تخمین دبی نفت تولیدی از چاه به‌وسیله روش‌های یادگیری ماشین با استفاده از داده‌های پمپ الکتریکی شناور (ESP)

محمدباقر صدیقی[۱]، مجید سیاوشی[۲*]، روح الدین میری[۳]

۱- دانشکده مهندسی مکانیک، دانشگاه علم و صنعت ایران

۲- دانشیار، دانشکده مهندسی مکانیک، دانشگاه علم و صنعت ایران

۳- استادیار، دانشکده مهندسی شیمی، نفت و گاز، دانشگاه علم و صنعت ایران

* نویسنده مسئول: msiavashi@iust.ac.ir

چکیده

تخمین دبی جریان در چاه‌های یک میدان نفتی، یک فرایند حیاتی و کاربردی است. بااین‌حال جریان‌های استخراج‌شده از چاه‌های نفتی، چند فازی بوده و تخمین دقیق دبی آن‌ها، بسیار چالش‌برانگیز و پرهزینه است. دبی سنج‌های مجازی در مقایسه با دبی سنج‌های چند فازی و روش‌های چاه آزمایی، ازنظر اقتصادی گزینه بسیار مناسبی هستند که با بهره‌گیری از داده‌های موجود و استفاده از روش‌های هوش مصنوعی، قادر به پیش‌بینی دقیق دبی در آینده هستند؛ بنابراین، اخیراً به دبی سنج‌های مجازی داده محور توجه زیادی شده است. در این مقاله تخمین دبی تولیدی یک چاه با استفاده از سه روش یادگیری ماشین: ۱- k همسایه نزدیک‌تر (k-NN)؛ ۲- تقویت گرادیان (GBR)؛ و ۳- درخت تصمیم (DT)، با استفاده از داده‌های پمپ انجام‌شده است. به‌منظور انتخاب ویژگی‌های مناسب به‌عنوان ورودی روش‌ها، از تحلیل‌های آماری پیرسون و اسپیرمن استفاده‌شده است. مجموعه داده موردبررسی مربوط به یکی از چاه‌های یک میدان نفتی در جنوب ایران است. مجموعه داده موجود دارای حجم کم و تنوع ناکافی است، اما بااین‌وجود نتایج نشان می‌دهند که روش‌های پیشنهادی عملکرد مناسبی دارند. روش k-NN با دقت ۰.۹۴۹۴ نسبت به دو روش دیگر عملکرد بهتری در تخمین دبی نفت داشته است. برای بررسی عملکرد روش‌ها در برابر داده‌های دارای نویز، یک درصد انحراف معیار نویز به داده‌های ورودی اضافه شد. بررسی‌ها نشان داد که مدل k همسایه نزدیک‌تر با دقت ۰.۹۲۵۷ در مقایسه با دو روش دیگر عملکرد بهتری داشته و کمترین تأثیر را از نویزها گرفته است.

کلیدواژگان

دبی سنج مجازی داده محور؛ یادگیری ماشین؛ k همسایه نزدیک‌تر ؛ تقویت گرادیان؛ درخت تصمیم.

Estimation of oil flow production of well employing machine learning algorithms using electrical submersible pump (ESP) data

Abstract

Estimating the flow rate in oil wells of a field is a vital and practical process. However, the flows extracted from oil wells are multiphase, and their accurate estimation is highly challenging and costly. Virtual flow meters, compared to multiphase flow meters and well-testing methods, are an economically viable option that can accurately predict future flow rates by leveraging existing data and artificial intelligence algorithms. Therefore, data-driven virtual flow meters have recently received significant attention. This paper estimates the production flow rate of a well using three machine learning algorithms: 1- k-nearest neighbors; 2- gradient boosting; and 3- decision tree, using pump data. Pearson and

Spearman statistical analyses were used to select appropriate features as the algorithm inputs. The dataset under investigation pertains to one of the wells of a southern oil field in Iran. The available dataset has a small volume and insufficient diversity, but despite this, the results show that the proposed algorithms perform well. The k-NN method, with an accuracy of 0.9494, performed better than the other two methods in estimating oil flow rate. To examine the performance of the algorithms against noisy data, one percent of standard deviation noise was added to the input data. The investigations showed that the k-NN model, with an accuracy of 0.9257, performed better than the other two methods and was least affected by the noise.

Keywords
Data driven virtual flow meter, machine learning, k-nearest neighbor, gradient boosting, decision tree

۱- مقدمه

آگاهی از نرخ تولید سیالات استخراج‌شده از یک چاه در صنعت نفت و گاز به‌منظور بهینه‌سازی تولید و مدیریت مخزن ضروری است [۱]. روش‌های چاه آزمایی و استفاده از دبی سنج چند فازی (MPFM[1]) دو روش متداول برای اندازه‌گیری دبی در میدان‌های نفتی به‌حساب می‌آیند. بااین‌حال این روش‌ها به دلیل نیاز به نصب دستگاه‌های فیزیکی ازلحاظ عملیاتی پرهزینه هستند. پژوهشگران باهدف یافتن راهی برای کاهش هزینه‌های عملیاتی برداشت نفت و گاز، دبی سنج‌های مجازی (VFM[2]) را توسعه دادند. دبی سنج مجازی درواقع یک بسته نرم‌افزاری است که بدون نیاز به تجهیزات فیزیکی و تنها با استفاده از داده‌های در دسترس چاه در بخش‌های مختلف (مانند فشار و دما)، دبی جریان را پیش‌بینی می‌نماید. هزینه نصب، راه‌اندازی، نگهداری و اجرای دبی سنج‌های مجازی در مقایسه با سایر روش‌ها بسیار کمتر است [۲]. استفاده از فناوری VFM می‌تواند به عملکرد بهتر مجموعه‌های تولید نفت و گاز کمک کند، زیرا این فناوری به کاربران این امکان را می‌دهد تا با دقت و سرعت بیشتری به داده‌های نرخ جریان دسترسی داشته باشند و عملیات تولید را برای بهبود بازدهی بهینه کنند. همچنین دبی سنج مجازی می‌تواند از طریق محاسبه‌ی برخطِ نرخ جریانِ سیالات به کاربران در تشخیص و پیش‌بینی مشکلات چاه کمک کرده، فرآیند نگهداری از چاه را ارتقا داده و به‌تبع آن، نیاز به تعمیرات و هزینه‌ها را کاهش دهد.

دبی سنج‌های مجازی ازنظر رویکرد محاسبه دبی به دو دسته دبی سنج مجازی فیزیک محور و دبی سنج مجازی داده محور تقسیم می‌شوند [۳]. یکی از روش‌های محاسبه دبی سیال به‌صورت مجازی، استفاده از مدل‌سازی فیزیکی بخش‌های مختلف یک مجموعه تولید است. مدل‌های ارائه‌شده برای جریان‌های چند فازی بسیار پیچیده هستند و حل آن‌ها بدون ساده‌سازی امکان‌پذیر نیست. از طرفی حل این معادلات نیازمند صرف هزینه محاسباتی و زمان اجرا بالا است [۲]. در روش فیزیک محور، جریان سیال با استفاده از معادلات بقا تحلیل‌شده و دبی تخمین زده می‌شود. با توجه به اینکه معادلات حاکم بر جریان سیالات چند فازی پیچیده هستند و حل تحلیلی برای برخی از آن‌ها وجود ندارد، درنتیجه معمولاً برای حل این معادلات ساده‌سازی زیادی بر روی آن‌ها انجام می‌شود یا از روش‌های حل عددی استفاده می‌شود که همین موجب بروز خطا و افزایش هزینه محاسبات می‌شود. عملکرد این نوع دبی سنج‌ها وابستگی زیادی به مقدار فشار، حجم و دما دارد. تنظیم این نوع دبی سنج‌ها کار دشواری است زیرا در این روش کاربر باید دانش کافی در زمینه عملیاتی و نرم‌افزاری داشته باشد [۲].

دومین رویکرد در دبی سنج‌های مجازی استفاده از داده‌های میدانی برای تخمین دبی است. در این روش تلاش بر این است که با استفاده از داده‌های جمع‌آوری‌شده از میدان شامل داده‌های به‌دست‌آمده از حسگرهای درون‌چاهی و چاه آزمایی، یک مدل ریاضی به کمک روش‌های پیشرفته یادگیری ماشین جهت تخمین دبی ایجاد می‌شود [۲]. اگر مدل مبتنی برداده به‌خوبی آموزش داده شود می‌تواند دبی فازها را به‌صورت آنی و دقیق پیش‌بینی کند. مزیت اصلی این روش نسبت به دبی سنج‌های مجازی فیزیک محور این است که از مدل‌سازی‌های فیزیکی-دقیق سیستم‌ها استفاده نمی‌کنند زیرا این مدل‌سازی‌ها معمولاً غیرخطی بوده و حل دقیق آن‌ها بسیار دشوار است. روش‌های داده محور بر این اصل تکیه‌دارند که داده‌های تجربی و عملکردی، سیستم را به‌خوبی نمایندگی می‌کنند و سعی می‌کنند روابط فیزیکی سیستم را مستقیماً از داده‌ها یاد بگیرند.

