

II Всеукраїнська науково-практична конференція

# **«Системні науки та інформатика»**

**Збірник доповідей**

4–8 грудня 2023 року

Київ, Україна

**Системні науки та інформатика:** збірник доповідей II науково-практичної конференції «Системні науки та інформатика», 4–8 грудня 2023 року, Київ. – К., НН ІПСА КПІ ім. Ігоря Сікорського, 2023. – 416 с.

© Навчально-науковий Інститут прикладного системного аналізу КПІ ім. Ігоря Сікорського, 2023.

Співголови програмного комітету конференції:

Касьянов П.О.	Романенко В.Д.	Панкратова Н.Д.
---------------	----------------	-----------------

Члени програмного комітету конференції:

Бідюк П.І.	Калініна І.О.	Пишнограєв І.О.
Гавриленко В.В.	Кисельов Г.Д.	Положаєнко С.А.
Гожий О.П.	Корабльов М.М.	Савченко І.О.
Данилов В.Я.	Литвиненко В.І.	Тимошук О.Л.
Джигирей І.М.	Мухін В.Є.	Цюцюра С.В.
Єфремов К.В.	Петренко А.І.	

Організаційний комітет конференції:

Пишнограєв І.О. – голова	Савченко І.О.
Левенчук Л.Б.	Кисельов Г.Д.

Верстка збірника: Савченко І.О.

# ЗМІСТ

## Секція 01

### Системний аналіз і управління

<b>Методи заповнення пропусків у даних по електроспоживанню</b>	<b>8</b>
<i>Бондаренко Ю.Г., Губарев В.Ф.</i>	
<b>Лінійні моделі для прогнозування та керування фінансово-економічними показниками на основі когнітивних карт</b>	<b>14</b>
<i>Зуєвський Ю.В., Мілявський Ю.Л.</i>	
<b>СППР для дослідження ринкових ризиків</b>	<b>21</b>
<i>Квашук І.О., Кузнєцова Н.В.</i>	
<b>Стабілізація демографічних процесів за допомогою управління по прогнозуючій моделі</b>	<b>27</b>
<i>Нетудихата А.С., Губарев В.Ф.</i>	
<b>Інформаційна система для кластеризації стану країн за показниками сталого розвитку</b>	<b>33</b>
<i>Самсонюк М.В., Бідюк П.І.</i>	
<b>Інтелектуальна система підтримки прийняття рішень для дослідження актуарних фінансових ризиків</b>	<b>38</b>
<i>Чеманова А.О., Кузнєцова Н.В.</i>	
<b>Моделювання впливу чат-ботів на основі штучного інтелекту на якість вищої освіти методами системного аналізу</b>	<b>45</b>
<i>Чернюк О.Р., Тимошук О.Л.</i>	

## Секція 02

### Системний аналіз фінансового ринку

<b>Прогнозування гетероскедастичних процесів для оцінювання фінансового ризику</b>	<b>58</b>
<i>Байбара А.Г., Кузнєцова Н.В.</i>	
<b>Підхід щодо визначення факторів впливу на ринок логістики США з використанням LLM</b>	<b>64</b>
<i>Балькін Я.Ю., Савастьянов В.В.</i>	
<b>Прогнозування ціни на золото методами машинного навчання</b>	<b>70</b>
<i>Білоус К.С., Кузнєцова Н.В.</i>	
<b>Порівняльний аналіз та програмна реалізація методів розв'язання задачі про максимальний потік у мережах</b>	<b>76</b>
<i>Боднар М.С., Статкевич В.М.</i>	
<b>Вибір та оцінка якості моделей прогнозування фінансових показників з урахуванням характеристик вхідних даних</b>	<b>84</b>
<i>Бойніцька С.В., Мілявський Ю.Л.</i>	

<b>Аналіз текстових повідомлень за допомогою методів машинного навчання</b> <i>Ведмедєв Д.О., Шаповал Н.В.</i>	<b>90</b>
<b>Система прогнозування енерговитрат будівель різного призначення</b> <i>Данилов В.Я., Дука О.О.</i>	<b>94</b>
<b>Математичні моделі суспільних процесів для аналізу впливу війни на динаміку розвитку економіки та економічних показників</b> <i>Діхтяр А.В., Лопатін О.К.</i>	<b>102</b>
<b>Система підтримки прийняття рішень для оптимізації рекламних кампаній підприємства на основі методу моделювання впливу з залежним представленням даних</b> <i>Заїка Б.Ю., Терентьєв О.М.</i>	<b>108</b>
<b>Модель управління ресурсами гетерогенних баз даних в хмарному середовищі</b> <i>Зайцев О.В., Мухін В.Є.</i>	<b>117</b>
<b>Система прогнозування метеорологічних умов на основі методів аналізу даних та штучного інтелекту</b> <i>Іванійчук А.П., Гуськова В.Г.</i>	<b>121</b>
<b>Кластеризація за допомогою OPTICS: аналіз та оптимізація з використанням графіків та метрик</b> <i>Іванюта О.О., Недашківська Н.І.</i>	<b>127</b>
<b>Моделювання сейсмічних хвиль з використанням машинного навчання</b> <i>Каніовська І.Ю., Кавара А.О.</i>	<b>134</b>
<b>Моделі інтелектуального аналізу даних для оцінювання фінансових моделей</b> <i>Коваленко О.М., Гуськова В.Г.</i>	<b>139</b>
<b>Сучасні моделі оцінювання фінансових ризиків</b> <i>Костенко М.О.І, Кузнєцова Н.В.</i>	<b>146</b>
<b>Система підтримки прийняття рішень для вибору S-моделей економічного зростання</b> <i>Кузьмінчук А.В., Лопатін О.К.</i>	<b>153</b>
<b>Розробка моделей оцінювання ризиків зелених проектів</b> <i>Кузнєцова Н.В., Шевчук О.С.</i>	<b>162</b>
<b>Аналіз відкритих даних про якість повітря в міському середовищі та розробка прогностичних моделей для прогнозу забруднення повітря в місті</b> <i>Луцкер Р.О., Гуськова В.Г.</i>	<b>168</b>
<b>Порівняльний аналіз моделей для методів прогнозування</b> <i>Макухін Є.І., Макаренко О.С., Бідюк П.І.</i>	<b>172</b>
<b>Підходи до прогнозування фінансового стану підприємства та оцінки інвестиційної привабливості сільськогосподарського підприємства на основі статистичних даних</b> <i>Мілявський Ю.Л., Павлуша А.О.</i>	<b>181</b>
<b>Порівняльний аналіз та покращення моделей прогнозування цін акцій на фінансовому ринку</b> <i>Муравльов А.Д., Гуськова В.Г.</i>	<b>187</b>
<b>Моделі оптимального розподілу даних</b> <i>Мухін В.Є., Яковлева А.П., Шмідт А.Є.</i>	<b>192</b>

<b>Прогнозування кредитної спроможності клієнтів банку на основі аналізу фінансових даних</b>	<b>198</b>
<i>Петровський В.Є., Гуськова В.Г.</i>	
<b>Розв'язання задачі заповнення пропусків даних альтернативними методами при створенні прогнозних моделей</b>	<b>201</b>
<i>Попов А.Ю., Макаренко О.С., Бідюк П.І.</i>	
<b>Система аналізу впливу кластеризації на якість рішень в моделях штучного інтелекту</b>	<b>207</b>
<i>Симонов Є.Д., Макаренко О.С.</i>	
<b>Системний підхід до аналізу кредитних ризиків в банківському секторі</b>	<b>213</b>
<i>Сумін О.О., Шубенкова І.А.</i>	
<b>Інтелектуальні засоби підтримки автоматизації управління бізнес процесами</b>	<b>218</b>
<i>Тагільцев Д.І., Мухін В.Є.</i>	
<b>Intellectual decision support system for estimation of financial risks</b>	<b>223</b>
<i>Tymoshchuk O.L., Levenchuk L.B., Vidyuk P.I.</i>	
<b>Чат-бот як середовище розгортання системи підтримки прийняття рішень: приклад телеграм-боту по наданню рекомендацій щодо вибіркових дисциплін</b>	<b>229</b>
<i>Харабара Д.В., Статкевич В.М.</i>	
<b>Автоматична сегментація об'єктів на зображеннях за допомогою напівкерованого навчання з використанням активного навчання та стабільної дифузії</b>	<b>237</b>
<i>Шаповал Н.В., Крутий І.В.</i>	

### Секція 03

#### Інтелектуальні сервіс-орієнтовані розподілені обчислювання

<b>Керування відносинами з клієнтом як сервіс в галузі закладів освіти</b>	<b>243</b>
<i>Бабіч К.О., Петренко А.І.</i>	
<b>Технології WASM, WASI та WAGI для розробки та розгортання мережевих застосунків</b>	<b>247</b>
<i>Булах Б.В., Бондаренко С.Д.</i>	
<b>Інтелектуальні агенти для системи автоматизації управління бізнес-процесами у сфері логістики</b>	<b>252</b>
<i>Вайнер Г.О., Мухін В.Є.</i>	
<b>Інструменти тестування цифрової доступності веб-сайтів</b>	<b>259</b>
<i>Волосожар Д.В., Кисельов Г.Д.</i>	
<b>Гарантування унікальної доставки повідомлень в розподілених системах</b>	<b>263</b>
<i>Гапонюк М.О., Письменний І.О.</i>	
<b>Резервне копіювання даних як послуга</b>	<b>268</b>
<i>Головін Б.О., Булах Б.В.</i>	
<b>Методика редуції обсягу інформації в системах обробки великих даних</b>	<b>273</b>
<i>Дзиговський В.І., Рогоза В.С.</i>	
<b>Дослідження бібліотеки React.js для побудови веб-інтерфейсів для людей з обмеженими можливостями</b>	<b>279</b>
<i>Забельський В.В., Кисельов Г.Д.</i>	

<b>Програмний сервіс для раціонального управління персоналом комерційної компанії</b>	<b>284</b>
<i>Кисельов Г.Д., Гречко Д.М.</i>	
<b>Розробка бібліотеки компонентів для веб інтерфейсів медичних платформ з використанням бібліотеки React.js</b>	<b>291</b>
<i>Кисельов Г.Д., Коваль Д.О.</i>	
<b>Розробка WordPress плагіна для керування файловою системою</b>	<b>297</b>
<i>Кисельов Г.Д., Таран А.І.</i>	
<b>Адаптивні засоби захисту комп'ютерних систем на основі апарата нейронних мереж</b>	<b>303</b>
<i>Коновал В.О., Мухін В.Є.</i>	
<b>Методології розробки мобільних додатків для медичних інформаційних систем на платформі iOS</b>	<b>309</b>
<i>Люлька Р.О., Харченко К.В.</i>	
<b>Розробка і конфігурація інтелектуальної розподіленої системи контролю споживання електроенергії в багатоквартирному будинку</b>	<b>314</b>
<i>Мунтян Д.М., Кисельов Г.Д.</i>	
<b>Сервер для мобільних застосунків як сервіс</b>	<b>320</b>
<i>Науменко Є.О., Булах Б.В.</i>	
<b>Методи аналізу новин</b>	<b>325</b>
<i>Орловський А.В., Кислий Р.В.</i>	
<b>Огляд пріоритезації трафіку в програмно визначених мережах</b>	<b>330</b>
<i>Переяславський С.К., Письменний І.О.</i>	
<b>Хмарний сервіс для дистанційного догляду за пацієнтами</b>	<b>335</b>
<i>Петренко А.І., Редька М.Ю.</i>	
<b>Оркестрування обчислень у мікросервісній архітектурі</b>	<b>341</b>
<i>Полещук В.О., Булах Б.В.</i>	
<b>Аналіз ризиків в задачах інформаційної безпеки</b>	<b>347</b>
<i>Северин М.С., Мухін В.Є.</i>	
<b>Платформи з низькою затримкою для обробки потоків даних у реальному часі</b>	<b>353</b>
<i>Сіркович А.І., Харченко К.В.</i>	
<b>Сервіс ведення фінансової звітності для ФОП</b>	<b>356</b>
<i>Скрипник А.В., Булах Б.В.</i>	
<b>Оптимізація передачі даних у багатокористувацьких відеоіграх</b>	<b>360</b>
<i>Сухарев О.М., Безносик О.Ю.</i>	
<b>Система розумного будинку з керуванням телеграм ботом та Google Assistant</b>	<b>366</b>
<i>Харченко К.В., Кушовий Д.І.</i>	
<b>Дослідження продуктивності мікросервісних архітектур через кешування даних</b>	<b>370</b>
<i>Хоміч Л.І., Яременко В.С.</i>	
<b>Геолокація позицій об'єктів робочої карти за даними зображень БПЛА</b>	<b>375</b>
<i>Хом'як К.В., Петренко А.І.</i>	

---

<b>Самонавчальна система розпізнавання людської діяльності з використанням навчання з підкріпленням</b>	<b>382</b>
<i>Цибін М.Д., Кислий Р.В.</i>	
<b>Інтернет речей в сфері охорони здоров'я</b>	<b>388</b>
<i>Цимбалюк Р.С.</i>	

---

#### **Секція 04**

##### Системи і методи штучного інтелекту

---

<b>Filter for confidential information for chats with LLM and using local texts databases</b>	<b>392</b>
<i>Bezumiannyi O.E., Sharoval N.V.</i>	
<b>Аналіз наукових статей за допомогою штучного інтелекту</b>	<b>396</b>
<i>Овчаренко О.С.</i>	
<b>Задача пошуку шляхів з використанням спектральної теорії графів</b>	<b>400</b>
<i>Сабітова Р.Р., Статкевич В.М.</i>	
<b>Довгострокове прогнозування попиту: використання ансамблю нейронних мереж для підвищення точності</b>	<b>406</b>
<i>Самошин А.О., Синєглазов В.М.</i>	
<b>Пошук відповідності між зображенням і його текстовим описом</b>	<b>412</b>
<i>Шаповал Н.В., Крижанівська О.В.</i>	

---

# МЕТОДИ ЗАПОВНЕННЯ ПРОПУСКІВ У ДАНИХ ПО ЕЛЕКТРОСПОЖИВАННЮ

Бондаренко Ю.Г.<sup>1</sup>, Губарев В.Ф.

Національний технічний університет України «Київський політехнічний інститут  
імені Ігоря Сікорського», Київ, Україна

<sup>1</sup> bondarenkojulia423@gmail.com

**Метою роботи є дослідження методів боротьби з пропущеними значеннями. Експериментально перевірити ефективність існуючих методів для вирішення задачі заповнення пропусків на даних по електроспоживанню. Порівняти отримані результати для кожного методу. Визначити які методи мають потенціал для ефективного практичного застосування в розглянутій галузі.**

**Ключові слова:** заповнення пропусків, електроспоживання, часові ряди, регресійні моделі.

## 1. ВСТУП

На сьогоднішній день важливим є питання ефективного аналізу, передбачення та прогнозування електроспоживання для подальшого запобігання можливості тимчасових відключень електроенергії чи навіть зatoryжних блекаутів. Тому важливо, щоб вхідні дані були якомога якіснішими та повними, тобто не містили пропусків.

Наявність пропущених значень в часових рядах – це проблема, яку досліджують вже не перше десятиліття науковці по всьому світу. Поява пропусків може бути із низки причин: технічні проблеми, зміни у методології збору інформації, неможливість отримати дані в певний період часу, тощо. Наявність відсутніх значень значно спотворює вхідну інформацію, що може призвести до неможливості коректного аналізу даних та подальшого прогнозування [1]. Вважається, що наявність пропусків менше 5 відсотків є досить несуттєвою та може не впливати на результат дослідження. Проте, якщо відсоток втрати даних становить 20 та більше, то результати такого дослідження потрібно ставити під сумнів [2].

Хоч методів для відновлення відсутніх значень існує досить багато, однак, універсальної методології немає. Тому важливо дослідити наявні способи боротьби з пропусками та оцінити їх ефективність для часового ряду електроспоживання.

## 2. МЕТОДИ ЗАПОВНЕННЯ ПРОПУСКІВ

Найпростішим методом обробки даних, що містять пропуски, є видалення відсутніх значень. Проте такий спосіб не є ефективним і лише спотворює наявні дані, тому використання його не є доцільним.

Метод медіани полягає у тому, що всі пропущені значення замінюються статичною мірою центральної тенденції – медіаною. Він простий у реалізації та використанні, але може призвести до заниження якості даних.

За схожою аналогією працює метод моди. Його принцип в тому, що замість пропусків записується значення, що найчастіше зустрічається в наборі даних. Цей метод корисний особливо у випадках категоріальних змінних.

Метод інтерполяції. Він допомагає оцінити відсутні значення між двома сусідніми точками даних, що відомі. Цей метод в основному можна вважати чудовим компромісом між

швидкістю роботи та точністю результатів. Проте ефективність цього методу залежить від характеру набору даних, тому в певних випадках він може бути недоцільним.

Регресійний метод. Основна його суть в тому, що заповнення пропусків виконується за допомогою прогнозування їх різними регресійними моделями. Для цього дані спостережуваної змінної, що не містять пропусків, використовуються як залежна змінна, а інші змінні використовуються як незалежні для побудови відповідної регресійної моделі. Наступний кроком є використання отриманої моделі для прогнозування відсутніх значень. Головною перевагою використання регресійного методу є те, що такий підхід враховує зв'язки між змінними та не змінює вихідний розподіл.

Для подальшого дослідження розглядалися два типи регресійних моделей: лінійні (Linear Regression, Bayesian Ridge) та нелінійні (Random Forest, Support Vector Regression, Adaptive Boosting).

Метод K-Nearest Neighbor. Його принцип роботи полягає в тому, що для кожного відсутнього значення вибираються значення k-найближчих сусідів, що не містять пропусків, які і будуть опорними значеннями для подальшого розрахунку. Далі кожен пропуск обраховується як функція від значень цих k-найближчих сусідів. Цей метод є досить ефективним, коли дані містять певні закономірності.

### 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для проведення досліджень використано дані про щоденне споживання електроенергії українцями за період з 2014 по 2022 рік, отримані з офіційних даних УкрЕнерго. Розмір датасету становить 3223 записи. Графічне представлення даних наведено на Рис. 1.

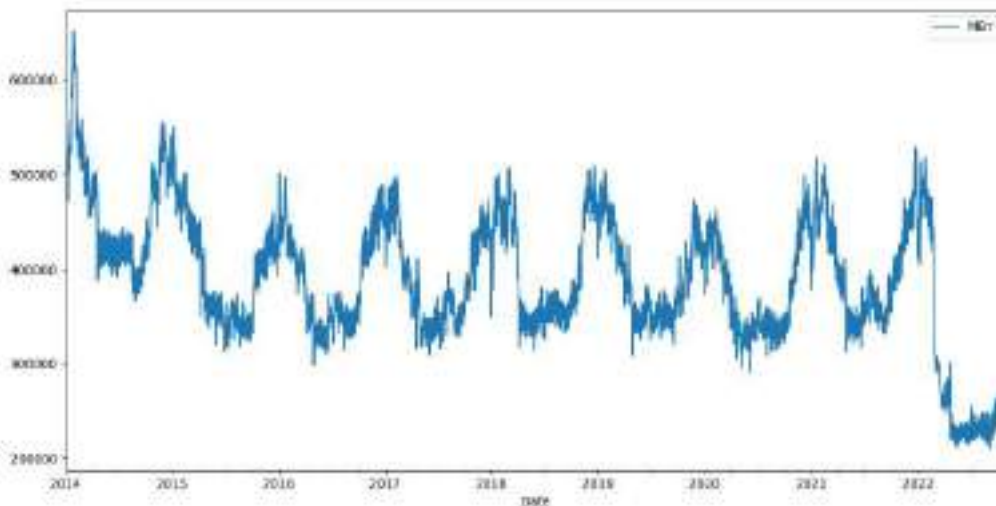


Рисунок 1. Споживання електроенергії 2014–2022 р.

Спочатку виконався попередній аналіз датасету. Для наявних даних було побудовано кореляційні матриці. Результати наведені в Табл. 1 демонструють кореляцію між споживанням електроенергії, днем, місяцем, роком та годиною. Результати наведені в Табл. 3 містять дані кореляції між споживанням, генерацією та температурою навколишнього повітря.

Таблиця 1. Кореляційна матриця 1

	Day	Month	Year	Hour	Consumption
Day	1,000000	0,008851	-0,003293	0,000494	0,020389
Month	0,008851	1,000000	-0,046340	-0,000223	-0,107189
Year	-0,003293	-0,046340	1,000000	0,000066	-0,354161
Hour	0,000494	-0,000223	0,000066	1,000000	0,370651
Consumption	0,020389	-0,107189	-0,354161	0,370651	1,000000

Таблиця 2. Кореляційна матриця 2

	Generation	Temperature	Consumption
Generation	1,000000	-0,673325	0,981036
Temperature	-0,673325	1,000000	-0,695876
Consumption	0,981036	-0,695876	1,000000

З таблиці 1 видно, що дата та час не мають значного впливу на обсяг спожитої електроенергії. Це пояснюється тим, що останні роки в країні не було вираженої сезонності, адже часто траплялося так, що зими були відносно теплі, а весни досить холодні. Однак відомо, що споживання електроенергії часто зростає у жаркі чи навпаки дуже холодні дні, адже люди активно починають використовувати кондиціонери, вентилятори, електрообігрівачі, тощо. Тому, як ми бачимо з таблиці 2, коефіцієнт кореляції між температурою та обсягом електроспоживання становить  $-0,695876$ , що означає те, що цей параметр може бути вагомим для побудови моделей. Тому доцільно буде перевірити його вплив на роботу методів.

Також з таблиці 2 видно, що обсяг згенерованої електроенергії має високий коефіцієнт кореляції, тому доцільно використовувати його як параметр для моделей.

Для перевірки методів було створено три різних датасети. Перший набір даних включає поля «Дата» та «Спожито, МВт». До нього застосовувалися прості методи заповнення пропусків (мода, медіана та інтерполяція). Другий та третій датасети – «Дата», «Згенеровано, МВт», «Спожито, МВт» та «Дата», «Згенеровано, МВт», «Спожито, МВт» та «Середня денна температура, С°» відповідно.

Для проведення експериментів штучним шляхом в усіх наборах даних було згенеровано пропуски із відсотком втрати інформації в полі «Спожито, МВт» 10%, 20%, 40% та 60%. Для порівняння ефективності методів в кожному випадку було обчислено коефіцієнт детермінації, середня абсолютна похибка та корінь середньоквадратичного відхилення.

Порівняння результатів методів моди, медіани та інтерполяції для трьох датасетів наведено у Табл. 3.

Порівняння результатів шести моделей для другого та третього датасетів наведено в Табл. 4.

Таблиця 3. Порівняння результатів для моди, медіани та інтерполяції

Похибка	Мода	Медіана	Інтерполяція
Перший датасет, 10%			
R2	0,81869714	0,8729142	0,9974808
RMSE	83,96228	77,28661	28,53677
MAE	7049,66449	5973,2203	814,3475067
Другий датасет, 10%			
R2	0,90443874393	0,95085209	0,9990900
RMSE	66,4259656	54,88019	16,4613682
MAE	4412,40890890	3011,83608	406,46496
Третій датасет, 10%			
R2	0,9362924	0,967234	0,999393
RMSE	54,236573	44,8094	16,4613682
MAE	2941,6059	2007,89072	270,97664
Перший датасет, 20%			
R2	0,6805497	0,76390084	0,994414
RMSE	111,8694	104,6918	40,963454
MAE	12514,7768	10960,384	1678,004
Другий датасет, 20%			
R2	0,79201352	0,887542	0,99771
RMSE	95,60291793	77,9939	29,803414
MAE	9139,9179	6083,0630	888,2435
Третій датасет, 20%			
R2	0,86134234	0,925028	0,9984792
RMSE	78,0594	63,681829	24,33438641
MAE	6093,2786	4055,3753	592,16236
Перший датасет, 40%			
R2	0,146221	0,33185800	0,9843825
RMSE	166,1744	149,0019	64,038066
MAE	27613,937	22201,57	4100,873
Другий датасет, 40%			
R2	0,3866518	0,6854566	0,99455428
RMSE	173,46507	111,89123	44,057018
MAE	30090,13	12519,6476	1941,02084
Третій датасет, 40%			
R2	0,5911012	0,7903044	0,9963695
RMSE	141,63363	91,3588078	35,9724
MAE	20060,0874	8346,43176	1294,013
Перший датасет, 60%			
R2	-0,601119	-0,3704258	0,96869
RMSE	195,543	182,021658	84,75633
MAE	38237,2045	33131,884	7183,6361
Другий датасет, 60%			
R2	-0,008068	0,30906	0,98832
RMSE	211,12529903	137,327202	58,896317
MAE	44573,891	18858,7605	3468,77624
Третій датасет, 60%			
R2	0,32795	0,53937485	0,992217
RMSE	172,38308	112,12719	48,0886
MAE	29715,927	12572,5070	2312,517497

Таблиця 4. Порівняння результатів для шести моделей

Похибка	Linear Regression	Random Forest	Bayesian Ridge	SVR	Ada Boost	KNN
Другий датасет, 10%						
R2	0,9935750	0,9852721	0,9935746	0,9852721	0,98527214	0,99815743
RMSE	38,61722	45,498611	1491,32252	80,221261	49,514787	29,0084641
MAE	1491,2899	2070,12365	2224042,868	6435,4508	2451,7141	841,49099099
Третій датасет, 10%						
R2	0,9922817	0,984748	0,992286	0,98474804	0,984748	0,9981645
RMSE	40,11553	44,59644	1608,8032	80,412648	47,57422	28,941684835
MAE	1609,2559	1988,8433	2588247,8	6466,19404	2263,3068	837,621121
Другий датасет, 20%						
R2	0,986310	0,9684657	0,9863098	0,96846574	0,968465	0,99468875
RMSE	55,65723	65,907300	3097,799486	116,921829	71,73040	44,51679
MAE	3097,72832	4343,77227	9596361,65	13670,7143	5145,2504	1981,74474
Третій датасет, 20%						
R2	0,9834775	0,9676315	0,98348760	0,9676315	0,9676315	0,9962415
RMSE	57,50567	64,6782397	3306,04511	117,186482	70,230005	41,053518623
MAE	3306,90220	4183,27469	10929934,27	13732,67161	4932,2536	1685,391391391
Другий датасет, 40%						
R2	0,971916	0,93806022	0,97191507	0,93806022	0,9380602	0,9898630
RMSE	79,6785807	92,4901632	6348,8339	167,048671	103,1290	61,86479817
MAE	6348,6762	8554,43029	40307692,5	27905,2587	10635,60085	3827,2532
Третій датасет, 40%						
R2	0,96749847	0,93494680	0,96751581	0,93494680	0,93541447	0,99337741
RMSE	81,6670297	91,1252390	6668,055854	167,470421	99,67827105	56,2941473958
MAE	6669,50374	8303,80918	44462968,88	28046,3422	9935,75772	3169,0310
Другий датасет, 60%						
R2	0,95884694	0,9052382	0,9588443	0,905238	0,90523821	0,9809668
RMSE	96,98512259	113,1810	9406,3757	206,49897	124,65692	78,6695678194
MAE	9406,114005	12809,9551	88479905,0	42641,8253	15539,34856	6188,9009
Третій датасет, 60%						
R2	0,95153	0,901126	0,951561	0,9011265	0,901126	0,98438803
RMSE	100,4346994	111,625299	100,4214981	206,96689	121,729431	74,164647
MAE	10087,128	12460,2074	10084,47728	42835,2951	14818,0544	5500,39489

#### 4. ВИСНОВКИ

З отриманих результатів видно, що найкращим методом виявився K-Nearest Neighbor. Він дозволяє ефективно боротися з пропусками навіть на високих відсотках втрати інформації. Метод інтерполяції також показує високі результати, проте у випадках, коли датасет містить пропуски на початку, він не завжди може виконатися, що не дозволяє йому бути універсальним, на відміну від KNN. Також, слід виділити метод Linear Regression. Хоч

показники цього методу трохи нижчі за показники попередніх двох, та все ж він веде себе досить стабільно не залежно від кількості пропусків.

Методи Random Forest, Ada Boost, SVR та Bayesian Ridge показують гарні результати для 10 та 20 відсотків пропусків, проте, при збільшенні відсотку до 40 і більше, отримані результати значно погіршуються в порівнянні з трьома попередніми методами.

Найгірше себе показали методи заповнення модою та медіаною. При відсотку втрачених даних більше 20 ці методи можна вважати неточними та неефективними.

Зазначимо, що додавання параметра «температура» покращило роботу більшості моделей, особливо на етапах з високим відсотком пропусків.

Отже, для заповнення пропусків в даних по електроспоживанню рекомендується використовувати метод K-Nearest Neighbor, інтерполяцію чи Linear Regression. Також, для великих наборів даних з високим відсотком пропущених значень доцільно буде враховувати показник температури зовнішнього середовища для отримання більш точних результатів.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методи заповнення пропусків даних у задачах прогнозного моделювання соціально-економічних процесів / П. І. Бідюк, О. М. Терентьев, Т. І. Просянкіна-Жарова // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту : матеріали міжнародної наукової конференції (ISDMCI-2017). – Херсон, ПП Вишемирський В. – С. 185-187. – Бібліогр.: 2 назви.

2. Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, 86(3), 343–358.

3. Balances - Energy Map. Energy Map. URL: <https://map.ua-energy.org/en/datasets/d1fb93e6-7751-4d70-9ee9-633747ff83d9/resources/> (date of access: 18.09.2023).

# ЛІНІЙНІ МОДЕЛІ ДЛЯ ПРОГНОЗУВАННЯ ТА КЕРУВАННЯ ФІНАНСОВО-ЕКОНОМІЧНИМИ ПОКАЗНИКАМИ НА ОСНОВІ КОГНІТИВНИХ КАРТ

Зуєвський Ю.В., Мілявський Ю.Л.

Національний технічний університет України «Київський політехнічний інститут  
ім. Ігоря Сікорського», Київ, Україна

**Фінансова сфера постійно зазнає впливу багатьох факторів, і точне прогнозування фінансових показників стає ключовим завданням для успішного управління. Одним із підходів до аналізу цих показників є побудова лінійних моделей когнітивних карт, моделювання, аналіз та керування на їх основі. Однією з проблем при побудові когнітивної карти є визначення вагових коефіцієнтів ребер карти на основі історичних даних системи, що моделюється. У даній роботі розглянуто деякі підходи до ідентифікації, а саме, на основі лінійної регресії та на основі кореляційного аналізу. Показано, що ці підходи можуть співіснувати і доповнювати одне одного у випадках, коли можна виділити залежну і незалежні змінні у моделі. Далі, пропонується відповідно два підходи до прогнозування - один на основі лінійної регресії для прогнозування залежної змінної, інший - на основі сценарного аналізу імпульсного процесу когнітивної карти для прогнозування решти змінних. Проведено моделювання побудованої таким чином когнітивної карти конкретної системи на історичних даних.**

**Ключові слова:** ідентифікація, когнітивні карти, лінійні моделі, регресія, кореляційний аналіз, фінансові ризики, сценарний аналіз, прогнозування.

## 1. ВСТУП

В сучасному світі фінансова діяльність відіграє важливу роль у глобальній економіці, і її результати мають значущий вплив на стабільність та розвиток різних сфер життя. Однак, незважаючи на значення фінансових операцій, вони завжди супроводжуються елементом невизначеності та ризику. Ризик є невід'ємною частиною будь-якої фінансової діяльності, і відсутність можливості передбачити його вплив може призвести до негативних наслідків для фірми, інвесторів та всієї економіки.

Фінансові ризики, як випадкові величини, виникають внаслідок впливу багатьох факторів, включаючи зміни в ринкових умовах, процентних ставках, валютних курсах, інфляції та багатьох інших. Тому, для забезпечення ефективного управління фінансовими ризиками, необхідно розробляти моделі, які допомагають прогнозувати їх вплив та вчасно приймати відповідні рішення.

Отже, **актуальність теми** полягає в проведенні досліджень спрямованих на вивчення можливостей використання лінійних моделей імпульсних процесів когнітивних карт для ефективного прогнозування та керування фінансово-економічними показниками. Наша робота спрямована на розвиток нових методів управління фінансами, які допомагають компаніям та інвесторам краще розуміти та керувати фінансовими ризиками, забезпечуючи стабільність та успішність їхньої діяльності.

Метою роботи є вдосконалення методів прогнозування та керування фінансово-економічними показниками на основі когнітивних карт. Для досягнення поставленої мети в роботі сформульовані наступні задачі:

- аналізу структурних залежностей між змінними та ідентифікація зв'язків між змінними та їх ваги
- математичне моделювання прогнозування електроспоживання населенням в часі з урахуванням зміни напруги в мережі з розробкою когнітивної карти

## **2. МЕТОДИ ПОБУДОВИ ТА АНАЛІЗУ КОГНІТИВНИХ КАРТ**

Припустимо, що у динамічній фінансово-економічній складній системі, яку можна моделювати у формі лінійного імпульсного процесу в когнітивній карті, існує залежна змінна, яка представляє найбільший практичний інтерес, і ряд незалежних змінних, пов'язаних із нею та між собою. Пропонується використати різні підходи до оцінювання ваг ребер когнітивної карти, які з'єднують вершину - залежну змінну з вершинами – незалежними змінними, і ребер, які з'єднують вершини - незалежні змінні між собою.

Перевагою лінійної регресії є простота інтерпретації, здатність визначати кількісний вплив незалежних змінних на залежну. Недоліком є вимога до лінійності зв'язку між змінними, чутливість до викидів і нелінійних взаємозв'язків.

Основна математична модель, яка використовується в лінійній регресії, представлена як

$$y = Xa + \varepsilon, \text{ де}$$

$y$  – вектор залежної змінної,

$X$  – матриця незалежних змінних,

$a$  – вектор невідомих коефіцієнтів,

$\varepsilon$  – вектор випадкових помилок.

Далі на основі побудованої лінійної регресії можна здійснювати прогнозування залежної змінної та аналіз її чутливості до змін у незалежних змінних. При цьому будуватимемо регресію між приростами незалежних і залежних змінних, наближаючись таким до підходу Ф. Робертса у когнітивних картах.

Але як відомо, сценарний аналіз імпульсних процесів когнітивних карт також має ряд переваг, зокрема, він дозволяє системно, комплексно аналізувати динаміку складної системи, бачити взаємовпливи між змінами в будь-яких вершинах карти і розуміти, як зміна однієї вершини вплине за всі інші. Для того, щоб знайти ваги ребер когнітивної карти, можна використати коефіцієнти лінійної регресії, побудовану перед цим. але це лише частина ребер карти. Для решти пропонується використати коефіцієнти кореляції між рядами приростів значень у вершинах – незалежних змінних, які можна легко отримати за наявних історичних даних.

Обидва ці підходи, застосовані разом, можуть надати більш повне уявлення про потенційні фінансові результати та ризики, даючи можливість краще управляти фінансово-економічними показниками. Лінійна регресія дозволяє здійснювати точні кількісні прогнози, тоді як сценарний аналіз допомагає адаптуватися до різних можливих умов ринку.

## **3. МОДЕЛЮВАННЯ ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ**

Для дослідження було обрано датасет, що називається "Household Power Consumption", що містить інформацію про споживання електроенергії в домогосподарствах. Як залежну змінну було обрано змінну `globalactiverpower`, що відображає загальну активну потужність, цей атрибут представляє собою середню потужність, споживану домогосподарством, в кіловаттах (кВт). Вибір лінійної регресії обґрунтований, тим що залежність між змінними лінійна. Цей метод добре підходить для простих, малих та середніх датасетів, де очікується, що зв'язок між

змінними можна апроксимувати лінійною функцією. Для обчислення матриці було використано метод найменших квадратів (МНК) для оцінки коефіцієнтів, що є класичним підходом в лінійній регресії. МНК мінімізує суму квадратів різниць між спостережуваними та прогнозованими значеннями.

Логіка полягає в аналізі та прогнозуванні змін в споживанні електроенергії, виходячи з різних параметрів, що відображаються у датасеті. Використовуючи лінійну регресію, можна зрозуміти, як різні фактори (наприклад, реактивна потужність, напруга тощо) впливають на активну потужність. Візуалізації допомагають інтерпретувати залежності та зв'язки між змінними.

Для оцінки моделі було визначено  $R^2$  – коефіцієнт детермінації, що вимірює, яка частина варіативності залежної змінної пояснюється незалежними змінними. Хороша модель матиме  $R^2$  близько до 1. Також використовуємо MSE – середньоквадратичну похибку, що вимірює середнє квадратичне відхилення між фактичними та прогнозованими значеннями. Чим нижче значення MSE, тим краще модель.

Використання когнітивної карти з вагами, визначеними на коефіцієнтах регресії та кореляціях, є корисним для візуалізації та аналізу структури залежностей між змінними. Когнітивна карта дозволяє легко ідентифікувати важливі зв'язки та потенційні джерела мультиколінеарності.

В результаті реалізації моделі було отримано когнітивну карту, що зображено на рис. 1.

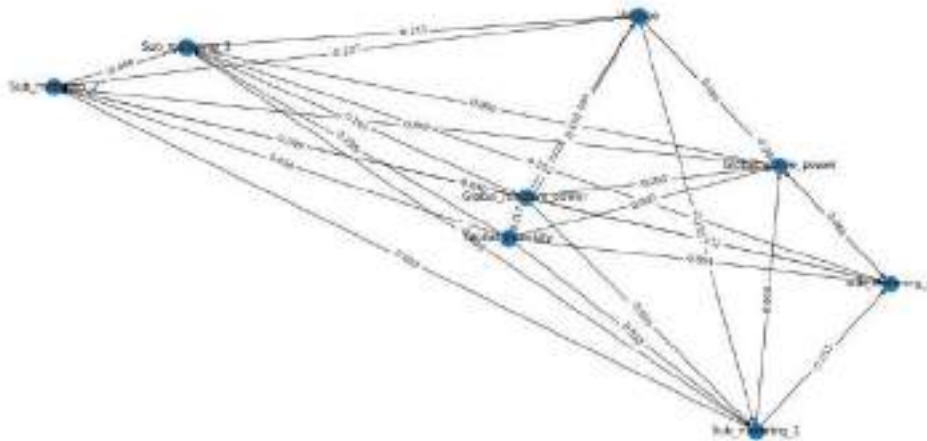


Рисунок 1. Когнітивна карта

Ми можемо спостерігати відображення когнітивної карти, на якій відображаються зв'язки між змінними та їх ваги. З рис. 2 можемо бачити теплокарту кореляційних зв'язків змінних. Теплокарта – це графічне зображення даних, де окремі значення в матриці відображаються за допомогою кольорів.

