
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی

رشته مهندسی کامپیوتر گرایش نرم افزار - هوش مصنوعی - شبکه های کامپیوتری

عنوان پایان‌نامه:

رویکرد ترکیبی یادگیری ماشین برای پیشرفت در شناسایی مناطق غنی از مواد معدنی با به‌کارگیری تقویت گرادیانی از طریق داده‌های دورسنجی

استاد راهنما:

دکتر فریا نصیری مفخم

پژوهشگران:

ملیکا آقاجانیان صباغ ۹۹۳۶۲۳۰۰۴

مهدیس فتحی ۹۹۳۶۱۳۰۴۹

شهریور ۱۴۰۳

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پروژه کارشناسی رشته‌ی مهندسی کامپیوتر گرایش نرم افزار - هوش مصنوعی - شبکه های کامپیوتری خانم‌ها ملیکا آقاجانیان صباغ و مهدیس فتحی تحت عنوان **رویکرد ترکیبی یادگیر ی ماشین برا ی پیشرفت در شناسایی مناطق غنی از مواد معدنی به کارگیری تقویت گرادیانی برا ی افزا یش کارایی در شناسایی مناطق غنی از آهن از طریق داده های دورسنجی** در تاریخ **۰۶/ / ۱۴۰۳** توسط هیأت داوران زیر بررسی و با نمره به تصویب نهایی رسید.

۱- استاد راهنمای پروژه:

دکتر فریا نصیری مفخم امضا

۲- استاد داور :

دکتر امضا

امضای مدیر گروه

تشکر و قدردانی

با تشکر و قدردانی از زحمات دکتر فریا نصیری مفخم، جناب آقای مهندس رحمتی و سرکارخانم مهندس رادگهر، برای تمامی راهنمایی‌ها و همکاری‌های بالقوه ایشان از ابتدا تا پایان انجام این پروژه.

با احترام و سپاس فراوان،

ملیکا آقاجانیان صباغ و مهدیس فتحی

تقدیم به

کودکان کار سرزمینان ایران

چکیده:

این پروژه به بررسی و شناسایی مناطق با پتانسیل معدنی، به‌ویژه منابع غنی از آهن، با استفاده از مدل‌های یادگیری ماشین از جمله تقویت گرادیانی (XGBoost)، درخت تصمیم‌گیری (Decision Tree) و الگوریتم خوشه‌بندی k-means می‌پردازد. داده‌های ورودی این مطالعه شامل تصاویر دورسنجی حاصل از ماهواره‌هایی مانند Aster و دیگر ماهواره‌ها است که ویژگی‌های طیفی و مکانی متنوعی را در باندهای مختلف پوشش می‌دهند. هدف اصلی این پروژه افزایش دقت پیش‌بینی الگوریتم‌ها در شناسایی کانسارهای آهن است و به‌منظور دستیابی به این هدف، پروژه در دو مرحله اصلی اجرا می‌شود: ۱) پیش‌پردازش داده‌ها و استخراج ویژگی‌ها، ۲) پیاده‌سازی و آموزش مدل نهایی.

در مرحله اول، داده‌ها با استفاده از تکنیک‌های مختلف پیش‌پردازش، مانند تبدیل داده‌های نامعتبر به صفر و مقیاس‌بندی مقادیر بین بازه صفر تا یک، آماده‌سازی می‌شوند. سپس ویژگی‌ها و شاخص‌های مورد نیاز از داده‌ها استخراج می‌شوند. در مرحله دوم، داده‌ها با استفاده از الگوریتم k-means به دسته‌های آهن و غیرآهن تقسیم می‌شوند. پس از آن مجموعه داده‌ها به دو بخش آموزشی ۷۰٪ و آزمون ۳۰٪ تقسیم شده و سپس، با تنظیم پارامترهای مدل، الگوریتم‌های درخت تصمیم‌گیری و تقویت گرادیانی پیاده‌سازی و ارزیابی می‌شوند. نتایج حاصل از این پیاده‌سازی دارای دقتی نزدیک به ۹۸٪ بوده که این نشان‌دهنده‌ی عملکرد موفقیت‌آمیز مدل در تشخیص مناطق معدنی از غیرمعدنی است. نتایج این پژوهش می‌تواند به کاهش هزینه‌های مالی و خطرات حفاری کمک کرده و تمرکز بیشتری بر شناسایی مناطق امیدبخش برای اکتشافات معدنی بگذارد.

واژگان کلیدی: شناسایی معادن آهن، داده‌های دورسنجی، خوشه بندی، یادگیری ماشین، تقویت گرادیانی، درخت تصمیم

فهرست مطالب

فصل اول – مقدمه ۱۱

۱-۱- هدف پروژه ۱۱

۱-۲- کاربردهای پروژه ۱۱

۱-۳- ساختار پایان نامه ۱۲

فصل دوم - مفاهیم ۱۳

۲-۱- مقدمه ۱۳

۲-۲- الگوریتم k-means ۱۳

۲-۲-۱- مراحل الگوریتم k-means ۱۳

۲-۳- الگوریتم Decision Tree ۱۵

۲-۳-۱- مراحل الگوریتم Decision Tree ۱۵

۲-۳-۲- شرایط پایان الگوریتم درخت تصمیم‌گیری ۱۶

۲-۴- الگوریتم XGBoost ۱۷

۲-۴-۱- مراحل الگوریتم XGBoost ۱۷

۲-۵- جمع‌بندی ۱۸

فصل سوم - شرح پروژه ۱۹

۳-۱- مقدمه ۱۹

۳-۲- معماری سیستم پیش‌بینی کانسار آهن با استفاده از تصاویر ماهواره‌ای ۱۹

۳-۳- روش استفاده شده در این پروژه برای پیش‌بینی مناطق دارای کانسار آهن ۲۰

۳-۳-۱- جمع آوری داده ۲۰

۳-۳-۲- پیش‌پردازش داده ۲۱

۳-۳-۳- استخراج ویژگی ۲۲

۳-۳-۴- برچسب گذاری داده ۲۴

۳-۳-۵- اجرا و آزمون مدل تقویت گرادیانی ۲۴

۳-۴- جمع‌بندی ۲۷

فصل چهارم - پیاده‌سازی و نتایج ۲۸

۴-۱- مقدمه ۲۸

۴-۲- جزئیات پیاده‌سازی ۲۸

۴-۳- بارگذاری تصویر و پردازش اولیه ۲۹
۴-۴- پیش‌پردازش داده‌ها ۳۱
۴-۴-۱- نرمال سازی تصویر ۳۱
۴-۴-۲- جمع‌آوری داده و استخراج ویژگی ۳۳
۴-۵- ایجاد نمونه ۳٤
۴-۶- برچسب‌گذاری داده‌ها ۳٦
۴-۷- پیاده سازی الگوریتم XGboost ۳۸
۴-۸- جمع‌بندی ٤۱
فصل پنجم - نتیجه‌گیری و پیشنهادات ٤۲
۵-۱- نتیجه‌گیری ٤۲
۵-۲- پیشنهادات برای بهبود و استفاده‌های آینده ٤۳
پیوست ۱ ..... ٤٤
منابع: ٤٥

فهرست تصاویر

تصویر ۱ پیاده‌سازی الگوریتم Kmeans ۱٤
تصویر ۲پیاده سازی الگوریتم Decision Tree ۱٦
تصویر ۳ پیاده‌سازی الگوریتم XGBoost ۱۸
تصویر ۴ نحوه محاسبه دقت مدل ۲٥
تصویر ۵ نحوه محاسبه دقت طبقه‌بندی ۲٦
تصویر ۶ نحوه محاسبه نرخ بازخوانی ۲٦
تصویر ۷ نحوه محاسبه f1-score ۲٦
تصویر ۸ کتابخانه‌های استفاده شده ۲۹
تصویر ۹ کد بررسی اطلاعات تصویر ۲۹
تصویر ۱۰ نمونه خروجی از اطلاعات تصویر ۳۰
تصویر ۱۱ بارگذاری یکی از تصاویر برای ادامه پروژه ۳۰
تصویر ۱۲ حذف مقادیر نامعتبر از تصویر و جایگزاری آن‌ها با مقدار صفر ۳۱
تصویر ۱۳ مقیاس‌بندی مقادیر باندها بین ۰ و ۱ ..... ۳۱
تصویر ۱٤ سه باند اول به صورت تصویر RGB ۳۲
تصویر ۱۵ معرفی و نام‌گذاری باندها ۳۳
تصویر۱۶ استخراج ویژگی‌ها ۳۳
تصویر ۱۷ شاخص‌های تصویر ۳٤
تصویر ۱۸ ایجاد نمونه ۳٤
تصویر۱۹ نمونه ایجاد شده ۳٥
تصویر۲۰ پیاده سازی Kmeans ۳٦
تصویر۲۱ نتایج خوشه‌بندی ۳٦
تصویر۲۲ تصویر خوشه‌بندی ۳۷
تصویر۲۳ پیاده سازی الگوریتم XGboost ۳۸
تصویر ۲۴ شاخص‌های ارزیابی ۳۹
تصویر۲۵ دقت و نتایج به دست آمده از مدل ۳۹
تصویر ۲۶ نتایج به دست آمده روی مجموعه آزمون ٤۰
تصویر۲۷ جدول نتایج ارزیابی مدل ٤۲

مخفف‌ها:

ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
SWIR	Short-wave infrared
NDVI	Normalized difference vegetation index
NDWI	Normalized Difference Water Index
NIR	Near-infrared spectroscopy
VNIR	Visible and Near-Infrared
TIR	Thermal Infrared

فصل اول – مقدمه

۱-۱- هدف پروژه

یکی از مهم‌ترین مراحل در اکتشاف و حفاری معادن و ذخایر معدنی، شناسایی مناطق امیدبخش با پتانسیل بالا برای ذخایر آهن است. برای دستیابی به این هدف، استفاده از الگوریتم‌های هوش مصنوعی و یادگیری ماشین، که قابلیت افزایش دقت پیش‌بینی و شناسایی مناطق معدنی را دارند، بسیار مورد توجه قرار گرفته است. این پروژه با استفاده از داده‌های دورسنجی با کیفیت بالا از ماهواره‌ی استر[1] که دارای ویژگی‌های طیفی، مکانی و باندهای مختلف است، به آموزش و بهبود مدل نهایی می‌پردازد.

هدف اصلی این پروژه، افزایش دقت پیش‌بینی در شناسایی مناطق دارای کانسار آهن است. به همین منظور، قبل از پیاده‌سازی مدل نهایی با استفاده از الگوریتم‌های تقویت گرادیانی[2] و درخت تصمیم‌گیری[3]، داده‌ها پیش‌پردازش شده و ویژگی‌های اصلی استخراج می‌شوند. سپس بهترین ویژگی‌ها برای پیاده‌سازی مدل و تنظیم پارامترهای اصلی انتخاب می‌شوند تا به دقت بالاتر و کارایی بهتر دست یابیم.

با این روش، امکان شناسایی دقیق‌تر و بهینه‌تر مناطق غنی از آهن فراهم می‌شود که به کاهش هزینه‌ها و ریسک‌های مرتبط با اکتشاف و حفاری منجر خواهد شد. این پروژه می‌تواند به عنوان یک نمونه موفق در استفاده از هوش‌مصنوعی برای بهبود فرآیندهای صنعتی و معدنی در نظر گرفته شود.

۱-۲- کاربردهای پروژه

نتایج این پروژه می‌تواند برای کارشناسان و شرکت‌های معدنی و اکتشافی بسیار ارزشمند و سودمند باشد. با استفاده از مدل‌های پیش‌بینی دقیق، این سازمان‌ها به‌ویژه نهادهای دولتی، قادر خواهند بود هزینه‌های زمانی، مالی و انسانی را کاهش داده و همچنین ریسک‌ها و خطرات ناشی از حفاری و اکتشافات معدنی را به حداقل برسانند.

به‌طور خاص، مدل‌های توسعه‌یافته می‌توانند به بهینه‌سازی تصمیمات در مراحل مختلف اکتشاف و حفاری کمک کنند. این دستاوردها باعث بهینه‌سازی فرآیندهای اکتشافی و افزایش بازدهی در شناسایی منابع معدنی می‌شوند. در نتیجه، استفاده از این روش‌ها نه تنها به بهبود دقت در شناسایی مناطق امیدبخش کمک می‌کند بلکه باعث کاهش هزینه‌های مربوط به عملیات‌های اکتشافی و حفاری نیز می‌شود.

