(پژوهشی)

تخمین تزریق‌پذیری خاک‌های دانه‌ای با به‌کارگیری داده‌های آزمایشگاهی و چند روش طبقه‌بندی هوشمند

هادی فتاحی*[1]، فاطمه جیریایی شراهی[1]

۱- دانشکده مهندسی علوم زمین، دانشگاه صنعتی اراک، اراک، ایران

(دریافت: فروردین ۱۴۰۰، پذیرش: بهمن ۱۴۰۰)

چکیده

تزریق‌پذیری یک پارامتر بااهمیت در عملیات تزریق است و پیش‌بینی صحیح آن منجر به انتخاب مناسب مواد سیال تزریق شونده می‌شود. این پارامتر در اکثر مواقع با روش‌های تجربی تخمین زده می‌شود و پیش‌بینی را با خطا همراه می‌کند. در این تحقیق سعی شد به‌منظور ساخت و صحت سنجی چند مدل داده‌کاوی در حوضه‌ی طبقه‌بندی، مجموعه‌ای از داده‌های آزمایشگاهی در عملیات تزریق موجود در چندین منبع به کار گرفته شود. مدل‌های طبقه‌بندی بکار گرفته‌شده در نرم‌افزار **Orange** شامل روش‌های ماشین بردار پشتیبان، شبکه عصبی مصنوعی، نزدیک‌ترین همسایگی، جنگل تصادفی و بیزین ساده می‌باشند. در این مدل‌ها، متغیرهای ورودی عبارت است از: نسبت آب به سیمان در دوغاب تزریق شونده، دانسیته نسبی خاک، فشار تزریق، درصد ریزدانه خاک، نسبت قطر ذرات خاک که ۱۵ درصد وزنی نمونه از آن کوچک‌تر است به قطر ذرات سیال تزریقی که ۸۵ درصد وزنی نمونه از آن کوچک‌تر است ($N_1=D_{15\ soil}/D_{85\ grout}$ و $N_2=D_{10\ soil}/D_{95\ grout}$). پس از مدل‌سازی، نتایج نشان می‌دهد که مدل‌های بکار گرفته‌شده به‌خوبی رابطه‌ی بین تزریق‌پذیری و عوامل مؤثر آن را تعریف می‌کنند و از دقت بالایی در تخمین تزریق‌پذیری خاک‌های دانه‌ای برخوردار هستند. با توجه به ماتریس کارایی مدل‌ها، مدل شبکه عصبی مصنوعی با دقت ۰/۸۶ درصد و مدل نزدیک‌ترین همسایگی با دقت ۰/۸۵ درصد عملکرد بهتری نسبت به سایر روش‌ها دارند. بعلاوه در بررسی اهمیت متغیرهای ورودی بر اساس شاخص‌های امتیازدهی، متغیرهای N_2 و N_1 تأثیرگذارترین متغیرها در روند پیش‌بینی صحیح تزریق‌پذیری هستند.

کلمات کلیدی

طبقه‌بندی، تزریق‌پذیری، خاک‌های دانه‌ای، نرم‌افزار **Orange**

*عهده‌دار مکاتبات: *h.fattahi@arakut.ac.ir*

۱- مقدمه

تزریق[۱] به فرآیند راندن مواد خارجی به نام دوغاب[۲] به داخل فضای خالی موجود در خاک و سنگ اطلاق می‌شود و هدف آن مقاوم‌سازی و بهبود خواص مکانیکی و هیدرولیکی محیط موردبررسی است [۱]. به این صورت خاک یا سنگ با اهدافی که مهندسین در پیش رو دارند سازگار می‌شود. از این عملیات در پروژه‌های مختلف ساختمانی، راه‌سازی و راه‌آهن، به‌منظور کنترل روانگرایی[۳] خاک نیز استفاده می‌شود [۲]. سیالی که به درون حفره‌ها و شکاف‌های محیط تزریق می‌شود مانند یک مایع ویسکوز متشکل از دانه‌هایی است که اندازه‌ی آنها در عملیات تزریق اهمیت پیدا می‌کند. به‌طوری‌که اگر برای خاک‌های ریزدانه از ذرات دانه‌درشت استفاده شود، فضای خالی موجود در خاک با این ذرات بسته شده و باعث افزایش کاذب فشار تزریق می‌شود و خوردند تزریق کاهش پیدا می‌کند. از طرفی اگر ذرات سیال بسیار دانه‌ریز و ابعاد ذرات خاک درشت‌دانه باشند، هزینه‌های تزریق بسیار افزایش می‌یابند. به همین دلیل تعیین نسبت تزریق‌پذیری[۴] در عملیات تزریق به‌عنوان یک پارامتر مهم تلقی می‌شود. تزریق‌پذیری در روابط تجربی توسط مقایسه‌ی ابعاد ذرات در محیط خاک یا سنگ میزبان و دوغاب یا سیال تزریق، انجام می‌شود. یکی از این روابط که در سال ۱۹۵۸ و توسط بوروول[۵] ارائه شده، در رابطه (۱) بیان شده است [۳].

$$GR = D_{15soil} / D_{85grout} \succ 25 \qquad (۱)$$

که در آن D_{15soil} اندازه‌ی قطر ذراتی است که ۱۵ درصد وزنی نمونه‌ی خاک از آن اندازه کوچک‌تر هستند و $D_{85grout}$ اندازه قطر ذراتی از سیال تزریقی است که ۸۵ درصد وزنی نمونه سیال از آن اندازه کوچک‌تر هستند. بررسی‌هایی که امروزه با به‌کارگیری علم داده‌کاوی در این خصوص انجام شده، نشان می‌دهد که تزریق‌پذیری خاک‌های دانه‌ای علاوه بر اندازه ذرات، متأثر از عوامل مختلفی از خاک و ماده تزریق شونده است که پیش‌بینی تزریق پذیری را با دقت بالاتری انجام می‌دهد. در طول تاریخ بسیاری از محققان به پیش‌بینی تزریق پذیری خاک از طریق روابط تجربی پرداخته‌اند؛ اما امروزه قابلیت روش‌های داده‌کاوی در پیش‌بینی‌های دقیق نشان داده که یک رویکرد در پیش‌بینی تزریق‌پذیری استفاده از انواع مدل‌های داده‌کاوی و سیستم‌های استنباطی است. در این راستا پژوهش‌های به شرح زیر صورت گرفته است.

