УДК 004.85:662.63 DOI 10.31548/ energiya3(67).2023.100

АНАЛІЗ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВИХОДУ БІОГАЗУ

В. П. Лисенко, доктор технічних наук, професор

Т. І. Лендєл, кандидат технічних наук, доцент

С. Г. Павлов, аспірант

Національний університет біоресурсів і природокористування України

E-mail: lysenko@nubip.edu.ua, taraslendel@gmail.com, sergpavlov89@gmail.com

Анотація. *Нині ефективне управління виробництвом біогазу залишається складною задачею.*

Мета дослідження – провести аналіз алгоритмів машинного навчання для прогнозування виходу біогазу залежно від характеристики виробництва біогазу.

Нині відсутній необхідний набір даних, аналізуючи які можна отримати показники для оптимізації виробництва біогазу в нашій установці. У той же час випробування різноманітних алгоритмів оптимізації та прийняття рішення щодо найкращого займає, як показує досвід, немало часу.

Розглянуто застосування машинних алгоритмів для прогнозування виробництва біогазу шляхом використання існуючих методів прогнозування. За умови, що системи управління типовими виробництвами біогазу укомплектовані необхідними сприймаючими елементами, все одно залишається задача, пов'язана з обробкою та аналізом даних для прийняття найкращого рішення щодо забезпечення відповідних технологічних вимог. Причиною цьому є великий обсяг даних та складність взаємодії процесів, що є складовими виробництва. У такому контексті, машинне навчання може бути корисним інструментом для оптимізації виробництва біогазу.

Ключові слова: ***машинне навчання, вихід біогазу, автоматизоване керування,алгоритми керування, математична модель***

Актуальність. У зв'язку з ростом популярності екологічної енергетики виробництво біогазу стає все більш актуальним. Однак, ефективне управління виробництвом біогазу залишається складною задачею.

Аналіз останніх досліджень та публікацій. На процес виробництва біогазу впливає ряд причин:

- процес виробництва залежить від багатьох факторів, які складно контролювати;

- відсутність можливості фіксувати параметри на всіх етапах виробництва.

Сировину зі складу (рис.1) підготовлюють і завантажується у спеціальну ємність (метатенк), що забезпечує надійний захист біомаси від кисню. Під впливом спеціальних бактерій в анаеробному середовищі починає відбуватися ферментація (процес перетворення органічної сировини в біогаз). Біомасу необхідно ретельно перемішувати для рівномірного розподілу кислотності та температури всередині органічної маси. Внаслідок цих маніпуляцій виробляється біогаз. Отриманий газ збирається в газгольдері, звідки трубами доставляється споживачеві. Біодобрива, отримані після переробки вихідної сировини, можна додавати до ґрунту.

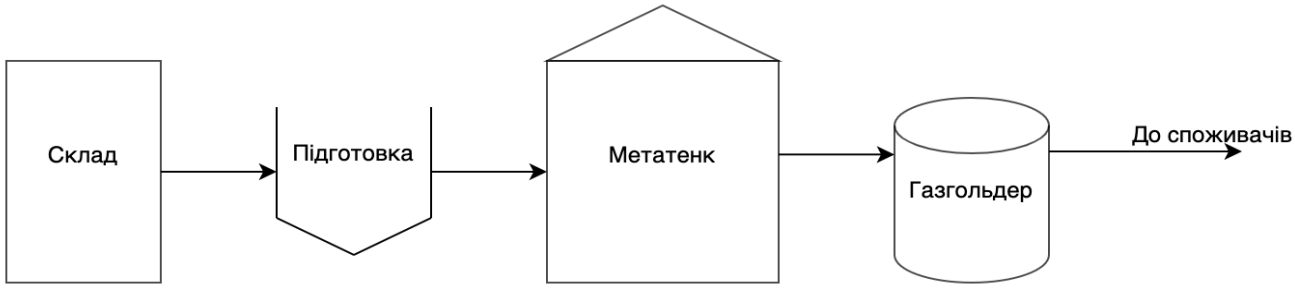

Рис. 1. Спрощена схема виробництва біогазу

Мета дослідження – провести аналіз алгоритмів машинного навчання для прогнозування виходу біогазу залежно від характеристики виробництва біогазу.

