Petroleum Research
Petroleum Research, 2025(February -March), Vol. 34, No. 139, 13-16
DOI:10.22078/pr.2024.5452.3426

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

Sepideh Atrbarmohammadi, Hossein Kheirollahi, Shahab Ayatollahi,* and Mahmoud Reza Pishvaie*
Department of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran
shahab@sharif.edu, pishvaie@sharif.edu
DOI:10.22078/pr.2024.5452.3426

Received: May 8, 2024 Accepted: October 12, 2024

Introduction

Acidizing is one of the most popular well-stimulation methods for damage removal, including clay swelling, sludge precipitation, grain movement, wettability change, etc. In this method, acid and additives are injected into the well to remove the precipitations or solve the rock for permeability recovery or improvement [1, 2]. One of the most common problems in acidizing implementation is the creation of induced damages like acid-oil emulsion due to the fluids' incompatibility with the formation fluid [3-5]. So, the best type and concentration of acids and additives for injection that are compatible with the formation fluid is determined by some laboratory tests, called acidizing static tests, before field operation [3, [5-7]. Since these tests are expensive, time-consuming, and dangerous, this research was done to find a machine-learning model to predict the results of emulsion static tests to optimize the number of required tests. In this research, some models such as Random Forest, Support Vector Machine, Multi-Layer Perceptron, and Extreme Gradient Boosting methods were trained with some fluid features such as HCl acid's concentration, anti-sludge additive's concentration, non-emulsion additive's concentration, iron reducer additive's concentration, IFT reducer additive's concentration, oil viscosity and API degree, and ferric ion's concentration.

Materials and Methods

A specific amount of ferric ions, acid, and additives are mixed in one vessel, while oil, bentonite, and microsilica are mixed in another. After thorough mixing, the contents are placed in a water bath for phase separation recording. Experiments vary ferric ion and hydrochloric acid concentrations while maintaining a 1:1 acid-to-oil ratio. A dataset of 140 samples is collected, including additives, acid concentration, oil properties, and iron ion concentration. Test results, indicating acid separation from oil, are categorized into "proper" and "improper" separation classes. Preprocessing includes outlier removal and deduplication, followed by a 4:1 split into training and testing sets. Machine learning models—Random Forest, Support Vector Machine, Multilayer Perceptron, and Gradient Boosting—are initially trained and evaluated on real data (Fig. 1, Right), then augmented using the SMOTE method for improved performance (Fig. 1, Left). Evaluation metrics include accuracy, recall, precision, F1 score, Cohen's Kappa, and Matthew's correlation coefficient, with Cohen's Kappa serving as the primary metric. A threshold of 0.8 is used for model suitability on both test and train set [8].

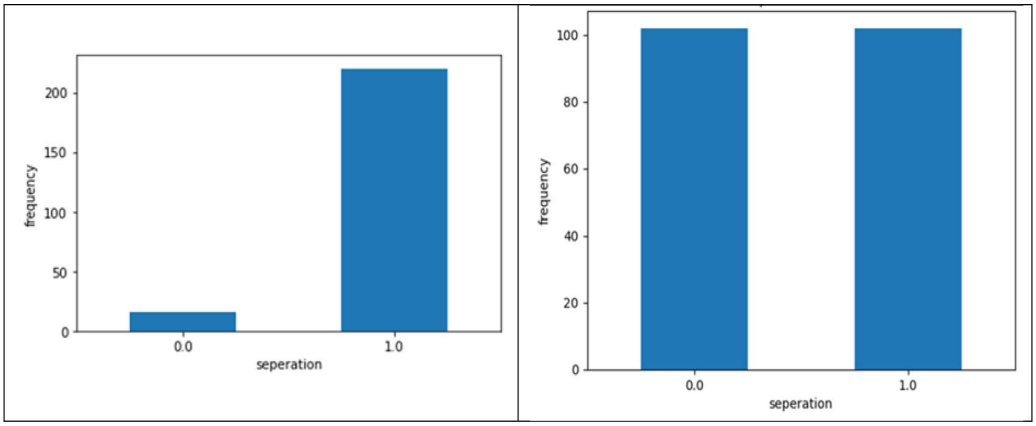


Fig. 1 Training data set outputs' bar plot before SMOTE (left) and after SMOTE method's implementation (right)

Results and Discussion

Each classification model was assessed by varying one parameter, and the performance of each model with different parameter values was compared through performance charts. Cohen's Kappa metric served as the criterion. However, when trained solely on real data, none of the models achieved a Cohen's Kappa metric greater than zero on the testing data. For instance, Fig. 2 illustrates Cohen's Kappa values for various support vector machine algorithms with different kernels trained solely on real data. The dataset's imbalance, with significantly more data points in the "proper separation" class (100 points) compared to the "improper separation" class (9 points), contributes to this challenge (Fig. 1, Left). This imbalance stems from extensive experience in conducting non-emulsion tests and effective anti-emulsion additives, leading to the majority of tests falling into the "proper separation" class. Consequently, each model tended to predict all data points as members of the majority class, thus failing to identify and predict data points from the minority class.

To address the poor performance of models trained solely on real data, synthetic training data were generated using the SMOTE method and added to the original training data. This augmentation process focused solely on the training data to ensure consistent evaluation using the existing testing data, comprised of real and laboratory data. The result is a new balanced training dataset comprising both real and synthetic data (Fig. 1, Right). Consequently, all models showed significantly improved performance on the testing data. The Extreme Gradient Boosting model with 5 estimators demonstrated the best performance, achieving Cohen's Kappa metrics of 0.79 on the training data and 0.523 on the testing data (Fig. 3, Left). Following this, the Support Vector Machine model with an "rbf" kernel achieved satisfactory performance, with Cohen's Kappa metrics of 0.93 and 0.46 on the training and testing data, respectively (Fig. 3, Right). This improvement is attributed to the balanced training dataset, with an equal number of data points in each class, "improper separation" and "proper separation."

The confusion matrix for the support vector machine algorithm with a polynomial degree of 10, trained solely on real data, highlights its limitations in predicting outcomes for minority class samples (Fig. 4). Conversely, the confusion matrix for the trained Extreme Gradient Boosting model with 5 estimators demonstrates the opposite pattern (Fig. 5).

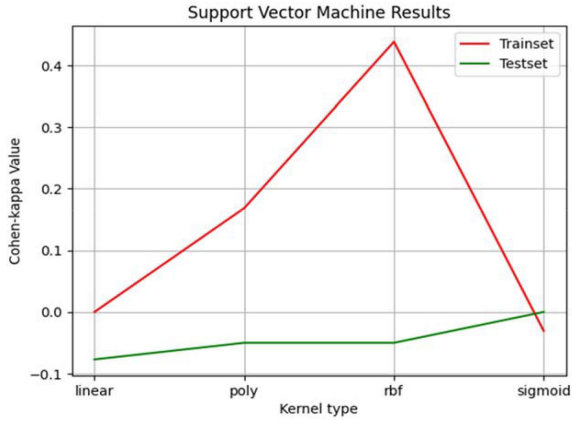


Fig. 2 The SVM performance on training data and test data trained with only real data

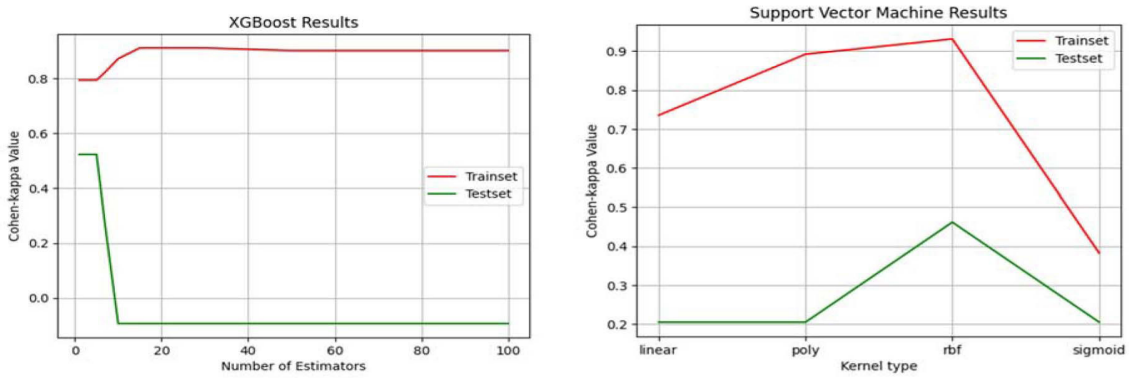


Fig. 3 The XGBoost (left) and SVC (right) algorithms' performance on training data and test data trained with combination of real and synthetic data (right).

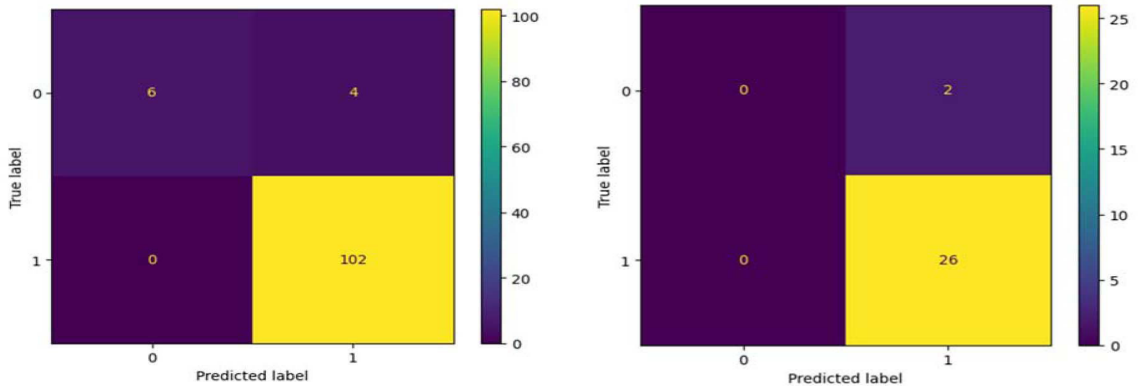


Fig. 4 Confusion matrix of SVC model with polynomial degree of 10 for train set (left) and test set (right)

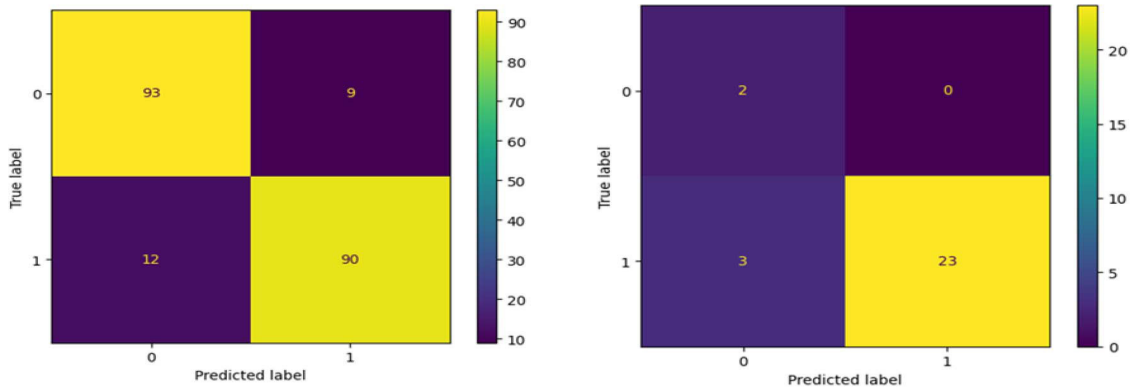


Fig. 5 Confusion matrix of XGB model with 5 estimators for train set (left) and test set (right).

