

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика»

Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «Прогнозування перемоги учасників у закупівлях Прозорро
методами машинного навчання»
(назва теми)

здобувача Беленка Дмитра Віталійовича
(ПІБ)

(підпис)

Науковий керівник: кандидат фізико-математичних наук,
професор Олена ЮНЬКОВА

(підпис)

**Робота допущена до захисту перед екзаменаційною комісією
з атестації здобувачів вищої освіти (ЕК)**

Завідувач кафедри: кандидат фізико-математичних наук,
професор Галина ВЕЛИКОІВАНЕНКО

(підпис)

Київ 2025

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика»

**Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»**

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми
Ольга ПРИТОМАНОВА

(підпис)

_____ 2025 р.

ЗАТВЕРДЖУЮ

Завідувач кафедри математичного
моделювання та статистики
Галина ВЕЛИКОІВАНЕНКО

(підпис)

_____ 2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

**здобувача вищої освіти Беленка Дмитра Віталійовича
очної (денної) форми здобуття освіти
на підготовку кваліфікаційної бакалаврської роботи**

на тему

«Прогнозування перемоги учасників у закупівлях Прозорро методами машинного навчання»

Тему затверджено наказом ректора Університету від «__» _____ 20__ р. № _____

Кваліфікаційна бакалаврська робота виконується на матеріалах

ТОВ «СМАРТТЕНДЕР»

План кваліфікаційної бакалаврської роботи

Розділ 1	Зasadничі принципи організації державних закупівель
Розділ 2	Аналіз даних системи Prozorro
Розділ 3	Реалізація та оцінювання моделей прогнозування перемоги учасників у системі Prozorro

Об'єкт дослідження:	Процес проведення публічних закупівель у системі ProZorro.
Предмет дослідження:	Методи машинного навчання, зокрема, алгоритми класифікації, та їх застосування для прогнозування перемоги учасників у державних закупівлях
Мета виконання кваліфікаційної бакалаврської роботи:	Налаштувати та впровадити моделі машинного навчання для прогнозування переможців у державних закупівлях електроенергії з метою підвищення ефективності участі в тендерах та оптимізації конкурентних стратегій

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1: Розкрити сутність та проаналізувати правові основи державних закупівель. Дослідити структуру та визначити ключових учасників процесу державних закупівель. Охарактеризувати особливості державних закупівель електроенергії та проаналізувати чинники, що впливають на їх проведення

У розділі 2: Здійснити попередню обробку та аналіз даних, зібраних із платформи ProZorro. Розглянути підходи до прогнозування перемоги учасників тендерів. Сформулювати задачу класифікації для моделі прогнозування. Визначити метрики для оцінювання якості моделей.

У розділі 3: Дослідити основні методи машинного навчання, що використовуються для класифікації. Розробити, побудувати та навчити модель прогнозування переможця тендера. Оптимізувати модель та застосувати її для прогнозування перемоги учасників у державних закупівлях електроенергії.

Завдання підготував
науковий керівник

(підпис)

Олена ЮНЬКОВА
(ім'я, прізвище)

«__» _____ 20__ р.

Завдання одержав
здобувач

(підпис)

Дмитро БЕЛЕНОК
(ім'я, прізвище)

«__» _____ 20__ р.

Реферат

Кваліфікаційна бакалаврська робота складається зі вступу, трьох розділів, висновків, переліку використаних джерел, що містить 30 найменувань на 3 сторінках, та 2 додатків на 5 сторінках. Загальний обсяг роботи складає 59 сторінок друкованого тексту, з яких 51 сторінка основної частини. Робота містить 18 рисунків, 19 формул та 10 таблиць.

Об'єктом дослідження виступає процес проведення публічних закупівель у системі ProZorro.

Предметом дослідження є вивчення методів машинного навчання, зокрема, алгоритми класифікації, та їх застосування для прогнозування перемоги учасників у державних закупівлях.

Метою дослідження є налаштування та впровадження моделей машинного навчання для прогнозування переможців у державних закупівлях електроенергії з метою підвищення ефективності участі в тендерах та оптимізації конкурентних стратегій.

Відповідно до окресленої мети були поставлені такі завдання, а саме:

- розгляд сутності та правових основ державних закупівель
- визначення ключових учасників державних торгів
- аналіз чинників, що впливають на проведення закупівель електроенергії
- вивчення методів машинного навчання для вирішення задач класифікації
- побудова моделей машинного навчання за допомогою Python та пакету *Scikit-learn*
- прогнозування перемоги учасників у закупівлях Prozorro
- порівняння отриманих результатів

Для розв'язання поставлених завдань застосовано методи класифікації, зокрема логістична регресія, дерева рішень та метод опорних векторів.

У результаті застосування вказаних методів здійснено аналіз- моделей прогнозу перемоги учасників у закупівлях електроенергії. За показниками точності прогнозу обрано кращу модель для прогнозування

Ключові слова: машинне навчання, державні закупівлі, прогнозування, логістична регресія, дерева рішень, метод опорних векторів, класифікація

Відгук
про кваліфікаційну бакалаврську роботу
здобувача освітньо-професійної програми «Економічна кібернетика»
Беленка Дмитра Віталійовича
на тему «Прогнозування перемоги учасників у закупівлях Прозорро методами
машинного навчання»»

1. Актуальність теми зумовлена значним обсягом державних закупівель через систему Прозорро та необхідністю підвищення ефективності та прозорості цього процесу. Застосування методів машинного навчання для прогнозування перемоги учасників у закупівлях є вкрай актуальним, оскільки дозволяє виявити ключові фактори успіху, оптимізувати стратегії участі та мінімізувати ризики. Отримані результати сприятимуть обґрунтуванню рекомендацій для учасників закупівель.

2. Позитивні риси кваліфікаційної бакалаврської роботи: робота виконана на високому теоретичному рівні, має логічно пов'язану структуру та містить усі необхідні компоненти: вступ із чітким обґрунтуванням актуальності, два змістовні розділи з докладним аналізом проблеми та сформульовані висновки. Особливо хочу відзначити глибокий огляд методів машинного навчання, застосованих до прогнозування у сфері закупівель, та ефективне поєднання якісного огляду літератури з емпіричним аналізом даних Прозорро.

3. Наявність самостійних розробок автора: автором виконано дослідження різних моделей машинного навчання для прогнозування перемоги учасників у закупівлях, що відображено у використанні метрик оцінки моделей та візуалізації результатів прогнозування. Також проведено крос-валідацію для оцінки стійкості моделей. Особливу увагу приділено детальному аналізу ринку закупівель електроенергії, а результати дослідження ефективно зображені за допомогою графіків.

4. Цінність теоретичних висновків полягає в поглибленні наукового розуміння механізмів формування успішності учасників у державних закупівлях та в обґрунтуванні методології застосування машинного навчання для такого прогнозування. Практичні рекомендації роботи спрямовані на розробку інструментів, здатних підвищити ефективність участі у закупівлях та покращити їх прозорість. Виконане дослідження може слугувати базою для подальших емпіричних і теоретичних розробок у сфері аналізу даних державних закупівель та застосування ШІ

5. Наявність недоліків: у роботі варто було б більш докладно аргументувати вибір конкретних ознак для моделювання. Доцільно розширити емпіричну базу, додавши аналіз впливу часових рядів на прогнозування, а також провести аналіз з іншими доступними моделями машинного навчання або ансамблевими методами.

6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: Кваліфікаційна бакалаврська робота Беленка Дмитра Віталійовича «Прогнозування перемоги учасників у закупівлях Прозорро методами машинного навчання» виконана на належному рівні, відповідає вимогам, що висуваються для даного типу робіт, може бути рекомендована до захисту перед ЕК. Робота свідчить про відповідність набутих здобувачем компетентностей вимогам освітньої програми «Економічна кібернетика», а її автор Беленок Д.В., за умови успішного захисту, заслуговує на присвоєння освітньої кваліфікації «Бакалавр з економіки».

Науковий керівник: професор кафедри
математичного моделювання і статистики
канд. фіз.-мат. наук, доцент _____

(підпис)

Олена ЮНЬКОВА
«__» _____ 20__ р

ЗМІСТ

ВСТУП	3
РОЗДІЛ 1 ЗАСАДНИЧІ ПРИНЦИПИ ОРГАНІЗАЦІЇ ДЕРЖАВНИХ ЗАКУПІВЕЛЬ	6
1.1. Сутність та правові основи державних закупівель в Україні	6
1.2. Структура та учасники процесу державних закупівель	8
1.3. Особливості закупівель електроенергії в системі ProZorro	12
РОЗДІЛ 2 АНАЛІЗ ДАНИХ СИСТЕМИ PROZORRO	16
2.1. Попередня обробка та аналіз даних	16
2.2. Підходи до прогнозування перемоги учасників у державних закупівлях	24
2.3. Постановка задач класифікації	26
2.4. Метрики для оцінювання якості моделей	30
РОЗДІЛ 3 РЕАЛІЗАЦІЯ ТА ОЦІНЮВАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ПЕРЕМОГИ УЧАСНИКІВ У СИСТЕМІ PROZORRO	34
3.1. Методи машинного навчання для вирішення задач класифікації	34
3.2. Вибір програмного забезпечення	41
3.3. Прогнозування перемоги учасників	44
ВИСНОВКИ	50
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	52
ДОДАТКИ	55

ВСТУП

Аналіз державних закупівель є важливою складовою економічних досліджень, оскільки він безпосередньо впливає на ефективне використання державних коштів і розвиток економіки країни. Публічні закупівлі, зокрема через систему ProZorro, дозволяють значно покращити прозорість та конкурентоспроможність державних тендерів, забезпечуючи більшу відкритість процесів для учасників та замовників. Завдяки цьому зростає інтерес до автоматизації та оптимізації процесу прогнозування результатів тендерів, що створює нові можливості для учасників ринку, особливо в сфері електроенергії.

Один із найперспективніших напрямків у цій сфері - застосування методів машинного навчання. Вони дозволяють здійснювати глибокий аналіз великої кількості даних, виявляти закономірності і на основі цього розробляти прогнози щодо ймовірності перемоги учасників. Оскільки участь у тендерах передбачає значну конкуренцію, здатність точно прогнозувати результат може суттєво підвищити ефективність бізнес-стратегій і оптимізувати бюджет компаній.

Використовуючи алгоритми класифікації та сучасні інструменти для аналізу даних, можна створити моделі, які допоможуть компаніям більш ефективно брати участь у тендерах, прогнозуючи результат на основі історичної інформації.

Актуальність даного дослідження обумовлена потребою в удосконаленні прийняття оптимальних рішень у державних закупівлях. Оскільки система ProZorro створена з метою зниження корупційних ризиків та підвищення ефективності витрат державних коштів, її використання вимагає глибокого аналізу даних для створення обґрунтованих прогнозів.

У даній дипломній роботі розглядаються методи машинного навчання, зокрема алгоритми класифікації, які використовують історичні дані публічних закупівель електроенергії в системі ProZorro, що включають інформацію про характеристики тендерів, пропозиції учасників та їх результати, для прогнозування потенційних переможців.

Питання аналізу та підвищення ефективності державних закупівель в Україні, а також застосування сучасних методів обробки даних для прогнозування активно досліджується вітчизняними науковцями. Зокрема, аспекти функціонування системи публічних закупівель, її реформування та економічне значення висвітлювали у своїх працях такі науковці, як С. П. Поліщук, А. Вальковець, К. Поляк, С. М. Науменко, Е. О. Юрій, О. М. Грубляк. Дослідженню та застосуванню методів машинного навчання, зокрема алгоритмів класифікації та вирішенню специфічних задач аналізу даних, присвячені роботи Д. Чернишова, Д. Ситнікова, С. М. Кравченка, Є. О. Гришкун, О. В. Власенка. Незважаючи на значний науковий доробок вітчизняних вчених у сферах державних закупівель та машинного навчання, ще досі лишаються невирішені питання щодо побудови найбільш адекватних та ефективних моделей, а також вибору оптимальних методів для прогнозування переможців у тендерах на постачання електроенергії в специфічних умовах функціонування системи ProZorro, що й визначає необхідність даного дослідження.

Метою цієї роботи є розробка моделей машинного навчання для прогнозування переможців у тендерах на постачання електроенергії, що проводяться через систему ProZorro.

Для досягнення поставленої мети визначено такі завдання:

- проаналізувати основні принципи організації та проведення державних закупівель в Україні, зокрема в системі ProZorro;
- здійснити збирання та попередню обробку даних публічних закупівель електроенергії з системи ProZorro;
- дослідити сучасні методи машинного навчання, що застосовуються для розв'язання задач класифікації;
- розробити та реалізувати моделі машинного навчання на основі алгоритмів класифікації для прогнозування переможців тендерів;
- оцінити ефективність розроблених моделей та порівняти їх результати.

Об'єктом дослідження є процес проведення публічних закупівель у системі

ProZorro.

Предметом дослідження є методи машинного навчання, зокрема алгоритми класифікації, їх застосування для прогнозування перемоги учасників у державних закупівлях.

Методи дослідження. Загальнонаукові - аналіз та синтез, порівняльний аналіз, а також спеціальні методи - методи машинного навчання, зокрема, алгоритми класифікації, включаючи Логістичну регресію, Древа рішень, Випадковий ліс, Метод опорних векторів.

Теоретична, методична та практична значущість отриманих результатів. Розроблені моделі прогнозування переможців можуть бути адаптовані та використані як методична основа для аналогічних досліджень стосовно інших категорій закупівель. Практичне застосування *може* стати інструментом для підвищення ефективності компаній та загальної ефективності державних витрат у цій сфері.

Інформаційна база дослідження. У даній роботі використані наукові статті та публікації зарубіжних та вітчизняних авторів, дані публічних закупівель електроенергії у системі закупівель ProZorro.

Структура дипломної роботи. Робота складається зі вступу, трьох основних розділів, висновків, списку використаних джерел та додатків.

РОЗДІЛ 1

ЗАСАДНИЧІ ПРИНЦИПИ ОРГАНІЗАЦІЇ ДЕРЖАВНИХ ЗАКУПІВЕЛЬ

1.1. Сутність та правові основи державних закупівель в Україні

Державні закупівлі - це організований та врегульований процес придбання товарів, робіт і послуг за рахунок публічних коштів для забезпечення функціонування органів державної влади, місцевого самоврядування, бюджетних установ, підприємств і організацій. Вони є важливим елементом економічної політики будь-якої держави, оскільки впливають на формування попиту в економіці, забезпечують потреби суспільства та слугують засобом реалізації державних і соціальних програм.

Важливим етапом у розвитку державних закупівель в Україні стало створення та впровадження електронної системи *ProZorro* [1]. Ця система, що функціонує з 2015 року та регламентується Законом України «Про публічні закупівлі», була розроблена з метою кардинального підвищення прозорості, забезпечення добросовісної конкуренції та ефективного використання публічних коштів. Завдяки електронному формату проведення торгів та обов'язковій публікації всієї інформації про закупівлі, *ProZorro* дозволила мінімізувати корупційні ризики, спростити участь у тендерах та забезпечити рівний доступ для всіх учасників ринку [2].

Публічні закупівлі мають здійснюватися за такими принципами:

- 1) добросовісна конкуренція серед учасників
- 2) максимальна економія, ефективність та пропорційність
- 3) відкритість та прозорість на всіх стадіях закупівель
- 4) недискримінація учасників та рівне ставлення до них
- 5) об'єктивне та неупереджене визначення переможця процедури закупівлі
- 6) запобігання корупційним діям і зловживанням

Контроль за дотриманням законодавства у сфері публічних закупівель здійснюється через кілька інституційних та інструментальних механізмів:

Уповноважений орган - Міністерство економіки України. Воно координує реалізацію державної політики у сфері закупівель, розробляє нормативно-правові акти, проводить моніторинг закупівель.

