‌اند و مجموعه داده‌ی آن‌ها در همان شرکت‌ها ثبت و نگهداری می‌شوند.

با پیشرفت حوزه‌ی کامپیوتر و ظهور علم هوش مصنوعی، مهندسان نفت از این تکنولوژی برای سرعت بخشدن به تصمیم‌گیری‌ها، داشتن پیش‌بینی‌های اولیه، کاهش زمان و نیز هزینه عملیات در زمینه‌های مختلف مهندسی نفت همچون اکتشاف، بهره‌برداری، ازدیاد برداشت و ترک چاه استفاده کرده‌اند [14]–[8]. بطور مثال، دانشمندان در دهه‌ی ۹۰ میلادی از مجموعه داده‌های آزمایشگاهی و عملیات‌های میدانی برای ساخت و آموزش سیستم‌های خبره جهت بهینه‌سازی و طراحی دقیق و صحیح عملیات اسیدکاری برای شرکت‌های مختلف استفاده کردند. این سیستم‌ها نوع و غلظت سیالات تزریقی از جمله اسید و انواع مختلف افزایه‌ها را با توجه به ویژگی‌های سازند و سیالات درون مخزن ارائه می‌دادند. در همین راستا مهندسان بسیاری سیستم‌های خبره و نیز شبکه‌های عصبی مصنوعی مختلفی را در دوره‌های متفاوت جهت بهینه سازی عملیات اسیدکاری و پیش‌بینی میزان امولسیون تشکیل‌شده طراحی و ساخته‌اند [21]–[15], [8], [26]–[22], [7], [5].

افزون بر استفاده از مدل‌های هوش مصنوعی برای طراحی کل عملیات اسیدکاری، از این روش‌ها برای پیش‌بینی نتایج تست‌های استاتیک اسیدزنی استفاده شده‌است. بطور مثال پوراکابریان و همکارانش در سال ۲۰۲۱ توانستند میزان لخته‌ی تشکیل‌شده در تست‌های آزمایشگاهی را بدون حضور افزایه‌ها و با استفاده از یک شبکه عصبی مصنوعی با سه لایه‌ی پنهان و هفت نرون پیش‌بینی کنند [5]. همچنین شکوری‌زاده و همکارانش در سال ۲۰۲۳ برای بررسی تأثیر افزایه‌ها از مدل‌های مختلف رگرسیونی برای پیش‌بینی میزان لخته‌ی تشکیل‌شده در تست‌های استاتیک ضد لجن اسیدی استفاده کردند [6]. با این وجود به علت پیچیده بودن این فرایند و در دسترس نبودن تعداد کافی از اطلاعات آزمایشگاهی، تا کنون مدل قابل اعتمادی در زمینه پیش‌بینی تشکیل امولسیون اسید در نفت در تست‌های آزمایشگاهی استاتیک اسیدکاری ارائه نشده‌است.

در این پژوهش سعی شده‌است که از مدل‌های مختلف یادگیری ماشین همچون شبکه‌های عصبی و مدل‌های طبقه‌بندی مختلف برای پیش‌بینی نتایج تست‌های استاتیک ضد امولسیون هیدروکلریک اسید و نفت در حضور افزایه‌ها استفاده شود و بدین ترتیب هزینه، زمان و خطرات ناشی از کار با اسید در آزمایشگاه اسیدزنی به طور قابل ملاحظه کاهش یابد. اعمال این روش‌ها بر روی داده‌های آزمایشگاهی منجر به کاهش تعداد تست‌های لازم برای اجرا شده و هزینه و طول مدت کارهای آزمایشگاهی کاهش

[1] Service Company

می‌یابد.

۲- تست‌های استاتیک بررسی افزایه‌ی ضد امولسیون

آزمایش استاتیک ضد امولسیون در آزمایشگاه اسیدزنی برای بررسی امکان جدایش فازهای اسید و نفت پس از تشکیل امولسیون اسید در نفت حین اجرای عملیات اسیدکاری انجام می‌شود. در این آزمایش در یک ظرف مخصوص (با سر آبی[1]) ۷۵ میلی‌لیتر اسید با غلظت معین به همراه یون آهن فریک ریخته شده و افزایه‌های مشخص شده به محلول اضافه می‌گردد (بعد از اضافه شدن هر افزایه بطری هم زده شده تا افزایه در سیستم به خوبی پخش شود). سپس ۷۵ میلی‌لیتر نفت مورد نظر به ظرفی که از قبل مقدار ۰/۳۷۵ گرم بنتونایت و ۳/۳۷۵ گرم میکروسیلیس در آن ریخته شده بود، اضافه می‌گردد. پس از آن محتویات ظرف حاوی سیستم اسید و ظرف حاوی نفت و جامدات به ظرف همزن همیلتون بیچ[2] ریخته می‌شود و محتویات ظرف به مدت ۳۰ ثانیه با سرعت ۱۸۰۰۰ دور بر دقیقه به خوبی مخلوط می‌شوند. در نهایت محتویات ظرف به استوانه مدرج منتقل شده و این استوانه در حمام آب در دمای ۹۴ درجه سانتی‌گراد قرار می‌گیرد و در بازه‌های زمانی ۱۰ و ۱۵ و ۳۰ دقیقه، ۱ ساعت و نیز ۲ ساعت پس از شروع تست، میزان جدایش فاز اسید در آن به میلی‌لیتر ثبت و گزارش می‌شود. داده آزمایش‌های مورد استفاده در این پژوهش، در حضور غلظت‌های ۱۰۰۰ ppm و ۳۰۰۰ ppm یون آهن فریک و نیز هیدروکلریک اسیدهای ۱۵٪ و ۲۸٪ جرمی انجام شده‌اند و نسبت اسید به نفت در تمامی آزمایش‌ها ۱:۱ بوده است. میزان افزایه‌ها بر اساس دستورالعمل مشخص شده از طرف سازنده این افزایه‌ها مورد استفاده قرار گرفت.

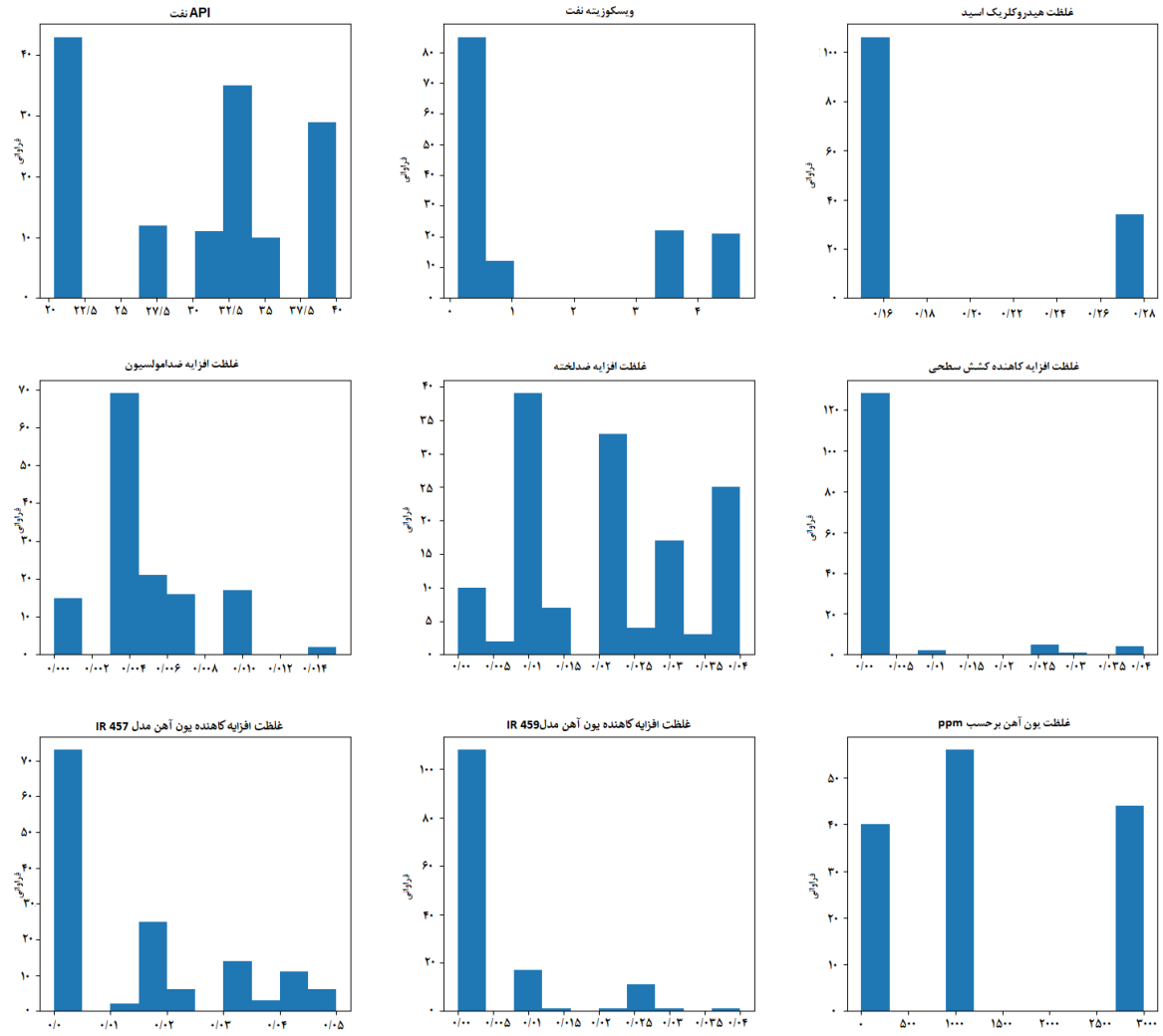
۳- جمع‌آوری داده‌ها

مجموعه داده‌ی گردآوری شده در این پژوهش، اطلاعات مربوط به اجرای ۱۴۰ تست استاتیک اسیدزنی با استفاده از هیدروکلریک اسید با غلظت‌های مختلف بر روی ۱۳ نوع نفت از میادین جنوب غربی ایران را شامل می‌شود. در این مجموعه داده، اطلاعات مربوط به نفت، اسید و افزایه‌های مورد استفاده در تست‌ها گزارش شده‌است. در میان اطلاعات ارائه شده، غلظت افزایه‌هایی همچون افزایه‌های ضد خوردگی، از بین برنده‌ی هیدروژن سولفید، کنترل‌کننده‌ی یون آهن، معلق نگهدارنده در تمامی تست‌های اجرا شده، ثابت و غلظت افزایه‌هایی همچون ضد امولسیون، ضد لخته، کاهنده‌ی کشش سطحی (سورفکتانت)، و نیز کاهنده‌ی یون آهن متغیر گزارش شده‌اند. همچنین خروجی‌های مربوط به تست‌های ضد امولسیون نیز به صورت حجم اسید جدا شده از نفت در زمان‌های مختلف گزارش شده‌است.

