Journal of Al-Turath University College (JATUC)
Open Access Journal
ISSN : 2074-5621
DOI: https://doi.org/10.63964/JATUC.43.2.2026.33
https://jatuc.uoturath.edu.iq/

تقدير انموذج الانحدار اللوجستي الجزائي باستعمال بعض الأساليب الإحصائية وتقنيات التعلم العميق

كرار عبد الرضا احمد [1]، صباح منفي رضا [2]

[1،2] قسم الاحصاء كلية الإدارة والاقصاد، جامعة بغداد، بغداد، العراق.

* *البريد الإلكتروني للتواصل:* karrar.abd2201@coadec.uobaghdad.edu.iq

الملخص

تهدف هذه الدراسة الى تصنيف تلوث مياه الأبار(المياه الجوفية) في العراق من خلال توظيف مجموعة من الأساليب الإحصائية المنتظمة وتقنيات الذكاء الاصطناعي الحديثة، اعتمادًا على بيانات بيئية حقيقية مقدمة من وزارة البيئة العراقية، إضافة إلى بيانات محاكاة مولدة باستخدام أسلوب مونت كارلو. وتم التركيز على نماذج الانحدار اللوجستي الجزائي، ولا سيما طرائق LASSO و Adaptive LASSOو SCADو Net Elastic، إلى جانب خوارزميات التعزيز التدريجي، وبخاصة خوارزميةXGBoost ، بهدف تقييم كفاءتها التفسيرية والتنبؤية عند التعامل مع بيانات بيئية متعددة المتغيرات وذات ارتباطات عالية.

اعتمدت الدراسة على معيار متوسط مربع الخطأ (MSE) للمقارنة بين الطرائق المختلفة، حيث أظهرت نتائج المحاكاة تفوق طريقة Net Elastic بين نماذج الانحدار اللوجستي المنتظم، محققةً أقل قيم لمعيار الخطأ عند مختلف أحجام العينات. كما بينت النتائج تفوق خوارزمية XGBoost على خوارزمية Gradient Boosting من حيث الدقة التنبؤية والاستقرار الإحصائي. وفي التطبيق العملي على بيانات تلوث مياه نهر دجلة، أظهرت جميع المتغيرات البيئية المدروسة دلالة إحصائية عالية في كل من نموذج Net Elastic وخوارزمية XGBoost، مع تقارب عام في اتجاه تأثير المتغيرات، وتميز XGBoost بقدرتها الأعلى على تمثيل العلاقات غير الخطية والتفاعلية.

أظهرت نتائج المقارنة أن نموذج Net Elastic أكثر ملاءمة للأغراض التفسيرية وتحديد المتغيرات المؤثرة في تلوث المياه، في حين تتميز خوارزمية XGBoost بكفاءة تنبؤية أعلى، مما يجعلها مناسبة للتطبيقات التي تتطلب دقة عالية في التنبؤ. وتؤكد الدراسة أن الدمج بين الأساليب الإحصائية المنتظمة وتقنيات الذكاء الاصطناعي يوفر إطارًا تحليليًا متكاملًا وفعالًا لدراسة تلوث المياه، ويسهم في دعم صناع القرار في مجال إدارة الموارد المائية وحماية البيئة.

الكلمات المفتاحية

تلوث المياه، الانحدار اللوجستي الجزائي، لاسو، سكاد، الشبكة المرنة، التصنيف الثنائي

© 2026.This is an open access article under the CC by licenses http://creativecommons.org/licenses/by/4.0

1. المقدمة

يُعد تلوث المياه من أخطر المشكلات البيئية المعاصرة التي تواجه العالم في الوقت الحاضر، لما له من انعكاسات مباشرة وغير مباشرة على صحة الإنسان، والتوازن البيئي، والتنمية الاقتصادية والاجتماعية المستدامة. فقد أدى التطور الصناعي المتسارع، والتوسع الزراعي غير المنظم، والنمو السكاني الكبير، إلى زيادة الضغوط على الموارد المائية السطحية والجوفية، الأمر الذي أسهم في تدهور نوعية المياه وارتفاع مستويات الملوثات الكيميائية والفيزيائية والبيولوجية فيها، ولا سيما في الدول النامية[2] [1].

وتُعد المياه موردًا استراتيجيًا لا غنى عنه في مختلف مناحي الحياة، إذ تعتمد عليها الأنشطة البشرية كافة، سواء في الشرب أو الزراعة أو الصناعة أو توليد الطاقة. ومن هذا المنطلق، أصبحت مراقبة جودة المياه وتحليل مستويات تلوثها من القضايا ذات الأولوية القصوى على المستويين المحلي والعالمي. وتشير التقارير البيئية الحديثة إلى أن تلوث المياه بات سببًا رئيسيًا في انتشار العديد من الأمراض المزمنة، وتدهور النظم البيئية المائية، وانخفاض الإنتاج الزراعي، فضلًا عن ارتفاع التكاليف الاقتصادية المرتبطة بمعالجة المياه والرعاية الصحية [3].

وفي السياق الإقليمي، يُعد العراق من الدول التي تواجه تحديات مائية كبيرة نتيجة شح الموارد المائية، والتغيرات المناخية، والتجاوزات على الأنهار، إضافة إلى ضعف أنظمة المعالجة والسيطرة البيئية. ويُعد نهر دجلة أحد أهم الموارد المائية الرئيسة في العراق، إذ يشكل شريانًا حيويًا تعتمد عليه المدن العراقية في تلبية احتياجاتها المائية المختلفة. إلا أن هذا النهر تعرض خلال العقود الأخيرة إلى مستويات متزايدة من التلوث نتيجة تصريف المياه الثقيلة غير المعالجة، والمخلفات الصناعية والزراعية، فضلاً عن الملوثات المنزلية، مما أدى إلى تدهور ملحوظ في نوعية مياهه وازدياد المخاطر البيئية والصحية المرتبطة به [4].

ومع تطور تقنيات الرصد البيئي، أصبحت بيانات جودة المياه تتسم بالحجم الكبير والتعقيد البنيوي، حيث تتضمن قياسات متعددة لمؤشرات فيزيائية وكيميائية وبيولوجية، غالبًا ما تكون مترابطة فيما بينها بدرجات متفاوتة. وقد أدى هذا التعقيد إلى بروز إشكاليات إحصائية ومنهجية عند استخدام الأساليب التقليدية في التحليل، لا سيما تلك التي تفترض استقلالية المتغيرات أو الخطية البسيطة في العلاقات، الأمر الذي يحد من قدرتها على تقديم نماذج دقيقة وقابلة للتعميم.[5]

في هذا الإطار، أصبحت نماذج الانحدار من الأدوات الأساسية في تحليل البيانات البيئية، وبخاصة نماذج الانحدار اللوجستي التي تُستخدم على نطاق واسع في حالات يكون فيها متغير الاستجابة ذا طبيعة تصنيفية، مثل تصنيف حالة المياه إلى ملوثة وغير ملوثة. إلا أن الانحدار اللوجستي التقليدي يواجه بدوره تحديات كبيرة عند التعامل مع عدد كبير من المتغيرات التوضيحية المتداخلة، مما قد يؤدي إلى مشكلة التعدد الخطي، وعدم استقرار التقديرات، وضعف القدرة التنبؤية للنموذج.

ومن هنا برزت الحاجة إلى استخدام طرائق الانحدار اللوجستي الجزائي(Penalized Logistic Regression) ، مثل طريقة LASSO وطريقة Elastic Net، التي تهدف إلى معالجة هذه الإشكاليات من خلال إدخال حدود جزائية تسهم في تقليل تعقيد النموذج، وتحسين اختيار المتغيرات، وتحقيق توازن أفضل بين الانحياز والتباين [7] [6]. قد أثبتت هذه الطرائق كفاءتها العالية في تحليل البيانات البيئية ذات الأبعاد العالية، ولا سيما في الحالات التي تتسم بوجود ارتباطات قوية بين المتغيرات التوضيحية

وبالتوازي مع ذلك، شهدت تقنيات الذكاء الاصطناعي وتعلم الآلة تطورًا ملحوظًا خلال السنوات الأخيرة، وأصبحت تُستخدم على نطاق واسع في تحليل البيانات الضخمة والمعقدة. ويُعد التعلم بالتجميع(Ensemble Learning) ، وبخاصة خوارزميات التعزيز التدريجي(Gradient Boosting) ، من أكثر التقنيات فعالية في هذا المجال، إذ تعتمد على دمج مجموعة من المتعلمين الضعفاء لبناء نموذج قوي يتمتع بدقة تنبؤية عالية. وقد قدم فريدمان [8]الإطار النظري العام لخوارزمية التعزيز التدريجي، التي شكلت أساسًا لتطوير نماذج أكثر تقدمًا.

وفي هذا السياق، طُوِّرت خوارزمية XGBoost بوصفها نسخة محسّنة من خوارزمية التعزيز التدريجي، حيث تمتاز بسرعتها العالية، وقدرتها على التعامل مع البيانات الكبيرة، وتضمينها لآليات تنظيم صريحة للحد من الإفراط في التعلم، فضلًا عن استخدامها لتقريب من الدرجة الثانية لدالة الخسارة، مما يجعلها أكثر كفاءة من الناحية الحسابية والتنبؤية [9].

وانطلاقًا مما سبق، تتجه الدراسات الحديثة إلى الدمج بين الأساليب الإحصائية المنتظمة وتقنيات الذكاء الاصطناعي، بوصفه مدخلًا منهجيًا متكاملًا لتحليل البيانات البيئية المعقدة، وتحسين دقة التنبؤ، وتعزيز قابلية تفسير النتائج. وفي هذا الإطار، تأتي هذه الرسالة لتسهم في دراسة تلوث مياه نهر دجلة من خلال توظيف طرائق الانحدار اللوجستي الجزائي وخوارزميات التعزيز التدريجي، وبخاصةXGBoost ، اعتمادًا على بيانات محاكاة وحقيقية، بهدف بناء نماذج دقيقة وقابلة للتفسير تسهم في دعم صناع القرار في مجال إدارة الموارد المائية وحماية البيئة.

1.1. مشكلة البحث

على الرغم من التطور الملحوظ في تقنيات رصد جودة المياه وتوافر بيانات بيئية غنية تتضمن عددًا كبيرًا من المتغيرات الفيزيائية والكيميائية، إلا أن تحليل هذه البيانات ما زال يمثل إشكالية علمية ومنهجية، ولا سيما في ظل وجود ارتباطات قوية بين المتغيرات وتداخل تأثيراتها. إذ تعاني النماذج الإحصائية التقليدية، مثل الانحدار اللوجستي الكلاسيكي، من محدودية الكفاءة عند تطبيقها على هذا النوع من البيانات، نتيجة لمشكلات التعدد الخطي وعدم استقرار التقديرات وضعف القدرة التنبؤية.ومن جهة أخرى، ورغم النجاح الذي حققته خوارزميات الذكاء الاصطناعي في تحسين دقة التنبؤ بمستويات تلوث المياه، إلا أن معظمها يفتقر إلى قابلية التفسير العلمي، مما يحد من الاستفادة منها في تحديد العوامل البيئية الأكثر تأثيرًا ودعم اتخاذ القرار. وبناءً على ذلك، تتمثل مشكلة البحث في الحاجة إلى بناء إطار تحليلي متكامل يجمع بين قابلية التفسير التي توفرها نماذج الانحدار اللوجستي الجزائي والكفاءة التنبؤية التي تتميز بها خوارزميات التعزيز التدريجي، ولا سيماXGBoost ، عند تحليل وتصنيف بيانات تلوث مياه نهر دجلة.

