نشریه علمی ژئومکانیک نفت
JOURNAL OF PETROLEUM GEOMECHANICS
(JPG)

مقاله پژوهشی

تعیین واحدهای ژئومکانیکی با استفاده از تجزیه و تحلیل الگوریتم‌های بدون نظارت یادگیری ماشین مبتنی بر نگاره‌های چاه‌های نفتی

حمید قالیباف محمدآبادی [۱]؛ ناصر حافظی مقدس [۲*]؛ غلامرضا لشکری پور [۳]؛ رئوف غلامی [۴]؛ حسین طالبی [۵]

۱. دانشجوی دکترای تخصصی؛ گروه زمین‌شناسی مهندسی، دانشکده علوم، دانشگاه فردوسی مشهد، مشهد، ایران
۲. استاد؛ گروه زمین‌شناسی مهندسی، دانشکده علوم، دانشگاه فردوسی مشهد، مشهد، ایران
۳. استاد؛ گروه زمین‌شناسی مهندسی، دانشکده علوم، دانشگاه فردوسی مشهد، مشهد، ایران
۴. استاد؛ دانشگاه منابع انرژی نروژ
۵. استاد؛ شرکت مناطق نفت جنوب

دریافت مقاله: ۱۴۰۰/۱۲/۲۳ پذیرش مقاله: ۱۴۰۱/۰۱/۱۹
شناسه دیجیتال (DOI): 10.22107/jpg.2022.329417.1158

واژگان کلیدی
الگوریتم‌های یادگیری ماشین، یادگیری بدون نظارت، یادگیری نظارت‌شده، مدل k- میانگین ، مدل آمیخته گوسی،

چکیده

در این تحقیق از روش یادگیری بدون نظارت جهت تعیین واحدهای ژئومکانیکی در یکی از چاه‌های نفتی جنوب ایران با استفاده از لاگ‌های داده‌های چاه‌نگاری شامل نگاره گاما طبیعی (*SGR*)، نگاره گاما اصلاح‌شده (*CGR*)، چگالی (*RHOB*)، تخلخل نوترونی (*NPHI*)، زمان موج برشی (*DTSM*) و زمان موج طولی (*DTCO*) استفاده شده است. برنامه‌نویسی مورد نیاز در محیط پایتون انجام‌گرفته است. در این راستا ابتدا بعد از پردازش داده‌های چاه‌نگاری از دو الگوریتم محبوب قدرتمند نظارت‌شده یادگیری ماشین ایکس جی بوست (*XGBoost*) و شبکه عصبی پرسپترون چندلایه (*Multi-Layer Perceptron Neural Network*) جهت بازیابی داده‌های گمشده استفاده گردید. سپس از روش‌های بدون نظارت یادگیری ماشین شامل مدل k- میانگین (*K-Means Clustering*)، الگوریتم خوشه‌بندی سلسله مراتبی (*HAC*)، الگوریتم خوشه‌بندی *DBSCAN* مبتنی بر غلظت و مدل آمیخته گوسی (*Gaussian Mixture Modelling*) جهت تعیین واحدهای ژئومکانیکی مخزنی پرفشار، آهک‌های نازک‌لایه و غیر مخزنی مسئله‌دار استفاده شد. در این روش‌ها الگوریتم‌ها خود الگوهای زیرسطحی را با استفاده از داده‌ها شناسایی می‌کنند که ممکن است به‌راحتی در طول کاوش داده قابل مشاهده نباشند. معیار ارزیابی دقت روش دقت در شناسایی آهک‌های نازک‌لایه، سازندهای غیر مخزنی مسئله‌دار و افق‌های پرفشار سازندهای مورد مطالعه در نظر گرفته شد. نتایج مطالعات نشان داد که از بین روش‌های مورد مطالعه روش *GMM* به‌جای اینکه بر اساس فاصله باشد، مبتنی بر توزیع است و از مرزهای خوشه/تصمیم بیضی استفاده می‌کند؛ بنابراین، منجر به طبقه‌بندی نرم‌تری می‌شود. علاوه براین، به خاطر قرار دادن الگوهای احتمالاتی مختلف برای شناسایی واحدهای ژئومکانیکی، روشی بهتر جهت تعیین واحدهای مخزنی پرفشار ایلام، سروک و آهک‌های نازک‌لایه می‌باشد.

* مشهد؛ دانشگاه فردوسی مشهد، گروه زمین شناسی مهندسی، دانشکده علوم، کدپستی: ۳۶۱۹۹۹۵۱۶۱؛ شماره‌ٔ تلفن: ۰۵۱۳۳۶۴۰۳۱۳؛ رایانامه: nhafezi@um.ac.ir

۱. پیشگفتار

در این مقاله، یادگیری نظارت‌شده *(Supervised Machine Learning)* رایج‌ترین و کاربردی‌ترین وظایف یادگیری ماشین است و طراحی آن بر اساس یادگیری از مدل با استفاده از داده‌های ورودی است که به خروجی دقیق نگاشت شده‌اند. در روش دیگر انجام مدل‌سازی با استفاده از یادگیری بدون نظارت *(Unsupervised Machine Learning)* است، جایی که به الگوریتم‌ها خود الگوهای زیرزمینی را با استفاده از داده‌ها شناسایی می‌کنند که ممکن است به‌راحتی در طول کاوش داده قابل مشاهده نباشند. یادگیری بدون نظارت یک تکنیک یادگیری ماشینی *(Machine Learning/ML)* است که نیازی به نظارت بر مدل‌ها توسط کاربران را ندارد. در یادگیری ماشینی بدون نظارت، از یک الگوریتم یادگیری برای کشف الگوهای ناشناخته در مجموعه داده‌های بدون علامت‌گذاری استفاده می‌کنیم. این برخلاف یادگیری ماشینی نظارت‌شده است که از داده‌های برچسب‌گذاری شده توسط انسان استفاده می‌کند. الگوریتم‌های یادگیری بدون نظارت از داده‌های بدون ساختار استفاده می‌کنند که بر اساس شباهت‌ها و الگوها گروه‌بندی شده‌اند. یادگیری بدون نظارت می‌تواند داده‌های پیچیده را برای ایجاد ویژگی‌های کمتر مرتبط تجزیه و تحلیل کند. سپس می‌توان با حذف این ویژگی‌ها با تأثیرات ناچیز بر بینش‌های ارزشمند، مدل را ساده کرد *(Hall, 2016; Bestagini et al., 2017)*.

در سال‌های اخیر استفاده از لاگ‌های چاه‌نگاری جهت طبقه‌بندی رخساره‌ها با استفاده از روش‌های یادگیری ماشین تحقیقات قابل توجهی را به خود دیده است *(Hall, 2016; Bestagini et al., 2017)*. دانشمندان بسیاری با استفاده از الگوریتم‌های محاسباتی جهت تخمین پارامترهای ناشناخته مانند رخساره‌ها ارائه داده‌اند *(Ashraf et al., 2019; Bestagini et al., 2017)*، گسل‌ها و شکستگی‌ها *(Ashraf et al., 2020a; Wu et al., 2019)*، ساخت یک مدل پتروفیزیکی برای ارزیابی مخزن ماسه‌سنگی *(Ali et al., 2020)*، در مطالعه دیگری با استفاده از لاگ‌های تصویری *(Image log)* و اطلاعات آزمایشگاهی از هر سازند و استفاده از ترکیب الگوریتم‌های نظارت‌شده و بدون نظارت در شناسایی و پیش‌بینی رخساره‌های مختلف به‌کار گرفته شد *(Marco et al., 2021)*. محمد و همکاران با استفاده از این الگوریتم‌ها لاگ‌های صوتی برشی را برای عمق‌هایی که فاقد این اطلاعات بودند پیش‌بینی نمودند *(Muhammad et al., 2010)*. تیاگو و همکاران با بررسی روش‌های طبقه‌بندی در یادگیری ماشین بهترین روش جهت شناخت لایه‌های سنگ‌شناسی، روش جنگل تصادفی *(Random forest method)* پیشنهاد داده‌اند *(Tiago et al., 2020)*. سانگ و همکاران با استفاده از روش خوشه‌بندی *K-means* جهت تعیین رخساره‌های لرزه‌ای پیشنهاد داده‌اند *(Song et al., 2021)*. دآن هام و همکاران جهت بهبود تعیین رخساره‌ها با توجه به کمبود داده‌های آموزشی جهت تعیین متغیر وابسته (رخساره) با استفاده از فرا پارامترها با استفاده از یک مدل مخلوط گوسی *(Gaussian mixture models)* مطرح کرده‌اند *(Dunham et al., 2020)*. فنگ و همکاران با استفاده از میدان‌های تصادفی مارکوف *(MRFs)* براساس مدل مخلوط گوسی جهت سنگ‌شناسی مخازن نفتی با توجه به تغییرات افقی و عمودی مخزن ارائه داده‌اند *(Fang et al., 2018)*. با توجه به اینکه در تعیین رخساره با استفاده از روش‌های نظارت‌شده چنانچه برچسب‌گذاری زون‌ها به‌صورت خیلی دقیق نباشد مدل ارائه‌شده کارایی لازم را ندارد. مارکو ایپولیتو و همکاران با ترکیب روش‌های نظارت‌شده با بدون نظارت جهت بهبود پیش‌بینی رخساره‌ها معرفی کرده‌اند *(Ippolito et al., 2021)*. از روش‌های بدون نظارت در تهیه نقشه‌های پهنه‌بندی با استفاده از پارامترهای ژئومکانیکی جهت بیشترین محدوده تزریق سیمان برای آب‌بندی پرده آب‌بند سد سرورد با استفاده از مدل سلسله مراتبی *(Analytic Hierarchy process, AHP)* به‌کار گرفته شد *(Ghalibaf. et al., 2020)*.