Матеріали та методи дослідження. Отримання достатньої кількості якісних даних є однією з найбільших проблем, що виникають при використанні алгоритмів машинного навчання для оптимізації виробництва біогазу. Збір та обробка даних може бути трудомістким та часозатратним процесом, а в якості джерела даних можуть використовуватись дані з різних сенсорів та пристроїв із різними розширеннями, що ускладнює їх обробку та аналіз.

Крім того, для успішної роботи алгоритмів машинного навчання необхідно мати достатньо велику вибірку даних, що є значною проблемою для всіх виробництв, оскільки зазначене потребує тривалих і спеціалізованих затрат. Також важливо враховувати, що дані, які були зібрані раніше, можуть не відображати поточної ситуації на виробництві, тому їх потрібно постійно накопичувати для

загального статистичного аналізу, що, без сумніву, позитивно вплине на результати оптимізації виробництва біогазу.

Нині відсутній необхідний набір даних, аналізуючи які ми б могли отримати показники для оптимізації виробництва біогазу в нашій установці. У той же час випробування різноманітних алгоритмів оптимізації та прийняття рішення щодо найкращого займає, як показує досвід, немало часу. Саме тому порівняльний аналіз машинних алгоритмів ми провели на тестовому масиві даних, що був опублікований у відкритих джерелах щодо енергоємності та складності процесу отримання біогазу [1].

Результати досліджень та їх обговорення. У нашій роботі ми використали сервіс платформи Amazon Web Services SageMaker Autopilot [2]. AWS SageMaker Autopilot - це сервіс машинного навчання, який дозволяє автоматизувати процес побудови моделей машинного навчання. Він використовує автоматичні методи для визначення оптимальних параметрів та архітектури моделі на основі вхідних даних. Робота сервісу розпочинається з автоматичного виконання EDA (Exploratory Data Analysis) - аналізу вхідних даних. Після цього виконується підготовка даних, включаючи зменшення розмірності, кодування категоріальних даних та видалення шумів. Далі SageMaker Autopilot використовує різні алгоритми машинного навчання, такі як глибокі нейронні мережі, ансамблевий мета-алгоритм (градієнтний бустінг), лінійні моделі тощо, для вибору оптимальної моделі та параметрів, значення яких використовується для керування процесом навчання. Після цього, кращі моделі порівнюються на основі використання метрик якості, таких як точність, втрати та інші.

Основна перевага SageMaker Autopilot полягає в тому, що він дозволяє прискорити та автоматизувати процес побудови моделей машинного навчання, зменшивши час, необхідний для визначення оптимальних параметрів та архітектури моделі. На кожному етапі роботи сервісу з даними є можливість вносити необхідні зміни для покращення результату.

Цей сервіс використовує різні методи машинного навчання для автоматизованої побудови та оптимізації моделей. Серед них можна виділити:

1) автоматичне підбір параметрів: автоматично налаштовує гіпер параметри моделі для досягнення максимальної точності;
2) стекінг моделей: автоматично комбінує результати кількох моделей для досягнення більшої точності;
3) автоматичний вибір алгоритмів: автоматично вибирає найкращі алгоритми для конкретної задачі або окремого етапу задачі;
4) глибинне навчання: використовує глибинні нейронні мережі для виявлення складних залежностей в даних;
5) автоматична обробка даних: автоматично обробляє і перетворює вхідні дані, щоб покращити точність моделі, може виконати необхідну очистку дублів або пропусків для зменшення викидів;
6) байєсівська оптимізація: використовує статистичні методи для ефективного підбору параметрів для керування процесом навчання моделі.

Ці методи дозволяють автоматизувати процес навчання моделей та підвищити точність результатів.

