

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

محمدباقر صدیقی^۱، مجید سیاوشی^{۲*} و روح الدین میری^۲

۱- دانشکده مهندسی مکانیک، دانشگاه علم و صنعت ایران

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

تاریخ دریافت: ۱۴۰۳/۰۳/۱۳ تاریخ پذیرش: ۱۴۰۳/۰۴/۱۰

چکیده

تخمین دبی جریان در چاه‌های یک میدان نفتی، یک فرآیند حیاتی و کاربردی است. با این حال جریان‌های استخراج‌شده از چاه‌های نفتی، چند فازی بوده و تخمین دقیق دبی آن‌ها، بسیار چالش‌برانگیز و پرهزینه است. دبی‌سنج‌های مجازی در مقایسه با دبی‌سنج‌های چند فازی و روش‌های چاه آزمایشی، از نظر 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 دبی‌سنج‌های مجازی داده محور توجه زیادی شده است. در این مقاله تخمین دبی تولیدی یک چاه با استفاده از سه روش یادگیری ماشین k همسایه نزدیک‌تر (k -NN)، تقویت گرادیان (GBR) و درخت تصمیم (DT) با استفاده از داده‌های پمپ انجام شده است. به منظور انتخاب ویژگی‌های مناسب به‌عنوان ورودی روش‌ها، از تحلیل‌های آماری پیرسون و اسپیرمن استفاده شده است. مجموعه داده مورد بررسی مربوط به یکی از چاه‌های یک میدان نفتی در جنوب ایران است. مجموعه داده موجود دارای حجم کم و تنوع ناکافی است، اما با این وجود نتایج نشان می‌دهند که روش‌های پیشنهادی عملکرد مناسبی دارند. روش k -NN با دقت 0.9494 نسبت به دو روش دیگر عملکرد بهتری در تخمین دبی نفت داشته است. برای بررسی عملکرد روش‌ها در برابر داده‌های دارای نوفه، یک درصد انحراف معیار نوفه به داده‌های ورودی اضافه شد. بررسی‌ها نشان داد که مدل k همسایه نزدیک‌تر با دقت 0.9257 در مقایسه با دو روش دیگر عملکرد بهتری داشته و کمترین تأثیر را از نوفه‌ها گرفته است.

کلمات کلیدی: دبی‌سنج مجازی داده محور، یادگیری ماشین، k همسایه نزدیک‌تر، تقویت گرادیان، درخت تصمیم

مقدمه

تولید و مدیریت مخزن ضروری است [۱]. روش‌های چاه آزمایشی و استفاده از دبی‌سنج چند فازی (MPFM)^۱ دو روش متداول برای اندازه‌گیری دبی در میدان‌های نفتی به حساب می‌آیند.

آگاهی از نرخ تولید سیالات استخراج‌شده از یک چاه در صنعت نفت و گاز به منظور بهینه‌سازی

*مسئول مکاتبات
آدرس الکترونیکی

msiavashi@iust.ac.ir

شناسه دیجیتال: (DOI:10.22078/pr.2024.5417.3415)

معمولاً برای حل این روابط 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 دبی‌سنج‌ها وابستگی زیادی به مقدار فشار، حجم و دما دارد. تنظیم این نوع دبی‌سنج‌ها کار دشواری است زیرا در این روش کاربر باید دانش کافی در زمینه عملیاتی و نرم‌افزاری داشته باشد [۲].

دومین رویکرد در دبی‌سنج‌های مجازی استفاده از داده‌های میدانی برای تخمین دبی است. در این روش تلاش بر این است که با استفاده از داده‌های جمع‌آوری شده از میدان شامل داده‌های به‌دست‌آمده از حسگرهای درون‌چاهی و چاه آزمایشی، یک مدل ریاضی به کمک روش‌های پیشرفته یادگیری ماشین جهت تخمین دبی ایجاد می‌شود [۲]. اگر مدل مبتنی بر داده به‌خوبی آموزش داده شود می‌تواند دبی فازها را به‌صورت آنی و دقیق پیش‌بینی کند. مزیت اصلی این روش نسبت به دبی‌سنج‌های مجازی فیزیکی 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 تکیه دارند که داده‌های تجربی و عملکردی، سیستم را به‌خوبی نمایندگی می‌کنند و سعی می‌کنند روابط فیزیکی سیستم را مستقیماً از داده‌ها یاد بگیرند. القاطمی و همکاران [۴] یک سیستم دبی‌سنج 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 دبی‌سنج‌های مجازی مبتنی بر داده در میدان‌هایی با داده‌های تولیدشده بسیار محدود بهبود یابد. دبی‌سنج مجازی توسعه‌یافته دقت بسیار مطلوبی دارد.

با این حال این روش‌ها به دلیل نیاز به نصب دستگاه‌های فیزیکی از لحاظ عملیاتی پرهزینه هستند. پژوهشگران با هدف یافتن راهی برای کاهش هزینه‌های عملیاتی برداشت نفت و گاز، دبی‌سنج‌های مجازی (VFM)^۱ را توسعه دادند. دبی‌سنج مجازی در واقع 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 دبی‌سنج‌های مجازی در مقایسه با سایر روش‌ها بسیار کمتر است [۲]. استفاده از فن‌آوری VFM می‌تواند به عملکرد بهتر مجموعه‌های تولید نفت و گاز کمک کند زیرا این فن‌آوری 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 دبی‌سنج مجازی می‌تواند از طریق محاسبه‌ی برخط نرخ جریان سیالات به کاربران در تشخیص و پیش‌بینی مشکلات چاه کمک کرده، فرآیند نگهداری از چاه را ارتقا داده و به تبع آن، نیاز به تعمیرات و هزینه‌ها را کاهش دهد. دبی‌سنج‌های مجازی از نظر رویکرد محاسبه دبی به دو دسته دبی‌سنج مجازی فیزیکی محور و دبی‌سنج 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 چندفازی بسیار پیچیده هستند و حل آن‌ها بدون ساده‌سازی امکان‌پذیر نیست. از طرفی، حل این روابط نیازمند صرف هزینه محاسباتی و زمان اجرا بالا است [۲]. در روش فیزیکی محور، جریان سیال با استفاده از روابط بقا تحلیل شده و دبی تخمین زده می‌شود. با توجه به اینکه روابط حاکم بر جریان سیالات چندفازی پیچیده هستند و حل تحلیلی برای برخی از آن‌ها وجود ندارد، در نتیجه

1. Virtual Flow Meter

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٪ برای کل نرخ جریان نفت و گاز بود. هاتودت و همکاران یک دبی‌سنج مجازی جدید با ترکیب روش فیزیک محور و روش داده محور ایجاد کردند. در این پژوهش از رویکرد ترکیبی برای مدل‌سازی شیر فشارشکن استفاده شده است. شیر فشارشکن با مجموعه‌ای ساده از روابط پایه و یک شبکه عصبی برای تخمین ضریب جریان درجه نشان داده می‌شود. آن‌ها نشان دادند که رویکرد ترکیبی عملکرد مناسبی دارد و ممکن است مزایایی نسبت به دبی‌سنج‌های 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 (از جمله یک همبستگی تجربی برای جریان بحرانی شیر فشارشکن) استفاده کردند. در مقایسه با تخمین‌های دبی جریان به دست آمده با استفاده از همبستگی‌های تجربی شیر فشارشکن، شبکه عصبی (NN) عملکرد نسبتاً بهتری را نشان داد. لازم به ذکر است که مدل‌های شیر فشارشکن مورد استفاده در این مطالعه تجربی بودند. سندس و همکاران [۱۱] یک دبی‌سنج مجازی مبتنی بر داده 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٪ کاهش می‌دهد.

