

Research Paper

Investigating water quality in wetlands using multivariate statistical analysis and machine learning models

Abdolsamad Davoodi¹, Reza Mohammadpour^{2*}, Touraj Sabzevari³

1. PhD student in Water and Hydraulic Structures, Islamic Azad University, Estahban Branch, Fars, Estahban. davoodi.asd16@yahoo.com

2*. Assistant Professor, Water Department, Estahban Branch, Islamic Azad University, Estahban Branch, reza564@gmail.com

3. Associate Professor, Water Department, Estahban Branch, Islamic Azad University, Estahban Branch tooraj.sabzevari@gmail.com

Received: 2024/12/08

Revised: 2025/01/11

Accepted: 2025/07/02

Use your device to scan and read the article online



DOI:

[10.71632/wej.2025.1192892](https://doi.org/10.71632/wej.2025.1192892)

Keywords:

hierarchical cluster analysis; machine learning; water quality factors; principal component analysis; constructed wetlands; random forest; XGBoost

Abstract

Introduction: Water is considered one of the main foundations of sustainable development of societies, while clean water resources are a major prerequisite for environmental protection and economic, political, social and cultural development. The increasing demand for water, increasing living standards and the spread of water resource pollution due to the development of agricultural, urban and industrial activities have led to a chaotic environmental situation and intensified water resource pollution, which will make it difficult to control.

Methods: Multivariate statistical methods and data mining have been used to investigate water quality in many studies. Cluster analysis (CA) and discriminant analysis (DA) were used to identify pollution sources in river basins. In order to systematically compare the assumptions of the analytical methods used, the theoretical foundations of each method were examined. Nonparametric methods such as percentage elimination (PR) and sign test (ST) were applicable without the need to assume a specific data distribution, while classical multivariate methods including PCA and FA were used with the assumption of multivariate normality and linear relationships between variables (as confirmed by KMO and Bartlett tests). Machine learning models including Random Forest and XGBoost with the ability to analyze nonlinear relationships and resist collinearity, SVM with sensitivity to feature scaling and the need for separable space, and regression methods such as PLS and Stepwise with the assumption of linear relationships and the need for cross-validation to prevent overfitting were used.

Findings: According to the results obtained from the statistical methods of percentage elimination and sign test, it was observed that the wetland plays a fundamental and key role in the entire drainage system; therefore, using the statistical methods of LDA, PCA/FA and HACA, all water quality factors in the wetland are examined. Also, principal component analysis (PCA) plays a positive role in prioritizing the importance of each factor in pollution, so that it places the more important factors in the first component and the less important factors in the subsequent components. The results obtained from the principal component analysis show that the components with more than one eigenvalue are considered the most important components that justify the variance.

Citation: Davoodi A, Mohammadpour R, Sabzevari T.. Investigating water quality in wetlands using multivariate statistical analysis and machine learning models. Water Resources Engineering Journal. 2025; 18(66): 35- 52.

***Corresponding author:** Reza Mohammadpour

Address: Assistant Professor, Water Department, Estahban Branch, Islamic Azad University, Estahban Branch, Iran, reza564@gmail.com

Tell: 09173134112

Email: reza564@gmail.com

Extended Abstract

Introduction

Water is a fundamental resource for the sustainable development of societies, and in recent years, there has been increasing attention to water quality alongside its quantity (Zhang et al., 2021). This is while clean water resources are the main prerequisite for environmental protection and economic, political, social and cultural development. The increasing demand for water, increasing living standards and the spread of water resource pollution due to the development of agricultural, urban and industrial activities have caused a chaotic environmental situation and intensified water resource pollution, which will make it difficult to control. And as a result, it has created health risks for people dependent on these water resources (M. Mirhashmi, 2023). Predicting water quality is a critical issue due to its significant impact on human health and ecosystems (Quoc Bao Pham, 2021). Surface waters, including rivers, lakes, wetlands, and other bodies of water, are among the first environmental interfaces that are highly affected by problems caused by chemical pollution. Among them, wetlands are considered one of these water resources and ecosystems (M. Mirhashmi, 2023). Also, Surface waters are vital resources for drinking, public use, irrigation, and supporting aquatic life. However, the discharge of urban and industrial waste into rivers, lakes, and reservoirs has significantly compromised the quality of these resources (Nguyen et al., 2021). Wetlands are one of the richest and most valuable ecosystems on Earth. As the most active ecosystem in the exchange of materials and energy on the Earth's surface, wetlands play an irreplaceable role in flood mitigation, carbon cycling, water cycling, climate regulation, and biodiversity conservation, etc (Wenlan Yang, 2023). Monitoring water quality is crucial to ensure safety and mitigate the negative impacts of pollutants (Guerrero et al., 2020). Any changes in the physical, chemical, and biological characteristics of water can indicate the presence of pollutants; thus, it is essential to monitor

factors such as conductivity, pH, turbidity, temperature, and dissolved oxygen (Casillas et al., 2021).

Materials and Methods

Data collection, encompassing eleven parameters, was conducted to evaluate water quality in the drainage system during 2010, 2011, and 2012. Sampling occurred once a month from the site. The measured parameters included temperature, conductivity, biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), ammonia (NH₃), phosphate (PO₄³⁻), nitrite (NO₂⁻), nitrate (NO₃⁻), pH, and dissolved oxygen (DO). Five multivariate statistical methods were employed to analyze the relationships between the physical and chemical elements of water: percentage removal (PR), sign test (ST), machine learning models, principal component analysis (PCA), factor analysis (FA), and hierarchical cluster analysis (HACA).

Findings and Discussion

To determine the water quality factors across the drainage system, nitrite (NO₂⁻), nitrate (NO₃⁻), ammonia (NH₃), and phosphate (PO₄³⁻) were analyzed. The water quality performance was assessed using the percentage removal method and the sign test, with pollutant concentrations measured at both the inlet and outlet of different drainage sections. A comprehensive evaluation and comparison of various machine learning and regression models—including Partial Least Squares (PLS), Bagging (BW), Stepwise (LM), Forward (FW), Support Vector Machines (SVM), XGBoost, and Random Forest—were conducted based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). The variable importance in projection (VIP) analysis revealed that phosphate, TSS, and ammonia (AN) were the most significant factors influencing both the random forest and XGBoost models. The agreement between these models underscores the

importance of these variables in predicting and modeling water quality. PCA proved beneficial in prioritizing factors affecting pollution; more significant factors were incorporated into the first principal component, while less significant factors were assigned to subsequent components.

Conclusion

Multivariate statistical methods, including percentage removal, sign test, HACA, and PCA/FA, were utilized to analyze the Water Quality Index (WQI) and identify its influencing factors. This research measured eleven water quality parameters within the drainage system. The findings indicated that while pollutant levels increased in certain areas, an overall decrease in pollution was observed for parameters such as nitrate, nitrite, and ammonia, highlighting the wetland's crucial role in mitigating these pollutants. For more precise predictions and modeling of nitrite—one of the critical parameters—random forest and XGBoost machine learning models were employed, demonstrating superior accuracy compared to traditional methods in modeling complex relationships. The results suggest that phosphate, ammonia, and TSS are significant factors affecting nitrite levels and thus should be considered in water quality management. Cluster analysis (HACA) classified the drainage system into three distinct clusters, each exhibiting unique characteristics and variations in water quality parameters. The first cluster encompassed the upper sections of the lagoon, the second the middle sections, and the third the farthest regions, with temperature showing the least variation and nitrite exhibiting the greatest. PCA also highlighted the most critical parameters within each cluster: four principal components in the first cluster explained 69.7% of the variance, four components in the second accounted for 66.6%, and three components in the third covered 58.9%. These analyses demonstrated that parameters such as nitrate, phosphate, and ammonia significantly impact water quality. Employing multivariate statistical methods facilitated a targeted analysis of extensive water quality data, enabling more accurate

identification of critical points and enhancing water resource management through cost reduction and optimized monitoring.

Conflicts of interest

The authors declared no conflict of interest.

مقاله پژوهشی

بررسی کیفیت آب در تالاب‌ها با استفاده از تجزیه و تحلیل روش‌های آماری چندمتغیره و مدل‌های یادگیری ماشین

عبدالصمد داودی^۱، رضا محمدپور^{۲*} و تورج سبزواری^۳

۱. دانشجوی دکتری آب و سازه های هیدرولیکی، دانشگاه آزاد اسلامی واحد استهبان، فارس، استهبان

۲. استادیار گروه آب واحد استهبان، دانشگاه آزاد اسلامی واحد استهبان

۳. دانشیار گروه آب واحد استهبان، دانشگاه آزاد اسلامی واحد استهبان

چکیده

مقدمه: آب به‌عنوان یکی از پایه‌های اصلی توسعه پایدار جوامع به‌شمار می‌رود این درحالی است که منابع آب سالم پیش نیاز اصلی برای حفظ محیط زیست و گسترش اقتصادی، سیاسی، اجتماعی و فرهنگی می باشد. بالا رفتن تقاضای آب، افزایش سطح زندگی و گسترش آلودگی منابع آب در اثر توسعه فعالیت های کشاورزی، شهری و صنعتی موجب وضعیت نابسامان زیست محیطی و تشدید آلودگی منابع آب شده که سبب سخت شدن کنترل آن خواهد شد.

روش: روش‌های آماری چندمتغیره و داده‌کاوی برای بررسی کیفیت آب در پژوهش‌های بسیاری به‌کار گرفته شده‌اند. با استفاده از تحلیل خوشه‌ای (CA) و تحلیل تشخیصی (DA) به شناسایی منابع آلودگی در حوضه‌های رودخانه‌ای پرداختند. به منظور مقایسه نظام‌مند فرضیات روش‌های تحلیلی به‌کاررفته، مبانی نظری هر روش مورد بررسی قرار گرفت. روش‌های غیرپارامتریک مانند حذف درصد (PR) و آزمون علامت (ST) بدون نیاز به فرض توزیع خاص داده‌ها قابل اجرا بودند، در حالی که روش‌های چندمتغیره کلاسیک شامل PCA و FA با فرض نرمال بودن چندمتغیره و روابط خطی بین متغیرها) با تأیید آزمون‌های KMO و بارتلت (به کار گرفته شدند. مدل‌های یادگیری ماشین شامل جنگل تصادفی و XGBoost با قابلیت تحلیل روابط غیرخطی و مقاومت در برابر هم‌خطی، SVM با حساسیت به مقیاس‌گذاری ویژگی‌ها و نیاز به فضای تفکیک‌پذیر، و روش‌های رگرسیونی نظیر PLS و Stepwise با فرض روابط خطی و نیاز به اعتبارسنجی متقابل برای جلوگیری از بیش‌برازش مورد استفاده قرار گرفتند.

یافته‌ها: باتوجه‌به نتایج به‌دست‌آمده از روش‌های آماری حذف درصد و آزمون علامت، مشاهده شد که تالاب نقش اساسی و کلیدی در کل سامانه زهکشی ایفا می‌کند؛ بنابراین با استفاده از روش‌های آماری LDA، PCA/FA و HACA کلیه عامل‌های کیفی آب را در تالاب مورد بررسی قرار می‌گیرد. همچنین تحلیل مؤلفه‌های اصلی (PCA) در اولویت‌بندی اهمیت هر یک از عامل‌ها در آلودگی نقش مثبتی ایفا می‌کند، به‌طوری‌که عامل‌های مهم‌تر را در مؤلفه اول و عامل‌های کم‌اهمیت را در مؤلفه‌های بعدی قرار می‌دهد. نتایج به‌دست‌آمده از تحلیل مؤلفه‌های اصلی نشان می‌دهد که مؤلفه با داشتن مقدار ویژه بیش از یک عنوان، مهم‌ترین مؤلفه‌های توجیه‌کننده واریانس در نظر گرفته شده‌اند.

نتیجه‌گیری: به‌طورکلی، این روش‌های آماری چندمتغیره کمک کردند که حجم گسترده‌ای از داده‌های کیفی آب به‌صورت هدفمند تحلیل شود. این رویکرد با کاهش هزینه‌ها و بهینه‌سازی پایش، امکان شناسایی دقیق‌تر نقاط حساس و بهبود مدیریت منابع آبی را فراهم می‌آورد.

