УДК: 004.67 DOI: **10.33099/2311-7249/2025-54-3-130-140**

КОВАЛЬ Ігор Михайлович,
Луцький національний технічний університет, Луцьк, Україна,
https://orcid.org/0009-0001-2083-1747

ГОЛОВНЯ Сергій Анатолійович,
Луцький національний технічний університет, Луцьк, Україна,
https://orcid.org/0009-0005-2997-9202

ЕКСПЕРИМЕНТАЛЬНЕ ОЦІНЮВАННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВИРОБНИЧИХ ПОКАЗНИКІВ ОБОРОННО-ПРОМИСЛОВОГО КОМПЛЕКСУ В СЕРЕДОВИЩІ ORANGE DATA MINING

У статті розглянуто прогнозування виробничих потужностей оборонно-промислового комплексу із застосуванням алгоритмів машинного навчання як інструментів інтелектуального аналізу даних у середовищі Orange Data Mining.

Мета. *Визначення доцільності використання, експериментальна перевірка та оцінювання точності алгоритмів наївного баєсівського класифікатора, логістичної регресії та методу випадкових лісів для прогнозування виробничих потужностей оборонно-промислового комплексу із використанням машинного навчання у середовищі Orange Data Mining.*

Методи дослідження. *Застосовано методи системного аналізу, інтелектуального аналізу даних та крос-валідації у середовищі Orange Data Mining. Використано алгоритми наївного баєсівського класифікатора, логістичної регресії та методу випадкових лісів. Оцінювання точності здійснено за метриками AUC, точність, F1 та коефіцієнт кореляції Метьюса.*

Отримані результати дослідження. *Сформовано експериментальний набір даних, що враховує вплив зовнішніх та внутрішніх факторів на виробничі процеси. Проведено порівняння алгоритмів машинного навчання, результати якого свідчать про суттєву перевагу алгоритму випадкових лісів, який досяг найвищих значень точності та збалансованості класифікації. Порівняння проводилося за допомогою крос-валідації та стандартних метрик точності. Матриці помилок засвідчили обмеження наївного баєсівського класифікатора та логістичної регресії при дисбалансі даних.*

Елементи наукової новизни. *Вперше для цієї предметної області показано відмінності у роботі алгоритмів наївного баєсівського класифікатора, логістичної регресії та випадкових лісів в умовах дисбалансу даних, що дало змогу визначити найбільш ефективний алгоритм прогнозування виробничих потужностей.*

Теоретичне та практичне значення викладеного у статті*. Теоретичне значення проведеного дослідження полягає у розширенні методологічних основ застосування алгоритмів машинного навчання для прогнозування виробничих процесів з урахуванням впливу зовнішніх та внутрішніх факторів. У роботі уточнено можливості використання інтегральних метрик оцінювання моделей, що підвищує достовірність результатів у випадках дисбалансу даних. Практичне значення отриманих результатів визначається можливістю використання алгоритму випадкових лісів як базового інструменту для прогнозування виробничих потужностей підприємств оборонно-промислового комплексу. Запропонований підхід дає змогу підвищити обґрунтованість управлінських рішень у сфері планування виробництва, оптимізації використання ресурсів та забезпечення стійкості оборонно-промислових процесів у складних умовах.*

Ключові слова: *машинне навчання, Naive Bayes classifier, Logistic Regression, Random Forest, Orange Data Mining, дисбаланс даних, оборонно-промисловий комплекс*.

Вступ

Сучасні умови ведення бойових дій висувають особливі вимоги до надійності та гнучкості оборонно-промислового комплексу. Ефективне планування та прогнозування виробничих потужностей підприємств оборонної сфери має вирішальне значення для забезпечення сектору безпеки і оборони держави потрібними засобами в умовах інтенсивного використання ресурсів та обмеженого часу. Традиційні методи планування часто не враховують комплекс впливових факторів, таких як перебої в електропостачанні, години повітряних тривог або рівень доступності персоналу, що ускладнює формування достовірних прогнозів. Останніми роками в наукових дослідженнях активно використовуються методи машинного навчання та інтелектуального аналізу даних для виконання завдань прогнозування у промислових і військових системах. Сьогодні активно використовуються статистичні моделі, нейронні

мережі, ансамблеві методи та алгоритми класифікації для підвищення точності прогнозів у різних сферах. Разом з тим, питання апробації класичних алгоритмів машинного навчання на специфічних даних оборонно-промислового комплексу (далі – ОПК) України досліджені недостатньо. Таким чином, актуальним є завдання перевірки ефективності різних алгоритмів машинного навчання для прогнозування виробничих потужностей з урахуванням впливу факторів середовища та дисбалансу даних. Це дасть змогу визначити доцільні підходи до автоматизації процесів прогнозування у сфері оборонної промисловості, вивести універсальний до застосування алгоритм та підвищити обґрунтованість управлінських рішень.

Постановка проблеми. Прогнозування виробничих потужностей оборонно-промислового комплексу пов'язане з обробкою даних, що характеризуються високим рівнем неоднорідності та дисбалансом класів. У вибірках переважають записи з високими показниками виробництва, тоді як прикладів з низькими значеннями критично мало. Це ускладнює використання стандартних статистичних методів і призводить до значних перекосів у результатах навчання моделей. Технічними чинниками, що впливають на точність прогнозування, є варіативність даних про тривалість відключень електроенергії, кількість годин повітряних тривог, кількість працівників та дефектності продукції. В таких умовах частина алгоритмів машинного навчання, наприклад наївний баєсівський класифікатор (Naive Bayes) демонструє високу чутливість до дисбалансу даних і формує прогнози зі значною кількістю хибних віднесень. Логістична регресія (Logistic Regression) обмежена у відображенні нелінійних залежностей, що знижує її придатність для моделювання складних виробничих процесів. Додатковою проблемою є неможливість використання окремих метрик як універсальних показників якості: при значному дисбалансі даних точність (accuracy) не відображає реальної ефективності моделі, а показники AUC (Area Under the Curve) – площа під кривою робочої характеристики приймача та F1 (**F1-score)** – *зважене гармонійне середнє між точністю і повнотою* суттєво змінюються залежно від структури вибірки. Це потребує застосування комплексного підходу до оцінювання моделей, з урахуванням коефіцієнта кореляції Метьюса та аналізом матриць помилок. Таким чином, існує потреба у дослідженні та порівнянні ефективності різних алгоритмів машинного навчання в умовах дисбалансованих виробничих даних оборонно-промислового комплексу, що дасть змогу визначити найбільш доцільні методи прогнозування для практичного використання.

Аналіз останніх досліджень і публікацій. Методи прогнозування виробничих процесів за допомогою алгоритмів машинного навчання активно досліджуються у працях як українських, так і зарубіжних авторів. Так, у роботі [1] розглядаються фундаментальні алгоритми класифікації та кластеризації, які заклали основу для сучасних методів аналізу даних. Науковці у [2] дослідили проблематику оцінювання якості кластеризації та вплив нерівномірності розподілу даних на результати моделювання. Значний внесок у розвиток машинного навчання продемонстровано у роботі [3], де описано можливості та обмеження логістичної регресії, зокрема її недостатню ефективність при моделюванні складних нелінійних залежностей. Варто зазначити, що наївний баєсівський класифікатор може працювати навіть при порушенні припущення про незалежність ознак [4], але водночас є чутливим до дисбалансу даних і потребує ретельної підготовки вибірки.