1.2. هدف البحث

تهدف هذه الدراسة إلى بناء إطار تحليلي متكامل لدراسة وتصنيف تلوث مياه نهر دجلة بالاعتماد على بيانات بيئية حقيقية وبيانات محاكاة مولدة، وذلك من خلال توظيف مجموعة من الأساليب الإحصائية وتقنيات الذكاء الاصطناعي الحديثة. وتسعى الدراسة إلى تطبيق نماذج الانحدار اللوجستي الجزائي، وخوارزميات التعزيز التدريجي بهدف تقييم كفاءتها في التعامل مع البيانات البيئية متعددة المتغيرات وذات الارتباطات العالية.

1.3. الاستعراض المرجعي

مع تزايد الاهتمام بتصنيف حالة المياه إلى فئات نوعية، مثل ملوثة وغير ملوثة، أو صالحة وغير صالحة للاستخدام، اتجهت العديد من الدراسات إلى استخدام نماذج الانحدار اللوجستي بوصفها أداة إحصائية مناسبة لتحليل المتغيرات التابعة ذات الطبيعة التصنيفية. فقد استُخدم الانحدار اللوجستي في عدد من الدراسات البيئية لتقدير احتمالية تلوث المياه اعتمادًا على مجموعة من المتغيرات الكيميائية والفيزيائية، وأثبت كفاءة مقبولة في الحالات التي يكون فيها عدد المتغيرات محدودًا والارتباط بينها ضعيفًا نسبيًا.غير أن العديد من الباحثين أشاروا إلى أن الانحدار اللوجستي التقليدي يعاني من مشكلات جوهرية عند تطبيقه على بيانات ذات أبعاد عالية أو عند وجود تعدد خطي بين المتغيرات التوضيحية، وهي سمات شائعة في بيانات تلوث المياه. وقد أدى ذلك إلى ظهور الحاجة إلى تطوير نماذج أكثر مرونة وقدرة على التعامل مع هذه الإشكاليات الإحصائية [10].

خلال السنوات الأخيرة، وبفعل التطور الكبير في تقنيات التحليل الإحصائي وتعلم الآلة، تزايد اهتمام الباحثين بتطبيق نماذج الانحدار اللوجستي الجزائي وخوارزميات التعزيز في معالجة البيانات المعقدة ذات الأبعاد العالية، ولا سيما في مجالي تلوث المياه والبيانات الطبية. ففي عام (2016) قدم [9]خوارزمية XGBoost بوصفها تطويرًا متقدمًا لخوارزميات التعزيز التدريجي، وقد أثبتت هذه الخوارزمية كفاءة عالية في مسائل التصنيف والتنبؤ، الأمر الذي شجع الباحثين على توظيفها لاحقًا في تطبيقات بيئية متعددة، من بينها نمذجة جودة المياه والتنبؤ بمؤشراتها.

وفي عام (2017) استخدم [11] خوارزميات التعزيز، ومنهاGradient Boosting ، في التنبؤ ببعض مؤشرات جودة المياه، وأظهرت نتائجهم تفوق نماذج التعزيز على النماذج الإحصائية التقليدية من حيث دقة التنبؤ وقدرتها على تمثيل العلاقات غير الخطية بين المتغيرات البيئية. وفي السياق نفسه، طبّق [12]خوارزمية XGBoost على بيانات جودة المياه السطحية، وتوصلوا إلى أن النموذج حقق قيم خطأ تنبؤي أقل مقارنةً بالشبكات العصبية والانحدار التقليدي.

أما في مجال الانحدار اللوجستي الجزائي، فقد استخدم [13] نماذج الانحدار اللوجستي المنتظم بأساليب مختلفة، مثل LASSO و Elastic Net وAdaptive LASSO، في تحليل بيانات ذات ارتباطات عالية، وأظهرت النتائج تحسنًا واضحًا في اختيار المتغيرات وتقليل متوسط مربعات الخطأ مقارنة بالانحدار اللوجستي الكلاسيكي. وفي عام (2019) استخدم [14] الانحدار اللوجستي باستخدام LASSO و Elastic Netفي تحليل بيانات طبية متعددة المؤشرات، وأشاروا إلى أن Elastic Net كانت أكثر استقرارًا في وجود ارتباط قوي بين المتغيرات، مع احتفاظها بقابلية تفسير جيدة.

وفي التطبيقات البيئية، استخدم [15] نماذج تعلم آلي متعددة، من بينهاXGBoost ، لتقييم وتصنيف جودة المياه الجوفية، وقد بينت نتائجهم تفوق XGBoostمن حيث الدقة التنبؤية عند مقارنتها بخوارزميات أخرى مثل SVM و .Random Forestكما استخدم [3] خوارزميات التعزيز في التنبؤ بمؤشر جودة المياه(WQI) ، وأكدوا أن نماذج التعزيز قادرة على التعامل بكفاءة مع التداخل المعقد بين المؤشرات الكيميائية والفيزيائية.

وفي المجال الطبي، استخدم [16] نماذج الانحدار اللوجستي الجزائي، بما في ذلك SCAD وMCP، في تصنيف الحالات المرضية اعتمادًا على عدد كبير من المتغيرات السريرية، وأظهرت النتائج أن طرائق الجزاء غير المحدب قدمت تقديرات أقل انحيازًا للمعاملات الكبيرة مقارنةً بـLASSO ، مع تحسن في دقة التصنيف. وفي عام (2022)أجرى [17] مقارنة بين الانحدار اللوجستي الجزائي وخوارزمية XGBoost في التنبؤ بنتائج سريرية، حيث أظهرت XGBoost أداءً تنبؤيًا أعلى، في حين تميز الانحدار اللوجستي الجزائي بسهولة تفسير النتائج وتحديد المتغيرات المؤثرة.

وخلال الفترة (2024–2023)، نُشرت عدة دراسات ومراجعات حديثة، من بينها دراسة [18] التي استخدمت XGBoost في تحليل بيانات جودة المياه السطحية، وأكدت النتائج تفوق النموذج في تقليل الخطأ التنبؤي مقارنة بالنماذج التقليدية. كما قدم [19] دراسة تطبيقية استخدمت Elastic Net والانحدار اللوجستي الجزائي في بيانات بيئية متعددة المتغيرات، وأظهرت النتائج أن طرائق الانحدار المنتظم توفر توازنًا مناسبًا بين الدقة التنبؤية وقابلية التفسير.

وبصورة عامة، تشير نتائج الدراسات التطبيقية الحديثة إلى أن نماذج الانحدار اللوجستي الجزائي تُعد أدوات فعّالة في اختيار المتغيرات وتحسين استقرار النماذج في البيانات الطبية والبيئية، في حين أثبتت خوارزميات التعزيز، ولا سيماXGBoost ، تفوقًا ملحوظًا من حيث الأداء التنبؤي عند التعامل مع العلاقات غير الخطية والبيانات المعقدة، مما يعزز الاتجاه نحو الجمع بين هذين المسارين في الدراسات التطبيقية الحديثة.

2. الجانب النظري

1.2 الانحدار اللوجستي

أصبحت طرق الانحدار جزءًا أساسيًا من أي تحليل بيانات يهدف إلى وصف العلاقة بين المتغير التابع والمتغيرات المستقلة .وغالبًا ما يكون المتغير التابع منفصلًا، يمكن أن يأخذ قيمتين أو أكثر .ويُعد نموذج الانحدار اللوجستي الأكثر استخدامًا لتحليل هذا النوع من البيانات. يُعد الانحدار اللوجستي أحد أنواع النماذج غير الخطية التي تصف العلاقة بين المتغير التابع ومجموعة من المتغيرات المستقلة، إذ تكون العلاقة بينهما غير خطية .وغالبًا ما يكون المتغير التابع ذا طبيعة نوعية، يأخذ قيمتين فقط (0) أو(1) ، وهو الأساس في الانحدار اللوجستي الثنائي الذي تمت دراسته في هذه الدراسة وفي حال كانت القيم أكثر من فئتين، يُطلق عليه الانحدار اللوجستي المتعدد الفئات.[20]

يفترض نموذج الانحدار اللوجستي أن المتغير التابع يتبع توزيع برنولي (Bernoulli Distribution) ، ويُعبّر عن دالة الاحتمال كما يلي:

$$(1)p(Y = y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

y_i=0 or 1.

$y_i = 1.$ عندما الاستجابة متغير احتمالية تمثل: $\pi(x_i)$

$y_i = 0.$ الفشل عند متغيرالاستجابة احتمالية تمثل: $1 - \pi(x_i)$

:) رياضيا من حيث المتغيرات التفسيرية والوظيفة اللوجستية كما في الصيغة التالية:-π_iيمكن تعريف الاحتمال (

$$(2)\pi(x_i) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

حيث ان:-

$\underline{\beta}$: *متجه المعاملات*

$X_i = \{1, x_{i1}, x_{i2}, \dots x_{ip}\}$: متجه المتغيرات المستقلة.

ولتبسيط الصيغة يتم استخدام الكمية $\pi(x) = E(Y/x)$ وبتبسيط الرموز، يُستخدم μ لتمثيل المتوسط الشرطي لـ Y كما في المعادلات .

$$(3)\text{logit}\,(\pi(x_i)) = \ln \frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{1-\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}$$

$$= \left(Bo + \sum_{j=1}^{p} Bj\ xij\right) \qquad (4)\text{logit}\,(\pi(x_i)) = \ln\left(e^{X_i\beta}\right) = X_i\beta$$

حيث ان:-

$\beta_0, \beta_1, \dots, \beta_p$:متجه المعلمات المقدرة.

X_{ij} : المتغيرات المستقلة .المعادلة (4) تمثل العلاقة الخطية للنموذج .

