Shahid Chamran University of Ahvaz
OPEN ACCESS
Adv. Appl. Geol. Spring 2024, Vol. 14(1): 167-185

Research Article

Using ensemble learning techniques based on feature selection to predict shear wave velocity

Abedin Hashemzade Kalvari, Iman Zahmatkesh*

Department of Geology, Faculty of Earth Science Shahid Chamran University of Ahvaz

Keywords: *Ensemble learning, Asmari reservoir, Mansouri oil field, Shear wave velocity, Machine learning.*

1-Introduction

Petrophysical logs are the most convenient and accessible tools for obtaining information about physical properties of reservoirs (Anemangely et al., 2019). Among different logs, shear wave is one of the geomechanical parameters of the reservoir, which is considered as one of the most important elements in the identification and development of reservoirs. Several studies have been devoted to the use of intelligent systems to predict shear wave velocity using petrophysical data (Rezaee et al., 2007; Khandelwal et al., 2010; Rajabi et al., 2010; Maleki et al., 2014; Onalo et al., 2019.). In this study, the shear wave velocity was estimated by using ensemble learning methods in the Asmari reservoir of the Mansouri oilfield.

2-Methodology

The ensemble learning paradigm is a new computational intelligence algorithm, where multiple base learners are built and multiple results are integrated with a certain strategy as the final result. The advantage of ensemble learning is combining the predictions of several base estimators to improve generalizability and increase model accuracy (Gao et al., 2019). Firstly, in the process of data preparation, all data were quality controlled and out of trend data were removed. Bad hole intervals were recognized based on the caliper and bit size logs and their data were not included in the models. As the petrophysical data used in this study have different scales, prior to construction of the models, they were normalized. Then to make an accurate estimate, it is important to choose appropriate inputs for the model. Accordingly, the RFE method determined conventional well logs such as density log, neutron porosity, acoustic log, gamma ray, spectral gamma ray and depth as optimal inputs for training models. Optimal values for parameters are also obtained by the grid search cv method. Grid search is the most widely used method for learning the hyperparameter configuration space, which tests and tunes all the hyperparameters given to the grid configuration (Belete et al., 2021). Then, the selected features were considered as input parameters for ensemble methods, including, voting, stacking, bagging and boosting. Basic learners such as, support vector regression, linear regression and the nearest neighbor algorithm were used to build Voting and stacking models. Also, the mentioned basic models were trained and evaluated for comparison with ensemble methods. In order to make a comparison, hybrid methods including neural network-genetic algorithm (ANN-GA), neural network-particle swarm optimization (ANN-PSO), and adaptive neuro-fuzzy inference system (ANFIS), were compared with ensemble learning methods.

3-Results and discussion

The obtained results for quality measurements including Correlation Coefficient (R) and MSE values of training and testing data were computed for each model. Considering the testing phase, among all of the methods applied for shear wave estimation, CatBoost ensemble learning method has provided the lowest

* Corresponding author: I.zahmatkesh@scu.ac.ir
DOI: 10.22055/AAG.2023.43470.2364
Received: 2023-04-30
Accepted: 2023-06-25

ISSN: 2717-0764

error and highest correlation coefficient. Moreover, comparison of the results shows that the proposed ensemble learning methods compared with individual intelligent systems and hybrid methods can sufficiently improve the computational efficiency and performance of the shear wave prediction.

4- Conclusion

This paper presents ensemble models to predict shear wave velocity by using well logs in Mansuri oilfield. It can be observed that the CatBoost ensemble learning method outperforms the other intelligent base learners in all cases and quality. Meanwhile, the developed CatBoost ensemble learning model can serve as an effective tool for estimation of other petrophysical rock properties.

References

Anemangely, M., Ramezanzadeh, A., Tokhmechi, B., 2017. Shear wave travel time estimation from petrophysical logs using ANFIS-PSO algorithm: A case study from Ab-Teymour Oilfield. Journal of Natural Gas Science and Engineering 38, 373-387. https://doi.org/10.1016/j.jngse.2017.01.003.

Belete, D.M., Huchaiah, M.D., 2022. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications 44(9), 875-886. https://doi.org/ 10.1080/1206212X.2021.1974663

Gao, X., C. Shan, C. Hu, Z. Niu and Z. J. I. A. Liu (2019). " Gao, X., Shan, C., Hu, C., Niu, Z. and Liu, Z., 2019. An adaptive ensemble machine learning model for intrusion detection. Ieee Access 7, 82512-82521. https://doi.org/10.1109/ACCESS.2019.2923640.

HOW TO CITE THIS ARTICLE:
Hashemzade Kalvari, A., Zahmatkesh, I., 2024. Using ensemble learning techniques based on feature selection to predict shear wave velocity. Adv. Appl. Geol. 14(1), 167-185.

DOI: 10.22055/AAG.2023.43470.2364
URL: https://aag.scu.ac.ir/article_18653.html

ISSN: 2717-0764

مقاله پژوهشی

استفاده از تکنیک‌های یادگیری جمعی بر پایه انتخاب ویژگی برای پیش‌بینی سرعت موج برشی

عابدین هاشم زاده کلواری

گروه زمین شناسی نفت و حوضه های رسوبی، دانشکده علوم زمین، دانشگاه شهید چمران اهواز

ایمان زحمت کش*

گروه زمین شناسی نفت و حوضه های رسوبی، دانشکده علوم زمین، دانشگاه شهید چمران اهواز

* *I.zahmatkesh@scu.ac.ir*

تاریخ دریافت: ۱۴۰۲/۰۲/۱۰ تاریخ پذیرش: ۱۴۰۲/۰۴/۰۴

چکیده

سرعت موج برشی یکی از پارامترهای مهم برای تعیین خواص مکانیکی و پتروفیزیکی در مخازن هیدروکربوری است. اندازه‌گیری موج برشی به کمک روش‌های آزمایشگاهی و بهره‌گیری از ابزار صوتی دوقطبی امکان‌پذیر است، با این‌حال، با توجه به هزینه بالای عملیات مغزه‌گیری و اخذ لاگ صوتی دوقطبی، داده‌های واقعی موج برشی تنها برای تعداد محدودی از چاه‌های یک میدان در دسترس می‌باشند. برای غلبه بر این محدودیت¬ها، روش های مختلف هوش مصنوعی به منظور تخمین پارامتر مذکور از طریق لاگ‌های معمول چاه به کار برده می‌شوند. در این مطالعه به تخمین سرعت موج برشی با استفاده از روش‌های یادگیری جمعی (ensemble learning) در مخزن آسماری میدان منصوری پرداخته شد. در این مطالعه، سرعت موج برشی با استفاده از روش‌های یادگیری جمعی مثل رای‌گیری (Voting)، برانبارش (Stacking) ، بسته‌بندی (Bagging) و تقویت (Boosting) در مخزن آسماری برآورد شد و نتایج با مدل های مرسوم مثل، رگرسیون خطی (LR)، رگرسیون بردار پشتیبان (SVR)، الگوریتم نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT)، شبکه عصبی (ANN) و روش‌های هیبریدی مثل ترکیب شبکه عصبی با الگوریتم ژنتیک (ANN-GA)، ازدحام ذرات (ANN-PSO) و سیستم‌های فازی (ANFIS) مقایسه شد. به منظور ارزیابی و اعتبارسنجی مدل‌ها از ضریب همبستگی (R2) و ریشه میانگین مربعات خطا (RMSE) استفاده شد. مقایسه مدل‌های مرسوم و روش‌های هیبریدی با روش‌های یادگیری جمعی نشان داد که الگوریتم‌های جمعی عملکرد بهتری در تخمین موج برشی دارند. از بین روش‌های یادگیری جمعی نیز مدل کت‌بوست (Catboost) با میزان R2 برابر ۰,۹۸۳ و RMSE برابر با ۰,۰۵۸ بهترین عملکرد را نشان داد و قادر به تعیین موج برشی با دقت بالاست. نتایج این تحقیق نشان می-دهد که مدل Catboost می‌تواند به عنوان ابزاری با دقت بالا جهت برآورد سایر ویژگی‌های مخزن نظیر تخلخل، تراوایی و غیره مورد استفاده قرار گیرد.

واژه‌های کلیدی: *یادگیری جمعی، مخزن آسماری، میدان نفتی منصوری، سرعت موج برشی، یادگیری ماشینی*

۱- مقدمه

سرعت موج برشی (Shear wave velocity) یک پارامتر مهم برای طیف گسترده‌ای از کاربردها در صنعت نفت است و به طور گسترده در مطالعات پتروفیزیک، ژئوفیزیک و ژئومکانیک استفاده می‌شود. به منظور اندازه‌گیری این پارامتر از ابزار تصویربرداری صوتی دو قطبی (Dipole shear sonic imager) و روش‌های تجزیه و تحلیل مغزه در آزمایشگاه استفاده می‌شود، اما به دلیل هزینه بالا و زمان بر بودن روش‌های مذکور، تلاش‌هایی برای یافتن روش‌های جدید اندازه‌گیری موج برشی صورت گرفته است. روابط تجربی مختلفی برای اندازه-گیری موج برشی از داده‌های چاه‌پیمایی معرفی شده‌اند (Tosaya and Nur, 1982; Castagna et al., 1985; Anselmetti et al., 1993; Koesoemadinata and McMechan, 2001; Eskandari et al., 2004). با این حال استفاده از این روابط با محدودیت‌هایی روبرو است. زیرا این روابط عمدتاً برای یک میدان خاص توسعه یافته‌اند و در سطح جهانی قابل اعتماد نیستند (Malekiet al., 2014; Nourafkan et al., 2015; Anemangely et al., 2017; Wang et al., 2020). همچنین این همبستگی‌ها فقط تعدادی از پارامترهای پتروفیزیکی تاثیرگذار بر سرعت موج برشی را

توضیح می‌دهند (Eberhart-Phillips et al., 1989; Al-Dousari et al., 2016). بنابراین برای غلبه بر این محدودیت-ها استفاده از سیستم‌های هوشمند محاسباتی و روش‌های مختلف یادگیری ماشینی، به منظور تعیین پارامترهای مخزنی پیشنهاد شده است (Rezaee et al., 2007; Khandelwal et al., 2010; Rajabi et al., 2010; Maleki et al., 2014; Onalo etal., 2019).

