نشریه علمی ژئومکانیک نفت
JOURNAL OF PETROLEUM GEOMECHANICS
(JPG)

مقاله پژوهشی

کاربرد مدل‌های پیشرفته یادگیری ماشین جهت پیش‌بینی رخساره‌های سنگی در یکی از میدان‌های نفتی جنوب ایران

حمید قالیباف محمدآبادی[۱]؛ ناصر حافظی‌مقدس[۲*]؛ الهام مهدی‌پور[۳]؛ مجتبی حیدری‌زاد[۴]؛ حسین طالبی[۵]

۱- دکتری؛ دپارتمان هوش مصنوعی و رباتیکز، دانشکده مهندسی کامپیوتر، دانشگاه فردوسی مشهد

۲- استاد، گروه زمین‌شناسی مهندسی دانشگاه فردوسی مشهد

۳- استادیار، گروه مهندسی کامپیوتر، موسسه آموزش عالی خاوران، مشهد

۴- محقق ارشد، دانشکده زمین‌شناسی دریایی، دانشگاه تونگ جی، شانگهای چین

۵- دکتری؛ دپارتمان زمین‌شناسی گسترشی، شرکت ملی مناطق نفتخیز جنوب

دریافت مقاله: ۱۴۰۲/۰۷/۲۱ پذیرش مقاله: ۱۴۰۲/۰۸/۲۰

شناسه دیجیتال (DOI): 10.22107/JPG.2023.420597.1213

واژگان کلیدی

رخساره‌های سنگی، *Trees*، *Extra Classifier*، *F1-SCORE*، *Gradient Boosting*، *Random Forest*، **ماتریس درهم ریختگی**، **نمودارهای** *ROC*

چکیده

به‌طور مرسوم طبقه‌بندی رخساره‌های سنگی در میدان‌های نفتی بر اساس مطالعات رسوب‌شناسی، چینه‌شناسی، پتروفیزیکی، نفوذپذیری و ژئومکانیکی انجام می‌گیرد. زیرا دانشمندان علوم زمین تفسیرهای متفاوتی را از یک مجموعه داده ارائه می‌دهند. اگرچه مطالعات آزمایشگاهی حاصل از نمونه گمانه‌های اکتشافی که در حین حفاری به‌دست می‌آیند، نمایش دقیق‌تری از زمین‌شناسی زیرسطحی به دست می‌دهند. اما به دلیل هزینه‌هایشان، همیشه برای همه چاه‌های یک میدان در دسترس نیستند. بنابراین، یک روش سریع و کارآمد برای طبقه‌بندی دقیق رخساره‌های سنگی در تعداد زیادی از چاه‌های بدون نمونه گمانه حفاری با حداقل سوگیری ذهنی ضروری است. هدف از این مطالعه استفاده از یادگیری ماشینی نظارت شده برای طبقه‌بندی رخساره‌های سنگی از لاگ‌های ژئوفیزیکی در چاه‌های بدون نمونه حفاری می‌باشد. برای این منظور از مجموعه داده چاه‌نگاری ۷ چاه آموزشی یکی از میدان‌های نفتی جنوب ایران که شامل نگاره گاما طبیعی(*SGR*)، نگاره گاما اصلاح‌شده (*CGR*)، چگالی(*RHOB*)، تخلخل نوترونی(*NPHI*)، کندی موج‌برشی(*DTSM*) و کندی موج طولی(*DTCO*) که مستقیماً در تعیین رخساره‌های ژئومکانیکی تأثیر دارند به‌عنوان داده‌های مستقل و واحدهای طبقه‌بندی شده رخساره، به‌عنوان متغیر وابسته استفاده شده است. این مجموعه داده از عمق ۳۰۰۰ تا ۴۰۰۰ هزار متری زمین مربوط به سازندهای آهکی ایلام و سروک (آهک بنگستان) تشکیل شده است. در مرحله اول این سازندها به‌وسیله روش‌های خوشه‌بندی هوش مصنوعی و مطالعات آزمایشگاهی به ۵ رخساره تفکیک شده است. بعد از این مرحله از ۸ روش یادگیری ماشین نظارت شده شامل *Regression*، *Logistic*، *K Nearest Neighbors*، *Decision Tree*، *Random Forest*، *Gaussian NB*، *Gradient Boosting*، *Extra Trees* و *SVM* جهت ساخت یک مدل مناسب بکار گرفته شد. مجموعه داده این چاه‌ها به‌وسیله هر یک از این الگوریتم‌ها مراحل آموزشی و آزمایشی جهت ساخت یک مدل مناسب بکار گرفته شد و برچسب‌های رخساره‌ها پیش‌بینی شد. جهت ارزیابی عملکرد مدل‌ها از چندین معیار ارزیابی شامل *Accuracy*، *Precision*، *F1-SCORE* و *Recall* به‌وسیله ماتریس درهم‌ریختگی و نمودارهای *ROC* استفاده شده است. از بین روش‌های مذکور الگوریتم *Extra Trees Classifier*، *Gradient Boosting*، *K-Nearest Neighbors* نتایج بهتری را نشان داده‌اند. در نهایت، عملکرد مدل جهت پیش‌بینی رخساره‌های سنگی چاه خارج از مدل یا چاه دیده نشده ارائه شده است.

* مشهد؛ دانشگاه فردوسی مشهد، دانشکده علوم، کدپستی: ۳۶۱۹۹۹۵۱۶۱؛ شماره تلفن: ۰۵۱۳۸۸۰۶۰۰۰؛ nhafezi@um.ac.ir

۱. پیش‌گفتار

پیش‌بینی رخساره‌ها با استفاده از روش‌های یادگیری ماشین یکی از موضوعات پرطرفدار در زمینه‌های مختلف مهندسی و علوم زمین است. رخساره‌ها، الگوهای مختلفی از سنگ‌ها و رسوبات در لایه‌های زمین هستند که معمولاً نشان‌دهنده شرایط محیطی و فرآیندهای زمین‌شناسی در گذشته هستند. با توجه به اهمیت تشخیص و تفسیر رخساره‌ها در اکتشافات نفت و گاز، مطالعات پیش‌بینی رخساره‌ها با استفاده از روش‌های یادگیری ماشین می‌تواند به بهبود فرآیندهای اکتشافی و توسعه نفت و گاز کمک شایانی کند.

در سال‌های اخیر استفاده از روش‌های یادگیری ماشین جهت طبقه‌بندی واحدهای ژئومکانیکی یا رخساره‌ها با استفاده از نگاره‌های چاه‌نگاری تحقیقات قابل توجهی را به خود دیده است [۲-۱]. محققان بسیاری با استفاده از الگوریتم‌های هوش مصنوعی راهکاری جهت تخمین پارامترهای ناشناخته مانند رخساره‌ها سنگی ارائه داده‌اند [۳-۲]، الگوریتم‌های یادگیری ماشین از تجزیه و آنالیز داده‌ها به‌وسیله اصل "مشابهت" جهت تعیین رخساره‌های ژئومکانیکی استفاده می‌کند. این الگوریتم‌ها شامل مدل‌های طبقه‌بندی، رگرسیون و خوشه‌بندی می‌باشد [۱]. تکنیک *ML* به‌طورکلی به دو دسته تقسیم می‌شود. گروه یادگیری ماشین نظارت شده و یادگیری ماشین بدون نظارت، برای یادگیری ماشین تحت نظارت متغیرهای مستقل و وابسته (رخساره‌ها) مشخص می‌باشند.

به‌طورکلی ارزیابی مشکلات زمین‌شناسی در اکتشاف نفت و گاز، توسعه و چالش‌های تولید در حال حاضر به‌طور معمول توسط برنامه‌های یادگیری ماشین با استفاده از نگاره‌های چاه‌های نفتی مورد استفاده قرار می‌گیرد. دانشمندان نتایج بسیاری را با استفاده از الگوریتم‌های محاسباتی جهت تخمین پارامترهای ناشناخته مانند طبقه‌بندی رخساره‌ها ارائه داده‌اند [۳-۲]، گسل‌ها و شکستگی‌ها [۴]، ساخت یک مدل پتروفیزیکی برای ارزیابی مخزن ماسه‌سنگی [۵]، در تحقیق دیگری به‌وسیله رویکرد یادگیری ماشین نیمه نظارت‌شده رخساره‌های لرزه‌ای پیش‌بینی شده و از داده‌های آموزشی با بهره‌گیری از داده‌های بدون برچسب استفاده شده است [۶]. در مقاله دیگری توسعه یک مدل مؤثر بر اساس یادگیری عمیق برای طبقه‌بندی رخساره‌های زمین‌شناسی در چاه‌ها بکار گرفته شد که با استفاده از اثر فوتوالکتریک، پرتو گاما، ثبت مقاومت، تفاوت تخلخل چگالی نوترون، میانگین تخلخل چگالی نوترون و متغیرهای محدودکننده زمین‌شناسی به‌عنوان داده‌های ورودی مدل در نظر گرفته شدند. دقت قابل‌قبول و استفاده از داده‌های گزارش چاه معمولی از مزایای اصلی مدل پیشنهادی آن می‌باشد. مدل پیشنهادی با مدل شبکه عصبی بازگشتی، مدل حافظه کوتاه مدت و بلندمدت، مدل ماشین بردار پشتیبان و مدل k نزدیک‌ترین همسایه مقایسه شده و نتایج دقیق‌تری را در مقایسه با آن‌ها نشان داده است [۷].

در مطالعه دیگری ارزیابی تجزیه و تحلیل مؤلفه اصلی برای کاهش ابعاد ویژگی‌های لرزه‌ای مفهومی برای طبقه‌بندی رخساره‌های لرزه‌ای نظارت شده یک مخزن رودخانه از حوضه مالایی، فراساحل مالزی انجام گرفته است. این مطالعه به بررسی استفاده از *PCA* برای کاهش تعداد ویژگی‌ها قبل از یادگیری تحت نظارت می‌پردازد. هدف از این تحقیق به حداکثر رساندن استفاده از ویژگی‌های لرزه‌ای، تجزیه و تحلیل داده‌ها، و کاربرد *ML* برای طبقه‌بندی مؤثر رخساره‌های لرزه‌ای ژئومورفولوژیک مخزن *I-X* از میدان *A*، فراساحل مالزی بوده است [۸]. در تحقیق دیگری با استفاده از لاگ‌های تصویری *(Image log)* و اطلاعات آزمایشگاهی از هر سازند و استفاده از ترکیب الگوریتم‌های نظارت شده و بدون نظارت در شناسایی و پیش‌بینی رخساره‌های مختلف بکار گرفته شد [۹]. محمد و همکاران نیز با استفاده از این الگوریتم‌ها لاگ‌های صوتی برشی را برای عمق‌هایی که فاقد این اطلاعات بودند پیش‌بینی کردند [۱۰].