[1] Multi-Phase Flow Meter
[2] Virtual Flow Meter

القطامی و همکاران [۴] یک سیستم دبی سنج مجازی مبتنی بر یادگیری گروهی ایجاد کردند. آن‌ها تلاش کردند با دست‌کاری داده‌های آموزشی، معماری شبکه عصبی و مسیر یادگیری یادگیرندگان، شبکه عصبی متنوعی را تولید کنند. هدف از این کار این بود که عملکرد دبی سنج‌های مجازی مبتنی برداده در میدان‌هایی با داده‌های تولیدشده بسیار محدود بهبود یابد. دبی سنج مجازی توسعه‌یافته دقت بسیار مطلوبی دارد. این سیستم می‌تواند دبی فاز مایع و گاز را به ترتیب با میانگین خطا ۴.۷ و ۲.۴ درصد تخمین بزند. القطامی و همکاران [۵] یادگیری یک مجموعه ترکیبی را با ترکیب شبکه عصبی و رویکردهای درخت رگرسیون[1] موردبحث قراردادند. ایشان رویکرد ترکیبی (NN-RTs[2]) را با رویکردهای مجموعه همگن (NN و RTs) مقایسه کردند و نشان دادند روش ترکیبی عملکرد دقیق‌تری را نشان می‌دهد. احمدی و همکاران [۶] در یک مسئله جریان سنجی چند فازی، روش جدیدی برای پیش‌بینی نرخ نفت تولیدی چاه‌ها با استفاده از منطق فازی، شبکه‌های عصبی مصنوعی و روش رقابت امپراتوری (ICA)[3] ارائه کردند. دماها و فشارهای خطوط به‌عنوان متغیر ورودی شبکه و نرخ جریان نفت به‌عنوان متغیر خروجی تنظیم شدند. یک مجموعه داده متشکل از ۱۶۰۰ داده‌ی مربوط به ۵۰ چاه دریکی از میدان‌های نفتی واقع در شمال خلیج‌فارس برای ساخت پایگاه داده استفاده‌شده است. نتایج به‌دست‌آمده کارآیی، قدرت و سازگاری مدل ICA-ANN[4] را نشان دادند. بیک مخامدوف و جاشکه [۷] عملکرد روش تقویت گرادیان را به‌عنوان جایگزینی برای شبکه‌های عصبی در یک مجموعه دبی سنج مجازی بررسی کردند. آن‌ها نشان دادند که این سازوکار به‌عنوان پشتیبان دبی سنج چند فازی و یا به‌طور مستقل حتی برای مجموعه داده آموزشی کوچک عملکرد خوبی دارد. نتایج همچنین نشان داد که با ترکیب روش تقویت گرادیان با اندازه‌گیری‌های دبی آزمایش چاه در محدوده عملیاتی گسترده‌ای از چاه، می‌توان پیش‌بینی دقیق دبی را از مرحله تولید اولیه انجام داد.

گوئز و همکاران [۸]، هاتودت و همکاران [۹] و العجمی و همکاران [۱۰] از داده‌های شیر فشارشکن برای توسعه مدل داده محور استفاده کردند. مدل ایجادشده توسط گوئز و همکاران دبی هر چاه را به‌عنوان تابعی از داده‌های جمع‌آوری‌شده مانند دما و فشار شیر فشارشکن و ویژگی سیالات در میدان نفتی بیان می‌کند. نتایج نشان می‌دهد که برای یک دوره ۲۴ ساعته، خطاهای نسبی بین نرخ جریان پیش‌بینی‌شده و اندازه‌گیری شده به ترتیب زیر ۳.۵ و ۳ درصد برای کل نرخ جریان نفت و گاز بود. هاتودت و همکاران یک دبی سنج مجازی جدید با ترکیب روش فیزیک محور و روش داده محور ایجاد کردند. در این پژوهش از رویکرد ترکیبی برای مدل‌سازی شیر فشارشکن استفاده‌شده است. شیر فشارشکن با مجموعه‌ای ساده از معادلات پایه و یک شبکه عصبی برای تخمین ضریب جریان دریچه نشان داده می‌شود. آن‌ها نشان دادند که رویکرد ترکیبی عملکرد مناسبی دارد و ممکن است مزایایی نسبت به دبی سنج‌های مجازی فیزیک محور و داده محور داشته باشد. العجمی و همکاران علاوه بر فشار، دما، اندازه شیر فشارشکن و داده‌های برش آب از برخی پارامترهای اضافی برای ورودی‌ها (ازجمله یک همبستگی تجربی برای جریان بحرانی شیر فشارشکن) استفاده کردند. در مقایسه با تخمین‌های دبی جریان به‌دست‌آمده با استفاده از همبستگی‌های تجربی شیر فشارشکن، شبکه عصبی (NN) عملکرد نسبتاً بهتری را نشان داد. لازم به ذکر است که مدل‌های شیر فشارشکن مورداستفاده در این مطالعه تجربی بودند.

[1] Decision Tree
[2] NN-RTE: Neural Network-Regression Trees
[3] Imperialist competitive algorithm
[4] ANN: Artificial Neural Networks

سندنس و همکاران [۱۱] یک دبی سنج مجازی مبتنی برداده با استفاده از معماری یادگیری چندوظیفه‌ای برای مطالعه ۵۵ حلقه چاه پیشنهاد کردند. نتایج با دو مدل پایه تک‌کاره مقایسه شده است. آن‌ها نشان دادند که الگوریتم یادگیری چند وظیفه‌ای استحکام را نسبت به روش‌های تک‌وظیفه‌ای، بدون به خطر انداختن عملکرد، بهبود می‌بخشد. این الگوریتم به‌طور متوسط برای دارایی‌هایی که در آن معماری تک‌وظیفه‌ای با مشکل مواجه هستند، خطا را به میزان ۲۵ تا ۵۰ درصد کاهش می‌دهد.

الجاسمی و همکاران [۱۲] یک مدل داده محور با استفاده از شبکه‌های عصبی برای پیش‌بینی دبی مایع و برش آب در یک مخزن کربناته دریایی با بیش از ۲۰ درصد برش آب را ارائه دادند. شبکه عصبی با استفاده از داده‌های تولید واقعی، داده‌های سطحی و ته-چاهی آموزش داده شد. داده‌هایی که برحسب زمان بودند به‌عنوان سری زمانی در نظر گرفته شدند تا به کاربران اجازه دهد سناریوهایی را با تغییر عملیات چاه ایجاد کنند. این رویکرد با تغییر متغیرهای کنترلی مانند فشار سر چاه و فرکانس پمپ، نتایج را شبیه‌سازی می‌کند. تغییر فشار سر چاه و فرکانس، به کاربران اجازه می‌دهد تولید را پیش‌بینی کرده و رویدادهای منفی پمپ را پیش‌بینی کنند. باوجود داده‌های محدود مخزن، نتایج نشان می‌دهد که شبکه‌های عصبی دقت قابل قبولی در پیش‌بینی دبی نفت و برش آب داشته است.

علاوه بر برنامه‌های کاربردی VFM توسط شبکه عصبی، شرکت بیکر هیوز[1] نرم‌افزار نئورال فلو[2] را توسعه داده است که مبتنی بر مدل شبکه عصبی است [۱۳]. این نرم‌افزار برای تخمین دبی در سیستم‌های دارای پمپ‌های شناور الکتریکی[3] (ESP) با استفاده از رویکرد شبکه عصبی استفاده می‌شود. مشابه شبکه‌های عصبی‌ای که قبلاً موردبحث قرار گرفت، این سازوکار فشار سر چاه، فشار ورودی و فشار تخلیه پمپ و همچنین سایر پارامترهای اندازه‌گیری شده مانند فرکانس و جریان در پمپ را به‌عنوان ورودی در نظر می‌گیرد و تخمین دبی را به‌عنوان خروجی شبکه تولید می‌کند.

در تحقیقات گذشته از پارامترهای مختلف درون‌چاهی مانند فشار و دمای سرچاهی و ته چاهی، فشار خط، کسر حجمی فازها و اطلاعات شیر فشارشکن مانند میزان بازشدگی شیر، فشار قبل و بعد شیر، دمای قبل و بعد شیر و ضریب تخلیه شیر ... برای تخمین نرخ دبی‌ها استفاده‌شده است و کمتر به تخمین دبی با استفاده از پمپ ESP به‌عنوان یک مدل داده محور پرداخته‌شده است. درحالی‌که تمامی این اطلاعات در چاه‌ها به‌صورت بر خط در دسترس نیست. ازآنجاکه اکثر چاه‌های ایران وارد دوره کاهش برداشت‌شده و برای استخراج نفت نیاز به پمپ ESP وجود دارد، بنابراین اطلاعات این پمپ‌ها در اکثر چاه‌ها موجود است و تخمین دبی از این طریق می‌تواند بسیار ساده و مفید باشد.

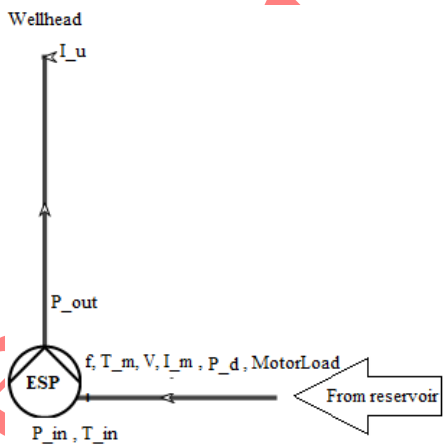
لازم به ذکر است که در کارهایی که دبی سنجی با اطلاعات پمپ انجام‌شده (مانند پژوهش دنی و همکاران [۱۳] که موجب ایجاد یک نرم‌افزار تجاری شده است) نیز از مدل‌های شبکه عصبی استفاده‌شده است. این درحالی است که چنین روش‌هایی به حجم بالایی از داده با تنوع بالا نیاز دارد که فراهم کردن این مورد در بسیاری از چاه‌ها ممکن نیست. این مشکل یک چالش در استفاده از روش‌های شبکه عصبی در چاه‌هایی است که حجم بالا داده با تنوع زیاد را تأمین نمی‌کنند. این تحقیق باهدف تخمین دبی نفت در یکی از چاه‌های مربوط به یک میدان نفتی واقع در جنوب ایران انجام‌شده است. چاه موردبررسی از حجم داده بالا و تنوع کافی برخوردار نیست بنابراین، مدل‌های شبکه عصبی در این چاه با خطا قابل‌توجهی

[1] Baker Hughes
[2] NeuraFlow
[3] Electrical Submersible Pump

روبه‌رو هستند. به همین دلیل در این مطالعه، از روش‌های k همسایه نزدیک‌تر (k-NN)[1]، درخت تصمیم (DT)[2] و تقویت گرادیان (GBR)[3] استفاده‌شده است. همچنین، داده‌های نویز دار که ممکن است به دلیل خرابی حسگرها در مجموعه داده پدیدار شوند، می‌توانند تأثیر معکوس قابل‌توجهی بر دقت و قابلیت اعتماد مدل‌ها داشته باشند. این نوع داده‌ها می‌توانند تأثیر منفی بر عملکرد دبی سنج مجازی داده محور داشته باشند؛ بنابراین، عملکرد دبی سنج مجازی داده محور در برابر داده‌های نویز دار نیز موردبررسی قرارگرفته است. در بخش دوم مدل‌های مورداستفاده توضیح داده‌شده‌اند. بررسی مجموعه داده در بخش سه آورده شده است. نتایج در بخش ۴ آورده شده و درنهایت نتیجه‌گیری در بخش ۵ آمده است.