Державна аудиторська служба України (ДАСУ) - здійснює моніторинг закупівель, виявляє порушення, направляє вимоги до замовника щодо усунення невідповідностей або відміни процедури.

Антимонопольний комітет України (АМКУ) - виступає органом оскарження, який розглядає скарги учасників на дії замовників. Рішення АМКУ є обов'язковими до виконання.

У разі виявлення грубих порушень, законодавство передбачає адміністративну, дисциплінарну або кримінальну відповідальність посадових осіб, які відповідають за закупівлі.

Сутність державних закупівель полягає у досягненні балансу між ефективним використанням бюджетних ресурсів і забезпеченням прозорості, чесної конкуренції та рівного доступу компаній до участі в тендерах. Цей процес не лише задовольняє потреби замовників, але й створює умови для розвитку національного бізнесу, зокрема малого та середнього підприємництва.

Значний обсяг державних закупівель у відношенні до валового внутрішнього продукту є характерною рисою розвинених економік. Наприклад, у Швейцарії цей показник становить близько 8%, тоді як у Нідерландах він сягає 21%. Такі суттєві обсяги підкреслюють, що державні закупівлі можуть бути інструментом для створення ринків товарів і послуг, впливу на баланс попиту та пропозиції, а також для заохочення розвитку інновацій та новітніх технологій [3].

У 2021 році електронна система ProZorro відіграла ключову роль у державних закупівлях України, обсяг яких перевищив 1,1 трлн грн, що становило близько 18% ВВП країни. Ефективність закупівель, що визначається своєчасним придбанням товарів, робіт або послуг належної якості за оптимальною ціною, вимірюється рівнем конкуренції та відносною економією.

2022 рік приніс суттєві зміни в практику закупівель через повномасштабне вторгнення, що призвело до різкого скорочення частки конкурентних процедур: з трьох чвертей у 2021 році (840 з 1,1 трлн грн) до лише 12% у 2022 році (100 млрд грн з 270 млрд грн). Таке падіння було спричинене тимчасовими дозволами замовникам здійснювати певні закупівлі без використання електронної системи.

Незважаючи на це, у 2022 році в ProZorro було проведено 160 тис. конкурентних закупівель на майже 300 млрд грн за участю 33 тис. учасників.

Дані аналітичного модуля ProZorro[4] демонстрували тимчасовий спад активності на початку війни, але подальше пожвавлення ринку та стабілізацію у перші три місяці 2023 року, коли 2366 успішних аукціонів принесли 897,6 млн грн до бюджетів різних рівнів. Серед найбільш поширених предметів закупівель за останні два роки лідирують будівельні матеріали, продукти харчування, послуги з ремонту, паливно-енергетичні товари (включно з електроенергією) та електротехнічне обладнання [5]

Через механізми державних закупівель уряд має змогу не лише забезпечувати свою діяльність, а й впливати на економічну активність, стимулювати інновації, підтримувати національних виробників та сприяти сталому розвитку. Крім того, система закупівель може бути інструментом антикризової політики, зокрема у періоди економічної нестабільності або надзвичайних ситуацій.

Державні закупівлі можуть виконувати також стимулювальну функцію - сприяти розвитку окремих галузей, підтримувати вітчизняного виробника, створювати робочі місця, запроваджувати екологічні або інноваційні рішення. Зокрема, інструментом такої політики є запровадження критеріїв «локалізації» або «зелених закупівель», коли перевага надається товарам, виготовленим із дотриманням екологічних стандартів або на території України.

1.2. Структура та учасники процесу державних закупівель

Система публічних закупівель в Україні побудована на взаємодії трьох ключових суб'єктів: замовників, учасників та електронних майданчиків, які взаємодіють через централізовану базу даних електронної системи ProZorro та виступають умовними посередниками.

Замовники - це державні органи, органи місцевого самоврядування, підприємства державної або комунальної форми власності, установи, організації, які здійснюють закупівлі за бюджетні кошти. Вони створюють тендер, формують тендерну документацію, оцінюють пропозиції та укладають договір з найбільш привабливим переможцем.

Учасники - це суб'єкти господарювання (юридичні або фізичні особи-підприємці), які подають свої тендерні пропозиції з метою отримання замовлення. Для них участь у закупівлях є можливістю реалізувати товари чи послуги державним структурам.

Електронні майданчики - це акредитовані веб-платформи, які забезпечують технічну взаємодію між користувачами (замовниками та учасниками) і центральною базою даних ProZorro.

Станом на 2025 рік в Україні функціонує понад 20 авторизованих майданчиків, серед яких найбільш відомі - Zakupki.Prom.ua, SmartTender, E-Tender тощо.

Проведення тендеру доволі трудомісний процес, який охоплює багато етапів, порядок яких наведений у табл. 1

Таблиця 1.1 - Загальний процес проведення державних закупівель[6]

Етап	Опис
Планування закупівлі	Замовник формує річний план закупівель, у якому зазначає предмет закупівлі, орієнтовну вартість, очікувану дату початку процедури

Продовження Таблиці 1.1

Оголошення процедури закупівлі	Замовник публікує оголошення про початок процедури, додаючи тендерну документацію з вимогами до товарів/послуг, кваліфікаційними критеріями до учасників, умовами договору
Період уточнень	Учасники можуть ставити запитання, вимагати роз'яснень або оскаржувати умови тендерної документації. Замовник надає відповіді через систему.
Подача тендерних пропозицій	Учасники завантажують у систему свої комерційні пропозиції, які зберігаються в зашифрованому вигляді до моменту розкриття.
Аукціон	Після завершення прийому пропозицій система автоматично призначає дату та час електронного аукціону. Учасники в режимі реального часу по черзі знижують свою цінову пропозицію.
Оцінка та кваліфікація	Замовник аналізує документи учасника з найнижчою ціною, перевіряє відповідність вимогам і приймає рішення щодо переможця.
Укладання договору	Після завершення оцінки переможець підписує договір із замовником. Договір та інформація про його виконання публікуються в системі.
Звітність	Замовник оприлюднює звіт про укладений договір, а згодом - звіт про його виконання, включаючи фінансові деталі.

Усі закупівлі поділяються на допорогові, надпорогові та спрощені, відповідно до вартісних меж [2]:

Допорогові закупівлі - це закупівлі, вартість яких не перевищує 100 тисяч гривень для товарів і послуг та 1,5 мільйона гривень для робіт. Для замовників, що здійснюють діяльність в окремих сферах господарювання, ці пороги вищі - до 1 мільйона гривень для товарів і послуг та до 5 мільйонів гривень для робіт

Надпорогові закупівлі, своєю чергою, починаються від 200 тисяч гривень для товарів і послуг і від 1,5 мільйона гривень для робіт. Для окремих сфер діяльності пороги встановлені на рівні 1 мільйона гривень для товарів і послуг та 5 мільйонів гривень для робіт. Усі надпорогові закупівлі проводяться з використанням однієї з трьох тендерних процедур: відкритих торгів, переговорної процедури та конкурентного діалогу [9].

Спрощена закупівля - дозволяє оперативного закуповувати ресурси на суми, що не перевищують встановлені пороги.

Найбільш близькою до звичних тендерів залишилася процедура відкритих торгів з особливостями, яка дозволяє укласти договір навіть за участю лише одного учасника. Таке послаблення виправдане необхідністю швидкого забезпечення армії та органів влади, але разом із тим зберігає базові принципи конкурентності. Прозора та конкурентна процедура, що здійснюється через електронну систему Prozorro. Переговорна процедура застосовується лише у виняткових випадках, передбачених законом, наприклад, у разі крайньої необхідності чи відсутності конкуренції. Конкурентний діалог, своєю чергою, дозволяє замовнику вести діалог з учасниками перед тим, як остаточно визначити технічні та якісні характеристики предмета закупівлі.

Таблиця 1.2 - Основні види процедур закупівель

Вид процедури	Порогове значення	Особливості
Відкриті торги	Надпорогові закупівлі	Найпоширеніший метод; передбачає максимальну конкуренцію
Спрощена закупівля	50 тис. - порогові межі	Проста процедура без етапу кваліфікації
Переговорна процедура	Особливі випадки	Застосовується за нагальної потреби або відсутності конкуренції
Електронний каталог	До 200 тис. грн	Швидка закупівля з каталогу Prozorro Market

Джерело: розроблено автором на основі [1]

Відкриті торги з особливостями залишаються основною процедурою у воєнний час, однак змінилися певні вимоги: наприклад, закупівля може відбутися

навіть за участю лише одного учасника. Це дозволяє пришвидшити процес придбання необхідних товарів або послуг, що є критичним під час війни. Строк на подання тендерних пропозицій складає щонайменше 7 календарних днів, а наявність тендерної документації залишається обов'язковою

Спрощена закупівля застосовується за ініціативою самого замовника у випадках, коли вартість закупівлі не досягає порогів, визначених для відкритих торгів з особливостями. Така форма дозволяє проводити конкурентну процедуру без обов'язковості торгів, якщо замовник вважає це за доцільне.

Закупівля без застосування електронної системи, тобто пряме укладання договору із постачальником, дозволяється лише у певних випадках. Серед таких умов - невелика вартість закупівлі або наявність документального підтвердження відсутності конкуренції, що виключає можливість проведення відкритих торгів чи використання електронного каталогу.

Останній тип - закупівля через електронний каталог, який дозволяє придбати стандартні товари, наявні в системі, без проведення повноцінного тендеру. Однак така форма застосовується лише для товарів (не для робіт), які мають спрощений характер постачання й експлуатації.

Таблиця 1.3 Розподіл лотів по типах процедур

Вид процедури	Частка від загальної кількості	Частка від загальної суми
Відкриті торги	55%	78%
Спрощена закупівля	35%	12%
Переговорна процедура	7%	9%
Інші	3%	1%

Джерело: розроблено автором на основі [9]

1.3. Особливості закупівель електроенергії в системі ProZorro

Починаючи з 2019 року, відповідно до Закону України «Про ринок електричної енергії» [7] основними учасниками ринку електричної енергії є:

Учасниками ринку електричної енергії є:

Споживач електричної енергії – це фізична або юридична особа, яка здійснює закупівлю електричної енергії для власного споживання.

Виробники електричної енергії – суб'єкти, які здійснюють виробництво електричної енергії, маючи відповідну ліцензію.

Оператор системи передачі – юридична особа, відповідальна за експлуатацію, диспетчеризацію, забезпечення технічного обслуговування, розвиток системи передачі та міждержавних ліній електропередачі, а також за забезпечення довгострокової спроможності системи передачі щодо задоволення обґрунтованого попиту на передачу електричної енергії.

Оператор системи розподілу – юридична особа, відповідальна за безпечну, надійну та ефективну експлуатацію, технічне обслуговування та розвиток системи розподілу і забезпечення довгострокової спроможності системи розподілу щодо задоволення обґрунтованого попиту на розподіл електричної енергії з урахуванням вимог щодо охорони навколишнього природного середовища та забезпечення енергоефективності.

Оператор ринку – юридична особа, що забезпечує функціонування ринку "на добу наперед" та внутрішньодобового ринку та організацію купівлі-продаж електричної енергії на цих ринках, а також має право організовувати та проводити електронні аукціони з купівлі-продажу електричної енергії за двосторонніми договорами на підставі відповідних ліцензій, що видаються Національною комісією з цінних паперів та фондового ринку.

Гарантований покупець – суб'єкт господарювання, зобов'язаний купувати електричну енергію у виробників, яким встановлено "зелений" тариф, а також у виробників за аукціонною ціною та виконувати інші функції, визначені чинним законодавством.

Електропостачальник – юридична особа, яка отримала відповідну ліцензію та здійснює постачання електричної енергії споживачам за укладеним із цим споживачем договором.

Треjder – юридична особа, яка здійснює купівлю-продаж електричної енергії за двосторонніми договорами та на організованих сегментах ринку

електричної енергії, крім її продажу за договором постачання електричної енергії споживачу.

Ключовою особливістю закупівель електроенергії, що відрізняє її від багатьох інших предметів закупівель, є динамічний характер ціноутворення на ринку.

Ціна електричної енергії на оптових сегментах (наприклад, ринку "на добу наперед") постійно змінюється під впливом попиту, пропозиції та інших чинників. З огляду на цю волатильність, договори на постачання електроенергії для публічних замовників укладаються з урахуванням специфічних умов.

Зазвичай вони передбачають формулу ціноутворення або механізми коригування ціни протягом терміну дії договору, що базуються на змінах індексів оптового ринку. Такий підхід дозволяє адаптувати вартість послуги до поточних ринкових умов, мінімізуючи ризики для обох сторін [8]

Результати тендерів на закупівлю електроенергії залежать від комплексу економічних, регуляторних, організаційних і технічних факторів:

- Найбільш важливий критерій — запропонована ціна за 1 кВт·год. Часто перемагає постачальник, який пропонує найнижчу загальну вартість договору, проте це не завжди гарантує стабільність постачання.
- Замовники перевіряють відповідність документації учасника, наявність ліцензії, підписання договору з оператором системи передачі.
- Важливим критерієм є відсутність податкової заборгованості, відсутність судових справ, стабільність роботи на ринку.
- Замовники надають перевагу учасникам, які пропонують відтермінування платежів, гнучкі графіки розрахунків або відсутність авансу.
- Постачальники з позитивним досвідом участі у ProZorro, стабільними відгуками та високою надійністю мають перевагу.
- Коливання ринкових цін або запровадження нових тарифів напряму впливають на тендерні пропозиції.
- Попит на електроенергію влітку/взимку змінюється, що також може впливати на рівень конкуренції та ціни в тендерах.

Система Prozorro демонструє ефективне поєднання цифрових технологій та принципів відкритого врядування. Її запровадження забезпечило якісно новий рівень організації державних закупівель, у якому прозорість, рівність доступу та публічний контроль стали базовими засадами. Результати функціонування системи свідчать про її здатність підвищувати економічну ефективність закупівель, мінімізувати ризики зловживань і забезпечувати доступ до ринку широкому колу учасників, зокрема представникам малого та середнього бізнесу.

Застосування Prozorro стало важливою складовою реформування державного управління в Україні. Завдяки електронізації процесів та автоматичному моніторингу платформа дозволяє виявляти типові порушення, формувати базу доказів та впроваджувати аналітичний підхід до прийняття рішень. Отже, система не лише забезпечує прозорість окремих закупівель, а й служить інструментом системного вдосконалення закупівельної політики держави.

Станом на сьогодні, Prozorro залишається однією з найуспішніших цифрових ініціатив в Україні, що здобула міжнародне визнання як приклад ефективної реалізації принципу “everyone sees everything”. Її подальший розвиток, зокрема інтеграція з аналітичними платформами, впровадження механізмів прогнозування та розширення функціоналу для користувачів, відкриває нові можливості для посилення прозорості, підвищення конкуренції та запобігання корупції в публічних закупівлях.

РОЗДІЛ 2

АНАЛІЗ ДАНИХ СИСТЕМИ PROZORRO

2.1. Попередня обробка та аналіз даних

Для успішного прогнозування перемог учасників у закупівлях Prozorro необхідно ретельно відібрати та проаналізувати дані, які мають найбільший вплив на результат тендерів.

Для того, щоб мати можливість правильно робити прогнозування, потрібні історичні дані з детальної інформацією про кожен лот, учасників та організаторів. Тому спершу сформовано такий датасет див. рис. 2.1, який містить дані з участями постачальників у торгах.

