МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
УКРАЇНСЬКА АКАДЕМІЯ ДРУКАРСТВА

КАЛІНІНА ІРИНА ОЛЕКСАНДРІВНА

УДК 004.852:004.6]:004.94-047.72](043.5)

МОДЕЛІ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЙМОВІРНІСНО-СТАТИСТИЧНОГО АНАЛІЗУ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ В ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

05.13.06 — Інформаційні технології

РЕФЕРАТ

дисертації на здобуття наукового ступеня
доктора технічних наук

Львів – 2023

Дисертацією є рукопис.

Робота виконана в Чорноморському національному університеті імені Петра Могили Міністерства освіти і науки України

Науковий консультант:

лауреат Державної премії в галузі науки і техніки,
доктор технічних наук, професор
Бідюк Петро Іванович,
Національний університет «КПІ ім. Ігоря Сікорського»,
професор кафедри математичних методів системного аналізу

Офіційні опоненти:

заслужений діяч науки і техніки України,
доктор технічних наук, професор
Литвиненко Володимир Іванович,
Херсонський національний технічний університет,
завідувач кафедри інформатики та комп'ютерних наук

доктор технічних наук, професор,
Піх Ірина Всеволодівна,
Українська академія друкарства,
професор кафедри комп'ютерних наук та інформаційних технологій

доктор технічних наук, доцент
Рак Тарас Євгенович,
Приватний заклад вищої освіти «ІТ СТЕП Університет»,
проректор з науково-педагогічної роботи

Захист дисертації відбудеться «26» січня 2024 р. о 12-00 на засіданні спеціалізованої вченої ради Д 35.101.01 в Українській академії друкарства за адресою: 79020, м. Львів, вул. Під Голоском, 19, ауд. 101.

Із дисертацією можна ознайомитися в бібліотеці Української академії друкарства за адресою: 79020, м. Львів, вул. Підвальна, 17.

Реферат розіслано «18» грудня 2023 р.

Учений секретар
спеціалізованої вченої ради Д 35.101.01
к.т.н., доцент В.Ц. Жидецький

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Головною задачею, яку вирішують в багатьох прикладних завданнях (економічних, фінансових, екологічних, медичних та інших), це отримання точних прогнозів, що стосуються поведінки складних систем та об'єктів на основі аналізу даних про їх минулу поведінку. Однак багато проблем, що виникають у практичному застосуванні, не можуть бути вирішені відомими методами та відповідними алгоритмами. Це відбувається через те, що механізми генерації вхідних даних точно не відомі або немає достатньо статистичних даних для побудови прогнозної моделі. Окрім того в реальних даних присутні нелінійності та нестаціонарності, які суттєво впливають на якість прогнозної моделі.

Існуючий рівень інформаційних технологій для вирішення завдань ймовірнісно-статистичного аналізу в контексті задач машинного навчання не задовольняє сучасним потребам аналітиків, експертів. Це пов'язано, насамперед, з відсутністю системної методології побудови та застосування інформаційних технологій для вирішення різних проблемно-орієнтованих задач машинного навчання, недостатньо розвиненим математичним забезпеченням, відсутністю адекватних математичних моделей, ефективних методів прогнозування та інформаційних технологій, що дозволяють вирішувати завдання ймовірнісно-статистичного аналізу даних, які мають нелінійний характер, та описують нестаціонарні процеси.

Створення інформаційних технологій та інформаційно-аналітичних систем для вирішення задач машинного навчання неможливе без застосування сучасних методів ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних, математичного моделювання, методів прогнозування. Все це дозволяє стверджувати, що тема досліджень *актуальна*.

Розвинута в даній роботі теорія вирішення завдань ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних, моделювання та прогнозування. Розроблені моделі, методи, алгоритми та засоби ґрунтуються на дослідженнях та теоретичних результатах, отриманих такими відомими зарубіжними вченими як Хасти Т. (Hastie T.), Тибширані Р. (Tibshirani R.) – теорія статистичного навчання; Ерл Т. (Erl T.), Хаттак В. (Khattak W.), Булер П. (Buhler P.) – методи обробки великих данних; Келлехер Д. (Kelleher J.), Мак-Нейми Б. (Mac Namee B.), Ланц Б. (Lantz B.) – методи машинного навчання; Фан Д. (Fan J.), Кантц Х. (Kantz H.), Шрайбер Т. (Schreiber T.) – методи обробки нелінійних нестаціонарних данних; Крушке Д. (Kruschke J.), Гельман А.(Gelman A.), Джейнса Е. (Jaynes E.) – теорія ймовірнісного аналізу; Кітагава Г. (Kitagawa G.), Нильсен Е. (Nilsen E.) – аналіз часових рядів, та вітчизняними вченими на основі праць Згуровського М.З., Панкратової Н.Д., Данилова В.Я. – теорія та прикладні методи системного аналізу; Крака Ю.В., Дурняка Б. В. – інтелектуальні технології моделювання складних процесів і систем; Бідюка П.І. – аналіз нелінійних систем та методи прогнозування, аналіз часових рядів; Бодянського Є.В. – нейромережеві та нечіткі методи штучного інтелекту; Кондратено Ю.П. – методи нечіткого аналізу складних систем; Литвиненко В.І. – методи інтелектуальних

обчислень та інтелектуального аналізу даних, Сеньківського В.М., Піх І.В., Тимченко О.В., Рак Т.Е., Гожого О.П. теорія та методи побудови інформаційних систем та інші.

Таким чином, актуальною **науково-прикладною проблемою** є підвищення ефективності ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в завданнях машинного навчання засобами сучасних інформаційних технологій з урахуванням нелінійності і нестаціонарності даних, а також можливих невизначеностей, що є характерними для них.

Зв'язок роботи із науковими програмами, планами, темами. Дисертація виконувалась в Чорноморському національному університеті ім. Петра Могили. Тема дисертаційної роботи повністю відповідає науковим напрямам, які виконуються на факультеті комп'ютерних наук, зокрема науковим дослідженням в області інтелектуальних систем і методів прийняття рішень, прогнозування, побудови інтелектуальних інформаційних систем. Робота виконувалась в рамках держбюджетних тем, госпдоговорів, зокрема таких: «Розроблення автоматизованої системи керування гібридним енергетичним комплексом із застосуванням засобів штучного інтелекту для забезпечення енергетичної безпеки України» *номер державної реєстрації* № 0120U102032 (*відповідальний виконавець*); «Розробка методів та алгоритмів інтелектуального аналізу даних на основі ймовірнісно-статистичних методів», *номер державної реєстрації* №0118U000862 (*відповідальний виконавець*); «Розробка моделей та інструментальних засобів для підвищення ефективності взаємодії web-сервісів інтелектуальних додатків» *номер державної реєстрації* №0118U000853 (*відповідальний виконавець*); «Теоретичні основи визначення індикаторів та коефіцієнтів вагомості індексів екологічної безпеки в системі сталого розвитку Південного регіону України», *номер державної реєстрації* №0114U004572 (*виконавець*); «Розробка інструментальних засобів для систем підтримки прийняття рішень на основі еволюційних методів і алгоритмів» *номер державної реєстрації* № 0112U001103 (*виконавець*).

Метою дисертаційного дослідження – є розроблення методологічних основ інформаційних технологій для підвищення якості та ефективності ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в завданнях машинного навчання завдяки створенню та вдосконаленню методів, моделей, інформаційних технологій для нелінійних нестаціонарних процесів.

Для досягнення поставленої мети необхідно розв'язати такі *задачі*:

1. Виконати аналіз існуючих методів, моделей і алгоритмів ймовірнісно-статистичного аналізу, моделювання та прогнозування, з урахуванням нелінійностей та нестаціонарностей даних і можливих супутніх невизначеностей, для розв'язання завдань машинного навчання.
2. Розробити метод синтезу нових інформаційних технологій для розв'язування задач ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в умовах наявності нелінійних нестаціонарних процесів.
3. Розробити метод обробки пропусків в даних, який ідентифікує пропуски в даних, виявляє закономірності їх появи та формує набори даних без пропусків.

4. Розробити метод виявлення та обробки аномальних значень в наборах даних, який ідентифікує екстремальні дані, аналізує причини їх появи, та здійснює їх обробку.

5. Розробити метод для вирішення завдань фільтрації даних на основі байєсівського підходу та методу гранулярної фільтрації.

6. Розробити підхід до нормалізації та стандартизації даних на основі системного поєднання методів перетворення даних та особливостей вирішення завдань машинного навчання.

7. Розробити метод до побудови прогнозних моделей на основі байєсівського підходу до аналізу часових рядів для вирішення завдань машинного навчання, який враховує нелінійності та нестаціонарності даних при оцінюванні прогнозу.

8. Розробити метод побудови імітаційних моделей систем зі складним стохастичним процесом обробки даних для аналізу та моделювання нелінійних нестаціонарних процесів на основі колірних мереж Петрі.

9. Розвинути метод синтезу параметрів нелінійної прогнозної моделі на основі генетичного алгоритму.

10. Дослідити та розробити системний підхід до моделювання, прогнозування та підтримки прийняття рішень в задачах машинного навчання.

11. Дослідити та розробити метод розв’язання задач прогнозування часових рядів за рахунок використання багатошарових нейронних мереж прямого розповсюдження.

12. Дослідити та розвинути метод побудови комбінованих прогнозів нелінійних та нестаціонарних даних з метою підвищення точності прогнозування.

13. Розробити метод побудови багаторівневих гетерогенних ансамблів прогнозних моделей для покращення точності прогнозування в задачах машинного навчання.