۲- تعریف مسئله

شماتیک چاه موردمطالعه در شکل ۱ آمده است. در این تحقیق فقط پمپ ESP مدل‌سازی شده است. این منجر به پیچیدگی کمتر مدل می‌شود و نیاز به داده‌های ته چاه و سر چاه را از بین می‌برد. این کار برای دارایی‌هایی که اندازه‌گیری‌های ته چاه و یا بالای چاه وجود ندارد یا معیوب هستند، سودمند است. شکل ۱ پارامترهای اندازه‌گیری شده پمپ برای تخمین دبی را نشان می‌دهد. این پارامترها شامل جریان کنتور کنترل‌کننده (I_u)، فشار ورودی پمپ (P_{in})، فشار خروجی پمپ (P_{out})، فشار تخلیه (P_d)، دمای ورودی به پمپ (T_{in})، دمای موتور (T_m)، جریان موتور (I_m)، ولتاژ (V)، فرکانس (f) و بار موتور ($MotorLoad$) هستند.



شکل ۱. شماتیک چاه موردمطالعه

برای ساخت یک دبی سنج مجازی داده محور ابتدا باید داده‌های مختلف جمع‌آوری شوند سپس از بین این داده‌ها آن‌هایی که ارتباط قوی‌تری با تابع هدف[4] دارند به‌عنوان ورودی انتخاب‌شده و آنگاه پیش‌پردازش روی آن‌ها انجام می‌گیرد یعنی داده‌های پرت حذف و نویزها فیلتر می‌شوند. در گام بعدی داده‌های پیش‌پردازش شده وارد مرحله توسعه مدل می‌شود. در این قسمت مجموعه داده به دو قسمت آموزش و آزمون تقسیم می‌گردد؛ روش سعی می‌کند با استفاده از داده‌های آموزشی یک الگوریتم تکرارشونده را بدون دخالت انسان یاد بگیرد. درنهایت مدل با استفاده از داده‌های بخش آزمون اعتبار سنجی می‌شود. اطلاعات مربوط به مدل داده محور در بخش ۳ آمده است.

[1] k-Nearest Neighbor.
[2] Decision Tree.
[3] gradient boosting

[4] تابع هدف پارامتری است که قرار است مقدار آن تخمین زده شود

۳- مدل داده محور

پیشرفت‌های سریع و شگفت‌انگیز روش‌های یادگیری ماشین باعث شده این روش‌ها در صنایع مختلف موردتوجه پژوهشگران قرار گیرد. اصل اساسی یادگیری ماشینی این است که رایانه‌ها را قادر می‌سازد از داده‌ها یاد بگیرند و رفتار خود را بر این اساس تطبیق دهند. این شامل آموزش روش‌های حل با استفاده از مجموعه داده‌های بزرگ برای شناسایی الگوها، همبستگی‌ها و بینش‌هایی است که می‌توانند برای پیش‌بینی یا تصمیم‌گیری دقیق استفاده شوند. در پژوهش حاضر از سه روش DT، k-NN و روش GBR به‌عنوان مدل‌های داده محور برای پیش‌بینی نرخ جریان استفاده‌شده است. در ادامه مراحل ایجاد مدل داده محور و همچنین پیشینه هریک از روش‌های انتخاب‌شده شرح داده می‌شود.

۳-۱- انتخاب ویژگی

در دیتاست‌های داده محور پارامترهای ورودی که به‌عنوان ویژگی‌ها یا متغیرها شناخته می‌شوند، بلوک‌های سازنده مدل‌های یادگیری ماشین هستند. این پارامترها ویژگی‌های داده‌ها را نشان می‌دهند که روش حل برای پیش‌بینی یا طبقه‌بندی از آن‌ها استفاده می‌کند. انتخاب پارامترهای ورودی مناسب می‌تواند به بهبود عملکرد مدل کمک کند، درحالی‌که پارامترهای نادرست یا نامربوط می‌تواند منجر به نتایج غیر بهینه و برازش بیش‌ازحد شود. انتخاب پارامترهای ورودی مناسب برای روش یادگیری ماشین می‌تواند یک فرایند پیچیده و چالش‌برانگیز باشد، زیرا داده‌ها ممکن است حاوی متغیرهای متعددی باشند که برخی از آن‌ها ممکن است نامربوط، زائد یا نویز دار باشند. با انتخاب ویژگی‌های مرتبط‌تر، می‌توانیم از تأثیرپذیری مدل از نویز و اطلاعات نامربوط داده‌ها جلوگیری کنیم که می‌تواند به تعمیم بهتر داده‌های دیده نشده کمک کند. همچنین، با تعداد ویژگی‌های کمتر، مدل به زمان و منابع کمتری برای آموزش نیاز دارد که آن را برای کاربردهای دنیای واقعی کارآمدتر و مقیاس‌پذیرتر می‌کند.

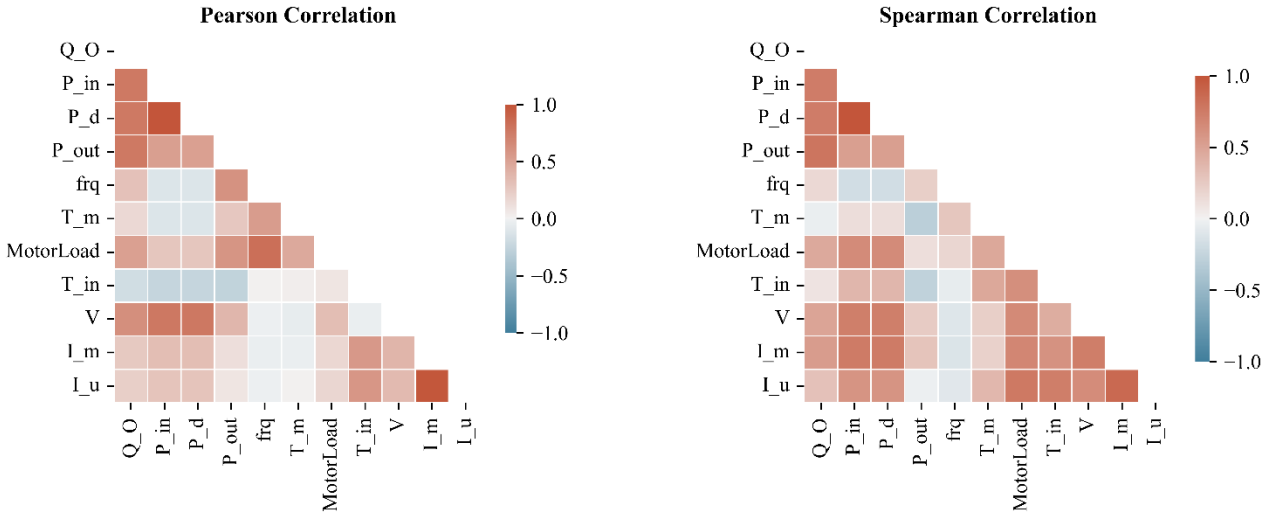
علاوه بر این، انتخاب ویژگی مناسب به افزایش تفسیرپذیری مدل کمک می‌کند. با تمرکز بر مهم‌ترین ویژگی‌ها، می‌توان به بینش‌هایی در مورد عواملی دست‌یافت که بیشترین تأثیر را بر پیش‌بینی‌های مدل دارند و برای درک و توضیح رفتار مدل برای ذینفعان و کارشناسان حوزه بسیار مهم است. روش‌های مختلفی برای انتخاب ویژگی مناسب وجود دارد که هرکدام نقاط قوت و ضعف خاص خود را دارند. روش‌های فیلتر یک رویکرد محبوب برای انتخاب ویژگی هستند، زیرا این روش‌ها از معیارهای آماری برای رتبه‌بندی و انتخاب ویژگی‌ها بر اساس ارتباط بین آن‌ها با متغیر هدف استفاده می‌کنند معیارهای رایج مورداستفاده در روش‌های فیلتر عبارتند از انتخاب ویژگی مبتنی بر همبستگی[1]، اطلاعات متقابل[2] و آزمون مجذور کای[3]. ویژگی‌ها بر اساس معیارهای از پیش تعریف‌شده مانند آستانه مشخصی از مقدار متریک انتخاب یا حذف می‌شوند.

ماتریس همبستگی اسپیرمن و پیرسون روش‌های آماری‌ای هستند که برای اندازه‌گیری قدرت و جهت رابطه بین دو متغیر استفاده می‌شوند. در زمینه انتخاب ویژگی، این روش‌ها می‌توانند برای شناسایی مرتبط‌ترین ویژگی‌ها برای یک کار پیش‌بینی معین استفاده شوند.

[1] Correlation-based
[2] mutual information
[3] Chi-Square

همبستگی پیرسون رابطه خطی بین دو متغیر پیوسته را اندازه‌گیری می‌کند، درحالی‌که همبستگی اسپیرمن رابطه یکنواخت بین دو متغیر را بدون توجه به خطی بودن یا نبودن آن اندازه‌گیری می‌کند. هر دو روش بینش‌های ارزشمندی را در مورد روابط بین ویژگی‌ها ارائه می‌دهند که می‌تواند برای شناسایی ویژگی‌های اضافی یا نامربوط برای پیش‌بینی استفاده شود.

هنگام استفاده از ماتریس همبستگی پیرسون و اسپیرمن برای انتخاب ویژگی، ویژگی‌هایی که با یکدیگر ضریب همبستگی بالا (نزدیک به ۱ یا ۱-) داشته باشند ممکن است برای پیش‌بینی اضافی باشند. از بین این ویژگی‌ها تنها یک مورد انتخاب می‌شود چون اگر دو ویژگی با هم مرتبط باشند، مدل فقط به یکی نیاز دارد، زیرا دومی اطلاعات جدیدی برای بهبود عملکرد آموزش مدل اضافه نمی‌کند. از سوی دیگر، ویژگی‌هایی با ضرایب همبستگی پایین (نزدیک به ۰) ممکن است کمتر برای پیش‌بینی مرتبط باشند و همچنین می‌توانند برای حذف در نظر گرفته شوند. در شکل ۲ ارتباط بین ویژگی‌های مختلف با معیارهای اسپیرمن و پیرسون نشان داده‌شده است. با توجه به نتایج می‌توان نتیجه گرفت که فشار ورودی (P_{in})، فشار خروجی (P_{out})، بار موتور ($MotorLoad$)، جریان موتور (I_m) و ولتاژ موتور (V) به‌عنوان ویژگی‌های مناسب برای تخمین دبی در نظر گرفته می‌شوند. کامیلری و همکاران در سال ۲۰۱۰ از معادله بقای انرژی در پمپ برای تخمین دبی استفاده کردند [۱۴]. طبق این معادله توان جذب‌شده توسط پمپ برابر با توان تولیدشده توسط موتور است. توان پمپ تابعی از اختلاف فشار دبی و بازدهی بوده و از سوی دیگر توان موتور تابعی از ولتاژ موتور، جریان موتور، ضریب توان و بازدهی موتور است؛ بنابراین در یک پمپ ESP دبی تابعی از اختلاف فشار قبل و بعد از پمپ، جریان و ولتاژ موتور است. نتایج این مطالعه نشان می‌دهد که ویژگی‌های انتخاب‌شده با نرخ دبی ارتباط دارند.