Conclusions

• Laboratory reports on anti-emulsion tests for acidizing emphasize the need for selecting key input features in machine learning models to reduce computational load. The most influential factors include hydrochloric acid concentration, injected additives (anti-emulsion, anti-sludge agents, IFT reducer, iron ion reducers), oil properties (viscosity, density), and ferric ion concentration.

• Due to dataset limitations such as small sample size and imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to overcome the high cost and time constraints of additional lab tests. The improved performance of classification models highlights the effectiveness of SMOTE in enhancing datasets for various oil industry applications.

• Given the data imbalance, Cohen's Kappa was used to determine the best model. After training and evaluating Random Forest, Multilayer Perceptron, Support Vector Machine, and Extreme Gradient Boosting (XGBoost), the XGBoost model with five estimators, trained on a combined dataset (laboratory data + SMOTE-generated data), was selected as the best-performing model. It achieved a Cohen's Kappa of 0.79 on the training set and 0.523 on the test set.

References

1. Economides, M. J., & Nolte, K. G. (1989). Reservoir stimulation, (Vol. 2, pp. 6-10). Englewood Cliffs, NJ: Prentice Hall.
2. Dai C. and Zhao F. (2019). Oilfield Chemistry, Oilfield, Chemistry 1–395, , doi: 10.1007/978-981-13-2950-0.
3. Shirazi, M. M., Ayatollahi, S., & Ghotbi, C. (2019). Damage evaluation of acid-oil emulsion and asphaltic sludge formation caused by acidizing of asphaltenic oil reservoir. Journal of Petroleum Science and Engineering, 174, 880-890. doi: 10.1016/j.petrol.2018.11.051.
4. Ganeeva, Y. M., Yusupova, T. N., Barskaya, E. E., Valiullova, A. K., Okhotnikova, E. S., Morozov, V. I., & Davletshina, L. F. (2020). The composition of acid/oil interface in acid oil emulsions. Petroleum Science, 17, 1345-1355. doi: org/10.1007/s12182-020-00447-9.
5. Pourakaberian, A., Ayatollahi, S., Shirazi, M. M., Ghotbi, C., & Sisakhti, H. (2021). A systematic study of asphaltic sludge and emulsion formation damage during acidizing process: Experimental and modeling approach. Journal of Petroleum Science and Engineering, 207, 109073, doi: 10.1016/j.petrol.2021.109073.
6. Shakouri, S., & Mohammadzadeh-Shirazi, M. (2023). Modeling of asphaltic sludge formation during acidizing process of oil well reservoir using machine learning methods. Energy, 285, 129433, doi: 10.1016/j.energy.2023.129433.
7. Chavanne, C., & Perthuis, H. G. (1992, November). A Fluid selection expert system for matrix treatments. In SPE Europec featured at EAGE Conference and Exhibition? (pp. SPE-24995). SPE, doi: 10.2523/24995-ms.
8. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big data, 7, 1-47, doi: 10.1186/s40537-020-00349-y.

به‌کارگیری مدل‌های یادگیری ماشین برای پیش‌بینی تشکیل امولسیون اسید و نفت در آزمایش‌های استاتیک اسیدکاری با استفاده از بانک اطلاعات ترکیبی

سپیده عطربرمحمدی، حسین خیرالهی، سید شهاب‌الدین آیت‌اللهی* و سید محمودرضا پیشوائی*

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

تاریخ دریافت: ۱۴۰۳/۰۲/۱۹ تاریخ پذیرش: ۱۴۰۳/۰۷/۲۱

چکیده

در طول مدت بهره‌برداری از مخازن نفتی، معمولاً نواحی نزدیک به دیواره چاه به دلایل مختلفی همچون جابه‌جایی ذرات سازندی، تورم رس و موارد دیگر در معرض آسیب‌های مختلف قرار گرفته و نرخ تولید/تزریق از آن‌ها به‌شدت 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. آزمایش‌های آزمایشگاهی که به‌منظور 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 آزمایش‌های 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 به‌عنوان 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 آزمایش‌های ضد امولسیون برروی مجموعه داده جمع‌آوری‌شده اعمال شدند. با توجه به کمبود داده‌های آزمایشگاهی از تکنیک آماری بیش نمونه‌گیری مصنوعی به منظور تولید نمونه داده مصنوعی و بهبود عملکرد مدل‌های هوش مصنوعی استفاده گردید. براساس نتایج به‌دست آمده، روش تقویت گرادیان شدید با پنج تخمین‌گر به‌ترتیب با مقادیر کوهن-کاپای ۰/۷۹ و ۰/۵۲۳ برای مجموعه داده‌های آموزش و آزمایش بهترین عمکرد را داشته است.

کلمات کلیدی: اسیدکاری، امولسیون نفت و اسید، آزمایش‌های استاتیک اسیدکاری، یادگیری ماشین، طبقه‌بندی، بیش نمونه‌گیری مصنوعی

*مسؤول مکاتبات
آدرس الکترونیکی shahab@sharif.edu
pishvaie@sharif.edu
شناسه دیجیتال: (DOI:10.22078/pr.2024.5452.3426)

پژوهش نفت شماره ۱۳۹، بهمن و اسفند ۱۴۰۳، صفحه ۹۱-۷۳

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، جابه‌جایی 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 [۱]. 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 برروی میادین نفتی در سراسر جهان، پژوهشگران نفتی متوجه شدند که درصورتی‌که عملیات اسیدکاری به‌درستی طراحی و اجرا نگردد می‌توانند 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 به‌طور کامل بسته شده و حجم عظیمی از منابع نفتی از دست می‌رود. به همین دلیل مواد شیمیایی تحت عنوان افزایه‌های اسیدکاری برای جلوگیری از به‌وجود آمدن چنین مشکلات ثانویه ساخته شدند و به‌همراه اسید به مخزن تزریق می‌گردند [۲].

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 [۳]. 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 ویسکوزیته سیال (حدود چهار برابر) می‌شود و بدین ترتیب حرکت نفت در داخل سازند را دشوار و بهره‌دهی چاه را به‌شدت کاهش می‌دهد [۳]. در برخی موارد امولسیون به‌حدی است که بعد از اسیدکاری امکان بهره‌برداری از چاه وجود ندارد. تحقیقات قبلی نشان داده است که تشکیل امولسیون در نفت به دلیل 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 mμ) 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[1] (ترکیبات اشباع، آروماتیک، رزین و آسفالتین موجود در نفت) و چگالی نفت در میزان امولسیون تشکیل شده در سازند مؤثر می‌باشند [۳–۵]. در صنعت نفت و گاز برای جلوگیری از به‌وجود 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 آزمایش‌های آزمایشگاهی صورت می‌پذیرد. این آزمایش‌های آزمایشگاهی که برای تعیین غلظت مناسب افزایه‌های تزریقی انجام می‌شوند به آزمایش‌های استاتیک اسیدزنی معروف هستند و انجام هر یک از آن‌ها هزینه‌بر، زمان‌بر و برای پرسنل آزمایشگاه خطرناک است [۳، ۵–۷]. همچنین تعداد آزمایش‌های 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 می‌گردد. شرط لازم در دستورالعمل‌های عملیاتی بر این تأکید دارد که در پایان انجام آزمایش، امولسیون اسید-نفت پس از گذشت ۳۰ min از زمان تشکیل آن باید از بین برود [۴]. این آزمایش‌ها در آزمایشگاه‌های شرکت‌های بزرگ سرویس در سراسر جهان از سالیان گذشته انجام شده‌اند و مجموعه داده آن‌ها در همان شرکت‌ها ثبت و نگهداری می‌شوند. با پیشرفت حوزه کامپیوتر، ظهور علم هوش مصنوعی و جمع‌آوری داده‌ها در اندازه‌های بزرگ، مهندسان نفت از تکنولوژی‌های هوش مصنوعی برای سرعت بخشیدن 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 [۸–۱۴]. در این راستا می‌توان به تلاش محققان در زمینه‌های مختلف همچون استفاده از روش‌های یادگیری ماشین جهت انتخاب بهترین ناحیه پیاده‌سازی پایلوت برای روش‌های تزریق آب‌پایه، غربالگری روش‌های ازدیاد برداشت، بازسازی تصاویر سنگ‌های مخزن متراکم با شبکه عصبی و پیش‌بینی رفتار رئولوژیکی سیال حفاری اشاره کرد [۱۵–۱۸]. در زمینه اسیدکاری نیز دانشمندان 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 متعدد و مختلفی را تا حال حاضر جهت بهینه‌سازی عملیات اسیدکاری و پیش‌بینی میزان امولسیون تشکیل‌شده طراحی و ساخته‌اند [۵، ۷، ۸، ۱۹–۳۰]. افزون بر استفاده از مدل‌های هوش مصنوعی برای طراحی کل عملیات اسیدکاری، از این روش‌ها برای پیش‌بینی نتایج آزمایش‌های استاتیک اسیدزنی استفاده شده‌است. به‌طور مثال پوراکابریان و همکارانش توانستند میزان لختگی تشکیل‌شده در آزمایش‌های آزمایشگاهی را بدون حضور افزایه‌ها و با استفاده از یک شبکه عصبی مصنوعی با سه لایه پنهان و هفت نرون پیش‌بینی کنند [۵]. همچنین شکوری‌زاده و همکارانش برای بررسی تأثیر افزایه‌ها از مدل‌های مختلف رگرسیونی برای پیش‌بینی میزان لختگی تشکیل‌شده در آزمایش‌های استاتیک ضد لجن اسیدی استفاده کردند [۶].