Спрощена схема послідовності роботи сервісу зображено на рис. 2

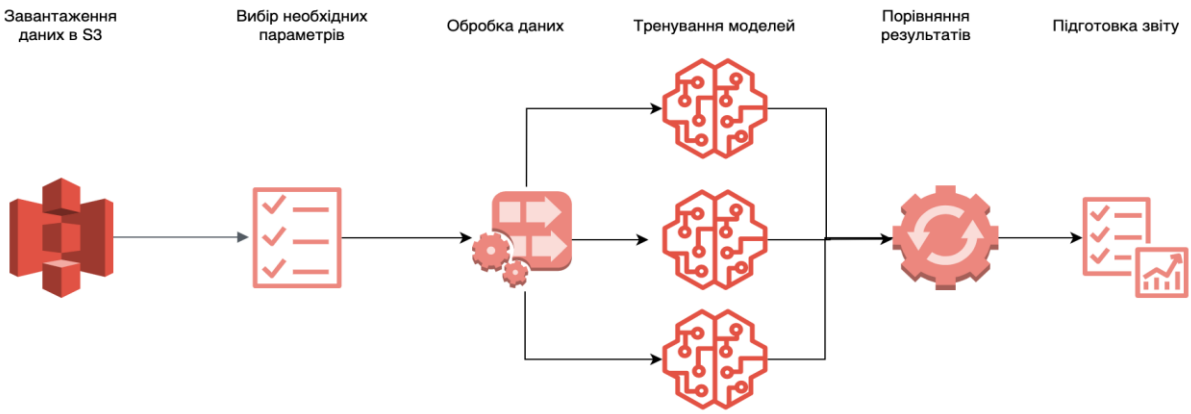

Рис. 2. Спрощена схема послідовності роботи SageMaker Autopilot

Для тесту ми обрали такі алгоритми LightGBM, CatBoost, XGBoost [3], Random Forest [4], Extra Trees [5], Linear Models [6], Neural Network built using fast.ai[7], Neural network built using PyTorch [8].

LightGBM, CatBoost, XGBoost, Random Forest і Extra Trees — це алгоритми на основі дерева рішень, які створюють ансамбль дерев рішень для прогнозування.

Лінійні моделі, з іншого боку, є типом алгоритму регресії, який функціонує шляхом підгонки лінійного рівняння до навчальних даних. Вони часто використовуються для таких завдань, як прогнозування тривалої вартості.

Нейронні мережі, створені на основі fast.ai і PyTorch, — це алгоритми глибокого навчання, які використовують багаторівневу нейронну мережу для вивчення складних шаблонів у даних. Зазвичай вони використовуються для таких завдань як розпізнавання зображень, обробка природної мови та розпізнавання мовлення.

З точки зору продуктивності, LightGBM, CatBoost і XGBoost часто використовуються в змаганнях з машинного навчання, і було встановлено, що вони мають високу точність прогнозування та швидкий час навчання. Random Forest і Extra Trees також широко використовуються і відомі своєю здатністю обробляти дані великої розмірності та дані з шумом.

Лінійні моделі прості та легкі для інтерпретації, але можуть не працювати так добре, як більш складні моделі для певних типів даних. Нейронні мережі можуть досягти високої точності у складних завданнях, але можуть бути обчислювально дорогими для навчання та можуть потребувати великої кількості даних.

Для порівняння моделей між собою ми будемо використовувати такі показники:

- MAE "Mean Absolute Error" (середня абсолютна помилка); це метрика оцінки точності моделі, яка вимірює різницю між прогнозованими значеннями моделі та фактичними значеннями цільової змінної; MAE обчислюється шляхом взяття середнього арифметичного з абсолютних значень різниць між прогнозованими та фактичними значеннями; значна величина MAE вказує на те, що модель дає великі помилки в прогнозуванні, тоді як мала величина MAE вказує на те, що модель добре прогнозує дані; MAE часто використовується в задачах регресії [9];

- MSE (Mean Squared Error) - це метрика оцінки точності моделі в машинному навчанні; вона вимірює середнє значення квадрата відхилень (різниць) між прогнозованими і справжніми значеннями вихідної змінної; це означає, що MSE показує, наскільки віддалені прогнозовані значення від справжніх значень; чим менше значення MSE, тим краще «працює» модель; використовують MSE, коли потрібно більш точно відображати великі відхилення між прогнозованими і справжніми значеннями [10];
- RMSE (англ. Root Mean Square Error) - це одна з найпоширеніших метрик оцінки точності моделей у машинному навчанні; вона представляє собою квадратний корінь з середньої квадратичної помилки (MSE), тобто вона вимірює середню відстань між прогнозованими значеннями моделі і фактичними значеннями; RMSE зазвичай використовують для порівняння точності різних моделей, де менший показник RMSE вказує на кращу точність моделі; RMSE вимірюється в тих же одиницях, що і вихідна змінна, тобто він дозволяє оцінити, на скільки одиниць в середньому прогнози моделі відрізняються від фактичних значень [9].

