Vol. 11, No. 1, Spring and Summer 2025, pp. 201-219 ISSN (Print): 2538-449X ISSN (Online): 2981-1619

Research Article
OPEN ACCESS
Kharazmi Journal of Earth Sciences
Journal homepage https://gnf.khu.ac.ir

A python-based framework for land cover classification in engineering geology: A comparative assessment of SVM, K-means, and spectral indices

Mehdi Talkhablou[1]*, Mahdi Farmahinifarahani[2], Saba Siah Mansouri[3]
1, 3. Department of Applied Geology, Faculty of Earth Sciences, Kharazmi University, Tehran, Iran.
2. Department of Geology, Faculty of Earth Sciences, Kharazmi University, Tehran, Iran.

Article info
Article history
Received: 12 July 2025
Accepted: 9 September 2025
Keywords:
Mass movements, K-means clustering, Support Vector Machine (SVM), spectral indices, machine learning, PCA, Python, land cover classification.

Abstract

Accurate land cover classification is a fundamental step in engineering geology studies, particularly for assessing slope instability and mass movements. With the growing availability of satellite data and machine learning tools, automated and reproducible classification frameworks have become essential. This study presents a comprehensive Python-based framework for land cover classification, comparing the performance of two machine learning algorithms, Support Vector Machine (SVM, supervised) and K-means clustering (unsupervised), against traditional spectral indices (NDVI, NDWI, UI, SAVI) using Landsat 8 imagery. The study area is located in East Azerbaijan Province, Iran, covering approximately 80×70 km with diverse land cover types, including vegetation, bare soil, urban areas, and surface water. Prior to classification, data underwent several preprocessing steps: gamma correction for visual enhancement, Min-Max normalization for data scaling, and Principal Component Analysis (PCA) for dimensionality reduction and multicollinearity mitigation. PCA retained components explaining at least 95% of total variance. Classification was performed on four main classes. Results were evaluated using Overall Accuracy (OA), Kappa Coefficient, and weighted Precision, Recall, and F1-Score. The SVM algorithm, using an RBF kernel, achieved the highest accuracy with 84% OA and a Kappa of 0.81, demonstrating superior ability in defining clear class boundaries, particularly in distinguishing urban areas from bare soil. In contrast, K-means clustering yielded 73% OA and a Kappa of 0.68, with noticeable class overlap. Spectral indices alone provided a baseline accuracy of ~65%, but their integration with machine learning models significantly improved performance. The findings confirm that supervised machine learning models, particularly SVM, outperform unsupervised clustering and standalone spectral indices. However, K-means remains viable in data-scarce scenarios. The proposed Python-based workflow offers a reproducible, transparent, and efficient approach for land cover analysis, making it a valuable tool for engineering geology applications such as landslide susceptibility mapping.

Introduction

In the era of rapid technological advancement, the integration of advanced computational tools in earth sciences has become essential. One of the most prominent trends is the application of machine learning in remote sensing and geospatial analysis. Depending on specific objectives, a range of machine learning-based algorithms both supervised and unsupervised have been developed and refined for land cover classification. Among these, K-means clustering and Support Vector Machine (SVM) stand out as widely used techniques. K-means, an unsupervised method, group's data based on similarity, while SVM, a supervised algorithm, finds an optimal hyperplane to separate classes with maximum margin (Pedregosa et al., 2011). Python, with its modularity and powerful libraries such as Scikit-learn, NumPy, and Pandas, has become the preferred programming language for implementing such models in earth sciences (Petrelli, 2021).

In engineering geology, particularly in studies of slope instability and mass mo vements like landslides, accurate land cover classification is critical. Traditional

DOI http://doi.org/10.22034/KJES.2025.11.1.103081
***Corresponding author:** Mehdi Talkhablou; **E-mail:** Talkhablou@khu.ac.ir
How to cite this article: Talkhablou, M., Farmahinifarahani, M., Siah Mansouri, S., 2025. A python-based framework for land cover classification in engineering geology: A comparative assessment of SVM, K-means, and spectral indices. Kharazmi Journal of Earth Sciences 11(1), 201- 219. http://doi.org/10.22034/KJES.2025.11.1.103081

spectral indices such as NDVI, NDWI, SAVI, and UI have long been used to identify vegetation, water, and urban areas (Javed et al., 2021; Xu, 2006). However, these indices may struggle to distinguish between spectrally similar classes (e.g., bare soil and urban areas). Machine learning algorithms, on the other hand, can learn complex patterns from data and offer higher classification accuracy.

This study aims to present a reproducible Python-based framework for land cover classification, evaluate the performance of SVM and K-means in engineering geology applications, compare them with spectral indices, apply Principal Component Analysis (PCA) to improve model efficiency, anduse advanced evaluation metrics (F1 Score, Precision, Recall).

The novelty of this research lies in the simultaneous comparison of supervised, unsupervised, and spectral based methods in engineering geology a comprehensive approach not commonly found in previous studies, especially in the context of Iranian landscapes.

Study area and data

The study area is located in East Azerbaijan Province, Iran, between latitudes 37.5° to 38.2°N and longitudes 46.1° to 46.8°E, covering an area of approximately 80×70 km. The region exhibits high spatial diversity in land cover, including forests, rangelands, bare soil, urban settlements, and surface water bodies such as rivers and reservoirs. The climate is semi-arid (BSk classification), which influences the seasonal dynamics of vegetation and soil moisture.

Landsat 8 OLI/TIRS imagery from July 20, 2023, was downloaded from the USGS Earth Explorer portal. Bands 2 (blue), 3 (green), 4 (red), 5 (NIR), and 7 (SWIR2) were used for spectral index calculation and subsequent classification. These bands are essential for computing key indices such as NDVI, NDWI, UI, and SAVI, which serve as primary features for model input.

Materials and methods

Spectral indices calculation

Four widely used spectral indices were computed to extract land cover features:

NDVI: Assesses vegetation health and density. NDWI: Identifies surface water bodies.UI (Urban Index): Highlights built-up areas. SAVI: Reduces soil brightness effects in low-vegetation regions.

These indices were calculated using standard mathematical formulas:

1. Normalized Difference Vegetation Index (NDVI):

$$NDVI = (NIR - Red) / (NIR + Red) \quad (1)$$

NDVI is widely used to assess vegetation health and density (Mróz and Sobieraj, 2004).

2. Normalized Difference Water Index (NDWI):

$$NDWI = (Green - NIR) / (Green + NIR) \quad (2)$$

NDWI is used to monitor water content in vegetation and soil (Xu, 2006).

3. Urban Index (UI):

$$UI = (SWIR2 - NIR) / (SWIR2 + NIR) \quad (3)$$

UI helps in identifying urban areas (Javed et al., 2021).

4. Soil Adjusted Vegetation Index (SAVI):

$$SAVI= ((NIR-Red) /(NIR+Red+L)) \times(1+L), L=0.5 \quad (4)$$

SAVI is used to minimize soil brightness influences on vegetation indices.

Machine learning algorithms

Two algorithms were employed:

K-means clustering (Unsupervised): Data were clustered into k=4 groups using the Elbow Method for optimal cluster selection. The objective is to minimize the within-cluster sum of squares:

$$J=j=1\sum ki=1\sum n\|xi\ (j)-\mu j\|2 \quad (5)$$

Support Vector Machine (SVM, Supervised): An RBF kernel was used to handle non-linear separability. The dual optimization problem is:

$$\alpha maxi=1\sum n\alpha i-21i,j\sum \alpha i\alpha jyiyjK\ (xi,xj) \quad (6)$$

Subject to: $\sum \alpha_i y_i = 0$, $0 \leq \alpha_i \leq C$

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification and regression problems. The goal of this algorithm is to find an optimal hyperplane that separates data belonging to different classes with the largest margin (Bishop, 2006).

In this research, the number of land cover classes was considered to be 4 and 50 labeled training samples were used for each class.

Data preprocessing

The data used in this study were extracted from Landsat 8 satellite images. To increase the accuracy of the analyses, the data underwent several preprocessing steps before applying machine learning algorithms.

1. Gamma Correction:

$$I_{corrected} = I^{\gamma}, \gamma \in [0.4, 1.2] \quad (7)$$

2. Min-Max Normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

3. Principal Component Analysis (PCA):

$$Y = XW \quad (9)$$

Retained components explaining ≥95% of total variance.

Implementation in Python

All steps including data preprocessing, spectral index calculation, machine learning model implementation, and result analysis were performed within the Python programming environment. To enhance accuracy, efficiency, and reproducibility, a suite of widely used and powerful libraries in data science and machine learning was employed.The entire workflow was implemented in Python 3.10 using the following libraries:

NumPy and Pandas for data manipulation and processing,

Scikit-learn for machine learning algorithms and Principal Component Analysis (PCA),

Matplotlib and Seaborn for data visualization,

Rasterio for reading and writing geospatial raster data.