شکل ۲. ارتباط بین نرخ جریان نفت و ویژگی‌ها با استفاده از معیارهای اسپیرمن (سمت راست) و پیرسون (سمت چپ)

۳-۲- روش‌های یادگیری ماشین

۳-۲-۱- روش GBR

روش تقویت‌کننده گرادیان(GBR) یک شیوه یادگیری ماشین قدرتمند است که هم برای مسائل رگرسیون و هم برای طبقه‌بندی استفاده می‌شود. این روش به روش‌های یادگیری گروهی تعلق دارد که چندین مدل را برای بهبود عملکرد پیش‌بینی ترکیب می‌کنند. در مورد GBR، مدل‌ها معمولاً درخت تصمیم هستند.

ایده اصلی رگرسیون GBR، آموزش مکرر مدل‌های جدید است که باقی‌مانده یا خطاهای مدل‌های قبلی را پیش‌بینی می‌کنند. سپس این مدل‌های جدید به مجموعه اضافه می‌شوند و هرکدام بر روی خطاهایی تمرکز می‌کنند که توسط مدل‌های قبلی به‌خوبی پیش‌بینی‌نشده بودند. با انجام این کار به‌صورت مکرر، گروه به‌تدریج عملکرد پیش‌بینی خود را بهبود می‌بخشد [۱۵]. یکی از مزایای کلیدی GBR، توانایی آن در مدیریت روابط پیچیده و غیرخطی در داده‌ها است. علاوه بر قدرت پیش‌بینی آن، GBR معیاری از اهمیت ویژگی را نیز ارائه می‌دهد که می‌تواند برای درک اینکه کدام ویژگی‌ها در پیش‌بینی‌ها بیشترین تأثیر را دارند مفید باشد. این می‌تواند برای به دست آوردن بینش در مورد فرآیندهای اساسی که داده‌ها را هدایت می‌کند و برای تصمیم‌گیری آگاهانه بر اساس خروجی مدل ارزشمند باشد. یک مدل GBR با تعداد M درخت را می‌توان به‌صورت زیر تعریف کرد.

$$f_M(x_j) = \sum_{m}^{M} \gamma_m h_m(x_j) \qquad (۱)$$

در رابطه فوق h_m نشان‌دهنده یک یادگیرنده ضعیف است که به‌تنهایی عملکرد ضعیفی دارد و به‌عنوان یک عامل مقیاس برای اضافه کردن سهم یک درخت به مدل عمل می‌کند. روش GBR از تابع تلفات نزولی گرادیان برای به حداقل رساندن خطاها با به‌روزرسانی تخمین اولیه با تخمین جدید استفاده می‌کند؛ که نتیجه آن ایجاد یک مدل نهایی با ادغام تمام تخمین‌های اولیه با وزن‌های مناسب است.

۳-۲-۲- روش k-NN

روش k همسایه نزدیک‌تر (k-NN) اولین بار در سال ۱۹۷۶ توسط دودانی [۱۶] توسعه یافت در سال ۱۹۸۳ جوزویک [۱۷] نسخه پیشرفته‌تر این روش را معرفی کرد. در زمینه یادگیری ماشین، k-NN یک روش ناپارامتریک[۱] پرکاربرد برای طبقه‌بندی و رگرسیون است. این روش بر این اصل استوار است که نقاط داده مشابه در یک فضای چندبعدی به یکدیگر نزدیک هستند. روش k-NN با یافتن k همسایه نزدیک یک نقطه داده بر اساس فاصله (مانند فاصله اقلیدسی یا فاصله منهتن) پیش‌بینی یا طبقه‌بندی برای نقطه داده جدید را انجام می‌دهد. مقدار k تعیین می‌کند که چه تعداد همسایه در فرآیند پیش‌بینی در نظر گرفته شود. یکی از شاخص‌ترین مزایای این روش، عملکرد مناسب در مدل‌هایی است که حجم داده آن‌ها کم است. البته باید توجه داشت که این روش قادر نیست مقادیر بزرگ‌تر از بیشترین مقدار مشاهده‌شده و کوچک‌تر از کمترین مقدار مشاهده‌شده را تولید کند. به‌عبارت‌دیگر این روش تنها توانایی درون‌یابی بین داده‌ها را دارد و قادر به انجام برونیابی نیست [۱۸].

k-NN با مقایسه یک نمونه آزمون داده‌شده (x, y) با یک مجموعه آموزشی $D = \left[(x_i, y_i)\right]$ یاد می‌گیرد. روش k-NN ابتدا فاصله بین x و هر نمونه x_i در D را محاسبه می‌کند. فاصله در اینجا با نماد d بیان‌شده است. مقدار d_i از رابطه‌های زیر محاسبه می‌شود.

$$d_{Ecu} = \sqrt{(x - x_i)^2} \qquad (۲)$$

[۱]. منظور این است که در این روش رابطه پارامتری از پیش تعیین شده ای میان متغیر های ورودی و خروجی برقرار نمی شود.

$$d_{Man} = |x - x_i| \quad (۳)$$

که در آن d_{Ecu} فاصله اقلیدسی و d_{Man} فاصله منهتن است.

در مرحله بعد k-NN داده‌ها را بر اساس مقدار d از کوچک به بزرگ مرتب می‌کند؛ پس از رتبه‌بندی، نمونه مرتبط با d_i به‌عنوان نزدیک‌ترین همسایه i نامیده می‌شود و خروجی با $y_i(x)$ نشان داده می‌شود. خروجی نهایی y_{pred} میانگین وزن‌دار خروجی k همسایه نزدیک آن است، همان‌طور که در معادله (۴) نشان داده‌شده است [۱۹].

$$y_{pred} = \frac{\sum_{i=1}^{k} w_i y_i(x)}{\sum_{i=1}^{k} w_i} \quad (۴)$$

در رابطه بالا y_{pred} مقدار تخمین زده‌شده و y_{pred} وزن هر همسایه است که برابر با معکوس فاصله آن است.

۳-۲-۳- روش DT

درخت تصمیم (Decision Tree) یک روش یادگیری ماشین ساده و محبوب است که برای طبقه‌بندی داده‌ها و رگرسیون (پیش‌بینی مقادیر عددی) استفاده می‌شود. این روش از یک نمودار درخت مانند برای مدل‌سازی تصمیمات و پیامدهای مختلف استفاده می‌کند [۲۰].

در این روش، فضای ورودی به مناطق کوچک‌تر تقسیم می‌شود و در هر منطقه، یک مدل ساده برای پیش‌بینی ایجاد می‌شود. در رگرسیون با درخت تصمیم، هدف پیش‌بینی مقدار یک متغیر هدف با استفاده از قوانین تصمیم‌گیری ساده است. درواقع، درخت تصمیم یک ساختار درخت مانند دارد که هر گره داخلی آن یک تصمیم بر اساس یک ویژگی از داده‌ها را نشان می‌دهد و هر گره برگ نشان‌دهنده مقدار پیش‌بینی‌شده متغیر هدف است. قوانین تصمیم‌گیری از داده‌های آموزشی یاد گرفته می‌شوند و ساختار درخت بر اساس ویژگی‌هایی تعیین می‌شود که به بهترین شکل داده‌ها را در هر گره تقسیم می‌کند. [۲۱]. یکی از مزایای اصلی این روش، سادگی و تفسیرپذیری آن است. درخت‌های تصمیم را می‌توان به‌راحتی تجسم کرد و حتی توسط افراد غیرمتخصص قابل‌درک است. این درخت‌ها به‌عنوان ابزاری مفید برای تحلیل داده‌ها و فرآیندهای تصمیم‌گیری پیچیده مورداستفاده قرار می‌گیرند. همچنین، درخت‌های تصمیم قادر به مدیریت داده‌های عددی و مقوله‌ای هستند و نسبت به مقادیر پرت و گمشده نیز مقاومت دارند.

۳-۳- تنظیم پارامترها

تنظیم پارامترها در الگوریتم‌های یادگیری ماشین اهمیت بالایی دارد، زیرا بر کیفیت مدل‌سازی و دقت پیش‌بینی‌ها تأثیر می‌گذارد. درواقع تنظیم پارامتر مناسب برای اطمینان از اینکه مدل به‌طور مؤثر از داده‌های آموزشی یاد می‌گیرد و به‌خوبی به داده‌های دیده نشده تعمیم داده می‌شود؛ ضروری است. یکی از رویکردهای رایج برای تنظیم پارامتر، جستجوی شبکه‌ای است که در آن مجموعه‌ای از مقادیر از پیش تعریف‌شده برای هر پارامتر مشخص می‌شود و مدل برای هر ترکیبی از مقادیر، آموزش و ارزیابی می‌گردد. روش دیگر جستجوی تصادفی است که از فضای پارامتر به‌صورت تصادفی نمونه‌برداری می‌کند. هر دو روش مزایا و معایب خود را دارند. در این پژوهش، روش جستجوی شبکه‌ای برای تنظیم پارامترها انتخاب‌شده

است. برای الگوریتم k-NN، پارامترهای کلیدی شامل تعداد همسایه‌ها (k)، نحوه وزن دهی و معیار فاصله (مانند فاصله اقلیدسی) هستند. در درخت تصمیم، عمق درخت و تعداد نمونه‌ها در هر برگ ازجمله پارامترهای مهم به شمار می‌روند. تقویت گرادیان نیز از پارامترهایی مانند نرخ یادگیری، تعداد درختان و عمق درختان استفاده می‌کند. این پارامترها هرکدام جنبه‌های مختلفی از الگوریتم را کنترل می‌کنند و برای دستیابی به بهترین عملکرد، باید با دقت تنظیم شوند. در جدول ۱، مقادیر پیشنهادی برای پارامترهای مذکور آورده شده است.