این سیستم می‌تواند دبی فاز مایع و گاز را به ترتیب با میانگین خطا ۴/۷٪ و ۲/۴٪ تخمین بزند. القطامی و همکاران [۵] یادگیری یک مجموعه ترکیبی را با ترکیب شبکه عصبی و رویکردهای درخت برآزش^۱ مورد بحث قرار دادند. ایشان رویکرد ترکیبی (NN-RTs) 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 فاز، روش جدیدی برای پیش‌بینی نرخ نفت تولیدی چاه‌ها با استفاده از منطق فازی، شبکه‌های عصبی مصنوعی و روش رقابت امپراتوری (ICA)^۲ 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 داده مربوط به ۵۰ چاه در یکی از میدان‌های نفتی واقع در شمال خلیج فارس برای ساخت پایگاه داده استفاده شده است. نتایج به دست آمده کارآیی، قدرت و سازگاری مدل ICA-ANN^۴ را نشان دادند. بیک مخامدوف و جاشکه [۷] عملکرد روش تقویت گرادیان را به‌عنوان جایگزینی برای شبکه‌های عصبی در یک مجموعه دبی‌سنج مجازی بررسی کردند. آن‌ها نشان دادند که این سازوکار به‌عنوان پشتیبان دبی‌سنج چند فاز 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 ایجاد شده توسط گوئز و همکاران دبی هر چاه را به‌عنوان تابعی از داده‌های جمع‌آوری شده مانند دما و فشار شیر فشارشکن و ویژگی سیالات در

1. Tree Regression

2. NN-RTE: Neural Network-Regression Trees

3. Imperialist Competitive Algorithm

4. ANN: Artificial Neural Networks

است. درحالی که تمامی این اطلاعات در چاه‌ها به صورت بر خط در دسترس نیست. از آنجا که اکثر چاه‌های ایران وارد دوره کاهش برداشت شده 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 دبی‌سنجی با اطلاعات پمپ انجام شده (مانند پژوهش دنی و همکاران [۱۳]) که موجب ایجاد یک نرم‌افزار تجاری شده است) نیز از مدل‌های شبکه عصبی استفاده شده 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 انجام شده است. چاه مورد بررسی از حجم داده بالا و تنوع کافی برخوردار نیست بنابراین، مدل‌های شبکه عصبی در این چاه با خطا قابل توجهی روبه‌رو هستند. به همین دلیل در این مطالعه، از روش‌های k همسایه نزدیک تر $(k\text{-NN})^4$ ، درخت تصمیم $(DT)^5$ و تقویت گرادیان $(GBR)^6$ استفاده شده است. همچنین، داده‌های نوفه‌دار^۷ که ممکن است به دلیل خرابی حسگرها در مجموعه داده پدیدار شوند، می‌توانند تأثیر معکوس قابل توجهی بر دقت و قابلیت اعتماد مدل‌ها داشته باشند. این نوع داده‌ها می‌توانند تأثیر منفی بر عملکرد دبی‌سنج مجازی داده محور داشته باشند؛ بنابراین، عملکرد دبی‌سنج مجازی داده محور در برابر داده‌های نوفه‌دار نیز مورد بررسی قرار گرفته است.

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٪ برش آب را ارائه دادند. شبکه عصبی با استفاده از داده‌های تولید واقعی، داده‌های سطحی و ته‌چاهی 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. با وجود 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^۱ نرم‌افزار نئورال فلو^۲ را توسعه داده است که مبتنی بر مدل شبکه عصبی است [۱۳]. این نرم‌افزار برای تخمین دبی در سیستم‌های دارای پمپ‌های شناور الکتریکی^۳ (ESP) با استفاده از رویکرد شبکه عصبی استفاده می‌شود. مشابه شبکه‌های عصبی‌ای که قبلاً مورد بحث 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 ته‌چاهی، 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 استفاده شده است و کمتر به تخمین دبی با استفاده از پمپ ESP به‌عنوان یک مدل داده محور پرداخته شده

1. Baker Hughes

2. NeuraFlow

3. Electrical Submersible Pump

4. K-Nearest Neighbor.

5. Decision Tree.

6. Gradient Boosting

7. Noise

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. در نهایت مدل با استفاده از داده‌های بخش آزمون اعتبارسنجی می‌شود. اطلاعات مربوط به مدل داده محور در بخش ۳ آمده است.

مدل داده محور

پیشرفت‌های سریع و شگفت‌انگیز روش‌های یادگیری ماشین باعث شده این روش‌ها در صنایع مختلف مورد توجه 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 GBR به‌عنوان مدل‌های داده محور برای پیش‌بینی نرخ جریان استفاده شده است. در ادامه مراحل ایجاد مدل داده محور و همچنین پیشینه هر یک از روش‌های انتخاب‌شده شرح داده می‌شود.

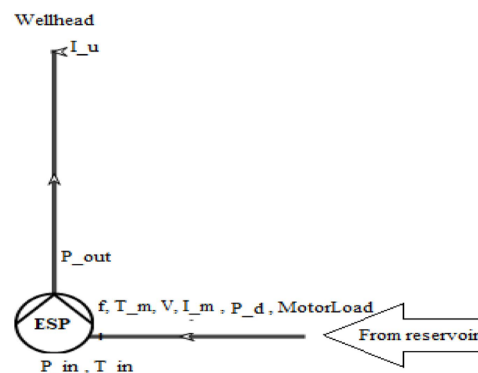
انتخاب ویژگی

در دبی‌سنج‌های 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، در حالی که پارامترهای نادرست یا نامربوط می‌تواند منجر به نتایج غیر بهینه و برآزش بیش‌ازحد شود.

در بخش دوم مدل‌های مورد استفاده توضیح داده شده است. بررسی مجموعه داده در بخش سه آورده شده است. نتایج در بخش ۴ آورده شده و در نهایت نتیجه‌گیری در بخش ۵ آمده است.

تعریف مسئله

نمای چاه مورد مطالعه 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 سرچاه 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 هستند.



شکل ۱ نمای چاه مورد مطالعه

برای ساخت یک دبی‌سنج مجازی داده محور ابتدا باید داده‌های مختلف جمع‌آوری شوند سپس از بین این داده‌ها آن‌هایی که ارتباط قوی‌تری با تابع هدف^۱ دارند به‌عنوان ورودی انتخاب‌شده و آنگاه پیش‌پردازش روی آن‌ها انجام می‌گیرد یعنی داده‌های پرت حذف و نوفه‌ها فیلتر می‌شوند. در گام بعدی داده‌های پیش‌پردازش شده وارد مرحله توسعه مدل می‌شود. در این قسمت مجموعه داده

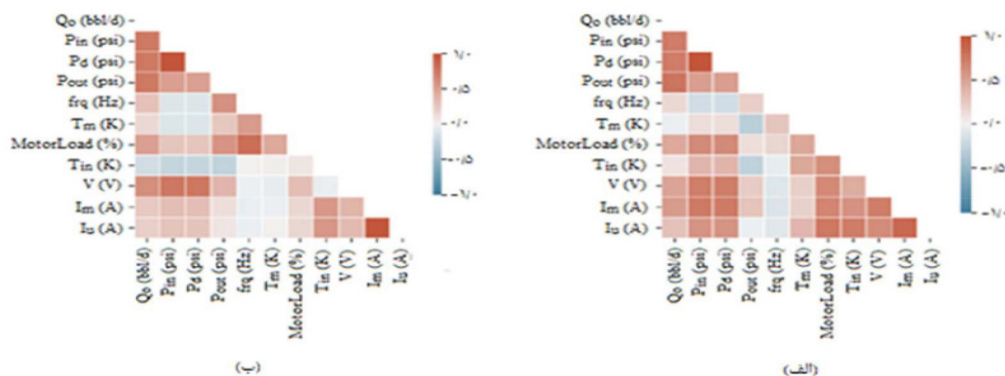
۱. تابع هدف پارامتری است که قرار است مقدار آن تخمین زده شود.

درحالی که 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 داده شده 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. کامیبری و همکاران از رابطه بقای انرژی در پمپ برای تخمین دبی استفاده کردند [۱۴]. طبق این رابطه 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.