Discussion and results

The study area, includes a diverse range of land cover types that are broadly classified into four main groups:

1. Vegetation (such as forests and fields)
2. Bare soil (areas with no or low cover)
3. Urban areas (construction, roads and infrastructure)
4. Surface water (rivers, lakes and reservoirs)

This spatial and structural diversity in land cover makes the study area an ideal testbed for evaluating machine learning algorithms in land cover classification. Figure 1 illustrates the land cover status before and after applying gamma correction. The gamma filter significantly enhanced image contrast, particularly in transitional zones between classes.

Use of spectral indices

To improve discrimination among land cover types and extract relevant features, several standard spectral indices were computed. As shown in Figure 2, these indices effectively highlight distinct surface characteristics. The indices used include:

NDVI (Normalized Difference Vegetation Index): for identifying and assessing vegetation health and density,

NDWI (Normalized Difference Water Index): for detecting surface water bodies,

SAVI (Soil-Adjusted Vegetation Index): for minimizing soil brightness effects in low-vegetation areas,

UI (Urban Index): for highlighting built-up and urban regions.

Applying support vector machine (SVM) and K-means

The Support Vector Machine (SVM) and K-means clustering algorithms were applied to the derived spectral indices; the classification outputs are presented in Figure 3. Analysis of the results reveals several key findings:

SVM demonstrated superior performance in accurately classifying land cover types. This is evident from the well-defined class boundaries and minimal overlap in the classification map. Quantitative evaluation shows that SVM achieved an overall accuracy of 84% in

this study. Its high performance stems from its supervised learning framework and its ability to model complex, non-linear decision boundaries using the RBF kernel.

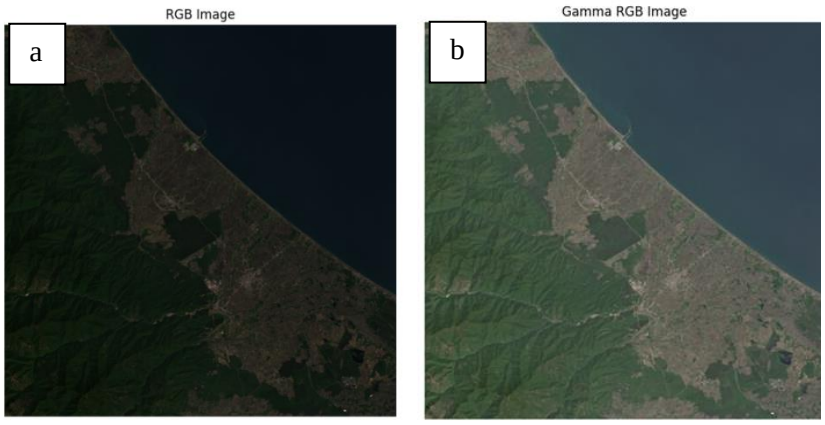


Fig. 1. Land cover status. a) RGB composite image of the study area before applying gamma correction b) RGB composite image of the study area after applying gamma correction.

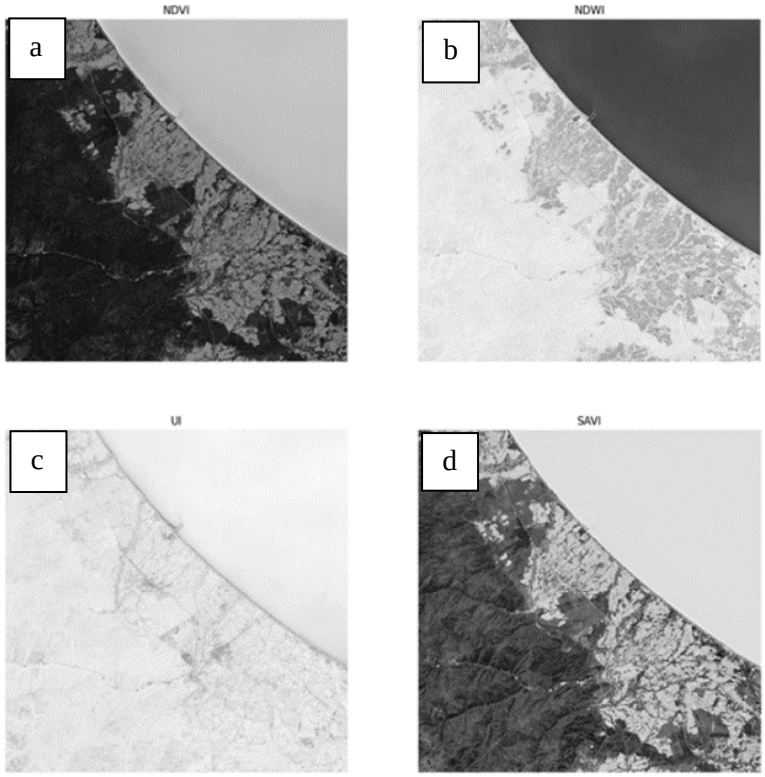


Fig. 2. The output of spectral indices. a) NDVI: High values correspond to dense vegetation; b) NDWI: Highlights surface water bodies; c) UI: Emphasizes built-up and urban areas; d) SAVI: Enhances vegetation signal in arid and low-vegetation zones.

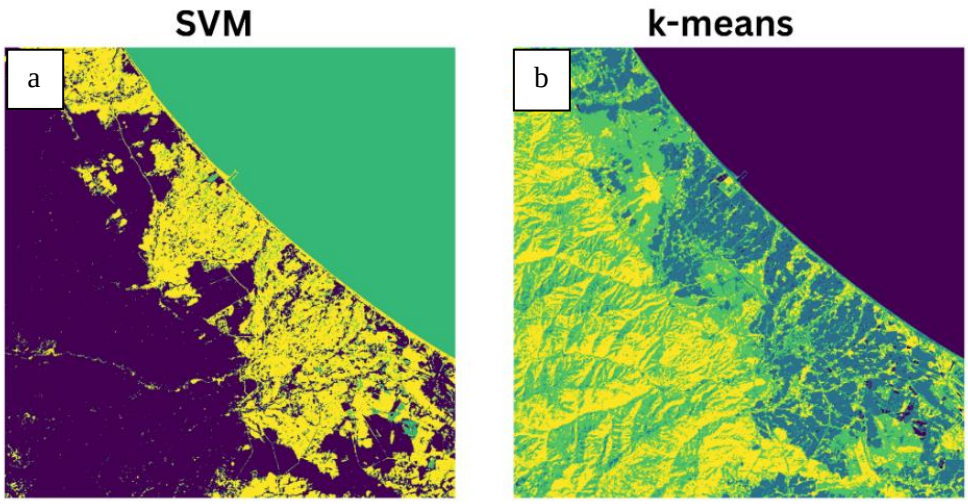


Fig. 3. Land cover classification results using machine learning algorithms applied to spectral indices (a) Support Vector Machine (SVM) (b) K-means clustering.

Model evaluation metrics

Model performance was evaluated using Overall Accuracy, Kappa Coefficient, and weighted averages of Precision, Recall, and F1-Score. The results for these evaluation metrics are presented in Table 1.

Table 1. Comparison of classification performance metrics between SVM and K-means

Evaluation Metric	SVM	K-means
Overall Accuracy (OA)	84%	73%
Kappa Coefficient	0.81	0.68
Weighted Precision	0.82	0.65
Weighted Recall	0.83	0.66
Weighted F1-Score	0.82	0.64

Conclusions

This study demonstrates that the Support Vector Machine (SVM) outperforms K-means clustering and standalone spectral indices in land cover classification, achieving an overall accuracy of 84% and a Kappa coefficient of 0.81. Key findings include:

The proposed Python-based framework is reproducible, transparent, and computationally efficient, making it suitable for operational use in engineering geology.

SVM with an RBF kernel is highly effective in modeling non-linear class boundaries, particularly in complex landscapes with spectrally similar land cover types.

The integration of spectral indices with machine learning models significantly enhances classification performance compared to using indices alone.

K-means clustering remains a viable option in unsupervised scenarios where labeled training data are unavailable or limited.

These findings confirm that supervised machine learning models, especially SVM, substantially improve the accuracy of land cover classification for critical engineering geology applications such as slope

instability assessment and landslide susceptibility mapping.

Future research should explore advanced techniques such as deep learning models (e.g., CNNs) and ensemble methods (e.g., XGBoost), as well as multi-temporal satellite analysis, to further enhance classification robustness and generalizability across diverse geological and climatic regions.

References

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

Javed, A., Cheng, Q., Peng, H., Altan, O., Li, Y., Ara, I., Saleem, N., 2021. Review of Spectral Indices for Urban Remote Sensing. Photogrammetric Engineering and Remote Sensing 87 (7), 513-524.