با توجه به اینکه در اکثر مخازن نفتی شناخت دقیقی از واحدهای ژئومکانیکی مخزن از لحاظ مکانیکی و فیزیکی وجود ندارد جهت شناسایی واحدها نیاز است از هر لایه چندین آزمایش سه محوری سنگ در شرایط مخزن از لحاظ فشار و دمای بالا انجام گیرد. حتی در صورت انجام این آزمایش‌ها بازهم سنگ موردمطالعه بکر است و فاقد قسمت‌های ازدست‌رفته و خورد شده است. علاوه بر این، انجام آزمایش به تعداد زیاد غیرممکن می‌باشد؛ بنابراین، استفاده از روش‌های نظارت‌شده اگر به‌طور دقیق جهت تعیین رخساره‌ها برچسب‌گذاری نشده باشد. الگوریتم به‌کار گرفته شده و پیشنهادی جهت پیش‌بینی زون‌های ژئومکانیکی همراه با خطا بیش از حد می‌باشد. با توجه به اینکه داده‌های چاه‌نگاری به‌صورت پیوسته در هر سانتی‌متر از چاه با دقت بالا پارامترهای ژئومکانیکی را اندازه‌گیری کرده است. بهتر است که از روش‌های بدون نظارت استفاده گردد تا بدون انجام آزمایش‌های پرهزینه و محدود، با استفاده از پارامترهای ژئومکانیکی به الگوریتم‌ها اجازه دهیم تا لایه‌ها و سازندهای مخزنی و غیر مخزنی را خود با توجه به شباهت‌ها از لحاظ خصوصیات ژئومکانیکی و پتروفیزیکی در کنار هم قرار دهد. بعد از این مرحله که لایه‌های مخزنی و غیر مخزنی مسئله‌دار مشخص و تأیید شد. می‌توان جهت پیش‌بینی

رخساره‌های میدان از روش‌های نظارت‌شده استفاده کرد. در این مرحله ابتدا بعد از به‌کارگیری الگوریتم موردنظر با استفاده از ماتریس درهم ریختی می‌توان پارامترهای صحت (*Accuracy score*)، دقت (*Precision*)، حساسیت یا نرخ مثبت مدل با مقدار واقعی (*Recall or Sensitivity*) و *F1- score* را به دست آورد. چنانچه نتایج این شاخص‌ها بالا باشد می‌توان بهترین الگوریتم را جهت پیش‌بینی رخساره‌ها معرفی کرد. با توجه به اینکه استفاده از هریک از روش‌های نظارت‌شده جهت ارائه الگوریتم موردنیاز، یک تحقیق جداگانه نیاز دارد، در این تحقیق ارائه نشده است.

در این مطالعه، در مرحله اول با استفاده از روش‌های یادگیری بدون نظارت ماشین مدل ژئومکانیکی (*MEM*) سازندهای مخزنی و غیر مخزنی مسئله‌دار (با سنگ‌شناسی پیچیده شامل انیدریت، گچ، مارن و لایه‌های نازک سنگ‌آهک) که مسئول بسیاری از مسائل پایداری چاه و زمان ازدست‌رفته غیر تولیدی (*NPT*) است شناسایی شدند. بعد از پردازش این روش‌ها بهترین روش جهت تعیین زون‌های غیر مخزنی مسئله‌دار (آهک‌های نازک‌لایه) و سازندهای مخزنی پرفشار ایلام و سروک که یکی از اهداف اصلی این مطالعه می‌باشد معرفی شد. شناسایی آهک‌های نازک‌لایه به این علت مهم است که در یک حوضه/مخزن بسته لایه‌های گچی در طول زمان به انیدریت تبدیل می‌شود، به علت قرارگیری این آهک‌های نازک‌لایه پر درز و شکاف در سازندهای تبخیری و ماسه‌سنگی آسماری در طی این فرایند آب آزادشده را جذب می‌کنند. درنتیجه، در زمان حفاری فشار بیش از حد انتظار باعث ریزش دیواره‌های چاه در این زون‌ها اتفاق خواهد افتاد. دانشمندان بسیاری شامل (*Aslannezhad et al., 2016*) روی کربنات‌ها، (*Kolawole et al., 2018*) با بررسی پایداری چاه بر روی ماسه‌سنگ‌ها و همچنین در سال‌های اخیر پایداری دیوار چاه بر روی شیل‌ها صورت پذیرفته است (*Wang et al., 2014*).

در مطالعه حاضر جهت تعیین واحدهای ژئومکانیکی سازندهای آسماری، جهرم، پابده، گورپی، ایلام، سروک و کژدمی از چهار روش خوشه‌بندی شامل: خوشه‌بندی k - میانگین (*K Means Clustering*)، مدل سلسله مراتبی (*HAC*)، مدلِ ترکیبیِ گوسی (*Gaussian Mixture Modelling*) و همین‌طور استفاده از خوشه‌بندی فضایی مبتنی بر چگالی در کاربردهای دارای نویز (*Density-Based Spatial Clustering of Applications with Noise /DBSCAN*) استفاده شده و با مقایسه نتایج مدل بهینه معرفی شده است.

۲. مبانی تئوری تحقیق

خوشه‌بندی داده‌ها شکل رایج تجزیه و تحلیل داده‌های اکتشافی (*exploratory data analysis/EDA*) است که برای تقسیم داده‌ها به گروه‌های مختلف بر اساس ویژگی‌ها یا ویژگی‌های مشترک استفاده می‌شود. نقاط داده‌ای که مشابه یکدیگر هستند در یک خوشه با هم گروه‌بندی می‌شوند و آن‌هایی که متفاوت هستند در خوشه دیگری قرار می‌گیرند. به‌طورکلی انواع اصلی خوشه‌بندی در یادگیری ماشینی بدون نظارت شامل کی میانگین، خوشه‌بندی سلسله مراتبی، خوشه‌بندی فضایی مبتنی بر چگالی برنامه‌ها با نویز (*DBSCAN*) و مدل مخلوط‌های گاوسی (*GMM*) است.

۲. ۱ خوشه‌بندی k – میانگین

خوشه‌بندی k - میانگین یک الگوریتم یادگیری ماشینی بدون نظارت بسیار رایج است. اصطلاح k - میانگین برای اولین بار توسط جیمز مک‌کوین در سال ۱۹۶۷ مورد استفاده قرار گرفت (*MacQueen, 1967*)، هرچند این ایده به هوگو استینگز در سال ۱۹۵۷ باز می‌گردد(*Steinhaus, 1957*). فورجی در سال ۱۹۶۵ الگوریتمی مشابه را منتشر کرد، به همین دلیل است که بعضی اوقات این الگوریتم، لوئید فورجی هم نامیده می‌شود (*Forgy, 1965*). با توجه به مجموعه‌ای از مشاهدات ($x_1, x_2, \ldots, x_n$) که در آن هر یک از مشاهدات یک بردار حقیقی d-بعدی است. خوشه‌بندی کی- میانگین با هدف پارتیشن‌بندی مشاهدات به $k \leq n$ مجموعه $s=\{s_1, s_2, s_3, \ldots sk\}$ است به‌طوری‌که مجموع مربع اختلاف از میانگین (یعنی واریانس) برای هر خوشه حداقل شود، تعریف دقیق ریاضی آن مطابق فرمول(*1*) زیر می‌باشد.

$$arg\ min \sum_{i=0}^{k} \sum_{X \in S_i} \|X - \mu_i\|^2 = arg\ min \sum_{i=1}^{k} |S_i|\, Var S_i \qquad (1)$$

که در آن μ_i میانگین نقاط در si است. این معادل است با به حداقل رساندن دو به دو مربع انحراف از نقاط در همان خوشه که مطابق با رابطه (*2*) زیر است.

$$\sum_{Cluster\ C_i} \sum_{Dimension\ d} \sum_{x,y \in C_i} (x_d - y_d)^2 \qquad (2)$$

چون کل واریانس ثابت است، از قانون واریانس کلی می‌توان نتیجه گرفت که این معادله برابر است با بیشینه کردن مربع انحرافات بین نقاط خوشه‌های مختلف(*Kriegel, 2016*) .

رایج‌ترین الگوریتم k- میانگین یا الگوریتم لوید با استفاده از یک تکرارشونده پالایش کار می‌کند(*MacKay, 2003*) . مراحل الگوریتم به‌صورت زیر می‌باشد (شکل ۱).

ابتدا k میانگین یعنی $(\mu_1^{(0)}, \mu_2^{(0)} \ldots . \mu_k^{(0)})$ را که نماینده خوشه‌ها هستند، به‌صورت تصادفی مقداردهی می‌کنیم.

سپس، این دو مرحله پایین را به‌تناوب چندین بار اجرا می‌کنیم تا میانگین‌ها به یک ثبات کافی برسند و یا مجموع واریانس‌های

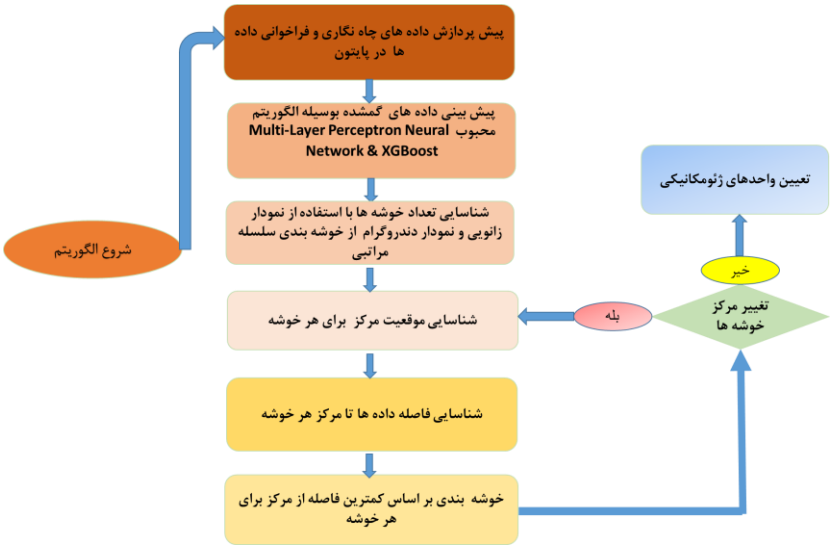
خوشه‌ها تغییر چندانی نکنند. از میانگین‌ها k خوشه می‌سازیم، خوشه i ام در زمان t تمام داده‌هایی هستند که از لحاظ اقلیدسی کمترین فاصله را با میانگین $\mu_1^{(t)}$ یعنی میانگین i ام در زمان t دارند زبان ریاضی خوشه i ام در زمان t برابر رابطه (۳) خواهد بود با:

$$S_i^{(t)} = \{x_p: \left\|x_p - \mu_i^{(t)}\right\|^2 \leq \left\|x_p - \mu_j^{(t)}\right\|^2 \forall j, 1 \leq j \leq k \quad (3)$$

حال میانگین‌ها را بر اساس این خوشه‌های جدید مطابق فرمول (۴) بروز می‌کنیم (*Kriegel, 2016*).

$$\mu_i^{(t+1)} = \frac{1}{\left|S_I^{(t)}\right|} \sum_{xj \in S_i^{(t)}} x_j \quad (4)$$

در نهایت میانگین‌های مرحله آخر (در زمان T یعنی $(\mu_1^{(T)}, \mu_2^{(T)} \ldots . \mu_k^{(T)})$ خوشه‌ها را نمایندگی خواهند کرد (*Kriegel et al., 2016*).