14. Розробити інформаційні технології (ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання і прогнозування) та інформаційно-аналітичні системи для вирішення завдань машинного навчання.

15. Створити інструментальні засоби для розв’язання прикладних завдань машинного навчання та прийняття рішень в різних галузях, що підтверджують достовірність наукових та практичних результатів.

Об’єкт дослідження – процеси ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних, моделювання та прогнозування при вирішенні завдань машинного навчання.

Предмет дослідження – методи, моделі, інформаційні технології та системи для ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання і прогнозування нелінійних нестаціонарних процесів в завданнях машинного навчання.

Методи дослідження. З урахуванням специфіки об’єкта досліджень й сформульованої мети методами дослідження є: для аналізу і обробки інформації – методи системного аналізу та методи ймовірнісно-статистичного аналізу, байєсівський аналіз даних; для моделювання і побудови прогнозних моделей – методи системного аналізу, методи побудови моделей в просторі станів,

ймовірнісно-статистичного моделювання і теорії графів, методи еволюційного програмування, методи байєсівського аналізу і колірні мережі Петрі; для побудови і оцінювання прогнозів в задачах машинного навчання – методи прогнозування на основі теорії часових рядів, методи регресійного аналізу, байєсівських структурних часових рядів, ансамблеві методи та методи комбінування прогнозів; для побудови практичних реалізацій – методи, засоби і технології сучасного прикладного програмування та створення інформаційно-аналітичних систем.

Наукова новизна одержаних результатів. На основі виконаних теоретичних і експериментальних досліджень вирішено важливу науково-прикладну проблему розширення можливостей та підвищення ефективності ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в завданнях машинного навчання засобами сучасних інформаційних технологій з урахуванням нелінійності і нестаціонарності даних, а також можливих невизначеностей. При цьому отримано такі нові результати:

вперше розроблено:

– метод синтезу нових інформаційних технологій для розв’язування завдань машинного навчання, який ґрунтується на системному використанні методів ймовірнісно-статистичного аналізу даних, математичного моделювання, методів прогнозування, що підвищує ефективність процесу машинного навчання в умовах наявності нелінійних нестаціонарних процесів, які досліджувались, та різних типів невизначеностей даних;

– метод пошуку викидів та аномалій в нелінійних нестаціонарних даних на основі системного використання методів виявлення аномальних значень, аналізу причин та методів обробки викидів, що дало змогу підвищити точність ідентифікації аномальних значень в наборах даних різного типу;

– метод обробки пропусків в наборах даних на основі системного використання методів пошуку закономірностей появи відсутніх значень та методів аналізу наборів даних без пропусків, що дало змогу значно підвищити ефективність попередньої обробки даних;

– метод побудови імітаційних моделей систем зі складним стохастичним процесом обробки даних на основі колірних мереж Петрі для розв’язування завдань аналізу та моделювання нелінійних та нестаціонарних процесів, який базується на системному використанні стохастичних, часових, ієрархічних мереж Петрі для імітаційного моделювання динамічних процесів, що дало можливість значно підвищити ефективність і точність побудови математичних моделей для аналізу складних нелінійних нестаціонарних процесів;

– системний підхід до моделювання та прогнозування в задачах машинного навчання, що враховує можливі невизначеності при аналізі процесу, нелінійності, нестаціонарності даних, особливості підбору й оцінки структури та параметрів прогнозних моделей та їх оцінювання, що дало змогу підвищити точність оцінювання якості прогнозу та визначити краще прогнозне рішення;

– метод побудови прогнозних моделей на основі байєсівських структурних часових рядів, якій базується на процесі навчання структурних моделей часових

рядів та на алгоритмі побудові BSTS-моделі, що дало можливість враховувати нелінійність та нестаціонарність процесів при аналізі часових рядів та підвищити точність прогнозування;

– метод зменшення похибки прогнозу за рахунок одночасного зменшення зміщення та дисперсії на основі використання багаторівневих гетерогенних ансамблів прогнозних моделей, що дало змогу підвищити якість та точність прогнозування;

– інформаційні технології (ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання і прогнозування) та архітектури інформаційно-аналітичних систем на основі розроблених методів, моделей ймовірнісно-статистичного аналізу, моделювання та прогнозування, з урахуванням нелінійностей та нестаціонарностей даних і можливих супутніх невизначеностей, що дозволило підвищити ефективність вирішення задач машинного навчання;

отримали подальший розвиток:

– метод фільтрації статистичних даних завдяки системному використанню різних типів фільтрів за рахунок модифікації алгоритмів гранулярної фільтрації, що дало можливість в процесі попередньої обробки врахувати нелінійність та нестаціонарність (гетероскедастичність) даних, і описати сам процес та динаміку його дисперсії;

– підхід до нормалізації та стандартизації даних на основі системного поєднання методів перетворення даних та особливостей вирішення завдань машинного навчання, що дало змогу спростити і прискорити процедури нормування складних наборів даних;

– метод синтезу параметрів нелінійної прогнозної моделі на основі використання генетичного алгоритму, що підвищило ефективність підбору параметрів прогнозних моделей;

удосконалено:

– метод розв’язання задач прогнозування для часових рядів за рахунок використання багатошарових нейронних мереж прямого розповсюдження, яка забезпечує підвищення точності прогнозування;

– метод побудови і використання комбінованих прогнозів за рахунок ітеративного оцінювання різних схем комбінування, що дає змогу підвищити ефективність та точність прогнозних рішень.

Практичне значення одержаних результатів полягає у тому, що розроблені інформаційні технології, методи і моделі забезпечують:

– зменшення часу, необхідного для розробки інформаційно-аналітичних систем з метою розв’язання задач ймовірнісно-статистичного аналізу та попередньої обробки, моделювання і прогнозування в різних завданнях машинного навчання;

– підвищення якості прогнозування за рахунок використання байєсівських структурних часових рядів, які базуються на процесі навчання структурних моделей часових рядів та на алгоритмі побудові BSTS-моделі (з предикторами та без предикторів) у середньому на 6-27% в порівнянні з різними статистичними

моделями ARIMA, що дало можливість враховувати нелінійність та нестаціонарність процесів при аналізі часових рядів;

– підвищення точності прогнозування на часових рядах за рахунок використання багатошарових нейронних мереж прямого розповсюдження, що забезпечує збільшення точності до 25% (по показнику RMSE) в порівнянні з кращою з альтернативних статистичних моделей ARIMA;

– підвищення якості прогнозних рішень за рахунок зменшення похибки прогнозу шляхом одночасного зменшення зміщення та дисперсії на основі використання багаторівневих гетерогенних ансамблів прогнозних моделей, що дало змогу підвищити якість та точність класифікації до 31%, по показнику F-міри до 29% в порівнянні з кращою з альтернативних моделей;

– удосконалено методику побудови і використання комбінованих прогнозів за рахунок ітеративного оцінювання різних схем комбінування, що дає змогу підвищити ефективність та точність прогнозних рішень у середньому на 22%.

Результати дисертаційної роботи впроваджені в Благодійній організації ОБФ «Регіональний фонд благочестя»; головному управлінні державної служби України з надзвичайних ситуацій у Чернівецький області; ТОВ «Торгівельної компанії «ЮВЕНТА» (м. Одеса); ТОВ «Артіль» ЛТД (м. Миколаїв); ПП (IT-компанія) «DataOx Incorporated» («IntroLabSystem»); ПП (IT-компанія) «TemplateMonster»; Чорноморському національному університеті (ЧНУ) ім. Петра Могили (м. Миколаїв).

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. Роботи [17-20, 49-51] опубліковано без співавторів. У друкованих працях, опублікованих у співавторстві, особисто здобувачу належать такі результати: [1,2,4,6,7,9,12,3,5,7-20,4,5,8,30,33,39,49,65-67] – розробка прогнозних моделей та їх елементів; [8,23] –побудова ансамблів моделей; [5,10,11,14,21,26,34,41,54,62,69] – розробка моделей за допомогою кольорових мереж Петрі; [4,9,35,39,57,68] – аналіз та подолання нелінійностей та нестаціонарностей; [22,31,32,40,42,44,56,57-63,68] – застосування підходу на основі Байєсівського аналізу даних; [28,43,46,47,53] – аналіз та класифікація невизначеностей; [4,24,25,66,68] – побудова моделей на основі методу байєсівських структурних часових рядів; [3,21] – вдосконалений метод ідентифікації пропусків; [3,15-18,36,37,38,46,48,52,55,64-67,70,71] – вирішення прикладних задач обробки даних на основі методів машинного навчання та аналізу даних.

Апробація результатів дисертації. Основні положення дисертації було подано й обговорено більш ніж на 25 міжнародних науково-технічних конференціях та наукових семінарах і симпозіумах, серед них: Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій, ISDMCI» (м. 2009, 2011, 2012, 2014, 2016-2022 м. Евпаторія - Залізний Порт,); Міжнародна наукова конференція «Інформаційні технології в металургії і машинобудуванні, ІТММ» (м. Дніпро, 2021, 2023); International conference on computer science and information technologies CSIT, (м. Львів, 2017-2023); Міжнародний науковий симпозіум «Інтелектуальні рішення»

(м. Ужгород, 2016, 2019, 2021); Міжнародна конференція «Ольвійський форум» (м. Ялта, 2008, 2009, 2011-2013, м. Миколаїв, 2017-2023). Всеукраїнська науково-методична конференція «Могилянські читання: досвід та тенденції розвитку суспільства в Україні» (м. Миколаїв, 2017-2022), IEEE Int. Conf. "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS–2017, 2019, 2021 (Bucharest, 2017, Metz, 2019, Krakow, 2021), IEEE Int. Scientific and Technical Conf. «Data Stream Mining & Processing DSMP–2016, 2018, 2020» (Львів , 2016, 2018, 2020), IEEE 2018 1st International Conference on Systems Analysis and Intelligent Computing (SAIC 2018), 2nd international conference «Computational Linguistics and Intelligent Systems» (CoLInS 2018, 2020), International Conference on Informatics & Data-Driven Medicine (IDDM-2020, Lviv), 1st International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk-2020, Kherson), International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT+DS 2022, 2023), Lviv, Ukraine.