У сучасних працях акцент робиться на використанні ансамблевих методів. Так, метод Random Forest (випадкових лісів), довів високу ефективність у задачах прогнозування [5] завдяки стійкості до зашумлених і дисбалансованих даних. Автори роботи [6] проаналізували внутрішні показники валідації моделей та показали переваги Random Forest у промислових застосуваннях. У класичному підручнику [7] систематизовано основні принципи машинного навчання та продемонстровано їх використання у прикладних задачах прогнозування. Водночас питання використання алгоритмів Naive Bayes (Наївний Байєс), Logistic Regression (Логістична регресія) та згаданий вище Random Forest саме для прогнозування виробничих потужностей оборонно-промислового комплексу залишаються малодослідженими. Особливу увагу необхідно приділити комплексному аналізу метрик у випадках дисбалансованих даних, зокрема використанню коефіцієнта кореляції Метьюса та інтерпретації матриць помилок, що є ключовим для об'єктивного оцінювання моделей. Враховуючи вищезазначене, актуальним науковим завданням є проведення порівняльного експериментального аналізу класичних алгоритмів машинного навчання з використанням інструменту візуального програмування (середовища) Orange із застосуванням методів Data Mining (добування даних) для прогнозування виробничих потужностей оборонно-промислового комплексу та визначення найбільш ефективного методу в умовах дисбалансованих даних.

Метою статті є розроблення та експериментальна перевірка автоматизованих алгоритмів прогнозування виробничих потужностей оборонно-промислового комплексу із застосуванням машинного навчання у середовищі Orange Data Mining, а також оцінювання точності та доцільності використання різних моделей для вирішення задачі класифікації рівнів виробництва за комплексом техніко-операційних факторів.

Виклад основного матеріалу дослідження

Зважаючи на дію правового режиму воєнного стану та у зв'язку з високою чутливістю інформації, що стосується ОПК для побудови експериментальної бази даних, що використовується у дослідженні методів прогнозування виробничих потужностей підприємств, було застосовано методику «Відкриті джерела інформації» (Open Source Intelligence (OSINT)). Така методика використовується для прийому збору,

аналізу та систематизації інформації з відкритих, загальнодоступних джерел, таких як соціальні мережі, публічні бази даних, медіа, веб-сайти та інші ресурси, доступні без порушення закону.

Під час досліджень було застосовано комбінацію реальних статистичних даних із відкритих джерел та змодельованих значень, що відображають реальні тенденції. Зокрема, у процесі формування змінних використовувалися такі джерела інформації:

1. Години повітряних тривог у регіонах України. Відповідні статистичні дані були взяті з відкритих аналітичних ресурсів, зокрема «Єдиний портал повітряних тривог» (alerts.in.ua) та дослідження аналітичного центру Texty.org.ua, які ведуть систематичний підрахунок тривалості тривог по регіонам [8]. Це дало змогу адекватно оцінити навантаження на виробничі процеси в умовах воєнних загроз.

2. Кількість ракетних і артилерійських атак по регіонах [9]. Для показника «Кількість атак в регіоні» (regional_attack_count) було використано узагальнені статистичні звіти Генерального штабу Збройних Сил України, а також дані міжнародних дослідницьких організацій, наприклад «Інститут вивчення війни» (Institute for the Study of War (ISW)), який щоденно публікує інформацію про масштаби атак.

3. Дані про енергетичні відключення. Показник «Години відключення електроенергії» (blackout_hours) формувався на основі офіційних повідомлень Національної енергетичної компанії «Укренерго» та профільних публікацій у відкритій пресі [10]. Найбільші значення цього параметра спостерігалися в осінньо-зимовий період 2022/23 та 2023/24 років, що узгоджується з реальними піками атак на енергетичну інфраструктуру України.

4. Виробничі показники «Кількість виробленої продукції» (output_qty), «Відсоток дефектних виробів» (defect_rate_pct), «Відсоток переробок» (rework_rate_pct) були змодельовані з урахуванням доступних оцінок з відкритих джерел. Зокрема, у публікаціях Міністерства оборони України та окремих інтерв'ю представників уряду неодноразово наголошувалося, що виробництво безпілотних літальних апаратів у 2023–2024 роках зросло у десятки разів порівняно з початком повномасштабного вторгнення [11]. Це було враховано при формуванні тренду зростання виробничих потужностей у базі даних. Отже, створена експериментальна база даних (далі – БД) є гібридною, оскільки враховує зовнішні та внутрішні фактори. Вона поєднує відкриті статистичні відомості з офіційних джерел та аналітичних платформ із математичним моделюванням. База даних містить часову ознаку, ідентифікатор підприємства, регіон, а також низку факторів, а саме: кількість годин повітряних тривог; тривалість відключень електроенергії; обсяг виробленої продукції; відсоток дефектності; рівень переробки; кількість атак у регіоні; відсоток доступності персоналу.

Для формування експериментальної БД, що поєднує реальні показники з відкритих джерел та змодельовані значення, використовувався інструмент Python 3.11 з бібліотеками Pandas та NumPy. Саме ці інструменти дали змогу згенерувати часовий ряд даних за період з червня 2022 року по червень 2025 року, відобразивши тренди розвитку виробництва безпілотних літальних апаратів у різних регіонах України.

Алгоритм генерації враховував сезонність (наприклад, пікові відключення електроенергії взимку), регіональні відмінності (різна інтенсивність атак і тривалості повітряних тривог), а також загальну тенденцію зростання виробництва дронів відповідно до офіційних заяв і прогнозів. Таким чином, сформована база даних відображає реалістичні умови функціонування оборонно-промислових підприємств у воєнний час та може бути використана для навчання і тестування алгоритмів машинного навчання в середовищі Orange Data Mining.

Цільовою змінною обрано категорію «Вихід» (output_category) виробництва, що має градацію Low (низька), Medium (середня), High (висока), та утворену на основі обсягів випуску продукції. Слід зазначити, що якість вхідних даних прямопропорційно впливає на точність роботи алгоритмів, та машинного навчання (зокрема передбачення). Перед практичним застосуванням будь якого з алгоритмів рекомендовано проводити попередню обробку даних як числових, так і текстових.

Експерименти виконано у середовищі Orange Data Mining, що забезпечує візуальне моделювання процесів машинного навчання. Для обробки даних використано стандартні графічні модулі, що розміщується в робочому просторі програми і призначені для швидкого доступу до інформації або послуги (віджети):

File (файл) – для імпорту набору даних;

Rank (ранг) – для оцінювання важливості ознак;

Test & Score (тест і оцінка) – для порівняння алгоритмів;

Confusion Matrix (матриця плутанини) – для відображення співвідношення між фактичними та передбаченими класами;

ROC Analysis (аналіз ROC) – для побудови ROC (Receiver Operating Characteristic curve – крива робочих характеристик приймача) кривої та обчислення площі під кривою;

Box Plot (бокс-діаграма) – для порівняння розподілу значень метрик різних алгоритмів та виявлення відхилень у результатах;

Scatter Plot (розсіяний графік) – для візуалізації результатів.