تکین و اکبس در سال ۲۰۱۱ به توسعه‌ی یک شبکه عصبی مصنوعی پرداختند و توسط مدل پیش‌بینی ساخته‌شده، تزریق‌پذیری نمونه خاک‌های دانه‌ای تست شده در آزمایشگاه را بررسی کردند [۴]. لیائو و همکاران در سال ۲۰۱۱ با چالش تخمین تزریق‌پذیری در یک سطح خطای معقول مواجه شدند و برای حل آن به‌جای استفاده از روابط تجربی، از شبکه عصبی با عملکرد شعاعی (RBFNN[۶]) بهره گرفتند [۵]. چنگ و همکاران در سال ۲۰۱۴ با استفاده از مدل جدید کمترین مربعات ماشین بردار پشتیبان (LS-SVM[۷]) نتایج عملیات تزریق را با دقت بالایی پیش‌بینی کردند [۶]. ترن و هانگ در سال ۲۰۱۴ پیش‌بینی تزریق‌پذیری با دوغاب سیمان دانه‌ریز را انجام دادند و در آن از روش‌های شبکه عصبی مصنوعی و ماشین بردار پشتیبان بهره گرفتند [۷]. چنگ و همکاران در سال ۲۰۱۴ برای پیش‌بینی تزریق‌پذیری در عملیات تزریق، شبکه بیزین و روش نزدیک‌ترین همسایگی را ادغام کردند. آن‌ها از تئوری بیزین برای تخمین احتمالات تزریق‌پذیری و از روش نزدیک‌ترین همسایگی برای تقریب توابع چگالی احتمال شرطی کمک گرفتند [۸]. چنگ و هانگ همچنین در سال ۲۰۱۴ با استفاده از تئوری استنتاج فازی در روش طبقه‌بندی نزدیک‌ترین همسایگی، به یک مدل جدید (EFKNIM[۸]) رسیدند و آن را در پیش‌بینی تزریق‌پذیری دوغاب سیمان دانه‌ریز در عملیات تزریق به کار بردند [۹]. هانگ و همکاران در سال ۲۰۱۶ با استفاده از یک مدل ترکیبی ماشین بردار پشتیبان به پیش‌بینی تزریق‌پذیری در فرآیند تزریق پرداختند [۱۰]. آزادی‌زاده و مجدی در سال ۲۰۱۹ از طریق سه سیستم استنتاج فازی، پیش‌بینی تزریق‌پذیری در خاک‌های دانه‌ای را انجام دادند [۱۱]. تکین و اکبس در سال ۲۰۱۹ با بهره‌گیری از سیستم استنتاج فازی به پیش‌بینی تزریق‌پذیری در خاک‌های دانه‌ای پرداختند [۱۲]. ونگ و همکاران در سال ۲۰۱۹ به‌منظور پیش‌بینی تزریق‌پذیری در توده سنگ درزه‌دار، با ترکیب الگوریتم بهینه‌سازی گرگ خاکستری و ماشین بردار پشتیبان یک مدل پیش‌بینی ارائه دادند [۱۳].

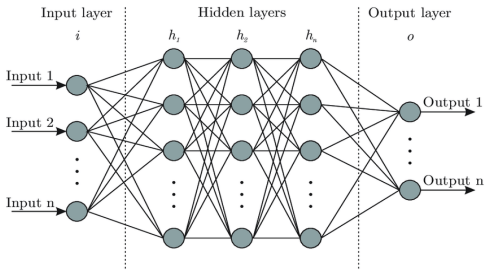
هدف از این تحقیق به‌کارگیری چند مدل مهم طبقه‌بندی داده‌کاوی شامل: شبکه عصبی مصنوعی، ماشین بردار پشتیبان، نزدیک‌ترین همسایگی، جنگل تصادفی و

بیزین ساده برای مدل‌سازی تزریق‌پذیری در خاک‌های دانه‌ای و همچنین معرفی نرم‌افزار Orange به‌عنوان یک راه ساده و سریع برای اهداف داده‌کاوی است. ضمن اینکه مدل‌سازی‌ها با دقت بالاتری انجام شود. به این منظور مجموعه‌ای از اطلاعات آزمایشگاهی مربوط به تزریق‌پذیری در چهار مطالعه که شامل ۸۷ داده است به کار گرفته می‌شود تا بتوان مدل‌هایی کارآمد برای پیش‌بینی تزریق‌پذیری ساخت که در بخش بعد این روش‌ها به‌طور اختصار توضیح داده می‌شوند.

۲- معرفی اجمالی روش‌های طبقه‌بندی استفاده‌شده در این تحقیق

۲-۱- شبکه عصبی مصنوعی (ANN[۹])

یک نرون عصبی از مغز انسان و عملکرد آن را می‌توان با الگوهای ریاضی شبیه‌سازی کرد. شبکه عصبی مصنوعی روشی است که با بهره‌گیری از این شبیه‌سازی به پردازش اطلاعات می‌پردازد. الگوهای الکترونیکی شبکه عصبی بر اساس تجربه و یادگیری بنا شده است. سامانه پردازش اطلاعات این روش از تعداد زیادی عناصر پردازشی به‌هم‌پیوسته به نام نرون‌ها تشکیل شده است که برای حل یک مسئله با هم هماهنگ عمل می‌کنند. این نرون‌ها ابتدا با داده‌های مشاهداتی فرآیند مسئله را یاد می‌گیرند و الگوهای آن را استخراج می‌کنند. این روش درواقع یک مدل پیچیده ریاضی است که قابلیت ساخت مدل و ایجاد روابط ریاضی غیرخطی برای درون‌یابی را دارد. به‌طورکلی شبکه عصبی شامل سه لایه ورودی، پنهان و خروجی است. شکل ۱ ساختار یک نرون مصنوعی و بخش‌های مختلف آن را نشان می‌دهد. تک‌تک ورودی‌ها در وزن‌های مخصوص خود ضرب و باهم جمع می‌شوند و در انتها توسط تابع‌هایی خاص، خروجی از روی ورودی تصمیم‌گیری می‌شود. وقتی خطای شبکه به حداقل برسد، خروجی شبکه نیز مشابه خروجی هدف خواهد شد [۱۴].