1. SARA (Saturate, Aromatic, Resin and Asphaltene)

با این وجود به‌علت پیچیده بودن این فرآیند و در دسترس نبودن تعداد کافی از اطلاعات آزمایشگاهی، تاکنون مدل قابل اعتمادی در زمینه پیش‌بینی تشکیل امولسیون اسید در نفت در آزمایش‌های 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 آزمایش‌های 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 به‌طور قابل ملاحظه کاهش یابد. اعمال این روش‌ها برروی داده‌های آزمایشگاهی منجر به کاهش تعداد آزمایش‌های لازم برای اجرا شده و هزینه و طول مدت کارهای آزمایشگاهی کاهش می‌یابد.

آزمایش‌های استاتیک بررسی افزایه 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 درب آبی رنگ[1]) ۷۵ mL اسید با غلظت معین به‌همراه یون آهن فریک ریخته شده و افزایه‌های مشخص شده به محلول اضافه می‌گردد (بعد از اضافه شدن هر افزایه بطری هم‌زده شده تا افزایه در سیستم به‌خوبی پخش شود). سپس ۷۵ mL نفت مورد نظر به ظرفی که از قبل مقدار ۰/۳۷۵ g بنتونایت و ۳/۳۷۵ g 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 ریخته می‌شود و محتویات ظرف به مدت ۳۰ s با سرعت ۱۸۰۰۰ rpm به‌خوبی مخلوط می‌شوند. در نهایت محتویات ظرف به استوانه مدرج منتقل شده و این استوانه در حمام آب در دمای ۹۴ °C قرار می‌گیرد و در بازه‌های زمانی ۱۰ min و ۱۵ و ۳۰ ، ۱ h و نیز ۲ h پس از شروع آزمایش، میزان جدایش فاز اسید در آن به میلی‌لیتر (mL) 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 %wt و ۲۸ %wt انجام شده‌اند و نسبت اسید به نفت در تمامی آزمایش‌ها ۱:۱ بوده است. میزان افزایه‌ها براساس دستورالعمل مشخص شده از طرف سازنده این افزایه‌ها مورد استفاده قرار گرفت.

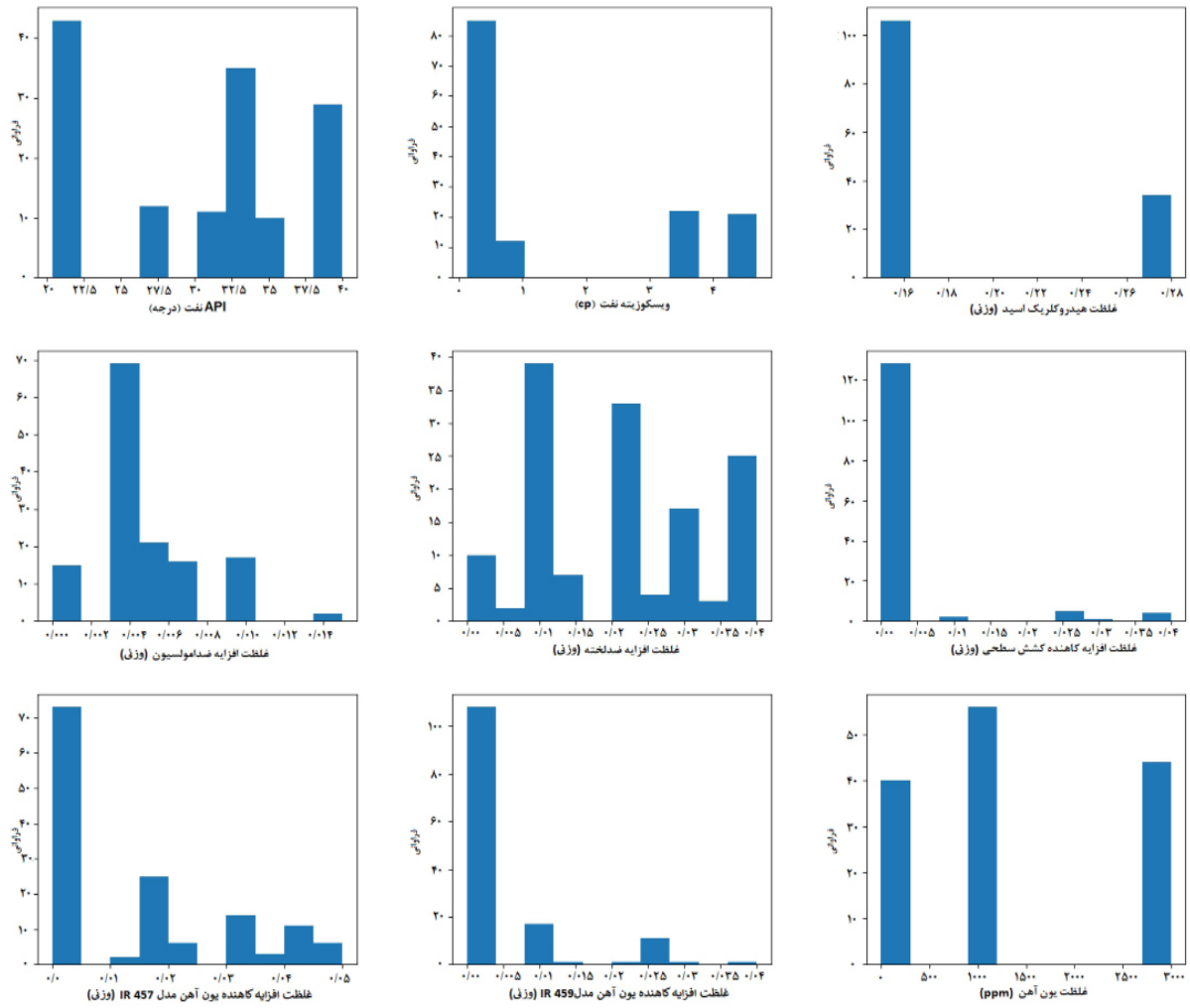
جمع‌آوری داده‌ها

مجموعه داده گردآوری شده در این پژوهش، اطلاعات مربوط به اجرای ۱۴۰ آزمایش استاتیک اسیدزنی با استفاده از هیدروکلریک اسید با غلظت‌های مختلف برروی 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 آزمایش‌ها گزارش شده‌است. در میان اطلاعات ارائه شده، غلظت افزایه‌هایی همچون افزایه‌های ضد خوردگی، از بین برنده هیدروژن سولفید، کنترل‌کننده یون آهن، معلق نگهدارنده در تمامی آزمایش‌های اجرا شده، ثابت و غلظت افزایه‌هایی همچون ضد امولسیون، ضد لخته، کاهنده‌ی کشش سطحی (سورفکتانت)، و نیز کاهنده یون آهن متغیر گزارش شده‌اند. همچنین خروجی‌های مربوط به آزمایش‌های ضد امولسیون نیز به‌صورت حجم اسید جدا شده از نفت در زمان‌های مختلف گزارش شده‌است. براساس تجربیات موجود، تعداد ویژگی‌ها و پارامترهای ورودی مدل‌های یادگیری ماشین در بار محاسباتی و نیز دقت مدل‌های آموزش دیده بسیار مؤثر است [۳۲].

1. Blue Cap

بنابراین بهینه‌سازی تعداد پارامترها برای کاهش بار محاسباتی و توانایی مدل برای پیش‌بینی خروجی‌ها از اهمیت ویژه‌ای برخوردار می‌باشد. در این راستا، پارامترهایی که در تمامی آزمایش‌ها 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 ضد امولسیون، ضد لخته و کاهنده کشش سطحی (سورفکتانت)، غلظت هیدروکلریک اسید، ویسکوزیته و API نفت مورد استفاده و نیز غلظت یون آهن فریک به‌عنوان پارامترهایی که در نتیجه آزمایش‌های ضد امولسیون تأثیر دارند به‌عنوان ورودی‌های مجموعه داده جمع‌آوری و ثبت گردیدند.

شکل ۱ تعداد آزمایش‌های انجام شده (فراوانی) با هر یک از مقادیر مختلف این ویژگی‌ها را به‌صورت نمودار هیستوگرامی نشان می‌دهد. با وجود اینکه رسیدن به‌دقت بالا در به‌کارگیری 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 آزمایش‌های انجام شده در محدوده مناسب آزمایشگاهی گزارش شده است و ویژگی‌های این مجموعه داده شامل تمامی مقادیر عددی مورد استفاده در آزمایشگاه اسیدزنی می‌باشند.