بر اساس تجربیات موجود، تعداد ویژگی‌ها و پارامترهای ورودی مدل‌های یادگیری ماشین در بار محاسباتی و نیز دقت مدل‌های آموزش دیده بسیار مؤثر است [27]. بنابراین بهینه‌سازی تعداد پارامترها برای کاهش بار محاسباتی از یکطرف و توانایی مدل برای پیش‌بینی خروجی‌ها از اهمیت ویژه ای برخوردار می‌باشد. در این راستا، پارامترهایی که در تمامی تست‌ها مقادیر ثابت و یکسانی داشته‌اند به این دلیل که تأثیری در پیش‌بینی خروجی مدل ندارند از مجموعه داده کنار گذاشته شدند و تنها غلظت افزایه‌های

[1] Blue Cap
[2] Hamilton Beach Mixer

ضد امولسیون، ضد لخته و کاهنده‌ی کشش سطحی (سورفکتانت)، غلظت هیدروکلریک اسید، ویسکوزیته و API نفت مورد استفاده و نیز غلظت یون آهن فریک بعنوان پارامترهایی که در نتیجه‌ی تست‌های ضد امولسیون تأثیر دارند بعنوان ورودی‌های مجموعه داده جمع‌آوری و ثبت گردیدند. شکل ۱ تعداد تست‌های انجام شده (فراوانی) با هر یک از مقادیر مختلف این ویژگی‌ها را بصورت نمودار هیستوگرامی نشان میدهد.

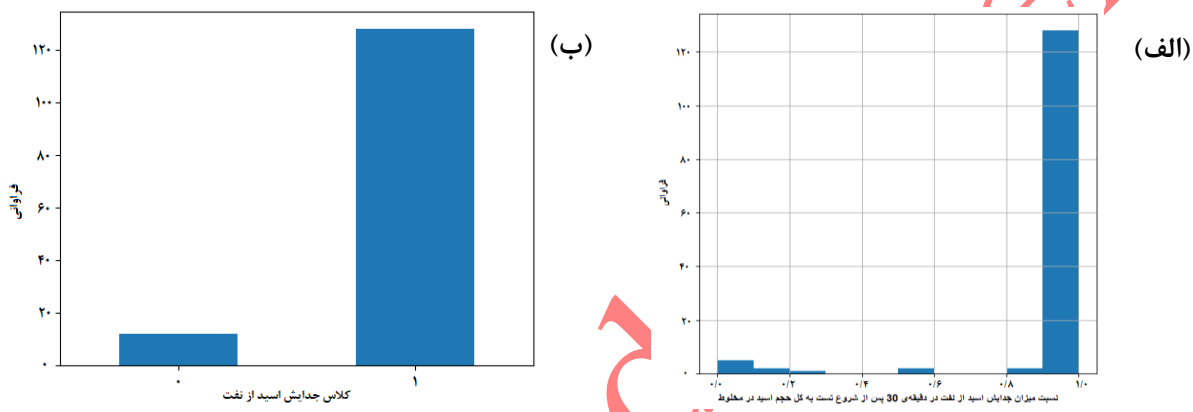


شکل ۱. توزیع داده‌های ورودی مجموعه داده‌ی تست‌های استاتیک ضد امولسیون

با وجود اینکه رسیدن به دقت بالا در بکارگیری روش‌ها و مدل‌های یادگیری ماشین مستلزم آموزش آن‌ها با طیف وسیعی از داده‌ها با توزیع ترجیحاً نرمال می‌باشد، هیچ یک از ویژگی‌های این مجموعه داده دارای توزیع نرمال نمی‌باشند (شکل ۱). با این حال، برای هر یک از ویژگی‌های ذکر شده تعداد مناسبی از تست‌های انجام شده در محدوده‌ی مناسب آزمایشگاهی گزارش شده‌است و ویژگی‌های این مجموعه داده شامل تمامی مقادیر عددی مورد استفاده در آزمایشگاه اسیدزنی می‌باشند.

در عملیات‌های اسیدکاری واقعی در میادین نفتی حد بحرانی جدایش اسید و نفت برای تعیین امکانپذیر بودن و یا نبودن یک عملیات اسیدکاری، جدایش حداقل ۹۰٪ اسید از نفت تا دقیقه‌ی ۳۰ پس از تشکیل امولسیون می‌باشد و در صورتیکه تستی به این آستانه برسد انجام عملیات در میدان امکان‌پذیر می‌باشد. فلذا در این پژوهش درصد جدایش اسید از نفت در مدت زمان ۳۰ دقیقه (نسبت حجم اسید جدا شده از نفت در دقیقه‌ی ۳۰ پس از شروع آزمایش به حجم کل اسید موجود در مخلوط اسید و

نفت) بعنوان خروجی تست‌ها و نیز مجموعه داده تعریف شده‌است. این مقادیر که اعدادی در بازه‌ی صفر تا یک بودند نهایتاً به دو کلاس "جدایش مناسب" (مقادیر صفر تا ۰/۹) و "جدایش نامناسب" (مقادیر ۰/۹ تا یک) تقسیم‌بندی گردید. بدین ترتیب این مسئله به یک مسئله‌ی طبقه‌بندی تبدیل گردید و نمودارهای توزیع مقادیر خروجی‌های این مجموعه داده در شکل ۲ ارائه شده‌اند. همانطور که از شکل ۲ مشخص است، توزیع نمونه‌ها در دو کلاس یکسان نیست بطوریکه کلاس "جدایش مناسب" دارای ۱۲۵ نمونه داده‌ی آزمایشگاهی و کلاس "جدایش نامناسب" دارای ۱۵ نمونه داده می‌باشد؛ به عبارت دیگر مجموعه داده‌ی در دسترس یک مجموعه‌ی نامتوازن می‌باشد. این عدم توازن بعلت وجود تجربه‌ی کافی در انجام تست‌های ضد امولسیون و استفاده از افزایه‌های بهینه و مناسب توصیه شده توسط شرکت‌های سازنده می‌باشد که منجر به جدایش مناسب اسید و نفت در اکثریت تست‌ها شده‌است.



شکل ۲. توزیع خروجی‌های درصد جدایش تست‌های استاتیک ضد امولسیون (الف) و توزیع خروجی‌های مجموعه داده‌ی تست‌های استاتیک ضد امولسیون به صورت دو کلاس "جدایش نامناسب" (کلاس ۰) و "جدایش مناسب" (کلاس ۱) (ب).

۴- روش تحقیق

در این مطالعه در وهله اول فرایند جمع‌آوری، آماده‌سازی و تهیه یک مجموعه داده آزمایشگاهی مرتبط با تست استاتیک اسیدکاری شامل اطلاعات مربوط به پارامترهای ورودی و خروجی تست‌های اسیدزنی ضد امولسیون پرداخته شد. در ادامه با به‌کارگیری تکنیک‌های آماری پیش پردازش داده، این مجموعه داده جهت آموزش و ارزیابی مدل‌های مختلف طبقه‌بندی یادگیری ماشین استفاده گردید. شرح جزئیات کار و نحوه پیاده‌سازی الگوریتم‌ها آورده شده‌است.

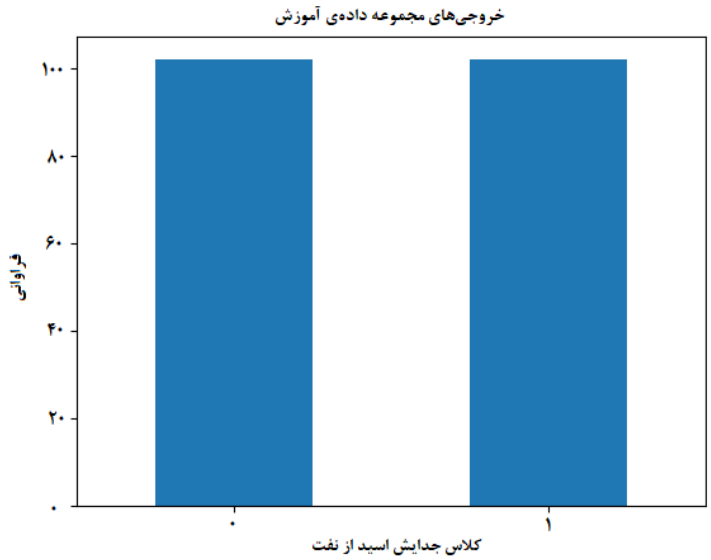
۴-۱- پیش‌پردازش داده‌ها[1]