تیاگو و همکاران با بررسی روش‌های طبقه‌بندی در یادگیری ماشین بهترین روش جهت شناخت لایه‌های سنگ‌شناسی را روش جنگل تصادفی*(Random forest method)* پیشنهاد داده‌اند [۱۱]. سانگ و همکاران روش خوشه‌بندی *K-means* را جهت تعیین رخساره‌های لرزه‌ای ارائه داده‌اند .[۱۲] دانهام و همکاران جهت بهبود تعیین رخساره‌ها با توجه به کمبود داده‌های آموزشی جهت تعیین متغیر وابسته (رخساره) فرا پارامترها در یک مدل مخلوط گوسی *(Gaussian mixture models)* را مطرح کرده‌اند [۱۳]. در تحقیق دیگری با توجه به اینکه تفسیر زمین‌شناسی و ژئوفیزیک با مجموعه داده‌های بزرگ بسیار گران است، از یادگیری نیمه نظارت شده عمیق با استفاده از شبکه‌های مولد تخاصمی *(Generative*

(Adversarial Network یا به‌اختصار *GAN* برای کاهش ابعاد داده جهت طبقه‌بندی رخساره‌های لرزه‌ای استفاده شده است [۱۴]. در مقاله دیگری به‌وسیله یک مدل شبکه عصبی بیزی *(BNN)* برای پیش‌بینی ضخامت مخزن و کمی کردن عدم قطعیت پیشنهاد شده است [۱۵]. فنگ و همکاران با استفاده از میدان‌های تصادفی مارکوف (*MRFs*) بر اساس مدل مخلوط گوسی جهت سنگ‌شناسی مخازن نفتی با توجه به تغییرات افقی و عمودی مخزن ارائه داده‌اند. با توجه به اینکه در تعیین رخساره با استفاده از روش‌های نظارت شده چنانچه برچسب-گذاری واحدها به‌صورت خیلی دقیق نباشد مدل ارائه‌شده کارایی لازم را ندارد؛ مارکو ایپولیتو و همکاران ترکیب روش‌های نظارت شده با بدون نظارت جهت بهبود پیش‌بینی رخساره‌ها معرفی کرده‌اند [۱۶]. از روش‌های بدون نظارت در تهیه نقشه‌های پهنه‌بندی با استفاده از پارامترهای ژئوتکنیکی جهت بیشترین محدوده تزریق سیمان برای آب‌بندی پرده آب‌بند سد سرورد با استفاده از مدل سلسله مراتبی *(Analytic Hierarchy process, AHP)* بکار گرفته شد [۱۷]. تحقیق دیگری با استفاده از طیف وسیعی از روش‌های بدون نظارت جهت تعیین واحدهای ژئومکانیکی با استفاده از داده‌های چاه‌نگاری در یکی از میدان‌های نفتی جنوب ایران ارائه شده است [۱۸].

در مطالعه حاضر، تشخیص رخساره سنگ‌شناسی با استفاده از چاه‌های مدل بر روی کل مجموعه داده به‌وسیله طیف وسیعی از روش‌های پیشرفته طبقه‌بندی نظارت شده بهبود یافته است و بهترین الگوریتم ارائه شده است. در این تحقیق، مراحل استاندارد و جامعی برای انتخاب بهترین مدل و فراپارامترها برای پیش‌بینی رخساره‌های سنگی به‌صورت مجموعه داده اجرا شده است. ابتدا داده‌ها برای مدل‌سازی آماده شد، مدل‌ها برازش شد و به‌وسیله اعتبارسنجی متقاطع، مدل صحت‌سنجی شده است، برچسب‌های رخساره پیش‌بینی شد و دقت مدل پیش‌بینی با چندین معیار ارزیابی شامل *Accuracy*، *Precision*، *F1-SCORE*، *Recall* و نمودارهای *ROC* ارزیابی شده است. در نهایت، عملکرد مدل جهت تعیین رخساره چاه پیش‌بینی[۱] بررسی شد.

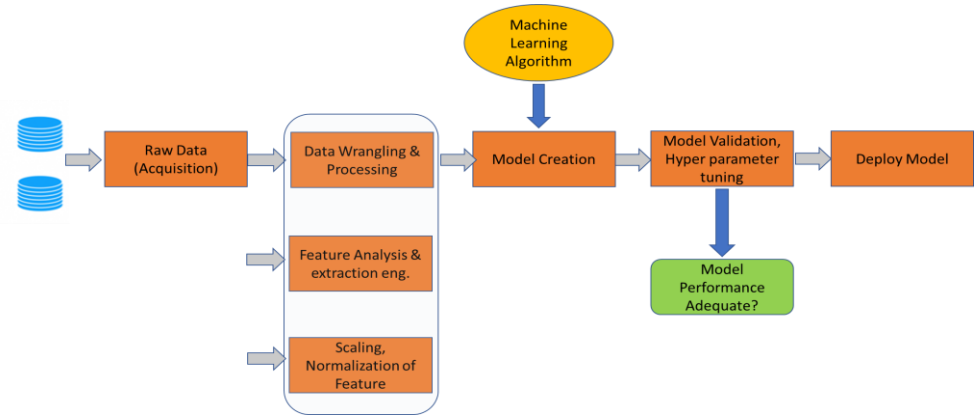
۲. مهندسی ویژگی‌ها جهت ساخت یک مدل

الگوریتم یادگیری ماشین

مهندسی ویژگی‌ها *(Feature engineering)* بخش بزرگی از یادگیری ماشین *(ML)* و یادگیری عمیق است. هر سیستم هوشمند، صرف‌نظر از پیچیدگی آن، باید بر اساس داده باشد. در قلب هر سیستم هوشمند، ما یک یا چند الگوریتم بینش داده‌ای را بر اساس مجموعه‌ای از داده‌های یادگیری، مانند یادگیری ماشین، یادگیری عمیق و یا روش‌های آماری استفاده می‌شود که این اطلاعات را برای جمع‌آوری دانش و ارائه بینش هوشمند بیش از یک دوره زمانی نیاز داریم. الگوریتم‌ها خودشان کاملاً مجزا کار می‌کنند و نمی‌توانند خارج از جعبه داده‌های خام که برای آن‌ها مشخص شده است کار کنند.

هر سیستم بینش اطلاعاتی هوشمند، اساساً شامل یک خط یا نقطه سربه‌سر با استفاده از داده‌های خام برای استفاده از تکنیک‌های پردازش داده‌ها جهت گردآوری، پردازش و خواص ویژگی‌های مهندسی از این داده‌ها است. معمولاً تکنیک‌هایی مانند مدل‌های آماری یا مدل‌های یادگیری ماشین برای مدل‌سازی بر روی این ویژگی‌ها استفاده می‌شود و در صورت لزوم برای استفاده آن‌ها در آینده بر اساس مشکلاتی که می‌توان به آن‌ها اشاره کرد به‌صورت دستی حل می‌شوند. یک سامانه یادگیری ماشین مبتنی بر فرایندهای استاندارد صنعت متقابل که به‌طور معمول برای داده‌کاوی استفاده می‌شود در شکل ۱ نشان داده شده است. مهندسی ویژگی فرایند تبدیل داده‌های خام به ویژگی‌هایی است که مشکلات پیش‌بینی شده مدل‌های اصلی را بهتر نشان می‌دهد، درنتیجه دقت مدل را در داده‌های غیرقابل مشاهده بهبود می‌بخشد. این به ما درک این را می‌دهد که مهندسی ویژگی لازم است تا با استفاده از فرایند تبدیل اطلاعات (داده‌ها) بتواند از یک ویژگی به‌عنوان ورودی برای مدل‌های یادگیری ماشین استفاده کند. یک ویژگی، به‌طور معمول، یک نمایش خاص در رأس داده‌های خام است که خصوصیات قابل اندازه‌گیری آن به‌صورت منحصربه‌فرد (خصوصی) است. که معمولاً در یک ستون از یک مجموعه داده نقش بسته‌اند. با توجه به مجموعه‌ای از داده‌های دوبعدی، هر مشاهده توسط یک ردیف و هر ویژگی توسط یک ستون نشان داده می‌شود که یک مقدار خاص برای مشاهده دارد.

[۱] Blind well



شکل ۱. مراحل الگوریتم مهندسی ویژگی‌ها جهت ساخت یک مدل ماشین لرنینگ

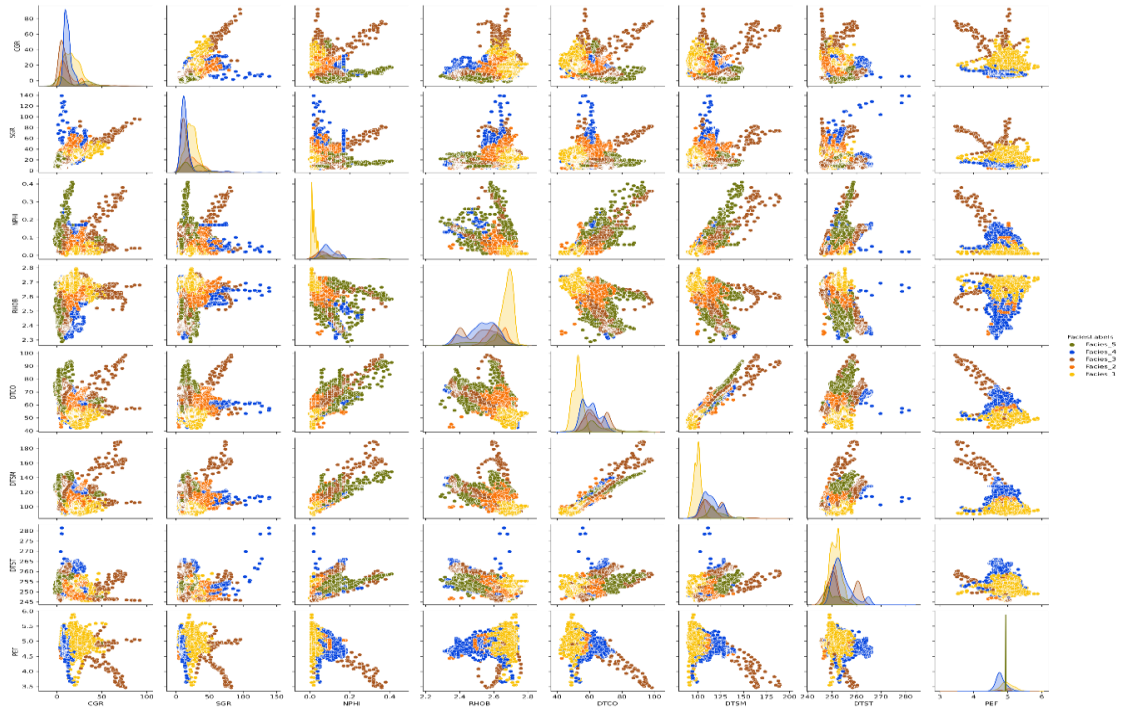
بنابراین، درباره نگاره‌های چاه‌های نفتی، هر سطر به‌طور خاص یک ویژگی از بردار را نشان می‌دهد و همه آن‌ها مجموعه‌ای از ویژگی‌ها در همه مشاهدات به شمار می‌آیند، همچنین یک ماتریس ویژگی دوبعدی است، که به‌عنوان مجموعه‌ای از ویژگی‌ها شناخته می‌شود. این شبیه به قاب داده‌ها یا صفحات گسترده‌ای است که داده‌های دوبعدی را نشان می‌دهند. به‌طور معمول، الگوریتم‌های یادگیری ماشین با این ماتریس‌های عددی یا تانسورها کار می‌کنند. ازاین‌رو بیشترین کاربرد مهندسی ویژگی‌ها، تبدیل داده‌های خام به‌عنوان نماینده‌ای از داده‌هایی که می‌توانند توسط این الگوریتم‌ها قابل فهم و درک باشند است. ویژگی‌ها می‌توانند از دو نوع اصلی بر اساس مجموعه داده‌ها باشند. ویژگی‌های خام (خالص) ذاتی مستقیماً از مجموعه داده‌ها و بدون دست‌کاری اطلاعات و یا مهندسی اضافی به دست می‌آیند. ویژگی‌های مشتق شده معمولاً از ویژگی‌های مهندسی به دست می‌آیند، جایی که ویژگی‌های داده‌های موجود از آن استخراج می‌شود.