تعدادی از کاربردهای موفقیت‌آمیز سیستم‌های هوشمند در تعیین خواص مخازن نفت گزارش شده‌اند، اما هنوز هم عدم قطعیت‌هایی در فرآیند توصیف مخازن و ایجاد مدل‌های بهینه در این زمینه وجود دارد. روش یادگیری جمعی (Ensemble learning) برای مقابله با این چالش‌ها به وجود آمده است.

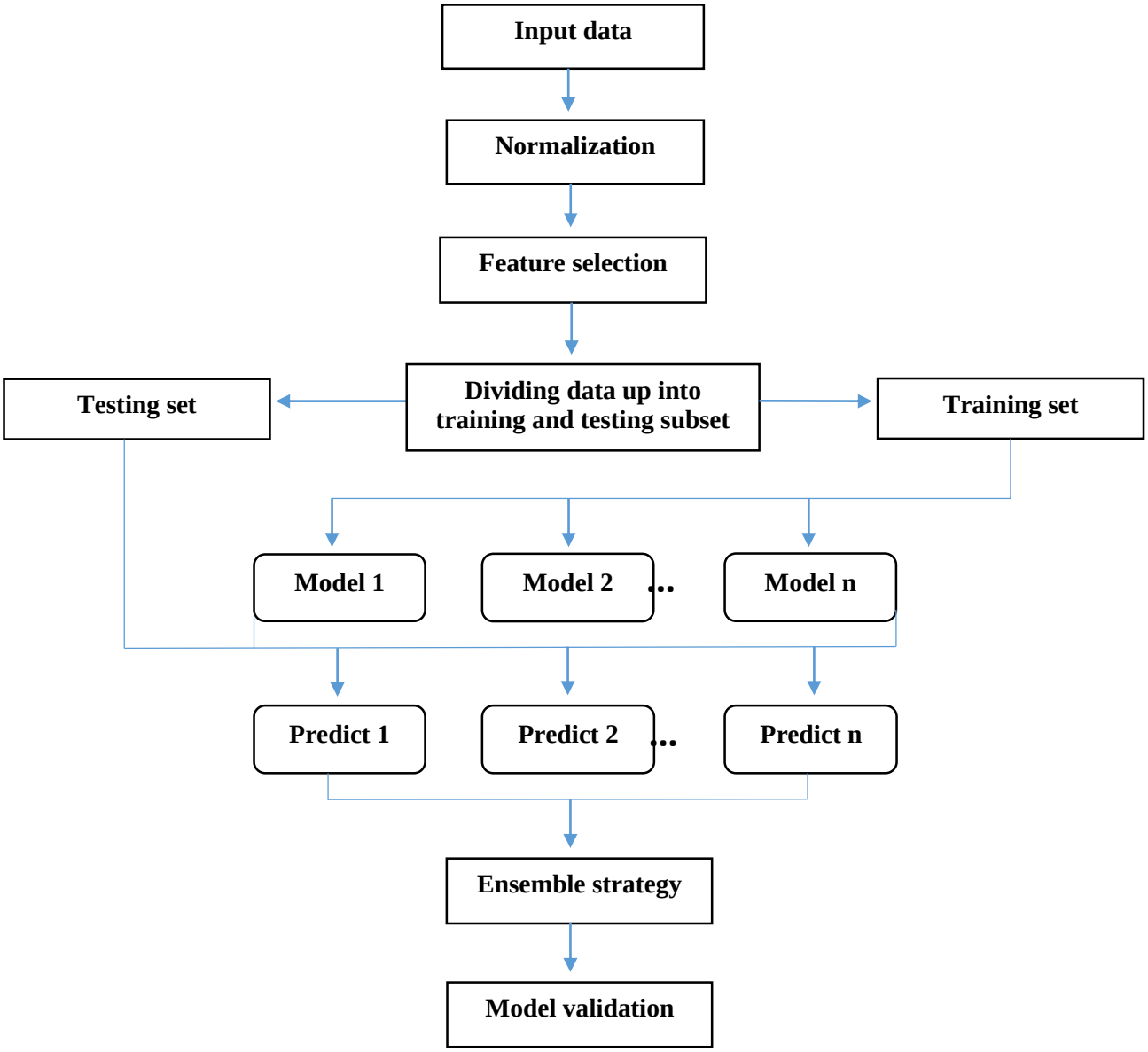
روش یادگیری جمعی ریشه در رفتار اجتماعی جوامع انسانی دارد که فرد قبل از اتخاذ هر تصمیم از نظرات سایر افراد استفاده می‌کند. در واقع این روش یک پارادایم جدید هوش محاسباتی است که از نظرات گروهی متخصصان برای بدست آوردن یک تصمیم استفاده می‌کند (Peng and Medicine, 2006; Polikar, 2012). یادگیری جمعی یکی از سیستم‌های هوشمند مبتنی بر یادگیری ماشینی است که از مجموع رای مدل‌ها برای حل یک مسئله استفاده می‌کند. همچنین این مدل ها روش‌های سریع و دقیق برای محاسبه پارامترهدف هستند (Chang et al., 2019).

این مطالعه با هدف اندازه‌گیری سرعت موج برشی با روش‌های هوشمند محاسباتی و استفاده از لاگ‌های معمول چاه انجام می‌شود. در ابتدا کارآمدترین روش‌های مرسوم یادگیری ماشینی شامل، رگرسیون خطی (Linear regression)، رگرسیون بردار پشتیبان (SVR)، الگوریتم نزدیک‌ترین همسایه (KNN)، درخت تصمیم (Decision tree)، شبکه‌های عصبی (ANN)، شبکه عصبی- الگوریتم ازدحام ذرات (ANN-PSO)، شبکه عصبی-الگوریتم ژنتیک (ANN-GA) و سیستم‌های فازی (ANFIS) برای پیش‌بینی موج برشی مورد استفاده قرار می‌گیرد. این روش‌ها بر اساس مطالعه گسترده و جامع تحقیقات قبلی انتخاب شدند. در گام بعد روش‌های نوین یادگیری جمعی برای پیش‌بینی موج برشی و مقایسه با روش‌های مرسوم مورد استفاده قرار می‌گیرند. در شکل ۱ نمودار جریانی مربوط به روش‌های یادگیری جمعی ارائه شده است. اولین روش یادگیری جمعی مورد ارزیابی، مدل رای‌گیری (Voting) است که استراتژی نهفته در آن میانگین‌گیری ساده در مرحله تصمیم-گیری مدل است. سپس به بررسی روش برانبارش (Stacking) پرداخته می‌شود که رای‌گیری هر مدل را به صورت وزن‌دار انجام می‌دهد و یک میانگین‌گیری وزن‌دار برای مدل‌ها ارائه می‌دهد. همچنین روش‌های بسته‌بندی (Bagging) و تقویت تطبیقی (Adaboost)، تکنیک تقویت گرادیان (Xgboost) و کت بوست (Catboost) با رویکرد استفاده از درخت تصمیم در مسائل رگرسیون مورد استفاده قرار می‌گیرند و به مقایسه این مدل‌ها پرداخته می‌شود.

۲- منطقه مورد مطالعه

در این مطالعه از داده‌های مربوط به سازند آسماری میدان نفتی منصوری، واقع در حوضه زاگرس ایران استفاده می‌شود. در اواخر الیگوسن، پیشروی آرام و محدود دریا، چرخه آسماری پایینی را تشکیل داده است، در ابتدای میوسن رسوبات محیط‌های کم ژرفا با نهشته شدن در تمامی حوضه‌ها چرخه آسماری میانی را تشکیل دادند و در اواخر میوسن با افت سطح دریا شرایط تبخیری حاکم و رسوب‌گذاری آسماری به پایان می‌رسد. سنگ شناسی لایه‌ها در سازند آسماری میدان منصوری شامل رسوبات دریایی عمیق در قاعده آسماری زیرین و رسوبات نیم عمیق و کم عمق تا سبخایی در آسماری میانی و بالایی می‌باشد. سنگ منشا و پوش سنگ ذخایر مخزن آسماری به ترتیب سازندهای کژدمی و گچساران هستند مخزن آسماری میدان منصوری به ۸ زون تقسیم شده است که تنها زون‌های ۱، ۲ و ۳ مخزن حاوی نفت بوده و زون‌های مخزنی دیگر علیرغم تخلخل زمینه‌ای خوب، حاوی آب هستند (Zahmatkesh et al., 2018). ظرفیت تولید نفت خام میدان منصوری به طور متوسط مقدار ۱۰۰ هزار بشکه در روز بوده و میزان تولید نفت از مخزن آسماری این میدان ۵۵ هزار بشکه در روز می‌باشد.

مخزن آسماری میدان منصوری ۳۰ کیلومتر طول و ۳ کیلومتر عرض دارد. این میدان از شمال غربی در همسایگی میدان اهواز ، از غرب با میدان آب تیمور و از شمال شرقی در مجاورت میدان نفتی شادگان قرار دارد (Kavianpor Sangno et al., 2015). میدان مذکور در سال ۱۳۴۲ کشف و از سال ۱۳۵۳ به بهره‌برداری رسیده است. شکل ۲ موقعیت جغرافیایی این میدان را نشان می دهد.