شکل ۱ توزیع داده‌های ورودی آزمایش‌های استاتیک ضد امولسیون

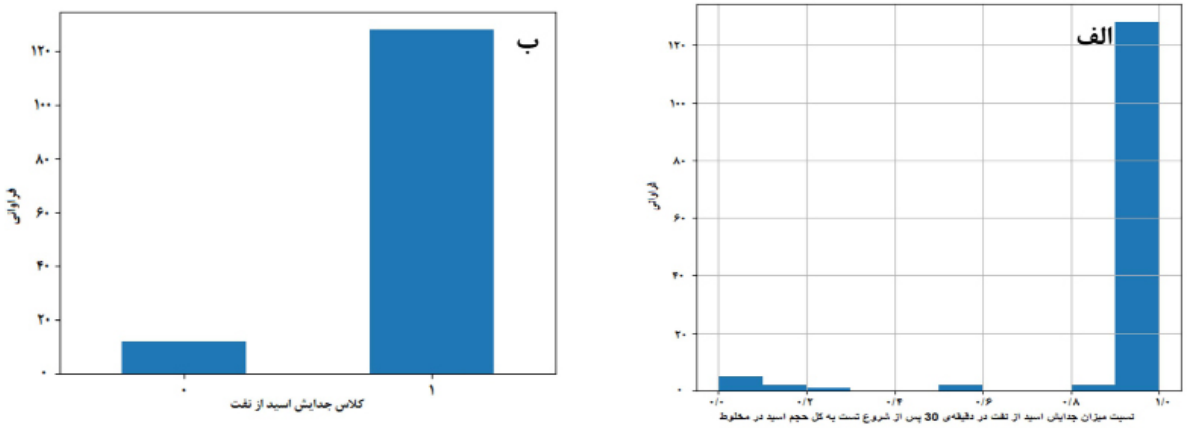
در عملیات‌های اسیدکاری واقعی در میادین نفتی حد بحرانی جدایش اسید و نفت برای تعیین امکان‌پذیر بودن و یا نبودن یک عملیات اسیدکاری، جدایش حداقل ۹۰٪ اسید از نفت تا ۳۰ min پس از تشکیل امولسیون می‌باشد و درصورتی‌که آزمایشی به‌این آستانه برسد انجام عملیات در میدان امکان‌پذیر می‌باشد. بنابراین در این پژوهش درصد جدایش اسید از نفت در مدت زمان ۳۰ min (نسبت حجم اسید جدا شده از نفت در ۳۰ min پس از شروع آزمایش به حجم کل اسید موجود در مخلوط اسید و نفت) به‌عنوان خروجی آزمایش‌ها و نیز مجموعه داده تعریف شده‌است. این مقادیر که اعدادی در بازه صفر تا یک بودند نهایتاً به دو کلاس "جدایش نامناسب" (مقادیر صفر تا ۰/۹) و "جدایش نامناسب" (مقادیر ۰/۹ تا یک) تقسیم‌بندی گردید. بدین ترتیب این مسئله به یک مسئله 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 به‌طوری‌که کلاس "جدایش مناسب " دارای ۱۲۵ نمونه داده آزمایشگاهی و کلاس "جدایش نامناسب" دارای ۱۵ نمونه داده می‌باشد؛ به‌عبارت دیگر مجموعه داده در دسترس یک مجموعه نامتوازن می‌باشد. این عدم توازن به‌علت وجود تجربه کافی در انجام آزمایش‌های 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 آزمایش‌ها شده‌است.

روش تحقیق

در این مطالعه در وهله اول فرآیند جمع‌آوری، آماده‌سازی و تهیه یک مجموعه داده آزمایشگاهی مرتبط با آزمایش استاتیک اسیدکاری شامل اطلاعات مربوط به پارامترهای ورودی و خروجی آزمایش‌های 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.

پیش‌پردازش داده‌ها[1]

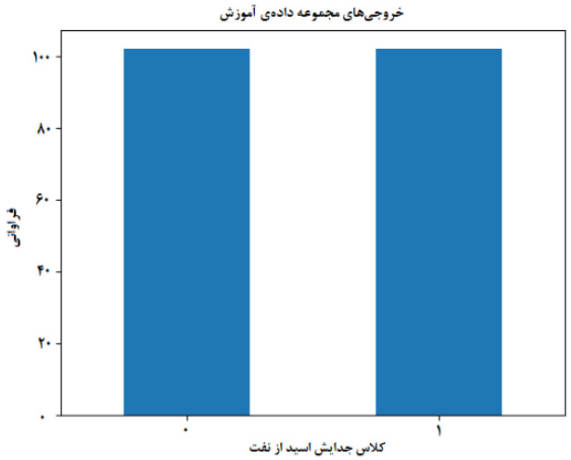
پس از آماده‌سازی مجموعه داده به‌صورت یک فایل اکسل و فراخوانی آن در محیط پایتون، روش‌های مختلف پیش‌پردازش داده‌ها همچون حذف داده‌های پرت[2]، داده‌های تکراری برروی مجموعه داده اعمال شدند.



شکل ۲ توزیع خروجی‌های درصد جدایش آزمایش‌های استاتیک ضد امولسیون (الف) و توزیع خروجی‌های مجموعه داده آزمایش‌های استاتیک ضد امولسیون به‌صورت دو کلاس «جدایش نامناسب» (کلاس ۰) و «جدایش مناسب» (کلاس ۱) (ب)

1. Data Preprocessing
2. Outlier

این مرحله جهت آماده‌سازی مجموعه داده جهت آموزش مدل‌های هوش مصنوعی با بالاترین میزان عملکرد انجام شد. پس از اجرای این مرحله، داده‌ها به‌صورت تصادفی (با امکان تکرارپذیری) به دو مجموعه آموزش و آزمایش با نسبت ۴ به ۱ تقسیم شدند. این تقسیم‌بندی به‌صورتی انجام شد که کلاس‌های خروجی در هر دو مجموعه توزیع یکسانی داشته باشند و مجموعه آموزش نمایندهٔ مناسبی از کل مجموعه داده و نیز مجموعه داده‌های آزمایش باشد. پس از آن ویژگی‌های هر دو مجموعه داده را به‌کمک 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 ماشین‌بردار پشتیبان نسبت به مقیاس ویژگی‌ها حساس بوده و ضرایب مربوط به ویژگی‌ها را براساس 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 براساس معیارهای مناسب برروی مجموعه دادهٔ آزمایش ارزیابی گردید. با توجه به ماهیت نامتوازن مجموعه داده واقعی در دسترس و تاثیر آن بر خروجی‌ها، 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 داده آموزش استفاده گردید. این روش از طریق درون‌یابی داده‌های واقعی، داده‌های مصنوعی می‌سازد و علاوه‌بر ایجاد توازن میان کلاس‌ها، باعث افزودن تعداد قابل توجهی نمونه داده به مجموعه داده اصلی و بهبود عملکرد مدل‌های یادگیری ماشین می‌گردد [۲۸ و ۲۹]. در این مسئله پس از اعمال بیش نمونه‌گیری مصنوعی برروی مجموعه داده‌ی آموزش، تعداد داده‌ها در هر یک از کلاس‌ها در این مجموعه داده به شرح شکل ۳ شده‌است.



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

مدل‌های طبقه‌بندی یادگیری ماشین

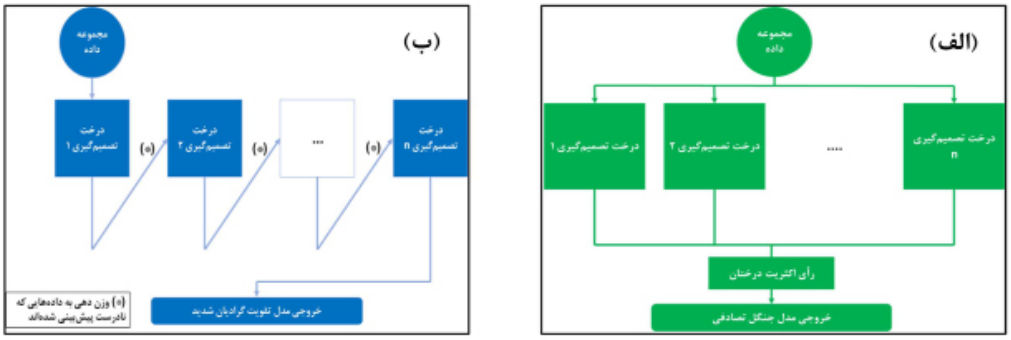
با توجه به نوع خروجی‌های مجموعه داده آزمایشگاهی که به‌صورت دو کلاس «جدایش مناسب» و «جدایش نامناسب» می‌باشد، در این پژوهش مدل‌های طبقه‌بندی یادگیری ماشین برروی داده‌های پیش‌پردازش شده اعمال شدند. در این راستا از چهار مدل یادگیری ماشین شامل مدل‌های جنگل تصادفی [2]، تقویت گرادیان شدید، پرسپترون چند لایه و ماشین بردار پشتیبان که نتایج دقیق‌تری نسبت به سایر روش‌های طبقه‌بندی دارند استفاده شده‌است. مدل‌های جنگل تصادفی و تقویت گرادیان شدید نوعی از مدل‌های یادگیری جمعی هستند که برپایه‌ی درختان تصمیم‌گیری کار می‌کنند. در مدل جنگل تصادفی تعداد معینی درخت تصمیم‌گیری به‌صورت مجزا و به موازات یکدیگر توسط داده‌ها آموزش می‌بینند و پس از آن خروجی یک نمونه داده جدید را براساس رأی اکثریت درختان آموزش‌دیده تعیین می‌کند.

1. Synthetic Minority Oversampling Technique (SMOTE)
2. Random Forest Classifier

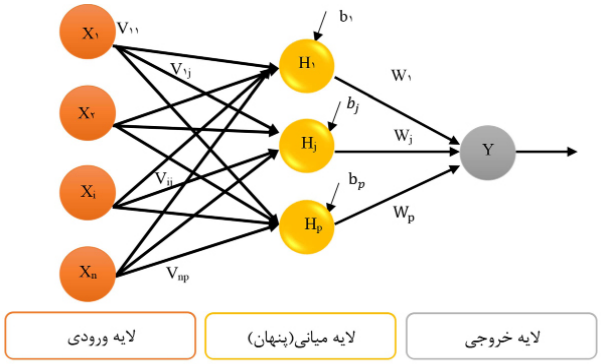
همچنین در روش تقویت گرادیان شدید، تعداد معینی از درختان تصمیم‌گیری یکی پس از دیگری آموزش می‌بینند به‌طوری‌که در این فرآیند وزن اهمیت کلاسی که نمونه‌های آن توسط درخت قبلی نادرست پیش‌بینی شده بودند افزایش می‌یابد تا بدین ترتیب درخت جدیدتر قادر به پیش‌بینی خروجی مربوط به نمونه‌های این کلاس باشد (شکل ۴). مدل پرسپترون چند لایه ساختاری شبیه به سیستم عصبی انسان دارد به‌طوری‌که این شبکه از لایه‌های مختلفی شامل یک لایه‌ی ورودی، یک یا چند لایه پنهان و یک لایه‌ی خروجی تشکیل می‌شود. هر یک از این لایه‌ها از تعدادی گره یا نرون ساخته شده‌اند که وظیفه اعمال عملیات ریاضی برروی 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 ماشین‌بردار 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.