۳. پیش‌پردازش داده‌های چاه‌نگاری

۱،۳. تجزیه و تحلیل اکتشافی داده‌ها[۲]

در مطالعه حاضر از ۷ داده چاه‌نگاری یکی از میدان‌های نفتی جنوب ایران شامل نگاره گاما طبیعی *(SGR)*، نگاره گاما اصلاح‌شده*(CGR)*، چگالی*(RHOB)*، تخلخل نوترونی *(NPHI)*، کندی موج برشی*(DTSM)* و کندی موج طولی *(DTCO)* که مستقیماً در تعیین رخساره‌های ژئومکانیکی تأثیر دارند استفاده شده است. ابتدا فایل داده‌های رقومی*(Las)* در پایتون فراخوانی شد و یک دیتا فریم*(Data Frame)* ساخته شد. بعد از این مرحله تمام مراحل پیش‌پردازش *(Data Preprocessing)* جهت ساخت مدل به‌درستی صورت پذیرفت. رخساره‌های سنگی در هر چاه با استفاده از روش‌های بدون نظارت و مطالعات آزمایشگاهی نمونه‌های سنگی واقعی چاه صحت‌سنجی شد. سپس نام رخساره‌ها تعیین شد. اما با توجه به اینکه این واحدهای سنگی بر اساس نگاره‌های تخلخل، کندی موج‌های صوتی، چگالی و دیگر نگاره‌ها استفاده شده است؛ غیرمعقول است نام خاصی گرفته شود، زیرا هر سنگی می‌تواند این خصوصیات را داشته باشد. بنابراین در این تحقیق رخساره‌های سنگی به نام رخساره با پسوند اعداد نام‌گذاری شده است. در این تحقیق جهت ارزیابی و تجسم داده‌ها[۳] به‌وسیله نمودار هیستوگرام[۴]، ستونی[۵] و متقاطع/ پراکنده[۶] استفاده شده است. در این نمودار پراکنده/متقاطع کل مجموعه داده‌ها (چاه‌ها) به‌وسیله متغیرهای مستقل با برچسب‌گذاری متغیر وابسته (رخساره‌ها) نشان داده می‌شود (شکل ۲).

[۲] Data Exploratory Analysis
[3] Data visualization
[4] Bar plot
[5] log-plot
[6] Cross-plot

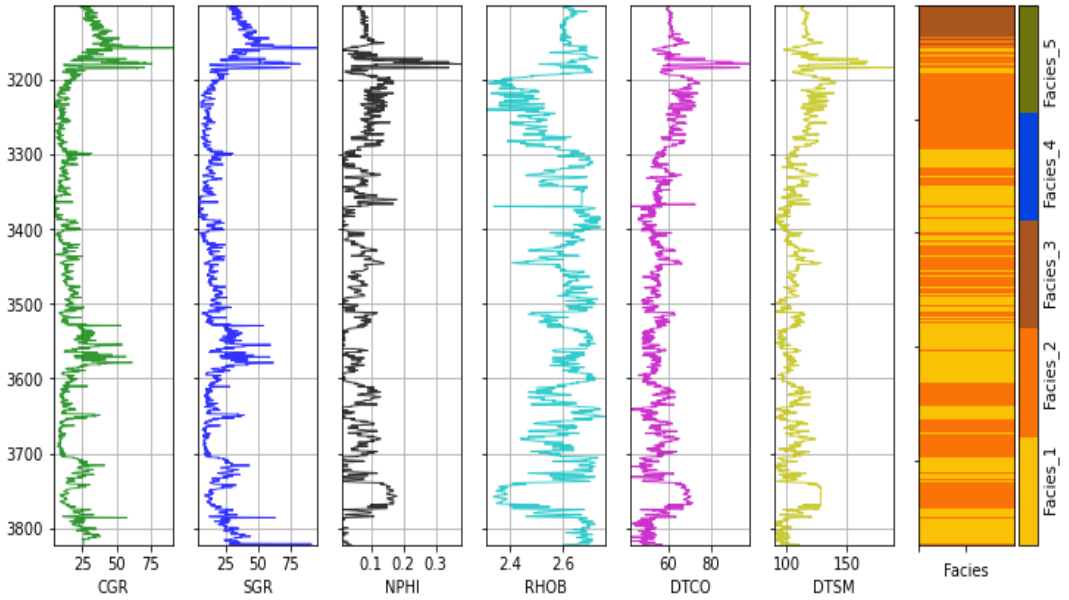


شکل ۲. در این نمودار پراکندگی کل مجموعه داده‌ها (چاه‌ها) به‌وسیله متغیرهای مستقل با برچسب‌گذاری متغیر وابسته (رخساره‌ها) نشان داده شده است.

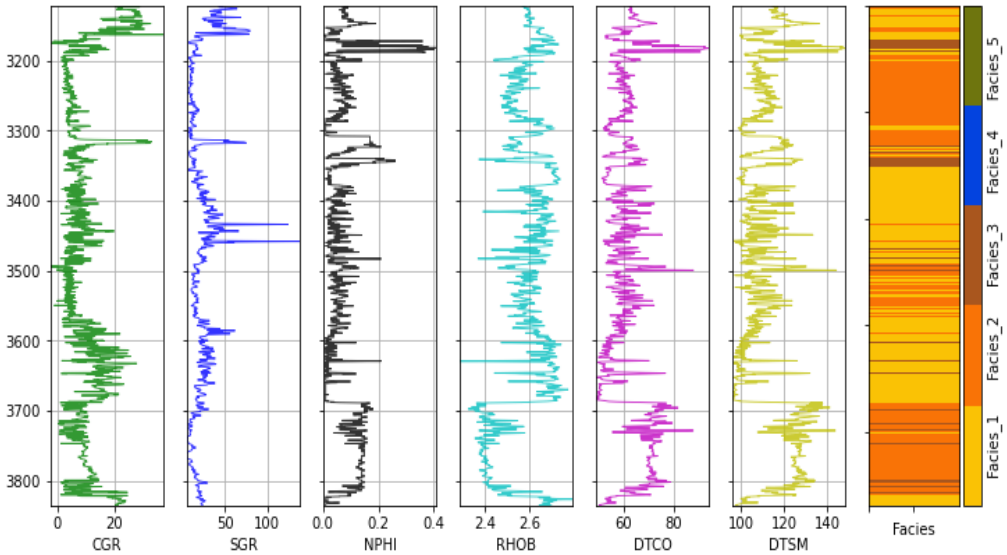
به‌وسیله نمودارهای ستونی متغیرهای مستقل/ ویژگی‌ها و متغیر وابسته (رخساره) از دو چاه جهت ساخت یک مدل مناسب یادگیری ماشین به‌وسیله شکل ۳ و ۴ نشان داده می‌شود.

در نمودار هیستوگرام[۷] درصد و فراوانی رخساره‌های سنگی در کل چاه‌های آموزشی این میدان به‌وسیله نمودار هیستوگرام شکل ۵ نشان داده شده است.

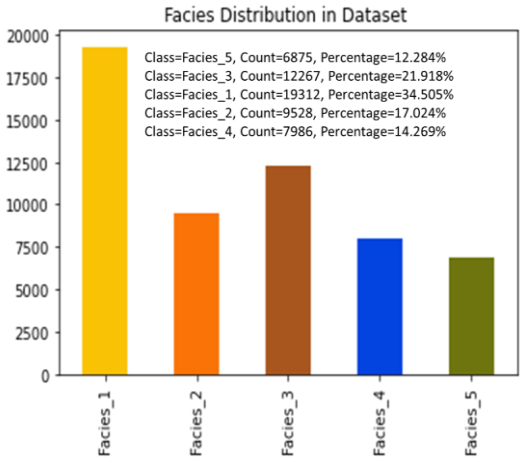
[۷] Bar plot



شکل ۳. نمودار ستونی متغیرهای مستقل (نگاره‌ها) و متغیر وابسته (رخساره) یکی از چاه‌های مدل



شکل ۴. نمودار ستونی متغیرهای مستقل (نگاره‌ها) و متغیر وابسته (رخساره) یکی از چاه‌های مدل



شکل ۵. درصد و فراوانی واحدهای ژئومکانیکی(رخساره‌ها) در چاه‌های آموزشی این میدان

۲،۳. نسبت و استخراج داده گم‌شده

با توجه به اینکه، الگوریتم‌های یادگیری ماشین با داده‌های گمشده در یک ستون از نگاره ناسازگار است. لازم است این داده‌ها استخراج گردد. به‌طور معمول جهت بازیابی داده‌های گمشده از روش‌های حذف ردیف داده، میانگین‌گیری، ماکزیمم و مینیمم یک ستون داده، و عدد ثابت به کار گرفته می‌شود. اما این روش‌ها جهت داده‌های زمین‌شناسی کاربرد مناسبی ندارد چون در داده‌های زمین‌شناسی هدف شناسایی خصوصیات واقعی است. بنابراین، جهت پیش‌بینی این داده‌ها از طیف وسیعی از روش‌های یادگیری ماشین استفاده شده است.

راه‌های مختلفی برای مقابله با مقادیر *Null* در مجموعه داده وجود دارد. ساده‌ترین روش حذف ردیف‌هایی است که حداقل یک مقدار تهی دارند. این می‌تواند با مجموعه داده‌های با اندازه بزرگ‌تر منطقی باشد، اما با قاب‌های داده با اندازه کوچک، هر ویژگی در هر ردیف مهم هستند. مقادیر تهی را می‌توان با میانگین یا استفاده از نقاط داده مجاور در ستون‌ها به دست آورد. پر کردن با مقدار میانگین بر واریانس داده‌ها تأثیر نمی‌گذارد و بنابراین تأثیری بر دقت پیش‌بینی نخواهد داشت، اگرچه می‌تواند سوگیری داده ایجاد کند. رویکرد دیگری که در این تحقیق پیاده‌سازی شده است، استفاده از مدل‌های یادگیری ماشین برای پیش‌بینی مقادیر داده گمشده است. این بهترین راه برای مقابله با این مجموعه داده است، زیرا فقط یک ویژگی در مجموعه داده مورد بررسی قرار نگرفته است، در این مطالعه متغیر مستقل نگاره کندی موج برشی *(DTSM)* به‌وسیله طیف وسیعی از روش‌های هوش مصنوعی در چاه‌های فاقد این نگاره به دست آمده است [۱۹].