جدول ۱. مقدار پارامترهای مهم الگوریتم‌های یادگیری ماشین استفاده‌شده

الگوریتم	پارامتر	مقدار پیشنهادی
k-NN	تعداد همسایه‌ها (k)	۵
	نوع وزن دهی	فاصله
	نوع فاصله	اقلیدسی
DT	عمق درخت	۷
	نمونه‌ها	۵
GBR	نرخ یادگیری	۰.۱
	تعداد درختان	۵۰۰
	عمق درختان	۲

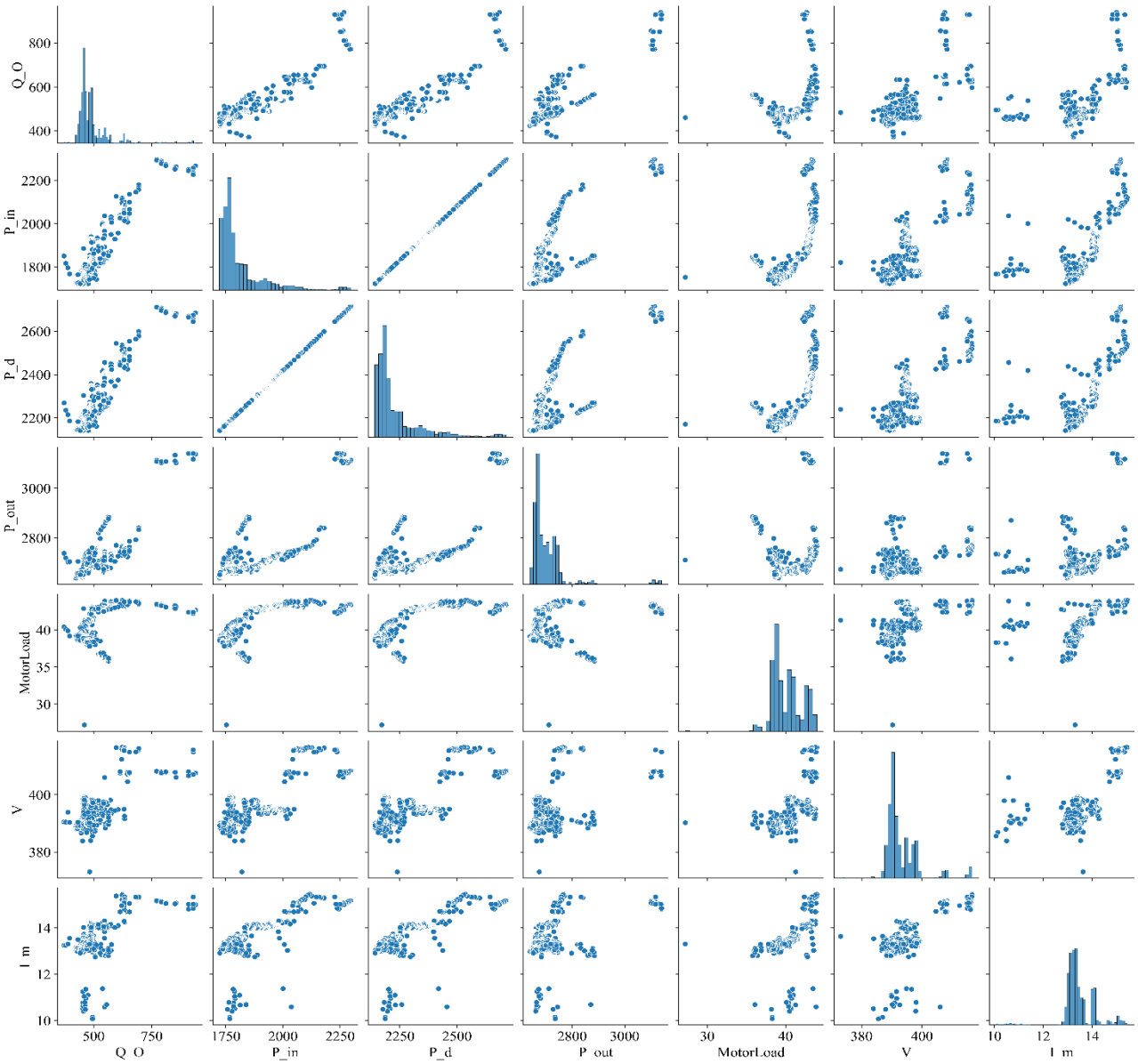
۴- توضیحات داده‌ها

اطلاعات مورداستفاده در این مطالعه به‌صورت یک روز در میان یا دو روز در میان و از تاریخ ۱۰ اکتبر ۲۰۲۰ تا ۸ جولای ۲۰۲۳ جمع‌آوری‌شده است و تعداد کل داده¬ها ۹۸۰ عدد است. در این بازه زمانی تغییرات زیادی در چاه‌های میدان صورت نگرفته و همچنین عملیات تزریق آب و یا گاز و سایر فرآیندهایی که می‌تواند منجر به تغییر جدی در مکانیزم های تولیدی گردد، گزارش نشده است. در بخش پیش‌پردازش داده‌ها ابتدا داده‌های گمشده از مجموعه داده حذف گردید. سپس با بررسی بیشتر داده‌ها مشخص شد که برخی از داده‌ها در فاصله بسیار زیاد از میانگین قرار دارند. به‌عنوان‌مثال دبی نفت در بعضی از روزها صفر گزارش‌شده بود، ازآنجاکه این بخش از داده‌ها فاصله قابل‌توجهی با سایر داده‌ها و میانگین داشتند از مجموعه داده حذف شدند. پس از حذف داده‌های پرت و گمشده حجم داده‌ها به ۸۵۰ مورد کاهش یافت. جدول ۲ اطلاعات آماری جامعی از مجموعه داده را نشان می‌دهد.

جدول ۲. اطلاعات آماری مجموعه داده پس از انجام پیش‌پردازش

پارامترها	فشار ورودی	فشار خروجی	بار موتور	جریان موتور	ولتاژ موتور	دبی نفت
نماد	P_{in}	P_{out}	$MotorLoad$	I_m	V	Q_o
کمترین مقدار	۱۷۲۲	۲۶۴۰	۲۷.۲	۱۰.۰۷	۳۷۳.۰۶	۳۷۲
بیشترین مقدار	۲۲۹۷	۳۱۴۰	۴۴	۱۵.۴۶	۴۱۶.۵۶	۹۳۹
میانگین	۱۸۱۳.۵۶	۲۷۰۴.۶۳	۴۰.۱۹	۱۳.۴۶	۳۹۳	۴۹۱.۶۷
انحراف معیار	۱۰۶.۸۳	۷۱.۶۷	۱.۸۷	۰.۶۵	۵.۷۷	۷۰.۹۹
واریانس	۱۱۳۹۹.۳۸	۵۱۳۶.۴۵	۳.۵۲	۰.۴۲	۳۳.۳۳	۵۰۴۰.۳۵

شکل ۳ رابطه بین ویژگی‌های انتخاب‌شده و تابع هدف و همچنین نمودار توزیع آن‌ها را نشان می‌دهد. در این شکل نمودارهای قطر اصلی توزیع فراوانی هر پارامتر را نشان می‌دهند. در دیگر ستون‌ها نیز پارامترها برحسب یکدیگر رسم شده‌اند. سطر اول مربوط به دبی نفت (Q_o) اندازه‌گیری شده است. اولین نمودار در این سطر توزیع فراوانی دبی است همان‌طور که مشاهده می‌شود دبی نفت از توزیع نرمال برخوردار نیست و دبی در ناحیه ۴۰۰ تا ۶۰۰ بشکه در روز بیشترین فراوانی را دارند و حجم داده‌ها در ناحیه بالاتر از این نرخ کم است. همان‌طور که در شکل ۳ مشخص است دبی نفت با فشار تخلیه نیز ارتباط قوی دارد بااین‌حال ازآنجاکه بین فشار ورودی پمپ (P_{in}) و فشار تخلیه (P_d) خطی است بنابراین تنها یکی از این دو مورد به‌عنوان ورودی مدل داده محور انتخاب‌شده است. توزیع فراوانی دیگر ورودی‌ها نیز غیر نرمال هستند.



شکل ۳. رسم نمودارهای دوتایی بین ویژگی‌های انتخاب‌شده و نرخ جریان نفت

۵- ارائه نتایج و بحث و بررسی

در این بخش عملکرد هرکدام از روش‌های DT، GBR و k-NN در دو حالت داده‌های با نویز کم و داده‌های با نویز شدید موردبررسی قرار می‌گیرد و با یکدیگر مقایسه می‌گردد. معیارهای آماری مختلفی برای بررسی عملکرد روش‌های یادگیری ماشین وجود دارد. این معیارها می‌توانند دقت مدل‌های ارائه‌شده را به‌خوبی نشان دهند. در اینجا از معیارهای جذر میانگین مربعات خطا، میانگین قدر مطلق خطاها، میانگین درصد قدر مطلق خطا و ضریب تعیین استفاده‌شده و روابط مربوط به آن‌ها، به ترتیب در زیر آمده است.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{test,i} - y_{pred,i}\right)^2} \quad (۵)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{test,i} - y_{pred,i}\right| \quad (۶)$$

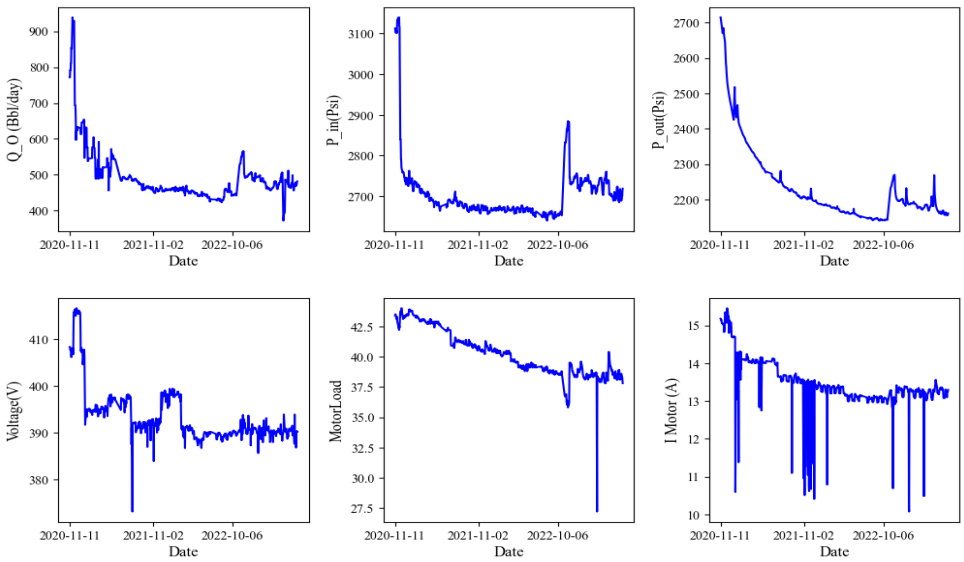
$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_{test,i} - y_{pred,i}}{y_{test,i}}\right| \quad (۷)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{test,i} - y_{pred,i}\right)^2}{\sum_{i=1}^{n}\left(y_{test,i} - \bar{y}\right)^2} \quad (۸)$$

در این روابط y_{test} مقادیر داده‌های آزمون، y_{pred} مقادیر پیش‌بینی‌شده، $\bar{y}$ مقدار متوسط نمونه و n تعداد داده‌های آزمون است. در یادگیری ماشین معمولاً مجموعه داده به دو بخش داده‌های آموزش و آزمون تقسیم می‌شوند. مدل‌های ارائه‌شده روابط موجود بین ویژگی‌ها و تابع هدف را با استفاده از داده‌های آموزش یاد گرفته و سپس دقت آن‌ها با استفاده از داده‌های آزمون بررسی می‌گردد. در پژوهش حاضر نیز از ۷۵ درصد داده‌ها برای آموزش و ۲۵ درصد برای آزمون استفاده‌شده است.