	LotId	LotAmount	OrganizerOKPO	ParticipantOKPO	OrganizerRegion	LotCost	LotQu.	VAT	TenderType	supply..
1	15767807	105620.0000	34500133	39587271	Донецька область	340.00000	24000.000...	0	Фіксована ціна	11
2	15767807	105620.0000	34500133	42086719	Донецька область	340.00000	24000.000...	0	Фіксована ціна	11
3	15768984	674400.0000	39072140	36716332	Київська область	510.00000	140500.00...	1	Фіксована ціна	10
4	15768984	674400.0000	39072140	41427817	Київська область	510.00000	140500.00...	1	Фіксована ціна	10
5	15768984	674400.0000	39072140	42423070	Київська область	510.00000	140500.00...	1	Фіксована ціна	10
6	15768985	189600.0000	39072140	36716332	Київська область	340.00000	39500.000...	1	Фіксована ціна	10
7	15768985	189600.0000	39072140	41427817	Київська область	340.00000	39500.000...	1	Фіксована ціна	10
8	15768985	189600.0000	39072140	42423070	Київська область	340.00000	39500.000...	1	Фіксована ціна	10
9	15769162	468000.0000	39091833	41427817	Житомирська область	510.00000	90000.000...	1	Фіксована ціна	11
1..	15769162	468000.0000	39091833	42095943	Житомирська область	510.00000	90000.000...	1	Фіксована ціна	11
1..	15769237	81950.0000	05462449	39594527	Тернопільська область	340.00000	14900.000...	1	Фіксована ціна	11
1..	15769237	81950.0000	05462449	42145798	Тернопільська область	340.00000	14900.000...	1	Фіксована ціна	11
1..	15769965	3414214.0000	01990683	36716332	Донецька область	1700.00000	569036.00...	1	Фіксована ціна	11
1..	15769965	3414214.0000	01990683	37877430	Донецька область	1700.00000	569036.00...	1	Фіксована ціна	11
1..	15769965	3414214.0000	01990683	39726263	Донецька область	1700.00000	569036.00...	1	Фіксована ціна	11
1..	15769965	3414214.0000	01990683	42129720	Донецька область	1700.00000	569036.00...	1	Фіксована ціна	11
1..	15769965	3414214.0000	01990683	42129888	Донецька область	1700.00000	569036.00...	1	Фіксована ціна	11
1..	15769966	585786.0000	01990683	36716332	Донецька область	510.00000	86145.000...	1	Фіксована ціна	11
1..	15769966	585786.0000	01990683	37877430	Донецька область	510.00000	86145.000...	1	Фіксована ціна	11
2..	15769966	585786.0000	01990683	42129720	Донецька область	510.00000	86145.000...	1	Фіксована ціна	11
2..	15769966	585786.0000	01990683	42129888	Донецька область	510.00000	86145.000...	1	Фіксована ціна	11

Рисунок. 2.1 - Учасники закупівель електроенергії за період 2022-2025 рр

Джерело: розроблено автором на основі [Додаток А]

Датасет учасників торгів містить таку інформацію:

lot_id - унікальний номер лоту

organizer_okpo - ЄДРПОУ Організатора

participant_okpo - ЄДРПОУ Учасника

organizer_region - регіон організатора

publish_date - дата публікації

lot_amount - бюджет лоту

lot_quantity - об'єм закупівлю

VAT - наявність ПДВ

supply_duration - час поставки

lot_payment_period - період оплати

participant_price - ціна учасника

lot_cost - ціна участі в лоті

status_bid - статус пропозиції учасника

tender_type - тип тендеру

prop_speed - швидкість подачі пропозиції

meet_with_org - кількість зустрічей учасника з організатором

wins_count_org - кількість перемог у конкретного організатора

win_rate_org - % перемоги учасника у конкретного організатора

part_in_region - участі постачальника в регіоні, в якому проводився лот

wins_in_region - перемоги учасника в регіоні, в якому проводився лот

win_rate_in_region - %, перемог учасника в регіоні, в якому проводився лот

diff_price - абсолютне зниження ціни учасника

diff_price_percent - % зниження ціни учасника

avg_diff_price_by_supply - середнє зниження ціни учасника

avg_lot_payment_period - середній період оплати учасника у переможних торгах

avg_supply_duration - середній період поставки учасника у переможних торгах

avg_proposal_speed - середня швидкість подачі пропозиції учасника у переможних торгах.

В окремому наборі даних розміщена детальна інформація про участь постачальників у державних/публічних закупівлях електроенергії CPV 09310000-5 [10] через систему Prozorro за період 2022-2025 років.

Для проведення аналізу обрана ця категорія, оскільки вона є ключовою для

енергетичної безпеки країни та характеризується високим рівнем конкуренції. Аналіз цих даних дозволить виявити закономірності та фактори, що впливають на успіх участі в тендерах.

З даного набору даних виключено прямі договори, укладені між замовником та постачальником товарів/надавачем послуг/виконавцем робіт, за умови, що не проводиться процедура закупівлі [11].

Аналіз таких договорів не дозволить виявити закономірності та розробити ефективні моделі прогнозування, які б враховували переваги учасників. Тому, для забезпечення релевантності та повноти даних, увагу зосереджено виключно на конкурентних процедурах закупівель, де учасники змагаються за право укласти договір [12].

Для правильного використання методів машинного навчання, потрібно дослідити пропущені значення, що наявні у вхідному датасеті.

У наборі даних виявлено пропущені значення у трьох стовпцях: AvgLotPaymentPeriod, AvgSupplyDuration та AvgProposalSpeed, кожен з яких має 117 пропусків. Всі пропущені значення заповнюємо 0, оскільки немає інформації щодо середніх значень по переможним торгам учасників торгів.

Отримано більш детальну інформацію, щодо певних тенденцій та паттернів поведінки учасників та організаторів на ринку електроенергії.

Публікації закупівель впродовж року розподіляються нерівномірно (рис. 2.2). Загалом помітна сильна сезонність, компанії переважно закупають електроенергію з листопаду по січень, саме в цей період найбільша кількість лотів. Але впродовж року компанії в невеликій кількості докупляють електроенергію, винятком став період 2022 року після початку війни, компанії мінімально проводили закупівлі.

Держзамовників доволі багато - 11455, оскільки в умовах війни та блекаутів компанії мають функціонувати та можуть закуплятися багато. За проаналізований період в середньому кожен замовник проводив 2,56 лотів.

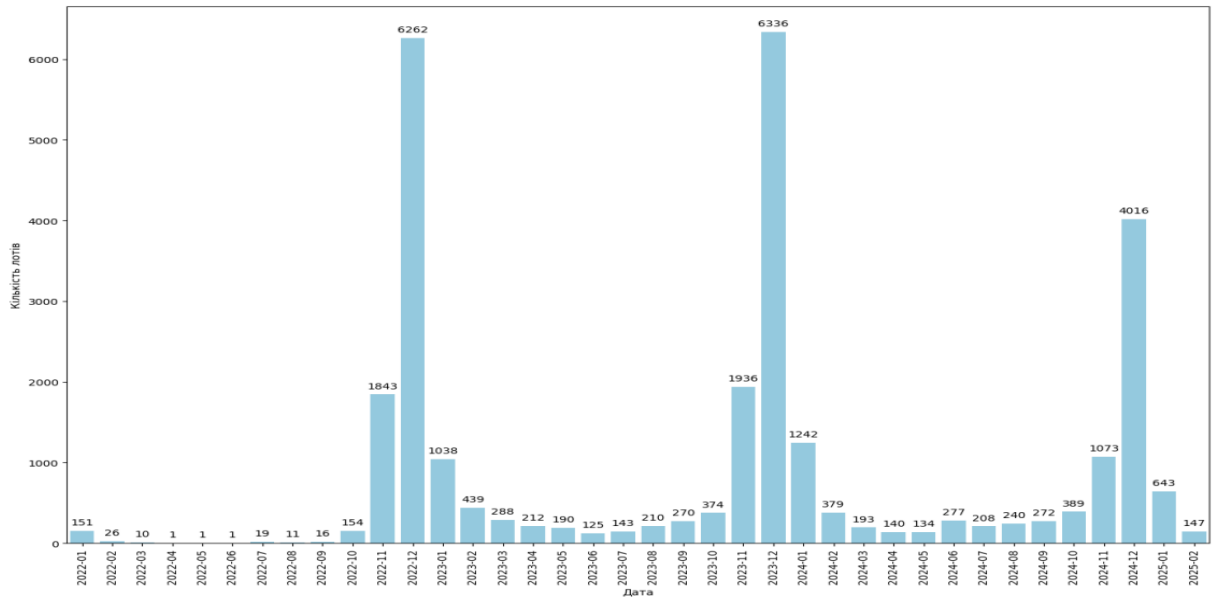


Рисунок 2.2 - Динаміка опублікованих лотів за 2022-2025 рр

Джерело: розроблено автором на основі [Додаток Б]

В табл 2.1 зображено організаторів, які проводили найбільшу кількість закупівель.

Таблиця 2.1 - Топ-10 організаторів за кількістю опублікованих лотів

Замовник	Організовано закупівель
КП "ЦЗО Івано-Франківської ОР"	92
КП "ЦЗО Харківської МР"	83
ДП "Укрводсервіс"	78
КП "ЦЗО Одеської ОР"	75
ДУ "Центр пробації"	69
ДУ "Професійні закупівлі"	66
КУ "Сумський обл. фонд підтримки підприємництва"	50
КП "Обласний центр екстреної медичної допомоги та медицини катастроф" Дніпропетровської ОР"	33

Джерело: розроблено автором на основі [Додаток Б]

Ринок електроенергії в Україні демонструє цікаву динаміку: лише 266 учасників беруть участь у 66 537 лотах, що в середньому становить близько 250 лотів на компанію. Цей факт свідчить про обмежену конкуренцію та потенційну можливість для нових гравців.

Проте, незважаючи на відносно невелику кількість учасників, розподіл участей, відображений на рис. 2.3, розкриває іншу картину. Топ-10 компаній акумулюють 24 563 участі, що становить майже 36,7% від загальної кількості. Це

підкреслює концентрацію ринку, де ключові гравці домінують, обмежуючи можливості для менших компаній та потенційних новачків.

Отже, хоча ринок може здаватися відкритим для нових учасників, реальна конкуренція зосереджена серед обмеженого кола лідерів. Це створює як можливості, так і виклики для тих, хто прагне увійти на ринок електроенергії в Україні.

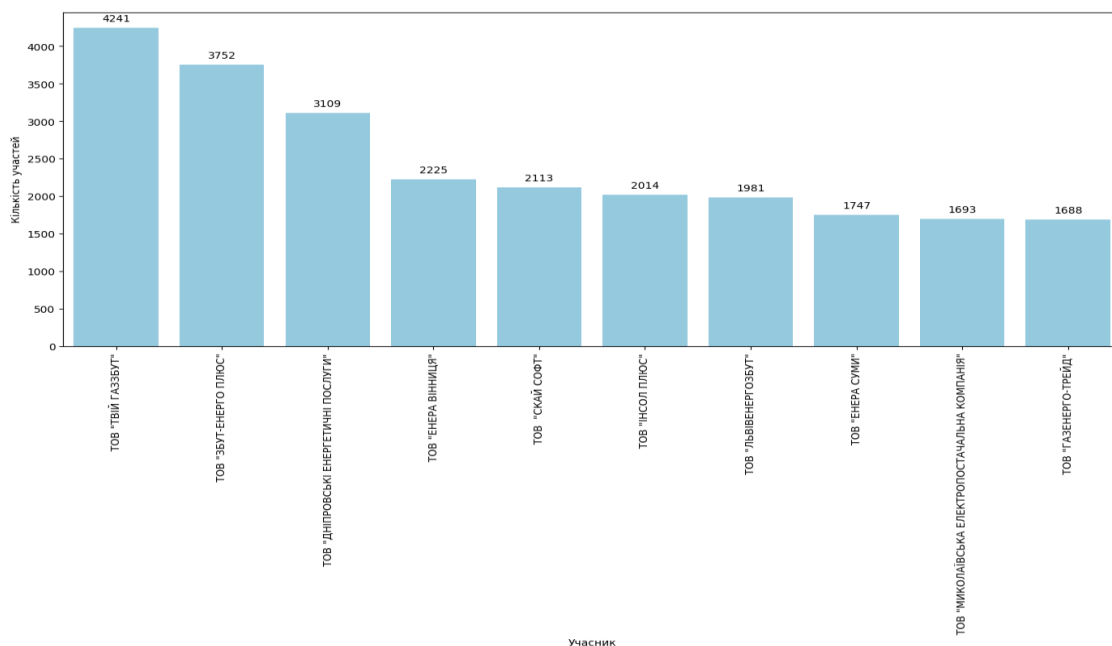


Рисунок 2.3 - Компанії за участю у закупівлях

Джерело: розроблено автором на основі [Додаток Б]

Розподіл перемог серед основних гравців ринку електроенергії в Україні є доволі рівномірним, явного лідера з великим відривом немає. Компанія ТОВ 'ДНІПРОВСЬКІ ЕНЕРГЕТИЧНІ ПОСЛУГИ' займає домінуючу позицію з часткою 18.2%, що не сильно, але все ж перевищує частки інших компаній.

Порівняно з розподілом учасностей, де перші 10 компаній охоплювали 36.7% від загальної кількості, розподіл перемог демонструє ще більшу концентрацію. Лідери ринку, зображені на рис. 2.4, разом охоплюють понад 60% ринку за кількістю перемог. Це підтверджує припущення про те, що невелика кількість компаній домінує на ринку.

Prozorro, демонструє широке розмаїття закупівель у різних бюджетних діапазонах, як показано на рис. 2.5. Шість основних категорій бюджетів дозволяють чітко візуалізувати розподіл лотів за вартістю.

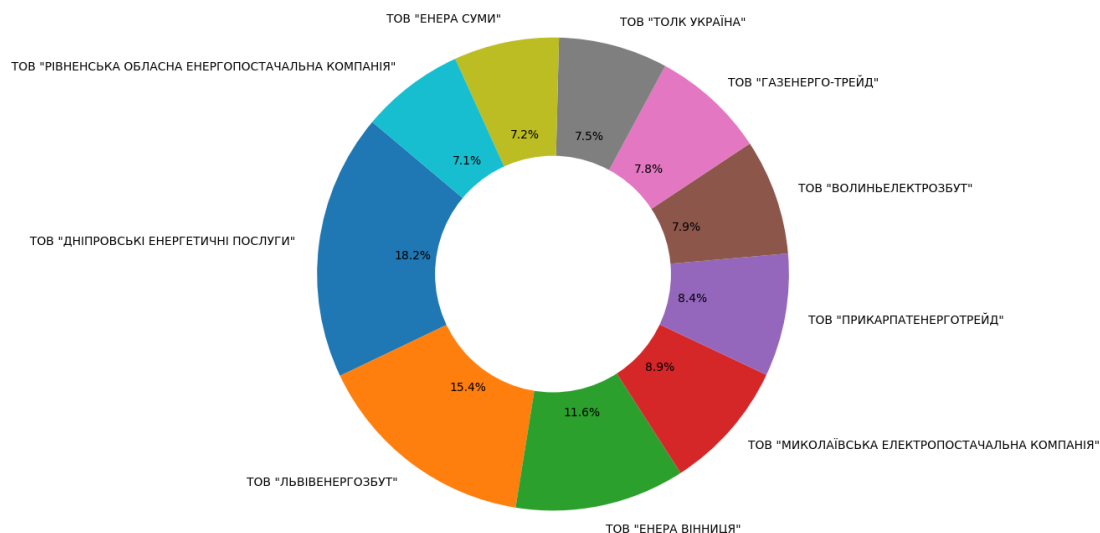


Рисунок 2.4 - Частка виграних лотів

Джерело: розроблено автором на основі [Додаток Б]

Аналіз даних виявляє, що найбільша концентрація лотів спостерігається в діапазоні від 200 000 до 1 000 000 грн. - 13894 закупівлі. Це свідчить про домінування закупівель середнього розміру, що, ймовірно, відображає типові потреби регіональних енергетичних компаній та муніципальних установ.