انتخاب پارامترهای ورودی مناسب برای روش یادگیری ماشین می‌تواند یک فرآیند پیچیده و چالش‌برانگیز باشد، زیرا داده‌ها ممکن است حاوی متغیرهای متعددی باشند که برخی از آن‌ها ممکن است نامربوط، زائد یا نوفه‌دار باشند. با انتخاب ویژگی‌های مرتبط‌تر، می‌توانیم از تأثیرپذیری مدل از نوفه 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 هر کدام 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 براساس ارتباط بین آن‌ها با متغیر هدف استفاده می‌کنند معیارهای رایج مورد استفاده در روش‌های فیلتر عبارتند از انتخاب ویژگی مبتنی بر همبستگی^۱، اطلاعات متقابل^۲ و آزمون مجذور کای^۳ ویژگی‌ها براساس معیارهای از پیش تعریف‌شده مانند آستانه مشخصی از مقدار متریک انتخاب یا حذف می‌شوند.

ماتریس همبستگی اسپیرمن و پیرسون روش‌های آماری 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. همبستگی پیرسون رابطه خطی بین دو متغیر پیوسته را اندازه‌گیری می‌کند،

1. Correlation-Based
2. Mutual Information
3. Chi-Square



شکل ۲ ارتباط بین نرخ جریان نفت و ویژگی‌ها با استفاده از معیارهای الف) اسپیرمن و ب) پیرسون

می‌توان به صورت زیر تعریف کرد.

$$f_M(x_j) = \sum_{m=1}^M \gamma_m h_m(x_j) \quad (1)$$

در رابطه فوق h_m نشان‌دهنده یک یادگیرنده ضعیف است که به تنهایی عملکرد ضعیفی دارد و به عنوان 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 به روزرسانی تخمین اولیه با تخمین جدید استفاده می‌کند؛ که نتیجه آن ایجاد یک مدل نهایی با ادغام تمام تخمین‌های اولیه با وزن‌های مناسب است.

روش k همسایه نزدیک‌تر

روش k همسایه نزدیک‌تر (k-NN) اولین بار توسط دودانی [۱۶] توسعه یافت. جوزویک [۱۷] نسخه پیشرفته‌تر این روش را معرفی کرد. در زمینه یادگیری ماشین، k-NN یک روش غیرمعیاری^۱ پرکاربرد برای طبقه‌بندی و برازش 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 براساس فاصله (مانند فاصله اقلیدسی یا فاصله منهتن) پیش‌بینی یا طبقه‌بندی برای نقطه داده جدید را انجام می‌دهد.

روش‌های یادگیری ماشین

روش GBR

روش تقویت‌کننده گرادیان (GBR) یک شیوه یادگیری ماشین قدرتمند است که هم برای مسائل برازش 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 برازش 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 به خوبی پیش‌بینی نشده بودند. با انجام این کار به صورت مکرر، گروه به تدریج 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 براساس خروجی مدل ارزشمند باشد. یک مدل GBR با تعداد M درخت را

1. منظور این است که در این روش رابطه پارامتری از پیش تعیین شده‌ای میان متغیرهای ورودی و خروجی برقرار نمی‌شود

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 برآزش با درخت تصمیم، هدف پیش‌بینی مقدار یک متغیر هدف با استفاده از قوانین تصمیم‌گیری ساده است. در واقع، درخت تصمیم یک ساختار درخت مانند دارد که هر گره داخلی آن یک تصمیم براساس یک ویژگی از داده‌ها را نشان می‌دهد و هر گره برگ نشان‌دهنده مقدار پیش‌بینی شده متغیر هدف است. قوانین تصمیم‌گیری از داده‌های آموزشی یاد گرفته می‌شوند و ساختار درخت براساس 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 به راحتی تجسم کرد و حتی توسط افراد غیرمتخصص قابل درک است. این درخت‌ها به‌عنوان ابزاری مفید برای تحلیل داده‌ها و فرآیندهای تصمیم‌گیری پیچیده مورد استفاده قرار می‌گیرند. همچنین، درخت‌های تصمیم قادر به مدیریت داده‌های عددی و مقوله‌ای هستند و نسبت به مقادیر پرت و گم‌شده نیز مقاومت دارند.

تنظیم پارامترها

تنظیم پارامترها در الگوریتم‌های یادگیری ماشین اهمیت بالایی دارد، زیرا بر کیفیت مدل‌سازی و دقت پیش‌بینی‌ها تأثیر می‌گذارد. در واقع 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.

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]. روش k -NN با مقایسه یک نمونه آزمون داده‌شده (x, y) با یک مجموعه آموزشی $D = [(x_i, y_i)]$ 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:

$$d_{Euc} = \sqrt{(x - x_i)^2} \quad (2)$$

$$d_{Man} = |x - x_i| \quad (3)$$

که در آن d_{Euc} فاصله اقلیدسی و d_{Man} فاصله منهتن است. در مرحله بعد، k -NN داده‌ها را براساس 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 رابطه ۴ نشان داده‌شده است [۱۹].

$$y_{pred} = \frac{\sum_{i=1}^k w_i y_i(x)}{\sum_{i=1}^k w_i} \quad (4)$$

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

روش درخت تصمیم

درخت تصمیم (Decision Tree) یک روش یادگیری ماشین ساده و محبوب است که برای طبقه‌بندی داده‌ها و برآزش (پیش‌بینی مقادیر عددی) استفاده می‌شود. این روش از یک نمودار درخت مانند برای

داده‌ها مشخص شد که برخی از داده‌ها در فاصله بسیار زیاد از میانگین قرار دارند. به‌عنوان مثال دبی نفت در بعضی از روزها صفر گزارش شده بود، از آنجاکه این بخش از داده‌ها فاصله قابل توجهی 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. **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 (bbl/d) ۴۰۰ تا ۶۰۰ (bbl/d) 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 باین‌حال از آنجاکه 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.

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 از جمله پارامترهای مهم به شمار می‌روند. تقویت گرادیان نیز از پارامترهایی مانند نرخ یادگیری، تعداد درختان و عمق درختان استفاده می‌کند. این پارامترها هر کدام 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.