Для сприйняття інформації було прийнято рішення проводити експеримент у три етапи:

обробка та візуалізація вхідних даних;

підключення алгоритмів і порівняння моделей;

візуалізація отриманих результатів.

Обробка та візуалізація вхідних даних починається з підключення віджету *File* та завантаження БД у форматі .csv. Після того як дані успішно завантажено та оброблено варто підключати вихідний зв'язок з віджетами *Data Table* (таблиця даних) та *Data Info* (інформація про дані) (рис.1).

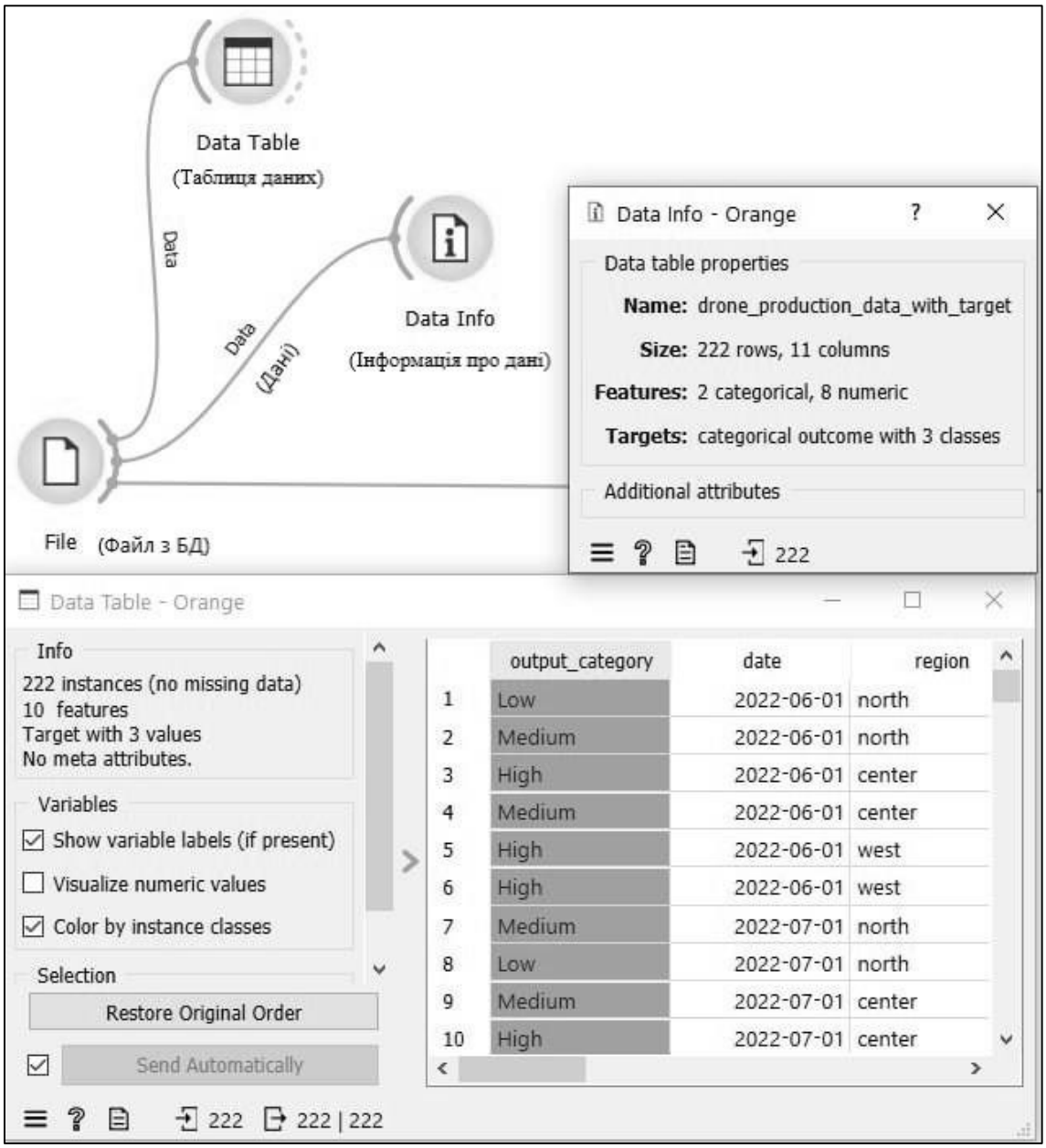

Рисунок 1 – Скріншот зв'язку між віджетами *File*, *Data Table* та *Data Info ma* огляд їх специфікацій

У розгорнотому вигляді віджети допомагають оцінити правильність розташування колонок, відповідність колонок і стовпців, а також кількість змінних та їхній тип для подальшої обробки. На рис.1 представлено віджети *Data Table* та *Data Info,* які використовуються для первинного контролю та аналізу вхідних даних. Поле Target with 3 values (цільова змінна з 3 значеннями) вказує на наявність трьох класів цільової категорії, яку потрібно передбачити. Опція Show variable labels (показати мітки змінних) відображає назви змінних, якщо вони задані у файлі. Опція Visualize numeric values (візуалізувати числові значення) дає змогу відображати числові дані у вигляді графічних індикаторів. Розділ Selection – Select full rows (вибір – виділити повні рядки) дає змогу вручну або автоматично обирати дані для подальшої обробки.

У правій частині зображено віджет *Data Info*, що подає короткі метадані щодо набору даних:

Name (назва) – ім'я завантаженого файлу;

Size (розмір) – кількість рядків і стовпців у таблиці;

Features (ознаки) – кількість категоріальних та числових змінних;

Targets (цілі) – характеристика вихідної змінної;

Additional attributes (додаткові атрибути) – кнопка для перегляду додаткових параметрів.

Наступним кроком було використання віджета *Rank*, який дає змогу оцінити інформативність змінних (рис. 2). Його застосування дало змогу визначити фактори, що найбільше впливають на класифікацію показників date, output_qty та defect_rate_pct. Водночас отримані дані з віджету інтерпретуються із застереженням: змінна date відображає часовий тренд, output_qty безпосередньо пов'язаний із формуванням цільової категорії, а defect_rate_pct характеризує якість продукції, а не виробничу потужність як таку. Тому попри високі рейтингові значення ці показники не розглядаються як ключові драйвери процесів, тоді як

технічні фактори (blackout_hours, air_alert_hours, staff_availability_pct) залишаються важливими для практичного аналізу.

На рис. 2 наведено результати роботи віджета Rank, що використовується для визначення вагомості (значущості) кожної змінної у формуванні прогнозу цільового показника. Gain Ratio (коефіцієнт приросту інформації) та Gini (індекс Джині) – стандартні критерії оцінювання інформативності змінних у машинному навчанні. Gain Ratio – показує, наскільки певна ознака зменшує невизначеність (ентропію) цільової змінної. Чим більше значення Gain Ratio, тим вагоміше змінна впливає на результат моделі. **Gini** – характеризує ступінь «чистоти» розподілу даних після поділу за певною ознакою. Менше значення Gini свідчить про більш точне розділення класів.