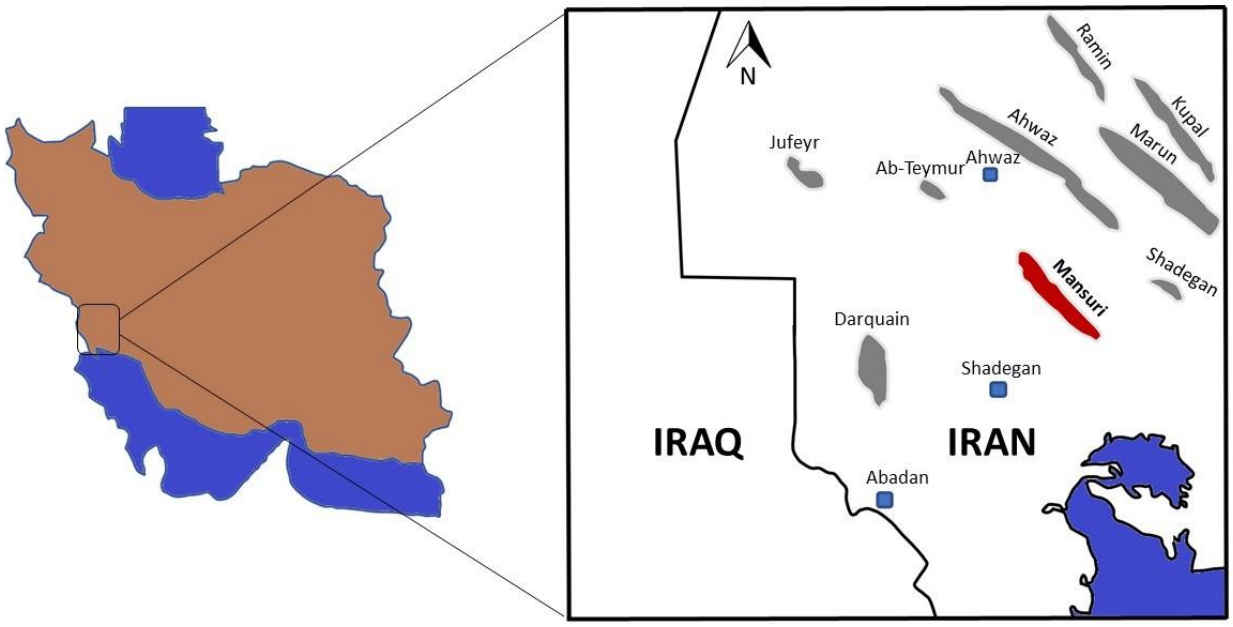
شکل ۱- فلوچارت روش یادگیری جمعی

Fig. 1. Ensemble learning flowchart

۳- روش پژوهش

در این بخش، به منظور تعیین موج برشی در میدان منصوری ، داده‌های چاه‌پیمایی حاصل از دو حلقه چاه شامل ۲۲۰۰۰ نقطه داده مثل، لاگ قطرسنجی (Caliper log)، لاگ صوتی (Sonic log)، چگالی ظاهری (RHOB)، تخلخل نوترونی (NPHI)، فاکتور فوتوالکتریک (PEF)، مقاومت واقعی سازند (RT)، گامای محاسبه شده (CGR)، پرتو گامای طیفی (SGR)، عمق (Depth)، قطر متهو موج برشی برای ساخت مدل‌ها در نظر گرفته شد. روش حذف ویژگی بازگشتی (Recursive feature elimination) با رتبه‌بندی ویژگی‌ها و اختصاص وزن به هرکدام از داده‌های چاه‌پیمایی ۱۴۰۰۰ نقطه داده را انتخاب کرد. لاگ‌های معمول چاه‌پیمایی نظیر، چگالی ظاهری، تخلخل نوترونی، لاگ صوتی، گامای محاسبه شده، پرتوی گامای طیفی، عمق و موج برشی به عنوان ورودی‌های بهینه جهت آموزش مدل‌ها تعیین شد. در گام بعد بین مجموعه داده‌های ورودی و خروجی تفکیک ایجاد شد. داده‌های چاه‌پیمایی انتخاب شده توسط RFE، به عنوان مقادیر ورودی و سرعت موج برشی با واحد کیلومتر بر ثانیه به عنوان مقدار خروجی انتخاب گردید.

پس از آشنایی و معرفی ساختمان مدل‌های مختلف یادگیری جمعی، مثل، برانبارش، بسته‌بندی و الگوریتم‌های تقویت، مدل-های مذکور به منظور تخمین سرعت موج برشی ساخته شدند. به دلیل کاربرد موفقیت آمیز روش‌های مرسوم در تخمین پارامترهای مخزنی، پس از ساخت مدل‌های مرسوم مثل شبکه-های عصبی، رگرسیون خطی، رگرسیون بردار پشتیبان، الگوریتم نزدیک ترین همسایه، درخت تصمیم و روش‌های هیبریدی مثل، ANN-PSO، ANN-GA و ANFIS، به مقایسه این مدل‌ها با روش‌های یادگیری جمعی پرداخته شد.



شکل ۲- موقعیت مکانی منطقه مورد مطالعه

Fig. 2. Location of the study area

۳-۱- آماده‌سازی داده‌ها

پس از عملیات بارگزاری داده‌های خام، به منظور ارزیابی‌های پتروفیزیکی باید تصحیحاتی بر روی داده‌های چاه‌پیمایی اعمال گردد. یکی از مواردی که در نمودارها مشاهده می‌شود عدم انطباق لاگ‌ها از نظر عمقی است. نمودار پرتوی گاما به علت تاثیرپذیری کم نسبت به عوامل محیطی به عنوان لاگ مرجع جهت هم عمق سازی، استفاده گردید. پس از تصحیحات مربوط به عمق، چارت‌های مربوطه، برای تصحیحات محیطی نمودارهای چاه‌پیمایی، به کار برده شد. همچنین به کمک محاسبات مقدماتی، اثر شرایط محیطی نمودارگیری مثل فشار ته چاه، دما، عوامل شیمیایی و غیره بر روی داده‌ها تصحیح گردید.

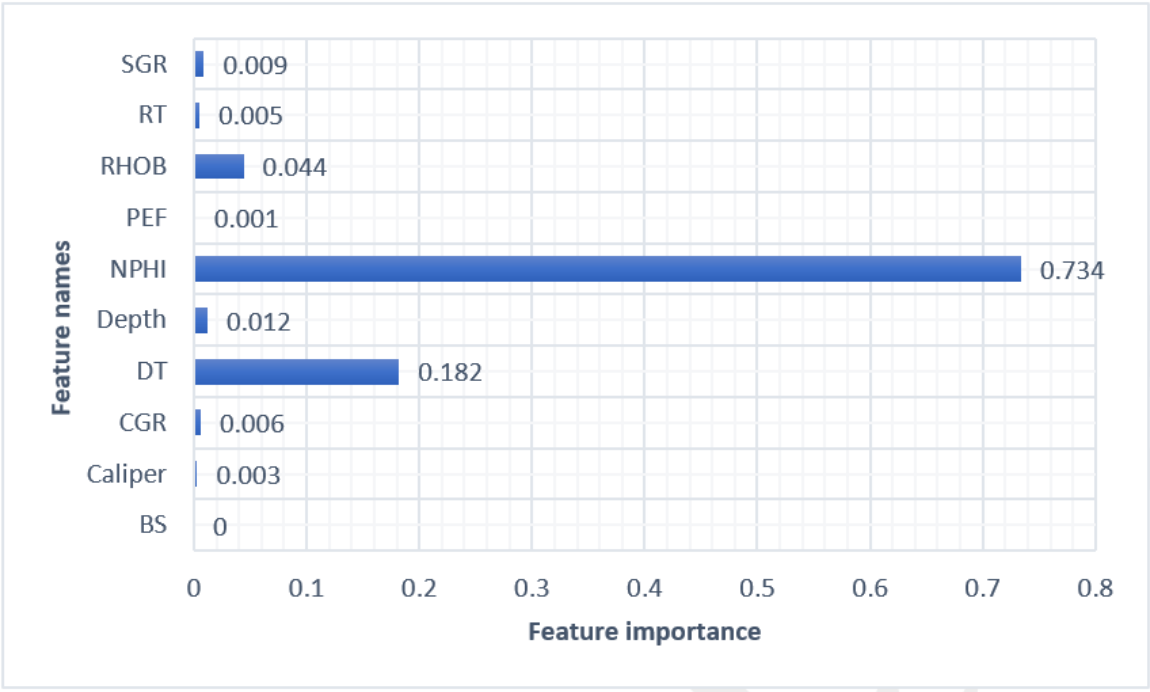
برای توسعه مدل باید پارامترهای ورودی به منظور آموزش به کمک الگوریتم‌های ماشینی آماده شوند، در فاز اولیه توسعه مدل، پارامترهایی نظیر لاگ قطرسنجی، لاگ صوتی، چگالی ظاهری، تخلخل نوترونی، فوتوالکتریک، مقاومت واقعی سازند، گامای محاسبه شده، پرتو گامای طیفی، عمق و قطر مته به عنوان ورودی‌های اولیه انتخاب شدند. تعداد ۲۲۰۰۰ نقطه داده از دو حلقه چاه میدان منصوری، برای اهداف آموزش و آزمایش مدل استفاده شد. از داده‌های یک چاه روند آموزش مدل‌ها انجام می‌گیرد. داده‌های مربوط به چاه دوم برای ارزیابی مدل به کار برده می‌شوند. به علت وجود مقادیر با اطلاعات پیش‌بینی اندک یا بدون اطلاعات، نمی‌توان از تمامی پارامترهای ورودی استفاده کرد. روش‌های انتخاب ویژگی (Feature selection) با کاهش ابعاد فضای ویژگی در طیف گسترده به انتخاب بهینه این پارامتر ها کمک می‌کنند.

حذف ویژگی بازگشتی یک روش انتخاب ویژگی پوششی است که توسط گایون و همکاران پیشنهاد شده است (Guyon et al., 2002). روش حذف ویژگی بازگشتی ابتدا تمامی داده های یک مجموعه را در نظر می‌گیرد و با رتبه‌بندی هر ویژگی بر اساس اهمیت آن، با روش بازگشتی، ویژگی‌های دارای کمترین امتیاز را به علت تاثیر کم در پیش‌بینی حذف می‌کند. این روند تا زمان رسیدن به ویژگی‌های مورد نظر تکرار می‌شود. در این مطالعه از یک روش رگرسیون خطی برای آموزش RFE استفاده می‌شود. بر اساس شکل ۳ روش حذف ویژگی بازگشتی به داده‌های چاه‌پیمایی وزن‌های معینی اختصاص داد. به نظر می‌رسد که همبستگی قوی بین سرعت موج برشی و لاگ نوترونی وجود دارد. علاوه بر این، گزارش لاگ صوتی، همبستگی قوی با سرعت موج برشی را نشان می‌دهد. اما حذف ویژگی بازگشتی کمترین وزن را به قطر مته، لاگ قطرسنجی، فاکتور فوتوالکتریک و مقاومت واقعی سازند اختصاص داد. بنابراین داده‌های چاه‌پیمایی مذکور، کم‌ترین همبستگی با سرعت موج برشی را نشان می‌دهند.