شکل ۱: ساختار بخش‌های یک سلول عصبی مصنوعی [۱۵]

۲-۲- ماشین بردار پشتیبان (SVM[۱۰])

روش ماشین بردار پشتیبان ازجمله روش‌های طبقه‌بندی و رگرسیون است که در آن یادگیری با نظارت انجام می‌شود. اساس کار این روش طبقه‌بندی خطی داده‌ها است، به‌طوری‌که در تقسیم‌بندی خطی داده‌ها سعی می‌شود خطی انتخاب شود که حاشیه اطمینان آن بیشتر باشد. این ترکیب خطی از تابع کرنل بر روی مجموعه‌ای از داده‌های آموزشی با نام بردارهای پشتیبان عمل می‌کند. یکی از ویژگی‌های مهم ماشین بردار پشتیبان این است که طوری عمل می‌کند که خطای تجربی طبقه‌بندی، کمینه و حاشیه‌های هندسی در بیشترین مقدار ممکن قرار می‌گیرد. به همین دلیل به این روش، طبقه‌بندی بیشینه کننده حاشیه نیز گفته می‌شود؛ بنابراین ابر صفحه‌ای که انتخاب می‌شود بایستی فاصله‌ی آن از نزدیک‌ترین داده‌ها در هر دو طرف جداکننده خطی بیشینه باشد. این ابر صفحه از طریق رابطه (۲) به دست می‌آید.

$$W^T \phi(x) + b = 0 \qquad (۲)$$

در این رابطه بردار وزن w برداری عمود بر ابر صفحه است. b بردار بایاس است که به‌منظور اندازه‌گیری فاصله ابر صفحه تا مبدأ استفاده می‌شود و کرنلی برای انتقال داده به فضایی با ابعاد بالاتر است [۱۶].

۲- ۳- نزدیک‌ترین همسایگی (KNN[۱۱])

روش نزدیک‌ترین همسایگی یکی از روش‌های داده‌کاوی است که هدف آن طبقه‌بندی و تخمین ویژگی‌های متغیر هدف در داده‌های مجهول است. پیش‌بینی و تخمین در این روش بر اساس شباهت داده‌های مجهول با داده‌های معلومی که در نزدیکی آن‌ها قرار دارد انجام می‌شود. به عبارتی در این روش گروهی شامل K رکورد از مجموعه رکوردهای آموزشی که در همسایگی رکوردهای آزمایشی قرار دارند، انتخاب می‌شود، سپس متغیر پاسخ در رکوردهای آزمایشی از طریق برخی معیارهای تشابه مانند توابع فاصله تخمین زده می‌شود. مزیت اصلی این مدل، نا پارامتری، اجرای ساده، قابلیت مدل‌سازی غیرخطی و عملکرد با بازدهی بالا در برخورد با تعداد دسته‌های زیاد از داده‌ها است. یک مسئله‌ی مهم در این روش انتخاب تابع فاصله برای محاسبه فاصله‌ی بین

داده‌ها است که معمولاً از فاصله اقلیدسی طبق رابطه (۳) استفاده می‌شود [۱۷].

$$d(p,q)=\sqrt{\sum_{i=1}^{k}(p_i-q_i)^2} \qquad (۳)$$

که در آن k تعداد فیلدهای هر رکود است و مقادیر ویژگی i ام برای رکودها است.

۲- ۴- جنگل تصادفی[۱۲]

جنگل تصادفی یکی از روش‌های یادگیری ماشین است که در علوم مختلف مورداستفاده قرار می‌گیرد. جنگل تصادفی جزء روش‌های درخت-پایه است و هر درخت از ریشه، گره‌ها و برگ‌ها تشکیل شده است. مجموعه‌ای از داده‌های آموزشی شامل متغیرهای ورودی و خروجی مانند رابطه (۴) برای آموزش هر درخت به کار گرفته می‌شود.

$$(X_i, Y_i) \qquad i=1,2,...,n \qquad (۴)$$

روش‌های درختی اگرچه در تفسیر نتایج ساده بوده اما باکمی پیچیدگی در مدل‌ها، متفاوت شده و مدل‌های ساده توانایی پوشش دادن پیچیدگی داده‌ها را ندارد. درحالی‌که جنگل تصادفی این نقص را برطرف کرده و صحت طبقه‌بندی را بالا می‌برد [۱۸].

۲- ۵- بیزین ساده[۱۳]

روش بیزن ساده یکی از روش‌های پرکاربردی است که با استفاده از احتمالات و تئوری بیزین، نمونه‌های جدید را طبقه‌بندی می‌کند. اساس طبقه‌بندی در این روش احتمال وقوع یا عدم وقوع یک پدیده است. بیزین ساده جزء ساده‌ترین روش‌ها است که دقت آن نیز قابل‌قبول است اما می‌توان با استفاده از برآورد چگالی کرنل دقت آن را بالا برد. این روش فرض بر مستقل بودن متغیرها دارد، به همین دلیل به بیز ساده‌لوح نیز معروف است. بیزین ساده برای مواقعی که تعداد متغیرها کم اما تعداد مشاهدات زیاد است برای تشخیص دسته‌ها مناسب است [۱۹].