۳،۳. افزایش تعداد داده کلیدی[۸]

داده‌های نامتوازن یکی از مشکلات موجود در طبقه‌بندی داده‌ها می‌باشد. داده‌های نامتوازن داده‌هایی هستند که نسبت کلاس‌ها در آن بسیار متفاوت با هم هستند. اگر ۹۰ درصد داده‌ها مربوط به یک کلاس و ۱۰درصد داده‌ها مربوط به کلاس دیگر (کلاس غالب یا اکثریت) باشد آنگاه داده‌ها نامتوازن هستند. در یادگیری ماشین نمونه‌گیری *Undersampling* و نمونه‌گیری *Oversampling* دو روش هستند که با برخورد با داده‌های نامتوازن به کار می‌روند. می‌توان از کلاس دارای اکثریت نمونه، کمتر نمونه‌گیری انجام داد یا از کلاس دارای اقلیت نمونه، بیشتر نمونه‌گیری انجام داد، یا از ترکیب هر دو روش استفاده نمود. دلیل اینکه طبقه-بندی نامتوازن به‌عنوان یک مشکل شناسایی می‌شود، این است که می‌تواند بر عملکرد الگوریتم‌های یادگیری ماشین تأثیر بگذارد. در مجموعه داده‌های نامتعادل، از روش نمونه‌گیری مجدد برای اضافه کردن چند نقطه داده بیشتر برای افزایش اعضای گروه‌های اقلیت استفاده شده است. این کار می‌تواند مفید باشد. دو روش برای مقابله با داده‌های نامتوازن وجود دارد: بیش نمونه‌گیری *(Oversampling)*؛ کم نمونه‌گیری *(Under sampling)*؛ در نمونه‌گیری *Oversampling* سعی می‌شود از کلاس اقلیت نمونه‌های بیشتر ایجاد شود تا نسبت کلاس‌ها به هم نزدیک شود. همچنین در *Under sampling* سعی می‌شود از کلاس حداکثر کمتر نمونه‌گیری کنیم. در واقع در این روش ما از همه نمونه‌ها در کلاس بیشتر استفاده نمی‌کنیم تا نسبت کلاس‌ها به یکدیگر نزدیک شود.

روش نمونه‌برداری بیش از حد اقلیت مصنوعی *(SMOTE)* این تکنیک برای انتخاب نزدیک‌ترین همسایگان در فضای ویژگی، جداسازی نمونه‌ها با افزودن یک خط و تولید نمونه‌های جدید در طول خط استفاده می‌شود. این روش صرفاً تکرارها را از

[۸] Oversampling

کلاس بیش‌تر تولید نمی‌کند، بلکه از نزدیک‌ترین همسایگان K برای تولید داده‌های مصنوعی استفاده می‌کند.

۴،۳. کاربرد تحلیل مولفه اساسی *(PCA)* و *(LDA)*

اندازه داده‌ها در عصر مدرن، نه فقط یک چالش برای سخت‌افزارهای کامپیوتر، بلکه تنگنایی برای کارایی بسیاری از الگوریتم‌های یادگیری ماشین *(Machine Learning)* محسوب می‌شود. هدف اصلی تحلیل‌های *PCA* شناسایی الگوهای موجود در داده‌ها است؛ تحلیل مولفه اساسی *(PCA)* قصد دارد همبستگی بین متغیرها را شناسایی کند. اگر یک همبستگی قوی بین متغیرها وجود داشت، تلاش‌ها برای کاهش ابعاد معنادار خواهد بود. به‌طور کل، آنچه در *PCA* به وقوع می‌پیوندد پیدا کردن جهت واریانس بیشینه در داده‌های ابعاد بالا و طرح‌ریزی کردن آن در زیرفضایی با ابعاد کمتر به‌طوری است که بیشترین اطلاعات حفظ شوند. هم تحلیل تشخیصی خطی *(Linear Discriminant Analysis (LDA))* و هم تحلیل مولفه اصلی *(PCA)* از جمله روش‌های تبدیل خطی هستند *PCA*، جهت‌هایی که واریانس داده‌ها را بیشینه می‌کنند (مولفه اصلی) می‌یابد، درحالی‌که هدف *LDA* پیدا کردن جهت‌هایی است که جداسازی (یا تمایز) بین دسته‌های گوناگون را بیشینه می‌کند و می‌تواند در مسائل دسته‌بندی الگو *PCA* از برچسب کلاس صرف‌نظر کند. به‌بیان‌دیگر، *PCA* کل مجموعه داده را در یک (زیر)فضای دیگر طرح‌ریزی می‌کند و *LDA* تلاش می‌کند تا یک ویژگی مناسب را به منظور ایجاد تمایز بین الگوهایی که به دسته‌های مختلف تعلق دارند تعیین کند. اغلب، هدف مورد انتظار کاهش ابعاد یک مجموعه داده d بعدی با طرح‌ریزی آن در یک زیرفضای k بعدی (که در آن $k < d$ است) به‌منظور افزایش بازدهی محاسباتی به صورتی است که بخش مهم اطلاعات باقی بماند. پرسش مهمی که در این وهله مطرح می‌شود آن است که سایز K که داده‌ها را به‌خوبی نشان می‌دهد چند است.

در ادامه، بردارهای ویژه *(Eigenvectors)* (مولفه‌های اصلی) برای مجموعه داده محاسبه و همه آن‌ها در یک ماتریس تصویر *(Projection Matrix)* گردآوری می‌شوند. به هر یک از این بردارهای ویژه یک مقدار ویژه تخصیص داده می‌شود که می‌تواند به‌عنوان طول یا بزرگنمایی بردار ویژه متناظر در نظر گرفته شود. اگر برخی از مقدارهای ویژه دارای بزرگنمایی به‌طور قابل‌توجهی بزرگ‌تر از دیگر موارد باشند، کاهش مجموعه داده با تحلیل مولفه اصلی *(PCA)* به یک زیرفضای ابعاد کوچک‌تر با حذف جفت ویژه‌هایی با اطلاعات کمتر معقول است. به‌طور خلاصه استفاده از رویکرد *PCA* شامل موارد زیر می‌باشد: استانداردسازی داده‌ها؛ به دست آوردن بردارهای ویژه و مقدارهای ویژه از ماتریس کواریانس *(Covariance matrix)* یا ماتریس همبستگی *(Correlation Matrix)*؛ یا انجام تجزیه مقدارهای منفرد *(Singular Vector Decomposition)*؛ مرتب‌سازی مقدارهای ویژه به ترتیب نزولی و انتخاب k بردار ویژه‌ای که متناظر با K بزرگ‌ترین مقدار ویژه هستند K تعداد ابعاد زیرفضای ویژگی جدید است.*(k≤d)*؛ ساخت ماتریس تصویر W از K بردار ویژه انتخاب شده؛ تبدیل مجموعه داده اصلی X به‌وسیله W، برای به دست آوردن زیرفضای K بعدی Y؛ خوشبختانه یک تکنیک فاکتورگیری استاندارد از ماتریس وجود دارد به نام *Singular Value Decomposition (SVD)* که می‌تواند ماتریس مجموعه‌ی آموزشی X را تجزیه کند. انتخاب ویژگی *(Feature Selection)*، یکی از مفاهیم کلیدی در یادگیری ماشین است. روش‌های انتخاب ویژگی نقش مهمی در عملکرد بهینه مدل‌های یادگیری دارند. روش حذف بازگشتی ویژگی *(Recursive Feature Elimination)*، یک روش حریصانه *(Greedy)* برای انتخاب ویژگی است. در این روش، ویژگی‌ها به‌طور بازگشتی و با در نظر گرفتن مجموعه‌های کوچک و کوچک‌تر از ویژگی‌ها (در هر مرحله) انتخاب می‌شوند. در این روش، ویژگی‌ها بر اساس مرتبه حذف شدن آن‌ها از فضای ویژگی رتبه‌بندی می‌شوند. در زبان برنامه‌نویسی پایتون، از بسته نرم‌افزاری *scikit-learn* برای حذف بازگشتی ویژگی‌های نامرتبط و انتخاب بهترین ویژگی‌ها استفاده می‌شود. در این تحقیق از روش‌های انتخاب ویژگی برای انتخاب بهترین ویژگی‌های ممکن جهت دسته‌بندی ارقام *(Digit Classification)* استفاده شده است [۱۹].

۵،۳. آماده‌سازی چاه پیش‌بینی [۹]

به‌طورکلی بعد از آموزش و تست مدل به‌وسیله مجموعه داده،

[۹] Blind well preparation

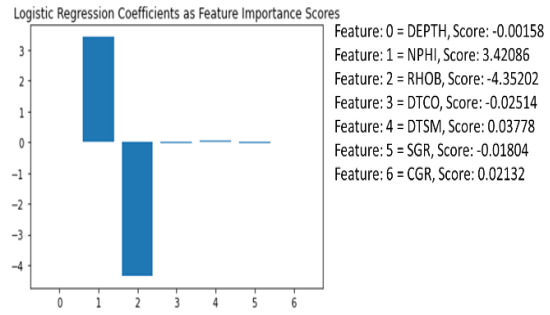
لازم است یک چاه که در آموزش نبوده است جهت پیش‌بینی و کارایی مدل مورد آزمایش قرار گیرد. به‌طورکلی الگوریتم یادگیری ماشین این داده‌ها را در فرآیند آموزش نمی‌بینند. بنابراین از این چاه جهت کارایی و پیش‌بینی رخساره‌ها استفاده می‌شود. همچنین بسیار مهم است چاه انتخاب شده شامل انواع واحدهای ژئومکانیکی (رخساره‌ها) باشد، در غیر این صورت در پیش‌بینی، ناسازگاری ابعاد داده‌ها باعث ایجاد خطا می‌شود.

۴. اهمیت ویژگی‌ها

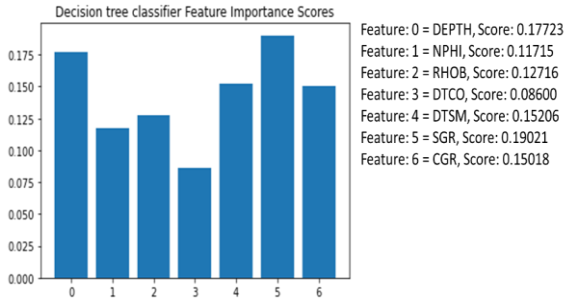
برخی از الگوریتم‌های یادگیری ماشین امتیاز اهمیت را برای کمک به کاربر در انتخاب کارآمدترین ویژگی‌ها برای پیش‌بینی ارائه می‌کنند. همان‌طور که مشخص است در الگوریتم رگرسیون لاجستیک متغیر تخلخل و چگالی رابطه معنی‌داری با هم دارند (شکل ۶). الگوریتم درخت تصمیم امتیازهای اهمیت را بر اساس کاهش معیار مورد استفاده برای تقسیم در هر گره مانند آنتروپی یا جینی ارائه می‌دهد (شکل ۷). *XGBoost* کتابخانه‌ای است که اجرای کارآمد و مؤثر الگوریتم تقویتی تصادفی[۱۰] را ارائه می‌دهد. این الگوریتم می‌تواند با استفاده از بسته *scikit-learn* از طریق کلاس‌های *XGB Regressor* و *XGB Classifier* پیاده‌سازی گردد (شکل ۸).