Основні положення дисертаційної роботи розглядалися на науково-технічних семінарах факультету комп'ютерних наук ЧДУ ім. П. Могили,

Публікації. Основні результати дисертаційної роботи опубліковано в 71 наукових працях, з них 32 роботи входять до наукометричних баз *Scopus* та/або *Web of Science* (з яких 1 стаття має 2-й квартиль, 2 статті – 3-й квартиль у *Scopus*). За результатами дисертаційного дослідження опубліковано: 2 монографії, 22 статей у фахових періодичних виданнях, із них 6 у фахових періодичних виданнях інших країн (включених до міжнародних наукометричних баз *Scopus* та/або *Web of Science*) і 16 в виданнях включених до переліку МОН України; 26 публікацій у виданнях включених до міжнародних наукометричних баз *Scopus* та/або *Web of Science* (10 у монографіях *Springer*, 16 в працях конференцій), 20 публікацій у збірниках праць міжнародних та національних конференцій, 1 навчальний посібник.

Структура та обсяг роботи. Дисертаційна робота складається з анотації, вступу, 6 розділів, висновків, 3 додатків на 34 сторінці та списку використаних джерел з 472 найменувань на 35 сторінках. Загальний обсяг дисертації становить 416 сторінок, з них 319 сторінок основного тексту, 142 рисунків, 43 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність обраної теми, наведено зв'язок роботи з науковими програмами, сформульовано мету, проблему та задачі дослідження, відображено наукову новизну, практичну цінність роботи та особистий внесок здобувача, наведено відомості про апробацію, публікації та використання результатів дослідження.

У **першому розділі** (*Системологічний аналіз проблеми обробки нелінійних та нестаціонарних даних в задачах машинного навчання*) проведено системологічний аналіз сучасного стану обробки нелінійних та нестаціонарних даних в задачах машинного навчання.

Описано різні типи задач та методів машинного навчання, проведена їх

класифікація. Проаналізовано основні методи машинного навчання. Проведено аналіз моделей простору станів у задачах ймовірнісно-статистичної обробки інформації. Моделі простору станів використовуються у детермінованих і в стохастичних додатках, застосовуються до безперервних та дискретних даних. Детально досліджені та проаналізовані наступні моделі і методи: фільтр Калмана для лінійної гаусової моделі, приховані марківські моделі, Байєсівські структурні часові ряди. Детально досліджені та проаналізовані методи аналізу та обробки нелінійностей у задачах обробки даних. Визначено основні типи нелінійностей, проаналізовані та досліджені методи їх ідентифікації. Згруповані та досліджені основні тести на нелінійність. Проведено аналіз типів стаціонарностей та нестаціонарностей, а також методів їх ідентифікації в задачах аналізу даних.

Проаналізовано методи прогнозування нелінійних нестаціонарних процесів в задачах машинного навчання. Доведено, що вибір та розробка нових ефективних методів прогнозування є основним фактором підвищення ефективності методів аналізу даних та машинного навчання в цілому. Це дало можливість сформувати загальні напрями дослідження.

Обґрунтовано доцільність та перспективність розробки і використання нових інформаційних технологій для вирішення завдань машинного навчання. Аналіз методів ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів показав наявність невирішених проблем. Доведено, що для вирішення визначених проблем необхідно розробити методи, інформаційні технології та їх елементи, опис, побудова і застосування яких наводиться в наступних розділах роботи. Сформульовано мету та основні задачі дисертаційного дослідження.

У другому розділі (*Теоретичні і методологічні основи побудови сучасних інформаційних технологій ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів в завданнях машинного навчання*) розглядається постановка загальних задач ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів, математичні моделі та методи формалізації цих задач в завданнях машинного навчання, побудова моделей, методів та реалізація на їх основі сучасних інформаційних технологій аналізу, моделювання та прогнозування.

Основні процедури аналізу й обробки інформації базуються на методах системного аналізу. Системний підхід до розв’язання прикладних задач машинного навчання та ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів використовується на кожному етапі обробки інформації.

Досліджено та представлено структуру обробки інформації при вирішенні задач машинного навчання (рис.1). Для первісної обробки інформації використовуються методи ймовірнісно-статистичні аналізу та попередньої обробки даних. Цей етап здійснюється за допомогою відповідної інформаційної технології. На наступному етапі будуються прогностичні моделі за допомогою інформаційної технології моделювання. Далі визначається структура прогнозних рішень і будуються відповідні прогнози. Це здійснюється за допомогою інформаційної технології прогнозування.

Таким чином, для розв’язання задач ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів в задачах машинного навчання необхідно побудувати наступні типи інформаційних технологій: інформаційну технологію

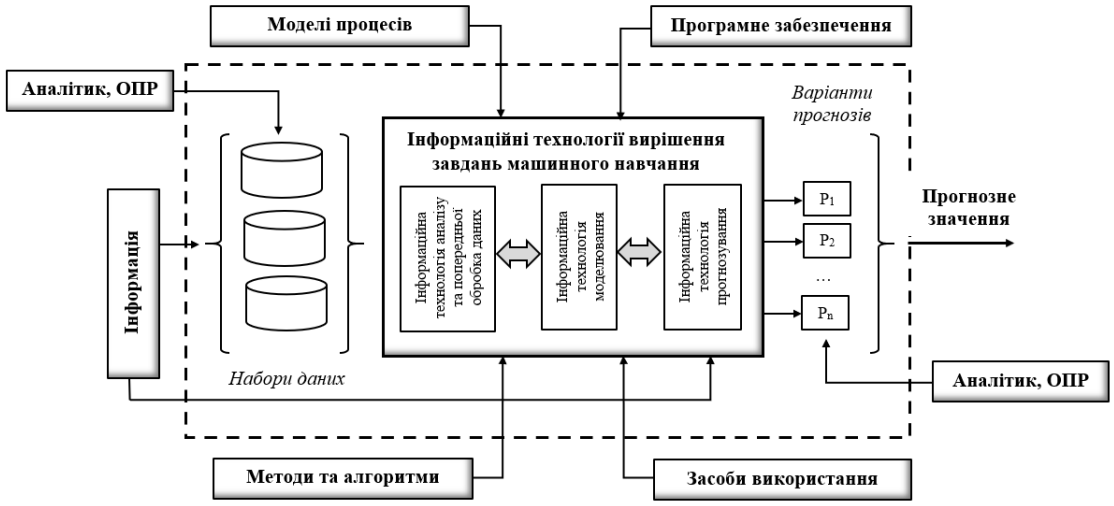

Рисунок 1 – Структура обробки інформації при розв’язанні задач

аналізу та попередньої обробки даних, інформаційну технологію моделювання, інформаційну технологію прогнозування.

Схему послідовності етапів обробки даних при розв’язанні задач машинного навчання, а саме класифікації і регресії, представлено на рис. 2. Вона складається з наступних п’яти етапів: збір даних, дослідження та підготовки даних, навчання прогнозних моделей, обчислення прогнозів, оцінювання та перевірка якості.

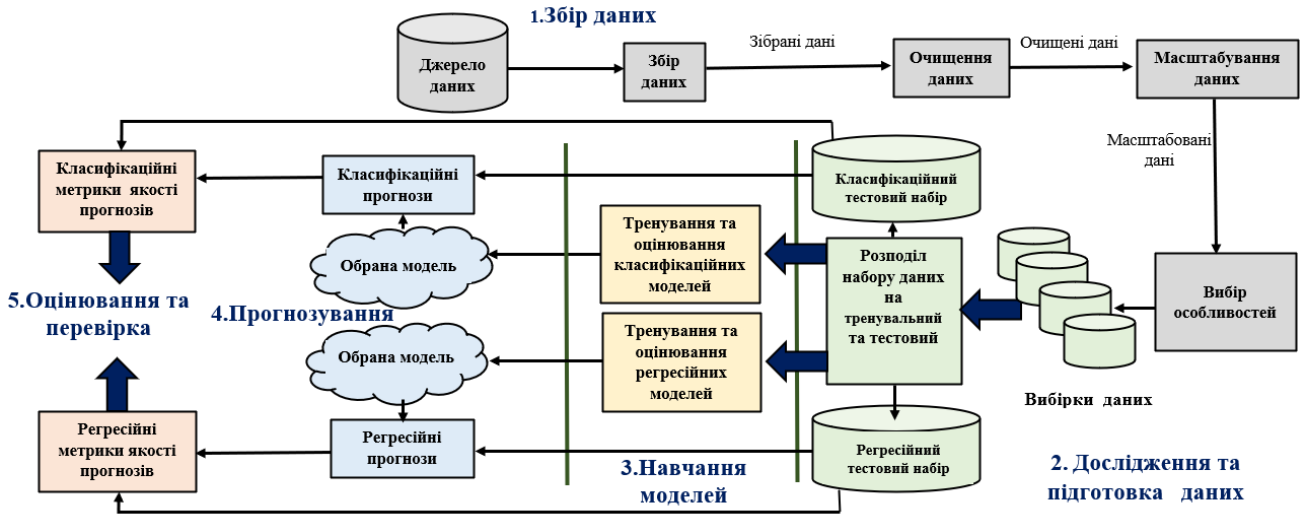

Рисунок 2 – Послідовність етапів обробки даних при розв’язанні задач машинного навчання

Розроблено інформаційну модель вирішення завдань машинного навчання, що дає можливість створення формалізованого опису завдань машинного навчання та їх використання у процесі побудові елементів інформаційних

технологій, а також для збереження даних про структуру задач машинного навчання.