Inference latency в машинному навчанні означає час, необхідний для отримання результату передбачення алгоритмом після подачі вхідних даних на вхід моделі. Іншими словами, це час затримки, який виникає між подачею запиту на передбачення та отриманням результату. Цей показник є важливим параметром, який впливає на продуктивність моделі, особливо коли вона використовується в реальному часі. Чим менше часу затримки, тим швидше можна отримати результат передбачення та виконати потрібну дію на основі цього результату [11].

Дані для тесту мають такий вигляд як це подано на рис.3.

Data	Hora	tempSubst	tempSolo	HumidAmb	tempAmb	HumidGas	tempGas	decDeBiogas	volumeTotal

Рис.3. Дані для тесту (Data - дата виміру, Hora - час вимірювання, tempSubst - температура матеріалу (субстракту), HumidAmb - вологість повітря, tempAmb - температура повітря, HumidGas - вологість газу, tempGas - температура газу, decDeBiogas - кількість виробленого біогазу, volumeTotal - загальна сума виробленого біогазу

Параметром який ми шукаємо буде decDeBiogas. Для порівняння були вибрані алгоритми XGBoost, LightGBM, CatBoost, ExtraTree, RandomForest. Найкращим варіантом виявився алгоритм WeightedEnsemble [12]. Це алгоритм ансамблювання (об'єднання) в машинному навчанні, який комбінує декілька моделей, надаючи різні «ваги» кожній моделі залежно від її ефективності. «Ваги» для кожної моделі визначаються шляхом оптимізації валідаційної метрики на тренувальному наборі даних. Результати, що їх показав даний алгоритм подані в таблиці.

Результати випробування алгоритму WeightedEnsemble

Показник	Прогнозоване значення	Стандартне відхилення,
MAE, $м^3$	0.000010049830731162501	5.9057183097363664e-8
MSE, $м^3$	1.0483145648558096e-10	2.1448057971034466e-12
RMSE, $м^3$	0.000010238723381632153	1.0508570575565888e-7