پس از آماده‌سازی مجموعه داده بصورت یک فایل اکسل و فراخوانی آن در محیط پایتون، روش‌های مختلف پیش‌پردازش داده‌ها همچون حذف داده‌های پرت[2]، داده‌های تکراری بر روی مجموعه داده اعمال شدند. این مرحله جهت آماده‌سازی مجموعه داده جهت آموزش مدل‌های هوش مصنوعی با بالاترین میزان عملکرد انجام شد. پس از اجرای این مرحله، داده‌ها بصورت تصادفی (با امکان تکرارپذیری) به دو مجموعه‌ی آموزش و تست با نسبت ۴ به ۱ تقسیم شدند. این تقسیم‌بندی بصورتی انجام شد که

[1] Data Preprocessing
[2] Outlier

کلاس‌های خروجی در هر دو مجموعه توزیع یکسانی داشته باشند و مجموعه‌ی آموزش نمایندهٔ مناسبی از کل مجموعه داده و نیز مجموعه داده‌های تست باشد. پس از آن ویژگی‌های هر دو مجموعه داده را به کمک کمترین و بیشترین مقادیر هر یک از ویژگی‌ها نرمالسازی کرده و در بازهٔ صفر تا یک قرار می‌دهیم. این مرحله را بدین علت انجام میدهیم که برخی از مدل‌های یادگیری ماشین همچون ماشین بردار پشتیبان نسبت به مقیاس ویژگی‌ها حساس بوده و ضرایب مربوط به ویژگی‌ها را بر اساس مقیاس و بازهٔ عددی آن‌ها تعیین می‌کنند. پس از آن مدل‌های طبقه‌بندی هوش مصنوعی شامل مدل‌های جنگل تصادفی، تقویت گرادیان شدید، ماشین بردار پشتیبان و شبکه عصبی مصنوعی با استفاده از داده‌های آموزش، آموزش دیده و سپس عملکرد آنها بر اساس معیارهای مناسب بر روی مجموعه دادهٔ تست ارزیابی گردید.

با توجه به ماهیت نامتوازن مجموعه داده‌ی واقعی در دسترس و تاثیر آن بر خروجی ها، در مرحله بعد داده‌های جدیدی تهیه و به بانک اطلاعاتی اضافه گردید تا یک بانک اطلاعات ترکیبی ساخته شود. در این زمینه از یک تکنیک آماری در حوزه هوش مصنوعی به نام بیش نمونه‌گیری مصنوعی (SMOTE[1]) جهت تولید و اضافه کردن داده‌های جدید در مجموعه داده‌ی آموزش استفاده گردید. این روش از طریق درونیابی داده‌های واقعی، داده‌های مصنوعی می‌سازد و علاوه بر ایجاد توازن میان کلاس‌ها، باعث افزودن تعداد قابل توجهی نمونه داده به مجموعه داده‌ی اصلی و بهبود عملکرد مدل‌های یادگیری ماشین می‌گردد [28 , 29]. در این مسئله پس از اعمال بیش نمونه‌گیری مصنوعی بر روی مجموعه داده‌ی آموزش، تعداد داده‌ها در هر یک از کلاس‌ها در این مجموعه داده به شرح شکل ۳ شده‌است.



شکل ۳. توزیع خروجی‌های مجموعه داده‌ی آموزش در دو کلاس "جدایش نامناسب" (کلاس ۰) و "جدایش مناسب" (کلاس ۱) پس از اعمال روش بیش نمونه‌گیری مصنوعی

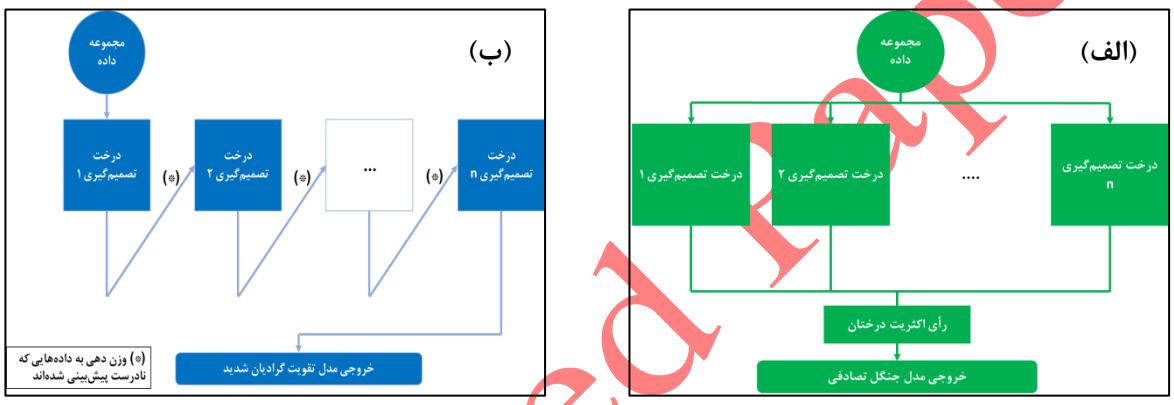
۴-۲- مدل‌های طبقه‌بندی یادگیری ماشین

با توجه به نوع خروجی‌های مجموعه داده‌ی آزمایشگاهی که بصورت دو کلاس "جدایش مناسب" و "جدایش نامناسب" می‌باشد، در این پژوهش مدل‌های طبقه‌بندی یادگیری ماشین بر روی داده‌های پیش‌پردازش شده اعمال شدند. در این راستا از چهار مدل

[1] Synthetic Minority Oversampling Technique(SMOTE)

یادگیری ماشین شامل مدل‌های جنگل تصادفی[1]، تقویت گرادیان شدید[2]، پرسپترون چند لایه[3] و ماشین بردار پشتیبان[4] که نتایج دقیق‌تری نسبت به سایر روش‌های طبقه‌بندی دارند استفاده شده‌است.

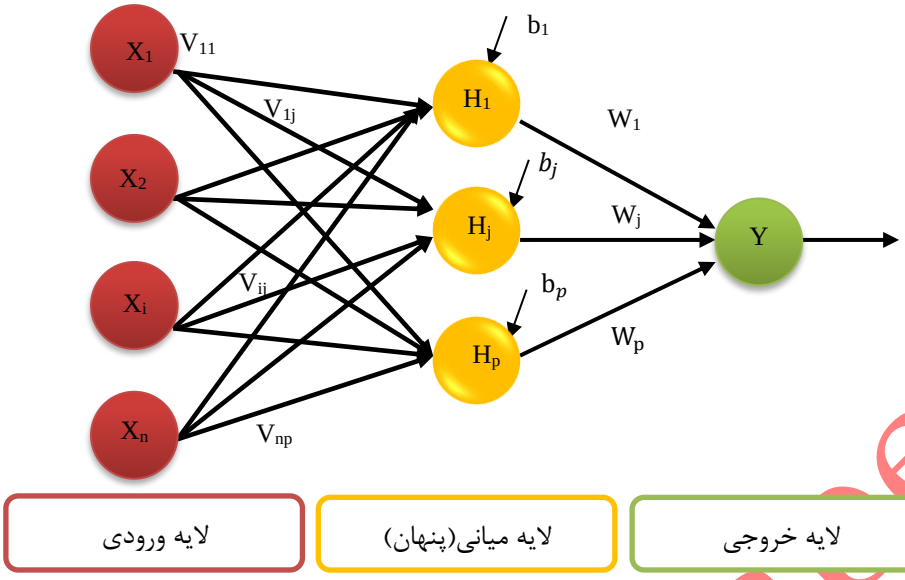
مدل‌های جنگل تصادفی و تقویت گرادیان شدید نوعی از مدل‌های یادگیری جمعی[5] هستند که برپایه‌ی درختان تصمیم‌گیری کار می‌کنند. در مدل جنگل تصادفی تعداد معینی درخت تصمیم‌گیری به صورت مجزا و به موازات یکدیگر توسط داده‌ها آموزش می‌بینند و پس از آن خروجی یک نمونه داده‌ی جدید را براساس رأی اکثریت درختان آموزش‌دیده تعیین می‌کند (**Error! Reference source not found.**، الف). همچنین در روش تقویت گرادیان شدید، تعداد معینی از درختان تصمیم‌گیری یکی پس از دیگری آموزش می‌بینند بطوریکه در این فرایند وزن اهمیت کلاسی که نمونه‌های آن توسط درخت قبلی نادرست پیش‌بینی شده بودند افزایش می‌یابد تا بدین ترتیب درخت جدیدتر قادر به پیش‌بینی خروجی مربوط به نمونه‌های این کلاس باشد (**Error! Reference source not found.**، ب).



شکل ۴. نحوه‌ی عملکرد مدل‌های یادگیری جمعی جنگل تصادفی (الف) و روش تقویت گرادیان شدید (ب)

مدل پرسپترون چند لایه ساختاری شبیه به سیستم عصبی انسان دارد بطوریکه این شبکه از لایه‌های مختلفی شامل یک لایه‌ی ورودی، یک یا چند لایه‌ی پنهان و یک لایه‌ی خروجی تشکیل میشود. هر یک از این لایه‌ها از تعدادی گره یا نرون ساخته شده‌اند که وظیفه‌ی اعمال عملیات ریاضی بر روی داده‌های ورودی مدل بر عهده دارند (شکل ۵). این مدل با یافتن مقادیر مناسب و بهینه برای ضرایب وزنی و مقادیر بایاس موجود در هر لایه، میان داده‌های ورودی و خروجی ارتباطی می‌یابد و سپس از این ضرایب برای پیش‌بینی خروجی داده‌های جدید استفاده می‌کند. مدل ماشین بردار پشتیبان نیز داده‌هایی با n ویژگی ورودی را در یک فضای n بُعدی ترسیم کرده و با تعیین خطوط مرزی میان کلاس‌ها، خروجی داده‌های جدید را پیش‌بینی کند.