---

[1] ASTER
[2] Gradient Boosting
[3] Decision Tree

۱-۳- ساختار پایان نامه

این پروژه شامل پنج فصل اصلی است که به تفصیل مراحل مختلف پروژه را پوشش می‌دهند. به‌طور کلی، فصل اول به کلیات و معرفی پروژه اختصاص دارد، در این فصل، به بیان اهداف و ضرورت‌های انجام این پروژه پرداخته شده است. در فصل دوم، به معرفی داده‌های دورسنجی، زبان برنامه‌نویسی، کتابخانه‌ها و الگوریتم‌های استفاده‌شده به‌ویژه الگوریتم درخت تصمیم و تقویت گرادیانی، پرداخته شده است. این فصل به تشریح ابزارها و روش‌هایی می‌پردازد که در فرآیند استخراج و تحلیل داده‌ها مورد استفاده قرار گرفته‌اند.

در فصل سوم مراحل مختلف پیش‌پردازش داده‌ها، از جمله تمیزسازی و نرمال‌سازی داده‌ها و همچنین روش‌های مورد استفاده برای استخراج ویژگی‌های مهم توضیح داده شده‌اند.

در فصل چهارم، الگوریتم نهایی پیاده‌سازی می‌شود و مدل انتخاب‌شده مورد ارزیابی قرار می‌گیرد. در این فصل، نتایج به دست آمده از آموزش و تست مدل‌ها مورد بحث قرار گرفته و عملکرد مدل‌ها بر اساس معیارهای مختلف ارزیابی شده است.

در نهایت، فصل پنجم به نتایج حاصل از این پروژه و گام‌های پیشنهادی برای ارتقای الگوریتم در آینده می‌پردازد. این فصل به بیان نقاط قوت و ضعف پروژه پرداخته و پیشنهاداتی برای کارهای آینده ارائه می‌دهد که می‌توانند به بهبود دقت و کارایی مدل‌های پیشنهادی کمک کنند.

فصل دوم - مفاهیم

۲-۱- مقدمه

در این فصل، مفاهیم اساسی و اولیه از ابتدا تا انتهای پیاده‌سازی پروژه معرفی خواهند شد. این فصل به معرفی و نحوه کارکرد الگوریتم‌های استفاده‌شده مانند تقویت گرادیانی، درخت تصمیم، K-means و همچنین بررسی ویژگی‌ها و مزیت‌های آن‌ها اختصاص دارد.

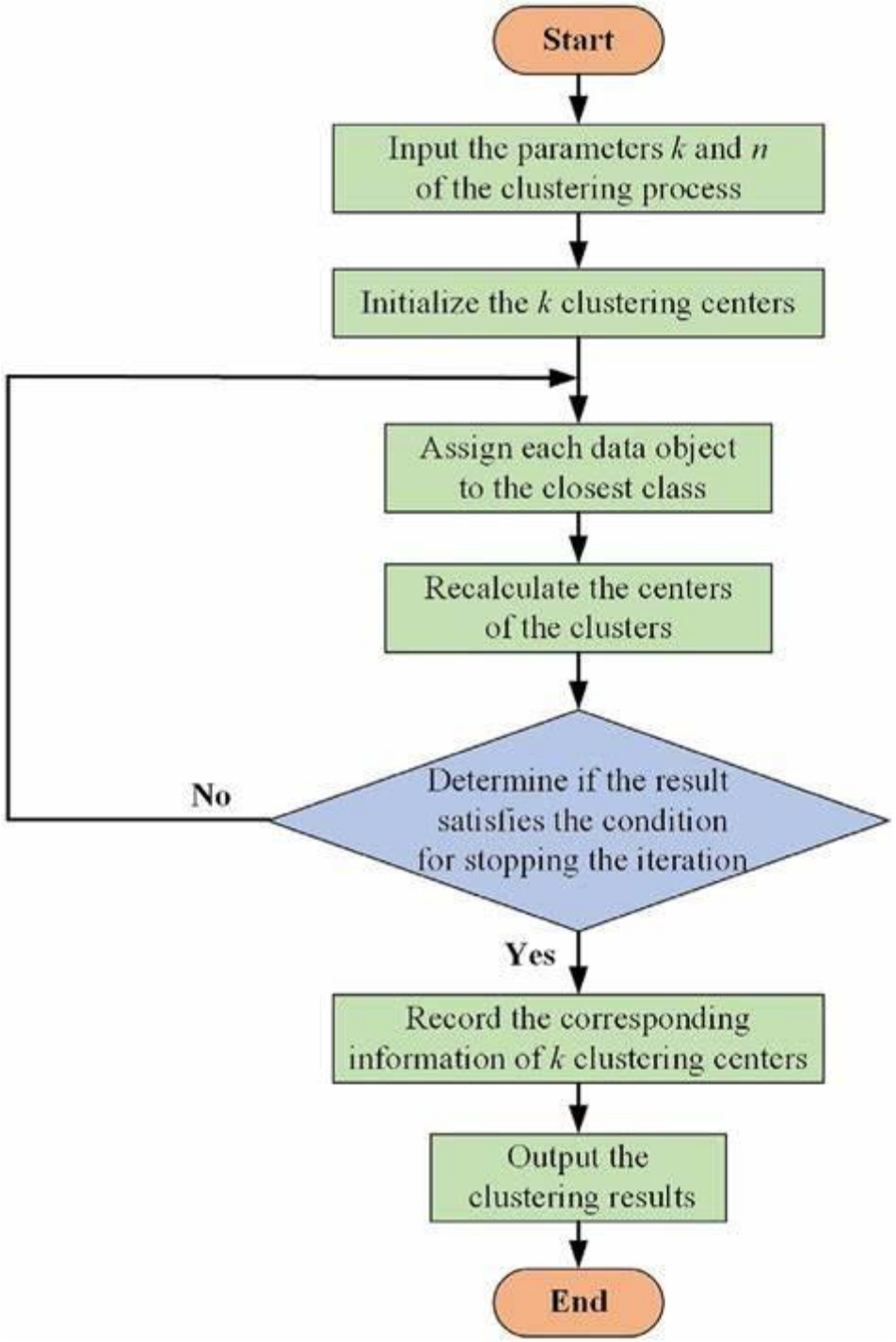
۲-۲- الگوریتم k-means

الگوریتم K-means یکی از معروف‌ترین و کارآمدترین الگوریتم‌های خوشه‌بندی[1] بدون نظارت/برچسب است. این الگوریتم، به دلیل سادگی و سرعت بالایی که دارد در پروژه‌های هوش مصنوعی بسیار محبوب است. با این حال، ضعف اصلی آن وابستگی شدید به نقاط ابتدایی (مراکز اولیه خوشه‌بندی) است. هدف اصلی این الگوریتم، حداکثر کردن شباهت درون‌گروهی و حداقل کردن شباهت بین‌گروهی است، به‌طوری‌که اعضای هر گروه بیشترین شباهت را با یکدیگر و کمترین شباهت را با اعضای سایر گروه‌ها داشته باشند. این الگوریتم مبتنی بر فاصله اقلیدسی است و هر عنصر از مجموعه داده به خوشه‌ای تعلق می‌یابد که کمترین فاصله را تا مرکز آن خوشه داشته باشد.[۳،۴]

۲-۲-۱- مراحل الگوریتم k-means

- انتخاب مقدار خوشه‌ها (k): در این مرحله باید تعداد خوشه‌ها تعیین شود.
- انتخاب مراکز اولیه خوشه‌ها: در این مرحله k نقطه به صورت تصادفی از بین داده‌ها انتخاب می‌شوند، این نقاط در تعیین و شکل گیری نهایی خوشه‌ها نقش به سزایی دارند.
- تخصیص داده‌ها به نزدیک‌ترین مرکز خوشه: در این مرحله فاصله اقلیدوسی هر داده از مرکز خوشه‌ها محاسبه شده و در نهایت این داده به خوشه‌ای که کمترین فاصله را تا مرکز آن دارد تخصیص می‌یابد.
- محاسبه مراکز جدید خوشه‌ها: در این مرحله پس از تخصیص داده‌ها به خوشه‌ها، مرکز هر خوشه با میانگین گرفتن از عناصر آن محاسبه و به روزرسانی می‌شود.
- تکرار مراحل ۳ و ۴: دو مرحله آخر تا جایی ادامه پیدا می‌کنند که مرکز خوشه‌ها تغییر نکنند یا دارای تغییرات ناچیزی باشند؛ این به معنای پایداری خوشه‌ها می‌باشد.

[1] Clustering



تصویر ۱ پیاده‌سازی الگوریتم Kmeans

۲-۳- الگوریتم Decision Tree

الگوریتم درخت تصمیم‌گیری یکی از پرکاربردترین الگوریتم‌های یادگیری ماشین است که برای دسته‌بندی[1] و رگرسیون[2] استفاده می‌شود. این الگوریتم با ساختار درختی عمل می‌کند؛ هر شاخه نشان‌دهنده نتیجه یک آزمون و هر برگ نشان‌دهنده یک برچسب یا مقدار است. سادگی و قابلیت تطبیق با داده‌های دسته‌بندی و عددی، آن را به یکی از کارآمدترین الگوریتم‌ها تبدیل کرده است. با این حال، تغییرات کوچک در داده‌های آموزشی می‌تواند به تغییرات قابل‌توجهی در ساختار درخت منجر شود، و در صورت عمیق شدن درخت، بیش‌برازش[3] و افزایش زمان محاسباتی رخ می‌دهد.[۴]

تعاریف اولیه این الگوریتم به صورت زیر می‌باشد:

- گره: گره‌ها نقاط تصمیم‌گیری در درخت هستند. هر گره نشان‌دهنده یک ویژگی از داده‌ها است که با آن تصمیم‌گیری برای تقسیم داده‌ها انجام می‌شود.
- برگ: گره‌هایی که دیگر تقسیم نمی‌شوند و به عنوان نتیجه نهایی درخت عمل می‌کنند. معمولا در دسته بندی دارای برچسب و در رگرسیون دارای مقدار عددی پیش‌بینی شده هستند.
- شاخه: شاخه‌ها مسیرهایی هستند که گره‌ها را به برگ‌ها متصل می‌کنند. به عبارتی، شاخه‌ها نتیجه آزمون یا پرسشی هستند که در گره انجام شده است.

۲-۳-۱- مراحل الگوریتم Decision Tree

- انتخاب بهترین ویژگی برای تقسیم داده‌ها: در این مرحله ابتدا باید از بین ویژگی‌های استخراج شده از مجموعه‌داده، بهترین ویژگی برای تقسیم بندی و تفکیک داده ها انتخاب شود.
- تقسیم داده‌ها بر اساس ویژگی انتخاب شده: در این مرحله مجموعه داده باید بر اساس ویژگی انتخاب شده به دو یا چند زیرمجموعه تقسیم شود. هر زیرمجموعه یک شاخه از درخت تصمیم‌گیری می‌باشد.
- تکرار فرآیند برای هر زیرمجموعه: برای هر زیرمجموعه، دو مرحله قبل تاز زمانی که یکی از شرایط پایان[4] برآورده شود ادامه می‌یابد.
- تعیین برچسب نهایی (برگ‌ها): هنگامی که شرایط پایان برآورده می‌شود، گره به یک برگ تبدیل می‌شود و برچسب نهایی به آن اختصاص داده می‌شود. در دسته‌بندی، این برچسب معمولاً اکثریت کلاس‌ها در آن گره است و در رگرسیون، میانگین مقادیر هدف در آن گره است.