З огляду на змінну date (календарна дата) має найвище значення Gain Ratio (0.103), що свідчить про сильний вплив часових факторів на зміну виробничих показників.

		#	Gain ratio	Gini
1	T date		0.103	0.046
2	N output_qty		0.101	0.045
3	N defect_rate_pct		0.077	0.033
4	C region	3	0.050	0.012
5	N air_alert_hours		0.048	0.014
6	N regional_attack_count		0.036	0.010
7	C plant_id	6	0.031	0.013
8	N rework_rate_pct		0.023	0.005

Рисунок 2 – Результати оцінювання важливості змінних у віджеті Rank

На другому етапі дослідження до підготовлених даних було застосовано три класичні алгоритми: Naive Bayes classifier (наївний баєсівський класифікатор), Logistic Regression (логістична регресія) та Random Forest (випадковий ліс) (рис. 3). Вибір саме цих алгоритмів пояснюється їх поширеним використанням у розв'язанні задач класифікації та прогнозування, різною природою (ймовірнісна, статистична та ансамблева) і можливістю здійснити порівняльний аналіз точності.

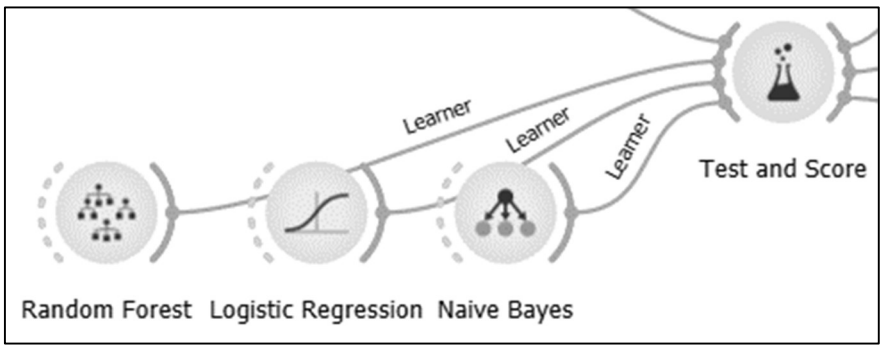

Рисунок 3 – Схема застосування віджета *Test & Score* у середовищі Orange Data Mining

Naive Bayes – це класифікаційний алгоритм, який спирається на ймовірнісний підхід [12] і теорему Байєса. Основна гіпотеза: усі ознаки (фактори) є незалежними одна від одної за умови відомого класу – це значно спрощує обчислення та дає змогу оцінювати параметри (апріорні ймовірності та умовні ймовірності ознак) простою підстановкою частот чи розподілів. Logistic Regression – це класифікаційний алгоритм, який належить до класу статистичних методів і ґрунтується на логістичній функції [13]. Його ідея полягає у моделюванні ймовірності належності об'єкта до певного класу як функції від лінійної комбінації вхідних ознак. На відміну від лінійної регресії, логістична регресія застосовує сигмоїдальне перетворення, що відображає вихід у діапазоні від 0 до 1, інтерпретуючи його як ймовірність. Основна гіпотеза полягає у припущенні, що лог-odds (логарифм відношення шансів) залежить від незалежних змінних лінійно. Random Forest – це ансамблевий алгоритм машинного навчання, який поєднує велику кількість дерев рішень для підвищення точності прогнозування [14]. Основна ідея полягає у побудові множини дерев на різних випадкових підвибірках даних і випадкових підмножинах ознак, після чого результати голосування об'єднуються у фінальне рішення. Такий підхід знижує ризик перенавчання, властивий окремим деревам рішень, і забезпечує високу стійкість до шуму та дисбалансу даних. Для порівняння точності роботи

алгоритмів застосовано віджет *Test & Score* (рис. 3), який забезпечує проведення стратифікованої крос-валідації [15] та визначення ключових метрик ефективності моделей. Використання цього інструменту дає змогу здійснити всебічне оцінювання якості класифікації даних та уникнути випадкових спотворень результатів, що можуть виникати при розподілі даних на тренувальні та тестові підмножини.

На основі підключених алгоритмів машинного навчання (Naive Bayes, Logistic Regression, Random Forest) було виконано оцінювання їхньої ефективності за допомогою віджета *Test & Score*. Цей інструмент дає змогу здійснити стратифіковану крос-валідацію та обчислити ключові метрики [16] якості класифікації: AUC, точність класифікації (CA), F1-показник, Precision (точність), Recall (відкликання) та коефіцієнт кореляції Метьюса (MCC). У таблиці 1 наведено середні значення метрик для трьох алгоритмів (Random Forest, Naive Bayes, Logistic Regression), а також наведено ймовірності статистично значущих відмінностей між ними за критерієм AUC.

Таблиця 1

Результати оцінювання алгоритмів у віджеті *Test & Score*

Алгоритм	AUC	CA	F1	Precision	Recall	MCC
Random Forest	0.998	0.991	0.987	0.983	0.991	0.946
Naive Bayes	0.907	0.014	0.008	0.041	0.014	0.071
Logistic Regression	0.949	0.910	0.867	0.828	0.910	0.000

У віджеті *Test & Score* для алгоритму Random Forest зафіксовано значення AUC – 0.998, що свідчить про майже ідеальне відокремлення класів. Для Logistic Regression відповідне значення становило 0.949, тоді як для Naive Bayes – 0.907. Аналогічна закономірність спостерігається і за метрикою точності класифікації (CA): для Random Forest показник досягнув 0.991, для Logistic Regression — 0.910, тоді як для Naive Bayes він був значно нижчим і становив лише 0.014. Для F1-показника, який інтегрує прецизійність і чутливість, зафіксовано значення 0.987 для Random Forest, 0.867 для Logistic Regression та 0.008 для Naive Bayes. Метрика Precision, що відображає частку істинно позитивних прогнозів серед усіх передбачених позитивних випадків, склала 0.983 у Random Forest, 0.828 у Logistic Regression та 0.041 у Naive Bayes. Показник Recall, який характеризує здатність алгоритму виявляти всі позитивні приклади, дорівнював 0.991 для Random Forest, 0.910 для Logistic Regression і лише 0.014 для Naive Bayes. Додатково обчислений коефіцієнт кореляції Метьюса (MCC) продемонстрував суттєві відмінності: 0.946 для Random Forest, 0.000 для Logistic Regression та 0.071 для Naive Bayes. Таким чином, таблиця результатів віджета Test & Score дає змогу простежити суттєву різницю в отриманих значеннях різних метрик для трьох досліджуваних алгоритмів, що створює підґрунтя для подальшого аналізу моделей з використанням інших методів візуалізації та оцінювання, зокрема Confusion Matrix та ROC Analysis.

Після етапу порівняння алгоритмів з використанням віджета *Test & Score* було здійснено перехід до візуалізації отриманих результатів, оскільки графічне подання даних дає змогу краще інтерпретувати якість роботи алгоритмів та зрозуміти специфіку їхніх помилок. У середовищі Orange Data Mining для цього використовуються кілька стандартних віджетів: *Box Plot*, *Scatter Plot* та *Confusion Matrix* (рис. 4).