شکل ۱. مراحل الگوریتم k - میانگین جهت تعیین واحدهای ژئومکانیکی

به‌طورکلی خوشه‌بندی کی- میانگین یک روش خوشه‌بندی سخت است که در آن یک نقطه داده یا متعلق به یک خوشه است یا نیست. همچنین با اعمال یک دایره (ابر کره در مجموعه داده‌های چندبعدی) روی داده‌ها، خوشه‌بندی را انجام می‌دهد. از نظر محاسباتی بسیار کارآمد است و می‌توان به‌راحتی پیاده‌سازی کرد.

۲,۲ خوشه‌بندی سلسله مراتبی (HAC)

یکی از روش‌های یادگیری ماشین *(Machine Learning)* که به آموزش بدون نظارت *(Unsupervised Learning)* شهرت دارد، تحلیل خوشه‌بندی *(Clustering Analysis)* است. در این روش، برعکس خوشه بندی k - میانگین، هر مشاهده ممکن است در بیش از یک خوشه قرار گیرد زیرا بر اساس سطوح مختلف فاصله، خوشه‌ها تشکیل می‌شود. بنابراین هر خوشه ممکن است زیرمجموعه خوشه دیگر در سطحی از فاصله قرار گیرد. خوشه‌بندی سلسله مراتبی با روش تجمیعی *(Hierarchical Agglomerative Clustering, or HAC)* اگر دیدگاه به این نمودار از پایین به بالا باشد *(Bottom-Up)*، برحسب ارتفاع نمودار *(Height)* در سطح پایینی، خوشه‌ها زیرمجموعه خوشه‌های سطح بالاتر هستند درنتیجه به نظر می‌رسد که خوشه‌های زیرین با یکدیگر ترکیب‌شده و خوشه‌های سطح بالاتر را ایجاد می‌کنند. این شیوه خوشه‌بندی سلسله مراتبی به روش تجمیعی *(Agglomerative)* معروف است. در الگوریتم *HAC* پیچیدگی زمانی برابر با $O(n^3)$ و فضای مورد نیاز حافظه نیز برابر با $O(n^2)$ است. بنابراین با افزایش حجم داده‌ها، سرعت و فضای حافظه برای اجرای عملیات خوشه‌بندی به‌شدت افزایش می‌یابد.

برای اجرای محاسبات مربوط به این روش خوشه‌بندی، به دو معیار فاصله (شباهت) احتیاج داریم. ۱- میزان فاصله بین زوج مشاهدات. ۲- میزان فاصله بین خوشه‌ها. در حالت اول توابع فاصله برای داده‌های کمی یا کیفی قابل استفاده هستند؛ بنابراین اگر داده‌ها کمی باشند برای مثال می‌توان از فاصله اقلیدسی یا فاصله منهتن استفاده کرد. همچنین برای داده‌های کیفی نیز می‌توان از میزان انطباق ساده *(Simple Matching)* و یا فاصله همینگ *(Hamming Distance)* برای داده‌های کمک گرفت. معمولاً قبل از شروع مراحل خوشه‌بندی سلسله مراتبی تجمیعی، از یک ماتریس فاصله *(Distance Matrix)* یا ماتریس شباهت *(Similarity Matrix)* برای تسریع در محاسبات کمک گرفته می‌شود. این ماتریس نشان می‌دهد که فاصله بین هر زوج از مشاهدات چقدر است. البته نوع تابعی که

باید به‌وسیله آن فاصله اندازه‌گیری شود، در مقدارهای موجود در این ماتریس تأثیرگذار است.

در خوشه‌بندی سلسله مراتبی تجمیعی یا *HAC* ، با توجه به مقدارهای این ماتریس، مشاهدات یا خوشه‌هایی که دارای کمترین فاصله (بیشترین شباهت) هستند با هم ادغام می‌شوند و خوشه جدیدی می‌سازند. در مرحله بعد باز هم فاصله بین مشاهدات و یا خوشه‌های جدید، توسط ماتریس فاصله که به‌روزرسانی شده، محاسبه و کار ادغام ادامه پیدا می‌کند تا تنها یک خوشه باقی بماند *(Hastie et al., 2009)*.

۳,۲ مدل‌های مخلوط گاوسی (GMM)

در این تکنیک خوشه‌بندی پیشرفته از ترکیبی از توزیع‌های گاوسی برای مدل‌سازی یک مجموعه داده استفاده می‌شود. این مدل‌های مخلوط احتمالی هستند و برای تولید نمونه‌های داده از مدل‌های خوشه‌بندی *GMM* استفاده می‌شود. در این مدل‌ها، هر نقطه داده عضوی از همه خوشه‌های مجموعه داده است، اما با درجات مختلف عضویت. احتمال عضویت در یک خوشه خاص بین ۰ و ۱ است. در مدل‌های مخلوط گاوسی، اطلاعات کلیدی شامل مراکز گاوسی پنهان و کوواریانس داده‌ها است. این باعث می‌شود که شبیه به *K-means* خوشه‌بندی شود. یک مدل آمیخته گاوسی در رابطه ۵ نشان داده می‌شود. در اینجا منظور از π_k وزن توزیع k ام یا فراوانی مشاهدات از آن توزیع است.

$$p(x) = \sum_{k=1}^{K} \pi_k N\,(x|\mu_k, \Sigma_k), \sum_{k=1}^{K} \pi_k = 1 \qquad (5)$$

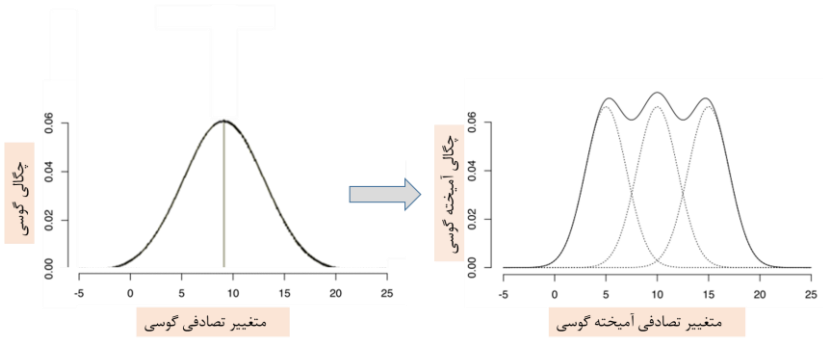
فرض کنید تعداد خوشه‌ها برابر با k باشد. در اینجا پارامترهای گاوسی (نرمال) نیز با θ نشان داده می‌شوند. متغیر پنهان که به‌صورت یک بردار، برچسب هر یک از مشاهدات را نشان می‌دهد با Z مشخص شده است. در این صورت مدل آمیخته گاوسی یک‌بعدی برای این خوشه‌های به‌صورت رابطه ۶ نوشته می‌شوند.

$$p(X|\theta) = \sum\nolimits_{Z} p(X, Z|\theta) \qquad (6)$$

درنتیجه هدف از اجرای خوشه‌بندی بر مبنای مدل، بیشینه‌سازی این تابع درست‌نمایی است. مقدار Z به دو صورت دو دویی یا باینری است به این معنی که اگر مشاهده nام به خوشه k ام تعلق داشته باشد مقدار این متغیر برابر با ۱ و در غیر این صورت برابر با صفر خواهد بود. البته ازآنجایی‌که بیشنه سازی لگاریتم آن مشابه است و لگاریتم گیری، فرم تابع احتمال آمیخته گاوسی (نرمال) را ساده‌تر می‌کند از رابطه ۷ برای اجرای الگوریتم خوشه‌بندی بر مبنای مدل استفاده خواهیم کرد.

$$Ln[(p(X|\theta)] = Ln\left\{\sum\nolimits_{z} p(X, Z|\theta)\right\} \qquad (7)$$

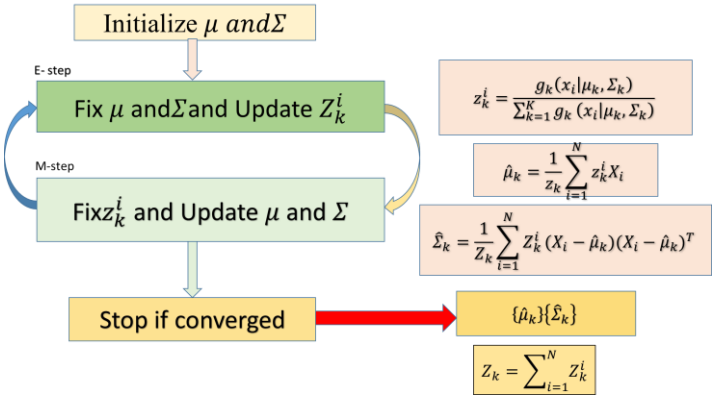
در نمودار شکل ۲، یک نمونه از توزیع آمیخته گاوسی (ترکیب سه متغیر تصادفی نرمال) نشان داده می‌شود.



شکل ۲. نمایش گرافیکی مدل گوسی و مدل‌های مخلوط گاوسی (Li ea al., 2011)

وجود پارامترهای زیاد در این تابع (مانند میانگین، واریانس و کوواریانس بین خوشه‌ها) امکان بیشینه‌سازی آن را با روش‌های تحلیلی نمی‌دهد؛ بنابراین در چنین مواقعی استفاده از تکنیک‌های عددی مانند الگوریتم *EM* کارساز است. در این تحقیق به‌منظور برآورد پارامترهای مدل آمیخته گاوسی از الگوریتم *EM* استفاده شد. این الگوریتم دارای دو بخش یا گام است. گام متوسط‌گیری *(Expectation)* و گام بیشینه‌سازی *(Maximization)* که به‌طور خلاصه مراحل این الگوریتم در نمودار شکل ۳ نشان داده می‌شود. در گام اول *(E-step)* یا گام متوسط گیری، هدف برآورد توزیع متغیر پنهان به‌شرط وزن (π)، میانگین‌ها *(Means)* و ماتریس کوواریانس (Σ) توزیع آمیخته نرمال است. بردار پارامترها را در اینجا با نماد θ نشان می‌دهیم. برای این کار ابتدا یک حدس اولیه برای این پارامترها زده می‌شود. سپس گام‌ها به ترتیب برداشته می‌شوند. حدس‌های اولیه را می‌توان به‌وسیله الگوریتم *k-means* به دست آورد. به این معنی که برای خوشه‌های حاصل از الگوریتم *k-means*، میانگین، ماتریس کوواریانس و وزن‌ها محاسبه می‌شود. منظور از وزن، درصدی (احتمال) از داده‌ها

است که در یک خوشه قرار دارند. در گام دوم (*M- step*) با استفاده از متغیرهای پنهان، سعی می‌شود تابع درست‌نمایی نسبت به پارامترهای θ بیشینه شود. مراحل گام E و گام M تا زمانی که الگوریتم همگرا شود (مقدار تابع درست‌نمایی تغییر نکند)، ادامه خواهد داشت. به‌این‌ترتیب الگوریتم *EM* علاوه بر برآورد پارامترهای توزیع آمیخته گاوسی، برچسب‌ها یا مقدار پنهان را نیز مشخص می‌کند.