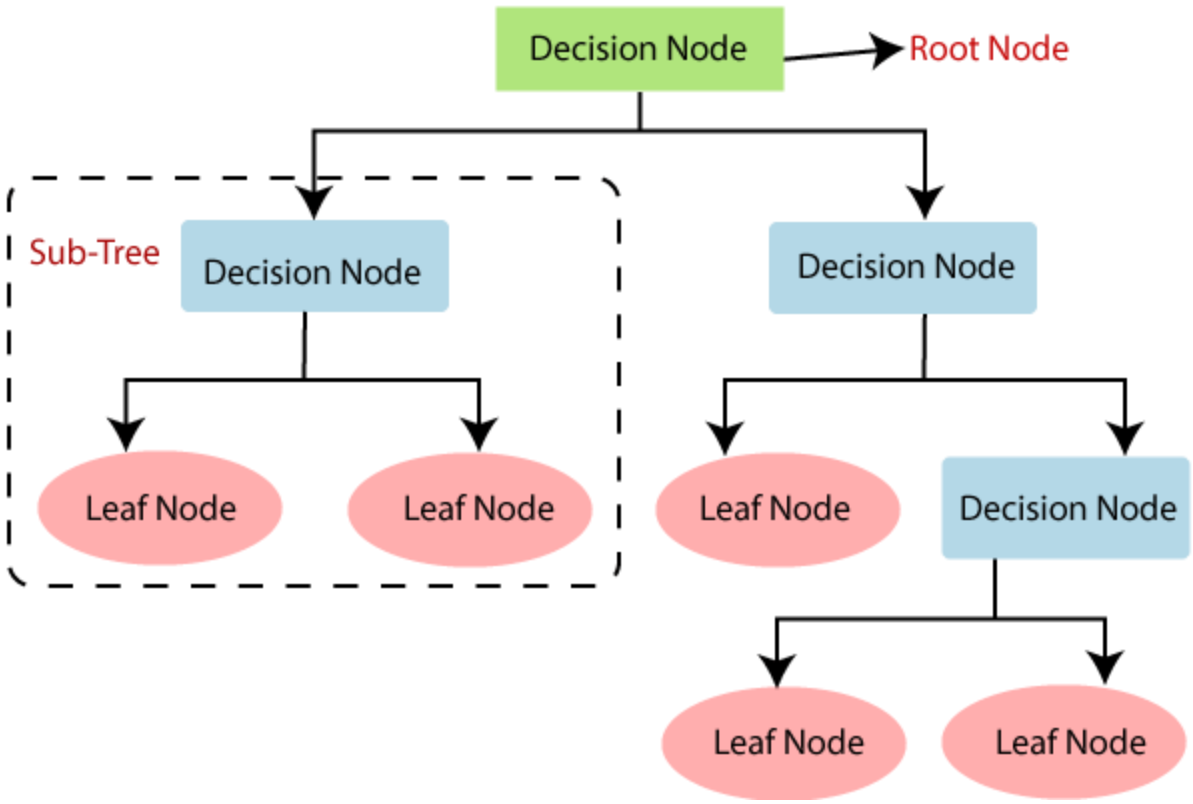
---

[1] Classification
[2] Regression
[3] Overfitting
[4] Stopping Criteria

- پیش‌بینی: پیش‌بینی یک نمونه جدید از ریشه درخت شروع می‌شود و بر اساس ویژگی‌های نمونه و آزمون‌های موجود در گره‌های درخت، به سمت برگ‌های نهایی حرکت می‌کند. برچسب یا مقدار موجود در برگ، نتیجه پیش‌بینی خواهد بود.

۲-۳-۲- شرایط پایان الگوریتم درخت تصمیم‌گیری

- هیچ ویژگی دیگری برای تقسیم باقی نمانده است؛
- عمق درخت به حداکثر مقدار تعیین‌شده رسیده است؛
- تمامی داده‌های یک گره به یک کلاس تعلق دارند (در حالت دسته‌بندی)؛
- تعداد داده‌های باقی‌مانده در یک گره کمتر از یک حد آستانه مشخص است.



تصویر ۲ پیاده سازی الگوریتم Decision Tree

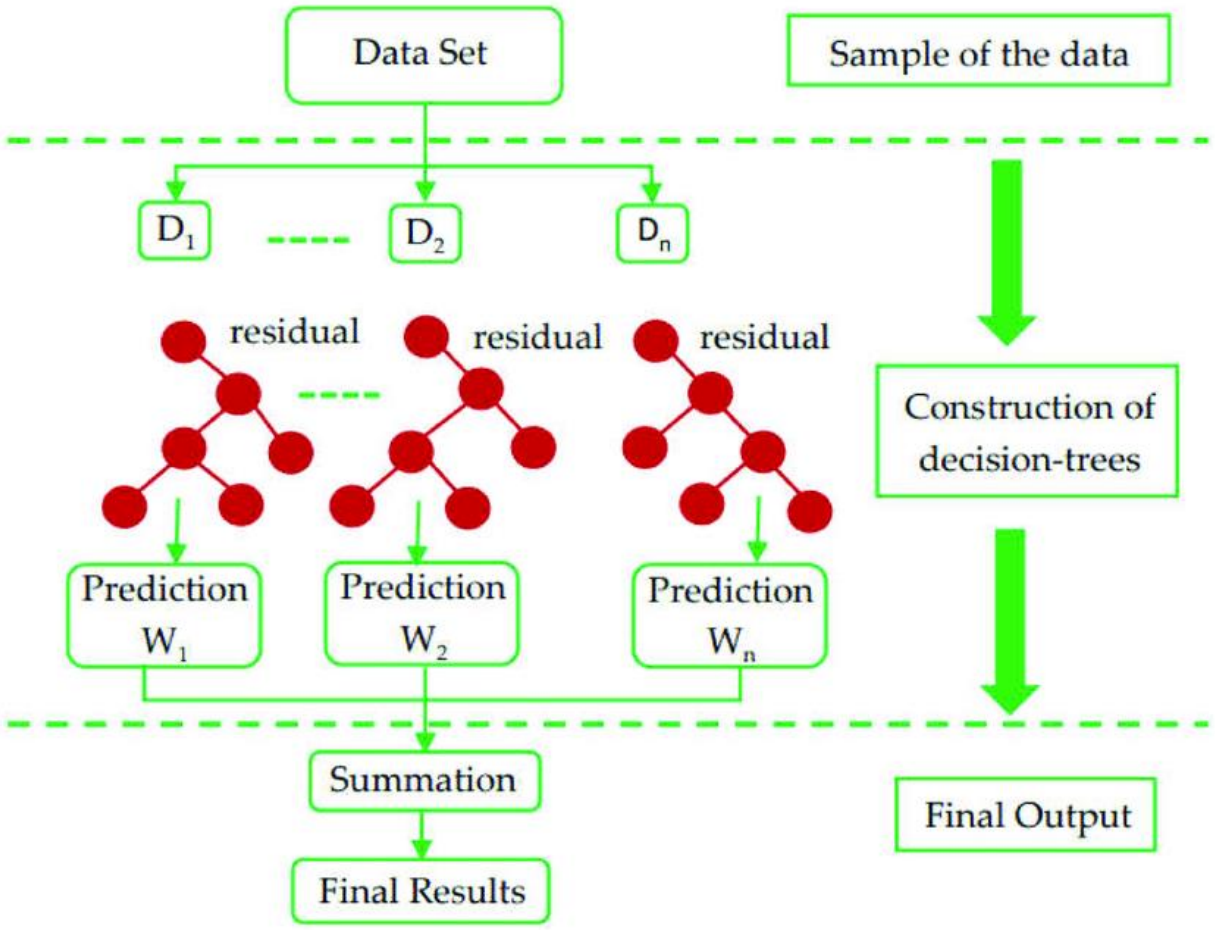
۲-۴- الگوریتم XGBoost

الگوریتم XGBoost یک پیاده‌سازی خاص و بهینه‌شده از الگوریتم تقویت گرادیانی است که برای افزایش کارایی و دقت مدل طراحی شده است. این الگوریتم به‌ویژه در مسائل دسته‌بندی و رگرسیون، از محبوب‌ترین الگوریتم‌ها محسوب می‌شود. XGBoost بر اساس تکنیک "گرادیان بوستینگ" عمل می‌کند که شامل بهبود مدل‌های ضعیف (مانند درخت‌های تصمیم) به‌صورت مرحله‌ای و ترتیبی است. ویژگی‌های منحصر به‌فرد این الگوریتم، به‌ویژه قابلیت جلوگیری از بیش‌برازش داده‌ها و کنترل پیچیدگی مدل از طریق regularization و early stopping، موجب محبوبیت آن شده‌اند.[۱۷]

۲-۴-۱- مراحل الگوریتم XGBoost

- ساخت مدل پایه: الگوریتم با ایجاد یک مدل پایه ساده به وسیله میانگین یا مد شروع می‌شود.
- محاسبه خطا: برای مدل پایه ایجاد شده، خطا یعنی فاصله آن تا هدف اندازه گیری می‌شود.
- ساخت مدل‌های جدید (درخت‌ها): یک درخت تصمیم جدید برای کاهش خطاهای باقی‌مانده[1]ساخته می‌شود. این درخت بر اساس داده‌های باقی‌مانده آموزش داده می‌شود تا خطاها را به حداقل برساند.
- بروزرسانی مدل: مدل جدید با ترکیب مدل قبلی و درخت جدید بر اساس نرخ یادگیری[2]ایجاد می‌شود. نرخ یادگیری کنترل می‌کند که چقدر از مدل جدید به مدل قبلی اضافه شود.
- تکرار مراحل ۲ تا ۴: سه مرحله قبلی تا زمانی که بهبودی در مدل ایجاد نشود ادامه می‌یابند.
- تنظیم پارامترها: پارامترهای مختلفی مانند عمق درخت،[3] نرخ یادگیری، تعداد تکرارها،[4] و پارامترهای regularization باید به‌طور مناسب تنظیم شوند تا به مدل بهینه دست یابید.
- پیش‌بینی نهایی: پس از اتمام تکرار ها و به دست آمدن مدل بهینه می‌توان از آن برای پیش‌بینی استفاده نمود.

[1] Residuals
[2] learning rate
[3] max_depth
[4] n_estimators



تصویر ۳ پیاده‌سازی الگوریتم XGBoost

۲-۵- جمع‌بندی

در این فصل، به معرفی الگوریتم‌های مؤثر و استفاده‌شده در این پروژه پرداخته شد. ابتدا الگوریتم K-means که برای دسته‌بندی و گروه‌بندی اولیه داده‌ها به‌کار می‌رود، معرفی شد و سپس، الگوریتم XGBoost که پیش‌نیاز آن الگوریتم درخت تصمیم‌گیری است معرفی گردید. این دو الگوریتم با همکاری یکدیگر به ایجاد مدلی بهینه و با دقت بالا برای پیش‌بینی نهایی کمک می‌کنند. فصل‌های بعدی به جزئیات بیشتری از نحوه به‌کارگیری و استفاده از این الگوریتم‌ها در پیش‌بینی و شناسایی مناطق امیدبخش دارای پتانسیل آهن خواهند پرداخت.

فصل سوم - شرح پروژه

۳-۱- مقدمه

صنعت معدن در زندگی انسان مدرن جایگاه ویژه‌ای دارد زیرا نقطه‌ی شروع زنجیره‌ی تعمین مواد اولیه برای ساخت انواع تجهیزات، دستگاه‌ها و دیگر وسایل می‌باشد؛ همین امر باعث شده تا امروزه این صنعت نقش مهمی در اقتصاد جهانی ایفا کند. یکی از مهم‌ترین منابعی که همواره با تقاضای زیاد مواجه بوده، آهن است که به دلیل کابرد زیاد در صنعت فولاد، خودروسازی، ماشین‌آلات صنعتی و... به عنوان یکی از ضروری‌ترین مواد معدنی شناخته می‌شود. با استخراج سریع‌تر و مطمئن‌تر آهن می‌توان به صنعت و اقتصاد کشور جان تازه‌ای بخشید.

با این‌حال صنعت معدن از دیرباز با چالش‌های متنوعی نظیر امنیت محیط کار، اثرات زیست محیطی، هزینه‌های بالا و بهره‌وری مواجه بوده است. کشف معدن فرایندی زمان‌بر و هزینه‌بر است که می‌تواند زیان بسیاری را به اقتصاد یک کشور وارد کند. در سال‌های اخیر و با توسعه‌ی فناوری و پیشرفت هوش مصنوعی و الگوریتم‌های یادگیری ماشین، روش‌های نوینی برای مقابله با این چالش‌ها پیشنهاد شده‌اند که اگر به درستی پیاده‌سازی و اجرا شوند می‌توانند انقلابی در صنعت معدن ایجاد کنند.

در این پروژه سعی شده تا با تجزیه و پردازش داده‌های دورسنجی ماهواره‌ای استر و اجرای الگوریتم تقویت گرادیانی روی داده‌های پردازش شده، وجود ذخایر سنگ آهن پیش‌بینی شود. تقویت گرادیانی یک الگوریتم قدرتمند است که می‌تواند به خوبی روابط بین داده‌ها را کشف کند و با استفاده از تعداد زیادی درخت به عنوان مدل پایه، دقت مدل را تا حد خوبی بالا ببرد.

۳-۲- معماری سیستم پیش‌بینی کانسار آهن با استفاده از تصاویر ماهواره‌ای

معماری سیستم، فرایند کامل ایجاد مدل و پیش‌بینی با استفاده از تصاویر خام اولیه را نشان می‌دهد. مراحل این فرایند به شکل زیر می‌باشند:

۱. جمع‌آوری داده: در این مرحله داده‌های دورسنجی و تصاویر زمین‌شناسی از ماهواره‌ی استر گرفته می‌شوند.
۲. پیش‌پردازش داده: در این مرحله داده‌های جمع‌آوری شده در قسمت قبل پیش‌پردازش می‌شوند تا برای اجرای مراحل بعدی آماده شوند. پیش‌پردازش داده شامل تمیزسازی، حذف نویز، نرمال‌سازی، تغییر شکل[1] و برگرداندن[2] عکس‌ها می‌باشد.
۳. استخراج ویژگی: در این مرحله باندهای مختلف (مانند مادون قرمز، طیف مرئی، و غیره) برای استخراج ویژگی‌های خاصی مانند پوشش گیاهی، رطوبت خاک و سایر پارامترهای محیطی استخراج می‌شوند و ویژگی‌های مختلف

[1] reshape
[2] transpose

مانند شاخص‌های کانی‌های معدنی، پوشش گیاهی، بافت‌ها، شکل‌ها و سایر ویژگی‌ها که می‌توانند به طبقه‌بندی یا پیش‌بینی کمک کنند محاسبه می‌شوند.