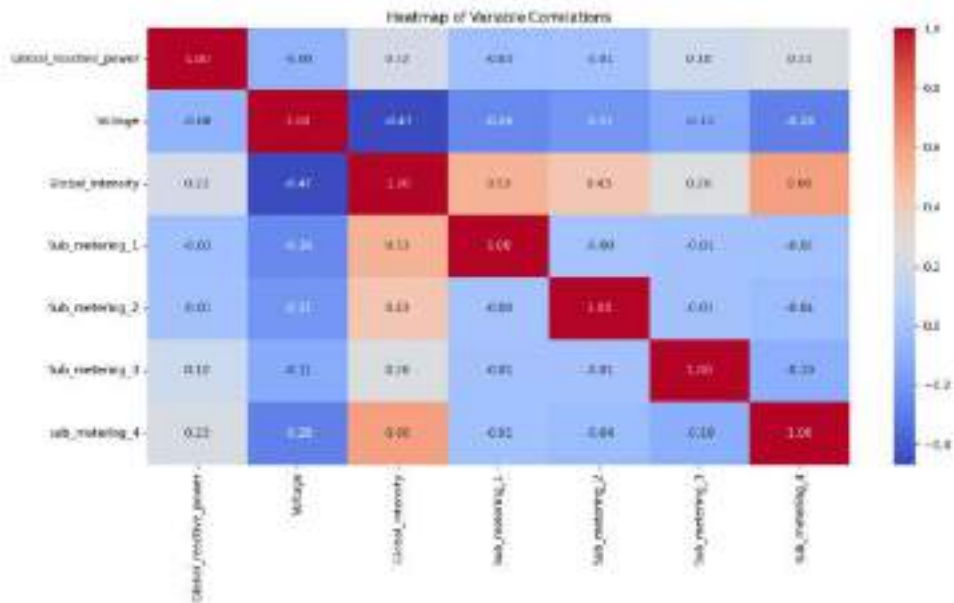


Рисунок 2. Теплокарта значень кореляції змінних

Також було отримано результат передбачення зміни середнього використання електроенергії, результат зображено на рис. 3.

```
C:\Users\Hp\Desktop\linear>python pr.py
R-square (R2): 0.9999999999999537
Mean Squared Error (MSE): 3.3046878919759077e-15
Predicted change in Global_active_power: [0.93200002]
```

Рисунок 3. Результати програми

Показник  $R^2$  є майже ідеальним, оскільки він наближається до 1. Такий високий  $R^2$  означає, що модель майже ідеально відповідає даним. MSE є мірою середньої квадратичної помилки між фактичними та прогнозованими значеннями. Дуже низьке значення MSE, як у вашому випадку, вказує на те, що помилка між прогнозованими та фактичними значеннями є мінімальною, що є хорошим показником ефективності моделі.

Прогнозоване змінення у `Global_active_power` становить приблизно 0.932 одиниць. Цей результат означає, що згідно з лінійною регресійною моделлю та вхідними даними для прогнозу, очікується, що `Global_active_power` збільшиться на 0.932 одиниць порівняно з останнім відомим значенням у датасеті.

Також було здійснено сценарний аналіз для відображення впливу зміни показників на прогнозоване значення використання електроенергії. На рис. 4–6 відображено вплив зміни показників.

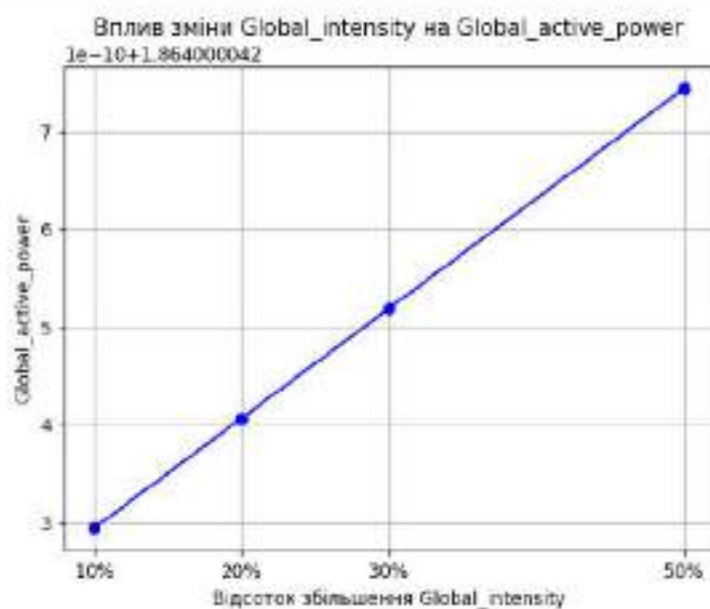


Рисунок 4. Зміна показника Global Intensity

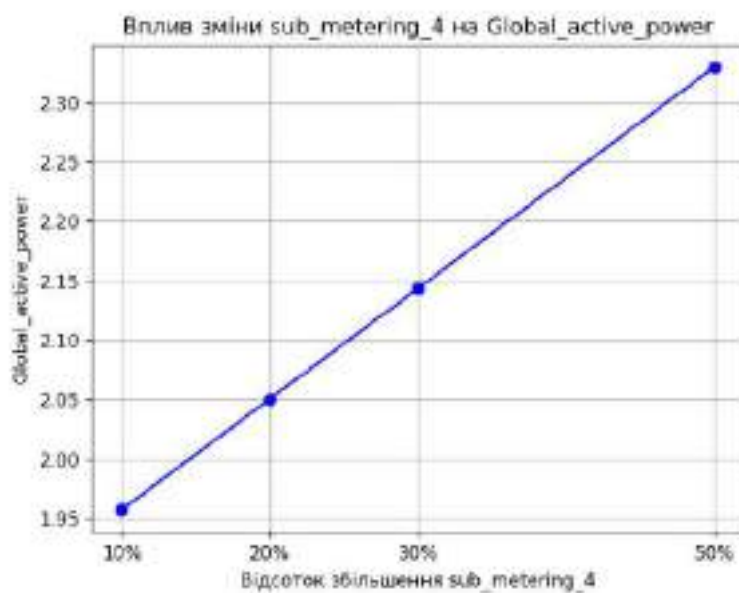


Рисунок 5. Зміна показника Sub Metering 4

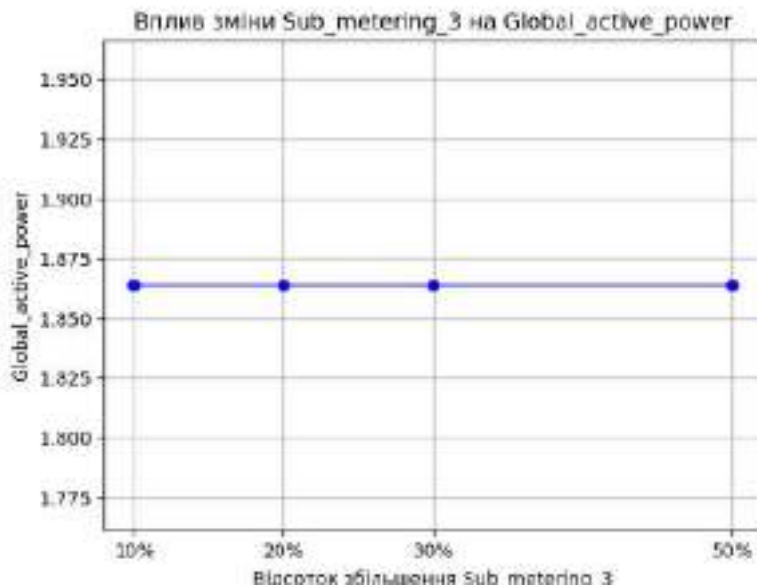


Рисунок 6. Зміна показника Sub Metering 3

Збільшення використання різних видів електроенергії дав наступні результати на зміну загального використання електроенергії (рис. 7).

Прогнозоване значення Global\_active\_power: 1.864000042182632

Рисунок 7. Прогнозоване значення

Даний прогноз може бути корисним у різних сценаріях, таких як планування навантаження на електромережу, аналіз енергоефективності тощо. В залежності від контексту, таке збільшення може бути важливим. Даний прогноз є індикатором можливих змін у споживанні електроенергії за заданими умовами.

#### 4. ВИСНОВКИ

Прогнозування та побудова когнітивних карт є важливими інструментами для аналізу фінансових даних та прийняття рішень. Вони допомагають розуміти складну інформацію, визначати взаємозв'язки та розробляти стратегії для подальшого розвитку. У ході дослідження було досліджено питання побудови когнітивних карт на основі фінансових даних та прогнозування цих даних на основі побудованих когнітивних карт. Це відіграє важливу роль у сучасному бізнесі та науці, допомагаючи покращити прийняття рішень та аналізувати складну інформацію. Було побудовано когнітивні карти та здійснено прогнозування за допомогою двох методів – лінійної регресії та сценарного аналізу імпульсного процесу когнітивної карти. Сценарний аналіз дає більші можливості для відслідковування зміни та впливу показників на середнє використання електроенергії. Висока точність моделі та здатність виявляти зв'язки між різними параметрами споживання електроенергії може бути корисною для розуміння та оптимізації використання енергоресурсів у домогосподарствах.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Mikhail Z. Zgurowsky, Victor D. Romanenko, Yuriy L. Milyavskiy. Principles and Methods of Impulse Processes Control in Cognitive Maps of Complex Systems. Part I // Journal of Automation and Information Sciences. – 2016. – Vol. 48, No. 3. – P. 36-45.
2. Mikhail Z. Zgurowsky, Victor D. Romanenko, Yuriy L. Milyavskiy. Principles and Methods of Impulse Processes Control in Cognitive Maps of Complex Systems. Part II // Journal of Automation and Information Sciences. – 2016. – Vol. 48, No. 7. – P. 4-16.
3. Mykhailo Zgurovsky, Victor Romanenko, Yuriy Milyavsky. Adaptive Control of Impulse Processes in Complex Systems Cognitive Maps with Multirate Coordinates Sampling // Advances in Dynamical Systems and Control, Studies in Systems, Decision and Control, 69. – Springer International Publishing . - Switzerland, 2016. - P. 363 – 374.
4. V. Gubarev, V. Romanenko, Y. Miliavskiy. Identification and Control Automation of Cognitive Maps in Impulse Process mode / Kuntsevich, Gubarev, Kondratenko, Lebedev, Lysenko (eds.), *Control Systems: Theory and Applications* – River Publishers, 2018. - P. 43 – 64.
5. Miliavskiy Yu.L. Identification in cognitive maps in impulse process mode with incomplete measurement of nodes coordinates // Кибернетика и вычислительная техника. – 2019. – № 1 (195). – С. 49–63.
6. V. F. Gubarev, V. D. Romanenko, Yu. L. Miliavskiy. Methods for Finding a Regularized Solution When Identifying Linear Multivariable Multiconnected Discrete Systems // Cybernetics and Systems Analysis. – 2019. - Volume 55, Issue 6. – P. 881–893.

# СППР ДЛЯ ДОСЛІДЖЕННЯ РИНКОВИХ РИЗИКІВ

Квашук І.О.<sup>1</sup>, Кузнецова Н.В.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

<sup>1</sup> illiakvashuk@gmail.com [0009-0006-5585-3045], <sup>2</sup> natalia-kpi@ukr.net [0000-0002-1662-1974]

**Створення системи підтримки прийняття рішень є особливо актуальним для ринкових фінансових операцій. Можливість прогнозування акцій, оцінка основних параметрів та розрахунок оптимального портфелю є ключовими задачами, які можуть бути виконані СППР. В роботі було виконано тестування системи на наборі даних Apple, Google, Amazon та Tesla. Було виконано прогнозування акцій методами лінійної регресії, випадкового лісу та SVM. Було виконано моделювання величини VaR методом Монте-Карло та розрахований оптимальний портфель на основі спрогнозованих акцій.**

**Ключові слова:** ринковий ризик, СППР, VaR, балансування портфелю

## 1. ВСТУП

Фінансові операції на ринку пов'язані з ризиком, що виникає в результаті дій інших гравців та агентів. Разом вони формують тренд або зміну на фінансовому ринку: коливання акцій, інший склад і прибутковість оптимального портфелю. Все це разом призводить до того, що дії на фінансовому ринку пов'язані з невизначеністю в майбутньому та мають тенденцію до змін. Кожний гравець на ринку прагне до максимізації власних прибутків та оптимальності дій. Інформація про стан та умови на ринку зараз та в майбутньому дозволяє оцінити величину ризику та визначити найкращу модель поведінки або навіть конкретну дію: продаж акції, купівля, збільшення інвестицій. Для цього потрібно провести розрахунки, виконати оцінку та підбір параметрів. З метою пришвидшення процесу прийняття рішення і зручної роботи спеціалісту, розроблюються системи підтримки прийняття рішень. Вони містять як модуль введення даних, так і модель виведення – графіки, прогнози та інформацію, що буде надана користувачу.

Системи підтримки прийняття рішень та їх модулі широко представлені та доступні. Разом з тим кожна окрема комбінація модуля та закладеного в ньому функціоналу, варіативність моделей та можливості вирізняються [1, 2]. Кожний підхід до розробки подібної системи підтримки прийняття рішень дозволить надати трейдерам або іншим користувачам більшу варіативність в виборі оптимальної стратегії інвестування.

## 2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Метою магістерської роботи є розробка системи підтримки прийняття рішень, яка буде надавати інформацію щодо фінансових даних відповідно до введеної користувачем інформації. СППР повинно розв'язувати наступні задачі: прогнозування цін акцій, що були обрані користувачем, моделювання стану ринку з метою отримання величини можливих втрат. Окремою задачею є побудова портфелю та розрахунок його оптимального складу на основі спрогнозованих результатів.

### 3. ОСНОВНІ МЕТОДИ

Акції компанії залежать від великої кількості внутрішніх змінних – стан компанії, новини, зміна керівників та CEO, рівні продажів – та від значної кількості зовнішніх факторів – стану сектору та ринку, нових правил регулювання. Економічні та політичні кризи також значно впливають на стан акцій. Це робить прогнозування складною задачею, адже через складеність процесів – великої кількості факторів, що лежать в основні змін – неможливо врахувати всі значення та важливі змінні для надання точного та безпомилкового прогнозу. Розрахунок стану акцій – це основа для всіх інших дій. Для розв’язання задачі прогнозування акцій було використано декілька моделей. Це дозволить знайти найбільш оптимальну модель для конкретного проміжку часу. Поведінка акцій під час різних періодів часу може мати різний характер та описуватися різними моделями.

Однією з обраних моделей є модель, що будується за допомогою SVM (Support Vector Machine) і застосовує методи регресії та класифікації. Основна ідея методу базується на знаходженні гіперплощини, що буде застосовуватися до розділення класів. Такий підхід використовується для розв’язання задачі класифікації, проте після модифікацій він може використовуватися і для задачі розв’язання регресії. Особливістю методу є те, що він може не обмежуватися розмірами простору початкових даних, а використовуючи техніку Kernel Trick, може переходити до просторів більшої розмірності. Це дозволяє розв’язати проблему, коли в даному просторі елементи не роздільні гіперплощиною [3, 4].

Наступною моделлю було вирішено обрати модель лінійної регресії. Модель лінійної регресії – це універсальний інструмент, що використовується в багатьох задачах в реальному світі, включаючи і фінансові операції. Не зважаючи на свою простоту, ця модель – це потужний інструмент розв’язання задач, та її поведінка та особливості добре вивчені та дослідженні в багатьох роботах [5, 6]. Суть лінійної регресії полягає в оптимальному виборі параметрів, що мінімізують певний критерій.

Модель випадкового лісу – це модель, що належить до сфери машинного навчання. Особливістю моделі є використання наборів дерев рішень, кожне з яких навчається на окремій складовій даних, та прийнятті рішення на засадах ранжування пропозицій наданих кожним окремим деревом. Пропозиція, за яку голосує найбільша кількість дерев, обирається як результат роботи моделі.

Всі моделі будуть використовуватися для прогнозування акцій. Водночас, учасники ринку можуть надавати перевагу іншому підходу, що базується на розрахунках певної, майже гарантованої, величини втрат. На основі цього можна проводити оцінку стану та рівню ризику для окремих операцій. Розрахована величина – це VaR і для її моделювання можна застосовувати різні підходи. Ми зупинили свій вибір на моделі Монте-Карло.

VaR (Value-at-Risk) – це величина, що знаходиться під ризиком втрати. Це один з основних фінансових показників, яка розраховується для отримання розуміння про величину потенційних (можливих) втрат.

Метод Монте-Карло – це метод, що застосовується для генерацій вибірок. В даному випадку метод використовується для отримання можливих траєкторій розвитку акцій з метою оцінки показників. Метод може використовувати різні розподіли і найкращим [7].

### 4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Реалізація системи підтримки прийняття рішень з метою розв’язання поставленої задачі вимагає перевірки роботи моделі. Для цього було вирішено обрати набір певних акцій і для них провести прогнозування, балансування портфелю та моделювання значення VaR. Дані можуть бути отримані в реальному часі, оскільки ринкові показники компаній наявні на біржах, де відбувається їх лістинг. З множини всіх доступних акцій було прийнято рішення

взяти акції великих технологічних компаній, оскільки багато учасників ринку купують їх акції, і перевірка роботи системи дозволить отримати максимально наближені до реальних.

Для перевірки якості моделей було обрано акції компанії Apple та прогнозувалась ціна їх закриття (Рис. 1). Існує декілька характеристик, що описують стан акцій – ціна відкриття, ціна закриття, проте було вирішено зупинитися на ціні закриття по тій причині, що прогноз закриття включає в себе зміни протягом одного дня, а також додаткову величину варіативності, що залишає трейдерам можливість розраховувати стан акцій та планувати власні дії.

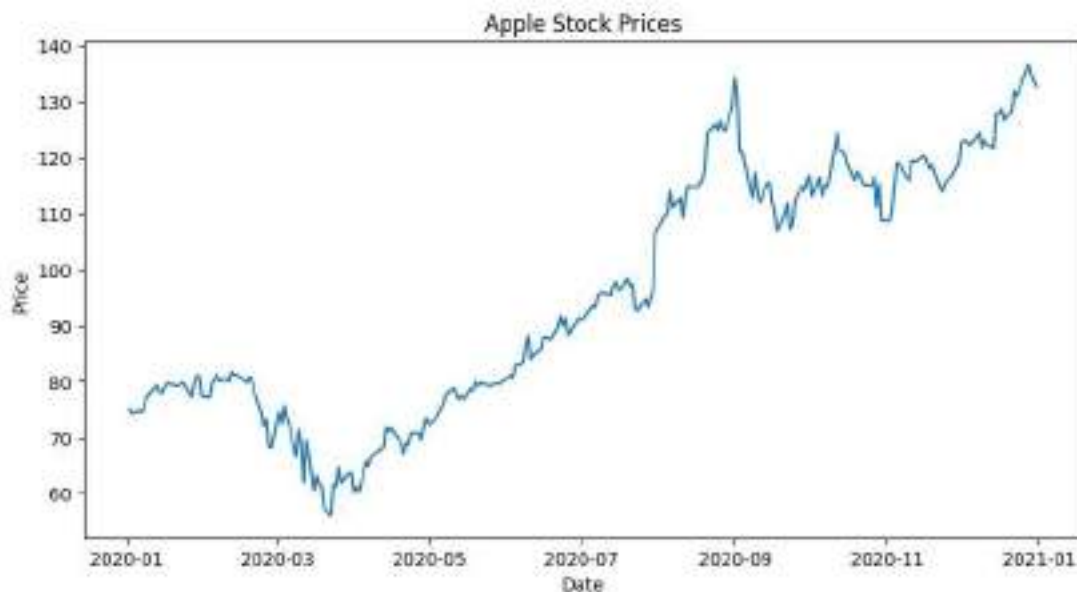


Рисунок 1. Графік акцій Apple

Для моделювання було обрано період в один рік. В цінах на акцій наявні всі історичні данні, тому для навчання моделі можна використовувати будь-який проміжок. Було виконано моделювання трьома моделями – випадковий ліс, лінійна регресія та метод SVM.

На Рис. 2 представлено графічні результати роботи моделей. Як видно, всі побудовані моделі відображають зміни руху в акціях. Разом з тим, очевидно що SVM в цьому випадку був найменш вдалою моделлю.

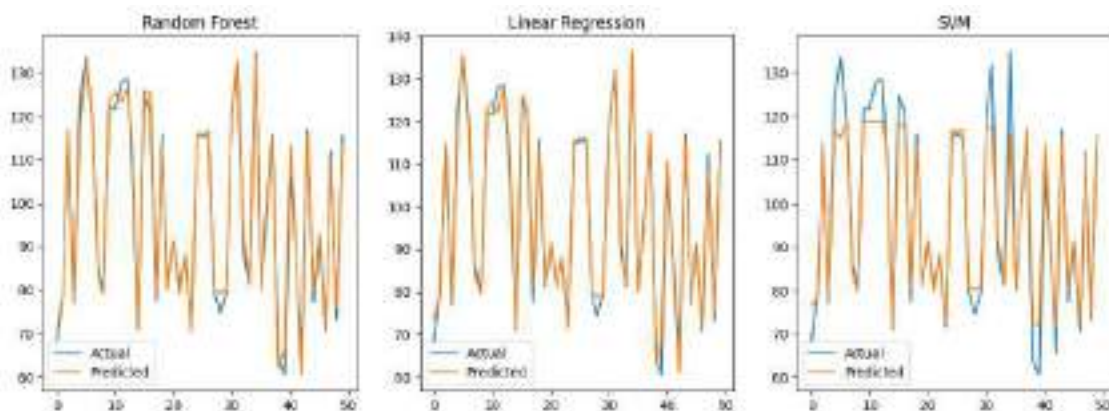


Рисунок 2. Графіки роботи моделей

Для порівняння наведемо таблицю результатів (Табл. 1) і порівняємо тільки дві найкращі моделі. Отримано значення VaR, а також величина помилки відповідно до різних критеріїв – MAE, MSE, RMSE. Найбільш вдалою є модель лінійної регресії, що має найкращі показники для помилки, а, отже, і точність. Загалом, лінійна регресія найкраще показує себе для задачі прогнозування стану акцій.

Таблиця 1. Параметри моделі

Metric	Random Forest	Linear Regression
VaR	3538.68	3518.17
MAE	2.15	1.91
MSE	7.78	6.99
RMSE	2.79	2.64
R <sup>2</sup>	0.98	0.99

Отримані результати дозволяють розрахувати значення VaR. Альтернативним підходом до розрахунку VaR може слугувати моделювання методом Монте-Карло. Тут було обрано більший проміжок для дослідження ціни та виконане моделювання (Рис. 3). Метод Монте-Карло може використовувати різні розподіли під час моделювання, при цьому відомо, що розподіли акцій в більшій мірі слідує розподілу Стюдента, аніж нормальному розподілу.

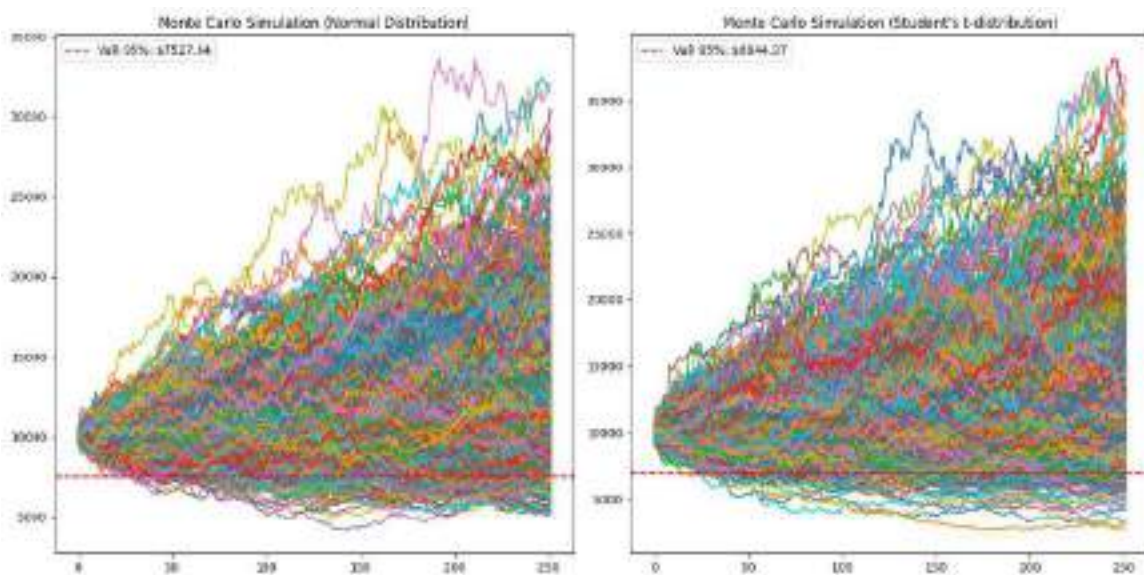


Рисунок 3. Моделювання методом Монте-Карло з різними розподілами

Розуміння поведінки акцій, які отримуються в результаті моделювання та прогнозування, надає можливість створення збалансованого портфолію з декількох акцій. Задача його розрахунку є ключовою для отримання прибутків.

Нами було обрано 4 акції – Apple, Google, Amazon та Tesla. Для них буде виконано оптимальний розрахунок співвідношення в портфелі протягом трьох років. Як видно з Рис. 4, оптимальне співвідношення змінювалося протягом часу разом зі зміною прибутковості акцій.

Під час певного періоду найбільш прибутковими були акції компанії Google, що майже повністю повинні скласти портфель при його оптимальній побудові.

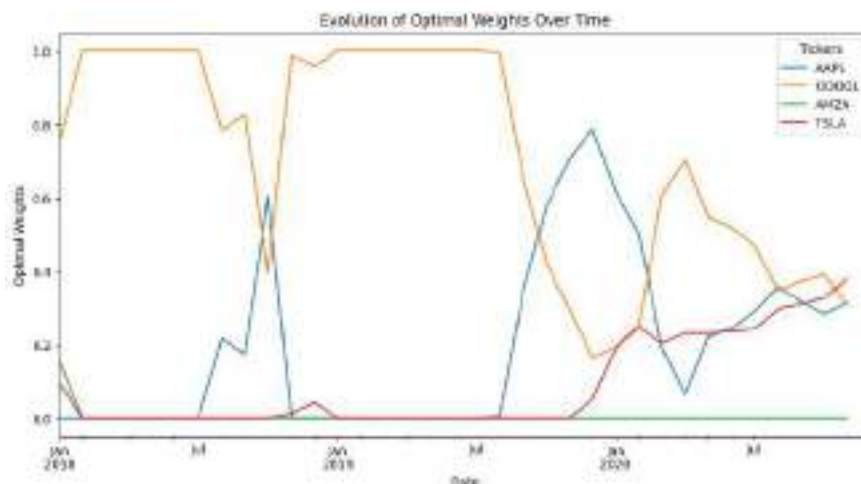


Рисунок 4. Графік зміни оптимальних пропорцій портфелю

Починаючи з середини 2020 року його прибутковість падає (Рис. 5). Акції компанії Amazon мають частку 0 для портфелю. Це пояснюється величиною зростання інших акцій, для яких прибуток був більший. Відповідно до складеного портфелю можна отримати значення портфелю за весь час.

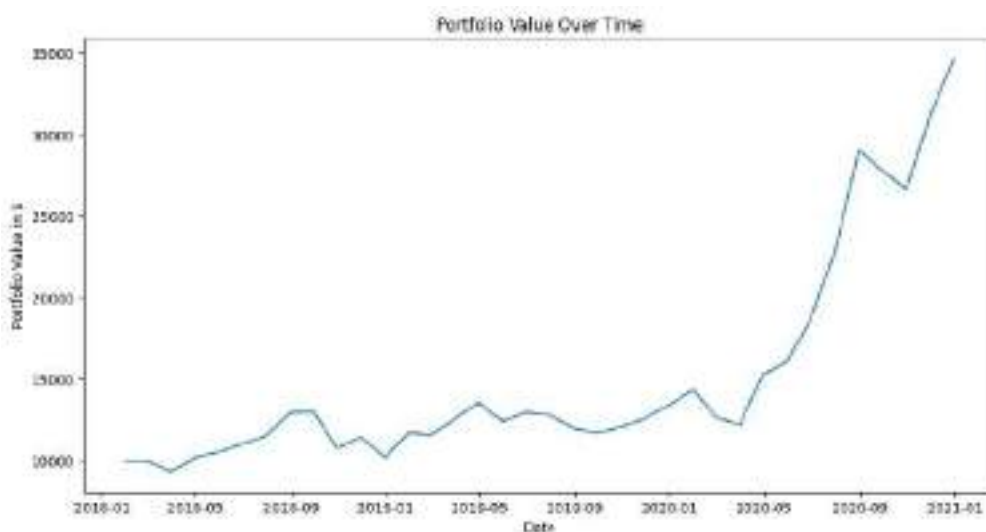


Рисунок 5. Графік прибутковості акцій

## 5. ВИСНОВКИ

Для проведення безбиткових операцій на фондових ринках потрібно опиратися на достовірний прогноз стану акцій. Для максимізації доходів інвестора необхідно розробити оптимальний портфель інвестора, що включатиме якнайбільше прибуткових акцій та найменше спадаючих акцій під час кризи. Розроблена СППР допомагає виконати отримані необхідні дані. Три розроблених моделі дозволяють отримати варіативність, що разом з модулем моделювання дозволяє отримати значення ключового значення параметру VaR різними способами.

Отримані за результатом моделі задовільно справились з задачею прогнозування, показали достатньо високу точність та низьку величину помилки. Шляхом відбору найкращої моделі можливо отримати найбільш підходящий для даної акції модель та прогноз.

Альтернативний спосіб обчислення VaR методом Монте-Карло дозволяє використовувати різні розподіли та надає додаткову варіативність. При наявності інформації про стан ринку та можливий розподіл – точність розрахунку даним методом зростає.

Метод побудови оптимального портфелю акцій дозволяє інтегрувати результати роботи модулів прогнозування та розрахувати найбільш прибуткове співвідношення для портфелю.

Загалом, побудована СППР виконує поставлені задачі та реалізує необхідний для роботи з фінансовими даними функціонал. Можливими шляхами до вдосконалення є автоматичний підбір розподілу для методу Монте-Карло використовуючи наявну інформацію про акції. Більша кількість моделей для прогнозування акцій також дозволить, при наявності відповідних можливостей для покриття апаратних вимог, покращити якість та точність прогнозування.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. G. H. Van Bruggen, A. Smidts, and B. Wierenga, “Improving decision making by means of a marketing decision support system”, *Management Science*, 44(5), pp. 645-658, 1998. doi: 10.1287/mnsc.44.5.645.

2. J. D. Little, “Decision support systems for marketing managers”, *Journal of Marketing*, 43(3), pp. 9-26, 1979. doi: 10.1177/002224297904300302.

3. R. Yang, L. et al., “Big data analytics for financial Market volatility forecast based on support vector machine”, *International Journal of Information Management*, 50, pp. 452-462, 2020. doi: 10.1016/j.ijinfomgt.2019.05.027.

4. W. Huang, Y. Nakamori, and S. Y. Wang, “Forecasting stock market movement direction with support vector machine”, *Computers & operations research*, 32(10), 2513-2522, 2005. doi: 10.1016/j.cor.2004.03.016.

5. T. L. Lai, and H. Xing, “Linear regression models”, *Statistical Models and Methods for Financial Markets*, pp.3-35, 2008. doi: 10.1007/978-0-387-77827-3\_1.

6. R. Seethalakshmi, “Analysis of stock market predictor variables using Linear Regression”, *International Journal of Pure and Applied Mathematics*, 119(15), pp. 369-378. 2018. doi: 10.1109/eSmarTA52612.2021.9515762.

7. M. C. Fu, and J. Q. Hu, “Sensitivity analysis for Monte Carlo simulation of option pricing”, *Probability in the Engineering and Informational Sciences*, 9(3), pp. 417-446, 1995. doi: 10.1017/S0269964800003958.

# **СТАБІЛІЗАЦІЯ ДЕМОГРАФІЧНИХ ПРОЦЕСІВ ЗА ДОПОМОГОЮ УПРАВЛІННЯ ПО ПРОГНОЗУЮЧІЙ МОДЕЛІ**

Нетудихата А.С.<sup>1</sup>, Губарев В.Ф.

Національний технічний університет України «Київський політехнічний інститут  
ім. Ігоря Сікорського», Київ, Україна

<sup>1</sup> netudykhata.arina@lil.kpi.ua

**Мета дослідження** полягає в стабілізації демографічних процесів через застосування управління по прогнозуючій моделі. Дослідження демографічних процесів здійснюється на основі моделі когнітивної карти. Отримана модель є нестійкою, тому при відсутності управління параметри демографічних процесів можуть необмежено збільшуватись. Тому необхідна стабілізація таких процесів. Модель було створено аналітичним методом на основі реальних даних. Методами моделювання досліджено поведінку демографічних процесів по отриманій моделі. Наукова новизна полягає в ефективному застосуванні прогнозного управління для стабілізації демографічних процесів. Практична значимість полягає в тому, що за допомогою запропонованого методу вдалося вирішити демографічну проблему в Україні. В роботі проведено багато чисельних експериментів, які демонструють ефективність прогнозного управління.

**Ключові слова:** демографічні показники, управління по прогнозуючій моделі, когнітивне моделювання, стабілізація, нестійкі процеси.

## **1. ВСТУП**

Згідно з даними, опублікованими на офіційному веб-сайті The World Bank Group [1], показник загального населення України скоротився з 52 мільйонів до 38 мільйонів людей. Цей демографічний зсув створює необхідність у глибокому розумінні основних демографічних показників України для розробки ефективних стратегій управління населенням.

У контексті когнітивного моделювання важливо враховувати, що система може бути нестабільною, що ускладнює аналіз та передбачення демографічних процесів. Для досягнення стабільності та ефективного управління демографічними змінами, використання управління по прогнозуючій моделі [2] може виявитися важливим інструментом. У даному дослідженні розглядається можливість застосування цього підходу для стабілізації системи демографічних процесів в Україні.

Враховуючи нестабільність системи та важливість демографічних показників, дослідження спрямовано на розуміння та вирішення проблем, пов'язаних зі змінами у населенні. Підкреслюється актуальність вивчення демографічних тенденцій та їх впливу на суспільство для розробки ефективних стратегій управління населенням.

## **2. КОГНІТИВНЕ МОДЕЛЮВАННЯ ТА СТАБІЛІЗАЦІЯ**

### **2.1. Стабілізація за допомогою управління по прогнозуючій моделі**

Керовані демографічні процеси описуються рівняннями

$$\Delta\bar{Y}(k+1) = A\Delta\bar{Y}(k) + B\Delta\bar{U}(k), \quad (1)$$

де  $A$  і  $B$  – відомі матриці розмірностей  $11 \times 11$  та  $11 \times 6$ ;  $\Delta\bar{Y}$  і  $\Delta\bar{U}$  – вектори стану та керування розмірностей  $11$  та  $6$  відповідно.

Вважається, що  $\Delta\bar{Y}$  або вимірюється в кожний момент часу, або оцінюється за допомогою оцінювача стану. Параметри матриці  $A$  такі, що система (1) є нестійкою (існують власні числа, що лежать за межами одиничного кола на комплексній площині).

Ставиться задача: синтезувати керування  $\Delta\bar{U}$  таке, щоб де б система не знаходилася в момент часу  $k$ , існує керування, яке за скінчений час приводить систему до нульового стану  $\Delta Y(k+s) = 0$ , де  $s$  – скінчений час або горизонт керування. Іноді розглядають задачу стабілізації, яку формулюють як асимптотична стабілізація, коли  $s$  прямує до нескінченності.

Таку задачу будемо вирішувати не апіорі для всіх довільних станів, а безпосередньо у реальному часі, де система за результатами вимірювань опинилася. Таким чином реалізується зворотний зв'язок, що забезпечує стабілізацію. Якщо в точці  $k$  ми його знайшли, то в точці  $k+1$  за результатами прогнозного розв'язку система повинна мати певне значення стану. За результатами вимірювань перевіряємо, чи співпадає прогнозне значення з вимірюваним чи ні. Якщо співпадає, то обчислене керування залишається незмінним. В протилежному випадку вирішується задача прогнозу для часу  $k+1$  так, щоб система на горизонті  $s$  мала нульовий стан. Так робиться на кожному кроці функціонування системи у реальному часі. Це означає, що синтез керування робиться не апіорі, а у кожен наступний час або через декілька кроків у реальному процесі. В цьому полягає зміст управління по прогнозуючій моделі зі зворотним зв'язком.

Для реалізації переходимо до траєкторного опису моделі (1):

Нехай  $k$  – початкова точка, що змінюється у часі в процесі функціонування системи або є ковзною точкою. Траєкторію системи на ковзному інтервалі  $[k, k+s]$  згідно з (1) запишемо як

$$\begin{aligned} \Delta\bar{Y}(k+1) &= A\Delta\bar{Y}(k) + B\Delta\bar{U}(k), \\ \Delta\bar{Y}(k+2) &= A^2\Delta\bar{Y}(k) + AB\Delta\bar{U}(k) + B\Delta\bar{U}(k+1), \\ &\vdots \\ \Delta\bar{Y}(k+s) &= A^s\Delta\bar{Y}(k) + A^{s-1}B\Delta\bar{U}(k) + \dots + B\Delta\bar{U}(k+s-1). \end{aligned} \quad (2)$$

Вирішується задача термінального керування: щоб за  $s$  кроків система мала нульовий стан. Перепишемо останнє рівняння в (2) у наступному вигляді

$$[B \ AB \ A^2B \ \dots \ A^{s-1}B] \cdot U(k, s) = \Delta\bar{Y}(k+s) - A^s\Delta\bar{Y}(k), \quad (3)$$

де  $U(k, s)$  – це каскадний вектор, що має вигляд  $U(k, s) = [\Delta\bar{U}^T(k+s-1) \ \Delta\bar{U}^T(k+s-2) \ \dots \ \Delta\bar{U}^T(k)]^T$ , де «Т» – операція транспонування.

Вирішується задача стабілізації, коли термінальний стан в кінці горизонту повинен бути нульовим, тоді  $\Delta\bar{Y}(k+s) = 0$  і рівняння (3) можна записати як

$$\Omega \cdot U(k, s) = -A^s\Delta\bar{Y}(k),$$

де  $\Omega$  – матриця керуваності системи, тобто

$$\Omega = [B \ AB \ A^2B \ \dots \ A^{s-1}B]. \quad (4)$$

При синтезі керування на основі управління по прогнозуючій моделі зі зворотнім зв'язком спочатку обчислюється невизначеність у вигляді похибки кожної із компонент поточного вектора стану. В результаті замість точного значення  $\{\Delta\bar{Y}(k)\}$  у кожній точці процесу керування маємо

$$\Delta\tilde{Y}(k) = \Delta\bar{Y}(k) + \delta\bar{Y}(k)$$

де  $\Delta\bar{Y}(k)$  – точне прогнозне значення вектора стану у точці  $k$ , а  $\delta\bar{Y}(k)$  – сумарний вектор похибки від різних факторів (похибка вимірювань, зовнішні збурення та інші). Опишемо, як

будемо визначати  $\delta \bar{Y}(k)$ . Для цього в розглядуваній точці знаходимо середнє по компонентах значення вектора  $\Delta \bar{Y}(k)$ , тобто

$$\Delta \bar{Y}_{\text{сеп}}(k) = \frac{\sum_{i=1}^n \Delta \bar{Y}_i(k)}{n},$$

де  $\Delta \bar{Y}_i(k)$  –  $i$ -а компонента вектора  $\Delta \bar{Y}(k)$ , а  $n$  його розмірність ( $n=11$ ).

Формуємо скалярну випадкову послідовність  $\{\delta_i\}$ , елементи якої рівномірно розподілені на інтервалі

$$-\varepsilon \leq \delta_i \leq \varepsilon,$$

де  $\varepsilon = 0,1 \cdot \Delta \bar{Y}_{\text{сеп}}(k)$ .

Розмірність випадкової послідовності  $\{\delta_i\}$  повинна бути більша або дорівнювати  $u_n$  ( $n = 11$ ), щоб визначити всі компоненти вектора  $\delta \Delta \bar{Y}(k)$ . Вони обираються з  $\{\delta_i\}$  довільно.

Після цього синтезується управління по прогнозуючій моделі із зворотним зв'язком.

