ارزیابی پارامترهای موثرجهت پیش‌بینی عیار پتاسیم شورابه با استفاده از الگوریتم‌های ماشین بردار پشتیبان و جنگل تصادفی (مطالعه موردی: پلایای شهرستان خورو بیابانک، استان اصفهان)

چکیده

اهمیت پتاسیم دربالا بردن کمیت و کیفیت محصولات کشاورزی، تقاضا را برای کودهای پتاسیمی افزایش داده است. تضمین استخراج پتاسیم از شورابه‌های زیرزمینی مقدار عیار پتاسیم درآنهاست. هدف این پژوهش استفاده از الگوریتم‌های جنگل تصادفی (RF) و ماشین‌بردار پشتیبان(SVM) به‌منظور اولویت‌بندی پارامترهای مؤثر بر عیار پتاسیم شورابه زیرزمینی در پلایای خور و بیابانک استان اصفهان است. به همین منظور تعداد ۵۵ پارامتر در ۱۲ گمانه حفاری اندازه گیری شد. پارامترهای اندازه‌گیری شده به عنوان متغیرهای مستقل شامل درصد رطوبت اشباع مغزه در ۱۵عمق مختلف، جرم مخصوص ظاهری مغزه در ۱۵عمق مختلف، تخلخل مغزه در ۱۵عمق مختلف، مساحت پلی‌گون، عمق آب زیرزمینی، عمق لایه نمک، پتاسیم لایه سطحی، دانسیته شورابه و میزان عناصر کلسیم، منیزیم، سدیم، کلر و عیار پتاسیم به عنوان متغیر وابسته وارد مدل شدند. در مدلRF برای اولویت‌بندی، پارامترها از روش‌های اهمیت ویژگی جایگشت(PFI) و حذف ویژگی جایگشتی(RFE) استفاده شد. درکرنل‌های مختلف الگوریتم SVM به منظور جلوگیری از هم‌خطی پارامترهای مستقل، تمام ترکیب‌های حاصل از متغیرهای مستقل با درنظر گرفتن ضریب تورم واریانس کمتر از ۸ و بالاترین ضریب تعیین وکمترین خطای MSE بررسی و به عنوان بهترین ترکیب انتخاب شدند. پارامترهای مؤثر در پیش‌بینی عیار پتاسیم شورابه در الگوریتم RF و تابع خطی الگوریتم SVM به ترتیب sp، ap، duw، slp، SAR و n، sp، duw و SAR بودند که منجر به بهترین نتیجه (ضریب تعیین زیاد و خطای کم) شدند. ضریب تعیین برای هر دو مدل به ترتیب ۰/۹۹ و ۰/۹۷ که نشان‌دهنده دقت خوب هر دو الگوریتم است.

واژه‌های کلیدی: *پیش‌بینی عیار، جنگل تصادفی، شورابه، ماشین بردار پشتیبان.*

ABSTRACT

The importance of potassium in agricultural products has increased the demand for potassium fertilizers.Potassium grade in aquifers ensures its extraction.The purpose of this research is to use RF and SVM algorithms in order to prioritize the effective parameters on the potassium grade of saline water groundwater in playa Khoor and Biabank in Isfahan province.For this purpose, 55parameters were measured in12 drilling holes.The parameters measured as independent variables include the percentage of saturated moisture, the apparent specific gravity and the porosity of the core at 15 different depths ,the area polygon, the depth of the underground water, the depth of the salt layer, the potassium of the surface layer, the density of the brine and the amount Elements of calcium, magnesium, sodium, chlorine and grade potassium were included in the model as dependent variables.In the RF model, the of (PFI) and (RFE) were used for prioritization.In the different kernels of the SVM algorithm, in order to prevent the collinearity of the independent parameters,all the combinations of the independent variables, considering the variance inflation factor less than 8 and the highest coefficient of determination and the lowest MSE error,were examined and selected as the best combination.The effective parameters in predicting the grade potassium of the brine in the RF algorithm and the linear function of the SVM algorithm are sp, ap,duw,slp,SAR and n, sp,duw,and SAR respectively,which led to the best results.The coefficient of determination for both models is 0.99 and 0.97,respectively,which indicates the good accuracy of both algorithms.

Keywords: *grade prediction, random forest, saline water, support vector machine*.

مقدمه

با توجه به رشد پیوسته جمعیت جهانی و افزایش نیازهای جدید برای تأمین مواد اصلی در کود، ارزش و اهمیت مواد معدنی، به‌ویژه پتاسیم، روشن‌تر از همیشه مشخص می‌شود. افزایش جمعیت در سطح جهان و کاهش اراضی زراعی قابل‌کشت، نیاز به منابع معدنی، به‌طور خاص پتاسیم را افزایش داده است (Zörb, 2014). در سطح دنیا شورابه‌های زیرزمینی و پتاسیم سنگی به‌عنوان ذخایر اساسی پتاسیم شناخته می‌شوند. ایران، یکی از ۱۰ کشور برتر دنیا در زمینه ذخایر پتاسیم، محسوب می‌شود که دارای ۵۵۰ میلیون تن ذخیره پتاسیم به‌صورت شورابه و ۴ میلیون تن ذخیره پتاسیم سنگی (گنبدهای نمکی) در کشور است (منوچهری، ۱۳۸۲).

پتاسیم یکی از عناصر بسیار ضروری برای امنیت غذایی که ذخایر آن به‌طور عمده در مناطقی خاص تجمع دارد که کمبود آن منجر به کاهش عملکرد تولید و بهره‌وری در بخش کشاورزی می‌گردد. در بسیاری از مناطق خشک و نیمه‌خشک دنیا، پلایا یک پدیده ژئومورفولوژیک مهم شناخته می‌شود. این پدیده به‌ویژه در مناطقی مانند غرب آمریکا، استرالیا و مرکز ایران با تنوع و تعداد بیشتری مشاهده می‌گردد (Shaw & Bryant, 2011). پلایاها و شورابه‌های زیرزمینی آن‌ها حاوی مقادیر بالایی از ذخایر پتاسیم هستند. ازآنجایی‌که روش کشف و استخراج پتاسیم در گنبدهای نمکی با چالش‌هایی نظیر هزینه بالا، پیچیدگی در فرآیند استخراج، نیاز به تجهیزات پیشرفته و همچنین داشتن بازدهی پایین در برخی سازندها همراه است، توصیه می‌شود از روش‌های جدیدی مانند استخراج پتاسیم از شورابه‌های کویری و ذخایر پتاسیم موجود در پلایاها استفاده شود (منوچهری، ۱۳۸۲). یکی از موارد مهم در استخراج شورابه از پلایاها که بهره‌برداری مؤثر و بهینه از این منابع را تضمین می‌کند، عیار پتاسیم در شورابه است. جهت تخمین عیار مواد معدنی، روش‌های متنوعی مانند روش‌های هندسی، مبتنی بر فواصل و زمین‌آماری به کار گرفته می‌شود (حسنی پاک، ۱۳۸۴)؛ که روش‌های زمین‌آماری، بسته به توزیع داده‌ها و تعداد آن‌ها، ممکن است در مواردی که تعداد داده‌ها کم باشد یا وابستگی مکانی میان آن‌ها ضعیف باشد، نتوانند به‌درستی تخمین مواد معدنی را انجام دهند (Dutta et al., 2010)؛ بنابراین جهت تخمین مقادیر مواد معدنی استفاده از روش‌های نوین با دقت بالا می‌تواند بسیار مؤثر و به‌صرفه باشد که یکی از روش‌های نوین به‌منظور تخمین عیار مواد معدنی استفاده از الگوریتم‌های یادگیری ماشین است. استفاده از الگوریتم‌های یادگیری ماشین در علوم زمین دارای کاربردهای فراوانی، ازجمله تشخیص نواحی ناهنجار ژئوشیمیایی تا ارزیابی منابع معدنی است.

ازآنجایی‌که الگوریتم‌های یادگیری ماشین به‌عنوان یک شاخه از هوش مصنوعی شناخته می‌شوند بنابراین برای ایجاد پاسخ‌های قابل‌قبول با درصد اطمینان بالا استفاده می‌شوند. هدف اصلی آن‌ها بهبود عملکرد سیستم با استفاده از تجربیات گذشته و داده‌های آموزشی است. این الگوریتم‌ها بر اساس نوع آموزش به چهار دسته شامل یادگیری با نظارت[1]، بدون نظارت[2]، نیمه نظارتی[3] و تقویتی[4] تقسیم می‌شوند. الگوریتم‌های جنگل‌های تصادفی[5] و ماشین بردار پشتیبان[6] نمونه‌هایی از یادگیری با نظارت هستند در یادگیری با نظارت، دو نوع متغیر وجود دارد: متغیرهای مستقل و متغیرهای وابسته. این الگوریتم‌ها متغیرهای مستقل را به‌عنوان ورودی در نظر گرفته تا متغیرهای وابسته را مدل‌سازی و پیش‌بینی نمایند؛ بنابراین برای پیش‌بینی متغیرهای وابسته، یک تابع تخمین‌گر[7] ایجاد می‌شود که از متغیرهای مستقل به‌عنوان ورودی استفاده کرده و مقادیر متغیرهای وابسته را مدل‌سازی می‌کند. این فرآیند که به‌عنوان فرآیند

---

[1] Supervised
[2] Unsupervised
[3] Semi supervised
[4] Reinforcement
[5] Random forest
[6] Support vector machine
[7] Estimator

تعلیم یا آموزش (واسنجی) شناخته می‌شود، روی داده‌های موجود (داده‌هایی که هر دو متغیرهای مستقل وابسته آن‌ها مشخص هستند) اعمال می‌شود و با تنظیم فرا پارامترهای[1] مدل تا رسیدن به‌دقت لازم ادامه می‌یابد (Ray, 2019).