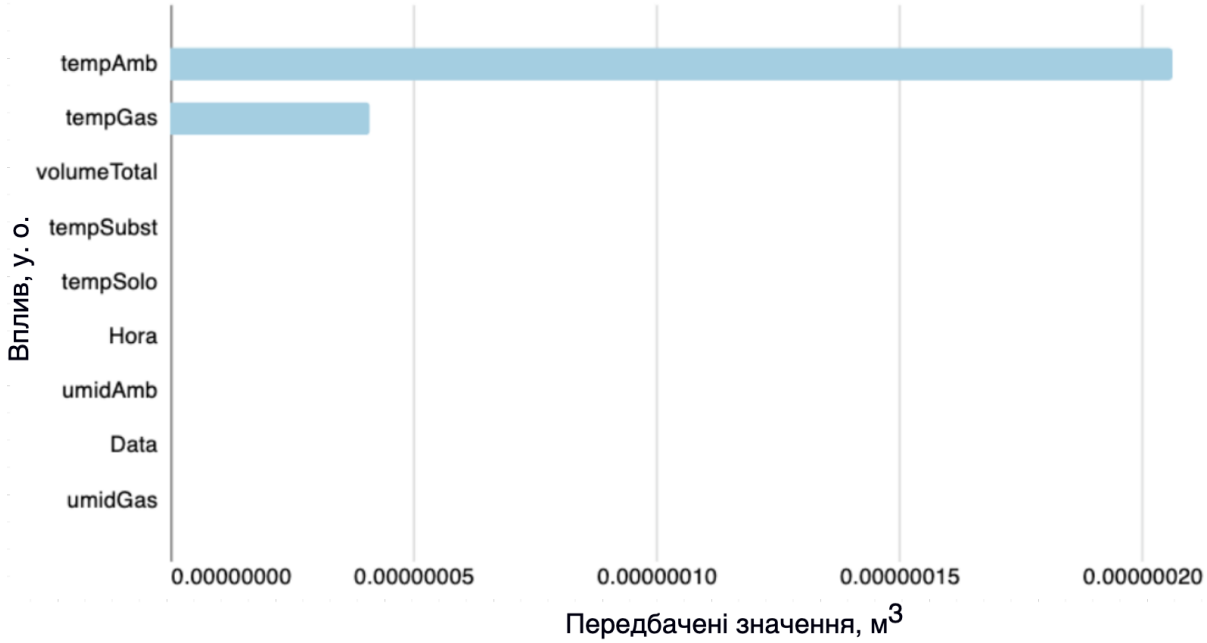

Рис. 4. Вплив параметрів на точність моделювання

На рис. 4 зображено графік впливу параметрів на точність вимірювання (Feature Impact plot) - це візуалізація, яка показує вплив кожного параметру (feature) на прогнозну змінну. Цей графік допомагає визначити, які параметри є найбільш важливими для моделі, що може бути корисно для подальшого вдосконалення моделі. Чим вища значимість параметру, тим більший вплив вона має на прогнозну

змінну і тим більше буде її відображення на графіку. Відповідно нашому графіку найбільшого впливу завдала температура повітря та газу. При подальшій роботі над збором параметрів нашої установки ми будемо приділяти найбільше уваги цим параметрам. Також будемо працювати над розширенням кількості параметрів і перевіркою їх впливу на модель.

Стандартизований залишок - це значення, що отримується шляхом ділення залишку на стандартне відхилення моделі. Це дає змогу отримати значення та порівняти його з іншими значеннями відносно їх розкиду. Стандартизовані залишки використовуються для виявлення аномальних даних, невідповідностей між моделлю та даними, а також для оцінки роботи моделі. Графік стандартизованих залишків нам дозволяє візуалізувати розподіл залишків та виявити потенційні проблеми з моделлю [13].

Зазвичай, залишки повинні розподілятися нормально з нульовим середнім значенням і однаковим стандартним відхиленням для всіх значеннях прогнозування.

Графік стандартизованого залишку дозволяє визначити систематичні зміщення у прогнозах, такі як недооцінка або переоцінка результатів.

У нашому випадку є кілька відхилень, тому необхідно буде детально вивчати, які дані на це впливають.

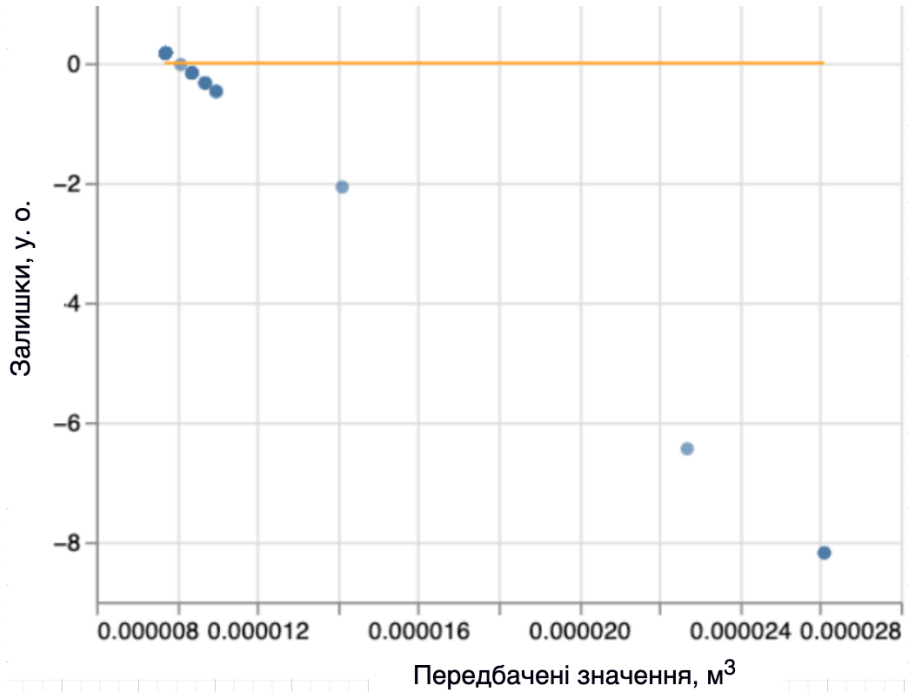

Рис. 5. Графік стандартизованого залишку

Гістограма залишкових значень (Residual histogram) показує кількість спостережень з відповідним значенням залишкових значень (residuals) між прогнозованими та фактичними значеннями в наборі даних. Залишкові значення - це різниця між фактичним значенням та прогнозованим значенням.