2.2 انموذج الانحدار الخطي الجزائي

يستعمل انموذج الانحدار الخطي لوصف العلاقة بين المتغير المعتمد والمتغيرات التوضيحية، إذ يوفر نموذج الانحدار الخطي طريقة بسيطة وسهلة لدراسة العلاقات بين المتغيرات. يعرف نموذج الانحدار الخطي بالشكل الرياضي التالي:

$$y = X\beta + \varepsilon \qquad \dots (5)$$

حيث ان $\mathrm{y} = (\mathrm{y}_1, \dots, \mathrm{y}_n)^T \in \mathbb{R}^n$ متجه متغير الاستجابة،و ان مصفوفة المشاهدات $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_p)^T \in \mathbb{R}^{n\times(p+1)}$ ، حيث ان الصف الاول للمصفوفة يستعمل لحساب متجه معاملات الانحدار $\beta_0, \beta = (\beta_1; \beta_2, \dots, \beta_p)^T \in \mathbb{R}^{p+1}$، وان متجه الاخطاء العشوائية $\varepsilon = (\varepsilon_1, \varepsilon_1, \dots, \varepsilon_1)^T$ التي تكون (i.i.d) مستقلة ولها نفس التوزيع (independently and identically distributed)، حيث انها تتوزع توزيعاً طبيعياً بمتوسط $(\mu = 0)$ وتباين (σ_ε^2). عادة ما تستخدم طريقة المربعات الصغرى الاعتيادية (Ordinary Least Square(OLS)) لتقدير المعادلة (6)، ويتم الحصول على عملية التقدير من خلال تعظيم مجموع مربعات البواقي (Residual Sum of Squares(RSS)) وبالشكل الاتي:

$$RSS = (\mathrm{y} - \mathrm{X}\beta)^T\,(\mathrm{y} - \mathrm{X}\beta) \qquad \dots (6)$$

وبأخذ المشتقة الاولى للمعادلة (6) وجعلها تساوي صفراً، فان مقدر (OLS) لمتجه معلمات الانحدار يعطى بالشكل الاتي :

$$\hat{\beta}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad \dots (7)$$

ان المقدر $\hat{\beta}^{OLS}$ يمتلك خاصية مميزة وهي انه مقدر غير متحيز للمعلمة β .

في العديد من الدراسات التطبيقية، عادة ما يحتوي انموذج الانحدار الخطي على عدد كبير من المتغيرات التوضيحية. وعلى الرغم من إستخدام طريقة (OLS) بشكل واسع لكن يؤخذ عليها عيبين [6];[7]. اولاً: إن تقديرات (OLS) على الرغم من أنها غالبا ما توفر مقدر غير متحيز، الا انها سوف تعطي تباين كبير، والسبب في ذلك يعود إلى العدد الكبير من المتغيرات التوضيحية بالنسبة إلى حجم العينة، أو عندما يكون هناك ارتباط بين المتغيرات التوضيحية، مما يؤثر على دقة التنبؤ لأنموذج الانحدار اولا، وثانيا: سيكون من الصعب تفسير أنموذج الانحدار بسبب وجود عددا كبيرا من المتغيرات التوضيحية.

لقد ظهرت في السنوات الاخيرة طرائق جديدة اكتسبت شعبية واسعة بين الاحصائيين تدعى بالطرائق الجزائية، حيث تعد هذه الطرائق المفتاح المهم الأداء اختيار المتغيرات وتقدير المعلمات في آن واحد. ونتيجة لذلك، اقترحت مجموعة من الطرائق الجزائية يتم من خلالها اضافة قيود جزائية الى نماذج الانحدار. والهدف وراء اضافة القيد الجزائي توفير معيار لاختيار المتغيرات والسيطرة على تعقيد الأنموذج، من خلال إدخال بعض القيود على المعاملات، والتي تفرض على بعض المعاملات ان تكون قيمتها مساوية للصفر مما يستوجب حذفها من النموذج.

ان كمية مقدار الجزاء يعمل على التوازن بين التباين والتحيز في الأنموذج المختار[21]. حيث عندما يكون مقدار الجزاء صغيراً فانه يؤدي إلى تحديد المتغيرات التوضيحية مع تحيز قليل ولكن سيكون التباين عالياً. على العكس من ذلك، فإن كمية الجزاء الكبيرة تؤدي إلى تحديد بعض المتغيرات التوضيحية مع تحيز عالي، ولكن بتباين أقل.

يعرف الانحدار الخطي الجزائي (Penalized Linear Regression (PLR)) بصورة عامة بالشكل التالي:

$$\mathrm{PLR}(\beta;\ \lambda) = (\mathrm{y} - \mathrm{X}\beta)^T\,(\mathrm{y} - \mathrm{X}\beta) + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \qquad \dots (8)$$

إذ ان المقدار $(P_\lambda(.))$ يمثل حد الجزاء، الذي هو دالة من المعاملات, وان (λ) تمثل معلمة التوليف (Tuning parameter) حيث ان $\lambda \in [0, \infty]$. وان حد الجزاء يعتمد تماما على قيمة المعلمة λ حيث انها تسيطر على كمية تقليص قيم المعلمات. فعندما تكون $(\lambda = 0)$ فأننا نحصل على مقدر (OLS). وعلى العكس من ذلك، يزداد عدد المتغيرات المستبعدة كلما تزداد قيمة (λ).

في الانحدار الخطي الجزائي، يتم الحصول على تقديرات المعلمات عن طريق حل المعادلة (9) وكالتالي:

$$\hat{\beta}_{\text{PLR}} = \arg\min_{\beta} \quad \text{PLR}(\beta;\, \lambda) \qquad \dots (9)$$

وعلى نحو مكافئ يمكن كتابة المعادلة (9) بالشكل التالي:

$$\hat{\beta}_{\text{PLR}} = \arg\min_{\beta} \ (\text{y} - \text{X}\beta)^T\,(\text{y} - \text{X}\beta) + \lambda \sum_{j=1}^{p} (|\beta_j|) \qquad \dots (10)$$

لقد تم التطرق الى عدد من الطرائق الجزائية وتم دراسة خصائص كل طريقة بصورة مفصلة.

3.2 طرائق الجزاء

1.1.2 طريقة لاسو (LASSO)

تعد دالة الجزاء (LASSO) والتي تمثل مختصر (Least Absolute Shrinkage and Selection Operator) من أشهر وأكثر الطرائق استخداما والتي تم اقتراحها من قبل [6] تستخدم دالة الجزاء LASSO المقدار $L_1 - norm$ وذلك من خلال اضافته الى مجموع مربعات البواقي. حيث اكتسبت دالة جزاء LASSO على شعبية واسعة واصبحت من الطرائق الجزائية الاساسية في اختيار المتغيرات. وذلك بسبب قدرتها على أداء كل من تقليص قيمة المعلمات واختيار المتغيرات في وقت واحد. حيث تعمل على تقليص معاملات الانحدار وجعلها تساوي الصفر بالضبط.

ويمكن تعريف الجزاء للانحدار الخطي باستخدام (LASSO) كالاتي:

$$\text{PLR}(\boldsymbol{\beta},\lambda)^{\text{LASSO}} = \left\{ (\boldsymbol{y} - \boldsymbol{x\beta})^T\,(\boldsymbol{y} - \boldsymbol{x\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad \dots (11)$$

حيث ان حد الجزاء $\sum_{j=1}^{p} |\beta_j|$ يمثل المقدار (L1-norm) وان(λ) هي معلمة توليف موجبة. وعندما $\lambda \to 0$، يتم الحصول على مقدرات (OLS)، وعندما تزداد قيمة (λ) فإن الكثير من مقدرات المعلمات تتجه نحو قيمة الصفر.

ويتم الحصول على معاملات الانحدار من الانحدار الخطي الجزائي باستخدام (LASSO) على النحو التالي:

$$\hat{\beta}_{PLR}^{\text{LASSO}} = \arg\min_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x\beta})^T\,(\boldsymbol{y} - \boldsymbol{x\beta}) + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \right\} \dots (12)$$

حيث ان المعادلة (12) ليست شكل تربيعي. اضافة الى ذلك، ان المعادلة (12) غير قابلة للاشتقاق عندما $\boldsymbol{\beta} = 0$. ليكن $L(\boldsymbol{\beta})$، وهو دالة الإمكان، فانه سيكون تقدير معاملات دالة الجزاء LASSO في المعادلة (13-2) من خلال اخذ المشتقة الاولى و على النحو الاتي:

$$\frac{\partial L(\beta)}{\partial \boldsymbol{\beta}_j} = -2\boldsymbol{x}_j^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\, sign\,(\boldsymbol{\beta}_j) \qquad \dots (13)$$

وبمساواة المعادلة اعلاه بالصفر نحصل على :

$$\frac{\partial L(\beta)}{\partial \boldsymbol{\beta}_j} = 0 = \begin{cases} -2\boldsymbol{x}_j^{\boldsymbol{T}}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\big) = \lambda\, sign\,(\boldsymbol{\beta}_j) & \forall_j; \hat{\beta}_j \neq 0 \\ \big|-2\boldsymbol{x}_j^{\boldsymbol{T}}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\big)\big| \leq \lambda & \forall_j; \hat{\beta}_j = 0 \end{cases} \dots (14)$$

عندما يكون:

$$sign\,(\boldsymbol{\beta}_j) = \begin{cases} -1 & if\ \ \hat{\beta}_j < 0 \\ 0 & if\ \ \hat{\beta}_j = 0 \\ 1 & if\ \ \hat{\beta}_j > 0 \end{cases} \qquad \dots (15)$$

حيث $\hat{\beta} = \hat{\beta}_{PLR}^{\text{LASSO}}$. وبافتراض ان تصميم المصفوفة متعامدة عندما $\mathbf{X}^T\mathbf{X} = \mathrm{I}_{p\times p}$. اذن حل المعادلة (12) سيكون سهلاً وباعتبارها تقديرات(Soft-Threshold). ويمكن تعريف $\mathrm{S}(\beta;\, \lambda)$ على النحو الاتي :

$$\hat{\beta}_j^{\text{LASSO}} = sign\,(\hat{\beta}_j^{\text{OLS}})\left(|\hat{\beta}_j^{\text{OLS}}| - \frac{\lambda}{2}\right)_+, \qquad j = 1,2,\dots,p \qquad \dots (16)$$

1.2.2 طريقة لاسو المكيف (Adaptive LASSO)

اقترحت طريقة Adaptive Lasso (ALASSO) من قبل[22] وذلك للتخفيف من مشاكل الاتساق في LASSO، في هذه الطريقة يتم استخدام الاوزان والتي يمكن من خلالها تقليص المعاملات في المقدار $L_1 - norm$ ، وان اختيار هذه الاوزان يعتمد على البيانات بشكل تام. حيث ان ALASSO يستعمل المقدار $L_1 - norm$ الموزون بدلاً من المقدار $L_1 - norm$. ان الفكرة الأساسية وراء ALASSO هي أنه يعمل على السماح باستخدام كمية انكماش كبيرة مع المعاملات الصفرية، وكمية صغيرة مع المعاملات غير الصفرية، وفي هذه الحالة فانه يكون من الممكن الحد من التحيز في التقدير وكذلك تحسين الاتساق من اجل اختيار المتغيرات.

ويمكن تعريف الجزاء للانحدار الخطي باستخدام (ALASSO) كالاتي:

$$\mathrm{PLR}(\boldsymbol{\beta},\lambda)^{\mathrm{ALASSO}} = (\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda\sum_{j=1}^{p} w_j|\beta_j| \qquad \dots(17)$$

إذ ان $\mathrm{w} = (\mathrm{w_j};\ \mathrm{j} = 1,2,\dots,\mathrm{p})^{\mathrm{T}} \in \mathbb{R}^p$ متجه الاوزان الذي يمكن الحصول عليه اعتمادا على البيانات, ويمكن الحصول عليه من خلال :

$$\widehat{\mathrm{w}}_{\mathrm{j}} = [abs(\hat{\beta}_j^{\mathrm{initial}})]^{-\gamma}\ ,\mathrm{j} = 1,2,\dots,\mathrm{p} \qquad \dots(18)$$

حيث γ ثابت موجب و $\hat{\beta}^{\mathrm{initial}}$ هي القيمة الاولية, التي تمثل المقدر β، وعلى سبيل المثال, فان مقدرات OLS ومقدرات انحدار الحرف تعتبر قيم اولية.