تاریخ دریافت: ۱۴۰۳/۰۹/۱۸

تاریخ داوری: ۱۴۰۳/۱۰/۲۲

تاریخ پذیرش: ۱۴۰۴/۰۴/۱۱

از دستگاه خود برای اسکن و خواندن مقاله به صورت آنلاین استفاده کنید



DOI:

[10.71632/wej.2025.1192892](https://doi.org/10.71632/wej.2025.1192892)

واژه‌های کلیدی:

تحلیل خوشه‌ای سلسله‌مراتبی، یادگیری ماشین، عامل‌های کیفی آب، تحلیل مؤلفه‌های اصلی، تالاب‌های ساخته‌شده، جنگل تصادفی، XGBoost، جامدات معلق کل.

* نویسنده مسئول: رضا محمدپور

نشانی: استادیار گروه آب واحد استهبان، دانشگاه آزاد اسلامی واحد استهبان، استهبان، ایران.

تلفن: ۰۹۱۷۳۱۳۴۱۱۲

پست الکترونیکی: reza564@gmail.com

مقدمه

آب به‌عنوان یکی از پایه‌های اصلی توسعه پایدار جوامع به‌شمار می‌رود، و در سال‌های اخیر توجه به کیفیت آب، علاوه بر کمیت آن، افزایش یافته است (۳۲). آب‌های سطحی منابع مهمی هستند که برای آب آشامیدنی، استفاده عمومی، آبیاری و حیات آبریان به‌کار می‌روند. با این حال، تخلیه زیاده‌های شهری و صنعتی به رودخانه‌ها، دریاچه‌ها و مخازن باعث کاهش کیفیت این منابع شده است (۲۶). این درحالی است که منابع آب سالم پیش نیاز اصلی برای حفظ محیط زیست و گسترش اقتصادی، سیاسی، اجتماعی و فرهنگی می‌باشد. بالا رفتن تقاضای آب، افزایش سطح زندگی و گسترش آلودگی منابع آب در اثر توسعه فعالیت‌های کشاورزی، شهری و صنعتی موجب وضعیت نابسامان زیست محیطی و تشدید آلودگی منابع آب شده که سبب سخت شدن کنترل آن خواهد شد. و در پی آن مخاطرات بهداشتی برای افراد وابسته به این منابع آب را به وجود آورده است (۱). پیش‌بینی کیفیت آب به دلیل تأثیر قابل توجه آن بر سلامت انسان و اکوسیستم، یک مسئله حیاتی است (۳۱). عوامل طبیعی که بر کیفیت آب تأثیر منفی می‌گذارند، شامل عوامل اقلیمی، هیدرولوژیکی، سنگ‌شناسی و توپوگرافی هستند همچنین فعالیت‌های انسانی مانند گسترش شهرنشینی، افزایش فعالیت‌های صنعتی، بالا رفتن روان آب‌ها و فعالیت‌های مرتبط با گردشگری می‌تواند سبب کاهش کیفیت مطلوب آب شوند (۱۹) لذا نظارت بر کیفیت آب برای اطمینان از ایمنی آن و جلوگیری از تأثیرات منفی آلاینده‌ها اهمیت بسیاری دارد (۱۶). هرگونه تغییر در خصوصیات فیزیکی، شیمیایی و بیولوژیکی آب، وجود آلاینده‌ها را نشان می‌دهد؛ بنابراین نظارت بر عواملی مانند رسانایی، pH، کدورت، دما و اکسیژن محلول ضروری است (۷).

آب‌های سطحی شامل رودخانه‌ها، دریاچه‌ها، تالاب‌ها و دیگر پهنه‌های آبی از نخستین محیط‌های واسطه زیست محیطی به‌شمار می‌آیند که بسیار تحت تأثیر مشکلات ناشی از آلودگی‌های شیمیایی هستند. در این میان، تالاب‌ها یکی از این منابع آبی و اکوسیستمی با اهمیت به حساب می‌روند (۱).

تالاب‌ها یکی از غنی‌ترین و ارزشمندترین اکوسیستم‌های روی زمین هستند. تالاب‌ها به‌عنوان فعال‌ترین اکوسیستم در تبادل مواد و انرژی در سطح زمین، نقشی غیرقابل جایگزین در کاهش سیل، چرخه کربن، چرخه آب، تنظیم آب و هوا و حفظ تنوع زیستی و غیره ایفا می‌کنند (۳۶). در مناطق شهری، به دلیل کاهش پوشش گیاهی و تأثیرات آلودگی صنعتی، آلودگی منابع آب به یک معضل جدی تبدیل شده است (۲۴). کیفیت پایین آب‌های سطحی، علاوه بر سلامت انسان، اکوسیستم‌ها و حیات گیاهان و حیوانات را نیز تهدید می‌کند و در نتیجه، کیفیت آب (WQ) به یکی از نگرانی‌های اصلی در مدیریت منابع آب و سامانه‌های زیست محیطی تبدیل شده است (۵). ارزیابی کیفیت آب به تحلیل خواص آن با توجه به تأثیرات سلامتی انسان و محیط زیست کمک می‌کند (۲). به همین دلیل، استفاده از روش‌های چندمتغیره و تحلیل‌های کاوشی داده‌ها در نظارت و تحلیل کیفیت آب اهمیت یافته‌اند (۲۸).

هرچند مطالعه حاضر تلاش کرده با بهره‌گیری از داده‌های گسترده و روش‌های تحلیلی پیشرفته، تصویری دقیق از کیفیت آب در تالاب نیبونگ تبال ارائه دهد، اما دارای محدودیت‌هایی نیز هست. محدود بودن بازه زمانی به سال‌های ۲۰۱۰ تا ۲۰۱۲، تمرکز صرف بر ۱۱ پارامتر کیفی، عدم بررسی فلزات سنگین یا میکروارگانیسم‌ها، و وابستگی به داده‌های مکانی خاص، قابلیت تعمیم‌پذیری نتایج را محدود می‌سازد. همچنین، مدل‌های مورد استفاده دارای پیچیدگی محاسباتی (مانند XGBoost یا حساسیت به مقیاس ویژگی‌ها) مانند (SVM) هستند که باید در تفسیر نتایج مدنظر قرار گیرد.

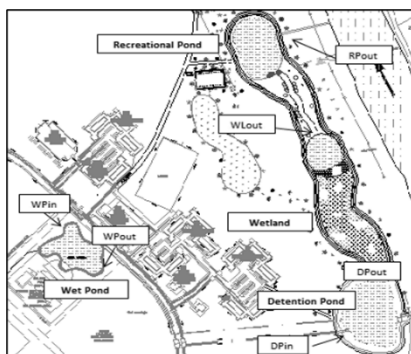
مواد و روش‌ها

روش‌های آماری چندمتغیره و داده‌کاوی برای بررسی کیفیت آب در پژوهش‌های بسیاری به‌کار گرفته شده‌اند (۳۲). با استفاده از تحلیل خوشه‌ای (CA) و تحلیل تشخیصی (DA) به شناسایی منابع آلودگی در حوضه‌های رودخانه‌ای پرداختند (۱۶). در مطالعه‌ای بر روی مخازن آب آتن از تکنیک‌های PCA، FA، CA و DA استفاده کردند و نتایج مثبتی در شناسایی و طبقه‌بندی نقاط نمونه‌برداری به‌دست آوردند (۳). کیفیت آب زیرزمینی دشت فسا را با روش‌های CA، PCA و LDA بررسی کردند و موفق به شناسایی پارامترهای اصلی مؤثر بر کیفیت آب شدند (۲۴). نیز به تحلیل مکانی و زمانی کیفیت آب سطحی در استان دوتنگ تاپ ویتنام پرداخته و از روش‌های CA و PCA برای ارزیابی همبستگی بین پارامترهای کیفیت آب بهره بردند (۱۸). به بررسی اثر تالاب‌های ساخته‌شده بر کاهش نیتروژن با استفاده از روش‌های DFA و HACA پرداختند (۲۶). تغییرات مکانی کیفیت آب‌های سطحی در مالاکا، مالزی را با استفاده از PCA و LDA تحلیل کرده و به شناسایی الگوهای کلیدی دست یافتند (۷). کیفیت آب رودخانه سایگون-دوتنگ نای را با استفاده از PCA، LDA و مدل MIKE 11 تحلیل کرده و منابع آلودگی و الگوهای مکانی-زمانی را شناسایی کردند.

ژنگ و همکاران (۲۰۲۱) با استفاده از PCA و LDA به بررسی شاخص کیفیت آب برای حفاظت از آبریان در رودخانه سانتیاگو-گوادالاخارا پرداختند (۳۲). آیدین و همکاران (۲۰۲۱) با استفاده از PCA، CA و FA به ارزیابی کیفیت آب و شناسایی منابع آلودگی در رودخانه شوانگجی چین پرداختند (۲). Aydin et al. (2021) با استفاده از شاخص کیفیت آب و روش‌های آماری چندمتغیره، کیفیت آب رودخانه‌های شمال شرق ترکیه را ارزیابی کردند و عوامل مؤثر بر آلودگی را شناسایی نمودند. همچنین، یوما و مکاران (۲۰۱۸) در بررسی عملکرد تالاب‌های پایین‌دست رودخانه‌ها نشان دادند که تالاب‌ها به‌طور مؤثری می‌توانند نیتروژن را کاهش دهند (۲۸).

هدف اصلی تحقیق حاضر ارزیابی تغییرات مکانی و زمانی کیفیت آب‌های سطحی و شناسایی پارامترهای کلیدی مؤثر بر کیفیت آب در تالاب نیبونگ تبال است.

منطقه مورد مطالعه همان‌طور که در شکل ۱ نشان داده شده است، با موقعیت جغرافیایی $5^{\circ}8'40.9023''$ و $100^{\circ}29'32.9273''$ در ۲ کیلومتری جنوب شرقی مرکز شهر Nibong Tebal و ۱/۵



شکل ۲- ورودی و خروجی قسمت‌های مختلف سیستم زهکشی
Beh, C. H. (2014)

جمع‌آوری داده‌ها که شامل ۱۱ پارامتر می باشد برای ارزیابی کیفیت آب در سیستم زهکشی در سال ۲۰۱۱، ۲۰۱۰ و ۲۰۱۲ انجام شد. نمونه گیری ها یک بار در ماه از محل انجام شد. پارامترهای اندازه‌گیری شده عبارتند از: COD, BOD, Conductivity, Temperature, DO, pH, NO₃⁻, NO₂⁻, PO₄³⁻, NH₃, TSS می باشند. در ادامه با استفاده از پنج روش آماری چندمتغیره، رابطه بین عناصر شیمیایی - فیزیکی آب به دست آمد. این پنج روش عبارتند از: حذف درصد (PR)، آزمون علامت (ST)، مدل های یادگیری ماشین، تحلیل مؤلفه اصلی (PCA) و تحلیل عاملی (FA) و تحلیل خوشه‌ای سلسله‌مراتبی (HACA).

در این مطالعه، به منظور مقایسه نظام‌مند فرضیات روش‌های تحلیلی به کاررفته، مبانی نظری هر روش مورد بررسی قرار گرفت. روش‌های غیرپارامتریک مانند حذف درصد (PR) و آزمون علامت (ST) بدون نیاز به فرض توزیع خاص داده‌ها قابل اجرا بودند، در حالی که روش‌های چندمتغیره کلاسیک شامل PCA و FA با فرض نرمال بودن چندمتغیره و روابط خطی بین متغیرها (با تأیید آزمون‌های KMO و بارتلت (به کار گرفته شدند. مدل‌های یادگیری ماشین شامل جنگل تصادفی و XGBoost با قابلیت تحلیل روابط غیرخطی و مقاومت در برابر هم‌خطی، SVM با حساسیت به مقیاس‌گذاری ویژگی‌ها و نیاز به فضای تفکیک‌پذیر، و روش‌های رگرسیونی نظیر PLS و Stepwise با فرض روابط خطی و نیاز به اعتبارسنجی متقابل برای جلوگیری از بیش‌برازش مورد استفاده قرار گرفتند. این تنوع روش‌شناسی با در نظر گرفتن نقاط قوت و محدودیت‌های هر روش، امکان تحلیل جامع داده‌های کیفیت آب را فراهم نمود.