مناطق خشک و نیمه‌خشک به دلیل شرایط خاص محیطی حاکم بر آن‌ها ازجمله تبخیر که در گسترش دریاچه‌های نمکی نقش اساسی ایفا می‌کند، به‌عنوان مهم‌ترین مناطق برای یافتن ذخایر پتاسیم محسوب می‌شوند. با توجه به ویژگی‌های اقلیمی و توپوگرافی کشور ایران، مناطق مشابه به‌صورت گسترده در مرکز ایران یافت می‌شوند؛ که منطقه خور و بیابانک در استان اصفهان یکی از این مناطق است. با توجه به اهمیت پتاسیم در بخش کشاورزی، امنیت غذایی کشور و سایر بخش‌های صنعت، هدف این پژوهش ارزیابی پارامترهای مؤثر جهت برآورد و تخمین عیار پتاسیم شورابه در پلایای شهرستان خور و بیابانک با استفاده از الگوریتم‌های یادگیری ماشین (جنگل تصادفی و ماشین بردار پشتیبان) به‌عنوان روش های نوین، کم هزینه و مقرون‌به‌صرفه و تعیین پارامترهای مؤثر (متغیرهای مستقل مورداستفاده) با بیشترین تأثیرگذاری در جهت سنجش عیار پتاسیم به‌منظور بهبود بهره‌برداری از ذخایر پتاسیم و کاهش هزینه‌های اجرایی، عملیاتی و آزمایشگاهی است.

پیشینه پژوهش

استفاده از الگوریتم‌های یادگیری ماشین در جهت تخمین عیار منابع معدنی در پژوهش‌هایی به کار گرفته‌شده است. این الگوریتم‌ها در تخمین عیار کانسارهای اسلیت (Matías et al., 2004)، تهیه نقشه مواد رادیواکتیویته طبیعی (Pozdnoukhov, 2005)، تخمین غلظت آرسنیک در سنگ‌بستر مشتق شده از رسوبات رودخانه‌ای (Twarakavi et al., 2006)، تحلیل داده‌های مکانی (Dutta, 2006)، طلا (Dutta et al., 2010)، طبقه‌بندی تصاویر ماهواره‌ای (Moorthi et al., 2011؛ Soliman&Mahmoud, 2012)، طبقه‌بندی پوشش زمین (Rodriguez-Galiano et al., 2012)، آهن (Maleki et al., 2014)، پلاتین نوع یخچالی (Tenorio et al., 2015)، مدل‌سازی آب‌های زیرزمینی (Naghibi et al., 2016؛ Baudron et al., 2013)، آینده‌نگری مواد معدنی (Harris, 2016؛ McKay & Rodriguez-Galiano et al., 2015)، طلا (Zhang et al., 2019)، مس (Wang et al., 2020). اشاره کرد. Sheng et al. (2015) ترکیبی از طیف‌سنجی شکست ناشی از لیزر[2] و الگوریتم جنگل تصادفی را برای طبقه‌بندی نمونه‌های سنگ‌آهن استفاده کردند که نتایج نشان داد که مدل از دقت بسیار بالاتری (۱۰۰ درصد) در جهت طبقه‌بندی نمونه‌های سنگ‌آهن برخوردار است. Jafrasteh et al. (2018) پتانسیل یک مدل مبتنی بر جنگل تصادفی را برای ارزیابی عیار ذخایر مس پورفیری و عملکرد مدل با سایر الگوریتم‌های یادگیری ماشینی (مانند شبکه عصبی چندلایه‌ای) و تکنیک زمین‌آماری (مانند کریجینگ نشانگر و کرجینگ معمولی) به کاربردند. نتایج نشان داد مدل مبتنی بر جنگل تصادفی با پارامترهای مورداستفاده شده با ۵۰۰ درخت تصمیم برای اطمینان از همگرایی مدل، از کارایی بهتری برخوردار است. نتایج پژوهش Schnitzler et al. (2019) نشان داد که استفاده از الگوریتم جنگل تصادفی جهت ارزیابی غلظت سدیم در منطقه معدنی ماتاگامی کانادا می‌تواند یک ابزار کارآمد برای تخمین عناصر ژئوشیمیایی اندازه‌گیری نشده در یک پایگاه داده اکتشافی باشد.

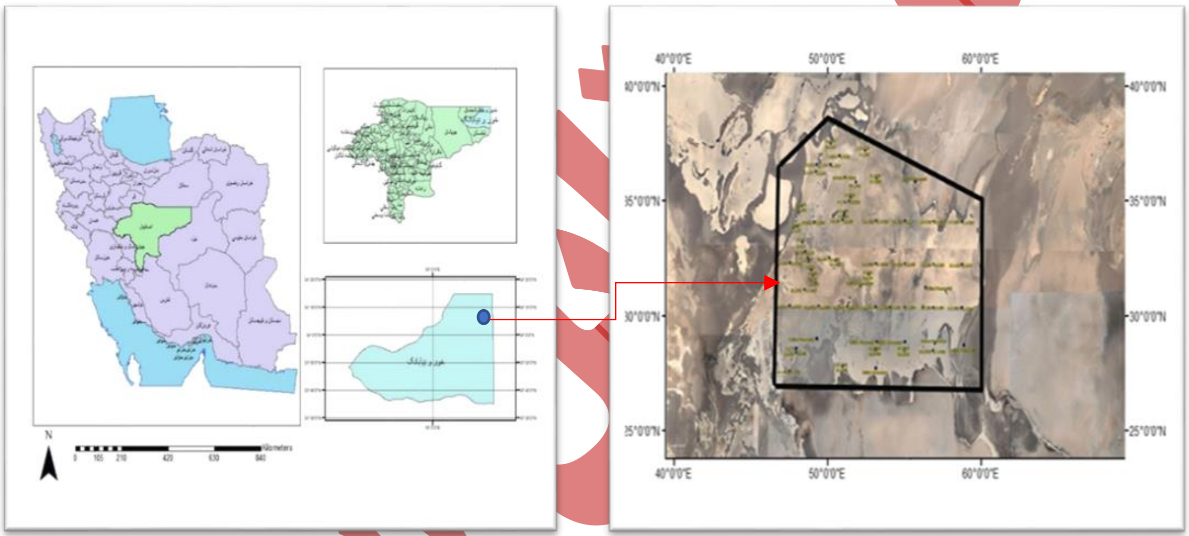
مواد و روش‌ها

منطقه مورد مطالعه

[1] Hyperparameter tuning
[2] LIBS

غیر قابل استناد

خور و بیابانک، منطقه‌ای خشک و بیابانی از توابع استان اصفهان است که در جنوب‌غربی کویر بزرگ نمک قرار دارد. این منطقه با وسعت ۳۷۹۴/۲۷ کیلومترمربع، با ارتفاع متوسط ۷۰۵ متر از سطح دریا، در مختصات عرض جغرافیایی ۳۳ درجه و ۳۷ دقیقه تا ۳۴ درجه و ۱۱ دقیقه شمالی و طول جغرافیایی ۵۴ درجه و ۱۶ دقیقه تا ۵۵ درجه و ۱۴ دقیقه شرقی گسترده شده است. سازندهای زمین‌شناسی منطقه گچی، نمکی، آهکی و با افق سطحی کلسیک است. کوه‌های آن جزء کوه‌های ایران مرکزی است که به چین‌خوردگی‌های آلپی تعلق دارد. خور و بیابانک در کمربند بیابانی نیمکره شمالی قرار دارد و اقلیم آن طبق روش طبقه‌بندی دومارتن فراخشک گرم است (موسوی و همکاران، ۱۳۹۵). طبق داده‌های اداره هواشناسی خور و بیابانک حداقل و حداکثر دمای مطلق منطقه به ترتیب ۱۴− و ۴۶/۶ درجه سانتی‌گراد، میانگین بلندمدت بارش آن ۸۳/۹ میلی‌متر و تعداد ساعات آفتابی ۳۳۰۰ ساعت در سال است. شکل ۱ موقعیت جغرافیایی منطقه موردمطالعه را با وسعت ۴۹/۰۵ کیلومترمربع که در قسمت شرقی پلایای خور و بیابانک واقع‌شده است؛ نشان می‌دهد. پلایای موجود در این منطقه یکی از مهم‌ترین و شاخص‌ترین پلایاهای ایران از منظر وجود معادن پتاسیم است.



شکل ۱. موقعیت جغرافیایی منطقه نمونه‌برداری شده منطقه موردمطالعه در استان اصفهان، شهرستان خور و بیابانک

پارامترهای اندازه‌گیری شده

این پژوهش از نوع کاربردی بوده و روش‌شناسی آن تلفیقی از نمونه‌برداری میدانی، تجزیه‌وتحلیل‌های آزمایشگاهی و مدل‌سازی آماری است. در محدوده موردنظر ۶ پلیگون بزرگ و ۶ پلیگون کوچک انتخاب و مغزه‌گیری (گمانه حفاری) در این پلیگون‌ها صورت گرفت. تعداد ۵۵ پارامتر در هر ۱۲ گمانه حفاری اندازه‌گیری شدند. شکل ۲ نمونه‌ای از پلیگون مغزه‌گیری شده و مغزه‌های برداشت‌شده را نشان می‌دهد. در پلایا به دلیل شرایط سخت محیطی و هزینه‌های بالای حفاری محدودیت‌های زیادی برای نمونه‌برداری وجود دارد؛ بنابراین به همین میزان نمونه بسنده شد. در هر گمانه حفاری سه پارامتر تخلخل، (n1-n15، درصد)، جرم مخصوص ظاهری، (pb1-pb15، گرم بر سانتی متر مکعب) و درصد رطوبت اشباع، (sp1-sp15، درصد) تا عمق ۱۰ متری به ترتیب به فواصل، از عمق صفر تا ۱/۵ متری به فواصل ۳۰ سانتی‌متری، از عمق ۱/۵ تا ۴ متری به فواصل ۵۰ سانتی‌متری و از عمق ۴ تا ۱۰ متری به فواصل ۱۰۰ سانتی‌متری جداسازی و اندازه‌گیری سه پارامتر مذکور در ۱۵ عمق مختلف انجام شد. تخلخل به روش اشباع و غوطه‌وری داخل اتیلن گلیکول، جرم مخصوص ظاهری با غوطه‌وری داخل پارافین مذاب و درصد رطوبت اشباع با روش رطوبت وزنی، اندازه‌گیری شد.