معیارهای ارزیابی مدل‌های طبقه‌بندی

با توجه به نوع مسئله طبقه‌بندی، در این پژوهش از معیارهای دقت، عدد f1 [1]، کوهن-کاپا [2] و ضریب رابطه متیو برای ارزیابی و مقایسهٔ مدل‌های آموزش دیده استفاده شده‌است [۳۵]. این معیارها با استفاده از ماتریس کانفیوژن ارائه شده در شکل ۶ و روابط ارائه شده 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).



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



شکل ۵ طرحواره ساختار شبکه عصبی مصنوعی پرسپترون چند لایه

1. Score
2. Cohen-Kappa

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

شکل ۶ ماتریس در هم‌ریختگی برای یک مسئله دودویی

• دقت: معیاری است که نسبت تعداد نمونه‌هایی را که به‌درستی پیش‌بینی شده‌اند به کل نمونه‌ها را نشان می‌دهند. این معیار در مسائل طبقه‌بندی که در آنها مجموعه داده یک مجموعه داده 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.

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

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

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

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

$$f_1 = (\text{یادآوری}^{*}\text{درستی})/(\text{یادآوری}+\text{درستی})^{*2} \quad (۴)$$

• 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 بازه 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 به‌عنوان 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 = (مثبت واقعی، مثبت پیش‌بینی شده)

(۶) احتمال مثبت پیش‌بینی شده × احتمال مبت واقعی

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

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

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

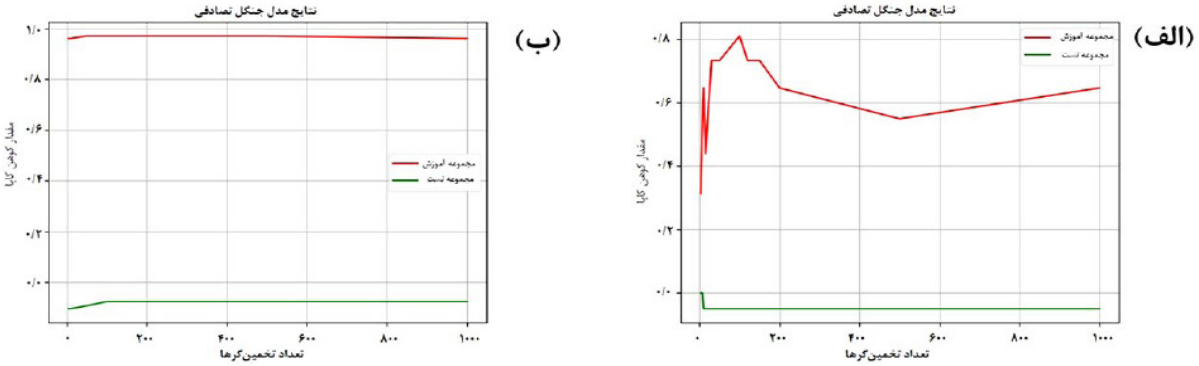
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (۹)$$

بحث و نتایج

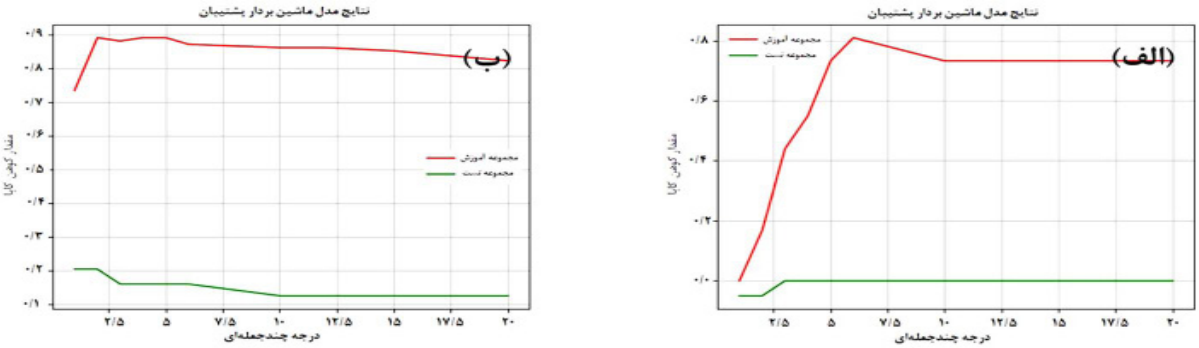
پس از آماده‌سازی داده‌ها، چهار مدل مختلف یادگیری ماشین شامل جنگل تصادفی، ماشین‌بردار پشتیبان، پرسپترون چند لایه و تقویت گرادیان شدید با مجموعه داده آموزش، آموزش دیدند. جهت انتخاب دقیق‌ترین مدل، مؤثرترین پارامتر هر یک از مدل‌های یادگیری ماشین که تغییر آنها تأثیر بسیار زیادی بر دقت نهایی مدل دارد انتخاب 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 جنگل تصادفی، پارامترهای هسته[1] و درجه 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 به‌عنوان 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 جمله‌ای) به‌کار می‌رود.

1. Kernel

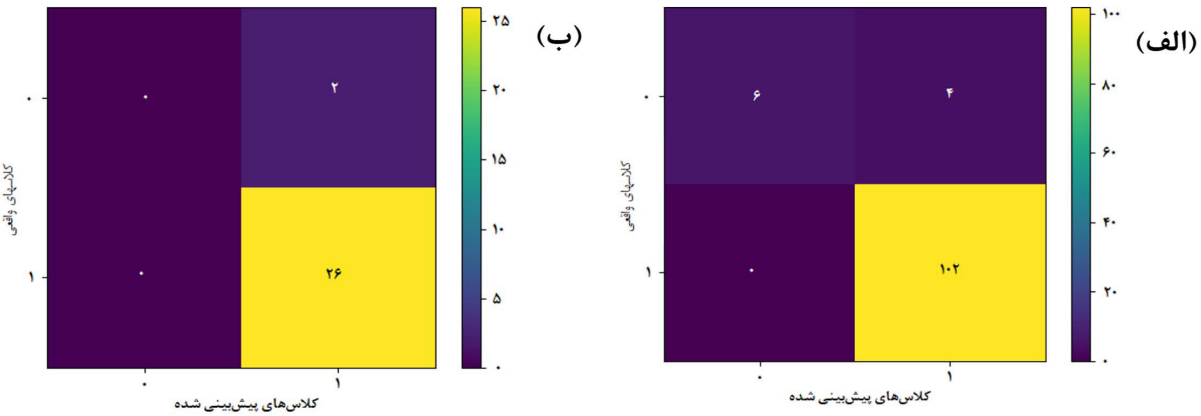
در این پژوهش، هر یک از مدل‌های مذکور در دو مرحله شامل آموزش با داده‌های صرفاً واقعی و آموزش با داده‌های واقعی و مصنوعی آموزش دیدند و براساس معیار کوهن-کاپا ارزیابی شدند. نتایج ارائه شده در شکل‌های ۷، ۸، ۹، ۱۰، ۱۱ و ۱۲ به‌ترتیب نشان‌دهنده نتایج اعمال مدل‌های جنگل تصادفی، ماشین بردار پشتیبان، تقویت گرادیان شدید و پرسپترون چند لایه با پارامترهای مختلف برروی مجموعه داده‌های آموزش می‌باشند.



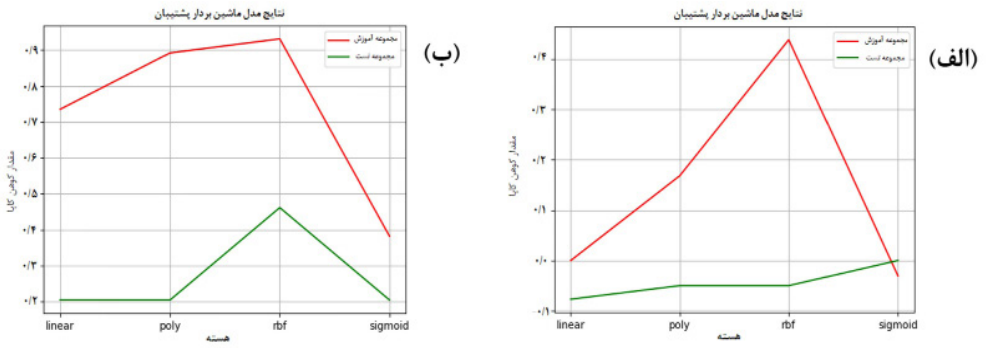
شکل ۷ نتایج اعمال روش جنگل تصادفی برروی داده‌های آموزش و آزمایش زمانیکه تنها با داده‌های واقعی آموزش دیده‌اند (الف) و زمانی‌که با داده‌های واقعی و مصنوعی آموزش دیده‌اند (ب)



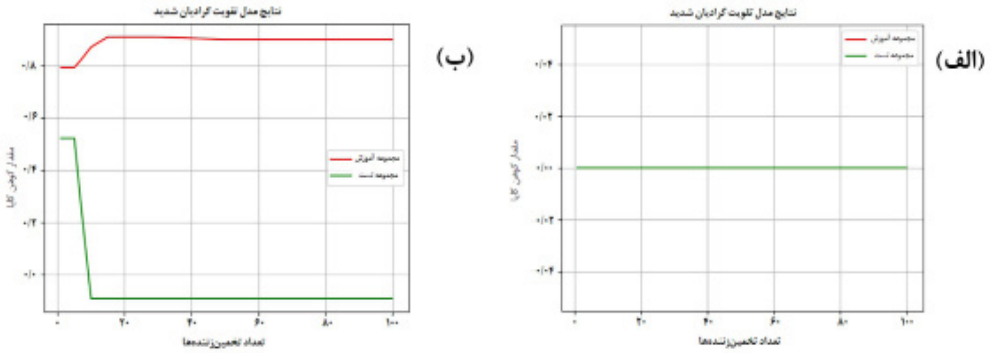
شکل ۸ نتایج اعمال روش ماشین بردار پشتیبان با توابع از درجات مختلف برروی داده‌های آموزش و آزمایش زمانی‌که تنها با داده‌های واقعی آموزش دیده (الف) و زمانی‌که با داده‌های واقعی و مصنوعی آموزش دیده (ب)