همچنین به منظور اطمینان از قابلیت استفاده داده‌ها در تحلیل‌های آماری، پیش از اجرای مدل‌ها، آزمون‌های آماری اولیه انجام شد. شاخص KMO برابر با ۰.۸۷۰ و نتایج معنی‌دار آزمون بارتلت نشان داد که داده‌ها برای اجرای تحلیل‌های PCA و FA مناسب هستند. همچنین، داده‌ها برای مدل‌های حساس به مقیاس (مانند SVM) استانداردسازی شدند.

۲-۱. حذف درصد (Percent Removal): معمولی ترین روش اندازه گیری برای طراحی و تجزیه و تحلیل کیفیت آب می باشد. عملکرد

کیلومتری شمال شرقی از مرکز شهر Parit Buntar در کشور مالزی واقع شده است. موقعیت جغرافیایی 47N 117E 296 (644296 , 592117 47N) (سیستم زهکشی مورد استفاده در پژوهش، در محیط مهندسی USM که در Nibong Tebal مالزی طراحی و اجرا شده است در شکل ۱ آمده است.

با وجود اینکه داده‌های تبخیر مستقیم در این مطالعه اندازه‌گیری نشده‌اند، بر اساس گزارش‌های اقلیمی اداره هواشناسی مالزی، میزان تبخیر سالانه در منطقه نیبونگ تبال حدود ۱۵۰۰ تا ۲۰۰۰ میلی‌متر در سال برآورد می‌شود. این مقدار می‌تواند منجر به افزایش نسبی غلظت آلاینده‌ها در فصول خشک شود و باید در تفسیر کیفیت آب لحاظ گردد. پیشنهاد می‌شود در مطالعات آتی، داده‌های دقیق‌تر تبخیر در سطح محلی نیز اندازه‌گیری شود.



شکل ۱- سایت تحقیقاتی در منطقه USM

قسمت‌های مختلف سامانه زهکشی که در مطالعه استفاده شده به‌صورت تصویر در شکل ۲ به طور کامل آورده شده که شامل گودال پوشش گیاهی (Grassed Swale)، حوضچه مرطوب (Wet Pond)، حوضچه تاخیری (Detention Pond)، تالاب (Wetland) و حوضچه تفریحی (Recreational Pond) است که در انتها به رودخانه منتهی می‌شود. مطابق شکل ۲ هر قسمت یک ورودی و یک خروجی دارد که رواناب را به قسمت بعدی منتقل می‌کنند.

هدف این تحقیق بررسی و ارزیابی کارایی حوضچه‌های مرطوب و تفریحی در بهبود کیفیت آب و کاهش آلاینده‌های مهمی همچون نیتریت و نیترات است. حوضچه‌های مرطوب به عنوان سامانه‌های تصفیه آب مبتنی بر فرآیندهای طبیعی عمل می‌کنند که با استفاده از گیاهان آبری و فعالیت میکروارگانیسم‌ها، آلاینده‌ها را کاهش می‌دهند. در کنار آن، حوضچه‌های تفریحی علاوه بر نقش زیست‌محیطی خود، برای ایجاد فضاهای تفریحی نیز طراحی شده‌اند. مسئله اصلی در این تحقیق این است که چگونه می‌توان با بهینه‌سازی طراحی و بهره‌برداری از این حوضچه‌ها، هم عملکرد تصفیه آب را بهبود بخشید و هم میزان آلاینده‌های خطرناک را به‌طور کارآمدتری کاهش داد. این تحقیق به دنبال یافتن راهکارهای علمی برای ارتقاء اثربخشی این حوضچه‌ها در حفظ محیط‌زیست و تأمین منابع آب پاکیزه است.

سیستم زهکشی در مقادیر کمیته، امکان مقایسه آسان بین سیستم‌های زهکشی و همچنین ترکیب آسان در طرح‌ها را فراهم می‌کند. روش حذف درصد با توجه به تفاوت غلظت جریان ورودی و غلظت جریان خروجی سیستم زهکشی صورت می‌گیرد که در معادله زیر آمده است:

$$C_{in}: \text{غلظت جریان ورودی} - C_{out}: \text{غلظت جریان خروجی}$$

$$PR = \frac{C_{in} - C_{out}}{C_{in}} \times 100 \quad (1)$$

۲-۲. آزمون علامت (Sign Test): آزمون علامت، متعلق به گروه روش‌های استنباط آماری ناپارامتری است و برای تشخیص اختلاف بین زوج مشاهدات مناسب است. طبق آزمون‌های ناپارامتری دیگر در مورد آزمون علامت نیز هیچ محدودیتی راجع به نوع توزیع متغیر مورد نظر وجود ندارد، اما مقادیر متغیر مورد نظر باید پیوستگی داشته و مقیاس آن از نوع ترتیبی باشد. عبارت دیگر اجرای این آزمون برای متغیرهایی که دارای مقوله‌های محدودی هستند امکان پذیر نیست.

۲-۳. مدل‌های آماری و یادگیری ماشین: در این مطالعه، برای مدل‌سازی و پیش‌بینی متغیر نیتريت به‌عنوان متغیری پیوسته، از ترکیبی از مدل‌های آماری و یادگیری ماشین بهره گرفته شده است. مدل‌های آماری مانند رگرسیون خطی (LM) و رگرسیون گام‌به‌گام (Stepwise) بر اساس مفاهیم آماری کلاسیک طراحی شده‌اند که امکان بررسی رابطه بین متغیر وابسته و متغیرهای مستقل را فراهم می‌کنند. رگرسیون خطی به شکل استاندارد در تحلیل‌های علمی استفاده می‌شود و با یافتن رابطه خطی بین متغیرها، به پیش‌بینی مقادیر متغیر وابسته کمک می‌کند (Montgomery et al., 2020). مدل گام‌به‌گام نیز با افزودن و حذف متغیرها در هر گام، مجموعه‌ای از مهم‌ترین متغیرهای مستقل را شناسایی می‌کند (James et al., 2013).

در این میان، Partial Least Squares (PLS) نیز به‌عنوان یک تکنیک مدرن برای تحلیل‌هایی با تعداد زیادی متغیر مستقل مورد استفاده قرار گرفته است. این روش ترکیبی از رگرسیون و کاهش ابعاد است که همبستگی متغیرهای مستقل با متغیر وابسته را بیشتر می‌کند و در مواقعی که متغیرهای مستقل زیادی وجود دارند، مؤثر است. در کنار این روش‌های آماری، مدل‌های یادگیری ماشین از جمله جنگل تصادفی (Random Forest) و XGBoost به دلیل توانایی آن‌ها در شناسایی روابط پیچیده و بهینه‌سازی دقت، به کار گرفته شده‌اند (۳۲). جنگل تصادفی با ترکیب چندین درخت تصمیم، خطای مدل را کاهش داده و از بیش‌برازش جلوگیری می‌کند، در حالی که XGBoost به دلیل بهره‌گیری از تقویت گرادینان، به‌ویژه در داده‌های پیچیده و بزرگ، سرعت و دقت بالایی را فراهم می‌کند. چن و گسترین (۲۰۱۶) همچنین، ماشین بردار پشتیبان (SVM) به‌عنوان مدلی برای تحلیل داده‌های غیرخطی، با کرنل‌های مختلفی که امکان ایجاد خط تفکیک مناسب در فضای ویژگی‌های پیچیده را فراهم می‌کند، در این پژوهش به کار رفته است (۸).

در نتیجه، نتایج نشان داد که مدل‌های جنگل تصادفی و XGBoost عملکرد بهتری در پیش‌بینی مقادیر نیتريت دارند و قادر به شناسایی متغیرهای کلیدی مانند فسفات، TSS و آمونیاک نیتروژن (AN) به‌عنوان مهم‌ترین عوامل تأثیرگذار بر نیتريت بودند.

۲-۴. تجزیه و تحلیل مؤلفه‌های اصلی (PCA): این روش یکی از تکنیک‌های چندمتغیره آماری است که در مواردی که حجم داده‌ها زیاد باشد، می‌توان به‌عنوان راهکاری مناسب برای کاهش تعداد ورودی‌ها از آن استفاده نمود. تعداد ورودی‌ها قادرند تغییرات سیستم را به خوبی توصیف نمایند. قبل از انجام تحلیل عاملی داده‌ها بایستی شرایط لازم برای تحلیل عاملی و به منظور تناسب داده‌های هر پژوهش جهت بکارگیری روش تحلیل مؤلفه‌های اصلی، آزمون‌های KMO و آزمون کرویت بارتلت استفاده می‌شود. در تجزیه و تحلیل مؤلفه‌های اصلی دو معیار، آزمون KMO برای اندازه‌گیری کیفیت نمونه‌گیری و آزمون کرویت بارتلت برای تأیید اجرای تحلیل عاملی انجام می‌شوند. آزمون بارتلت نیز نشان می‌دهد که آیا ماتریس همبستگی یک ماتریس واحد است و توانایی نشان دادن متغیرهای نامرتبط را دارد یا خیر. در مرحله بعد برای کاهش تعداد عوامل استخراج شده، بارهای عاملی باید از طریق مقادیر ویژه ارزیابی شوند. با استفاده از PCA متغیرهای اولیه به متغیرهای جدید که ترکیب خطی از آنها هستند، تبدیل می‌شوند. در این روش هر مؤلفه اصلی می‌تواند با دنباله‌ای به صورت رابطه ۲ مشخص شود (۲۰؛ ۲۱):

$$Z_i = a_{i1} X_1 + a_{i2} X_2 + \dots + a_{ij} X_j \quad (2)$$

در فرمول (۲)، Z_i معرف مؤلفه موردنظر، a_{ij} ضرایب مربوط به متغیرهای اولیه و X_i نیز متغیرهای اولیه هستند. ضرایب مربوط به متغیرهای اولیه از حل رابطه ۳ بدست می‌آید (۲۰؛ ۲۱):

$$|R - \lambda I| = 0 \quad (3)$$

در فرمول (۳)، I ماتریس واحد، R ماتریس همبستگی بین متغیرهای اولیه و λ نیز مقادیر ویژه هستند.

از مقادیر ویژه، بردارهای ویژه به دست می‌آیند. در PCA مقادیر ویژه ماتریس تشابه بر اساس روند نزولی استخراج می‌شود، به‌طوری‌که محورهای متناظر PCA به طور متوالی مقادیر تغییرات را از بزرگ‌ترین تا کوچک‌ترین در ماتریس نشان می‌دهند. این تحلیل به‌عنوان یک تکنیک برای مواردی از جمله کاهش متغیرها، روندهای دوره‌ای داده‌ها، کیفیت آب و فاکتورهای مؤثر در آن و نیز استخراج عامل‌های مهم در آلودگی استفاده می‌شود.

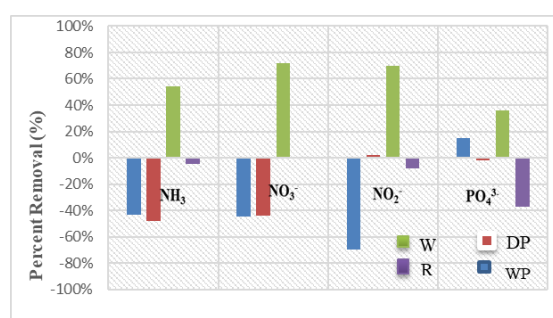
۲-۵. تجزیه و تحلیل خوشه‌ای سلسله مراتبی (HACA): تحلیل خوشه‌ای مجموعه‌ای از متغیرها را در داخل خوشه‌های یکسان جای می‌دهد. خوشه بندی مترکم سلسله مراتبی، رایجترین روش تحلیل خوشه‌ای است که ارتباطات همسان ذاتی را بین هرکدام از نمونه‌ها و نیز بین همه داده‌ها فراهم می‌کند و معمولاً با یک نمودار درختی نشان داده می‌شود (Nosrati et al. 2011). نمودار درختی، خلاصه‌ای از فرایند خوشه بندی، تصویر خوشه‌ها و مجاورت آنها را به همراه کاهش قابل توجه ابعاد داده‌های اولیه، ارائه می‌دهد.

غلظت جریان ورودی حوضچه مرطوب و غلظت جریان خروجی حوضچه تفریحی به ترتیب برای NO_3^- ، $1.89 \pm 1.18 \text{ mg/l}$ به $0.66 \pm 0.20 \text{ mg/l}$ ، برای NO_2^- ، $0.010 \pm 0.005 \text{ mg/l}$ به $0.004 \pm 0.002 \text{ mg/l}$ و برای NH_3 ، $0.35 \pm 0.26 \text{ mg/l}$ به $0.17 \pm 0.05 \text{ mg/l}$ کاهش یافته است.