Цікаво відзначити, що діапазон понад 4 000 000 грн. також демонструє значну кількість лотів - 3622, що вказує на наявність великих проектів або стратегічних закупівель. Це може включати будівництво нових підстанцій, розширення мереж або закупівлю великих обсягів електроенергії для критичних потреб.

З іншого боку, категорії з нижчими бюджетами (менше 20 000 грн та 20 000 - 50 000 грн) мають незначну кількість лотів. Це свідчить про те, що дрібні закупівлі не користуються великим попитом на ринку електроенергії, тобто якщо компанії закупляються то одразу у великій кількості.

Такий розподіл бюджетів відображає структуру ринку електроенергії, де основна активність зосереджена в середньому та великому сегментах. Це важливо для розуміння стратегій учасників ринку, планування закупівель та оцінки потенційних можливостей для нових гравців. Також це вказує на те, що ринок електроенергії вимагає значних фінансових ресурсів для участі в тендерах, що може бути бар'єром для малих компаній.

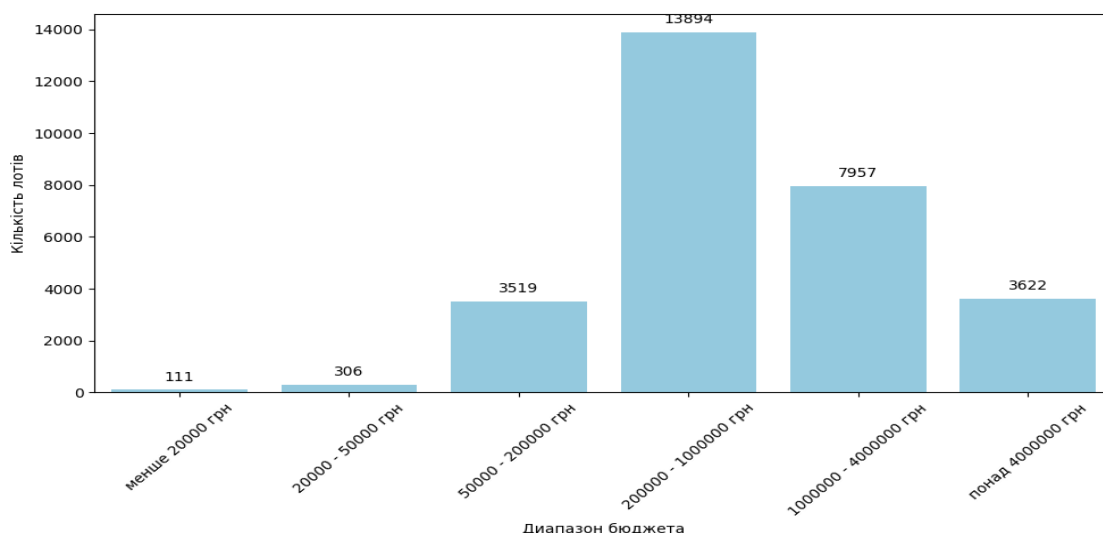


Рисунок 2.5 - Розподіл лотів за бюджетами закупівлі

Джерело: розроблено автором на основі [Додаток Б]

Розподіл сум проведених лотів за регіонами демонструє значну концентрацію закупівель у Київській області, яка лідирує з часткою 28.4%. регіоні (рис 2.6).

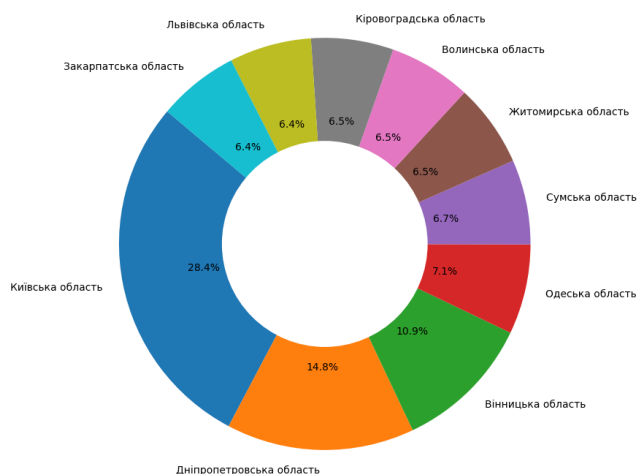


Рисунок 2.6 - Розподіл суми лотів за областями

Джерело: розроблено автором на основі [Додаток Б]

Дніпропетровська та Вінницька області також відзначаються значними частками, 14.8% та 10.9% відповідно, що може бути пов'язано з розвиненим промисловим та агропромисловим секторами.

Решта регіонів, включаючи Одеську, Сумську, Житомирську, Волинську, Кіровоградську, Львівську та Закарпатську області, мають відносно рівномірний розподіл часток, коливаючись в межах 6.4-7.1%, що свідчить про активне проведення закупівель електроенергії по всій країні, хоча й у менших обсягах

порівняно з лідерами.

Однією з проблем в даних є дисбаланс класів, який представляє собою значну різницю представлених класів, що виникає через внутрішні фактори, такі як природний розподіл даних [13].

Згідно з рис 2.7, 37127 разів компанії програвали, що становить 55.8% від загальної кількості ставок. Виграли компанії 29409 разів, що складає 44.2%. Це означає, що можна застосовувати стандартні методи машинного навчання без необхідності вживати спеціальних заходів для боротьби з дисбалансом класів.

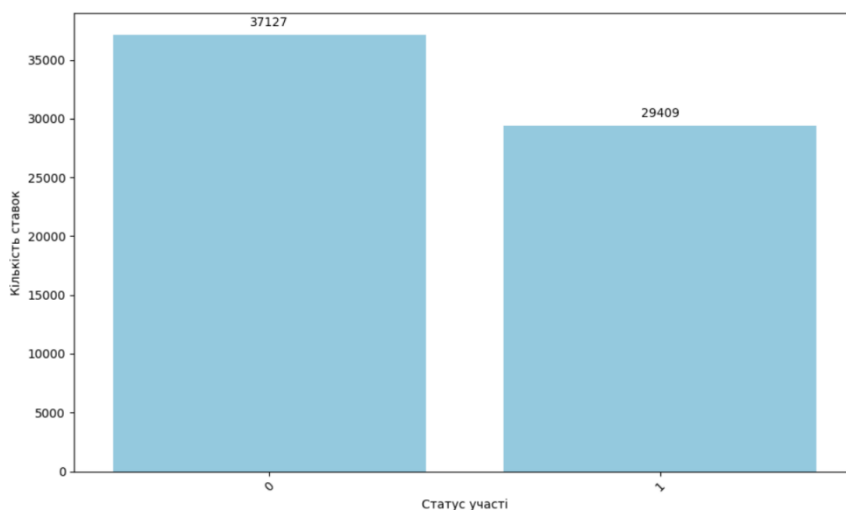


Рисунок 2.7 - Перевірка дисбалансу класів

Джерело: розроблено автором на основі [Додаток Б]

Кореляційна матриця Пірсона (рис. 2.8) відображає зв'язок між змінними.

По-перше, спостерігаються сильні позитивні кореляції між статусом пропозиції та перемогою учасника в організатора торгів, це свідчить, що минулі перемоги учасника у конкретного замовника дуже сильною впливають на результат торгів. Також присутня висока кореляція між відсотком перемоги у конкретному регіоні.

Високий зв'язок має бути між ціною та статусом пропозиції, оскільки у конкурентних торгах перемагають компанії з найменшою ціною пропозиції. Але між ціною, зміною ціни та статусом пропозиції дуже маленька кореляція.

Для подальшої підготовки даних необхідно перетворити категоріальні змінні на числові, щоб адаптувати їх до моделей машинного навчання, які використовують лише числові дані.

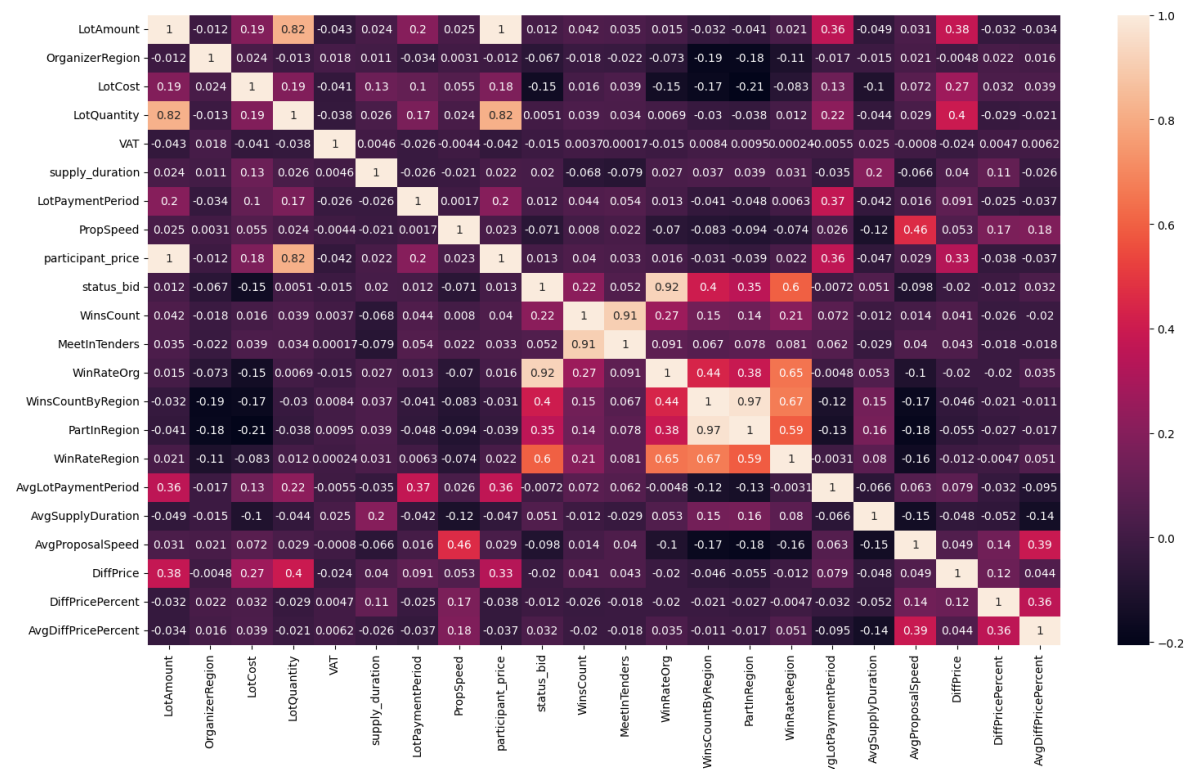


Рисунок 2.8 - Кореляційна матриця

Джерело: розроблено автором на основі [Додаток Б]

Для запобігання викривлення результатів моделювання числові ознаки необхідно стандартизувати - привести їх до єдиного масштабу. Це допоможе алгоритмам машинного навчання краще обробляти викиди і розсіювання даних.

2.2. Підходи до прогнозування перемоги учасників у державних закупівлях

Прогнозування результатів державних закупівель є актуальним завданням, спрямованим на оптимізацію процесів та ресурсів. Дослідження в цій сфері пропонують різні підходи до аналізу даних та передбачення ключових аспектів закупівель.

Одна з дослідницьких ідей полягає у вивченні впливу структурованих факторів на поведінку учасників тендерів та їхні цінові пропозиції [14]. В рамках

цього підходу аналізуються такі характеристики, як параметри самого тендеру (наприклад, місце проведення, час на підготовку), характеристики учасників (досвід, розмір) та особливості предмету закупівлі. Для моделювання та прогнозування рівня участі (кількості поданих заявок) та очікуваних цін застосовуються традиційні статистичні моделі (регресійний аналіз) та базові методи машинного навчання. Мета полягає у виявленні залежностей між цими факторами та результатами тендерів, що дозволяє передбачати конкурентне середовище та потенційні цінові рівні.

На противагу цьому, інша дослідницька ідея сфокусована на використанні неструктурованих даних, зокрема текстових описів тендерних оголошень [15]. Суть підходу полягає в тому, що спосіб формулювання умов та вимог у тексті оголошення містить важливу інформацію, яка може впливати на рішення потенційних учасників та, як наслідок, на результат закупівлі. Застосовуючи передові методи обробки природної мови (NLP) та глибокого навчання, дослідники аналізують зміст оголошень для прогнозування різноманітних показників результативності, таких як загальна кількість отриманих пропозицій, ймовірність наявності лише одного учасника, а також співвідношення між фінальною ціною контракту та його очікуваною вартістю. Цей підхід демонструє силу текстових даних та сучасних аналітичних інструментів для отримання прогностичних інсайтів.

Бачимо принципову різницю у підходах. Перший опирається на формалізовані атрибути закупівель та учасників і використовує кількісні моделі для прогнозування поведінкових аспектів. Інший зосереджується на змістовому наповненні текстових оголошень і застосовує NLP-методи для передбачення наслідків та ефективності закупівель, що можуть бути приховані у способі опису предмету торгів.

Також ключову роль у структуруванні та аналізі складних даних публічних закупівель відіграють онтології та семантичні графи. Завдяки їхньому застосуванню, можливості аналізу структурованих тендерних даних значно розширюються, інтегруючи машинне навчання для глибшого розуміння та

виявлення прихованих закономірностей [16]. Це створює основу для більш інтелектуальних систем моніторингу, що дозволяють ефективніше контролювати процес закупівель.

Особливо перспективними є *zero-shot* моделі для класифікації тендерів. Запропоновано модель для класифікації публічних тендерів за CPV-таксономією з використанням мовної моделі без попереднього навчання [17]. Цей підхід значно спрощує категоризацію нових тендерів, що є критично важливим для оперативного аналізу великих обсягів даних. Крім того, машинне навчання також успішно застосовується для виявлення потенційної змови у торгах інфраструктурних проєктів, що підкреслює універсальність цього підходу у виявленні аномалій [18].

Машинне навчання демонструє високу ефективність у виявленні корупційних схем та змов на торгах. Модель на основі ансамблю випадкових лісів дозволяє ідентифікувати підозрілі контракти за даними державних закупівель [19]. Запропоновано також інноваційний підхід до виявлення змов через згорткові нейронні мережі, які навчаються на візуалізаціях тендерних пропозицій [20].

Таким чином, інтеграція онтологій, семантичних графів та різноманітних моделей машинного навчання надає можливість для підвищення прозорості, ефективності та зниження ризиків корупції в публічних закупівлях по всьому світу.

У даній роботі розглядаються саме методи машинного навчання для вирішення задач класифікації.

2.3. Постановка задач класифікації

У сучасному цифровому середовищі обробка та аналіз великих обсягів даних набули широкого значення для прийняття data-driven рішень у різних сферах — від медицини до економіки.

Однією з ключових задач, що вирішуються в рамках машинного навчання, є задача *класифікації*, сутність якої полягає у віднесенні об'єктів до одного з

наперед визначених класів на основі їх параметрів.

Задачі класифікації мають багато практичних застосувань, зокрема:

- класифікація пацієнтів за типами захворювань на основі результатів аналізів чи медичних зображень;
- визначення кредитоспроможності клієнтів або виявлення шахрайства;
- сегментація клієнтів за різними ознаками поведінки або передбачення ймовірності здійснення покупки;
- класифікація мережових атак чи спам-повідомлень.

Класифікація відноситься до навчання з учителем, тобто потрібен набір даних, на якому наша модель буде навчатися та шукати закономірності.