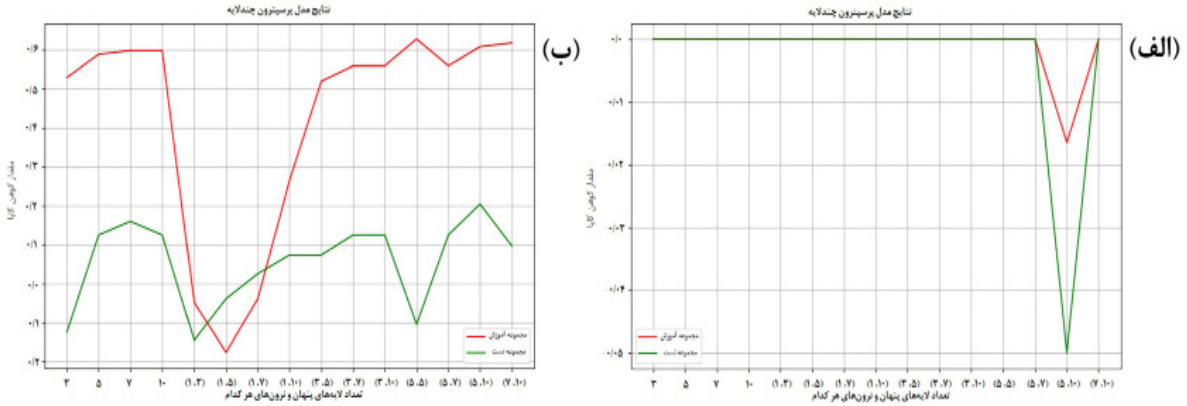
شکل ۹ ماتریس درهم‌ریختگی عملکرد مدل ماشین بردار پشتیبان با چندجمله‌ای درجه ۱۰ برروی داده‌های آموزش متشکل از داده‌های صرفاً واقعی (الف) و داده‌های آزمایش (ب)



شکل ۱۰ نتایج اعمال روش ماشین بردار پشتیبان با هسته‌های مختلف برروی داده‌های آموزش و آزمایش زمانی‌که تنها با داده‌های واقعی آموزش دیده (الف) و زمانی‌که با داده‌های واقعی و مصنوعی آموزش دیده (ب)



شکل ۱۱ نتایج اعمال روش تقویت گرادیان شدید برروی داده‌های آموزش و آزمایش زمانی‌که تنها با داده‌های واقعی آموزش دیده (الف) و زمانی‌که با داده‌های واقعی و مصنوعی آموزش دیده (ب)



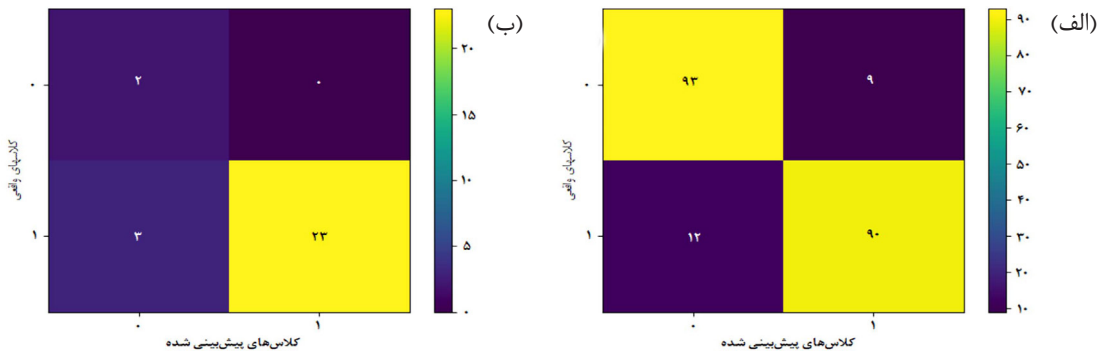
شکل ۱۲ نتایج اعمال روش پرسپترون چند لایه برروی داده‌های آموزش و آزمایش زمانی‌که تنها با داده‌های واقعی آموزش دیده (الف) و زمانیکه با داده‌های واقعی و مصنوعی آموزش دیده (ب)

بخش الف در شکل ۷ نشان می‌دهد که با افزایش تعداد درختان تصمیم‌گیری در مدل جنگل تصادفی، معیار کوهن-کاپا برروی 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 برروی داده‌های آزمایش با افزایش تعداد درختان تصمیم‌گیری از مقدار اولیه صفر به مقادیر کمتر کاهش می‌یابد.

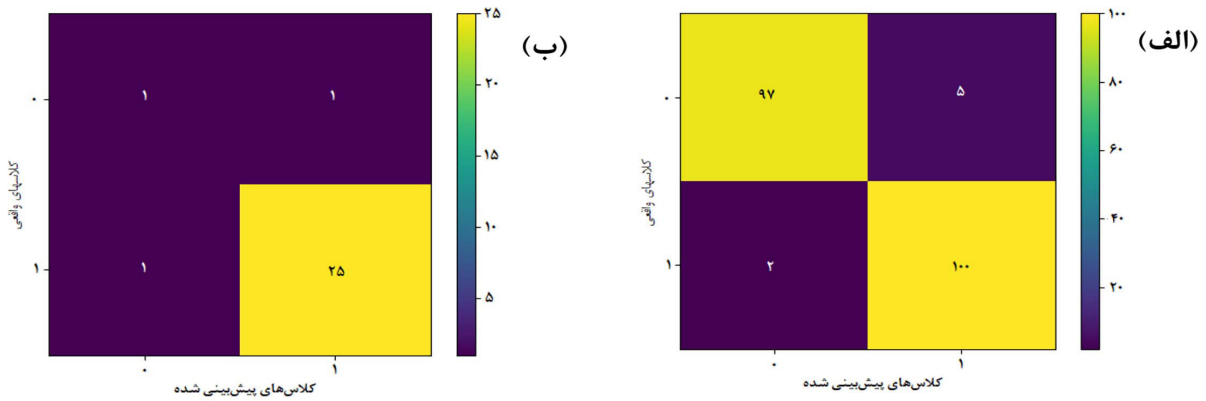
این موضوع نشان‌دهنده ضعیف‌تر شدن عملکرد مدل‌های آموزش دیده برروی داده‌های آزمایش بر اثر برازش بیش از حد[1] مدل برروی داده‌های آموزش می‌باشد. با توجه به این موضوع که مهم‌ترین معیار ارزیابی یک مدل یادگیری ماشین میزان عملکرد آن برروی داده‌های آزمایش است که آستانه آن نیز معیار کوهن-کاپای بالاتر از ۰/۸ در 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 آزمایش‌های آزمایشگاهی انجام نشده را که داده‌های آن‌ها را تاکنون ندیده است، پیش‌بینی کند. همچنین نتایج ارائه شده در بخش (الف) شکل‌های ۸، ۱۰، ۱۱ و ۱۲ نشان می‌دهند که تمامی مدل‌های آموزش دیده معیار کوهن-کاپا و عملکرد نسبتاً خوبی برروی داده‌های آموزش دارند اما هیچ یک از آن‌ها عملکرد مناسبی برروی داده‌های آزمایش ندارند و این موضوع از مقدار کوهن-کاپای منفی و صفر برروی داده‌های آزمایش 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 شده درست به کل داده‌ها)، تمامی داده‌ها در کلاس اقلیت را به‌عنوان نمونه‌هایی از کلاس اکثریت پیش‌بینی می‌کند. زیرا داده‌هایی در کلاس اکثریت هستند و در نهایت به‌درستی پیش‌بینی می‌شوند، درصد بالایی از کل داده‌ها را شکل می‌دهند. مشکل مجموعه داده نامتعادل یکی از اساسی‌ترین مشکلات مسائل طبقه‌بندی در حوزه هوش مصنوعی می‌باشد که عدم کنترل آن و نیز ارزیابی مدل‌ها براساس معیار دقت می‌تواند منجر به خطای بسیار بزرگ و ارزیابی غیرقابل اعتمادی گردد. افزون بر معیار کوهن-کاپا، ماتریس درهم‌ریختگی مدل‌های آموزش دیده نشان‌دهنده عملکرد بسیار ضعیف و عدم توانایی آن‌ها در شناسایی و پیش‌بینی نمونه داده از کلاس اقلیت می‌باشند. ماتریس درهم‌ریختگی مدل ماشین‌بردار پشتیبان با چندجمله‌ای درجه ۱۰ به‌عنوان نمونه در شکل ۹ ارائه شده‌است. همان‌طور که در شکل ۹، ب مشاهده می‌شود، مدل آموزش دیده ماشین‌بردار پشتیبان با هسته چندجمله‌ای با درجه ۱۰ تمامی داده‌ها را به‌عنوان نمونه از کلاس اکثریت (جدایش مناسب) پیش‌بینی کرده و قادر به پیش‌بینی هیچ نمونه‌ای از داده‌های کلاس اقلیت نبوده است. معیار کوهن-کاپای صفر برروی داده‌های آزمایش در بخش الف شکل ۸ برای مدل ماشین بردار پشتیبان با چندجمله ای درجه ۱۰ تأئیدی بر این موضوع است. بنا به عملکرد نامناسب مدل‌های آموزش دیده برروی مجموعه داده آموزش شامل داده‌های صرفاً واقعی (به داده‌های آزمایشگاهی)، داده‌های مصنوعی جدید با استفاده از روش بیش نمونه‌گیری مصنوعی ساخته‌شده و به داده‌های آموزش قبلی اضافه گردیدند. این کار صرفاً برروی داده‌های آموزش اعمال شده‌است تا ارزیابی مدل‌های آموزش دیده با این مجموعه داده‌ی جدید نیز با استفاده از همان داده‌های آزمایش قبلی (داده‌های صرفاً آزمایشگاهی) صورت گرفته و نتایج قابل اعتمادتر باشد. نتایج ارزیابی مدل‌های مختلف آموزش دیده با داده‌های آموزش واقعی و مصنوعی براساس معیار کوهن-کاپا در بخش (ب) شکل‌های ۷، ۸، ۱۰، ۱۱ و ۱۲ ارائه شده‌است. نتایج ارائه شده نشان می‌دهند که عملکرد تمامی مدل‌ها برروی داده‌های آزمایش بسیار بهبود یافته است.