همچنین در هر ۱۲ گمانه حفاری، پارامترهای سطحی، مساحت پلی گون، (ap، متر مربع) با استفاده از متر، پتاسیم لایه سطحی، (slp، درصد) با استفاده از دستگاه فلیم فتومتر و پارامترهای زیرسطحی عمق آب زیرزمینی، (duw، سانتی متر)، بعد از عملیات مغزه‌گیری عمق آن اندازه‌گیری شد. عمق لایه نمک، (dsl، متر) نیز با متر و در شورابه زیرزمینی، دانسیته شورابه، (d، گرم بر سانتی متر مکعب) با استفاده از پیکنومتر که دقیق‌ترین وسیله برای اندازه‌گیری دانسیته مایعات است و میزان سدیم شورابه، (na، گرم بر لیتر) و پتاسیم شورابه، (Potassium grade، گرم بر لیتر) با استفاده از دستگاه فلیم فتومتر، کلسیم شورابه، (ca، گرم بر لیتر) و منیزیم شورابه، (mg، گرم بر لیتر) به روش تیتراسیون با EDTA و کلر شورابه، (cl، گرم بر لیتر) به روش تیتراسیون با نیترات نقره اندازه گرفته شد (Estefan, 2013).

شکل ۲. پلی‌گون‌های انتخابی و مغزه‌های برداشت‌شده به‌منظور اندازه‌گیری پارامترهای موردنیاز

الگوریتم‌های یادگیری ماشین

الگوریتم‌های یادگیری ماشین زمانی مورداستفاده قرار می‌گیرند که با مسائلی پیچیده مواجه هستیم، به این معنی که روش‌های سنتی پیش‌بینی کننده نمی‌توانند مسئله را به‌خوبی حل کنند. این الگوریتم‌ها به دلیل دقت و توانایی بالاتر در مدل‌سازی نسبت به روش‌های سنتی برتری دارند. از دو الگوریتم معروف به نام جنگل تصادفی (RF) و ماشین بردار پشتیبان (SVM) بهره گرفته می‌شود. در ادامه، توضیحات مختصری در مورد هر یک از این الگوریتم‌ها ارائه‌شده است.

الگوریتم جنگل تصادفی

جنگل تصادفی (RF) یک الگوریتم یادگیری ماشین ساده و بسیار کارآمد است که به‌طورمعمول حتی بدون نیاز به تنظیم پیچیده پارامترها نتایج عالی را به ما می‌دهد. این الگوریتم هم برای مسائل دسته‌بندی و هم رگرسیون مناسب است. RF با ایجاد یک گروه از درخت‌های تصمیم به‌صورت تصادفی عمل می‌کند (امینی‌خویی و عبداله‌پوری، ۱۳۹۵). در این روش، از تعداد زیادی درخت تصمیمی برای

رسیدن به نتیجه نهایی استفاده می‌شود. هر درخت به‌صورت تصادفی با استفاده از یک نمونه تصادفی از داده‌ها انتخاب می‌شود. سپس یک زیرمجموعه تصادفی از ویژگی‌ها و ورودی‌ها برای هر درخت ایجاد می‌شود. این کار باعث تقسیم داده‌ها به گروه‌های جدید می‌شود. درواقع، هر درخت به‌صورت جداگانه یک پیش‌بینی انجام می‌دهد؛ به‌عبارت‌دیگر، هر درخت به تفکیک نمونه‌ها به گروه‌های مختلف کمک می‌کند. درنهایت، با تحلیل بیشترین تعداد رای داده‌شده توسط این درخت‌ها، برچسب مناسب برای هر نمونه مشخص می‌شود. این الگوریتم بسیار سازگار با تغییرات و نویز در داده‌ها بوده و بهبود دقت و عملکرد مدل را به ارمغان می‌آورد (Nitze et al., 2012).

از این روش معمولاً برای جلوگیری از مشکلات بیش برازش در مدل‌های پیش‌بینی استفاده می‌شود. همچنین، RF سریع‌ترین الگوریتم‌ها در دسته‌بندی داده‌ها است و با بهره‌گیری از تعداد زیادی درخت، از اطلاعات بیشتری در داده‌ها بهره می‌برد تا پیش‌بینی دقیق‌تری انجام دهد. این الگوریتم به‌ویژه زمانی کارآمد است که تعداد نمونه‌های داده نسبت به تعداد ویژگی‌ها کم باشد (Mohri et al., 2018). از جنگل تصادفی برای بسیاری از مجموعه‌های داده استفاده می‌شود و این الگوریتم دسته‌بندی را با سرعت بالا و با دقت بالا انجام می‌دهد. برخلاف روش‌های سنتی مانند رگرسیون که بر پایه‌یک مدل تکیه‌دارند، RF از صدها یا حتی هزاران درخت برای مدل‌سازی استفاده می‌کند. این تعداد زیاد از درخت‌های تصمیمی به این الگوریتم امکان می‌دهد تا از اطلاعات بیشتری در داده‌ها بهره‌برداری کند تا بتوان به استنباط‌های دقیق‌تری از ویژگی‌ها و روابط داده‌ها دست یابد (Chen et al., 2020).

الگوریتم ماشین بردار پشتیبان

ماشین بردار پشتیبان (SVM) یک روش یادگیری ماشین قدرتمند است که برای مسائل طبقه‌بندی و رگرسیون به کار می‌رود. این الگوریتم، به علت کارایی بالا و نتایج برتری که در انواع مسائل ارائه می‌دهد، توجه زیادی از پژوهشگران جلب کرده است. در SVM، یک ابرصفحه[1] جداسازی به‌گونه‌ای ایجاد می‌شود که داده‌های آموزشی به دو کلاس خاص جداگانه و معین تقسیم می‌شوند. این ابرصفحه به‌گونه‌ای انتخاب می‌شود که فاصله بین نزدیک‌ترین نقاط دو کلاس به حداکثر برسد. این نقاط که به آن‌ها بردارهای پشتیبان[2] گفته می‌شود، موقعیت ابرصفحه را تعیین می‌کنند و در فرآیند طبقه‌بندی از آن‌ها استفاده می‌شود. این الگوریتم توانایی دقت بالا در طبقه‌بندی داده‌های جدید را با تعداد کمی بردار پشتیبان نشان می‌دهد که این مزیت نسبت به روش‌های دیگر مهم است. در مسائل رگرسیون نیز، SVM قادر است بهترین خط را برای داده‌های ورودی و خروجی پیداکرده و با استفاده از آن، مقدار خروجی برای داده‌های جدید را پیش‌بینی کند (Yang et al., 2008). یکی از ویژگی‌های برجسته SVM، توانایی تفکیک داده‌های غیرخطی با استفاده از توابع کرنل است. توابع کرنل متعددی وجود دارند که بر داده‌های ورودی اعمال‌شده و ابعاد بیشتری به آن‌ها تخصیص می‌دهند. این ابعاد اضافی به ورودی‌ها افزوده می‌شوند و در فضای ابعاد بالاتر، ابرصفحه جداکننده‌ای برای داده‌ها ایجاد می‌شود. اغلب در الگوریتم ماشین بردار پشتیبان از توابع کرنل خطی، چندجمله‌ای، تابع پایه شعاعی و تابع سیگموئید بیشترین استفاده را می‌شود. این توابع[3] به کمک SVM، توانایی تبدیل داده‌های ابتدایی به فضاهای با بعد بالاتر و تفکیک داده‌ها در این فضاها رادارند. ازاین‌جهت، آن‌ها به مدل SVM امکان تفکیک داده‌های غیرخطی را می‌دهند. روابط این توابع در جدول ۱ آورده شده است.