شکل ۳. مراحل الگوریتم EM برای مدل آمیخته گاوسی (GMM)

به‌طورکلی، روش *GMM* اجازه می‌دهد تا نقاط، داده را خوشه‌بندی کنند با این تفاوت که واریانس داده‌ها را محاسبه می‌کند، منجر به طبقه‌بندی نرم‌تری می‌شود و به‌جای اینکه بر اساس فاصله باشد، مبتنی بر توزیع است. همچنین، نقطه داده‌ای که طبقه‌بندی می‌شود، احتمال بودن در دو گروه خوشه‌بندی را دارد. درحالی‌که اگر خوشه‌های داده دایره‌ای باشند، خوشه‌بندی *K-Means* عالی عمل می‌کند، اما در موقعیت‌های پتروفیزیکی و زمین‌شناسی داده‌ها به‌ندرت الگوهای دایره‌ای خوبی را تشکیل می‌دهند. بنابراین مدل‌سازی *GMM* از مرزهای خوشه/تصمیم **بیضی** شکل استفاده می‌کند و بنابراین انعطاف‌پذیرتر است(*Jinghua et al.,2021*) . برای این منظور از داده‌های چاه‌نگاری یکی از مخازن نفتی واقع در جنوب ایران جهت تعیین زون‌های ژئومکانیکی استفاده شده است.

۳. پیش‌پردازش داده‌ها

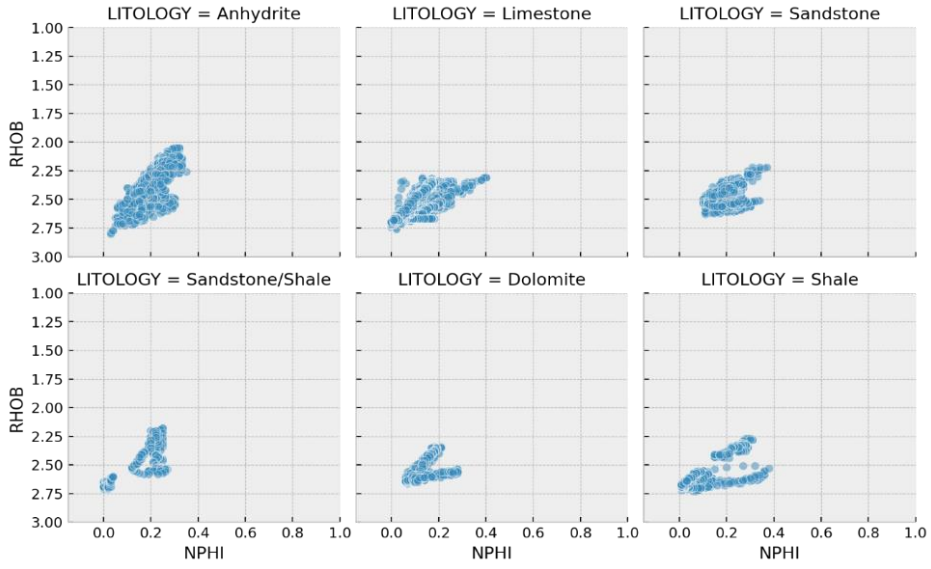
در مطالعه حاضر از داده‌های چاه‌نگاری یکی از میدان نفتی جنوب ایران شامل نگاره گاما طبیعی *(SGR)*، نگاره گاما اصلاح‌شده*(CGR)* ، چگالی *(RHOB)*، تخلخل نوترونی *(NPHI)*، زمان موج برشی *(DTSM)* و زمان موج طولی *(DTCO)* که مستقیماً در تعیین واحدهای ژئومکانیکی تأثیر دارند استفاده شده است. ابتدا فایل داده‌های رقومی *(Las)* این داده‌ها در پایتون فراخوانی شد و یک دیتا فریم *(Data Frame)* ساخته شد بعد از این مرحله تمام مراحل پیش‌پردازش *(Data Preprocessing)* جهت ساخت مدل شامل موارد زیر به‌درستی صورت پذیرفت:

- اندازه‌گیری کیفیت داده‌ها *Measures for data (quality)*: نمای چندبعدی، دقت (درست یا غلط، دقیق یا نه)، کامل بودن (ثبت‌نشده، غیرقابل دسترس)، سازگاری*(Consistency)* (برخی اصلاح‌شده اما برخی نه، نامشخص)، به‌موقع بودن (به‌روز رسانی به‌موقع)، باورپذیری (داده‌ها چقدر قابل اعتماد هستند؟)، تفسیرپذیری (به چه راحتی می‌توان داده‌ها را درک کرد).
- جایگزینی داده‌های گمشده: مقادیر از دست رفته با استفاده از دو الگوریتم محبوب قدرتمند نظارت‌شده یادگیری ماشین ایکس جی بوست *(XGBoost)* و شبکه عصبی پرسپترون چندلایه *Multi-Layer Percepteron Neural Network* تعیین شد. این الگوریتم یک روش عمیق یادگیری ماشین نظارت شده است. استفاده از این روش برای پیش‌بینی طیف وسیعی از علوم رو به گسترش می‌باشد. این روش زمانی که شناخت خوبی از پیش‌بینی مدل وجود دارد استفاده می‌گردد. به‌طور مثال در مطالعات ژئومکانیک ما به مدل تغییرات سنگ‌شناسی در عمق‌های مختلف را می‌دهیم تا بهترین الگوریتم را آنالیز و پردازش کند. معمولاً جهت پرکردن داده‌های گمشده از روش‌های میانگین‌گیری یا استفاده از داده‌هایی بالا و پایین داده گمشده (بالا و پایین واحد ژئومکانیکی) استفاده می‌گردد. ارائه رابطه خطی نیز به دلیل محاسبه مقدار میانگین متغیر وابسته همیشه همراه با خطا بوده و مدل مناسبی ارائه نمی‌دهد. اما الگوریتم‌های نظارت شده

داده‌های گمشده را با در نظر گرفتن تمام داده‌های ورودی به مدل در نظر می‌گیرد. به‌طور مثال در عمقی از لاگ که ما داده تخلخل را نداریم این داده گمشده با بررسی تمام متغیرها از جمله سرعت موج طولی، موج برشی، چگالی و غیره پر می‌شود. درنتیجه این داده پرشده بهترین حالت ممکن برای نبود داده گمشده ارائه می‌دهد. قابل ذکر است داده‌های گمشده با *Accuracy=99%* به‌دست‌آمده است.

- پاک‌سازی داده‌ها: در این بخش داده‌های نویز دار صاف شد (*smooth noisy data*)، موارد پرت شناسایی یا حذف شد و ناسازگاری‌ها و خطاها را برطرف شد.
- یکپارچه‌سازی داده‌ها (*Data integration*): ادغام چندین پایگاه داده، یکپارچگی داده یا فایل، تبدیل داده‌ها، عادی‌سازی و تجمیع
- کاهش داده‌ها: نمایش کاهش‌یافته در حجم را به دست می‌آورد اما نتایج تحلیلی یکسان یا مشابهی را تولید می‌کند.
- گسسته‌سازی داده‌ها (*Data discretization*): بخشی از کاهش داده‌ها اما با اهمیت ویژه، به‌ویژه برای داده‌های عددی

بعد از این مرحله با توجه به اینکه در مطالعات زمین‌شناسی شناخت کاملی از نوع سازندهای مخازن وجود دارد، در این مدل‌سازی از یک ستون رخساره‌ای تحت عنوان لیتولوژی جهت شناسایی زون‌های ژئومکانیکی حاصل از مدل استفاده گردیده است. برای انجام این کار در محیط پایتون از روش *Seaborn-FacetGrid* برای رسم نمودارهای پراکنده چگالی-نوترون (*scatterplots*) برای هر سنگ‌شناسی که در ستون رخساره‌ای مشخص‌شده استفاده گردید. *FacetGrid* برای ایجاد یک ساختار زیرسطحی برای نمودار استفاده می‌شود. در این مثال، هنگامی که 6 ستون وجود دارد، داده‌ها تقسیم‌بندی می‌شوند. سپس می‌توانیم یک کراس پلات نوترونی چگالی را روی آن *FacetGrid* ترسیم کنیم. در این تحقیق با توجه به توالی رسوبی و چینه‌شناسی، به ترتیب از بالا به پایین بودن سازندهای گچساران، آسماری، جهرم، پابده، گورپی، ایلام، سروک و کژدمی از پنج رخساره قابل پیش‌بینی شامل انیدریت، کلسیت، دولومیت، ماسه‌سنگ، شیل، ماسه‌سنگ/ شیل استفاده شده است. (شکل ۴).