## 2.2. Визначення показників вершин для побудови когнітивної карти

Проаналізувавши показники з офіційного сайту Державної статистики України [3] на 2021 рік, було обрано наступні вершини для побудови когнітивної карти: 1 – середньомісячна заробітна плата одного працівника, 2 – індекс споживчих цін, 3 – експорт товарів, 4 – імпорт товарів, 5 – чисельність населення, 6 – народжуваність населення, 7 – смертність населення, 8 – міграційний рух населення, 9 – ВВП на душу населення, 10 – загальний рівень безробіття, 11 – інфляція.

Експертним шляхом було визначено причинно-наслідкові зв'язки, виділені у таблиці 1:

Таблиця 1. Причинно-наслідкові зв'язки між показниками для побудови когнітивної карти

№ вершини відповідно назві	1	2	3	4	5	6	7	8	9	10	11
1	0	0	0	0	0,02	0,1	0	0	<b>0,7</b>	-0,2	0,3
2	0,4	0	0	0	0	0,2	-0,3	0,3	0,4	0,2	<b>0,5</b>
3	0,3	-0,3	0	0	0	0	0	-0,2	0,4	0	0
4	-0,3	<b>0,8</b>	-0,2	0	0	0	0	0,3	-0,4	0	<b>0,5</b>
5	0	0	0	0	0	0	0	0,4	0	<b>0,85</b>	0
6	0	0	0	0	<b>1</b>	0	0	0	0,3	0	0
7	0	0	0	0	<b>-1</b>	0	0	<b>-0,7</b>	<b>-0,55</b>	0	0
8	0	0,3	0,2	0,2	<b>0,5</b>	0,1	0	0	0	0	0
9	<b>0,7</b>	<b>0,5</b>	0,1	-0,1	0	0,2	-0,2	-0,3	0	-0,45	0,3
10	<b>-0,8</b>	-0,3	0	0	0	0,2	0,4	0	-0,4	0	0
11	<b>0,8</b>	<b>0,9</b>	0,3	<b>-0,9</b>	0	-0,15	0	<b>0,7</b>	0,3	<b>0,7</b>	0

Примітка: напівжирним шрифтом в таблиці 1 виділено сильні абсолютні зв'язки з пороговим значенням 0,5.

Значення таблиці 1 сформували матрицю  $A$ . З даної матриці отримано наступні власні числа:  $\lambda_1 \approx -0,286$ ,  $\lambda_2 \approx 0,142$ ,  $\lambda_3 \approx 1,389$ ,  $\lambda_{4,5} \approx -0,359 \pm i \cdot (0,205)$ ,  $\lambda_{6,7} \approx 0,689 \pm i \cdot (0,331)$ ,  $\lambda_{8,9} \approx -0,883 \pm i \cdot (0,361)$ ,  $\lambda_{10,11} \approx -0,068 \pm i \cdot (0,511)$ . Серед власних чисел матриці  $A$  є числа більші за одиницю (наприклад  $\lambda_3$ ). Таким чином можна зробити висновок, що система є нестійкою.

Було проведено дослідження обумовленості матриці керованості  $\Omega$  (4). Було обчислено число обумовленості для різних значень горизонту  $s$ , починаючи від  $s = 2$ . Визначено, що найкраще число обумовленості отримано при  $s = 2$ , при наступних значеннях горизонту число обумовленості зростає.

### 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

#### 3.1. Когнітивне моделювання без застосування управління по прогнозуючій моделі

Візьмемо початковий стан системи [0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0] і подамо імпульс  $-0.5$  на вершини 10 і 11 та імпульс  $+1$  на вершину 6. Зробимо 10 ітерацій.

На рисунку 1 зображено результати. Як бачимо, всі показники мають не дуже стабільну поведінку, при цьому тренд смертності має плавний зріст, а тренд народжуваності навпаки вийшов в нуль.

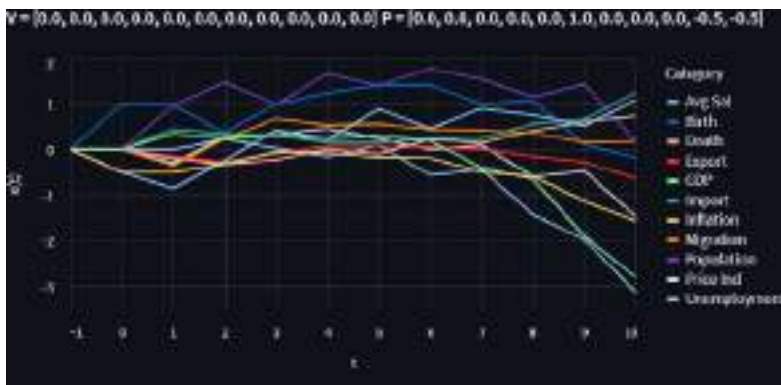


Рисунок 1. Імпульс  $+1$  на народжуваність населення та  $-0.5$  на безробіття та інфляцію

Візьмемо аналогічний початковий стан системи та подамо імпульс на вершини 1 та 6. Зробимо 8 ітерацій. З рисунку 2 можемо побачити, що всі параметри відповідають очікуванням і є більш-менш стабільними. Також варто зазначити, що показники народжуваності та чисельності населення, експорту товарів та ВВП зростають, показники імпорту товарів та смертності спадають. При цьому показники експорту товарів, міграційного руху та безробіття мають незначний зріст. Мінусом є те зростає інфляція, проте це очікувано відповідно до сильного зросту ВВП.

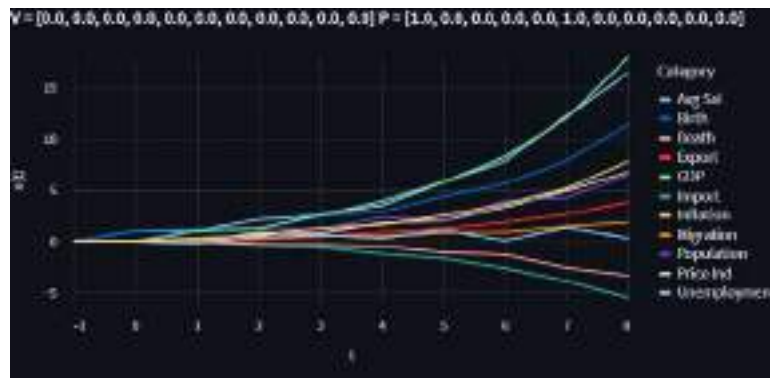


Рисунок 2. Імпульс  $+1$  на народжуваність населення та середньомісячну заробітну плату одного працівника

#### 3.2. Когнітивне моделювання із застосуванням управління по прогнозуючій моделі без зворотного зв'язку

Візьмемо компоненти початкового стану всі рівні 10, подамо імпульс на вершини 1 та 6.

На рисунку 3 можемо спостерігати, що система проявляє нестабільність, ускладнюючи можливість знаходження ефективного розв'язку. Графіки, які відображають поведінку системи, демонструють синусоїдальний характер, що свідчить про циклічні коливання та непередбачувані зміни у демографічних процесах, проте система приходиться в нуль. Очевидно,

що значення компонент початкового вектора завеликі та могли мати вплив на отриманий результат.

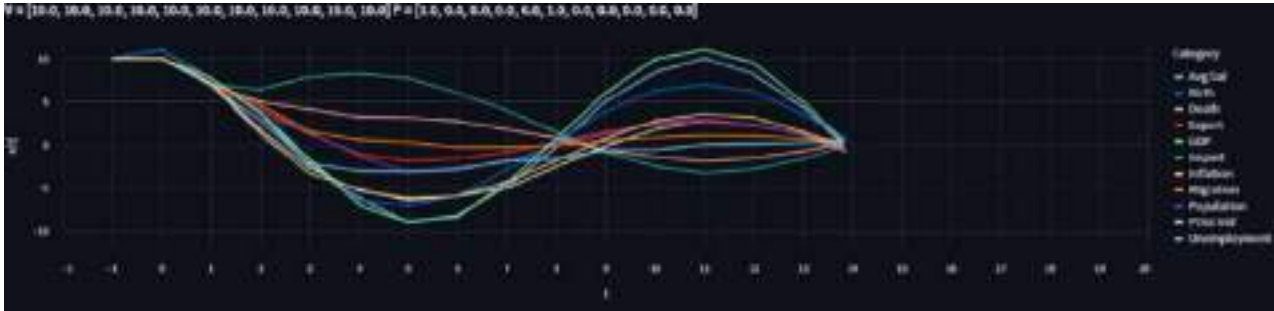


Рисунок 3. Нестабільне керування на основі управління по прогнозуючій моделі

Візьмемо початковий стан системи, де всі компоненти рівні 0. Застосуємо той самий імпульс на показники 1 та 6. Зменшимо довжину горизонту до оптимального, що виконуватиме умову найкращої обумовленості матриці керованості.

На рисунку 4 можемо спостерігати стабілізовану поведінку демографічних процесів. Графіки, що відображають поведінку системи, показують плавні та прогнозовані коливання, що свідчить про успішну стабілізацію демографічних процесів. Застосування управління по прогнозуючій моделі дозволило зменшити синусоїдальний характер графіків та забезпечити більш плавну та передбачувану динаміку. Система досягає нульового стану.

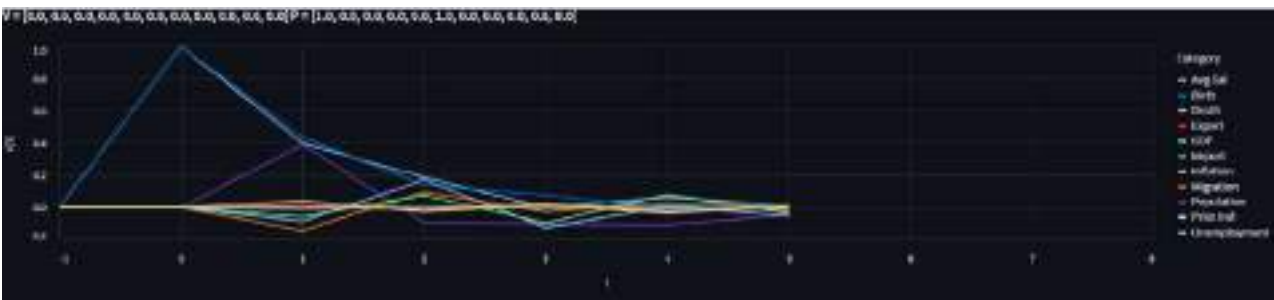


Рисунок 4. Стабільне керування на основі управління по прогнозуючій моделі

### 3.3. Когнітивне моделювання із застосуванням управління по прогнозуючій моделі зі зворотнім зв'язком

Візьмемо компоненти початкового стану всі рівні 10, подамо імпульс на вершини 1 та 6. На рисунку 5 можемо спостерігати графіки реального та прогнозованого значень для компоненти середньої заробітної плати. Система стабілізується, навіть враховуючи великі значення компонент початкового вектора, що свідчить про кращі результати за рахунок використання зворотного зв'язку.

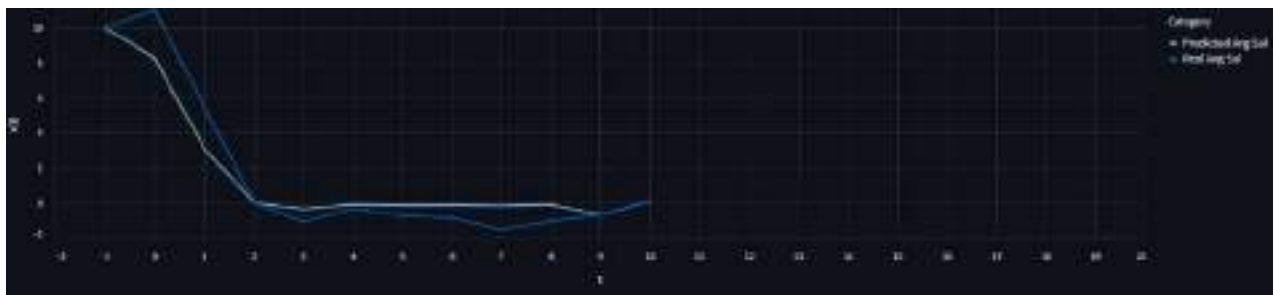


Рисунок 5. Керування на основі управління по прогнозуючій моделі

Візьмемо початковий стан системи, де всі компоненти рівні 0. Застосуємо той самий імпульс на показники 1 та 6. Зменшимо довжину горизонту до оптимального, що виконуватиме умову найкращої обумовленості матриці керованості.

На рисунку 6 бачимо графіки реального та прогнозованого значень для компоненти середньої заробітної плати – графіки відтворюють оптимальний результат вирішення проблеми. Система є достатньо стабільною, показники кращі ніж при моделюванні без зворотнього зв'язку, система плавно досягає нульового стану за скінченний час.



Рисунок 6. Керування на основі управління по прогнозуючій моделі

#### 4. ВИСНОВКИ

В роботі було досліджено когнітивне моделювання демографічних показників України, за допомогою методу модального керування з декількома керуючими сигналами, управління по прогнозуючій моделі без зворотного зв'язку та зі зворотнім зв'язком.

Було визначено, що використання управління по прогнозуючій моделі, особливо зі зворотнім зв'язком, дозволяє досягти стабілізації демографічних процесів. Оптимальна ініціалізація та врахування умов найкращої обумовленості матриці керованості, горизонту, початкових компонент вектора  $\Delta \bar{Y}(0)$  грають важливу роль у досягненні ефективних результатів. Управління по прогнозуючій моделі з врахуванням зворотного зв'язку є перспективним методом для стабілізації та оптимізації демографічних систем.

Результати дослідження підтверджують, що управління по прогнозуючій моделі є перспективним інструментом для стабілізації та оптимізації демографічних процесів. Порівняно з альтернативними методами, цей підхід виявляється більш ефективним та пристосованим до змін в системі.

#### ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. The World Bank Group. World Bank Open Data Ukraine. World Bank Open Data. URL: <https://data.worldbank.org/country/ukraine> (date of access: 05.10.2023).
2. Camacho E. F. Model predictive control. Berlin : Springer, 1999. 280 p.
3. Держстат України. Основні показники соціально-економічного розвитку України. Державна служба статистики України. URL: [https://ukrstat.gov.ua/operativ/operativ2021/mp/arh\\_op2021.html](https://ukrstat.gov.ua/operativ/operativ2021/mp/arh_op2021.html) (дата звернення: 12.09.2023).
4. Романенко В. Д., Мілявський Ю. Л. КОГНІТИВНЕ МОДЕЛЮВАННЯ ДИНАМІКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ СТАБІЛІЗАЦІЇ НЕСТІЙКИХ РЕЖИМІВ У СОЦІАЛЬНО-НАВЧАЛЬНОМУ ПРОЦЕСІ СТУДЕНТА. Наукові вісті НТУУ "КПІ". 2016. № 5. С. 48–53. URL: <https://doi.org/10.20535/1810-0546.2016.5.67264> (дата звернення: 15.08.2023).
5. V. Romanenko and Y. Milyavskiy, "Methods of impulse processes control for cognitive maps with delays", Naukovi visti NTUU "KPI", no. 5, pp. 57 – 63, 2015 (in Ukrainian).
6. В.Д. Романенко, Ю.Л. Мілявський. Методи керування імпульсними процесами когнітивних карт із запізненнями // Наукові вісті НТУУ «КПІ». 2015. – № 5. – С. 57 – 63.

# ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ КЛАСТЕРИЗАЦІЇ СТАНУ КРАЇН ЗА ПОКАЗНИКАМИ СТАЛОГО РОЗВИТКУ

Самсонюк М.В.<sup>1</sup>, Бідюк П.І.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

<sup>1</sup> ms-0807@ukr.net [0000-0001-6593-3504], <sup>2</sup> pbidyuke\_00@ukr.net [0000-0002-7421-3565]

**Метою дослідження є побудова інформаційної системи підтримки прийняття рішень (ІСППР), що ґрунтується на методах інтелектуального аналізу даних (ІАД). Вона дозволить провести кластеризацію стану країн за індикаторами сталого розвитку, а також порівняти існуючі моделі кластеризації за допомогою спеціальних критеріїв. У роботі використані експериментальні дослідження та методи математичного моделювання.**

**Ключові слова:** кластеризація, сталий розвиток, критерії якості, математичні моделі, інтелектуальний аналіз даних, машинне навчання

## 1. ВСТУП

Сталий розвиток є дуже важливим для стабільного та щасливого життя людей на нашій планеті. У випадку, якщо такі проблеми людства як військові конфлікти, розрив між бідними і багатими, руйнування навколишнього середовища не будуть вирішені, людству загрожує повна загибель. Отже, дуже важливою задачею є побудова математичних моделей на основі даних сталого розвитку. Це допоможе оцінити те, у яких саме сферах є найбільші проблеми, які цілі сталого розвитку не досягаються, який вплив на систему керування державою ми можемо здійснити, щоб досягти цілей сталого розвитку та як саме прийти до мінімізації ймовірності настання військових конфліктів. Для України дана тема є особливо актуальною, оскільки вплив пандемії COVID-19 та повномасштабного вторгнення на життя людей є дуже значним. У рамках даного дослідження вирішується задача кластеризації стану країн за індексами сталого розвитку шляхом створення інформаційної системи, що базується на основі методів та моделей інтелектуального аналізу даних. Для дослідження були обрані наступні типи моделей: –  $k$ -means, agglomerative, fuzzy  $c$ -means, dbscan. Ці моделі були об'єднані в єдиний програмний продукт – інформаційну систему, яка за допомогою критеріїв оцінювання кластеризації (критерій компактності кластерів та Silhouette criterion) дала можливість визначити найкращі моделі.

## 2. МОДЕЛІ КЛАСТЕРИЗАЦІЇ І КРИТЕРІЇ ЯКОСТІ

За останні десятиліття було створено велику кількість моделей та методів для кластеризації, які застосовуються у багатьох сферах (медицина, бізнес, екологія, економіка тощо). Опишемо основні методи кластеризації, які були застосовані у даному дослідженні, а також методи для порівняння адекватності таких моделей.

### 2.1. Моделі інтелектуального аналізу для кластеризації даних

$K$ -means або метод  $k$ -середніх є одним із найбільш популярних моделей. Він працює наступним чином. Обирається початкова кількість кластерів  $k$ . Для цього застосовується ліктьовий (elbow) метод. Він полягає у тому, що будується певна кількість моделей (для різної

кількості кластерів) та обчислюється метрика WCSS (within cluster sum of squares) – сума квадратів відстаней від точок до центрів кластерів:

$$WCSS = \sum_{P_i \text{ in Cluster } 1} distance(P_i, C_1)^2 + \dots + \sum_{P_i \text{ in Cluster } N} distance(P_i, C_N)^2,$$

де  $C_i$  – відповідний центроїд і-того кластеру.

Зі збільшенням кількості кластерів дана метрика прямуватиме до 0, але починаючи з певної кількості кластерів зміна буде незначною. Відповідно, така кількість кластерів і буде оптимальною.

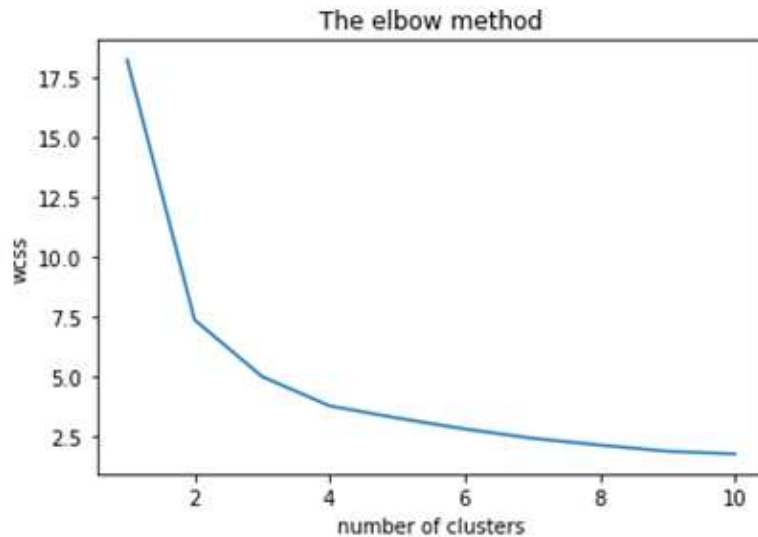


Рисунок 1. Приклад визначення кількості кластерів

Загалом, існують й інші методи визначення оптимальної кількості кластерів (наприклад, метод дендрограм), але в даному дослідженні застосовується саме ліктювий метод. Наступним етапом алгоритму є обчислення для кожної точки даних його евклідова відстань до кожного центру кластера. У якості відстані береться середньоквадратична норма  $l_2$ , тобто цільовою функцією вважається:

$$S = \sum_{j=1}^k \sum \{|x_i - \mu_j|^2 | x_i \in c_j\},$$

де  $x_i$  – і-тий об'єкт,  $c_j$  – j-тий кластер з центром  $\mu_j$  [1].

Алгоритм обчислює центроїди (centroids) - центри мас кластерів. Кожен центроїд - це вектор, елементи якого являють собою середні значення характеристик, обчислені по всіх точках кластера. Центр кластера зміщується в його центр ваги. Точки заново призначаються найближчого центру кластера. Етапи зміни центрів кластерів і перепризначення точок ітераційно повторюються до тих пір, поки кордони кластерів і розташування центроїдів не перестануть змінюватися, тобто, на кожній ітерації в кожен кластер будуть потрапляти одні і ті ж точки даних.

Agglomerative clustering відноситься до сімейства алгоритмів кластеризації, в основі яких лежать однакові принципи: алгоритм починає свою роботу з того, що кожен точку даних заносить в свій власний кластер і по мірі виконання об'єднує два найбільш схожих між собою кластера до тих пір, поки не буде задоволено певний критерій зупинки. В основі цих критеріїв лежить відстань між двома існуючими кластерами.

$$\tilde{d}(c_i, c_j) = \frac{1}{n_i n_j} \sum_{x \in c_i} \sum_{y \in c_j} \|x - y\|,$$

де  $n_i, n_j$  – кількість об'єктів відповідно у і-тому та j-тому кластері.

Нечіткі методи (fuzzy c-means) відрізняються тим, що в результаті роботи алгоритму ми отримуємо розподілення точок на кластери з деяким значенням рівня належності:  $w_{k,j} \in [0,1]$  k-го вектору ознак до j-го кластеру. При цьому виконується розрахунок  $N \times m$  матриці  $W = \{w_{k,j}\}$ , яка називається матрицею нечіткого розбиття. Алгоритми фаззи-кластеризації, що засновані на цільових функціях, призначені для вирішення задач шляхом оптимізації деякого наперед заданого критерію якості кластеризації і являються найбільш строгими з математичної точки зору. Цільова функція, що мінімізується, має наступний вигляд:

$$E(w_{k,j}, c_j) = \sum_{k=1}^N \sum_{j=1}^m w_{k,j}^{\beta} d^2(x_k, c_j)$$

при обмеженнях:

$$\sum_{j=1}^m w_{k,j} = 1, \quad k = 1, \dots, N$$

$$0 < \sum_{k=1}^N w_{k,j} < N, \quad j = 1, \dots, m,$$

де  $c_j$  – прототип (центр) j-го кластера,  $\beta$  – невід’ємний параметр «фаззифікатор» (зазвичай,  $\beta = 2$ ),  $d^2(x_k, c_j)$  – відстань між  $x_k$  і  $c_j$  [3].

## 2.2. Критерії оцінки якості кластеризації

Так як ми достеменно не знаємо, яке розбиття даних є правильним, досить важко оцінити адекватність побудованих моделей кластеризації. На сьогоднішній день таких методик існує не так багато, проте є деякі метрики, що допомагають вирішити це питання. В рамках дослідження було застосовано два підходи – компактність кластерів та критерій силуету.

Ідея компактності кластерів полягає в тому, що чим ближче один до одного знаходяться об’єкти всередині кластерів, тим краще поділ. Таким чином, необхідно мінімізувати суму квадратів відхилень:

$$WSS = \sum_{j=1}^M \sum_{i=1}^{|c_j|} (x_{ij} - \bar{c}_j)^2,$$

де  $M$  – кількість кластерів [4].

Критерій силуету в свою чергу вказує на те, наскільки об’єкти схожі на свій кластер в порівнянні з іншими кластерами. Критерій силуету визначається наступним чином:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Інакше можна записати:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases},$$

де  $a(i)$  – середня відстань між точкою і та іншими точками у тому ж кластері,  $b(i)$  – середня відстань і до всіх точок в будь-якому іншому кластері (не в тому, де знаходиться точка і).

## 3. РЕЗУЛЬТАТИ МОДЕЛЮВАННЯ

Масив даних для дослідження складається з таких ознак: індикатор GINI, індекс стану здоров’я населення, показник рівня життя, індекс сталого розвитку. Використовуючи описані вище алгоритми, була розроблена інформаційна система для визначення оптимальної кількості кластерів, побудови моделей, обчислення критеріїв адекватності. Далі розглянемо отримані результати моделювання.

Таблиця 1. Порівняння реалізованих алгоритмів кластеризації

Алгоритм	WSS	Silhouette criterion
KMeans (K=3)	4.969	0.419
KMeans (K=4)	3.742	0.431
Agglomerative clustering (K=3)	5.032	0.422
Agglomerative clustering (K=4)	4.101	0.402
Нечіткий метод (K=3)	5.013	0.419
Нечіткий метод (K=4)	3.867	0.425

Отже, можемо бачити, що всі алгоритми з приблизно однаковою якістю розбили вхідні дані на кластери. Серед найкращих по метрикам можемо виділити метод K-середніх та нечіткий метод Бездека при кількості кластерів K=4. Далі опишемо отримані кластери та визначимо, які країни потрапили до відповідних кластерів.

Таблиця 2. Результат кластеризації методом k-середніх (K=4)

Номер кластеру	Країни
1	Australia, Austria, Belgium, Canada, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea Republic, Latvia, Lithuania, Luxembourg, Niderlandi, New Zeland, Norway, Poland, Portugal, Rumuniya, Slovachchina, Slovenia, Spain, Sweden, Switzerland, Great Britain, USA, Uruguay
2	Albania, Algeria, Armenia, Azerbaijan, Bosnia and Herzegovina, China, Egypt, Georgia, Indonesia, Yordaniya, Kazakhstan, Kyrgyzstan, Moldova..Republic, Mongolia, Morocco, Filippini, Russian Federation, Sri Lanka, Trinidad, Tunisia, Turkey, Ukraine, United by Arabskiye Emirati, Venezuela
3	Bangladesh, Benin, Cambodia, Cameroon, Ethiopia, Gambia, India, Kenya, Madagascar, Malawi, Mozambique, Nepal, Nicaragua, Niger, Pakistan, Senegal, Tajikistan, Tanzania., Uganda, Uzbekistan, Vjetnam, Zambia, Zimbabwe
4	Argentina, Bolivia, Botswana, Brazil, Bulgaria, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Jamaica, Malaysia, Mexico, Namibia, Panama, Paraguay, Peru, South Africa, Thailand

При розбитті на 4 кластери отримано такі результати. До першого кластеру потрапили найбільш розвинені країни з найвищими показниками (Західна Європа, Америка, Північна Європа). До другого кластеру потрапили країни з середніми значеннями усіх показників. Переважно сюди потрапили країни, які й раніше були у другому кластері (Albania, Algeria, Armenia, Turkey, Ukraine,..) – Східна Європа, країни Балканського півострову тощо. До третього кластеру віднесли країни з найнижчими показниками та середнім показником Джині. Це такі країни як Bangladesh, Benin, Cambodia, Cameroon, Ethiopia, Gambia, Honduras, India. До четвертого кластеру належать країни з найнижчим показником Gini та середніми іншими індексами. В основному, це країни Латинської Америки (Argentina, Bolivia, Botswana, Brazil, Bulgaria, Chile, Colombia, Costa Rica, Dominican Republic, ...).

## 4. ВИСНОВКИ

Отже, у ході виконання дослідження розроблена інформаційна система підтримки прийняття рішень, яка базується на основі методів інтелектуального аналізу даних та дає можливість виконувати кластеризацію країн світу за показниками сталого розвитку. Запропонована система виконує попередню обробку даних, побудову математичних моделей, перевірку якості (адекватності) моделей за спеціальними статистичними критеріями. Система розроблена мовою програмування Python, використовуючи бібліотеку scikit-learn. В ході роботи виконано експериментальне дослідження на основі даних за показниками сталого розвитку вибраних країн світу. В результаті було визначено, що серед представлених та апробованих моделей найкращими для застосування до заданого набору даних є модель  $k$ -середніх з кількістю кластерів  $k = 4$ . Для подальшого покращення результатів планується використання більш складних методів нечіткої кластеризації (Густафона-Кесселя, Гета-Леві), а також розширення масиву вхідних даних (додавання більшої кількості економічних та соціальних показників сталого розвитку).

### ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Шумейко А.А., Сотник С.Л. Інтелектуальний аналіз даних. Введення в data mining. Дніпро: Біла Е.А., 2012. 212 с.
2. Sustainable Development Report 2021. URL: <https://dashboards.sdgindex.org/rankings> (дата звернення: 20.11.2023).
3. Зайченко Ю.П., Гончар М.А. Нечіткі методи кластерного аналізу в задачах автоматичної класифікації в економіці. – Київ: Вісник НТУУ «КПІ», 2007. – с.197-204.
4. Алгоритми кластеризації на службі Data Mining. URL: <https://loginom.ru/blog/data-mining-clustering> (дата звернення: 25.11.2023).

# **ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ДОСЛІДЖЕННЯ АКТУАРНИХ ФІНАНСОВИХ РИЗИКІВ**

Чеманова А.О.<sup>1</sup>, Кузнецова Н.В.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

<sup>1</sup> ankachemanova@gmail.com [0009-0008-6647-9364], <sup>2</sup> natalia-kpi@ukr.net [0000-0002-1662-1974]

Дослідження актуарних (страхових) ризиків є однією з ключових задач у галузі страхування. Дослідження актуарних процесів є доволі складним процесом, оскільки потребує врахування різних показників і параметрів, використання різних підходів та методик оцінювання, а також достатньої статистичної бази. Метою нашого дослідження є розробка інтелектуальної системи підтримки прийняття рішень (ІСППР), яка буде надавати необхідну користувачу інформацію щодо конкретного страхового випадку з точки зору доцільності надання полісу. Для реалізації даної системи було обрано модульну структуру, а також три різні за принципом та структурою – інтелектуальної системи підтримки прийняття рішень. При цьому основною особливістю розробленої ІСППР є надання користувачеві конкретних рішень та порад щодо видачі полісу страхування.

**Ключові слова:** актуарні фінансові ризики, інтелектуальна система підтримки прийняття рішень, лінійна регресія, випадковий ліс, метод екстремального градієнтного підсилення, скорингова карта.

## **1. ВСТУП**

У сучасному світі страхування широко представлене в багатьох сферах людського життя. Процес страхування невід’ємно пов’язаний з ризиком та включає в себе два протилежних за моделями поведінки учасники: страхувальника та застрахованої особи. При чому страхувальник прагне до страхування виключно тих випадків, де не очікуються збитки, а людина, що має поліс, не може спрогнозувати величину збитку або ж кількість таких випадків і час їх настання [1]. Це робить дослідження актуарних процесів складним – необхідно врахувати фактори та виконувати оцінку збиткам, імовірностям їх настання [2]. Окремою проблемою є проблема інтерпретації результатів. Особа, що приймає рішення (ОПР), використовує результати та оцінки, однак разом з тим повинна враховувати лише їх наближену точність, при цьому спосіб отримання та фактори, що впливають, залишаються ключовими, не зважаючи на лише наближену величину їх оцінки. Популярним способом розв’язання поставлених задач є використання інтелектуальної системи підтримки прийняття рішень [3]. Вона дозволяє отримати можливості для швидшої обробки даних, варіативності та легкої взаємодії з користувачем.

## **2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ**

Задачею даної роботи є побудова інтелектуальної системи підтримки прийняття рішень з набором необхідних модулів для виконання обробки, аналізу та перевірки даних. Побудована

система повинна надавати необхідну користувачу інформацію та мати можливість бути розширеною при необхідності додатковими модулями, тобто мати модульну структуру.

Метою роботи системи є надання користувачеві рішень та рекомендацій на основі прогнозів полісів. Це дозволить вирішити проблему автоматизованої видачі полісів та надасть ОПР готові результати.

В системі мають бути передбачені наступні модулі – модуль введення (завантаження) інформації, модуль виведення, модуль попередньої обробки даних, модуль побудови моделі та модуль побудови скорингової карти.

### 3. ОСНОВНІ МЕТОДИ

Для врахування можливих варіантів зв'язків між змінними, в тому числі і нелінійних, було вирішено побудувати декілька класів обрати моделей, що за своїм принципом і структурою повинні відрізнятися один від одної та належати до різних сімейств моделей. Було обрано наступні моделі: лінійна регресія, випадковий ліс та метод екстремального градієнтного підсилення.

Лінійна регресія – це метод, що базується на ідентифікації незалежних змінних, від яких залежить шукана цільова змінна, яка в свою чергу вважається залежною. Ключовим аспектом моделі є дві складові: визначення незалежних змінних, що будуть використанні для моделювання, та способу оптимізації параметрів моделі з метою найбільш оптимального, з точки зору критерію, підбору параметрів. Для розв'язання першої проблеми можливі два підходи, що будуть використовуватися в системі. Перший підхід – це пошук найбільш корельованих із залежною змінною незалежних змінних шляхом побудови матриці кореляцій. Слід також зазначити, що деякі змінні з набору даних можуть мати великий коефіцієнт кореляції з цільовою змінною, а також один з одним. Такі змінні було прийнято рішення видаляти. Другий підхід – це підхід експертний з метою надання особі, що відповідає за прийняття рішення, можливості вибрати деякі змінні, що точно будуть використовуватися для побудови моделі [4].

Метод випадкового лісу – це метод машинного навчання, що представляє собою комбінацію дерев рішень, що були побудовані на окремих частинах набору даних, та за допомогою яких приймається єдине рішення шляхом врахування та подальшого зваження рішення з кожного дерева. Цей підхід набирає популярність та застосовується у фінансовій сфері та не тільки, оскільки дозволяє моделювати нелінійні зв'язки між змінними [5].

Метод екстремального градієнтного підсилення – це ще один метод машинного навчання, що також використовує дерева, проте на відміну від методу випадкового лісу, він базується на поступовому додаванні нової складової до моделі – слабкої моделі. На кожному кроці навчання основна модель отримує додаткове дерево, що позитивно впливає на точність моделі. Існує цілий клас методів підсилення, зокрема метод екстремального градієнтного підсилення є одним з найбільш популярних [6, 7].

Окремим модулем є побудова скорингової карти. Скорингова карта є набором значень з кожної змінної-ознаки (для категорійних змінних – це окрема категорія, для неперервних – це інтервал) з певним числовим значенням. Кожне спостереження відповідно до карти отримує загальний бал або оцінку. Цей підхід є основним під час видання кредитів та є ключовим інструментом, яким користуються під час оцінки та класифікації клієнта. Це ефективний та легко зрозумілий метод, результати використовуються прямо на місці під час взаємодії з клієнтом [8].

Термінологія скорингових карт пов'язана з двома поняттями: WoE та IV.

WoE (1Weight of Evidence або вага (величина) спостереження (доказу)) – це коефіцієнт наявності залежності між певною групою в середині даних та цільовою змінною. Він розраховується за формулою (1).

$$WoE = \ln\left(\frac{Bad\ Results}{Good\ Results}\right), \quad (1)$$

де Bad Results – це кількість подій (позовів, скарг) у групі, тобто записів, в яких наявний позов або певна його критична величина, а Good Results – це кількість записів, по яким не було події, тобто поліси без скарг або такі, по яким не було збитків.

IV (Information Value або Величина інформації) – це коефіцієнт величини зв'язку між незалежною та цільовою змінною, що розраховується за формулою (2).

$$IV = \sum(\% \text{ of Good Results} - \% \text{ of Bad Results}) * WoE. \quad (2)$$

Ці величини будуть використанні для визначення змінних та категорій, що не несуть корисної інформації для побудови карти та можуть бути відкинуті.

#### 4. РЕЗУЛЬТАТИ ДСЛІДЖЕННЯ

З метою перевірки роботи побудованої системи підтримки прийняття рішення було обрано тестовий набір даних, запропонований на Kaggle. [9] Це набір даних звернень від осіб, що мали страховий поліс та страховий випадок, пов'язаний з травмуванням. Слід зазначити, що подібні поліси представляють собою окрему категорію, оскільки для них не виконується припущення про незалежність кількості випадків та їх величини збитку.

Набір даних включає в себе 54 000 записів для побудови моделі та 36 000 записів для перевірки. Цільовою змінною є кінцева величина затрат по полісу.

На Рис. 1 побудований розподіл категоріальних змінних, змінної заробітку та змінних витрат по полісам, а на Рис. 2 – числовий розподіл цільової та фінансових змінних. Попередній огляд цих даних вказує на певну величину дисбалансу класів категорійних змінних та центрованості значень.