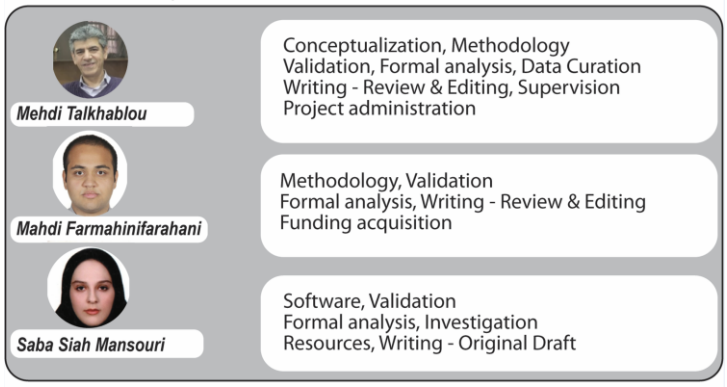
Mróz, M., Sobieraj, A., 2004. Comparison of several vegetation indices calculated on the basis of a seasonal SPOT XS time series and their suitability for land cover and agricultural crop identification. Technical Sciences 7, 39–66.

Pedregosa, F., 2011. Scikit- learn: Machine learning in python Fabian. Journal of machine learning research 12, 2825.

Petrelli, M., 2021. Introduction to Python in Earth Science Data Analysis: From Descriptive Statistics to Machine Learning. Springer Nature.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International journal of remote sensing 27 (14), 3025-3033.

CRediT authorship contribution statement



جلد ۱۱، شماره ۱، بهار و تابستان ۱۴۰۴، صفحه ۲۰۱ تا ۲۱۹

شاپا چاپی: ۴۴۹x- ۲۵۳۸ شاپا الکترونیکی: ۱۶۱۹-۲۹۸۱

مقاله پژوهشی

دسترسی آزاد

مجله علوم زمین خوارزمی

Journal homepage https://gnf.khu.ac.ir

چارچوب مبتنی بر پایتون برای طبقه‌بندی پوشش زمین در مطالعات زمین‌شناسی مهندسی مقایسه SVM، K-means و شاخص‌های طیفی

مهدی تلخابلو*[۱]، مهدی فرمهینی فراهانی[۲]، سبا سیاه منصوری[۳]

۱، ۳. گروه زمین‌شناسی کاربردی، دانشکده علوم زمین، دانشگاه خوارزمی، تهران، ایران.

۲، گروه زمین‌شناسی، دانشکده علوم زمین، دانشگاه خوارزمی، تهران، ایران.

اطلاعات مقاله

تاریخچه مقاله

دریافت: ۱۴۰۴/۰۴/۲۱

پذیرش: ۱۴۰۴/۰۶/۱۸

واژه‌های کلیدی

حرکات توده‌ای، خوشه‌بندی K-means، ماشین بردار پشتیبان (SVM)، شاخص‌های طیفی، یادگیری ماشین، PCA، پایتون.

چکیده

طبقه‌بندی پوشش زمین از اهمیت بالایی در مطالعات زمین‌شناسی مهندسی، به‌ویژه در ارزیابی ناپایداری دامنه‌ها و حرکات توده‌ای برخوردار است. این پژوهش به ارائه یک چارچوب مبتنی بر پایتون برای طبقه‌بندی پوشش زمین پرداخته و عملکرد دو الگوریتم یادگیری ماشین شامل خوشه‌بندی (K-means) بدون نظارت و ماشین بردار پشتیبان (SVM) با نظارت را در مقایسه با روش‌های متداول مبتنی بر شاخص‌های طیفی NDVI, NDWI, UI, SAVI ارزیابی کرده است. منطقه مورد مطالعه در استان آذربایجان شرقی ایران، با تنوع بالایی در پوشش زمین شامل پوشش گیاهی، خاک برهنه، مناطق شهری و آب‌های سطحی است. داده‌های ماهواره‌ای لندست ۸ پس از اعمال مراحل پیش‌پردازش (اصلاح گاما، نرمال‌سازی Min-Max و تحلیل مؤلفه‌های اصلی PCA) برای طبقه‌بندی چهار کلاس اصلی به کار گرفته شدند. نتایج نشان داد که الگوریتم SVM با دقت کلی ۸۴٪ و ضریب کاپا ۰/۸۱ عملکرد بهتری نسبت به الگوریتم K-means با دقت ۷۳٪ و ضریب کاپا ۰/۶۸ دارد. شاخص‌های طیفی به تنهایی دقتی حدود ۶۵٪ فراهم کردند، اما در ترکیب با الگوریتم‌های یادگیری ماشین، کارایی آن‌ها به‌طور معنی‌داری افزایش یافت. یافته‌ها نشان می‌دهد که الگوریتم‌های یادگیری ماشین، به‌ویژه SVM، می‌توانند به‌عنوان ابزاری قدرتمند در مطالعات زمین‌شناسی مهندسی از منظر طبقه‌بندی پوشش زمین در ارزیابی ناپایداری دامنه‌ها مورد استفاده قرار گیرند. علاوه بر این، چارچوب مبتنی بر پایتون ارائه‌شده در این پژوهش، قابلیت تکرار، شفافیت و کارایی بالایی دارد و می‌تواند به‌عنوان یک راهکار عملیاتی در مطالعات مشابه به‌کار گرفته شود.

مقدمه

در عصر پیشرفت سریع فناوری، استفاده از ابزارها و روش‌های پیشرفته در زمینه‌های مختلف از جمله علوم زمین، ضروری شده است. غلامی درگاهی و همکاران (Gholami Dargahi et al., 2021) میدان تنش دیرین با استفاده از روشی جدید در تعیین جهت لغزش در یکی از میادین هیدروکربوری جنوب ایران را برآورد نمودند. تلخابلو و همکاران (Talkhablou et al., 2023) با مقایسه روش های معتبر طبقه بندی توده سنگ با نگرشی بر تعیین پوشش اولیه روابط جدیدی را ارائه نمودند. آذری (Azari, 2025) در مطالعه‌ای از یک ماشین مرکب هوش مصنوعی نظارت شده جهت از بین بردن خطا و تخمین دقیق پارامترهای هیدرودینامیکی آبخوان های محبوس استفاده نمود. یکی از برجسته‌ترین روندهای اخیر، کاربرد یادگیری ماشین در حوزه‌های مختلف است (Petrelli, 2021). بسته به اهداف خاص، مجموعه‌ای از الگوریتم‌های مبتنی بر یادگیری ماشین، از جمله

DOI http://doi.org/10.22034/KJES.2025.11.1.103081

***نویسنده مسئول:** مهدی تلخابلو Talkhablou@khu.ac.ir

استناد به این مقاله: تلخابلو، م.، فرمهینی فراهانی، م.، سیاه منصوری، س. (۱۴۰۴). چارچوب مبتنی بر پایتون برای طبقه‌بندی پوشش زمین در مطالعات زمین‌شناسی مهندسی: مقایسه SVM، K-means و شاخص‌های طیفی. مجله علوم زمین خوارزمی. جلد ۱۱، شماره ۱، صفحه ۲۰۱ تا ۲۱۹. http://doi.org/10.22034/KJES.2025.11.1.103081

روش‌های نظارت‌شده و بدون نظارت، توسعه و بهبود یافته‌اند. در میان این الگوریتم‌ها، خوشه‌بندی K-means و ماشین بردار پشتیبان (SVM) به عنوان الگوریتم‌های شناخته‌شده و پرکاربرد برجسته می‌شوند. این روش‌ها در مطالعات متعددی برای حل چالش‌های مختلف به کار گرفته شده‌اند. پایتون، به دلیل ماژولار بودن، قابلیت‌های سطح بالا و کتابخانه‌های قدرتمند، به عنوان زبان برنامه‌نویسی مورد ترجیح در بسیاری از زمینه‌های علمی، از جمله علوم زمین، شناخته شده است (Petrelli, 2021).

در مطالعات زمین‌شناسی مهندسی، به‌ویژه در بررسی پدیده‌هایی مانند زمین‌لغزش و ناپایداری دامنه‌ها، شناسایی دقیق پوشش زمین از طریق سنجش از دور، نقش کلیدی دارد (Van Westen et al., 2008). در این راستا به منظور تعیین پوشش زمین، چندین شاخص طیفی برای افزایش دقت و کارایی توسعه یافته‌اند. شاخص‌های برجسته شامل شاخص تفاوت نرمال شده پوشش گیاهی (NDVI)، شاخص تفاوت نرمال شده (NDWI)، شاخص شهری (UI) و شاخص تنظیم شده پوشش گیاهی خاک (SAVI) می‌باشند. این شاخص‌ها نقش مهمی در طبقه‌بندی و تحلیل انواع پوشش زمین ایفا می‌کنند و بینش‌های ارزشمندی برای نظارت و مدیریت بهتر در مطالعات حرکات دامنه‌ای از جمله زمین لغزش‌ها فراهم می‌کنند (Petrelli, 2023).

شاخص‌های طیفی به‌عنوان روش‌های سنتی برای شناسایی پوشش گیاهی، آب و مناطق شهری مورد استفاده قرار گرفته‌اند (Javed et al., 2021). با این حال، این روش‌ها ممکن است در تفکیک کلاس‌های مشابه(مانند خاک برهنه و مناطق شهری) با محدودیت مواجه شوند. در مقابل، الگوریتم‌های یادگیری ماشین قادر به یادگیری الگوهای پیچیده از داده‌ها و ارائه طبقه‌بندی دقیق‌تر هستند (Zhang et al., 2020).