شکل ۴. پراکندگی داده‌های نوترون- چگالی به تفکیک خصوصیات سنگ‌شناسی در ستون چاه مورد مطالعه

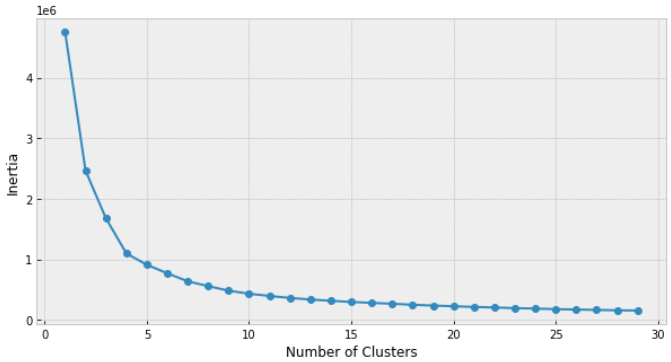
۴. خوشه‌بندی - بدون نظارت

۱,۴ یافتن تعداد بهینه خوشه‌ها

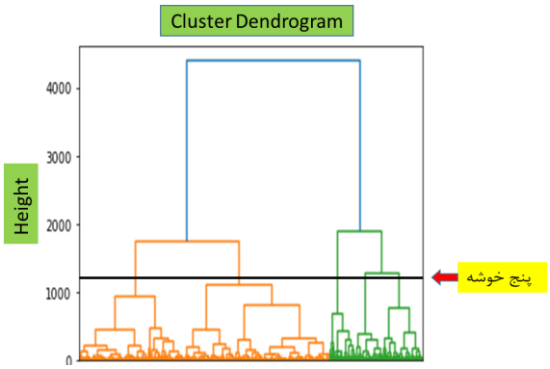
برای اطمینان از اینکه مدل‌های *HAC*، *K-Means* و *Gaussian Mixture Modeling* به‌طور مؤثر کار می‌کنند، باید تعداد اولیه خوشه‌ها را برای آن‌ها فراهم کنیم. اگر تعداد خوشه‌ها اشتباه انتخاب‌شده باشد، الگوریتم‌ها ممکن است عملکرد خوبی نداشته باشند یا رفع آن‌ها بیشتر طول بکشد. از آنجایی که برخلاف عملیات تحلیل رده‌بندی (*Classification*)، خوشه‌بندی (*Clustering*) یک فرآیند

بدون نظارت (*Unsupervised*) است، بررسی صحت نتیجه خوشه‌بندی به‌راحتی امکان‌پذیر نیست. بنابراین احتیاج به معیارهای مناسب، هم برای بررسی کارایی یک روش خوشه‌بندی در بازیابی خوشه‌ها و هم برای مقایسه عملکرد روش‌های مختلف خوشه‌بندی، ضروری به نظر می‌رسد. در این میان دو گونه معیار ارزیابی نتایج خوشه‌بندی وجود دارد. معیارهای ارزیابی درونی (*Internal Criteria Index*) و معیارهای ارزیابی بیرونی (*External Criteria index*). از روش زانویی (*elbow*) برای تعیین تعداد صحیح خوشه‌ها در این دیتاست استفاده شده است. در این روش مقادیر افزایشی k بر روی محور افقی و مجموع خطاهایی که در هنگام استفاده از k میانگین رخداده بر روی محور عمودی ترسیم می‌شود (شکل ۵).

در این تحقیق از روش خوشه‌بندی سلسله مراتبی (*Agglomerative hierarchical Clustering*) جهت مقایسه با روش زانویی استفاده گردیده است. انتخاب تعداد خوشه‌ها در این روش اساساً با مدل *k-means* متفاوت است. در این خوشه‌بندی بجای انتخاب تعداد خوشه‌ها و آغاز با مرکز وارهای تصادفی، کار بدین‌صورت است که هر نقطه درمجموع داده یک خوشه است. سپس، دو نقطه نزدیک‌تر به هم پیدا و با یکدیگر در یک خوشه ادغام می‌شوند. پس از آن، دو نقطه نزدیک به هم بعدی یافت و با هم در یک خوشه قرار می‌گیرند. این فرآیند تا هنگامی انجام می‌شود که تنها یک خوشه بسیار بزرگ باقی بماند. در طول فرآیند بیان‌شده، چیزی ساخته می‌شود که به آن دندروگرام (*Dendrogram*) گفته می‌شود. در نمودار شکل ۶ زیر تعداد خوشه‌ها با تقاطع خط افقی با محور عمودی در هر جای نمودار بسته به نظر متخصص می‌توان انتخاب کرد. در اینجا عدد ۵ در نظر گرفته شده است. می‌توان مشابه روش زانویی این عدد را ۵ تا ۱۰ انتخاب کرد.



شکل ۵. نمودار زانویی جهت تعیین تعداد بهینه خوشه‌ها، در این روش مقادیر افزایشی k بر روی محور افقی و مجموع خطاهایی که در هنگام استفاده از k میانگین رخداده بر روی محور عمودی ترسیم می‌شود در این تحقیق تعداد خوشه‌ها ۵ انتخاب شده است.



شکل ۶. نمودار دندروگرام جهت تعیین تعداد بهینه خوشه‌ها که در اینجا یک خط افقی از هر نقطه نمودار عبور داده می‌شود. تعداد بهینه خوشه‌ها از ۵ تا ۱۰ عدد می‌توان انتخاب کرد. در اینجا عدد پنج انتخاب شده است.

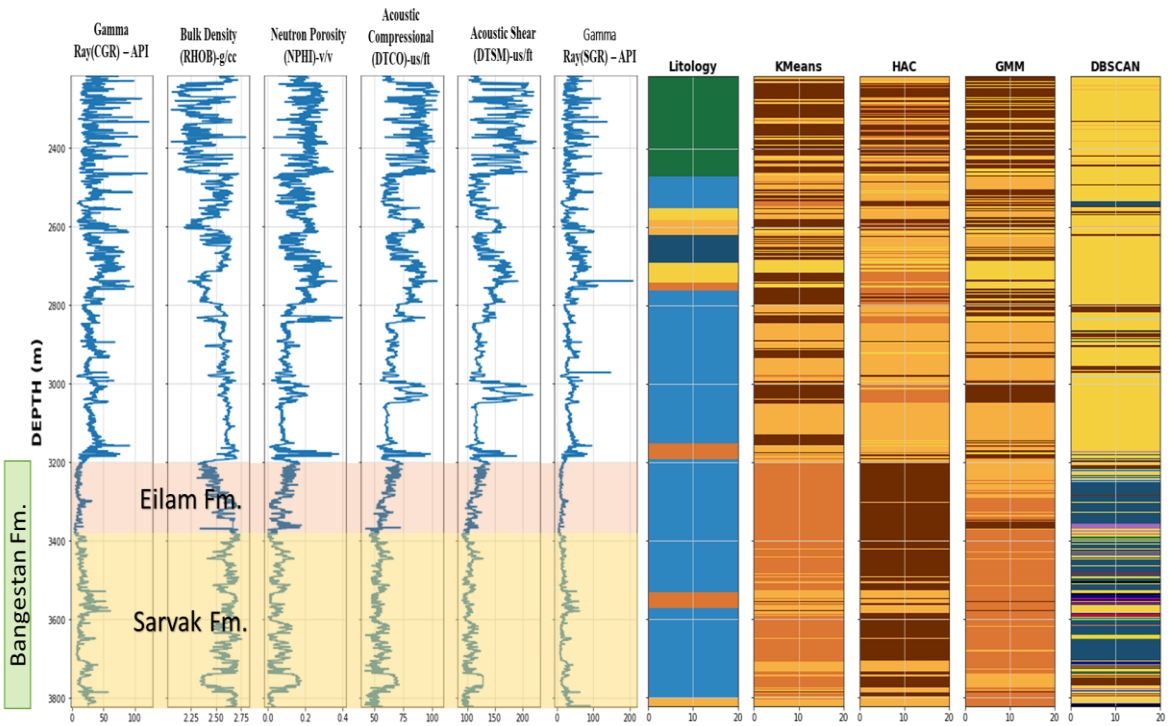
در نمودار بالا می‌بینیم که تابع هزینه (مجموع مجذور فواصل تا نزدیک‌ترین مرکز خوشه) با افزایش تعداد خوشه‌ها کاهش می‌یابد. هیچ شکاف مشخصی در این مجموعه داده وجود ندارد، با این حال، می‌توانیم ببینیم که شیب از حدود ۵ خوشه به بعد کم می‌شود. درنتیجه، مقدار تابع هزینه ثابت می‌شود. انتخاب این مقدار به مفسر بستگی دارد و می‌تواند از ۵ تا ۱۰ متغیر باشد. بنابراین برای این مثال ما ۵ را به‌عنوان تعداد بهینه خوشه‌ها در نظر گرفته شده است.

۵. ارزیابی نتایج

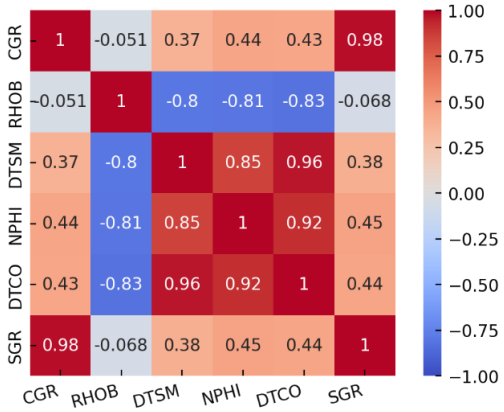
اکنون که خوشه‌ها با استفاده از روش‌های *KMeans*، *HAC*، *DBSCAN* و *GMM* محاسبه شده‌اند، می‌توانیم نتایج را رسم و تحلیل کنیم. لازم به ذکر است که این روش‌ها بدون نظارت هستند و از داده‌های برچسب‌گذاری شده برای آموزش استفاده نمی‌کنند لذا جهت مقایسه نتایج با نمودارهای لیتولوژی چاه تطبیق داده شده است. در نمودار (شکل ۷) رخساره‌های سنگی (*Lithology*) و نتایج *DBSCAN*، *HAC*، *KMeans* و *GMM* جهت مقایسه نشان داده می‌شود. همچنین در شکل ۸

با ماتریس در هم‌ریختی *(Confusion matrix)* ارتباط مثبت و منفی داده‌های چاه‌نگاری نشان داده می‌شود که در این شکل

زمان عبوری موج فشاری و برشی با تخلخل ارتباط مستقیم دارد و با چگالی که قابل انتظار است ارتباط منفی را نشان می‌دهد.