---
[1] Hyperplane
[2] Support Vectors
[3] Kernel functions

جدول ۱. توابع متداول الگوریتم SVM

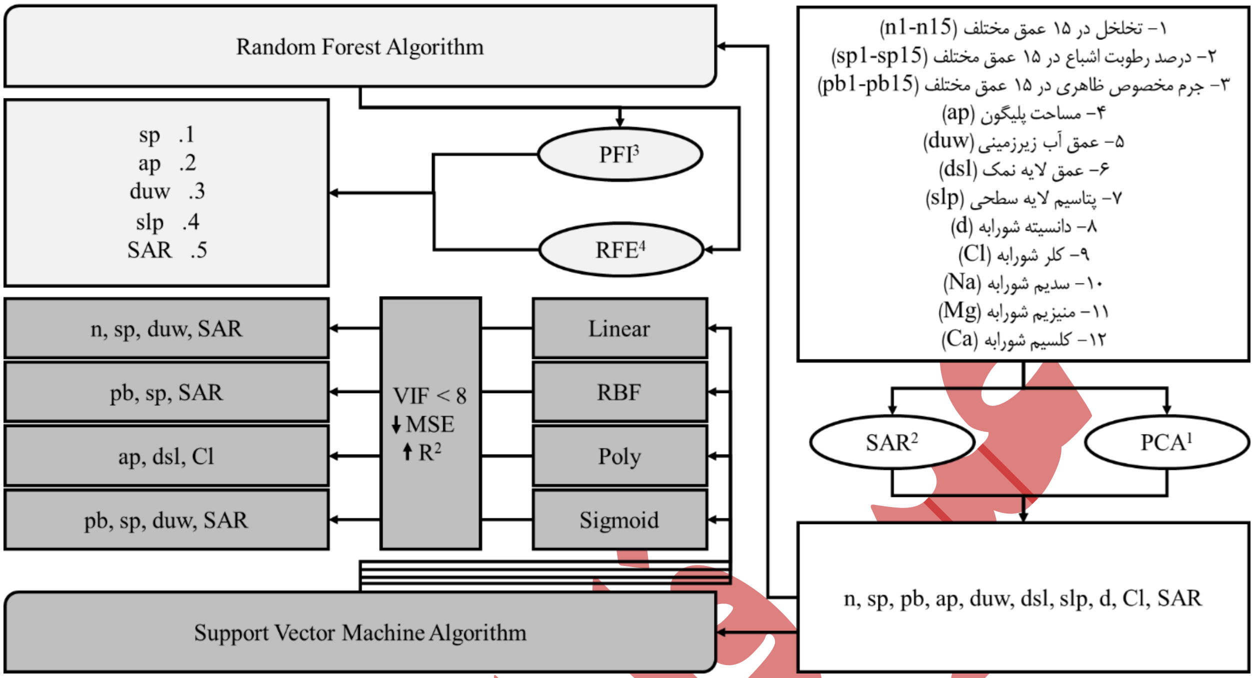
نوع تابع	فرمول
سیگموئید	$\tanh(\gamma x_i x_j + \text{cofficient})$
پایه شعاعی	$\exp(-\gamma \|x_i - x_j\|^2)$
چندجمله‌ای	$[\gamma x_i x_j + \text{cofficient}]^d$
خطی	$x_i \times x_j$

مراحل مدل‌سازی

جهت کاهش تعداد مجموعه داده‌های ورودی به مدل برای سه پارامتر n ,sp و pb که در ۱۵ عمق مختلف اندازه گیری شده بودند از تکنیک کاهش ابعاد داده استفاده شد. یکی از کارآمدترین و دقیق‌ترین روش‌ها برای کاهش ابعاد داده‌ها تجزیه‌وتحلیل مؤلفه اصلی (PCA[1]) است. روش مذکور ویژگی‌ها را به تعدادی از خصوصیت‌های موردنظر در ابعاد کوچک‌تر کاهش می‌دهد و تا آنجا که ممکن است تغییرات موجود در داده‌ها را حفظ می‌کند و ویژگی‌های مهم را که با یکدیگر همبستگی ندارند؛ استخراج می‌کند (García-Gil., 2018). در سه پارامتر ذکرشده از روش PCA جهت تمرکز بر ویژگی‌هایی که اهمیت بیشتری در پیش‌بینی عیار پتاسیم دارند و این ویژگی‌ها کم‌ترین همبستگی را با یکدیگر دارند استفاده شد و هر پارامتر به یک عمق معادل تبدیل شد. همچنین سه متغیر Na، Mg و Ca با شاخص نسبت جذبی سدیم (SAR، میلی اکی والان بر لیتر) با استفاده از رابطه ۱ در مدل سازی استفاده شدند. در نهایت ۱۰ پارامتر n، sp، pb، duw، dsl، slp، d، Cl و SAR برای مدل‌سازی نهایی انتخاب شدند. شکل ۳ مراحل آماده‌سازی داده‌ها و روند مدل‌سازی را به صورت شماتیکی و خلاصه نشان می‌دهد. در مرحله بعدی و بر اساس ملاحظات مربوط به هر الگوریتم به مدل‌سازی پرداخته شد که نتایج و توضیحات بیشتر در قسمت نتایج و بحث ارائه می‌شود.

$$SAR = \frac{Na^+}{\sqrt{\frac{1}{2}(Ca^{2+} + Mg^{2+})}}$$ رابطه ۱)

[1] Principial Component Analysis



1. Principal Component Analysis, 2. Sodium Absorption Ratio, 3. Permutation Feature Importance, 4. Recursive Feature Elimination

شکل ۳. خلاصه، مراحل و توابع استفاده شده در مدل‌سازی عیار پتاسیم در مطالعه حاضر

کلیه فرآیند مدل‌سازی با استفاده از زبان برنامه‌نویسی Python انجام‌شده است و از کتابخانه‌های Sklearn، Numpy، Pandas و Matplotlib استفاده شد. در هر دو الگوریتم SVM و RF، ۷۰ درصد از داده‌ها به‌عنوان داده‌های آموزشی و به‌منظور ایجاد مدل و ۳۰ درصد دیگر داده‌ها، به‌طور کامل مستقل از داده‌های آموزشی به منظور داده‌های آزمایش و ارزیابی اعتبار مدل‌های تولیدشده به کار رفت (Kisi et al., 2006). محققان نشان داده‌اند که این ترکیب بهترین عملکرد به همراه داشته است (Ghorbanzadeh et al., 2018؛ Tiwari et al., 2018). با توجه به اینکه رنج داده‌ها و واحد اندازه‌گیری آن‌ها متفاوت می‌باشند، در الگوریتم‌های یادگیری ماشین عملکرد تابع با محدودیت سخت‌افزاری و نرم‌افزاری همراه است و نمی‌تواند عملکرد مثبتی داشته باشد، بنابراین نیاز است تا مقادیر پردازش‌شده نرمالایز[1] شوند. نرمال‌سازی داده‌ها روش مؤثر برای افزایش دقت مدل‌های یادگیری ماشین است (Ioffe & Szegedy, 2015). در این الگوریتم‌ها از روش MinMaxScaler و بر اساس رابطه ۲ استفاده شد.

$$\hat{x}=\frac{x-\min(x)}{\max(x)-\min(x)}$$ رابطه (۲)

ارزیابی مدل‌ها

برای ارزیابی عملکرد مدل‌های RF و SVM در پیش‌بینی عیار پتاسیم شورابه پلایاها از آماره‌های مجذور میانگین مربعات خطا (RMSE)، قدر مطلق خطای نسبی (MAE) خطای میانگین مربعات (MSE) و ضریب تعیین (R^2) بر اساس روابط ۳ تا ۶ استفاده شد (Moriasi et al., 2007).

[1] Normalize

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$ رابطه ۳)

$$MAE = \sqrt{\frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}}$$ رابطه ۴)

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$ رابطه ۵)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$ رابطه ۶)

که در رابطه‌های بالا y_i و $\hat{y}$ به ترتیب مقادیر اندازه‌گیری واقعی و پیش‌بینی‌شده است و n تعداد نمونه‌ها و $\bar{y}$ میانگین مقادیر اندازه‌گیری شده واقعی است؛ درمجموع، مدلی که بیش‌ترین مقدار ضریب تبیین R^2 و کمترین مقادیر خطاها RMSE، MAE و MSE را داشته باشد، از عملکرد بهتری در پیش‌بینی و برآورد عیار پتاسیم شورابه پلایا برخوردار است.