در حوضچه مرطوب، غلظت جریان ورودی NH_3 و غلظت جریان خروجی اندازه گیری شد که وقتی با داده های زاکاریا در سال ۲۰۰۷ مقایسه شد به این نتیجه منجر شد که کمی کمتر از آن نتایج واقعی است درحالی که در حوضچه مرطوب تمرکز NH_3 نزدیک به نتایج Zakaria et. al (2007) بود.

در تالاب، غلظت جریان ورودی و خروجی NO_3^- مشاهده شده در مقایسه با نتایج Ismail et. al (2008) بالاتر هست درحالی که غلظت جریان ورودی و خروجی برای NO_2^- مشاهده شده، کمتر است. جدول ۱ میانگین غلظت جریان ورودی و خروجی PO_4^{3-} در سامانه زهکشی را نشان می دهد. سامانه، افزایش غلظت PO_4^{3-} از WP (0.19 ± 0.09) به RP (0.21 ± 0.13) را نشان می دهد. آزمون حذف درصد برای تعیین مقدار حذف آلاینده ها در سامانه زهکشی استفاده می شود.

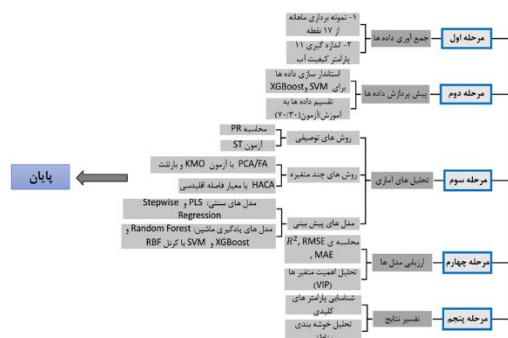
در شکل ۳ متوسط حذف درصد برای NH_3 ، NO_3^- ، NO_2^- و PO_4^{3-} در کل سامانه نشان داده شده است. در حوضچه مرطوب متوسط حذف درصد برای NH_3 ، NO_3^- و NO_2^- به ترتیب ۴۳-، ۴۵-، درصد و ۷۰- درصد است. در حوضچه تأخیر، متوسط حذف درصد برای NH_3 ، NO_3^- و NO_2^- به ترتیب ۴۸-، ۴۴-، درصد و ۲ درصد است. در تالاب، متوسط حذف درصد برای NH_3 ، NO_3^- و NO_2^- به ترتیب ۵۴ درصد، ۷۲ درصد و ۷۰ درصد است. در حوضچه تفریحی، متوسط حذف درصد برای NH_3 ، NO_3^- و NO_2^- به ترتیب ۵-، درصد، ۱ درصد و ۸- درصد است. متوسط حذف درصد برای PO_4^{3-} در DP، WL و RP به ترتیب ۱۵ درصد، ۲-، درصد، ۳۶ درصد و ۳۷- درصد است.



شکل ۳- متوسط حذف درصد NO_2^- ، NO_3^- ، NH_3 در Detention Pond، Wet Pond، Relation Pond، Wetland

جدول ۲، تأثیر حوضچه ها را بر روی پارامتر های NH_3 ، NO_3^- ، NO_2^- و PO_4^{3-} با استفاده از آزمون علامت نشان می دهد. به همینطور که مشاهده می شود در قسمت تالاب کاهش چشمگیر و قابل توجهی صورت گرفته است که تأثیر و عملکرد تالاب را در کل سیستم نشان می دهد.

همه متغیرها در تحلیل خوشه ای سلسله مراتبی برای طبقه بندی تالاب در مناطق مختلف با ویژگی های مشابه استفاده می شود. فلوجارت روش ها به شرح زیر می باشد:



نتایج و بحث

۳-۱. آزمون علامت

برای تعیین عامل های کیفی آب در کل سامانه زهکشی، از عامل های NO_2^- ، NO_3^- ، NH_3 و PO_4^{3-} استفاده شد. در این تحقیق عملکرد کیفی آب در کل سامانه زهکشی با استفاده از روش حذف درصد و آزمون علامت ارزیابی شد. غلظت آلاینده ها در ورودی و خروجی در قسمت های مختلف سامانه زهکشی برداشت شده است. مواد مغذی شامل نیترات (NO_3^-)، نیتريت (NO_2^-) و نیتروژن آمونیوم (NH_3) می باشند.

در جدول ۱ میانگین غلظت جریان ورودی و غلظت جریان خروجی برای هر ۴ عامل NO_3^- ، NO_2^- ، NH_3 و PO_4^{3-} در کل سامانه زهکشی ارائه شده است. در بعضی از قسمت های سامانه افزایش آلودگی مشاهده شده است؛ ولی به طور کلی در بیشتر قسمت های سامانه، کاهش آلودگی برای NO_3^- ، NO_2^- و NH_3 رخ داده است.

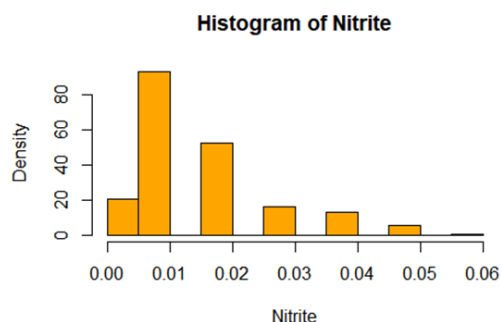
جدول ۱- اولویت دلایل تضاد آب کشاورزی

Recreation al Pond	Wetlan d	Detentio n Pond	Wet Pond		
Conc (mg/l)	Conc (mg/l)	Conc (mg/l)	Conc (mg/l)		
0.75±0.29	3.02±0.82	2.61±1.28	1.89±1.18	Inflo w	NO ₃ ⁻
0.66±0.20	0.75±0.29	3.02±0.82	2.38±1.18	Outfl ow	
0.004±0.002	0.016±0.004	0.013±0.006	0.010±0.005	Inflo w	NO ₂ ⁻
0.004±0.002	0.004±0.002	0.016±0.004	0.014±0.007	Outfl ow	
0.16±0.04	0.37±0.08	0.39±0.12	0.35±0.26	Infl ow	NH ₃
0.17±0.05	0.16±0.04	0.37±0.08	0.34±0.12	Outfl ow	
0.13±0.05	0.22±0.11	0.28±0.21	0.19±0.09	Inflo w	PO ₄ ³⁻
0.21±0.13	0.13±0.05	0.22±0.11	0.15±0.12	Outfl ow	

جدول ۲- مقادیر کیفیت آب در سیستم زهکشی با استفاده از آزمون علامت

Parameters	Exam Sign
NH ₃	WP 0.118
	DP 1.000
	WL 0.000
	RP 0.302
NO ₃ ⁻	WP 0.302
	DP 0.302
	WL 0.000
	RP 0.791
NO ₂ ⁻	WP 0.118
	DP 0.092
	WL 0.000
	RP 1.000
PO ₄ ³⁻	WP 0.424
	DP 0.424
	WL 0.000
	RP 0.092

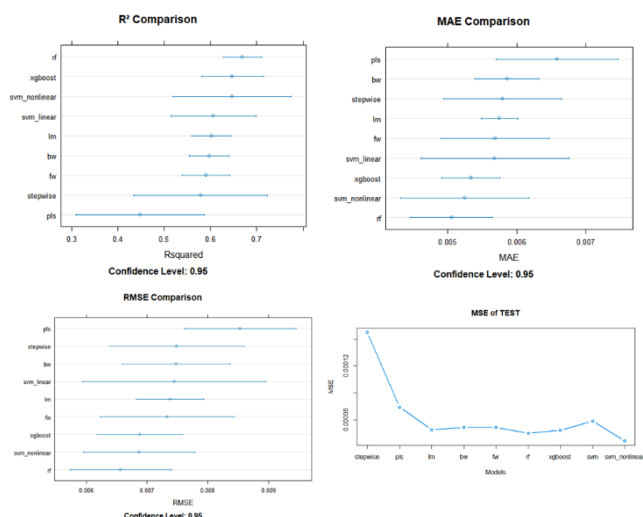
آن‌ها را بر روی داده‌های آزمایشی ارزیابی کنیم. بعد از این مرحله، با انتخاب پارامترهای مناسب و تنظیم مدل، می‌توان به پیش‌بینی‌های دقیق‌تر و بهینه‌سازی مدل‌ها دست یافت. همچنین، برای ارزیابی عملکرد مدل‌ها از معیارهایی مانند MAE، RMSE، و R^2 استفاده می‌شود تا دقت پیش‌بینی‌ها سنجیده شود. با توجه به شکل ۴ نمودار هیستوگرام توزیع متغیر نیتريت چوله به راست بوده و مقادیر آن متمایل به مقادیر پایین هستند.



شکل ۴- نمودار هیستوگرام متغیر نیتريت

برای اجرای مدل‌های آماری از نرم‌افزار R و کتابخانه یادگیری ماشین CARET استفاده شده است (۲۱). این کتابخانه ابزارهای متنوعی برای کنونی و بهینه‌سازی مدل‌های آماری و یادگیری ماشین فراهم می‌کند. خروجی‌های به‌دست آمده با استفاده از این ابزارها در ادامه تفسیر شده است.

خلاصه نتایج برازش مدل‌های مذکور را در نمودارهای شکل ۵ ارائه شده است.



شکل ۵- نمودارهای مقایسه ای ارزیابی مدل‌های برازش داده شده به داده های آموزشی

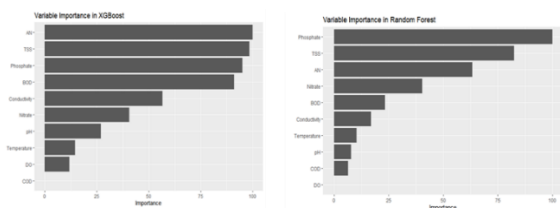
باتوجه به نتایج به‌دست آمده از روش‌های آماری حذف درصد و آزمون علامت، مشاهده شد که تالاب نقش اساسی و کلیدی در کل سامانه زهکشی ایفا می‌کند؛ بنابراین با استفاده از روش‌های آماری LDA، PCA/FA و HACA کلیه عامل‌های کیفی آب را در تالاب مورد بررسی قرار می‌گیرد.

۳-۲. تحلیل تشخیصی با مدل های یادگیری ماشین

تحلیل تشخیص روشی است که مؤثرترین متغیرها برای پیش‌بینی متغیر وابسته را مشخص می‌سازد. معمولاً در حالتی که کمیت پاسخ گسسته باشد، از رگرسیون لجستیک و روش‌های ممیزی خطی (LDA) استفاده می‌شود. اما اگر کمیت وابسته، پیوسته باشد، معمولاً از روش‌های رگرسیون خطی و مدل‌های یادگیری ماشین مانند رگرسیون خطی ساده و چندگانه، جنگل تصادفی، XGBoost، و ماشین بردار پشتیبان (SVM) استفاده می‌شود. این مدل‌ها توانایی پیش‌بینی مقادیر پیوسته را دارند و می‌توانند رابطه خطی یا غیرخطی بین متغیرهای مستقل و متغیر وابسته را شناسایی و مدل‌سازی کنند. انتخاب مدل مناسب به نوع داده‌ها، پیچیدگی روابط بین متغیرها، و دقت مورد نیاز در پیش‌بینی بستگی دارد. حال برای پیش‌بینی متغیر نیتريت که یک متغیر پیوسته است، استفاده از مدل‌های یادگیری ماشین که برای پیش‌بینی متغیرهای پیوسته مناسب هستند، توصیه می‌شود.