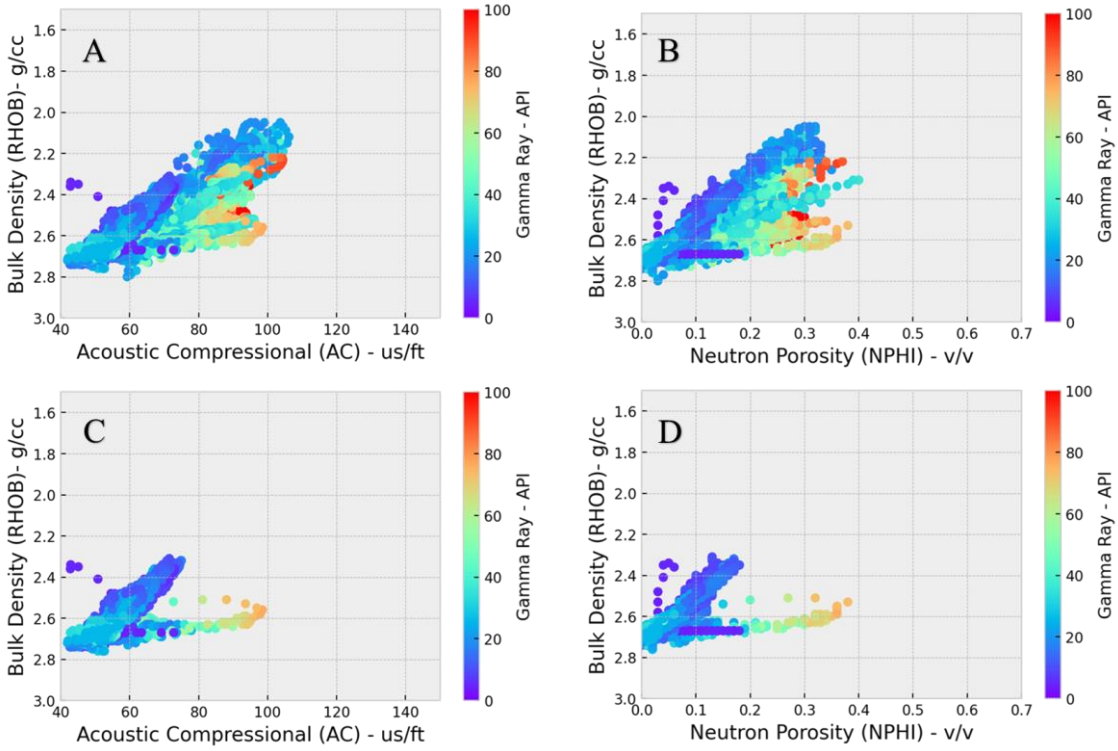
شکل ۷. اولین چیزی که باید به آن توجه داشت این است که رنگ‌ها لزوماً به این معنی نیستند که داده‌ها از یک گروه در هر روش هستند. در این شکل مقایسه روش‌های ماشین لرنینگ شامل DBSCAN، GMM،HAC،k-means و یک ستون تحت عنوان لیتولوژی جهت شناسایی سازندهای مدل که از عمق ۳۲۰۰ تا ۳۸۰۰ سازند بنگستان می‌باشد



شکل ۸. در این ماتریس در هم‌ریختی (Confusion matrix) **ارتباط مثبت و منفی داده‌های چاه‌نگاری نشان داده می‌شود**

در نمودار شکل ۷، اولین چیزی که باید به آن توجه داشت این است که رنگ‌ها لزوماً به این معنی نیستند که داده‌ها از یک گروه در هر روش هستند. در این شکل مقایسه روش‌های ماشین لرنینگ شامل *DBSCAN، GMM،HAC،k-means* و یک ستون تحت عنوان لیتولوژی جهت شناسایی سازندهای مدل که از عمق ۳۲۰۰ تا ۳۸۰۰ سازند بنگستان می‌باشد. این یک سازند آهکی - دولومیتی با رخساره‌های پکستون و گرینستون دارای تخلخل حفره‌ای و شکافی تعیین شده است که جهت مقایسه با مدل با رنگ آبی کدگذاری شده است. همان‌طور که مشخص است در روش خوشه‌بندی *k* - میانگین

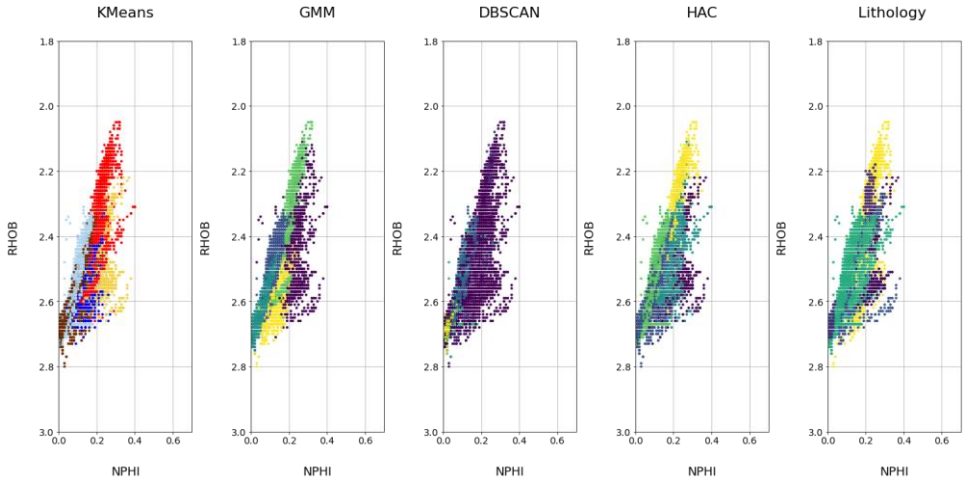
و خوشه‌بندی سلسله مراتبی پیش‌بینی سازند بنگستان به‌صورت یک سازند تقریباً مجزا به‌درستی نشان داده می‌شود با این وجود تفکیک بهتری را نشان نمی‌دهد. اما در روش *GMM* رخساره بنگستان را به‌صورت مشخص‌تر که در قسمت بالا و پایین سازند بنگستان با رنگ زرد متمایل به قهوه‌ای روشن شامل لایه‌های آهکی - دولومیتی دارای تخلخل بالا که سازندهای مخزنی ایلام و سروک را مشخص کرده است و رنگ قهوه‌ای روشن لایه‌های آهکی - دولومیتی با تخلخل کمتر (وکستون) و یک رخساره با رنگ قهوه‌ای تیره تفکیک کرده است. همچنین شناسایی لایه‌های ماسه‌سنگی سازند آسماری و میان لایه‌های آهک‌های نازک‌لایه در این سازند که مسئول ناپایداری چاه است به‌درستی شناسایی شد. روش *DBSCAN* جهت مقایسه با این روش‌ها انتخاب گردیده است. اما نتایج آن دل‌چسب نیست علت این است این روش برای داده‌های پراکنده دارای نویز بسیار خوب عمل می‌کند در الگوریتمِ *DBSCAN* دو پارامتر وجود دارد. یکی از آن‌ها شعاع است که به آن *Epsilon* نیز می‌گویند و دومی حداقل نقاط موجود در یک خوشه است که به آن *MinPoints* می‌گویند که برای نتایج بهتر باید این اعداد با تکرارهای مختلف عوض گردد تا بهترین نتیجه را به کاربر بدهد. مشاهده می‌گردد که زمان موج p با تخلخل رابطه مستقیم و با چگالی رابطه معکوس دارد که تناسب خوبی را نشان می‌دهد. اما این پارامترها با لاگ گاما رابطه معناداری را نشان نمی‌دهد. به‌طور مثال در شیل‌ها تخلخل بالا/نفوذپذیری پایین است. درنتیجه لاگ گاما بالا ولی در آهک‌های سازند بنگستان مقدار گاما پایین و تخلخل بالا است. درنتیجه، سنگ‌شناسی بیشتر از تخلخل و چگالی تبعیت می‌کند. جهت شناخت بیشتر مخزن بنگستان با استفاده از نمودار نوترون-چگالی و کندی موج فشاری- چگالی این قسمت از مخزن را از بقیه مخازن جدا شده است. همان‌طور که مشخص است در این نمودار مقدار گاما کم می‌باشد که نشان‌دهنده مخزن آهکی – دولومیتی می‌باشد (شکل ۹).



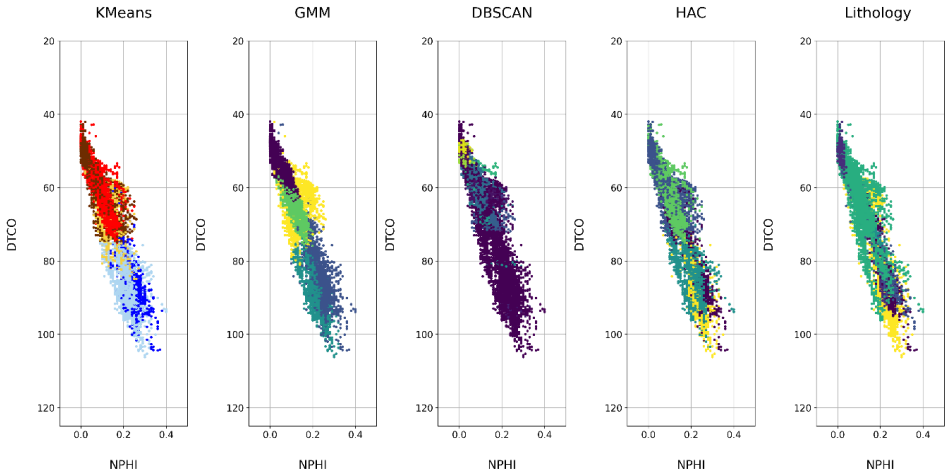
شکل ۹. نمودار A و B مربوط به کل ستون چاه می‌باشد. اما در نمودار C و D جهت شناخت بیشتر مخزن بنگستان از عمق ۳۲۰۰ تا ۳۸۰۰ جدا شده است

۵،۱ نمودارهای پراکنده (Scatter plots / Cross plots)

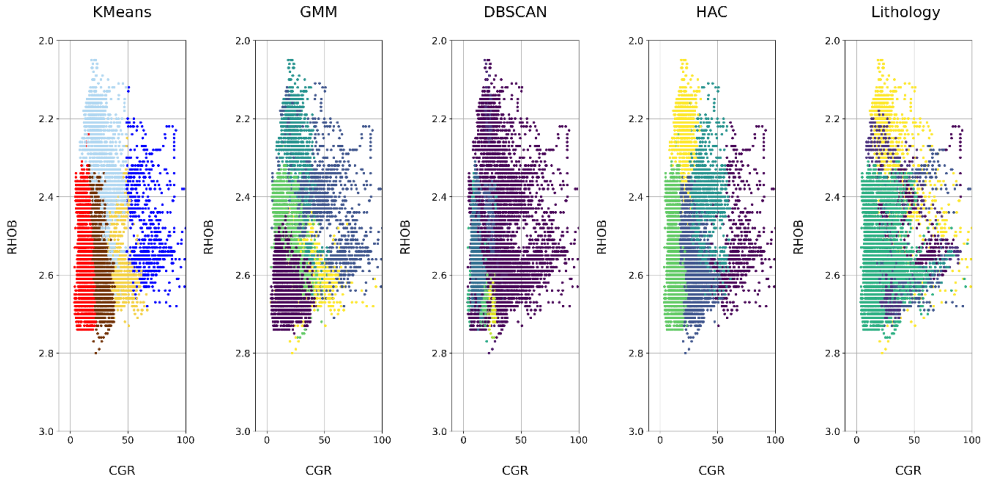
راه دیگر برای مشاهده عملکرد خوشه‌بندی از طریق نمودارهای پراکنده *(Scatter plots / Cross plots)* است. این کار را با استفاده از نمودارهای پراکنده متقاطع چگالی _تخلخل نوترونی (شکل ۱۰)، ارتباط تخلخل با زمان موج P (شکل ۱۱)، ارتباط گاما با چگالی (شکل ۱۲) و ارتباط تخلخل- نوترونی با گاما نشان داده می‌شود (شکل ۱۳).