[1] Random Forest Classifier
[2] Extreme Gradient Boosting (XGBoost)
[3] Multi-Layer Perceptron Classifier
[4] Support Vector Machine
[5] Ensemble Learning Algorithms



شکل ۵. طرح‌واره ساختار شبکه عصبی مصنوعی پرسپترون چند لایه

۳-۴- معیارهای ارزیابی مدل‌های طبقه‌بندی

با توجه به نوع مسئله‌ی طبقه‌بندی، در این پژوهش از معیارهای دقت[1]، عدد f1[2]، کوهن-کاپا[3] و ضریب رابطه‌ی متیو[4] برای ارزیابی و مقایسهٔ مدل‌های آموزش دیده استفاده شده‌است [30]. این معیارها با استفاده از ماتریس کانفیوژن[5] ارائه شده در شکل ۶ و روابط ارائه شده‌ی ۱ الی ۹ محاسبه می‌شوند. در روابط ارائه شده از نمادهای TP (مثبت درست)، TN (منفی درست)، FP (مثبت نادرست) و FN (منفی نادرست) با تعاریف ارائه شده در ماتریس درهم‌ریختگی استفاده شده‌است (شکل ۶).

		کلاس پیش بینی شده	
		مثبت	منفی
کلاس واقعی	مثبت	TP	FN
	منفی	FP	TN

شکل ۶. ماتریس درهم ریختگی برای یک مسئله دودویی.

- دقت: معیاری است که نسبت تعداد نمونه‌هایی را که به درستی پیش‌بینی شده‌اند به کل نمونه‌ها را نشان می‌دهند. این معیار در مسائل طبقه‌بندی که در آنها مجموعه داده یک مجموعه داده‌ی نامتعادل است و تفاوت فاحشی در تعداد داده‌های مربوط به چند کلاس مختلف خروجی وجود دارد، مناسب نیست و نتایج قابل اعتمادی را نشان نمی‌دهد.

$$\text{دقت} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۱)$$

[1] Accuracy
[2] F1 Score
[3] Cohen-Kappa
[4] Mathew's Correlation Coefficient (MCC)
[5] Confusion Matrix

- عدد f1: میانگین هارمونیک دو معیار درستی[1] و یادآوری[2] است. هرچه مقدار این پارامتر بیشتر باشد به معنای این است که مدل عملکرد مناسبتری دارد.

$$\text{درستی} = \frac{TP}{TP + FP} \quad (۲)$$

$$\text{یادآوری} = \frac{TP}{TP + FN} \quad (۳)$$

$$\text{پارامتر } f1 = \frac{2 * \text{درستی} * \text{یادآوری}}{\text{درستی} + \text{یادآوری}} \quad (۴)$$

- کوهن-کاپا: این معیار احتمالات موجود در پیش‌بینی موارد را در نظر بگیرد. این معیار بسیار قابل اعتمادتر از دقت است. مقدار این معیار در بازه‌ی ۰ تا ۱ است و هرچه این معیار بیشتر باشد به معنای این است که مدل عملکرد بهتری دارد. مدلی که معیار کوهن کاپای بیشتر از ۰/۸ داشته‌باشد بعنوان مدل مناسب شناخته می‌شود.

$$cohen - kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad (۵)$$

در این رابطه p_0 همان دقت است و p_e با استفاده از رابطه‌های ۶ الی ۸ محاسبه میشود.

$$p_e(\text{تبثم شیپ ینیب هدش ، تبثم یعقاو}) = \text{احتمال مثبت پیش بینی شده} \times \text{احتمال مثبت واقعی} \quad (۶)$$

$$\text{احتمال مثبت واقعی} = \frac{TP+FN}{TP+TN+FP+FN} \quad (۷)$$

$$\text{احتمال مثبت پیش بینی شده} = \frac{\text{TP} + FP}{TP + TN + FP + FN} \quad (۸)$$

- ضریب رابطه میتو: تمام چهار مقدار موجود در ماتریس کانفیوژن را در رابطه‌ی خود دارد. این پارامتر در بازه‌ی ۱- تا ۱+ می‌باشد و هرچه مقدار آن بیشتر باشد نشانگر از عملکرد مناسب مدل است. بصورت کلی اگر مقدار این پارامتر برای مدلی بیشتر از ۰/۷ باشد آن مدل، مدل مناسبی می‌باشد.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(\text{TP} + \text{FP})(TP + FN)(TN + FP)(TN + FN)}} \quad (۹)$$

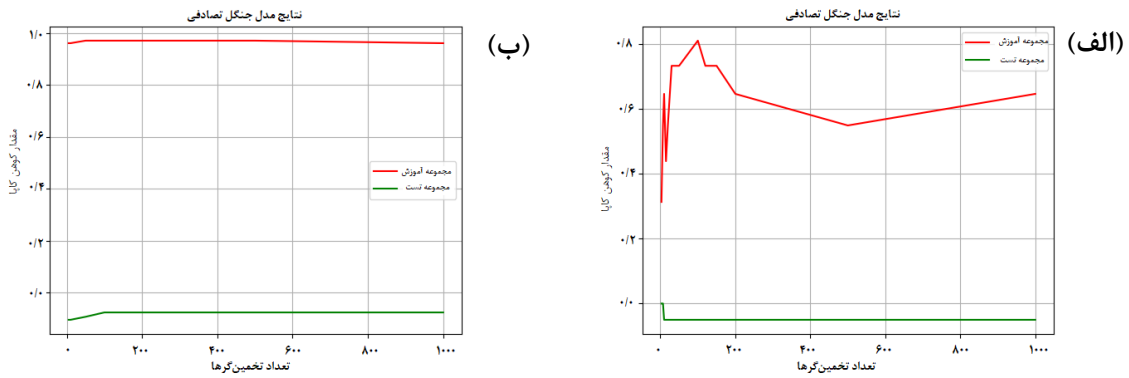
۵- بحث و نتایج

پس از آماده سازی داده ها، چهار مدل مختلف یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، پرسپترون چند لایه و تقویت گرادیان شدید با مجموعه داده‌ی آموزش، آموزش دیدند. جهت انتخاب دقیق‌ترین مدل، مؤثرترین پارامتر هر یک از مدل‌های یادگیری ماشین که تغییر آنها تأثیر بسیار زیادی بر دقت نهایی مدل دارد انتخاب شدند و هر یک از مدل‌ها با مقادیر مختلفی از پارامترهای انتخاب شده‌شان مورد آموزش و ارزیابی قرار گرفتند. در این راستا، تعداد درختان تصمیم‌گیری در مدل‌های

[1] Precision
[2] Recall

جنگل تصادفی، پارامترهای هسته[1] و درجه‌ی چند جمله‌ای در مدل ماشین بردار پشتیبان، تعداد لایه‌های پنهان و تعداد نرون‌های واقع در این لایه‌ها در مدل پرسپترون چند لایه و تعداد تخمین‌زننده‌ها (تخمین‌گرها) در مدل تقویت گرادیان شدید بعنوان پارامترهای بسیار مؤثر انتخاب شدند. لازم به ذکر است که پارامتر هسته در ماشین بردار پشتیبان برای تعیین کردن شکل خطوط مرزی میان کلاس‌ها (خطی یا چند جمله ای) بکار می‌رود. در این پژوهش، هر یک از مدل‌های مذکور در دو مرحله شامل آموزش با داده‌های صرفاً واقعی و آموزش با داده‌های واقعی و مصنوعی آموزش دیدند و بر اساس معیار کوهن-کاپا ارزیابی شدند. نتایج ارائه شده در شکل ۷، شکل ۸، شکل ۱۰، شکل ۱۱ و شکل ۱۲ به ترتیب نشان‌دهنده‌ی نتایج اعمال مدل‌های جنگل تصادفی، ماشین بردار پشتیبان، تقویت گرادیان شدید و پرسپترون چند لایه با پارامترهای مختلف بر روی مجموعه داده‌های آموزش می‌باشند.

بخش الف در شکل ۷ نشان می‌دهد که با افزایش تعداد درختان تصمیم‌گیری در مدل جنگل تصادفی، معیار کوهن-کاپا بر روی داده‌های آموزش افزایش می‌یابد، بدان معنا که مدل‌های آموزش دیده قادر به پیش‌بینی خروجی داده‌های آموزش با دقت بالاتر می‌شوند. با این وجود، معیار کوهن-کاپا بر روی داده‌های تست با افزایش تعداد درختان تصمیم‌گیری از مقدار اولیه‌ی صفر به مقادیر کمتر کاهش می‌یابد. این موضوع نشان‌دهنده‌ی ضعیف‌تر شدن عملکرد مدل‌های آموزش دیده بر روی داده‌های تست بر اثر برازش بیش از حد[2] مدل بر روی داده‌های آموزش می‌باشد. با توجه به این موضوع که مهم‌ترین معیار ارزیابی یک مدل یادگیری ماشین میزان عملکرد آن بر روی داده‌های تست است که آستانه آن نیز معیار کوهن-کاپای بالاتر از ۰/۸ در مسائل طبقه‌بندی می‌باشد، می‌توان نتیجه‌گیری کرد که مدل جنگل تصادفی آموزش دیده با تعداد درختان تصمیم‌گیری مختلف مدل مناسبی جهت پیش‌بینی خروجی داده‌های آزمایشگاهی نمی‌باشد. این موضوع به این دلیل است که هدف آموزش مدل‌های یادگیری ماشین رسیدن به سیستمی است که بتواند نتایج تست‌های آزمایشگاهی انجام نشده را که داده‌های آن‌ها را تاکنون ندیده است، پیش‌بینی کند.