1. Over-Fitting

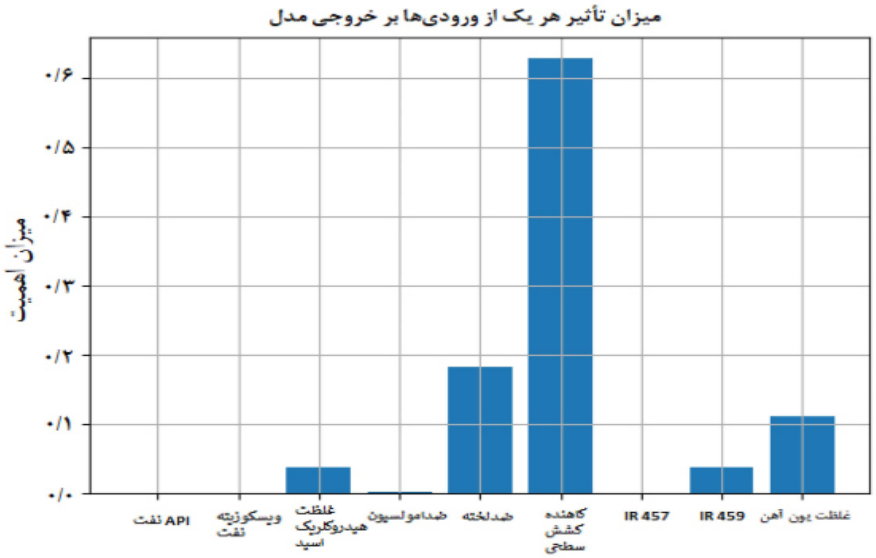
به‌طوری‌که مقدار عددی معیار کوهن-کاپا برروی داده‌های آزمایش مقادیر بیش از صفر می‌باشد. این بدان سبب است که مجموعه داده‌ی آموزش در این مرحله متعادل بوده و تعداد داده‌های موجود در هر یک از کلاس‌های «جدایش نامناسب» و «جدایش مناسب» یکی می‌باشد. در این میان، مدل تقویت گرادیان شدید با ۵ تخمین زننده، با دارا بودن بالاترین مقادیر کوهن-کاپا به میزان ۰/۷۹ برروی داده‌های آموزش و ۰/۵۲۳ برروی داده‌های آزمایش 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 به‌ترتیب برروی داده‌های آموزش و آزمایش عملکرد مناسبی در پیش‌بینی خروجی‌های مجموعه داده دارد (شکل ۱۰، ب). شکل ۱۱، ب نشان‌دهنده این موضوع است که با افزایش تعداد درختان تصمیم‌گیری در مدل تقویت گرادیان شدید، عملکرد مدل برروی داده‌های آموزش افزایش می‌یابد، حال آنکه دقت مدل در پیش‌بینی خروجی‌های داده‌های آزمایش کاهش می‌یابد. زیرا با افزایش تعداد درختان تصمیم‌گیری، درختان در جهت برازش بیش از حد برروی داده‌های آموزش می‌شوند و بدین ترتیب صرفاً می‌توانند عملکرد مناسبی برروی داده‌هایی که با آنها آموزش دیده‌اند داشته باشند و قادر به پیش‌بینی خروجی داده‌های جدید و دیده نشده 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 برروی داده‌های آموزش و هم برروی داده‌های آزمایش دارد. شکل‌های ۱۳ و ۱۴ به‌ترتیب ماتریس درهم‌ریختگی مدل‌های تقویت گرادیان شدید با ۵ تخمین زننده و مدل ماشین بردار پشتیبان با هسته rbf را که توسط مجموعه داده آموزش شامل داده‌های واقعی و مصنوعی آموزش‌دیده و توسط داده‌های آزمایش واقعی مورد ارزیابی قرار گرفته است را نشان می‌دهند. مقادیر عددی غیرصفر در قطر اصلی ماتریس درهم‌ریختگی در شکل‌های ۱۳، ب و 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 آزمایش را دارا هستند. 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 به‌دست آمد. مطابق با نتایج به‌دست‌آمده در شکل ۱۵، میزان غلظت افزایه کاهنده کشش سطحی، ضد لخته و غلظت یون آهن بیشترین میزان تأثیر را برروی نتیجه آزمایش‌های ضد امولسیون در آزمایشگاه اسیدزنی دارند.



شکل ۱۳ ماتریس درهم‌ریختگی عملکرد مدل تقویت گرادیان شدید با ۵ تخمین‌زننده برروی داده‌های آموزش متشکل از داده‌های واقعی و مصنوعی (الف) و داده‌های آزمایش (ب)



شکل ۱۴ ماتریس درهم‌ریختگی عملکرد مدل ماشین بردار پشتیبان با هسته «rbf» برروی داده‌های آموزش متشکل از داده‌های واقعی و مصنوعی (الف) و داده‌های آزمایش (ب)



شکل ۱۵ آنالیز حساسیت سنجی میزان تأثیر هر یک از ورودی‌ها بر خروجی مدل (نتیجه آزمایش ضد امولسیون)

با توجه به این مسئله که امولسیون اسید در نفت به‌دلیل پایه آبی بودن اسید و وجود سورفکتانت‌های طبیعی در نفت به‌وجود 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 امولسیون 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.

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 آزمایش 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

بتوانـد نتایـج آزمایش‌هـای آزمایشـگاهی را براسـاس داده‌هـای ورودی مربـوط بـه اطلاعـات نفـت، اسـید و افزایه‌هـای مورداسـتفاده، پیش‌بینـی کنـد. ایـن کار بـرای بهینـه کـردن تعـداد آزمایش‌هـای موردنیـاز در آزمایشـگاه اسـتاتیک ضـد امولسـیون در حوزه اسـیدکاری صـورت گرفتـه اسـت کـه در ادامـه خلاصـه‌ای از نتایـج حاصـل شـده، ارائـه می‌شـود:

• گزارش‌هـای آزمایشـگاهی از آزمایش‌هـای ضـد امولسـیون بـرای عملیـات اسـیدکاری بیانگـر ایـن موضوع بـوده اسـت کـه داده‌هـای ایـن مجموعـه از آزمایش‌هـا دارای تعـداد متنوعـی از پارامترهـای ورودی می‌باشـند. بـا توجـه بـه اهمیـت اسـتفاده از تعـداد ورودی‌هـای بهینـه در مدل‌هـای یادگیـری ماشـین جهـت جلوگیـری از ایجـاد بـار محاسـباتی مـازاد، مؤثرتریـن ویژگی‌هـا بـر خروجـی آزمایش‌هـای ضـد امولسـیون شـامل داده‌هـای مربـوط بـه غلظـت هیدروکلریـک اسـید و افزایه‌هـای تزریقـی همچـون افزایه‌هـای ضـد امولسـیون، ضـد لختـه، کاهنـده کشـش سـطحی و کاهنـده یـون آهـن، ویژگی‌هـای نفـت همچـون گرانـروی و دانسـیته و نیـز غلظـت یـون فریـک به‌عنـوان ورودی‌هـای مدل‌هـای یادگیـری ماشـین انتخـاب و مـورد اسـتفاده قـرار گرفتند.

• در ایـن کار تحقیقاتـی، بـا توجـه بـه مشـکلات مربوط به بانـک اطلاعاتـی در دسـترس (همچـون تعـداد داده‌هـای کـم و نیـز عـدم تـوازن مجموعـه داده)، هزینـه سـنگین و زمانبـر بـودن آزمایش‌هـای مـورد نیـاز بـرای تکمیـل بانـک اطلاعاتـی از روش بیـش نمونه‌گیـری مصنوعـی (SMOTE) اسـتفاده گردیـد. بهبـود عملکـرد مدل‌هـای طبقه‌بنـدی بـا اجـرای ایـن روش، نشـان‌دهنده اهمیـت روش بیـش نمونه‌گیـری مصنوعـی در بهبـود مجموعـه بانـک اطلاعـات داده‌هـای سـایر کارهـای مختلـف در صنعـت نفـت می‌باشـد.

• بـا توجـه بـه نـوع مسـئله در ایـن پژوهـش و نیـز وجـود مشـکل عـدم تـوازن در مجموعـه داده، از معیـار کوهـن-کاپـا جهـت تعییـن بهتریـن مـدل اسـتفاده شده‌اسـت. پـس از آمـوزش و ارزیابـی مدل‌هـای جنـگل تصادفـی، پرسـپترون چندلایـه، ماشـین بـردار پشـتیبان و مـدل تقویـت گرادیـان شـدید، مـدل تقویـت گرادیـان شـدید بـا ۵ تخمین‌زننـده و آمـوزش دیـده بـا داده‌هـای آمـوزش ترکیبـی (شـامل داده‌هـای آزمایشـگاهی و داده‌هـای سـاخته شـده بـا روش بیـش نمونه‌گیـری مصنوعـی) به‌عنـوان بهتریـن مـدل آمـوزش دیـده در ایـن مسـئله انتخـاب شـد. مقادیـر کوهن-کاپـا در ایـن مـدل ۰/۷۹ بـرروی داده‌هـای آمـوزش و ۰/۵۲۳ بـرروی داده‌هـای آزمایـش می‌باشـد.