توضیحات داده‌ها

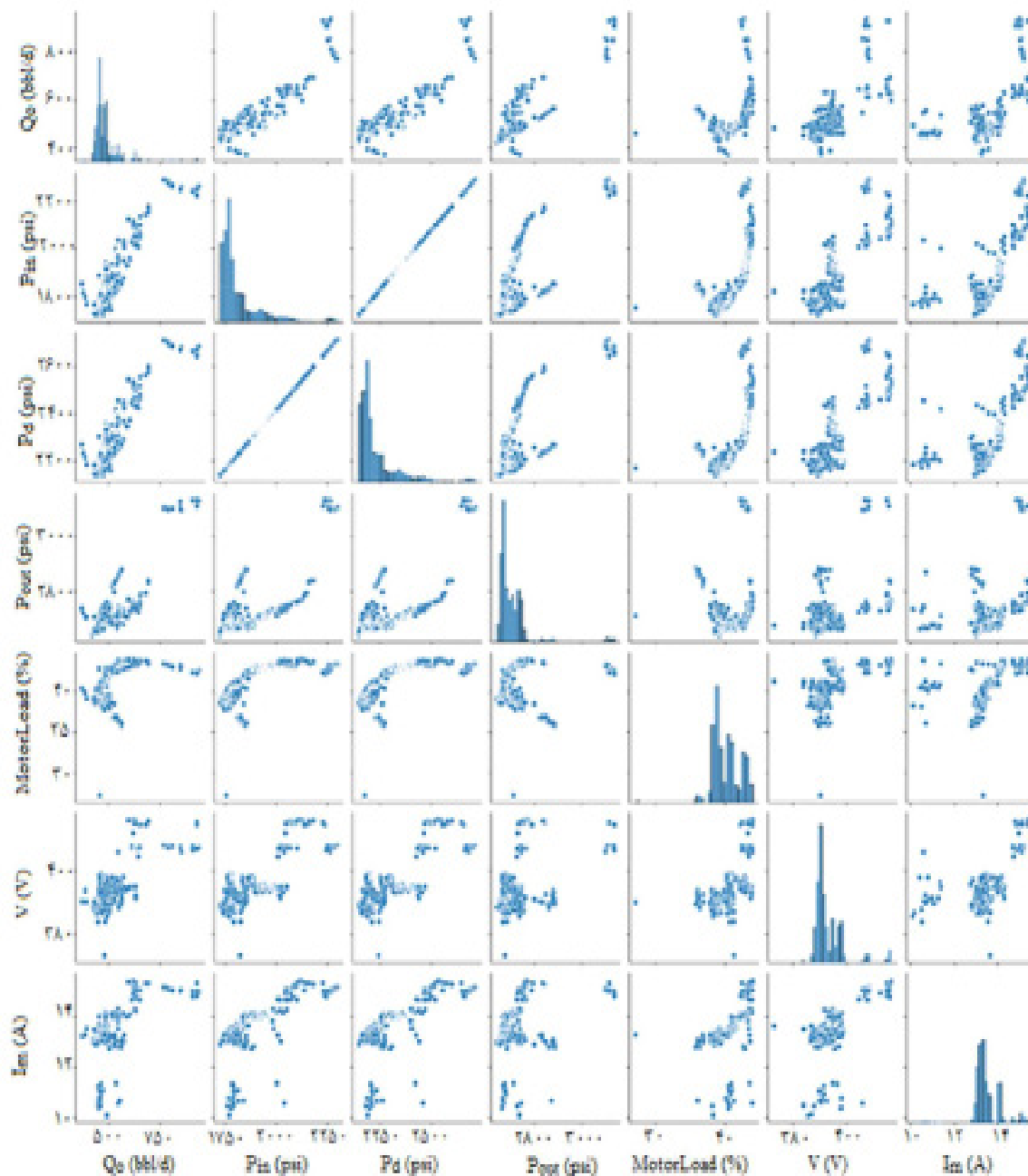
اطلاعات مورد استفاده در این مطالعه به‌صورت یک روز در میان یا دو روز در میان و از تاریخ ۱۰ اکتبر ۲۰۲۰ تا ۸ جولای ۲۰۲۳ جمع‌آوری شده است و تعداد کل داده‌ها ۹۸۰ عدد است. در این بازه زمانی تغییرات زیادی در چاه‌های میدان صورت گرفته و همچنین عملیات تزریق آب و یا گاز و سایر فرآیندهایی که می‌تواند منجر به تغییر جدی در مکانیزم‌های تولیدی گردد، گزارش نشده است. در بخش پیش‌پردازش داده‌ها ابتدا داده‌های گمشده از مجموعه داده حذف گردید. سپس با بررسی بیشتر

جدول ۱ مقدار پارامترهای مهم الگوریتم‌های یادگیری ماشین استفاده‌شده

مقدار پیشنهادی	پارامتر	الگوریتم
۵	تعداد همسایه‌ها (k)	k -NN
فاصله	نوع وزن دهی	
اقلیدسی	نوع فاصله	
۷	عمق درخت	DT
۵	نمونه‌ها	
۰/۱	نرخ یادگیری	GBR
۵۰۰	تعداد درختان	
۲	عمق درختان	

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}	Motor Load	I_m	V	Q_e
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 رسم نمودارهای دوتایی بین ویژگی‌های انتخاب‌شده و نرخ جریان نفت

ارائه نتایج و بحث و بررسی

در این بخش عملکرد هر کدام از روش‌های DT، GBR و k-NN در دو حالت داده‌های با نوفه کم و داده‌های با نوفه شدید مورد بررسی 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 به خوبی 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:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{test,i} - y_{pred,i})^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{test,i} - y_{pred,i}| \quad (6)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_{test,i} - y_{pred,i}}{y_{test,i}} \right| \quad (7)$$

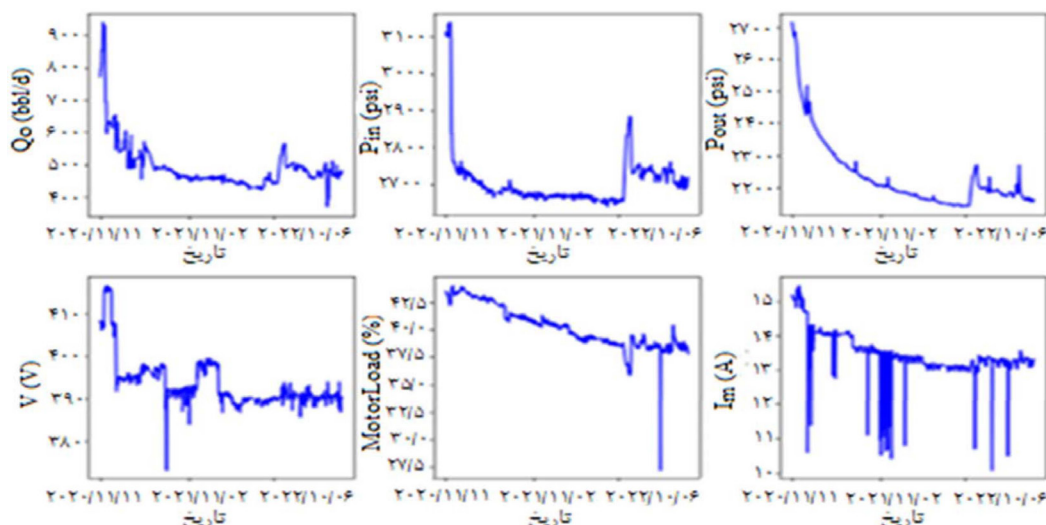
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{test,i} - y_{pred,i})^2}{\sum_{i=1}^n (y_{test,i} - \bar{y})^2} \quad (8)$$