وان القيمة الثابتة الى n والقيمة الكبيرة الى λ في المعادلة اعلاه تؤدي الى زيادة في التحيز واقل تغير في القيمة المتوقعة [23]

وقد تم إجراء العديد من الدراسات النظرية على مقدر (ALASSO)، عندما p > n، وأيضاً تم اقتراح العديد من الأوزان الأولية. [24]قدم دراسة في خصائص المتناظرة لمقدرات (ALASSO)، حيث اقترح اختيار الوزن الأولي من خلال (marginal regression estimator). وأيضاً تم دراسة توزيع مقدر (ALASSO) من قبل (Potscher and Schneider, 2009) وبالإضافة إلى ذلك، يمكن إستخدام مقدر (LASSO) نفسه كمقدر أولي [25]

يتم تعريف مقدر معلمات (ALASSO)على النحو الاتي:

$$\hat{\beta}_{PLR}^{\mathrm{ALASSO}} = \arg\min_{\beta}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda\sum_{j=1}^{p}\widehat{w}_j|\beta_j|\right\} \quad \dots(19)$$

بالنسبة (orthonormal design)، حيث ان شكل مصفوفة $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$ إذ ان $\widetilde{\boldsymbol{x}}_{\boldsymbol{j}} = \boldsymbol{x_j}/\widehat{w}_j$, $j = 1,2,\dots p$ ويتم تعريف حل (ALASSO) بواسطة دالة (orthonormal design) على النحو التالي:

$$\hat{\beta}_{PLR}^{\mathrm{ALASSO}} = sign(\hat{\beta}_j^{\mathrm{OLS}})\left(|\hat{\beta}_j^{\mathrm{OLS}}| - \left(\frac{\lambda}{2}\right)\right)_+ \qquad j = 1,2,\dots p \qquad \dots(20)$$

حيث ان $(b)_+$ يشير الى الجزء الموجب.

1.3.2 طريقة سكاد (SCAD)

The Smoothly Clipped Absolute Deviation

تعد احدى الطرائق الجزائية المستعملة في تقدير معلمات نماذج الانحدار الخطية وقد تم اقتراحه من قبل [26]، هذه الطريقة تقوم باختيار المتغيرات المهمة وتقدير معلمات الانحدار في نفس الوقت، ويمكن تعريف الجزاء للانحدار الخطي باستخدام (SCAD) كالاتي :

$$\mathrm{PLR}(\boldsymbol{\beta},\lambda)^{\mathrm{SCAD}} = \{(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})\} + \lambda\sum_{i=1}^{p} p_\lambda(|\beta_i|) \qquad \dots(21)$$

ويتم الحصول على معاملات الانحدار من الانحدار الخطي الجزائي باستخدام SCAD وكالاتي:

$$\hat{\beta}^{\text{SCAD}} = \arg\min_{\beta}.\{(\boldsymbol{y}-\boldsymbol{X\beta})^T\,(\boldsymbol{y}-\boldsymbol{X\beta})\} + \lambda\sum_{i=1}^{p} p_\lambda(|\beta_i|) \qquad \ldots(22)$$

حيث ان $p_\lambda(.)$ هي حد جزاء (SCAD) وان (λ) هي معلمة التوليف. وهذه الطريقة تضمن استمرارية النماذج المختارة (لاستقرار اختيار معلمات النموذج) وإعطاء تقديرات غير متحيزة للمعاملات الكبيرة.

وكما لاحظنا في مقدر (LASSO) انه يتم تقليص اغلب قيم $\boldsymbol{\beta}$ الى الصفر، لذلك فانه سوف يتم نقصان عدد كبير من $\boldsymbol{\beta}$، من ناحية اخرى فان مقدر (SCAD) ليس لديه هذا العيب. وكما في (Lasso) فان مقدر (SCAD) يعمل على حل مشكلة الاتساق. وبالرغم من ان (SCAD) ليس دالة محدبة، الا انه تتوفر خوارزميات لحسابه.

1.4.2 طريقة الشبكة المرنة (Elastic Net)

تعد Elastic Net احدى طرائق الامكان الجزائية المهمة، الذي تم اقتراحها من قبل [7]، وذلك للتعامل مع أول اثنين من عيوب LASSO. وتحاول طريقة الشبكة المرنة دمج المقدارين ($\boldsymbol{L_1-norm}$) و ($\boldsymbol{L_2-norm}$)، حيث تستخدم طريقة انحدار الحرف للتعامل مع مشكلة الارتباط بين المتغيرات مع الاستفادة من طريقة LASSO في عملية اختيار المتغيرات. ولتوضيح الفروق بين دالة جزاء Elastic Net و LASSO، يعرض الشكل (2-4) تمثيل ثنائي الأبعاد(2D) في منطقتها المحددة وكذلك طريقة الانحدار.

ويعرف الانحدار الخطي الجزائي بإستخدام Elastic Net كالاتي:

$$\text{PLR}(\beta;\,\lambda_1,\lambda_2)^{\text{Elastic}} = (\mathbf{y}-\mathbf{X\beta})^T\,(\mathbf{y}-\mathbf{X\beta}) + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \qquad \ldots \quad (23)$$

وكما يلاحظ من المعادلة (23) ان طريقة Elastic Net تعتمد على اثنين من معاملات التوليف الموجبة λ_1 و λ_2 . ان الحد الثاني من المعادلة ($26-2$) يعمل على اختيار المتغيرات، حيث ان زيادة كمية انكماش المعلمة λ_1 سوف تعمل على تقليل عدد المتغيرات التوضيحية المختارة، في حين ان الحد الثالث من المعادلة اعلاه يدعم المجاميع الفرعية المؤثرة، حيث ان زيادة كميةλ_2 يزيد من السيطرة على اختيار المجاميع الفرعية المؤثرة.

وعلاوة على ذلك يمكن كتابة المعادلة (24)، بتركيبة خطية من كل من المقدارين (L_1-norm) و (L_2-norm)، بافتراض ان $\alpha = \lambda_2/(\lambda_1+\lambda_2)$:

$$\text{PLR}(\beta;\,\lambda_1,\lambda_2)^{\text{Elastic}} = (\mathbf{y}-\mathbf{X\beta})^T\,(\mathbf{y}-\mathbf{X\beta}) + (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2 \qquad (24)$$

حيث ان $\alpha \in [0,1]$، ويمكن التحكم في كمية الوزن المعطى لكل من (L_1-norm) و (L_2-norm)، عندما $\alpha = 1$ فان المعادلة (24) سوف تكون طريقة انحدار الحرف، وعندما $\alpha = 0$ فان المعادلة (24) سوف تكون طريقة LASSO ، وعندما تكون $0 < \alpha < 1$ ، فانه يؤدي الى عمل كل من طريقة LASSO و طريقة انحدار الحرف بنفس الوقت.

يتم الحصول على تقديرات معاملات الانحدار للانحدار الخطي الجزائي باستخدام Elastic Net على النحو التالي:

$$\hat{\beta}_{PLR}^{\text{Elastic}} = \arg\min_{\beta}\ \text{PLR}(\beta;\,\alpha)^{\text{Elastic}} \qquad \ldots(25)$$

2.3 انموذج الانحدار اللوجستي الجزائي

Penalized logistic Regression Model

يُعد الانحدار اللوجستي طريقة إحصائية مهمة وشائعة الاستخدام في الطب والعلوم الأخرى لمعالجة مشكلات التصنيف الثنائي، حيث يمكن أن يأخذ المتغير التابع قيمتين فقط (0) :أو .(1) ويواجه الباحثون تحديات كبيرة عند تطبيق الانحدار اللوجستي على بيانات عالية الأبعاد مثل بيانات التصوير الطبي الحيوي، وشرائح الـDNA، وعلم الجينوم، خصوصًا عندما يكون عدد المتغيرات (p) أكبر من عدد العينات.(n) .

في هذا العمل، تم استخدام الانحدار اللوجستي الجزائي (PLR) لمعالجة مشكلات تصنيف التعبير الجيني .إذ يقوم النموذج بفرض عقوبة (Penalty)على النموذج في حال وجود عدد كبير من المتغيرات، مما يساعد في التغلب على مشكلة الأبعاد العالية وتحسين أداء التصنيف وتحديد الجينات الأكثر تأثيرًا وفاعلية من حيث التكلفة.

من خلال ربط حد الجزاء $\lambda P(\boldsymbol{\beta})$ مع دالة الامكان الاعظم اللوغاريتمية لانموذج الانحدار اللوجستي، يمكن كتابة انموذج الانحدار اللوجستي الجزائي Penalized logistic regression model (PLR) بالشكل التالي:

$$PLR = \ell(\beta) + \lambda P(\boldsymbol{\beta}) \qquad \ldots(26)$$

حيث يتم تعريف λ كمعلمة توليف ($\lambda \geq 0$). فهي تسيطر على قوة تقليص المتغيرات التوضيحية، عندما تكون قيمة λ كبيرة، فانها ستعطى وزنا أكبر للحد الجزائي. حيث ان قيمة λ تعتمد على البيانات، و يمكن حسابها باستخدام طريقة العبور الشرعي (Cross-Validation) . من الجدير بالذكر ان قبل الشروع بحل PPR، يفضل اجراء تحويل قياسي للمتغيرات التوضيحية لإزالة تأثير وحدات القياس الخاصة بهذه المتغيرات.