۳- شاخص‌های ارزیابی مدل‌های طبقه‌بندی

یکی از کاربردی‌ترین ابزارها برای ارزیابی کارآیی و دقت مدل‌ها در مسائل طبقه‌بندی، ماتریس درهم‌ریختگی[۱۴] است. جدول ۱ نمونه‌ای از این ماتریس را برای طبقه‌بندی‌های دو کلاسه نشان می‌دهد. در این ماتریس TP تعداد کلاس‌های مثبتی است که به‌درستی پیش‌بینی شده‌اند، FP تعداد کلاس‌های منفی است که کلاس‌بندی به اشتباه مثبت پیش‌بینی کرده است، FN تعداد کلاس‌های مثبتی است که کلاس‌بندی به اشتباه منفی پیش‌بینی کرده است و TN تعداد کلاس‌های منفی است که به‌درستی، منفی پیش‌بینی شده‌اند.P و N نیز به ترتیب تعداد کلاس‌های مثبت و منفی است. با استفاده از این ماتریس معیارهای آماری گوناگونی را می‌توان برای ارزیابی کارآیی مدل‌ها محاسبه کرد [۲۰].

جدول ۱: ماتریس درهم‌ریختگی برای طبقه‌بندی دو کلاسه

		کلاس پیش‌بینی شده		
	-	Yes	No	Total
کلاس واقعی	Yes	TP	FN	P
	No	FP	TN	N
	Total	p´	N´	P + N

معیار دقت طبقه‌بندی طبق رابطه‌ی (۵) عبارت است از نسبت نمونه‌های طبقه‌بندی شده مثبت صحیح به داده‌هایی که مثبت پیش‌بینی شده‌اند.

$$\text{Precision} = (TP) / (TP + FP) \qquad (۵)$$

صحت طبقه‌بندی مطابق رابطه‌ی (۶) از نسبت نمونه‌های طبقه‌بندی شده‌ی صحیح به تعداد کل نمونه‌ها به دست می‌آید.

$$\text{Accuracy} = (TP + TN) / (P + N) \qquad (۶)$$

شاخص بازیابی در رابطه‌ی (۷) عبارت است از نسبت نمونه‌های طبقه‌بندی شده مثبت صحیح به کل نمونه‌هایی که به‌طور واقعی مثبت هستند.

$$\text{Recall} = (TP) / (TP + FN) \qquad (۷)$$

در رابطه‌ی (۸)، F_1 میانگین هارمونیک وزن‌دار شده‌ی دو شاخص بازیابی و دقت است.

$$F_1=2\times\text{Recall}\times\text{Precision}/(\text{Recall}+\text{Precision}) \qquad (۸)$$

علاوه بر شاخص‌های ذکرشده شاخص دیگری نیز به نام نمودار مشخصه عملکرد[۱۵] وجود دارد که برای مقایسه عملکرد مدل‌های طبقه‌بندی دودوئی کاربرد دارد. این نمودار به‌طور هم‌زمان دو شاخص حساسیت و بازیابی را بررسی می‌کند. بر محور افقی این نمودار نرخ مثبت کاذب (FPR[۱۶]) و بر محور عمودی نرخ مثبت صحیح (TPR[۱۷]) قرار می‌گیرد. خط قطری در نمودار آن را به دو ناحیه‌ی بالا

یعنی مطلوب و پایین یعنی نامطلوب تقسیم می‌کند. منحنی مدل‌ها هرچه به بالا و سمت چپ نمودار نزدیک‌تر باشند عملکرد بهتری دارند. می‌توان گفت مدلی که سطح زیر منحنی (AUC[۱۸]) آن بیشتر باشد بهتر است.

۴- معرفی نرم‌افزار Orange

نرم‌افزار Orange یک ابزار متن‌باز برای تصویرسازی داده‌ها و تجزیه‌وتحلیل بر پایه پایتون است و همچنین دارای قطعاتی برای یادگیری ماشین و متن کاوی است. نرم‌افزار در اواخر دهه ۱۹۹۰ در آزمایشگاه بیوانفورماتیک، دانشکده علوم کامپیوتر و اطلاعات در دانشگاه Ljubljana اسلوونی توسعه یافته است و یکی از قدیمی‌ترین ابزار داده‌کاوی محسوب می‌شود. کار با این نرم‌افزار به صورت تعاملی بوده و می‌توان عملیات داده‌کاوی را بدون کدنویسی انجام داد. یکی از ویژگی‌های این نرم‌افزار گرافیکی بودن آن است که کار با آن را بسیار ساده و قابل فهم کرده است. نرم‌افزار Orange شامل مجموعه‌ای از الگوریتم‌های یادگیری ماشین تحت نظارت برای طبقه‌بندی و رگرسیون، روش‌های اعتبار سنجی بر اساس نمونه‌برداری و ارزیابی قابل‌اطمینان، الگوریتم‌های بدون نظارت یادگیری برای خوشه‌بندی، الگوریتم‌های قواعد انجمنی، الگوریتم‌هایی برای پردازش زبان طبیعی و استخراج متن و الگوریتم‌هایی برای تجزیه‌وتحلیل سری‌های زمانی و غیره است.

۵- تحلیل و نتایج

بررسی تزریق‌پذیری خاک به عبارتی یک مقایسه بین ابعاد و توزیع دانه‌بندی ذرات خاک یا سنگ مورد تزریق و سیال تزریق شونده است، به‌طوری‌که شرایط برای عملیات تزریق با دقت قابل قبولی تخمین زده شود. در این بررسی اگر علاوه بر اندازه ذرات تشکیل‌دهنده محیط و سیال تزریق شونده، سایر متغیرها نیز در نظر گرفته شود، نتایج مطلوب‌تر و دقیق‌تری به دست خواهد آمد. در این قسمت به‌منظور ارزیابی چند مدل از روش‌های داده‌کاوی، از اطلاعات تزریق‌پذیری آزمایشگاهی چند مطالعه موردی استفاده شده است [۴, ۲۱-۲۳]. پنج مدل طبقه‌بندی نزدیک‌ترین همسایگی، ماشین بردار پشتیبان، جنگل تصادفی، شبکه عصبی مصنوعی و بیزین ساده در نرم‌افزار Orange برای پیش‌بینی تزریق‌پذیری خاک‌های دانه‌ای با دوغاب سیمان ساخته شده است. متغیر خروجی، ویژگی تزریق‌پذیری است که به‌عنوان یک متغیر دودوئی دارای دو حالت صفر یعنی تزریق ناپذیر و ۱ یعنی تزریق پذیر است. متغیرهای ورودی نیز شامل نسبت آب به سیمان در دوغاب یا ویسکوزیته (W/C)، دانسیته نسبی خاک (Dr)، فشار تزریق (P)، درصد نمونه خاکی که از سرند ۰/۶ میلی‌متر عبور می‌کند یا همان درصد ریزدانه خاک (FC)، $N_1=D_{15soil}/D_{85grou}$ و $N_2=D_{10soil}/D_{95grout}$ است. مشخصات آماری داده‌های مربوط به متغیرهای ورودی در جدول ۲ آورده شده است. این داده‌ها برای آماده‌سازی و ورود به ساخت مدل‌های طبقه‌بندی بایستی بین مقادیر ۱ و ۱- نرمالایز شوند.