۴. برچسب‌گذاری داده: با استفاده از یک الگوریتم خوشه‌بندی، داده‌ها برچسب‌گذاری می‌شوند تا برای آموزش و آزمون یک مدل یادگیری ماشین آماده شوند.
۵. اجرا و آزمون مدل تقویت گرادیانی: در این مرحله مدل یادگیری ماشین ساخته می‌شود و با داده‌های آماده شده آموزش داده شده و سپس آزمایش می‌شود تا دقت آن مشخص شود. مدل XGBoost از تعداد زیادی درخت برای بالا بردن دقت خود استفاده می‌کند.
۶. نتیجه‌گیری و بصری‌سازی نتایج: در آخرین قدم، دقت مدل براساس شاخص‌های مختلف محاسبه شده و نتایج پیش‌بینی مدل روی تصویر ماهواره‌ای نشان داده می‌شوند تا بتوان به‌طور بصری نتایج پیش‌بینی را بررسی کرد.

۳-۳- روش استفاده شده در این پروژه برای پیش‌بینی مناطق دارای کانسار آهن

در این پروژه برای تشخیص نقاط دارای کانسار آهن از مراحل نامبرده در قسمت قبل استفاده شده که در زیر یک‌به‌یک شرح داده شده‌اند:

۳-۳-۱- جمع آوری داده

تصاویر ماهواره‌ای ابزارهای بسیار قدرتمندی در زمینه‌های مختلف علمی، صنعتی و مدیریتی هستند که به خاطر وضوح مکانی، وضوح طیفی، پوشش جهانی و تصاویر استریو ابزارهای مهم و کابردی‌ای در زمینه‌ی سنجش از دور هستند. ماهواره‌ی استر قادر به ثبت تصاویر در چندین باند طیفی مختلف است که هر کدام برای تحلیل‌های خاصی در علوم زمین و محیط‌زیست کاربرد دارند. تصاویر گرفته شده توسط این ماهواره هرکدام دارای ۱۴ باند می باشند که به شرح زیر اند:

- باندهای مرئی و نزدیک به مادون قرمز (VNIR - Visible and Near-Infrared):

تعداد این باندها ۳ عدد است و وضوح مکانی هریک ۱۵ متر می باشد. این باندها برای مشاهده پوشش گیاهی، آب، خاک، و سایر ویژگی‌های سطحی به کار می‌روند. باندهای VNIR به دلیل توانایی تشخیص تغییرات در پوشش گیاهی و دیگر ویژگی‌های زمینی، در مطالعات محیط‌زیست و کشاورزی بسیار مفید هستند.

- باندهای مادون قرمز کوتاه‌موج (SWIR - Shortwave Infrared):

تعداد این باندها ۶ عدد بوده و وضوح مکانی هریک ۳۰ متر می باشد. این باندها برای شناسایی ترکیبات معدنی، رطوبت خاک و سایر ویژگی‌های سطح زمین استفاده می‌شوند. باندهای SWIR به دلیل حساسیت به مواد معدنی و آب، در نقشه‌برداری زمین‌شناسی و ارزیابی منابع طبیعی کاربرد دارند.

- باندهای مادون قرمز حرارتی (TIR - Thermal Infrared):
- تعداد این باندها ۵ عدد بوده و وضوح مکانی هریک ۹۰ متر می باشد. این باندها برای اندازه‌گیری دمای سطح زمین و تحلیل تغییرات حرارتی استفاده می‌شوند. باندهای TIR برای بررسی فعالیت‌های حرارتی مانند

آتش‌سوزی‌ها، فعالیت‌های آتشفشانی و شناسایی مناطق با تغییرات حرارتی (مانند نشت گرمایی) بسیار مفید هستند.

تصاویر با وضح بالایی که استر ارائه می‌دهد، امکان شناسایی جزئیات سطح زمین را فراهم می‌کنند. این ماهواره با داشتن ۱۴ باند طیفی مختلف، قادر است اطلاعات دقیق‌تری از ویژگی‌های سطحی را ارائه دهد که برای شناسایی انواع مختلف پوشش زمینی و مواد معدنی ضروری است.

این ماهواره به دلیل دقت زیاد و پوشش مناسب در عکاسی در کاربردهای مختلفی از جمله نقشه‌برداری زمین‌شناسی، مدیریت منابع طبیعی، مطالعات زیست‌محیطی و مدیریت بلایای طبیعی مورد استفاده قرار می‌گیرد. داده‌های استر به‌طور گسترده‌ای در دسترس محققان و سازمان‌ها قرار دارد. این داده‌ها از طریق پلتفرم‌های مختلفی مانند وبسایت ناسا قابل دسترسی و دانلود هستند. این ویژگی، در کنار دیگر توانایی‌های استر، موجب استفاده‌ی گسترده از داده‌های این ماهواره شده است.

۳-۳-۲- پیش‌پردازش داده

برای استفاده از داده‌ها و اطمینان از نتایج به‌دست آمده توسط مدل یادگیری ماشین، پیش‌پردازش داده‌های ضروری ست. در این قدم، مراحل زیر طی می‌شوند:

۱. تمیزسازی داده‌ها:

تمیزسازی داده‌ها شامل شناسایی و تصحیح یا حذف داده‌های ناسازگار، گمشده، تکراری یا نادرست است.

۲. نرمال‌سازی داده‌ها:

نرمال‌سازی یکی از مراحل مهم در پیش‌پردازش داده‌ها است که هدف آن مقیاس‌بندی داده‌ها در یک محدوده خاص است، به طوری که تمام ویژگی‌ها یا متغیرها تأثیر یکسانی در مدل یادگیری ماشین داشته باشند.

۳. تغییر ابعاد داده‌ها:

تغییر ابعاد داده‌ها به معنی تغییر ساختار داده‌ها با جابجایی ردیف‌ها و ستون‌ها است. این کار به ویژه در مواقعی که داده‌ها باید به فرمت مناسب برای تحلیل تبدیل شوند کاربرد دارد.

۳-۳-۳- استخراج ویژگی

استخراج ویژگی یکی از مراحل مهم در یافتن کانسار آهن در تصاویر ماهواره‌ای و داده‌های دورسنجی است. این فرآیند شامل شناسایی و استخراج اطلاعاتی از تصاویر است که می‌توانند به تشخیص وجود کانسارهای آهن (مانند مگنتیت و هماتیت) کمک کنند. برای یافتن کانسار آهن با استفاده از باندهای مختلف تصاویر ماهواره‌ای استر، باید از اطلاعات طیفی موجود در باندهای مختلف این ماهواره بهره برد. تصاویر استر به دلیل داشتن باندهای متعدد در نواحی مختلف طیفی، ابزار مناسبی برای شناسایی مواد معدنی و بررسی ویژگی‌های زمین‌شناسی هستند.[۱۸] در ادامه، نحوه استفاده از باندهای مختلف برای شناسایی کانسارهای آهن توضیح داده شده است:

۱. باندهای مرئی و نزدیک به مادون قرمز (VNIR):

- **باندها:**
  - باند ۱: ۰,۵۲–۰,۶۰ میکرومتر (سبز)
  - باند ۲: ۰,۶۳–۰,۶۹ میکرومتر (قرمز)
  - باند ۳: ۰,۷۶–۰,۸۶ میکرومتر (نزدیک به مادون قرمز)

این باندها برای تحلیل پوشش گیاهی، رطوبت خاک و ویژگی‌های سطحی دیگر استفاده می‌شوند. اگرچه تمرکز اصلی باندهای VNIR بر روی ویژگی‌های بیولوژیکی است، اما می‌توان از آن‌ها برای شناسایی تغییرات عمده در سطح زمین که ممکن است به وجود کانسارهای آهن مرتبط باشد، استفاده کرد.

۲. باندهای مادون قرمز کوتاه‌موج (SWIR):

- **باندها:**
  - باند ۴: ۱,۶۰–۱,۷۰ میکرومتر
  - باند ۵: ۲,۱۴–۲,۲۹ میکرومتر
  - باند ۶: ۲,۳۱–۲,۳۶ میکرومتر
  - باند ۷: ۲,۴۸–۲,۶۸ میکرومتر

باندهای SWIR برای شناسایی ترکیبات معدنی مانند مگنتیت، هماتیت و لیمونیت بسیار مفید هستند. این باندها حساس به ویژگی‌های بازتابشی مواد معدنی هستند و می‌توانند به تشخیص وجود آهن کمک کنند.

همچنین از باندهای SWIR می‌توان برای محاسبه شاخص‌های طیفی مانند شاخص آهن[1] استفاده کرد که می‌تواند به شناسایی و نقشه‌برداری کانسارهای آهن کمک کند.

[1] Iron Oxide Index

۳. باندهای مادون قرمز حرارتی (TIR):

- **باندها:**
  - باند ۱۰: ۸,۱۲۵–۸,۴۷۵ میکرومتر
  - باند ۱۱: ۸,۴۷۵–۸,۸۲۵ میکرومتر
  - باند ۱۲: ۸,۹۲۵–۹,۲۷۵ میکرومتر
  - باند ۱۳: ۱۰,۲۵–۱۰,۹۵ میکرومتر
  - باند ۱۴: ۱۰,۹۵–۱۱,۶۵ میکرومتر

باندهای TIR بیشتر برای اندازه‌گیری دمای سطح زمین و تحلیل ویژگی‌های حرارتی مفید هستند. این باندها می‌توانند به شناسایی مناطق با فعالیت‌های حرارتی یا تغییرات دما که ممکن است با وجود کانسارهای معدنی مرتبط باشد، کمک کنند.

نسبت‌های باندی در تحلیل تصاویر ماهواره‌ای ابزار مهمی هستند که به تشخیص ویژگی‌های سطح زمین و شناسایی مواد معدنی کمک می‌کنند.[۱۸] در این پروژه از باندهای شماره ۴ تا ۹ برای شناسایی کانسار آهن استفاده شده که در ادامه توضیح داده شده‌اند:

- نسبت باند ۶ به باند ۴: این نسبت، بین باند SWIR (باند ۶) و باند قرمز (باند ۴) قرار دارد. در مناطقی که کانی‌های حاوی آهن مانند هماتیت یا گوتیت وجود دارد، این نسبت می‌تواند افزایش یابد. مواد معدنی حاوی آهن معمولاً در باند قرمز (۴) بازتاب بیشتری دارند و در باند SWIR (۶) بازتاب کمتری دارند، بنابراین این نسبت می‌تواند به شناسایی مناطق غنی از آهن کمک کند.
- نسبت باند ۷ به باند ۵: این نسبت، بین دو باند SWIR (باند ۷) و NIR (باند ۵) قرار دارد. باند NIR برای شناسایی تفاوت‌های رطوبتی و پوشش گیاهی استفاده می‌شود، در حالی که باند SWIR بیشتر به تفاوت‌های معدنی حساس است. بنابراین، این نسبت می‌تواند به شناسایی مناطق معدنی بدون پوشش گیاهی کمک کند، که ممکن است حاوی آهن باشند.
- نسبت باند ۸ به باند ۷: این نسبت به شناسایی تغییرات جزئی در سطح زمین کمک می‌کند. در مناطق معدنی، نسبت بالای باند ۸ (پانکروماتیک) به باند ۷ (SWIR) ممکن است نشان‌دهنده وجود کانسارهای آهن باشد، زیرا این نسبت به شناسایی تغییرات ساختاری سطح زمین کمک می‌کند که ممکن است ناشی از وجود مواد معدنی باشد.
- نسبت باند ۵ به باند ۴: این نسبت بین باند NIR (باند ۵) و باند قرمز (باند ۴) قرار دارد. در مناطقی که مواد معدنی حاوی آهن وجود دارند، این نسبت می‌تواند تغییرات در بازتاب باند قرمز را به خوبی نشان دهد. نسبت‌های پایین‌تر ممکن است نشان‌دهنده وجود آهن باشد.