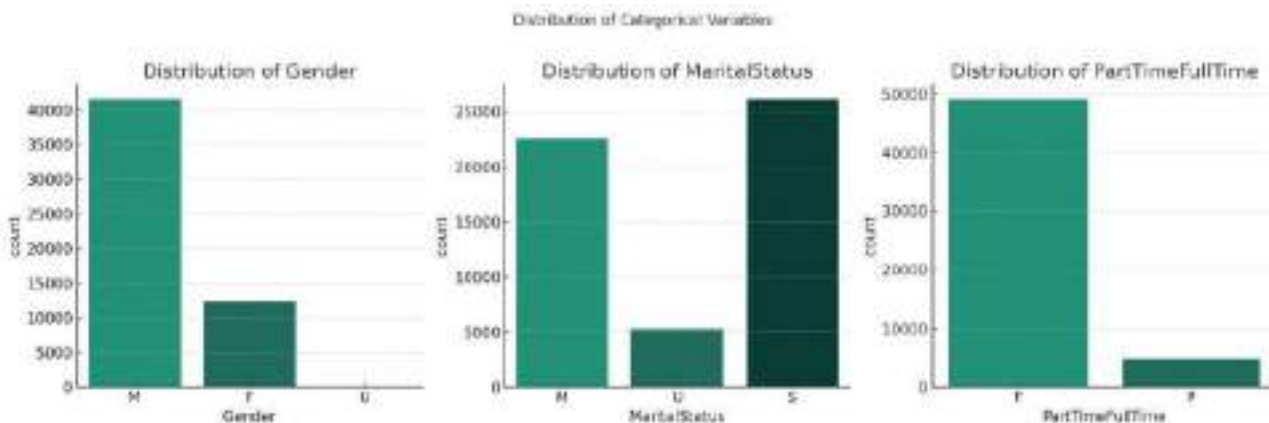


Рисунок 1. Розподіл категоріальних змінних, змінної заробітку та змінних витрат за полісами

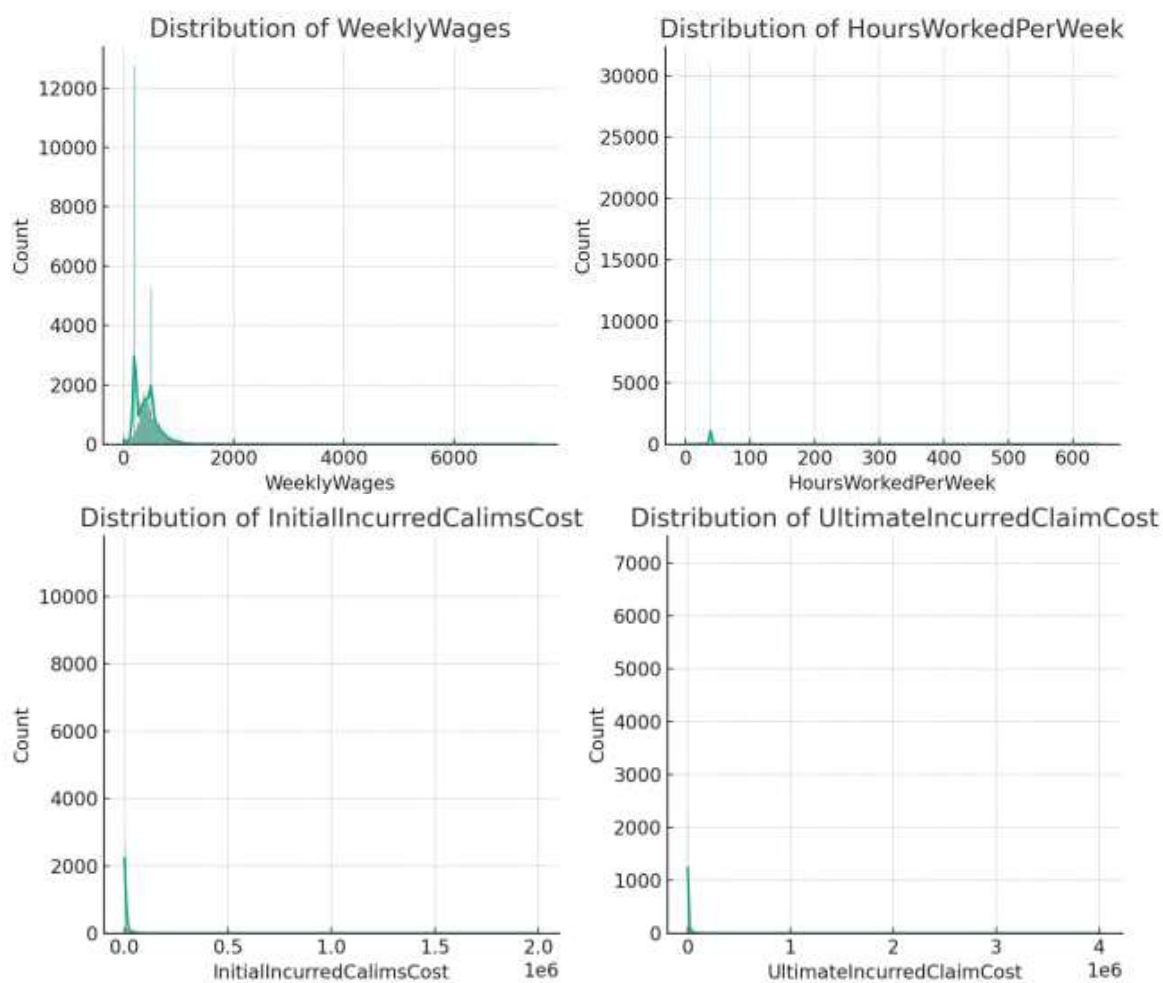


Рисунок 2. Числовий розподіл цільової та фінансових змінних

Проблема наявності незбалансованих класів є актуальною для проведення аналізу будь-яких вибірок. Значна диспропорція у класах, як, наприклад, видно для змінної *PartTimeFullTime*, може призвести до невірної оцінки та інтерпретації параметрів. Важливо використовувати моделі, що нечутливі до диспропорцій та будуть враховувати малозначущі класи.

Матриця кореляцій (Рис. 3) вказує про наявність набору ознак, що має відносно великий коефіцієнт кореляції.

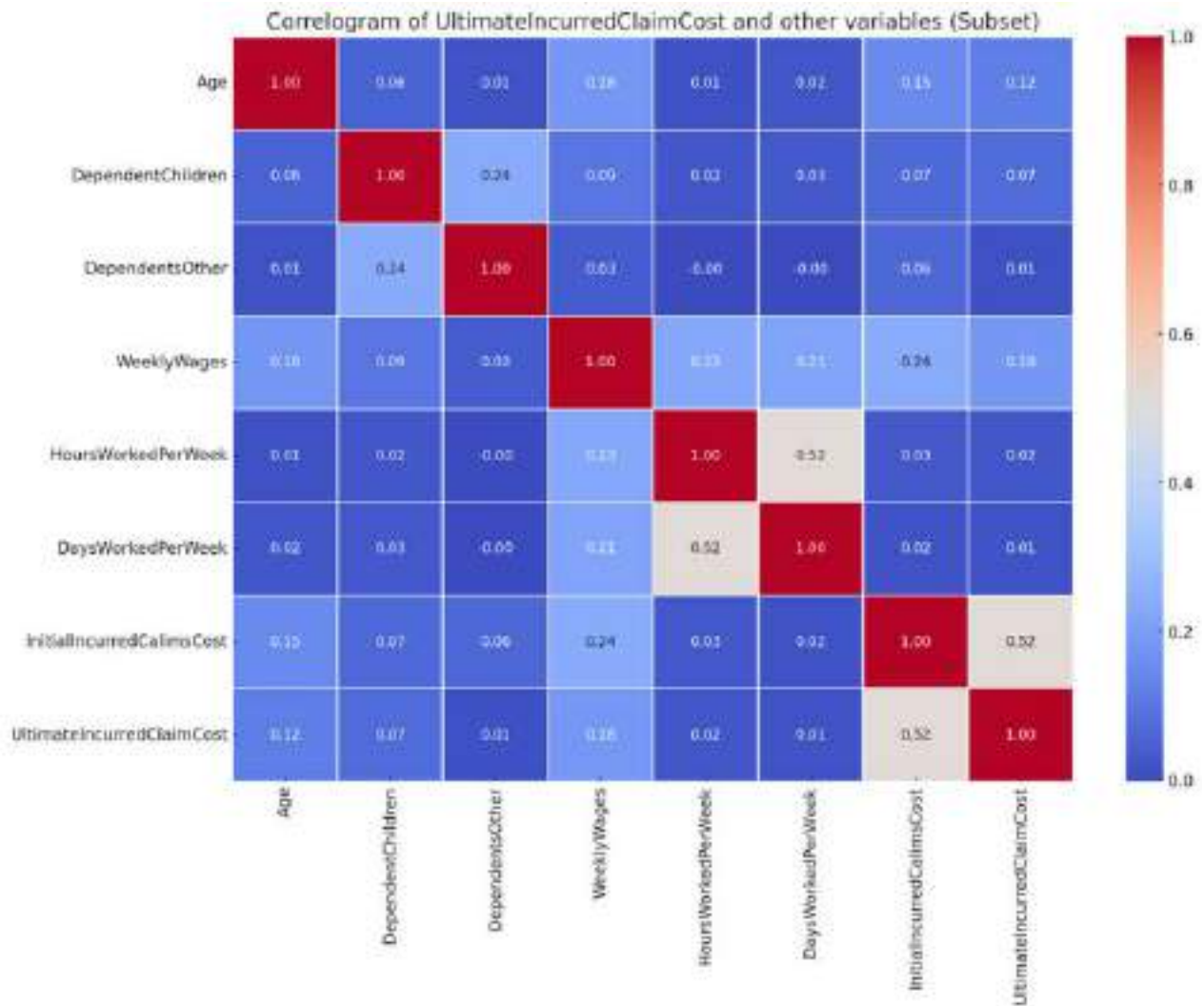


Рисунок 3. Матриця кореляцій

Для подальшої роботи системи було виконано синтез додаткової ознаки з метою покращення результатів моделі: кількість днів між подією та зверненням за полісом, а також було розраховано зміну відхилення початкової оцінки звернення та кінцевої величини збитків. Це дозволить визначити та розрахувати величину ризику, яка буде величиною помилки.

Результати роботи моделей представлені у Табл. 1.

Таблиця 1. Оцінки прогнозів побудованих моделей

Модель	MSE	RMSE	MAE	R2
Linear Regression	712949507,6221465	26701,11435169226	8668,957348620825	0,2146899381959123
Random Forest	32846600,81837761	5731,195409194979	361,3268974754968	0,9638196452301848
XGBoost	132286446,40038218	11501,584516942967	1006,4390728725784	0,8542871882400046

Найкращою моделлю виявилася модель випадкового лісу.

Результати роботи модулю з розрахунку скорингових оцінок представлені на Рис. 4–5.

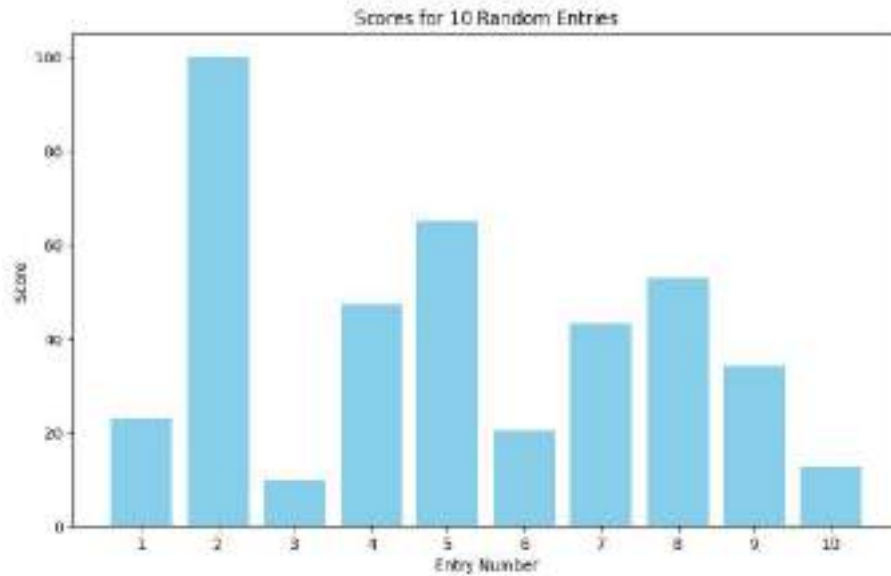


Рисунок 4. Розрахунок значення оцінки для 10 полісів

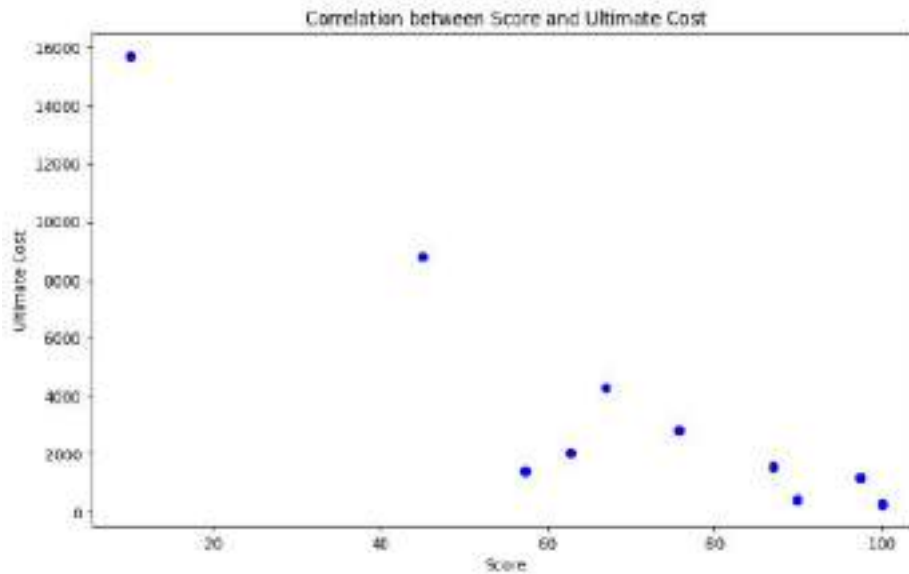


Рисунок 5. Зв'язок між величиною витрат та оцінкою (скором)

Результати показують ефективність скорингової моделі, проте разом з тим наявні і випадки розходження результатів. Позитивним результатом застосування моделі є те, що найменшу кількість балів отримують поліси з найбільшою величиною потенційного збитку. Це дозволяє відразу отримати уявлення про потенційну величину витрат, орієнтуючись на отриманий результат скорингової оцінки.

## 5. ВИСНОВКИ

Актуарні операції становлять значну частку всіх фінансових операцій. Для надання можливості користувачам проводити швидкий аналіз даних та функціоналу для розрахунку ключових показників, було побудовано інтелектуальну систему підтримки прийняття рішень.

Запропонована ІСППР надає можливість проводити розрахунки найбільш ключових параметрів та дозволяє експертну оцінку з метою врахування параметрів, які хоч і мають малу величину кореляції, проте мають відомий, потенційно, нелінійний зв'язок з цільовою змінною.

Використані та застосовані моделі дозволяють підходити до задачі прогнозування більш комплексно, враховувати можливі нелінійні зв'язки та шукати найкращу модель для прогнозування. Отримані результати для моделі випадкового лісу говорять про можливість застосування розробленого модуля на реальних даних.

Окремим модулем є модуль побудови скорингової карти. Сам концепт розроблявся з метою покращення розуміння результатів моделі та надання працівникам на місцях швидкого способу інтерпретації стану клієнта та потенційної величини збитків.

Все разом, реалізує прикладну систему, що виконує аналіз даних та надає всю необхідну для прийняття рішень інформацію.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Н. М. Внукова, В. І. Успенко, та Л. В. Временко, *Страховання: теорія та практика : навч. посіб. / за заг. ред. проф. Н. М. Внукової*. Харків, Україна: Бурун Книга, 2004.
2. Н.В Кузнєцова, *Теорія і практика аналізу фінансових ризиків: системний підхід: монографія / Н. В. Кузнєцова, П. І. Бідюк*, Київ, Україна: Ліра-К, 2020.
3. О. В. Нестеренко, О. І. Савенков, та О. О. Фаловський, *Інтелектуальні системи прийняття рішень: Навч. посібник*. Київ, Україна: Національна академія управління, 2016.
4. В. В. Головачко, та В. М. Головачко, "Поняття лінійної регресії", на *III Міжнародній науково-практичній конференції Сучасні тенденції розвитку науки й освіти в умовах поглиблення євроінтеграційних процесів*, Мукачево, 2023, с. 348-350.
5. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees", *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 5, No 3, 2012.
6. T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, ACM Press, 2016, pp. 785-794. doi: 10.1145/2939672.2939785.pdf.
7. S. F. N. Islam, A. Sholahuddin, and A. S. Abdullah, "Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah", *Journal of Physics: Conference Series* 1722 (2021) 012016, 2021, pp. 1-11. doi:10.1088/1742-6596/1722/1/012016.
8. N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, New Jersey, Canada: John Wiley & Sons, Inc., Hoboken, 2006.
9. Actuarial loss prediction. [Електронний ресурс]. Доступно: <https://www.kaggle.com/competitions/actuarial-loss-estimation/overview>. Дата звернення: листопад 2023.

# МОДЕЛЮВАННЯ ВПЛИВУ ЧАТ-БОТІВ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ НА ЯКІСТЬ ВИЩОЇ ОСВІТИ МЕТОДАМИ СИСТЕМНОГО АНАЛІЗУ

Чернюк О.Р.<sup>1</sup>, Тимошук О.Л.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

<sup>1</sup> cherniuk.oleksii@lll.kpi.ua, <sup>2</sup> o.tymoshchuk@kpi.ua [0000-0003-1863-3095]

**У дослідженні запропоновано ефективні стратегії для покращення якості вищої освіти шляхом мінімізації негативних наслідків, пов'язаних із зловживанням чат-ботами студентами, і максимізації навчальних, практичних і наукових переваг, які можна отримати від взаємодії учасників навчального процесу зі штучним інтелектом чат-ботів. У дослідженні використано двоетапний метод модифікованого морфологічного аналізу та метод когнітивного моделювання. В результаті було розроблено дві взаємопов'язані моделі, які виявили чисельну ієрархію ефективності освітніх втручань для покращення якості вищої освіти. Ця ієрархія дій може служити цінним інструментом для педагогів і бути рекомендованою для впровадження в систему вищої освіти.**

**Ключові слова:** якість вищої освіти, морфологічний аналіз, когнітивна карта, когнітивне моделювання, прийняття рішень, системний аналіз.

## 1. ВСТУП

Вища освіта є критично важливим компонентом суспільного розвитку, що надає людям знання, навички та кваліфікацію, необхідні для особистого та професійного зростання. Одним із визначних аспектів галузі вищої освіти, яка розвивається, є інтеграція технологій в освітні процеси.

Чат-боти, що базуються на технологіях штучного інтелекту (ШІ) і обробки природної мови (NLP), стали цінними інструментами в різних сферах, включаючи вищу освіту. Чат-боти – це комп'ютерні програми, призначені для імітації людської розмови та надання автоматичних відповідей на запити користувачів. Значний технологічний прорив здійснили генеративні чат-боти на основі штучного інтелекту, такі як Chat GPT (OpenAI), Bard (Google), Bing Chat (Microsoft), Perplexity AI, YouChat, Chatsonic, Aria та багато інших. Хоча основною функцією генеративних чат-ботів є імітація співрозмовника-людини, вони можуть виконувати дуже багато завдань. Наприклад, писати і налагоджувати код програм; відповідати на тестові питання з поясненнями; генерувати бізнес-ідеї; писати вірші, твори, статті, тексти пісень; перекладати, переписувати та резюмувати текст; емулювати систему Linux; моделювати цілі чати; грати в такі ігри, як «хрестики-нулики»; змоделювати банкомат; надавати психологічні консультації; розпізнавати зображення; вирішувати математичні задачі та багато іншого [1–5].

Зв'язок чат-ботів на основі штучного інтелекту з вищою освітою є неминучим, тому існує потреба критично вивчити їхній вплив на якість освіти. Це дослідження прагне надати інформацію та рекомендації для навчальних закладів, які, хочуть цього чи ні, змушені

взаємодіяти з новою реальністю, коли студент стає одним цілим з чат-ботом, і стає незрозуміло, де справжні знання студента, і в чому вони полягають.

## 2. ШКАЛА ВИМІРЮВАННЯ ВЗАЄМОВПЛИВУ АЛЬТЕРНАТИВ

Усі впливи вершин одна на одну відбуваються на інтервалі  $[-1; +1]$ . Вагу кожного впливу будемо розуміти так (Табл. 1):

Таблиця 1. Міри впливу

Дуже сильно негативний вплив	$[-1; -0,8)$
Сильно негативний вплив	$[-0,8; -0,6)$
Досить негативний вплив	$[-0,6; -0,4)$
Помірно негативний вплив	$[-0,4; -0,2)$
Легко негативний вплив	$[-0,2; -0)$
Прямого впливу немає	$\{0\}$
Легко позитивний вплив	$(0; +0,2]$
Помірно позитивний вплив	$(+0,2; +0,4]$
Досить позитивний вплив	$(+0,4; +0,6]$
Сильно позитивний вплив	$(+0,6; +0,8]$
Дуже сильно позитивний вплив	$(+0,8; +1]$

## 3. МОДЕЛЬ НА ОСНОВІ МЕТОДУ ДВОЕТАПНОГО МОДИФІКОВАНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

### 3.1. Характеристика вхідних даних

Для дослідження нашої тематики застосуємо метод двоетапного модифікованого морфологічного аналізу (МММА) [6–9, 17]. Експериментальна модель об'єкта буде заснована на морфологічних таблицях з двох характеристичних параметрів, кожен з яких має свої альтернативні сценарії.

Усі параметри морфологічної таблиці є якісними за своєю природою, тобто альтернативи таких параметрів принципово (якісно) відрізняються між собою, і між такими альтернативами неможливо встановити відношення переваги, як для кількісних параметрів.

Усі параметри є релевантними, тобто параметр повинен залежати або впливати на хоча б один інший параметр. В рамках деталізації, яка обрана для задачі. з параметрів, на основі яких визначається вплив на якість вищої освіти, було вилучено усі нерелевантні альтернативи.

У державних та університетських документах, підручниках та інших джерелах [10–12] можна знайти декілька десятків параметрів для оцінки якості вищої освіти, але в контексті нашої задачі було підібрано тринадцять релевантних показників ефективності студентської навчальної діяльності (Табл. 2).

Таблиця 2. Морфологічна таблиця першого етапу МММА (морфологічна таблиця сценаріїв)

Вплив чат-ботів на основі штучного інтелекту на якість вищої освіти	
Мета використання чат-ботів студентами	Студентські показники ефективності навчальної діяльності
1.1. Списування і халатність у навчанні	2.1. Оцінки студентів
1.2. Саморозвиток, навчальна комунікація, удосконалення навичок	2.2. Розуміння навчального матеріалу
	2.3. Базові практичні навички, здобуті під час навчання
	2.4. Навички критичного, творчого та незалежного мислення
	2.5. Запам'ятовування інформації
	2.6. Швидкість виконання навчальних завдань
	2.7. Дослідницька майстерність
	2.8. Відвідуваність занять та активність на них
	2.9. Здатність ефективно передавати думки усно і письмово
	2.10. Співпраця та командна робота
	2.11. Адаптивність та стійкість до нових ситуацій та викликів
	2.12. Організація часу (тайм-менеджмент)
	2.13. Технологічна компетентність

На основі джерел про генеративні чат-боти на основі ШІ [3–5] та експертного актуального навчального студентського досвіду було відібрано дев'ять способів контролю якості вищої освіти (Табл. 3) в контексті використання студентами чат-ботів. Альтернативи параметрів є взаємовиключними, що було враховано при проведенні процедури експертного оцінювання початкових наближень параметрів при експертному оцінюванні попарними порівняннями. Множина альтернатив є повною.

Таблиця 3. Морфологічна таблиця другого етапу МММА (морфологічна таблиця стратегій)

Способи контролю якості вищої освіти в контексті використання студентами чат-ботів
3.1 Використання чат-ботів в навчальних програмах предметів
3.2 Заборона використання гаджетів під час очних контрольних робіт і суворий контроль
3.3 Усні опитування віч на віч
3.4 Оцінювання за творчими індивідуальними унікальними роботами
3.5 Використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт
3.6 Усунення гуманітарних або тестових завдань в якості способів оцінювання студентів
3.7 Вимога дотримання спеціального стилю оформлення робіт, який чат-боти не зможуть повторити
3.8 Написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя
3.9 Суворі часові обмеження для написання контрольних робіт

### 3.2. Перший етап МММА

Для початку оцінимо початкові ймовірності альтернатив. Головна мета першого етапу морфологічного аналізу - отримати початкові наближення  $p_j^{(i)}$  для ймовірностей кожної з альтернатив  $a_j^{(i)}$  характеристичних параметрів. Для альтернатив  $a_j^{(i)}, j \in \overline{1, n}$  параметра «Мета використання чат-ботів студентами» експертами надається оцінка  $\widetilde{p}_j^{(i)}$ . Для альтернативи «1.1» обрали значення 0,8, знаючи що в основному студенти використовують чат-боти для списування і халтури, для «2» відповідно значення 0,2. Для альтернатив параметра «Студентські показники ефективності навчальної діяльності» використали рівномірний розподіл, тобто однакову ймовірність для кожної його альтернативи. Оскільки ми не можемо апріорно отримати адекватні оцінки ймовірностей, використання експертної процедури для цього не є раціональним через значну невизначеність оцінок і через їх близькість. У такому випадку результат роботи МММА над параметром «Студентські показники ефективності навчальної діяльності» буде базуватись виключно на використанні матриці взаємозв'язків альтернатив параметрів і стане одним із важливих результатів даного дослідження.

Наведемо оцінену уже нормовану морфологічну таблицю (табл. 4).

Таблиця 4. Початкові ймовірності альтернатив

Мета використання чат-ботів студентами		Студентські показники ефективності навчальної діяльності	
Номер альтернативи	Ймовірність альтернативи	Номер альтернативи	Ймовірність альтернативи
1.1	0,8	2.1	0,076
1.2	0,2	2.2	0,077
		2.3	0,077
		2.4	0,077
		2.5	0,077
		2.6	0,077
		2.7	0,077
		2.8	0,077
		2.9	0,077
		2.10	0,077
		2.11	0,077
		2.12	0,077
		2.13	0,077

Далі оцінимо матрицю взаємозв'язків параметрів першого етапу.

Для врахування зв'язків між параметрами морфологічної таблиці (МТ) на основі таблиці 1 побудуємо числову матрицю взаємозв'язків альтернатив параметрів (Табл. 5).

Таблиця 5. Матриця взаємозв'язків параметрів першого етапу

	1.1	1.2
2.1	0,4	0,5
2.2	-0,8	0,3
2.3	-0,5	0,15
2.4	-0,3	0,15
2.5	-0,7	-0,2
2.6	0,8	0,2
2.7	-0,7	-0,2
2.8	-0,4	-0,2
2.9	-0,25	0,05
2.10	-0,35	-0,1
2.11	0,4	0,4
2.12	0,15	0,3
2.13	0,5	0,8

Нарешті, проведемо процедури з розрахунку ймовірностей альтернатив і конфігурацій. Щоб отримати остаточні значення ймовірності, необхідно розв'язати задачу розрахунку ймовірностей альтернатив параметрів. Подібні розрахунки повторюємо для всіх 26 конфігурацій, отримуємо результат першого етапу морфологічного аналізу – оцінки ймовірностей альтернатив з урахуванням зв'язків між ними (Табл. 6).

Таблиця 6. Результат першого етапу МММА

Мета використання чат-ботів студентами		Студентські показники ефективності навчальної діяльності	
Номер альтернативи	Ймовірність альтернативи	Номер альтернативи	Ймовірність альтернативи
1.1	0,74806734	2.1	0,11668415
1.2	0,25193266	2.2	0,03496632
		2.3	0,05244948
		2.4	0,06576998
		2.5	0,03330126
		2.6	0,13986528
		2.7	0,03330126
		2.8	0,05328201
		2.9	0,06743505
		2.10	0,0582772
		2.11	0,1165544
		2.12	0,09823871
		2.13	0,1298749

### 3.3. Другий етап МММА

Специфіка другого етапу МММА полягає в тому, що вибір альтернатив параметрів морфологічної таблиці стратегій залежить не від випадкових зовнішніх факторів, а від особи,

що приймає рішення. Тому на другому етапі для оцінки альтернатив і конфігурацій використовується величина очікуваної результативності, тобто вірогідності того, що вибір цієї альтернативи або конфігурації призведе до бажаних результатів.

Матриця зв'язків, яка співставляє кожен пару альтернатив першого і другого етапів наведена в таблиці 7.

Таблиця 7. Матриця зв'язків альтернатив першого та другого етапів

	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
<b>1.1</b>	0,5	1	1	0,65	0,9	0,6	0,15	1	0,4
<b>1.2</b>	0,5	0	0	0	0	0	0,1	0	0
<b>2.1</b>	0	-0,5	-0,4	0	-0,4	0	-0,1	-0,5	-0,5
<b>2.2</b>	0,3	0,5	0,4	0,6	0,3	-0,1	0	0,5	0,3
<b>2.3</b>	-0,25	0,7	0,3	0,5	0,3	-0,1	0	0,5	0,3
<b>2.4</b>	-0,15	0,2	0,3	0,6	0,1	0,2	0	0,2	-0,45
<b>2.5</b>	-0,35	0,5	0,3	0,15	0,3	0,2	0	0,4	0,2
<b>2.6</b>	0,7	-0,5	0	-0,25	0	-0,2	-0,2	-0,4	0,5
<b>2.7</b>	-0,1	-0,2	0	0,4	0	0,3	0	0	0
<b>2.8</b>	0,2	0,5	0,8	-0,2	0	0	0,1	0	0
<b>2.9</b>	-0,2	0	0,45	0,4	0	-0,25	0	0	0
<b>2.10</b>	0,2	0,3	0,15	0	-0,1	0	0,1	0	-0,4
<b>2.11</b>	0,6	0,35	0,4	0,1	0,1	0	0,1	0,25	0,45
<b>2.12</b>	0,55	0,3	0	0,25	0,3	0	0	0,2	0,7
<b>2.13</b>	0,85	-0,1	0	0,4	0,15	0	0,15	0,05	0

У результаті процедури з розрахунку оцінок альтернатив другого етапу отримуємо таблицю 8.

Таблиця 8. Розраховані на другому етапі МММА оцінки альтернатив

3.1. Використання чат-ботів в навчальних програмах предметів	0,13713944
3.3. Усні опитування віч на віч	0,129650906
3.4. Оцінювання за творчими індивідуальними унікальними роботами	0,116804045
3.2. Заборона використання гаджетів під час очних контрольних робіт і суворий контроль	0,116340931
3.5. Використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт	0,114872605
3.8. Написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя	0,112691262
3.6. Усунення гуманітарних або тестових завдань в якості способів оцінювання студентів	0,096571773
3.9. Суворі часові обмеження для написання контрольних робіт	0,096144714
3.7. Вимога дотримання спеціального стилю оформлення робіт, який чат-боти не зможуть повторити	0,079784324

### 3.4. Результати моделювання МММА

У результаті першого етапу МММА отримали початкові наближення для ймовірностей кожної з альтернатив «1.1-1.2», «2.1-2.13». З отриманих даних можна зробити висновки, що приблизно три використання студентами генеративного чат-бота із чотирьох відбувається для списування і халтури. Було отримано різноманітну ієрархію показників студентської ефективності навчальної діяльності саме в контексті використання чат-ботів студентами (Рис. 1). Пам'ятаємо що вибіркове середнє рівномірного розподілу появи будь-якого параметра альтернативи «Студентські показники ефективності навчальної діяльності» дорівнює 0,077. Згідно з таблицею 8 можна зробити висновок, що використання чат-ботів студентами позитивно вплинуло на швидкість виконання навчальних завдань, технологічну компетентність, оцінки студентів, адаптивність та стійкість до нових ситуацій, організацію часу. І негативно - на запам'ятовування інформації, дослідницьку майстерність, розуміння навчального матеріалу, базові практичні навички, відвідуваність занять та активність на них, співпрацю та командну роботу, навички критичного, творчого та незалежного мислення, здатність ефективно передавати думки усно і письмово. Порівнявши кожне числове значення ефективності з початковим вибіркочним середнім рівномірного розподілу, можна поррахувати приблизно, у скільки разів покращилася або погіршилася ситуація з відповідним показником студентської ефективності.

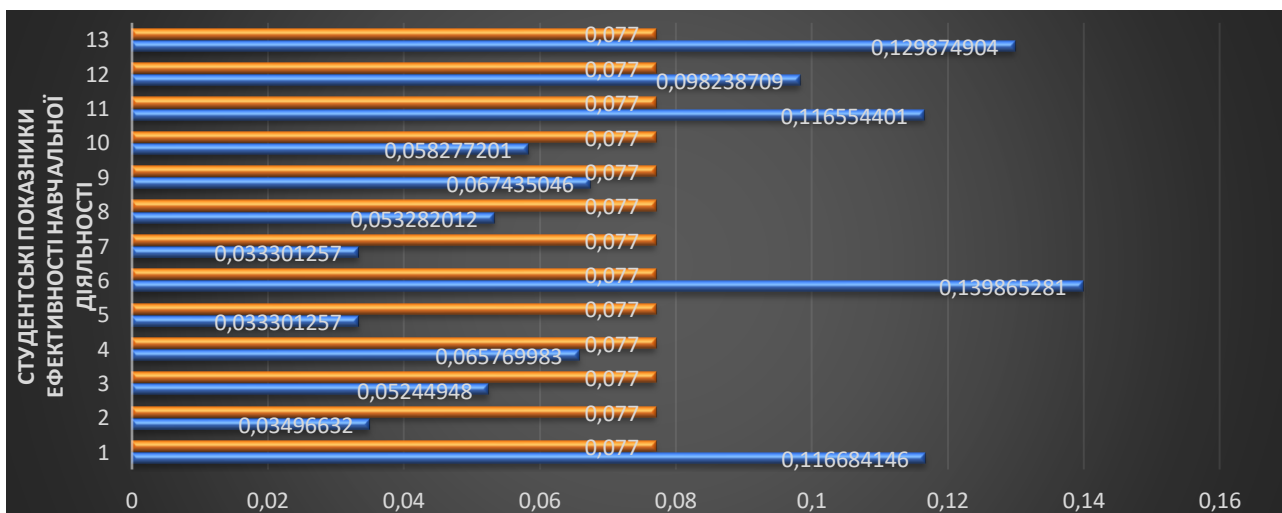


Рисунок 1. Порівняння ймовірностей альтернатив за рівнем позитивно-негативного впливу генеративних чат-ботів на студентські показники ефективності навчальної діяльності

У результаті другого етапу МММА отримали рейтинг способів контролю якості вищої освіти (Рис. 2).

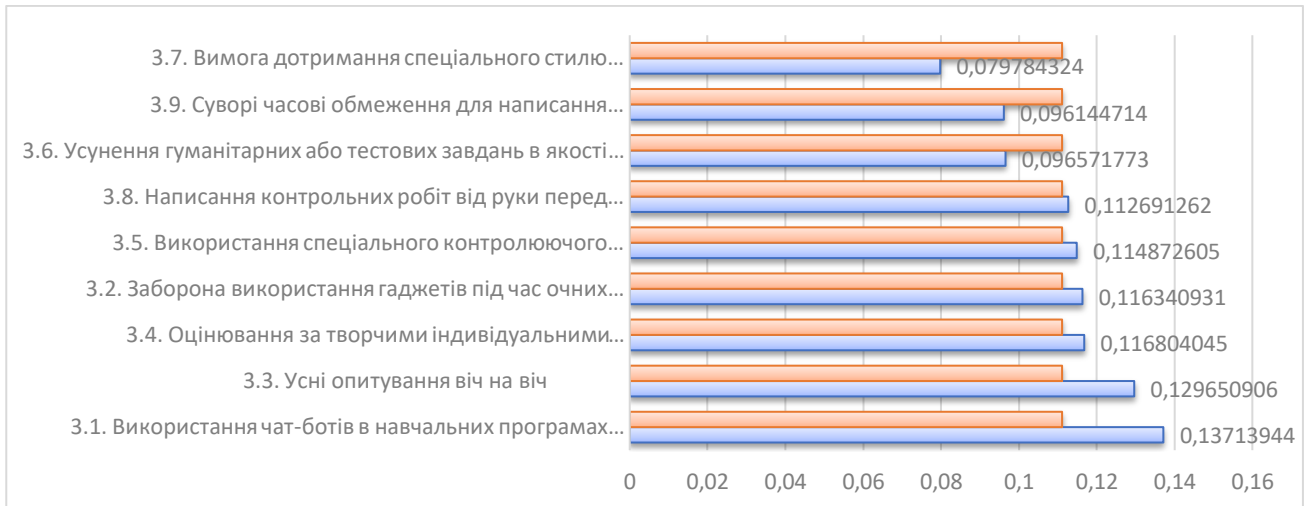


Рисунок 2. Порівняння ймовірностей результативності альтернатив для способів контролю якості вищої освіти

З діаграми можемо бачити, що майже усі величини очікуваної результативності близькі до вибіркового середнього 0,(11), тобто усі способи мають деякий позитивний вплив на якість вищої освіти в умовах використання чат-ботів студентами. Проте можемо виділити 6 найефективніших способів контролю якості вищої освіти в контексті використання студентами чат-ботів, а саме: використання чат-ботів в навчальних програмах предметів, усні опитування віч на віч, оцінювання за творчими індивідуальними унікальними роботами, заборона використання гаджетів під час очних контрольних робіт і суворий контроль, використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт, написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя. Тобто 4 із 6 найефективніших способів є способами-обмежувачами, які спрямовані на посиленій контроль студента. Спосіб використання чат-ботів в навчальних програмах предметів набрав високий рівень ефективності через те, що на відміну від інших способів він має значний вплив на альтернативу «1.2»: студенти активно використовують чат-боти на основі штучного інтелекту для саморозвитку, навчальної комунікації і удосконалення своїх навичок, тобто таке рішення не лише допомагає зменшити шкідливий вплив чат-ботів, а також надає можливість отримувати від них значну користь.

#### 4. КОГНІТИВНЕ МОДЕЛЮВАННЯ

В роботі також було застосоване когнітивне моделювання [13–17]. Вершини когнітивної карти (КК) повністю співпадають з альтернативами параметрів методу МММА з додаванням однієї цільової вершини «4. Якість вищої освіти», нумерація вершин також співпадає (Рис. 3).



проведено експериментальне тестування шляхом направлення на кожну керуючу вершину 3.1 – 3.9 одиничного додатного імпульсу (Рис. 5).



Рисунок 5. Ієрархія впливу способів контролю якості вищої освіти в контексті використання студентами чат-ботів на основі ШІ

Отже, створена когнітивна модель реагує на зовнішні зміни адекватно, адже при моделюванні вона показала цілком реалістичні закономірності. За результатами моделювання можемо виділити 5 найефективніших способів контролю якості вищої освіти, а саме: використання чат-ботів в навчальних програмах предметів, усні опитування віч-на-віч, заборона використання гаджетів під час очних контрольних робіт і суворий контроль, використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт, написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя.

## 5. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати роботи двох методів для порівняння і наочності були зведені в єдину таблицю (Табл. 9), в якій можемо бачити ієрархію способів максимізації позитивного впливу чат-ботів на основі ШІ на якість вищої освіти.

Таблиця 9. Порівняння результатів моделювання методами морфологічного аналізу та когнітивного моделювання

Ієрархія пріоритетності	Метод двоетапного модифікованого морфологічного аналізу	Метод когнітивного моделювання
1	Використання чат-ботів в навчальних програмах предметів	Використання чат-ботів в навчальних програмах предметів
2	Усні опитування віч на віч	Усні опитування віч на віч
3	Оцінювання за творчими індивідуальними унікальними роботами	Написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя
4	Заборона використання гаджетів під час очних контрольних робіт і суровий контроль	Заборона використання гаджетів під час очних контрольних робіт і суровий контроль
5	Використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт	Використання спеціального контролюючого програмного забезпечення в умовах дистанційного навчання під час контрольних робіт
6	Написання контрольних робіт від руки перед камерою, яка фіксує руки та обличчя	Оцінювання за творчими індивідуальними унікальними роботами
7	Усунення гуманітарних або тестових завдань в якості способів оцінювання студентів	Усунення гуманітарних або тестових завдань в якості способів оцінювання студентів
8	Суворі часові обмеження для написання контрольних робіт	Суворі часові обмеження для написання контрольних робіт
9	Вимога дотримання спеціального стилю оформлення робіт, який чат-боти не зможуть повторити	Вимога дотримання спеціального стилю оформлення робіт, який чат-боти не зможуть повторити

## 6. ВИСНОВКИ

Отримані результати є досить реалістичними, хоч і засновані на суб'єктивній оцінці. Дане дослідження є досить наочним і актуальним, адже побудоване на оцінці студента, який сам знаходиться у розглянутій ситуації і знає, як це працює зсередини. Була розроблена модель, яка може допомогти проаналізувати способи контролю якості вищої освіти в контексті використання студентами генеративних чат-ботів на основі ШІ.

Новизна роботи полягає в тому, що було досліджено конкретну складову впливу чат-ботів на якість вищої освіти і застосовано конкретні рішення, які дозволять підвищити якість вищої освіти. Особливістю даної роботи є те, що в ній для двох різних методів системного аналізу було використано майже однакові базові дані. Результати виявилися вражаюче подібними не дивлячись на різні способи їх отримання. Обидва методи показали майже однакові пріоритетні напрями для покращення якості вищої освіти. Такого результату вдалося досягти через експериментальне комбінування властивостей когнітивної і морфологічної моделей. Завдяки комплексному підходу визначення і оцінки параметрів результуючі моделі стали більш структурованими, репрезентативними, повноцінними, логічними, зрозумілими і адекватними.