Формально, задача класифікації визначається як пошук функції:

(2.1)

$$F: X \rightarrow Y$$

де X — простір входних ознак,

Y скінченна множина класів

Метою задачі класифікації є побудова цієї функції, яка для будь-якого нового входного об'єкта X , зможе з достатньою точністю передбачити мітку класу Y .

Класифікація ділиться на декілька типів:

Бінарна класифікація

Один з найпростіших та найпоширеніших типів задач класифікації в машинному навчанні, у якому кожен об'єкт відноситься до одного з двох можливих класів.

Формально задача класифікації може бути представлена наступною функцією:

(2.2)

$$X \rightarrow \{0,1\}$$

де X - простір входних ознак,

0,1 - два можливі класи

На вхід моделі подається певна кількість ознак, на основі них робиться передбачення, до якого класу буде віднесений вхідний вектор X .

На рис. 2.9 можна побачити приклад бінарної класифікації.

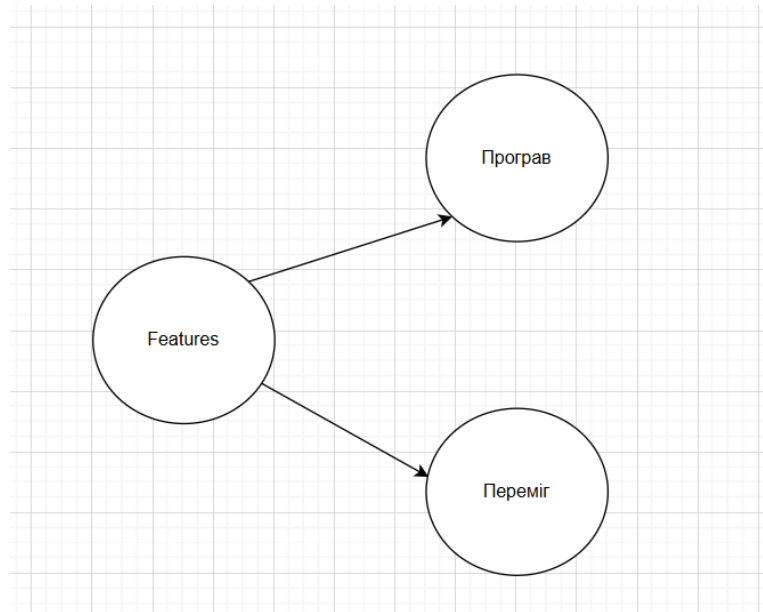


Рисунок 2.9 - Бінарна класифікація

Джерело: розроблено автором на основі [21]

Саме методи бінарної класифікації використовуються для прогнозування перемоги у тендерах.

Найбільш популярними моделями, які вирішують завдання класифікації, є логістична регресія, дерева рішень, випадковий ліс, XBoost та застосування нейронних мереж.

Багатокласова класифікація

Порівняно з попередньо розглянутим типом класифікації, в багатокласовій класифікації об'єкт може бути віднесений до одного з трьох і більше класів.

Тоді математично задача формулюється як побудова функції:

$$X \rightarrow \{0, 1, 2, \dots, K - 1\}$$

де X - простір вхідних ознак,

K - кількість класів

На рис. 2.10 зображений приклад розпізнавання друкованих цифр, на вхід подається зображення цифри, далі воно класифікується

(2.3)

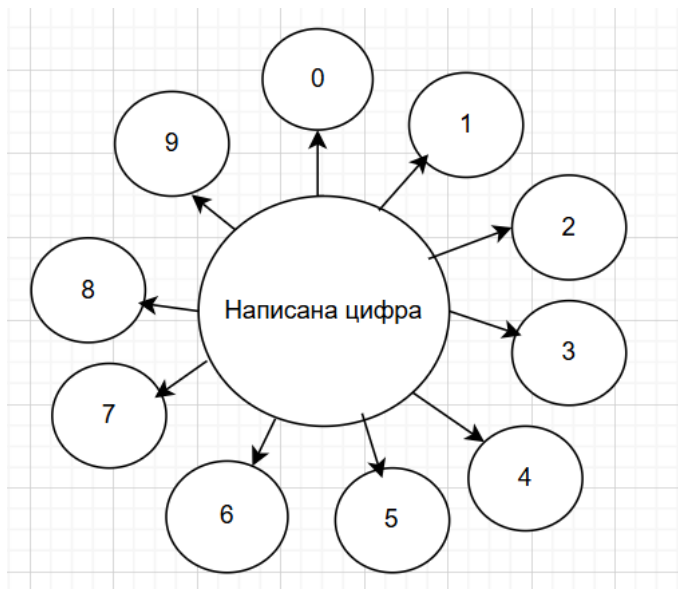


Рисунок 2.10 - Багатокласова класифікація

Джерело: розроблено автором на основі[22]

Багатокласова класифікація підтримується у моделях Decision Tree, Random Forest, XGBoost, а також застосовується у нейронних мережах типу багатошаровий перцептрон.

Мульти-лейблова класифікація

Тип класифікації, де кожному об'єкту може відповідати декілька класів одночасно.

Математична задача формулюється як:

$$X \rightarrow \{0,1\}^k \quad (2.4)$$

де X - простір входних ознак,

$\{0,1\}^k$ - вектор біток довжини K , де кожен елемент вказує на наявність або відсутність відповідного класу для об'єкта.

На рис. 2.11 зображено приклад того, як для певної статті можна пропоставляти теги, залежно від того, чи належить на основі змісту текст статті певній категорії.

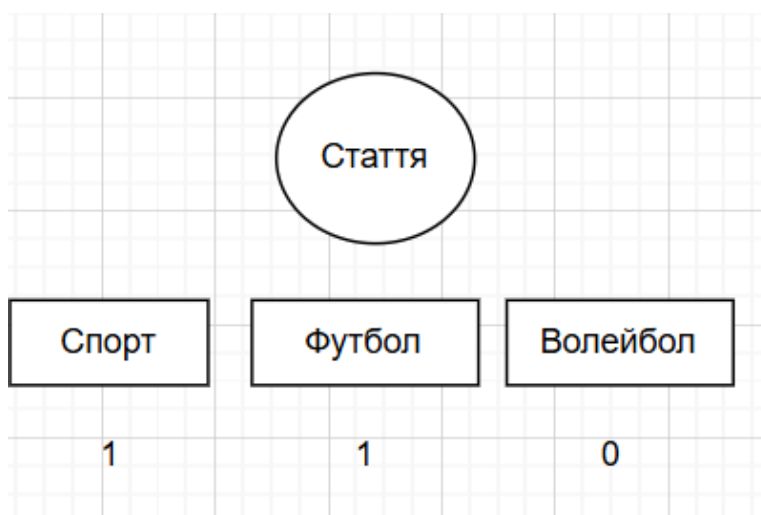


Рисунок 2.11 Мульти-лейблова класифікація

Джерело: розроблено автором на основі[23]

Класифікація як задача машинного навчання передбачає побудову моделей, здатних розпізнавати шаблони у даних та прогнозувати належність об'єктів до певних класів. Вона є особливо корисною у випадках, коли необхідно автоматизувати процес прийняття рішень на основі історичних даних. Завдяки різноманітним алгоритмам, таким як логістична регресія, дерева рішень чи нейронні мережі, можливо адаптувати підхід до конкретної предметної області. Зростаюча якість та доступність даних робить задачі класифікації все більш затребуваними для вирішення практичних проблем у державному, комерційному та науковому секторах.

2.4. Метрики для оцінювання якості моделей

Створення та навчання моделей машинного навчання для прогнозування перемоги учасників у державних закупівлях електроенергії включає в себе вибір відповідних алгоритмів, вибір потрібних ознак та оцінку ефективності моделей, що дозволить розробити надійний інструмент для прогнозування їхнього успіху.

Для оцінювання якості різних моделей застосовують різні критерії якості.

Одним з показників оцінювання продуктивності моделі класифікації є

accuracy, що забезпечує швидкий аналіз того, наскільки правильно модель прогнозує результат:

(2.5)

$$Accuracy = \frac{\text{Кількість правильних передбачень}}{\text{Загальна кількість передбачень}}$$

Найбільш відомим методом є матриця плутанини - це таблиця, яка показує співвідношення правильних і неправильних прогнозів. Вона має таку назву, оскільки показує, де модель «заплутується» під час прогнозування. Матриця містить 4 показники, які використовуються для інтерпретації результатів.

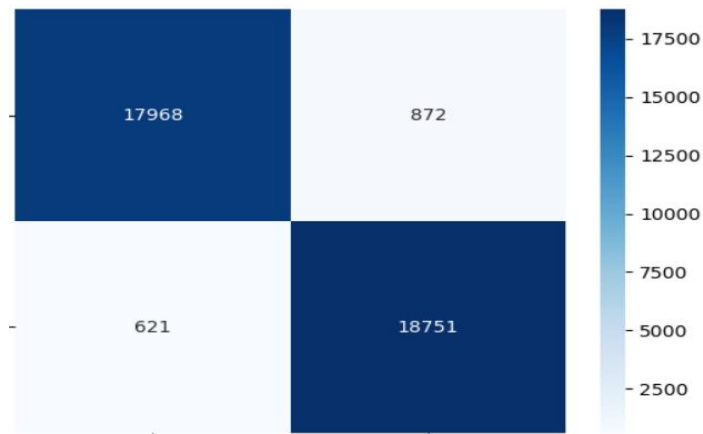


Рисунок 2.12 – Матриця плутанини

Джерело: розроблено автором на основі [Додаток Б]

TP -об'єкт є класом 1, і алгоритм його ідентифікує як клас 1.

FP - об'єкт є класом 0, алгоритм його ідентифікує як клас 1

TN - об'єкт є класом 0, і алгоритм його ідентифікує як клас 0.

FN -об'єкт є класом 1, алгоритм його ідентифікує як клас 0.

На основі матриці плутанин можна визначити точність моделі.

(2.6)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Формула показує, як враховуються правильно класифіковані позитивні та негативні об'єкти, а також неправильно класифіковані об'єкти. Вона більш детальна і дозволяє краще зрозуміти, як точність пов'язана з результатами класифікації.

Метрика *Precision*, повідомляє, скільки позитивних прогнозів, зроблених

моделлю, насправді правильні.

(2.7)

$$Precision = \frac{TP}{TP + FP}$$

Також однією з основних метрик є *Recall*, яка показує відношення істинно позитивних результатів до суми істинно позитивних результатів та хибно негативних результатів. Іншими словами, це вимірює, наскільки повно алгоритм виявив позитивні випадки.

(2.8)

$$Recall = \frac{TP}{TP + FN}$$

F1-Score є середнім гармонійним показником точності та запам'ятовування моделі. Обчислюється за формулою

(2.9)

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Показник *F1-Score*, ближчий до 1, вказує на високу точність. Однак досягнення *F1* оцінки 1 рідко зустрічається, це може свідчити про перенавчання тренувальних даних.

Крива *AUC-ROC* є важливим інструментом, який використовується для оцінки ефективності моделей бінарної класифікації. Він будує графік між істинним позитивним результатом і помилковим позитивним результатом при різних порогових значеннях, показуючи, наскільки добре модель може розрізнити два класи, наприклад позитивні та негативні результати.

AUC вимірює площу під кривою *ROC*. Більше значення *AUC* вказує на кращу продуктивність моделі, оскільки передбачає більшу здатність розрізняти класи. Значення *AUC* 1,0 вказує на ідеальну продуктивність, тоді як 0,5 означає, що це випадкове вгадування [24].

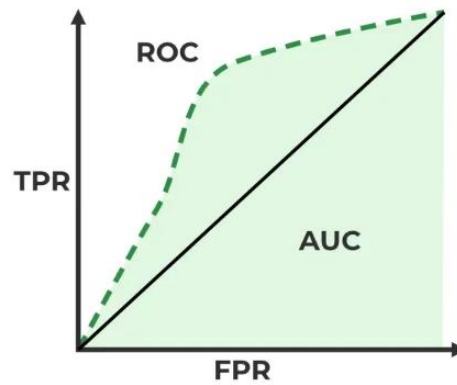


Рисунок 2.13 - ROC-AUC Крива[24]

Cross-Validation - це статистичний метод, який використовується в машинному навчанні для оцінки того, наскільки добре модель працює на незалежному наборі даних. Це передбачає поділ доступних даних на кілька підмножин, використання однієї з них як набору перевірки та навчання моделі на решті складок. Цей процес повторюється кілька разів кожного разу, використовуючи іншу складку як набір перевірки. Результати кожного етапу перевірки усереднюються для отримання більш надійної оцінки моделі [25].

Метрики оцінювання якості моделей класифікації відіграють ключову роль у виборі найбільш ефективного алгоритму для розв'язання конкретного завдання. Вони надають можливість не лише оцінити модель на тестових вибірках, а й виявити її слабкі місця та оптимізувати параметри. Показники точності (accuracy), повноти (recall), точності передбачень (precision), F1-міри та площі під ROC-кривою (AUC-ROC) дозволяють комплексно оцінити здатність моделі розрізняти класи, особливо у випадках з нерівномірним розподілом класів. Вони є необхідним інструментом для побудови моделей, здатних надійно працювати в умовах реальних задач.

Вибір відповідної метрики залежить від особливостей задачі та наслідків помилкових класифікацій. Саме аналіз метрик дозволяє уникнути перенавчання, покращити якість прогнозування та забезпечити надійність моделі при застосуванні до нових даних.

РОЗДІЛ 3

РЕАЛІЗАЦІЯ ТА ОЦІНЮВАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ПЕРЕМОГИ УЧАСНИКІВ У СИСТЕМІ PROZORRO

3.1. Методи машинного навчання для вирішення задач класифікації

Машинне навчання є важливою складовою сучасного інформаційного світу, що забезпечує здатність комп'ютерних систем навчатися по прикладам, виявляти закономірності і навіть приймати рішення. Сучасні методи успішно застосовуються для вирішення широкого спектру завдань.

Наприклад, регресія дозволяє передбачати числові значення на основі наявних ознак, кластеризація, для поділу об'єктів на групи без попереднього знання про їх приналежність до певних класів.

Значну роль відіграють методи зниження розмірності, які допомагають спростити аналіз високовимірних даних, а також побудова рекомендаційних систем, що дозволяють прогнозувати вподобання користувачів.

У цьому розділі розглядаються основні методи для вирішення задач класифікації.

Наївний баєсівський класифікатор

Статистичний метод класифікації, який базується на теоремі Байєса з припущенням про незалежність ознак [26].

Теорема Баєса формулюється так:

$$P(C_k|x) = \frac{P(C_k)P(x|C_k)}{P(x)} \quad (3.1)$$

$P(C_k|x)$ - апостеріорна ймовірність класу C_k при заданому векторі ознак x ,

$P(x|C_k)$ - ймовірність спостереження x за умови, що клас - C_k ,

$P(C_k)$ - апіорна ймовірність класу C_k ,

$P(x)$ - повна ймовірність спостереження x для всіх класів.

Припущення, на якому базується наївний баєсівський класифікатор, полягає в тому, що ознаки є незалежними одна від одної. Це означає, що для кожної ознаки X_i , ймовірність спостереження $X = (X_1, X_2, X_3, \dots, X_n)$ за умови класу C_k , можна розкласти як добуток ймовірностей для кожної окремої ознаки:

(3.2)

$$P(C_k|x_1 \dots x_n) = P(C_k)P(x_1|C_k) \dots P(x_n|C_k) = \prod_n P(x_i|C_k)$$

Цей підхід дозволяє ефективно розраховувати ймовірності навіть у високовимірних просторах, що робить його придатним для задач, де кількість ознак значно перевищує кількість прикладів. Приклад моделі зображений на рис.3.1.