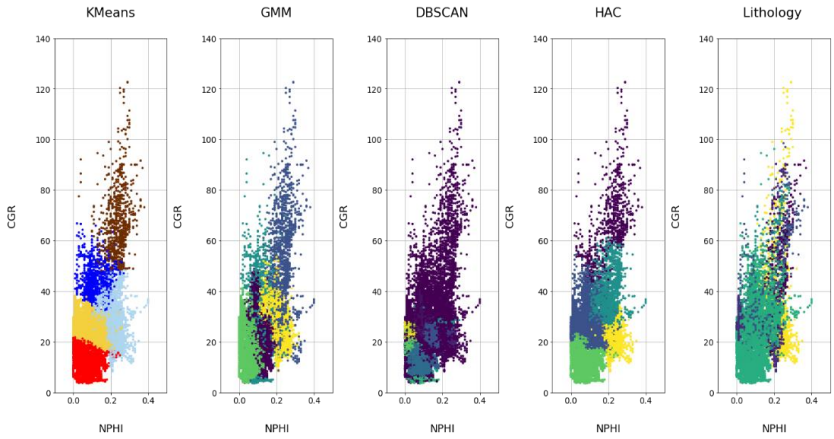
شکل ۱۰. مقایسه روش‌های ماشین لرنینگ به‌طور مثال در روش KMeans و HAC درگوشه پایین سمت چپ جایی که چگالی بالا و نوترون پایین است اختلاط دیده می‌شود ولی در روش GMM با تفکیک بهتر و مطابقت بیشتری نشان داده می‌شود.



شکل ۱۱. از بین روش‌های فوق روش خوشه‌بندی سلسله مراتبی، k- میانگین و مدل آمیخته گوسی بهتر از بقیه عمل کرده است



شکل ۱۲. از بین روش‌های فوق روش خوشه‌بندی سلسله مراتبی، k- میانگین و مدل آمیخته گوسی بهتر از بقیه عمل کرده است. با این وجود در روش GMM تفکیک واحدهای ژئومکانیکی و شناسایی مبتنی بر واریانس می‌باشد.

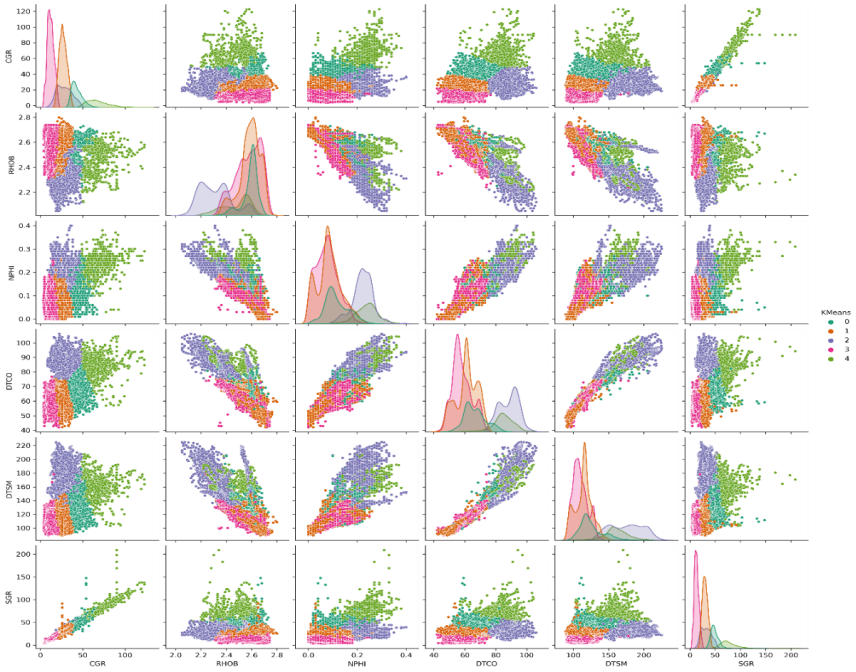


شکل ۱۳. از بین روش‌های فوق روش خوشه‌بندی سلسله مراتبی، k- میانگین و مدل آمیخته گوسی بهتر از بقیه عمل کرده است. مزیت بهتر روش GMM انتخاب چندین الگو احتمالاتی مختلف در زمان خروجی از پایتون می‌باشد.

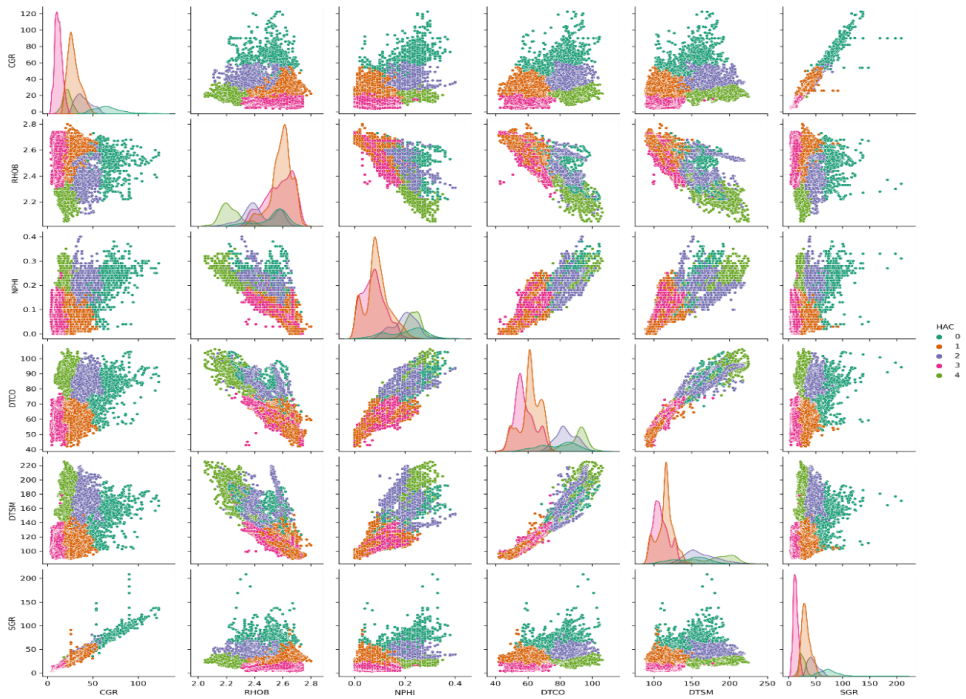
۲،۵ نمودارهای جفتی (Pair plot)

در این نمودارها از شش منحنی ورودی برای مدل‌سازی استفاده شده است تا تفاوت خوشه‌ها نسبت به هم و ارتباط هر خوشه با پارامترهای به‌کار گرفته شده مشخص‌تر باشد. بهترین راه برای انجام این کار استفاده از نمودارهای زوجی می‌باشد. این نمودار روابط بین داده‌های مجموعه هر داده را در یک شبکه نمایش می‌دهد. همان‌طور که مشخص است این یک راه سریع و آسان برای شناسایی و تجسم داده‌ها می‌باشد. در امتداد مورب توزیع داده‌های تقسیم‌شده توسط خوشه نیز رسم شده است. در نمودار *k* - میانگین شکل ۱۴ و خوشه‌بندی سلسله مراتبی شکل ۱۵ زون‌های با رنگ قهوه‌ای کمرنگ، صورتی و آبی به ترتیب بیشترین حجم سازندها را به خود اختصاص داده‌اند و اختلاط خوشه‌ها در بعضی از این نمودارها به چشم می‌خورد. اما در شکل ۱۶ نمودار *GMM* تفکیک خوشه خیلی مشخص‌تر و بهتر نمایان شده است. ملاحظه می‌شود که مدل *GMM* بهبودهایی را در تعریف خوشه‌ها به‌خصوص در نمودار *DTC* در مقابل *RHOB*ارائه می‌دهد. بنابراین با توجه به این نمودارها و تحقیقات آزمایشگاهی سازند بنگستان از سه نوع لیتولوژی آهکی- دولومیتی با تخلخل بالا، متوسط و پایین تقسیم‌بندی می‌شود که در مقایسه با رخساره‌های رسوب‌شناسی به ترتیب شامل پکستون- گرینستون دارای تخلخل حفره‌ای و شکافی، گرینستون دارای تخلخل شکافی و وکستون- پکستون دارای تخلخل کمتر معادل است.