Результати даного дослідження дозволять викладачам поставити пріоритети в тому, які методи взаємодії зі студентами будуть найбільш ефективними для покращення якості вищої освіти.

Ці моделі можна покращити шляхом уточнення початкових оцінок за допомогою великої кількості компетентних у даній області експертів, або за допомогою використання певних статистичних даних. У методі морфологічного аналізу можна додати нові контекстні параметри. У методі когнітивного моделювання можливо застосувати методи керування імпульсними процесами.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Abdullahi A. 10 Best AI Chatbots 2023. eWEEK. URL: <https://www.eweek.com/artificial-intelligence/best-ai-chatbots/#comparison-chart> (дата звернення: 10.11.2023).
2. Contributors to Wikimedia projects. ChatGPT. Wikipedia, the free encyclopedia. URL: <https://en.wikipedia.org/wiki/ChatGPT> (дата звернення: 15.11.2023).
3. Fengchun M., Wayne H. Guidance for generative AI in education and research. 7, place de Fontenoy, 75352 Paris 07 SP, France : UNESCO, 2023. 44 с. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000386693> (дата звернення: 13.11.2023).
4. Hulick K. How ChatGPT and similar AI will disrupt education [Електронний ресурс] / Kathryn Hulick // ScienceNews. – 2023. – Режим доступу до ресурсу: <https://www.sciencenews.org/article/chatgpt-ai-artificial-intelligence-education-cheating-accuracy>.
5. Kamalov F., Santandreu Calonge D., Gurrib I. New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*. 2023. Т. 15, № 16. С. 12451. URL: <https://doi.org/10.3390/su151612451> (дата звернення: 11.11.2023).
6. Морфологічний аналіз. Теорія, проблеми, застосування : навчальний посібник / Н.Д. Панкратова, І.О. Савченко ; М-во освіти і науки України, НТУУ "КПІ", Ін-т прикладного системного аналізу. Київ : Наукова думка, 2015. - 244 с.
7. Методологічне і математичне забезпечення розв'язання задач передбачення на основі модифікованого методу морфологічного аналізу / І.О. Савченко // Систем. дослідж. та інформ. технології. — 2011. — № 3. — С. 18-28. — Бібліогр.: 14 назв. — укр. URL: <http://dspace.nbuv.gov.ua/handle/123456789/50108>
8. Pankratova, N. & Naiko, Hennadii & Savchenko, Illia. (2021). Morphological model for underground crossings of water objects. *System research and information technologies*. 53-67. 10.20535/SRIT.2308-8893.2021.4.04. URL: [https://www.researchgate.net/publication/358359378\\_Morphological\\_model\\_for\\_underground\\_crossings\\_of\\_water\\_objects](https://www.researchgate.net/publication/358359378_Morphological_model_for_underground_crossings_of_water_objects)
9. Панкратова, Н. Д. Стратегія застосування методу морфологічного аналізу в процесі технологічного передбачення / Н. Д. Панкратова, І. О. Савченко // Наукові вісті НТУУ «КПІ» : науково-технічний журнал. – 2009. – № 2(64). – С. 35–44. – Бібліогр.: 16 назв.
10. Педагогіка вищої школи [Електронний ресурс] : підручник / В. П. Головенкін ; КПІ ім. Ігоря Сікорського. – 2-ге вид., переробл. і доповн. – Електронні текстові дані (1 файл: 3,6 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2019. – 290 с.
11. Анненкова І. П. Критерії і показники якості освіти у ВНЗ. *Наука і освіта*. 2011. URL: [https://scienceandeducation.pdpu.edu.ua/doc/2011/8\\_2011/1.pdf](https://scienceandeducation.pdpu.edu.ua/doc/2011/8_2011/1.pdf) (дата звернення: 01.11.2023).
12. Гапон Л. О. Показники ефективності освітньої діяльності педагога. Методист ТКМЦНОІМ Гапон Л. О. Блог учителів української мови і літератури міста Тернополя. URL: <https://gapon.te.ua/rubryka-metodysta/dorobok-metodysta/metodychni->

rekomendatsii/item/1360-mekhanizm-pobudovy-u-zakladi-osvity-vnutrishnoyi-systemy-otsinyuvannya-yakosti-osvity (дата звернення: 21.10.2023).

13. Мілявський, Ю. Л. Ідентифікація та керування складними системами на основі моделей імпульсних процесів когнітивних карт : дис. ... д-ра техн. наук. : 01.05.04 Системний аналіз і теорія оптимальних рішень / Мілявський Юрій Леонідович. – Київ, 2019. – 297 с. URL: <https://ela.kpi.ua/handle/123456789/43829>

14. Roberts F. Discrete Mathematical Models with Applications to Social, Biological, and Environmental Problems. – Englewood Cliffs, Prentice-Hall, 1976. – 559 p.

15. Романенко В. Д. Когнітивне моделювання динаміки прийняття рішень для стабілізації нестійких режимів у соціально-навчальному процесі студента / В. Д. Романенко, Ю. Л. Мілявський // Наукові вісті НТУУ «КПІ» : науково-технічний журнал. – 2016. – № 5(109). – С. 48–53. – Бібліогр.: 10 назв. URL: <https://ela.kpi.ua/handle/123456789/18873>

16. Метод проектування когнітивної карти для оптимізації профорієнтаційної діяльності ЗВО. ВБ Мокін, ОВ Бурдейна, КО Коваль, АР Ящолт. ВНТУ, 2018. URL: [https://www.researchgate.net/publication/330089909\\_METOD\\_PROEKTUVANNA\\_KOGNITIVNOI\\_KARTI\\_DLA\\_OPTIMIZACII\\_PROFORIENTACIINOI\\_DIALNOSTI\\_ZVO](https://www.researchgate.net/publication/330089909_METOD_PROEKTUVANNA_KOGNITIVNOI_KARTI_DLA_OPTIMIZACII_PROFORIENTACIINOI_DIALNOSTI_ZVO)

17. Системный анализ : проблемы, методология, приложения: монография / М.З. Згуровский, Н.Д. Панкратова; Министерство образования и науки, молодежи и спорта Украины, НАН Украины, Институт прикладного системного анализа. Киев : Наукова думка, 2011 - 726 с.

# ПРОГНОЗУВАННЯ ГЕТЕРОСКЕДАСТИЧНИХ ПРОЦЕСІВ ДЛЯ ОЦІНЮВАННЯ ФІНАНСОВОГО РИЗИКУ

Байбара А.Г.<sup>1</sup>, Кузнєцова Н.В.

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

<sup>1</sup> angelinabaibara@gmail.com

**Ризики є елементами ведення фінансової діяльності та є наслідком невизначеності, тому для зменшення і компенсації їх негативних ефектів використовують усталені методи оцінки фінансового ризику. Метою роботи є оцінювання ризиків різними методами та їх порівняння для виявлення найбільш ефективних. В роботі представлені моделі для оцінки ризиків VaR та CVaR, а також гетероскедастичні моделі для опису динаміки волатильності фінансових процесів. Такий аналіз фінансово-економічних процесів та ризиків дає можливість для обґрунтованого прийняття рішень та управління ризиками в реальному світі.**

**Ключові слова: фінансові ризики, системний підхід, VAR, CVAR, аналіз ризиків, ринковий ризик.**

## 1. ВСТУП

В сучасному фінансовому середовищі, де ринкові коливання можуть мати значущий вплив на інвестиційні портфелі та фінансові установи, точна оцінка ризику є надзвичайно важливим завданням для інвесторів та гравців на фінансових ринках при прийнятті рішень та побудови стратегій інвестування. Тож для всіх учасників ринку постає задача точної оцінки ризиків в умовах, коли більшість фінансово-економічних процесів характеризуються нестаціонарністю, нелінійністю та сильною волатильністю, а також сильною залежністю від багатьох зовнішніх факторів. Це, в свою чергу, ще більше обумовлює необхідність і важливість мати точне розуміння реального становище своїх позицій на ринку та можливих втрат. Дане дослідження спрямоване на вдосконалення методів прогнозування ризиків (VaR та CVaR), а також на їхнє застосування в контексті ефективного управління ризиком та його контролю на фінансових ринках.

## 2. МЕТОДИ ОЦІНЮВАННЯ ФІНАНСОВИХ РИЗИКІВ

Фінансові ризики зазвичай поділяють на чотири основні категорії: операційний ризик, ризик ліквідності, кредитний ризик і ринковий ризик. Під час формування вартості портфеля та розрахунку ризикового капіталу враховують всі ці ризики, зазвичай вони розглядаються окремо і працюють з ними незалежно. Ефективне управління ризиками виявляється критичним для визначення можливих втрат та захисту інвесторів від потенційних збитків у негативних сценаріях, або навіть від настання банкрутства фінансових установ та фондів.

Оскільки фінансовий ризик обумовлений невизначеністю щодо результатів фінансових операцій у майбутньому, отримання прийнятних оцінок не завжди є тривіальним завданням. Тому розвиток ефективних та науково обґрунтованих методів аналізу та оцінювання фінансових ризиків визначається як важливий напрямок науки.

Фінансово-економічні процеси характеризуються нестационарністю, нелінійністю та волатильністю. Традиційно волатильність відіграє важливу роль у вимірюванні ризику: щоб оцінити волатильність, необхідно розробити модель, що враховує зміни волатильності в часовому ряді на основі гетероскедастичних процесів [1]. Методи оцінювання фінансового ризику, такі як методологія VaR (Value at Risk) та моделі прогнозування дисперсії втрат, включаючи ARCH (Autoregressive Conditional Heteroskedasticity) та GARCH (Generalized Autoregressive Conditional Heteroskedasticity), стають об'єктом дослідження для розробки точних прогнозів та стратегій мінімізації ризику.

Вартість під ризиком (VaR) — метод, який використовується для оцінки фінансового ризику шляхом оцінки потенційно можливої несприятливої зміни в ринковій вартості портфеля із заданим рівнем довірчої ймовірності за певний період часу [2]. Перевагами даного підходу були простота і легкість в інтерпретації, а також наявність всього одного агрегованого показника. VaR показує величину збитків, яку компанія не перевищить з певною ймовірністю за певний період часу.

Прогнозування VaR та CVaR визначається необхідністю адекватної кількісної оцінки ризиків для прийняття обґрунтованих фінансових рішень [3]. У цьому контексті, прогнозування VaR надає можливість визначити верхню межу можливого збитку з певним рівнем ймовірності, що стає ключовим фактором для управління портфелем та прийняття рішень щодо розподілу активів, проте VaR може не дати повної картини ризиків, оскільки ця міра байдужа до всього, що виходить за межі її власного порогу. Спільно з цим, прогнозування CVaR допомагає врахувати не тільки екстремальні величини збитків, але й їхню середню інтенсивність, що дозволяє ліпше розуміти загальний ризик у разі великих втрат.

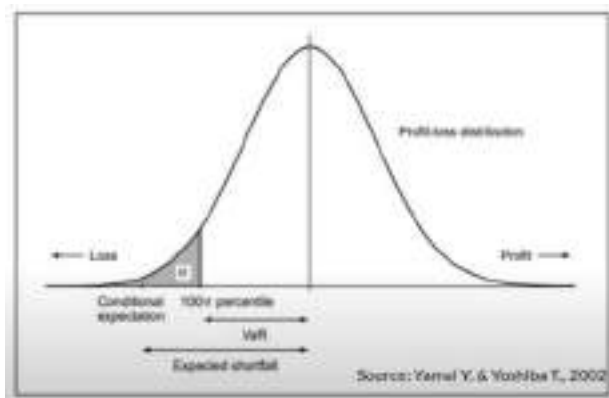


Рисунок 1. Візуалізація VaR та CVaR на графіку розподілу портфелю

Методи оцінювання VaR поділяються на дві групи: параметричні та непараметричні. До непараметричних відносяться історичний метод та метод моделювання Монте Карло, який відрізняється від попередніх більшою складністю обчислень, однак показує більш точні результати.

Основною проблемою під час обчислення VaR параметричними методами є моделювання та прогнозування волатильності доходності інструментів. Використання дельта-нормального методу для VaR передбачає оцінку волатильності доходності фінансових інструментів, для чого важливо врахувати змінність дисперсії в часі.

Динаміка дисперсії фінансових процесів може бути адекватно описана моделлю авторегресії з умовною гетероскедастичністю (ARCH) [4]. Гетероскедастичність вказує на змінність дисперсії у часі, і використання ARCH моделі дозволяє побудувати математичну структуру, яка враховує цю змінність і на її основі можна робити прогнози на наступні кроки. Також для моделювання волатильності широко використовується узагальнена

модель авторегресії з умовною гетероскедастичністю (GARCH модель) [5], та їх модифікації, такі як EGARCH, FIGARCH моделі [6].

Створивши адекватну модель, ми зможемо суттєво покращити якість управління ризиком, оскільки це дозволить нам краще розуміти та передбачати зміни в динаміці ризику у фінансових процесах

Для впевненості в адекватності використовуваної моделі оцінки ризику, необхідно провести її верифікацію, що включає процес "бек-тестування" [7]. Верифікація дозволяє визначити, наскільки точно модель VaR відображає реальні умови ринку.

### 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для застосування методологій оцінювання ринкового ризику було сформовано інвестиційний портфель з акцій компаній Apple (AAPL), META, The Walt Disney Company (DIS), Citigroup Inc (C). Будемо оцінювати ризик на основі щоденних цін закриття акцій за 2 роки. Таким чином, часовий період для значень акцій компаній — 4 роки (з 1 січня 2019 по 1 січня 2023), а кількість спостережень — 1008.

Найпростішим методом розрахунку VaR є історичний метод, згідно з яким модель історичної прибутковості вказує на модель майбутніх прибутків. Припущення щодо розподілу доходів не робиться, а моделюється на основі історичних емпіричних даних. Візуалізуємо для цього методу на гістограмі розподілу доходностей портфелю підраховані значення VaR та CVaR для рівня довіри 95%, який використовується в системі RiskMetrics (Рис. 2).

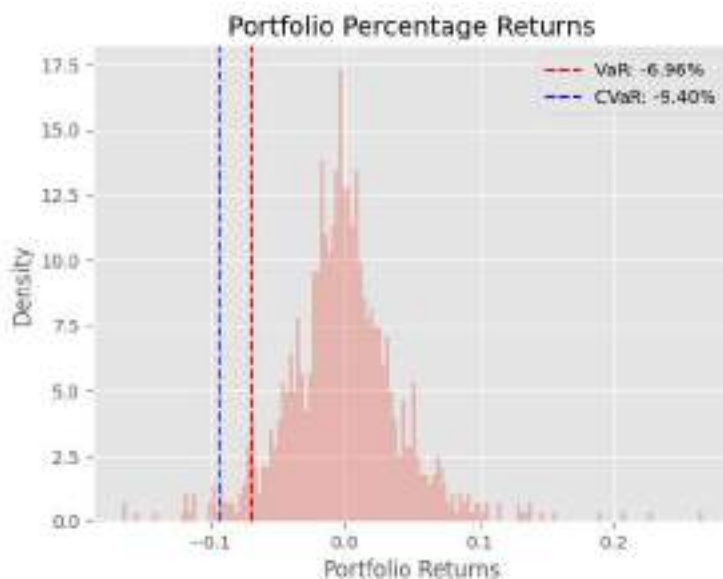


Рисунок 2. Розподіл прибутковості портфелю з VaR та Cvar за історичним методом

Для розрахунку VaR варіаційно-коваріаційна методом припущення полягає в тому, що доходи розподілені нормально і використовуються історичні доходи портфелю та стандартні відхилення (оцінки) для визначення параметрів моделі. Визначаються параметри нормального розподілу, що найкращим чином апроксимує фактичний розподіл розглянутого ринкового показника. Далі необхідно визначити значення зворотного нормального розподілу – для довгої позиції, а отже, негативної зміни, на основі параметрів, отриманих раніше, і зворотного рівня довіри.

Наступним методом оцінки VaR є параметричний метод на основі прогнозованого значення волатильності ряду. Для моделювання дисперсії, що є гетероскедастичним процесом, було застосовано такі моделі – ARCH, GARCH, EGARCH, FIGARCH.

Результати порівняння якості побудованих моделей зведено до Таблиці 1.

Таблиця 1. Характеристики різних моделей прогнозування волатильності портфелю акцій

Тип моделі	AIC	BIC
ARCH(1)	-6213,287764	-6203,013207
GARCH(1,1)	-6372,040331	-6356,628496
EARCH(1,1)	-6194,814760	-6184,540203
FIGARCH(1,1)	-6362,575627	-6342,026513

За результатами моделювання всі моделі виявились задовільними та можуть використовуватись для подальших оцінок рівня ризику, проте за інформаційними критеріями Акайке і Баеса-Шварца, найкращою моделлю для волатильності і найбільшим точним описом серії є GARCH(1,1).

Останнім методом оцінки ризику є метод Монте-Карло. Метод Монте-Карло є найбільш складним методом розрахунку VaR, проте його точність може бути значно вищою, ніж у інших методів. Метод Монте-Карло передбачає здійснення великої кількості випробувань – разових моделювань розвитку ситуації на ринку з розрахунком одержаного результату за портфелем. За підсумком випробувань формується розподіл можливих результатів. Відповідно до обраного рівня вірогідності відсікаються найгірші варіанти і одержується VaR-оцінка. Для найбільшої точності, характеристики стохастичного процесу, покладеного в основу симуляції, мають збігатись з аналогічними показниками процесу, який ми досліджуємо, але це не обов'язково.

Оскільки оцінка VaR методом Монте-Карло майже завжди виконується за допомогою програмних засобів, які вимагають великої кількості симуляцій для формування прогнозу. Іншими словами, метод Монте-Карло дозволяє використовувати моделі ризиків практично будь-якої складності при розрахунках. При цьому, збільшення кількості симуляцій підвищує точність обчислень. Застосувавши 400 симуляцій, було визначено VaR для довірчого інтервалу – 95%, а також визначено показники очікуваних втрат CVaR при початковій інвестиції у розмірі 100000\$ (Рис. 3).

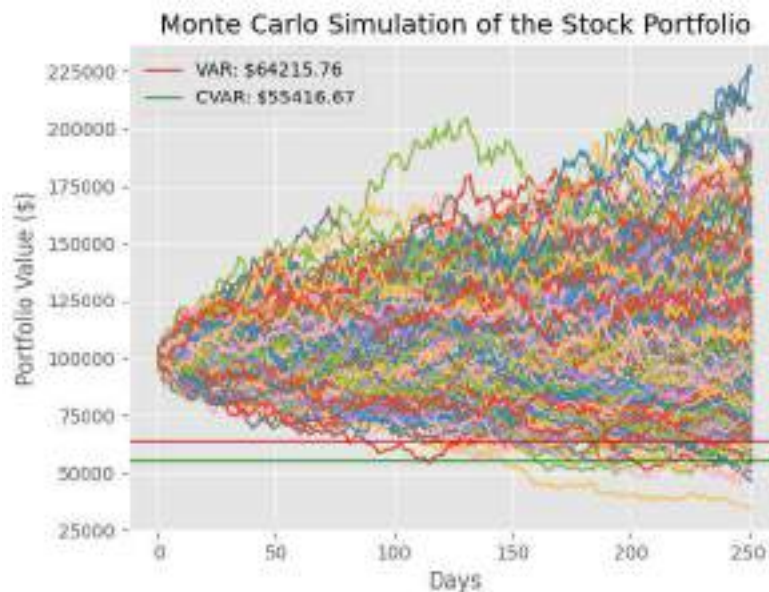


Рисунок 3. Значення VaR та CVaR для методу Монте Карло alpha = 5%

Результати обчислень VaR та CVaR різними методами для нашого портфелю наведені у таблиці нижче (Табл. 2).

Таблиця 2. Значення VaR та CVaR для різних рівнів довіри та різних методів обчислення

Довірчий інтервал	Метод оцінки ризику	Історичний	Параметричний N(0,1)	Параметричний GARCH(1,1)	Монте Карло
95%	VaR	-4,1%	-5,12%	-5,01%	-4,83%
	CVaR	-6,85%	-6,32%	-6,26%	-5,56%
99%	VaR	-5,55%	-5,63%	-6,32%	-5,06%
	CVaR	-9,4%	-6,42%	-7,15	-5,97%

Результати верифікації моделей для оцінювання ризику за допомогою процедури бек-тестування подано у Таблиці 3.

Таблиця 3. Результати процедури бек-тестування для оцінювання VaR

Довірчий інтервал	Метод оцінки ризику	Історичний	Параметричний N(0,1)	Параметричний GARCH(1,1)	Монте Карло
95%	VaR	94,9%	96,4%	96,54%	95,3%
	CVaR	97,5%	97,3%	97,1%	96,7%
99%	VaR	98,4%	98,45%	98,67%	98%
	CVaR	99,1%	98,64%	99,21%	98,56%

Загалом, можна зробити висновок, що усі моделі є у прийнятній мірі адекватними. Історичний метод дає гіршу оцінку ризику, що цілком закономірно, враховуючи, що цей метод базується тільки на історичних даних, таким чином історичний метод може не зреагувати на раптові зміни на ринку. Найкращими виявились параметричний метод на основі GARCH(1,1) та метод Монте Карло. Вони не завищують ризику і не занижують ризику, і, як показує бектестування, є гнучкими до змін на ринку.

Методи оцінки VaR з довірчим інтервалом 99% можуть дещо недооцінювати ризик при настанні екстремальних ситуацій. Саме в таких випадках треба звертати увагу на значення CVaR, яке є більш прийнятним і забезпечує точнішу оцінку ризику при 99% довірчому інтервалі. Таким чином, ми надаємо інвестору більш адекватну картину щодо його інвестицій, щоб він міг, по-перше, прийняти рішення щодо інвестування, або навпаки, про виведення інвестиції вчасно з урахуванням всіх ризиків, а по друге, мати достатній капітал для покриття збитків.

Враховуючи складність в обчисленні ризиків за методом Монте Карло, метод оцінки ризиків на основі прогнозування волатильності GARCH(1,1) моделлю можна вважати хорошою альтернативою. На основі оцінки VaR та CVaR даним методом з різними довірчими інтервалами, можна приймати рішення щодо інвестувань та стратегію, яка вам найбільше підходить на ринку.

## 4. ВИСНОВКИ

В даній роботі було розглянуто фінансові ризики, які є важливою складовою на фінансовому ринку, та дозволяють комплексно оцінити можливі майбутні втрати. Складання оптимального портфеля цінних паперів є важливою практичною задачею на фондовому ринку. Велике значення мають наукові дослідження в галузі математичного моделювання процесів оцінювання фінансових ризиків та управління ними. Тож використання ймовірнісних функціоналів VaR та CVaR і розрахунок їх за допомогою математичних моделей буде дуже корисним для даної практичної задачі.

У ході дослідження було реалізовано методології VaR та CVaR оцінки ринкового ризику на прикладі інвестиційного портфелю, сформованого з акцій, представлених на біржі NASDAQ. Було виконано порівняння отриманих результатів для низки методів, а саме історичного, параметричного, методу Монте-Карло та методу на основі прогнозування волатильності за допомогою моделювання гетероскадестичних процесів, зокрема GARCH моделлю. Беручи за основу GARCH-модель волатильності для опису змінної в часі дисперсії була досягнута висока точність прогнозів, придатна для прийняття рішень під час виконання фінансових операцій. Також була проведена процедура бек-тестування для верифікації використаних моделей оцінки ризику, що дозволяє зрозуміти наскільки точно модель VaR та CVaR відображає реальні умови ринку. Це дозволило комплексно оцінити ризики, а також порівняти точність різних методів оцінки VaR та CVaR, що дозволить надалі інвесторам приймати більш обґрунтовані рішення під час гри на фінансовому ринку.

### ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Кузнєцова Н.В., Бідюк П.І. Теорія і практика аналізу фінансових ризиків: системний підхід, монографія. Київ 2020, 400 с.
2. Башкіров О.В. Порівняльний аналіз VAR-методів оцінки ризику активів банку О.В. Башкіров, Проблеми і перспективи розвитку банківської системи України : зб. наук. праць ДВНЗ «УАБС НБУ». Вип. 14. С. 302–309.
3. Moraux F. How valuable is your VaR? Large sample confidence intervals for normal VaR. F. Moraux. Journal of risk management in financial institutions. 2011. № 4.2. P. 189–200. – [Електронний ресурс]. – Режим доступу : <https://perso.univ-rennes1.fr/franck.moraux/research/JRMFI.pdf>
4. Bollerslev T., Chow R., Kroner K. ARCH modeling in finance: a review of the theory and empirical evidence. Journal of Econometrics. 1992. Vol. 52. P. 5–59.
5. Bollerslev T. General autoregressive conditional heteroscedasticity. Journal of Econometrics. 1986. Vol. 31. P. 518–537.
6. Kilic, R. Conditional Volatility and Distribution of Exchange Rates: GARCH and FIGARCH Models with NIG Distribution. Studies in Nonlinear and Econometrics. Vol. 11. P. 1-31
7. Longerstae J. Risk Metrics TM – Technical Document . J. Longerstae, M. Spencer. Morgan Guaranty Trust Company of New York: New York, 1996. [Електронний ресурс]. – Режим доступу : <http://yats.free.fr/papers/td4e.pdf>

# ПІДХІД ЩОДО ВИЗНАЧЕННЯ ФАКТОРІВ ВПЛИВУ НА РИНОК ЛОГІСТИКИ США З ВИКОРИСТАННЯМ LLM

Балькін Я.Ю.<sup>1</sup>, Савастьянов В.В.

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

<sup>1</sup> yarik.balkin@gmail.com

**Ринок логістики одним з найвразливіших ринків в економіці США, тому для ефективної оцінки та якісного прогнозування потрібно знаходити фактори впливу, та працювати з цими факторами. Метою роботи є розробка підходу задля ефективного визначення факторів впливу з застосуванням потужностей штучного інтелекту. Результатом дослідження є основні фактори впливу інших галузей на галузь логістики. У роботі було використано теоретичні та емпіричні методи дослідження, а також Large Language Model.**

**Ключові слова:** логістика, вантажні перевезення, LLM, TSI, текстовий аналіз, фактори впливу.

## 1. ВСТУП

Сектор вантажних перевезень в США займає значне місце в формуванні економіки. На перевезення вантажів витрачається сотні мільярдів доларів на рік, а вплив інших сфер на цей сектор непомірно великий. Через це вантажні перевезення і викликають непомірний інтерес для аналітиків з усього світу. Ціноутворення на логістичні послуги має дуже складну систему через вплив сотень факторів, і дослідження цих факторів є необхідним задля моделювання ризиків та прогнозування.

На сьогоднішній день були проведені дослідження щодо ціноутворення в окремих галузях вантажних перевезень, а також дослідження які вивчали фактори впливу між вузькоспрямованими вантажними перевезеннями та іншими галузями. Ми ж спробуємо використовуючи більш новітні методи аналізу знайти та оцінити фактори більш широкого спектру впливу.

## 2. МЕТОДИ ПОШУШУ ФАКТОРІВ ВПЛИВУ

Задля більш точного та кваліфіковано пошуку факторів впливу було вирішено пропрацювати методи описані в роботах інших дослідників [1–5]. Так наприклад в роботах [4, 5] аналітика вантажних перевезень була проведена завдяки створенню та пропрацюванню часових рядів різними методами. Також були взяті роботи Українських дослідників [1–3] для більш точного розуміння сфери вантажних перевезень.

Стаття [1] досліджує темпи розвитку логістики в Україні. Акцент цієї роботи припадає на оцінку географічне положення України і розглядання динаміки росту та падіння об'єму вантажних перевезення за 2010–2018 роки. Також в роботі було розглянуто сфери які мають вплив на логістику та описана сила впливу цих сфер.

В наступній статті [2] були розглянуті тенденції розвитку Українського ринку логістики. Задля спостереження динаміки було введено індекс ефективності логістики. Були розглянуті основні країни експортери та імпортери, була проведена оцінка сили впливу експорту та імпорту на логістичну сферу. Також було розглянуто зв'язок ВВП країни та логістичної сфери.

У результаті цього дослідження було виділено фактори зростання попиту на логістичні послуги, проведено сегментацію ринку логістики в Україні та класифікацію цих послуг.

Стаття [3] розглядала логістику як галузь прикладних наук. В ній було проведено паралель між «минулим» та «теперішнім» логістики. Розглядалась еволюція ринку логістики, які питання були актуальні на той час, які фактори впливу були у 1960х роках та їх еволюціонування у сучасні проблеми цієї сфери. Тенденції та темпи розвитку того часу і їх розвиток з часом. Також було розглянуто нові ідеї для визначення сил факторів впливу на основі минулих досліджень та спостережень ринку перевезень.

Статті [4, 5] використовують більш новітні методи прогнозування та дослідження ринку вантажних та пасажирських перевезень. В них запроваджені моделі ARIMA, ARMA, SARIMA та інші для прогнозування попиту та ціноутворення логістичних послуг в США. А для аналізу були взяті індекси TSI для вантажних та пасажирських перевезень, який характеризує об'єми цих перевезень. Також у статті [4] результатами досліджень були модельні прогнози ціни на транспортні перевезення вантажівками Ці дані були отримані на основі аналізу TSI та сьогodenних цін на транспортні перевезення вантажними машинами (рис. 1). Але ціноутворення вантажних перевезення сильно вразливе до попиту, тому треба враховувати що пристосування перевізників до зміни цін займає час до бти місяців.

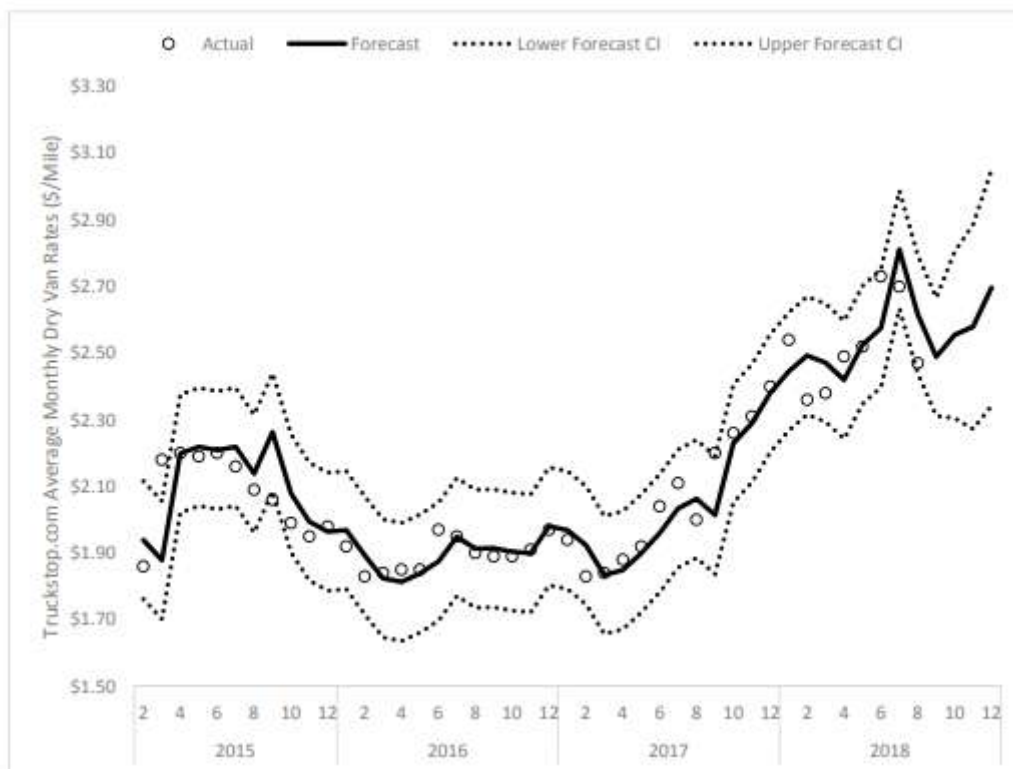


Рисунок 1. Модельні прогнози на вантажні перевезення

На основі аналізу робіт інших авторів було вирішено використовувати більш новітню методику аналітики, а саме LLM. Завдяки LLM ми можемо урахувати більше факторів впливу, а також сезонність та вразливість до інших нестандартних факторів впливу. LLM дозволить нам обробити великі об'єми даних, а також допоможе їх правильно структурувати та класифікувати. Ще однією з причин щодо вибору LLM було врахування всіх недоліків всіх попередніх методів, та можливість їх модифікування. Так як LLM дуже гарно працює з

текстами, було вирішено шукати фактори впливу в новинних джерелах, і так як LLM дозволяє гарно сортувати та відсіювати непотрібне, ми можемо дозволити собі працювати з несумісними, на перший погляд, текстовими даними. Також для пошуку факторів впливу будемо розглядати числовий регресор TSI для вантажних перевезень. Цей індекс допоможе нам якісніше оцінити фактори впливу та кореляцію між галузями. Також цей індекс являє собою готові проаналізовані дані і завдяки ньому ми можемо бачити зріст та падіння логістичної сфери в різних проміжках часу (Рис. 2).

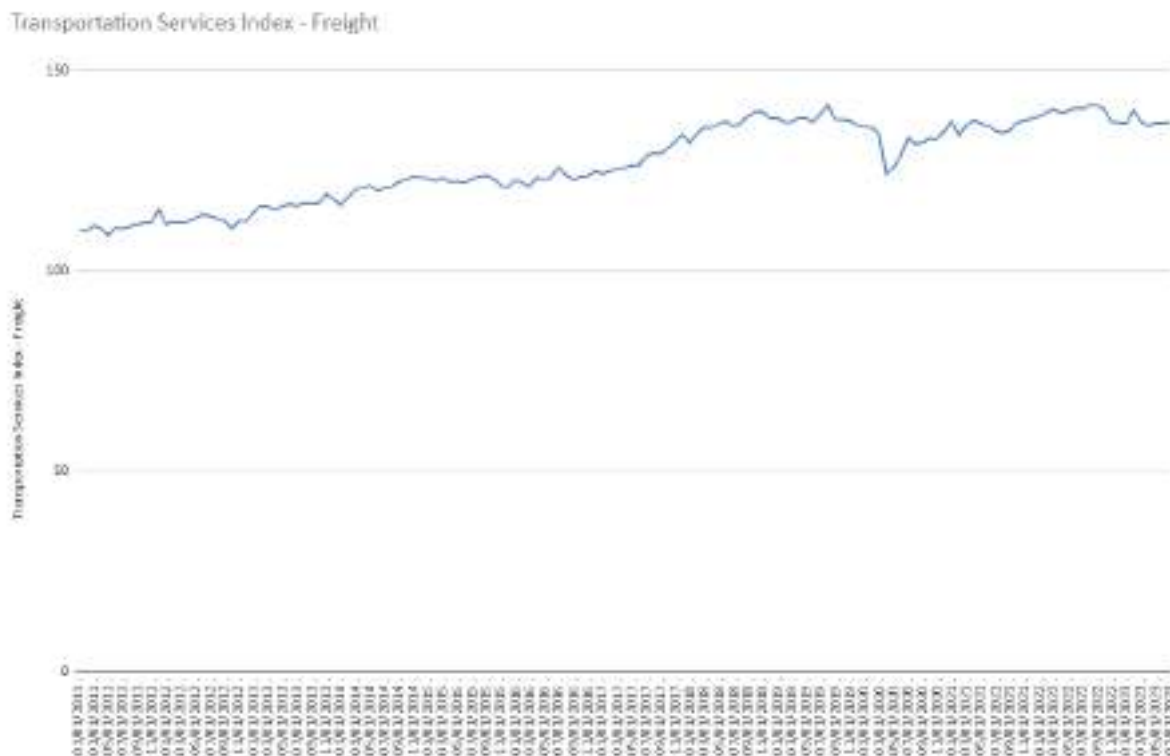


Рисунок 2. Графік змін TSI

### 3. ВХІДНІ ДАНІ ТА ОЧІКУВАНІ РЕЗУЛЬТАТИ

Для дослідження факторів впливу було сформовано базу даних текстової інформації, яка являє заголовки новин (Табл. 2), та часовий ряд числової інформації, яка являє собою TSI – Freight індекс (Табл. 1). Інформацію було взято за останні 12 років задля підвищення точності. Для пошуку залежностей було проведено накладання подій, на наш графік змін TSI(рис ). Це дозволить нам визначати час, коли відбувались значущі події для сфери логістики і тим самим визначити ці події.

Таблиця 1. TSI – Freight індекс за 2011 рік

Date	Transportation Services Index - Freight
01/01/2011 12:00:00 AM	110,3
02/01/2011 12:00:00 AM	109,7
03/01/2011 12:00:00 AM	111,1
04/01/2011 12:00:00 AM	110,6
05/01/2011 12:00:00 AM	108,7
06/01/2011 12:00:00 AM	110,6
07/01/2011 12:00:00 AM	110,4
08/01/2011 12:00:00 AM	110,9
09/01/2011 12:00:00 AM	111,4
10/01/2011 12:00:00 AM	111,7
11/01/2011 12:00:00 AM	112
12/01/2011 12:00:00 AM	115,1

Таблиця 2. Приклад текстових даних

June 2, 2023	Why Americans Want Part-Time Jobs Again
June 1, 2023	IEA Head Wants Fossil Fuel Industry To Set Climate Targets
June 1, 2023	Airlines Are Weighing Passengers
June 1, 2023	How Grab Became a Super-App
May 31, 2023	Why Janet Yellen Doesn't Lose Sleep Over U.S. Borrowing
May 31, 2023	Twitter Now Worth One-Third of What Musk Paid for It
May 30, 2023	AI Is as Risky as Nuclear War, Top CEOs Say
May 27, 2023	Twitter Withdraws From EU Disinformation Code, Commissioner Says
May 26, 2023	Why Gas Prices Are Cheaper Right Now
May 25, 2023	Why Europe's New Climate Rules Matter to American Companies
May 25, 2023	Germany Endures First Recession Since COVID on Consumers
May 24, 2023	Yellen: Treasury 'Not Involved' in Planning With Investors for Default

Текстові дані потребують попередньої обробки, тому в першу чергу треба зробити відсів даних які зовсім не стосуються логістики, наприклад стаття з заголовком «Philippines' Duterte Subpoenaed Over Alleged Death Threat» ніяк не стосується логістики навіть опосередковано. В той час як стаття «China Invests \$5.4 Billion in Chipmaker» цілком може вплинути на коливання сфери логістики. Початковий дата сет має в собі більш ніж 200 тисяч рядків, коли після обробки цей список зменшується до 175 тис. (Рис. 3).