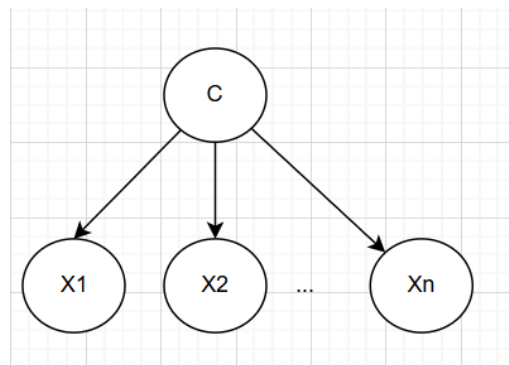


Рисунок 3.1 - Наївна баєсівська модель

Джерело: розроблено автором на основі [26]

Змінна C представляє собою мітку, яку необхідно передбачити для кожного об'єкта. Ознаки $X_1, X_2, \dots, X_n$ - це випадкові величини, які описують характеристики об'єкта та використовуються у процесі класифікації.

Наївний баєсівський підхід базується на припущенні про умовну незалежність ознак: якщо відомо значення змінної класу C , то всі ознаки X_i вважаються незалежними між собою. Проте в реальних умовах це припущення рідко виконується, адже між ознаками часто існують залежності.

Розглянемо приклад нашої майбутньої моделі, яка прогнозує, переможець тендеру. До ознак можна віднести: тип товару або послуги, стартову ціну, термін виконання, кваліфікацію учасників, а також історію попередніх тендерів компанії.

Змінна класу С може набувати двох значень: виграв або програв.

У цьому разі ознаки є взаємозалежними: наприклад, якщо тендер стосується будівельних матеріалів, то, ймовірно, що ціна буде вищою, а кваліфікація учасників може вимагати більш серйозного досвіду у будівельній сфері. Також, якщо компанія вже має позитивну історію виграних тендерів, це збільшує ймовірність виграшу в майбутньому тендері, що залежить від попередніх результатів. Отже, припущення про повну незалежність ознак є порушеним, оскільки між ознаками існують залежності.

Незважаючи на це, наївний баєсівський класифікатор демонструє високу ефективність у багатьох практичних задачах, зокрема — у фільтрації спаму, медичній діагностиці та аналізі текстів. Це пояснюється тим, що навіть за умов порушення припущення незалежності, оцінки ймовірностей залишаються досить інформативними для прийняття рішень.

Логістична регресія

Типовий метод машинного навчання з учителем, який вирішує задачі бінарної класифікації шляхом прогнозування ймовірності результату.

У разі, якщо вихідна змінна приймає лише два значення, логістична регресія найкраще підходить під постановку даної задачі.

Логістична регресія, використовується для оцінки ймовірності того, чи належить участь до класу перемога або програш. Пороговим значенням вважається - 0,5. Якщо ж ймовірність більша за 0,5, модель передбачає, що об'єкт належить до позитивного класу, а якщо вона менша за 0,5, вона передбачає, що він належить до негативного класу[27]

Логістична функція або як її ще називають - сигмоїдна функція розраховується за формулою:

(3.3)

$$f(x) = \frac{1}{1 + e^{-x}}$$

Переваги логістичної регресії:

- Коефіцієнти моделі відображають вплив кожної ознаки на логарифм

шансів результату, що дозволяє зрозуміти, як кожна ознака впливає на ймовірність результату.

- Логістична регресія є ефективним алгоритмом для завдань бінарної класифікації, тобто завдань, де потрібно передбачити один з двох можливих результатів.
- Логістична регресія не лише передбачає клас, але й оцінює ймовірність того, що об'єкт належить до певного класу.
- Це дає змогу використовувати модель для оцінки ризиків та прийняття рішень на основі ймовірностей.

Недоліки логістичної регресії:

- Логістична регресія припускає, що існує лінійна залежність між ознаками та логарифмом шансів результату.
- Логістична регресія може бути чутливою до викидів, тобто до аномальних значень у даних.
- Логістична регресія призначена для бінарної класифікації, тобто для завдань не більше двох класів.
- Як і будь-яка модель, логістична регресія може бути схильною до перенавчання, якщо модель занадто складна для наявних даних.

Метод опорних векторів

Керований алгоритм машинного навчання, який загалом використовують для класифікації. За допомогою даного методу розв'язуються задачі бінарної класифікації.

Основна ідея полягає в тому, щоб знайти гіперплощину, яка найкраще розділяє класи даних, потім цю гіперплощину можна використовувати для класифікації нових точок даних на основі їхнього положення відносно неї.

Зокрема, метод опорних векторів шукає поділяючу поверхню, максимально віддалену від будь-яких точок даних. Відстань між цією поверхнею і найближчою точкою даних називається зазором класифікатора. У методі опорних векторів обов'язково маєтись на увазі, що вирішальна функція цілком визначається підмножиною даних, що впливають на положення роздільника. Ці точки

називаються опорними векторами, тому що у векторному просторі точку можна розглядати як вектор між початком координат і цією точкою. На рис. 3.2 зображено приклад методу опорних векторів.

В алгоритмі будується кожен елемент вибірки даних, як точку в n -мірному просторі, де n - кількість ознак, які містяться у тренувальному наборі, кожен з елементів якого має власну координату у цьому просторі.

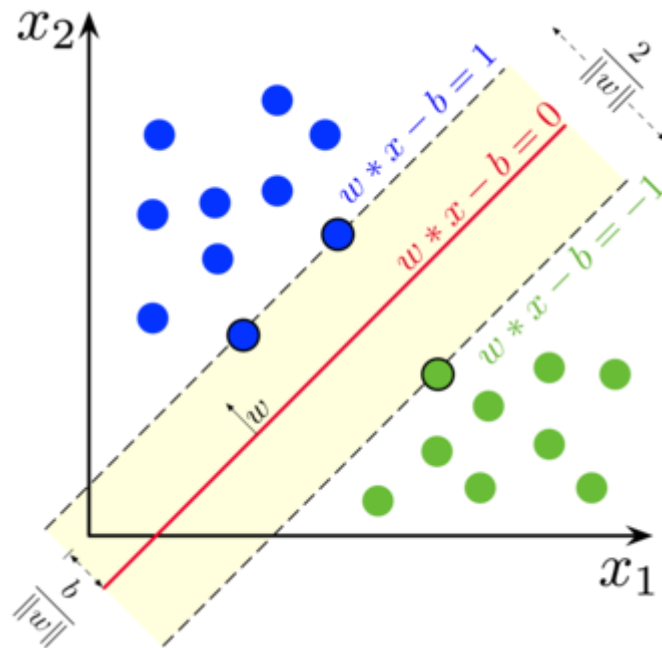


Рисунок 3.2 - Опорні вектори [28]

Метод опорних векторів будує деяку гіперплощину або їх набір у просторі тренувальних даних, які можуть бути використані для класифікації або виявлення викидів. Слід зазначити, що розмірність таких гіперплощин буде складати $n - 1$, де n - кількість об'єктів у навчальній вибірці.

Площину можна записати так:

$$w * x - b = 0$$

w — вектор ваг, x — вектор ознак, b — зміщення.

З геометричної точки зору, відстанню між цими двома гіперплощинами є:

$$L = \frac{2}{\|w\|}$$

Якщо вибірку можна лінійно розподілити, то є можливість побудувати

дві паралельні площини, які поділять простір на дві частини, які будуть відповідати двом класам, в такий спосіб, щоб максимальна різниця між довжинами векторів була якомога більшою. Область, яка обмежена площинами, називається поділом, а максимально роздільна площина є гіперплощиною, яка знаходиться посередині між ними. Вони описуються системою рівнянь:

(3.6)

$$\begin{cases} \bar{w} * \bar{x} - b = -1 \\ \bar{w} * \bar{x} - b = 1 \end{cases}$$

Тож для її максимізації потрібно мінімізувати $\|w\|$, за умови виконання двох нерівностей:

(3.7)

$$\begin{cases} \bar{w} * \bar{x}_i - b \geq 1, \text{ якщо } y_i = 1 \\ \bar{w} * \bar{x}_i - b \leq -1, \text{ якщо } y_i = -1 \end{cases}$$

Мінімізувати $\|w\|$ за умови $y_i(\bar{w}\bar{x}_i - b)$, для всіх $i = 1, \dots, n$.

Важливим покращенням алгоритму опорних векторів є заміна скалярного добутку на нелінійну ядрову функцію, що дає змогу поширити даний метод на вибірки з лінійно нерозподіленими класами, яких більшість у сучасних умовах.

Найбільш поширеними функціями ядра є:

- Полімііальна

(3.8)

$$k(x_i, x_j) = (x_i * x_j + 1)^d$$

- Гаусова радіальна базисна функція

(3.9)

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

- Сигмоїдна(гіперболічний тангенс)

(3.10)

$$k(x_i, x_j) = \tanh(kx_i * x_j + c) \text{ для } k > 0, c > 0$$

Ефективність методу опорних векторів сильно залежить від підбору

параметрів. Тобто береться кожен можливий варіант параметрів, запускають його на існуючих даних, і обираються ті параметри що показують найкращу точність.

Дерева рішень

Дерево рішень є фундаментальним алгоритмом машинного навчання з учителем, що використовується для задач класифікації. Цей метод створює деревоподібну модель, яка відображає рішення та їх можливі наслідки.

Метою використання цього алгоритму є створення навчальної моделі, яка використовується для прогнозування класу або значення цільової змінної шляхом вивчення простих правил прийняття рішень, отриманих з навчальних даних. У деревах рішень прогнозування мітки класу для запису починається з кореня дерева. В результаті порівнюється значення кореневого атрибуту з атрибутом запису. На основі порівняння відбувається пересування по гілці, яка відповідає обраному значенню і переходить до наступного вузла. Побудова дерева рішень включає рекурсивне розбиття даних на основі ознак [29].

Основними компонентами дерева рішень є:

- Вузли , що представляють атрибути або ознаки, на основі яких приймаються рішення.
- Гілки - відображають можливі значення або діапазони значень атрибутів.
- Листки - представляють кінцеві результати або прогнози.

Дерево починається з кореневого вузла, який представляє найбільш важливий атрибут. Кожен вузол розбивається на гілки, що ведуть до інших вузлів або листків, утворюючи ієрархічну структуру. Процес розбиття триває до тих пір, поки не буде досягнуто листового вузла, який представляє кінцевий результат.

Існує кілька алгоритмів для побудови дерев рішень, включаючи ID3, C4.5 та CART. Ці алгоритми використовують різні критерії для вибору найкращого атрибута для розбиття на кожному вузлі.

Найбільш поширеними критеріями для вибору найкращого атрибута розбиття є :

- Ентропія - вимірює невпорядкованість або неоднорідність набору даних.

- Приріст інформації - вимірює зменшення ентропії після розбиття на основі атрибута.
- Індекс Джині - вимірює ймовірність неправильної класифікації випадково обраного елемента

Мета полягає в тому, щоб вибрати атрибут, який максимізує приріст інформації або мінімізує індекс Джині, створюючи найбільш однорідні підвузли.

Дерева рішень мають кілька переваг:

- Легко зрозуміти та інтерпретувати, що робить їх цінними для аналізу даних та прийняття рішень.
- Можуть обробляти як категорійні, так і числові дані.
- Не потребують нормалізації або масштабування даних.
- Здатність обробляти пропущені дані: Можуть обробляти набори даних з пропущеними значеннями

Однак дерева рішень також мають деякі недоліки:

- Схильні до перенавчання, особливо при побудові глибоких дерев.
- Невеликі зміни в даних можуть призвести до значних змін у структурі дерева.
- Можуть створювати різні дерева для різних вибірок одного й того ж набору даних.

3.2. Вибір програмного забезпечення

Для реалізації завдань класифікації обрано мову програмування Python, що має доволі простий синтаксис в порівнянні з такими мовами як C/C++, Java, Javascript або R.

Використання Python охоплює різні сфери застосування від веб-розробки, тестування до аналізу даних, машинного навчання та створення великих кросплатформених систем.

Основні переваги:

- Інтуїтивно зрозумілий синтаксис, що дозволяє швидко освоїти мову та зосередитися на вирішенні задачі, а не на складнощах програмування.
- Наявність підтримки великої кількості бібліотек для аналізу, обробки та візуалізації даних. Вбудовані моделі для машинного навчання та інтерпретації результатів.
- Велика кількість готових рішень, що дозволяють швидко розробляти прототипи та експериментувати з різними моделями та їх варіаціями
- Постійно розвивається та доповнюється новими можливостями, що робить її актуальною для новітніх тенденцій в галузі машинного навчання і не тільки.

Недоліки:

- Python є інтерпретованою мовою, що означає, що код виконується по рядках, а не компілюється в машинний код. Це може призвести до повільнішого виконання порівняно з компільованими мовами, такими як C++ або Java.
- Глобальне блокування інтерпретатора обмежує виконання потоків, що може ускладнити використання багатопоточності для підвищення продуктивності на багатоядерних процесорах. Це може бути обмеженням для задач, що вимагають паралельного виконання.
- Використовує динамічну типізацію, що означає, що типи змінних визначаються під час виконання, цей підхід може призвести до повільнішого часу кодування через вилучення кожної змінної під час виконання.

Для виконання чисельних обчислень застосовано бібліотеку *NumPy*, яка забезпечує велику підтримку багатомірних масивів, матриць та велику кількість математичних функцій для операцій з ними

Створена на основі *NumPy* бібліотека *Pandas* основною структурою для роботи з даними є *Dataframe*, який складений із одновимірних масивів, об'єднаних

під однією назвою. Основна перевага, зображення даних у табличному вигляді за допомогою рядків та стовбців.

Окрім цього, Pandas також дозволяє читання і запис даних, наприклад, таблиць Csv або Excel, групування даних та навіть має вбудовані засоби візуалізації [30]

Matplotlib - це фундаментальна бібліотека для візуалізації даних вона надає широкий набір інструментів для створення статичних, анімованих та інтерактивних візуалізацій. Дозволяє створювати різні типи графіків, дозволяючи низькорівнево контролювати побудову складних візуалізацій.

Seaborn - це бібліотека для візуалізації даних, побудована на основі *Matplotlib*. *Seaborn* надає високорівневий інтерфейс для створення графіків, що спрощує процес візуалізації складних даних.

За допомогою цих інструментів можна створювати теплові карти, діаграми парних зв'язків та графіки розподілу. Також *Seaborn* пропонує вбудовані теми та стилі, що роблять графіки більш естетично привабливими.

Оскільки *Seaborn* будується на основі *Matplotlib*, можна використовувати його можливості для налаштування та доопрацювання графіків, створених за допомогою *Seaborn*.

Scikit-learn - це бібліотека для машинного навчання. Вона надає можливість просто та ефективно використовувати інструменти для вирішення задач класифікації, кластеризації, регресії, що базуються на сучасних алгоритмах. Також надає можливість передоброби даних, очищення та перекодування категоріальних змінних. Містить інструменти для інтерпретації, оцінки моделей та засоби для покращення гіперпараметрів [25].

Програмне забезпечення для машинного навчання виконує ключову роль у перетворенні сирих даних на аналітичні висновки, забезпечуючи зручні інтерфейси для моделювання, обробки даних та оцінювання результатів. Його використання дозволяє дослідникам і практикам швидко реалізовувати алгоритми, масштабувати рішення та адаптувати їх до конкретних завдань класифікації чи прогнозування. Крім того, інтеграція цих інструментів у

прикладні аналітичні середовища дозволяє швидко адаптувати моделі до нових завдань, забезпечуючи гнучкість і масштабованість у процесі прийняття рішень.

3.3. Прогнозування перемоги учасників

Логістична регресія без налаштування параметрів та ознак

Модель без відбору кореляційних ознак і застосування будь-яких параметрів, розглядається як модель за замовчанням.