بـا توجـه بـه اهمیـت هوشمندسـازی عملیـات اسـیدزنی چاه‌هـای نفـت و گاز، امـکان اسـتفاده از ایـن مـدل بـرروی سـایر آزمایش‌هـای اسـتاتیک وجـود دارد امـا مطابـق بـا نتایـج به‌دسـت آمـده، دقـت مـدل به‌دسـت آمـده‌ی نهایـی بدلیـل کـم و محـدود بـودن داده‌هـای در دسـترس بسـیار زیـاد نیسـت. بنابرایـن توصیـه میشـود در پژوهش‌هـای آتـی، بـا رفـع مشـکل محدودیـت در دسترسـی بـه داده‌هـای کافـی، غنی‌سـازی بانـک اطلاعاتـی بـا اطلاعـات بومـی و قابـل اطمینـان و بـالا بـردن دقـت و عملکـرد مـدل از کیفیـت نتایـج اطمینـان حاصـل کـرد و سـپس از آن بـرای پیش‌بینـی نتیجـه‌ی سـایر آزمایش‌هـای اسـتاتیک در آزمایشـگاه‌های اسـیدزنی کشـور اسـتفاده نمـود.

تشکر و قدردانی

از شـرکت ملـی نفـت ایـران بخـش مدیریـت اکتشـاف جهـت همـکاری و تأمیـن اطلاعـات لازم بـرای انجـام ایـن پژوهـش نهایـت قدردانـی و سـپاس‌گزاری به‌عمـل می‌آیـد. همچنیـن از آقایـان مهنـدس روشـنی و موسـوی جهـت یـاری رسـاندن در جمع‌آوری داده‌هـای آزمایشـگاهی تشـکر و قدردانـی می‌گـردد.

مراجع

[1]. Economides, M. J., & Nolte, K. G. (1989). Reservoir stimulation. 2, 6-10. Englewood Cliffs, NJ: Prentice Hall.
[2]. Dai C. and Zhao F. (2019). Oilfield Chemistry, Oilfield, Chemistry 1–395, , doi: 10.1007/978-981-13-2950-0.
[3]. Shirazi, M. M., Ayatollahi, S., & Ghotbi, C. (2019). Damage evaluation of acid-oil emulsion and asphaltic sludge formation caused by acidizing of asphaltenic oil reservoir. Journal of Petroleum Science and Engineering, 174, 880-890. doi: 10.1016/j.petrol.2018.11.051.
[4]. Ganeeva, Y. M., Yusupova, T. N., Barskaya, E. E., Valiullova, A. K., Okhotnikova, E. S., Morozov, V. I., & Davletshina, L. F. (2020). The composition of acid/oil interface in acid oil emulsions. Petroleum Science, 17, 1345-1355. doi.org/10.1007/s12182-020-00447-9.
[5]. Pourakaberian, A., Ayatollahi, S., Shirazi, M. M., Ghotbi, C., & Sisakhti, H. (2021). A systematic study of asphaltic sludge and emulsion formation damage during acidizing process: Experimental and modeling approach. Journal of Petroleum Science and Engineering, 207, 109073, doi: 10.1016/j.petrol.2021.109073.
[6]. Shakouri, S., & Mohammadzadeh-Shirazi, M. (2023). Modeling of asphaltic sludge formation during acidizing process of oil well reservoir using machine learning methods. Energy, 285, 129433, doi: 10.1016/j.energy.2023.129433.
[7]. Chavanne, C., & Perthuis, H. G. (1992, November). A Fluid selection expert system for matrix treatments. In SPE Europec featured at EAGE Conference and Exhibition? (pp. SPE-24995). SPE. doi: 10.2118/24995-MS.
[8]. Ebrahim, A. S., Garrouch, A. A., & Lababidi, H. M. (2014). Automating sandstone acidizing using a rule-based system. Journal of Petroleum Exploration and Production Technology, 4, 381-396. doi: 10.1007/s13202-014-0104-3.
[9].Koroteev, D., & Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. Energy and AI, 3, 100041, doi: 10.1016/j.egyai.2020.100041.
[10]. Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research, 6(4), 379-391, doi: 10.1016/j.ptlrs.2021.05.009.
[11]. Mohaghegh, S. D. (2005). Recent developments in application of artificial intelligence in petroleum engineering. Journal of Petroleum Technology, 57(04), 86-91.doi.org/10.2118/89033-JPT.
[12]. Alkinani, H. H., Al-Hameedi, A. T., Dunn-Norman, S., Flori, R. E., Alsaba, M. T., & Amer, A. S. (2019). Applications of artificial neural networks in the petroleum industry. A review. In SPE Middle East oil and gas show and conference (p. D032S063R002). SPE.doi.org/10.2118/195072-MS.
[13]. Mohaghegh, S. (2000). Virtual-intelligence applications in petroleum engineering: Part 1—Artificial neural networks. Journal of Petroleum Technology, 52(09), 64-73. doi.org/10.2118/58046-JPT.
[14]. Mohaghegh, S., Arefi, R., Ameri, S., Aminiand, K., & Nutter, R. (1996). Petroleum reservoir characterization with the aid of artificial neural networks. Journal of Petroleum Science and Engineering, 16(4), 263-274.doi.org/10.1016/S0920-4105(96)00028-9.
[15]. Ebrahim, A. S., Garrouch, A. A., & Lababidi, H. M. (2014). Automating sandstone acidizing using a rule-based system. Journal of Petroleum Exploration and Production Technology, 4, 381-396, doi: 10.1007/s13202-014-0104-3.

[۱۶]. خیرالهی، ح.، چهاردولی، م. و سیم‌جو م. (۱۴۰۳). انتخاب بهترین ناحیه پیادهسازی پایلوت برای روش‌های تزریق آب پایه با استفاده از الگوریتم‌های تصمیم‌گیری چند شاخصه، پژوهش نفت، ۳۴(۵)، ۱۹-۳. doi: 10.22078/pr.2024.5315.3361.

[۱۷]. خیرالهی، ح.، زاید، م.، سبحانی، ص.، چهاردولی، م. و سیم‌جو، م. (۱۴۰۲). غربال‌گری روش‌های ازدیادبرداشت از مخازن نفتی با استفاده از تلفیق روش‌های هوش‌مصنوعی، پژوهش نفت، ۳۳(۵)، ۶۲-۵۱. doi: 10.22078/pr.2023.5151.3284.

[۱۸]. کریمی، ع. و صادق‌نژاد، س. (۱۴۰۱). بازسازی تصویر سنگ مخزن متراکم با شبکه عصبی مولد رقابتی، پژوهش نفت، ۳۲(۵)، ۹۴-۸۳. doi: 10.22078/pr.2022.4843.3165.

[۱۹]. رجبی هشتجین، م. و جعفری‌بهبهانی، ت. (۱۳۹۶). بهبود مدل پیش‌بینی رفتار رئولوژیکی سیال حفاری با استفاده از شبکه عصبی، پژوهش نفت، ۲۷(۶)، ۵۸-۴۶. doi: 10.22078/pr.2017.2619.2211.

[20]. Sumotarto, U., Hill, A. D., & Sepehrnoori, K. (1995, October). An integrated sandstone acidizing fluid selection and simulation to optimize treatment design. In SPE Annual Technical Conference and Exhibition? (pp. SPE-30520). SPE., doi: 10.2118/30520-ms.
[21]. Kellogg, R. P., Chessum, W., & Kwong, R. (2018, April). Machine learning application for wellbore damage

removal in the wilmington field. In SPE Western Regional Meeting (p. D041S011R001). SPE., doi: 10.2118/190037-ms.
[22]. Sidaoui, Z., Abdulraheem, A., & Abbad, M. (2018, April). Prediction of optimum injection rate for carbonate acidizing using machine learning. In SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition (pp. SPE-192344). SPE., doi: 10.2118/192344-ms.
[23]. Xue, H., Liu, P. L., Li, N. Y., Luo, Z. F., & Zhao, L. Q. (2012). Expert system for acidizing based on BP neural network. Advanced Materials Research, 548, 438-443, doi: 10.4028/www.scientific.net/AMR.548.438.
[24]. Van Domelen, M. S., Ford, W. G. F., & Chiu, T. J. (1992, October). An expert system for matrix acidizing treatment design. In SPE Annual Technical Conference and Exhibition? (pp. SPE-24779). SPE. doi: 10.2523/24779-ms.
[25]. Alkathim, M., Aljawad, M. S., Hassan, A., Alarifi, S. A., & Mahmoud, M. (2023). A data-driven model to estimate the pore volume to breakthrough for carbonate acidizing. Journal of Petroleum Exploration and Production Technology, 13(8), 1789-1806, doi: 10.1007/s13202-023-01642-1.
[26]. Hassan, A., Aljawad, M. S., & Mahmoud, M. (2021). An artificial intelligence-based model for performance prediction of acid fracturing in naturally fractured reservoirs. ACS Omega, 6(21), 13654-13670.doi: 10.1021/acsomega.1c00809.
[27]. Dargi, M., Khamehchi, E., & Mahdavi Kalatehno, J. (2023). Optimizing acidizing design and effectiveness assessment with machine learning for predicting post-acidizing permeability. Scientific Reports, 13(1), 11851, doi: 10.1038/s41598-023-39156-9.
[28]. 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.
[29]. Kurniawan, C., Azis, M. M., & Ariyanto, T. (2023). Supervised machine learning and multiple regression approach to predict successfulness of matrix acidizing in hydraulic fractured sandstone formation. ASEAN Journal of Chemical Engineering, 23(1), 113-127.doi.org/10.22146/ajche.78255.
[30]. Blackburn, C. R., Abel, J. C., & Day, R. (1990). An expert system to design and evaluate matrix acidizing. SPE Computer Applications, 2(06), 15-17, doi: https://doi.org/10.2118/20337-PA.
[31]. Chiu, T. J., Caudell, E. A., & Wu, F. L. (1993). Development of an expert system to assist with complex fluid design. SPE Computer Applications, 5(01), 18-20. doi.org/10.2118/24416-PA.
[32]. Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21(8), 1509-1515. doi: 10.1093/bioinformatics/bti171.
[33]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of artificial intelligence research, 16, 321-357. doi: 10.1613/jair.953.
[34]. Blagus, R. and Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics, 14. 2013. doi: 10.1186/1471-2105-14-106.
[35]. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data, 7, 1-47, doi: 10.1186/s40537-020-00349-y.

پژوهش نفت شماره ۱۳۹، بهمن و اسفند ۱۴۰۳، صفحه ۹۱-۷۳

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’
	درجه 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]