الگوریتم K-means به‌عنوان یک روش بدون نظارت، قادر به گروه‌بندی داده‌ها بر اساس شباهت است، در حالی که SVM به‌عنوان روش با نظارت، با استفاده از نمونه‌های برچسب‌دار، مرزهای بهینه بین کلاس‌ها را تشخیص می‌دهند (Pedregosa et al., 2011). پایتون نیز با کتابخانه‌های قدرتمندی مانند Scikit-learn وNumPy ، به یکی از پرکاربردترین ابزارها در این حوزه تبدیل شده است (Petrelli, 2023).

اهداف اصلی این پژوهش عبارت‌اند از: ارزیابی عملکرد الگوریتم‌های K-means، SVM در طبقه‌بندی پوشش زمین، مقایسه نتایج این روش‌ها با شاخص‌های طیفی سنتی، استفاده از تکنیک PCA برای بهبود عملکرد مدل‌ها، گزارش معیارهای جامع ارزیابی شامل Precision، Recall، F1-Score و Kappa و ارائه یک چارچوب کامل و قابل تکرار با استفاده از پایتون است.

نوآوری پژوهش حاضر در آن است که برای نخستین بار، عملکرد هم‌زمان الگوریتم‌های نظارت‌شده و بدون نظارت در کنار شاخص‌های طیفی در مطالعات زمین‌شناسی مهندسی بررسی شده است. در مطالعات پیشین معمولاً یکی از این رویکردها به‌صورت منفرد استفاده می‌شد و مقایسه جامعی از عملکرد این الگوریتم‌ها در کنار شاخص‌های طیفی و با استفاده از معیارهای دقیق ارزیابی در مناطق با تنوع پوشش زمین در ایران انجام نشده است. همچنین، استفاده از روش‌های کاهش بعد و گزارش معیارهای چندگانه ارزیابی (مانند F1-Score و Kappa) در مطالعات پیشین محدود بوده است.

منطقه مورد مطالعه

منطقه مورد مطالعه در شمال غرب ایران، در استان آذربایجان شرقی قرار دارد و مختصات آن بین ۳۷/۵° تا ۳۸/۲° شمالی و ۴۶/۱° تا ۴۶/۸° شرقی است. این منطقه دارای ابعاد تقریبی ۸۰ در ۷۰ کیلومتر است و از تنوع بالایی در پوشش زمین شامل جنگل، مراتع، خاک برهنه، مناطق شهری و آب‌های سطحی برخوردار است. اقلیم منطقه از نوع نیمه‌خشک(BSk) است. داده‌های ماهواره‌ای Landsat 8 OLI/TIRS با تاریخ تصویربرداری ۲۰ ژوئیه ۲۰۲۳ از پایگاه USGS Earth Explorer دانلود شدند. از باندهای ۲ (آبی)، ۳ (سبز)، ۴ (قرمز)، ۵ (مادون قرمز NIR) و۷ (مادون قرمز کوتاه موج دوم SWIR2) برای محاسبه شاخص‌های طیفی و انجام پردازش‌ها استفاده شده است.

مواد و روش‌ها

شاخص‌های طیفی در طبقه‌بندی پوشش اراضی

در مطالعات سنجش از دور، استفاده از شاخص‌های طیفی (Spectral Indices) یکی از روش‌های مؤثر برای استخراج اطلاعات پوشش زمین و تحلیل ویژگی‌های سطحی مانند پوشش گیاهی، آب، مناطق شهری و غیره است. این شاخص‌ها با استفاده از ترکیب باندهای مختلف طیفی تصاویر ماهواره‌ای محاسبه می‌شوند و امکان تفسیر بهتر و دقیق‌تر از داده‌های سنجش از دور را فراهم می‌سازند. در این پژوهش، چهار شاخص پرکاربرد و معتبر طیفی برای طبقه‌بندی پوشش اراضی مورد استفاده قرار گرفته‌اند که در ادامه معرفی شده‌اند:

۱. شاخص تفاضل نرمال شده پوشش گیاهی (NDVI)

$$NDVI = (NIR - Red) / (NIR + Red) \quad (۱)$$

شاخص NDVI یکی از رایج‌ترین و پرکاربردترین شاخص‌های پوشش گیاهی در علم سنجش از دور است که برای اندازه‌گیری تراکم و سلامت پوشش گیاهی به کار می‌رود (Mróz and Sobieraj, 2004). این شاخص بر مبنای تفاوت بازتاب طیفی پوشش گیاهی در باندهای نزدیک به مادون قرمز (NIR) و قرمز (red) عمل می‌کند. گیاهان سالم نور قرمز را به شدت جذب و نور مادون قرمز را به شدت بازتاب می‌دهند، بنابراین مقدار NDVI برای مناطق پوشیده از پوشش گیاهی بالا خواهد بود. مقادیر NDVI در بازه‌ای بین ۱- تا ۱+ قرار می‌گیرند که مقادیر نزدیک به ۱+ نشان‌دهنده‌ی پوشش گیاهی متراکم و سالم، و مقادیر نزدیک به صفر یا منفی نشان‌دهنده‌ی خاک بدون پوشش، مناطق شهری یا سطوح آبی هستند.

۲. شاخص تفاضل نرمال شده آب (NDWI)

$$NDWI = (Green - NIR) / (Green + NIR) \quad (۲)$$

شاخص NDWI توسط Xu به منظور شناسایی و پایش محتوای آب در پوشش گیاهی، خاک و دیگر سطوح توسعه داده شد (Xu, 2006). این شاخص به‌ویژه در مناطق مرطوب، تالابی و کشاورزی کاربرد فراوان دارد. در این رابطه، باند سبز (green) و باند نزدیک به مادون قرمز (NIR) مورد استفاده قرار می‌گیرند. آب در مقایسه با دیگر سطوح، نور NIR را کمتر بازتاب می‌دهد و نور سبز را بیشتر، بنابراین شاخص NDWI به‌طور مؤثری سطوح دارای آب را از سایر انواع پوشش زمین تفکیک می‌کند.

۳. شاخص شهری(UI)

$$UI = (SWIR2 - NIR) / (SWIR2 + NIR) \quad (۳)$$

شاخص UI (Urban Index) با هدف شناسایی مناطق شهری و ساخت‌وساز شده معرفی شده است (Javed et al., 2021). در این شاخص از باند مادون قرمز کوتاه موج دوم (SWIR2) و باند نزدیک به مادون قرمز (NIR) استفاده می‌شود. سطوح شهری مانند آسفالت و بتن معمولاً مقادیر بالاتری از بازتاب در باند SWIR2 نسبت به NIR دارند، در حالی که سطوح طبیعی مانند گیاهان چنین ویژگی‌ای ندارند. از این‌رو، UI می‌تواند به‌طور مؤثری مناطق ساخته‌شده و انسانی را از سایر کاربری‌های اراضی تفکیک کند.

۴. شاخص پوشش گیاهی تعدیل‌شده با خاک(SAVI)

$$SAVI= ((NIR−Red) /(NIR+Red+L)) ×(1+L), L=0.5 \quad (۴)$$

شاخص SAVI برای رفع تأثیر روشنایی خاک در مناطق با پوشش گیاهی کم معرفی شده است (Mróz and Sobieraj, 2004). این شاخص مشابه NDVI است با این تفاوت که در آن یک عامل تعدیل خاک (L) اضافه شده است که معمولاً مقدار آن برابر با ۰/۵ در نظر گرفته می‌شود. این اصلاح باعث کاهش تأثیر خاک‌های روشن یا تیره در نتایج شاخص شده و امکان تحلیل دقیق‌تری از پوشش گیاهی در مناطقی با تراکم گیاهی پایین را فراهم می‌سازد.

الگوریتم‌های یادگیری ماشین برای طبقه‌بندی پوشش زمین

در این پژوهش به‌منظور طبقه‌بندی انواع پوشش زمین از دو الگوریتم شناخته‌شده در حوزه یادگیری ماشین استفاده شده است: خوشه‌بندی K-means و ماشین بردار پشتیبان (SVM). این دو

الگوریتم با رویکردهای متفاوت (بدون نظارتی و با نظارتی) به تحلیل داده‌های سنجش از دور پرداخته و در جهت دستیابی به دقت بالاتر در طبقه‌بندی سطوح مختلف زمین به‌کار گرفته شده‌اند. اگرچه این پژوهش بر روی الگوریتم‌های کلاسیک متمرکز است، روش‌های نوظهور و قدرتمندی مانند شبکه‌های عصبی پیچشی (CNN) نیز نتایج امیدوارکننده‌ای در این حوزه نشان داده‌اند.