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

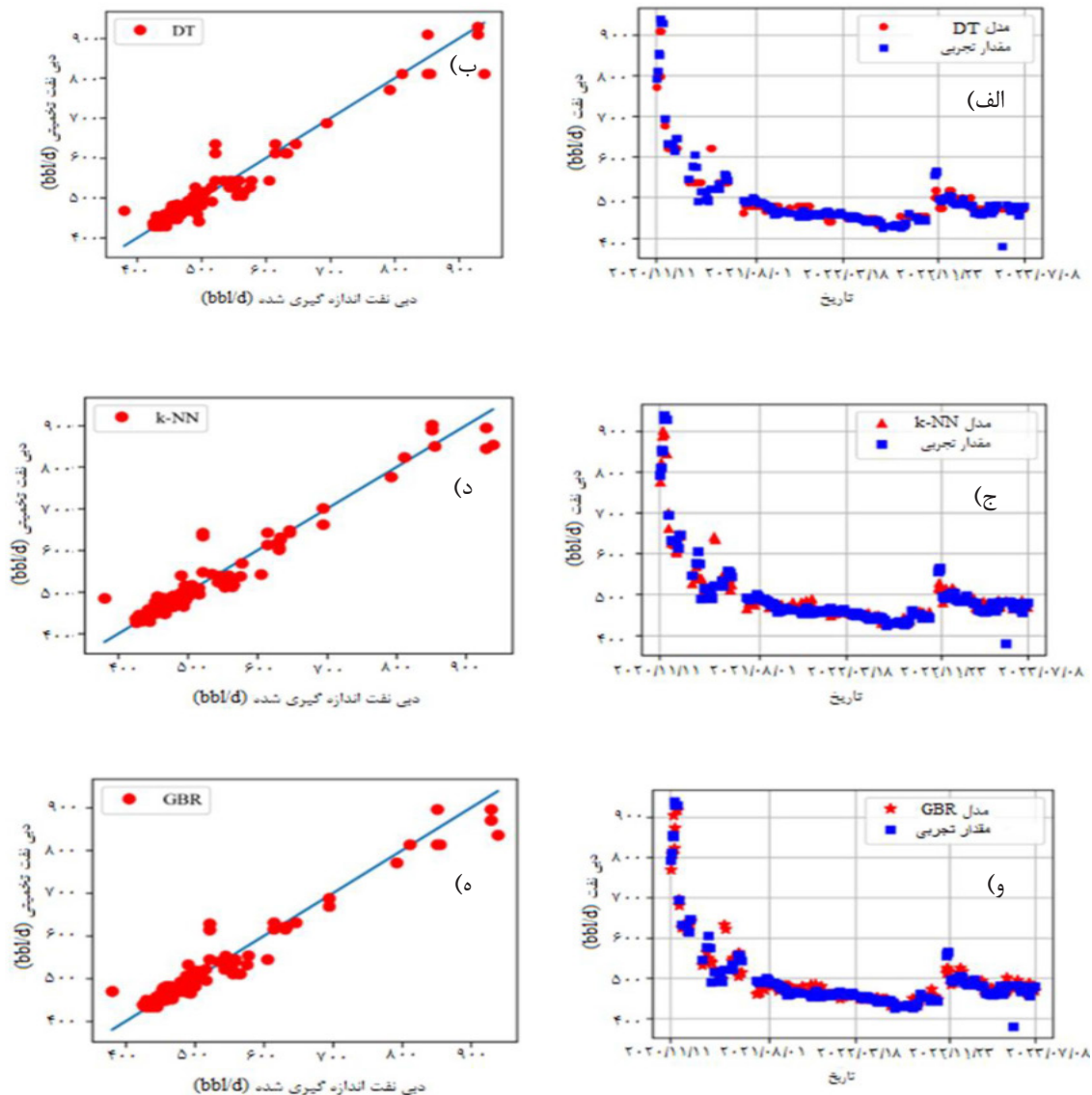
موجود بین ویژگی‌ها و تابع هدف را با استفاده از داده‌های آموزش یاد گرفته و سپس دقت آن‌ها با استفاده از داده‌های آزمون بررسی می‌گردد. در پژوهش حاضر نیز از ۷۵٪ داده‌ها برای آموزش و ۲۵٪ برای آزمون استفاده‌شده است.

داده‌های با نوفه کم

داده‌های واقعی مورد استفاده در مدل داده محور در شکل ۴ نشان داده‌شده است. شکل ۴ نشان می‌دهد ویژگی‌های ورودی دارای نوفه شدیدی نیستند که این مورد می‌تواند به مدل‌های ارائه‌شده برای پیش‌بینی دقیق‌تر کمک کند. البته برخی از ویژگی‌ها مانند جریان و ولتاژ موتور دارای نوفه هستند. شکل ۵ الف، ج و و) دبی تخمینی نفت توسط مدل‌های داده محور و دبی تجربی نفت را برحسب زمان نمایش می‌دهد. همان‌طور که مشاهده می‌شود مدل‌های دبی‌سنج داده محور در دبی‌های بیشتر از ۵۰۰ (bbl/d) نسبت به بازه کمتر از ۵۰۰ (bbl/d) خطای بالاتری دارند که دلیل آن حجم کمتر داده‌ها در این بازه نسبت به بازه کمتر از ۵۰۰ (bbl/d) است. از مقایسه شکل ۵ الف، ج و و) استنباط می‌شود که بین روش‌های ارائه‌شده با قاطعیت نمی‌توان درباره برتری یک روش نسبت به روش دیگر صحبت کرد.



شکل ۴ رسم ورودی‌های مدل داده محور مجموعه داده با نوفه کم



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

عملکرد بسیار نزدیکی دارند باین حال مدل k-NN با اختلاف کمی نسبت به دو مدل دیگر عملکرد بهتری دارد. مدل DT نیز دقت پایین تری نسبت به دو مدل دیگر دارد.

داده‌های نوفه‌دار

نوفه 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.

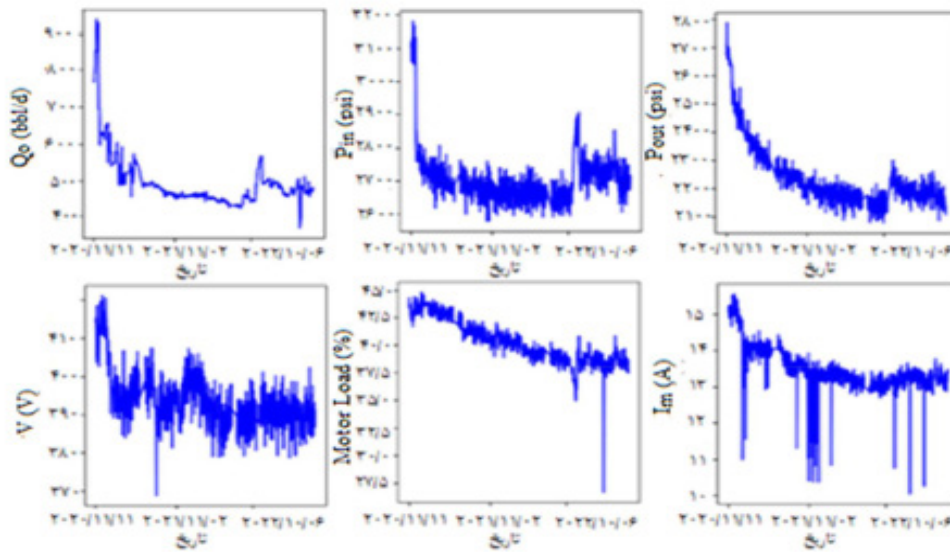
شکل ۵- (ب، د) و (ه) نشان می‌دهند که مدل‌های داده محور ارائه‌شده در محدوده کمتر از ۵۰۰ (bbl/d) عملکرد خوبی دارند و نمی‌توان گفت که کدام مدل عملکرد بهتری داشته است. باین حال در محدوده بالاتر از ۵۰۰ (bbl/d) به ترتیب مدل k-NN، DT و GBR بهترین عملکرد را دارند. برای بررسی دقیق‌تر عملکرد هر روش، دقت آن‌ها براساس معیارهای ارزیابی در جدول ۳ آمده است. جدول ۳ نشان می‌دهد که مدل‌های داده محور ارائه‌شده

جدول ۳ ارزیابی عملکرد مدل‌های ارائه‌شده در مجموعه داده با نوفه کم برای داده‌ها آزمون

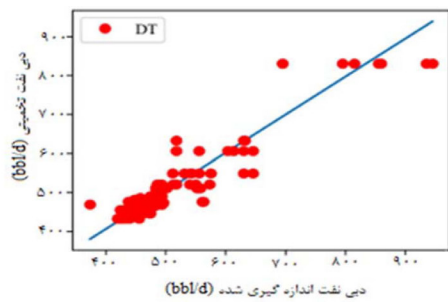
روش یادگیری ماشین	MAE	MAPE	RMSE	R ²
DT	۱۲/۹۵۶۳	۲/۴۷۷۳	۲۱/۵۶۶۹	۰/۹۴۱۳
k-NN	۱۰/۹۵۶۵	۲/۰۵۹۲	۲۰/۳۱۴۰	۰/۹۴۷۹
GBR	۱۲/۱۸۳۴	۲/۳۱۹۱	۲۰/۲۸۱۵	۰/۹۴۷۷