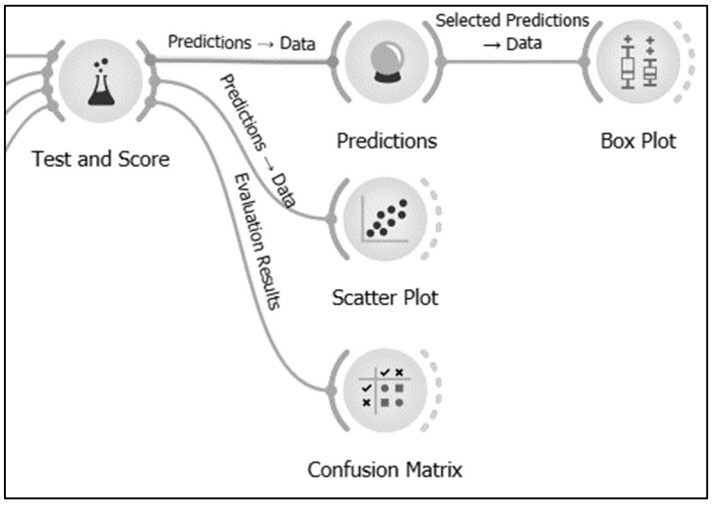

Рисунок 4 – Набір віджетів для візуалізації результатів оцінювання алгоритмів Naive Bayes classifier, Logistic Regression та Random Forest

Box Plot застосовується у дослідженні для відображення розподілу метрик (наприклад, точності чи AUC) між різними алгоритмами. Дає змогу швидко оцінити середні значення, варіацію та наявність викидів. Основна перевага цього віджета – наочність при порівнянні алгоритмів; недолік – обмежена деталізація щодо конкретних помилок.

Scatter Plot використовується у схемі для аналізу просторового розподілу об'єктів за ознаками. Він дає змогу побачити, як моделі класифікують дані, які групи утворюють об'єкти та наскільки чітко відокремлюються класи. Перевага – візуальне уявлення про структуру даних та межі між класами; недолік – придатний лише для роботи з двома-трьома ознаками одночасно, що не завжди повністю відображає багатовимірні залежності.

Confusion Matrix як інструмент візуалізації у нашій моделі відображає кількість правильних і помилкових класифікацій для кожного класу. Це найбільш інформативний інструмент для оцінки того, які саме класи плутає алгоритм. Перевага – точність і деталізація; недолік – менш наочний у випадку великої кількості класів.

Таким чином, оцінивши всі переваги та недоліки, було зроблено висновок про доцільність використання саме такого набору, оскільки усі три віджети доповнюють один одного: *Box Plot* може забезпечити загальне порівняння моделей, *Scatter Plot* – аналіз структури даних і меж класифікації, тоді як *Confusion Matrix* дає змогу детально розібратися у характері помилок кожного алгоритму.

Наступним кроком стало детальне дослідження обраних інструментів візуалізації. На першому етапі було використано віджет *Predictions*, який використовується в парі з *Box Plot*, що дає змогу зіставити передбачені алгоритмами класи з фактичними значеннями, а також оцінити якість роботи кожної моделі на рівні окремих записів бази даних. Це дало змогу виявити випадки правильної та хибної класифікації.

Віджет *Predictions* у середовищі Orange Data Mining використовується для отримання прогнозів класів на основі моделей, побудованих за допомогою алгоритмів Naive Bayes, Logistic Regression та Random Forest. Цей віджет дає змогу переглядати передбачені класи для кожного об'єкта, що є проміжним етапом перед візуалізацією результатів і може бути використаний для відбору підмножин даних.

Основна функція віджету *Predictions* – показати, як кожен об'єкт із набору даних класифікується машинним навчання і наскільки впевнено алгоритм відніс його до того чи іншого класу. Функціонал використовується – після крос-валідації для аналізу «сирих» результатів класифікації, верифікації правильності класифікації окремих записів та вибору даних з найменшою кількістю помилкових прогнозів. У свою чергу, без додаткових метрик оцінити якість важко, а за великого об'єму (датасету) дані стають громіздкими, що робить аналіз вручну майже неможливим.

Віджет *Box Plot* застосовується для візуалізації розподілу прогнозованих значень між класами [17]. Він використовується:

для інтерпретації результатів *Predictions*;

на етапі перевірки рівномірності розподілу прогнозів;

для порівняння результатів різних моделей.

Віджет показує медіану, квартилі та аномальні значення. Цей тип візуалізації має такі переваги як: наочність – легко виявляє проблемні класи; добре працює для виявлення дисбалансу; дозволяє зосередитись на аномальних випадках, а також дає змогу швидко оцінити:

чи є сильний розкид прогнозів;

чи трапляються аномалії;

наскільки добре алгоритми відокремлюють класи.

З недоліків можна виокремити таке: не показує деталізованих метрик (наприклад, точність, Recall); менш інформативний при невеликій кількості спостережень. При побудові графіка, для отримання максимальної інформативності, віджету, було встановлено такі налаштування:

Variable (ліва панель) – метрики з *Predictions*;

Subgroups (підгрупи) – використовувати output_category (Low, Medium, High), що показують розподіл прогнозів за класами;

Display – Compare means (порівняння середніх), що дає змогу побачити відмінності між групами.

Серед всіх значень метрик було обрано категорію Medium (середнє значення спостережень), а саме – Random Forest Medium (рис. 5) оскільки вона є найбільш показовою та містить збалансовану кількість спостережень і дає змогу об'єктивно оцінити роботу алгоритма. Класи Low і High зазвичай містять більш крайні значення й не завжди показують повну картину роботи алгоритмів, тоді як Medium краще підходить для узагальнення.

Для Random Forest у класі Medium результати значно відрізняються. Середнє значення перебуває на рівні 0.753, з невеликим розкидом, а також добре вираженим інтервалом довіри. Це означає, що алгоритм не лише впевнено відокремлює клас Medium, а й робить це з високою точністю. Порівняно з іншими алгоритмами, Random Forest демонструє кращу узгодженість прогнозів і мінімальну похибку, що робить його найбільш ефективним інструментом для задач із дисбалансованими даними. Крім того, вузький діапазон розподілу прогнозів свідчить про стійкість до шумів та випадкових відхилень у даних.

Віджет *Scatter Plot* – допомагає виявити кластери, аномалії та закономірності, що впливають на класифікацію [18]. Це метод графічного подання даних, який відображає окремі спостереження у вигляді точок на площині за двома вибраними змінними. Використовується для виявлення взаємозв'язків між ознаками, спостереження тенденцій та перевірки того, як алгоритми розділяють об'єкти на класи

Для інформативності графіків у *Scatter Plot* (рис. 6) доцільно підбирати змінні, які найбільше впливають на результат прогнозу.