Гістограма залишкових значень показує частоту кожного інтервалу значень у порівнянні із залишковими значеннями. Чим ближче значення залишків до нуля, тим краще модель прогнозує дані. Гістограма залишків дозволяє оцінити, наскільки рівномірно розподілені значення залишків протягом всього дослідження. [14]. На іншій гістограмі спостерігаються відхилення, які необхідно нормалізувати шляхом збільшення та покращення якості даних.

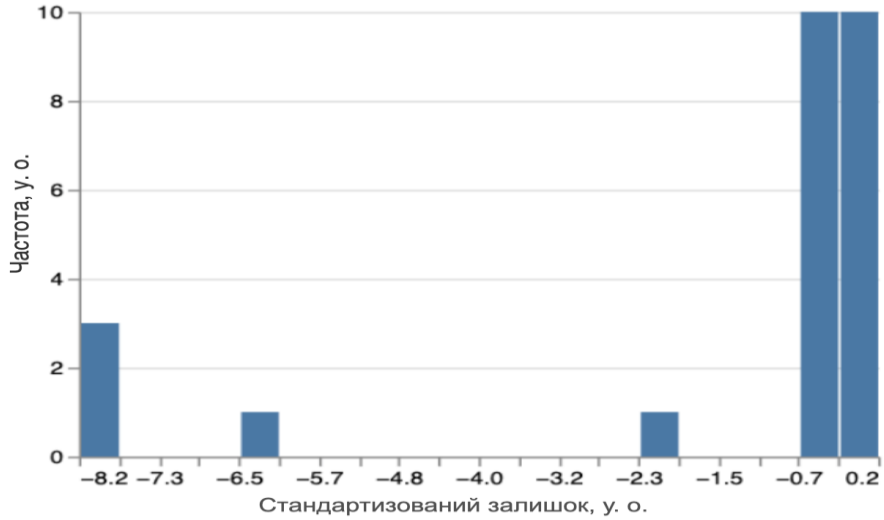

Рис. 6. Гістограма залишкових значень (Residual histogram)

За результатами аналізу цього експерименту можна зробити висновок, що тестових даних, які були використані, недостатньо для прийняття правильного рішення щодо покращення виробництва біогазу. Але в експерименті були випробувані інструменти, які в подальшому можуть прискорити виявленням і уточнення зв'язків між параметрами біогазової установки.

У цілому алгоритм WeightedEnsemble показав себе перспективним для подальшого вивчення і застосування: його використання виявилось досить ефективним для покращення точності прогнозування порівняно з використанням окремих моделей. Особливо позитивно цей алгоритм проявляє властивості, коли існує суттєва різниця в ефективності між моделями. У WeightedEnsemble ваги, з

якими враховуються прогнози кожної моделі, визначаються автоматично в процесі навчання, що дозволяє адаптувати ваги до конкретного набору даних і моделей. Однією з головних переваг WeightedEnsemble є його гнучкість. Вказаний алгоритм можна апробовувати з будь-якими типами моделей, включаючи лінійні моделі, моделі на основі дерев рішень, нейронні мережі та інші. Окрім того, як уже зазначалось, його застосування дозволяє автоматично визначати оптимальні ваги для кожної моделі, що зменшує необхідність вручну налаштовувати параметри ансамблю.

Процес побудови моделі WeightedEnsemble можна розділити на кілька етапів:

- першим етапом є навчання кожної моделі на тренувальних даних;
- наступним етапом є використання цих моделей для прогнозування на валідаційному наборі даних;
- потім прогнози кожної моделі комбінуються із допомогою ваг, що визначаються автоматично;
- завершальним етапом є використання зважених прогнозів моделей для створення завершальних прогнозів.