Інформаційна модель для вирішення завдань машинного навчання описується за допомогою наступної множини елементів:

$$ML_T = \{NData_T, NL_{Data}, NS_{Data}, MA_T, Mod_T, MP_T, MQp\}$$

де $NData_T$ – множина наборів даних, які обробляються при вирішенні задачі машинного навчання; NL_{Data} – множина нелінійностей наборів даних, які враховуються при вирішенні задачі машинного навчання; NS_{Data} – множина нетаціонарностей процессу, який описують набори даних, що враховуються при вирішенні з задачі машинного навчання; MA_T – множина методів аналізу та попередньої обробки наборів даних; Mod_T – множина методів побудови моделей та методів моделювання для вирішення задачі машинного навчання; MP_T – множина методів прогнозування на основі ймовірнісно-статистичного аналізу з врахуванням нелінійностей та нестаціонарностей даних; Us_T – множина невизначеностей при вирішенні задачі машинного навчання.

Множина набору даних $NData_T = Data_{TR} \cup Data_{TST}$ поєднує дві підмножини. Перша підмножина $Data_{TR}$ поєднує дані для тренування, друга підмножина $Data_{TST}$ – набір тестових даних. В множину невизначеностей Us_T , входять невизначеності статистичного типу, невизначеності через відсутність спостережень, невизначеність параметрів моделей, невизначеності структури моделі, невизначеності амплітудного та ймовірнісного типу.

В машинному навчанні існують три типи навчання: *з вчителем, без вчителя* та *з підкріпленням*. Задачі навчання при любому типі навчання ділиться на три етапи: підготовка (обробка) даних, побудова моделі, вирішення задач машинного навчання (класифікація, регресія, прогнозування)

Перший етап породжує ряд підзадач: визначення способу завдання даних (об'єктів); визначення способу опису даних; визначення способу формування відповідей; визначення способу побудови функції $\boldsymbol{a}$; визначення способу наближення для $\boldsymbol{a}_i$. Всі ці задачі вирішуються послідовно, і результати рішення попередньої, як правило, впливають на рішення поточної. Але найважливішою серед них буде визначення способу опису даних, від цього залежить опис результату. Можна виділити наступні типи вхідних даних при навчанні: координати об'єктів у просторі ознак; часовий ряд; сигнал; зображення; відеоряд; опис взаємовідносин між об'єктами.

На основі вище сказаного представлено формальний опис процесу машинного навчання.

Дано $\boldsymbol{X} = \{x_1,\ldots,x_k\}$ – дані (об’єкти), які досліджуються. $\boldsymbol{F} = \{f_1,..,f_N\}$ – простір ознак в якому об’єкти функціонують, при цьому $f_i(x_j)$, $i \in [1\ldots N], j \in [1\ldots k]$ – значення i-го признака, j-го об’єкта. Тоді ознаковий опис об’єкту задається у вигляді матриці:

$$\boldsymbol{D} = || f_i(x_j) ||_{k\times N} = \begin{pmatrix} f_1(x_1) \ldots f_N(x_1) \\ f_1(x_k) \ldots f_N(x_k) \end{pmatrix}. \quad (1)$$

В кінцевому результаті вхідна інформація – навчальна множина:

$$\langle \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}} \rangle = \{\tilde{x}_i, \tilde{y}_i\}, i = 1, \dots, k,$$

де $\tilde{x}_i \in \mathbb{R}^n$ – вхідний вектор з *i*-го прикладу, $\tilde{y}_i \in \mathbb{R}^r$ – вектор відповідних вказівок.

$$\left[\begin{array}{cccc|cccc} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1n} & \tilde{y}_{11} & \tilde{y}_{12} & \cdots & \tilde{y}_{1r} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2n} & \tilde{y}_{21} & \tilde{y}_{22} & \cdots & \tilde{y}_{2r} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \tilde{x}_{k1} & \tilde{x}_{k2} & \cdots & \tilde{x}_{kn} & \tilde{y}_{k1} & \tilde{y}_{k2} & \cdots & \tilde{y}_{kr} \end{array}\right],$$

де *k* прикладів (спостережень), *n* входів та *r* виходів.

Наступний етап: побудова моделі та її навчання.

Етапи побудови (ідентифікації) формальних математичних моделей:

– *структурна* (визначається різноманіття/клас моделей, які залежать від параметрів *w*);

– *параметрична* (у термінології машинного навчання та теорії нейронних мереж «навчання»).

Відхилення виходу моделі y_i на *i*-м прикладі та відповідної вказівки вчителя $\tilde{y}_i$ при поточному вхідному векторі $x_i \in \mathbb{R}^n$ (нев'язка, *residual*):

$$\varepsilon_i(w) = y_i(w) - \tilde{y}_i. \qquad (2)$$

Вектор нев'язок:

$$\varepsilon = \varepsilon(w) = y(w) - \tilde{y}. \qquad (3)$$

Нехай відомі значення вектора параметрів $w \in \mathbb{R}^s$. Миттєвий функціонал якості для прикладу *i*:

$$Q_i = Q_i(w) = Q\big(\varepsilon_i(w)\big) \qquad (4)$$

Ступінь відповідності моделі даних навчальної множини: інтегральний функціонал якості навчання:

$$Q = Q(w) = \sum_{i=1}^{k} Q_i(w) \qquad (5)$$

Мета навчання:

$$w^* = \arg\min_{w \in \mathbb{R}^s} Q(w) \qquad (6)$$

Процес навчання може бути формалізовано. Метод навчання $\mu: (X \times Y)_k \rightarrow A$ по вибірці $\boldsymbol{XY}_k = (x_i, y_i)$, $i \in [1 \dots k]$ будує алгоритм $\boldsymbol{a} = \mu(\boldsymbol{X}\ \boldsymbol{Y}_k)$:

$$\begin{pmatrix} f_1(x_1) \dots f_N(x_1) \\ f_1(x_k) \dots f_N(x_k) \end{pmatrix} y \ \rightarrow \ \begin{pmatrix} y_1 \\ y_k \end{pmatrix} \mu \ \rightarrow \ a \qquad (7)$$

Наступний етап: застосування навченої моделі, він формалізується наступним чином. Алгоритм $\boldsymbol{a}$ для нових об'єктів $\boldsymbol{X}' = \{x_1' \dots x_k'\}$ видає відповіді $y_i' = \boldsymbol{a}(x'_i), i \in [1 \dots k]$:

$$\begin{pmatrix} f_1(x'_1) \dots f_N(x'_1) \\ f_1(x'_k) \dots f_N(x'_k) \end{pmatrix} a \ \rightarrow \ \begin{pmatrix} y'_1 \\ y'_k \end{pmatrix} \qquad (8)$$

Для ймовірнісної постановки задач машинного навчання були визначені наступні етапи:

1. Відновлення розподілу ймовірності: $p(x, y)$ – для генеративних моделей, $p(y|x)$ - для дискретних моделей.
2. Визначення розподілу із параметричного сімейства $p(x, y) = \varphi(x, y, w)$, де параметри w відновлюються за навчальною вибіркою X^l.

3. Застосування принципу максимальної правдоподібності для знаходження таких параметрів w, при яких ймовірність усієї вибірки максимальна

$$p(X^l) = p\left((x^1, y^1), (x^2, y^2), \dots, (x^l, y^l)\right) = p(x^1, y^1) \dots p(x^l, y^l)$$

$$Likelyhood(w, X^l) \prod_{i=1}^{l} \varphi(x_i, y_i, w) \rightarrow \min_w L$$

$$-\ln\left(Likelyhood(w, X^l)\right) = -\sum_{i=1}^{l} \ln\left(\varphi(x_i, y_i, w)\right) \rightarrow \min_w L$$

При використанні методів ймовірнісно-статистичного аналізу для вирішення задач машинного навчання зустрічаються обмеження пов'язані з наявністю різного типу невизначеностей у вхідних даних. Вони залежать від ряду факторів і не дають можливість зробити відповідні припущення і встановити закони розподілу невизначених чинників, та зробити висновок про вплив окремих вхідних величин на результат. В задачах аналізу даних присутні невизначеності різного типу, такі як неточність і невизначеність різних параметрів системи, недостатність інформації про систему, нелінійність, нестаціонарність і стохастичність процесів, що відбуваються в системі, вирішення завдань ймовірнісно-статистичного аналізу погано структурованою та важкою для формалізації.

Ймовірнісно-статистичний аналіз процесів, подій, даних різних типів передбачає два підходи:

– *частотний*, який ґрунтується на ймовірнісному аналізі за класичним підходом – дані накопичуються у процесі виконання експериментів (або збору статистики) і обробляються так званими частотними методами теорії ймовірностей;

– *байєсівський*, в основу якого покладається той чи інший варіант теореми Байєса; цей підхід не виключає використання класичних методів, а інформація може бути подана у вигляді статистичних (експериментальних) даних, експертних оцінок, окремих фактів і та інші.

Основою підходу до вирішення задач ймовірнісно-статистичного аналізу даних є *байєсівський* аналіз. Байєсівський підхід до побудови моделей та формування ймовірнісного висновку передбачає використання етапів, представлених на рисунку 3. У байєсівській статистиці передбачається, що інформація надходить з двох джерел: апріорна інформація від дослідника стосовно досліджуваної задач і статистичні дані, отримані в результаті виконання експериментів.