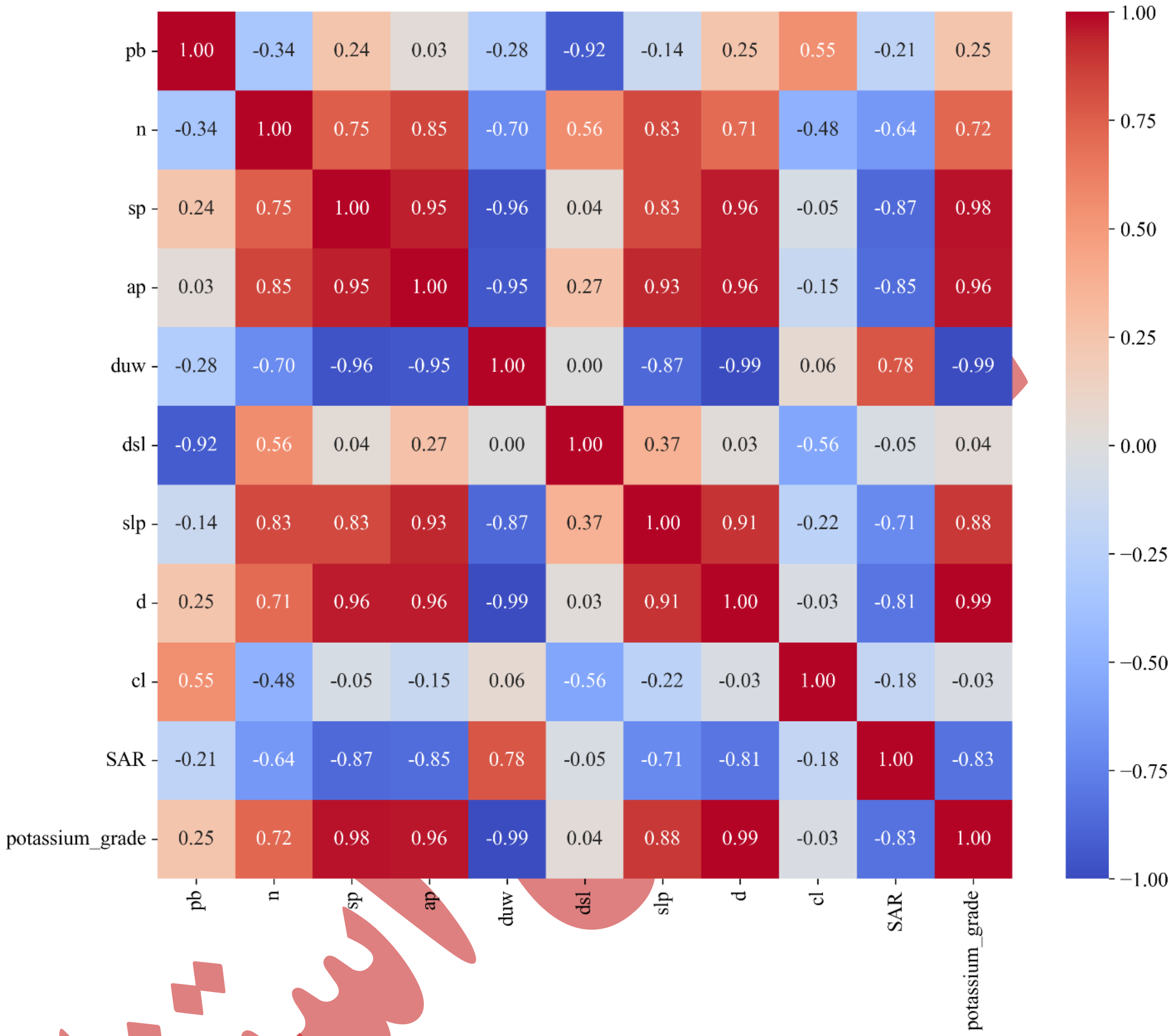
نتایج و بحث

الگوریتم ماشین بردار پشتیبان

بر اساس ضریب همبستگی پیرسون[1] بالای موجود بین متغیرهای مستقل موجود که نتایج آن در شکل ۴ ارائه شده است و همچنین به منظور بررسی چندخطی بین متغیرهای مستقل که در استفاده از مدل‌های رگرسیونی (همانند SVM) مشکل‌ساز بوده و باعث برازش بیش از حد می‌شود؛ از عامل تورم واریانس (VIF[2]) استفاده شد. در الگوریتم رگرسیونی SVM و مخصوصاً تابع خطی آن که به هم‌خطی و چندخطی متغیرهای مستقل بسیار حساس است؛ از VIF به عنوان یک عامل جهت جلوگیری از این مشکل استفاده شد. به این صورت که با نوشتن یک الگوریتم جدید (مختص مطالعه) در هر تابع (چهار کرنل موجود در SVM) با توجه به تعداد متغیرهای مستقل (۱۰ پارامتر اندازه‌گیری شده n، sp، pb، duw، dsl، slp، d، Cl و SAR) در کل می‌توان 2^{10} (برابر ۱۰۲۴) ترکیب مختلف از ویژگی‌های اندازه‌گیری شده داشت؛ بررسی شدند و آن مجموعه از ویژگی‌ها که دارای VIF کوچک‌تر از ۸، کمترین خطای میانگین مربعات (MSE) و بالاترین ضریب تعیین (R^2) بودند؛ برای مدل‌سازی انتخاب شدند (شکل ۳). بنابراین هیچ‌گونه هم‌خطی و چندخطی بین متغیرهای مستقل انتخاب شده در هر کرنل برای مدل‌سازی وجود نداشت.

---

[1] Pearson Correlation Coefficient
[2] Variance Factor Inflation



شکل ۴. همبستگی موجود بین متغیرهای مستقل و وابسته بر اساس ضریب همبستگی پیرسون

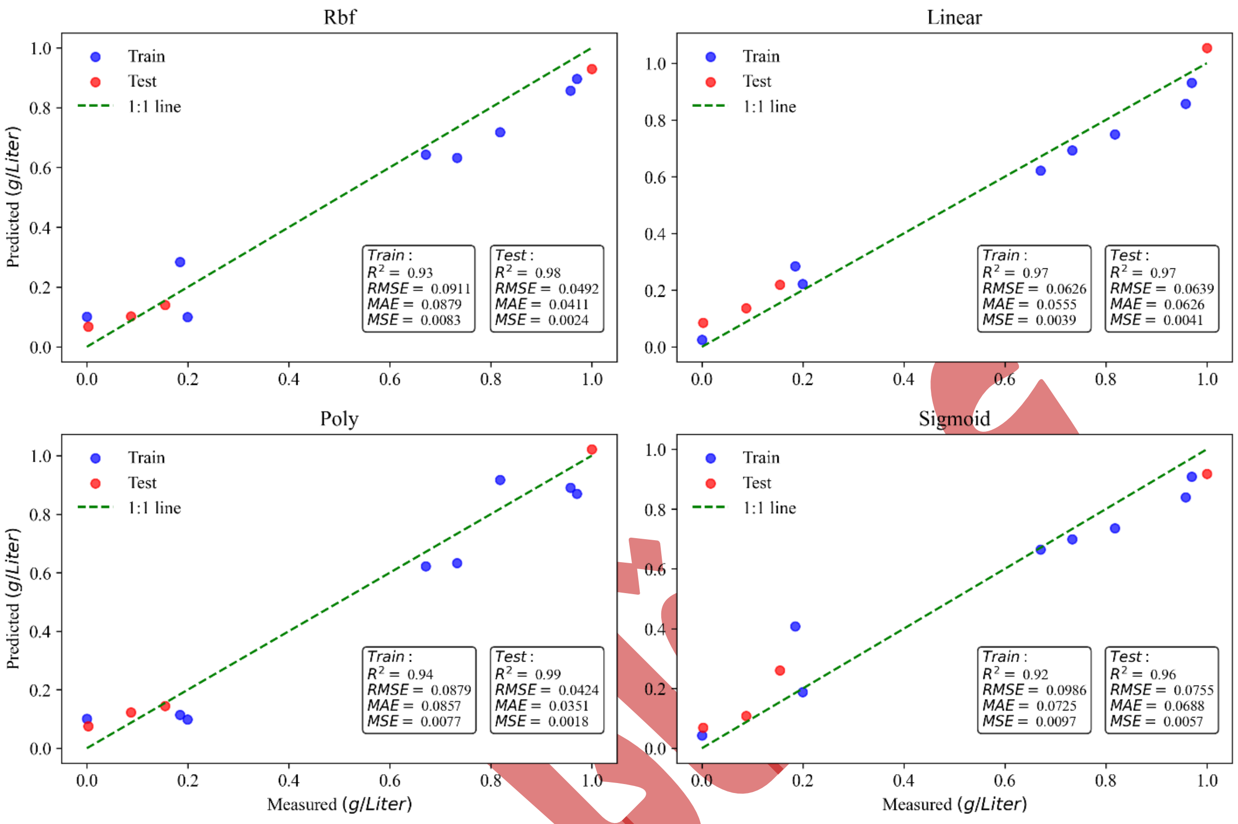
جدول ۲ ویژگی‌های که سه شرط مذکور (VIF کوچک‌تر از ۸، بیشترین R^2 و کمترین MSE) را در هر کرنل SVM داشته باشند، را نشان می‌دهد. بر اساس نتایج ارائه شده در جدول ۲، به عنوان مثال برای کرنل Linear پارامترهای n، sp، duw و SAR منجر به بهترین دقت برای مدل‌سازی انتخاب شدند. نتایج دیگر کرنل‌ها در جدول ۲ ارائه شده است.

جدول ۲. نتایج انتخاب بهترین مجموعه ویژگی‌ها برای کرنل‌های مختلف الگوریتم SVM

کرنل		VIF	R^2	R^2
Linear	n	۴/۶۲	۰/۹۷	۰/۹۷
	sp	۴/۹۷		
	duw	۶/۷۷		
	SAR	۶/۰۴		
RBF	pb	۲/۳۹	۰/۹۳	۰/۹۸
	sp	۲/۲۴		
	SAR	۱/۱۴		
Poly	ap	۵/۲۳	۰/۹۴	۰/۹۹
	dsl	۲/۵۸		
	Cl	۳/۶۹		
Sigmoid	pb	۲/۴۱	۰/۹۲	۰/۹۶
	sp	۲/۵۸		
	duw	۶/۸۱		
	SAR	۶/۲۵		

در شکل ۵ نمودار مقادیر واقعی اندازه‌گیری شده (Measured) و پیش‌بینی‌شده (Predicted) عیار پتاسیم شورابه در مدل SVM برای داده‌های واسنجی (Train) و صحت‌سنجی (Test) در چهار تابع خطی (Linear)، چندجمله‌ای (Poly)، تابع پایه شعاعی (RBF) و سیگموئیدی (Sigmoid) ترسیم‌شده است. نتایج نشان می‌دهد در میان توابع ذکرشده، در دو مرحله واسنجی و صحت‌سنجی تابع خطی با در نظر گرفتن پارامترهای تخلخل (n)، درصد رطوبت اشباع (sp)، عمق آب زیرزمینی (duw) و نسبت جذب سدیم (SAR) عیار پتاسیم شورابه را با دقت خوبی پیش‌بینی کرده است؛ و مقادیر شاخص‌های آماری R^2، MSE، RMSE و MAE در تابع خطی برای سری داده‌های واسنجی به ترتیب ۰/۹۷، ۰/۰۰۳۹، ۰/۰۶۲۶ و ۰/۰۵۵۵ به دست آمد. همچنین شاخص‌های آماری R^2، MSE، RMSE و MAE برای داده‌های صحت‌سنجی مدل به ترتیب برابر با ۰/۹۷، ۰/۰۰۴۱، ۰/۰۶۳۹ و ۰/۰۶۲۶ گرم بر لیتر است. بعد از تابع خطی به ترتیب کرنل‌های چندجمله‌ای، پایه شعاعی و سیگموئید دارای بالاترین دقت بوده‌اند. نتایج نشان‌دهنده نتایج خوب و قابل‌اعتماد مدل SVM است. نتایج دیگر توابع مدل SVM در شکل ۵ ارائه‌شده است.