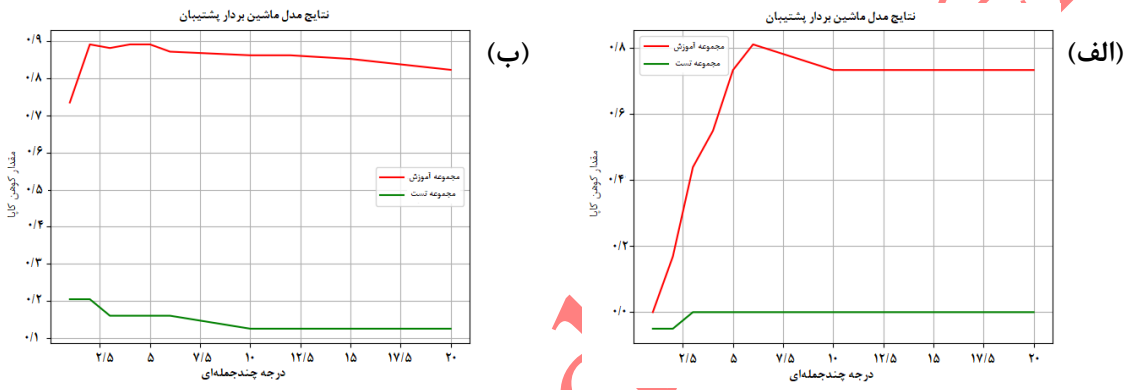


شکل ۷. نتایج اعمال روش جنگل تصادفی بر روی داده‌های آموزش و تست زمانیکه تنها با داده‌های واقعی آموزش دیده اند (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده اند (ب)

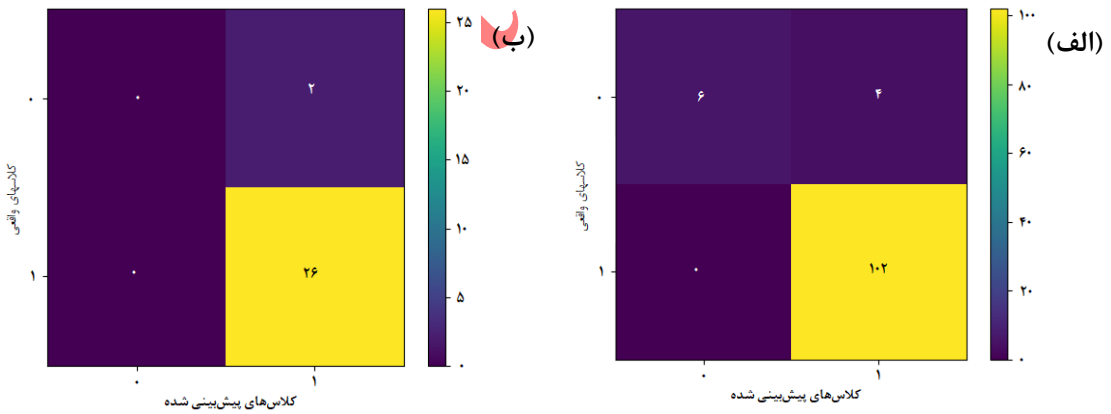
همچنین نتایج ارائه شده در بخش (الف)شکل ۸، شکل ۱۰، شکل ۱۱ و شکل ۱۲ نشان می‌دهند که تمامی مدل‌های آموزش دیده معیار کوهن-کاپا و عملکرد نسبتاً خوبی بر روی داده‌های آموزش دارند اما هیچ یک از آن‌ها عملکرد مناسبی بر روی داده‌های

[1] Kernel
[2] Over-fitting

تست ندارند و این موضوع از مقدار کوهن-کاپای منفی و صفر بر روی داده‌های تست نتیجه‌گیری می‌شود. عملکرد بسیار ضعیف این مدل‌ها بعلت وجود تعداد بسیار زیادی نمونه در کلاس "جدایش مناسب" به تعداد ۱۰۰ داده و تعداد بسیار کمی داده در کلاس "جدایش نامناسب" به تعداد ۹ داده می‌باشد (شکل ۲، ب). در طی وجود این مشکل که اصطلاحاً مشکل مجموعه داده نامتعادل نامیده می‌شود، مدل آموزش دیده توسط این مجموعه داده در جهت رسیدن به بالاترین میزان دقت (نسبت تعداد داده‌های پیش‌بینی شده‌ی درست به کل داده‌ها)، تمامی داده‌ها در کلاس اقلیت را بعنوان نمونه‌هایی از کلاس اکثریت پیش‌بینی می‌کند. زیرا داده‌هایی در کلاس اکثریت هستند و در نهایت به درستی پیش‌بینی می‌شوند، درصد بالایی از کل داده‌ها را شکل می‌دهند. مشکل مجموعه داده نامتعادل یکی از اساسی‌ترین مشکلات مسائل طبقه‌بندی در حوزه‌ی هوش مصنوعی می‌باشد که عدم کنترل آن و نیز ارزیابی مدل‌ها بر اساس معیار دقت می‌تواند منجر به خطای بسیار بزرگ و ارزیابی غیرقابل اعتمادی گردد.



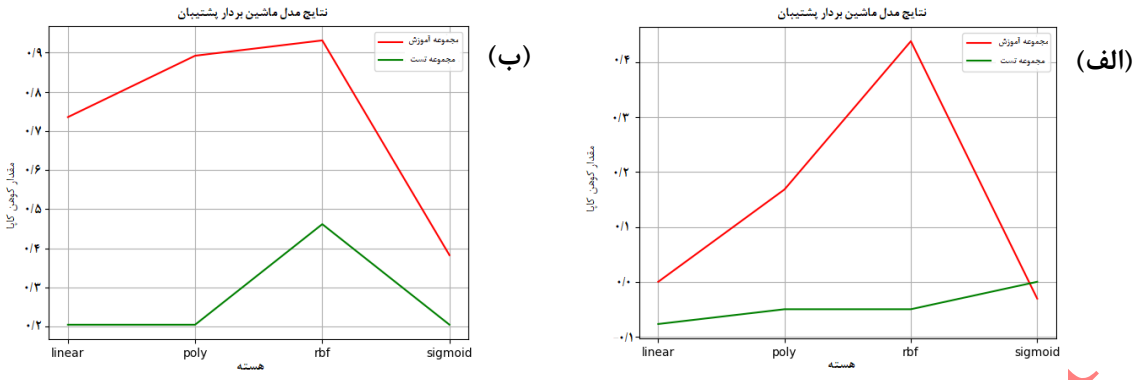
شکل ۸. نتایج اعمال روش ماشین بردار پشتیبان با توابع از درجات مختلف بر روی داده‌های آموزش و تست زمانیکه تنها با داده‌های واقعی آموزش دیده (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده (ب)



شکل ۹. ماتریس درهم‌ریختگی عملکرد مدل ماشین بردار پشتیبان با چندجمله ای درجه ۱۰ بر روی داده‌های آموزش متشکل از داده‌های صرفاً واقعی (الف) و داده‌های تست (ب)

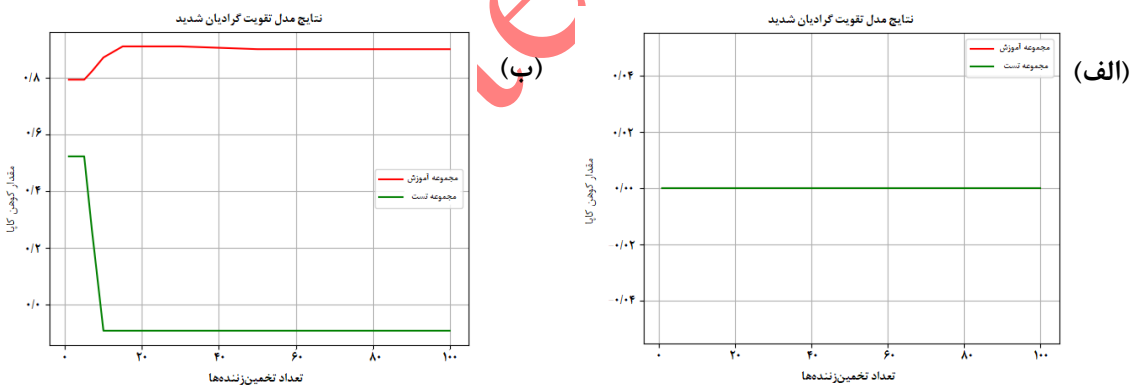
افزون بر معیار کوهن-کاپا، ماتریس درهم‌ریختگی مدل‌های آموزش دیده نشان‌دهنده‌ی عملکرد بسیار ضعیف و عدم توانایی آن‌ها در شناسایی و پیش‌بینی نمونه داده‌ای از کلاس اقلیت می‌باشند. ماتریس درهم‌ریختگی مدل ماشین بردار پشتیبان با چندجمله‌ای درجه ۱۰ به عنوان نمونه در شکل ۹ ارائه شده‌است. همانطور که از بخش ب در شکل ۹ مشاهده می‌شود، مدل آموزش دیده‌ی ماشین بردار پشتیبان با هسته چندجمله‌ای با درجه ۱۰ تمامی داده‌ها را بعنوان نمونه از کلاس اکثریت (جدایش مناسب) پیش‌بینی کرده و قادر به پیش‌بینی هیچ نمونه ای از داده‌های کلاس اقلیت نبوده است. معیار کوهن-کاپای صفر بر روی داده‌های

تست در بخش الف شکل ۸ برای مدل ماشین بردار پشتیبان با چندجمله ای درجه ۱۰ تأئیدی بر این موضوع است.



شکل ۱۰. نتایج اعمال روش ماشین بردار پشتیبان با هسته‌های مختلف بر روی داده‌های آموزش و تست زمانیکه تنها با داده‌های واقعی آموزش دیده (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده (ب)

بنا به عملکرد نامناسب مدل‌های آموزش دیده بر روی مجموعه داده‌ی آموزش شامل داده‌های صرفاً واقعی (بداده‌های آزمایشگاهی)، داده‌های مصنوعی جدید با استفاده از روش بیش نمونه‌گیری مصنوعی ساخته‌شده و به داده‌های آموزش قبلی اضافه گردیدند. این کار صرفاً بر روی داده‌های آموزش اعمال شده‌است تا ارزیابی مدل‌های آموزش دیده با این مجموعه داده‌ی جدید نیز با استفاده از همان داده‌های تست قبلی (داده‌های صرفاً آزمایشگاهی) صورت گرفته و نتایج قابل اعتمادتر باشد. نتایج ارزیابی مدل‌های مختلف آموزش دیده با داده‌های آموزش واقعی و مصنوعی بر اساس معیار کوهن-کاپا در بخش (ب) شکل-هایشکل **۷**،شکل **۸**،شکل **۱۰**،شکل **۱۱** وشکل **۱۲** ارائه شده‌است.