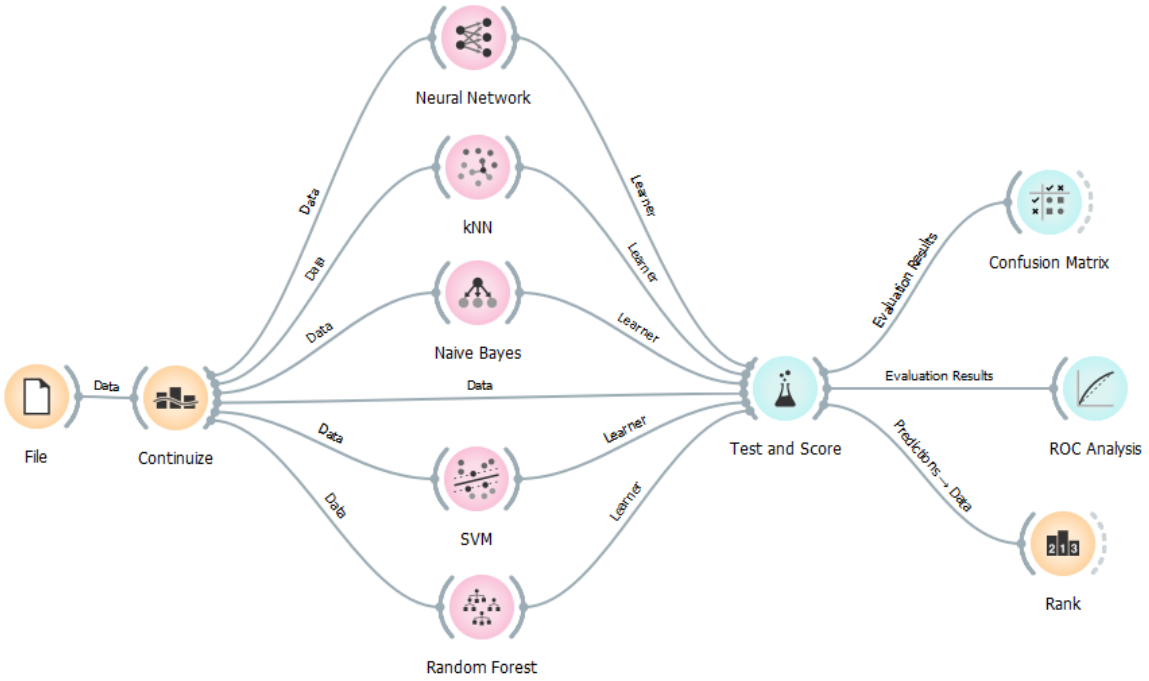
جدول ۲: مشخصات آماری متغیرهای ورودی مدل‌ها

متغیرها	مینیمم	ماکزیمم	میانگین	انحراف معیار
ویسکوزیته	۱	۶	۱٫۷۷	۱٫۳۱
دانسیته نسبی خاک	۲۷	۸۰	۵۳٫۲	۲۲٫۳۴
فشار تزریق	۵۰	۶۹۰	۲۷۹٫۶	۱۹۴٫۹۹
درصد ریزدانه خاک	۱	۱۰۰	۴۲٫۲۵	۳۳٫۵۴
N_1	۱۰	۷۶۲	۶۸	۹۸٫۴۲
N_2	۴	۴۳۳٫۳۳	۳۵٫۴۵	۵۴٫۲۸

در نرم‌افزار Orange برای ساخت مدل‌های دلخواه، نماد آن‌ها به‌صورت گرافیکی در صفحه جاگذاری می‌شوند. شکل ۲ شمای کلی عملیات انجام‌شده برای ساخت مدل‌ها در نرم‌افزار را نشان می‌دهد. به‌این‌ترتیب که ابتدا فایل داده‌ها فراخوانی شده و پس از نرمالایز شدن وارد مدل‌ها شده‌اند. هر مدل شامل چندین پارامتر تنظیم می‌باشند که برای مجموعه داده‌های مختلف این پارامترها می‌توانند تغییر کنند تا بهترین نتایج با بالاترین دقت به دست آید و یک مدل کارآمد ایجاد شود. به‌منظور صحت سنجی و مقایسه مدل‌ها، مدل‌سازی وارد مرحله‌ی Test and Score شده و با شاخص‌های ارزیابی مختلف بررسی می‌شوند. درنهایت نیز انواع مختلفی از ارزیابی جمعی داده‌ها گرفته می‌شود که در ادامه به آن‌ها پرداخته خواهد شد.