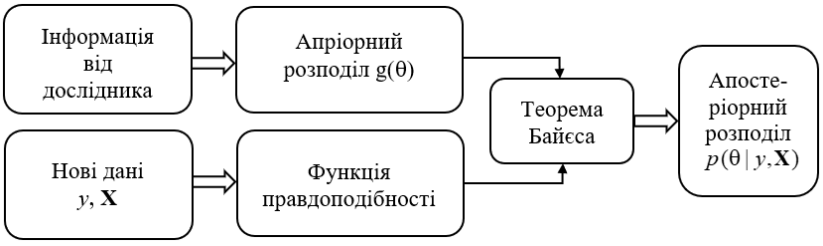

Рисунок 3 – Схема моделювання на основі теореми Байєса

Узагальнено процес байєсівського аналізу даних представлено наступними кроками:

Крок 1. Створення моделі повної ймовірності.

Крок 2. Обробка даних, які спостерігаються.

*Крок3.*Оцінка відповідності моделі та наслідків отриманого апостеріорного розподілу.

Для більш ефективного вирішення задач машинного навчання з врахування нелінійностей та нестаціонарностей даних, та для системного використання сукупності інформаційних технологій: інформаційних технологій ймовірнісно-статистичного аналізу та попередньої обробки даних, інформаційних технологій моделювання, інформаційних технологій прогнозування вперше розроблено метод синтезу інформаційних технологій (рис.4).

Рисунок 4 – Метод синтезу інформаційних технологій для розв’язування задач машинного навчання з врахуванням нелінійностей та нестаціонарностей

Залежно від постановки завдання, яке вирішуються, метод синтезу інформаційних технологій можливо застосовувати для різних задач машинного навчання. Врахування особливостей кожної задачі дозволяє використовувати різні елементи інформаційних технологій. Кожна інформаційна технологія заснована на системному використанні інструментальних методів, які можуть вирішувати окремі задачі машинного навчання з врахуванням нелінійностей та нестаціонарностей даних.

У третьому розділі (*Створення інформаційних технологій ймовірнісно-статистичного аналізу та попередньої обробки даних*) розглянуто головні задачі і методи аналізу та попередньої обробки даних в задачах машинного навчання.

Рисунок 5 – Схема методу обробки пропусків

Визначено, що основними задачами аналізу та попередньої обробки даних є обробка пропусків в наборах даних, виявлення та обробка екстремальних значень в наборах даних, фільтрація, згладжування, нормалізація та стандартизація даних.

Розроблено метод обробки пропусків, який ідентифікує пропуски в даних, виявляє закономірності їх появи та формує набори даних без пропусків. Метод обробки пропусків даних складається з трьох етапів (рис.5). Першим етапом – є ідентифікація відсутніх даних. Якщо вихідний набір даних повністю укомплектований, то, минаючи процедуру обробки пропусків, переходять до наступної процедури аналізу даних. Аналіз, чому дані відсутні, залежить від розуміння процесів, які відтворюють експериментальну інформацію. Однак, за наявності пропусків, наступним (другим) етапом процедури є дослідження закономірностей появи пропущених значень. На цьому етапі розглядаються три підходи до оцінювання можливих закономірностей пропусків: матричний підхід, графічний підхід та елементи кореляційного аналізу. Додаткову статистичну інформацію про пропуски надає матричний підхід, а графічний підхід дозволяє відобразити на графіках закономірності появи відсутніх спостережень. Третім, останнім, етапом процедури обробки пропусків в даних є безпосереднє формування набору даних без пропусків за рахунок використання одного з відомих підходів (видалення даних/змінної, ігнорування пропусків, методи заміни, прогнозування відсутніх значень, для часових рядів: на основі регресійних моделей, на основі метода LOCF та на основі методу згладжування Калмана).

Розроблено метод виявлення та обробки аномальних значень в наборах даних, який ідентифікує аномалії в даних, аналізує причини їх появи, та здійснює їх обробку. При формуванні початкової вибірки виникають проблеми наявності

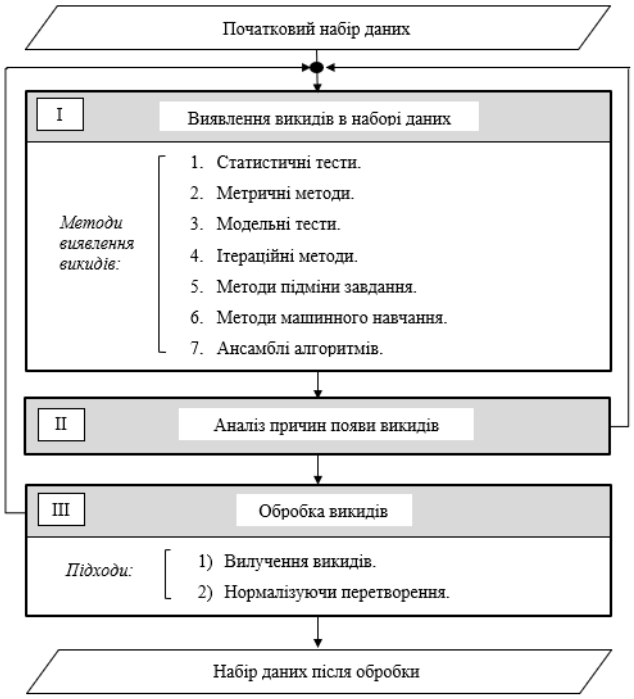

Рисунок 6 – Схема методу пошуку викидів

частини даних, які відрізняються від загальної вибірки та знаходяться на далекій відстані. Такі дані називають викидами. Вони можуть відповідати реальним відхиленням, але можуть бути помилками.

Метод ідентифікації та обробки викидів представлено на рисунку 6. Він складається з трьох етапів. На *першому* етапі процедури виявляється наявність викидів в наборі даних. Для ідентифікації та виявлення тестів використовуються: статистичні тести, метричні методи, модельні тести, ітераційні методи, методи підміни завдання, методи машинного навчання та ансамблі алгоритмів. На *другому* етапі відбувається аналіз причин появи викидів. На *третьому* етапі здійснюється обробка наборів даних з викидами. Для цього використовують два підходи: видалення аномальних значень або виконання нормалізуючих перетворень.

Одним із підходів, який використовується при аналізі і попередньої обробки даних є підхід на основі системного використання методів фільтрації даних. Головна мета етапу фільтрації – виділити корисну частину спектру даних для подальшої обробки та моделювання і затримати шумову або просто непотрібну для аналізу складову частину набора даних. Найбільш розповсюджені наступні типи фільтрів: цифрові, оптимальні і ймовірнісні фільтри (див. рис.7).

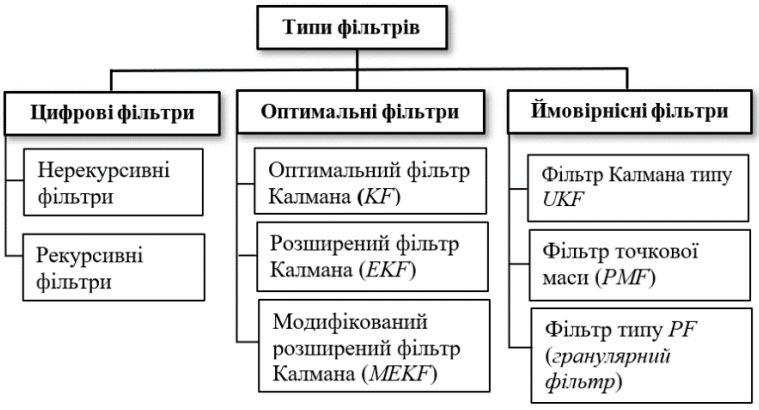

Рисунок 7 – Класифікація фільтрів

Запропоновано метод для вирішення завдань фільтрації статистичних даних на основі використання різних типів фільтрів за рахунок модифікації алгоритмів гранулярної фільтрації (табл.1).

Таблиця 1 – Алгоритми гранулярної фільтрації

№	Назва алгоритму	Псевдокод алгоритму	№	Назва алгоритму	Псевдокод алгоритму
1	Алгоритм послідовної вибірки за важливістю (SIS). *Базовий*	Algorithm 1: SIS Particle Filter $[\{x^i(k), w^i(k)\}_{i=1}^{N_S}]$ = SIS $[\{x^i(k-1), w^i(k-1)\}_{i=1}^{N_S}, z(k)]$ FOR – generate $x^i(k) \sim q(x(k) \mid x^i(k-1), z(k))$. – assign the particle $x^i(k)$ weight $w^i(k)$ according to (28) END FOR	3	Алгоритм послідовної вибірки за важливістю (SIS). *Повторний відбір проб частинок*	Algorithm 2: Resampling Algorithm $[\{x^{j*}(k), w^j(k), i^j\}_{j=1}^{N_S}]$ = RESAMPLE $[\{x^i(k), w^i(k)\}_{i=1}^{N_S}]$ Initialize distribution function (DF): $c(1) = 0$ FOR $i = \overline{2, N_S}$ – Construct DF: $c(i) = c(i-1) + w^i(k)$ END FOR Start DF from beginning: $i = 1$ Generate initial point: $u(1) \sim U[0, N_S^{-1}]$. FOR $j = \overline{1, N_S}$ – Move along DF: $u(j) = u(1) + N_S^{-1}(j-1)$ – WHILE $u(j) \succ c(i)$ – – $i = i+1$ – END WHILE – Assign new value: $x^{j*}(k) = x^j(k)$ – Assign weight: $w^j(k) = N_S^{-1}$ – Assign basic index: $i^j = i$ END FOR
2	Послідовна вибірка за важливістю з фільтром повторної вибірки (SISR)	Algorithm 3: SIR Particle Filter $[\{x^i(k), w^i(k)\}_{i=1}^{N_S}]$ = SIR$[\{x^i(k), w^i(k)\}_{i=1}^{N_S}, z(k)]$ FOR $i = \overline{1, N_S}$ – Generate $x^i(k) \sim p(x(k) \mid x^i(k-1))$ – Compute $w^i(k) = p(z(k) \mid x^i(k))$ END FOR Compute total weight: $t = \sum_{i=1}^{N_S} w^i(k)$ FOR $i = \overline{1, N_S}$ – Normalize i-th weight: $w^i(k) = t^{-1} w^i(k)$ END FOR Perform resampling using algorithm 2 (Resampling Algorithm): – $[\{x^i(k), w^i(k), -\}_{i=1}^{N_S}]$ = RESAMPLE $[\{x^i(k), w^i(k)\}_{i=1}^{N_S}]$			