همچنین می‌تواند منجر به برآزش بیش‌ازحد شود. برآزش بیش‌ازحد زمانی اتفاق می‌افتد که یک مدل از نوفه موجود در داده‌ها به‌جای الگوهای زیربنایی یاد می‌گیرد و در نتیجه 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 نوفه در داده‌ها وجود دارند که از جمله آن‌ها می‌توان به خطاهای اندازه‌گیری، مقادیر از دست‌رفته، 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 نوفه با توزیع گاوسی به پارامترهای اندازه‌گیری شده ورودی، عملکرد مدل‌ها در شرایط دارای نوفه بررسی شده است. **شکل ۶** داده‌های ورودی مورد مطالعه در این بخش از مقاله را نشان می‌دهد. در این حالت نیز مجموعه داده به دو قسمت بخش آموزش و بخش آزمون تقسیم شده است. **شکل ۷ الف، ج و و** دبی تخمینی نفت توسط مدل‌های داده محور و دبی تجربی نفت را برحسب زمان نمایش می‌دهند. همان‌طور که مشاهده می‌شود نوفه موجود در داده‌ها تأثیر منفی بر عملکرد دبی‌سنج‌های مجازی داده محور داشته است. تأثیر نوفه‌ها در مناطقی که حجم داده کمتری دارند بیشتر است. **شکل ۷ ب، د و ه** نشان می‌دهند که مدل داده محور k-NN بهترین عملکرد را در برابر ورودی‌های دارای نوفه دارد. دلیل این امر این است که این مدل از همسایه‌ها برای

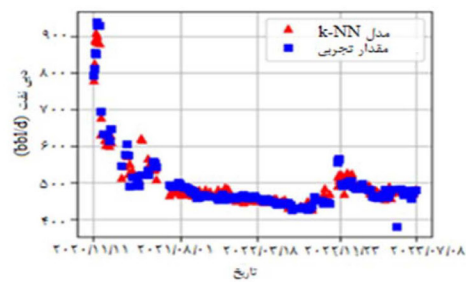
تخمین مقدار استفاده می‌کند. مدل DT نیز بیشتر از مدل‌های k-NN و GBR تحت تأثیر نوفه موجود در ورودی‌ها بوده است. مدل GBR توانسته با ترکیب چندین مدل DT عملکرد این مدل را بهبود دهد. **جدول ۴** میزان دقت هر کدام از مدل‌ها را در برابر داده‌های نوفه‌دار نشان می‌دهد. با توجه به نتایج ارائه‌شده در این جدول مشاهده می‌شود که مدل k-NN بهترین عملکرد را در برابر داده‌های نوفه‌دار داشته است و نوفه تأثیر کمی بر عملکرد این روش داشته است. مدل DT نیز بیشترین تأثیر را از نوفه گرفته است. بررسی نتایج به‌دست‌آمده برای ورودی‌های نوفه‌دار و بدون نوفه نشان داد که مدل DT یک مدل ساده است که در یک مجموعه داده با نوفه کم، تنوع ناکافی و حجم محدود می‌تواند تخمین قابل قبولی داشته باشد. باین حال این مدل در برابر نوفه 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 با نوفه کم عملکرد نزدیک به عملکرد DT دارد. باین حال در مجموعه داده نوفه‌دار عملکرد مدل GBR بهتر از مدل DT است. الگوریتم k-NN عملکرد بسیار خوبی در هر دو مجموعه داده با نوفه کم و نوفه‌دار 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.



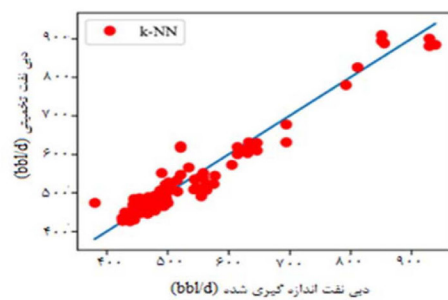
شکل ۶ داده‌های ورودی با یک درصد انحراف معیار نوفه



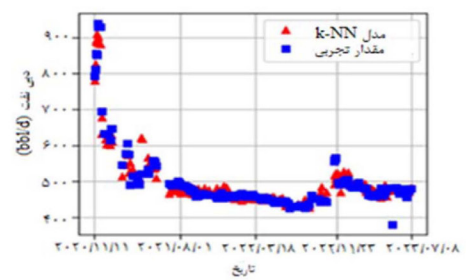
(ب)



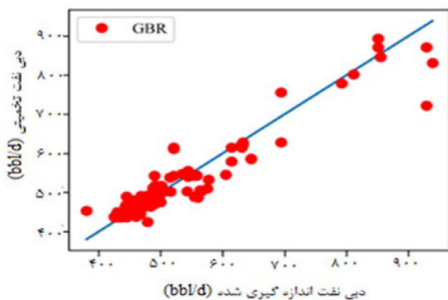
(الف)



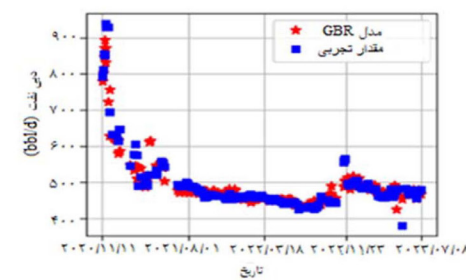
(د)



(ج)



(ه)



(و)

شکل ۷ دبی تخمینی نفت و دبی تجربی نفت نسبت به زمان (الف) DT (ج) k-NN و GBR و دبی تخمینی نفت در برابر دبی تجربی نفت (ب) DT (د) k-NN (ه) GBR در حالت ورودی با یک درصد انحراف معیار نوفه توسط مدل‌های داده محور

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

روش مورد بررسی	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

دبی نفت آن‌ها در محدوده بیش‌تر از ۵۰۰ (bbl/d) بوده کم است. همین امر چالشی در برابر استفاده از دبی‌سنج‌های مجازی داده محور ایجاد کرده است، با این حال 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 دارای نوفه با اضافه کردن یک درصد انحراف معیار نوفه با استفاده از تابع توزیع گاوسی به داده‌های ورودی مورد ارزیابی قرار گرفت. مشاهده شد که روش‌های یادگیری ماشین استفاده‌شده در اثر وجود نوفه در اطلاعات ورودی، خصوصاً در محدوده‌های با داده ناکافی دچار خطا می‌شوند. روش k-NN بهترین عملکرد را در برابر داده‌های دارای نوفه در مقایسه با دو روش دیگر به نمایش گذاشت که نشان از مدیریت مناسب این مدل در حضور داده‌های دارای نوفه است. روش DT نیز در مقایسه با دو روش دیگر، بیشترین خطا را در حضور داده‌های دارای نوفه نشان داد. نتایج این مطالعه نشان داد که انتخاب داده ورودی و مدل یادگیری ماشین مناسب می‌تواند عملکرد بسیار مناسبی را برای دبی‌سنج مجازی به ارمغان بیاورد. در ادامه پیشنهاد می‌شود که از داده‌های ته چاهی و سر چاهی در کنار داده‌های پمپ به‌عنوان دبی‌سنج مجازی داده محور استفاده شود.

تشکر و قدردانی

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.

باوجود مزایایی که برای مدل دبی‌سنج 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 درون چاهی در نظر گرفته نشده است. به همین دلیل نیاز است که دقت مدل در طول زمان به‌صورت مکرر امتحان شود. در این پژوهش دبی‌سنج مجازی داده محور برای تخمین دبی نیازمند اطلاع از شرایط درون چاه و مخزن نیست. باین‌حال در صورتی که شرایط چاه و مخزن تغییر کنند مدل داده محور نیازمند آموزش مجدد است.