- نسبت باند ۶ به باند ۵ : این نسبت بین دو باند SWIR (باند ۶) و NIR (باند ۵) قرار دارد. مواد معدنی غنی از آهن معمولاً در باند SWIR جذب بیشتری دارند و در NIR بازتاب کمتری دارند، بنابراین نسبت بالا در این شاخص می‌تواند به شناسایی مناطق معدنی حاوی آهن کمک کند.
- نسبت باند ۷ به باند ۶: این نسبت نیز بین دو باند SWIR قرار دارد و می‌تواند به شناسایی تفاوت‌های جزئی در ترکیبات معدنی سطح زمین کمک کند. در مناطقی که کانی‌های آهنی وجود دارند، این نسبت می‌تواند نشان‌دهنده تفاوت در ترکیبات معدنی باشد که ممکن است به کشف کانسارهای آهن کمک کند.

از ویژگی‌های استخراج شده برای برچسب‌گذاری و آموزش و آزمون مدل یادگیری ماشین استفاده می‌شود. قبل از آموزش مدل، مقادیر ویژگی‌ها استانداردسازی می‌شوند تا از قرار گرفتن تمام مقادیر در یک مقیاس یکسان اطمینان حاصل شود.

۳-۳-۴- برچسب گذاری داده

داده‌های مورد استفاده در ورودی مدل XGBoost داده‌های برچسب‌گذاری شده هستند؛ بنابراین لازم است تا بعد از استخراج ویژگی‌ها، با استفاده از یک الگوریتم خوشه‌بندی، برچسب‌گذاری شوند. در این پروژه از الگوریتم Kmeans استفاده شده است.

برای اجرای Kmeans ابتدا تعداد خوشه‌ها یا دسته‌هایی که باید برچسب‌گذاری شوند، تعیین می‌شود؛ در اینجا تنها به دو خوشه‌ی "دارای آهن" و "فاقد آهن" نیاز است. سپس الگوریتم اجرا می‌شود. پس از اتمام Kmeans، هر پیکسل یا داده‌ای که ویژگی‌های آن استخراج شده است، به یک خوشه یا دسته مشخص برچسب‌گذاری می‌شود.

۳-۳-۵- اجرا و آزمون مدل تقویت گرادیانی

مدل تقویت گرادیانی یک تکنیک قدرتمند در یادگیری ماشین است که به‌ویژه برای مسائل طبقه‌بندی و رگرسیون پیچیده کاربرد دارد. این مدل با ترکیب چندین مدل ضعیف‌تر، معمولاً درخت‌های تصمیم‌گیری، به تدریج یک مدل قوی‌تر و دقیق‌تر می‌سازد. در زمینه پیش‌بینی وجود یا عدم وجود کانسار آهن با استفاده از داده‌های استخراج‌شده از تصاویر ماهواره‌ای استر، این مدل می‌تواند به شکل موثری عمل کند.

در زیر مراحل طی شده در این پروژه برای پیش‌بینی مناطق دارای کانسار آهن توسط این مدل توضیح داده شده:

۱. آماده‌سازی داده‌ها: در این مرحله داده های مربوط به تصاویر زمین شناسی و مناطق معدنی جمع‌آوری و پیش‌پردازش می‌شوند. در این مرحله داده‌های تصویری تمیز و نرمال‌سازی شده و برای استفاده توسط مدل به یک فرمت مناسب تبدیل می‌شوند.

۲. ورود داده‌ها به مدل: داده‌های برچسب‌گذاری‌شده پس از تقسیم شدن، به مدل تقویت گرادیانی داده می‌شوند. این مدل با یک سری از درخت‌های تصمیم‌گیری کوچک و ضعیف[1] شروع می‌کند. ۷۰٪ داده‌ها برای آموزش مدل و ۳۰٪ باقی‌مانده برای آزمون آن استفاده می‌شود.
۳. آموزش مدل: مدل در هر مرحله، یک درخت تصمیم‌گیری جدید می‌سازد که تلاش می‌کند خطاهای مدل قبلی را تصحیح کند. این فرآیند تکرار می‌شود تا مدل نهایی به‌صورت تدریجی بهتر و دقیق‌تر شود. پارامترهای مختلفی مانند تعداد درخت‌ها، عمق درخت‌ها و نرخ یادگیری باید تنظیم شوند تا مدل به بهترین عملکرد خود برسد.
۴. ارزیابی مدل: پس از آموزش، مدل بر روی داده‌های آزمون که در فرآیند آموزش استفاده نشده‌اند، اعمال می‌شود تا عملکرد آن ارزیابی شود. معیارهایی مانند دقت[2]، دقت طبقه‌بندی[3]، نرخ بازخوانی[4] و F1-Score محاسبه می‌شوند تا نشان دهند که مدل تا چه حد در پیش‌بینی وجود یا عدم وجود کانسار آهن موفق عمل کرده است.

مدل آموزش‌دیده می‌تواند برای پیش‌بینی وجود یا عدم وجود کانسار آهن در نواحی دیگر تصویر یا تصاویر جدید استر که برچسب‌گذاری نشده‌اند، استفاده شود. نتایج پیش‌بینی می‌توانند به اکتشافات معدنی و تصمیم‌گیری‌های عملی کمک کنند. مناطقی که مدل آن‌ها را به‌عنوان "دارای آهن" پیش‌بینی کرده است، می‌توانند به عنوان نقاط هدف برای بررسی‌های بیشتر میدانی در نظر گرفته شوند.

۳-۳-۵-۱- روش‌های ارزیابی مدل

- دقت: دقت معیاری است که نشان می‌دهد چند درصد از پیش‌بینی‌های مدل درست بوده است. این معیار نسبت تعداد کل پیش‌بینی‌های صحیح به تعداد کل نمونه‌ها را اندازه‌گیری می‌کند.

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Samples}} = \text{Accuracy}$$

تصویر ۴ نحوه محاسبه دقت مدل

این معیار زمانی مفید است که داده‌ها به صورت متعادل باشند، یعنی تعداد نمونه‌های کلاس‌های مختلف تقریباً برابر باشد.

[1] weak learners
[2] accuracy
[3] precision
[4] recall

- دقت طبقه‌بندی: این معیار نشان می‌دهد که از میان نمونه‌هایی که مدل به عنوان مثبت پیش‌بینی کرده است، چند درصد واقعاً مثبت هستند. این معیار به کاهش نرخ پیش‌بینی‌های مثبت اشتباه[1] کمک می‌کند.

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \text{Precision}$$

تصویر ۵ نحوه محاسبه دقت طبقه‌بندی

دقت طبقه‌بندی بیشتر زمانی اهمیت دارد که هزینه‌ی پیش‌بینی اشتباه مثبت بالا باشد.

- نرخ بازخوانی: این معیار نشان می‌دهد که مدل از میان تمام نمونه‌های مثبت موجود، چه تعداد را به درستی شناسایی کرده است. این معیار به کاهش نرخ پیش‌بینی‌های منفی اشتباه[2] کمک می‌کند.

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \text{Recall}$$

تصویر ۶ نحوه محاسبه نرخ بازخوانی

نرخ بازخوانی زمانی اهمیت بیشتری پیدا می‌کند که شناسایی تمام نمونه‌های مثبت اولویت داشته باشد.

- F1-score: F1-Score میانگینی از دقت طبقه‌بندی و نرخ بازخوانی است که برای مواقعی که تعادلی بین این دو معیار مورد نیاز است، کاربرد دارد. این معیار به ویژه زمانی مفید است که با مجموعه داده‌های نامتعادل کار می‌کنیم.

$$\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 = \text{F1-Score}$$

تصویر ۷ نحوه محاسبه f1-score

[1] False Positives
[2] False negatives

۳-۴- جمع‌بندی

در این بخش مراحل طی شده در پروژه به‌طور کلی توضیح داده شدند. ابتدا نوع داده‌های مورد استفاده معرفی شد که تصاویر دورسنجی گرفته شده توسط ماهواره‌ی استر هستند. پس از معرفی ویژگی‌های این ماهواره، پردازش عکس‌ها و برچسب‌گذاری آن‌ها توسط الگوریتم Kmeans انجام می‌شود. سپس الگوریتم تقویت گرادیانی روی داده‌های برچسب‌گذاری شده اجرا شده و روی داده‌های آزمون، پیش‌بینی انجام می‌دهد تا دقت مدل بررسی شود.

فصل چهارم - پیاده‌سازی و نتایج

۴-۱- مقدمه

در این فصل الگوریتم معرفی شده یعنی تقویت گرادیانی با استفاده از درخت تصمیم پیاده سازی شده و نتایج مورد بحث قرار می‌گیرند.

۴-۲- جزئیات پیاده‌سازی

در جهت پیاده سازی این پروژه، از ابزارها، کتابخانه‌ها و معیارهایی که در زیر معرفی شده‌اند استفاده گردیده است.

- زبان برنامه نویسی پایتون: زبان برنامه‌نویسی پایتون به علت داشتن کتابخانه‌های متعدد یکی از بهترین زبان‌های برنامه‌نویسی برای پیاده‌سازی الگوریتم‌های هوش مصنوعی می‌باشد.
- کتابخانه numpy: این کتابخانه برای پردازش و محاسبات عددی و تحلیل داده‌های چندبعدی استفاده می‌شود. از این کتابخانه برای محاسبه میانگین، انحراف‌معیار، نرمال‌سازی و از بین بردن نویز استفاده شده است.
- کتابخانه matplotlib: از این کتابخانه برای تجسم داده‌ها و پردازش تصاویر ماهواره‌ای استفاده شده است.
- کتابخانه‌ی rasterio: از این کتابخانه برای خواندن و نوشتن داده‌های raster استفاده می‌شود. تصاویری که در این پروژه استفاده شده‌اند، همگی تصاویر تولید شده از ماهواره استر می‌باشند که اطلاعات را به صورت raster ذخیره می‌کنند.
- کتابخانه xgboost: از این کتابخانه برای پیاده‌سازی الگوریتم پیشنهادی مبتنی بر تقویت گرادیانی استفاده شده است.
- کتابخانه sklearn: این کتابخانه دارای بخش‌های مختلفی از جمله preprocessing، cluster، model_selection و metrics می‌باشد که در طول پیاده‌سازی از هر یک به صورت جداگانه استفاده شده است:
- Preprocessing: استفاده از StandardScaler برای نرمال‌سازی یا مقیاس‌بندی مجدد داده‌ها به صورتی که میانگین هر ویژگی صفر و انحراف معیار آن یک باشد. این کار باعث بهبود عملکرد الگوریتم‌های یادگیری ماشین از جمله XGboost و kmeans می‌شود.
- Cluster: از ابزار Kmeans برای خوشه‌بندی و ایجاد برچسب‌های موقت استفاده می‌شود. هر خوشه می‌تواند نمایانگر یک کلاس باشد برای مثال معدن بودن یا نبودن یک منطقه.
    - model_selection: استفاده از train_test_split به منظور تقسیم‌بندی داده‌ها به دسته‌های آموزشی و آزمایشی. هنگام پیاده‌سازی الگوریتم‌های پیش‌بینی‌کننده هوش‌مصنوعی باید مجموعه داده به دو گروه آموزشی و آزمایشی تقسیم‌بندی شود تا مدل با مجموعه‌ی آموزشی، آموزش ببیند و با مجموعه‌ی آزمایشی دقت آن بررسی شود.

- Metrics: برای بررسی نتایج الگوریتم از معیارهایی مانند classification_report و accuracy_score استفاده می‌شود. accuracy_score دقت کلی مدل را محاسبه می‌کند و classification_report گزارش کاملی از معیارهای ارزیابی مانند دقت، دقت طبقه‌بندی و F1-Score برای هر کلاس ارائه می‌دهد.