Приклад комплексного використання методів фільтрації в умовах стохастичних невизначеностей представлено на рисунку 8. Блок фільтрації даних є частиною підсистеми аналізу та попередньої обробки даних. В блоку фільтрації передбачено комплексне використання всіх типів фільтрів: цифрових, оптимальних та ймовірнісних. В таблицях 2-3 підкреслено підвищення показників якості прогнозу при застосуванні блока фільтрації.

Рисунок 8 – Схема застосування фільтрів у системі обробки

Таблиця 2 – Якість моделей та прогнозів без застосування блока фільтрації

Тип моделі	Якість моделі			Якість прогнозу			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0.99	26655.77	2.21	49.93	43.57	8.49	0.047
AR(1,4)	0.99	25487.25	2.18	49.12	40.18	8.28	0.046
AR(1) + 1st order trend	0.99	25391.39	2.13	34.31	24.26	4.31	0.030
AR(1) + 4th order trend	0.99	25088.74	2.11	24.89	18.32	3.05	0.022

Таблиця 3 Якість моделей та прогнозів із застосуванням блока фільтрації

Тип моделі	Якість моделі			Якість прогнозу			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0.99	24376.32	2.11	45.21	39.73	7.58	0.037
AR(1,4)	0.99	24141.17	2.09	47.29	38.75	7.06	0.035
AR(1) + 1st order trend	0.99	23964.73	2.08	31.15	22.11	3.27	0.029
AR(1) + 4th order trend	0.99	22396.83	2.04	21.35	13.52	2.71	0.019

Розвинуто підхід до нормалізації та стандартизації даних на основі системного поєднання методів перетворення даних та особливостей вирішення завдань машинного навчання. Схематично підхід представлено на рисунку 9. Необхідність нормалізації вибірок даних обумовлена природою алгоритмів і моделей машинного навчання.

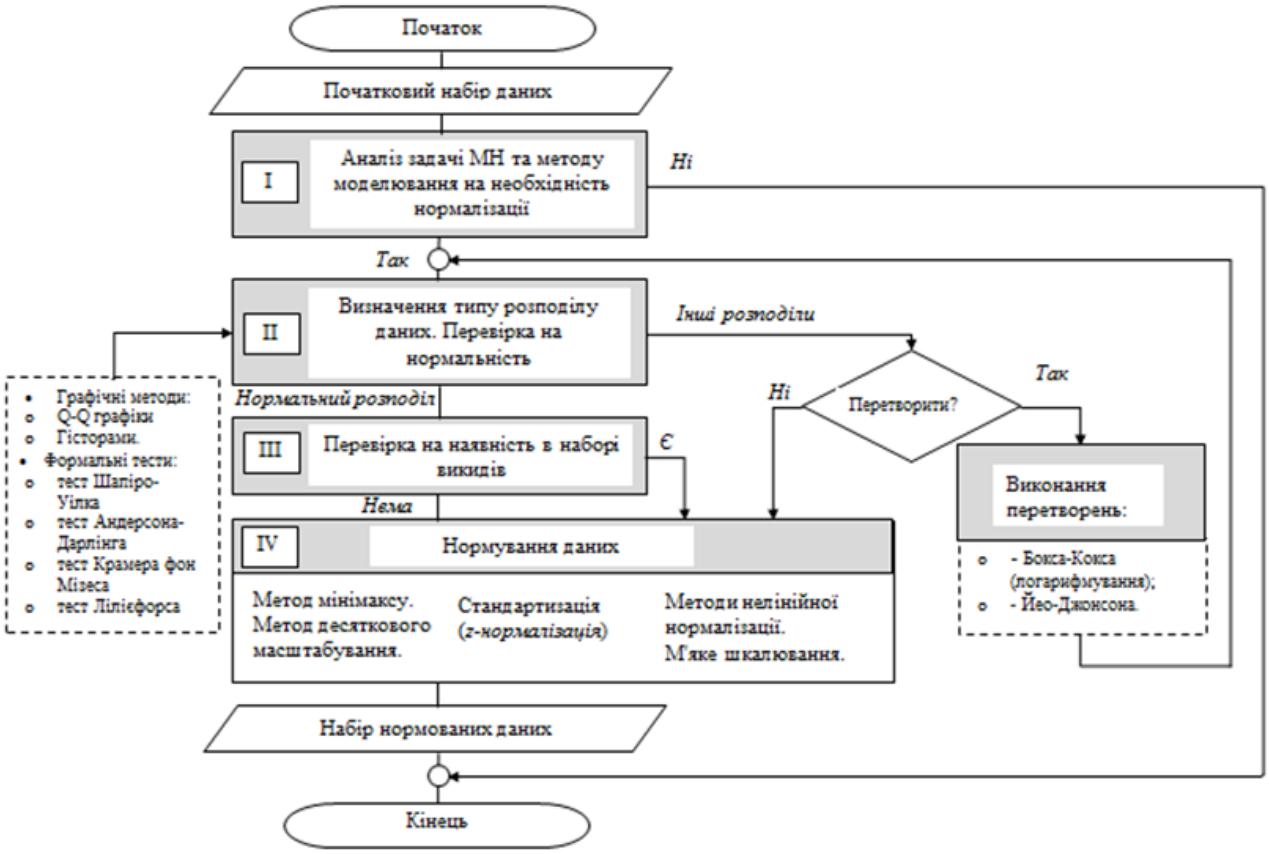

Рисунок 9 – Системний підхід для вирішення завдань нормалізації

Підхід складається з чотирьох етапів. *На першому етапі* аналізується початковий набір даних, задача машинного навчання та метод моделювання на необхідність нормалізації. *На другому етапі* визначається тип розподілу даних. Та проводиться перевірка на нормальність. Тип розподілу визначається за допомогою графічних методів: квантильних Q-Q графіків та гістограм. Перевірка на нормальність здійснюється за допомогою формальних тестів: теста Шапіро-Уілка, непараметричного критерія і теста Андерсона-Дарлінга, теста Крамера фон Мізеса, та теста Лілієфорса. Якщо розподіл відрізняється від нормального, то виконуються перетворення Бокса-Кокса або Йео-Джонсона. Якщо перетворення не дасть результатів, то необхідно використати методи нелінійної нормалізації. *На третьому етапі* здійснюється перевірка на наявність в наборі викидів. Якщо викиди існують, то використовують методи нелінійної нормалізації. *На четвертому етапі* безпосередньо відбувається нормалізація даних.

Методи нормалізації можуть бути *лінійними* та *нелінійними.* Лінійна нормалізація використовується переважно у тому випадку, коли значення змінної рівномірно заповнюють певний інтервал. Якщо в даних є *аномалії*, які значно

перевищують типову різницю, то в цих випадках при нормуванні орієнтуються не на мінімальне та максимальне значення, а на *середнє* і *дисперсію.*

Для вирішення задач ймовірнісно-статистичного аналізу та попередньої обробки даних в задачах машинного навчання розроблено інформаційна технологія (рис.10) для аналізу та попередньої обробки даних, яка ґрунтується на системному використанні запропонованих методів та підходів: обробки пропусків даних, ідентифікації та обробки екстремальних значень, фільтрації, згладжування, нормалізації та стандартизації даних. Інформаційна технологія ймовірнісно-статистичного аналізу та попередньої обробки даних об'єднує групи методів та методологічних підходів, які згруповані по функціональному призначенню в процесі аналізу та підготовки даних до моделювання.