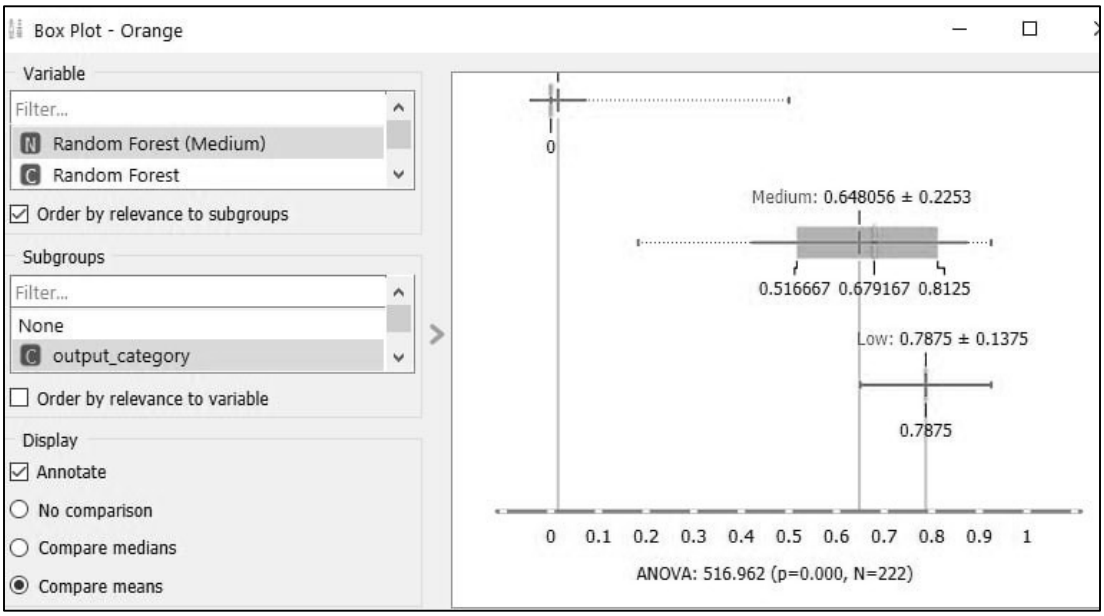

Рисунок 5 – Зображення бокс-діаграми розподілу прогнозованої категорій Medium для алгоритму Random Forest

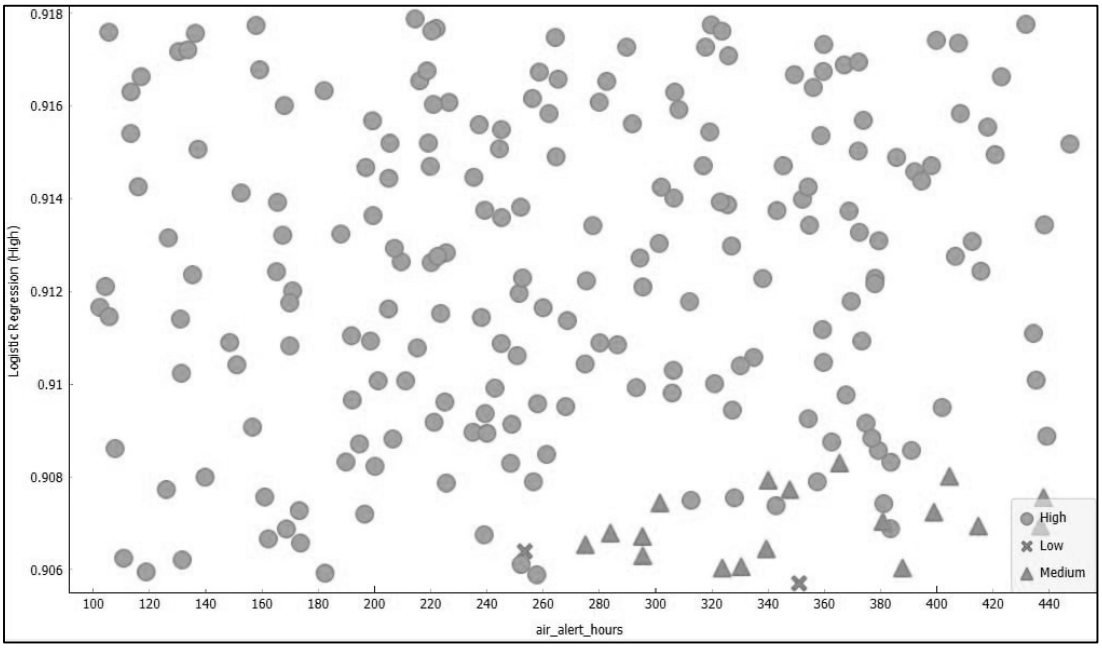

Рисунок 6 – Взаємозв’язок між тривалістю повітряних тривог та результатами виробництва

На рис. 6 візуалізовано дані про взаємозв’язок між тривалістю повітряних тривог та результатами виробництва отриманих за допомогою віджета *Scatter Plot*. Для побудови графіка обрано змінні air_alert_hours (кількість годин повітряних тривог) по осі X та output_qty (обсяг виробництва) по осі Y. Геметричні фігури відповідають категоріям цільової змінної (output_category): High (круг), Medium (трикутник), Low (перехрестя). Вибір саме цих змінних є обґрунтованим, оскільки вони дають змогу дослідити можливий вплив інтенсивності повітряних тривог на кінцеві показники виробництва.

З рис. 6 видно, що більшість спостережень із категорією High (круги) зосереджені у верхніх ділянках осі Y, що свідчить про відносно високі обсяги виробництва навіть для різних значень тривалості тривог. Категорія Medium локалізується у нижчих зонах осі Y при значних значеннях air_alert_hours, що може вказувати на певний вплив тривалих тривог на зниження виробничих результатів. Поодинокі приклади категорії Low трапляються при середніх значеннях обох змінних, утворюючи невеликі кластери, що виділяються на фоні загального розподілу.

Отже, візуалізація, наведена на рис. 6, допомагає простежити взаємозв’язок між тривалістю повітряних тривог та результатами виробництва. Використання віджету *Scatter Plot* у цьому випадку є доречним, адже

воно дає змогу не лише оцінити загальну тенденцію, а й побачити локальні відхилення та кластери в даних.

Віджет *Confusion Matrix* – це інструмент оцінювання класифікаційних моделей, який подає результати у вигляді матриці зіставлення фактичних і передбачених значень. Використовується для детального аналізу, які саме класи алгоритм визначає правильно, а де виникають помилки класифікації [19].

Таблиця 2

Результати класифікацій за алгоритмом Naive Bayes

Actual (фактично)	Predicted (прогнозовано)			
	High	Low	Medium	Σ
High	100.0 %	91.2 %	34.8 %	202
Low	0.0 %	1.8 %	4.3 %	2
Medium	0.0 %	7.0 %	60.9 %	18
Σ	142	57	23	222

В таблиці 2 зображено результати класифікацій за алгоритмом Naive Bayes та наведено кількість правильних і помилкових класифікації, отриманих завдяки використанню алгоритму Naive Bayes, відображені у вигляді *Confusion Matrix*. По горизонталі подано передбачені класи (Predicted), а по вертикалі – фактичні значення (Actual). Для зручності інтерпретації результати відображаються у відсотках, що дозволяє оцінити не лише кількість, а й пропорцію правильних та помилкових передбачень.

Аналізуючи матрицю, можна зазначити, що модель достаньо добре прогнозує категорії High, досягаючи 100% правильних класифікацій. Водночас у категоріях Medium та Low точність є нижчою. Так для класу Medium модель вірно класифікує 60,9% прикладів, тоді як решта помилково віднесені до класів High чи Low. Особливо складними виявилися спостереження класу Low, яких у вибірці було всього два, і їх передбачення виявилося ненадійним.