نتیجه‌گیری

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 (Mo-)، 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 دبی‌سنج‌های 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

مراجع

- [1]. Dayev, Z. A. (2020). Application of artificial neural networks instead of the orifice plate discharge coefficient. *Flow Measurement and Instrumentation*, 71, 101674. doi.org/10.1016/j.flowmeasinst.2019.101674.
- [2]. Bismukhametov, T., & Jäschke, J. (2020). First principles and machine learning virtual flow metering: a literature review. *Journal of Petroleum Science and Engineering*, 184, 106487. doi.org/10.1016/j.petrol.2019.106487.
- [3]. Mercante, R., & Netto, T. A. (2022). Virtual flow predictor using deep neural networks. *Journal of Petroleum Science and Engineering*, 213, 110338. doi.org/10.1016/j.petrol.2022.110338.
- [4]. Al-Qutami, T. A., Ibrahim, R., Ismail, I., & Ishak, M. A. (2018). Virtual multiphase flow metering using diverse neural network ensemble and adaptive simulated annealing. *Expert Systems with Applications*, 93, 72-85. doi.org/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. doi.org/10.1016/j.asoc.2012.10.009
- [7]. Bismukhametov, T., & Jäschke, J. (2019). Oil production monitoring using gradient boosting machine learning algorithm. *Ifac-Papersonline*, 52(1), 514-519. doi.org/10.1016/j.ifacol.2019.06.114.
- [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. doi.org/10.1016/j.petrol.2021.108586.
- [9]. Hotvedt, M., Grimstad, B., Ljungquist, D., & Imsland, L. (2022). On gray-box modeling for virtual flow metering. *Control Engineering Practice*, 118, 104974. doi.org/10.1016/j.conengprac.2021.104974.
- [10]. AlAjmi, M. D., Alarif, 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. doi.org/SPE-173394-MS.
- [11]. Sandnes, A. T., Grimstad, B., & Kolbjørnsen, O. (2021). Multi-task learning for virtual flow metering. *Knowledge-Based Systems*, 232, 107458. doi.org/10.1016/j.knosys.2021.107458.
- [12]. Al-Jasmi, A., Goel, H.K., Nasr, H., Querales, M., Rebeschini, J., Villamizar, M.A., Carvajal, G.A., Knabe, S., Rivas, F. and 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. doi.org/10.2118/164813-MS.
- [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. doi.org/10.2118/163704-MS.
- [14]. Camilleri, L. A., Banciu, T., Ditoiu, G., & Petrom, S. A. (2010, March). First Installation of 5 ESPs Offshore Romania-A Case Study and Lessons Learned. In *SPE Intelligent Energy International Conference and Exhibition* (pp. SPE-127593). SPE. doi: 10.2118/127593-MS.
- [15]. Rao, H., Shi, X., Rodrigue, A.K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X. and Gu, L., (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, 74, 634-642. doi.org/10.1016/j.asoc.2018.10.036.
- [16]. Dudani, S. A. (1978). The distance-weighted k-nearest neighbor rule. *IEEE trans. on systems, man and cybernetics*, 8(4), 311-313.
- [17]. Józwiak, 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. doi.org/10.1016/0167-8655(83)90064-8.
- [19]. Devroye, L., Györfi, L., Krzyżak, A., & Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3), 1371-1385. doi.org/10.1214/aos/1176325633.
- [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. doi.org/10.1016/j.rse.2005.05.008.
- [21]. Breiman, L. (2017). *Classification and regression trees*. Routledge. 1st Edition. 25 October 2017. New York. Chapman and Hall/CRC. doi.org/10.1201/9781315139470.



Estimation of Oil Flow Production of Well Employing Machine Learning Algorithms Using Electrical Submersible Pump (ESP)

Mohammad Bagher Seddighi¹, Majid Siavashi¹ and Rohaldin Miri²

1. School of Mechanical Engineering, Iran University of Science and Technology, Tehran, Iran

2. Faculty of Chemical, Oil and Gas Engineering, Iran University of Science and Technology, Tehran, Iran

msiavashi@iust.ac.ir

DOI:10.22078/pr.2024.5417.3415

Received: June/02/2024

Accepted: June/30/2024

Introduction

In the oil and gas industry, the measurement of the flow rate of extracted fluids from wells is essential for optimizing production and reservoir management [1]. Conventional methods include well testing and multiphase flow metering (MPFM), both of which require physical equipment and are costly. Consequently, virtual flow meters (VFMs) have been developed that predict flow rates using existing well data, without the need for physical equipment [2].

Virtual flow meters are divided into two categories: physics-based and data-driven. In the physics-based approach, the flow rate is estimated using conservation equations. However, these equations are complex, and analytical solutions do not exist for some of them. In the data-driven approach, a mathematical model is created using field data and machine learning methods to estimate the flow rate [2,3].

This research aims to estimate the oil flow rate in one of the wells in a southern Iranian oil field. Due to the small volume of data and insufficient diversity, neural network algorithms cannot accurately estimate the flow rate. Therefore, in this study, the k-nearest neighbors (k-NN), decision tree (DT), and gradient boosting (GBR) methods have been used.

Problem Definition

The schematic of the well under study is depicted in Fig. 1. In this research, only the ESP pump has been modeled. This leads to a less complex model and

eliminates the need for bottom-hole and wellhead data. This is beneficial for assets that lack or have faulty bottom-hole or wellhead measurements.

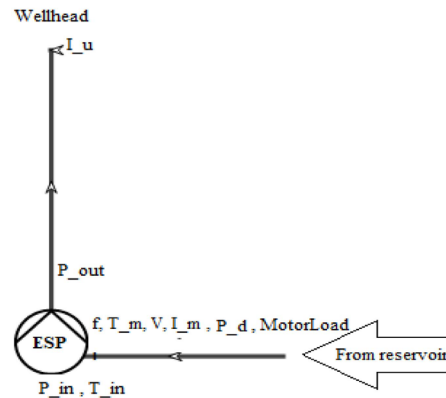


Fig. 1 Schematic of the studied well.

Data-Driven Model

machine learning models have garnered significant attention across various industries due to their rapid advancements. These methodologies empower computers to learn from data and adapt their behavior accordingly. In the present study, three machine learning models, namely DT, k-NN, and GBR, were employed for flow rate prediction.

Input parameters play a crucial role in these models, and their appropriate selection can aid in enhancing the model's performance. However, this process can be challenging as it necessitates avoiding irrelevant

or noisy variables. By selecting suitable features, it is possible to better generalize overlooked data.

In data-driven flowmeters, input parameters, recognized as features or variables, are the building blocks of machine learning models. These parameters represent data features that the solution method uses for prediction or classification. There are numerous methods for selecting an appropriate feature. Filter methods are a popular approach for feature selection, as they utilize statistical criteria to rank and select features based on their correlation with the target variable. The Spearman and Pearson correlation matrices are used to measure the relationship between

two variables. Pearson correlation measures linear relationships, while Spearman correlation measures monotonic relationships. Features with high correlation coefficients may be redundant, and only one of them is selected. Features with low correlation coefficients may be less relevant and are considered for removal. Fig. 2 shows the relationship between different characteristics with Spearman and Pearson criteria. According to this figure, it can be concluded that inlet pressure (P_{in}), outlet pressure (P_{out}), motor load (Motor Load), motor current (I_m) and motor voltage (V) are considered as suitable features for flow estimation.

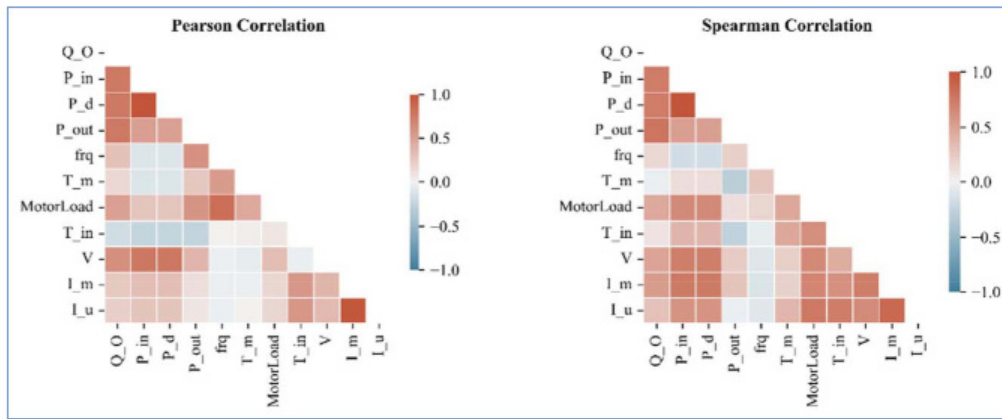


Fig. 2 Plotting the inputs of the data-driven model of the low-noise dataset.