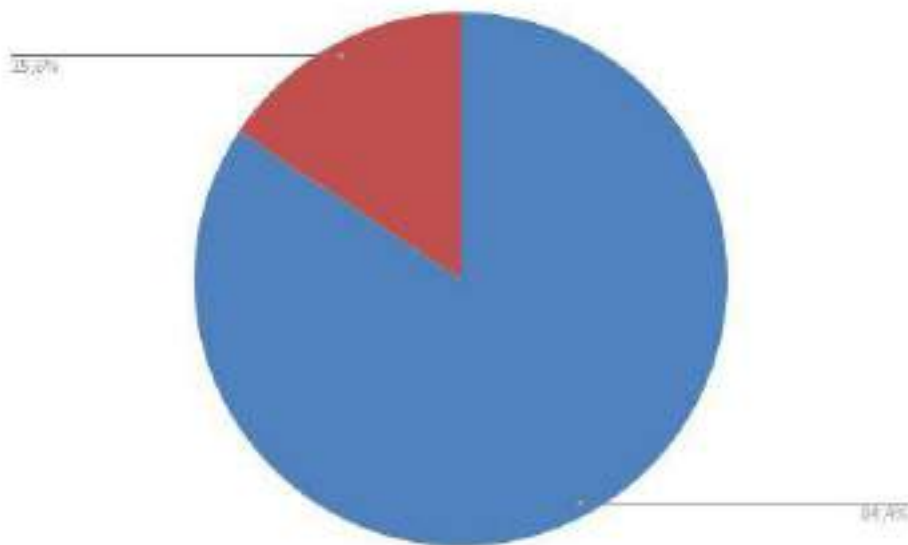


Рисунок 3. Відношення ефективних та неефективних текстових даних

Наступним етапом нашої роботи йде класифікація. Для початкової класифікації нам підійде модель STEEP(Social, Technological, Economical, Environmental and Political), для початкової класифікації. Для уточнення будемо користуватись підходом «few-shot prompting». На вхід будуть подаватись заголовки – після цього кожний заголовок буде оброблюватись моделлю GPT-3.5, використовуючи заздалегідь прописані правила класифікації(на цьому етапі застосовується підхід «few-shot prompting», який окрім правил використовує ще й підказки) – на виході отримуємо класифікований заголовок.

Основна задача цієї роботи – це знайти фактори впливу на сферу логістики, та визначити їх силу. Щоб цього досягти кожна класифікована новина буде подаватися до моделі GPT-3.5, оброблюватися завдяки правилам, прописаним заздалегідь, та буде отримувати свою силу впливу. Після всієї обробки планується обрахувати які сектори мають більший вплив на сферу логістики, а також які саме події викликали різкі збільшення/зменшення об'єму вантажоперевезень.

#### 4. ВИСНОВКИ

Об'єми вантажних перевезень збільшуються із року в рік, тому для працівників цієї галузі дуже важливим є розуміння ринку, його оцінка та аналітика економічної ситуації. Також вважаючи те що ринок логістики є одним найбільшим у США, його кардинальні зміни приведуть до змін майже у всіх економічних сферах.

Завдяки новітнім технологіям, а саме LLM ми можемо розглядати та оцінювати такі великі структури. Займатися прогнозуванням та плануванням. А завдяки тому що LLM має велику бібліотеку даних та вміє в ній орієнтуватись, це може стати заміною експертної думки одразу у багатьох галузях. Завдяки цьому LLM може пов'язувати факти, опосередковано, або й зовсім не зв'язаних з логістикою. Також через гарне розуміння LLM текстової інформації наш підхід буде не тільки інноваційним, але й досить точним порівняно з іншими методами аналізу(використання часових рядів, тощо)

У ході дослідження планується розробити універсальний продукт виявляючий фактори впливу на логістичний ринок. Продукт який буде підлаштований до нештатних ситуацій, та буде використовувати не тільки статистичну модель, а також LLM.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Статистичний аналіз структури та тенденцій розвитку логістичного ринку України / Р. В. Ціщик, Н. В. Котис // Проблеми системного підходу в економіці. - 2018. - Вип. 3(1). - С. 54-59. - [Електронний ресурс] - Режим доступу: [http://nbuv.gov.ua/UJRN/PSPE\\_print\\_2018\\_3%281%29\\_\\_11](http://nbuv.gov.ua/UJRN/PSPE_print_2018_3%281%29__11)
2. АНАЛІЗ РИНКУ ЛОГІСТИЧНИХ ПОСЛУГ В УКРАЇНІ ЗА 2010-2018 РОКИ / Наталія Ю. Кирлик - 2019. – С. 83-95. - [Електронний ресурс]. – Режим доступу <https://dspace.uzhnu.edu.ua/jspui/bitstream/lib/31105/1/83--95.pdf>
3. Logistics research: A 50 years' march of ideas/ Peter Klaus – 2008. – С. 53-65. - [Електронний ресурс]. – Режим доступу [https://www.researchgate.net/publication/220232983\\_Logistics\\_research\\_A\\_50\\_years%27\\_march\\_of\\_ideas](https://www.researchgate.net/publication/220232983_Logistics_research_A_50_years%27_march_of_ideas)
4. ARIMA Time Series Models for Full Truckload Transportation Prices/Jason W. Miller – 2019. – [Електронний ресурс]. – Режим доступу <https://www.mdpi.com/2571-9394/1/1/9>
5. Forecasting Daily and Weekly Passenger Demand for Urban Rail Transit Stations Based on a Time Series Model Approach/ Dung David Chuwang, Weiya Chen – 2022. - [Електронний ресурс]. – Режим доступу <https://www.mdpi.com/2571-9394/4/4/49>
6. ChatGPT takes the stand in its defense/DAN LUCARINI – 2023. - [Електронний ресурс]. – Режим доступу <https://www.deep-analysis.net/chatgpt-takes-the-stand-in-its-defense/>
7. What is a STEEP Analysis? - [Електронний ресурс]. – Режим доступу <https://www.utsdesignindex.com/researchmethod/steep-analysis/>

# **ПРОГНОЗУВАННЯ ЦІНИ НА ЗОЛОТО МЕТОДАМИ МАШИННОГО НАВЧАННЯ**

Білоус К.С.<sup>1</sup>, Кузнєцова Н.В.

Національний технічний університет України «Київський політехнічний інститут  
ім. Ігоря Сікорського»

<sup>1</sup> katyabelousal@gmail.com

**Золото - є ключовим дорогоцінним металом для економіки світу, хоча його ціна постійно коливається. Саме тому своєчасна актуальна інформація про його ціну є вкрай важливою. Методами машинного навчання можна виконати прогнозування ціни на золото з високою точністю, що дозволяє забезпечити краще управління ризиками та приймати обґрунтовані рішення щодо продажу чи придбання дорогоцінного металу. Метою роботи є розробка системи для прогнозування цін на золото. Результатом дослідження є розроблена система прогнозування цін на золото.**

**Ключові слова:** золото, машинне навчання, прогнозування, дорогоцінні метали, прогнозування цін.

## **1. ВСТУП**

Золото – це дорогоцінний метал, що відомий своєю стійкістю до корозії, високою пластичністю та хорошою електропровідністю. Саме через його фізичні властивості його високо цінують і використовують у ювелірній сфері, при виробництві електроніки, медичного обладнання, а також золото має статус "безризикового активу" і використовують як спосіб зберігання грошей [1].

Золото як спосіб зберігання вартості можуть використовувати як приватні особи, так і країни. Такими прикладами є золотовалютні резерви – запас золота та іноземної валюти, що утримується центральним банком країни. Золотовалютний резерв впливає на загальну економічну динаміку країни, а його правильне управління є важливим елементом для забезпечення стабільності та розвитку [2].

Зважаючи на це золото є важливим елементом економіки країн і бізнесу. Відповідно, є важливим вміти прогнозувати ціни на золото.

## **2. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ**

Мета даного дослідження – розробка системи прогнозування ціни на золото методами машинного навчання. В роботі розглядаються різні існуючі методи роботи з прогнозуванням цін дорогоцінних металів. Виконується порівняння конкретних методів для роботи з даною задачею на основі методів машинного навчання. Об'єкт дослідження – ціни на золото. Предмет дослідження – методи та моделі машинного навчання для прогнозування цін на золото.

## **3. ОГЛЯД КЛАСИЧНИХ МЕТОДІВ ПРОГНОЗУВАННЯ ЦІН НА ЗОЛОТО**

Існують різні методики прогнозування цін на дорогоцінні метали, такі як золото, срібло, платина та інші. Основними напрямками в даній задачі є наступні методи:

- Аналіз фундаментальних факторів.

Цей підхід враховує фактори, які можуть впливати на попит і пропозицію на ринку дорогоцінних металів. Це може включати політичні події, економічний стан країн, геополітичні конфлікти, інфляцію, торгові відносини між країнами, зміни відсоткових ставок та інші фактори. Аналізуючи ці дані, можна робити припущення щодо майбутньої динаміки цін [3].

Зміни у політиці, такі як введення нових законодавчих актів щодо фінансової політики, можуть вплинути на ціни дорогоцінних металів. Економічний стан країн, зокрема країн-виробників і споживачів, також має велике значення. Нестабільність у виробничих регіонах, можливі воєнні конфлікти або санкції можуть призвести до змін в постачанні металів і вплинути на їхні ціни. Зміни у торгових угодах між країнами можуть мати великий вплив на експорт та імпорт дорогоцінних металів. Зміни в рівні інфляції або відсоткових ставках також можуть впливати на ціни металів.

- Технічний аналіз.

Цей метод полягає у вивченні та аналізі графіків цін, обсягів торгівлі, патернів поведінки цін та інших технічних показників. Технічний аналіз може допомогти виявити тренди, підтримку та опір на ринку, що може бути використано для прогнозування майбутніх цін на метали.

Аналіз графіків допомагає виявити тренди цін, патерни поведінки цін, рівні підтримки та опору на ринку. Використання індикаторів, таких як ковзне середнє, стохастичний осцилятор, MACD тощо, допомагає прогнозувати можливі рухи цін на метали.

- Модель ARIMA (Autoregressive Integrated Moving Average).

Це статистична модель, яка використовується для аналізу часових рядів. ARIMA дозволяє прогнозувати майбутні значення на основі попередніх даних. Застосовується до історичних даних цін на дорогоцінні метали для прогнозування їх майбутнього руху.

ARIMA використовує попередні значення часових рядів для прогнозування майбутніх значень. Він може враховувати тренди, сезонність та інші закономірності в даних для прогнозування цін.

- Моделювання з використанням машинного навчання.

Методи машинного навчання, такі як нейронні мережі, випадкові ліси, градієнтний бустінг і інші, можуть бути використані для аналізу великої кількості даних і прогнозування цінових рухів [4].

Ці методи машинного навчання можуть аналізувати великі обсяги даних та виявляти складні зв'язки між різними факторами для прогнозування цін на метали.

- Аналіз геополітичних та макроекономічних подій.

Події на світовій арені, такі як конфлікти, зміни режимів, торгові угоди, а також макроекономічні показники, такі як зміни у ВВП, безробіття, можуть мати великий вплив на ціни дорогоцінних металів.

Події на світовій арені, такі як військові конфлікти або зміни у владних режимах, можуть призвести до нестабільності на ринку дорогоцінних металів. Зміни у макроекономічних показниках, таких як ВВП, безробіття, інфляція, можуть впливати на попит і пропозицію на ринку металів [5].

Ці методи можуть використовуватися окремо або комбінуватися для отримання більш точного прогнозу цін на дорогоцінні метали. Жоден метод не може гарантувати точний прогноз цін на дорогоцінні метали. Зазвичай, найефективніше використовувати комбінацію кількох підходів для отримання більш точного прогнозу.

### 3. ОГЛЯД МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ЦІН НА ЗОЛОТО

Методи машинного навчання можуть бути потужним інструментом для прогнозування цін на золото, використовуючи різні моделі та алгоритми. Вони можуть описувати складну структуру взаємозв'язків між факторами, що описують закон формування цін на золото. Ось деякі з них:

- Нейронні мережі.

Це можуть бути різні типи нейронних мереж, такі як рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) або глибокі нейронні мережі (DNN) [6]. Вони можуть аналізувати складність та неоднорідність даних для прогнозування цін на золото.

- Випадкові ліси.

Випадкові ліси – це тип алгоритму, що базується на деревах рішень. Вони використовуються для аналізу великих обсягів даних та можуть добре прогнозувати цінові рухи на основі різних факторів [7].

- Градієнтний бустінг.

Градієнтний бустінг – це метод, що поєднує декілька слабких моделей для створення потужної ансамблевої моделі. Він може виявляти складні зв'язки між різними факторами та прогнозувати ціни на золото з високою точністю.

- Метод опорних векторів (SVM).

Цей метод може бути застосований до прогнозування цін на золото, шукаючи оптимальну гіперплощину для розділення даних на категорії та прогнозування майбутніх значень.

- Лінійна регресія.

Лінійна регресія – це один з найпростіших методів, який моделює залежність між залежною та незалежними змінними [8]. Він може бути використаний для простого прогнозування цін на золото на основі історичних даних.

Ці методи машинного навчання використовують різні підходи та алгоритми для аналізу великих обсягів даних про ціни на золото, враховуючи різноманітні фактори, які можуть впливати на ціни. Точність прогнозів може залежати від якості даних, використаної моделі та правильно підібраних параметрів.

### 4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для даного дослідження було використано дані про ціни на золото за період з 18 листопада 2011 року до 1 січня 2019 року. Дані було зібрано з різних джерел. В датасеті містяться 1718 рядків даних та 80 колонок характеристик. До даних входить інформація про ціни на нафту та інші дорогоцінні метали, ставки облігацій США, обмінні курси євро і долара США та індекси Standart and Poor's і Доу-Джонса.

Спершу дані оброблялись, перевірялись і аналізувались з допомогою статистичних показників і методів. Було розраховано додаткові технічні індикатори для покращення прогнозування цін на золото: MACD, RSI, SMA, UpperBand, LowerBand, DIFF, Open-Close, High-Low. Також дані було нормалізовано для роботи з методами машинного навчання.

В роботі використовувались наступні методи машинного навчання: метод Бенчмарк, SVR, Random Forest, Lasso, Ridge, Bayesian Ridge, Gradient Boosting, Stochastic Gradient Descent, LGBM Regressor with Repeated stratified K fold, XGBRegressor, GRU NN, LSTM NN. Впродовж дослідження проводився відбір показників для покращення методів машинного навчання. На наступних Рисунках 1–2 представлено графіки прогнозів на рік вказаними вище методами.

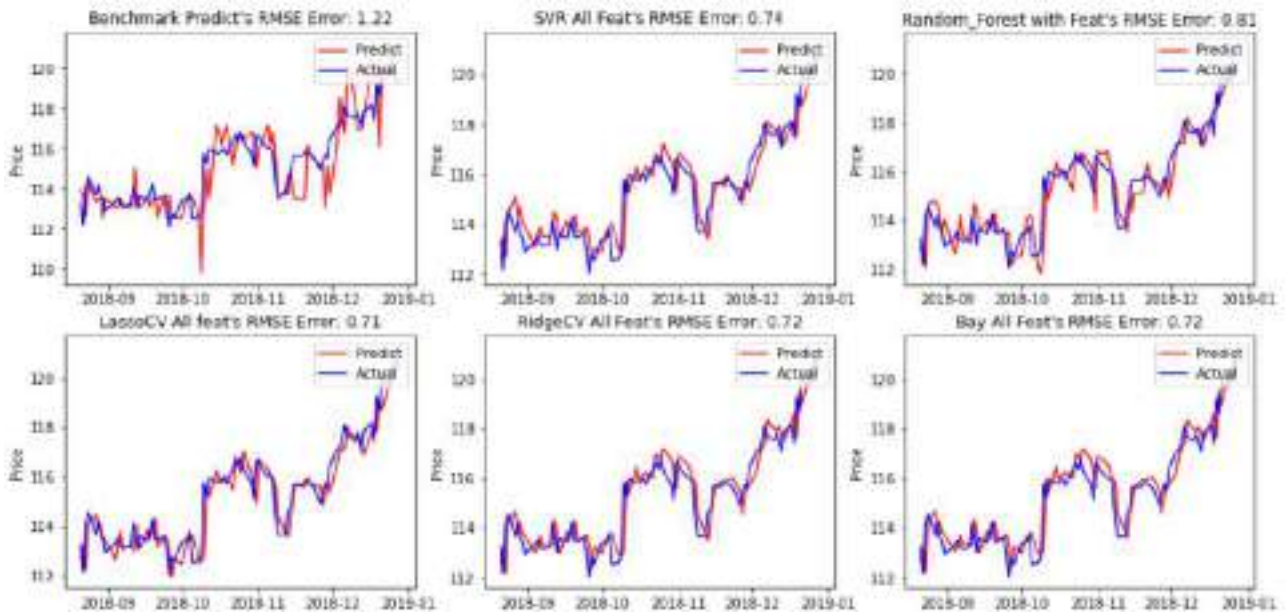


Рисунок 1. Порівняльні графіки прогнозування моделей в порівнянні з реальними даними

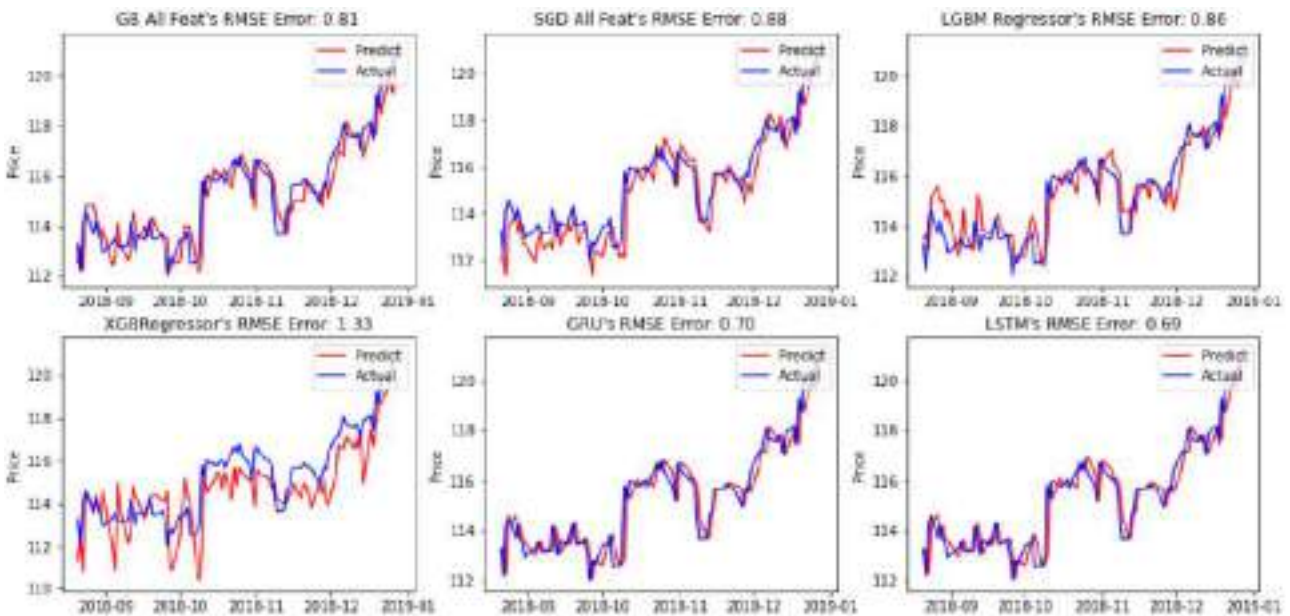


Рисунок 2. Порівняльні графіки прогнозування моделей в порівнянні з реальними даними

На Рисунку 3 представлено графік RMSE для оцінки найкращої моделі для методів навчених на всіх показниках і з відібраними параметрами.

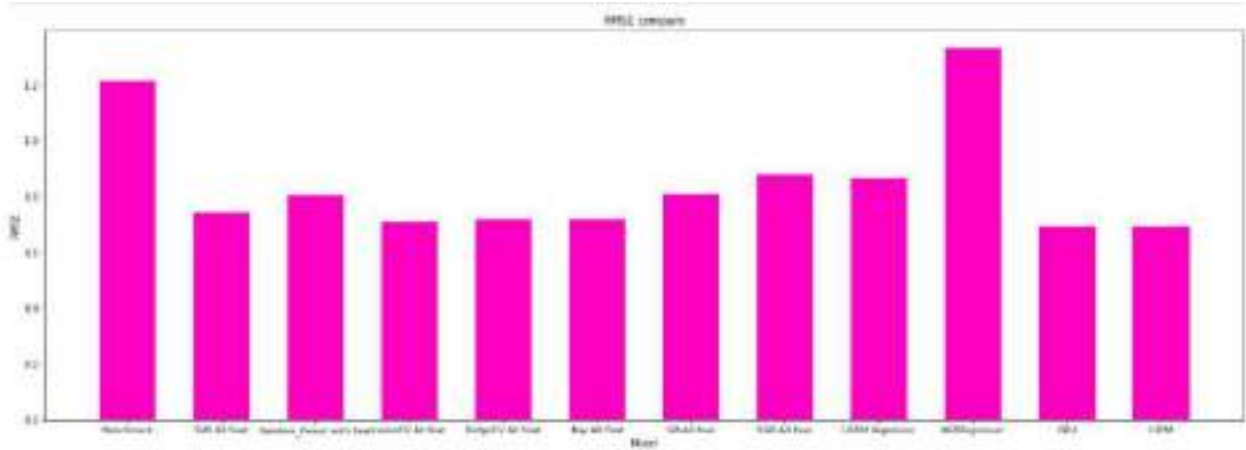


Рисунок 3. Гістограма RMSE для всіх методів

Відповідно до Рисунок 3, найкраще показали себе моделі навчені на відібраних параметрах. Для уточнення роботи моделей оцінимо усереднені похибки кожного методу прогнавши їх 50 разів. В наступній Таблиці 1 представлено результати дослідження.

Таблиця 1. Порівняльна таблиця результатів кращих моделей

Модель	RMSE	R2
Benchmark Model	1,2166	0,6590
LSVR	0,8136	0,8474
Random Forest	0,8078	0,8496
Lasso	0,7117	0,8833
Ridge	0,7186	0,8810
Bayesian Ridge	0,7195	0,8807
Gradient Boosting	0,8094	0,8490
Stochastic Gradient Descent	0,8763	0,8231
LGBM Regressor with Repeated stratified K fold	0,8643	0,8279
XGBRegressor	1,3319	0,5913
LSTM NN	0,6942	0,8889
GRU NN	0,6955	0,8885

Отже, відповідно до результатів представлених в Таблиці 1 найкращими виявились моделі LSTM NN. Але варто зазначити, що модель GRU NN є співставною до LSTM NN, і різниця RMSE між ними лежить у межах статистичної похибки.

## 5. ВИСНОВКИ

В даному дослідженні було розглянуто математичні методи і моделі машинного навчання для прогнозування цін на золото. В процесі дослідження було розглянуто багато напрямків роботи над поставленою задачею і відібрано основний, на основі якого і проводились експерименти. Проведено аналіз даних і обробка даних перед використанням методів машинного навчання для максимізації точності методів.

В результаті дослідження було обрано модель, яка найкраще описує закон ціноутворення на золото. Модель створена на базі методу LSTM NN, яка дала найкращі результати.

### ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Диба М.І., Бахтарі Е.А. Еволюція золота: історія і сучасність : Вісник НБУ Жовтень 2008. С. 20–28.
2. Морозов А.С. Еволюція економічної ролі золота. Економіка та держава. – №12. – 2010. С. 53 – 57.
3. Johnston J., DiNardo J. Econometric methods. New York: McGraw-Hill, Inc., 1997. 530 p.
4. Бідюк П.І. Методи прогнозування. Луганськ: Альма Матер, 2008. С. 605.
5. Ксенжук, О. С. Тенденції функціонування світового ринку дорогоцінних металів в умовах глобальної фінансової нестабільності. Економічний аналіз: зб. наук. праць. Тернопільський національний економічний університет. 2017. Т. 27. № 4. – С. 289-298
6. П. Бідюк, Є. Гуць, В. Гавриленко, Н. Рудоман. Прогнозування цін акцій з використанням рекурентної нейронної мережі lstm / Системи управління, навігації та зв'язку. Збірник наукових праць. – Полтава: ПНТУ, 2021. – Т. 3 (65). – С. 64-68. – doi:<https://doi.org/10.26906/SUNZ.2021.3.064>.
7. Leo Breiman. Random Forests. Berkeley: Statistics Department, University of California, 2001. 33 p.
8. Altman, N. and Krzywinski, M. (2015). Simple linear regression. Nature Methods, 2(11), 999–1000 ст. [Електронний ресурс]. – Режим доступу : <https://www.nature.com/articles/nmeth.3627>

# ПОРІВНЯЛЬНИЙ АНАЛІЗ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДІВ РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРО МАКСИМАЛЬНИЙ ПОТІК У МЕРЕЖАХ

Боднар М.С.<sup>1</sup>, Статкевич В.М.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Навчально-науковий інститут прикладного системного аналізу, кафедра математичних методів системного аналізу, Київ, Україна

<sup>1</sup> poit25@ukr.net, <sup>2</sup> mstatkevich@yahoo.com [0000-0001-5210-9890]

**Розглянуто задачу пошуку максимального потоку в мережах. Для алгоритмів Форда-Фалкерсона, Едмондса-Карпа та просування передпотуку наведено асимптотичну складність, виконано порівняльний аналіз за часом залежно від розміру та характеристик мережі. У результаті, алгоритм Форда-Фалкерсона зазвичай є найповільнішим, а просування передпотуку – найшвидшим, проте для розрідженого графа з великою кількістю вершин алгоритм Едмондса-Карпа може бути швидшим. Розроблено програмне забезпечення, яке реалізує вказані алгоритми та графічно відображає процес розв'язку, проведено експерименти на різних множинах тестових мереж.**

**Ключові слова:** максимальний потік, задача про максимальний потік, алгоритм Форда-Фалкерсона, алгоритм Едмондса-Карпа, алгоритм просування передпотуку.

## 1. ВСТУП

Необхідність оптимізації поточкових мереж стає все більш актуальною в різних сферах, включаючи транспортні системи, телекомунікації та комп'ютерні мережі [1]. Ефективне управління потоками ресурсів, інформації або товарів є життєво важливим для підвищення загальної продуктивності системи, мінімізації вузьких місць і забезпечення оптимального використання ресурсів [2].

Це дослідження має на меті зробити внесок в існуючу базу знань шляхом аналізу, порівняння та реалізації трьох відомих алгоритмів для вирішення задач про максимальний потік: алгоритмів Форда-Фалкерсона, Едмондса-Карпа та алгоритму "просування передпотуку" (англ. *Push-Relabel*).

Задача про максимальний потік (англ. *Maximum Flow Problem*) є однією з фундаментальних проблем в теорії оптимізації, яка має багато практичних застосувань у різних галузях, таких як проектування мереж, транспорт, логістика, планування та розподіл ресурсів. Наприклад, ця задача може використовуватися для моделювання потоків транспорту в дорожніх мережах, потоків даних в комунікаційних мережах, мережевій маршрутизації, потоків товарів у ланцюгах постачання, потоків крові в кровоносній системі, потоків електрики в електричних мережах та інших [3].

## 2. ЗАДАЧА ПРО МАКСИМАЛЬНИЙ ПОТІК

Задача про максимальний потік може бути сформульована так: дано орієнтований граф (також називається мережею) з джерелом та стоком, де кожне ребро має додатну пропускну здатність, яка визначає максимальну кількість потоку, що може проходити через нього, знайти

спосіб надсилання якомога більшого потоку від джерела до стоку з урахуванням обмежень на пропускну здатність. Потік на кожному ребрі повинен бути невід'ємним і задовольняти умови збереження потоку: загальний потік, що входить у вузол (крім джерела та стоку), повинен бути рівним загальному потоку, що виходить з цього вузла (див. рис. 1). Більш формально, **потік** (англ. *flow*) у  $G$  – це двовимірна дійснозначна функція на декартовому квадраті множини вершин  $f: V \times V \rightarrow \mathbb{R}$ , яка визначає кількість потоку, що проходить по кожному ребру. Така функція задовольняє наступним умовам:

- 1) **обмеження пропускну здатності** (*capacity constraint*):

$$\forall u, v \in V: f(u, v) \leq c(u, v),$$

- 2) **антисиметричність** (*skew symmetry*):

$$\forall u, v \in V: f(u, v) = -f(v, u),$$

- 3) **збереження потоку** (*flow conservation*):

$$\forall u \in V \setminus \{s, t\}: \sum_{v \in V} f(u, v) = 0.$$

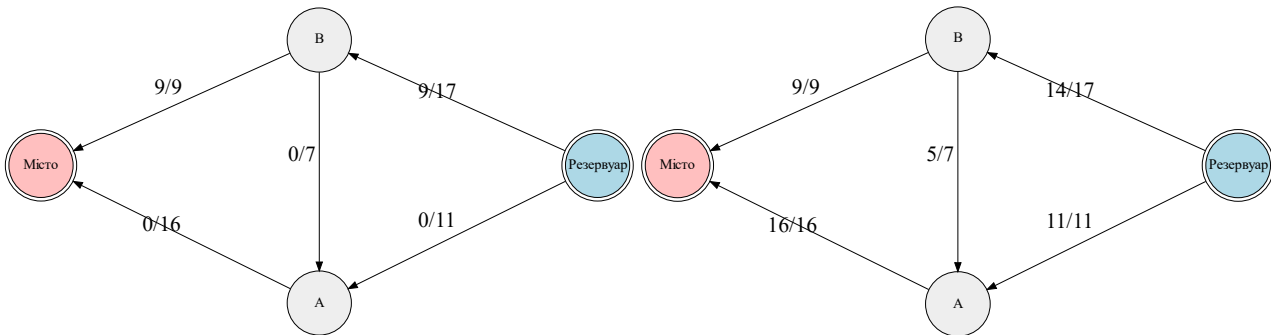


Рисунок 1. Зліва – можливий потік по мережі, справа – максимальний

Можемо збільшити потік, знайшовши збільшувачий шлях – це шлях від джерела до стоку, що має деяку невикористану пропускну здатність на своїх ребрах. Збільшувачий шлях також називають доповнюючим шляхом або шляхом, що збільшує (англ. *augmenting path*). Теорема про максимальний потік та мінімальний розріз стверджує, що максимальна величина потоку від  $s$  до  $t$  дорівнює мінімальній пропускну здатності розрізу від  $s$  до  $t$  у мережі.

Задача про максимальний потік була вперше сформульована у 1954 році Т. Е. Гаррісом та Ф. С. Россом як спрощена модель радянського залізничного транспортного потоку [4]. У 1956 році Лестер Рендольф Форд-молодший та Делберт Реймонд Фалкерсон створили перший відомий алгоритм – алгоритм Форда-Фалкерсона [5], який базується на пошуку збільшувачих шляхів.

З тих пір було розроблено різні покращені алгоритми для вирішення задачі максимального потоку, такі як алгоритм Едмондса-Карпа, алгоритм "просування передпотіку" (Push-Relabel), алгоритм Голдберга-Рао та інші [3]. Ці алгоритми мають різну складність виконання та характеристики продуктивності залежно від структури та розміру мережі. Ці алгоритми використовують різні техніки для пошуку збільшувачих шляхів, різні структури даних та методи для відстеження залишкової мережі.

**Залишкова мережа**  $G_f$  – це граф, який показує можливий додатковий потік у мережі  $G$ . Якщо є шлях із джерела в стік у залишковій мережі, то можна збільшити потік у вихідній мережі. Кожне ребро залишкової мережі має значення, зване залишковою пропускну здатністю, яка дорівнює різниці вихідної пропускну здатності ребра та поточного потоку по ньому. Залишкова пропускну здатність – це, по суті, поточна невикористана пропускну здатність ребра.

## 2.1. Алгоритм Форда-Фалкерсона

Це один із найкласичніших і найпростіших алгоритмів розв'язання задачі про максимальний потік. Основна ідея алгоритму полягає в наступному. На кожній ітерації алгоритму знаходимо збільшуючий шлях із джерела до стоку у залишковій мережі, знаходимо ребро з найменшим значенням залишкової пропускної здатності та за допомогою цього значення збільшуємо потік уздовж цього шляху. Потім шукаємо інший збільшуючий шлях, і так далі (ітеративно), доки в отриманій залишковій мережі не буде доступного збільшувального шляху. Алгоритм завершується, коли в залишковій мережі більше не існує збільшувальних шляхів. Складність за часом (або асимптотична складність) даного алгоритму залежить від способу вибору збільшувальних шляхів, і може бути як поліноміальною, так і експоненціальною щодо розміру вхідних даних.

**Приклад роботи.** Розглянемо наступний граф  $G$  з вершинами  $V = \{s, a, b, c, d, t\}$ , де  $s$  – джерело,  $t$  – стік. Ребра та їх пропускні здатності зображені на рис. 2.

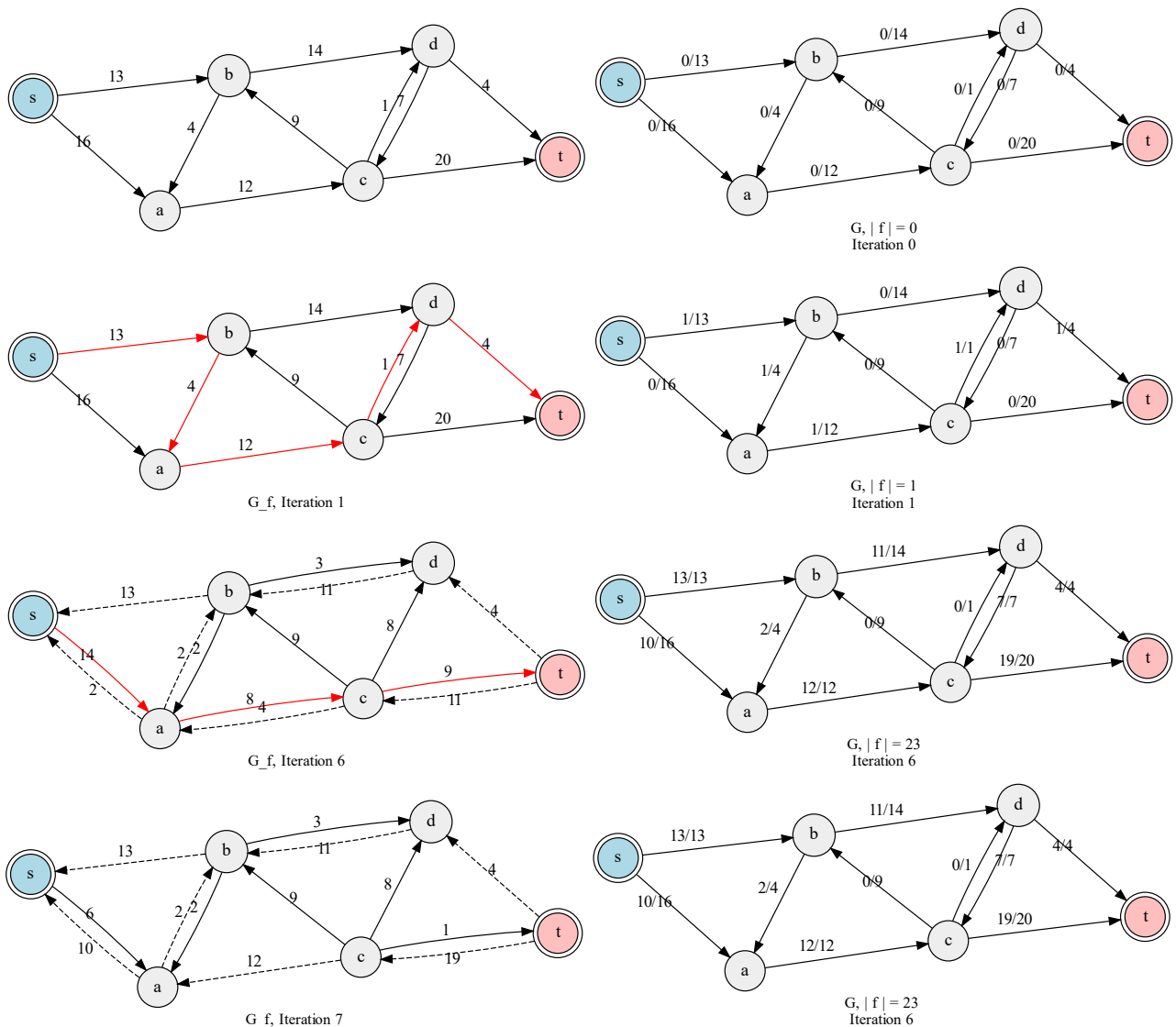


Рисунок 2. Мережа  $G$ , ітерації 0, 1, 6 та 7 (ітерації 2–5 випущено)

## 2.2. Алгоритм Едмондса-Карпа

Це версія алгоритма Форда-Фалкерсона, яку модифікували Джек Едмондс та Річард Меннінг Карп у своїй роботі 1972 року [6], вона гарантує поліноміальну складність за часом алгоритма. Відмінність полягає в тому, що для вибору збільшуючих шляхів використовується алгоритм пошуку в ширину (англ. *Breadth First Search* або скорочено *BFS*), який знаходить найкоротший за кількістю ребер збільшуючий шлях у залишковій мережі (вважаючи, що кожне ребро має одиничну довжину). Нагадаємо, що пошук у ширину – це один із найдавніших і найвідоміших алгоритмів обходу графа, який починається із заданої вершини та відвідує всі вершини, які можна досягти з неї, у порядку зростання відстані до неї. Складність алгоритму Едмондса-Карпа за часом становить  $O(|V||E|^2)$ , де  $|V|$  – кількість вершин, а  $|E|$  – кількість ребер у вихідному графі.

**Приклад роботи.** Розглянемо приклад роботи алгоритму Едмондса-Карпа в транспортній мережі, яка була розглянута на рис. 2. Наступні кроки аналогічні крокам, які наведені на рис. 2, лише за відмінністю, що треба шукати **найкоротший** збільшуючий шлях на кожній ітерації (див. рис. 3).

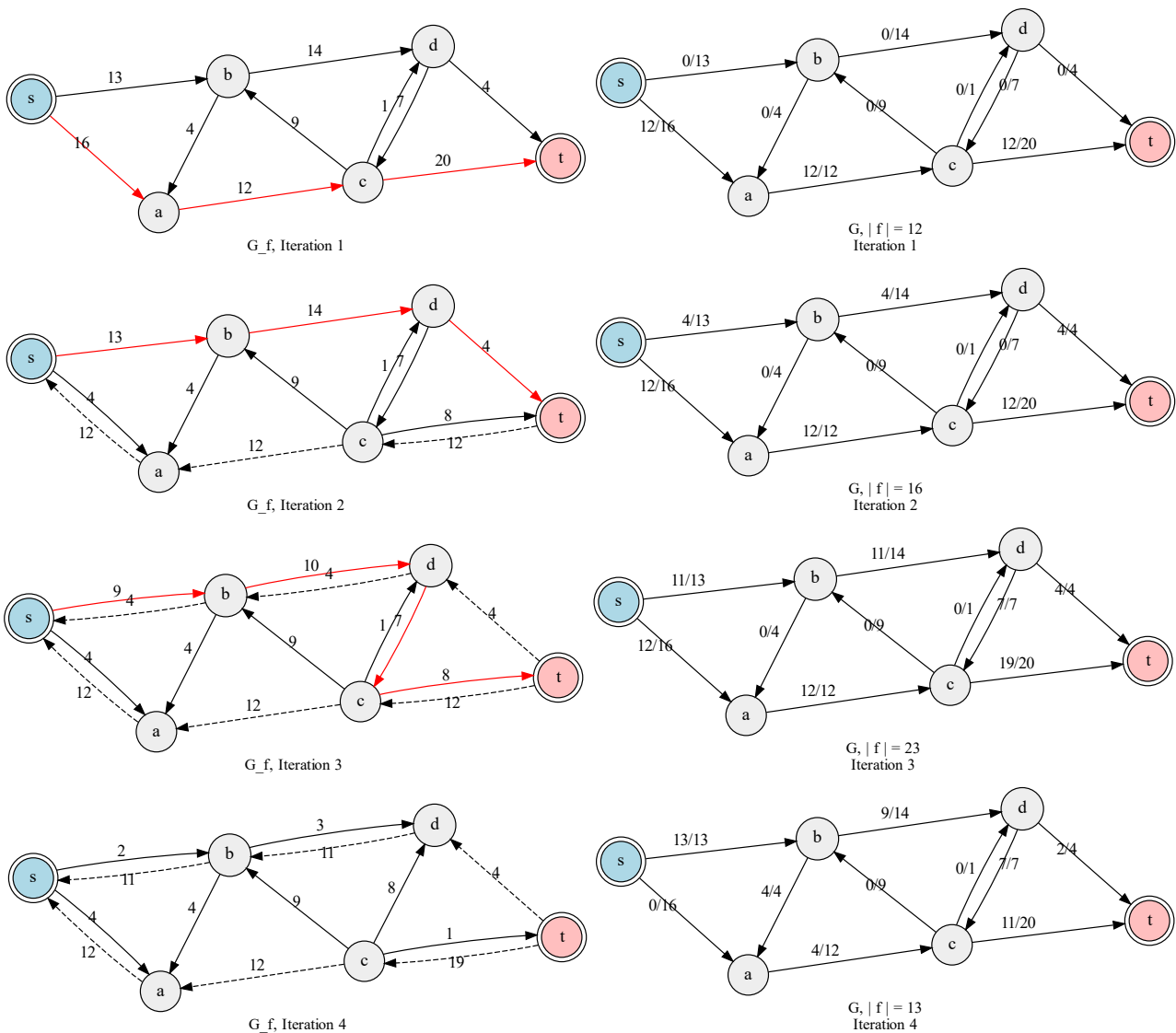


Рисунок 3. Мережа G, ітерації 0–4

Як можна побачити, алгоритмом Едмондса-Карпа максимальний потік знайдений швидше, ніж алгоритмом Форда-Фалкерсона, якому знадобилось 7 ітерацій. Тобто краща ефективність алгоритма Едмондса-Карпа у порівнянні з алгоритмом Фордом-Фалкерсоном помітна неозброєним оком. Це все за рахунок зміни вибору збільшуючого шляху.