۵. ساخت مدل و اعتبارسنجی[۱۱]

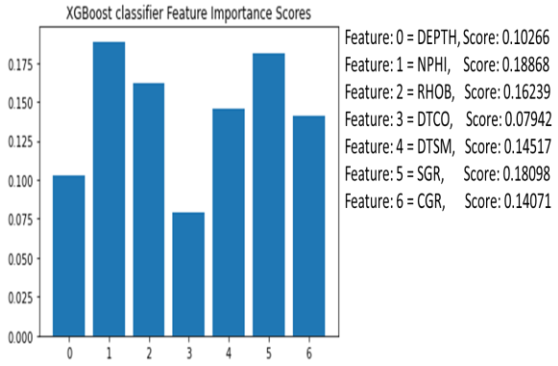
جهت پیش‌بینی رخساره‌های سنگی در چاه‌های فاقد نمونه حفاری و آزمایشگاهی ابتدا باید یک مدل تعریف شود و سپس با ویژگی استخراج شده مقایسه گردد. در این تحقیق عملکرد مدل به‌وسیله اعتبارسنجی متقاطع بررسی شده است. داده‌ها به ۱۰ زیرگروه تقسیم می‌شوند و فرآیند ۳ بار تکرار می‌شود. در اینجا، می‌توان بررسی کرد که آیا استخراج ویژگی می‌تواند عملکرد مدل را بهبود بخشد. رویکردهای زیادی وجود دارد، درحالی‌که در اینجا از برخی تبدیل‌ها برای زنجیره‌بندی توزیع متغیرهای ورودی مانند *Quantile Transformer* و *KBins Discretizer* استفاده شده است. سپس، وابستگی‌های خطی بین متغیرهای ورودی با استفاده از *PCA* و *Truncated SVD* شناسایی شده است. با استفاده از کلاس واحد ویژگی، لیستی از تبدیل‌ها برای انجام نتایج تجمیع شده تعریف شده است. این یک مجموعه داده با تعداد زیادی از ستون‌های ویژگی ایجاد می‌کند درحالی‌که باید ابعاد برای عملکرد سریع‌تر و بهتر کاهش یابند. در نهایت، تکنیک *Recursive Feature Elimination* یا *RFE* برای انتخاب مرتبط‌ترین ویژگی‌ها استفاده شده است.



شکل ۶. اهمیت متغیرهای مستقل (نگاره‌ها) در مدل رگرسیون لاجستیک



شکل ۷. اهمیت متغیرهای مستقل (نگاره‌ها) در مدل درخت تصمیم

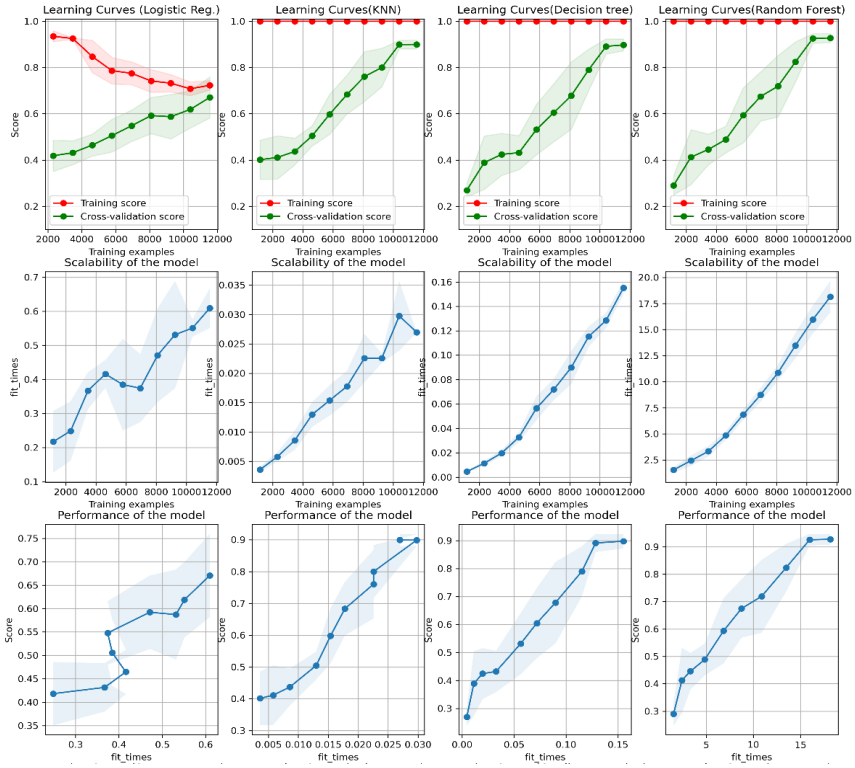


شکل ۸. اهمیت متغیرهای مستقل (نگاره‌ها) در مدل گرادیان تقویتی تصادفی

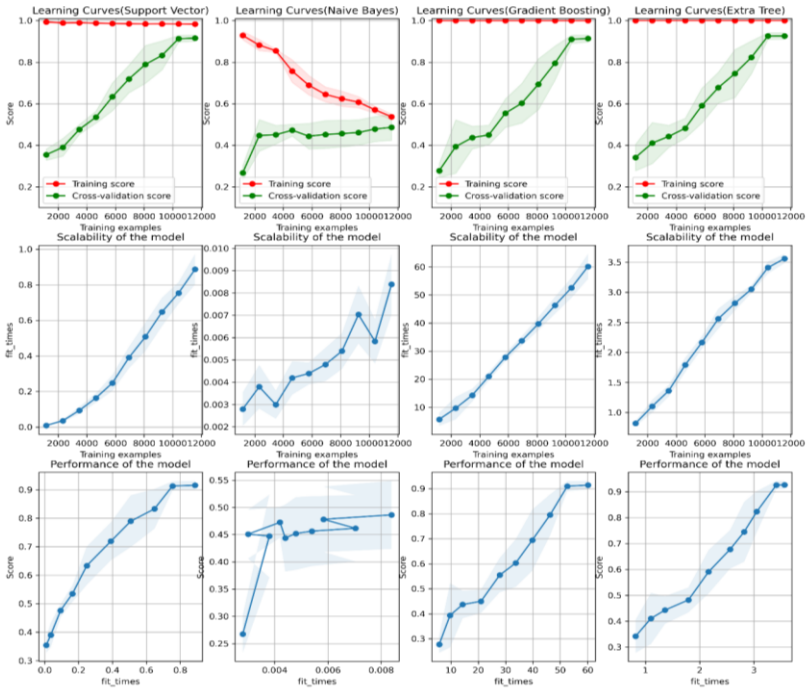
[۱۰] XGBoost

[۱۱] Build Model & Validate

در این تحقیق برای همه الگوریتم‌ها، دیده می‌شود که نمرات آموزشی همیشه بالاتر از نمونه‌های آزمایشی یا نمرات اعتبارسنجی متقابل است این تقریباً استاندارد *ML* است. در شکل ۹ و ۱۰ برای رگرسیون لجستیک، و *Naive Bayes* یک الگوی خاص دیده می‌شود. امتیاز داده *Traning* با افزایش نمونه‌ها در مجموعه داده آموزشی کاهش می‌یابد و امتیاز اعتبارسنجی ابتدا کاهش یا سوگیری زیاد (*high bias*) ولی بعد افزایش می‌یابد. برای بقیه الگوریتم‌های طبقه‌بندی‌کننده، امتیاز آموزشی هنوز در حدود حداکثر است و اعتبارسنجی را می‌توان با نمونه‌های داده جدید بیشتر افزایش داد. در این نمودارهای شماتیک، نقطه عطف منحنی اعتبارسنجی در تعدادی از روش‌ها دیده می‌شود. از جمله این روش‌های طبقه‌بندی (مانند الگوریتم جنگل تصادفی، درختان اضافی) بالاترین عملکرد را نشان می‌دهند. زیرا این منحنی‌های اعتبارسنجی در پایان فرآیند آموزش مماس شده‌اند. نمودارهای ردیف دوم زمان‌های مورد نیاز مدل‌ها را برای آموزش با اندازه‌های مختلف مجموعه داده‌های آموزشی نشان می‌دهند. همان‌طور که مشاهده می‌شود، مدل برای داده منحنی قرمز یا آموزشی خیلی خوب عمل کرده است که این استاندارد روش‌های *ML* است. برای مدل منحنی سبز یا اعتبارسنجی ابتدا پایین است به معنی دیگر مدل، بیش‌برازش(*Overfitting*) شده یا به عبارتی واریانس بالا می‌باشد یعنی اگر داده از بیرون در داخل داده‌های آموزشی قرار گیرد داخل مدل قرار نمی‌گیرد. به عبارتی تعداد متغیرهای مستقل (نگاره‌ها) یا تعداد داده‌های آموزش کم است. این مشکل با افزایش تعداد داده‌ها به نقطه عطف خود نزدیک می‌گردد.



شکل ۹. همان‌طور که مشاهده می‌شود مدل برای داده‌ی منحنی قرمز یا آموزشی خیلی خوب عمل کرده است که این استاندارد روش‌های *ML* است. در این شکل در مدل *Logistic Regression* امتیاز اعتبارسنجی بعد از افزایش داده‌ها پایین‌تر از بقیه است که مدل ارائه شده دارای بایاس بالا یا *Underfitting* است یعنی مدل ارائه شده دقت پایین‌تری دارد و داده‌ها به‌راحتی در مدل قرار می‌گیرند. درنتیجه، نتایج پیش‌بینی رخساره سنگی به‌وسیله این الگوریتم مناسب نمی‌باشد.



شکل ۱۰. در این تصویر مدل برای داده‌ی منحنی قرمز یا آموزشی به‌وسیله الگوریتم‌های *Extra tree*، *Gradient Boosting* و *SVM*خوب عمل کرده است. در این شکل مدل *Navise Bayes* امتیاز اعتبارسنجی بعد از افزایش داده‌ها پایین‌تر از بقیه است. یعنی مدل ارائه شده دقت پایین‌تری دارد و داده‌ها به‌راحتی در مدل قرار می‌گیرند.

در شکل ۹ مدل برای*Logistic Regression* و در شکل ۱۰ مدل برای *Navise Bayes* نشان داده شده است. امتیاز اعتبارسنجی بعد از افزایش داده‌ها پایین‌تر از بقیه است که مدل ارائه شده بایاس بالا یا *Underfitting* را تجربه کرده است یعنی مدل ارائه شده دقت پایین‌تری دارد و داده‌ها به‌راحتی در مدل قرار می‌گیرند. نمودارهای ردیف سوم نشان می‌دهد که برای آموزش مدل‌ها با هر اندازه داده و افزایش مقدار نمره چقدر زمان لازم است [۲۲].