وبإعادة كتابة طرائق الجزاء، التي ذكرت سابقا، بدلالة دالة الامكان الاعظم اللوغاريتمية لأنموذج الانحدار اللوجستي نحصل على مجموعة من نماذج انحدار لوجستي الجزائي والتي تضم طريقة LASSO ، SCAD ، ALASSO ، Elastic Net ، MCP والمعرفة بالشكل الرياضي التالي وعلى الترتيب:

1- طريقة لاسو (LASSO)

يتم الحصول على الانحدار اللوجستي الجزائي باستخدام طريقة (LASSO) على النحو التالي:

$$\mathrm{PLR}(\boldsymbol{\beta},\lambda)^{\mathrm{LASSO}} = \left\{ -\sum_{i=1}^{n}\{y_i \ln\pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \quad \ldots (27)$$

2- طريقة لاسو المكيف (ALASSO)

ويمكن تعريف معادلة الانحدار اللوجستي الجزائي باستخدام طريقة (ALASSO) كالاتي:

$$\mathrm{PLR}(\boldsymbol{\beta},\lambda)^{\mathrm{ALASSO}} = \left\{ -\sum_{i=1}^{n}\{y_i \ln\pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda\sum_{j=1}^{p} w_j|\beta_j| \right\} \quad \ldots (28)$$

3- طريقة سكاد (SCAD)

يمكن كتابة معادلة الانحدار اللوجستي الجزائي باستخدام طريقة (SCAD) وكالاتي:

$$\mathrm{PLR}(\boldsymbol{\beta},\lambda)^{\mathrm{SCAD}} = \left\{ -\sum_{i=1}^{n}\{y_i \ln\pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda\sum_{i=1}^{p} p_\lambda(|\beta_i|) \right\} \quad \ldots (29)$$

4- طريقة الشبكة المرنة (Elastic Net)

وتعرف معادلة انحدار لوجستي الجزائي باستخدام طريقة (Elastic Net) كالاتي:

$$\mathrm{PLR}(\boldsymbol{\beta};\ \lambda_1,\lambda_2)^{\mathrm{Elastic}} = \left\{ -\sum_{i=1}^{n}\{y_i \ln\pi(x_{ij}) + (1-y_i)\ln(1-\pi(x_{ij})\} + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \right\} \quad \ldots (30)$$

5- طريقة الجزاء المحدبة (MCP)

وتعرف معادلة انحدار لوجستي الجزائي باستخدام طريقة (MCP) كالاتي:

$$\hat{\beta}_{MCP} = \arg\min_{\beta}\left[-\sum_{i=1}^{n}\left\{\mathbf{y}_i \ln\pi(\mathbf{x}_{ij}) + (1-\mathbf{y}_i)\ln(1-\pi(\mathbf{x}_{ij})\right\} + \lambda\sum_{i=1}^{p} p_\lambda\left(|\beta_i|\right)\right]$$

حيث ان $p_\lambda(.)$ هي دالة الجزاء الخاصة بطريقة الجزاء المحدبة:

$$\frac{2c\lambda|\beta_j| - \beta_j^2}{2c} I\left(|\beta_j| \leq c\lambda\right) + \frac{c\lambda^2}{2} I\left(|\beta_j| > c\lambda\right), c > 1 and\ \lambda \geq 0.$$

2.4 خوارزميات التعزيز التدريجي (Gradient Boosting) وخوارزمية XGBoost في التصنيف

A. مقدمة عن خوارزميات التعزيز(Boosting)

يُعدّ التعلّم بالتجميع (Ensemble Learning) من الاتجاهات الحديثة في مجال الذكاء الاصطناعي وتعلّم الآلة، إذ يعتمد على دمج عدد من المتعلّمين الضعفاء (Weak Learners) لتكوين نموذج قويّ قادر على تحقيق دقة أعلى في التنبؤ .تُعدّ خوارزميات التعزيز (Boosting) من أكثر طرق التجميع كفاءة، حيث تُبنى النماذج بشكل تسلسلي بحيث يعمل كل نموذج جديد على تصحيح أخطاء النماذج السابقة.
قدّم الباحث فريدمان [8]الإطار الرياضي العام لخوارزمية التعزيز التدرّجي (Gradient Boosting) بوصفها عملية تحسين تدرّجي (Gradient Descent)في فضاء الدوال، مما جعلها أساسًا للعديد من النماذج المتقدمة في مهام الانحدار والتصنيف.

B. خوارزمية التعزيز التدرّجي (Gradient Boosting) للتصنيف

في مهام التصنيف الثنائي، تُستخدم عادة دالة الخسارة اللوجستية (Logistic Loss) المعرّفة كما يلي:

$$L(y, F(x)) = \log(1 + e^{-2yF(x)}), \quad y \in \{-1, +1\}$$

الهدف هو إيجاد دالة التنبؤ المثلى التي تُقلّل متوسط الخسارة:

$$\hat{F} = \arg\min_{F} \sum_{i=1}^{n} L(y_i, F(x_i))$$

يبدأ النموذج بقيمة ابتدائية$F0(x)$ ، ثم في كل مرحلة m يتم حساب المخلفات المزيفة،
ثم بناء متعلّم جديد $hm(x)$ يحاكي هذه المخلفات، وبعد ذلك يتم إيجاد الوزن γm الأمثل وتحديث النموذج كالتالي:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x)$$

وفي النهاية يتم تحويل القيم الناتجة إلى احتمالات تصنيف باستخدام دالة سيغمويد

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-2F_M(x)}}$$

C. خوارزمية(XGBoost (Extreme Gradient Boosting

قدّم تشين وغوسترين [9] خوارزمية (XGBoost) كنسخة محسّنة من خوارزمية التعزيز التدرّجي(Gradient Boosting) ، تهدف إلى زيادة السرعة والدقة وتطبيق تنظيم صريح لتجنّب الإفراط في التعلّم . تستخدم الخوارزمية تقريبًا من الدرجة الثانية لدالة الخسارة وتضيف دالة تنظيم (Regularization Function) للتحكم في تعقيد النموذج .

النتائج والمناقشة

3. الجانب التجريبي

يعرض هذا الجانب نتائج ومناقشة ومقارنة أداء الطرائق الجزائية لنماذج الانحدار اللوجستي بالاضافة الى خوارزميات التحسين المعتمدة على الذكاء الاصطناعي. إذ تم اختبار فعالية كل طريقة من خلال تطبيق بيانات المحاكاة والبيانات الحقيقية على نماذج الانحدار اللوجستي الجزائي مع خوارزمتي الذكاء الاصطناعي، إذ تم فيه التطرق الى مفهوم المحاكاة، ومن ثم اقامة تجربة محاكاة ووصفها لتحقيق هدف الرسالة بالمقارنة بين طرائق التقدير الجزائية المختلفة وبيان أفضليتها، وان جميع العمليات الحسابية لبيانات المحاكاة والبيانات الحقيقية قد تم تنفيذها من خلال برامج كتبت باستخدام برنامج ماتلاب . وتم الاعتماد على معيار متوسط مربعات الخطأ لتقييم طرائق الجزاء وطرائق خوارزميات الذكاء الاصطناعي كمعيار للمفاضلة بين الطرائق المستعملة.

3.1 وصف تجربة المحاكاة

لقد تم تصميم تجربة ومحاكاتها باستخدام برنامج ماتلاب إذ تم توليد المتغير (y_i) في انموذج الانحدار لوجستي الثنائي الذي يتبع توزيع برنولي(0و1)، إذ تم استخدام اسلوب مونت كارلو (Monte Carlo) في المحاكاة وكالاتي:
تم تعيين قيم عدد المتغيرات التوضيحية (p) ، وفضلا عن العينات (n) إذ تم استخدام سبعة احجام مختلفة من العينات وهي (30,50,100,300,500,700,1000) وذلك لأجل دراسة المقارنة وفق العينات باختلاف أنواعها (صغيرة، متوسطة، كبيرة).

3.2 خوارزمية المحاكاة

اولا: تم توليد بيانات المتغير y التي تتبع انموذج الأنحدار اللوجستي وكالاتي :

$$y \sim Ber(\mathrm{P}(0,1))$$

ثانياً: تم توليد مصفوفة المتغيرات التوضيحية **X** ذات ابعاد $(n \times p)$ التي تتبع التوزيع الطبيعي المتعدد **(Multivariate Normal Distribution)** كالاتي **:**

$$X \sim \mathrm{MN}(\mu, \sum)$$

إذ ان $\sum$ هي مصفوفة التباين المشترك, إذ ان $\sum_{ij} = r^{|i-j|}$, عندما $(i, j = 1,2, \dots, p)$ إذ ان المتغيرات التوضيحية تكون مرتبطة.

ثالثاً : تم تكرار التجربة (100) مرة وذلك لغرض تقليل التحيز في تجارب مونت كارلو (Monte Carlo).

3.2.1 نتائج المحاكاة لنماذج الانحدار اللوجستي الجزائي(المنتظم)

يبيّن جدول (1) نتائج المحاكاة الخاصة بقيم متوسط مربع الخطأ (MSE) لمجموعة من طرائق الانحدار اللوجستي الجزائي(المنتظم)، وهي: LASSO، Adaptive LASSO (ALASSO)، SCAD، Elastic Net، وMCP، وذلك عند مستويات مختلفة من أحجام العينات n وقد استُخدم معيار MSE لكونه من أكثر المعايير شيوعاً في تقييم دقة النماذج الإحصائية وقدرتها على التنبؤ، حيث تشير القيم الأصغر إلى أداء أفضل للنموذج.

جدول 1. يبين Mse لكل طريقة

n	LASSO	ALASSO	SCAD	Elastic Net	MCP	Best
30	0.9023	0.9010	0.8801	0.3362	0.7741	Elastic Net
50	0.7902	0.8921	0.8154	0.2114	0.7015	Elastic Net
100	0.5970	0.7703	0.5533	0.1110	0.5566	Elastic Net
300	0.3654	0.6554	0.4018	0.1008	0.4982	Elastic Net
500	0.2155	0.5411	0.2155	0.0946	0.4910	Elastic Net
700	0.1924	0.2501	0.1011	0.0019	0.2591	Elastic Net
1000	0.0954	0.1088	0.0345	0.0004	0.1002	Elastic Net

من خلال النتائج الواردة في الجدول، يُلاحظ بوضوح أن طريقة Elastic Net حققت أقل قيم لـ MSE عند جميع أحجام العينات المدروسة، مما يجعلها الطريقة الأفضل مقارنةً ببقية الطرائق، وهو ما ينعكس في عمود(Best) .

عند حجم عينة صغير نسبيًا n=30، سجلت طريقة Elastic Net قيمة MSE منخفضة بشكل ملحوظ بلغت (0.3362)، مقارنةً ببقية الطرائق التي أظهرت قيماً أعلى بكثير، لا سيما LASSO وALASSO، مما يشير إلى تفوق Elastic Net في التعامل مع صِغر حجم العينة والتقليل من التباين الناتج عن التقدير.

ومع زيادة حجم العينة إلى n=50وn=100، استمر هذا التفوق، حيث انخفضت قيمة MSE لطريقة Elastic Net إلى (0.2114) و(0.1110) على التوالي، في حين أظهرت بقية الطرائق تحسنًا تدريجيًا ولكن بدرجات أقل. ويلاحظ أن طريقة SCAD بدأت بإظهار أداء أفضل نسبيًا مقارنة بـ LASSOو ALASSOعند n=100 = 100n=100، إلا أنها بقيت أقل كفاءة منElastic Net .

عند أحجام العينات المتوسطة n=300 300وn=500، لوحظ انخفاض عام في قيم MSE لجميع الطرائق، مما يعكس الأثر الإيجابي لزيادة حجم العينة على دقة التقدير. ومع ذلك، استمرت Elastic Net في تحقيق أقل القيم، حيث بلغت (0.1008) و(0.0946) على التوالي، في حين بقيت قيم MSE لطرائق MCP و ALASSOمرتفعة نسبيًا.

أما عند أحجام العينات الكبيرة n=700 وn=1000، فقد ظهر تفوق Elastic Net بشكل أكثر وضوحًا، إذ وصلت قيمة MSE إلى (0.0019) و(0.0004) على التوالي، وهي قيم صغيرة جدًا مقارنة ببقية الطرائق. كما يُلاحظ أن طريقة SCAD أظهرت تحسنًا ملحوظًا عند هذه الأحجام، إلا أنها لم تصل إلى مستوى الدقة الذي حققتهElastic Net .