غیر قابل استناد



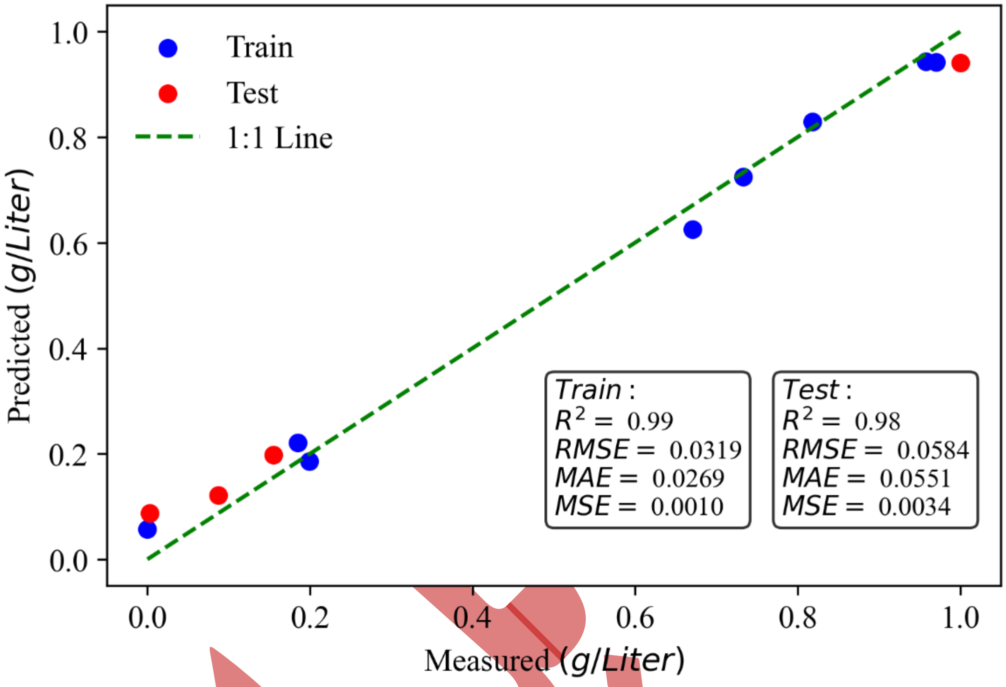
شکل ۵. نتایج واسنجی و صحت‌سنجی کرنل‌های مختلف الگوریتم SVM در برآورد و پیش‌بینی میزان عیار پتاسیم پلایاهای موردمطالعه

الگوریتم جنگل تصادفی

در الگوریتم RF که یک روش یادگیری گروهی مبتنی بر درخت‌های تصمیم است و هر درخت با استفاده از زیرمجموعه‌ای تصادفی از ویژگی‌ها ساخته می‌شود و پیش‌بینی نهایی، مجموعه‌ای از پیش‌بینی‌های انجام‌شده توسط درختان است که بر مفروضات مشابه رگرسیون خطی تکیه نمی‌کند و به همان دلیل مشکل چندخطی در آن وجود ندارد و VIF در یک مدل جنگل تصادفی معنادار نیست و VIF در این الگوریتم محاسبه نشد. با توجه به مراحل نشان داده شده در شکل ۳، در الگوریتم RF ابتدا با همه ویژگی‌های انتخاب شده (n، sp، pb، duw، dsl، slp، d، Cl و SAR) به مدل‌سازی پرداخته شده که نتایج آن در شکل ۶ ارائه شده است. سپس با استفاده از روش‌های Permutation Feature Importance (PFI) و Recursive Feature Elimination (RFE) ویژگی‌هایی که بیشترین تأثیر در میزان عیار پتاسیم داشتند؛ انتخاب شده و دوباره به مدل‌سازی با استفاده از پارامترهای انتخابی این روش‌ها پرداخته شد (نتایج این مدل‌سازی در شکل ۸ ارائه شده است).

در شکل ۶ نمودار مقادیر واقعی اندازه‌گیری شده و مقادیر پیش‌بینی‌شده عیار پتاسیم شورابه با مدل RF برای داده‌های واسنجی و صحت‌سنجی ترسیم‌شده است. نتایج نشان داد، مدل داده‌های واسنجی با دارا بودن R^2 معادل ۰/۹۹ و MSE، RMSE و MAE به ترتیب معادل ۰/۰۰۱۰، ۰/۰۳۱۹ و ۰/۰۲۶۹ و همچنین شاخص‌های آماری R^2، MSE، RMSE و MAE برای داده‌های صحت‌سنجی مدل به ترتیب ۰/۹۸، ۰/۰۰۳۴، ۰/۰۵۸۴ و ۰/۰۵۵۱ گرم بر لیتر است که از دقت بالایی برخوردار است. همان‌طور که در شکل ۶ نیز به‌خوبی دیده می‌شود، نتایج مدل RF به خط نیم‌ساز نزدیک هستند و می‌توان گفت مقادیر را با دقت بالایی پیش‌بینی کرده است. از

لحاظ تئوری اگر R^2 برابر ۱ درصد شود تمامی مقادیر مشاهده‌شده با مقادیر برازش شده یکسان خواهند بود و همه نقاط داده‌ها بر روی خط برازش شده قرار دارند (Devore, 2015).

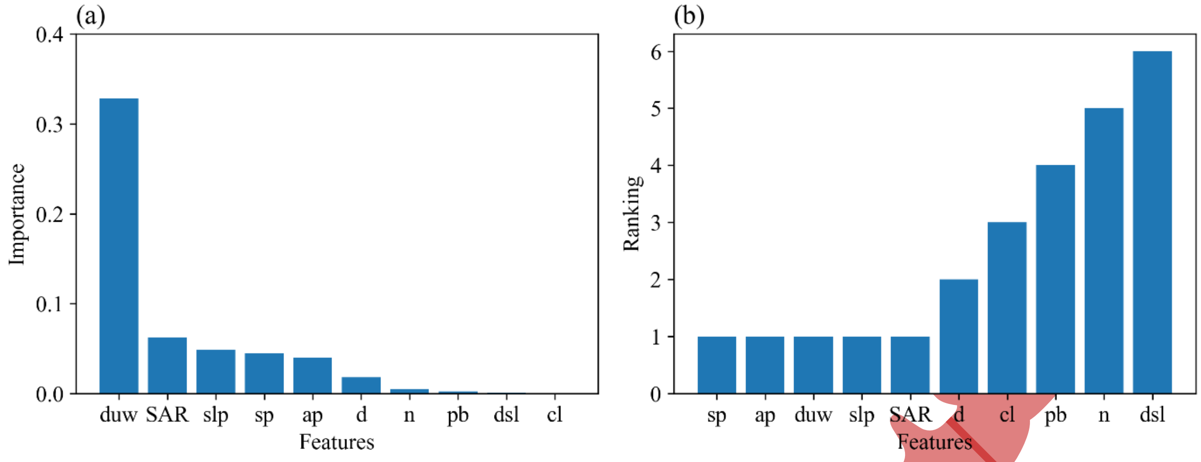


شکل ۶. نتایج واسنجی و صحت‌سنجی الگوریتم RF با استفاده از تمام پارامترهای اندازه‌گیری شده در پیش‌بینی میزان عیار پتاسیم

اولویت‌بندی پارامترهای مؤثر بر عیار پتاسیم شورابه در روش RF

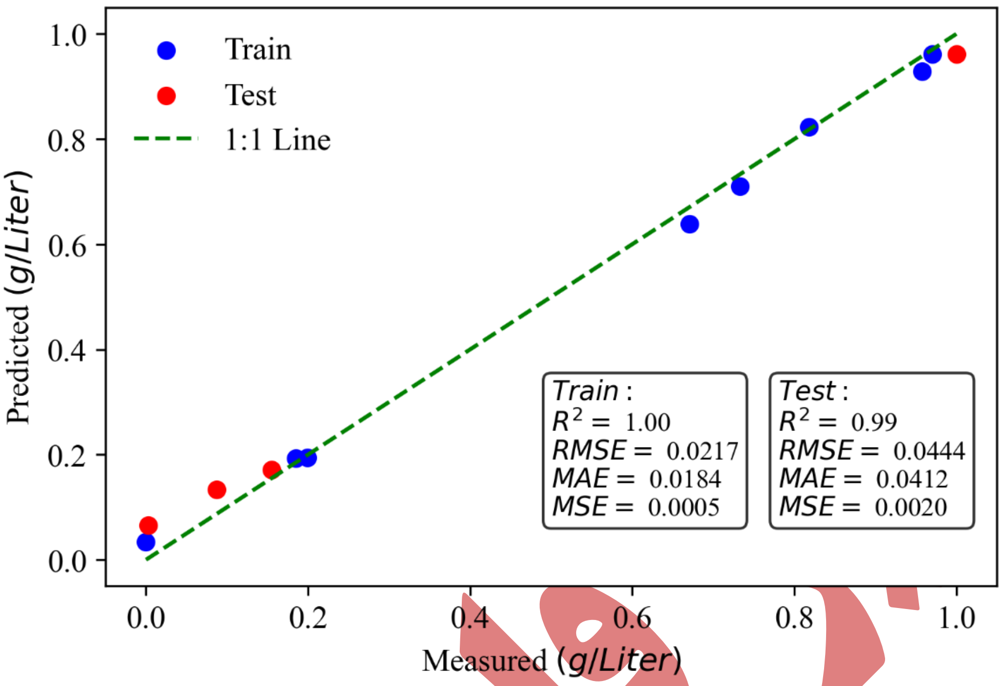
مطابق نتایج به‌دست‌آمده و ارائه‌شده در شکل ۶ برای انتخاب و اولویت‌بندی مهم‌ترین پارامترهایی که بیشترین اثر بر عیار پتاسیم شورابه را داشته‌اند از روش‌های اهمیت ویژگی جایگشت[1] (Kaneko, 2022) و حذف ویژگی بازگشتی[2] (Jeon & Oh, 2020) استفاده شد؛ که نتایج آن در شکل ۷ آورده شده است. سپس برای مدل‌سازی مجدد از ۵ پارامتر مشخص‌شده در شکل ۷ که دارای رتبه ۱ هستند، استفاده‌شده است (شکل 7b).