Отже, *Confusion Matrix* допомагає не лише побачити загальну точність моделі, а й оцінити її поведінку для окремих категорій. Це важливо у випадках, коли класи є нерівномірно представленими у вибірці, адже традиційні метрики (наприклад, Accuracy) можуть приховувати слабкі місця моделі. У цьому випадку видно, що Naive Bayes краще працює для класу High, але має значні труднощі з відокремленням Low і Medium, що пояснює потребу у використанні додаткових алгоритмів для підвищення збалансованості результатів.

Висновки й перспективи подальших досліджень

У статті досягнуто поставленої мети – здійснено експериментальне оцінювання алгоритмів машинного навчання для прогнозування виробничих показників оборонно-промислового комплексу у середовищі Orange Data Mining. Проведено порівняльний аналіз ефективності трьох алгоритмів – Naive Bayes, Logistic Regression та Random Forest – з використанням набору інтегральних і локальних метрик (AUC, Accuracy, Precision, Recall, F1, MCC) та інструментів візуального аналізу.

Результати дослідження підтвердили, що Random Forest є найбільш ефективним і стабільним методом (найвищі значення AUC, CA, F1, MCC) в умовах, наближених до операційної невизначеності та дисбалансу класів; Naive Bayes і Logistic Regression показали обмеження, зумовлені припущеннями моделей і нелінійністю зв'язків у даних.

Результати статті вперше систематично демонструють відмінності у поведінці Naive Bayes, Logistic Regression та Random Forest у контексті прогнозування виробничих потужностей з урахуванням специфічних зовнішніх чинників (екстрені події, енергопостачання, доступність персоналу) та вираженого дисбалансу класів. Запропоновано методологію комбінованого використання стандартних метрик та візуалізаційних віджетів Orange для комплексної оцінки якості моделей у прикладних умовах оборонної промисловості.

Отримані результати створюють методичне підґрунтя для подальшої апробації і впровадження алгоритмічних рішень у системах підтримки прийняття рішень для підприємств оборонно-промислового комплексу. Робота розширює підхід до оцінювання класичних алгоритмів у задачах з дисбалансованими класами та демонструє роль зовнішніх техніко-операційних факторів у формуванні прогнозів. Створений набір інструментів (побудова набору даних, візуальне моделювання у середовищі Orange, набір метрик і візуалізацій) можуть бути використані як робоча база для створення аналітичних систем підтримки прийняття рішень на підприємствах оборонно-промислового комплексу (планування виробництва, управління ризиками, сценарний аналіз).

Напрямами подальших досліджень слід вважати:

1. Проведення валідації на реальних виробничих даних оборонних підприємств (за дотримання вимог безпеки й конфіденційності) для оцінки переносимості висновків.
2. Дослідження впливу методів балансування на якість моделей у задачі з нерівномірним розподілом класів.
3. Дослідження глибинною інженерією ознак: створення лагових, агрегованих та комбінованих показників, що враховують накопичувальні ефекти (кумулятивні години тривог, взаємодії blackout_hours × staff_availability тощо).
4. Дослідження підходів калібрування

ймовірностей (Platt scaling (пасштабування Платта), isotonic regression (ізотонічна регресія)) і кількісного оцінювання невизначеності прогнозів для ризик-орієнтованого прийняття рішень.

5. Розроблення сценаріїв інтеграції алгоритмів у оперативні інформаційні системи підприємств (автоматизовані панелі моніторингу, модулі планування виробництва, модулі раннього попередження) і виконати економіко-операційний аналіз їхнього впливу.

Реалізація перелічених напрямів дослідження дасть змогу створити високоефективну автоматизовану систему прогнозування та оцінювання виробничих потужностей оборонно-промислового комплексу, що підвищить якість й оперативність виготовлення військової техніки та озброєння, сприяючи зміцненню обороноздатності України.

Список бібліографічних посилань

1. Jain A., Dubes R. Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall, 1988. 320 p. URL: https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf (accessed: 10.08.2025). **2. Kaufman L., Rousseeuw P.** Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 1990. 368 p. DOI: 10.2307/2532178. **3. Bishop C.** Pattern Recognition and Machine Learning. New York : Springer, 2006. 738 p. DOI: 10.1117/1.2819119. **4. Domingos P., Pazzani M.** On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*. 1997. Vol. 29. № 2–3. P. 103–130. DOI: 10.1023/A:1007413511361. **5. Liu Y., Li Z., Xiong H., Gao X., Wu J.** Understanding of Internal Clustering Validation Measures. *Proceedings of the IEEE International Conference on Data Mining*. Sydney, 2010. P. 911–916. DOI: 10.1109/ICDM.2010.35. **6. Breiman L.** Random Forests. *Machine Learning*. 2001. Vol. 45. P. 5–32. DOI: 10.1023/A:1010933404324. **7. Mitchell T.** Machine Learning. New York : McGraw-Hill, 1997. 414 p. URL: https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf (дата звернення: 25.08.2025). **8**. **Texty.org.ua.** Статистика повітряних тривог в Україні. 2025. URL: https://texty.org.ua (дата звернення: 16.09.2025). **9**. **Institute for the Study of War (ISW).** Daily Campaign Assessments. 2025. URL: https://www.understandingwar.org (дата звернення: 15.09.2025). **10**. **НЕК «Укренерго»**. Офіційні повідомлення про роботу енергосистеми. URL: https://ua.energy (дата звернення: 10.09.2025). **11**. **Міністерство оборони України.** Офіційні публікації та прес-релізи щодо розвитку виробництва БПЛА. URL: https://www.mil.gov.ua (дата звернення: 01.09.2025). **12. Torrijos J., Cao D., Casado-Vara R., Prieto J.** Federated Learning with Discriminative Naive Bayes Classifier. *Lecture Notes in Computer Science*. 2025. Vol. 15347. P. 355–369. DOI: 10.1007/978-3-031-77738-7_27. **13. Hosmer D., Lemeshow S., Sturdivant R.** Applied Logistic Regression. Wiley, 2013. DOI:10.1002/9781118548387. **14. Liaw A., Wiener M.** Classification and Regression by randomForest. *R News*. 2002. Vol. 2(3). P. 18–22. URL: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf (дата звернення: 05.09.2025). **15. Arlot S., Celisse A.** A survey of cross-validation procedures for model selection. *Statistics Surveys*. 2010. Vol. 4. P. 40–79. DOI: 10.1214/09-SS054. **16. Chicco D., Jurman G.** The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020. Vol. 21(6). P. 1-13. DOI: 10.1186/s12864-019-6413-7. **17**. **Correll M.** Teru Teru Bōzu: Defensive Raincloud Plots. *Computer Graphics Forum*. 2023. Vol. 42, Is. 3. P. 235–246. DOI: 10.1111/cgf.14826. **18**. **Quadri G. J., Wong L., Dunne C., Lee B.** Automatic Scatterplot Design Optimization for Clustering Identification. *IEEE Transactions on Visualization and Computer Graphics*. 2022. Vol. 29. No. 10. P. 4312–4327. DOI: 10.1109/TVCG.2022.3189883. **19**. **Tharwat A.** Classification assessment methods. *Applied Computing and Informatics*. 2021. Vol. 17. No. 1. P. 168–192. DOI: 10.1016/j.aci.2018.08.003.