۶. نتایج مدل پایه و بهبود مدل به‌وسیله فراپارامترها[۱۲] در میدان مورد مطالعه

فلسفه ساخت یک مدل پایه ساده است ولی با تغییرات تنظیمات بر روی پارامترهای داده و مدل می‌توان عملکرد مدل را افزایش داد. در این تحقیق از هایپرپارامترها که باعث بهبود عملکرد الگوریتم می‌شود استفاده شده است که یکی از وظایف چالش‌برانگیز در آموزش مدل‌های یادگیری ماشین است. به‌طورکلی سه راه برای بهبود مدل‌های یادگیری ماشین وجود دارد: استفاده از داده‌های بیشتر و مهندسی ویژگی‌ها؛ استفاده از الگوریتم‌های یادگیری ماشین دیگر یا استفاده از روش‌های ترکیبی یادگیری ماشین؛ تنظیم هایپرپارامترهای الگوریتم می باشد.

پارامترهای مدل، ویژگی‌هایی از داده آموزشی هستند که در طول آموزش توسط الگوریتم‌های یادگیری ماشین، یاد گرفته می‌شوند، مانند شیب و عرض از مبدأ در رگرسیون خطی. پارامترهای مدل در هر آزمایشی متفاوت هستند و بستگی به داده‌ها و نوع مسئله دارند. درحالی‌که هایپرپارامترها باید توسط دانشمند داده، قبل از آموزش مشخص شوند. برای مثال تعداد و اندازه لایه‌های مخفی در شبکه عصبی، در الگوریتم جنگل تصادفی *(RandomForest)*، هایپرپارامترها تعداد درختان تصمیم در جنگل است. این‌ها به‌عنوان هایپرپارامترها شناخته می‌شوند؛ اما پارامترهای الگوریتم جنگل تصادفی، متغیرها و حد آستانه‌هایی *(thresholds)* هستند که برای

[۱۲] Hyper-parameters

تقسیم هر گره بکار می‌رود و در هنگام آموزش، یاد گرفته می‌شوند.

۱،۶. جستجوی شبکه‌ای[۱۳]

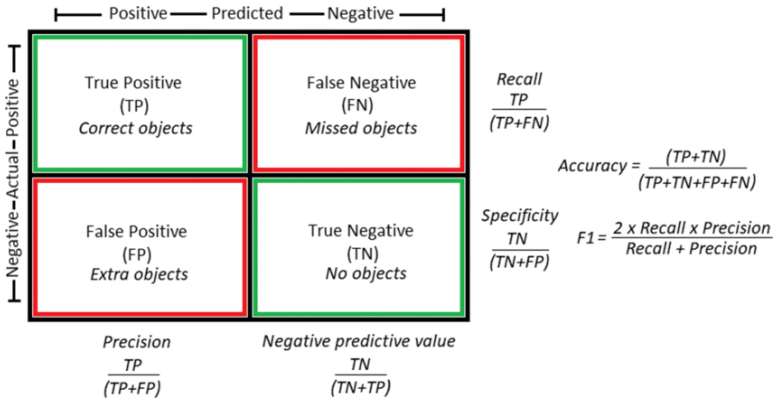
به‌طور رایج یک مدل پایه بدون تعدیل پارامترهای فوق (پارامترهایی که توسط کاربر قابل تنظیم هستند) ایجاد می‌شود. گاهی اوقات، انتخاب دقیق این پارامترها می‌تواند نتایج مدل را به‌طور قابل‌توجهی بهبود بخشد. جستجوی شبکه‌ای برای استفاده از محدوده عددی/رشته‌ای برای هایپرپارامترهای خاصی بدون نیاز به کد کردن تابع حلقه در این تحقیق طراحی شده است. نتایج استفاده از فراپارامترها جهت بهبود یک مدل نظارت شده یادگیری ماشین در این مطالعه تنظیم شده است. دقت مدل پایه و مدل اجرا شده با فراپارامترها در جدول ۱ نشان داده شده است [۲۰].

۷. ماتریس درهم‌ریختگی[۱۴]

برای ارزیابی کیفیت عملکرد مدل، معیارهای مختلفی برای هر رگرسیون و طبقه‌بندی وجود دارد. در این تحقیق که از طبقه‌بندی چند کلاسه استفاده شده است، ماتریس درهم‌ریختگی برچسب‌های کلاس پیش‌بینی شده را در برابر داده‌های برچسب واقعی نشان می‌دهد. در این تحقیق، از معیارهای ارزیابی مدل در شکل ۱۱ استفاده شده است.

جدول ۱. مقایسه مدل پایه با مدل اجرا شده با فراپارامترها همان‌طور که مشخص است در بعضی از الگوریتم‌های یادگیری مدل بهبود یافته است.

Row	*Model*	*Baseline model*	*Hyper-parameters model*
1	*LogR*	*0.76*	*0.77*
2	*Knn*	*0.971*	*0.982*
3	*Dtree*	*0.962*	*0.963*
4	*Rforest*	*0.983*	*0.984*
5	*SVM*	*0.919*	*0.974*
6	*GNB*	*0.528*	*0.55*
7	*GBC*	*0.938*	*0.974*
8	*Etree*	*0.986*	*0.986*



شکل ۱۱. معیارهای انتخاب شده ارزیابی مدل در این تحقیق بوسیله ماتریس درهم‌ریختگی(*Confusion matrix*)

۱،۷. ارزیابی مدل بوسیله ماتریس در هم‌ریختگی و نمودارهای *ROC*

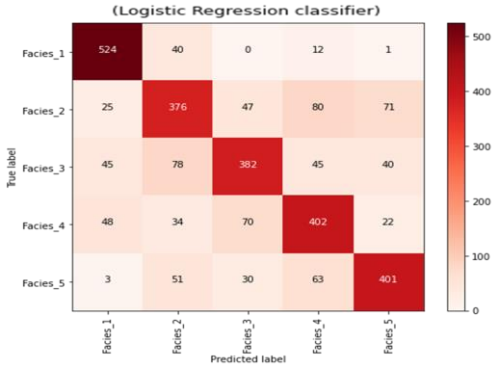
در این تحقیق، همه مدل‌ها با هایپرپارامترها بهینه شده است. سپس میانگین معیارهای ارزیابی بوسیله ماتریس درهم‌ریختی/آشفتگی برای همه الگوریتم‌های یادگیری ماشین در مرحله تست نشان داده شده است (جدول ۲). بعد از استفاده بوسیله میانگین‌گیری معیارهای ارزیابی مدل برای الگوریتم‌های به کار گرفته شده در این مطالعه، حالا نیاز است برای درک بهتر تک تک رخساره‌های سنگی معیارهای ارزیابی مدل بر هر واحد نشان داده شود. در شکل‌های ۱۲ تا ۱۹

[۱۳] Grid search

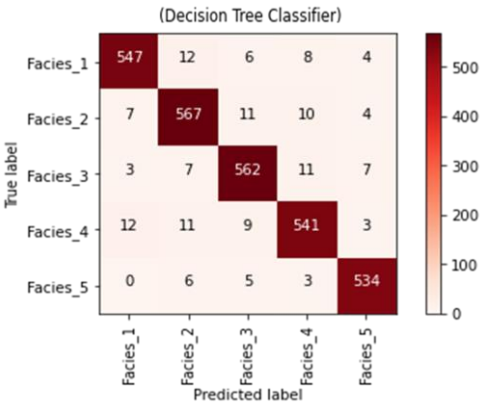
[14] Confusion matrix

بوسیله ماتریس درهم‌ریختگی برای همه روش‌ها نتایج مقدار واقعی با نتایج پیش‌بینی شده در مرحله تست نشان داده می‌شود [۲۲].

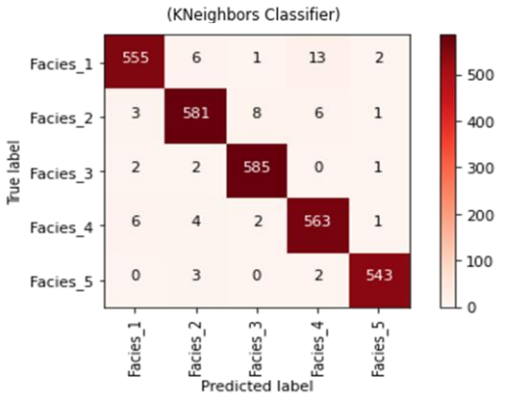
جدول ۲. میانگین معیارهای ارزیابی مدل برای روش‌های نظارت شده یادگیری ماشین



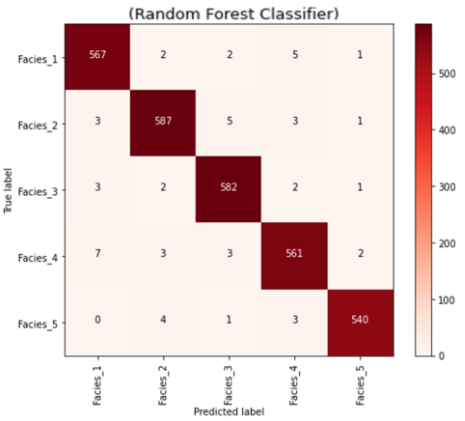
شکل ۱۲. پیش‌بینی رخساره‌های سنگی براساس مدل *Logistic Regression Classifier* در این حالت پیش‌بینی در بعضی رخساره‌ها به اشتباه در رخساره دیگر نشان داده می‌شود. در مدل‌های پیچیده روش *Logistic Regression* خوب نیست. بنابراین جهت پیش‌بینی این روش پیشنهاد نمی‌گردد.



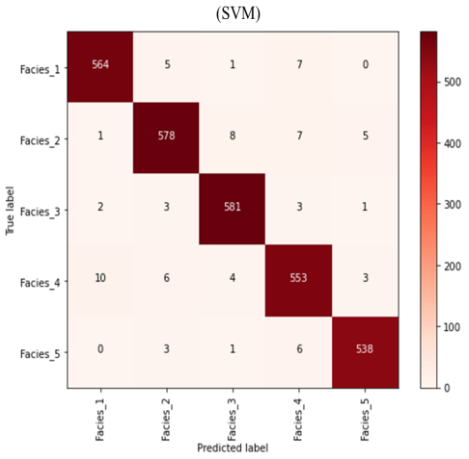
شکل۱۴: پیش‌بینی رخساره‌های سنگی براساس مدل طبقه‌بندی درخت تصمیم *(Decision Tree Classifier)*



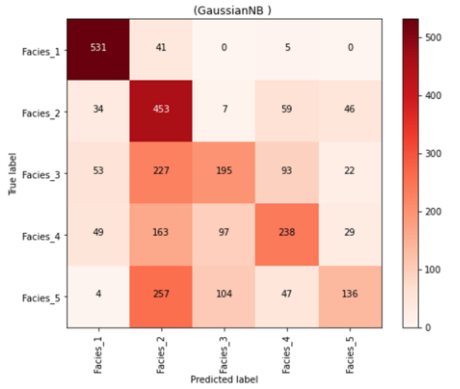
شکل۱۳. پیش‌بینی رخساره‌های سنگی براساس مدل *K Nearest Neighbors Classifier* همین‌طور که نشان داده شده نتایج پیش‌بینی مدل عملکرد خوبی دارد. به طور مثال رخساره واقعی شماره ۳ هیچ رخساره ۴ در طبقه آن قرار نگرفته است.