### 2.3. Алгоритм просування передпоток

*Алгоритм просування передпоток* (або «алгоритм підняття передпоток», англ. *Push-Relabel*), розроблений Ендрю Голдбергом і Робертом Тар'яном у 1986–1988 роках [7], представляє собою важливий прорив у теорії потоків у графах. Автори ввели поняття передпоток і запропонували дві основні операції: *просування* і *підняття*, які дають змогу поступово перетворювати передпотік на максимальний потік. Вони також довели, що цей алгоритм має сильно поліноміальну складність  $O(|V|^2|E|)$ , що є кращою за складність алгоритму Форда-Фалкерсона та його модифікацій (Едмондс-Карп, Дініц). Вони використовували ідею міток висоти у кожній вершині для управління потоком. Це один з найефективніших алгоритмів розв'язання задачі про максимальний потік.

На відміну від класичного методу Форда-Фалкерсона, де аналізують усю залишкову мережу для пошуку шляху, що збільшує потік, алгоритми просування передпоток діють більш локально. Тобто опрацювання вершин відбувається по одній, розглядаючи тільки їх сусідів у залишковій мережі. Це дає змогу скоротити обсяг обчислень і прискорити процес знаходження максимального потоку. У кожній ітерації алгоритму вибирають вершину з *надлишковим потоком* і просовують зайвий потік у її сусідні вершини. Таким чином, алгоритми просування передпоток працюють поетапно, збільшуючи потік від джерела до стоку, доки не досягнуто межі пропускної здатності або не буде досягнуто максимального потоку.

**Надлишковий потік** (англ. *excess flow*) представляє собою кількість потоку, яка перевищує пропускну здатність вихідної вершини в мережі, тобто це означає, що алгоритм просування передпоток не забезпечує збереження потоку під час свого виконання, на відміну від методу Форда-Фалкерсона. Натомість він підтримує **передпотік** (англ. *preflow*), який являє собою функцію  $f: V \times V \rightarrow \mathbb{R}$ , яка схожа на потік, за виключенням того, що замість звичайної умови збереження потоку вона задовольняє **послаблену умову збереження потоку**:  $f(V, u) \geq 0$  для всіх вершин  $u \in V \setminus \{s\}$ . На кожному кроці алгоритма, коли відбувається просування надлишкового потоку, функція передпоток оновлюється відповідно до перерозподілу потоку.

Також вводиться поняття **висоти вершини**, що представляє собою функцію  $h: V \rightarrow \mathbb{N} \cup \{0\}$ , причому  $h(s) = |V|$ ,  $h(t) = 0$  та  $h(u) \leq h(v) + 1$  для  $\forall (u, v) \in E_f$ . Тут варто зазначити, що  $f$  є саме передпоток у  $G$ , а не просто потоком, як у попередніх алгоритмах.

Алгоритм заснований на двох основних операціях: *просуванні* та *піднятті*.

**Просування** (англ. *push*) – це операція, під час якої вершина  $u$  пересилає частину свого надлишкового потоку (тобто різниці між вхідним і вихідним потоком) по одному з ребер до сусідньої вершини  $v$ . Просування можливе лише за наступних умов:

- 1) вершина  $u$  має бути переповненою, тобто  $e(u) > 0$ ;
- 2) ребро  $(u, v)$  має належати залишковій мережі,  $(u, v) \in G_f$ ;
- 3)  $h(u) = h(v) + 1$ .

**Підняття** (англ. *relabel*) – це операція, за якої для вершини  $u$  збільшуємо висоту до мінімально можливої, щоб зробити хоча б одне ребро допустимим для просування надлишкового потоку в бік стоку  $t$ . Підняття можливе лише за наступних умов:

- 1)  $e(u) > 0$ ;
- 2)  $h(u) \leq h(v)$  для всіх ребер  $(u, v)$  залишкової мережі.

Ці операції повторюються в циклах до досягнення максимального потоку, алгоритм завершується, коли більше немає переповнених вершин або можливих просувань.

Таким чином, алгоритм просування передпотіку працює поступово збільшуючи потік і оптимізуючи його розподіл у мережі. У результаті, він забезпечує ефективне розв'язання задачі максимального потоку, особливо у великих і складних мережах.

Алгоритм працює наступним чином.

1. Ініціалізація передпотіку: встановлюємо висоту витіку, що дорівнює кількості вершин у мережі, а висоту решти вершин рівною нулю. Потім просуваємо максимально можливий потік із витіку по всіх вихідних ребрах. Таким чином, витік стає порожнім, а всі його сусіди – переповненими.

2. Поки існує переповнена вершина (крім стоку), обираємо її і виконуємо одну з операцій: просування або підняття. Якщо у вершини є допустиме ребро для просування, то просуваємо по ньому частину або весь надлишковий потік. Якщо у вершини немає допустимих ребер для просування, то піднімаємо її на мінімально можливу висоту.

3. Коли всі вершини (крім стоку) стають порожніми, алгоритм завершується. Поточний передпотік є максимальним потоком у мережі.

Приклад роботи зображено на рис. 4.

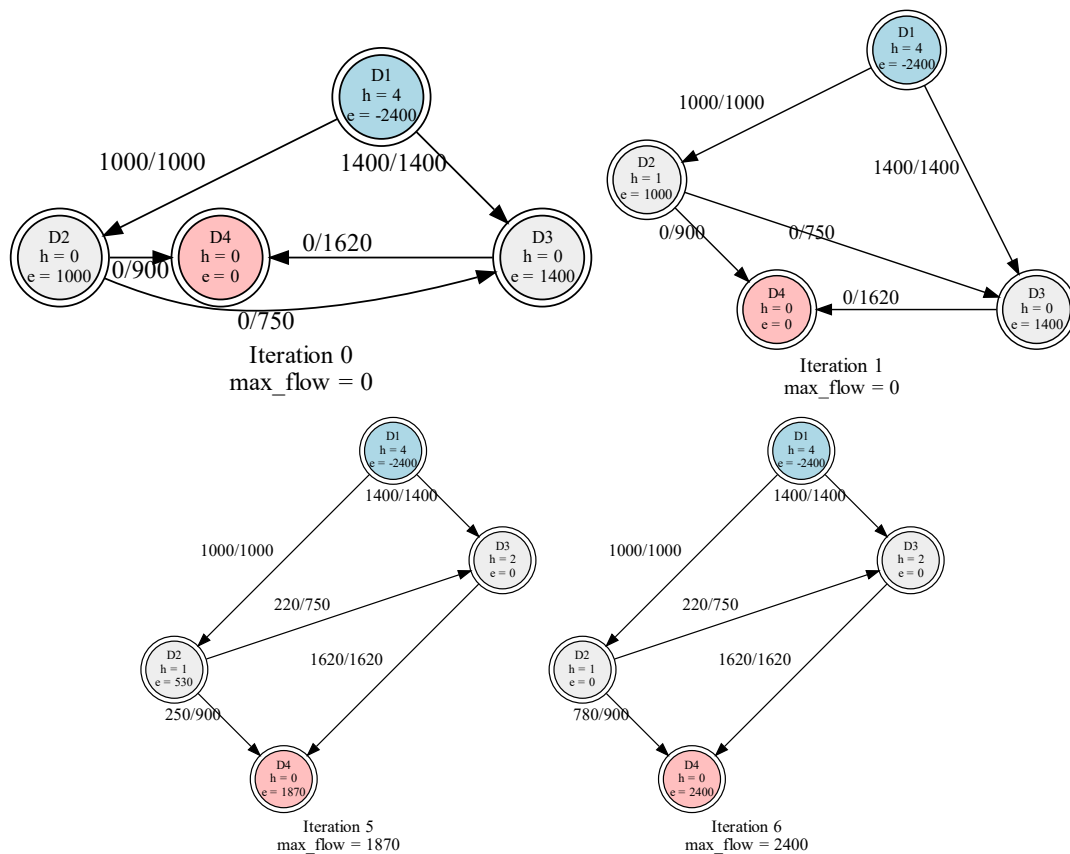


Рисунок 4. Алгоритм просування передпотіку

Алгоритм просування передпотіку може бути застосовано для розв'язання різних завдань, пов'язаних з оптимізацією потоків у мережах, таких як задача про мінімальний вартісний потік, задача про розбиття графа, задача про максимальне паросполучення.

## 2.4. Порівняння алгоритмів

Загалом, алгоритм Форда-Фалкерсона є досить простим, але дієвим алгоритмом для розв'язання задач про максимальний потік, але він має певні обмеження, зокрема, при роботі з великими графами та необхідності ретельної реалізації для забезпечення завершення та

ефективності. Також вагомим недоліком є дуже висока залежність від значень максимального потоку та пропускних здатностей мережі, через свою асимптотичну складність.

Алгоритм Едмондса-Карпа має переваги, такі як краща складність за часом, простота реалізації та гарантоване завершення роботи. Його асимптотична складність у найгіршому випадку є значно кращою за часом, ніж у алгоритму Форда-Фалкерсона, що покращує швидкість виконання. Проте він має деякі недоліки, зокрема, повільну роботу у щільних графах, можливість виконувати неефективну роботу і чутливість до топології графа. При виборі алгоритму слід враховувати специфіку задачі та особливості графа.

Алгоритм просування передпотуку має кілька значних переваг. По-перше, він має кращу асимптотичну складність порівняно з іншими алгоритмами, що дозволяє ефективно вирішувати задачі потоку в графах. Крім того, його локальний підхід і простота розуміння роблять його придатним для паралельних та розподілених реалізацій. Більш того, алгоритм може бути легко модифікований для вирішення різних проблем, пов'язаних з потоком, забезпечуючи гнучкість в його застосуванні.

Однак, варто враховувати деякі недоліки алгоритму. Перш за все, його реалізація може бути складною і вимагати глибокого розуміння теорії графів і структур даних. Зокрема, ефективність алгоритму значно залежить від евристики, що потребує приділяти значну увагу до правила вибору вершини. Нестабільність алгоритму і можливість виконання зайвої роботи є ще двома недоліками, які можуть ускладнити оцінку його продуктивності на реальних даних. Проте, незважаючи на його недоліки, алгоритм просування передпотуку є потужним і ефективним інструментом для вирішення задач потоку в графах.

### 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Усі три алгоритми було програмно реалізовано з візуалізацією їх роботи. Для їх реалізації, була обрана мова програмування Python та середовище програмування Jupyter Lab. Для простішого розуміння роботи алгоритмів було реалізоване графічне представлення. Для цього була використана бібліотека Graphviz. Також Graphviz, на відміну від більш популярного NetworkX, дозволяє дуже просто графічно представляти графи з мультиребрами, у той час коли NetworkX з'єднує декілька таких ребер в одне, що унеможлиблює коректне відображення роботи алгоритмів. Також, у візуалізації роботи алгоритма просування передпотуку висота кожної вершини відповідає висоті розміщення цієї вершини на рисунку, що покращує сприйняття роботи алгоритма.

Для порівняння роботи алгоритмів будемо вимірювати час роботи кожного з них, в залежності від графа. Спочатку порівняємо час роботи в залежності від кількості ребер графу, при фіксованому значенні вершин, у кількості 100 вершин. Пропускна здатність кожного ребра буде випадкова, від 1 до 40.

Так як кількість вершин у нас 100 і граф зв'язний та орієнтований, то максимальна кількість ребер буде становити  $N(N - 1) - 2N = 100 \cdot 99 - 200 = 9700$ . Ми тут віднімаємо  $2N$  через те, що з джерела ребра можуть тільки виходити  $N$  ребер, а у стік тільки надходити.

На рис. 5 зліва показана залежність часу виконання у секундах від кількості ребер у графі, а справа – залежність від кількості вершин. Залежно від щільності графа, алгоритми показують різні результати.

На рис. 5 зліва показано, що в щільному графі найшвидшим є алгоритм просування передпотуку, а найповільнішим – Форда-Фалкерсона. На противагу справа видно, що в розрідженому графі ситуація змінюється: найшвидшим стає алгоритм Едмондса-Карпа, а найповільнішим – просування передпотуку. Це пояснюється квадратичною залежністю асимптотичної складності алгоритма просування передпотуку, яка дорівнює  $O(|V|^2|E|)$ .

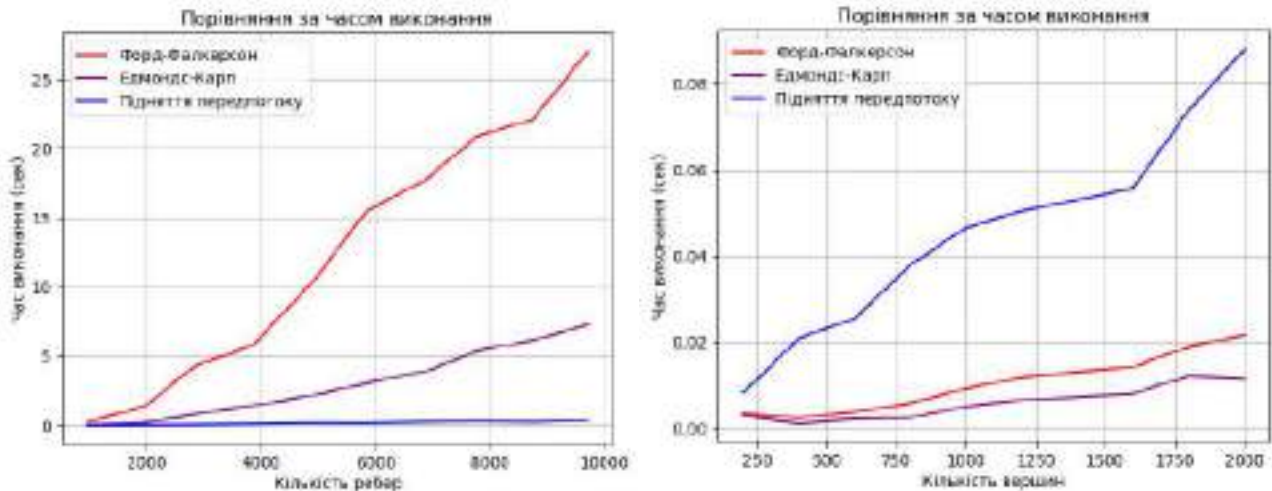


Рисунок 5. Залежність часу роботи від кількості ребер та вершин

Спираючись на результати, можемо зробити висновок, що алгоритм Форда-Фалкерсона завжди програє у швидкості алгоритму Едмондса-Карпа. В свою чергу, алгоритм просування передпотіку у загальному випадку виявився набагато краще за Едмондса-Карпа, за винятком ситуації з розрідженим графом, у якого багато вершин, але такі випадки не дуже розповсюджені та грають велику роль лише в окремих випадках.

#### 4. ВИСНОВКИ

У роботі проведено детальне аналітичне порівняння та продемонстрований результат програмної реалізації задачі про максимальний потік. Досліджені основні теоретичні поняття, визначення та важливі алгоритми, такі як алгоритми Форда-Фалкерсона, Едмондса-Карпа та просування передпотіку. Результати порівняння алгоритмів показали, що найшвидшим серед них є алгоритм просування передпотіку, проте кожен з них має свої переваги та недоліки, і вибір конкретного алгоритму залежить від вхідних умов та вимог ефективності.

#### ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Abd-Alsabour N. The maximum flow problem. *International journal of engineering research and technology*. 2020. Vol. 13, no. 7. P. 1534–1545.
2. Tayyebi J., Deaconu A. Inverse generalized maximum flow problems. *Mathematics*. 2019. Vol. 7, no. 10. P. 899.
3. Introduction to algorithms / T. H. Cormen et al. 3rd ed. Cambridge, MA, USA : The MIT Press, 2009. 1312 p.
4. Harris T. E., Ross F. S. Fundamentals of a method for evaluating rail net capacities. Santa Monica, CA : RAND Corporation, 1955. 63 p.
5. Ford L. R., Fulkerson D. R. Maximal flow through a network. *Canadian journal of mathematics*. 1956. Vol. 8. P. 399–404.
6. Edmonds J., Karp R. M. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*. 1972. Vol. 19, issue 2. P. 248–264.
7. Goldberg A. V., Tarjan R. E. A new approach to the maximum-flow problem. *Journal of the ACM*. 1988. Vol. 35, issue 4. P. 921–940.

# ВИБІР ТА ОЦІНКА ЯКОСТІ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ФІНАНСОВИХ ПОКАЗНИКІВ З УРАХУВАННЯМ ХАРАКТЕРИСТИК ВХІДНИХ ДАНИХ

Бойніцька С.В.<sup>1</sup>, Мілявський Ю.Л.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут  
імені Ігоря Сікорського», Київ, Україна

<sup>1</sup>sonia.boinitska@gmail.com, <sup>2</sup>yuriy.milyavsky@gmail.com

**Складність взаємодії різних факторів, які визначають фінансові показники, вимагає розробки методів прогнозування, здатних враховувати різноманітність вхідних даних. Головний внесок дослідження полягає в створенні універсального програмного продукту, який автоматично аналізує характеристики даних і на їхній основі рекомендує найкращу модель прогнозування. Використання такого підходу значно підвищує якість та швидкість прогнозів, що сприяє ефективному фінансовому плануванню.**

**Ключові слова:** прогнозування, фінансові показники, часові ряди, вибір моделей, характеристики, стратегічне управління.

## 1. ВСТУП

У сучасному світі, де обсяги фінансової інформації неспинно зростають, вирішення завдань прогнозування фінансових показників стає важливим етапом стратегічного управління. З метою вдосконалення цього процесу та надання ефективних інструментів для прийняття рішень, в рамках даного дослідження була створена програма, яка використовує інноваційний підхід до вибору моделей прогнозування фінансових показників.

Метою дослідження є розробка методології для оцінки якості моделей прогнозування фінансових показників, в залежності від характеристик, які притаманні вхідним даним, а також створення програмного продукту, який дозволить автоматизувати цей процес.

Завдання роботи полягають у аналізі теоретичної інформації щодо методів оцінювання моделей прогнозування та їх залежності від вхідних даних, а також у розробці методології для порівняння моделей.

Актуальність дослідження зумовлена швидкими темпами змін у фінансовому середовищі та необхідністю оперативної та точної інформації для ефективного управління ресурсами. Завдяки автоматизованому підходу до вибору та оцінки моделей прогнозування, створений продукт відповідає потребам ринку, дозволяючи зробити процес прогнозування більш ефективним, швидким і надійним.

Наукова новизна дослідження полягає в тому що у ньому вперше була розроблена методологія яка співвідносить характеристики вхідних часових рядів до найбільш доцільних моделей для прогнозування для кожного випадку, а також у створенні програмного забезпечення, універсального для різноманітних вхідних фінансових даних.

Розроблена програма, спираючись на результати теоретичного аналізу, систематично визначає ключові параметри вхідних часових рядів та на основі виділених характерних ознак формує рекомендації щодо вибору найбільш придатної моделі прогнозування.

Однією з ключових переваг розробленого продукту є універсальність підходу до моделювання різноманітних фінансових часових рядів. Програма враховує індивідуальні

особливості кожного датасету, автоматично конфігуруючи параметри моделей, беручи до уваги велику кількість факторів, що можуть впливати на його поведінку.

Це відкриває нові можливості для прогнозування, що робить дане дослідження інноваційним у сфері автоматизації аналізу та прогнозування часових рядів.

Розроблений підхід може знайти застосування у різних галузях, де прогнозування фінансових показників є ключовим елементом стратегічного управління – наприклад, в сферах економіки, інвестицій, торгівлі чи виробництва, де швидкі і точні прогнози є ключовим елементом для досягнення успішних стратегій та оптимізації ресурсів.

## 2. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

В результаті дослідження було створено універсальний алгоритм для вибору найкращої моделі для вхідного набору. Алгоритм наведений на рисунку 1:

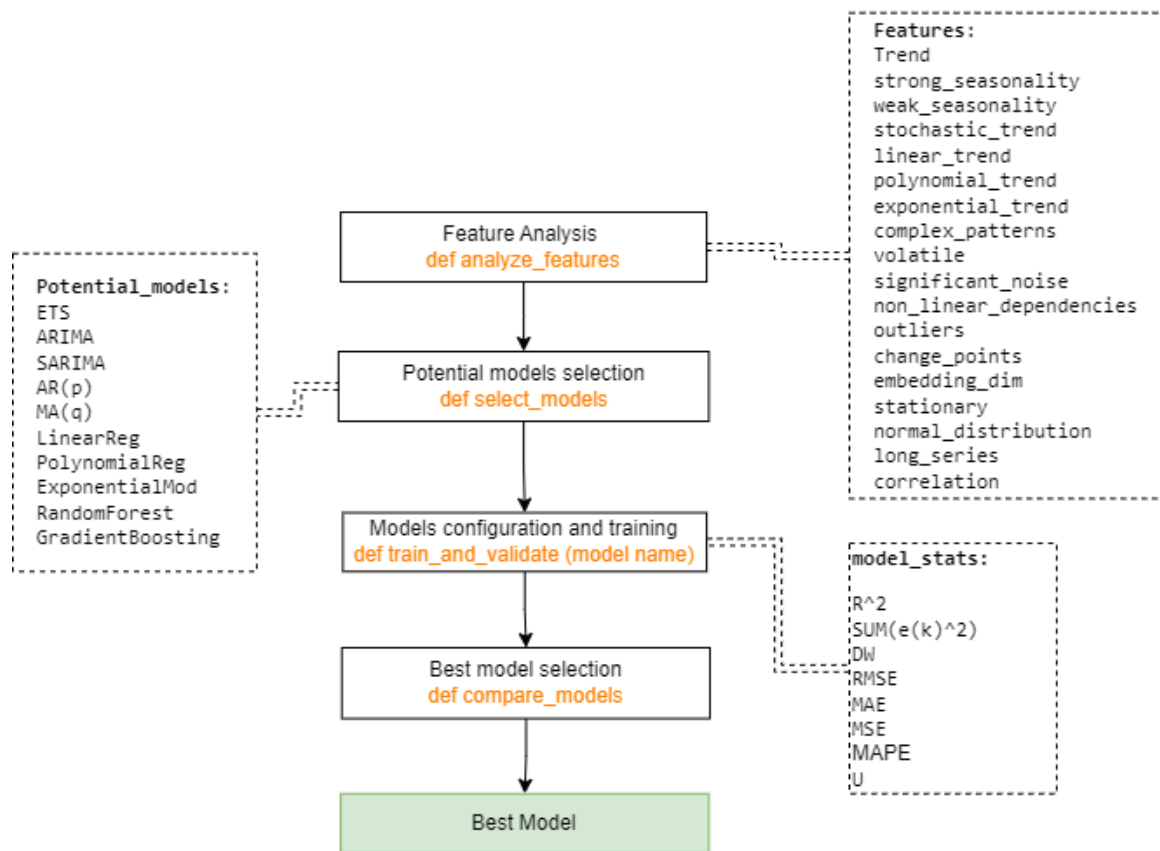


Рисунок 1. Алгоритм створеного програмного продукту

### Етап 1. Аналіз ознак вхідних даних (Feature Analysis)

Можна виділити основний перелік ознак, які впливають на вибір майбутньої моделі та які перевіряються для кожного вхідного датасету: стаціонарність, кореляція, наявність тренду, тип тренду: лінійний, поліноміальний, експоненційний чи стохастичний, сезонність, волатильність, шум та викиди, нелінійні залежності в даних, наявність точок зміни, нормальний розподіл залишків, довжина вибірки

### Етап 2. Вибір потенційних моделей (Potential models selection)

Тепер, коли вже відомі характеристики, які притаманні вхідному ряду, оберемо, які моделі будуть доцільними для прогнозування. Отже, на цьому етапі в залежності від ознак, що були обраховані на етапі 1, обирається набір потенційних моделей, які підходять до даного часового ряду. Кожна модель обирається в залежності від певних умов.

Загальний алгоритм вибору моделей зображено на рис. 2. Тут зеленим позначені моделі, які додаються до масиву «потенційних» при умові наявності відповідної характеристики у вхідних даних, а червоним навпаки позначені моделі, які видаляються.

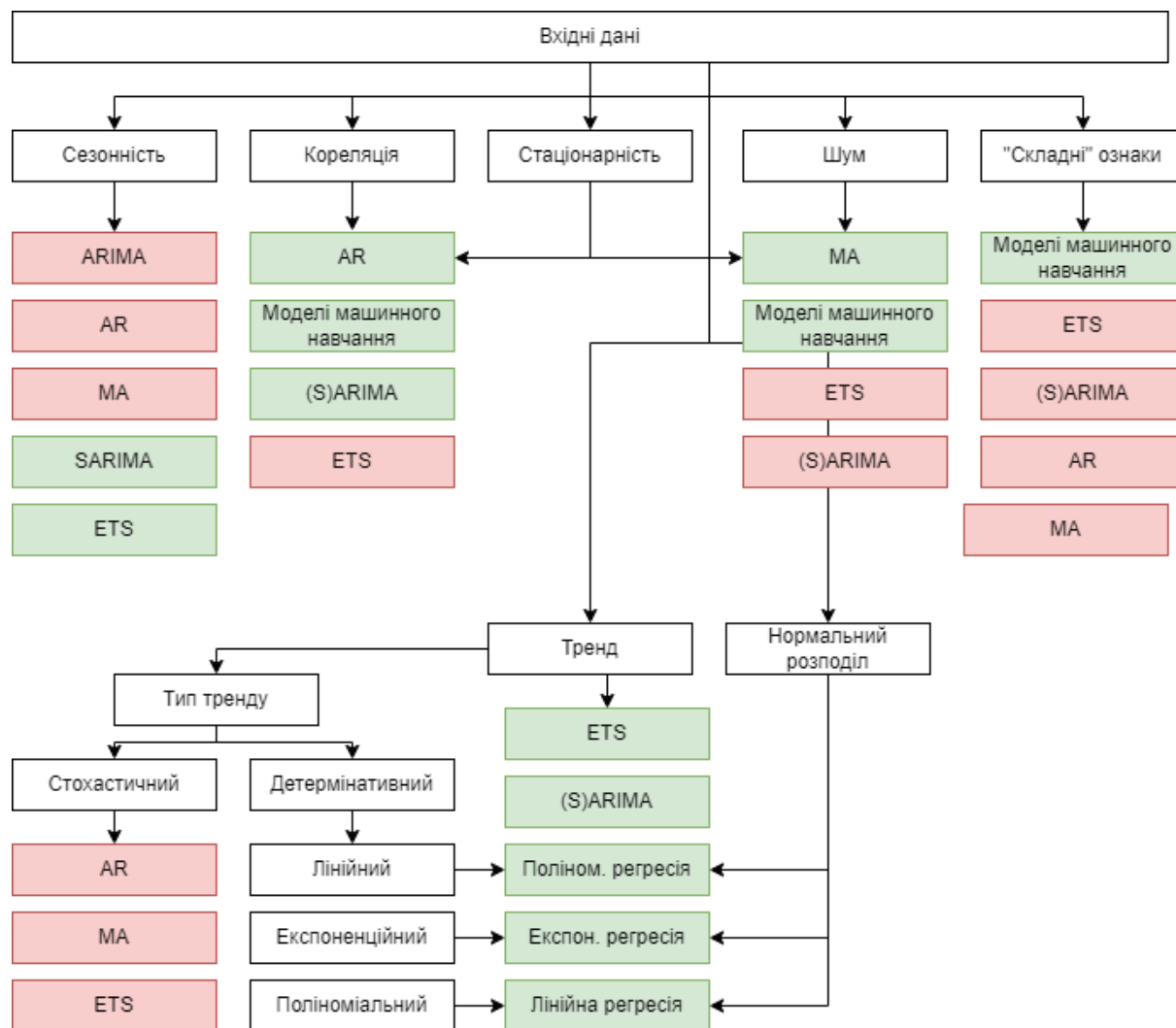


Рисунок 2. Залежність вибору потенційної моделі від наявних ознак

Розглянемо детальніше зв'язки, зображені на рисунку вище.

Умови на моделі Лінійної і Поліноміальної регресії доволі прості, їх пробуємо при наявності відповідного тренду – лінійного чи поліноміального.

Аналогічно Експоненційна модель будується для експоненційного тренду. Сама модель являє собою застосування логарифмічної трансформації для вхідних даних і потім лінійної регресії.

Також для моделей регресії дуже важливим є щоб залишки були розподілені за нормальним розподілом, тож цей параметр також додано до «визначальних» характеристик, необхідних для моделі.

Додавання цих моделей до переліку потенційних необхідно для того, щоб мати певну базу для порівняння роботи більш складних моделей. Також у випадку простих вхідних даних використання такої моделі може бути більш доцільним, ніж налаштування і використання більш складних систем.

Модель машинного навчання Gradient Boosting застосовується у випадках, коли наявні «складні» характеристики вхідних даних і звичайні моделі, як Експоненційне згладжування і ARIMA, можуть бути не доцільними. Ці моделі обираються, коли присутні такі характеристики як волатильність, шум, кореляція, викиди, нелінійні залежності, точки змін.

В той же час при виборі ETS, (S)ARIMA, AR і MA стоїть умова на відсутність цих ознак.

Основним рушієм в виборі моделі AR(p) є присутність кореляції. При виборі MA(q) важливим є присутність шуму. Відсутність сезонності та стохастичного тренду є критерієм для обох моделей. Також обидві моделі вимагають, щоб дані були стаціонарні.

ARIMA поєднує в собі ознаки AR(p) і MA(q), проте завдяки параметру d ще здатна працювати з трендом. А SARIMA здатна працювати з сезонністю. Також розглядається сценарій, за якого ARIMA обирається для нестационарних даних, адже вона може привести їх до стаціонарності за допомогою I компоненти.

ETS добре працює з трендом і сезонністю, проте, на відміну від ARIMA, в ETS стоїть умова на відсутність кореляції. Також ETS погано працює зі стохастичним трендом, при його наявності обирається ARIMA яка має I компоненту для роботи з трендом.

### **Етап 3. Вибір параметрів моделей, їх тренування і валідація (Models configuration and training)**

Після того, як на етапі 2 ми отримали масив потенційних моделей, починається робота з кожною з них. Моделі конфігуруються, навчаються і валідуються. Також на цьому етапі обчислюються метрики якості прогнозів, такі як  $R^2$ ,  $\text{SUM}(e(k)^2)$ , 'DW', 'RMSE', 'MAE', 'MSE', 'MAPE', 'U'.

### **Етап 4. Вибір найкращої з моделей (Best model selection)**

На цьому етапі для кожної моделі, на основі її метрик, обчислюється її score. Score враховує кожен метрику обчислену для моделі зважаючи на її вагу. Після обчислення кількості балів у моделі обирається модель з найнижчим score. Ця модель і вважається найкращою для вхідного набору даних.

## **3. РОБОТА ПРОГРАМИ**

Наведемо приклад роботи програмного продукту. В ролі вхідних даних використаємо акції компанії Google з 1 жовтня 2010 по 1 січня 2017.

Дані зображені на рис. 3.

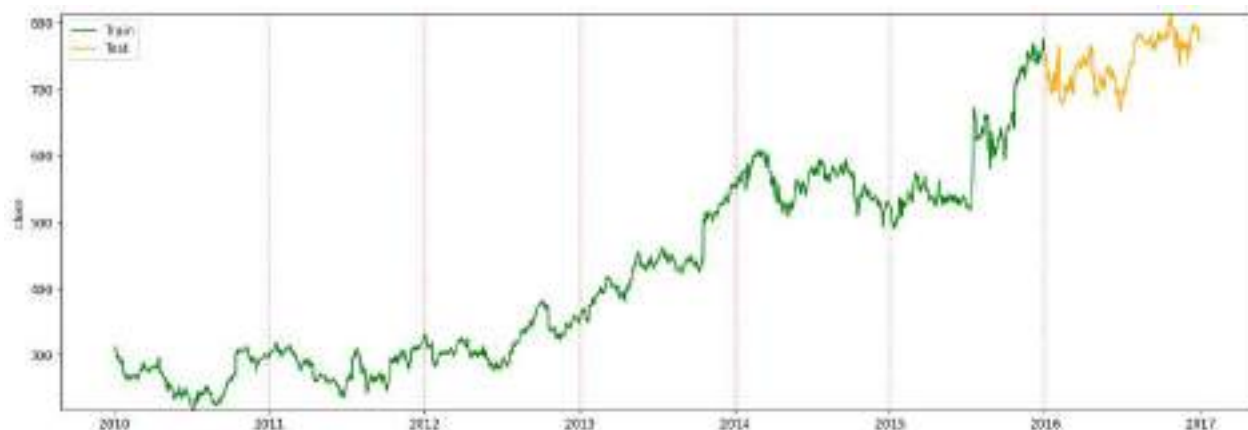


Рисунок 3. Графік часового ряду значення акцій Google

Застосуємо розроблений алгоритм до цих даних. Схема зображена на рисунку 4.

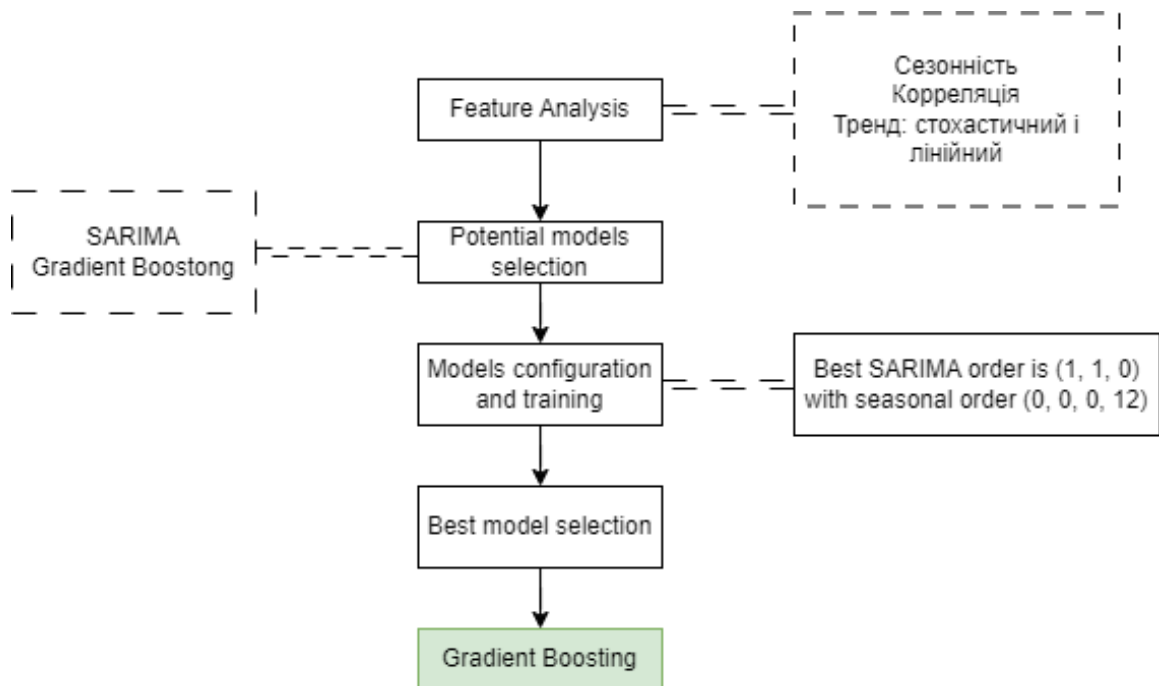


Рисунок 4. Робота алгоритму для даних з прикладу

**Етап 1.** Під час аналізу даних виділено такі характеристики, як наявність сезонності і тренду, а також кореляції в ряді.

**Етап 2.** На основі цих характеристик обрано моделі SARIMA (для роботи з трендом, сезонністю і кореляцією) та Gradient Boosting (для роботи з кореляцією)

**Етап 3.** Під час конфігурації моделей визначено, що найкращим варіантом параметрів SARIMA є порядок  $p = 1$ ,  $d = 1$ ,  $q = 0$ , а також сезонний період  $s = 12$ . Результати тренування моделей зображені на рис. 5 та 6 відповідно.

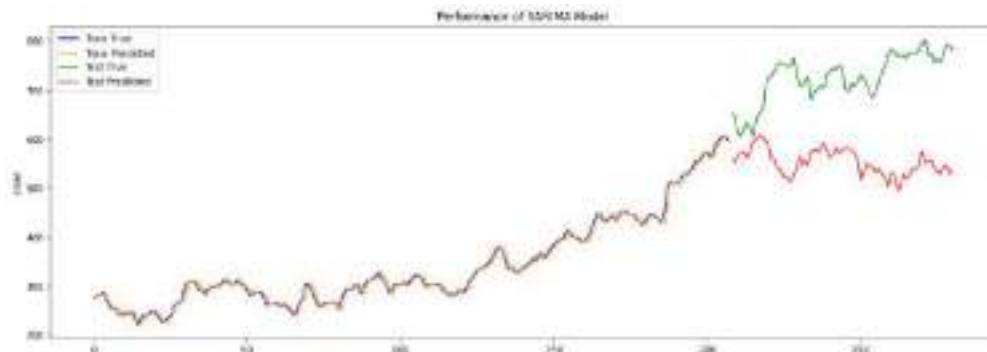


Рисунок 5. Робота моделі SARIMA на даних з прикладу



Рисунок 6. Робота моделі Gradient Boosting на даних з прикладу

Обчислені значення метрик наведені у таблиці 1.

Таблиця 1. Значення метрик для даних з прикладу

model_name	R <sup>2</sup>	SUM(e(k) <sup>2</sup> )	DW	MSE	RMSE	MAE	U
SARIMA(1, 1, 0)x(0, 0, 0, 12)	0,99	18735,55	1,846	34531,106	185,825	173,161	1,576
GradientBoosting	0,97	0,245	0,468	0,016	0,125	0,102	0,00035

**Етап 4.** Як результат, на етапі вибору найкращої моделі обрано модель Gradient Boosting.

## 4. ВИСНОВКИ

В даній роботі досліджено автоматизацію вибору оптимальних моделей прогнозування фінансових показників на основі характеристик вхідних даних. Розроблено програму та логіку вибору моделей, базовану на теоретичному дослідженні. Отримані результати підтверджують ефективність алгоритму та його універсальність для різних фінансових часових рядів. Застосування цього підходу має великий потенціал для автоматизації прогнозування та управління фінансовими ресурсами.