EXPERIMENTAL EVALUATION OF MACHINE LEARNING ALGORITHMS MACHINE LEARNING ALGORITHMS FOR FORECASTING PRODUCTION INDICATORS OF THE DEFENCE INDUSTRY IN THE ORANGE DATA MINING ENVIRONMENT

KOVAL Igor**,** Lutsk National Technical University, Lutsk, Ukraine, https://orcid.org/0009-0001-2083-1747

HOLOVNIA Serhii, Lutsk National Technical University, Lutsk, Ukraine, https://orcid.org/0009-0005-2997-9202

Formulation of the problem in general. *Accurate forecasting of defence industry production capacities is complicated by external disruptions and imbalanced data. Classical analytical approaches often fail to ensure reliable predictions in such conditions.* ***The purpose of the article*** *is to develop and experimentally test a model for forecasting the production capacities of the defence-industrial complex using machine learning in the Orange Data Mining environment, as well as to assess the accuracy and feasibility of using algorithms (naive Bayes classifier, logistic regression, and random forest) to solve the task at hand.*

Research methods. *The methods of system analysis, data mining, and cross-validation were applied in the Orange Data Mining environment. The algorithms of the Naive Bayes classifier, logistic regression, and the random forest method were used. The accuracy was assessed using the following metrics: AUC (Area Under the Receiver Operating Characteristic Curve), precision (positive predictive value), F1 (harmonic mean of precision and recall), and Matthews correlation coefficient (MCC) – a measure of the quality of binary and multiclass classifications.*

Literature review. *Forecasting production capacities in the defence industry remains a challenging task due to the influence of multiple operational and external factors. Existing studies on machine learning provide valuable insights into the application of classical algorithms; however, they often show limitations when dealing with imbalanced and*

domain-specific data. This highlights the importance of further research in this direction and confirms the relevance of the article's chosen topic.

Research results. *An experimental dataset was generated that takes into account the influence of external and internal factors on production processes. A comparison of machine learning algorithms was conducted, the results of which showed a significant advantage of the random forest method, which achieved the highest values of accuracy and classification balance. Error matrices demonstrated the limitations of the naive Bayes classifier and logistic regression in data imbalance.*

Research novelty*. For the first time in this subject area, differences in the performance of the naive Bayesian classifier, logistic regression, and random forest method under conditions of data imbalance have been shown, which made it possible to determine the most effective algorithm for forecasting production capacities.*

Theoretical and practical significance. *The theoretical significance of this research lies in expanding the methodological foundations for applying machine learning algorithms to forecast production processes, taking into account the influence of both external and internal factors. The paper clarifies the possibilities of using integral metrics for evaluating models, which increases the reliability of results in cases of data imbalance. The practical significance of the results obtained is determined by the possibility of using the random forest algorithm as a basic tool for forecasting production capacities of enterprises of the defence-industrial complex. The proposed approach enables the enhancement of management decision validity in the fields of production planning, resource optimisation, and ensuring the stability of defence-industrial processes under challenging conditions.*

Conclusions and future work. *The study demonstrated that Random Forest provides the highest accuracy and robustness in forecasting defence industry production capacities under data imbalance, compared to Naive Bayes and Logistic Regression. The results confirm the feasibility of applying machine learning methods in this domain and highlight the importance of using multiple evaluation metrics. Future research should focus on testing advanced ensemble models, balancing techniques, and real operational datasets to further improve the reliability and applicability of forecasting methods.*

Keywords: *machine learning, Naive Bayes classifier, Logistic Regression, random forest, Orange Data Mining, data imbalance, defence-industrial complex.*

References

1. Jain, A. and Dubes, R., (1988). *Algorithms for Clustering Data*. Englewood Cliffs: Prentice Hall. Available at: https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf [Accessed: 10 August 2025]. **2. Kaufman, L. and Rousseeuw, P.,** (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley. DOI: 10.2307/2532178. **3. Bishop, C.,** (2006). *Pattern Recognition and Machine Learning.* New York: Springer. 738 p. DOI: 10.1117/1.2819119. **4. Domingos, P. and Pazzani, M.,** (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning.* 29(2–3), 103–130. DOI: 10.1023/A:1007413511361. **5. Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J.,** (2010). Understanding of Internal Clustering Validation Measures. In: *Proceedings of the IEEE International Conference on Data Mining.* Sydney, DOI: 10.1109/ICDM.2010.35. **6. Breiman, L.,** (2001). Random Forests. *Machine Learning*. 45, 5–32. DOI: 10.1023/A:1010933404324. **7. Mitchell, T.,** (1997). *Machine Learning*. New York: McGraw-Hill [online] Available at: https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf [Accessed: 25 August 2025] **8.** ***Texty.org.ua***. *Statistics of air alerts in Ukraine* [online], (2025). Available at: https://texty.org.ua [Accessed: 16 September 2025]. **9. Institute for the Study of War (ISW).** *Daily Campaign Assessments* [online], (2025). Available at: https://www.understandingwar.org [Accessed: 15 September 2025]. **10. NEC «Ukrenergo».** *Official announcements on the operation of the power system* [online], (2025). Available at: https://ua.energy [Accessed: 10 September 2025]. **11. Ministry of Defence of Ukraine.** *Official publications and press releases on the development of UAV production* [online], (2025). Available at: https://www.mil.gov.ua [Accessed: 01 September 2025]. **12. Torrijos, J., Cao, D., Casado-Vara, R. and Prieto, J.,** (2025). Federated Learning with Discriminative Naive Bayes Classifier. In: *Lecture Notes in Computer Science*, 15347, 355–369. Springer. DOI: 10.1007/978-3-031-77738-7_27. **13. Hosmer, D., Lemeshow, S. and Sturdivant, R.,** (2013). *Applied Logistic Regression*. Hoboken, NJ: Wiley. DOI: 10.1002/9781118548387. **14. Liaw, A. and Wiener, M.,** (2002). Classification and Regression by randomForest. *R News*, [online]. Available at: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf [Accessed: 05 September 2025]. **15. Arlot, S. and Celisse, A.,** (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. DOI: 10.1214/09-SS054. **16. Chicco, D. and Jurman, G.,** (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6), 1–13. DOI: 10.1186/s12864-019-6413-7. **17. Correll, M.,** (2023). Teru Teru Bōzu: Defensive Raincloud Plots. *Computer Graphics Forum*, 42(3), 235–246. DOI: 10.1111/cgf.14826. **18. Quadri, G. J., Wong, L., Dunne, C. and Lee, B.,** (2022). Automatic Scatterplot Design Optimisation for Clustering Identification. *IEEE Transactions on Visualisation and Computer Graphics*. 29(10), 4312–4327. DOI: 10.1109/TVCG.2022.3189883. **19. Tharwat, A.,** (2021). Classification assessment methods. *Applied Computing and Informatics.* 17(1), 168–192. DOI: 10.1016/j.aci.2018.08.003.

Рукопис надійшов до редакції *22.09.2025*
Рукопис прийнято до друку після рецензування *17.11.2025*
Дата публікації *30.12.2025*