۴-۳- بارگذاری تصویر و پردازش اولیه

```python
import os
import rasterio
import numpy as np
import pandas as pd
import xgboost as xgb
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score
```

Executed at 2024.08.30 05:56:37 in 30s 788ms

تصویر ۸ کتابخانه‌های استفاده شده

در ابتدای پروژه اطلاعاتی از طول و عرض و تعداد باندهای تصویر به دست آورده می‌شود.

```python
def print_band_info(file_path):
    with rasterio.open(file_path) as src:
        print(f"\nFile: {os.path.basename(file_path)}")
        for band_idx in range(1, src.count + 1):
            description = band_descriptions.get(band_idx, f"Band {band_idx}")
            print(f"  Band {band_idx}: {description}")

print("Existing files in the directory:")
for file in existing_files:
    if file.endswith(".tif"):
        # Construct the full file path
        file_path = os.path.join(file_dir, file)

        # Open the file using Rasterio and print metadata and tags
        with rasterio.open(file_path) as src:
            metadata = src.meta
            tags = src.tags()

            # Print metadata and tags
            print(f"\nFile: {file}")
            print(f"Metadata: {metadata}")
            print(f"Tags: {tags}")

        # Print band information
        print_band_info(file_path)
```

تصویر ۹ کد بررسی اطلاعات تصویر

```
File: 4.tif
Metadata: {'driver': 'GTiff', 'dtype': 'float64', 'nodata': None, 'width': 2803, 'height': 9472, 'count': 6, 'crs': CRS.from_epsg(4326),
 'transform': Affine(0.00026949458523585647, 0.0, 54.75321488236896,
       0.0, -0.00026949458523585647, 34.524142830809865)}
Tags: {'AREA_OR_POINT': 'Area'}

File: 4.tif
  Band 1: Band 4 (SWIR, 1.60-1.70 µm)
  Band 2: Band 5 (SWIR, 2.145-2.185 µm)
  Band 3: Band 6 (SWIR, 2.185-2.225 µm)
  Band 4: Band 7 (SWIR, 2.235-2.285 µm)
  Band 5: Band 8 (SWIR, 2.295-2.365 µm)
  Band 6: Band 9 (SWIR, 2.360-2.430 µm)
```

تصویر ۱۰ نمونه خروجی از اطلاعات تصویر

در ادامه یکی از تصاویر برای پیاده سازی انتخاب و بارگذاری شده.

```python
# Define the file path
# file_path = r'E:\uni\project\ASTER\3.tif'
file_path = r'C:\Users\Melika\Downloads\ASTER\ASTER\3.tif'

# Open the TIFF file
with rasterio.open(file_path) as dataset:
    # Read the image data into a numpy array
    image = dataset.read()  # This will be an array of shape (bands, height, width)

# convert to (height, width, bands) shape for easier handling
image_transposed = np.transpose(image, (1, 2, 0))
print("\nTransposed image array shape:", image_transposed.shape)  # (height, width, bands)
```

Executed at 2024.08.30 06:12:58 in 16s 335ms

تصویر ۱۱ بارگذاری یکی از تصاویر برای ادامه پروژه

۴-۴- پیش‌پردازش داده‌ها

به منظور پیاده‌سازی الگوریتم ابتدا باید مراحل اولیه یعنی پیش‌پردازش داده‌ها انجام شود. در این مرحله به تجزیه و تحلیل و بررسی مجموعه داده پرداخته شده است.

۴-۴-۱- نرمال سازی تصویر

به منظور استخراج ویژگی‌ها ابتدا تمامی باندهای تصویر نرمال‌سازی شده و داده‌های نامعتبر nan و inf از تصویر حذف و با مقدار صفر جایگزین می‌شوند.

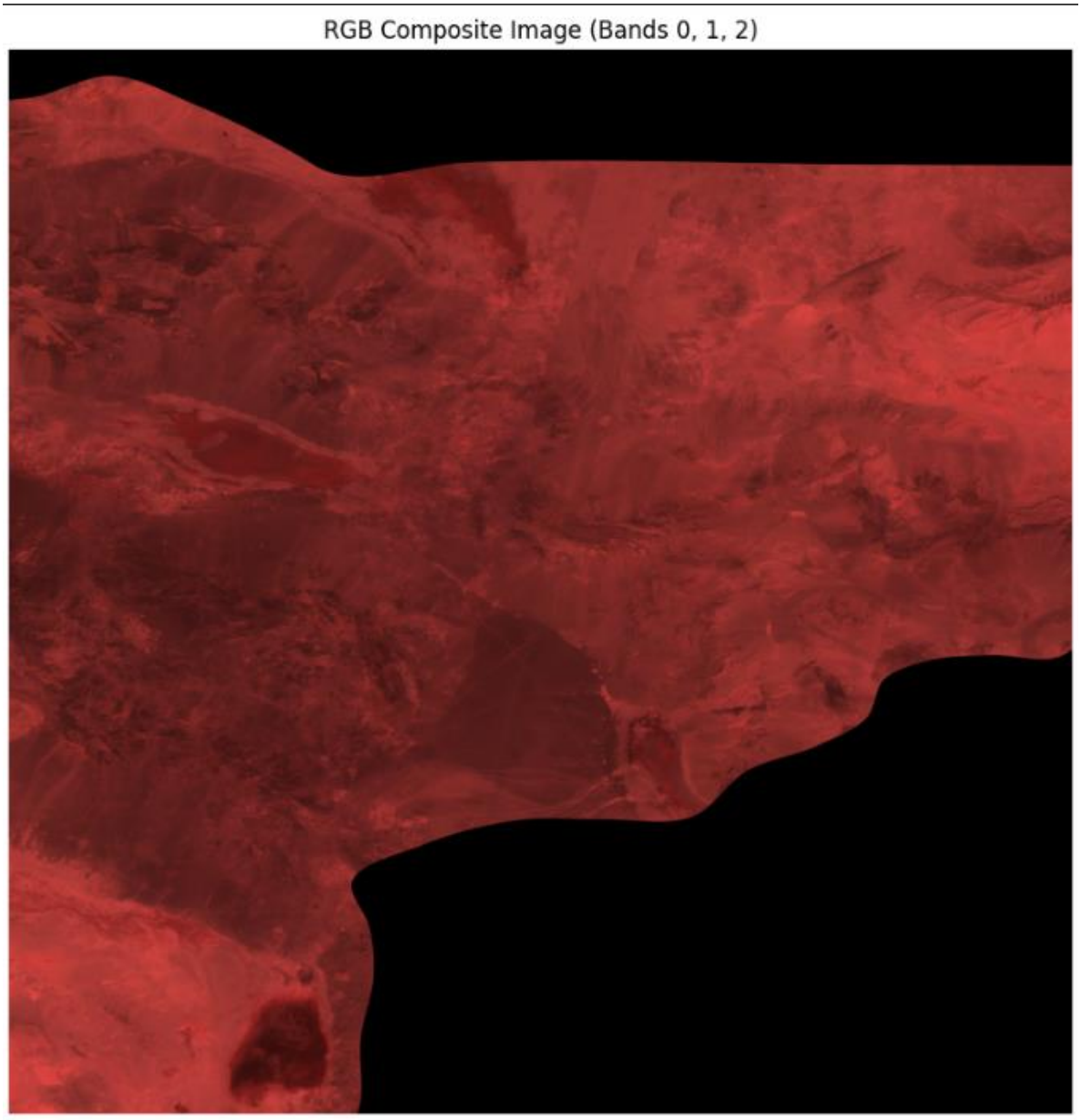
```
# Replace NaN, inf, and -inf values with 0 in the transposed image
image_transposed = np.nan_to_num(image_transposed, nan=0, posinf=0, neginf=0)
```

تصویر ۱۲ حذف مقادیر نامعتبر از تصویر و جایگزاری آن‌ها با مقدار صفر

پس از آن مقادیر تصویر مقیاس بندی مجدد می‌شود، ابتدا مقادیر بین بازه صفر تا یک مقیاس بندی می‌شوند سپس برای نمایش بهتر تصویر تمامی مقادیر در ۲۵۵ ضرب شده تا به مقیاس صفر تا ۲۵۵ برسند.

```
# Normalize the entire image to [0, 1]
min_val = image_transposed.min()
max_val = image_transposed.max()

if max_val - min_val != 0:
    image_normalized = (image_transposed - min_val) / (max_val - min_val)
else:
    image_normalized = image_transposed  # If all values are the same, avoid division by zero

# Clip values to ensure they are within the range [0, 1]
image_normalized = np.clip(image_normalized, 0, 1)

# Scale the normalized values to the range [0, 255]
image_scaled = image_normalized * 255

```

تصویر ۱۳ مقیاس‌بندی مقادیر باندها بین ۰ و ۱



تصویر ۱۴ سه باند اول به صورت تصویر RGB

۴-۴-۲- جمع‌آوری داده و استخراج ویژگی

همانطور که پیش‌تر گفته شد مجموعه داده استفاده شده در این پروژه مجموعه تصاویر گرفته شده از ماهواره‌ی استر می‌باشد. هر تصویر شامل ۶ باند می‌باشد که در تجزیه و تحلیل تصاویر و استخراج ویژگی‌ها استفاده شده‌اند.

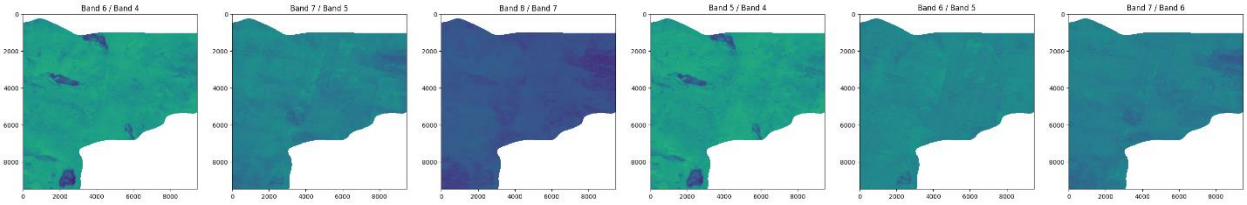
```
]  1  # Assuming image_transposed is defined earlier in the script
   2  band_4 = image_normalized[:, :, 0].astype('float64')
   3  band_5 = image_normalized[:, :, 1].astype('float64')
   4  band_6 = image_normalized[:, :, 2].astype('float64')
   5  band_7 = image_normalized[:, :, 3].astype('float64')
   6  band_8 = image_normalized[:, :, 4].astype('float64')
   7  band_9 = image_normalized[:, :, 5].astype('float64')
   8
```

تصویر ۱۵ معرفی و نام‌گذاری باندها

```
   8
   9  # Compute indices with error handling for divisions by zero
  10  with np.errstate(divide='ignore', invalid='ignore'):
  11      ratio_3_1 = band_6 / band_4
  12      ratio_4_2 = band_7 / band_5
  13      ratio_5_4 = band_8 / band_7
  14
  15      # Compute new indices
  16      swir_ratio_1 = band_5 / band_4
  17      swir_ratio_2 = band_6 / band_5
  18      swir_ratio_3 = band_7 / band_6
```

تصویر۱۶ استخراج ویژگی‌ها

هر یک از شاخص‌ها را می‌توان با ایجاد یک ماسک رنگی به نمایش در آورد برای مثال:



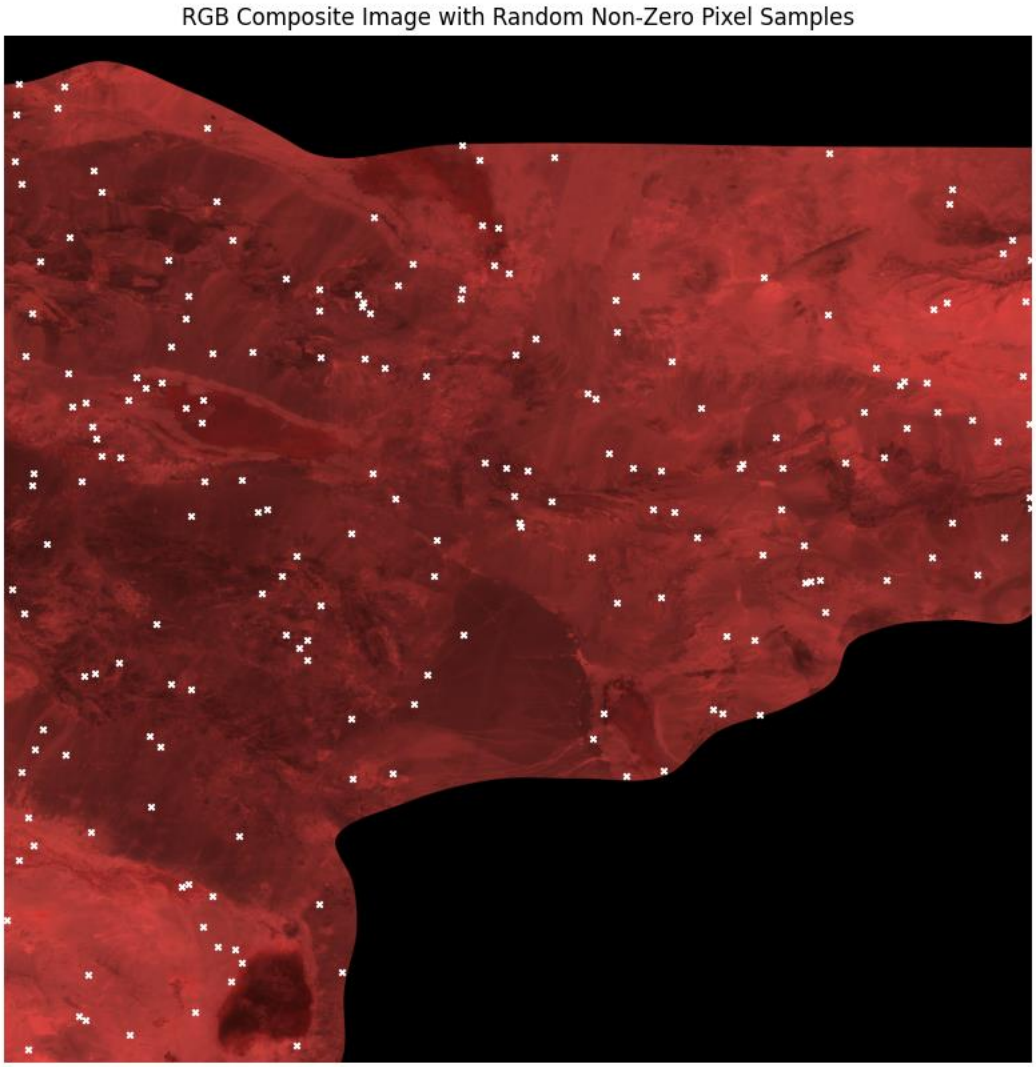
تصویر ۱۷ شاخص‌های تصویر

۴-۵- ایجاد نمونه

باتوجه به حجم بالای تصاویر ماهواره‌ای استر در جهت جلوگیری از خطای حافظه و بیش‌برازش مدل نهایی، یک نمونه داده از تصویر استخراج شده و در مراحل بعدی از آن استفاده می‌شود.

```
# Create a mask where all RGB bands are non-zero
non_zero_mask = np.all(image_normalized[:, :, :3] != 0, axis=2)

# Get the indices of non-zero pixels
non_zero_indices = np.argwhere(non_zero_mask)

# If there are enough non-zero pixels, randomly select
if non_zero_indices.shape[0] >= 200:
    selected_indices = non_zero_indices[np.random.choice(non_zero_indices.shape[0], 200, replace=False)]