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк П.І., Романенко В.Д., Тимошук О.Л. Аналіз часових рядів (навчальний посібник). Київ: Політехніка, 2010. 317 с.
2. Hyndman, R.J., & Athanasopoulos, G. Forecasting: principles and practice, 3rd edition, 2021 [Електронний ресурс] – Режим доступу до ресурсу: OTexts: Melbourne, Australia. URL: <https://otexts.com/fpp3/>
3. Box, G.E.P., Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
4. Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). Forecasting with exponential smoothing: The state space approach. Springer-Verlag. [Електронний ресурс] – Режим доступу до ресурсу: <https://robjhyndman.com/expsmooth/http://www.exponentialsMOOTHING.net>

# АНАЛІЗ ТЕКСТОВИХ ПОВІДОМЛЕНЬ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Ведмедєв Д.О.<sup>1</sup>, Шаповал Н.В.<sup>2</sup>

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського», ПСА, Київ, Україна

<sup>1</sup> vedmedevdanil@gmail.com, <sup>2</sup> shovgun@gmail.com [0000-0002-8509-6886]

**Ріст кількості текстових даних, зокрема СМС, вимагає розвитку аналізу та кластеризації для виявлення закономірностей і групування повідомлень. Дане дослідження об'єднує в собі використання вбудованих векторних представлень слів (Embedded Word2Vec), метод Mini Batch K-means та метод найбільшої спільної підпоследовності. Мета даного дослідження - розробити ефективний метод кластеризації СМС-повідомлень, яка може бути використана для автоматичного виявлення текстових шаблонів.**

**Ключові слова:** аналіз текстових повідомлень, Embedded Word2Vec, Mini Batch K-means, метод найбільшої спільної підпоследовності, кластеризація, СМС-повідомлення.

## 1. ВСТУП

З кожним днем у сучасному світі із розвитком соціальних мереж, розвиваються і методи комунікації. Одним із таких методів є короткі текстові повідомлення (СМС). Хочу наголосити, що саме короткі, тому що останнім часом у людей розвивається активно “кліпова увага”, і одним із головних атрибутів даного явища, це розсіяний фокус та відсутність довготривалої концентрації на одній події довгий час. Саме це явище безпосередньо впливає на актуальність саме коротких СМС.

Метою даного дослідження виявлення ефективних методів комунікації, а також виявлення спам розсилок серед “живого трафіку”. Тобто результатом роботи буде скрипт, що виконує аналіз та кластеризацію великого обсягу коротких текстових повідомлень за допомогою методів машинного навчання, та оцінка результатів кластеризації.

Ключовими аспектами дослідження є розробка вбудованих векторних представлень слів (Embedded Word2Vec), використання Mini Batch K-means для ефективною кластеризації великого обсягу даних, а також використання методу найбільшої спільної підпоследовності для аналізу схожості текстів.

## 2. АНАЛІЗ ТА КЛАСТЕРИЗАЦІЯ ТЕКСТІВ

Для аналізу вирішення поставленої задачі в першу чергу потрібно звернутись до постановки задачі. В якій сказано: “Із вхідного великого масиву коротких СМС-повідомлень виділити шаблони СМС-ної розсилки, отримавши на виході назву шаблону та кількість екземплярів вхідної вибірки, що до цього шаблону належать”. Отже, отримавши якісне формулювання задачі, можна загальну задачу розбити на декілька підзадач:

1. *Перетворення текстів у числові вектори*
2. *Кластеризація числових векторів*
3. *Оцінка кластеризації*

## 2.1. Перетворення текстів у числові вектори

Ця задача була сформульована таким чином, щоб можна було працювати із текстами, як із векторами ознак. Далі вектори ознак можна подати на вхід до підзадачі Кластеризації числових векторів.

Отже, для перетворення текстів у числові вектори був використаний метод вбудованих представлень слів **embedded Word2Vec**. Embedded Word2Vec – це техніка векторного представлення слів у вигляді числових векторів у просторі з низькою розмірністю. Основна ідея полягає в тому, щоб кожному слову призначити унікальний вектор, таким чином, що подібні слова будуть мати схожі вектори. Використання Embedded Word2Vec має декілька переваг:

- **Семантична схожість**: Слова зі схожим значенням розташовані близько одне до одного у векторному просторі, відображаючи їхню семантичну схожість.
- **Збереження контексту**: Враховує синтаксичні та семантичні відносини між словами, зберігаючи контекст використання слова у реченні.
- **Зменшення розмірності**: Зменшує розмірність векторів, зберігаючи при цьому важливу інформацію, що полегшує обробку та аналіз текстових даних.

Ця техніка є ефективним інструментом для роботи з текстовою інформацією, особливо в завданнях машинного навчання, де слова потрібно представити у вигляді числових векторів для подальшого використання у моделях.

## 2.2. Кластеризація числових векторів

Для кластеризації числових векторів був використаний давно всім відомий K-means, а точніше його модифікація для обробки великих обсягів даних Mini Batch K-means.

**Mini Batch K-means** – це варіант алгоритму K-means для кластеризації даних. Основна відмінність полягає в тому, що **Mini Batch K-means** використовує випадковий підмасив даних (міні-партію) для оновлення центрів кластерів, що робить його ефективнішим у великих наборах даних.

Переваги Mini Batch K-means:

- **Швидкість**: Mini Batch K-means часто працює швидше за класичний K-means, особливо на великих обсягах даних, оскільки використовує лише частину набору даних для кожного оновлення.
- **Ефективність великих даних**: Підходить для роботи з великими обсягами даних, оскільки не вимагає обробки всього набору даних на кожній ітерації.
- **Можливість онлайн-навчання**: Цей метод може використовуватися для онлайн-навчання, де нові дані можуть динамічно додаватися для постійного покращення моделі.
- **Зменшення вимог до пам'яті**: Завдяки використанню міні-партій, вимоги до обсягу пам'яті значно менше, що робить його більш ефективним для обробки великих наборів даних у вигляді потоків.
- **Тенденція до локальних мінімумів**: Може уникнути застрягання в локальних мінімумах через випадковий вибір міні-партій, що може призвести до кращого розв'язку.

## 2.3. Оцінка кластеризації

Після етапу кластеризації, її потрібно оцінити. На жаль емпіричним шляхом було виявлено досить неефективне оцінювання кластеризації за допомогою вже існуючих методів оцінки кластеризації:

Silhouette\_score вимірює, наскільки кожен об'єкт в кластері схожий на інші об'єкти свого кластера порівняно з кластером, до якого він не належить,

Silhouette\_samples – це функція, яка визначає коефіцієнти силуету для кожного окремого елемента в масиві даних. Коефіцієнт силуету вимірює, наскільки об'єкт добре вписується в свій власний кластер в порівнянні з сусідніми кластерами,

Inertia – це сума квадратів відстаней між кожною точкою в кластері та його центром. У контексті методу k-means, inertia визначає, наскільки кластери компактні,

Calinski-Harabasz – також відомий як критерій варіативності внутрішньокластерної суми квадратів, Суть індексу полягає в порівнянні "внутрішньокластерної" дисперсії з "міжкластерною" дисперсією. Ідеальний результат досягається, коли кластери компактні всередині і вони відокремлені один від одного.

Davies-Bouldin index – вимірює, наскільки кожен кластер відокремлений від інших та наскільки компактний у порівнянні з найближчими кластерами.

Отже, використавши підхід прикладного аналізу потрібно було задуматись над евристикою розуміння, що таке є задовільний результат. Задовільним результатом був шаблон тексту, та напроти нього число(кількість екземплярів СМС із таким самим текстом, або схожим). Провівши дослідження даного питання, знайшли метод найбільшої спільної підпоследовності((Longest Common Subsequence, LCS).

LCS – це метод інтуїтивно доволі легкий в розумінні(нагадує добре всім відоме поняття зі школи Найбільшого спільного дільника, НСД), а саме виконує пошук найбільшої спільної підпоследовності символів між двома текстами. Відрізняється від розглянутих метрик саме тим, що працює не з векторами, а саме безпосередньо з текстами що отримали в результаті кластеризації.

Тут слід зазначити, що розуміння "последовності" доволі гнучке, тому що можна текст сприймати як последовність слів, так і последовність символів. Але для нашої задачі більш доречно буде інтерпретувати СМС, як последовність символів, оскільки при написанні в методах комунікації існує людський фактор, через який, одне слово може мати купу різних варіацій написання.

LCS можна використати для порівняння всіх текстів в межах кожного кластеру, але попарне перевірка має складність алгоритму  $O(n^2)$ . На цьому етапі, можна полегшити обчислювальну складність алгоритму, знайшовши центр кожного кластеру. Центром кожного кластеру будемо вважати найближчу точку до геометричного центру кожного кластеру. І далі можна порівнювати кожен кластер із його центром, таким чином зменшили обчислювальну складність алгоритму до  $O(n)$ .

Далі за допомогою методу LCS, можна отримати відносну метрику схожості, яка вираховується наступним чином. Спершу розглядається кластер та його центр, а потім визначається найбільша спільна підпоследовність між ними. Ця довжина вимірюється в кількості символів, що є спільними для обох текстів. Далі, ця довжина спільної підпоследовності нормалізується, поділивши на загальну довжину екземпляру тексту (або відомостей у кластері). Це значення дозволяє отримати відсоток схожості між кластером та його центром. Тому було запропоновано метод LCS в якості метрики для аналізу схожості двох текстів. За допомогою цієї метрики "чистоту" кластерів можна знаходити за середнім кожної вибірки, нижнім квантилем вибірки кожного кластеру. Або ж регулювати також межами ступеня довіри до кожного елемента кластера. Тобто виставити, наприклад умову: "якщо екземпляр даного кластеру схожий на його центр на  $p\%$ , то цей екземпляр належить до цього кластеру", де  $p = \{60, 85, 90\}$ .

Для прикладу, якщо маємо тексти "кластер" і "центр кластера", то найбільша спільна підпоследовність буде "кластер", і після нормалізації відношення довжин, ми отримаємо

високий показник схожості. Тепер результат кластеризації можна оцінити з точністю до того, що саме мається на увазі під словосполученням “схожість двох текстів”.

### **3. ВИСНОВКИ**

В результаті дослідження було отримано доволі ефективний метод кластеризації текстових повідомлень для виявлення шаблонів та групування їх за схожістю. При цьому він є доволі гнучким та модульним, а саме кожна з наведених підзадач, можна вирішити різними методами машинного навчання або інтелектуального аналізу даних, що в свою чергу не закриває питання актуальності вирішення даної задачі, а навпаки спонукає до пошуку більш ефективних рішень, використання не тільки суцільного скрипту, а й його частин по окремої, або в різних конфігураціях і послідовностях, для вирішення інших задач у сфері аналізу текстових повідомлень.

### **4. СУМІЖНІ АКТУАЛЬНІ ЗАДАЧІ**

Насамперед, дане дослідження може бути корисним для вирішення таких задач як:

- Пошук *spin-words* тобто взаємозамінними словами та словосполученнями, що є різними по формі, та однаковими по суті. Наприклад: “Привіт”, “Вітаю”, “Доброго часу доби” – ці всі вирази мають однакову суть, але різну форму.
- Розпізнавання іменованих об’єктів (Named Entity Recognition, NER). Тобто в цій задачі на вхід отримуємо текстове повідомлення, а на виході повинні отримати Ім’я та прізвище, якщо такі є. Наприклад: “Доброго дня! Джоне Джонсонюк, пишу вам щодо нашої зустрічі.”-> “Джоне Джонсонюк”.

# СИСТЕМА ПРОГНОЗУВАННЯ ЕНЕРГОВИТРАТ БУДІВЕЛЬ РІЗНОГО ПРИЗНАЧЕННЯ

Данилов В.Я., Дука О.О.<sup>1</sup>

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

<sup>1</sup> olga.duka575@gmail.com

**Електроенергія є невід’ємною частиною сучасного життя та важливою для економіки усього світу, у тому числі України. Очікується, що в найближчому майбутньому електроенергія замінить інші джерела енергії як основне джерело для використання в будинках, на підприємствах і у транспорті. Для ефективного і правильного прогнозування доцільно створити Систему прогнозування енерговитрат будівель різного призначення. Метою роботи є аналіз існуючих методів прогнозування для покращення його точності. Результатом дослідження є система, що виконує прогнозування енерговитрат із застосуванням методів машинного навчання. У роботі було використано теоретичні та емпіричні методи дослідження.**

**Ключові слова:** машинне навчання, енергоспоживання, прогноз, RMSLE, аналіз енерговитрат.

## 1. ВСТУП

Електроенергія є невід’ємною частиною сучасного життя та важливою для економіки усього світу, у тому числі України. Очікується, що в найближчому майбутньому електроенергія замінить інші джерела енергії як основне джерело для використання в будинках, на підприємствах і у транспорті [1]. Це підкреслює, наскільки важливо правильно прогнозувати споживання електроенергії, оскільки воно має великий вплив на багато операційних і бізнес-операцій. Електроенергія стає головним аспектом нашого повсякденного життя.

Надзвичайний попит на електроенергію прискорюється останнім часом потужним економічним розвитком і швидкою урбанізацією. Прогнозування попиту на електроенергію стає критичним в електричному секторі, оскільки воно служить основою для прийняття важливих рішень у сфері експлуатації та управління енергосистемою. Через спад економічного розвитку, а також відносно помірну температуру в багатьох великих країнах світовий попит на енергію зріс меншими темпами в 2019 році порівняно з останніми роками (+0,7% порівняно із середнім показником 3% на рік у рік період 2000–2018 рр.) [1].

Незважаючи на це, глобальний попит на енергію впав до 2,5% у першому кварталі 2020 року, незважаючи на те, що карантинні заходи в більшості країн тривали лише близько місяця. Зміни в тому, де і як електроенергія використовувалася під час карантину, ще більше змінили структуру попиту на енергію протягом дня в певних районах, причому шаблони будніх днів тепер збігаються з моделями неділі [2]. Це ілюструє відсотковий зсув у попиті та споживанні енергії за різних умов.

За останні десятиліття попит на енергію в будівельному секторі значно зріс через збільшення кількості населення, швидку урбанізацію та соціальний попит. Будівлі зробили значний внесок у світове споживання енергії та викиди парникових газів. Таким чином, будівлі

повинні бути енергоефективними та сталими. Розуміння моделей енергоспоживання в будівлях є корисним для комунальних підприємств, користувачів та керівників об'єктів, оскільки це може допомогти підвищити енергоефективність.

## 2. ПОСТАНОВКА ЗАДАЧІ

Метою роботи є розробка програмного забезпечення системи прогнозування енерговитрат будівель різного призначення. Буде виконано аналіз і обробку набору даних та порівняння обраних методів машинного навчання для прогнозування енергоспоживання будівель. Методи машинного навчання будуть застосовані на наборі реальних даних. За вхідними параметрами необхідно буде визначити обсяг енерговитрат. Об'єктом дослідження є набір реальних даних, що описує параметри будівель, показники лічильників та погодні дані з найближчих метеорологічних станцій.

## 3. ОПИС НАБОРУ ДАНИХ

У задачах машинного навчання дуже важливо виконати правильну попередню обробку даних, для того, щоб результати прогнозування були максимально точними. У даній роботі використаний набір даних, зібраних Американським товариством інженерів з опалення, охолодження та кондиціювання повітря (American Society of Heating, Refrigerating & Air Conditioning Engineers; ASHRAE). Параметрами є характеристики будівель (призначення, площа, рік побудови, поверховість); показники лічильників (тип лічильника і час зняття виміру); погодні дані (температура, хмарність, тиск, напрямок і швидкість вітру).

Під час аналізу і обробки даних створена матриця кореляції для визначення зв'язків між змінними. Матрицю кореляції можна побачити на рис. 1:

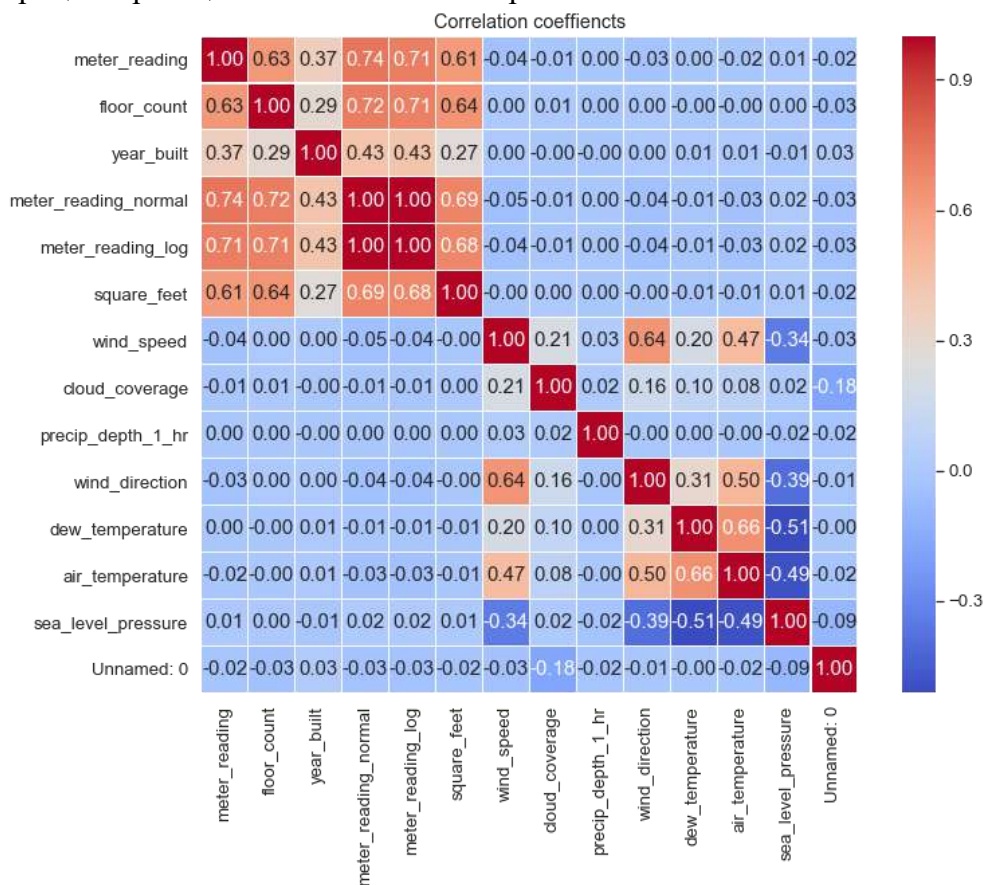


Рисунок 1. Матриця кореляції

Існує висока кореляція між наступними ознаками:

— **dew\_temperature & air\_temperature** – 0,75;

— **square\_feet & floor\_count** – 0,56;

— **site\_id & building\_id** – 0,98.

Решта ознак мають кореляцію менше 0,50.

Розмір будівлі визначає споживання електроенергії, що ми бачимо з наведеного графіка, де кореляція між квадратними футами та кількістю поверхів становить 0,56. Можна також помітити, що температура теж впливає на енергоспоживання.

#### 4. ОПИС АЛГОРИТМУ

У цю нову еру сучасного глибокого навчання дуже спокусливо використовувати сучасну модель глибокого навчання для вирішення будь-якої проблеми. Але завжди треба враховувати вищезазначені фактори, перш ніж відразу переходити до вибору моделі на основі DL.

Розглянемо приклад. Очевидно, що це регресія, оскільки ми прогнозуємо показання лічильників, які є реальними. Крім того, маємо приблизно 40–45 функцій, тому він потрапляє в категорію з низькою кількістю функцій. Ми маємо лише 2380 пар будівельних лічильників з даними за 1 рік для навчання, тому вибірка не є великою.

Як правило, моделі GBDT добре працюють з даними низької розмірності, хоча вони можуть зайняти багато часу через адитивну природу алгоритму.

Крім того, у нас є лише пара будівельних лічильників 2380, тому ми не повинні починати з моделі на основі DL, як LSTM.

Спочатку ми створюємо `u_predict`, використовуючи `train_data`. Потім розділяємо наші `train_data` на `x_train` і `x_cv` як набір для навчання та перехресної перевірки (*cross-validation*) за допомогою бібліотеки Scikit-Learn `train_test_split`. З цього ми створюємо `u_train` і `u_cv`. Тепер, використовуючи ці набори, запустимо різні моделі машинного навчання та зробимо прогнози для `test_data`.

Кросс-валідація (*cross-validation*) — це метод оцінки точності моделі машинного навчання, який використовується для зменшення перенавчання (*overfitting*) та недонавчання (*underfitting*). При кросс-валідації модель будується на частині даних, а потім оцінюється на іншій частині даних. Це дозволяє оцінити, як модель буде працювати на нових даних, які не були використані для її навчання.

Так як цільова змінна має широкий діапазон значень, то є сенс користуватися метрикою RMSLE (Root Mean Squared Logarithmic Error) (1):

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2} \quad (1)$$

де  $n$  – кількість прикладів у наборі даних;  $p_i$  – прогнозоване значення (прогноз);  $a_i$  – фактичне (спостережуване) значення.

Це дозволяє приділяти менше уваги великим аномальним значенням і фокусуватися на точності прогнозу в нормальному діапазоні.

Оскільки вже було взято логарифм цільової змінної, то визначається середня квадратична помилка або RMSE між прогнозованими та фактичними значеннями цільової змінної (**meter\_reading**) (2):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - a_i)^2} \quad (2)$$

де  $n$  – кількість прикладів у наборі даних;  $p_i$  – прогнозоване значення (прогноз);  $a_i$  – фактичне (спостережуване) значення.

### Decision Tree Regression Model

Застосовується регресійна модель Decision Tree до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$ . Помилки перехресної перевірки для цієї моделі наведено нижче на рис. 2:

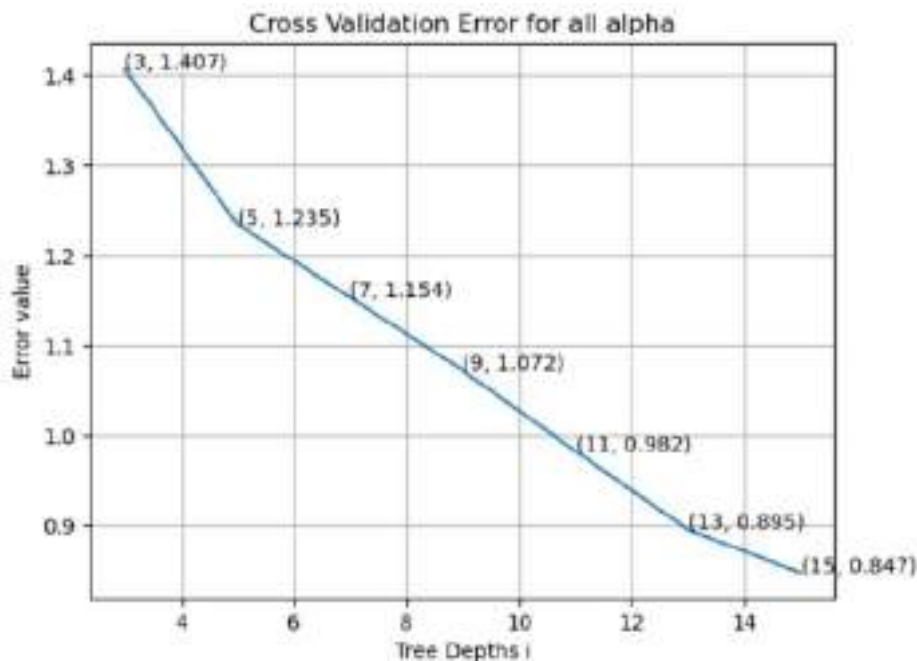


Рисунок 2. RMSLE для Decision Tree

Спостерігається, що найменша помилка перехресної перевірки припадає на глибину дерева 15 із середньоквадратичною помилкою 0,847.

Цю помилку можна ще більше зменшити в інших майбутніх моделях.

### Light Gradient Boosting Machine (LGBM) Decision Tree Regression Model

Тепер застосовуємо регресійну модель дерева рішень Light GBM до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$ . Помилки перехресної перевірки для цієї моделі наведено нижче на рис. 3:

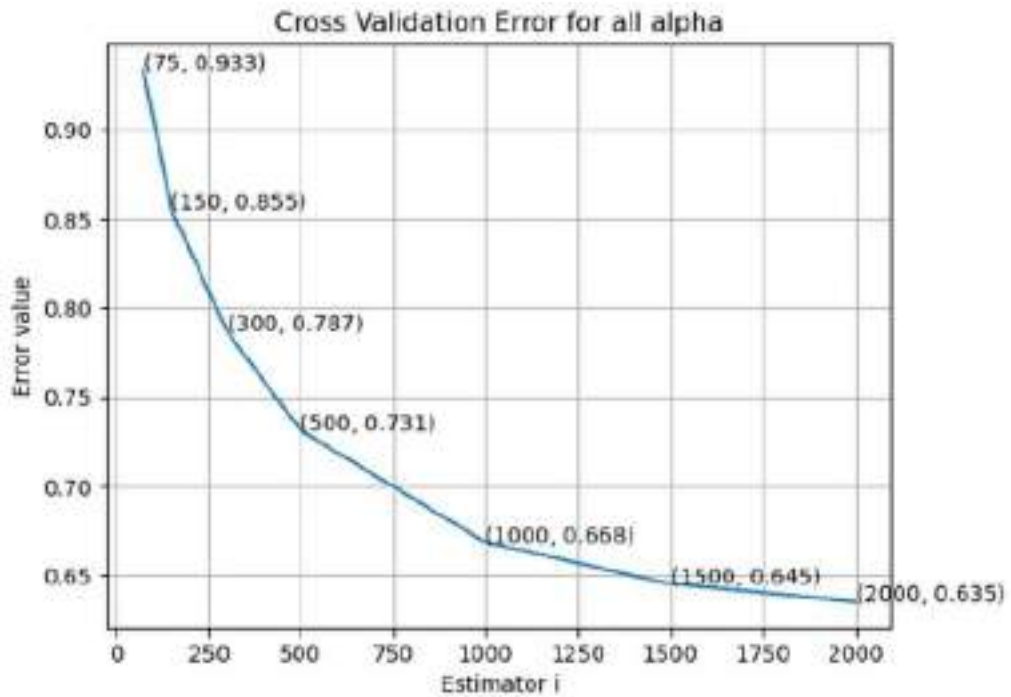


Рисунок 3. RMSLE для LGBM

### Random Forest Regression Model

Тепер ми застосуємо Random Forest Regression Model до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$  (рис. 4):

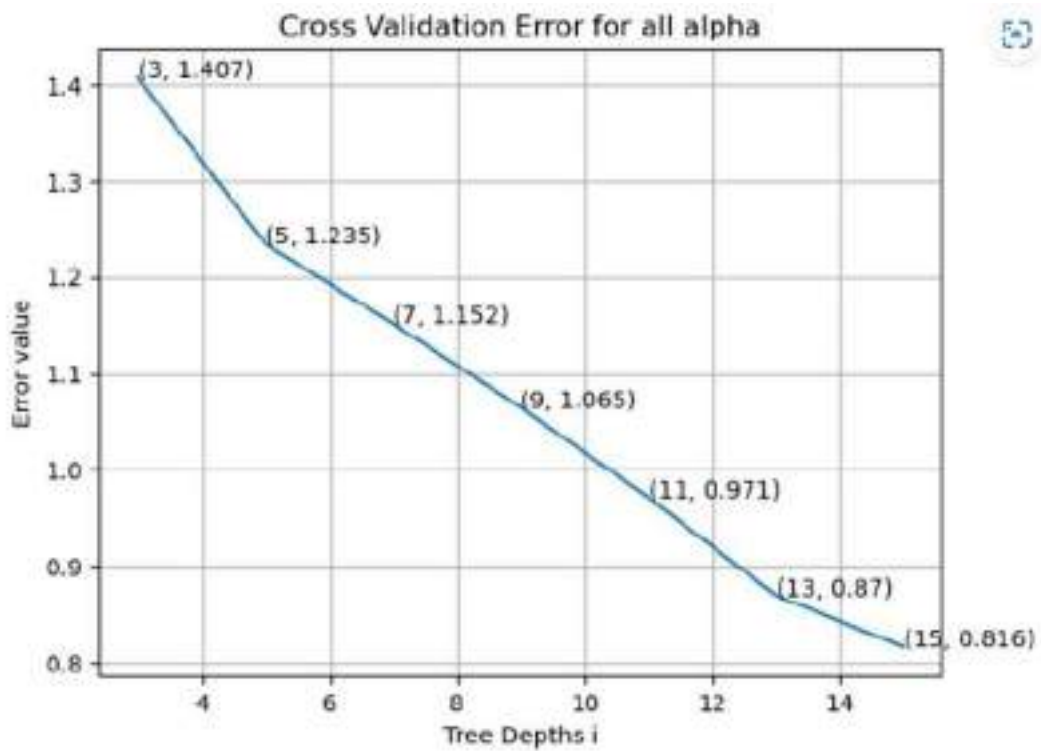


Рисунок 4. RMSLE для Decision Tree

### Light Gradient Boost Machine (LGBM) Random Forest Regression Model

Тепер ми застосовуємо регресійну модель випадкового лісу Light GBM до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$  (рис. 5):

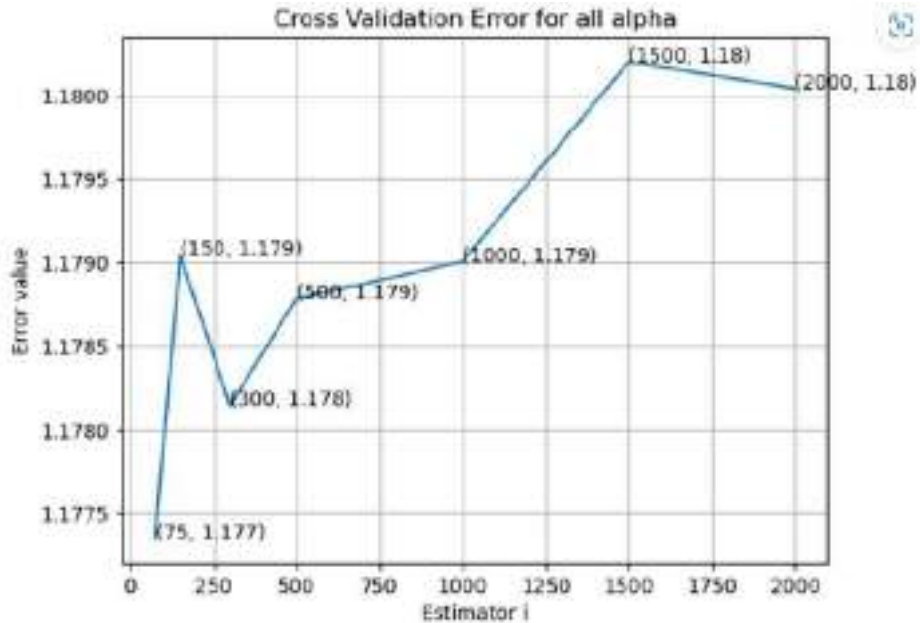


Рисунок 5. RMSLE для LGBM Random Forest

### CatBoost Regression Model

Тепер ми застосовуємо модель Categorical Boost або регресійну модель CatBoost до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$  (рис. 6):

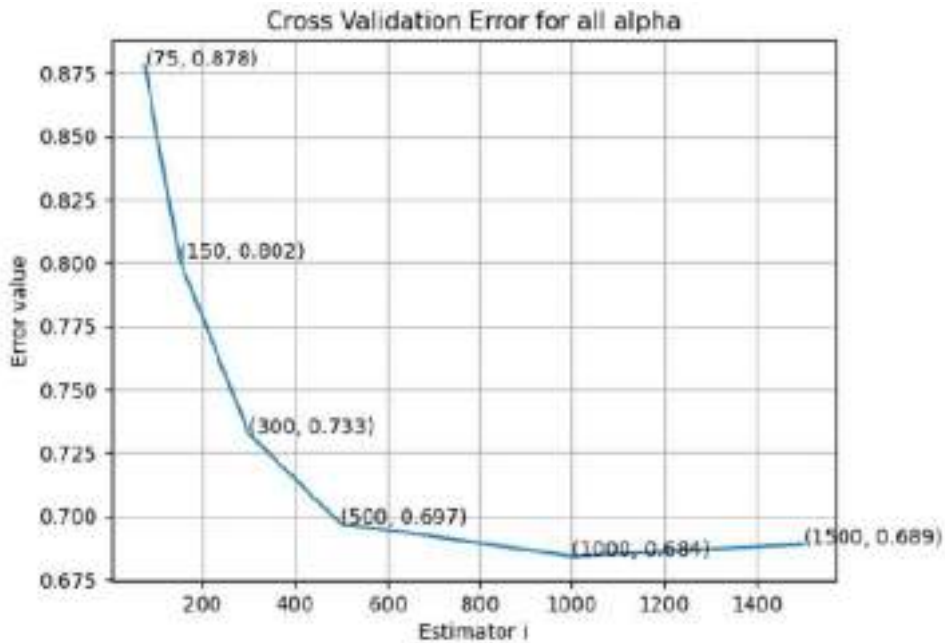


Рисунок 6. RMSLE для CatBoost

## Extreme Gradient Boost (XGBoost) Regression Model

Тепер застосовуємо Extreme Gradient Boost або регресійну модель XGBoost до наборів  $x_{train}$ ,  $x_{test}$ ,  $y_{train}$  і  $y_{test}$ .

Помилки перехресної перевірки для цієї моделі наведено нижче на рис. 7:

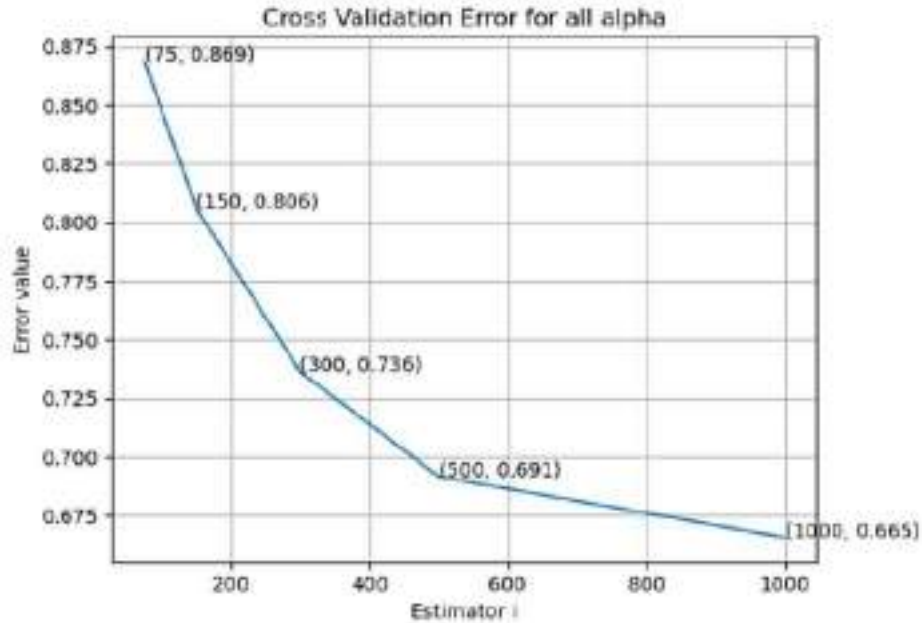


Рисунок 7. RMSLE для XGBoost

Ми зводимо в таблицю всі згенеровані помилки train і перехресної перевірки використаних моделей (табл. 1):

## 5. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результат виглядає наступним чином. Ми зводимо в таблицю всі згенеровані помилки train і перехресної перевірки використаних моделей (табл. 1):

Таблиця 1. Результати

Модель	Cross-Validation RMSLE	Train RMSLE
Decision Tree	1,502	1,533
LGBM GBDT	0,578	0,846
Random Forest	0,561	0,816
LGBM Random Forest	1,166	1,177
Catboost	0,575	0,684
XGBoost	0,497	0,665

Найкращою моделлю для прогнозування RMSLE є XGBoost. Вона має найнижчий показник RMSLE на обох наборах даних, тренувальному і крос-валідаційному. Інші моделі також мають порівняно низькі показники RMSLE. Random Forest і Catboost також мають хороші результати, але вони не такі хороші, як XGBoost. Decision Tree і LGBM Random Forest мають вищі показники RMSLE. Це може бути пов'язано з тим, що ці моделі не так добре

справляються з нелінійною залежністю між змінними. Decision Tree має найвищий показник RMSLE. Це може бути пов'язано з тим, що це простий алгоритм, який не може враховувати складні взаємозв'язки між змінними. LGBM GBDT має найнижчий показник RMSLE. Цей алгоритм є одним з найпотужніших алгоритмів машинного навчання для регресії. Він може ефективно враховувати складні взаємозв'язки між змінними. Random Forest також має хороший результат. Цей алгоритм є ансамблем декількох дерев рішень. Він може поліпшити точність прогнозів порівняно з одним деревом рішень. LGBM Random Forest має вищий показник RMSLE, ніж LGBM GBDT. Це може бути пов'язано з тим, що цей алгоритм є більш складним, ніж LGBM GBDT. Він може бути більш сприйнятливим до перенавчання. Catboost має хороший результат. Цей алгоритм є ансамблем декількох моделей CatBoost. Він може ефективно враховувати складні взаємозв'язки між змінними.

## 6. ВИСНОВКИ

Прогнозування енергоспоживання будівель за допомогою моделей машинного навчання є важливим для оцінки ефективності в контексті модернізації, вимірювання та верифікації, інтеграції відновлюваних джерел енергії, управління системами, виявлення несправностей, енергоспоживання в житловому секторі та моделювання енергетики в міському масштабі.

Висока частка енергії, що споживається в будівлях, спричинила появу багатьох екологічних проблем, які негативно впливають на існування людства. Прогнозування енергоспоживання будівель, по суті, проголошується методом енергозбереження та покращення прийняття рішень щодо зменшення споживання енергії. Крім того, будівництво енергоефективних будівель сприятиме зменшенню загального споживання енергії в новозбудованих будівлях.

Дослідники припускають, що наявність енергетичної системи будівлі з точним прогнозуванням може заощадити від 10 до 30% загального енергоспоживання в будівлях. Без виявлення алгоритму, який може точно прогнозувати енергоспоживання будівлі, це призведе до збільшення викидів парникових газів, будівництва більш неефективних будівель, попиту на енергію та зменшення фінансових заощаджень. У даній роботі було розглянуті методи машинного навчання для прогнозування енерговитрат. Методи машинного навчання було застосовано на наборі реальних даних. Результати прогнозу показали, що найкращою моделлю для прогнозування RMSLE є XGBoost. Вона має найнижчий показник RMSLE на обох наборах даних, тренувальному (0,497) і крос-валідаційному (0,665).

## ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. World power consumption | Electricity consumption | enerdata. World Energy Statistics Enerdata. Accessed 7 Dec 2020. <https://yearbook.enerdata.net/electricity/electricity-domestic-consumption-data.html>

2. Global energy review 2020. IEA. Accessed 7 Dec 2020. Available from: <https://www.iea.org/reports/global-energy-review-2020/electricity>