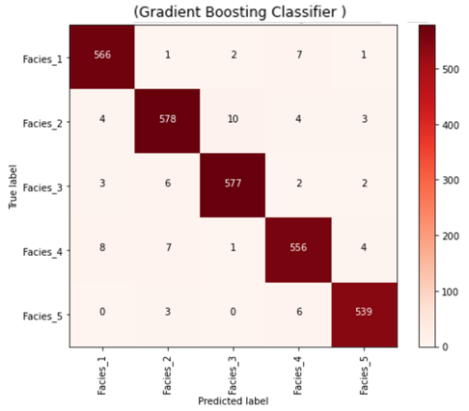
شکل ۱۵: پیش‌بینی رخساره‌های سنگی براساس مدل طبقه‌بندی جنگل تصادفی *(Random Forest Classifier)* در مدل الگوریتم طبقه‌بندی جنگل تصادفی عملکرد خوبی را نشان داده است.



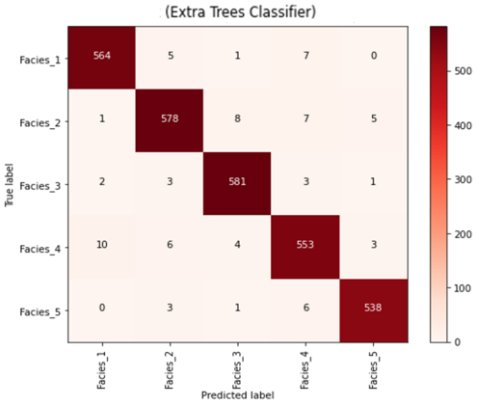
شکل ۱۶.پیش‌بینی رخساره‌های سنگی براساس مدل طبقه‌بندی ماشین بردار پشتیبان *(SVM)*



شکل ۱۷. پیش‌بینی رخساره‌های سنگی براساس مدل چگالی بیز *(Gaussian NB)*



شکل ۱۸. پیش‌بینی رخساره‌های سنگی براساس مدل گرادیان تصادفی ایکس جی بوست *(Gradient Boosting Classifier)* **تقویت گرادیان یک روش یادگیری ماشین برای مسائل رگرسیون و کلاسه‌بندی است که یک مدل پیش‌بینی‌کننده را در قالب مجموعه‌ای از مدل‌های پیش‌بینی‌کننده ضعیف ایجاد می‌کند.**



شکل ۱۹. پیش‌بینی رخساره‌های سنگی براساس مدل طبقه‌بندی درختان اضافی *(Extra Trees Classifier)* **این کلاس یک برآوردگر متا را پیاده‌سازی می‌کند که تعدادی درخت تصمیم تصادفی (معروف به درخت‌های اضافی) را در زیر نمونه‌های مختلف مجموعه داده برازش می‌دهد و از میانگین‌گیری برای بهبود دقت پیش‌بینی و کنترل بیش از حد برازش استفاده می‌کند.**

۲,۷. منحنی‌های یادگیری بوسیله نمودار *ROC*

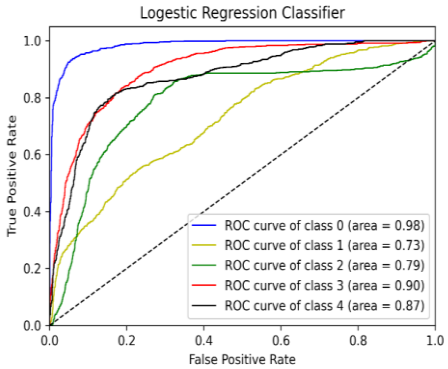
منحنی‌های *ROC* معمولاً از نرخ‌های مثبت واقعی در محور y و نرخ‌های مثبت کاذب در محور x تشکیل می‌شوند. این بدان معناست که گوشه سمت چپ بالای منطقه نمودار نقطه ایده‌آل است زیرا مثبت واقعی حداکثر و مثبت کاذب صفر است. با توجه به اینکه در یک مجموعه داده ممکن است داده‌های نویزدار و جود داشته باشد؛ بنابراین بسیار واقع‌بینانه نیست. با این حال، همیشه بهتر است که یک منطقه بزرگتر زیر منحنی داشته باشید در شکل ۲۰ کمترین مقدار مدل مربوط به الگوریتم *(Logistic Regression Classifier)* آورده شده است. در این شکل دقت کلاس‌های صفر یا اول در میان سایرین بالاتر است. در شکل ۲۱ برای الگوریتم *Extra Trees Classifier* نشان داده شده است؛ که کلاسه‌ها پیش‌بینی خوبی را نشان می‌دهد. قابل ذکر است این نمودار در مرحله تست مدل انجام گرفته است.

۸. ارزیابی چاه پیش‌بینی[۱۵]

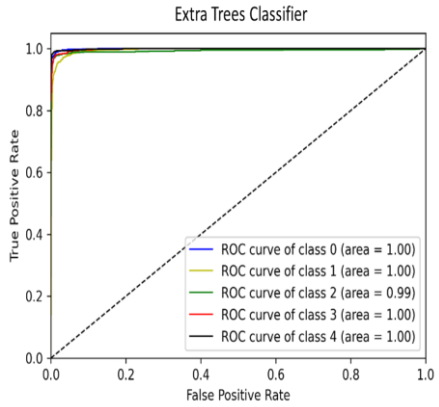
برای دقت مدل از بین این روش‌ها الگوریتم *Trees Classifier*

[۱۵] Blind well prediction and evaluation

Extra Classifier، *Gradient Boosting Classifier*، *K Neighbors* نتایج بهتری را نشان داده‌اند. اگرچه دقت پیش‌بینی مجموعه داده‌های آموزشی-آزمایشی برای همه مدل‌ها بالا است، اما در اعتبارسنجی چاه پیش‌بینی کاهش می‌یابد. بنابراین، بهتر است از تعداد نگاره‌ها وتعداد چاه‌های بیشتری استفاده گردد. می‌توان نتیجه گرفت که تعداد نمونه‌های داده و احتمالاً مقدار ویژگی‌ها به‌اندازه کافی زیاد نیست که مدل‌ها بتوانند تمام جنبه‌های پیچیدگی داده‌ها را به خوبی نشان دهند. در واقع، نمونه‌های بیشتر، ویژگی‌های داده بیشتر و تعداد بیشتر چاه‌های مجاور می‌تواند به بهبود دقت پیش‌بینی مدل کمک کند. همانطور که مشخص است در نمودار ستونی شکل ۲۲ برچسب‌های رخساره سنگی واقعی به عنوان معیاری برای مقایسه با نتایج مدل‌های مختلف پیش‌بینی ترسیم شده است. از نظر بصری، پیش‌بینی مدل‌های *Extra Trees*، *Gradient Boosting*، *K Nearest Neighbors* رابطه خوبی را نشان می‌دهند. یکی از نکات مهمی که در این تحقیق نشان داده شده است، این است که الگوریتم‌های مدل، لایه‌های نازک درون لایه‌های ضخیم سازند بنگستان را شناسایی وتشخیص داده‌اند.



شکل ۲۰. میانگین منحنی *ROC* یا *Receiver Operating Characteristic Curve* برای الگوریتم *Logistic Regression Classifier* محور عمودی نمودار *TPR/Recall/ Sensitivity* نمودار افقی *FPR=1-Specificity* می‌باشد مقدار ایده‌آل جایی است که منحنی *ROC* به نقطه *(1, 0)* نزدیکتر باشد و سطح زیر نمودار بیشتر باشد.

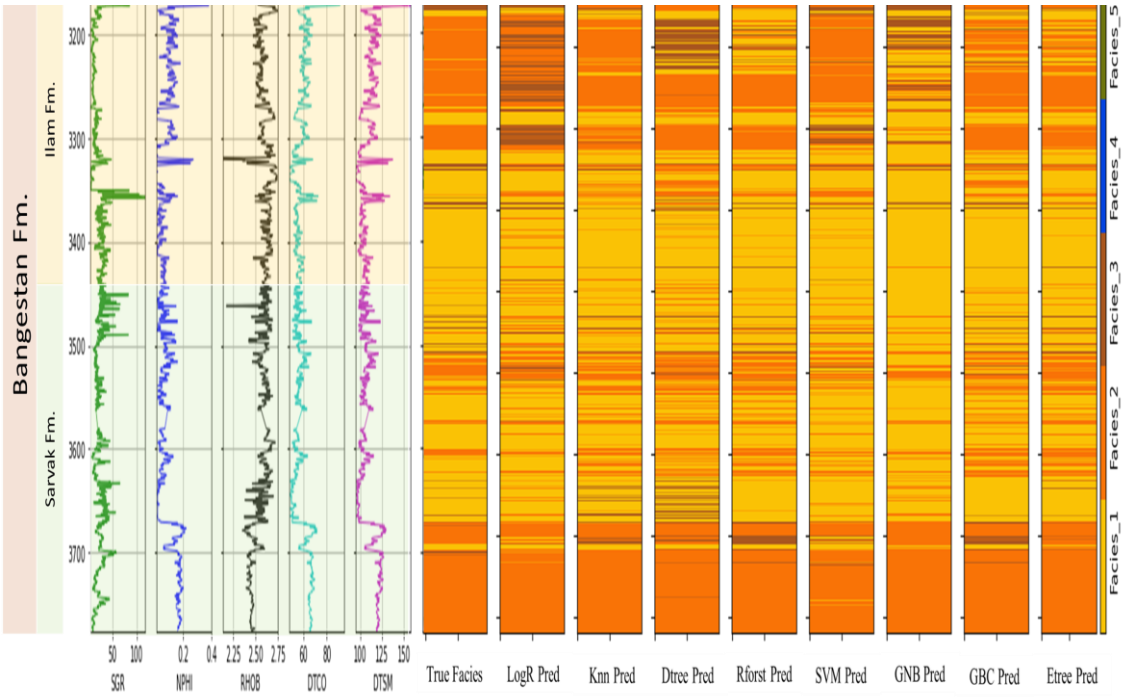


شکل ۲۱. میانگین منحنی *ROC* یا *Receiver Operating Characteristic Curve* برای الگوریتم *Extra Trees Classifier* محور عمودی نمودار *TPR/Recall/ Sensitivity* نمودار افقی *FPR=1-Specificity* می‌باشد مقدار ایده‌آل جایی است که منحنی *ROC* به نقطه *(1, 0)* نزدیکتر باشد و سطح زیر نمودار بیشتر باشد. همان‌طور که مشخص است برای همه کلاسه‌ها نتایج خوبی را نشان می‌دهد.