خوشه‌بندی K-means

الگوریتم K-means یک روش یادگیری ماشین بدون نظارت (Unsupervised learning) است که هدف آن گروه‌بندی داده‌ها بر اساس شباهت‌های موجود میان آن‌هاست. در این الگوریتم، داده‌ها به تعداد مشخصی از خوشه‌ها (k) تقسیم می‌شوند (Ishfagh et al., 2022). تعیین تعداد بهینه خوشه‌ها (k) در این پژوهش با استفاده از روش آرنج (Elbow Method) و نمودار Sum of Squared Errors (SSE) انجام شده است. بر این اساس مقدار k=4 به عنوان تعداد بهینه خوشه‌ها انتخاب گردید. به این معنا که داده‌ها به چهار گروه مجزا تقسیم می‌شوند که هر یک نمایانگر نوعی از پوشش زمین هستند.

تابع هدف الگوریتم K-means بر اساس کمینه‌سازی مجموع مربع فاصله نقاط داده از مرکز خوشه متناظر تعریف می‌شود. این تابع به صورت معادله ۵ بیان می‌گردد:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - \mu_j \|^2 \qquad (۵)$$

در این معادله $x_i^{(j)}$ نقطه داده متعلق به خوشه j و uj مرکز خوشه j است.

الگوریتم به صورت تکراری با به‌روزرسانی مراکز خوشه‌ها و تخصیص مجدد نقاط داده به نزدیک‌ترین مرکز، به‌سوی همگرایی حرکت می‌کند تا زمانی که تغییرات به حداقل برسد یا متوقف شود (Kwedlo, 2011).

ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان یا SVM یکی از الگوریتم‌های یادگیری ماشین با نظارت (Supervised learning) است که به‌طور گسترده برای مسائل طبقه‌بندی و رگرسیون به کار می‌رود. در این الگوریتم، هدف یافتن یک ابر صفحه (Hyperplane) بهینه است که داده‌های مربوط به کلاس‌های مختلف را با بیشترین حاشیه از یکدیگر جدا کند (Bishop, 2006).

در این پژوهش، تعداد دسته‌های پوشش زمین برابر با ۴ در نظر گرفته شده و برای هر کلاس، ۵۰ نمونه آموزشی برچسب‌دار استفاده شده است.

فرم دوگان (Dual Form) مسئله بهینه‌سازی SVM به‌صورت معادله ۶ تعریف می‌شود:

$$\max_{\{\alpha_i\}} g(\{\lambda_n\},\{\alpha_n\}) = \sum_n \alpha_n + \frac{1}{2}\sum_n \alpha_m \alpha_n y_m y_n k(x_m, x_n) \qquad (۶)$$

با قیود زیر:

$$\alpha_n, \lambda_n \geq 0, \forall_n; \sum_n \alpha_n y_n = 0;\ C - \alpha_n - \lambda_n = 0$$

در این معادلات α_n ضرایب لاگرانژ، λ_n برچسب‌های کلاس، و $k(x_m, x_n)$ تابع کرنل است (Berk, 2008).

در این پژوهش از تابع کرنل پایه شعاعی (RBF) در الگوریتم SVM استفاده شده است. دلیل انتخاب این کرنل، توانایی بالای آن در مدل‌سازی مسائل غیرخطی پیچیده که در داده‌های سنجش از دور معمول هستند و عملکرد بالا در داده‌های تصویری می‌باشد.

جمع‌آوری و پیش‌پردازش داده‌ها

داده‌های مورد استفاده در این پژوهش از تصاویر ماهواره‌ای Landsat 8 استخراج شده‌اند. برای افزایش دقت تحلیل‌ها، داده‌ها قبل از اعمال الگوریتم‌های یادگیری ماشین تحت چند مرحله پیش‌پردازش قرار گرفتند.

اصلاح گاما (Gamma correction)

برای بهبود کیفیت دیداری تصاویر شاخص‌ها و افزایش وضوح تفکیک بین کلاس‌های پوشش زمین، از اصلاح گاما بر روی خروجی‌های طیفی استفاده شد. اصلاح گاما با اعمال یک تابع غیرخطی بر روی سطوح روشنایی تصویر، کنتراست آن را به گونه‌ای تغییر می‌دهد

که ویژگی‌های پنهان یا با کنتراست پایین بهتر دیده شوند. این تکنیک به‌ویژه در تفکیک مناطق با پوشش متوسط یا مرزی بسیار مؤثر بوده و موجب افزایش دقت تفسیر بصری نقشه‌ها شده است.

فرمول کلی اصلاح گاما به صورت معادله ۷ است:

$$Icorrected = I\gamma \quad (۷)$$

که در آن I مقدار روشنایی نرمال شده هر پیکسل و γ مقدار گاما (معمولاً بین ۰/۴تا ۱/۲) است. در این پژوهش مقدار گاما به‌صورت تجربی برای هر شاخص انتخاب شده تا بهترین نتایج بصری و طبقه‌بندی حاصل شود.

نرمال‌سازی Min-Max

برای یک‌دست‌سازی داده‌های طیفی و ارتقای عملکرد مدل‌های یادگیری ماشین، از تکنیک نرمال‌سازی Min-Max استفاده شده است. این روش، مقادیر عددی هر باند را به بازه‌ای ثابت، معمولاً بین [۰, ۱]، مقیاس می‌کند تا از تأثیر تفاوت‌های مقیاس و واحد جلوگیری شود. فرمول این نرمال‌سازی به صورت معادله ۸ تعریف می‌شود:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (۸)$$

که در آن X مقدار اصلی، X_{min} حداقل مقدار در مجموعه داده و X_{max}حداکثر مقدار در مجموعه داده است. این فرآیند به کاهش تأثیر مقیاس‌ها و واحدهای مختلف در داده‌ها کمک می‌کند و آن‌ها را برای مدل‌های یادگیری ماشین مناسب‌تر می‌سازد (Peshawa, 2022).

تحلیل مؤلفه‌های اصلی - Principal Component Analysis (PCA)

برای کاهش ابعاد داده‌های طیفی و حذف هم خطی (Multicollinearity) بین باندهای ماهواره‌ای و شاخص‌های مشتق شده، از تکنیک تحلیل مؤلفه‌های اصلی (PCA) به‌عنوان یکی از روش‌های پیشرفته کاهش بعد و استخراج ویژگی استفاده شد. این روش با تبدیل مجموعه داده‌های هم‌بسته به مجموعه‌ای از مؤلفه‌های جدید متعامد(نامرتبط) و مرتب‌شده بر اساس واریانس تبیین شده، امکان حفظ بیشترین اطلاعات ممکن با تعداد کمتری از ویژگی‌ها را فراهم می‌کند (Jolliffe and Cadima, 2016).

این تکنیک به‌ویژه در پردازش تصاویر چند طیفی مانند داده‌های Landsat 8 که شامل چندین باند طیفی و شاخص‌های مشتق شده مانند NDVI, NDWI, NDBI و غیره هستند، بسیار مؤثر است؛ زیرا بین این شاخص‌ها همبستگی بالایی وجود دارد که می‌تواند عملکرد مدل‌های یادگیری ماشین را تحت تأثیر قرار دهد (Ashrafi Igder et al., 2022). با اعمال PCA، مؤلفه‌های اصلی (PCs) به گونه‌ای استخراج می‌شوند که اولین مؤلفه بیشترین واریانس داده‌ها را توجیه کند، مؤلفه دوم بیشترین واریانس باقیمانده را در جهتی عمود بر مؤلفه اول، و به همین ترتیب.

فرمول ریاضی PCA بر پایه تجزیه مقادیر ویژه (Eigenvalue Decomposition) ماتریس کوواریانس داده‌ها استوار است. فرآیند تبدیل داده‌ها به مؤلفه‌های اصلی به صورت معادله ۹ تعریف می‌شود:

$$Y = XW \quad (۹)$$

در معادله ۹، Xماتریس داده‌های نرمال شده با ابعاد $n \times p$ که n تعداد نمونه‌ها و p تعداد ویژگی‌هاست، W ماتریس بردارهای ویژه (Eigenvectors) متناظر با بزرگ‌ترین مقادیر ویژه (Eigenvalues) ماتریس کوواریانس، و Y ماتریس مؤلفه‌های اصلی با ابعاد کاهش‌یافته است.

در این پژوهش، پس از اعمال PCA، مؤلفه‌هایی که در مجموع حداقل ۹۵٪ از واریانس کل داده‌ها را توجیه می‌کردند، برای تحلیل‌های بعدی انتخاب شدند. این رویکرد نه‌تنها سرعت محاسبات الگوریتم‌های یادگیری ماشین را افزایش داد، بلکه از بیش برازش (Overfitting) جلوگیری کرده و عملکرد کلی مدل‌ها را بهبود بخشید. نتایج نشان داد که استفاده از مؤلفه‌های اصلی به‌جای داده‌های اولیه، دقت طبقه‌بندی را به‌ویژه در الگوریتم‌های مبتنی بر فاصله (مانند K-means) به‌صورت قابل‌توجهی افزایش می‌دهد.