بصورة عامة، تشير هذه النتائج إلى أن طريقة Elastic Net تمتلك قدرة عالية على تحقيق توازن فعّال بين الانكماش والاختيار المتغيري، مما يجعلها أكثر استقرارًا وأفضل أداءً عبر مختلف أحجام العينات، سواء كانت صغيرة أو كبيرة. كما تؤكد نتائج المحاكاة أن الاعتماد على Elastic Net يُعد خيارًا مناسبًا في النماذج التي تعاني من الارتباط العالي بين المتغيرات التفسيرية أو عند اختلاف أحجام العينات.

3.3.2 نتائج المحاكاة لخوارزميات الذكاء الاصطناعي

يعرض جدول (2) نتائج المحاكاة الخاصة بقيم متوسط مربع الخطأ (MSE) لخوارزميتي التعزيز التدريجي Gradient Boosting (GBoost) و Extreme Gradient Boosting (XGBoost)عند مستويات مختلفة من أحجام العينات n ويُعد معيار MSE من المقاييس الأساسية لتقييم دقة النماذج التنبؤية، إذ تعكس القيم الأصغر قدرة أعلى للنموذج على تمثيل العلاقة الحقيقية بين المتغيرات وتقليل الخطأ التنبؤي.

جدول 2. يبين Mse لكل طريقة

n	GBoost	XGBoost	Best
30	0.021566	0.011166	XGBoost
50	0.012336	0.010056	XGBoost
100	0.008547	0.002222	XGBoost
300	0.005210	0.001143	XGBoost
500	0.001253	0.000423	XGBoost
700	0.000589	0.000177	XGBoost
1000	0.000148	0.000021	XGBoost

تُظهر النتائج الواردة في الجدول اعلاه تفوقًا واضحًا ومستمرًا لخوارزمية **XGBoost** على خوارزمية GBoost عند جميع أحجام العينات المدروسة، حيث حققت XGBoost أقل قيم لمتوسط مربع الخطأ، وهو ما انعكس في عمود (Best) ويشير هذا التفوق إلى الكفاءة العالية لـ XGBoost في تحسين دقة التنبؤ مقارنة بخوارزمية التعزيز التقليدية.

عند أحجام العينات الصغيرة نسبيًا 30=و50=n، سجلت خوارزمية XGBoost قيماً منخفضة لـ MSE بلغت (0.011166) و(0.010056) على التوالي، مقارنة بقيم أعلى لخوارزمية GBoostويمكن تفسير ذلك من منظور إحصائي بأن XGBoost تعتمد على آليات تنظيم إضافية (Regularization)تشمل جزاءً على تعقيد النموذج، مما يقلل من التباين ويحد من مشكلة الإفراط في الملاءمة(Overfitting) ، وهي مشكلة شائعة عند التعامل مع أحجام عينات صغيرة.

ومع زيادة حجم العينة إلى 100=n و300=n، لوحظ انخفاض ملحوظ في قيم MSE لكلتا الخوارزميتين، الأمر الذي يعكس تحسن قدرة النماذج على التعلّم من البيانات وتقدير الأنماط الكامنة فيها. إلا أن الانخفاض كان أكثر حدة في خوارزميةXGBoost ، حيث وصلت قيمة MSE إلى (0.002222) و(0.001143)، مما يدل على كفاءة أعلى في استغلال المعلومات الإضافية التي يوفرها ازدياد حجم العينة.

أما عند أحجام العينات الكبيرة 500=n، 700=n و1000=n، فقد أظهرت خوارزمية XGBoost تفوقًا لافتًا، حيث انخفضت قيم MSE إلى مستويات صغيرة جدًا بلغت (0.000423)، (0.000177) و(0.000021) على التوالي. ويشير هذا السلوك إلى خاصية الاتساق (Consistency) العالية للخوارزمية، إذ تقترب التقديرات التنبؤية تدريجيًا من القيم الحقيقية مع تزايد حجم العينة، دون حدوث زيادة ملحوظة في التباين.

من الناحية الإحصائية، يمكن إرجاع هذا الأداء المتفوق لـ XGBoost إلى عدة عوامل، من أبرزها اعتمادها على التوسع من الدرجة الثانية لدالة الخسارة(Second-order Taylor Approximation) ، واستخدامها خوارزمية تحسين أكثر كفاءة، فضلًا عن قدرتها على التعامل مع القيم المفقودة والارتباطات المعقدة بين المتغيرات. كما أن الجمع بين التعزيز التدريجي والتنظيم الصريح يجعلها تحقق توازنًا أفضل بين الانحياز والتباين مقارنة بخوارزمية GBoost التقليدية.

وبصورة عامة، تؤكد نتائج المحاكاة أن خوارزمية XGBoost تُعد أكثر كفاءة واستقرارًا في تقليل الخطأ التنبؤي عبر مختلف أحجام العينات، مما يجعلها خيارًا مناسبًا في التطبيقات التي تتطلب دقة تنبؤية عالية وقابلية جيدة للتعميم، خاصة في البيئات التي تتسم بتعقيد البيانات أو كبر حجمها.

3.4 الجانب التطبيقي

تم الاعتماد في الجانب التطبيقي من هذه الدراسة على بيانات حقيقية لتلوث المياه، تم الحصول عليها من وزارة البيئة العراقية، حيث تم عمل كتاب تسهيل مهمة رسمي الى الجهة الخاصة بمراقبة نوعية مياه نهر دجلة في عدد من المواقع المختارة داخل العراق. وقد جُمعت هذه البيانات ضمن برنامج الرصد البيئي الدوري الذي تنفذه الوزارة، بهدف تقييم مستوى التلوث المائي ومتابعة التغيرات في الخصائص الفيزيائية والكيميائية للمياه.

تتضمن البيانات (714) مشاهدة تمثل قياسات فعلية لعينات مياه مأخوذة من نهر دجلة خلال فترة زمنية محددة، حيث جرى جمع العينات وفق الإجراءات المعتمدة من قبل الجهات المختصة، وباستخدام طرائق قياس مختبرية قياسية لضمان دقة النتائج وموثوقيتها. وقد شملت العينات مجموعة من المتغيرات البيئية التي يُعتقد أن لها تأثيرًا مباشرًا في مستوى تلوث المياه، وتمثل هذه المتغيرات الخصائص الفيزيائية والكيميائية الأساسية للمياه.

وقد تم تسجيل البيانات الخاصة بكل عينة ضمن استمارات رسمية أعدّت لهذا الغرض من قبل وزارة البيئة، وشملت قياسات تراكيز بعض الملوثات والعناصر الكيميائية، إضافة إلى مؤشرات جودة المياه المعتمدة في تقييم التلوث المائي. ويُعد نهر دجلة من أهم الموارد المائية في العراق، الأمر الذي يضفي أهمية خاصة على هذه البيانات من الناحية البيئية والصحية والاقتصادية.

تضمنت الدراسة عددًا من المتغيرات التوضيحية التي تم اختيارها استنادًا إلى الأدبيات البيئية والدراسات السابقة، والتي يُعتقد بأن لها تأثيرًا في متغير الاستجابة الذي يمثل هل ان المياه ملوثة وغير صالحة للاستهلاك البشري ام صالحة للاستهلاك البشري . وقد استُخدمت هذه المتغيرات كأساس لتطبيق النماذج الإحصائية وخوارزميات الذكاء الاصطناعي لغرض تحليل البيانات، وتحديد العوامل الأكثر تأثيرًا في تدهور نوعية المياه.

يوضح الجدول (3) المتغيرات التوضيحية والمتغير التابع المستخدمة في الدراسة، مع بيان وحدات القياس والحدود المسموح بها لكل متغير، حيث تم اعتماد هذه المعايير بالاستناد إلى إرشادات منظمة الصحة العالمية (WHO) لجودة مياه الشرب، لما تمثله من مرجع دولي معتمد في تقييم صلاحية المياه للاستهلاك البشري

جدول 3. وصف المتغير التابع والمتغيرات التوضيحية المستخدمة في الدراسة

المتغير	المدى/الحد المسموح	ملاحظات
Y_i	1 صالح للأستهلاك البشري 0 غير صالح للأستهلاك البشري	
نترات(NO3)	≤ 10 mg/L	وفق توصيات شائعة لجودة مياه الشرب
بيكربونات(HCO3)	≤ 300 mg/L	قيمة إرشادية للبيكربونات
كبريتات(SO4)	≤ 250 mg/L	حد تركيزي شائع للكبريتات
كلوريد(Cl)	≤ 250 mg/L	حد تركيزي شائع للكلوريد
كالسيوم(Ca)	≤ 75 mg/L	للصلابة الكلسية مساهمة
مغنيسيوم(Mg)	≤ 50 mg/L	يسهم في الصلابة الكلية
صوديوم(Na)	≤ 200 mg/L	اعتبارات الطعم والصحة
بوتاسيوم(K)	≤ 10 mg/L	عادة منخفض في المياه العذبة
الرقم الهيدروجيني(pH)	بين 6.5 و 8.5	النطاق المقبول للشرب

) TDSالأملاح الذائبة الكلية(	≤ 500 mg/L	مرتبطة بالطعم والتوصيل الكهربائي

تقدير معلمات أنموذج الانحدار اللوجستي المنتظم بطريقة الشبكة المرنة

جدول 4. تقديرات معلمات انموذج الانحدار اللوجستي المنتظم لطريقة الشبكة المرنة لتلوث مياه نهر دجلة

Variable	Estimate	SE	p-value
Intercent	1.3256	$3.23x10^{-6}$	<0.000
NO3	-0.7523	$4.22x10^{-6}$	<0.000
HCO3	0.8544	$2.11x10^{-6}$	<0.000
SO4	0.7080	$1.21x10^{-6}$	<0.000
Cl	-0.0214	$3.33x10^{-6}$	<0.000
Ca	0.0987	$4.56x10^{-6}$	<0.000
Mg	-0.6201	$5.81x10^{-6}$	<0.000
Na	-0.1142	$9.05x10^{-6}$	<0.000
K	0.0015	$8.88x10^{-6}$	<0.000
PH	0.0025	$7.99x10^{-6}$	<0.000
TDS	-0.8881	$6.10x10^{-6}$	<0.000

*معنوية عند $\alpha = 0.05$

وضح جدول (4) نتائج تقدير معاملات نموذج Elastic Net المطبق على بيانات تلوث مياه نهر دجلة، حيث تم تقدير معاملات المتغيرات التوضيحية، إضافة إلى الثابت(Intercept) ، مع بيان الخطأ المعياري (SE) وقيم الاحتمالية .(p-value) ويُلاحظ أن جميع المتغيرات كانت ذات دلالة إحصائية عالية عند مستوى معنوية أقل من (0.05>)، بل إن قيم الاحتمالية كانت أقل من (0.000>)، مما يدل على قوة العلاقة الإحصائية بين المتغيرات المدخلة ومتغير الاستجابة، ويعكس كفاءة النموذج في تفسير التغيرات في مستوى تلوث المياه.