مقادیر محاسبه‌شده برای معیارهای مختلف ارزیابی در جدول ۳ آورده شده است. همان‌طور که اعداد نشان می‌دهند، مقادیر شاخص‌های ارزیابی برای روش‌ها، تقریباً نزدیک به هم هستند. از نظر شاخص AUC، مدل جنگل

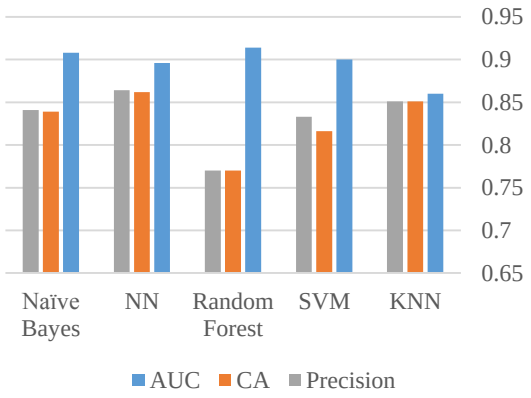
تصادفی با اختلاف اندکی بهترین روش معرفی‌شده و روش نزدیک‌ترین همسایگی، کمترین مقدار از این شاخص را به خود اختصاص داده است؛ اما از نظر سایر شاخص‌ها، شبکه عصبی مصنوعی از سایر روش‌ها بالاتر بوده و روش نزدیک‌ترین همسایگی نیز بسیار به آن نزدیک است. از طرفی روش جنگلی تصادفی دارای پایین‌ترین مقدار از شاخص‌ها است. شاخص‌های صحت طبقه‌بندی که میزان پیش‌بینی‌های صحیح نسبت به تعداد کل داده‌ها و دقت طبقه‌بندی که میزان پیش‌بینی مثبت صحیح نسبت به تعداد کل کلاس مثبت یا همان کلاس ۱ است، برای هر روش تقریباً مشابه هستند و برای روش‌های شبکه عصبی مصنوعی و نزدیک‌ترین همسایگی بیشترین مقدار را به خود اختصاص داده‌اند. این موضوع بیانگر عملکرد خوب روش‌های مذکور است. این مقایسه از نظر شاخص‌های مختلف را می‌توان در شکل ۳ نیز مشاهده کرد. با چشم‌پوشی از معیار AUC روش‌های شبکه عصبی مصنوعی و نزدیک‌ترین همسایگی بهترین روش‌ها هستند.



شکل ۲: شمای کلی مدل‌سازی در نرم‌افزار Orange

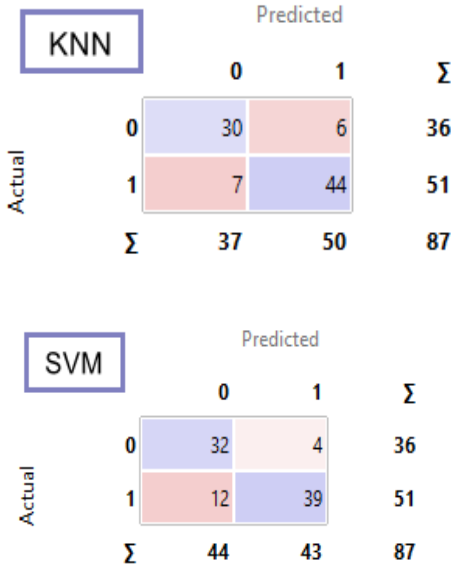
جدول ۳: مقدار شاخص‌های ارزیابی به ازای هر مدل طبقه‌بندی

مدل‌های طبقه‌بندی	مساحت زیر منحنی ROC	صحت طبقه‌بندی	F_1	دقت طبقه‌بندی	بازیابی مدل
نزدیک‌ترین همسایگی	۰٫۸۶۰	۰٫۸۵۱	۰٫۸۵۱	۰٫۸۵۱	۰٫۸۵۱
ماشین بردار پشتیبان	۰٫۹۰۰	۰٫۸۱۶	۰٫۸۱۷	۰٫۸۳۳	۰٫۸۱۶
جنگل تصادفی	۰٫۹۱۴	۰٫۷۷۰	۰٫۷۷۰	۰٫۷۷۰	۰٫۷۷۰
شبکه عصبی مصنوعی	۰٫۸۹۶	۰٫۸۶۲	۰٫۸۶۳	۰٫۸۶۴	۰٫۸۶۲
بیزین ساده	۰٫۹۰۸	۰٫۸۳۹	۰٫۸۴۰	۰٫۸۴۱	۰٫۸۳۹



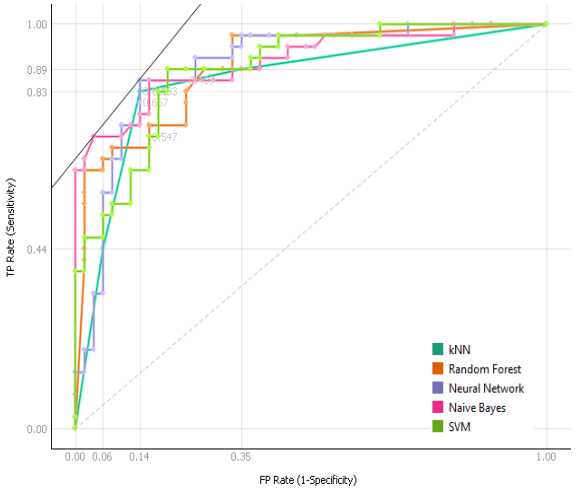
شکل ۳: نمودار ستونی شاخص ارزیابی در مدل‌ها

مقایسه‌ی پیش‌بینی انجام‌شده برای متغیر تزریق‌پذیری با استفاده از مدل‌های طبقه‌بندی و مقدار واقعی به‌دست‌آمده از آن در آزمایشگاه منجر به تشکیل ماتریس درهم‌ریختگی می‌شود که در شکل ۴ برای دو روش ماشین بردار پشتیبان و نزدیک‌ترین همسایگی نمایش داده شده است. معیارهای ارزیابی توسط این ماتریس‌ها محاسبه می‌شوند. عدد موجود در هر درایه بیانگر تعداد داده‌هایی است که در واقعیت و در پیش‌بینی در کلاس‌های صفر یا یک قرار می‌گیرند. جمع کل درایه، تعداد کل داده‌ها را نشان می‌دهد.



شکل ۴: ماتریس درهم‌ریختگی برای دو روش ماشین بردار پشتیبان و نزدیک‌ترین همسایگی

شکل ۵ نمودار ROC را نمایش می‌دهد و همان‌طور که مشخص است، منحنی مربوط به همه‌ی مدل‌ها در بالای خط قطری و در ناحیه‌ی مطلوب قرار گرفته‌اند. به این دلیل که عملکرد مدل‌های طبقه‌بندی تقریباً مشابه هم بوده است، منحنی آن‌ها نزدیک به هم و کمی پیچیده رسم شده است. منحنی مربوط به روش‌های ANN، جنگل تصادفی و بیزین ساده بالاتر از سایر روش‌ها قرار گرفته است و عملکرد بهتری دارد.