پیاده‌سازی در پایتون

تمام مراحل پردازش داده‌ها، محاسبات شاخص‌های طیفی، پیاده‌سازی الگوریتم‌های یادگیری ماشین و تجزیه‌وتحلیل نتایج در محیط برنامه‌نویسی پایتون انجام گرفت. به‌منظور افزایش دقت، کارایی و سهولت در اجرای مراحل مختلف، از مجموعه‌ای از کتابخانه‌های قدرتمند و رایج در حوزه علم داده و یادگیری ماشین استفاده شد.

کتابخانه‌های مورد استفاده

- NumPy برای انجام محاسبات عددی و کار با آرایه‌های چندبعدی
- Pandas برای مدیریت و پردازش ساختارهای جدولی داده‌ها
- Scikit-learn به‌عنوان کتابخانه اصلی برای پیاده‌سازی الگوریتم‌های یادگیری ماشین از جمله K-means وSVM
- Matplotlib و Seaborn برای ترسیم نمودارها، نمایش نتایج طبقه‌بندی و تحلیل بصری داده‌ها (Pedregosa, 2011).

ویژگی‌های مورد استفاده برای آموزش مدل‌های K-means و SVM، مقادیر چهار شاخص طیفی محاسبه شده (NDVI, NDWI, UI, SAVI) برای هر پیکسل بوده است.

محاسبه شاخص‌های طیفی

شاخص‌های طیفی مختلف با استفاده از روابط ریاضی تعریف‌شده در بخش قبلی (رابطه‌های ۱ تا ۴) در قالب توابع پایتون محاسبه شدند. اسکریپت‌های توسعه‌یافته به‌صورت خودکار از روی تصاویر ورودی باندهای طیفی را استخراج کرده، شاخص‌های مورد نظر را محاسبه کرده و در قالب داده‌های قابل تحلیل ذخیره می‌کنند.

پیاده‌سازی خوشه‌بندیK-means

برای خوشه‌بندی داده‌ها از کلاس KMeans در کتابخانه Scikit-learn استفاده شد. قبل از تعیین تعداد خوشه‌ها، روش آرنج (Elbow Method) برای تحلیل و یافتن مقدار بهینه k به کار رفت. نمودار مجموع مربعات درون خوشه‌ای (inertia) برای مقادیر مختلف k رسم شده و نقطه شکست به عنوان تعداد مناسب خوشه‌ها انتخاب شد.

پیاده‌سازی طبقه‌بندی باSVM

برای طبقه‌بندی داده‌های برچسب خورده از کلاس SVC (Support Vector Classification) در Scikit-learn استفاده شد.

بحث و تحلیل نتایج

منطقه مورد مطالعه و تحلیل پوشش اراضی

منطقه مورد مطالعه، شامل طیف متنوعی از انواع پوشش زمین می‌باشد که به‌طور کلی در چهار گروه اصلی طبقه‌بندی می‌شوند:

۱. پوشش گیاهی (مانند جنگل‌ها و مزارع)
۲. خاک برهنه (مناطق بدون پوشش یا با پوشش کم)
۳. نواحی شهری (ساخت‌وساز، جاده و زیرساخت‌ها)
۴. آب‌های سطحی (رودخانه‌ها، دریاچه‌ها و مخازن آبی)

وجود این تنوع مکانی و ساختاری در پوشش اراضی، منطقه مورد مطالعه را به محیطی ایده‌آل برای ارزیابی و آزمون الگوریتم‌های یادگیری ماشین در طبقه‌بندی زمین تبدیل کرده است. در شکل ۱وضعیت پوشش زمین قبل و بعد از اعمال اصلاح فیلتر گاما نشان داده شده است. اصلاح گاما کنتراست را به ویژه در مناطق انتقالی بهبود بخشید.

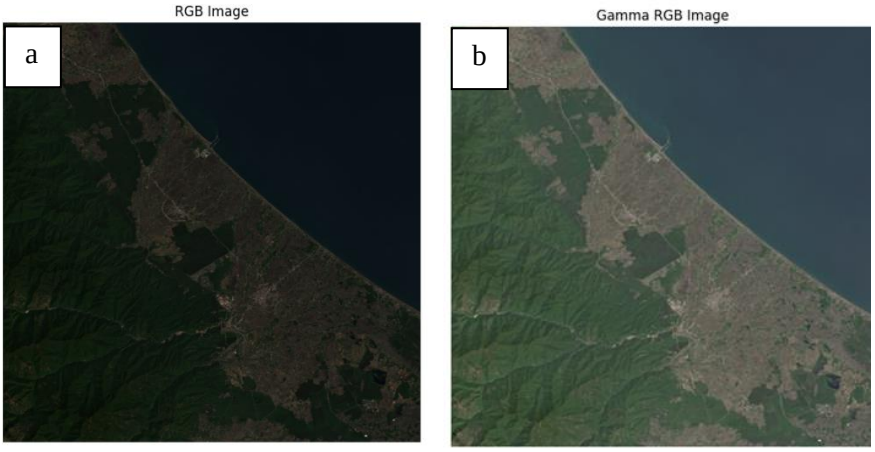
استفاده از شاخص‌های طیفی

برای تقویت قدرت تفکیک انواع پوشش زمین و استخراج ویژگی‌های مرتبط، از چندین شاخص طیفی استاندارد استفاده شد. شکل ۲ نتایج بدست آمده از شاخص‌های طیفی را نمایش می‌دهد که ویژگی‌های مختلف سطوح زمین را به‌خوبی نمایان می‌سازند. شاخص‌های استفاده‌شده عبارت‌اند از:

- NDVI شاخص پوشش گیاهی نرمال شده: (برای شناسایی و تحلیل سلامت و تراکم پوشش گیاهی)
- NDWI شاخص آب نرمال شده: (برای آشکارسازی منابع آبی سطحی)
- SAVI شاخص پوشش گیاهی با تعدیل خاک: (برای حذف اثر خاک در مناطق با پوشش گیاهی کم)

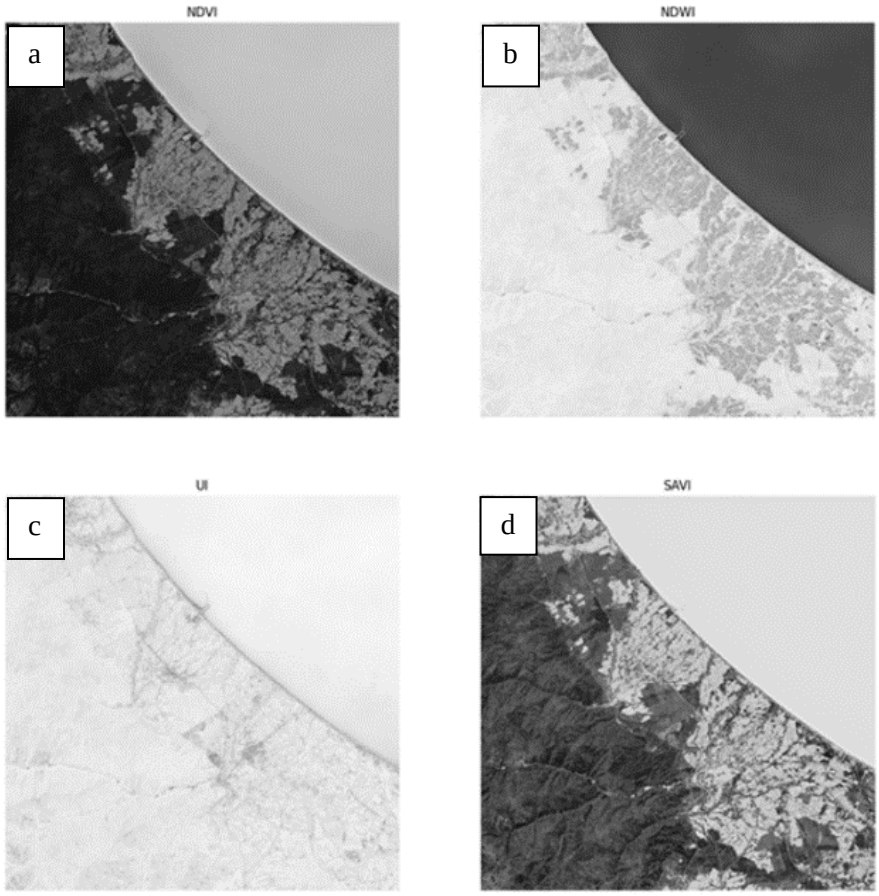
- UI شاخص شهری: (برای برجسته‌سازی نواحی ساخت‌وساز و شهری)

ترکیب این شاخص‌ها امکان تمایز بهتر بین کلاس‌های مختلف را فراهم کرده است. به‌عنوان نمونه، مناطق با مقادیر بالای NDVI نمایانگر پوشش گیاهی متراکم، در حالی‌که مقادیر بالای UI نشان‌دهنده ساختارهای انسانی یا مناطق شهری هستند.