برای اجرای مدل‌های مذکور با توجه به اینکه مدل SVM به مقیاس ویژگی‌ها حساس است، زیرا فاصله‌ها و بردارهای پشتیبان در فضای ویژگی‌ها بر اساس مقادیر مطلق محاسبه می‌شوند، لذا ابتدا داده‌ها را استاندارد سازی کرده و سپس داده‌ها را به مجموعه‌های آموزشی و آزمایشی با نسبت ۷۰ به ۳۰ تقسیم می‌کنیم. این کار به ما کمک می‌کند تا مدل‌ها را بر روی داده‌های آموزشی آموزش داده و سپس عملکرد

10.45%	35	4.41%	15	pH
4.48%	15	2.94%	10	COD
7.46%	25	1.47%	5	DO



شکل ۶- نمودارهای مقایسه متغیرهای مهم برای مدل‌های جنگل تصادفی و XGBoost

اهمیت متغیرها در مدل جنگل تصادفی

برای محاسبه اهمیت متغیرها در مدل جنگل تصادفی، از کتابخانه VIP (Variable Importance Plot) نرم‌افزار R استفاده شده است (Greenwell et al., 2019). که ابزار قدرتمندی برای تحلیل و نمایش اهمیت متغیرها در مدل‌های یادگیری ماشین محسوب می‌شود. نتایج حاصل از این کتابخانه در جدول ۳ نشان می‌دهد که متغیر فسفات (Phosphate) با امتیاز ۱۰۰ و TSS با امتیاز ۸۵، به‌عنوان مهم‌ترین عوامل تأثیرگذار شناسایی شدند. این متغیرها به ترتیب ۲۹٫۴۱٪ و ۲۵٫۰۰٪ از اهمیت کل مدل را تشکیل می‌دهند. علاوه بر این، متغیرهای AN (۲۰٫۵۹٪)، نیترات (Nitrate) (۱۴٫۷۱٪) و BOD (۱۳٫۲۴٪) نیز تأثیرات قابل توجهی در دقت مدل داشته‌اند. این نتایج اهمیت این پارامترها را در مدیریت کیفیت آب برجسته می‌کند. برای جزئیات بیشتر در رابطه با نحوه محاسبه و پیاده‌سازی اهمیت متغیرها، به منابع کتابخانه VIP مراجعه کنید (نمودار شکل ۶ را ملاحظه کنید).

اهمیت متغیرها در مدل XGBoost

در مدل XGBoost نیز با استفاده از کتابخانه VIP، متغیرهای AN و TSS به‌عنوان مهم‌ترین عوامل تأثیرگذار شناخته شدند. با توجه به نتایج جدول ۳ این دو متغیر به ترتیب ۲۹٫۸۲٪ و ۲۵٫۳۵٪ از اهمیت کلی مدل را تشکیل می‌دهند. همچنین، متغیرهای فسفات (Phosphate) با ۲۳٫۲۸٪ و BOD (۱۷٫۹۱٪) و Conductivity (۱۴٫۹۳٪) نیز جایگاه مهمی در مدل دارند. نتایج مشابه با مدل جنگل تصادفی، نشانگر تأثیر برجسته متغیرهای AN، TSS و فسفات در هر دو مدل است که نقش کلیدی آن‌ها را در مدیریت کیفیت آب تأیید می‌کند. برای دسترسی به جزئیات بیشتر در رابطه با متدولوژی و پیاده‌سازی این تحلیل، به مستندات کتابخانه VIP مراجعه نمایید (نمودار شکل ۳ را ملاحظه کنید).

در نتیجه، با استفاده از کتابخانه VIP، متغیرهای فسفات، TSS و AN در هر دو مدل جنگل تصادفی و XGBoost به‌عنوان مهم‌ترین عوامل تأثیرگذار شناسایی شدند. این تطابق در نتایج هر دو مدل بر اهمیت متغیرها در پیش‌بینی و مدل‌سازی کیفیت آب تأکید دارد.

۳-۳. تحلیل عاملی / تحلیل مولفه‌های اصلی (PCA/FA)

بر اساس نتایج نمودارهای ارائه شده در شکل ۵، ارزیابی و مقایسه دقیق مدل‌های مختلف یادگیری ماشین و رگرسیون انجام شد. در این بررسی، مدل‌های SVM، FW، LM، Stepwise، BW، PLS، و دو نوع مختلف خطی و غیرخطی (Linear و Nonlinear)، XGBoost و جنگل تصادفی از نظر معیارهای میانگین خطای مطلق (MAE)، جذر میانگین مربعات خطا (RMSE)، و ضریب تعیین (R^2) مورد تحلیل قرار گرفتند.

مقایسه معیار MAE

با توجه به نتایج، مدل‌های جنگل تصادفی و XGBoost کمترین میزان MAE را نشان دادند که به ترتیب ۰٫۰۰۵۰۶ و ۰٫۰۰۵۳۴ بود. این مقادیر نشان‌دهنده دقت بالاتر این مدل‌ها در پیش‌بینی داده‌های تست است. در مقابل، مدل‌های PLS و Stepwise دارای MAE بالاتری بودند (۰٫۰۰۶۵۸ و ۰٫۰۰۵۸۰)، که بیانگر عملکرد ضعیف‌تر آن‌ها در پیش‌بینی داده‌ها است. این نتایج نشان می‌دهند که مدل‌های جنگل تصادفی و XGBoost قادر به ارائه پیش‌بینی‌های دقیق‌تری هستند و در کاربردهای مرتبط با کیفیت آب می‌توانند ترجیح داده شوند. مقایسه R^2 (ضریب تعیین)

مقایسه ضریب تعیین (R^2) مدل‌ها نیز نشان داد که مدل Random Forest با مقدار حدود ۰٫۶۷ بهترین برازش را به داده‌ها ارائه می‌دهد. پس از آن، مدل‌های XGBoost با مقدار ۰٫۶۴ و SVM و Nonlinear با ۰٫۶۵ نیز عملکرد بسیار خوبی داشتند. در مقابل، مدل‌های PLS و Stepwise با R^2 ‌های پایین‌تر (۰٫۴۵ و ۰٫۵۷) ضعیف‌ترین عملکرد را در برازش داده‌ها نشان دادند. این اختلاف در مقدار R^2 نشان‌دهنده قدرت بالاتر مدل‌های مبتنی بر جنگل تصادفی و XGBoost در پیش‌بینی دقیق‌تر و برازش بهتر داده‌ها است.

ارزیابی RMSE و MSE

در بررسی خطای میانگین مربعات خطا (MSE) و جذر میانگین مربعات خطا (RMSE)، دوباره مدل‌های جنگل تصادفی و XGBoost به‌عنوان دقیق‌ترین مدل‌ها ظاهر شدند. مقدار RMSE مدل جنگل تصادفی برابر با ۰٫۰۰۶۵۷ و ۰٫۰۰۶۸۸ بود که به‌طور واضح نسبت به مدل‌های سنتی رگرسیون بهتر عمل کردند. در مقابل، مدل‌های PLS و Stepwise مقادیر بالاتری داشتند که نشان‌دهنده خطای بیشتر و دقت کمتر آن‌ها است.

جدول ۳- درصد اهمیت متغیرها برای هر دو مدل جنگل تصادفی و XGBoost

متغیر	اهمیت در جنگل تصادفی	درصد اهمیت در جنگل تصادفی	اهمیت در XGBoost	درصد اهمیت در XGBoost
Phosphate	100	29.41%	78	23.28%
TSS	85	25.00%	85	25.35%
AN	70	20.59%	100	29.82%
Nitrate	50	14.71%	45	13.43%
BOD	45	13.24%	60	17.91%
Conductivity	35	10.29%	50	14.93%
Temperature	25	7.35%	30	8.96%

اگر عامل‌های به‌دست‌آمده را با روش واریماکس چرخش دهید و باتوجه‌به نتایج مندرج در جدول ۷، مؤلفه‌های اول و دوم و سوم بر اساس معادلات زیر مشخص می‌شوند:

$$Z_1: \text{مؤلفه اول} \\ = 0.852X_1 + 0.847X_2 + 0.835X_3 + 0.830X_4 + 0.809X_5 + 0.82X_6 + 0.742X_7 + 0.321X_8 + 0.215X_9 + 0.19X_{10} + 0.04X_{11} \quad (4)$$

$$Z_2: \text{مؤلفه دوم} \\ = 0.134X_1 + 0.91X_2 + 0.93X_3 + 0.242X_4 + 0.91X_5 + 0.31X_6 + 0.191X_7 + 0.10X_8 + 0.843X_9 + 0.539X_{10} + 0.870X_{11} \quad (5)$$

$$Z_3: \text{مؤلفه سوم} \\ = 0.218X_1 + 0.141X_2 + 0.186X_3 - 0.17X_4 + 0.241X_5 - 0.70X_6 + 0.263X_7 + 0.829X_8 - 0.67X_9 + 0.616X_{10} - 0.187X_{11} \quad (6)$$

جدول ۷- بار عاملی برای هرکدام از مؤلفه‌های اصلی با

دوران واریماکس

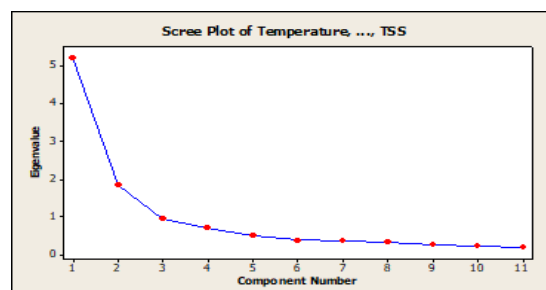
متغیر	مؤلفه اول	مؤلفه دوم	مؤلفه سوم
TSS	۰.۸۵۲	۰.۱۳۴	۰.۲۱۸
Nitrate	۰.۸۴۷	۰.۰۹۱	۰.۱۴۱
Phosphate	۰.۸۳۵	۰.۰۹۳	۰.۱۸۶
Nitrite	۰.۸۳۰	۰.۲۴۲	-۰.۰۱۷
AN	۰.۸۰۹	۰.۰۹۱	۰.۲۴۱
BOD	۰.۸۰۲	۰.۰۳۱	-۰.۰۷۰
COD	۰.۷۴۲	-۰.۱۹۱	۰.۲۶۳
Temperature	۰.۳۲۱	۰.۰۱۰	۰.۸۲۹
Conductivity	۰.۲۱۵	۰.۸۴۳	-۰.۰۶۷
DO	۰.۰۱۹	۰.۵۳۹	۰.۶۱۶
pH	-۰.۰۰۴	۰.۸۷۰	۰.۱۸۷
مقدار ویژه	۵,۲۷۲	۱,۸۴۵	۰,۹۵۰
درصد واریانس	۴۷,۹۲۸	۱۶,۷۷۰	۸,۶۳۸
درصد واریانس تجمعی	۴۷,۹۲۸	۶۴,۶۹۹	۷۳,۳۳۶

همچنین در شکل ۸ (الف) نمودار متغیرهای کیفی استراخ شده در تالاب با استفاده از روش PCA نمایش داده است. همانطور که مشاهده می‌شود بیشتر همبستگی بین نیترات، نیتريت، فسفات، TSS و AN وجود دارد. نتایج سه بعدی آنالیز آماری PCA در شکل ۸ (ب) ارائه شده است.

از بارگذاری عاملی تخمین زد. بارهای عاملی قوی (بزرگ‌تر از ۰/۶۰) در نظر گرفته می‌شود. در تالاب نیز PCA، ۱۱ مؤلفه را معرفی می‌کند که سه مؤلفه اول آن مؤلفه‌های اصلی بوده و در مجموع ۷۳,۳۳۶ درصد واریانس تجمعی را بیان می‌کند. مؤلفه اول دارای بیشترین مقدار ویژه با مقدار ۵,۲۷۲ و درصد واریانس ۴۷,۹۲۸ را دارد که عامل‌های TSS، نیترات، نیتريت، فسفات، AN، BOD و COD بیشترین همبستگی را با این مؤلفه داشتند. در میان عامل‌های موجود در مؤلفه اول، TSS دارای بیشترین بار عاملی است برابر (0.852) و به‌عنوان عامل اصلی این مؤلفه شناخته شده است. مقدار ویژه مؤلفه دوم نیز ۱,۸۴۵ درصد واریانس ۱۶,۷۷۰ درصد است که کدورت و pH دارای بیشترین همبستگی با این مؤلفه را دارند. مقدار ویژه مؤلفه سوم نیز ۰,۹۵۰ درصد واریانس ۸,۶۳۸ درصد است که متغیرهای دما و DO بیشترین همبستگی را با این مؤلفه دارند؛ بنابراین سه مؤلفه اصلی در توصیف تغییرات عامل‌های کیفیت آب در نظر گرفته می‌شود.