تشير قيمة الثابت (Intercept) البالغة (1.3256) إلى القيمة المتوقعة لمتغير الاستجابة عند ثبات جميع المتغيرات التوضيحية عند الصفر، وهي قيمة ذات دلالة إحصائية عالية، ما يعكس وجود مستوى أساسي من التلوث في مياه نهر دجلة حتى في غياب تأثير المتغيرات المدروسة، وهو ما يمكن تفسيره بوجود مصادر تلوث خلفية ناتجة عن عوامل طبيعية أو بشرية أخرى غير مدرجة في النموذج.

بالنسبة لمتغير النترات(NO_3) ، فقد ظهر بمعامل سالب قدره (0.7523-)، مما يشير إلى علاقة عكسية بين تركيز النترات ومتغير الاستجابة. ويُفسر ذلك بأن زيادة تركيز النترات قد ترتبط بتغيرات كيميائية في المياه تؤدي إلى خفض قيمة مؤشر التلوث المعتمد في الدراسة، أو قد تعكس تداخل النترات مع متغيرات أخرى ضمن آلية الانتظام التي يعتمدها نموذج.Elastic Net

أما البيكربونات **(HCO_3)** فقد سجلت معاملًا موجبًا (0.8544)، مما يدل على علاقة طردية قوية مع متغير الاستجابة، ويشير ذلك إلى أن ارتفاع تراكيز البيكربونات يسهم في زيادة مستوى تلوث المياه، وهو ما قد يرتبط بتصريف المياه الصناعية أو التفاعلات الجيولوجية التي تؤثر في التركيب الكيميائي للمياه.

كما أظهرت كبريتات (SO_4) تأثيرًا موجبًا ومعنويًا، حيث بلغ معاملها (0.7080)، مما يدل على أن زيادة تركيز الكبريتات تسهم في ارتفاع مستوى التلوث، وهو ما يتفق مع العديد من الدراسات البيئية التي تربط ارتفاع الكبريتات بتصريف المخلفات الصناعية والزراعية إلى المجاري المائية.

في المقابل، ظهر الكلوريد **(Cl)** بمعامل سالب (0.0214-)، مما يشير إلى تأثير عكسي ضعيف نسبيًا على متغير الاستجابة، رغم دلالته الإحصائية العالية. وقد يعزى ذلك إلى أن الكلوريد يعد من الأيونات الشائعة في المياه الطبيعية، وأن تأثيره في مستوى التلوث الكلي قد يكون أقل مقارنة بغيره من الملوثات.

أما بالنسبة للأيونات القلوية الترابية، فقد أظهر الكالسيوم **(Ca)** معاملًا موجبًا (0.0987)، في حين سجل المغنيسيوم **(Mg)** معاملًا سالبًا (0.6201-) من تأثير بعض الملوثات الأخرى.

وبخصوص العناصر القلوية، فقد سجل كل من الصوديوم (Na) معاملًا سالبًا (0.1142-)، في حين كان تأثير البوتاسيوم **(K)** ضعيفًا جدًا وبمعامل موجب صغير (0.0015)، مما يدل على أن تأثير البوتاسيوم في مستوى التلوث المدروس محدود مقارنة ببقية المتغيرات، رغم دلالته الإحصائية.

أما الأس الهيدروجيني**(pH)** ، فقد ظهر بمعامل موجب صغير (0.0025)، مما يشير إلى أن التغيرات الطفيفة في درجة الحموضة قد يكون لها تأثير محدود ولكن معنوي على مستوى التلوث، وهو ما يعكس حساسية النظام المائي للتغيرات الكيميائية حتى وإن كانت طفيفة.

وسجلت المواد الصلبة الذائبة الكلية **(TDS)** أعلى تأثير سلبي بين المتغيرات، حيث بلغ معاملها (0.8881-)، مما يدل على علاقة عكسية قوية مع متغير الاستجابة. ويعكس ذلك الدور الكبير لـ TDS في تحديد خصائص المياه، وقد يشير إلى أن ارتفاع تركيز المواد الذائبة يؤدي إلى تغيرات تؤثر في مؤشر التلوث المعتمد في الدراسة.

بصورة عامة، تُظهر نتائج نموذج Elastic Net قدرة عالية على تحديد المتغيرات الأكثر تأثيرًا في تلوث مياه نهر دجلة، مع تحقيق توازن فعّال بين تقليل التباين والانحياز، خاصة في ظل تعدد المتغيرات واحتمالية وجود ارتباطات فيما بينها. كما تؤكد النتائج أن استخدام Elastic Net يُعد مناسبًا لتحليل البيانات البيئية المعقدة، ويسهم في تقديم تفسير علمي دقيق للعوامل المؤثرة في تلوث المياه.

وبهذا تكون معادلة الانحدار اللوجستي المنتظم للشبكة المرنة لبيانات تلوث مياه نهر دجلة كما يلي:

$$log(p / (1 - p)) = 1.3256 - 0.7523\ NO3 + 0.8544\ HCO3 + 0.7080\ SO4 - 0.0214\ Cl + 0.0987\ Ca - 0.6201\ Mg - 0.1142\ Na + 0.0015\ K + 0.0025\ pH - 0.8881\ TDS$$

وكذلك من جانب اخر نتائج استخدام خوارزميات الذكاء الاصطناعي

جدول 5. تقديرات معلمات انموذج الانحدار اللوجستي المنتظم لطريقة XGBOOST لتلوث مياه نهر دجلة

Variable	Estimate	SE	p-value
Intercent	3.0225	$5.115x10^{-8}$	<0.000
NO3	-0.1361	$2.102x10^{-8}$	<0.000
HCO3	0.0214	$1.662x10^{-8}$	<0.000
SO4	0.0563	$3.851x10^{-8}$	<0.000
Cl	-0.0015	$4.996x10^{-8}$	<0.000
Ca	0.0321	$5.227x10^{-8}$	<0.000
Mg	-0.0159	$2.605x10^{-8}$	<0.000
Na	-0.0277	$6.509x10^{-8}$	<0.000
K	0.0022	$7.007x10^{-8}$	<0.000
PH	0.0003	$5.301x10^{-8}$	<0.000
TDS	-0.0018	$9.199x10^{-8}$	<0.000

يوضح جدول (5) نتائج تقدير معاملات خوارزمية XGBoostالمطبقة على بيانات تلوث مياه نهر دجلة، حيث عُرضت قيم المعاملات التقديرية (Estimate)، والانحرافات المعيارية(SE) ، وقيم الاحتمالية .(p-value) ويُلاحظ أن جميع المتغيرات، بما فيها الحد الثابت(Intercept) ، كانت ذات دلالة إحصائية عالية جدًا عند مستويات معنوية أقل من (0.05>)، إذ بلغت جميع قيم الاحتمالية أقل من (0.000>)، مما يدل على قوة النموذج في التقاط العلاقات بين المتغيرات البيئية ومتغير الاستجابة.

تشير قيمة الحد الثابت المرتفعة نسبيًا (3.0225) إلى وجود مستوى أساسي مرتفع من احتمالية التلوث في مياه نهر دجلة، حتى في حال ثبات جميع المتغيرات التوضيحية، وهو ما يعكس التأثير التراكمي لمصادر تلوث متعددة، بعضها قد لا يكون ممثلاً صراحةً ضمن المتغيرات المدخلة للنموذج، وهو ما تتميز خوارزميات التعلم المعزز بالقدرة على التقاطه ضمن بنيتها الشجرية.

أظهر متغير النترات (NO_3) تأثيرًا سالبًا ومعنويًا، إذ بلغ معامل التقدير (0.1361-)، مما يشير إلى علاقة عكسية بين تركيز النترات واحتمالية تلوث المياه وفق النموذج. ويُفسر ذلك بأن XGBoost تأخذ في الحسبان التأثيرات غير الخطية والتفاعلية بين المتغيرات، الأمر الذي قد يؤدي إلى اختلاف إشارة التأثير مقارنة بالنماذج الخطية المنتظمة.

أما البيكربونات (HCO_3) فقد سجلت تأثيرًا موجبًا (0.0214)، مما يدل على مساهمتها في رفع احتمالية التلوث، وإن كان تأثيرها أقل حجمًا مقارنة ببعض المتغيرات الأخرى، وهو ما يعكس قدرة النموذج على توزيع الأهمية النسبية للمتغيرات ضمن البنية الشجرية المعززة.

كما أظهرت الكبريتات (SO_4) تأثيرًا موجبًا ومعنويًا (0.0563)، مما يشير إلى دورها في زيادة مستويات التلوث، ويتوافق ذلك مع كون الكبريتات من المؤشرات المهمة لتصريف المخلفات الصناعية والزراعية في المسطحات المائية.

بالنسبة إلى الكلوريد(Cl) ، فقد ظهر بتأثير سالب ضعيف جدًا (0.0015-)، مما يدل على محدودية تأثيره المباشر في احتمالية التلوث مقارنة ببقية المتغيرات، رغم دلالته الإحصائية العالية، وهو ما يعكس حساسية XGBoost العالية في الكشف عن التأثيرات الدقيقة.

وسجل الكالسيوم (Ca) تأثيرًا موجبًا (0.0321)، في حين كان تأثير المغنيسيوم (Mg) سالبًا (-0.0159)، مما يشير إلى اختلاف دور الأيونات القلوية الترابية في تحديد خصائص المياه، حيث يتأثر تأثيرها بالتفاعلات المعقدة مع بقية المكونات الكيميائية للمياه.

كما أظهرت النتائج أن الصوديوم (Na) له تأثير سالب ومعنوي (0.0277-)، في حين كان تأثير البوتاسيوم (K) موجبًا ضعيفًا (0.0022)، مما يعكس التباين في دور العناصر القلوية في التأثير على جودة المياه واحتمالية تلوثها.

أما الأس الهيدروجيني(pH) ، فقد ظهر بتأثير موجب ضعيف جدًا (0.0003)، مما يدل على أن التغيرات الطفيفة في درجة الحموضة تؤثر بدرجة محدودة ولكن معنوية على احتمالية التلوث، وهو ما يتسق مع الطبيعة الحساسة للأنظمة المائية تجاه التغيرات الكيميائية.

وأخيرًا، سجلت المواد الصلبة الذائبة الكلية (TDS) تأثيرًا سالبًا ومعنويًا (0.0018-)، مما يشير إلى دورها في التأثير على خصائص المياه بطريقة غير مباشرة، ويعكس قدرة خوارزمية XGBoost على التقاط التأثيرات المعقدة وغير الخطية بين TDS وبقية المتغيرات البيئية.

4. المناقشة

بصورة عامة، تُظهر نتائج خوارزمية XGBoostقدرة عالية على نمذجة العلاقات غير الخطية والتفاعلية بين المتغيرات البيئية المختلفة، مما يجعلها أكثر مرونة ودقة في تمثيل ظاهرة تلوث المياه مقارنة بالنماذج الخطية التقليدية. كما تؤكد النتائج أن XGBoost تُعد أداة قوية في تحليل البيانات البيئية المعقدة، وتسهم في تقديم فهم أعمق للعوامل المؤثرة في تلوث مياه نهر دجله.