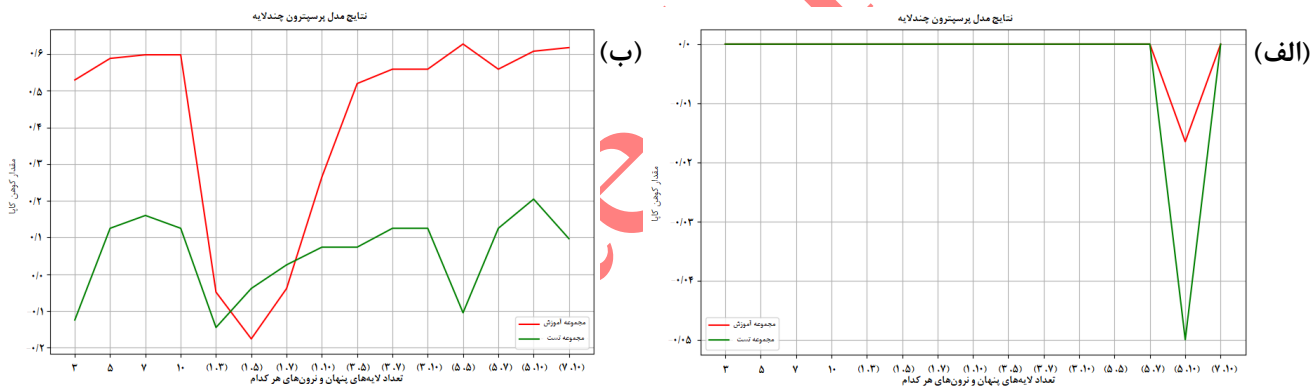


شکل ۱۱. نتایج اعمال روش تقویت گرادیان شدید بر روی داده‌های آموزش و تست زمانیکه تنها با داده‌های واقعی آموزش دیده (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده (ب)

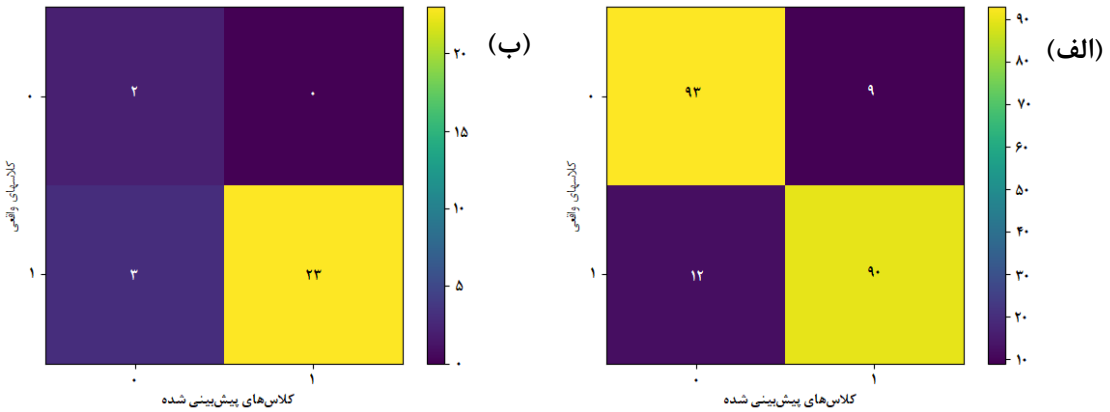
نتایج ارائه شده نشان می‌دهند که عملکرد تمامی مدل‌ها بر روی داده‌های تست بسیار بهبود یافته است بطوریکه مقدار عددی معیار کوهن-کاپا بر روی داده‌های تست مقادیر بیش از صفر می‌باشد. این بدان سبب است که مجموعه داده‌ی آموزش در این مرحله متعادل بوده و تعداد داده‌های موجود در هر یک از کلاس‌های "جدایش نامناسب" و "جدایش مناسب" یکی می‌باشد. در این میان، مدل تقویت گرادیان شدید با ۵ تخمین زننده، با دارا بودن بالاترین مقادیر کوهن-کاپا به میزان ۰/۷۹ بر روی داده‌های آموزش و ۰/۵۲۳ بر روی داده‌های تست بالاترین میزان دقت و عملکرد را در میان مدل‌های اعمال شده دارد (شکل ۱۱، ب). پس از مدل تقویت گرادیان شدید نیز مدل ماشین بردار پشتیبان با هسته "rbf" با داشتن مقادیر کوهن-کاپای ۰/۹۳ و ۰/۴۶

به ترتیب بر روی داده‌های آموزش و تست عملکرد مناسبی در پیش‌بینی خروجی‌های مجموعه داده دارد (شکل ۱۰، ب).

شکل ۱۱، ب نشان‌دهنده‌ی این موضوع است که با افزایش تعداد درختان تصمیم‌گیری در مدل تقویت گرادیان شدید، عملکرد مدل بر روی داده‌های آموزش افزایش می‌یابد، حال آنکه دقت مدل در پیش‌بینی خروجی‌های داده‌های تست کاهش می‌یابد. زیرا با افزایش تعداد درختان تصمیم‌گیری، درختان در جهت برازش بیش از حد بر روی داده‌های آموزش می‌شوند و بدین ترتیب صرفاً می‌توانند عملکرد مناسبی بر روی داده‌هایی که با آنها آموزش دیده‌اند داشته باشند و قادر به پیش‌بینی خروجی داده‌های جدید و دیده نشده[1] با دقت بالا نباشند. همچنین شکل ۱۰، ب نیز بیانگر این موضوع است که مدل ماشین بردار پشتیبان با هسته "rbf" که نوع خاصی از مرزبندی میان کلاس‌ها را اجرا می‌کند، بالاترین معیار کوهن-کاپا را هم بر روی داده‌های آموزش و هم بر روی داده‌های تست دارد. شکل ۱۳ و شکل **۱۴** به ترتیب ماتریس درهم‌ریختگی مدل‌های تقویت گرادیان شدید با ۵ تخمین زننده و مدل ماشین بردار پشتیبان با هسته rbf را که توسط مجموعه داده‌ی آموزش شامل داده‌های واقعی و مصنوعی آموزش‌دیده و توسط داده‌های تست واقعی مورد ارزیابی قرار گرفته است را نشان می‌دهند. مقادیر عددی غیرصفر در قطر اصلی ماتریس درهم‌ریختگی در شکل ۱۳، ب وشکل **۱۴**، ب نشانگر این موضوع هستند که مدل‌های آموزش دیده، توانایی نسبتاً مناسبی برای پیش‌بینی خروجی داده‌های کلاس "جدایش نامناسب" (کلاس ۰ در ماتریس درهم‌ریختگی) در مجموعه داده‌ی تست را دارا هستند.

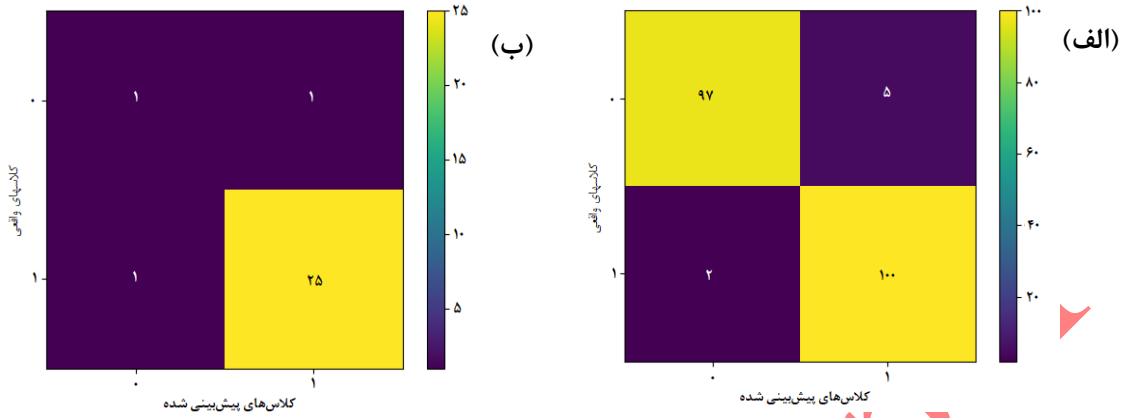


شکل ۱۲. نتایج اعمال روش پرسپترون چند لایه بر روی داده‌های آموزش و تست زمانیکه تنها با داده‌های واقعی آموزش دیده (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده (ب)



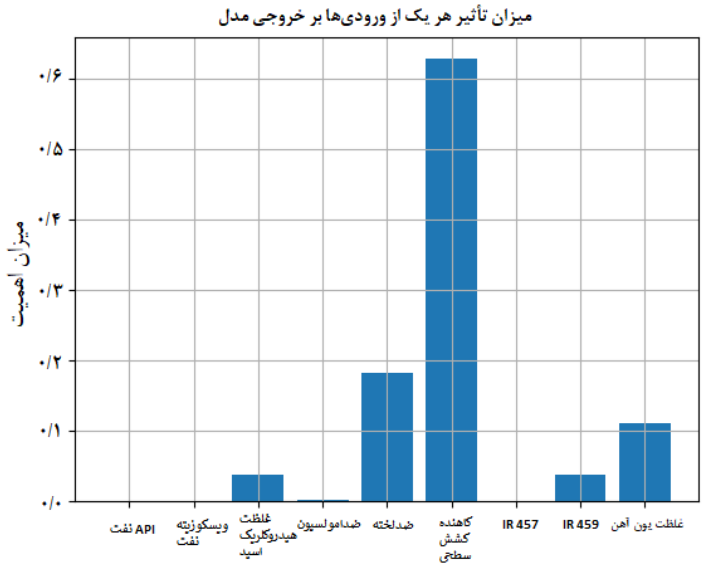
[1] Unseen

شکل ۱۳. ماتریس درهم‌ریختگی عملکرد مدل تقویت گرادیان شدید با ۵ تخمین‌زننده بر روی داده‌های آموزش متشکل از داده‌های واقعی و مصنوعی (الف) و داده‌های تست (ب)



شکل ۱۴. ماتریس درهم‌ریختگی عملکرد مدل ماشین بردار پشتیبان با هسته "rbf" بر روی داده‌های آموزش متشکل از داده‌های واقعی و مصنوعی (الف) و داده‌های تست (ب)

همچنین لازم به ذکر است که پس از آموزش مدل‌ها و ارزیابی آن‌ها، جهت آنالیز حساسیت سنجی میزان تأثیر هر یک از ورودی‌ها بر خروجی مدل توسط بهترین مدل ساخته شده (مدل تقویت گرادیان شدید با ۵ تخمین‌زننده) به شکل ۱۵ بدست آمد.