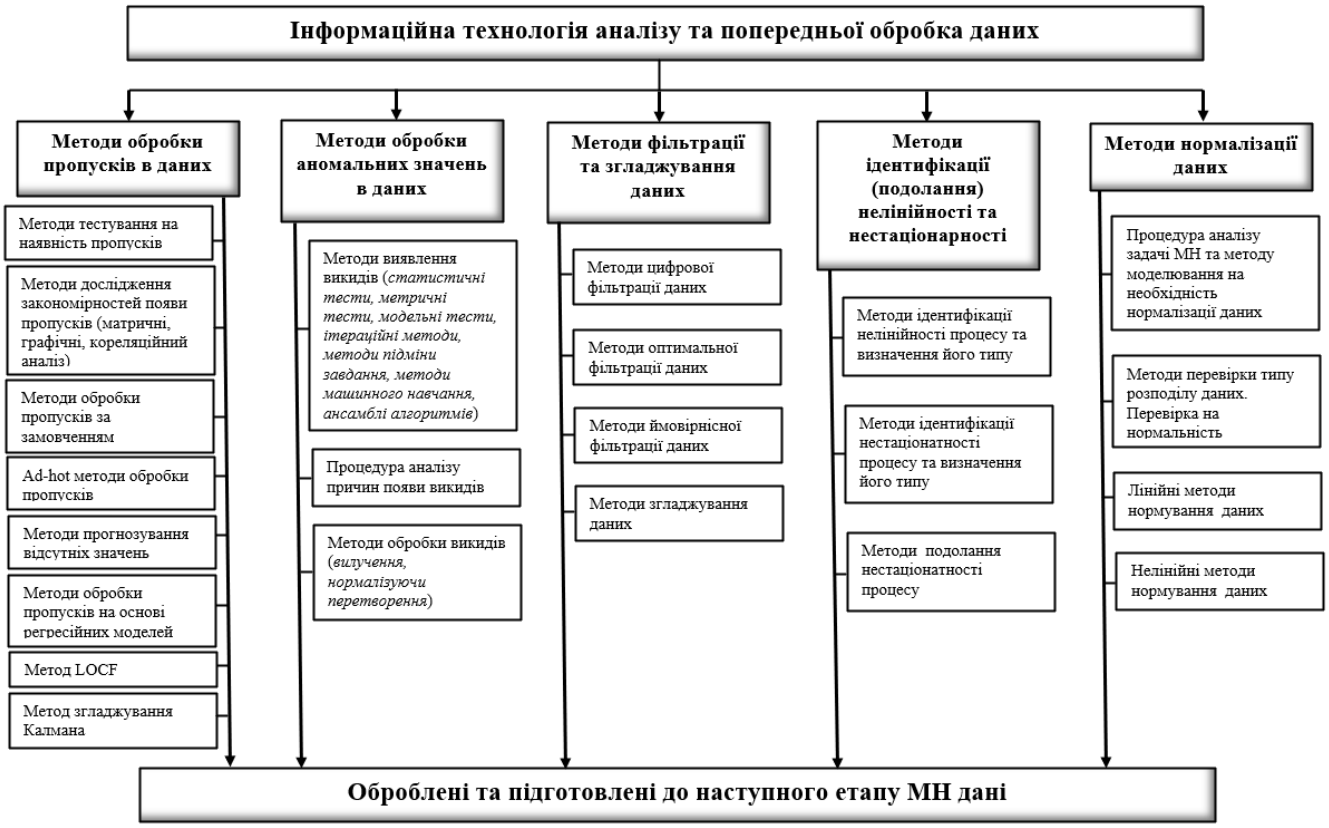

Рисунок 10 – Інформаційна технологія ймовірнісно-статистичного аналізу та попередньої обробки даних

Для практичного вирішення задач ймовірнісно-статистичного аналізу та попередньої обробки даних на основі інформаційної технології розроблено архітектуру інформаційно-аналітичної (рис. 11). Результатом використання методів і методологічних підходів інформаційної технології ймовірнісно-статистичного аналізу та попередньої обробки даних є оброблені та підготовлені до наступного етапу машинного навчання дані.

В **четвертому розділі** (*Створення інформаційних технологій моделювання нелінійних та нестаціонарних даних в процедурах машинного навчання*) представлені результати досліджень, виконаних при побудові моделей обробки

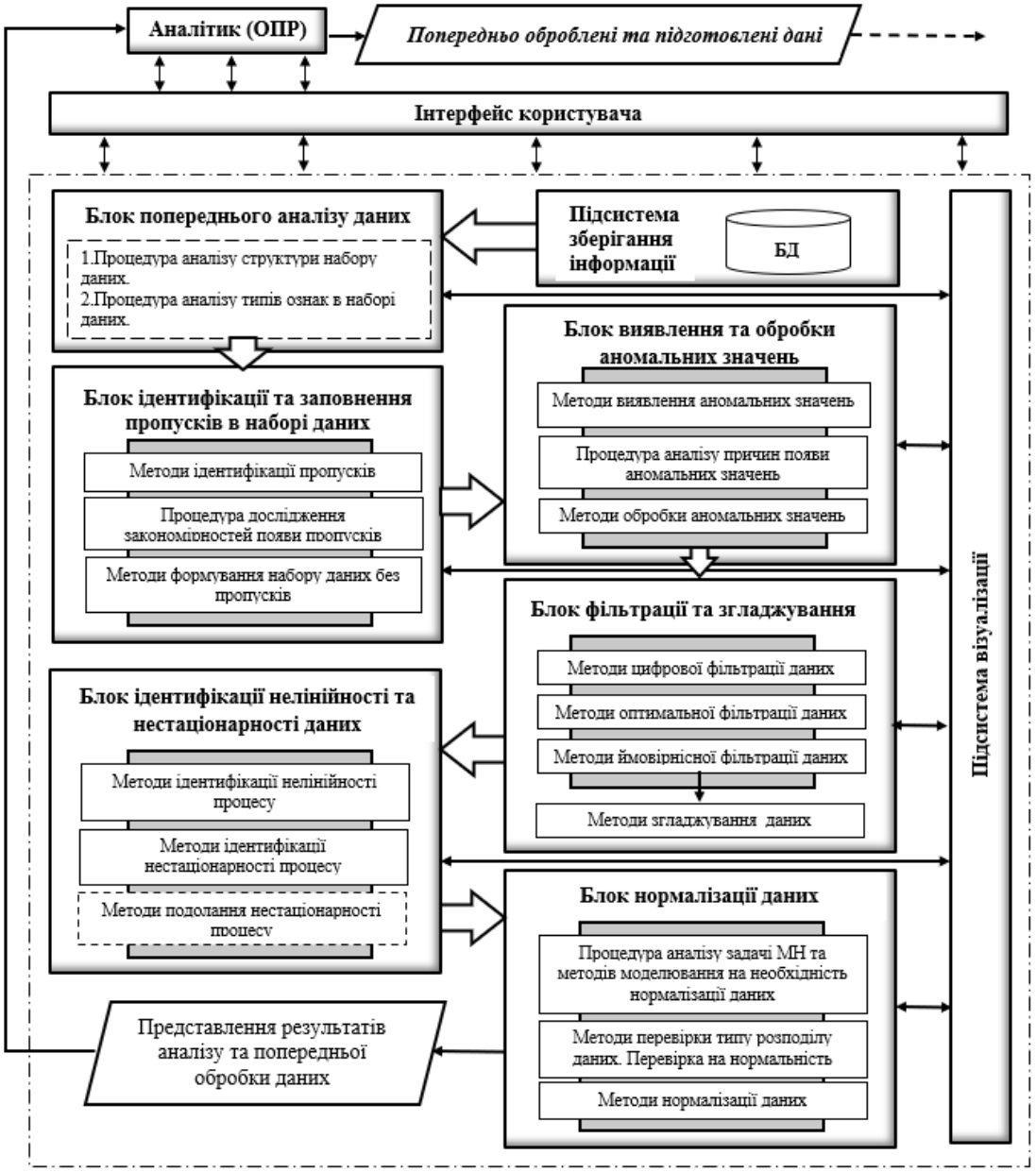

Рисунок 11 – Інформаційно-аналітична система ймовірнісно-статистичного аналізу та попередньої обробки даних

нелінійних та нестаціонарних даних та при використанні і побудові математичних моделей для вирішення завдань машинного навчання.

Вперше розроблено метод до побудови моделей на основі байєсівських часових рядів для вирішення задач машинного навчання. Структурні моделі часових рядів мають три ключові переваги для моделювання нелінійних нестаціонарних процесів:

– Можливість визначати невизначеність у прогнозах, у зв'язку з чим потім кількісно оцінити майбутні ризики.

– Прозорість для розуміння механізму роботи моделі.

– Можливість включення зовнішньої інформації для відомих факторів, коли не проглядається зв'язок у наявних даних.

Модель структурних часових рядів може бути описана парою рівнянь (9). Перше, *рівняння спостереження*, пов'язує дані y_t з вектором латентних змінних α_t, який називають «станом». Друге, *рівняння переходу* описує, як латентний стан розвивається в часі:

$$y_t = Z_t^T \alpha_t + \epsilon_t, \quad \epsilon_t \sim N(0, H_t), \qquad (9)$$
$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t).$$

Матриці моделі Z_t, T_t та R_t є структурними параметрами. Вони містять суміш відомих значень (часто 0 та 1) і невідомих параметрів. Матриця переходу T_t це квадратна матриця, а матриця R_t може бути прямокутною, якщо частина переходів станів будуть детермінованими. Наявність R_t у рівнянні дозволяє працювати з матрицею дисперсії повного рангу Q_t, оскільки будь-які лінійні залежності у векторі стану можна перемістити з Q_t R_t. Часто під час реалізації H_t – позитивна скалярна величина. Залишки ϵ_t та μ_t незалежні один від одного і мають нормальний розподіл із середнім 0. Прийнято вважати, що модель, яка може бути описана такою парою рівнянь та знаходиться у формі простору станів. У формі простору станів може бути виражений великий клас моделей, включаючи всі різновиди моделей ARIMA і VARMA.

Підхід до побудови моделей на основі байєсівських часових рядів для вирішення завдань машинного навчання складається з наступних етапів. На першому етапі відбувається процес навчання байєсівської структурної моделі часових рядів виконується у чотири кроки:

1. Завдання структури моделі та апріорних ймовірностей.
2. Застосування фільтра Калмана для оновлення оцінок стану на основі даних, що спостерігаються.
3. Застосування методу «spike-and-slab» для вибору змінних у структурній моделі.
4. Усереднення за байєсівською моделлю для об'єднання результатів з метою складання прогнозу.