جدول ۶- نتایج واریانس و مقادیر ویژه هر یک از مؤلفه‌ها در آنالیز PCA

مؤلفه	مقدار ویژه	درصد واریانس	درصد واریانس تجمعی
۱	۵,۲۷۲	۴۷,۹۲۸	۴۷,۹۲۸
۲	۱,۸۴۵	۱۶,۷۷۰	۶۴,۶۹۹
۳	۰,۹۵۰	۸,۶۳۸	۷۳,۳۳۶
۴	۰,۶۹۸	۶,۳۴۴	۷۹,۶۸۱
۵	۰,۵۱۲	۴,۶۵۲	۸۴,۳۳۳
۶	۰,۳۸۶	۳,۵۰۸	۸۷,۸۴۱
۷	۰,۳۶۹	۳,۳۵۳	۹۱,۱۹۴
۸	۰,۳۰۳	۲,۷۵۲	۹۳,۹۴۶
۹	۰,۲۶۵	۲,۴۰۸	۹۶,۳۵۴
۱۰	۰,۲۱۰	۱,۹۱۰	۹۸,۲۶۴
۱۱	۰,۱۹۱	۱,۷۳۶	۱۰۰,۰۰۰



شکل ۷- نمودار اسکری گراف برای تعیین مقادیر ویژه عامل‌ها

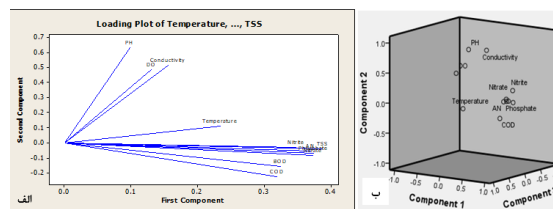
با استفاده از تحلیل PCA اثر هر کدام از این خوشه‌ها بر روی عامل‌های موجود در تالاب و کیفیت آب مورد بررسی قرار گرفت ضمناً اینکه در هر خوشه کدام متغیرها نقش بسزایی در کیفیت آب ایفا می‌کنند نیز مورد تحلیل قرار گرفت. باتوجه به مقادیر جدول ۸ که اطلاعات آمار توصیفی داده‌های کیفی آب را نشان می‌دهد، ضریب تغییرات عامل بدون واحدی است که خطای متغیرها را به صورت درصد نشان می‌دهد. در Cluster-1 کمترین خطا (پراکندگی حول میانگین) را متغیر Temperature با میزان ۴۶ درصد دارد و بیشترین خطا را متغیر Nitrite با میزان ۳۰ درصد دارد. در Cluster-2 کمترین خطا را متغیر Temperature با میزان ۴۶٫۷ درصد دارد و بیشترین خطا را متغیر Nitrite با میزان ۴۳٫۴۲ درصد دارد. در Cluster-3 کمترین خطا را متغیر Temperature با میزان ۴۳٫۳۴ درصد دارد و بیشترین خطا را متغیر Nitrite با میزان ۴۹٫۳۵ درصد دارد.

جدول ۸- خصوصیات آماری نتایج آنالیز نمونه‌ها

عامل	حداکثر	میانگین	انحراف	ضریب	عامل	حداکثر	میانگین	انحراف	ضریب
تغییرات	معیار	تغییرات	معیار	تغییرات	تغییرات	معیار	تغییرات	معیار	تغییرات
Temperature	۲۸٫۵۰۰	۲۵٫۱۴۰	۲۲٫۰۱۵	۱٫۶۹۲	Phosphate	۴٫۶۶۵	۱٫۶۹۲	۰٫۰۶۶	۱۶٫۲۶۵
pH	۶٫۲۴۰	۶٫۱۰۰	۶٫۶۷۱	۰٫۲۲۴	AN	۹٫۳۲۲	۰٫۲۲۴	۰٫۰۶۱	۱۹٫۵۲۲
DO	۶٫۶۶۰	۱۰٫۲۶۰	۸٫۳۵۵	۰٫۵۹۹	BOD	۱۱٫۲۲۰	۰٫۵۹۹	۰٫۰۸۲	۱۶٫۴۲۵
Conductivity	۹۸٫۰۰۰	۲۰۶٫۰۰۰	۱۳۹٫۲۸۰	۲۷٫۳۹۰	COD	۱۹٫۹۰۰	۲۷٫۳۹۰	۳۹٫۷۷۰	۱۵۸٫۸۸۲
Nitrite	۰٫۰۱۵	۰٫۰۵۵	۰٫۰۲۳	۰٫۰۱۰	TSS	۳۰۰۰۰	۰٫۰۱۰	۴۸۰۰۰	۱۷۰۰۰
Nitrate	۱٫۶۰۰	۴٫۹۰۰	۳٫۱۸۲	۰٫۶۸۰		۲۱٫۶۶۲	۰٫۶۸۰	۳۸٫۱۳۵	۲۱۰۰۰
Temperature	۲۸٫۵۰۰	۲۵٫۱۴۰	۲۲٫۰۱۵	۱٫۶۹۲	Phosphate	۴٫۶۶۵	۱٫۶۹۲	۰٫۰۶۶	۱۶٫۲۶۵
pH	۶٫۲۴۰	۶٫۱۰۰	۶٫۶۷۱	۰٫۲۲۴	AN	۹٫۳۲۲	۰٫۲۲۴	۰٫۰۶۱	۱۹٫۵۲۲
DO	۶٫۶۶۰	۱۰٫۲۶۰	۸٫۳۵۵	۰٫۵۹۹	BOD	۱۱٫۲۲۰	۰٫۵۹۹	۰٫۰۸۲	۱۶٫۴۲۵
Conductivity	۹۸٫۰۰۰	۲۰۶٫۰۰۰	۱۳۹٫۲۸۰	۲۷٫۳۹۰	COD	۱۹٫۹۰۰	۲۷٫۳۹۰	۳۹٫۷۷۰	۱۵۸٫۸۸۲
Nitrite	۰٫۰۱۵	۰٫۰۵۵	۰٫۰۲۳	۰٫۰۱۰	TSS	۳۰۰۰۰	۰٫۰۱۰	۴۸۰۰۰	۱۷۰۰۰
Nitrate	۱٫۶۰۰	۴٫۹۰۰	۳٫۱۸۲	۰٫۶۸۰		۲۱٫۶۶۲	۰٫۶۸۰	۳۸٫۱۳۵	۲۱۰۰۰
Temperature	۲۸٫۵۰۰	۲۵٫۱۴۰	۲۲٫۰۱۵	۱٫۶۹۲	Phosphate	۴٫۶۶۵	۱٫۶۹۲	۰٫۰۶۶	۱۶٫۲۶۵
pH	۶٫۲۴۰	۶٫۱۰۰	۶٫۶۷۱	۰٫۲۲۴	AN	۹٫۳۲۲	۰٫۲۲۴	۰٫۰۶۱	۱۹٫۵۲۲
DO	۶٫۶۶۰	۱۰٫۲۶۰	۸٫۳۵۵	۰٫۵۹۹	BOD	۱۱٫۲۲۰	۰٫۵۹۹	۰٫۰۸۲	۱۶٫۴۲۵
Conductivity	۹۸٫۰۰۰	۲۰۶٫۰۰۰	۱۳۹٫۲۸۰	۲۷٫۳۹۰	COD	۱۹٫۹۰۰	۲۷٫۳۹۰	۳۹٫۷۷۰	۱۵۸٫۸۸۲
Nitrite	۰٫۰۱۵	۰٫۰۵۵	۰٫۰۲۳	۰٫۰۱۰	TSS	۳۰۰۰۰	۰٫۰۱۰	۴۸۰۰۰	۱۷۰۰۰
Nitrate	۱٫۶۰۰	۴٫۹۰۰	۳٫۱۸۲	۰٫۶۸۰		۲۱٫۶۶۲	۰٫۶۸۰	۳۸٫۱۳۵	۲۱۰۰۰

ضرایب KMO محاسبه شده در Cluster-1، ۰٫۵۶۷، و Cluster-2، ۰٫۶۴۶ و در Cluster-3، ۰٫۷۱۹ هستند. باتوجه به اینکه احتمال معناداری برای این آزمون کمتر از ۰٫۰۵ است، پس نتایج آزمون بارتلت در هر سه Cluster نیز معنی دار است، به این مفهوم که فرض مخالف تایید می‌شود، پس داده‌های مورد نظر برای انجام تحلیل عاملی مناسب اند.

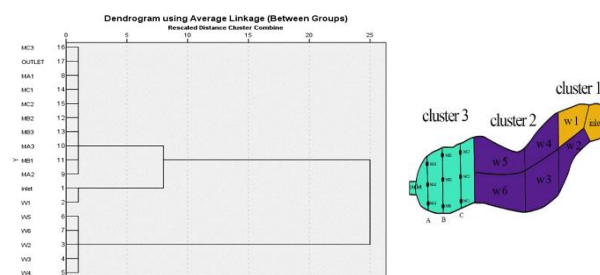
نتایج به‌دست‌آمده از تحلیل مؤلفه‌های اصلی نشان می‌دهد که مؤلفه با داشتن مقدار ویژه بیش از یک عنوان، مهم‌ترین مؤلفه‌های توجیه‌کننده واریانس در نظر گرفته شده‌اند. بر اساس نتایج جدول ۹، در Cluster-1 چهار عامل مهم با مقادیر ویژه تقریباً بیشتر از یک انتخاب شد که ۶۹٫۶۶۹ درصد از واریانس داده‌ها را توصیف کرد. مؤلفه اول دارای بیشترین مقدار ویژه با مقدار ۲٫۶۵۵ و درصد واریانس ۲۴٫۱۴ را دارد. همچنین مقدار ویژه مؤلفه دوم نیز ۱٫۹۵۴ با درصد واریانس ۱۷٫۷۶ درصد، مقدار ویژه سوم نیز ۱٫۶۹۸ با درصد واریانس



شکل ۸- الف) متغیرهای کیفی استخراج شده در تالاب با استفاده از روش PCA، ب) نتایج آنالیز PCA به صورت سه بعدی

۳-۴ تحلیل خوشه‌ای سلسله‌مراتبی (HACA)

نتایج حاصل از تحلیل خوشه‌ای در تالاب در شکل ۶ (الف) نشان داده شده است. با انجام تحلیل خوشه‌ای، تالاب در سه خوشه همگن قرار گرفت. فاصله بین هر یک از قسمت‌های تالاب با توجه از تحلیل خوشه‌ای، همبستگی و خود همبستگی بین پارامترهای کیفی آب را نشان می‌دهد. همه متغیرها در تحلیل خوشه‌ای سلسله‌مراتبی برای طبقه‌بندی تالاب در مناطق مختلف با ویژگی‌های مشابه استفاده شد. شکل ۶ (الف) نشان می‌دهد که تحلیل خوشه‌ای سلسله‌مراتبی همه ۱۷ منطقه را به صورت معنی داری طبقه‌بندی کرده است. ایستگاه‌های نمونه برداری به سه خوشه آماری معنی دار، اولین خوشه (Cluster-I)، در قسمت بالای تالاب، از دو قسمت تالاب، ورودی و W1 تشکیل شده است. Cluster-II در قسمت وسط تالاب واقع شده است و شامل پنج قسمت از تالاب بود: W2، W3، W4، W5 و W6 و قسمت باقیمانده تالاب، میکروپول و خروجی، به‌عنوان Cluster-III طبقه‌بندی شده است. خروجی از هر خوشه به‌عنوان ورودی بعدی انتخاب شد. خوشه بعدی، ورودی و خروجی تالاب به‌عنوان ورودی و خروجی برای Cluster-I و Cluster-III به ترتیب انتخاب شدند. دندروگرام‌ها یا نمودارهای درختی برای نشان دادن روابط تولید شده بوسیله روش خوشه‌بندی سلسله‌مراتبی ایجاد شده‌اند. با توجه به شکل ۹ (الف) می‌توان فهمید که W2 تا W6 از ضریب همبستگی مثبت بالایی برخوردارند و از این رو در یک خوشه قرار می‌گیرند. همچنین قسمت ورودی و W1 از منظر ضریب همبستگی مشابهت زیادی دارند و در یک خوشه قرار می‌گیرند، بنابراین قسمت‌های دیگر تالاب که عمدتاً میکروپول هستند و قسمت خروجی، دارای ضریب همبستگی بیشتری است و در یک خوشه قرار می‌گیرند. موقعیت سه خوشه در تالاب در شکل ۶ (ب) نشان داده شده است.