4.1. الاستنتاجات

أظهرت نتائج التحليل أن كلا النموذجين Elastic Net و XGBoostقدما معاملات ذات دلالة إحصائية عالية، مما يدل على أهمية المتغيرات الكيميائية والفيزيائية في تفسير تلوث مياه نهر دجلة .ويُعد Elastic Net نموذجًا خطيًا منتظمًا مناسبًا للتفسير الإحصائي المباشر، بينما تتميز خوارزمية XGBoostبقدرتها العالية على نمذجة العلاقات غير الخطية والتفاعلية، الأمر الذي يجعلها أكثر كفاءة من الناحية التنبؤية .وتشير نتائج المقارنة إلى أن

Elastic Netيخدم الأهداف التفسيرية، في حين أن XGBoost تتفوق عند التركيز على دقة التنبؤ، مما يعزز الاعتماد على النموذجين معًا لتحقيق فهم شامل لظاهرة تلوث المياه.

أظهرت نتائج دراسة المحاكاة تفوق طريقة الشبكة المرنة (Elastic Net) على بقية طرائق الانحدار اللوجستي المنتظم، إذ حققت أدنى قيم لمعيار متوسط مربع الخطأ (MSE) عند جميع أحجام العينات، مما يدل على كفاءتها في التعامل مع البيانات ذات الأبعاد المختلفة ووجود الترابط بين المتغيرات التوضيحية. كما بيّنت نتائج المحاكاة الخاصة بخوارزميات الذكاء الاصطناعي تفوق خوارزمية XGBoost على خوارزمية التعزيز التدرّجي التقليدية، حيث حققت أقل قيم لمعيار (MSE) عبر مختلف أحجام العينات.

4.2 التوصيات

وفي التطبيق العملي على بيانات تلوث مياه نهر دجلة، أظهرت النتائج أن جميع المتغيرات البيئية المدروسة كانت ذات دلالة إحصائية عالية في كل من نموذج Elastic Net وخوارزميةXGBoost ، مما يؤكد أهمية هذه المتغيرات في تفسير مستوى تلوث المياه. كما أظهرت نتائج نموذج Elastic Netأن بعض المتغيرات كان لها تأثير موجب في مستوى التلوث، في حين كان لبعضها الآخر تأثير سالب، مع اتفاق عام في اتجاه التأثير عند استخدام خوارزميةXGBoost ، التي تميزت بقدرتها الأعلى على تمثيل العلاقات غير الخطية.

وبصورة عامة، بينت المقارنة أن نموذج Elastic Net أكثر ملاءمة للأغراض التفسيرية الإحصائية، في حين تتفوق خوارزمية XGBoost من الناحية التنبؤية، مما يؤكد أن الدمج بين الأساليب الإحصائية وتقنيات الذكاء الاصطناعي يوفر إطارًا تحليليًا متكاملًا يجمع بين التفسير والدقة التنبؤية .

4.2.1 للباحثين

يُوصى بتبني نهج مزدوج في دراسات تلوث المياه: استخدام Elastic Net لتحديد المتغيرات المؤثرة وتفسير العلاقات السببية، واستخدام XGBoost لبناء أنظمة التنبؤ عالية الدقة.

4.2.2 لوزارة البيئة العراقية

تطوير نظام رصد ذكي لمياه نهر دجلة يعتمد على خوارزمية XGBoost للتنبؤ المبكر بحالات التلوث، مع تكثيف الرقابة على المتغيرات الحرجة (TDS، NO_3، HCO_3، SO_4) التي أظهرت دلالة إحصائية عالية.

4.2.3 للدراسات المستقبلية

توسيع نطاق التطبيق على بيانات زمنية أطول لدراسة التباين الموسمي، وتغطية مناطق جغرافية إضافية (نهر الفرات، شط العرب)، مع زيادة حجم البيانات الحقيقية بدلاً من الاعتماد الكبير على بيانات المحاكاة.

4.2.4 على المستوى المؤسسي

بناء القدرات من خلال تدريب الكوادر الفنية في وزارة البيئة على استخدام الطرائق الجزائية وخوارزميات التعلم الآلي، وإتاحة البيانات والأكواد البرمجية للباحثين لتشجيع البحث التطبيقي في المجال البيئي.

تضارب المصالح
لا يوجد أي تضارب في المصالح.

الموافقة على النشر
"لقد قرأنا ووافقنا على النسخة النهائية من المخطوطة للنشر".

توافر البيانات والمواد
جميع البيانات قد تم تضمينها ضمن المخطوطة.

التمويل
لا يوجد اي جهة ممولة للبحث، سواء كانت شركة أو مؤسسة أو منظمة.

الشكر والتقدير
شكر لكل الأشخاص أو المؤسسات التي قدّمت لنا مساعدة علمية أو مادية في إنجاز الدراسة وكتابة المقال.

5. المراجع

[1] WHO (2022). *Guidelines for drinking-water quality*. Geneva: World Health Organization.
[2] UNEP (2023). *Global Environment Outlook*. Nairobi: United Nations Environment Programme.
[3] Khan, R., Zhang, Y. and Kumar, A. (2021). *Water quality prediction using machine learning techniques: A comparative study*. Journal of Hydrology, 603, pp. 1–12.

[4] Al-Jubouri, A.A. and Al-Khafaji, M.S. (2023). *Assessment of water quality of the Tigris River using statistical and environmental indicators*. Iraqi Journal of Science, 64(3), pp. 1123–1138.
[5] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. 5th ed. New York: Wiley.
[6] Tibshirani, R. (1996). *Regression shrinkage and selection via the LASSO*. Journal of the Royal Statistical Society: Series B, 58(1), pp. 267–288.
[7] Zou, H. and Hastie, T. (2005). *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B, 67(2), pp. 301–320.
[8] Friedman, J.H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), pp. 1189–1232.
[9] Chen, T. and Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
[10] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regression*. 3rd ed. New York: Wiley.
[11] Kisi, O. and Parmar, K.S. (2017). *Application of least square support vector machine and multivariate adaptive regression spline models in long-term river water quality prediction*. Journal of Hydrology, 534, pp. 104–112.
[12] Zhang, Y., Li, M. and Wang, J. (2018). *Water quality prediction using XGBoost algorithm*. Water Resources Management, 32(14), pp. 4681–4695.
[13] Algamal, Z.Y. (2018). *Penalized logistic regression with the adaptive LASSO for gene selection*. Communications in Statistics – Simulation and Computation, 47(2), pp. 347–360.
[14] Saha, S., Gupta, A. and Banerjee, S. (2019). *Penalized logistic regression for high-dimensional biomedical data*. Statistical Applications in Genetics and Molecular Biology, 18(4), pp. 1–14.
[15] Rahman, M.A., Hassani, A. and Khan, A. (2020). *Groundwater quality classification using machine learning algorithms*. Environmental Monitoring and Assessment, 192(6), pp. 1–18.
[16] Wang, L., Li, H. and Huang, J. (2021). *Nonconvex penalized logistic regression for disease classification*. Statistics in Medicine, 40(12), pp. 2845–2860.
[17] Li, X., Wang, Y. and Zhao, Z. (2022). *Comparison of penalized logistic regression and XGBoost in medical outcome prediction*. BMC Medical Research Methodology, 22(1), pp. 1–14.
[18] Al-Mukhtar, M., Al-Ansari, N. and Knutsson, S. (2023). *Application of machine learning techniques for surface water quality assessment*. Environmental Earth Sciences, 82(4), pp. 1–15.
[19] Zhao, Y., Liu, H. and Chen, X. (2024). *Penalized logistic regression for environmental data analysis*. Environmental Modelling & Software, 172, pp. 1–12.
[20] Harrell, F.E. (2010). *Regression Modeling Strategies*. New York: Springer.
[21] Yang, Y., Zou, H. and Bickel, P.J. (2015). *Adaptive elastic net for high-dimensional regression*. Journal of Statistical Planning and Inference, 156, pp. 18–32.
[22] Zou, H. (2006). *The adaptive LASSO and its oracle properties*. Journal of the American Statistical Association, 101(476), pp. 1418–1429.
[23] Hui, F.K.C., Warton, D.I. and Foster, S.D. (2015). *Multivariate logistic regression with adaptive LASSO penalty*. Journal of Computational and Graphical Statistics, 24(2), pp. 469–488.
[24] Huang, J., Ma, S. and Zhang, C.H. (2008). *Adaptive LASSO for sparse high-dimensional regression models*. Statistica Sinica, 18(4), pp. 1603–1618.
[25] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer.
[26] Fan, J. and Li, R. (2001). *Variable selection via nonconcave penalized likelihood and its oracle properties*. Journal of the American Statistical Association, 96(456), pp. 1348–1360.

Journal of Al-Turath University College (JATUC)
Open Access Journal
ISSN : 2074-5621
DOI: https://doi.org/10.63964/JATUC.43.2.2026.33
https://jatuc.uoturath.edu.iq/

Estimation of Penalized Logistic Regression Models Using Selected Statistical Methods and Deep Learning Techniques

Karrar Abdul Redha Ahmed[1], Sabah Munfi Redha[2]

[1,2] Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq.

Corresponding E-mail: karrar.abd2201@coadec.uobaghdad.edu.iq

Keywords

Water Pollution, Penalized Logistic Regression, LASSO, SCAD, Elastic Net, Binary Classification.

Abstract

This study aims to investigate and classify water pollution in the Groundwater in Iraq by employing a combination of penalized statistical methods and modern artificial intelligence techniques. The study is based on real environmental data provided by the Iraqi Ministry of Environment, in addition to simulated data generated using the Monte Carlo simulation approach. The research focuses on penalized logistic regression models, including LASSO, Adaptive LASSO, SCAD, and Elastic Net, alongside gradient boosting algorithms, particularly the Extreme Gradient Boosting (XGBoost) algorithm, in order to evaluate their interpretative and predictive performance when dealing with high-dimensional and highly correlated environmental data.

The Mean Squared Error (MSE) criterion was adopted to compare the performance of the proposed methods. Simulation results demonstrated that the Elastic Net method achieved the best performance among penalized logistic regression models, yielding the lowest MSE values across all sample sizes. Moreover, the results indicated that the XGBoost algorithm outperformed the traditional Gradient Boosting algorithm, showing superior predictive accuracy and stability. In the practical application to real water quality data from the Tigris River, all studied environmental variables were found to be statistically significant in both the Elastic Net model and the XGBoost algorithm. While the two approaches showed general agreement in the direction of variable effects, XGBoost exhibited a greater ability to capture nonlinear and complex relationships among variables.

The comparative analysis revealed that the Elastic Net model is more suitable for interpretative purposes and identifying the most influential environmental factors, whereas XGBoost provides higher predictive accuracy. These findings confirm that integrating penalized statistical models with artificial intelligence techniques offers a comprehensive and effective analytical framework for studying water pollution.

© 2026.This is an open access article under the CC by licenses http://creativecommons.org/licenses/by/4.0