شکل ۱۵. آنالیز حساسیت سنجی میزان تأثیر هر یک از ورودی‌ها بر خروجی مدل (نتیجه تست ضد اموسیون)

مطابق با نتایج به‌دست‌آمده در شکل ۱۵، میزان غلظت افزایه کاهنده کشش سطحی، ضد لخته و غلظت یون آهن بیشترین میزان تأثیر را بر روی نتیجه‌ی تست‌های ضد اموسیون در آزمایشگاه اسیدزنی دارند. با توجه به این مسئله که اموسیون اسید در نفت به دلیل پایه آبی بودن اسید و وجود سورفکتانت‌های طبیعی در نفت به وجود می‌آید، می‌توان گفت افزایه کاهنده کشش سطحی با کاهش نیروی میان این بخش‌ها تأثیر بسزایی در جدایش فازها دارد. همچنین می‌توان گفت که لجن اسیدی (لخته) نوعی اموسیون غلیظ و پایدار اسید در نفت می‌باشد. از آنجاییکه غلظت یون آهن فریک و افزایه ضد لخته بر میزان تشکیل این

امولسیون پایدار تأثیر دارد، می‌توان نتایج به‌دست‌آمده در شکل ۱۵ را توجیه کرد.

۶- نتیجه‌گیری

در این کار تحقیقاتی که با استفاده از اطلاعات واقعی آزمایشگاهی مربوط به چندین نوع نفت و افزایه‌های مختلف انجام شد، چندین مدل طبقه‌بندی از مدل‌های یادگیری ماشین آموزش و مورد ارزیابی قرار گرفتند. در این پژوهش از داده‌های مربوط به ۱۴۰ تست استاتیک اسیدزنی اجرا شده در آزمایشگاه انگیزش چاه دانشگاه صنعتی شریف با استفاده از هیدروکلریک اسید با غلظت‌های مختلف و ۱۳ نوع نفت از میادین جنوب غربی ایران استفاده شده‌است. هدف این تحقیق یافتن یک مدل یادگیری ماشین دقیق با عملکرد مناسب بود که بتواند نتایج تست‌های آزمایشگاهی را بر اساس داده‌های ورودی مربوط به اطلاعات نفت، اسید و افزایه‌های مورداستفاده، پیش‌بینی کند. این کار برای بهینه کردن تعداد آزمایش‌های موردنیاز در آزمایشگاه استاتیک ضد امولسیون در حوزه‌ی اسیدکاری صورت گرفته است که در ادامه خلاصه‌ای از نتایج حاصل شده، ارائه می‌شود:

- گزارش‌های آزمایشگاهی از آزمایش‌های ضد امولسیون برای عملیات اسیدکاری بیانگر این موضوع بوده است که داده‌های این مجموعه از تست‌ها دارای تعداد متنوعی از پارامترهای ورودی می‌باشند. با توجه به اهمیت استفاده از تعداد ورودی‌های بهینه در مدل‌های یادگیری ماشین جهت جلوگیری از ایجاد بار محاسباتی مازاد، مؤثرترین ویژگی‌ها بر خروجی تست‌های ضد امولسیون شامل داده‌های مربوط به غلظت هیدروکلریک اسید و افزایه‌های تزریقی همچون افزایه‌های ضد امولسیون، ضد لخته، کاهنده کشش سطحی و کاهنده یون آهن، ویژگی‌های نفت همچون گرانروی و دانسیته و نیز غلظت یون فریک بعنوان ورودی‌های مدل‌های یادگیری ماشین انتخاب و مورد استفاده قرار گرفتند.
- در این کار تحقیقاتی، با توجه به مشکلات مربوط به بانک اطلاعاتی در دسترس (همچون تعداد داده‌های کم و نیز عدم توازن مجموعه داده)، هزینه سنگین و زمانبر بودن آزمایش‌های مورد نیاز برای تکمیل بانک اطلاعاتی از روش بیش نمونه‌گیری مصنوعی (SMOTE) استفاده گردید. بهبود عملکرد مدل‌های طبقه‌بندی با اجرای این روش، نشان-دهنده‌ی اهمیت روش بیش نمونه‌گیری مصنوعی در بهبود مجموعه بانک اطلاعات داده‌های سایر کارهای مختلف در صنعت نفت می‌باشد.
- با توجه به نوع مسئله در این پژوهش و نیز وجود مشکل عدم توازن در مجموعه داده، از معیار کوهن-کاپا جهت تعیین بهترین مدل استفاده شده‌است. پس از آموزش و ارزیابی مدل‌های جنگل تصادفی، پرسپترون چندلایه، ماشین بردار پشتیبان و مدل توقیت گرادیان شدید، مدل تقویت گرادیان شدید با ۵ تخمین‌زننده و آموزش دیده با داده‌های آموزش ترکیبی (شامل داده‌های آزمایشگاهی و داده‌های ساخته شده با روش بیش نمونه‌گیری مصنوعی) به عنوان بهترین مدل آموزش دیده در این مسئله انتخاب شد. مقادیر کوهن-کاپا در این مدل ۰/۷۹ بر روی داده‌های آموزش و ۰/۵۲۳ بر روی داده‌های تست می‌باشد.

با توجه به اهمیت هوشمند سازی عملیات اسیدزنی چاه‌های نفت و گاز، امکان استفاده از این مدل بر روی سایر تست‌های استاتیک وجود دارد اما مطابق با نتایج بدست آمده، دقت مدل بدست آمده‌ی نهایی بدلیل کم و محدود بودن داده‌های در دسترس بسیار زیاد نیست. بنابراین توصیه میشود در پژوهش‌های آتی، با رفع مشکل محدودیت در دسترسی به داده‌های کافی، غنی‌سازی بانک اطلاعاتی با اطلاعات بومی و قابل اطمینان و بالا بردن دقت و عملکرد مدل از کیفیت نتایج اطمینان حاصل کرد و سپس از

آن برای پیش‌بینی نتیجه‌ی سایر تست‌های استاتیک در آزمایشگاه‌های اسیدزنی کشور استفاده نمود.

تشکر و قدردانی

از شرکت ملی نفت ایران بخش مدیریت اکتشاف جهت همکاری و تأمین اطلاعات لازم برای انجام این پژوهش نهایت قدردانی و سپاسگزاری به عمل می‌آید. همچنین از آقایان مهندس روشنی و موسوی جهت یاری رساندن در جمع‌آوری داده‌های آزمایشگاهی تشکر و قدردانی می‌گردد.

مراجع

[1] M. J. Economides and K. G. Nolte, *Reservoir Stimulation*, 3rd ed. Chichester, England: John Wiley & Sons, 2000.

[2] C. Dai and F. Zhao, "Oilfield Chemistry," *Oilf. Chem.*, no. King 1986, pp. 1–395, 2019, doi: 10.1007/978-981-13-2950-0.

[3] M. Mohammadzadeh Shirazi, S. Ayatollahi, and C. Ghotbi, "Damage evaluation of acid-oil emulsion and asphaltic sludge formation caused by acidizing of asphaltenic oil reservoir," *Journal of Petroleum Science and Engineering*, vol. 174. pp. 880–890, 2019. doi: 10.1016/j.petrol.2018.11.051.

[4] Y. M. Ganeeva *et al.*, "The composition of acid/oil interface in acid oil emulsions Yulia," *Petroleum Science*, vol. 17. pp. 1345–1355, 2020. doi: https://doi.org/10.1007/s12182-020-00447-9.

[5] A. Pourakaberian, S. Ayatollahi, M. Mohammadzadeh Shirazi, C. Ghotbi, and H. Sisakhti, "A systematic study of asphaltic sludge and emulsion formation damage during acidizing process: Experimental and modeling approach," *J. Pet. Sci. Eng.*, vol. 207, no. 109073, 2021, doi: 10.1016/j.petrol.2021.109073.

[6] S. Shakouri and M. Mohammadzadeh-Shirazi, "Modeling of asphaltic sludge formation during acidizing process of oil well reservoir using machine learning methods," *Energy (Oxf.)*, vol. 285, no. 129433, p. 129433, 2023, doi: 10.1016/j.energy.2023.129433.

[7] C. Chavanne and H. G. Perthuis, "A fluid selection expert system for matrix treatments," *SPE Repr. Ser.*, no. 41, pp. 135–140, 1996, doi: 10.2523/24995-ms.

[8] A. A. S. Ebrahim, A. A. Garrouch, and H. M. S. Lababidi, "Automating sandstone acidizing using a rule-based system," *J. Pet. Explor. Prod. Technol.*, vol. 4, no. 4, pp. 381–396, 2014, doi: 10.1007/s13202-014-0104-3.