Results and Discussion

In this section, the performance of each of the DT, GBR, and k-NN methods is examined under two conditions: low-noise data and high-noise data, and compared with each other. Various statistical metrics exist for evaluating the performance of machine learning methods. These metrics can effectively demonstrate the accuracy of the proposed models. Here, the metrics of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2) are used, and their respective relations are presented below:

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{test,i} - y_{pred,i}| \quad (2)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_{test,i} - y_{pred,i}}{y_{test,i}} \right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{test,i} - y_{pred,i})^2}{\sum_{i=1}^n (y_{test,i} - \bar{y})^2} \quad (4)$$

while $y_{test,i}$ is the test data, $y_{pred,i}$ is the predicted value, \bar{y} is the average value of the sample and n is the given number

of the test. In machine learning, the data set is usually divided into two parts: training data and test data. The proposed models learn the existing relationships between features and the target function using the training data, and then their accuracy is examined using the test data. In the present study, 75 percent of the data is used for training and 25 percent for testing.

Low-Noise Data

The actual data used in the data-driven model is depicted in Fig. 3. Moreover, Fig. 3 demonstrates that the input features do not possess significant noise, which can aid the proposed models in making more accurate predictions. However, some features, such as motor current and voltage, are noisy.

Fig. 4 (a, b and c) presents the estimated oil flow rate by the data-driven models and the experimental oil flow rate for the test data set. As observed, this model has performed well against these data. One of the reasons for the error in the flow rates of 500 to 600 could be the low volume of training data in this range. From the comparison between the proposed methods, it is not definitively possible to speak about the superiority of one method over another. Therefore, to examine the precise performance of each method, their accuracy based on the evaluation metrics is provided in Table 1.

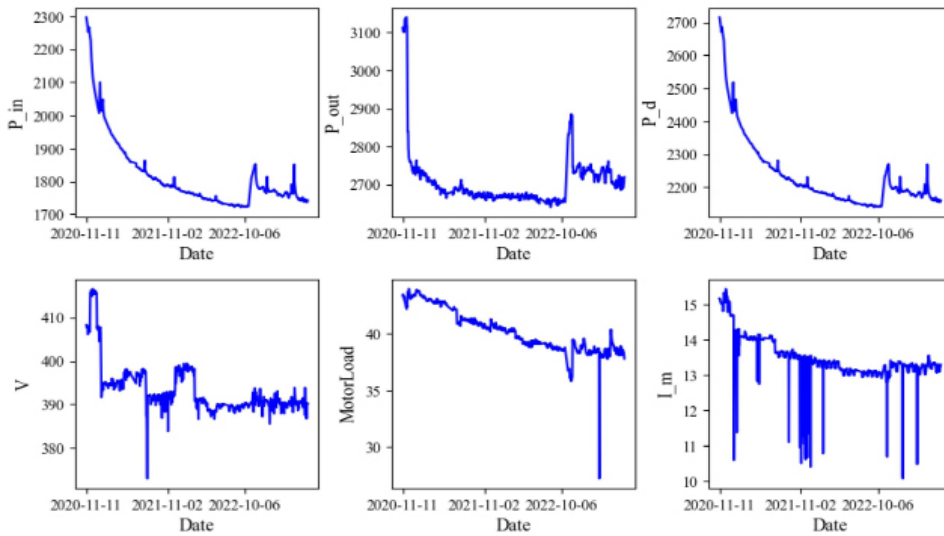


Fig. 3 Plotting the inputs of the data-driven model of the low-noise dataset.

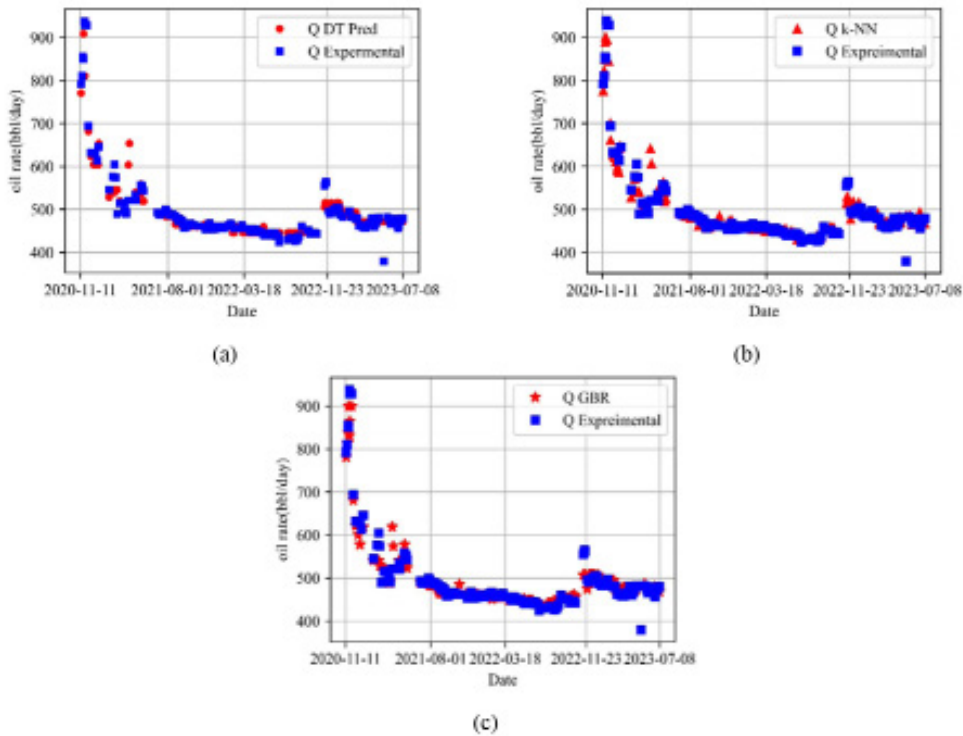


Fig. 4 Experimental oil flow rate and predicted oil flow rate for test data set in low noise input mode by data-driven models (a) DT (b) k-NN (c) GBR.

Table 1 Evaluation of the performance of the presented models in the data set with missing noise for the test data.

ML	MAE	MAPE	RMSE	R ²
DT	12.9563	2.4773	21.5669	0.9413
k-NN	10.9565	2.0592	20.3140	0.9479
GBR	12.1834	2.3191	20.2815	0.9477

Table 1 indicates that the presented data-driven models have very similar performances. However, the k-NN model performs slightly better than the other two models. The DT model also has a lower accuracy compared to the other two models.

Conclusions

In the present study, machine learning methods including Gradient Boosting Regressor (GBR), Decision Tree (DT), and k-Nearest Neighbors (k-NN) were used to estimate the oil flow rate from an oil well located in southern Iran. Suitable features for model training were selected using Spearman and Pearson statistical criteria. Moreover, this study utilized input and output data from Electric Submersible Pumps (ESP). In addition, the results suggest that pump input and output data can serve as inputs for data-driven flow meters. Furthermore, the frequency distribution of the data indicates a scarcity of instances with an oil flow

rate exceeding 500 barrels per day, posing a challenge for the use of data-driven virtual flow meters. However, the results demonstrate that the proposed models had acceptable performance. Moreover, the k-NN model showed the best performance compared to the other two models.

References

1. Dayev, Z. A. (2020). Application of artificial neural networks instead of the orifice plate discharge coefficient. *Flow Measurement and Instrumentation*, 71, 101674. doi.org/10.1016/j.flowmeasinst.2019.101674.
2. Bismukhametov, T., & Jäschke, J. (2020). First principles and machine learning virtual flow metering: a literature review. *Journal of Petroleum Science and Engineering*, 184, 106487. doi.org/10.1016/j.petrol.2019.106487.
3. Mercante, R., & Netto, T. A. (2022). Virtual flow predictor using deep neural networks. *Journal of Petroleum Science and Engineering*, 213, 110338. doi.org/10.1016/j.petrol.2022.110338.