پس از انتخاب ویژگی‌های ورودی، نرمال‌سازی داده‌ها امری بسیار ضروری است. زیرا پارامترهای ورودی واحدهای اندازه-گیری متفاوتی دارند و وجود مقادیر بزرگ داده‌ها، تاثیر قابل توجهی بر خروجی مدل خواهد داشت. بنابراین با استفاده از معادله زیر مقادیر داده‌ها بین صفر و یک نگاشت شدند:

$$X_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad \text{(رابطه ۱)}$$

که x داده قبل از نرمال‌سازی، xmin حداقل مقدار داده‌ها، xmax حداکثر مقدار داده‌ها و xnorm مقدار نرمال شده است.



شکل ۳- اهمیت ویژگی مربوط به داده‌های چاه‌پیمایی

Fig. 3. The feature importance of well logs

۳-۲- مدل‌های پایه

در این مطالعه روش‌های یادگیری جمعی نظیر رای‌گیری و برانبارش به کمک الگوریتم‌های رگرسیون خطی، رگرسیون بردار پشتیبان و نزدیک‌ترین همسایه مرحله آموزش را طی می-کنند. به منظور آموزش روش‌های بسته‌بندی و تقویت، مدل‌های پایه درخت تصمیم به کار برده می‌شوند. در ادامه این مدل‌های پایه معرفی خواهند شد.

۳-۲-۱- رگرسیون خطی

رگرسیون خطی ساده یک مدل موردی با یک متغیر مستقل است. رگرسیون خطی ساده وابستگی متغیر را تعریف می‌کند و تأثیر متغیرهای مستقل را از تأثیر متقابل متغیرهای وابسته متمایز می‌کند. هدف از رگرسیون خطی مدل‌سازی رابطه خطی بین متغیرهای مستقل و متغیر وابسته است (Maulud et al., 2020).

۳-۲-۲- رگرسیون بردار پشتیبان

رگرسیون بردار پشتیبان زیرمجموعه‌ای از روش‌های یادگیری نظارت شده است که برای طبقه بندی و وظایف رگرسیون استفاده می‌شود. رگرسیون بردار پشتیبان می‌تواند روابط غیرخطی بین داده‌های ورودی و خروجی را در ابعادی بالاتر بیاموزد و مدل‌سازی کند، در نتیجه خطای آموزشی مشاهده‌شده و خطای توزیع را به‌اندازه کافی برای دستیابی به کارایی رگرسیون تعمیم‌یافته به حداقل می‌رساند. این مدل یک سیستم یادگیری کارآمد بر اساس یک تئوری بهینه‌سازی مؤثر است که اصل کمینه‌سازی القایی خطای ساختاری را برای دستیابی به یک راه‌حل بهینه کلی پیاده‌سازی می‌کند (Yu et al., 2018).

۳-۲-۳- الگوریتم نزدیک‌ترین همسایه

مدل رگرسیونی نزدیک‌ترین همسایه یکی از ساده‌ترین و قدیمی‌ترین روش‌های مورد استفاده برای مسائل رگرسیون و طبقه‌بندی است و اغلب عملکرد کارآمدی دارد. این مدل از نزدیک‌ترین نقاط داده و مشابه‌ترین نمونه‌ها در مجموعه داده آموزشی برای تخمین ارزش یک مشاهده جدید استفاده می‌کند. عملکرد مدل به طور قابل توجهی به متریک فاصله مورد استفاده بستگی دارد. مدل رگرسیون نزدیک‌ترین همسایه با تعیین فاصله بین یک مشاهده جدید و همه مشاهدات موجود در داده-های آموزشی عمل می‌کند (Sumayli, 2023).

۳-۲-۴- درخت تصمیم

درخت تصمیم مدل مبتنی بر درخت است و یک الگوریتم یادگیری نظارت شده است که می تواند برای هر دو مدل طبقه-بندی و رگرسیون استفاده شود. این مدل یک ساختار درختی است که از تعداد دلخواه گره و شاخه در هر گره تشکیل شده است. گره‌های بیرونی گره داخلی نامیده می‌شود. گره‌های دیگر برگ نامیده می‌شوند. مقادیر متغیرهای ورودی یک تابع خاص را در مرحله آموزش در نظر می‌گیرند. محرک درخت تصمیم الگوریتمی است که درخت تصمیم را از نمونه‌های داده شده تولید می‌کند. هدف این الگوریتم یافتن درخت تصمیم بهینه با به حداقل رساندن تابع تناسب است (Loh and Discovery, 2011).

۳-۳- روش یادگیری جمعی

یادگیری جمعی اصطلاحی کلی برای روش‌هایی است که مدل های متعدد را برای تصمیم‌گیری ترکیب می‌کنند. یک روش که به عنوان یادگیرنده پایه نیز نامیده می‌شود، الگوریتمی است که مجموعه‌ای از مثال‌ها را به عنوان ورودی می‌گیرد و مدلی تولید می‌کند که این مثال‌ها را تعمیم می‌دهد. با استفاده از مدل تولید شده می‌توان برای نمونه‌های جدید پیش‌بینی را انجام داد. فرض اصلی یادگیری جمعی این است که با ترکیب چند مدل، خطاهای یک مدل واحد احتمالاً توسط مدل‌های دیگر جبران می‌شود و در نتیجه عملکرد کلی پیش‌بینی گروه بهتر از یک مدل منفرد خواهد بود. در ادامه روش‌های مختلف یادگیری جمعی معرفی خواهند شد.

۳-۳-۱- برانبارش

روش برانبارش یکی از تکنیک‌های یادگیری جمعی است که توسط والپرت معرفی شد (Wolpert, 1992). در ساختار مدل دولایه وجود دارد. در لایه یا سطح صفر، مدل‌های پایه مختلف (Base learner) با یک نوع داده آموزش داده می‌شوند، سپس پیش‌بینی هر مدل نسبت به پارامتر هدف به عنوان ورودی به لایه اول وارد می‌شود. در لایه اول، یک فرامدل (Meta learner) آموزش می‌بیند و پیش‌بینی نهایی به کمک این مدل صورت می‌گیرد (Da Silva et al., 2021). در این مطالعه الگوریتم های نزدیک‌ترین همسایه، رگرسیون بردار پشتیبان و رگرسیون خطی به عنوان مدل‌های سطح صفر آموزش داده می‌شوند و به کمک فرامدل، پیش‌بینی نهایی انجام می‌گیرد.

زمانی که ساختار مدل دارای یک لایه باشد تمامی مدل‌ها در تصمیم‌گیری سهم یکسان دارند. روش Voting از این رویکرد استفاده می‌کند. در واقع این مدل فاقد لایه دوم در ساختار خود می‌باشد. در این مطالعه روش Voting با پارامترهای یکسان به کار برده شده برای Stacking آموزش می‌بیند تا تاثیر وجود فرامدل در ساختار Stacking آشکار شود.

۳-۲-۲- بسته‌بندی

روش بسته‌بندی با هدف راهی برای کاهش واریانس مدل توسط بریمن معرفی شد (Breiman, 1996). این روش برای بهبود واریانس از رویکردی به نام بوت استرپ (Bootstrap) بهره می‌برد که زیر مجموعه‌ای با تکرار از داده‌های آموزشی به صورت تصادفی ایجاد می‌کند. این زیرمجموعه قابلیت جایگزینی دارد و تعداد عضوهای آن برابر با داده‌های آموزشی است (Anifowose et al., 2017). نحوه آموزش این روش به صورت موازی است و همه مدل‌ها به صورت همزمان مرحله آموزش را طی می‌کنند، به این مفهوم که هر مدل به صورت فردی، مستقل از سایر مدل‌ها آموزش داده می‌شود و در روند آموزش سایر مدل‌ها دخالتی نمی‌کند (Xu et al., 2020). باید توجه داشت که تمامی مدل‌های پایه از یک نوع هستند اما تعداد متعددی از این مدل‌ها برای بهبود عملکرد پیش‌بینی به کار برده می‌شوند. این روش سهم تمامی یادگیرندگان را در مرحله تصمیم‌گیری یکسان در نظر می‌گیرد و پیش‌بینی خروجی نهایی که پیش‌بینی تمامی مدل‌هاست، میانگین‌گیری می‌شود.

۳-۳-۳- تقویت تطبیقی

تقویت تطبیقی یکی از پرکاربردترین و قدرتمندترین روش‌های یادگیری جمعی است. این روش از آموزش متوالی (Seuential) برای ساخت مدل‌ها استفاده می‌کند. در این روش در آغاز پروسه آموزش، با ایجاد یک درخت تصمیم (Decision tree)، به تمامی داده‌ها وزن مساوی داده می‌شود. سپس وزن‌های پیش‌بینی شده توسط درخت تصمیم در مرحله قبل بررسی می‌شوند. پیش‌بینی‌های دقیق وزن‌های ثابت می‌گیرند، اما وزن پیش‌بینی‌های نادرست افزایش می‌یابد تا درخت تصمیم بعدی در مرحله آموزش به این پیش‌بینی‌های نادرست توجه بیشتری کند و خطاهای مدل قبل را اطلاح کند. این روند تا زمانی ادامه می‌یابد که تمام یادگیرندگان مرحله آموزش را طی کرده و وزن‌های معینی به آن‌ها اختصاص یابد. در پایان، راه‌حل بهینه براساس ترکیب همه درختان بدست می‌آید (Tüysüzoğlu and Birant, 2020; Jafarzadeh et al., 2021).