Представлені у табл. 3.1 метрики демонструють невисоку загальну точність у 67%, тобто правильно спрогнозовано такий відсоток значень. Модель визначає лише 55% правильних значень, майже половину, що не є прийнятним значенням.

Таблиця 3.1 - Метрики для оцінки базової логістичної моделі

Метрика	Значення
Accuracy	0.67
Precision	0.67
Recall	0.55
ROC-AUC	0.70
F1	0.60

Джерело: розроблено автором на основі [Додаток Б]

Логістична регресія з підібраними ознаками

Для покращення точності моделі потрібно відібрати ті ознаки, які будуть прогнозувати цільову змінну з найбільшою точністю. Використовуючи перебір всіх пар параметрів порівнюємо точність кожної моделі і отримаємо найкращі ознаки з найбільш можливою точністю моделі.

В табл. 3.2 відображено метрики оцінки моделі, вона демонструє дуже високу якість класифікації. Усі метрики мають значення вище 0.9, що свідчить про високу точність, повноту та здатність до розрізнення.

Значення F1-міри 0.99 вказує на те, що модель має відмінний баланс між precision та recall.

Таблиця 3.2 - Метрики для оцінки моделі з підібраними ознаками

Метрика	Значення
Accuracy	0.93
Precision	0.93
Recall	0.92
ROC-AUC	0.92
F1	0.99

Джерело: розроблено автором на основі [Додаток Б]

Логістична регресія з налаштуванням гіперпараметрів

За основу взято попередню модель, яка вже показала високу точність прогнозування, але з вибором гіперпараметрів для більшої повноти картини.

Для пошуку найкращих параметрів використано GridSearch, що виконує вибір моделі та налаштування гіперпараметрів одночасно, навчаючи модель для кожної комбінації гіперпараметрів, а потім обирає модель, яка працює найкраще.

Таблиця 3.3 Метрики для оцінки моделі з налаштованими параметрами

Метрика	Значення
Accuracy	0.93
Precision	0.90
Recall	0.96
ROC-AUC	0.99
F1	0.93

Джерело: розроблено автором на основі [Додаток Б]

Модель продемонструвала значне покращення здатності виявляти потенційних переможців що є критично важливим у державних закупівлях, де важливо не пропустити саме кандидата переможця.

Також, модель стала набагато краще розрізняти переможців від тих, хто не

виграв за рахунок збільшення показника ROC-AUC. Це свідчить про те, що налаштування гіперпараметрів зробило модель більш чутливою до ключових факторів, що впливають на перемогу в тендері.

Однак, ці покращення відбулися за рахунок невеликого зменшення точності. Це означає, що модель стала дещо частіше помилково прогнозувати перемогу учасників, які насправді не виграли. У контексті державних закупівель, де важлива не лише повнота, але й точність, це може мати значення.

Дерева рішень з випадково обраними параметрами моделі

Для створення базової моделі використано такі параметри при побудові моделі для вибору найкращого атрибута розбиття був обраний критерій Джині. Максимальна глибина навчання обмежена 5 рівнями, це має допомогти запобігти перенавчанню моделі. Також для того аби модель не тренувалась тривалий час, обмежено її ознаки до 8, вони обираються випадковим шляхом. Параметри обрано випадковим чином, оскільки без вказаних параметрів модель перенавчалась, та показувала 100% точність навчання при параметрах за замовченням.

На основі табл. 3.4 можна зробити висновки, що модель показує високу ефективність. Загальна точність у 98% свідчить про здатність моделі правильно класифікувати більшість випадків. Високі значення Precision 0.97 та Recall 0.99 вказують на збалансованість моделі, яка однаково добре справляється з виявленням як переможців, так і тих, хто не переміг. Показник ROC-AUC 0.99, демонструє майже ідеальне розрізнення між класами. Модель відображає типовий приклад перенавчання.

Таблиця 3.4 Оцінка якості моделі дерева рішень

Метрика	Значення
Accuracy	0.98
Precision	0.99
Recall	0.98
ROC-AUC	0.99
F1	0.98

Джерело: розроблено автором на основі [Додаток Б]

Спробуємо застосувати GridSearch для підбору оптимальних параметрів. Для списку параметрів додаємо всі можливі варіанти для покращення моделі. Згідно рис. 3.3, 100% точність моделі, свідчить про її перенавчання.

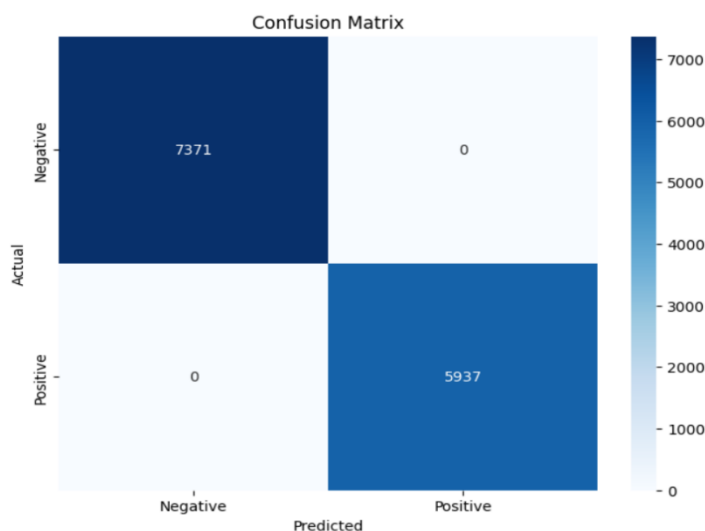


Рисунок 3.3 - Матриця плутанини для дерева рішень з підібраними параметрами моделі

Джерело: розроблено автором на основі [Додаток Б]

Останній спосіб, який буде використаний аби запобігти перенавчанню, це простий вибір ознак з найбільшою кореляцією з цільовою змінною.

Таких ознак 4 - кількість перемог учасника у минулих торгах організатора, % цих перемог, кількість участей в регіоні організатора та його кількість перемог у цьому регіоні. Також п'ятою змінною для подальшої можливості використання моделі на реальних прикладах, має бути ціни учасника, оскільки це дуже вагома змінна, зниження якої в реальних закупівлях збільшує шанси на перемогу.

Тепер видно, що модель може помилятись і виглядає більш надійною для використання на реальних закупівлях.

Таблиця 3.5 - Оцінка якості моделі дерева з підібраними параметрами

Метрика	Значення
Accuracy	0.94
Precision	0.93
Recall	0.92
ROC-AUC	0.99

F1	0.93
----	------

Джерело: розроблено автором на основі [Додаток Б]

Метод опорних векторів

Модель демонструє вкрай нерівномірну продуктивність. Вона має дуже високу здатність виявляти один із класів, практично безпомилково його ідентифікуючи. Однак, при цьому, модель абсолютно не здатна правильно класифікувати інший клас. Це говорить про сильну упередженість моделі в бік одного з класів.

На рис. 3.4 бачимо, що модель чудово справляється з виявленням переможців, практично безпомилково його ідентифікуючи. Однак при цьому вона повністю ігнорує клас поразка, не виявляючи жодного його представника. Це говорить про те, що модель сильно упереджена в бік класу перемога і не здатна до узагальнення.

Така поведінка моделі мала б вказувати на дисбаланс класів, але у п.п 2.1. визначено, що дисбалансу у даних не виявлено. Модель демонструє жахливі результати, тому вона потребує подальшого доопрацювання.

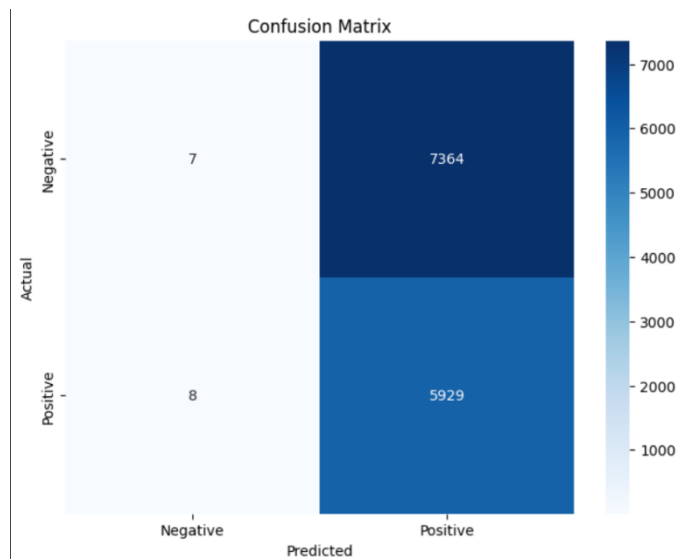


Рисунок 3.4 - Матриця плутанини для методу опорних векторів

Джерело: розроблено автором на основі [Додаток Б]

Спроба покращити модель за допомогою прищвидшеного пошуку параметрів за допомогою RandomizedSearchCV, як видно з рис. 3.5, не допомогла покращити результат моделі. Проблема з визначенням правильних класів залишилася.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	7371
1	0.45	1.00	0.62	5937
accuracy			0.45	13308
macro avg	0.22	0.50	0.31	13308
weighted avg	0.20	0.45	0.28	13308

Рисунок 3.5 - Оцінювання якості моделі

Джерело: розроблено автором на основі [Додаток Б]

Порівняння використаних моделей для прогнозування наведено у табл. 3.6.

Таблиця 3.6 - Порівняння моделей

Метрика	Логістична регресія	Метод опорних векторів	Дерева рішень
Accuracy	0.93	0.44	0.94
Precision	0.90	0.44	0.93
Recall	0.96	0.99	0.92
ROC-AUC	0.99	-	0.99
F1	0.93	0.61	0.93

Джерело: розроблено автором на основі [Додаток Б]

Логістична регресія та дерева рішень, демонструють високу ефективність з точки зору метрик Recall, Precision та F1-score. Вони мають подібні показники, що свідчить про їхню здатність добре класифікувати дані. Високий ROC-AUC для обох моделей підтверджує їхню здатність ефективно розрізняти класи.

В свою чергу метод опорних векторів потребує значного покращення для можливості практичного використання цієї моделі.

Результати прогнозування переможців тендеру методами машинного навчання засвідчили ефективність використання сучасних аналітичних підходів у сфері публічних закупівель. Застосування цих методів створює передумови для підвищення прозорості закупівель і стратегічного планування участі постачальників у тендерах з обґрунтуванням шансів на перемогу.

ВИСНОВКИ

Аналіз державних закупівель є критично важливим для ефективного використання державних коштів та економічного розвитку країни. Впровадження електронної системи ProZorro значно підвищило прозорість та конкурентоспроможність, мінімізуючи корупційні ризики та забезпечуючи рівний доступ учасникам ринку.

Актуальність дослідження зумовлена потребою вдосконалення методів аналізу та прогнозування для прийняття оптимальних рішень у сфері публічних закупівель, зокрема електроенергії.

В рамках дослідження проведено попередню обробку та аналіз даних закупівель електроенергії, виявлено особливості ринку, такі як нерівномірний розподіл закупівель протягом року та концентрацію участей і перемог серед обмеженого кола лідерів.

Для прогнозування перемоги учасників протестовано моделі логістичної регресії, дерева рішень та метод опорних векторів.

Базова модель логістичної регресії показала невисоку точність у 67%. Використання підібраних ознак значно покращило її якість до 93% та показником, F1-score 99%. Налаштування гіперпараметрів логістичної регресії ще більше підвищило здатність моделі виявляти потенційних переможців Recall- 96%, хоча дещо зменшило Precision до 90%.

Модель дерева рішень з випадково обраними параметрами продемонструвала високу точність – 98%, але мала ознаки перенавчання. Застосування GridSearch для підбору параметрів також призвело до перенавчання. Модель дерева рішень з ознаками, відібраними вручну на основі кореляції з цільовою змінною, показала надійніші результати загальна точність – 94% , F1-score – 93% та ROC-AUC 99%

Метод опорних векторів продемонстрував вкрай невдалі результати, успішно визначаючи один клас, але повністю ігноруючи інший, що свідчить про

сильну упередженість попри відсутність дисбалансу класів.

Порівняння використаних моделей показало, що найбільш вдалою для практичного застосування є модель логістичної регресії, оскільки вона демонструє високу точність та немає ознак перенавчання.

В ході роботи успішно розроблено та протестовано різні підходи для прогнозування переможців у тендерах на постачання електроенергії в системі ProZorro, що підтверджує потенціал використання даних методів для підвищення ефективності участі компаній у державних закупівлях, подальше дослідження теми та використання більш складних методів аналізу, зокрема нейронних мереж дасть можливість для побудови більш гнучких моделей.

Отримані результати підтверджують доцільність використання методів машинного навчання для аналізу та прогнозування успіху учасників у сфері публічних закупівель. Побудовані моделі дозволяють виявити приховані закономірності в поведінці учасників торгів та забезпечити вищу точність передбачення їх успішності. Це відкриває можливості для підвищення прозорості та ефективності закупівельних процесів. Застосування подібних підходів може стати основою для подальшої автоматизації та оптимізації управлінських рішень у сфері державних фінансів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. ProZorro URL: <https://prozorro.gov.ua/uk> (дата звернення: 13.05.2025)
2. Про публічні закупівлі: Закон України від 25.12.2015 № 922-VIII URL: <https://zakon.rada.gov.ua/laws/show/922-19> (дата звернення: 13.05.2025)
3. Поліщук, С. П. Реформа публічних закупівель як важіль економічного розвитку. Актуальні проблеми модернізації економіки України: виклики, реалії, перспективи : колективна монографія / за заг. ред. В. Г. Андрущенко. Київ : Видавничий дім «Гельветика», 2021. С. 143-15
4. Публічний модуль аналітики Prozorro. URL: <https://bi.prozorro.org> (дата звернення: 13.05.2025)
5. Вальковець А., Поляк К. Сучасний стан та перспективи розвитку електронних публічних закупівель в Україні. Економіка та суспільство. 2023. № 49. URL: <https://doi.org/10.32782/2524-0072/2023-49-1> (дата звернення: 13.05.2025).
6. Про здійснення державних закупівель : Закон України від 10.04.2014 №1197-VII URL: <https://zakon.rada.gov.ua/laws/show/1197-18> (дата звернення: 13.05.2025)
7. Про ринок електричної енергії : Закон України від 13.04.2017 № 2019-VIII URL: <https://zakon.rada.gov.ua/laws/show/2019-19>
8. Блінов І.В. Теоретичні та практичні засади функціонування конкурентного ринку електроенергії. К.: Наукова думка, 2015. 250 с
9. Види закупівель у системі Prozorro. E-tender.ua URL: <https://e-tender.ua/training-tenders/teoriya-zakupivel-3/vidi-zakupivel-v-sistemi-prozorro-6>
10. Класифікатор ДК 021:2015(CPV, Єдиний закупівельний словник). URL: <https://infobox.prozorro.org/articles/pro-klasifikator-dk-021-2015-yediniy-zakupivelniiy-slovník>
11. Юрій Е. О., Грубляк О. М. (2024). Дослідження вартісних показників

функціонування системи публічних закупівель в Україні. Галицький економічний вісник, 88(3), 104-112. (Фінанси, банківська справа, страхування та фондовий ринок).