شکل ۹- الف) نمودار دندروگرام خوشه‌بندی تالاب، ب) موقعیت سه خوشه

جدول ۱۰- بار عاملی برای هر کدام از مؤلفه های اصلی با دوران واریماکس

DO	Conductivity	Temperature	COD	BOD	AN	Nitrite	Phosphate	Nitrate	TSS	
0.381	0.507	0.395	0.369	0.0028	0.686	0.110	0.614	0.988	0.123	مؤلفه اول
0.009	0.574	0.398	0.0057	0.588	0.399	0.00	0.352	0.390	0.200	مؤلفه دوم
0.004	0.591	0.307	0.899	0.003	0.310	0.378	0.118	0.00	0.507	مؤلفه سوم
0.006	0.596	0.600	0.002	0.00	0.350	0.149	0.811	0.004	0.601	مؤلفه چهارم
0.106	0.590	0.826	0.396	0.608	0.606	0.107	0.004	0.008	0.000	مؤلفه اول
0.068	0.0067	0.009	0.397	0.0006	0.300	0.007	0.017	0.014	0.008	مؤلفه دوم
0.000	0.377	0.379	0.686	0.387	0.002	0.107	0.366	0.008	0.399	مؤلفه سوم
0.089	0.388	0.391	0.007	0.002	0.191	0.154	0.002	0.007	0.896	مؤلفه چهارم
0.007	0.388	0.112	0.009	0.377	0.001	0.350	0.819	0.622	0.589	مؤلفه اول
0.009	0.007	0.362	0.812	0.006	0.960	0.966	0.161	0.000	0.008	مؤلفه دوم
0.819	0.310	0.631	0.326	0.600	0.108	0.002	0.190	0.008	0.362	مؤلفه سوم

در این تحقیق، از روش‌های آماری چندمتغیره حذف درصد، آزمون علامت، HACA و PCA/FA برای تحلیل شاخص کیفیت آب (WQI) و تشخیص عوامل تأثیرگذار بر آن استفاده شد. در این پژوهش، ۱۱ پارامتر کیفیت آب در سامانه زهکشی سنجیده شد. نتایج نشان داد که در برخی قسمت‌های سیستم، میزان آلاینده‌ها افزایش داشته، اما به‌طور کلی کاهش آلودگی در پارامترهایی مانند-NO₃، NO₂-و NH₃ مشاهده شد و تالاب نقش اساسی در کنترل این آلودگی‌ها ایفا کرد.

۱۵,۳۳۷ درصد و مقدار ویژه چهارم نیز ۱,۳۵۷ با درصد واریانس ۱۲,۳۳۷ درصد است. در Cluster-2 چهار عامل مهم با مقادیر ویژه تقریباً بیشتر از یک انتخاب شد که ۶۶,۶۰۲ درصد از واریانس داده ها را توصیف میکند. مولفه اول دارای بیشترین مقدار ویژه با مقدار ۲,۷۸۴ و درصد واریانس ۲۵,۳۰۶ را دارد. همچنین مقدار ویژه مولفه دوم نیز ۲,۰۹۵ با درصد واریانس ۱۹,۰۴۷ درصد است. مقدار ویژه مولفه سوم نیز ۱,۳۶۱ با درصد واریانس ۱۲,۳۷ درصد است. همچنین در Cluster-3 سه عامل مهم با مقادیر ویژه تقریباً بیشتر از ۱,۹۵ انتخاب شد که ۵۸,۹۲۷ درصد از واریانس داده ها را توصیف کرد. مولفه اول دارای بیشترین مقدار ویژه با مقدار ۳,۷۵۱ و درصد واریانس ۳۴,۰۹۷ را دارد. همچنین مقدار ویژه مولفه دوم نیز ۱,۴۵۹ با درصد واریانس ۱۳,۲۶۳ درصد است. مقدار ویژه مولفه سوم نیز ۱,۲۷۲ با درصد واریانس ۱۱,۵۶۷ درصد است. جدول ۱۰ بارهای عاملی مربوط به هر پارامتر در هر سه Cluster به طور جداگانه بیان شده است. همبستگی بین مولفه ها و متغیرها را می توان با استفاده از بارگذاری عاملی تخمین زد. برای هر سه Cluster بارهای عاملی قوی (بزرگتر از ۰/۶۰) در نظر گرفته شد. در Cluster-1 چهار مولفه اصلی، Cluster-2 چهار مولفه اصلی و در Cluster-3 سه مولفه اصلی شناخته شد که بیشترین بارهای عاملی در مولفه اول مربوط به خوشه اول نیتريت، نیترا، AN، فسفات با مقادیر بترتیب ۰/۷۹۸، ۰/۸۱۵، ۰/۸۱۶، ۰/۶۱۴، درخوشه دوم دما، BOD، AN با مقادیر بترتیب ۰/۸۲۶، ۰/۶۵۸، ۰/۶۵۶ و در خوشه سوم فسفات، TSS، نیترا با مقادیر بترتیب ۰/۸۱۶، ۰/۷۵۹، ۰/۶۳۳ مشخص شده اند. به طور کلی مهمترین پارامترهای کیفیت آب که توسط PCA بر روی هر سه Cluster تعیین شدند در خوشه اول پارامترهای نیترا، فسفات، نیتريت و AN با درصد واریانس کل (۶۶/۶۶۹)، در خوشه دوم AN، BOD و دما با درصد واریانس کل (۶۶/۶۰۲) و در خوشه سوم پارامترهای TSS، نیترا و فسفات با درصد واریانس کل (۶۷/۸۶۵) به عنوان موثرترین پارامترها انتخاب شده اند. براساس نتایج بدست آمده می توان نتیجه گرفت که نیترا، فسفات و AN نقش کلیدی و مهمترین نسبت به متغیرهای دیگر دارند. شکل ۱۰ تغییرات مقادیر ویژه به ازای افزایش تعداد مولفه ها را در هر سه Cluster نشان می دهد. این نمودار برای تعیین تعداد بهینه مولفه ها به کار می رود.

هله	مقدار	هیزه	درصد	درصد واریاس	مؤله	مقدار	درصد	درصد واریاس	تجمعی
تجمعی	واریاس	درصد	درصد واریاس	تجمعی	تجمعی	واریاس	درصد	درصد واریاس	تجمعی
۱	۲,۶۵۵	۲۴,۱۰۰	۲۴,۱۰۰	۱۰۰	۵	۸,۴۴۴	۷۸,۰۱۳	۹	۳,۵۲۶
۲	۱,۹۵۴	۱۷,۳۶۰	۴۱,۹۰۰	۶	۶,۰۸۳	۶۶,۹۲۶	۱۰	۲,۵۲۳	۹۸,۴۲۳
۳	۱,۶۸۸	۱۵,۳۳۳	۵۷,۳۳۳	۷	۴,۳۲۳	۶۲,۶۰۳	۱۱	۱,۵۸۷	۱۰۰,۰۰۰
۴	۱,۳۵۷	۱۳,۳۳۷	۷۵,۹۴۰	۸	۳,۵۵۵	۶۶,۱۵۸	۱۲	۱,۲۶۶	۱۰۰,۰۰۰
۵	۲,۷۸۴	۲۵,۳۰۶	۱۰۱,۶۶۲	۹	۸,۴۶۶	۷۵,۵۲۴	۱۳	۹۰۰	۱۰۰,۰۰۰
۶	۲,۰۵۵	۱۹,۰۴۷	۱۲۰,۷۰۹	۱۰	۵,۶۳۲	۸۱,۱۵۶	۱۴	۶۰۰	۱۰۰,۰۰۰
۷	۱,۳۶۱	۱۳,۳۷۰	۱۳۴,۰۷۹	۱۱	۴,۸۵۷	۸۶,۰۱۳	۱۵	۳۰۰	۱۰۰,۰۰۰
۸	۱,۰۸۷	۹,۸۷۸	۱۴۳,۹۵۷	۱۲	۴,۱۷۲	۹۰,۱۸۵	۱۶	۲۰۰	۱۰۰,۰۰۰
۹	۲,۷۵۱	۲۴,۰۰۷	۱۶۸,۰۰۰	۱۳	۷,۸۲۳	۹۸,۰۰۰	۱۷	۱۰۰	۱۰۰,۰۰۰
۱۰	۱,۵۵۹	۱۳,۳۶۳	۱۸۱,۳۶۳	۱۴	۶,۶۳۳	۱۰۴,۶۳۳	۱۸	۵۰	۱۰۰,۰۰۰
۱۱	۱,۳۷۲	۱۱,۶۲۷	۱۹۲,۹۶۰	۱۵	۵,۰۵۳	۱۱۰,۶۸۳	۱۹	۲۵	۱۰۰,۰۰۰
۱۲	۹۰۰	۸,۴۸۸	۲۰۱,۴۵۰	۱۶	۴,۱۰۰	۱۱۴,۷۸۳	۲۰	۱۰	۱۰۰,۰۰۰
۱۳	۸۰۰	۷,۶۸۸	۲۰۹,۱۳۸	۱۷	۳,۱۰۰	۱۱۷,۸۸۳	۲۱	۵	۱۰۰,۰۰۰
۱۴	۷۰۰	۶,۹۸۸	۲۱۶,۱۲۳	۱۸	۲,۱۰۰	۱۱۹,۹۸۳	۲۲	۵	۱۰۰,۰۰۰
۱۵	۶۰۰	۶,۲۸۸	۲۲۲,۴۱۱	۱۹	۱,۱۰۰	۱۲۱,۰۸۳	۲۳	۵	۱۰۰,۰۰۰
۱۶	۵۰۰	۵,۵۸۸	۲۲۷,۹۹۹	۲۰	۱,۰۰۰	۱۲۲,۰۸۳	۲۴	۵	۱۰۰,۰۰۰
۱۷	۴۰۰	۴,۸۸۸	۲۳۲,۸۸۳	۲۱	۱,۰۰۰	۱۲۳,۰۸۳	۲۵	۵	۱۰۰,۰۰۰
۱۸	۳۰۰	۴,۱۸۸	۲۳۷,۰۷۱	۲۲	۱,۰۰۰	۱۲۴,۰۸۳	۲۶	۵	۱۰۰,۰۰۰
۱۹	۲۰۰	۳,۴۸۸	۲۴۰,۵۵۹	۲۳	۱,۰۰۰	۱۲۵,۰۸۳	۲۷	۵	۱۰۰,۰۰۰
۲۰	۱۰۰	۲,۷۸۸	۲۴۳,۳۴۷	۲۴	۱,۰۰۰	۱۲۶,۰۸۳	۲۸	۵	۱۰۰,۰۰۰
۲۱	۱۰۰	۲,۰۸۸	۲۴۵,۴۳۵	۲۵	۱,۰۰۰	۱۲۷,۰۸۳	۲۹	۵	۱۰۰,۰۰۰
۲۲	۱۰۰	۱,۳۸۸	۲۴۶,۸۲۳	۲۶	۱,۰۰۰	۱۲۸,۰۸۳	۳۰	۵	۱۰۰,۰۰۰
۲۳	۱۰۰	۶,۲۸۸	۲۵۳,۱۱۱	۲۷	۱,۰۰۰	۱۲۹,۰۸۳	۳۱	۵	۱۰۰,۰۰۰
۲۴	۱۰۰	۵,۵۸۸	۲۵۸,۷۰۰	۲۸	۱,۰۰۰	۱۳۰,۰۸۳	۳۲	۵	۱۰۰,۰۰۰
۲۵	۱۰۰	۴,۸۸۸	۲۶۳,۵۸۸	۲۹	۱,۰۰۰	۱۳۱,۰۸۳	۳۳	۵	۱۰۰,۰۰۰
۲۶	۱۰۰	۴,۱۸۸	۲۶۷,۷۷۶	۳۰	۱,۰۰۰	۱۳۲,۰۸۳	۳۴	۵	۱۰۰,۰۰۰
۲۷	۱۰۰	۳,۴۸۸	۲۷۱,۲۶۴	۳۱	۱,۰۰۰	۱۳۳,۰۸۳	۳۵	۵	۱۰۰,۰۰۰
۲۸	۱۰۰	۲,۷۸۸	۲۷۴,۰۵۲	۳۲	۱,۰۰۰	۱۳۴,۰۸۳	۳۶	۵	۱۰۰,۰۰۰
۲۹	۱۰۰	۲,۰۸۸	۲۷۶,۱۴۰	۳۳	۱,۰۰۰	۱۳۵,۰۸۳	۳۷	۵	۱۰۰,۰۰۰

به‌طورکلی، این روش‌های آماری چندمتغیره کمک کردند که حجم گسترده‌ای از داده‌های کیفی آب به‌صورت هدفمند تحلیل شود. این رویکرد با کاهش هزینه‌ها و بهینه‌سازی پایش، امکان شناسایی دقیق‌تر نقاط حساس و بهبود مدیریت منابع آبی را فراهم می‌آورد.