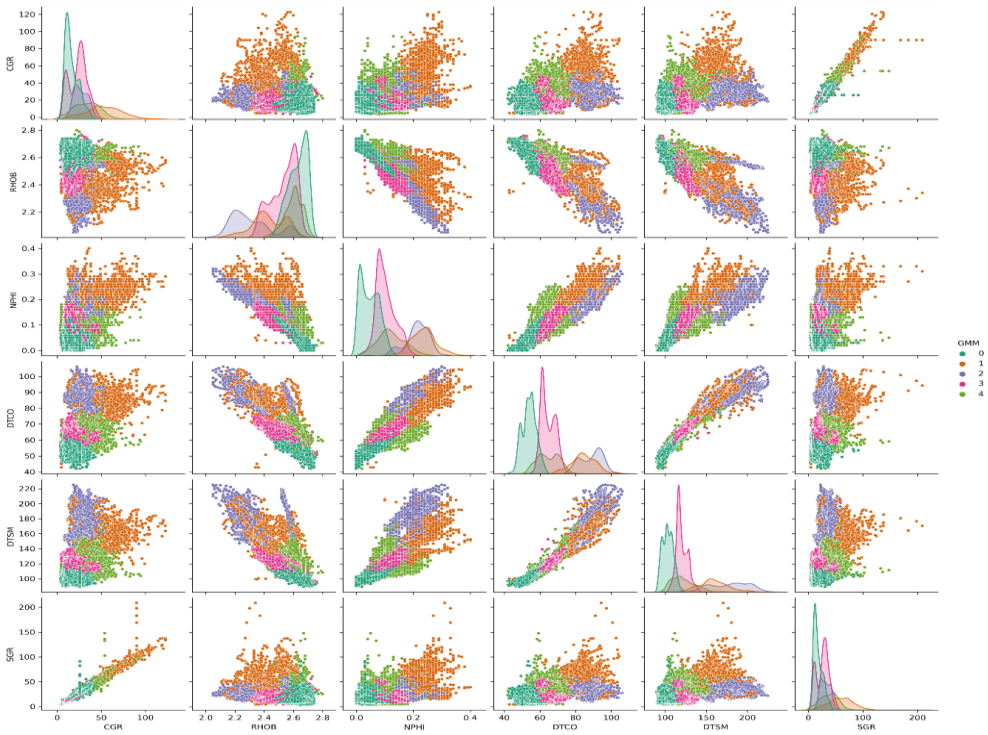
الگوریتم‌های *K-Means*، *Hierarchical Clustering Agglomerative Clustering* و *Gaussian Mixture Modeling* سه روش کاربردی در تحلیل خوشه‌بندی بدون نظارت یادگیری ماشین هستند اما در واقعیت‌های پتروفیزیکی، زمین‌شناسی و ژئومکانیکی داده‌ها به‌ندرت الگوهای دایره‌ای خوبی را تشکیل می‌دهند. درحالی‌که اگر خوشه‌های داده دایره‌ای باشند، خوشه‌بندی *K-Means* عالی عمل می‌کند. بنابراین بهتر است از روش مدل‌های مخلوط گاوسی (*GMM*) که در تشخیص لیتولوژی تفکیک بهتر و مطابقت بیشتری با سازند زمین‌شناسی دارد. استفاده شود. درنتیجه با توجه مطالب گفته‌شده، شکل ۱۷ جهت شناسایی لایه زمین‌شناسی ارائه شده است. در این شکل لایه‌های زمین به پنج واحد تقسیم‌بندی شده است. با توجه به بررسی‌های زمین‌شناسی *Facies1* با رنگ زرد معادل پکستون های با تخلخل بالا مربوط به آهک و دولومیت‌های سازند آسماری، جهرم، ایلام و سروک می‌باشد. لایه‌های مخزنی پرفشار ایلام و سروک بعد از بررسی و اعتبار سنجی با تحقیقات میدانی به‌درستی تشخیص داده شده است، *Facies2* با رنگ سبز کم‌رنگ مربوط به شیل و ماسه‌سنگ‌های گورپی می‌باشد، *Facies3* بارنگ سبز تیره با توجه به درصد کم رس مربوط به میان لایه‌های آهک نازک‌لایه سازند آسماری و گرینستون‌های محیط سدی سازند سروک می‌باشد *Facies4* مربوط به وکستون‌های سازند سروک و *Facies5* مربوط به ماسه‌سنگ‌های ناپایدار سازند آسماری می‌باشد. همان‌طور که مشخص است میان لایه‌های آهک نازک‌لایه دارای درز و شکاف که مسئول اصلی ناپایداری دیواره چاه می‌باشد در این واحد به‌درستی تشخیص داده شده است علت ریزش دیواره چاه (ناپایداری) در لاگ کالیپر به‌درستی در سازند آسماری دیده می‌شود و همین‌طور لایه‌های مخزنی پرفشار ایلام و سروک به‌درستی نشان داده شده است. درنتیجه با توجه به این تحقیق چهار روش یادگیری ماشین به‌کار گرفته شد که با وجود عملکرد خوب روش خوشه‌بندی سلسله مراتبی، مدل کی میانگین و مدل آمیخته گوسی، روش مدل مخلوط گوسی جهت پیش‌بینی دقیق‌تر زون‌های ژئومکانیکی پیشنهاد داده می‌شود. جهت شناخت بیشتر این زون‌ها درصد و فراوانی این زون‌ها در شکل ۱۸ نشان داده می‌شود.



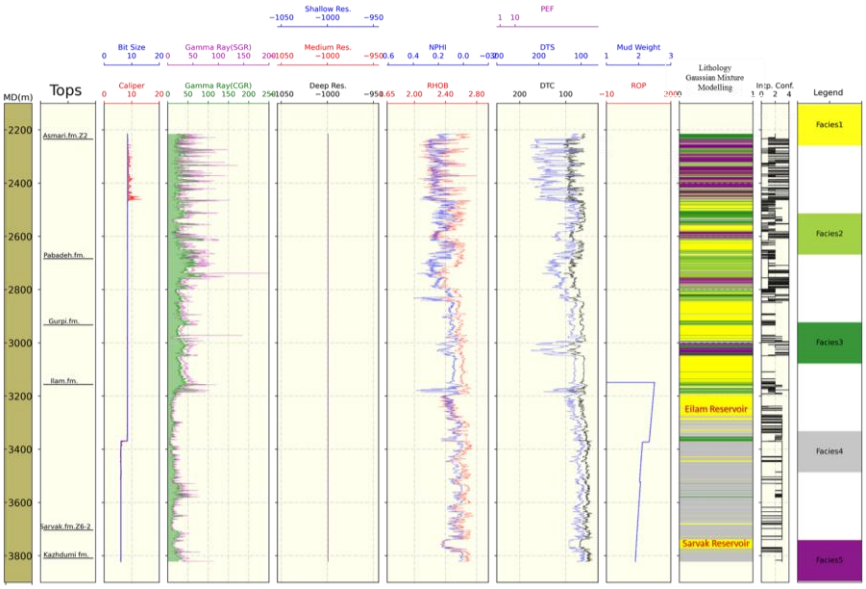
شکل ۱۴. با توجه به نمودار کی میانگین تفکیک خوشه‌ها را خوب نشان داده است. اما در بعضی نمودارها مثل DTCO- RHOB اختلاط خوشه‌ها دیده می‌شود.



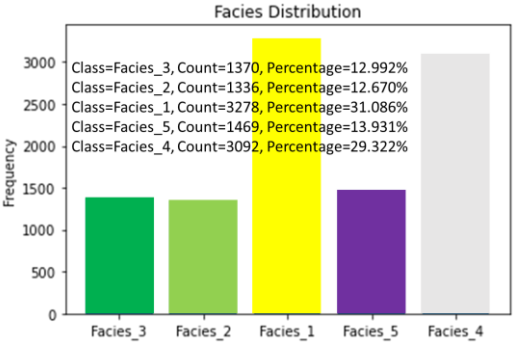
شکل ۱۵. با توجه به نمودار خوشه‌بندی سلسله مراتبی (HAC) تفکیک خوشه‌ها خوب نشان داده شده است.



شکل ۱۶. در این روش داده‌ها بر اساس چندین احتمال مختلف خوشه‌بندی می‌شوند و نسبت به روش‌های دیگری انعطاف‌پذیر می‌باشد



شکل ۱۷. استفاده از روش (GMM) در تعیین واحدهای ژئومکانیکی در یکی از چاه‌های جنوب ایران. در این تصویر میان لایه‌های آهک نازک‌لایه دارای درز و شکاف با رنگ سبز و خاکستری که مسئول اصلی ناپایداری دیواره چاه در سازند آسماری (Facies 5) می‌باشد. همین‌طور سازندهای مخزنی پرفشار ایلام و سروک با رنگ زرد به‌درستی شناسایی شدند.



شکل ۱۸. در این شکل درصد و فراوانی واحدهای ژئومکانیکی نشان داده می‌شود.

۶. نتیجه‌گیری

این مقاله روش انجام تحلیل خوشه‌ای بدون نظارت با استفاده از چهار الگوریتم *DBSCAN*، *HAC*، *K-Means Clustering* و *Gaussian Mixture Modeling* پوشش داده شده است. در مرحله اول، بعد از پردازش داده‌های چاه‌نگاری از دو الگوریتم محبوب قدرتمند نظارت‌شده عمیق یادگیری ماشین ایکس جی بوست *(XGBoost)* و شبکه عصبی پرسپترون چندلایه *Percepteron Neural Multi-Layer Network* جهت بازیابی داده‌های گمشده استفاده گردید. سپس از روش‌های بدون نظارت یادگیری ماشین شامل مدل خوشه‌بندی سلسله مراتبی (*HAC*)، مدل کی میانگین (*K-Means Clustering*)، الگوریتم خوشه بندی *DBSCAN* مبتنی بر غلظت و مدل آمیخته گوسی *(Gaussian Mixture Modelling)* جهت تعیین واحدهای ژئومکانیکی غیر مخزنی مسئله‌دار، آهک‌های نازک‌لایه و مخزنی پرفشار استفاده شد. در اولین قدم با استفاده از یک روش بهینه‌سازی و با استفاده از نمودار زانویی و همین‌طور استفاده از نمودار دندروگرام *(Dendrogram)* از روش خوشه‌بندی سلسله مراتبی (*Agglomerative hierarchical Clustering*) بهترین تعداد خوشه‌ها تعیین شد که تعداد بهینه خوشه‌ها عدد پنج انتخاب گردید. مقایسه نتایج نشان داد ازآنجایی‌که در روش خوشه‌بندی *HAC* و *K-Means* جهت پیش‌بینی واحدهای ژئومکانیکی از خوشه‌های کروی استفاده می‌کنند، جهت مدل‌سازی مناسب‌تر است از الگوریتم *GMM* که از الگو مرزهای خوشه/تصمیم بیضی شکل استفاده می‌کند، انعطاف‌پذیرتر است و سازگاری بیشتری با داده‌های مطالعه حاضر نشان می‌دهد. بنابراین مدل *GMM* به‌عنوان روش مناسب‌تر جهت تعیین واحدهای غیر مخزنی مسئله‌دار، آهک‌های نازک‌لایه در سازند آسماری و سازندهای مخزنی پرفشار ایلام و سروک معرفی شد. همین‌طور با این مدل بخش‌های پرفشار سازندهای مخزنی ایلام و سروک با دقت خوب شناسایی شدند.

۷. مراجع:

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

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

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

Ashraf, U., Zhang, H., Anees, A., Ali, M., Zhang, X., Abbasi, S.S., Mangi, H. N., (2020b). Controls on Reservoir Heterogeneity of a Shallow-Marine Reservoir in Sawan Gas Field, SE Pakistan: Implications for Reservoir Quality Prediction Using Acoustic Impedance Inversion. Water 12 (11), 2972.

https://doi.org/10.3390/w12112972. In this issue

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.

Forgy, E. W., (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics.

Jinghua, W., Jianmin J., (2021). Unsupervised deep clustering via adaptive GMM modeling and optimization. Neurocomputing 433 (2021) 199-211

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

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

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

Kriegel, H.P., Schubert, E., Zimek, A., (2016). The (black) art of runtime evaluation: Are we comparing algorithms or implementations. Knowledge and Information Systems. 52: 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377

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

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

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281&ndash, 297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

MacKay, D., (2003). "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284&ndash, 292. ISBN 0-521-64298-1. MR 2012999.

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

Steinhaus, H. (1957). "Sur la division des corps matériels en parties". Bull. Acad. Polon. Sci. (French). 4 (12): 801&ndash, 804. MR 0090073. Zbl 0079.16403.

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

W. Dunham, 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