[1] Permutation Feature Importance
[2] Recursive Feature Elimination



شکل ۷. نتایج بررسی اهمیت ویژگی‌ها (a) و انتخاب آن‌ها (b) برای مدل‌سازی مجدد با استفاده از الگوریتم RF

بعد از مدل‌سازی مجدد با استفاده از ۵ پارامتر پراهمیت که در شکل ۷ نیز مشخص است و شامل، عمق آب زیرزمینی، نسبت جذب سدیم، پتاسیم لایه سطحی، درصد رطوبت اشباع و مساحت پلی‌گون است، مدل توسعه داده‌شده، مورد ارزیابی قرار گرفت که نتایج آن در شکل ۸ ارائه‌شده است. با توجه به نتایج، با کاهش پارامترهای ورودی مدل و انتخاب ویژگی‌های پراهمیت بازهم می‌توان به مدلی با دقت بسیار خوب دست‌یافت. به‌طوری‌که نتایج نشان داد، مدل دوم نیز در داده های واسنجی با دارا بودن R^2 معادل ۱/۰۰ و MSE، RMSE و MAE به ترتیب معادل ۰/۰۰۰۵، ۰/۰۲۱۷ و ۰/۰۱۸۴ و همچنین شاخص‌های آماری R^2، MSE، RMSE و MAE برای داده‌های صحت‌سنجی مدل به ترتیب ۰/۹۹، ۰/۰۰۲۰، ۰/۰۴۴۴ و ۰/۰۴۱۲ است که از دقت بالایی برخوردار است.



شکل ۸. نتایج واسنجی و صحت‌سنجی الگوریتم RF با استفاده از داده‌های انتخاب‌شده (نتایج شکل ۶) در برآورد و پیش‌بینی میزان عیار پتاسیم

در کل، بررسی نتایج حاصل نشان‌دهنده کارآمدی مدل‌های استفاده‌شده برای پیش‌بینی عیار پتاسیم شورابه است که روش RF عملکرد بهتری را نسبت به SVM داراست. از طرفی الگوریتم RF نسبت به چندخطی بین پارامترهای مستقل حساس نیست. همچنین محققان دیگری نقش الگوریتم‌ها دربرآورد منابع معدنی نشان دادند. تلفیق تخمین عیار با استفاده از مدل ترکیبی (Li et al., 2013)، تخمین منبع پلاتین (Chatterjee & Bandopadhyay, 2011)، تلفیق شبکه‌های عصبی مصنوعی و زمین‌آمار در تخمین ذخیره معدنی (Jalloh et al., 2016). مطابق نتایج به‌دست‌آمده الگوریتم‌های یادگیری ماشین، دارای این توانایی است که عیار را به‌خوبی در نقاط دیگر پلایا تخمین بزند. درواقع با استفاده از یک الگوریتم مناسب، امکان اظهارنظر در مورد برنامه‌های آتی استخراج مواد معدنی، به صورتی که امکان برنامه‌ریزی درازمدت وجود داشته باشد، ممکن خواهد بود.

نتیجه‌گیری

در مدل‌سازی با استفاده از یادگیری ماشین، نیازی به یادگیری و نوشتن فرمول‌های پیچیده مدل‌سازی نیست. هدف از استفاده الگوریتم یادگیری ماشین و استفاده از این الگوریتم‌ها در مدل‌سازی؛ تکمیل، ساده‌سازی و دقیق‌تر کردن مدل‌سازی در برآورد و پیش‌بینی عیار پتاسیم شورابه پلایاها است. در مطالعه حاضر ازآنجاکه در سیستم پلایا به علت آنکه شرایط غیرقابل‌پیش‌بینی است و این محیط دارای پیچیدگی بالا است، جهت جلوگیری از هزینه‌های بالای حفاری و آزمایشگاهی با استفاده از الگوریتم مناسب می‌توان پارامترهای مؤثر در عیار پتاسیم بالاتر را از طریق الگوریتم یادگیری ماشین مشخص کرد. به همین منظور استفاده از سیستم‌هایی که دارای خاصیت یادگیری باشند و انتخاب بهینه این پارامترها را انجام بدهند، احساس شد. بنابراین، از دو الگوریتم SVM و RF استفاده شد. نتایج نشان داد از بین دو الگوریتم RF و SVM، مدل RF با دقت بالایی توانسته است پیش‌بینی عیار پتاسیم شورابه را انجام دهد. اما در عمل از هر دو الگوریتم می‌توان با دقت مناسب و خوب استفاده کرد. در الگوریتم RF متغیرهای عمق لایه نمک، تخلخل، جرم مخصوص ظاهری، میزان کلر شورابه و دانسیته شورابه بیش‌ترین نقش را در برآورد عیار شورابه بالاتر داشته است و مقدار R^2 در داده واسنجی

برابر ۰/۹۹ و در داده صحت‌سنجی R^2 معادل ۰/۹۸ بوده است. در الگوریتم SVM از بین توابع مختلف، توابع خطی، چندجمله‌ای، تابع پایه شعاعی و تابع سیگموئید استفاده شد که در بین این توابع، تابع خطی بالاترین دقت را داشته است و متغیرهای که بیشترین اثر بر عیار پتاسیم شورابه را در این کرنل داشته‌اند، شامل پارامترهای تخلخل، درصد رطوبت اشباع، عمق آب زیرزمینی و نسبت جذب سدیم (SAR) بوده‌اند و مقدار R^2 در داده آموزشی برابر ۰/۹۷ و در داده آزمون R^2 معادل ۰/۹۷ بوده است. با توجه به‌دقت الگوریتم‌های بررسی‌شده، با استفاده از این الگوریتم‌ها می‌توان پارامترهایی که بیشترین تأثیر در عیار پتاسیم شورابه پلایا دارند؛ را پیش‌بینی نمود. این پارامترها کمک زیادی به اکتشاف ذخایر پتاس در سایر پلایا خواهد کرد و از هزینه‌های حفاری و آزمایشگاهی زیاد جلوگیری می‌کند.

منابع

امینی خویی، زهره.، و عبدالله پوری، علیرضا. (۱۳۹۶). طبقه‌بندی ترافیک شبکه با استفاده از الگوریتم جنگل تصادفی بهبودیافته. *نشریه علوم رایانشی*. ۲ (۲) : ۳۸-۲۴.

حسنی پاک، علی‌اصغر (۱۳۸۴). تحلیل داده‌های اکتشافی. چاپ دوم. تهران: مؤسسهٔ انتشارات دانشگاه تهران.

موسوی، سید علی.، رنجبرفردوئی، ابوالفضل.، موسوی، سید حجت.، و ساداتی نژاد، سید جواد. (۱۳۹۵). مدل‌سازی فرسایش‌پذیری خاک در منطقه خور و بیابانک با استفاده از شاخص‌های دورسنجی. *مهندسی اکوسیستم بیابان*. ۵ (۱۳) :۸۰-۶۷ .

منوچهری، شهرزاد (۱۳۸۲) پتاس، دانشنامه مواد و صنایع معدنی ایران، شرکت تحقیق و گسترش صنایع معدنی ایران.

REFERENCES

Amini Khoei, Z., & Abdullah Puri, A. (2017). Network traffic classification using improved random forest algorithm. *Journal of Computer Science*, 2 (2): 24 38.(In Persian).

Baudron, P., Alonso-Sarría, F., García-Aróstegui, J. L., Cánovas-García, F., Martínez-Vicente, D., & Moreno-Brotóns, J. (2013). Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification. *Journal of Hydrology*, *499*, 303-315.

Chatterjee, S., & Bandopadhyay, S. (2011). Goodnews Bay Platinum resource estimation using least squares support vector regression with selection of input space dimension and hyperparameters. *Natural Resources Research*, *20*, 117-129.

Chen, H., Huang, J. J., & McBean, E. (2020). Partitioning of daily evapotranspiration using a modified shuttleworth-wallace model, random Forest and support vector regression, for a cabbage farmland. *Agricultural Water Management*, *228*, 105923.

Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Dutta, S. (2006). *Predictive performance of machine learning algorithms for ore reserve estimation in sparse and imprecise data*. University of Alaska Fairbanks.

Dutta, S., Bandopadhyay, S., Ganguli, R., & Misra, D. (2010). Machine learning algorithms and their application to ore reserve estimation of sparse and imprecise data. *Journal of Intelligent Learning Systems and Applications*, *2*(02), 86-96.

Estefan, G., Sommer, R., & Ryan, J. (2013). Methods of soil, plant, and water analysis. *A manual for the West Asia and North Africa region*, *3*, 65-119.

García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2018). Principal components analysis random discretization ensemble for big data. *Knowledge-Based Systems*, *150*, 166-174.

Ghorbanzadeh, O., Rostamzadeh, H., Blaschke, T., Gholaminia, K., & Aryal, J. (2018). A new GIS-based data mining technique using an adaptive neuro-fuzzy inference system (ANFIS) and k-fold cross-validation approach for land subsidence susceptibility mapping. *Natural Hazards*, *94*, 497-517.