شکل۱- وضعیت پوشش زمین: (a) تصویر ترکیبی RGB منطقه مورد مطالعه پیش از اعمال تصحیح گاما؛ (b) تصویر ترکیبی RGB منطقه مورد مطالعه پس از اعمال تصحیح گاما

Fig. 1. Land cover status. a) RGB composite image of the study area before applying gamma correction b) RGB composite image of the study area after applying gamma correction.

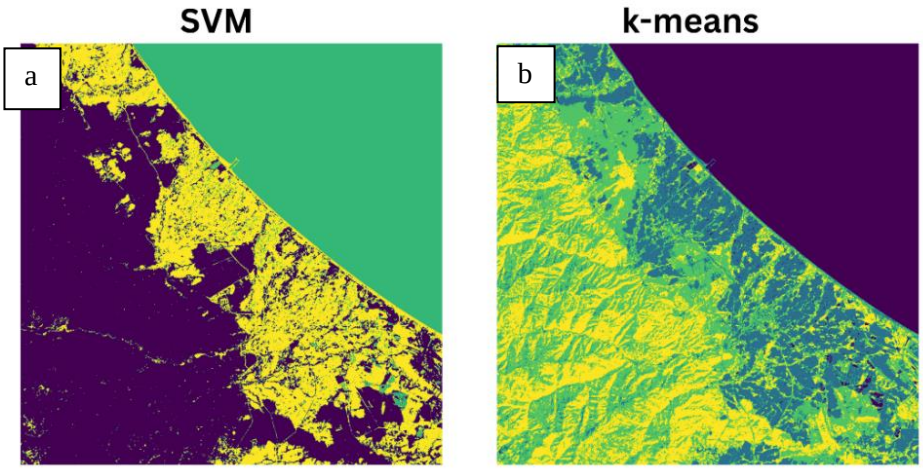


شکل ۲- نتایج به‌دست‌آمده از شاخص‌های طیفی. a)NDVI: مقادیر بالای شاخص، نشان‌دهنده پوشش گیاهی انبوه؛ b)NDWI: برجسته‌سازی پهنه‌های آب سطحی؛ c)UI: برجسته‌سازی مناطق شهری و ساخته‌شده؛ d)SAVI: پوشش گیاهی بهبود یافته در مناطق خشک

Fig. 2. The output of spectral indices. a) NDVI: High values correspond to dense vegetation; b) NDWI: Highlights surface water bodies; c) UI: Emphasizes built-up and urban areas; d) SAVI: Enhances vegetation signal in arid and low-vegetation zones.

اعمال الگوریتم‌های ماشین بردار پشتیبان (SVM) و خوشه‌بندی (K-means)

با اعمال الگوریتم‌های ماشین بردار پشتیبان (SVM) و خوشه‌بندی (K-means) به شاخص‌های طیفی، نتایج طبقه‌بندی پوشش زمین در شکل ۳ نشان داده شده است. تحلیل این نتایج چند نکته کلیدی را آشکار می‌کند.



شکل ۳- نتایج طبقه‌بندی پوشش زمین با استفاده از الگوریتم‌های یادگیری ماشین اعمال‌شده بر شاخص‌های طیفی. a) ماشین بردار پشتیبان (SVM)؛ b) خوشه‌بندی K-means.

Fig. 3. Land cover classification results using machine learning algorithms applied to spectral indices (a) Support Vector Machine (SVM) (b) K-means clustering.

SVM عملکرد برتری در طبقه‌بندی دقیق انواع پوشش زمین نشان داده است. این امر از مرزهای واضح و خوشه‌های متمایز تشکیل شده در خروجی SVM مشهود است. با محاسبه میزان صحت شناسایی صورت گرفته می‌توان دریافت که الگوریتم SVM در ۸۴ درصد مواقع در این مورد مطالعاتی دقت داشته است.

معیارهای ارزیابی

برای ارزیابی عملکرد مدل‌ها از دقت کلی (Overall Accuracy) و ضریب کاپا (Kappa Coefficient) و میانگین‌های وزنی Precision، Recall، F1-Score استفاده شد که در زیر به بیان روابط هر یک از آن‌ها پرداخته شده است.

- دقت کلی (Overall Accuracy, OA): بر اساس معادله ۱۰ محاسبه می‌شود.

$$OA = (\Sigma\ xii) / N \qquad (۱۰)$$

که در آن:

K: تعداد کلاس‌ها

xii: تعداد نمونه‌های صحیح طبقه‌بندی‌شده در کلاس i (درایه‌های قطر اصلی ماتریس اشتباه)

N: کل تعداد نمونه‌ها

-ضریب کاپا (Kappa Coefficient, κ): بر اساس معادله ۱۱ بیان می‌شود.

$$\kappa = (Po - Pe) / (1 - Pe)$$

$$Po = \Sigma\ xii / N$$
$$Pe = \Sigma\ (xi+ * x+i) / N^2 \qquad (۱۱)$$

Po: احتمال توافق واقعی = دقت کلی

Pe: احتمال توافق تصادفی

xi+: جمع سطر i ماتریس اشتباه (تعداد نمونه‌های طبقه‌بندی شده در کلاس i)

x+i: جمع ستون i ماتریس اشتباه (تعداد نمونه‌های واقعی کلاس i)

N: کل تعداد نمونه‌ها

- میانگین وزنی Precision، Recall، F1-Score: که به ترتیب طبق معادله‌های ۱۲ تا ۱۴ محاسبه می‌شود.

$$\text{Precision weighted} = \Sigma \left((TP\ i) / (TP\ i + FP\ i) \times (n\ i / N) \right) \quad (۱۲)$$

$$\text{Recall weighted} = \Sigma \left((TP\ i) / (TP\ i + FN\ i) \times (n\ i / N) \right) \quad (۱۳)$$

$$F1\ \text{weighted} = \Sigma \left((2 \times \text{Precision i} \times \text{Recall i}) / (\text{Precision i} + \text{Recall i}) \times (n\ i / N) \right) \quad (۱۴)$$

که در آن‌ها:

K: تعداد کلاس‌ها

TPi: تعداد نمونه‌های مثبت درست (True Positives) در کلاس i

FPi: نمونه‌های مثبت اشتباه (False Positives) در کلاس i

FNi: تعداد نمونه‌های منفی اشتباه (False Negatives) در کلاس i

ni: تعداد نمونه‌های واقعی در کلاس i

$N = \Sigma\ ni$: کل نمونه‌ها در همه کلاس‌ها

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

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

نتایج مقایسه عملکرد SVM و K-means بر اساس معیارهای ارزیابی در جدول ۱ ارائه شده است.

جدول ۱. مقایسه عملکرد SVM و K-means بر اساس معیارهای ارزیابی

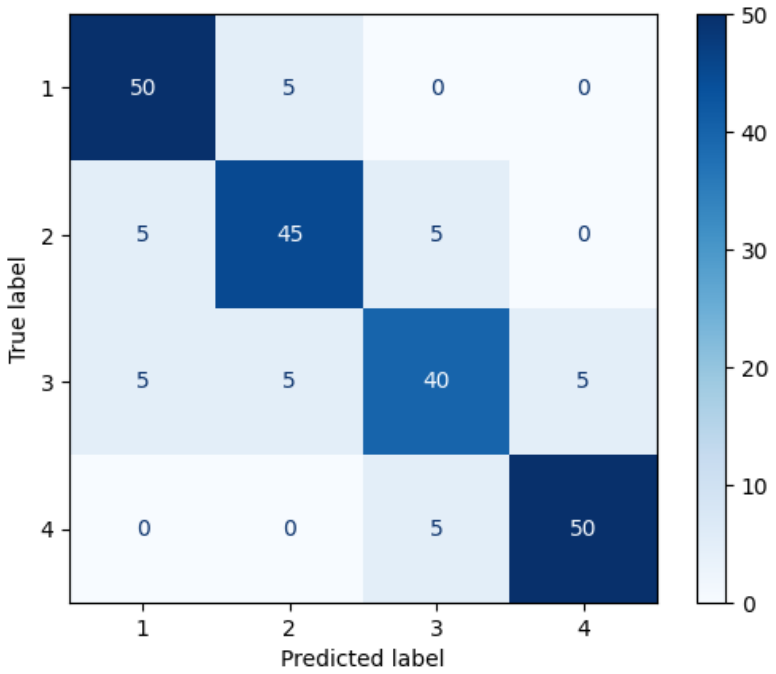
Table 1. Comparison of classification performance metrics between SVM and K-means

Evaluation Metric	SVM	K-means
Overall Accuracy (OA)	84%	73%
Kappa Coefficient	0.81	0.68
Weighted Precision	0.82	0.65
Weighted Recall	0.83	0.66
Weighted F1-Score	0.82	0.64