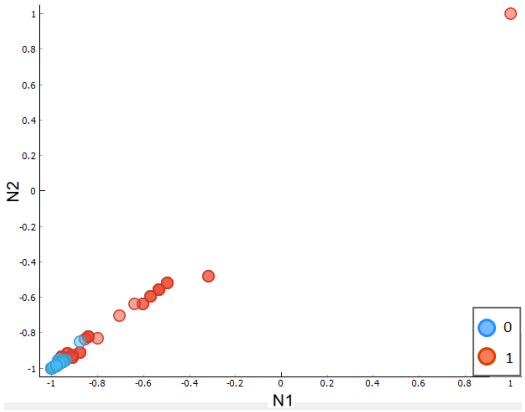
شکل ۵: نمودار ROC مدل‌های طبقه‌بندی

یکی از قابلیت‌های نرم‌افزار Orange مطالعه‌ی تأثیرگذاری و اهمیت متغیرهای ورودی بر پیش‌بینی متغیر هدف است و به عبارتی حساسیت متغیر خروجی نسبت به متغیرهای ورودی را بررسی می‌کنند. شکل ۶ نشان می‌دهد متغیر N_2 بر اساس سه معیار بهره اطلاعاتی[۱۹]، بهره اطلاعاتی نسبی[۲۰] و شاخص جینی[۲۱] در جایگاه اول قرار دارد و متغیر N_1 نیز با اختلاف بسیار کمی در مقادیر شاخص‌ها در اولویت دوم قرار می‌گیرد. بعلاوه در ردیف آخر W/C کمترین مقدار از شاخص‌ها را دارا است و نشان می‌دهد نقش کمی در پیش‌بینی صحیح تزریق‌پذیری دارد. معیارهای نام‌برده شده، روش‌های مختلف ارزش‌گذاری یا انتخاب شایسته‌ترین صفت‌ها در انواع درخت‌های تصمیم هستند. معیار بهره اطلاعاتی در نوع درخت تصمیم ID3 استفاده می‌شود و هرچه مقدار آن برای یک ویژگی یا صفت بیشتر باشد، آن ویژگی مهم‌تر است و در سطوح بالاتر درخت تصمیم و نزدیک به ریشه قرار می‌گیرد. شاخص جینی نیز در درخت تصمیم CART به کار می‌رود و هرچه مقدار آن برای یک ویژگی کمتر باشد، آن ویژگی اهمیت بیشتری دارد. در بهره اطلاعاتی نسبی نیز، از بین ویژگی‌ها آن‌که نسبت بهره اطلاعاتی به آنتروپی آن بزرگ‌تر باشد، وزن بیشتری خواهد داشت.

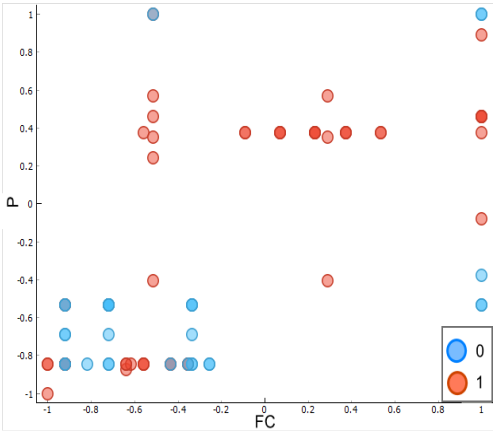
	#	Info. gain	Gain ratio	Gini
N2		0.592	0.296	0.305
N1		0.542	0.273	0.280
P		0.327	0.170	0.189
Dr		0.176	0.097	0.100
FC		0.104	0.052	0.065
W/C		0.057	0.040	0.032

شکل ۶: رتبه‌بندی و امتیازدهی متغیرها ورودی

نمودار پراکندگی یک راه بررسی چگونگی ارتباط دوبه‌دوی متغیرها به‌صورت گرافیکی است؛ بنابراین به‌منظور بررسی مهم‌ترین متغیرها، نمودار پراکندگی دو متغیر که همبستگی بالایی دارند یعنی N_1 و N_2 و همچنین دو متغیر که همبستگی پایینی دارند، یعنی FC و P در شکل‌های ۷ و ۸ رسم شده است. دو متغیر N_1 و N_2 یک رابطه‌ی بسیار منظمی داشته و تغییرات خطی نسبت به یکدیگر دارند. اما در نمودار پراکندگی دو متغیر FC و P داده‌ها بسیار پراکنده هستند و هیچ نظم خاصی در ارتباط بین آنها وجود ندارد و همبستگی ضعیفی دارند. به عبارتی تغییرات آنها نسبت به یکدیگر از قانون خاصی پیروی نمی‌کند.



شکل ۷: نمودار پراکندگی دو متغیر N_1 و N_2



شکل ۸: نمودار پراکندگی دو متغیر FC و P

۶- نتیجه‌گیری

پیش‌بینی تزریق‌پذیری خاک‌های دانه‌ای با دوغاب سیمان، با ساخت پنج مدل طبقه‌بندی نزدیک‌ترین همسایگی، ماشین بردار پشتیبان، جنگل تصادفی، شبکه عصبی مصنوعی و بیزین ساده در نرم‌افزار Orange انجام شد. نتایج نشان داد:

- به‌طورکلی مدل‌های داده‌کاوی در بحث طبقه‌بندی که در این تحقیق به کار گرفته شد، دارای قابلیت خوبی برای پیش‌بینی تزریق‌پذیری هستند و این کار را با دقت بالایی انجام می‌دهند و می‌توانند از این نظر جایگزین روش‌های تجربی و روابط قدیمی باشند.
- بر اساس اکثر شاخص‌های ارزیابی مدل‌های طبقه‌بندی، به ترتیب مدل‌های شبکه عصبی مصنوعی، نزدیک‌ترین همسایگی و بیزین ساده، عملکرد بهتری داشته و با دقت بالایی تزریق‌پذیری را پیش‌بینی می‌کنند. هرچند همه‌ی روش‌ها عملکردی نزدیک به هم دارند، اما مدل جنگلی تصادفی نسبت به سایر روش‌ها ضعیف‌تر عمل کرده است.
- اهمیت متغیرهای N_1 و N_2 در تخمین تزریق‌پذیری در سه معیار بهره اطلاعاتی، بهره اطلاعاتی نسبی و شاخص جینی دارای مقادیر بالاتری نسبت به سایر متغیرها هستند و تأثیر زیادی در روند پیش‌بینی صحیح تزریق‌پذیری دارند. بعلاوه متغیر W/C با پایین‌ترین مقادیر از شاخص‌ها تأثیر چندانی در عملیات طبقه‌بندی ندارد و می‌توان از آن چشم‌پوشی کرد. همچنین نمودار پراکندگی متغیرها نشان داد دو متغیر N1 و N2 همبستگی بالایی باهم دارند.

مراجع

[1] Huang, C., J. Fan, and W.J.S.-G. Yang, A study of applying microfine cement grout to sandy silt soil. 2007. 111(1): p. 71-82.

[2] Miller, E.A., G.A.J.J.o.G. Roycroft, and G. Engineering, Compaction grouting test program for liquefaction control. 2004. 130(4): p. 355-361.

[3] Burwell jr, E.B.J.J.o.t.S.M. and F. Division, Cement and clay grouting of foundations: practice of the corps of engineers. 1958. 84(1): p. 1551-1-1551-22.

[4] Tekin, E., S.J.B.o.E.G. Akbas, and t. Environment, Artificial neural networks approach

for estimating the groutability of granular soils with cement-based grouts. 2011. 70(1): p. 153-161.

[5] Liao, K.-W., et al., An artificial neural network for groutability prediction of permeation grouting with microfine cement grouts. 2011. 38(8): p. 978-986.

[6] Cheng, M.-Y., N.-D.J.J.o.C.E. Hoang, and Management, Groutability prediction of microfine cement based soil improvement using evolutionary LS-SVM inference model. 2014. 20(6): p. 839-848.

[7] Tran, H.-H. and N.-D.J.J.o.C.E. Hoang, An artificial intelligence approach for groutability estimation based on autotuning support vector machine. 2014. 2014: p. 1-9.

[8] Cheng, M.-Y., N.-D.J.T. Hoang, and u.s. technology, A novel groutability estimation model for ground improvement projects in sandy silt soil based on Bayesian framework. 2014. 43: p. 453-458.

[9] Cheng, M.-Y. and N.-D.J.J.o.C.i.C.E. Hoang, Groutability estimation of grouting processes with microfine cements using an evolutionary instance-based learning approach. 2014. 28(4): p. 04014014.

[10] Hoang, N.-D., D.T. Bui, and K.-W.J.A.S.C. Liao, Groutability estimation of grouting processes with cement grouts using differential flower pollination optimized support vector machine. 2016. 45: p. 173-186.

[11] Asadizadeh, M., A.J.I.J.o.M. Majdi, and Geo-Engineering, Developing new Adaptive Neuro-Fuzzy Inference System models to predict granular soil groutability. 2019. 53(2): p. 133-142.

[12] Tekin, E., S.O.J.N.C. Akbas, and Applications, Predicting groutability of granular soils using adaptive neuro-fuzzy inference system. 2019. 31(4): p. 1091-1101.

[13] Deng, S., et al., Hybrid Grey Wolf Optimization Algorithm–Based Support Vector Machine for Groutability Prediction of Fractured Rock Mass. 2019. 33(2): p. 04018065.

[14] Mashrei, M.A.J.F.I.S.-T. and Applications, Neural network and adaptive neuro-fuzzy inference system applied to civil engineering problems. 2012.

[15] Bre, F., et al., Prediction of wind pressure coefficients on building surfaces using artificial neural networks. 2018. 158: p. 1429-1441.

[16] Suthaharan, S., Support vector machine, in Machine learning models and algorithms for big data classification. 2016, Springer. p. 207-235.

[17] Peterson, L.E.J.S., K-nearest neighbor. 2009. 4(2): p. 1883.

[18] Speiser, J.L., et al., A comparison of random forest variable selection methods for classification prediction modeling. 2019. 134: p. 93-101.

[19] Saritas, M.M., A.J.I.J.o.I.S. Yasar, and A.i. Engineering, Performance analysis of ANN and Naive Bayes classification algorithm for data classification. 2019. 7(2): p. 88-91.

[20] Stehman, S.V.J.R.s.o.E., Selecting and interpreting measures of thematic classification accuracy. 1997. 62(1): p. 77-89.

[21] Naudts, A., et al., Additives and admixtures in cement-based grouts, in Grouting and Ground Treatment. 2003. p. 1180-1191.

[22] Zebovitz, S., R.J. Krizek, and D.J.J.o.g.e. Atmatzidis, Injection of fine sands with very fine cement grout. 1989. 115(12): p. 1717-1733.

[23] Tekin, E.J.G.U., Experimental studies on the groutability of microfine cement (Rheocem 900) grouts to sands having various gradations. 2004.

[۱] Grouting
[۲] Grout
[۳] Liquefaction
[٤] Groutability Rate
[٥] Burwell
[٦] Radial basis Function Neural Network
[۷] Least Square Support Vector Machine
[۸] Evolutionary Fuzzy K-Nearest Neighbor Inference Model
[۹] Artificial Neural Network
[۱۰] Support Vector Machine
[۱۱] K-Nearest Neighbor
[۱۲] Random Forest
[۱۳] Naive Bayes
[۱٤] Confusion Matrix
[۱٥] Receiver Operating Characteristic
[۱٦] False Positive Rate
[۱۷] True Positive Rate
[۱۸] Area Under (ROC) Curve
[۱۹] Information Gain
[۲۰] Relative Information Gain
[۲۱] Gini Index