Висновки та перспективи.

Проведений аналіз застосування машинних алгоритмів для прогнозування виробництва біогазу шляхом використання однієї і тієї ж вибірки, як характеристики виробництва біогазу, показав суттєву перевагу алгоритму WeightedEnsemble перед іншими. Повнота вибірки в значній мірі впливає як на результат прогнозування, так і на результат оптимізації виробництва.

References

1. Kaggle.com. JM Biogas production experiment analysis jumpstart. Available at: https://www.kaggle.com/code/ivandaudt/jm-biogas-production-experiment-analysis-jumpstart/notebook.

2. Amazon SageMaker Autopilot. Available at: [https://aws.amazon.com/sagemaker/autopilot/?nc1=h_ls&sagemaker-data-wrangler-whats-new.sort-by=item.additionalFields.postDateTime&sagemaker-data-wrangler-whats-new.sort-order=desc].

3. Essam Al Daoud. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataseю Available at: [https://publications.waset.org/10009954/comparison-between-xgboost-lightgbm-and-catboost-using-a-home-credit-dataset].

4. Muhammad Waseem Ahmad, Jonathan Reynolds, Yacine Rezgui. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. Available at: [https://www.sciencedirect.com/science/article/pii/S0959652618325551].

5. Pierre Geurts, Damien Ernst, Louis Wehenkel. Extremely randomized trees. Available at: [https://link.springer.com/article/10.1007/s10994-006-6226-1].

6. J. A. Nelder, R. W. M. Wedderburn. Generalized Linear Models. Available at: [https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2344614].

7. Rita Yi Man Li, Herru Ching Yu Li, Beiqi Tang, WaiCheung Au. Fast AI classification for analyzing construction accidents claims. Available at[https://dl.acm.org/doi/abs/10.1145/3407703.3407705].

8. Sagar Imambi, Kolla Bhanu Prakash, G. R. Kanagachidambaresan. PyTorch. Available at: [https://link.springer.com/chapter/10.1007/978-3-030-57077-4_10].

9. Weijie Wang, Yanmin Lu . Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. Available at:[https://iopscience.iop.org/article/10.1088/1757-899X/324/1/012049/meta].

10. Kalyan Das, Jiming Jiang, J. N. K. Rao. Mean squared error of empirical predictor. Available at: [https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-2/Mean-squared-error-of-empirical-predictor/10.1214/009053604000000201.full].

11. Alon Brutzkus, Ran Gilad-Bachrach, Oren Elisha. Low Latency Privacy Preserving Inference. Available at [http://proceedings.mlr.press/v97/brutzkus19a].

12. Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, Alex Ksikes. Ensemble selection from libraries of models. Available at: [https://dl.acm.org/doi/abs/10.1145/1015330.1015432].

13. Robert E. Weiss, Carlos G. Lazaro. Residual plots for repeated measures. Available at: [https://doi.org/10.1002/sim.4780110110].

14. Zhila Esna Ashari, Nairanjana Dasgupta, Kelly A Brayton, Shira L Broschat. An optimal set of features for predicting type IV secretion system effector proteins for a subset of species based on a multi-level feature selection approach. Available at: [http://dx.doi.org/10.1371/journal.pone.0197041].

ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR BIOGAS YIELD PREDICTION

V. Lysenko, T. Lendyel, S. Pavlov

Abstract. *Currently, effective management of biogas production remains a difficult task.*

The purpose of the research is to analyze machine learning algorithms for predicting biogas output depending on the characteristics of biogas production.

Currently, there is no necessary set of data, analyzing which indicators can be obtained to optimize biogas production in our installation. At the same time, testing various optimization algorithms and deciding on the best takes a lot of time, as experience shows.

The application of machine algorithms for forecasting biogas production by using existing forecasting methods is considered. Provided that the control systems of typical

biogas productions are equipped with the necessary sensing elements, there still remains the task of processing and analyzing data to make the best decision to meet the relevant technological requirements. The reason for this is the large volume of data and the complexity of the interaction of processes that are components of production. In this context, machine learning can be a useful tool for optimizing biogas production.

Key words: ***machine learning, biogas output, automated control, control algorithms, mathematical model***