یکی از جنبه‌های مهم در تفسیر نتایج، توجه به منابع احتمالی خطاست که ممکن است بر کیفیت داده‌ها و خروجی مدل‌ها تأثیرگذار باشند. خطاهای اندازه‌گیری ناشی از تنوع در روش‌های آزمایشگاهی یا خطاهای انسانی در اندازه‌گیری پارامترهایی چون TSS، نیترات یا نیتريت، از مهم‌ترین عوامل هستند. نوسانات محیطی و فصلی (مانند تغییرات دما یا شدت بارندگی) نیز ممکن است باعث تغییرات موقتی در کیفیت آب شده باشند که در نمونه‌برداری ماهانه ثبت کامل نشده است. علاوه بر آن، برخی فرضیات مدل‌ها، نظیر فرض خطی بودن در PCA یا خطر بیش‌برازش در مدل‌های یادگیری ماشین، ممکن است محدودیت‌هایی ایجاد کرده باشند. همچنین، پوشش مکانی محدود (۱۷ نقطه نمونه‌برداری) کل سامانه را به‌طور کامل نمایندگی نمی‌کند. با این حال، تلاش شده تا با استفاده از پروتکل‌های دقیق نمونه‌برداری، نرمال‌سازی داده‌ها، و اعتبارسنجی متقابل، این منابع خطا به حداقل برسند.

تعارض منافع

بنابر اظهار نویسندگان، مقاله حاضر فاقد هرگونه تعارض منافع بوده است.

مقدار نیتريت دارند و می‌توانند در پیش‌بینی و مدیریت کیفیت آب به‌عنوان عوامل مهم در نظر گرفته شوند.

برای بررسی تأثیر متغیرهای کلیدی بر عملکرد مدل‌ها، تحلیل حساسیت برای دو مدل Random Forest و XGBoost انجام شد. در این تحلیل، حذف تدریجی متغیرها و بررسی تغییرات MAE، RMSE و R^2 نشان داد که متغیرهای فسفات، TSS و آمونیاک بیشترین تأثیر را بر خروجی مدل دارند. این نتایج با تحلیل VIP کاملاً سازگار هستند و اهمیت متغیرهای مذکور در پیش‌بینی نیتريت را تأیید می‌کنند.

در ادامه، تحلیل خوشه‌ای (HACA) برای تفکیک بخش‌های مختلف سامانه به سه خوشه انجام شد. خوشه اول شامل قسمت‌های بالایی تالاب، خوشه دوم در قسمت‌های میانی و خوشه سوم در انتهای‌ترین قسمت قرار داشت. هر خوشه، ویژگی‌ها و خطاهای متفاوتی در پارامترهای کیفیت آب داشت، به طوری که کمترین میزان پراکندگی در پارامتر Temperature و بیشترین پراکندگی در Nitrite مشاهده شد.

تحلیل مؤلفه‌های اصلی (PCA) نیز برای شناسایی مهم‌ترین پارامترهای هر خوشه به کار رفت. در خوشه اول، چهار مؤلفه اصلی ۶۹،۷ درصد واریانس داده‌ها را توضیح دادند؛ در خوشه دوم، چهار مؤلفه اصلی ۶۶،۶ درصد واریانس را و در خوشه سوم، سه مؤلفه اصلی ۵۸،۹ درصد واریانس داده‌ها را پوشش دادند. این تحلیل‌ها نشان دادند که پارامترهایی چون نیترات، فسفات و AN در کیفیت آب تأثیر بالایی دارند.

References

1. و نصیر شاهنظری، علی میرهاشمی، مریم. احمدی، کامران. ۱۴۰۲. ارزیابی آلودگی نیترات و WASP و فسفات تالاب میانکاله با استفاده از مدل تحقیقات منابع آب ایران، سال نوزدهم، شماره ۴.
2. Aydin, H., Ustaoglu, F., Tepe, Y., & Soyulu, E. N. (2021). Assessment of water quality of streams in northeast Turkey by water quality index and multiple statistical methods. *Environmental Forensics*, 22(1-2), 270-287.
3. Ayub, K. R., Sidek, L. M., Ainan, A., Zakaria, N. A., Ghani, A. A., & Abdullah, R. (2005). Storm water treatment using bio-ecological drainage system. *International Journal of River Basin Management*, 3(3), 215-221.
4. Azhar, S. C., Aris, A. Z., Yusoff, M. K., et al. (2015). Classification of river water quality using multivariate analysis. *Procedia Environmental Sciences*, 30, 79-84.
5. Bahrami, M., Khaksar, E., & Khaksar, E. (2020). Spatial variation assessment of groundwater quality using multivariate statistical analysis (Case Study: Fasa Plain, Iran). *Journal of Groundwater Science and Engineering*, 8(3), 230-243.
6. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
7. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
8. Casillas-García, L. F., de Anda, J., Yebra-Montes, C., Shear, H., Díaz-Vázquez, D., & Gradilla-Hernández, M. S. (2021). Development of a specific water quality index for the protection of aquatic life of a highly

- polluted urban river. *Ecological Indicators*, 129, 107899.
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
10. Dornelas, F. L., Machado, M. B., & von Sperling, M. (2009). Performance evaluation of planted and unplanted subsurface-flow constructed wetlands for the post-treatment of UASB reactor effluents. *Water Science and Technology*, 60(12), 3025-3033.
11. Espejo, L., Kretschmer, N., Oyarzún, J., Meza, F., Núñez, J., Maturana, H., Soto, G., Oyarzo, P., Garrido, M., Suckel, F., Amezcaga, J., & Oyarzún, R. (2012). Application of water quality indices and analysis of the surface water quality monitoring network in semiarid North-Central Chile. *Environmental Monitoring and Assessment*, 184, 5571-5588.
12. Farmaki, E. G., Thomaidis, N. S., Simeonov, V., & Efstathiou, C. E. (2012). A comparative chemometric study for water quality expertise of the Athenian water reservoirs. *Environmental Monitoring and Assessment*, 184(12), 7635-7652.
13. Fernández, N., Ramírez, A., & Solano, F. (2004). Physico-chemical water quality indices—a comparative review. *Bistua: Revista de la Facultad de Ciencias Básicas*, 19-30.
14. Gazzaz, N. M., Yusoff, M. K., Ramli, M. F., Aris, A. Z., & Juahir, H. (2012). Characterization of spatial patterns in river water quality using chemometric pattern recognition techniques. *Marine Pollution Bulletin*, 64(4), 688-698.
15. Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *Journal of Open Source Software*, 3(30), 1051.
16. Guerrero, J., Mahmoud, A., Alam, T., Chowdhury, M. A., Adetayo, A., Ernest, A., & Jones, K. D. (2020). Water quality improvement and pollutant removal by two regional detention facilities with constructed wetlands in South Texas. *Sustainability*, 12(7), 2844.
17. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.
18. Ismail, W. R., Z. A. Rahaman, N. A. Zakaria, A. Ab. Ghani, R. Abdullah & M. Mansor. 2008. Nutrients and water quality of the ecological components of the BioEcological Drainage System (BIOECODS), USM, Penang, Malaysia. In *Asian Wetland Symposium 2008 Ha Noi, Vietnam*.
19. Ismail, M. , SK Ariful, H. , Sujit Kumar ,R. , Jay, K. ,(2024). Assessing intra and interannual variability of water quality in the Sundarban mangrove dominated estuarine ecosystem using remote sensing and hybrid machine learning models, *Journal of Cleaner Production*, Volume 442, 25 February 2024, 140889
20. James, J., Sandhya, L. and Thomas, C., 2013, December. Detection of phishing URLs using machine learning techniques. In *2013 international conference on control communication and computing (ICCC)* (pp. 304-309). IEEE.
21. Johari N. E., Changc. K., Wahid M. A., Ghani A. A. and Talib S. A. (2015). Water Quality Level in Stormwater Runoff through Constructed

- Wetland under Tropical Climate. E-proceedings of the 36th IAHR World Congress 28 June – 3 July, 2015, The Hague, the Netherlands, 1-6.
22. Kadlec, R. H., & Wallace, S. D. (2008). Treatment wetlands (2nd ed.). CRC Press.
 23. Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1-26.
<https://doi.org/10.18637/jss.v028.i05>
 24. Manly, B. F. J. (1986). *Multivariate statistical methods: A primer* (3rd ed.). Chapman & Hall.
 25. Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., & Kaufman, L. (1988). *Chemometrics: Textbook*. Elsevier.
 26. Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., & Ghani, A. A. (2014). Spatial pattern analysis for water quality in free-surface constructed wetland. *Water Science and Technology*, 70(7), 1161-1167.
 27. Montgomery, L., Miège, C., Miller, J., Scambos, T.A., Wallin, B., Miller, O., Solomon, D.K., Forster, R. and Koenig, L., 2020. Hydrologic properties of a highly permeable firn aquifer in the Wilkins Ice Shelf, Antarctica. *Geophysical Research Letters*, 47(22), p.e2020GL089552.
 28. Murdoch, P. S., Baron, J. S., & Miller, T. L. (2000). Potential effects of climate change on surface water quality in North America. *Journal of the American Water Resources Association*, 36, 347-366.
 29. Nguyen, H. D., Hong Quan, N., Quang, N. X., Hieu, N. D., & Thang, L. V. (2021). Spatio-temporal pattern of water quality in the Saigon-Dong Nai river system due to waste water pollution sources. *International Journal of River Basin Management*, 19(2), 221-243.
 30. Nosrati, K., Derafshi, Kh., Ghrachahi, S. and Rahimi, Kh. (2011). Assessment of Surface Water Quality in Haraz-Ghara Soo Watershed using Multivariate Statistical Techniques. *Research Earth Science Journal*, 5, 41-55.
 31. Quoc Bao Pham, Reza Mohammadpour, Nguyen Thi Thuy Linh, Meriam Mohajane, Ameneh Pourjaseem, Saad Sh Sammen, Duong Tran Anh & Van Thai Nam.(2021). Application of soft computing to predict water quality in wetland. Volume 28, pages 185–200.
 32. Razmkhah, H., Abrishamchi, A., & Torkian, A. (2010). Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: A case study on Jajrood River (Tehran, Iran). *Journal of Environmental Management*, 91, 852-860.
 33. Uuemaa, E., Palliser, C. C., Hughes, A. O., & Tanner, C. C. (2018). Effectiveness of a natural headwater wetland for reducing agricultural nitrogen loads. *Water*, 10(3), 287.
 34. Vymazal, J. (2007). Removal of nutrients in various types of constructed wetlands. *Science of the Total Environment*, 380(1-3), 48-65.
 35. Wang, L., Li, X., & Cui, W. (2012). Fuzzy neural networks enhanced evaluation of wetland surface water quality. *International Journal of Computer Applications in Technology*, 44, 235-240.
 36. Wenlan Yang, Bolin Fu, Sunzhe Li, Zhinan Lao, Tengfang Deng, Wen He, Hongchang He, Zhikun Chen.(2023). Monitoring multi-water quality of internationally important karst wetland through deep learning, multi-sensor and multi-platform

- remote sensing images: A case study of Guilin, China. *Ecological Indicators*, Volume 154, October 2023, 110755.
37. Wold, H. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (Vol. 6, pp. 581–591). Wiley.
38. Zakaria, N. A., A. Ab Ghani & K. R. Ayub. 2007. Efficiency of Ecological Pond for Stormwater Pollutants Removal. In 9th Symposium of the Malaysian Society of Applied Biology: Exploring The Science of Life As a Catalyst for Technological Advancement, 0143-0150. Bayview Georgetown, Penang.
39. Zhang, Liu, J., , D., Tang, Q., Xu, H., Huang, S., Shang, D., & Liu, R. (2021). Water quality assessment and source identification of the Shuangji River (China) using multivariate statistical methods. *PloS One*, 16(1), e0245525.