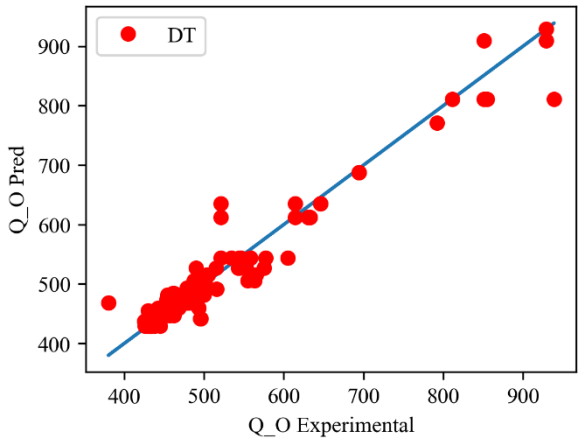
۵-۱- داده‌های با نویز کم

داده‌های واقعی مورداستفاده در مدل داده محور در شکل ۴ نشان داده‌شده است. شکل ۴ نشان می‌دهد ویژگی‌های ورودی دارای نویز شدیدی نیستند که این مورد می‌تواند به مدل‌های ارائه‌شده برای پیش‌بینی دقیق‌تر کمک کند. البته برخی از ویژگی‌ها مانند جریان و ولتاژ موتور دارای نویزند.

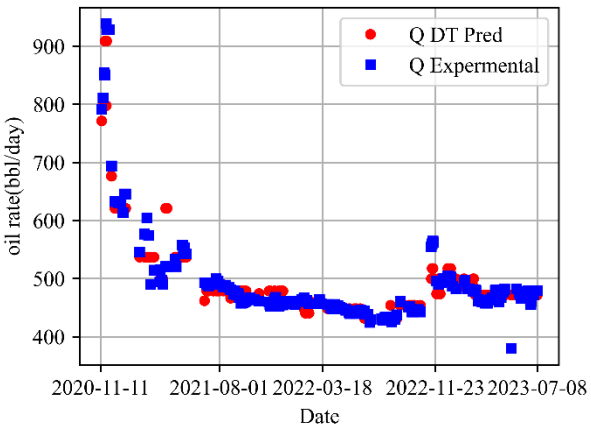


شکل ۴. رسم ورودی‌های مدل داده محور مجموعه داده کم نویز

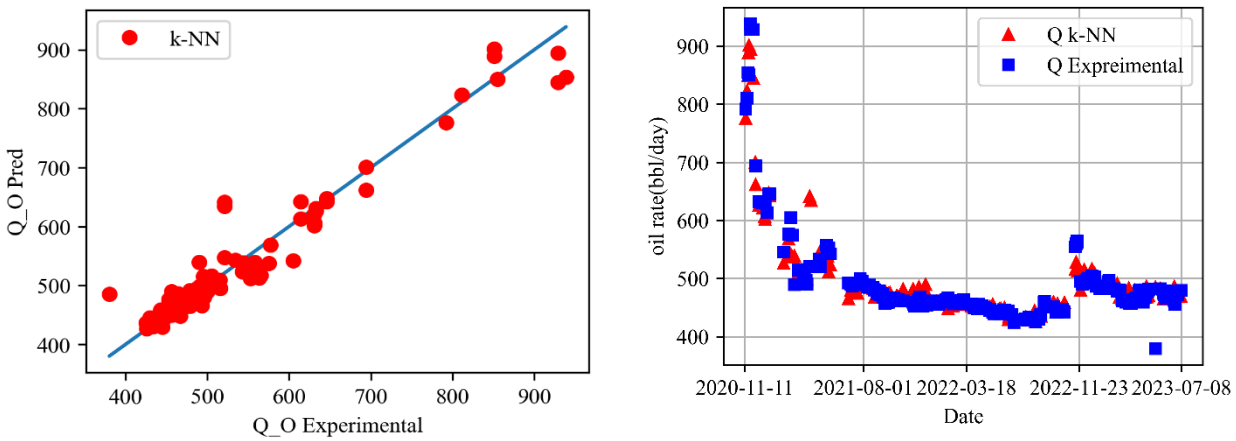
شکل ۵ (الف)، (ب) و (ج) سمت راست دبی تخمینی نفت توسط مدل‌های داده محور و دبی تجربی نفت را برحسب زمان نمایش می¬دهد. همان‌طور که مشاهده می‌شود مدل‌های دبی سنج داده محور در دبی‌های بیشتر از ۵۰۰ بشکه در روز نسبت به بازه کمتر از ۵۰۰ بشکه در روز خطای بالاتری دارند که دلیل آن حجم کمتر داده‌ها در این بازه نسبت به بازه کمتر از ۵۰۰ بشکه در روز است. از مقایسه شکل‌های ۵ (الف)، (ب) و (ج) سمت راست استنباط می‌شود که بین روش‌های ارائه‌شده با قاطعیت نمی‌توان درباره برتری یک روش نسبت به روش دیگر صحبت کرد. شکل‌های ۵ (الف)، (ب) و (ج) سمت چپ نشان می‌دهند که مدل‌های داده محور ارائه‌شده در محدوده کمتر از ۵۰۰ بشکه در روز عملکرد خوبی دارند و نمی‌توان گفت که کدام مدل عملکرد بهتری داشته است. بااین‌حال در محدوده بالاتر از ۵۰۰ بشکه در روز به ترتیب مدل k-NN، GBR و DT بهترین عملکرد را دارند. برای بررسی دقیق‌تر عملکرد هر روش، دقت آن‌ها بر اساس معیارهای ارزیابی در جدول ۳ آمده است.



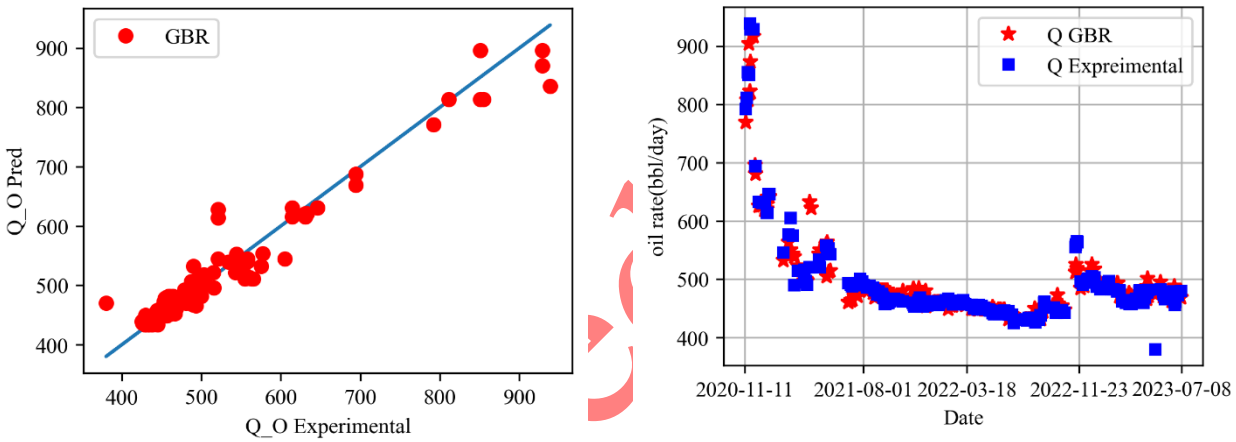
(الف) چپ



(الف) راست



(ب) چپ (ب) راست



(ج) چپ (ج) راست

شکل ۵. دبی تخمینی نفت و دبی تجربی نفت نسبت به زمان (راست) و دبی تخمینی نفت در برابر دبی تجربی نفت (چپ) در حالت ورودی با نویز کم توسط مدل‌های داده محور (الف) DT (ب) k-NN (ج) GBR

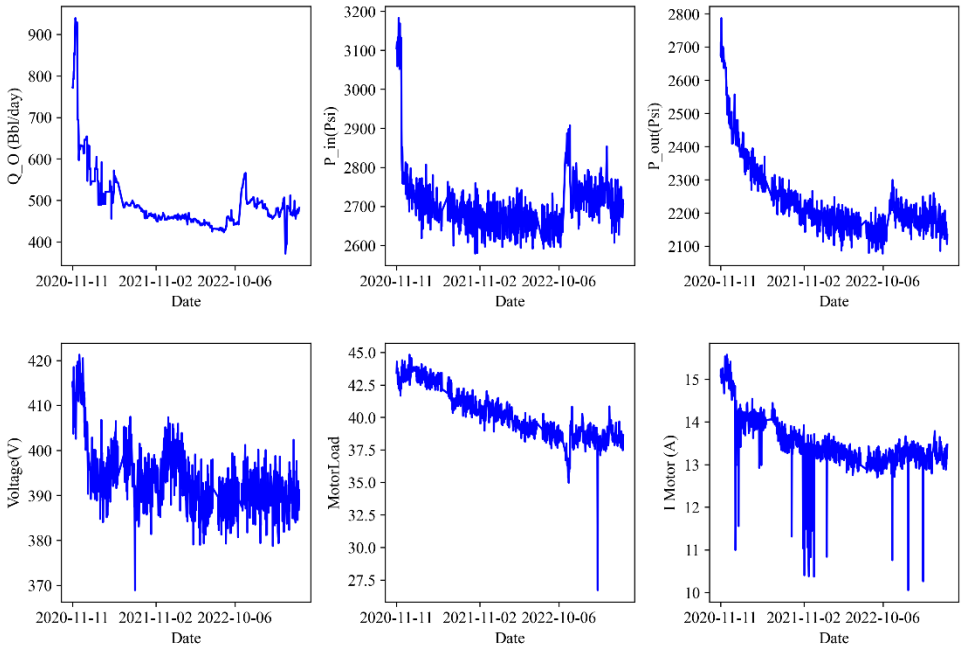
جدول ۳ نشان می‌دهد که مدل‌های داده محور ارائه‌شده عملکرد بسیار نزدیکی دارند بااین‌حال مدل k-NN با اختلاف کمی نسبت به دو مدل دیگر عملکرد بهتری دارد. مدل DT نیز دقت پایین‌تری نسبت به دو مدل دیگر دارد.