۳-۳-۴- Xgboost

روش Xgboost از زیرمجموعه‌های روش تقویت گرادیان (Gradient boosting) به شمار می‌آید که توسط چن و گاسترین پیشنهاد شد (Chen and Guestrin, 2016). در مرحله آموزش این مدل، درخت تصمیم ایجاد شده خطای مقادیر واقعی و پیش‌بینی شده که همان تابع هزینه یا تابع زیان است را جستجو کرده و بهینه‌سازی تابع صورت می‌گیرد. سپس درخت‌ها بر اساس خطاهای پیش‌بینی یا باقیمانده‌های درخت قبلی ایجاد می‌شوند (Abedi et al., 2022).

Xgboost عمدتا به منظور افزایش سرعت و بهبود عملکرد با بهره‌گیری از تقویت گرادیان توسعه یافته است. این ابزار با دارا بودن ویژگی‌هایی مثل، پشتیبانی از ساختار موازی ایجاد درخت و استفاده بهینه از منابع حافظه، افزایش سرعت مدل را شامل می‌شود. همچنین برای جلوگیری از برازش بیش از حد مدل‌های درختی ، ضریب تعیینی را در محدوده‌های مشخص تنظیم می‌کند.

۳-۳-۵- Catboost

الگوریتم Catboost روشی نوین است که از تقویت گرادیان مبتنی بر درخت برای حل مسائل رگرسیون و طبقه‌بندی استفاده می‌کند. این روش تقویت گرادیان، الگوریتمی را آموزش می‌دهد که در هر تکرار با کمینه‌سازی گرادیان، بهینه‌سازی تابع زیان را انجام می‌دهد. اما این فرایند ممکن است سبب برازش بیش از حد مدل شود. به منظور حل این مشکل Catboost به کمک روش تقویت سفارشی (Ordered boosting) برای تخمین گرادیان، بایاس مدل را کاهش می‌دهد (Zhang et al., 2020). همچنین به منظور کنترل برازش بیش از حد مدل (Overfitting) در ساختار درختی، با ایجاد جایگشت‌های تصادفی تخمین مقادیر درختان تصمیم را انجام می‌دهد (Jabeur et al., 2021).

۳-۴- ارزیابی مدل

یکی از حیاتی‌ترین مراحل در ساخت مدل‌ها، ارزیابی آن‌هاست. ارزیابی مدل‌ها تعیین می‌کند که آیا هدف مدل‌سازی محقق شده است یا خیر. همچنین مرحله ارزیابی با مقایسه رویکردهای مدل‌سازی مختلف، هدایت تحقیقات آینده را فراهم می‌کند. در این مطالعه برای ارزیابی پیش‌بینی مدل‌های مرسوم مثل، شبکه‌های عصبی، رگرسیون خطی، رگرسیون بردار پشتیبان، الگوریتم نزدیک ترین همسایه، درخت تصمیم و روش‌های هیبریدی مثل، ANN-PSO، ANN-GA و ANFIS و انواع مدل‌های جمعی از شاخص‌های آماری ریشه میانگین مربعات خطا و ضریب همبستگی طبق روابط زیر استفاده شد:

(رابطه ۲)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y_i)^2}$$

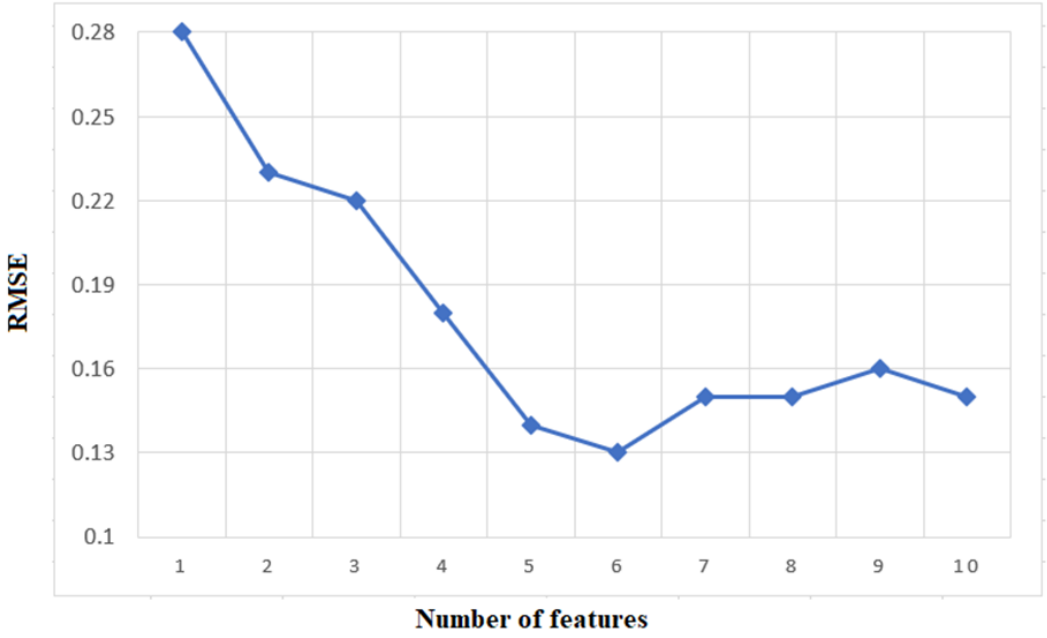
(رابطه ۳)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y - y_i)^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

که در آن،y مقدار واقعی موج برشی، yi مقدار پیش‌بینی شده، $\bar{y}$ میانگین مقادیر واقعی و n تعداد نمونه‌هاست. یک مدل مناسب مقدار R^2 نزدیک به یک ولی مقدار RMSE پایینی دارد.

۴- بحث و نتایج

در این مطالعه ساخت و ارزیابی مدل‌های مرسوم و انواع روش-های یادگیری جمعی با استفاده از داده‌های چاه‌پیمایی و سرعت موج برشی در میدان منصوری انجام گرفت. به این منظور در گام نخست برای انتخاب ویژگی‌های بهینه، روش حذف ویژگی بازگشتی با رگرسیون خطی به کار برده شد. این الگوریتم با رتبه‌بندی ویژگی‌ها، مقادیر ضعیف را حذف می‌کند تا به ویژگی-های بهینه دست یابد. بر اساس شکل ۳ روش حذف ویژگی بازگشتی برای داده‌های چاه‌پیمایی مثل، لاگ قطرسنجی، لاگ صوتی، چگالی ظاهری، تخلخل نوترونی، فاکتور فوتوالکتریک، مقاومت واقعی سازند، گامای محاسبه شده، پرتو گامای طیفی، عمق و قطر مته، رتبه‌بندی ویژگی‌ها را انجام داد. طبق شکل ۴، با افزایش تعداد ورودی، میزان مجذور میانگین مربعات خطا تا مقدار ۶ ورودی، کاهش یافته و بعد از آن افزایش می‌یابد. روش RFE لاگ‌های معمول چاه‌پیمایی نظیر چگالی ظاهری، تخلخل نوترونی، لاگ صوتی، گامای محاسبه شده، پرتوی گامای طیفی و عمق را به عنوان ورودی‌های بهینه جهت آموزش مدل‌ها تعیین کرد. تعیین مقادیر بهینه ورودی با چندین بار آزمایش و اجرای مدل RFE صورت گرفت.



شکل ۴- انتخاب تعداد ورودی‌های بهینه

Fig. 4. The optimal number of inputs

پس از تعیین تعداد بهینه ورودی، مدل‌های مرسوم و همچنین انواع مدل‌های یادگیری جمعی برای تخمین موج برشی ساخته شده و آموزش داده شدند. قبل از آموزش مدل‌ها، مجموعه داده‌ها به دو مجموعه مختلف، یعنی مجموعه‌های آموزشی و آزمایشی تقسیم شدند. لاگ‌های مربوط به چاه اول برای ساخت مدل هوشمند در فرآیند آموزشی مورد استفاده قرار گرفتند. علاوه بر این، چاه دوم میدان، به عنوان مجموعه

آزمایشی برای ارزیابی مدل‌های مختلف و اجتناب از برازش بیش از حد مدل‌ها در مجموعه آموزشی به کار برده شد.

در مدل‌های هیبریدی مثل شبکه عصبی- ازدحام ذرات، شبکه عصبی- الگوریتم فازی و شبکه عصبی-الگوریتم ژنتیک، پارامترها و متغیرهای مهم تنظیم شدند. در شبکه‌های عصبی برای دستیابی به حداقل خطای خروجی باید وزن‌ها و بایاس‌های مدل تنظیم شوند. بنابراین در روش ANN، مقادیر بهینه برای تعداد نورون‌های لایه ورودی، لایه پنهان و خروجی تعیین شد. در مدل نزدیک‌ترین همسایه مقدار بهینه برای تعداد همسایه‌ها تعیین شد. همچنین برای رگرسیون بردار پشتیبان تنظیم پارامتر جریمه C و گاما به بهینه‌سازی مدل کمک کرد. در مدل هیبریدی شبکه عصبی-ازدحام ذرات، آموزش و به روزرسانی وزن‌های شبکه عصبی توسط الگوریتم ازدحام ذرات صورت می‌گیرد. بنابراین برای تعیین مدل بهینه در روش ANN-PSO، مقدار بهینه جمعیت اولیه ذرات، تعداد تکرار و پارامترهای ثابت C1 و C2 تعیین شد. همچنین روش فازی-عصبی و شبکه عصبی-ژنتیک برای تعیین سرعت موج برشی و مقایسه با نتایج سایر مدل‌ها آموزش داده شد و پارامترهای مهم این روش‌ها نیز بهینه‌سازی شدند. جدول ۱ مقادیر بهینه‌سازی شده برای روش‌های مرسوم و هیبریدی را نشان می‌دهد.