۹. نتیجه‌گیری

هدف اصلی از این پژوهش ساخت یک مدل یادگیری ماشین جهت پیش‌بینی رخساره‌های سنگی در چاه‌های نفت با دید، ارتقاء دقت در تخمین و پیش‌بینی خصوصیات سنگی مخزن نفت در چاه‌های فاقد نمونه حفاری می‌باشد. این اهمیت از جانب مهندسانِ ژئومکانیک و ژئوفیزیک‌دانان برجسته می‌شود چرا که دقت بالا در پیش‌بینی ویژگی‌های سنگی مخزن نفت، می‌تواند بهبود برنامه‌ریزی و بهره‌وری در فرآیندهای حفاری، تولید و بهره‌برداری از مخزن نفت منجر شود. با توجه به اینکه الگوریتم‌های یادگیری ماشین و یادگیری عمیق، توانمندی‌های پردازشی پیشرفته‌ای دارند که به تحلیل دقیق‌تر و پیش‌بینی بهتر خصوصیات سنگی مخزن نفت کمک می‌کنند این مدل‌ها در این تحقیق با استفاده از طیف وسیعی از روش‌های پیشرفته هوش مصنوعی شامل *Logistic Regression*، *K Nearest Neighbors*، *Decision Tree*، *Random Forest*، *Gaussian NB*، *Gradient Boosting*، *Extra Trees* و *SVM* جهت پیش‌بینی چاه‌های بدون طبقه‌بندی رخساره‌های سنگی استفاده شده است. بدین منظور از مجموعه داده، داده‌های ژئوفیزیکی و لاگ‌های چاه‌نگاری ۷ چاه آموزشی یکی از میدان‌های نفتی جنوب ایران که شامل نگاره گاما طبیعی *(SGR)*، نگاره گاما اصلاح

شده(*CGR*)، چگالی(*RHOB*)، تخلخل نوترونی(*NPHI*)، زمان موج‌برشی(*DTSM*) و زمان موج طولی (*DTCO*)که مستقیماً در تعیین رخساره‌ای ژئومکانیکی تأثیر دارند به عنوان داده‌های مستقل و واحدهای طبقه‌بندی شده رخساره به عنوان متغیر وابسته که بوسیله روش‌های دسته‌بندی هوش مصنوعی و مطالعات آزمایشگاهی به ۵ رخساره تفکیک شده بدست آماده است؛ در چاه‌های مدل به کار گرفته شد. این مجموعه داده از عمق ۳۰۰۰ تا ۴۰۰۰ هزار متری زمین مربوط به سازندهای آهک‌های ایلام و سروک (آهک بنگستان) است.



شکل۲۲. نمودار ستونی رخساره‌های سنگی پیش‌بینی شده بوسیله مدل‌های یادگیری ماشین با چاه هدف(*True Facies*)، همان‌طور که مشخص است الگوریتم‌های یادگیری ماشین از جمله الگوریتم درختان اضافی، گرادیان تقویتی تصادفی و $k-$نزدیک‌ترین همسایه نتایج بهتری نسبت به سایر الگوریتم‌ها جهت شناسایی رخساره‌های سنگی واقعی ارائه می‌دهد.

در این تحقیق برای ساخت مدل در گام اول، آماده‌سازی مجموعه داده شامل تجسم داده‌ها، مهندسی ویژگی‌ها، مدیریت مقادیر تهی و استخراج ویژگی‌های مهم که بخش مهمی از آماده‌سازی داده‌ها بود انجام گرفت. دومین مرحله مهم این پژوهش، ساخت یک مدل و اعتبارسنجی آن است؛ که با استفاده از یک مدل پایه و به کارگیری فراپارامترها برای عملکرد کارآمد مدل با دقت انتخاب شد. در این مطالعه، از یک جستجوی شبکه‌ای برای یافتن پارامترهای بهینه استفاده شده است. در نهایت، ارزیابی مدل، مهم‌ترین وظیفه در تولید مدل *ML* انجام گرفت و برچسب‌های رخساره‌ها پیش‌بینی بوسیله داده‌های تست پیش‌بینی شد. عملکرد مدل‌ها با چندین معیار ارزیابی شامل *Accuracy*، *Precision*، *F1-SCORE* و *Recall* بوسیله ماتریس درهم‌ریختگی و نمودارهای *ROC* مورد ارزیابی قرار گرفت. از بین این روش‌ها الگوریتم *Extra Trees Classifier*، *Gradient Boosting*، *K Nearest Neighbors Classifier* نتایج بهتری را نشان داده‌اند. در نهایت، عملکرد مدل جهت پیش‌بینی رخساره‌های سنگی چاه خارج از مدل یا چاه دیده نشده ارائه شد.

در نتیجه، برچسب‌های رخساره سنگی واقعی به‌عنوان معیاری برای مقایسه با نتایج مدل‌های مختلف پیش‌بینی ترسیم شده است. از نظر بصری، پیش‌بینی مدل *Extra Trees Classifier*،

K Nearest Neighbors ،*Gradient Boosting Classifier* *Classifier* رابطه خوبی را نشان می‌دهند. یکی از نکات برجسته در این تحقیق این است که الگوریتم‌های مدل، لایه‌های نازک درون لایه‌های ضخیم سازند بنگستان را شناسایی و تشخیص داده است.

۱۰. سپاس‌گزاری

با تشکر فراوان از محقق ارشد جناب آقای مردانی که ما را در زمینه نوشتن کدها و اجرای الگوریتم‌های به‌کاربرده شده در این تحقیق یاری کردند؛ کمال تشکر را داریم.

۱۱. مراجع

[1] Hall, B., (oct, 2016). Facies classification using machine learning. Lead. Edge 35 (10), 906–909. https://doi.org/10.1190/tle35100906.1

[2] Bestagini, P., Lipari, V., Tubaro, S., aug, (2017). A machine learning approach to facies classification using well logs. In: SEG Technical Program Expanded Abstracts 2017. Society of Exploration Geophysicists, pp. 2137–2142. https://doi.org/10.1190/ segam2017-17729805.1.

[3] Ashraf, U., Zhu, P., Yasin, Q., Anees, A., Imraz, M., Mangi, H.N., Shakeel, S., (2019). Classification of reservoir facies using well log and 3D seismic attributes for prospect evaluation and field development: a case study of Sawan gas field, Pakistan. J. Petrol. Sci. Eng. 175, 338–351. https://doi.org/10.1016/j.petrol.2018.12.060

[4] Ashraf, U., Zhang, H., Anees, A., Mangi, H.N., Ali, M., Ullah, Z., Zhang, X., (2020a). Application of unconventional seismic attributes and unsupervised machine learning for the identification of fault and fracture network. Appl. Sci. https://doi.org/ 10.3390/app10113864.

[5] Ali, M., Ma, H., Pan, H., Ashraf, U., Jiang, R., (2020). Building a rock physics model for the formation evaluation of the Lower Goru sand reservoir of the Southern Indus Basin in Pakistan. J. Petrol. Sci. Eng. https://doi.org/10.1016/j.petrol.2020.107461

[6] Asghar, S., Choi, J., Yoon, D., & Byun, J. (2020). Spatial pseudo-labeling for semi-supervised facies classification. Journal of Petroleum Science and Engineering, 195, 107834. https://doi.org/10.1016/j.petrol.2020.107834

[7] Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering, 174, 216-228. https://doi.org/10.1016/j.petrol.2018.11.023

[8] Babikir, I., Elsaadany, M., Sajid, M., & Laudon, C. (2022). Evaluation of principal component analysis for reducing seismic attributes dimensions: Implication for supervised seismic facies classification of a fluvial reservoir from the Malay Basin, offshore Malaysia. Journal of Petroleum Science and Engineering, 217, 110911. https://doi.org/10.1016/j.petrol.2022.110911

[9] Marco, I., John, F., Fred, J., 2021. Improving facies prediction by combining supervised and unsupervised learning methods: Journal of Petroleum Science and Engineering 200 (2021) 108300

[10] Muhammad Ali, A., Ren Jiang, B., Huolin Ma, A., Heping Pan, A., Khizar Abbas, C., Umar Ashraf, D., (2021). Machine learning – A novel approach of well logs similarity based on synchronization measures to predict shear sonic logs: Journal of Petroleum Science and Engineering 203 (2021) 108602

[11] Thiago Santi, B., Marcelo Kehl, D., Tiago, J., Girelli, F., Chemale, J., (2020). Evaluation of machine learning methods for lithology classification using geophysical data: Computers and Geosciences 139(2020)104475

[12] Song, C., Li, L., Li, K., (20201). Robust K-means algorithm with weighted window for seismic facies analysis: Journal of Geophysics and Engineering (2021) 18, 618–626. https://doi.org/10.1093/jge/gxab039.

[13] Dunham, W.M., Malcolm, A., Welford, J. K., (2020). Improved well log classification using semisupervised Gaussian mixture models and a new hyper-parameter selection strategy: Computers and Geosciences Volume 140, July 2020, 104501. https://doi.org/10.1016/j.cageo.2020.104501.

[14] Xu, R., Puzyrev, V., Elders, C., Fathi Salmi, E., & Sellers, E. (2023). Deep semi-supervised learning using generative adversarial networks for automated seismic facies classification of mass transport complex. Computers & Geosciences, 180, 105450. https://doi.org/10.1016/j.c ageo.2023.105450.

[15] Bao, L., Zhang, J., Zhang, C., Guo, R., Wei, X., & Jiang, Z. (2023). A reliable Bayesian neural network for the prediction of reservoir thickness with quantified uncertainty. Computers & Geosciences, 178, 105409. https://doi.org/10.1016/j.c ageo.2023.105409.

[16] Lppolito, M., Ferguson, J., Jenson, F., (2021). Improving facies prediction by combining supervised and unsupervised learning methods. Journal of

Petroleum Science and Engineering,Volume 200, May 2021, 108300. https://doi.org/10.1016/j.petrol.2020.108300.

[17] Ghalibaf, H., Ghafoori, M., Lashkaripoor, G.R., Hafezi Moghaddas, N.,(2020). Preparation of Zoning Maps for Cut-off Wall Using the Geotechnical parameters and Analytic Hierarchal Process (AHP). Case study: Sarroud Dam "Journal of Dam and Hydroelectric PowerPlant 7th Year / No. 26 / December 2020

[18] Ghalibaf, H., Hafezi Moghaddas, N., Lashkaripoor, G.R., Raoof G., Hossin T., (2022). Determination of geomechanical zones based on evaluation of Unsupervised Machine Learning algorithm methods "JOURNAL OF PETROLEUM GEOMECHANICS (JPG). (DOI): 10.22107/jpg.2022.329417.1158.

[19] Ghalibaf, H., Hafezi Moghaddas, N., Lashkaripoor, G.R., Raoof G., Hossin T., (2022). Estimation of Geomechanical Parameters, In Situ Stress Measurement Techniques, and Determination of Safe Mud Weight Windows Using Machine Learning Algorithm Methods "JOURNAL OF PETROLEUM GEOMECHANICS (JPG). 10.22107/JPG.2022.345905.1168

[20] Scikit Learn, (2020). K Nearest Neighbors Documentation, sklearn, viewed 20 April 2020,

[21] Scikit Learn, (2020), GridSearchCV Documentation, sklearn, viewed 20 April 2020,

[22] Mardani, R., (2020) https://github.com/mardani72/Facies-ClassificationMachine-Learning