جدول ۳. ارزیابی عملکرد مدل‌های ارائه‌شده در مجموعه داده با نویز گم برای داده‌ها آزمون

روش یادگیری ماشین	MAE	MAPE	RMSE	R^2
DT	۱۲.۹۵۶۳	۲.۴۷۷۳	۲۱.۵۶۶۹	۰.۹۴۱۳
k-NN	۱۰.۹۵۶۵	۲.۰۵۹۲	۲۰.۳۱۴۰	۰.۹۴۷۹
GBR	۱۲.۱۸۳۴	۲.۳۱۹۱	۲۰.۲۸۱۵	۰.۹۴۷۷

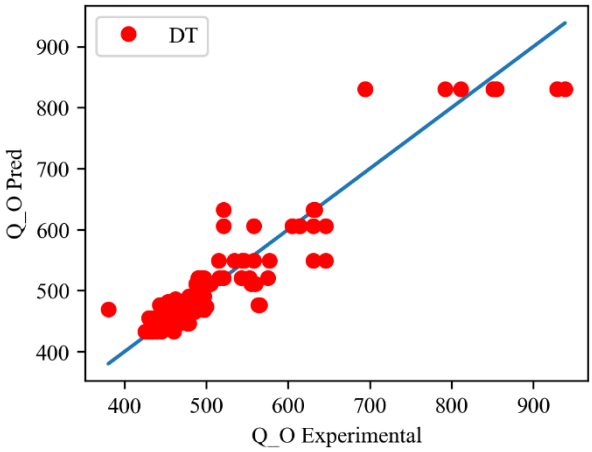
۵-۲- داده‌های نویز دار

نویز موجود در مجموعه داده می‌تواند تأثیر مخربی بر عملکرد روش‌های یادگیری ماشین داشته باشد و منجر به سوگیری شده و دقت پیش‌بینی‌ها را کاهش دهد. همچنین می‌تواند منجر به برازش بیش‌ازحد شود. برازش بیش‌ازحد زمانی اتفاق می‌افتد که یک مدل از نویز موجود در داده‌ها به‌جای الگوهای زیربنایی یاد می‌گیرد و درنتیجه تعمیم ضعیفی به داده‌های جدید و نامرئی ایجاد می‌کند. این می‌تواند قابلیت اطمینان و اثربخشی مدل‌های یادگیری ماشین را به خطر بیندازد و منجر به نتیجه‌گیری و تصمیم‌گیری اشتباه شود. چندین منبع نویز در داده‌ها وجود دارند که ازجمله آن‌ها می‌توان به خطاهای اندازه‌گیری، مقادیر ازدست‌رفته، نقاط پرت و ویژگی‌های نامربوط اشاره کرد. این‌ها می‌توانند از عوامل مختلفی مانند نقص حسگرها، خطای انسانی یا تنوع ذاتی در فرآیند جمع‌آوری داده‌ها ناشی شوند. در این قسمت با افزودن مقدار یک درصد انحراف معیار نویز با توزیع گاوسی به پارامترهای اندازه‌گیری شده ورودی، عملکرد مدل‌ها در شرایط نویزی بررسی‌شده است. شکل ۶ داده‌های ورودی موردمطالعه در این بخش از مقاله را نشان می‌دهد.

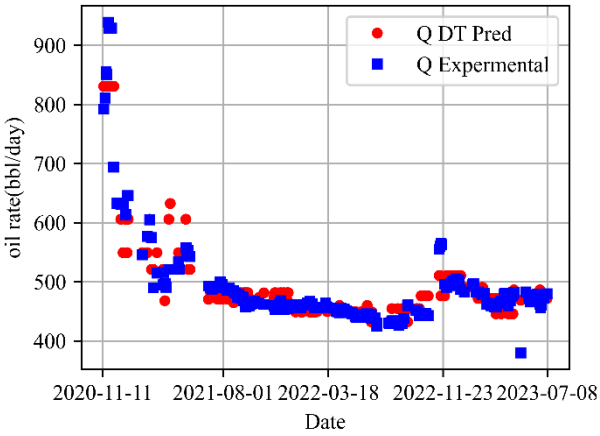


شکل ۶. داده‌های ورودی با یک درصد انحراف معیار نویز

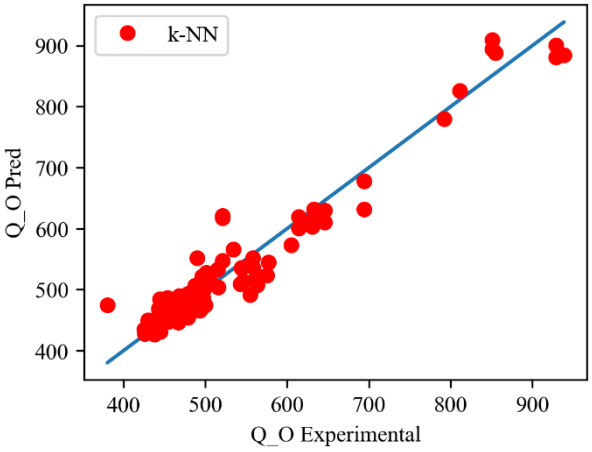
در این حالت نیز مجموعه داده به دو قسمت بخش آموزش و بخش آزمون تقسیم‌شده است. شکل ۷ (الف)، (ب) و (ج) سمت راست دبی تخمینی نفت توسط مدل‌های داده محور و دبی تجربی نفت را برحسب زمان نمایش می‌دهد. همان‌طور که مشاهده می‌شود نویز موجود در داده‌ها تأثیر منفی بر عملکرد دبی سنج‌های مجازی داده محور داشته است. تأثیر نویزها در مناطقی که حجم داده کمتری دارند بیشتر است. شکل‌های ۷ (الف)، (ب) و (ج) سمت چپ نشان می‌دهند که مدل داده محور k-NN بهترین عملکرد را در برابر ورودی‌های دارای نویز دارد. دلیل این امر این است که این مدل از همسایه‌ها برای تخمین مقدار استفاده می‌کند. مدل DT نیز بیشتر از مدل‌های k-NN و GBR تحت تأثیر نویز موجود در ورودی‌ها بوده است. مدل GBR توانسته با ترکیب چندین مدل DT عملکرد این مدل را بهبود دهد.



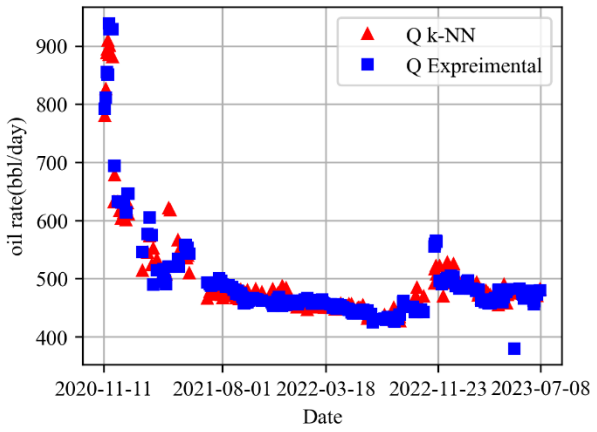
(الف) چپ



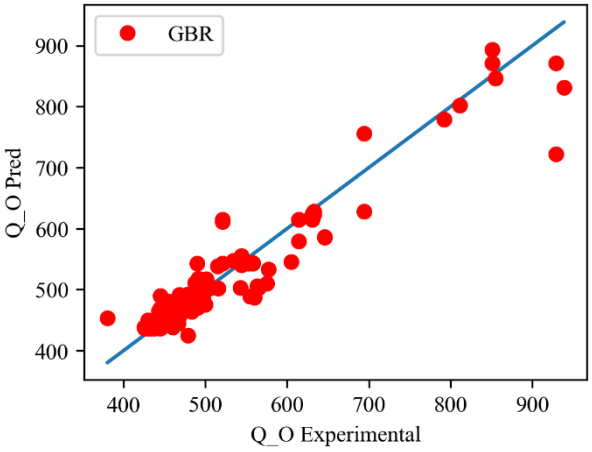
(الف) راست



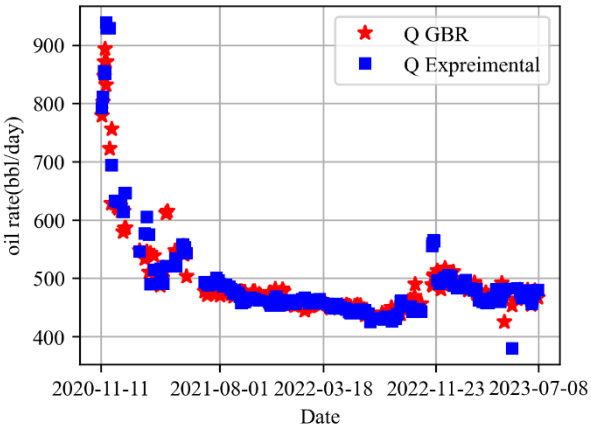
(ب) چپ



(ب) راست



(ج) چپ



(ج) راست

شکل ۷. دبی تخمینی نفت و دبی تجربی نفت نسبت به زمان (راست) و دبی تخمینی نفت در برابر دبی تجربی نفت (چپ) در حالت ورودی با نویز کم توسط مدل‌های داده محور (الف) DT (ب) k-NN (ج) GBR

جدول ۴ عملکرد مدل‌های مختلف را در حالت با نویز نشان می‌دهد. جدول ۴ میزان دقت هرکدام از مدل‌ها را در برابر داده‌های نویز دار نشان می‌دهد. با توجه به نتایج ارائه‌شده در این جدول مشاهده می‌شود که مدل k-NN بهترین عملکرد را در برابر داده‌های نویز دار داشته است و نویز تأثیر کمی بر عملکرد این روش داشته است. مدل DT نیز بیشترین تأثیر را از نویز گرفته است.