تنظیم دقیق هایپرپارامترها می‌تواند عملکرد پیش‌بینی مدل را به مقدار قابل توجهی بهبود بخشد. در تمامی مدل‌های جمعی از جستجوی شبکه‌ای (Grid search cv) برای یافتن بهترین مقادیر پارامترها استفاده شد. برای کاهش فضای جستجو فقط تعدادی از مهمترین پارامترها تنظیم شدند. در روش‌های voting و stacking، الگوریتم‌های نزدیک‌ترین همسایه، رگرسیون بردار پشتیبان و رگرسیون خطی برای آموزش در نظر گرفته شدند. برای مقایسه بین این دو روش فضاهای جستجوی یکسان برای آن‌ها به کار برده شد. هایپرپارامترهای تنظیم شده دو مدل Voting و Stacking شامل تعداد همسایه‌ها در الگوریتم نزدیک‌ترین همسایه، گاما و پارامتر جریمه C در رگرسیون بردار پشتیبان هستند. برای مدل bagging پارامترهای تخمین‌گر پایه، تعداد تخمین‌گر و حداکثر نمونه تنظیم شدند. در الگوریتم Adaboost مقادیر بهینه برای هایپرپارامترهای تعداد تخمین‌گر، تخمین‌گر پایه و نرخ یادگیری بدست آمد. همچنین در الگوریتم Xgboost برای ایجاد مدل بهینه از هایپرپارامترهای حداکثر عمق درخت، تعداد تخمین‌گر و نرخ یادگیری استفاده شد و برای مدل Catboost پارامترهای حداکثر عمق درخت، نرخ یادگیری و تعداد تکرار تنظیم شدند. مجموعه هایپرپارامترها و مقادیر بهینه آن‌ها در جدول ۲ گزارش شده‌اند.

پس از انتخاب ویژگی‌های بهینه و تنظیم هایپرپارامترها، مقایسه عملکرد پیش‌بینی مدل‌ها برای تعیین موج برشی صورت گرفت. جدول ۳ و ۴ به ترتیب نشان‌دهنده نتایج آماری روش‌های مرسوم و روش‌های یادگیری جمعی در تخمین موج برشی است. از بین روش‌های مرسوم، الگوریتم شبکه عصبی با مجذور میانگین مربعات خطا برابر ۰,۱۲۴ و ضریب هبستگی ۰,۹۶۲ عملکرد بهتری را ارائه داد. نتایج نشان داد تمامی الگوریتم‌های جمعی عملکردی بهتری نسبت به مدل‌های مرسوم داشتند. شکل ۵ تطابق بین مقادیر واقعی و براورد شده موج برشی را برای مدل‌های جمعی نشان می‌دهد. همچنین از بین تمامی مدل‌ها، روش Catboost بیش‌ترین تطابق را با مقادیر واقعی موج برشی نشان داد. شکل ۶ و ۷ تطابق بالای روش Catboost در تخمین موج برشی نسبت به مقادیر واقعی را نشان می‌دهد. Catboost با بیش‌ترین مقدار R2 و کمترین میزان RMSE در مرحله آزمایش مدل، بهترین عملکرد را در تخمین موج برشی داشت. روش رگرسیون خطی، با بالاترین مقادیر RMSE و کم‌ترین میزان R2، ضعیف‌ترین میزان عملکرد پیش‌بینی را در بین تمامی مدل‌ها نشان داد. رگرسیون خطی، رگرسیون بردار پشتیبان و الگوریتم نزدیک‌ترین همسایه نتایج نسبتا ضعیفی در پیش‌بینی موج برشی داشتند، اما با به کارگیری این الگوریتم‌ها در ساختار مدل‌های تجمیعی مثل، رای‌گیری و برانبارش در پیش‌بینی موج برشی بهبود قابل ملاحظه‌ای حاصل شد. وجود فرامدل در ساختار stacking و میانگین‌گیری وزن‌دار سبب بهبود عملکرد قابل توجه مدل نسبت به روش Voting شد. الگوریتم‌های تقویت مثل Adaboost و Xgboost در مقایسه با Bagging عملکرد بهتری نمایش دادند و برای پیش‌بینی موج برشی می‌توانند قابل اعتمادتر باشند.

جدول ۱- هایپرپارامترهای تنظیم شده برای مدل‌های مرسوم و روش‌های هیبریدی

Table 1. Tuned hyperparameters for conventional models and hybrid methods

Model	Parameters	Optimal Values
SVR	Penalty parameter, C	0.5
	Gamma (SVR)	0.2
KNN	N_neighbors	3
ANN	Input layer	6
	Hidden layer	7
	Output layer	1
ANFIS	Number of inputs	6
	Number of outputs	1
ANN-GA	Population Size	150
	Generation (Iteration)	50
ANN-PSO	Number of particles	50
	Number of iteration	100
	C_1, C_2	4

جدول ۲- هایپرپارامترهای تنظیم شده برای مدل‌های یادگیری جمعی

Table 2. Tuned hyperparameters for ensemble learning models

Model	Parameters	Range of parameters	Optimal parameters
Voting	Penalty parameter, C (SVR)	0.1-4.0	4.0
	Gamma (SVR)	0.1-0.9	0.1
	Number of neighbors (KNN)	3-10	3
Stacking	Penalty parameter, C (SVR)	0.1-4.0	0.5
	Gamma (SVR)	0.1-0.9	0.1
	Number of neighbors (KNN)	3-10	3
Bagging	Base estimator	-	Decision tree
	Number of estimators	10-1000	150
	Max sample	0.1-1.0	0.9
Adaboost	Base estimator	-	Decision tree
	Number of estimators	10-1000	1000
	Learning rate	0.1-0.9	0.5
Xgboost	Maximum tree depth	5-20	5
	Number of estimators	10-1000	500
	Learning rate	0.1-0.9	0.1
Catboost	Maximum tree depth	5-20	9
	Iterations	10-2000	1500
	Learning rate	0.1-0.9	0.1

جدول ۳- مقایسه ریشه میانگین مربعات خطا (RMSE) و ضریب همبستگی (R2) مربوط به داده‌های آموزشی و آزمایشی برای مدل‌های مرسوم و روش‌های هیبریدی

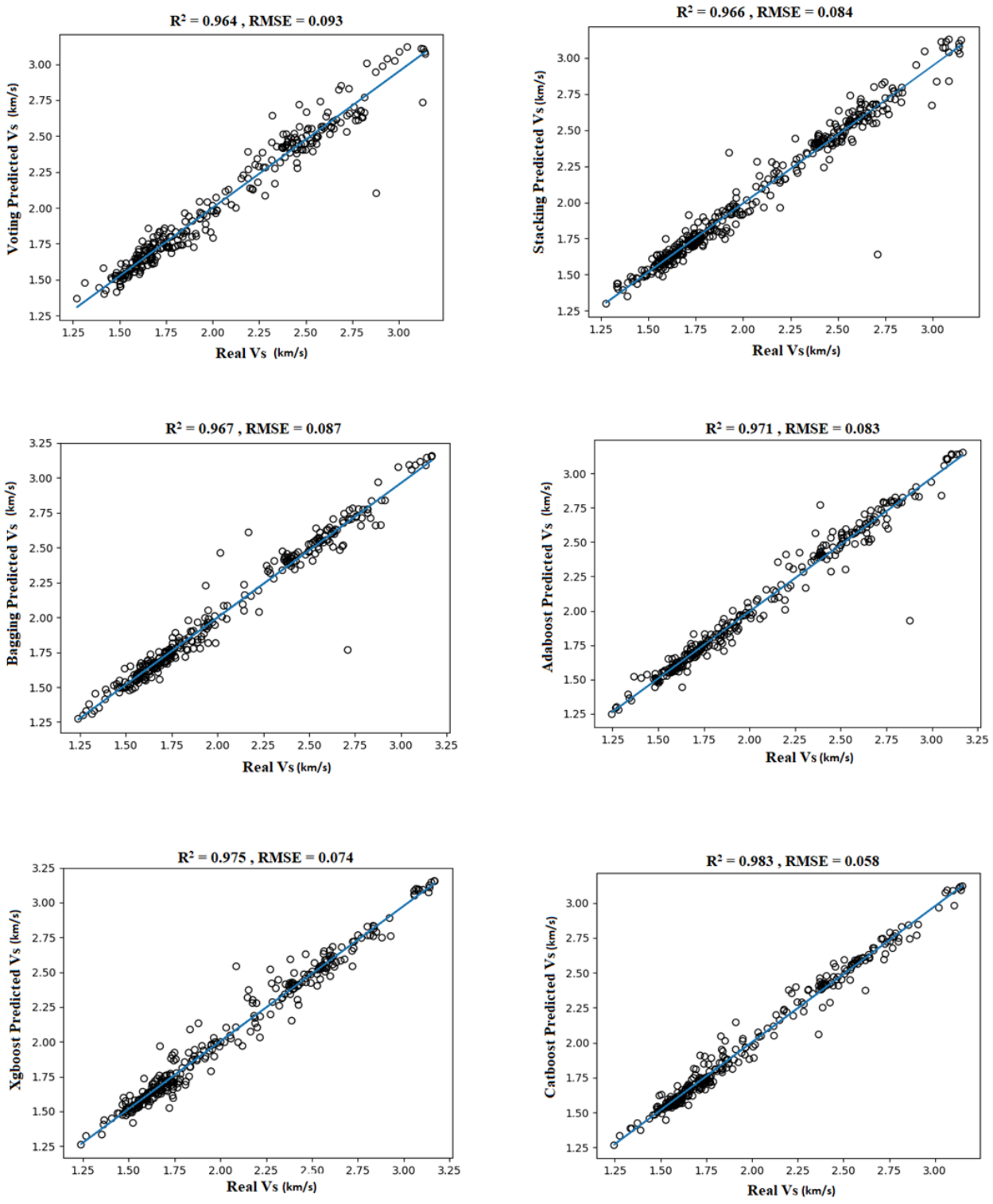
Table 3. Comparison of root mean square error and correlation coefficient related to training and testing data for conventional models and hybrid methods