```

تصویر ۱۸ ایجاد نمونه



تصویر۱۹ نمونه ایجاد شده

۴-۶- برچسب‌گذاری داده‌ها

از آنجایی که تصاویر استفاده شده در این پروژه فاقد هرگونه برچسب می‌باشند، لذا برای استفاده از الگوریتم‌های پیش‌بینی‌کننده هوش‌مصنوعی مانند XGboost ابتدا باید مجموعه داده با استفاده از یک الگوریتم خوشه‌بندی برچسب‌گذاری شود.

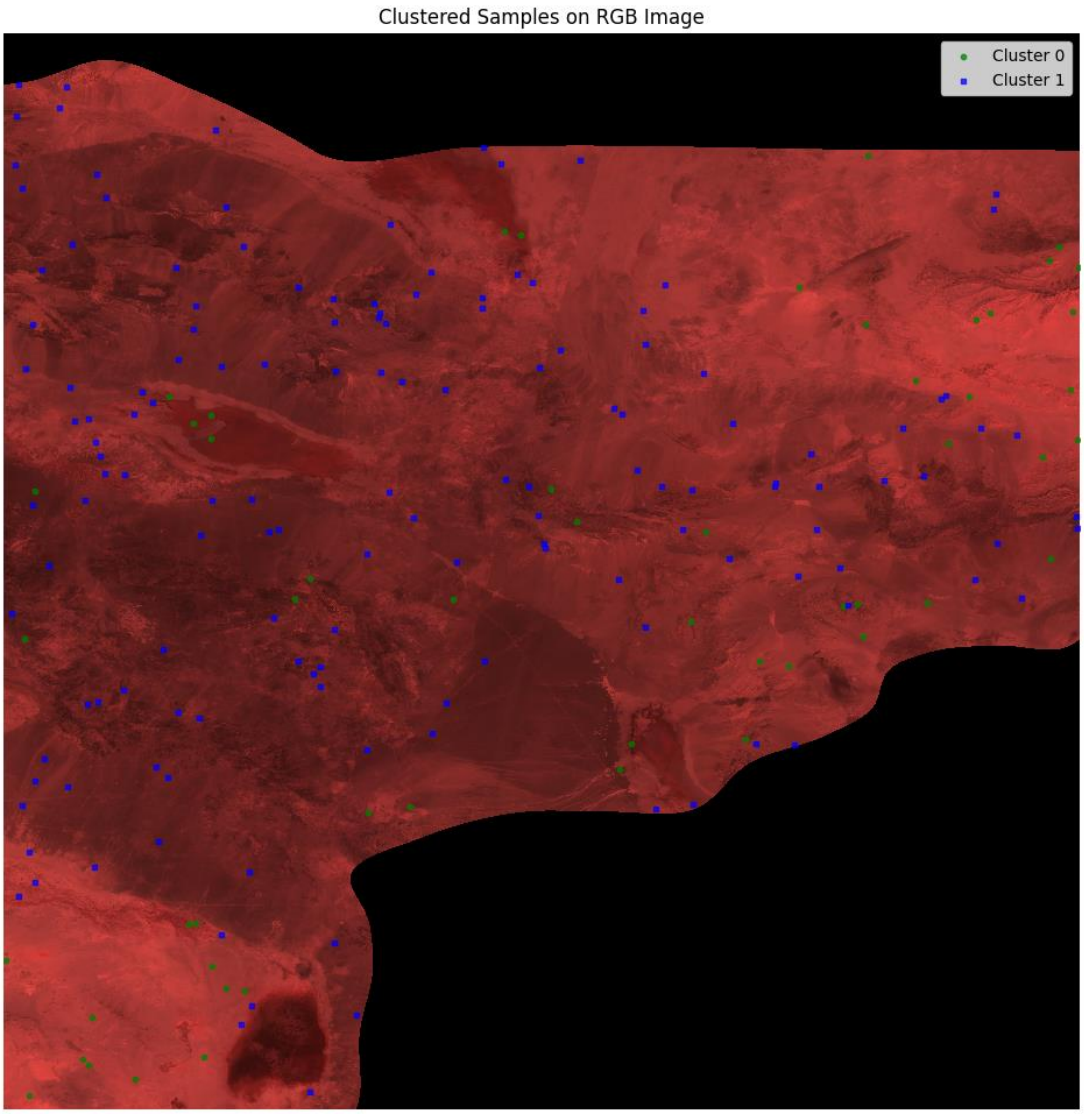
در این پروژه از الگوریتم Kmeans استفاده شده تا مجموعه‌ی داده به دو خوشه‌ی "معدنی" و "غیرمعدنی" دسته‌بندی شود. این دسته‌بندی طبق شاخص‌های اصلی که در بخش قبل معرفی شده است، انجام می‌گیرد.

```python
# Extract features for the selected pixels
selected_features = np.stack([
    ratio_3_1[selected_indices[:, 0], selected_indices[:, 1]],
    ratio_4_2[selected_indices[:, 0], selected_indices[:, 1]],
    ratio_5_4[selected_indices[:, 0], selected_indices[:, 1]],
    swir_ratio_1[selected_indices[:, 0], selected_indices[:, 1]],
    swir_ratio_2[selected_indices[:, 0], selected_indices[:, 1]],
    swir_ratio_3[selected_indices[:, 0], selected_indices[:, 1]],
], axis=-1)

# Reshape the selected features into a matrix for clustering
X = selected_features.reshape((selected_features.shape[0], -1))  # Flatten last dimension

# Perform k-means clustering into 2 groups
kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
labels = kmeans.labels_
```

تصویر۲۰ پیاده سازی Kmeans

```
Cluster 0 count: 54
Cluster 0 color: green
Cluster 1 count: 146
Cluster 1 color: blue
```

تصویر۲۱ نتایج خوشه‌بندی



تصویر۲۲ تصویر خوشه‌بندی

۴-۷- پیاده سازی الگوریتم XGboost

در این مرحله به پیاده‌سازی نهایی الگوریتم پرداخته شده‌است. در این قسمت، ابتدا داده برای ورود به مدل یادگیری ماشین آماده می‌شود. سپس داده‌ها به دو قسمت تقسیم شده، ٪۷۰ برای آموزش و ٪۳۰ برای آزمون استفاده می‌شوند. مدل XGBoost با استفاده از کتابخانه‌ی xgboost پایتون ساخته شده و با مجموعه‌ی آموزشی، آموزش داده می‌شود. در این مدل از ۱۰۰ درخت تصمیم به عنوان مدل پایه استفاده شده و عمق هر درخت حداکثر ۳ می‌باشد. برای جلوگیری از بیش‌برازش مدل از درختان کم‌عمق استفاده شده‌است.[۱۶]

آرگومان colsample_bytree = 0.8 درصد ویژگی‌هایی که برای ساخت هر درخت استفاده شده و subsample=0.6 درصد داده‌ای که از کل مجموعه‌ی آموزشی برای آموزش هر درخت استفاده می‌شود را نشان می‌دهند. با تنظیم این پارامترها در هر دور تکرار الگوریتم، بخشی از داده به‌طور تصادفی برای آموزش مدل انتخاب شده و به‌این ترتیب احتمال بیش‌برازش مدل پایین می‌آید.

```
e data for XGBoost
ls
es = X  # Use the features from the selected pixels

the data into training and testing sets, while keeping track of the indices
 X_test, y_train, y_test, train_idx, test_idx = train_test_split(
atures, y, selected_indices, test_size=0.3, random_state=0)

the XGBoost model
xgb.XGBClassifier(eval_metric='mlogloss', n_estimators=100, max_depth=3, subsample=0.6, colsample_b
t(X_train, y_train)
```

تصویر۲۳ پیاده سازی الگوریتم XGboost

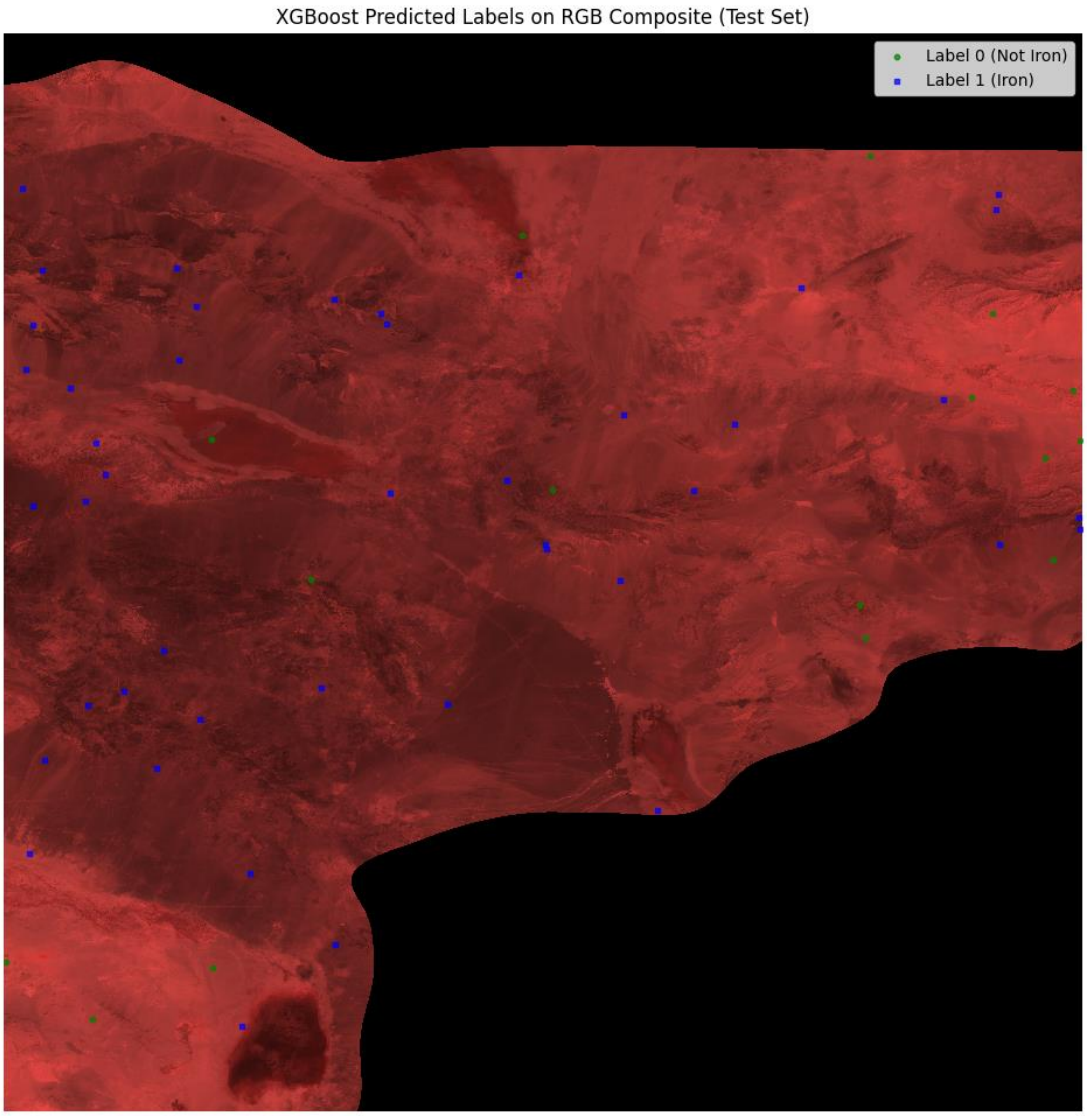
پس از ساخت و آموزش مدل XGBoost، با استفاده از داده‌ی آزمون، دقت مدل بررسی می‌شود. تصاویر زیر معیارها و نتایج ارزیابی مدل را نشان می‌دهند.

```
# Predict and evaluate the model
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
report = classification_report(y_test, y_pred)
```

تصویر ۲۴ شاخص‌های ارزیابی

```
Accuracy: 0.9833
Precision: 0.9837
Recall: 0.9833
F1-Score: 0.9832
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        17
           1       0.98      1.00      0.99        43

    accuracy                           0.98        60
   macro avg       0.99      0.97      0.98        60
weighted avg       0.98      0.98      0.98        60
```

تصویر۲۵ دقت و نتایج به دست آمده از مدل



تصویر ۲۶ نتایج به دست آمده روی مجموعه آزمون

همان‌طور که در تصویر مشاهده می‌شود نتایج آزمون مدل حدود ۹۸ درصد بوده که نشان از دقت خوب مدل دارد. این یعنی مدل به خوبی توانایی شناسایی الگوها و روابط بین ویژگی‌ها را دارد و می‌تواند به‌عنوان یک ابزار قدرتمند در شناسایی معادن و کانسارهای آهن استفاده شود.