[9] D. Koroteev and Z. Tekic, "Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future," *Energy AI*, vol. 3, no. 100041, 2021, doi: 10.1016/j.egyai.2020.100041.

[10] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, "Application of machine learning and artificial intelligence in oil and gas industry," *Pet. Res.*, vol. 6, no. 4, pp. 379–391, 2021, doi: 10.1016/j.ptlrs.2021.05.009.

[11] S. D. Mohaghegh, "Recent developments in application of artificial intelligence in petroleum engineering." Journal of Petroleum Technology, 2005. doi: https://doi.org/10.2118/89033-JPT.

[12] H. H. Alkinani, A. T. T. Al-hameedi, S. Dunn-norman, and R. E. Flori, "Applications of Artificial Neural Networks in the Petroleum Industry : A Review," no. 1957, 2019, doi: https://doi.org/10.2118/195072-MS.

[13] S. Mohaghegh, *Virtual-intelligence applications in petroleum engineering: Part 1—Artificial neural networks*. Journal of Petroleum Technology, 2000. doi: https://doi.org/10.2118/58046-JPT.

[14] S. Mohaghegh, R. Arefi, S. Ameri, K. Aminiand, and N. Roy, "Petroleum Reservoir Characterization with the Aid of Artificial Neural Network." Journal of Petroleum Science and Engineering, 1996. doi: https://doi.org/10.1016/S0920-4105(96)00028-9.

[15] U. Sumotarto, A. D. Hill, and K. Sepehrnoori, "Integrated sandstone acidizing fluid selection and simulation to optimize treatment design," *Proc. - SPE Annu. Tech. Conf. Exhib.*, vol. Delta, pp. 717–724, 1995, doi: 10.2118/30520-ms.

[16] R. P. Kellogg, W. Chessum, and R. Kwong, “Machine learning application for wellbore damage removal in the wilmington field,” *SPE West. Reg. Meet. Proc.*, vol. 2018-April, 2018, doi: 10.2118/190037-ms.

[17] Z. Sidaoui, A. Abdulraheem, and M. Abbad, “Prediction of optimum injection rate for carbonate acidizing using machine learning,” *Soc. Pet. Eng. - SPE Kingdom Saudi Arab. Annu. Tech. Symp. Exhib. 2018, SATS 2018*, 2018, doi: 10.2118/192344-ms.

[18] H. Xue, P. Liu, N. Li, Z. Luo, and L. Zhao, “Expert system for acidizing based on BP neural network,” *Adv. Mater. Res.*, vol. 548, pp. 438–443, 2012, doi: 10.4028/www.scientific.net/AMR.548.438.

[19] M. S. Van Domelen, W. G. F. Ford, and T. J. Chiu, “An expert system for matrix acidizing treatment design,” *SPE Repr. Ser.*, no. 41, pp. 179–188, 1992, doi: 10.2523/24779-ms.

[20] M. Alkathim, M. S. Aljawad, A. Hassan, S. A. Alarifi, and M. Mahmoud, “A data-driven model to estimate the pore volume to breakthrough for carbonate acidizing,” *J. Pet. Explor. Prod. Technol.*, vol. 13, no. 8, pp. 1789–1806, 2023, doi: 10.1007/s13202-023-01642-1.

[21] A. Hassan, M. S. Aljawad, and M. Mahmoud, “An Artificial Intelligence-Based Model for Performance Prediction of Acid Fracturing in Naturally Fractured Reservoirs,” *ACS Omega*, vol. 6, pp. 13654–13670, 2021, doi: doi: 10.1021/acsomega.1c00809.

[22] M. Dargi, E. Khamehchi, and J. M. Kalatehno, “Optimizing acidizing design and effectiveness assessment with machine learning for predicting post-acidizing permeability,” *Sci. Rep.*, vol. 13, no. 11851, pp. 1–16, 2023, doi: 10.1038/s41598-023-39156-9.

[23] O. Sanni, O. Adeleke, K. Ukoba, J. Ren, and T.-C. Jen, “Prediction of inhibition performance of agro-waste extract in simulated acidizing media via machine learning,” *Fuel*, vol. 356, 2024, doi: https://doi.org/10.1016/j.fuel.2023.129527.

[24] C. Kurniawan, M. M. Azis, and T. Ariyanto, “Supervised Machine Learning and Multiple Regression Approaches to Predict the Successfulness of Matrix Acidizing in Hydraulic Fractured Sandstone Formation,” *ASEAN J. Chem. Eng.*, vol. 23, no. 1, pp. 113 – 127, 2023, doi: https://doi.org/10.22146/ajche.78255.

[25] C. R. Blackburn, J. C. Abel, and R. Day, “An Expert System to Design and Evaluate Matrix Acidizing,” *SPE Comput. Appl.*, vol. 2, pp. 15–17, 1990, doi: https://doi.org/10.2118/20337-PA.

[26] T.-J. Chiu, E. A. Caudell, and F.-L. Wu, “Development of an Expert System to Assist with Complex Fluid Design,” *SPE Comput. Appl.*, vol. 5, pp. 18–20, 1993, doi: https://doi.org/10.2118/24416-PA.

[27] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8. pp. 1509–1515, 2005. doi: 10.1093/bioinformatics/bti171.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. Sept. 28. pp. 321–357, 2002. doi: 10.1613/jair.953.

[29] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14. 2013. doi: 10.1186/1471-2105-14-106.

[30] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting methods for multi-class imbalanced data classification: an experimental review,” *J. Big Data*, 2020, doi: 10.1186/s40537-020-00349-y.

پیوست

جدول ۱. شبه کد الگوریتم مورد استفاده در مقاله

```
BEGIN
  #   Step 1: Read the data file
  INPUT data_file
  Read data_file into dataset
  #   Step 2: Handle missing values
  IF dataset contains missing values THEN
    REMOVE missing values from dataset
  #   Step 3: Handle outliers
  IF dataset contains outliers THEN
    REMOVE outliers from dataset
  #   Step 4: Handle duplicate values
  IF dataset contains duplicate values THEN
    REMOVE duplicate values from dataset
  #   Step 5: Split dataset into training and testing sets
  CALL split the data set into train set and test set
  #   Step 6: Apply SMOTE to training set
  CALL SMOTE oversampling technique only on train set
  #  Step 7: Normalize input values
  CALL minimum, maximum normalizing on feature values of train set and test set
  #   Step 8: Shuffle data order
  CALL pandas sampling on train set samples
  #   Step 9: Train model
  CALL fit the model on train set
  #   step 10: Predict test set results
  CALL predict the test set results by trained model
  #   Step 11: Evaluate model performance by calculating metrics
  CALL calculate confusion matrix for test set real results and the predicted ones by trained
model
  CALL cohen kappa score for test set real results and the predicted ones by trained model
  #  Step 12: Output the performance metrics
  OUTPUT confusion matrix for each trained model
  OUTPUT cohen kappa score for each trained model
END
```

جدول ۲. پارامترهای مورد بررسی در هر یک از مدل‌های یادگیری ماشین آموزش‌دیده

مدل	پارامتر	مقادیر پیشنهادی
جنگل تصادفی	تعداد درختان تصمیم‌گیری	[۳, ۵, ۷, ۱۰, ۱۵, ۲۰, ۳۰, ۵۰, ۱۰۰, ۱۲۰, ۱۵۰, ۲۰۰, ۵۰۰, ۱۰۰۰]

پرسپترون چندلایه	تعداد لایه‌های پنهان	[۱، ۲]
	تعداد نرون‌های واقع در هر لایه پنهان	[(۳), (۵), (۷), (۱۰), (۱,۳), (۱,۵), (۱,۷), (۱,۱۰), (۳,۵), (۳,۷), (۳,۱۰), (۵,۵), (۵,۷), (۵,۱۰), (۷,۱۰)]
ماشین بردار پشتیبان	هسته	‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’
	درجه‌ی چند جمله‌ای	[۱, ۲, ۳, ۴, ۵, ۶, ۱۰, ۱۲, ۱۵, ۲۰]
تقویت گرادیان شدید	تعداد تخمین‌گر (تخمین‌زننده)	[۱,۳,۵,۷,۱۰,۱۵,۲۰,۳۰,۵۰,۱۰۰]

Utilizing Machine Learning Classification Models for Acid-Oil Emulsion Prediction in Laboratory Acidizing Static Tests by Using a Hybrid Databank

Abstract

During the lifetime of an oil well, the near wellbore areas are usually exposed to formation damage due to factors such as fines migration, clay swelling, etc., significantly reducing the oil well's productivity and injectivity rates. One of the widely used well-stimulation methods to remove formation damage is acidizing in which the acid and chemicals (additives) are injected into the formation to increase the permeability of the formation by dissolving carbonate rocks. However, the lack of laboratory examination of the compatibility of injection fluids with formation fluids at the design stage results in induced damage such as acid emulsion in oil in formation. The conduction of laboratory tests in order to execute compatibility between fluids is time-consuming, expensive, and has issues related to safety. This research aims to predict the primary results of anti-emulsion tests using data-driven models in a short time. For this purpose, the most influential data on the results of these tests, including type and concentration of acid, and additives like anti-emulsion, anti-sludge, surface tension reducer, and iron ion reducer, as well as properties of 13 different types of oil from various reservoirs, such as viscosity, density, and ferric ion concentration, were collected and recorded as inputs to a data set. Then, some supervised classification models including random forest, support vector machine, multi-layer perceptron, and extreme gradient boosting algorithms have been implemented to predict the output of anti-emulsion tests. Additionally, the statistical technique SMOTE was employed to generate artificial data samples and enhance AI models' performance. Results indicate that the extreme gradient boosting with five estimators achieved the best performance with Cohen's kappa values of 0.79 and 0.523 for training and testing datasets, respectively.

Keywords: Acidizing, Acid-Oil Emulsion, Acidizing Static Tests, Machine Learning, Classification, SMOTE.