12. Наumenko С. М. Система тендерних закупівель: теоретико-методологічні підходи до дефініції і класифікації. Теоретичні і практичні аспекти економіки та інтелектуальної власності. 2014. Вип. 1 (10), Т. 2. С. 242-247
13. Кучер М. О., Бабаков Р. М. Методи боротьби з дисбалансом класів в датасетах для навчання нейронних мереж : Донецький національний університет імені Василя Стуса, м. Вінниця
14. Stiti K., Yape S. J. (2019). Bid Forecasting in Public Procurement: Master of Science Thesis. KTH Royal Institute of Technology, Stockholm, Sweden.
15. Acikalin U. U., Gorgun M. K., Kutlu M., Tas B. K. O. (2023). How you describe procurement calls matters: Predicting outcome of public procurement using call descriptions. Natural Language Engineering, 3, 1-22.
16. Ristov B., Eftimov S., Trajanoska M., & Trajanov, D. (2025). Ontology-Based Structuring and Analysis of North Macedonian Public Procurement Contracts.
17. Moiraghi F., Palmonari M., Allavena, D., Morando, F. (2024). Zero-Shot Hierarchical Classification on the Common Procurement Vocabulary Taxonomy.
18. Wallimann H., Sticher S. (2023). On Suspicious Tracks: Machine-Learning Based Approaches to Detect Cartels in Railway-Infrastructure Procurement.
19. Aldana A., Falcón-Cortés A., Larralde H. (2022). A Machine Learning Model to Identify Corruption in México's Public Procurement Contracts.
20. Huber M., & Imhof D. (2021). Deep Learning for Detecting Bid Rigging: Flagging Cartel Participants Based on CNN.
21. Чернишов Д., Ситніков Д. Бінарна класифікація на основі поєднання теорії приблизних множин і дерев рішень *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. №26.

22. Jelina J., Sasikala R. Multi Class Classification Methods on Data Analysis Using Data Mining Techniques'', IOSR Journal of Engineering ,2018.
23. Tsoumakas G., Katakis I. Multi-label classification: An overview. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications, 2008, 64-74.
24. Кравченко С.М., Гришкун Є.О., Власенко О.В. Методи класифікації машинного навчання з використанням бібліотеки Scikit-learn // Інформатика, обчислювальна техніка та автоматизація. 1 2020. № 3-1. С. 19-28
25. Handelman G.S., Kok H.K., Chandra R.V., Razavi A.H., Huang S., Brooks M., Lee M.J., Asadi H. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods - 2019. - Vol. 212, No. 1
26. Koller D., Friedman N. Probabilistic graphical models: Principals and techniques MIT Press. ISBN 978-0262013192, 2009
27. Liu J., Cui J., Chen C. (2023). Online Efficient Secure Logistic Regression based on Function Secret Sharing
28. Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press. ISBN 978-1-139-64363-4.
29. Hastie T.J., Tibshirani R.J., Friedman J.H. The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition. Springer; 2009.
30. Карабін Р. М., Литвиненко Я. В. Огляд бібліотек машинного навчання для мови Python: Тернопільський національний технічний університет імені Івана Пулюя, 2023

ДОДАТКИ

Додаток А

SQL Запит для отримання потрібних даних

```

SELECT DISTINCT
    GPBID.IDLOT as LotId,GPBID.INITAMOUNT as LotAmount,
    GPBID.ORGANIZER_OKPO as OrganizerOKPO, GPBID.SUPPLIER_OKPO as
    ParticipantOKPO,GPBID.ORGANIZER_REGION as OrganizerRegion,
    GPBID.LOT_COST as LotCost, GPBID.LOT_QUANTITY as LotQuantity,
    GPBID.WITH_VAT as VAT, GPBID.TENDER_TYPE as TenderType,
    GPBID.SUPPLY_DURATION as supply_duration,
    GPBID.LOT_PAYMENT_PERIOD as LotPaymentPeriod,
    GPBID.PROPOSAL_SPEED as PropSpeed, GPBID.PARTICIPANT_PRICE as
    participant_price, GPBID.STATUS_BID as status_bid,
    SUM(CASE WHEN GPBID.STATUS_BID = 1 THEN 1 ELSE 0 END)
    OVER (PARTITION BY GPBID.ORGANIZER_OKPO, GPBID.SUPPLIER_OKPO)
    AS WinsCount, COUNT(GPBID.IDLOT) OVER (PARTITION BY
    GPBID.ORGANIZER_OKPO, GPBID.SUPPLIER_OKPO) AS MeetInTenders, CASE
    WHEN SUM(CASE WHEN GPBID.STATUS_BID = 1 THEN 1 ELSE 0 END) OVER
    (PARTITION BY GPBID.ORGANIZER_OKPO, GPBID.SUPPLIER_OKPO) = 0 OR
    COUNT(GPBID.IDLOT) OVER (PARTITION BY GPBID.ORGANIZER_OKPO,
    GPBID.SUPPLIER_OKPO) = 0 THEN 0 ELSE ROUND(CAST(SUM(CASE WHEN
    GPBID.STATUS_BID = 1 THEN 1 ELSE 0 END) OVER (PARTITION BY
    GPBID.ORGANIZER_OKPO, GPBID.SUPPLIER_OKPO) AS FLOAT) /
    CAST(COUNT(GPBID.IDLOT) OVER (PARTITION BY
    GPBID.ORGANIZER_OKPO, GPBID.SUPPLIER_OKPO) AS FLOAT), 2) END AS
    WinRateOrg, SUM(CASE WHEN GPBID.STATUS_BID = 1 THEN 1 ELSE 0 END)
    OVER (PARTITION BY GPBID.ORGANIZER_REGION,
    GPBID.SUPPLIER_OKPO) AS WinsCountByRegion, COUNT(GPBID.IDLOT)
    OVER (PARTITION BY GPBID.ORGANIZER_REGION,
    GPBID.SUPPLIER_OKPO) AS PartInRegion,

```

```

CASE WHEN SUM(CASE WHEN GPBID.STATUS_BID = 1 THEN 1
ELSE 0 END) OVER (PARTITION BY GPBID.ORGANIZER_REGION,
GPBID.SUPPLIER_OKPO) = 0 OR COUNT(GPBID.IDLOT) OVER (PARTITION
BY GPBID.ORGANIZER_REGION, GPBID.SUPPLIER_OKPO) = 0 THEN 0 ELSE
ROUND(CAST(SUM(CASE WHEN GPBID.STATUS_BID = 1 THEN 1 ELSE 0
END) OVER (PARTITION BY GPBID.ORGANIZER_REGION,
GPBID.SUPPLIER_OKPO) AS FLOAT) / CAST(COUNT(GPBID.IDLOT) OVER
(PARTITION BY GPBID.ORGANIZER_REGION, GPBID.SUPPLIER_OKPO) AS
FLOAT), 2) END AS WinRateRegion, AVG(CASE WHEN GPBID.STATUS_BID = 1
THEN GPBID.LOT_PAYMENT_PERIOD ELSE NULL END) OVER (PARTITION
BY GPBID.SUPPLIER_OKPO) as AvgLotPaymentPeriod,
AVG(CASE WHEN GPBID.STATUS_BID = 1 THEN
GPBID.SUPPLY_DURATION ELSE NULL END) OVER (PARTITION BY
GPBID.SUPPLIER_OKPO) as AvgSupplyDuration,
AVG(CASE WHEN GPBID.STATUS_BID = 1 THEN
GPBID.PROPOSAL_SPEED ELSE NULL END) OVER (PARTITION BY
GPBID.SUPPLIER_OKPO) as AvgProposalSpeed,
GPBID.INITAMOUNT - GPBID.PARTICIPANT_PRICE as DiffPrice,
(GPBID.INITAMOUNT - GPBID.PARTICIPANT_PRICE) /
GPBID.INITAMOUNT as DiffPricePercent,
AVG((GPBID.INITAMOUNT - GPBID.PARTICIPANT_PRICE) /
GPBID.INITAMOUNT) OVER (PARTITION BY GPBID.SUPPLIER_OKPO) as
AvgDiffPricePercent
FROM GPBID;

```

Додаток Б

```

from sklearn.model_selection import GridSearchCV, RandomizedSearchCV,
cross_val_score, train_test_split

from sklearn.linear_model import LogisticRegression, SGDClassifier
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.tree import DecisionTreeClassifier

```

```

from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score, roc_auc_score, confusion_matrix, classification_report

import pandas as pd, seaborn as sns, matplotlib.pyplot as plt, numpy as np

data = pd.read_excel("train.xlsx"); filled_na = data.fillna(0);
filled_na['ParticipantName'] = filled_na['ParticipantName'].str.replace('ТОВАРИСТВО
З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ', 'TOB', regex=False)

def plot_bar(data, x, y, title, xlabel, ylabel, rotation=0): plt.figure(figsize=(10,
6)); ax = sns.barplot(x=x, y=y, data=data, color='skyblue'); plt.title(title);
plt.xlabel(xlabel); plt.ylabel(ylabel); plt.xticks(rotation=rotation);
[ax.annotate(f'{int(p.get_height())}', (p.get_x() + p.get_width() / 2., p.get_height()),
ha='center', va='center', xytext=(0, 10), textcoords='offset points') for p in ax.patches];
plt.tight_layout(); plt.show()

def plot_pie(data, values, labels, title): plt.figure(figsize=(8, 8)); plt.pie(values,
labels=labels, autopct='%1.1f%%', startangle=140); plt.gca().add_artist(plt.Circle((0, 0),
0.50, fc='white')); plt.axis('equal'); plt.title(title); plt.show()

data_type = filled_na.groupby("TenderType")['LotId'].nunique().reset_index();
plot_bar(data_type, 'TenderType', 'LotId', 'Розподіл лотів за Типом тендеру', 'Тип
тендеру', 'Кількість лотів')

participations =
filled_na.groupby('ParticipantName')['LotId'].nunique().sort_values(ascending=False).re
set_index().head(10); plot_bar(participations, 'ParticipantName', 'LotId', 'Найбільші
компанії за участю у закупівлях', 'Учасник', 'Кількість участей', rotation=90)

filled_na['PublishDate'] = pd.to_datetime(filled_na['PublishDate']);
filled_na['YearMonth'] = filled_na['PublishDate'].dt.to_period('M'); org_counts_ym =
filled_na.groupby('YearMonth')['LotId'].nunique().reset_index();
plot_bar(org_counts_ym, 'YearMonth', 'LotId', 'Динаміка по кількості лотів', 'Дата',
'Кількість лотів', rotation=90)

unique_lots = filled_na[['LotId', 'LotAmount']].drop_duplicates();

def get_budget_range(amount): return 'менше 20000 грн' if amount < 20000
else '20000 - 50000 грн' if 20000 <= amount < 50000 else '50000 - 200000 грн' if

```

```
50000 <= amount < 200000 else '200000 - 1000000 грн' if 200000 <= amount <
1000000 else '1000000 - 4000000 грн' if 1000000 <= amount < 4000000 else 'понад
4000000 грн';
```

```
unique_lots['BudgetRange']=nunique_lots['LotAmount'].apply(get_budget_range);
budget_categories = ['менше 20000 грн', '20000 - 50000 грн', '50000 - 200000 грн',
'200000 - 1000000 грн', '1000000 - 4000000 грн', 'понад 4000000 грн'];
unique_lots['BudgetRange'] = pd.Categorical(unique_lots['BudgetRange'],
categories=budget_categories,ordered=True);budget_range_sums=unique_lots.groupby(
'BudgetRange')['LotId'].count().reset_index();plot_bar(budget_range_sums,
'BudgetRange', 'LotId', 'Розподіл лотів по бюджетах закупівлі', 'Діапазон бюджету',
'Кількість лотів', rotation=45)
```

```
winners = filled_na[filled_na['status_bid'] == 1][['ParticipantName',
'LotId']].groupby('ParticipantName')['LotId'].count().sort_values(ascending=False).reset_index().head(10);
plot_pie(winners, winners['LotId'], winners['ParticipantName'], 'Топ 10 переможців')
```

```
regions = filled_na[["LotId", "OrganizerRegion", "LotAmount"]].drop_duplicates().groupby('OrganizerRegion')['LotAmount'].sum().sort_values(ascending=False).reset_index().head(10);
plot_pie(regions, regions['LotAmount'], regions['OrganizerRegion'], 'К-сть лотів за регіонами')
```

```
status = filled_na.groupby('status_bid')['LotId'].count().reset_index();
plot_bar(status, 'status_bid', 'LotId', 'Розподіл ставок', 'Статус участі', 'Кількість ставок')
```

```
def encode_categorical_columns(data, categorical_columns):
    label_encoders = {}
    for feature in categorical_columns:
        label_encoders[feature] = LabelEncoder().fit(data[feature].astype(str))
    data.update({feature: label_encoders[feature].transform(data[feature].astype(str)) for feature in categorical_columns})
    return data
```

```
def generate_all_combinations(features):
    return [list(features[start_index:end_index]) for start_index in range(len(features)) for end_index in range(start_index + 1, len(features) + 1)]
```

```

def accuracy_metrics(model, X_train, y_train, y_test, y_pred, y_pred_proba,
name, cv=5): accuracy, precision, recall, f1 = accuracy_score(y_test, y_pred),
precision_score(y_test, y_pred), recall_score(y_test, y_pred), f1_score(y_test, y_pred);
conf_matrix = confusion_matrix(y_test, y_pred); roc_auc = roc_auc_score(y_test,
y_pred_proba) if y_pred_proba is not None else "N/A"; cv_mean =
np.mean(cross_val_score(model, X_train, y_train, cv=cv, scoring='accuracy')); cv_std =
np.std(cross_val_score(model, X_train, y_train, cv=cv, scoring='accuracy'));
plt.figure(figsize=(8, 6)); sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
xticklabels=['Negative', 'Positive'], yticklabels=['Negative', 'Positive']);
plt.xlabel('Predicted'); plt.ylabel('Actual'); plt.title('Confusion Matrix'); plt.show();
print(classification_report(y_test, y_pred)); return {'name': name, 'accuracy': accuracy,
'precision': precision, 'recall': recall, 'f1': f1, 'roc_auc': roc_auc, 'cv_mean_accuracy':
cv_mean, 'cv_std': cv_std}

```

```

def default_model_training(features, model_data, target, model, name): X, y =
model_data[list(features)], model_data[target]; X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42); model.fit(X_train, y_train);
y_pred = model.predict(X_test); y_pred_proba = model.predict_proba(X_test)[:, 1] if
hasattr(model, "predict_proba") else None; return accuracy_metrics(model, X_train,
y_train, y_test, y_pred, y_pred_proba, name), X_train, X_test, y_train, y_test, y_pred, y

```

```

def model_tuning(X_train, X_test, y_train, y_test, name, model, param_grid):
grid_search = GridSearchCV(model, param_grid, cv=3, scoring='roc_auc', n_jobs=-
1); grid_search.fit(X_train, y_train); best_model = grid_search.best_estimator_; y_pred =
best_model.predict(X_test); y_pred_proba = best_model.predict_proba(X_test)[: ,
1]; metrics = accuracy_metrics (model, X_train, y_train, y_test, y_pred, y_pred_proba,
name); return metrics, X_train, X_test, y_train, y_test, y_pred, y

```



Звіт подібності

метадані

Назва організації
Kyiv National Economic University named after Vadym Hetman KNEU
Заголовок
Беленок_Юнькова_плагіат
Автор Науковий керівник / Експерт
БеленокЮнькова
підрозділ
кафедра математичного моделювання та статистики

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.



Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		3
Інтервали		0
Мікропробіли		42
Білі знаки		0
Парафрази (SmartMarks)		27

Подібності за списком джерел

Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Копію тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

10 найдовших фраз		Копія тексту
ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КОЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	Спосіб виявлення неоптимального використання програмного коду мовою Kotlin 3/15/2025 National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute (National Technical University of Ukraine Igor Sikorskyi Kyiv Politech Institute)	66 0.72 %
2	2022_M_ІНФ_ІНФм-21-1_Івашенко_О_О 5/30/2024 Kharkiv National University of Radio Electronics (Kharkiv National University of Radio Electronics)	63 0.69 %