Hosni Pak, A (2005). Exploratory data analysis. second edition. Tehran: Tehran University Press.(In Persian).

Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

Jafrasteh, B., Fathianpour, N., & Suárez, A. (2018). Comparison of machine learning methods for copper ore grade estimation. *Computational Geosciences*, *22*, 1371-1388.

Jalloh, A. B., Kyuro, S., Jalloh, Y., & Barrie, A. K. (2016). Integrating artificial neural networks and geostatistics for optimum 3D geological block modeling in mineral reserve estimation: A case study. *International Journal of Mining Science and Technology*, *26*(4), 581-585.

Jeon, H., & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences*, *10*(9), 3211.

Kaneko, H.(2022). Cross□validated permutation feature importance considering correlation between features. *Analytical Science Advances*, *3*(9-10), 278-287.

Kisi, O., Karahan, M. E., & Şen, Z. (2006). River suspended sediment modelling using a fuzzy logic approach. *Hydrological Processes: An International Journal*, *20*(20), 4351-4362.

Li, X. L., Li, L. H., Zhang, B. L., & Guo, Q. J. (2013). Hybrid self-adaptive learning based particle swarm optimization and support vector regression model for grade estimation. *Neurocomputing*, *118*, 179-190.

Maleki, S., Ramazia, H. R., & Moradi, S. (2014). Estimation of Iron concentration by using a support vector machineand an artificial neural network-the case study of the Choghart deposit southeast of Yazd, Yazd, Iran. *Geopersia*, *4*(2), 201-212.

Manouchehri, Sh. (2003) Potash, Encyclopaedia of Mineral Materials and Industries of Iran, Iran Mineral Industries Research and Development Company.(In Persian).

Matías, J. M., Vaamonde, A., Taboada, J., & Gonzalez-Manteiga, W. (2004). Support vector machines and gradient boosting for graphical estimation of a slate deposit. *Stochastic Environmental Research and Risk Assessment*, *18*, 309-323.

McKay, G., & Harris, J. R. (2016). Comparison of the data-driven random forests model and a knowledge-driven method for mineral prospectivity mapping: A case study for gold deposits around the Huritz Group and Nueltin Suite, Nunavut, Canada. *Natural Resources Research*, *25*(2), 125-143.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Moorthi, S. M., Misra, I., Kaur, R., Darji, N. P., & Ramakrishnan, R. (2011). Kernel based learning approach for satellite image classification using support vector machine. In *2011 IEEE Recent Advances in Intelligent Computational Systems* (pp. 107-110). IEEE.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900.

Mousavi, S.A., RanjbarFardoi, A., Mousavi, S. H.(2022). Modeling soil erodibility in Khoor and Biabank region using remote sensing indicators. *Desert Ecosystem Engineering*, 5(13): 67 80.(In Persian).

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental monitoring and assessment*, *188*, 1-27.

Nitze, I., Schulthess, U., & Asche, H. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil*, *79*, 3540.

Oke, J., Akinkunmi, W. B., & Etebefia, S. O. (2019). Use of correlation, tolerance and variance inflation factor for multicollinearity test. *GSJ*, *7*(5).

Pozdnoukhov, A. (2005). Support vector regression for automated robust spatial mapping of natural radioactivity. *automatic mapping algorithms*, 57.

Ray, S. (2019). A quick review of machine learning algorithms. *International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39).

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, *67*, 93-104.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804-818.

Schnitzler, N., Ross, P. S., & Gloaguen, E. (2019). Using machine learning to estimate a key missing geochemical variable in mining exploration: Application of the Random Forest algorithm to multi-sensor core logging data. *Journal of Geochemical Exploration*, *205*, 106344.

Shaw, P. A., & Bryant, R. G. (2011). Pans, playas and salt lakes. *Arid zone geomorphology: process, form and change in drylands*, 373-401.

Sheng, L., Zhang, T., Niu, G., Wang, K., Tang, H., Duan, Y., & Li, H. (2015). Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF). *Journal of Analytical Atomic Spectrometry*, *30*(2), 453-458.

Soliman, O. S., & Mahmoud, A. S. (2012, May). A classification system for remote sensing satellite images using support vector machine with non-linear kernel functions. In *2012 8th International Conference on Informatics and Systems (INFOS)* (pp. BIO-181). IEEE.

Tenorio, V. O., Bandopadhyay, S., Misra, D., Naidu, S., & Kelley, J. (2015). Support vector machines applied for resource estimation of underwater glacier-type platinum deposits. *Application Of Computers an d Operations Research in the Mineral Industry*, 889-902.

Tiwari, S., Babbar, R., & Kaur, G. (2018). Performance evaluation of two ANFIS models for predicting water quality index of River Satluj (India). *Advances in Civil Engineering*, *2018*.

Twarakavi, N. K., Misra, D., & Bandopadhyay, S. (2006). Prediction of arsenic in bedrock derived stream sediments at a gold mine site under conditions of sparse data. *Natural Resources Research*, *15*, 15-26.

Wang, C., Pan, Y., Chen, J., Ouyang, Y., Rao, J., & Jiang, Q. (2020). Indicator element selection and geochemical anomaly mapping using recursive feature elimination and random forest methods in the Jingdezhen region of Jiangxi Province, South China. *Applied Geochemistry*, *122*, 104760.

Yang, Q., Li, X., & Shi, X. (2008). Cellular automata for simulating land use changes based on support vector machines. *Computers & geosciences*, *34*(6), 592-602.

Zhang, S., Xiao, K., Carranza, E. J. M., & Yang, F. (2019). Maximum entropy and random forest modeling of mineral potential: Analysis of gold prospectivity in the Hezuo–Meiwu district, west Qinling Orogen, China. *Natural Resources Research*, *28*, 645-664.

Zörb, C., Senbayram, M., & Peiter, E. (2014). Potassium in agriculture–status and perspectives. *Journal of plant physiology*, *171*(9), 656-669.

ماهان استاد

Evaluation of effective parameters for predicting the potassium grade of saline water by using support vector machine and random forest algorithms (case study: playa of Khoor and Biabank area city, Isfahan province)

EXTENDED ABSTRACT

Introduction:

With the increase in the world population, one of the important issues in the field of agriculture is increasing the production of agricultural products, and potassium is one of the most widely used elements to increase crop yield. For this reason, the demand for potassium fertilizers increases. One of the main sources of potassium fertilizers is underground water one of the important issues in saline water extraction is the amount of potassium grade of saline water. conventional methods of grade estimation, such as geometric and geostatistics techniques, cannot accurately estimate the grade value and have low accuracy. one of the novel solutions to estimate the grade of minerals is Machine learning algorithm, which perform evaluation and determination of the grade of mineral resources with high accuracy.

Objective:

The aim of this research is to Evaluation of effective parameters for predicting the potassium grade of saline water by using of machine learning algorithms (random forest and support vector machine) as new, low cost and cost effective methods and determining the effective parameters (independent variables used) with the greatest influence measuring the potassium grade in order to improve the utilization of potassium reserves and reduce executive, operational and laboratory costs.

Materials and method:

In this research is to use support vector machine (SVM) and random forest (RF) algorithms in order to predict and prioritize the effective parameters on the potassium grade of groundwater in playa Khoor and Biabank in Isfahan province. For this purpose, 55 different parameters were measured in 12 boreholes (sampling locations).The parameters measured as independent variables include the percentage of saturation moisture core at 15 different depths (sp1 sp15), the apparent specific gravity of the core at 15 different depths (pb1 pb15), the porosity of the core at 15 different depths (n1 n15), polygon area (ap), underground water depth (duw), salt layer depth (dsl), surface layer potassium (slp), brine density (d) and the amount of calcium (Ca), magnesium (Mg), sodium (Na), chlorine (Cl) and the dependent variable were also the potassium grade in the brine (Potassium Grade). three parameters n, sp and pb which were measured in 15 different depths; They were converted into an equivalent parameter using the principal component analysis (PCA) method. Also, three measured parameters, Ca, Mg, and Na were entered into the model with the sodium absorption ratio (SAR) formula. A total of 10 measured parameters were entered into the model as independent variables to predict the grade of potassium. Both RF and SVM models were implemented in Python programming language based on the relationship between dependent variable and independent variables. In different kernels of the SVM algorithm, in order to prevent the collinearity of independent parameters, all the different combinations of independent variables (2 to the power of 10 different combinations) considering the variance inflation factor (VIF) less than 8 and the highest coefficient of determination and the lowest MSE error are checked and the best combination were chosen. Permutation Feature Importance (PFI) and Recursive Feature Elimination (RFE) methods were used in the RF model to prioritize and select parameters for modeling.

Results and discussion:

The parameters effective in predicting the potassium grade of the both in the RF algorithm and the linear function of the SVM algorithm were sp, ap, duw, slp, SAR and n, sp, duw, and SAR respectively, which led to the best results (high determination coefficient and low error). Based on the results, the accuracy of the model (explanation coefficient) for the RF model and SVM (linear function) was 0.99 and 0.97, respectively, which indicates the good accuracy of both algorithms. Effective parameters in choosing suitable areas for drilling in order to extract potassium from saline water play a significant role and prevent repeated and time consuming tests in the laboratory, and the developed models can be used for this purpose

Conclusion:

Machine learning algorithms are one of the most important techniques for evaluating mineral grade estimation . Given that, a large part of the country consists of arid and semi arid areas, where there are many playas that are rich in underground saline water that have good and suitable reserves of potassium and because in playa, the conditions are unpredictable and the environment has high complexity ،Effective parameters in choosing suitable areas for drilling in order to extract potassium from saline water play a significant role.

Keywords: grade prediction, random forest, saline water, support vector machine

غیر قابل استناد