۴-۸- جمع‌بندی

در این فصل، به معرفی جزئیات الگوریتم‌های استفاده شده در پیش‌پردازش تصویر و پیاده‌سازی نهایی پروژه و بررسی نتایج به دست آمده از آن پرداخته شد. برای پیاده سازی این پروژه از زبان برنامه‌نویسی پایتون و کتابخانه‌های معروف آن استفاده شد. از جمله:

- numpy برای پردازش و محاسبات عددی؛
- rasterio برای پردازش تصاویر ماهواره‌ای استر؛
- xgboost برای پیاده سازی الگوریتم پیش‌بینی کننده مبتنی بر تقویت گرادیانی؛
- sklearn، برای خوشه بندی داده‌ها به کمک Kmeans؛
- model_selection برای تقسیم بندی مجموعه داده به دو گروه آموزشی و آزمایشی؛
- Metrics برای بررسی نتایج مدل نهایی.

پس از پیش‌پردازش اولیه مجموعه داده و نرمال‌سازی آن‌ها، با استفاده از الگوریتم خوشه‌بندی kmeans مجموعه داده را برچسب‌گذاری کرده و پس از آن مجموعه داده را به دو گروه آموزشی و آزمایشی تقسیم کرده و مجموعه آزمایشی را به عنوان ورودی به الگوریتم XGboost داده و نتایج حاصل از آن با معیارهایی مانند دقت، دقت طبقه‌بندی و... بررسی می‌شود.

فصل پنجم - نتیجه‌گیری و پیشنهادات

۵-۱- نتیجه‌گیری

هدف این پروژه، توسعه یک الگوریتم تقویت گرادیانی برای شناسایی معادن آهن با استفاده از داده‌های دورسنجی به‌دست آمده از ماهواره‌ی استر بود. با توجه به حجیم بودن داده‌ها و طیف وسیع باندهای تصاویر استر که اطلاعات گوناگونی را نمایش می‌دهند، رویکرد اصلی در این پروژه بر استفاده از نسبت‌های باندهای خاص برای استخراج ویژگی و برچسب‌گذاری داده‌ها متمرکز شد. نسبت‌های باندهای ۶ به ۴، ۷ به ۵، ۸ به ۷، ۵ به ۴، ۶ به ۵ و ۷ به ۶ که برای شناسایی کانسارهای آهن و پوشش گیاهی و ترکیبات معدنی سطح زمین بسیار مفید هستند، به‌عنوان ویژگی‌های اصلی در نظر گرفته شدند.

برای پردازش این داده‌ها، ابتدا نقاطی از تصویر که دارای مقدار غیرصفر در باندهای RGB بودند، انتخاب شدند. سپس با استفاده از الگوریتم KMeans این نقاط به دو دسته‌ی "دارای آهن" و "فاقد آهن" خوشه‌بندی شدند. این دسته‌ها به‌عنوان نمونه‌های اولیه برای ایجاد برچسب‌ها استفاده شدند که در مرحله بعد به عنوان ورودی به مدل XGBoost داده شدند.

پس از اعمال الگوریتم XGBoost و تقسیم داده‌ها به دو مجموعه آموزشی و آزمون (۳۰ درصد برای آزمون)، مدل توانست به دقت نهایی ۹۸ درصد دست یابد. این نتیجه نشان‌دهنده عملکرد بسیار خوب مدل در تشخیص صحیح مناطق آهن‌دار از سایر نقاط تصویر است. دستیابی به دقت ۹۸ درصد نشان می‌دهد که ویژگی‌های انتخاب شده و رویکردی که برای خوشه‌بندی و سپس طبقه‌بندی استفاده شده، توانسته‌اند به‌خوبی تفاوت‌ها و الگوهای موجود در داده‌ها را تشخیص دهند.

```
Accuracy: 0.9833
Precision: 0.9837
Recall: 0.9833
F1-Score: 0.9832
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        17
           1       0.98      1.00      0.99        43

    accuracy                           0.98        60
   macro avg       0.99      0.97      0.98        60
weighted avg       0.98      0.98      0.98        60
```

تصویر۲۷ جدول نتایج ارزیابی مدل

نتایج این پروژه نشان می‌دهد که استفاده از KMeans به‌عنوان یک روش بدون نظارت برای خوشه‌بندی اولیه و سپس استفاده از XGBoost برای طبقه‌بندی نهایی، ترکیبی موثر بوده است که توانسته به دقت قابل قبولی دست یابد.

این پروژه نشان داد که درصورتی‌که داده‌های اولیه به‌خوبی پردازش و برچسب‌گذاری شوند، استفاده از یادگیری ماشین برای فعالیت‌های اکتشافی می‌تواند امری ممکن و بسیار و موثر باشد که هزینه‌ها و خطرات کشف معدن را به حداقل می‌رساند.

۵-۲- پیشنهادات برای بهبود و استفاده‌های آینده

با توجه به گستردگی این پژوهش و نتایج به دست آمده از آن، روش‌های زیر برای بهبود کارهای آینده پیشنهاد می‌شوند:

- **استفاده از روش‌های ترکیبی**: ترکیب روش‌های دیگری مانند شبکه‌های عصبی عمیق با مدل‌های فعلی می‌تواند منجر به افزایش دقت مدل شود.
- **استفاده از داده‌های ترکیبی**: استفاده از داده‌های ژئوشیمیایی به دست آمده در آزمایشگاه‌ها در کنار ویژگی‌های استخراج شده از داده‌های دورسنجی می‌تواند دقت و اطمینان نتایج را بالا ببرد.
- **استفاده از یادگیری عمیق**: با استفاده از یادگیری عمیق و روش‌های پیشرفته‌تر یادگیری ماشین می‌توان مدل‌های هوشمندتری ساخت که می‌توانند الگوهای پیچیده‌تر و غیرخطی را در داده‌ها کشف کنند.
- **آموزش شرکت‌ها و متخصصان**: برای استفاده‌ی واقعی از این مدل و مدل‌های مشابه و بهبود کاربرد آن در صنعت که هدف نهایی این تحقیقات است پیشنهاد می‌شود به تدریج نیروهای متخصص زمین‌شناسی و هوش مصنوعی در این رابطه آموزش بینند.
- **استفاده از مدل فعلی در مناطق دیگر**: برای بهبود بیشتر مدل می‌توان آن را در مناطق مختلف جغرافیایی اجرا کرد و ضعف‌ها و قوت‌های آن را شناسایی کرد و در رفع آن‌ها کوشید.

پیوست ۱

مجموعه‌ی داده (شامل تصویر استفاده شده در اجرای پروژه)، دفترچه یادداشت‌های ژوپیتر (شامل منبع کد پروژه) و نتایج اجرای مدل در یک پیوست قرار دارند و در پوشه‌ی enclosure ذخیره شده‌اند.

منابع:

۱. کاربرد روش آنالیز تمایز و ماشین بردار پشتیبان مرحله ای در مدل سازی کانی زایی کانسارهای طلای داشکسن. حمید گرانیان، سید حسن طباطبائی، هوشنگ اسدی هارونی، آرمان محمدی.

۲. مدل سازی تصویری اکتشاف پتانسیل های معدنی با استفاده از ماشین بردار پشتیبان. ماندانا طهمورثی، بهنام بابایی، سعید دهقان.

۳. سیستم پیشنهاد دهنده برای شناسایی مکان مناسب برای اکتشاف معدن با استفاده از تجزیه مقدار تکین.

۴. کاربرد الگوریتم درخت تصمیم گیری در شناسایی مناطق امیدبخش معدنی کانسار پلی متال طلا در محدوده جانجا، سیستان و بلوچستان.

5. Abubakar, F., *Investigation of iron ore potential in north-central Nigeria, using high-resolution aeromagnetic dataset and remote sensing approach.* Heliyon, 2024. 10 (1 .)

6. Li, S., C. Liu, and J. Chen, *Mineral Prospecting Prediction via Transfer Learning Based on Geological Big Data: A Case Study of Huayuan, Hunan, China.* Minerals, 2023.

7. Liu, C., et al., *A deep-learning-based mineral prospectivity modeling framework and workflow in prediction of porphyry–epithermal mineralization in the Duolong Ore District, Tibet.* Ore Geology Reviews, 2023.

8. Park, S. and Y. Choi, *Applications of unmanned aerial vehicles in mining from exploration to reclamation: A review.* Minerals, 2020.

9. Tahmooresi, M., B. Babaei, and S .Dehghan, *Mineral exploration modeling by convolutional neural network and continuous genetic algorithm: a case study in Khorasan Razavi, Iran.* Arabian Journal of Geosciences, 2022.

10. chatgpt.com

11. picterra.ch/blog/how-ai-machine-learning-are-revolutionizing-mining-efficiency

12. groundhogapps.com/machine-learning-in-mining

13. Investigation of iron ore potential in north-central Nigeria, using high-resolution aeromagnetic dataset and remote sensing approach

14. Visual Modeling of Mineral Potential Exploration Using Support Vector Machine

15. Estimation of Fe Grade at an Ore Deposit Using Extreme Gradient Boosting Trees (XGBoost)

16. Iron Ore Grade Modeling using a Gradient Booster Model
17. Residual geochemical gold grade prediction using extreme gradient boosting

18. ASTER-Based Remote Sensing Image Analysis for Prospection –Criteria of Podiform Chromite at the Khoy Ophiolite (NW Iran)

University of Isfahan

Faculty of Computer Engineering

Department of Computer Engineering

BSc thesis

Computer engineering majoring in software - artificial intelligence - computer networks

Title:

A hybrid machine learning approach to advance the identification of mineral-rich areas by applying gradient boosting through remote sensing data

Supervisor:

Dr. Faria Nasiri-Mofakham

By:

Melika Aghajanian Sabagh

Mahdis Fathi

August 2024

Abstract:

This project investigates and identifies areas with mineral potential, especially iron-rich resources, using machine learning models including gradient boosting (XGBoost), decision tree and k-means clustering algorithm. The input data of this study includes remote sensing images from satellites such as Aster and other satellites that cover various spectral and spatial features in different bands. The main objective of this project is to increase the prediction accuracy of algorithms in identifying iron deposits and in order to achieve that, the project is implemented in two main stages: 1) data preprocessing and feature extraction, 2) implementation and training of the final model.

In the first step, the data is prepared using various preprocessing techniques such as converting invalid data to zero and scaling values between zero and one. Then the required features and indicators are extracted from the data. In the second step, the data are divided into ferrous and non-ferrous categories using the k-means algorithm. After that, the data set is divided into two parts, 70% training and 30% testing, and then, by adjusting the model parameters, decision tree and gradient boosting algorithms are implemented and evaluated. The results of this implementation have an accuracy of nearly 98% which means that it has succeeded to a large extent in distinguishing mineral from non-mineral areas. The results of this research can help reduce financial costs and drilling risks and focus more on identifying promising areas for mineral exploration.

Keywords: identification of iron mines, remote sensing data, clustering, machine learning, gradient boosting, decision tree