	Method	Training Result		Testing Result	
		R^2	RMSE	R^2	RMSE
Shear Wave	Linear regression	0.842	0.267	0.836	0.244
	SVR	0.904	0.203	0.891	0.188
	KNN	0.933	0.176	0.916	0.165
	Decision tree	0.911	0.197	0.904	0.176
	ANN	0.968	0.118	0.962	0.124
	ANFIS	0.956	0.145	0.941	0.138
	GA-ANN	0.945	0.157	0.942	0.152
	PSO-ANN	0.966	0.122	0.960	0.129

جدول ۴- مقایسه ریشه میانگین مربعات خطا (RMSE) و ضریب همبستگی (R2) مربوط به داده‌های آموزشی و آزمایشی برای مدل‌های یادگیری جمعی

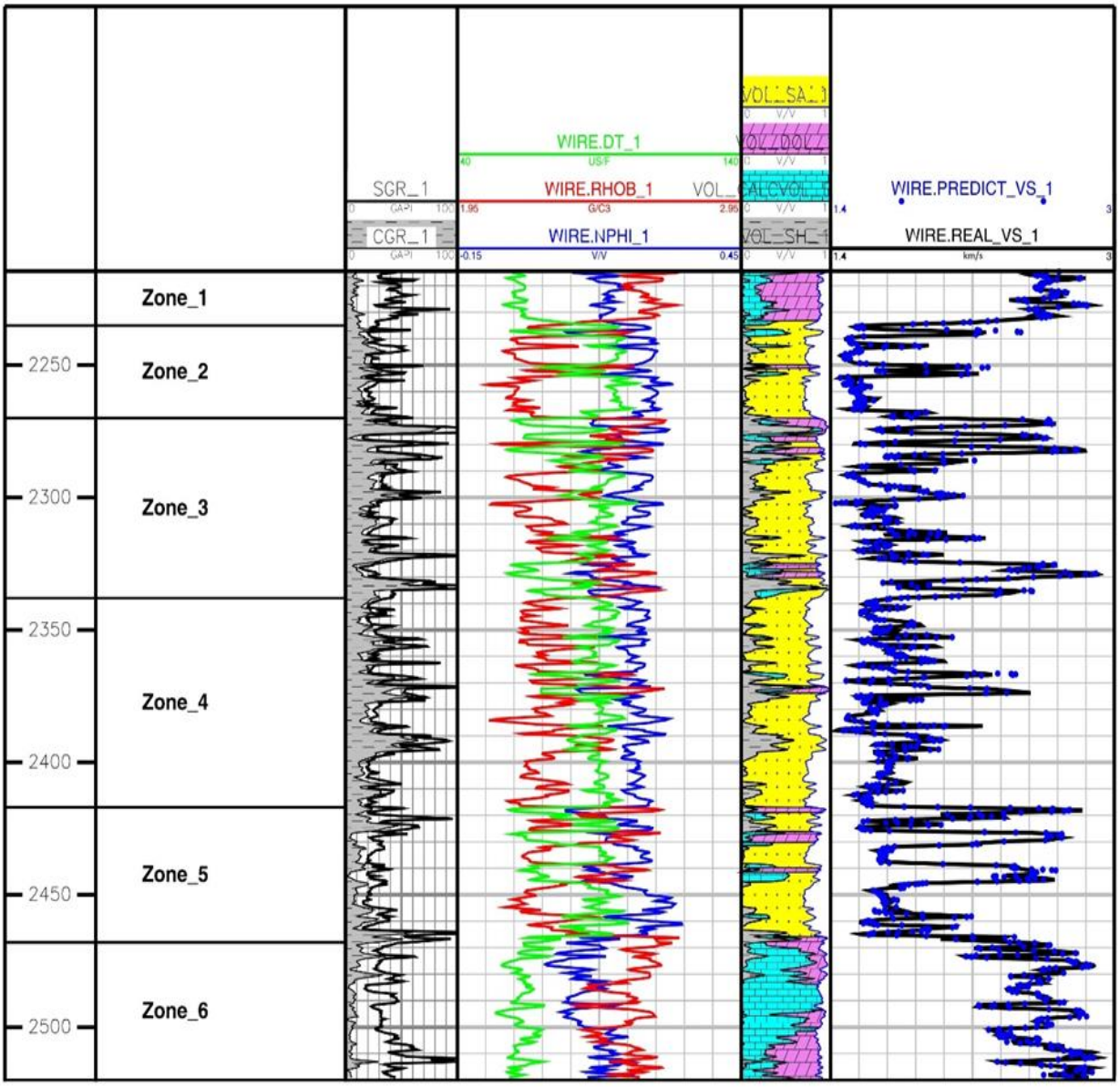
Table 4. Comparison of root mean square error and correlation coefficient related to training and testing data for ensemble learning models

	Method	Training Result		Testing Result	
		R^2	RMSE	R^2	RMSE
Shear wave	Voting	0.973	0.078	0.964	0.093
	Stacking	0.974	0.073	0.966	0.084
	Bagging	0.993	0.033	0.967	0.087
	Adaboost	0.999	0.005	0.971	0.083
	Xgboost	0.999	0.001	0.975	0.074
	Catboost	**0.998**	**0.019**	**0.983**	**0.058**



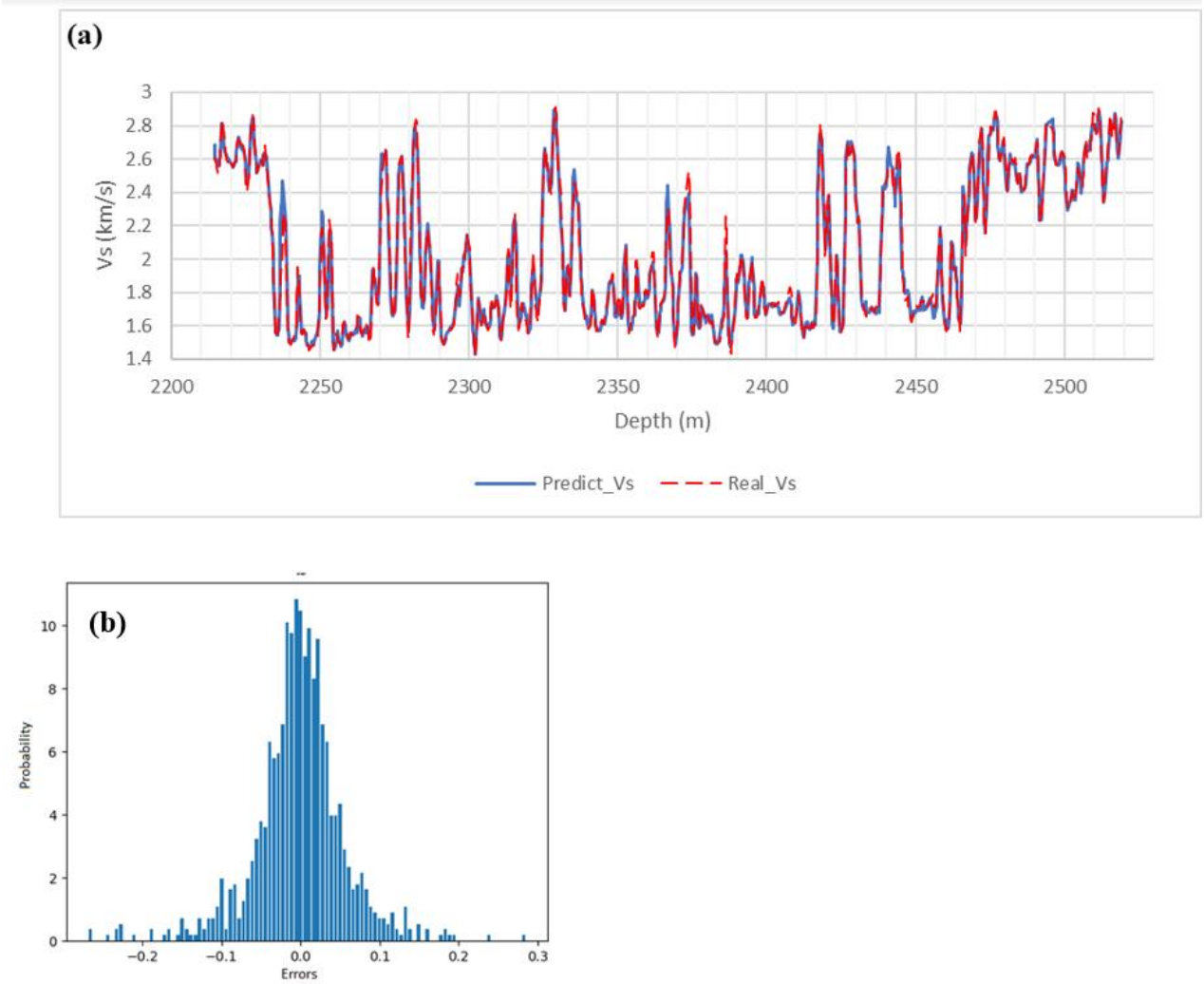
شکل ۵- مقایسه سرعت موج برشی واقعی (Real) و مقادیر تخمین زده شده (Predicted) توسط روش‌های یادگیری جمعی برای داده آزمایش

Fig. 5. Comparison of real and predicted shear wave velocity ensemble learning for test data



شکل ۶- تطابق خوب بین مقادیر پیش‌بینی شده (Predicted) و واقعی (Real) موج برشی توسط روش Catboost برای داده آزمایش

Fig. 6. The Catboost model demonstrates a strong correlation between predicted and actual shear wave velocity values for the test data.



شکل ۷- (a) مطابقت موج برشی پیش‌بینی شده (Predicted) توسط مدل کت بوست و مقادیر واقعی موج برشی (Real) در داده‌های آزمایش، (b) هیستوگرام خطای مربوط به روش کت بوست.

Fig. 7. (a) Correspondence of the predicted shear wave by the Catboost model and the real shear wave in the test data, (b) Error histogram of the Catboost method.