در ادامه ماتریس درهم‌ریختگی این مدل در شکل ۴ ارائه شده و همانطور که مشاهده می‌شود مقادیر قطری بالایی را نشان می‌دهد که نشان‌دهنده طبقه‌بندی دقیق است. توانایی الگوریتم در مدیریت تبدیل داده‌های پیچیده و استحکام آن در تعریف مرزهای بهینه بین کلاس‌های مختلف به دقت بالاتر آن کمک کرده است. در حالی که خوشه‌بندی K-means نیز عملکرد خوبی داشت، اما دقت به مراتب کمتری در خصوص تفکیک مناطقی با پوشش گیاهی از خاک در مقایسه با SVM دارد. خوشه‌های تشکیل شده توسط K-means به اندازه کافی متمایز نبودند و نشان‌دهنده همپوشانی بین انواع مختلف پوشش ارضی بودند. به عنوان یک الگوریتم بدون نظارت، K-means به انتخاب اولیه مراکز خوشه‌ها وابسته است که می‌تواند بر نتیجه نهایی خوشه‌بندی تأثیر بگذارد. با این حال، K-means همچنان طبقه‌بندی قابل قبولی را در خصوص تفکیک آب از خشکی فراهم می‌کند. هر دو الگوریتم SVM و خوشه‌بندی K-means نسبت به شاخص‌های طیفی به تنهایی دارای دقت بالاتری در طبقه‌بندی هستند. این نتیجه بیانگر وجود پتانسیل بهبود الگوریتم‌های یادگیری ماشین جهت تقویت طبقه‌بندی پوشش اراضی با شاخص‌های طیفی را برجسته می‌کند. بنابراین در یک تحلیل کلی می‌توان بیان کرد که:

SVM با استفاده از کرنل غیرخطیRBF، با دقت ۸۴٪ و ضریب کاپا ۰/۸۱ قادر به تفکیک دقیق‌تر کلاس‌ها بوده و K-means به دلیل

وابستگی به انتخاب اولیه مراکز، با دقت کلی ۷۳٪ و ضریب کاپا ۰/۶۸دقت پایین‌تری داشت. از دلایل عملکرد بهتر SVM می‌توان به ماهیت نظارت‌شده و نیز پیدا کردن مرز بهینه و از محدودیت‌های K-means به حساسیت به مراکز اولیه و نبود برچسب اشاره کرد. SVM با دقت بالا در تفکیک مناطق شهری از خاک برهنه عملکرد مطلوبی داشت، در حالی که K-means در این موارد همپوشانی بیشتری نشان داد. ترکیب شاخص‌های طیفی با الگوریتم‌های یادگیری ماشین، به‌ویژه با استفاده از PCA، دقت طبقه‌بندی را بهبود بخشید. این یافته‌ها نشان می‌دهند که ادغام روش‌های پیشرفته یادگیری ماشین با داده‌های سنجش از دور، ابزار قدرتمندی برای مطالعات زمین‌شناسی مهندسی است.



شکل ۴- ماتریس درهم‌ریختگی SVM برای طبقه‌بندی پوشش زمین

Fig. 4. Confusion matrix of the SVM classifier

مقایسه با مطالعات مشابه

نتایج این پژوهش با یافته‌های مروز و سوبیرای (Mróz and Sobieraj, 2004) و جاود و همکاران (Javed et al., 2021) هم‌خوانی دارد که نشان دادند ترکیب شاخص‌های طیفی و یادگیری ماشین موجب بهبود طبقه‌بندی می‌شود. تفاوت عمده مطالعه حاضر، دستیابی به دقت بالاتر SVM است که می‌تواند ناشی از پیش‌پردازش داده‌ها و استفاده از تصاویر Landsat 8 باشد.

نتیجه‌گیری

این پژوهش نشان می‌دهد که الگوریتم‌های یادگیری ماشین، به‌ویژه SVM، می‌توانند به‌طور قابل‌توجهی دقت طبقه‌بندی پوشش زمین را در مطالعات زمین‌شناسی مهندسی بهبود بخشند. یافته‌های کلیدی نشان می‌دهد که SVM با دقت کلی ۸۴٪ و ضریب کاپا ۰/۸۱عملکرد بهتری نسبت به K-means با دقت ۷۳٪ و ضریب کاپا ۰/۶۸دارد. این برتری ناشی از ماهیت نظارت‌شده SVM و توانایی آن در یادگیری الگوهای پیچیده و تعریف مرزهای بهینه بین کلاس‌هاست.

اگرچه شاخص‌های طیفی (NDVI, NDWI, SAVI, UI) پایه‌ای محکم برای طبقه‌بندی پوشش زمین فراهم می‌کنند، اما به تنهایی دقت محدودی دارند. با این حال، ترکیب این شاخص‌ها با الگوریتم‌های یادگیری ماشین، به‌ویژه در چارچوبی یکپارچه و قابل تکرار، دقت طبقه‌بندی را به‌طور چشمگیری افزایش می‌دهد.

نکته برجسته این پژوهش، ارائه یک چارچوب مبتنی بر پایتون است که تمام مراحل پیش‌پردازش، محاسبه شاخص‌ها، اجرای الگوریتم‌ها و ارزیابی نتایج را به‌صورت شفاف و قابل تکرار انجام می‌دهد. این چارچوب می‌تواند به‌عنوان یک ابزار عملیاتی در مطالعات آینده در حوزه زمین‌شناسی مهندسی، به‌ویژه در طبقه‌بندی پوشش زمین به منظور ارزیابی ناپایداری دامنه‌ها، مورد استفاده قرار گیرد.

اگرچه دقت K-means کمتر از SVM است، اما همچنان یک روش قابل قبول برای طبقه‌بندی پوشش زمین در سناریوهایی است که داده‌های برچسب‌دار در دسترس نیستند. به‌طور کلی، ادغام الگوریتم‌های یادگیری ماشین با شاخص‌های طیفی یک رویکرد امیدوارکننده برای طبقه‌بندی دقیق و کارآمد پوشش زمین است.

برای تحقیقات آتی، استفاده از الگوریتم‌های پیشرفته‌تر مانند XGBoost یا شبکه‌های عصبی پیچشی (CNN) به همراه شاخص‌های طیفی و تکنیک‌های کاهش بعد پیشنهاد می‌شود تا عملکرد طبقه‌بندی بهبود یابد.

References

Ashrafi Igder, M., Liang, X., Mitolo, M., 2022. Service Restoration Through Microgrid Formation in Distribution Networks: A Review. IEEE Access 10, 46618-46632.

Azari, T., 2025. Design of a supervised artificial intelligence committee machine to estimate hydrodynamic parameters of confined aquifers. Kharazmi Journal of Earth Sciences 11 (1), 28-46.

Berk, R. A., 2008. Support vector machines. Statistical Learning from a Regression Perspective. Springer Nature Switzerland AG.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

Gholami Dargahi, M., Pourbeyranvand, Sh., Talkhablou, M., 2021. Stress inversion studies using 3D seismic interpretation data and earthquakes focal mechanism. Kharazmi Journal of Earth Sciences 7 (2), 317-341.

Ishfagh, A., Atiq ur, R., Khan DM, Khan Z, Shafiq M, Choi J-G., 2022. Model Selection Using K-means Clustering Algorithm for the Symmetrical Segmentation of Remote Sensing Datasets. Symmetry 14(6), 1149.

Javed, A., Cheng, Q., Peng, H., Altan, O., Li, Y., Ara, I., Saleem, N., 2021. Review of Spectral Indices for Urban Remote Sensing. Photogrammetric Engineering and Remote Sensing 87(7), 513-524.

Jolliffe, I. T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065), 20150202.

Kwedlo, W., 2011. A clustering method combining differential evolution with the Kmeans algorithm. Pattern Recognition Letters 32(12), 1613-1621.

Mróz, M., Sobieraj, A., 2004. Comparison of several vegetation indices calculated on the basis of a seasonal SPOT XS time series and their suitability for land cover and agricultural crop identification. Technical Sciences 7, 39–66

Pedregosa, F., 2011. Scikit-learn: Machine learning in python Fabian. Journal of machine learning research 12, 2825.

Peshawa, J.M.A., 2022. Investigating the Impact of min-max data normalization on the regression performance of K-nearest neighbor with different similarity measurements. ARO-The Scientific Journal of Koya University 10(1), 85-91.

Petrelli, M., 2021. Introduction to Python in Earth Science Data Analysis: From Descriptive Statistics to Machine Learning. Springer Nature.

Petrelli, M., 2023. Machine Learning for Earth Sciences Using Python to Solve Geological Problem. Springer Textbooks in Earth Sciences, Geography and Environment.

Talkhablou, M., Fatemi Aghda, S.M., Milani Chegooshi, H., 2023. Comparison of main rock mass classification methods with an attitude on determining the initial support. Kharazmi Journal of Earth Sciences 8 (2), 94-116.

Van Westen, C. J., Castellanos, E., Kuriakose, S. L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Engineering Geology 102(3-4), 112-131.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International journal of remote sensing 27(14), 3025-3033.

Zhang, C., Ma, X., Li, W., 2020. A review on deep learning in remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 162, 1-14.