جدول ۴. عملکرد روش‌های ارائه‌شده در مجموعه داده با یک درصد انحراف معیار نویز برای داده‌های آزمون

روش موردبررسی	MAE	MAPE	RMSE	R^2
DT	۱۶.۹۲۶۸	۳.۰۹۶۲	۳۲.۵۳۲۴	۰.۸۶۶۵
k-NN	۱۴.۹۱۹۷	۲.۸۳۰۵	۲۴.۲۵۷۹	۰.۹۲۵۷
GBR	۱۴.۶۶۱۱	۳.۴۲۵۳	۲۶.۸۰۸۸	۰.۸۹۶۵

بررسی نتایج به‌دست‌آمده برای ورودی‌های نویز دار و بدون نویز نشان داد که مدل DT یک مدل ساده است که در یک مجموعه داده با نویز کم، تنوع ناکافی و حجم محدود می‌تواند تخمین قابل قبولی داشته باشد بااین‌حال این مدل در برابر نویز بسیار حساس است. الگوریتم GBR یک مدل پیشرفته‌تر از مدل DT است. این مدل با ایجاد چند درخت تصمیم مختلف سعی می‌کند که عملکرد مدل DT را بهبود دهد. نتایج به‌دست‌آمده در این مطالعه به‌خوبی نشان می‌دهد که مدل GBR در شرایط مجموعه داده با حجم محدود، تنوع ناکافی و کم نویز عملکرد نزدیک به عملکرد DT دارد. بااین‌حال در مجموعه داده نویز دار عملکرد مدل GBR بهتر از مدل DT است. الگوریتم k-NN عملکرد بسیار خوبی در هر دو مجموعه داده کم نویز و نویز دار داشته است. این الگوریتم می‌تواند در شرایط مجموعه داده با حجم کم و تنوع ناکافی عملکرد خوبی داشته است.

در این مطالعه سعی شده است که قابلیت‌های الگوریتم‌های k-NN، DT و GBR نشان داده شود. باوجود مزایایی که برای مدل دبی سنج مجازی داده محور ارائه‌شده یادشده است، این مدل‌ها دارای محدودیت‌هایی نیز هستند. در این مدل تنها از اطلاعات پمپ استفاده‌شده است و تأثیرات دیگر پارامترهای درون‌چاهی در نظر گرفته نشده است. به همین دلیل نیاز است که دقت مدل در طول زمان به‌صورت مکرر امتحان شود. در این پژوهش دبی سنج مجازی داده محور برای تخمین دبی نیازمند اطلاع از شرایط درون چاه و مخزن نیست. بااین‌حال درصورتی‌که شرایط چاه و مخزن تغییر کنند مدل داده محور نیازمند آموزش مجدد است.

۶- جمع‌بندی

در پژوهش حاضر از روش‌های یادگیری ماشین شامل روش تقویت گرادیان(GBR)، درخت تصمیم (DT) و k همسایه نزدیک‌تر (k-NN) برای تخمین نرخ دبی نفت خروجی از یک چاه نفتی واقع در جنوب ایران استفاده شد. انتخاب ویژگی‌های مناسب برای آموزش مدل‌ها با استفاده از معیارهای آماری اسپیرمن و پیرسون انجام شد. در این مطالعه از داده‌های ورودی و خروجی پمپ‌های شناور الکتریکی (ESP) شامل فشار ورودی (P_{in})، فشار خروجی (P_{out})، بار موتور ($MotorLoad$)، جریان موتور (I_m) و ولتاژ موتور (V) استفاده شد و از نتایج به‌دست‌آمده می‌توان به این نتیجه رسید که اطلاعات ورودی و خروجی پمپ‌ها می‌توانند به‌عنوان ورودی‌های دبی سنج‌های داده محور استفاده شوند. تعداد داده‌های موجود در مجموعه

داده ۹۸۰ عدد بود که پس از حذف داده‌های پرت و مقادیر گمشده تعداد داده‌ها به ۸۵۰ عدد کاهش یافت. توزیع فراوانی داده‌ها نشان می‌دهد که تعداد داده‌هایی که دبی نفت آن در محدوده بیش‌تر از ۵۰۰ بشکه در روز بوده کم است. همین امر چالشی در برابر استفاده از دبی سنج‌های مجازی داده محور ایجاد کرده است، بااین‌حال نتایج نشان می‌دهد که مدل‌های ارائه‌شده عملکرد قابل قبولی داشتند. مدل k-NN بهترین عملکرد را نسبت به دو مدل دیگر نشان داد. در ادامه عملکرد مدل‌های ارائه‌شده در برابر داده‌های نویزی با اضافه کردن یک درصد انحراف معیار نویز با استفاده از تابع توزیع گاوسی به داده‌های ورودی مورد ارزیابی قرار گرفت. مشاهده شد که روش‌های یادگیری ماشین استفاده‌شده در اثر وجود نویز در اطلاعات ورودی، خصوصاً در محدوده‌های با داده ناکافی دچار خطا می‌شوند. روش k-NN بهترین عملکرد را در برابر داده‌های نویزی در مقایسه با دو روش دیگر به نمایش گذاشت که نشان از مدیریت مناسب این مدل در حضور داده‌های دارای نویز است. روش DT نیز در مقایسه با دو روش دیگر، بیشترین خطا را در حضور داده‌های دارای نویز نشان داد. نتایج این مطالعه نشان داد که انتخاب داده ورودی و مدل یادگیری ماشین مناسب می‌تواند عملکرد بسیار مناسبی را برای دبی سنج مجازی به ارمغان بیاورد. در ادامه پیشنهاد می‌شود که از داده‌های ته چاهی و سر چاهی در کنار داده‌های پمپ به‌عنوان دبی سنج مجازی داده محور استفاده شود.

تقدیر و تشکر و پیوست‌ها

داده‌های این مطالعه توسط شرکت توسعه نفت و گاز پرشیا جمع‌آوری و ارائه‌شده است. لذا نویسندگان از این شرکت بابت همکاری ایشان، صمیمانه تشکر می‌نمایند.

References

[1] Dayev, Z. A. (2020). Application of artificial neural networks instead of the orifice plate discharge coefficient. Flow Measurement and Instrumentation, 71, 101674. .

[2] Bikmukhametov, T. & Jäschke, J. (2020). First principles and machine learning virtual flow metering: a literature review. Journal of Petroleum Science and Engineering, 184, 106487.

[3] Mercante, R., & Netto, T. A. (2022). Virtual flow predictor using deep neural networks. Journal of Petroleum Science and Engineering, 213, 110338..

[4] T. A. AL-Qutami, R. Ibrahim, I. Ismail, and M. A. Ishak, “Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing,” Expert Syst. Appl., vol. 93, pp. 72–85, Mar. 2018, doi: 10.1016/j.eswa.2017.10.014..

[5] Al-Qutami, T. A., Ibrahim, R., & Ismail, I. (2017, September). Hybrid neural network and regression tree ensemble pruned by simulated annealing for virtual flow metering application. In 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA) (pp. 304-309). IEEE.

[6] Ahmadi, M. A., Ebadi, M., Shokrollahi, A., & Majidi, S. M. J. (2013). Evolving artificial neural network and imperialist competitive algorithm for prediction oil flow rate of the reservoir. Applied Soft Computing, 13(2), 1085-1098.

[7] Bikmukhametov, T., & Jäschke, J. (2019). Oil production monitoring using gradient boosting machine learning algorithm. Ifac-Papersonline, 52(1), 514-519.

[8] Góes, M. R. R., Guedes, T. A., d’Avila, T. C., Vieira, B. F., Ribeiro, L. D., de Campos, M. C., & Secchi, A. R. (2021). Virtual flow metering of oil wells for a pre-salt field. Journal of Petroleum Science and Engineering, 203, 108586.

[9] Hotvedt, M., Grimstad, B., Ljungquist, D., & Imsland, L. (2022). On gray-box modeling for virtual flow metering. Control Engineering Practice, 118, 104974.

[10] AlAjmi, M. D., Alarifi, S. A., & Mahsoon, A. H. (2015, March). Improving multiphase choke performance prediction and well production test validation using artificial intelligence: a new milestone. In SPE digital energy conference and exhibition (p. D031S022R003). SPE.

[11] A. T. Sandnes, B. Grimstad, and O. Kolbjørnsen, “Multi-task learning for virtual flow metering,” Knowl.-Based Syst., vol. 232, p. 107458, Nov. 2021, doi: 10.1016/j.knosys.2021.107458.

[12] Al-Jasmi, A., Goel, H. K., Nasr, H., Querales, M., Rebeschini, J., Villamizar, M. A., ... & Saputelli, L. (2013, June). Short-term production prediction in real time using intelligent techniques. In SPE Europec featured at EAGE Conference and Exhibition? (pp. SPE-164813). SPE.

[13] Denney, T., Wolfe, B., & Zhu, D. (2013, March). Benefit evaluation of keeping an integrated model during real-time ESP operations. In SPE Digital Energy Conference and Exhibition (pp. SPE-163704). SPE..

[14] L. A. Camilleri, T. Banciu, G. Ditoiu, and S. A. Petrom, “First Installation of 5 ESPs Offshore Romania - A Case Study and Lessons Learned,” presented at the SPE Intelligent Energy Conference and Exhibition, Mar. 2010, p. SPE-127593-MS. doi: 10.2118/127593-MS."

[15] Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., ... & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 74, 634-642..

[16] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, (4), 325-327.

[17] Jóźwik, A. (1983). A learning scheme for a fuzzy k-NN rule. Pattern Recognition Letters, 1(5-6), 287-289.

[18] Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26-34. .

[19] Devroye, L., Gyorfi, L., Krzyzak, A., & Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. The annals of Statistics, 22(3), 1371-1385.

[20] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), 322-336.

[21] L. Breiman, Classification and regression trees. Routledge, 2017.