۵- نتیجه‌گیری

نتایج نشان داد که استراتژی یادگیری جمعی، روشی دقیق، سریع و مقرون به صرفه برای پیش‌بینی اهداف ارائه داد. هر یک از مدل‌های یادگیری جمعی مورد استفاده برای پیش‌بینی سرعت موج برشی، مفاهیم، روش‌ها و ساختار متفاوتی برای حل مسائل دارند، اما نتایج این مطالعه بسیار قابل اعتماد و نزدیک به یکدیگر بود که شاخص خوبی برای تأیید مفاهیم اساسی آنها برای حل مسائل است. مدل‌های مرسوم مثل شبکه‌های عصبی و مدل‌های هیبریدی در پیش‌بینی موج برشی عملکرد نسبتا خوبی ارائه دادند. بین مدل‌های مرسوم، شبکه عصبی بهترین نتایج را ارائه داد. خطای اندازه‌گیری شده برای شبکه عصبی با استفاده از پارامتر ارزیابی RMSE در چاه تست، میزان ۰,۱۲۴ ثبت شد. اما تمامی روش‌های یادگیری جمعی میزان RMSE کمتر از ۰,۱ را نشان دادند. همچنین الگوریتم‌های یادگیری جمعی نسبت به روش‌های مرسوم ضریب همبستگی بالاتری نشان دادند. نتایج نشان داد آموزش مدل‌های تجمیعی مثل رای‌گیری و برانبارش می تواند ضعف روش‌های انفرادی معمول مثل رگرسیون بردار پشتیبان، رگرسیون خطی و الگوریتم

نزدیک‌ترین همسایه را پوشش دهد و نتایج بهتری در پیش‌بینی موج برشی حاصل شود. در بین تمامی روش‌های یادگیری ماشینی، Catboost نسبت به سایر الگوریتم‌ها عملکرد بهتری داشت. روش معرفی شده در این مطالعه قادر به تخمین سرعت امواج برشی برای چاه‌هایی است که داده‌های تصویرگر صوتی دو قطبی ندارند. همچنین این الگوریتم می‌تواند به عنوان روشی با قدرت اطمینان و تعمیم بالا برای برآورد سایر خواص مخزن استفاده شود.

مراجع

Abedi, R., Costache, R., Shafizadeh-Moghadam, H., Pham, Q.B., 2022. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto International 37, 5479-5496. https://doi.org/10.108 /10106049.2021.1920636.

Al-Dousari, M., Garrouch, A.A., Al-Omair, O., 2016. Investigating the dependence of shear wave velocity on petrophysical parameters. Journal of Petroleum Science and Engineering 146, 286-296. https://doi.org/10.1016/ j.petrol.2016.04.036.

Anemangely, M., Ramezanzadeh, A., Tokhmechi, B., 2017. Shear wave travel time estimation from petrophysical logs using ANFIS-PSO algorithm: A case study from Ab-Teymour Oilfield. Journal of Natural Gas Science and Engineering 38, 373-387. https://doi.org/10.1016/j.jngse.2017.01.003.

Anifowose, F.A., Labadin, J., Abdulraheem, A., 2017. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. Journal of Petroleum Science and Engineering 151, 480-487. https://doi.org/10.1016 /j.petrol.2017.01.024.

Anselmetti, F.S., Eberli, G.P., 1993. Controls on sonic velocity in carbonates. Pure and Applied geophysics 141, 287-323. https://doi.org/10.1007/BF00998333.

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123-140. https://doi.org/10.1007/BF00058655.

Castagna, J.P., Batzle, M.L., Eastwood, R.L., 1985. Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks. geophysics 50, 571-581. https://doi.org/10.1190/1.1441933.

Chang, J.F., Dong, N., Ip, W.H., Yung, K.L., 2019. An ensemble learning model based on Bayesian model combination for solar energy prediction. Journal of Renewable and Sustainable Energy 11, 043702. https://doi.org/10.1063/1.5094534.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 785-794. https://doi.org/10.1145/2939672.2939785.

Da Silva, R.G., Ribeiro, M.H.D.M., Moreno, S.R., Mariani, V.C., dos Santos Coelho, L., 2021. A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting. Energy 216, 119174. https:// doi.org/10.1016/j.energy.2020.119174.

Eberhart-Phillips, D., Han, D.H., Zoback, M.D., 1989. Empirical relationships among seismic velocity, effective pressure, porosity, and clay content in sandstone. Geophysics 54, 82-89. https://doi.org/10.1190/1.1442580.

Eskandari, H., Rezaee, M.R., Mohammadnia, M., 2004. Application of multiple regression and artificial neural network techniques to predict shear wave velocity from wireline log data for a carbonate reservoir South-West Iran. CSEG recorder 42, 48. https://doi.org/10.4236/ojg.2014.47023.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389-422. https://doi.org/10.1023/A:1012487302797.

Jabeur, S.B., Gharib, C., Mefteh-Wali, S., Arfi, W.B., 2021. CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change 166, 120658. https://doi.org/10.1016/j.techfore.2021.120658.

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., Homayouni, S., 2021. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. Remote Sensing 13, 4405. https://doi.org/10.3390/rs13214405.

Kavianpor Sangno, M., Namdarian, A., Mousavi-Harami, S.R., Mahboubi, A. and Omidpour, A., 2015. The Study of Role and Texture of Anhydrite in Production Zone of Asmari Formation in Mansuri Oil Field, Zagros, Iran. Scientific Quarterly Journal of Geosciences 24, 203-216. https://doi.org/10.22071/gsj.2015.42659.

Khandelwal, M., Singh, T.N., 2010. Artificial neural networks as a valuable tool for well log interpretation. Petroleum Science and Technology 28, 1381-1393. https://doi.org/10.1080/10916460903030482.

Koesoemadinata, A.P., McMechan, G.A., 2001. Empirical estimation of viscoelastic seismic parameters from petrophysical properties of sandstoneFrom Petrophysical to Seismic Parameters. Geophysics 66, 1457-1470. https ://doi.org/10.1190/1.1487091.

Loh, W.Y., 2011. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery 1, 14-23. https://doi.org/10.1002/widm.8.

Maleki, S., Moradzadeh, A., Riabi, R.G., Gholami, R., Sadeghzadeh, F., 2014. Prediction of shear wave velocity using empirical correlations and artificial intelligence methods. NRIAG Journal of Astronomy and Geophysics 3, 70-81. https://doi.org/10.1016/j.nrjag.2014.05.001.

Maulud, D., Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends 1, 140-147. https://doi.org/1038094/jastt1457.

Nourafkan, A., Kadkhodaie-Ilkhchi, A., 2015. Shear wave velocity estimation from conventional well log data by using a hybrid ant colony–fuzzy inference system: A case study from Cheshmeh–Khosh oilfield. Journal of Petroleum Science and Engineering 127, 459-468. https://doi.org/10.1016/j.petrol.2015.02.001.

Onalo, D., Oloruntobi, O., Adedigba, S., Khan, F., James, L., Butt, S., 2019. Dynamic data driven sonic well log model for formation evaluation. Journal of Petroleum Science and Engineering 175, 1049-1062. https://doi.org/10.1016/j.petrol. 2019. 01.042.

Peng, Y., 2006. A novel ensemble machine learning for robust microarray data classification. Computers in Biology and Medicine 36, 553-573. https://doi.org/10.1016/j.compbiomed.2005.04.001.

Polikar, R., 2012. Ensemble learning. Ensemble machine learning: Methods and applications, pp. 1-34. https://doi.org/10.1007/978-1-4 419-9326-7_1.

Rajabi, M., Bohloli, B., Ahangar, E.G., 2010. Intelligent approaches for prediction of compressional, shear and Stoneley wave velocities from conventional well log data: A case study from the Sarvak carbonate reservoir in the Abadan Plain (Southwestern Iran). Computers & Geosciences 36, 647-664. https://doi.org/10. 101 6/j.cageo. 2009.09.008.

Rezaee, M.R., Ilkhchi, A.K., Barabadi, A., 2007. Prediction of shear wave velocity from petrophysical data utilizing intelligent systems: An example from a sandstone reservoir of Carnarvon Basin, Australia. Journal of Petroleum Science and Engineering 55, 201-212. https://doi.org/10.1016/j.petrol.2006.08.008.

Sumayli, A., 2023. Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. Arabian Journal of Chemistry 16, 104833. https://doi.org/10.1016/j.arabjc.2023.104833.

Tosaya, C., Nur, A., 1982. Effects of diagenesis and clays on compressional velocities in rocks. Geophysical Research Letters 9, 5-8. https://doi.org/10.1029/GL009i001p00005.

Tüysüzoğlu, G.Ö.K.S.U., Birant, D., 2020. Enhanced bagging (eBagging): A novel approach for ensemble learning. International Arab Journal of Information Technology 17, 515-528. https://doi.org/10.34028/iajit/17/4/10.

Wang, J., Cao, J., Yuan, S., 2020. Shear wave velocity prediction based on adaptive particle swarm optimization optimized recurrent neural network. Journal of Petroleum Science and Engineering 194, 107466. https://doi.org/10.1016/ j. petrol.2020. 107466.

Wolpert, D.H., 1992. Stacked generalization. Neural networks 5, 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Xu, S.B., Huang, S.Y., Yuan, Z.G., Deng, X.H., Jiang, K., 2020. Prediction of the dst index with bagging ensemble-learning algorithm. The Astrophysical Journal Supplement Series 248, 14. https://doi.org/10.3847/1538-4365/ab880e.

Yu, X., Zhang, X., Qin, H., 2018. A data-driven model based on Fourier transform and support vector regression for monthly reservoir inflow forecasting. Journal of Hydro-environment Research 18, 12-24. https://doi.org/10.1016/j.jher.2017.10.005.

Zahmatkesh, I., Kadkhodaie, A., Soleimani, B., Golalzadeh, A., Azarpour, M., 2018. Estimating Vsand and reservoir properties from seismic attributes and acoustic impedance inversion: A case study from the

Mansuri oilfield, SW Iran. Journal of Petroleum Science and Engineering 161, 259-274. https://doi.org/10.1016/j.petrol.2017.11.060.

Zhang, Y., Zhao, Z., Zheng, J., 2020. CatBoost: A new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. Journal of Hydrology 588, 125087. https://doi.org/10.1016/j. jhydrol.2020.125087.