Гнучкість моделі BSTS, заснованої на виборі модульних компонентів, проявляється на перших двох кроках. На наступних кроках модель навчається на наявних даних за допомогою байєсівського методу, що оновлює оцінку параметрів з часом. На другому етапі набору даних становлять у відповідність кілька альтернативних BSTS-моделей, спираючись на результати попереднього аналізу та обробки даних. Кожна з моделей комплектується компонентами, здатними відобразити характер змін у даних. На третьому етапі відбувається побудова BSTS-моделі за наступним алгоритмом:

1. Визначення набору компонентів моделі.
2. Якщо предикторів у моделі немає, список апріорних ймовірностей відповідає апріорному розподілу стандартного розподілу залишків моделі. Якщо модель з предикторами, то апріорні розподіли створюються методом «spike-and-slab».
3. Завдання числа ітерацій алгоритму MCMC та параметрів генератора випадкових чисел (для відтворюваності результатів обчислень).
4. Побудова BSTS-моделі.

5. Оцінка якості моделі та перевірка її адекватності: за допомогою метрик, візуалізації результатів припасування моделі та її компонентів, перевірка відсутності автокореляції в залишках моделі.

Визначено різні компоненти моделі BSTS, за допомогою яких формуються структури альтернативних прогнозних моделей. Всі готові альтернативні BSTS-моделі зазнають порівняння та оцінки для відбору найбільш адекватних вихідному набору даних. Відібрані моделі є основою для складання прогнозів.

Вперше розроблено метод побудови моделей аналізу та моделювання нелінійних нестаціонарних процесів на основі колірних мереж Петрі. При моделюванні складних процесів генерації та обробки даних використовують методи імітаційного моделювання. Для розширення можливостей аналізу складних систем та процесів, в тому числі для моделювання систем з складним стохастичним процесом обробки заявок використовується моделювання на основі колірних мереж Петрі, головною особливістю яких передбачається врахування змінних різного типу та умов спрацювання переходів

Модель на основі колірної мережі Петрі має наступний вигляд:

$$PN_{col} = \{S, T, F, M_0, Type, Type\_T, Type\_M, In\}$$

де $S = \{S_1, S_2, \dots, S_g\}$ – множина станів; $T = \{t_1, t_2, \dots, t_v\}$ – множина переходів; F – множина дуг, яка включає підмножини вхідних та вихідних дуг по відношенню до переходу; M_0 – множина, в якої задається початкове маркування мережі Петрі, *Type* – множина типів даних; *Type_T* – множина, яка відображає доступну множину типів даних у позиціях мережі; *Type_M* – множина типів маркерів, що ініціюють перехід ; *In* – множина умов ініціації переходів.

Для моделювання задач обробки даних також можливе використання *часових* мереж Петрі, що є вдосконаленням мережі Петрі і пов'язано з додаванням до кожного з переходів інформації про часові межі. Це дозволяє визначити і детально описати часові проміжки процесу, який моделюється. Часова мережа описується за допомогою наступного виразу:

$$N_{time} = \{S, T, F, Eft, Lft, M_0\},$$

де Eft, Lft – функції, що ставляться у відповідність кожному з переходів і визначають нижню (Eft) та верхню (Lft) часові межі, які задовольняють наступним умовам: $Eft \leq Lft$. Модель, яка враховує пріоритети, включає множину пріоритетів для кожного з переходів і має наступний вигляд:
$PN_{pr\_time} = \{S, T, F, Eft, Lft, PR, M_0\}$, де $PR = \{Pr_1, Pr_2, \dots, Pr_v\}$ – множина пріоритетів, а Pr_{1-v} – величини пріоритетів.

Алгоритм побудови імітаційних моделей на основі колірних мереж Петрі складається з наступних кроків:

1.Визначається множина станів $S = \{S_1, S_2, \dots, S_g\}$, та множина переходів $T = \{t_1, t_2, \dots, t_v\}$, які відповідають процесу, що моделюється.

2.Визначається *F*, та задається початкове маркування мережі M_0.

3.Визначаються можливі зміни ситуації, та відображаються у *Type_T* множині типів у позиціях мережі.

4.Визначаються маркери, які ініціюють переходи, і умови ініціалізації переходів *In* (часові).

5. Формується остаточно структура мережі.

На основі колірних мереж Петрі доцільно розробляти імітаційні моделі нелінійних та нестаціонарних процесів генерації та обробки даних, приймати рішення на підставі моделювання, а також перевіряти результати моделювання.

Отримав подальший розвиток метод синтезу параметрів нелінійної прогнозної моделі за допомогою генетичного алгоритму. Метод складається з 7 кроків:

1. Аналіз вхідного набору даних.
2. Вибір структури і показників моделі на основі аналізу кластерів вхідного набору даних.
3. Визначення аналітичних залежностей між показниками.
4. Завдання fitness-функції та параметрів генетичного алгоритму для підбору параметрів нелінійної залежності між ознаками.
5. Виконання генетичного алгоритму.
6. Візуальний аналіз підбору параметрів моделі.
7. Оцінювання якості моделі.

Запропонована схема інформаційних технологій для моделювання в задачах машинного навчання Інформаційна технологія моделювання побудована на основі системного підходу і об'єднує групи методів та методологічних підходів, які згруповані по функціональному призначенню в процесі побудови моделей для вирішення завдань машинного навчання. Структура інформаційної технології моделювання представлена на рис. 12.

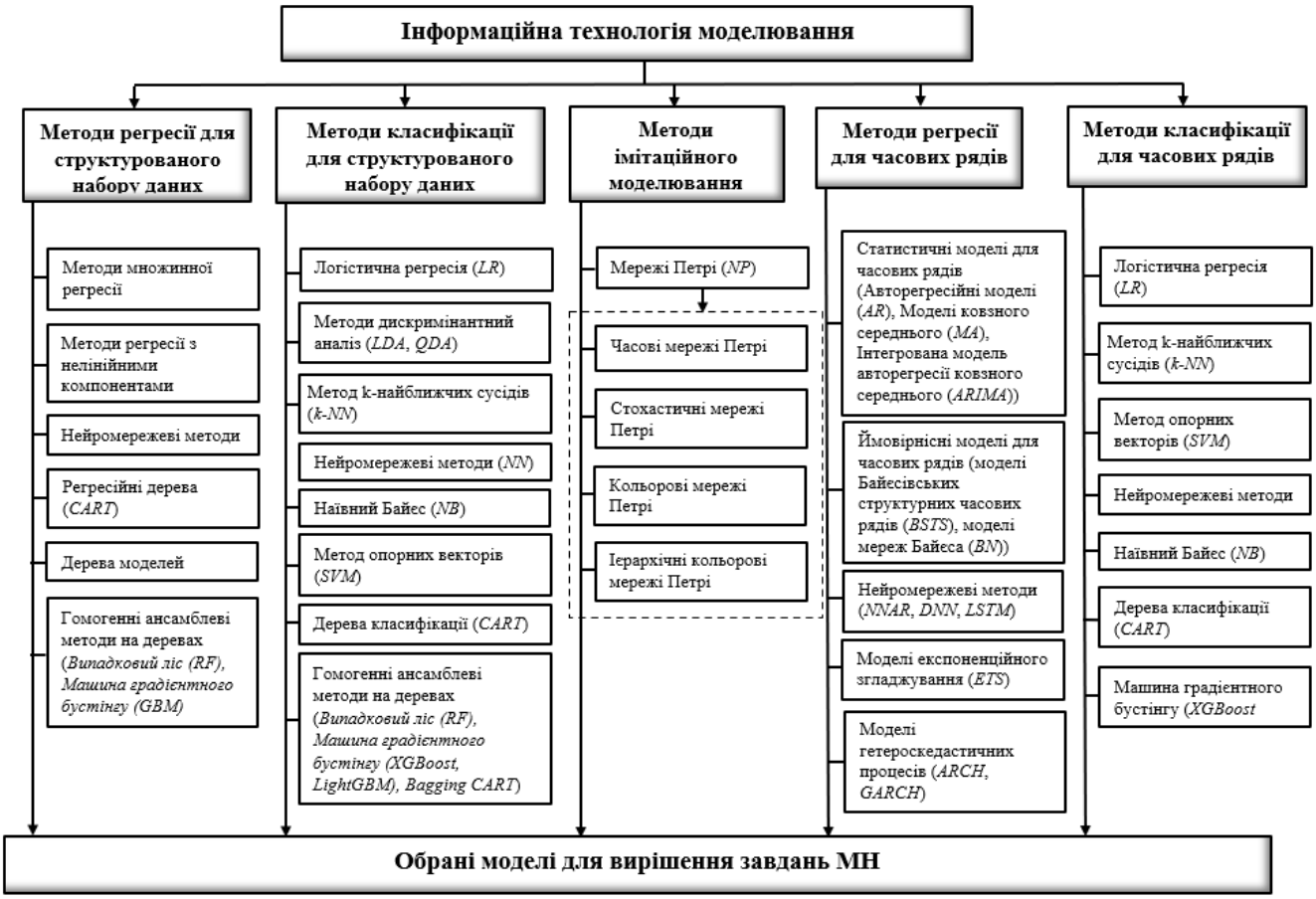

Рисунок 12 – Інформаційна технологія моделювання

Інформаційна технологія об’єднує наступні групи методів: методи регресії для структурованих наборів даних; методи класифікації для структурованих наборів даних; методи імітаційного моделювання; методи регресії для часових рядів; методи класифікації для часових рядів.

Результатом використання методів і методологічних підходів інформаційної технології моделювання є підготовлені моделі для вирішення задач машинного навчання. На основі інформаційної технології розроблено архітектура інформаційно-аналітичної системи моделювання для вирішення задач машинного навчання (рис.13).

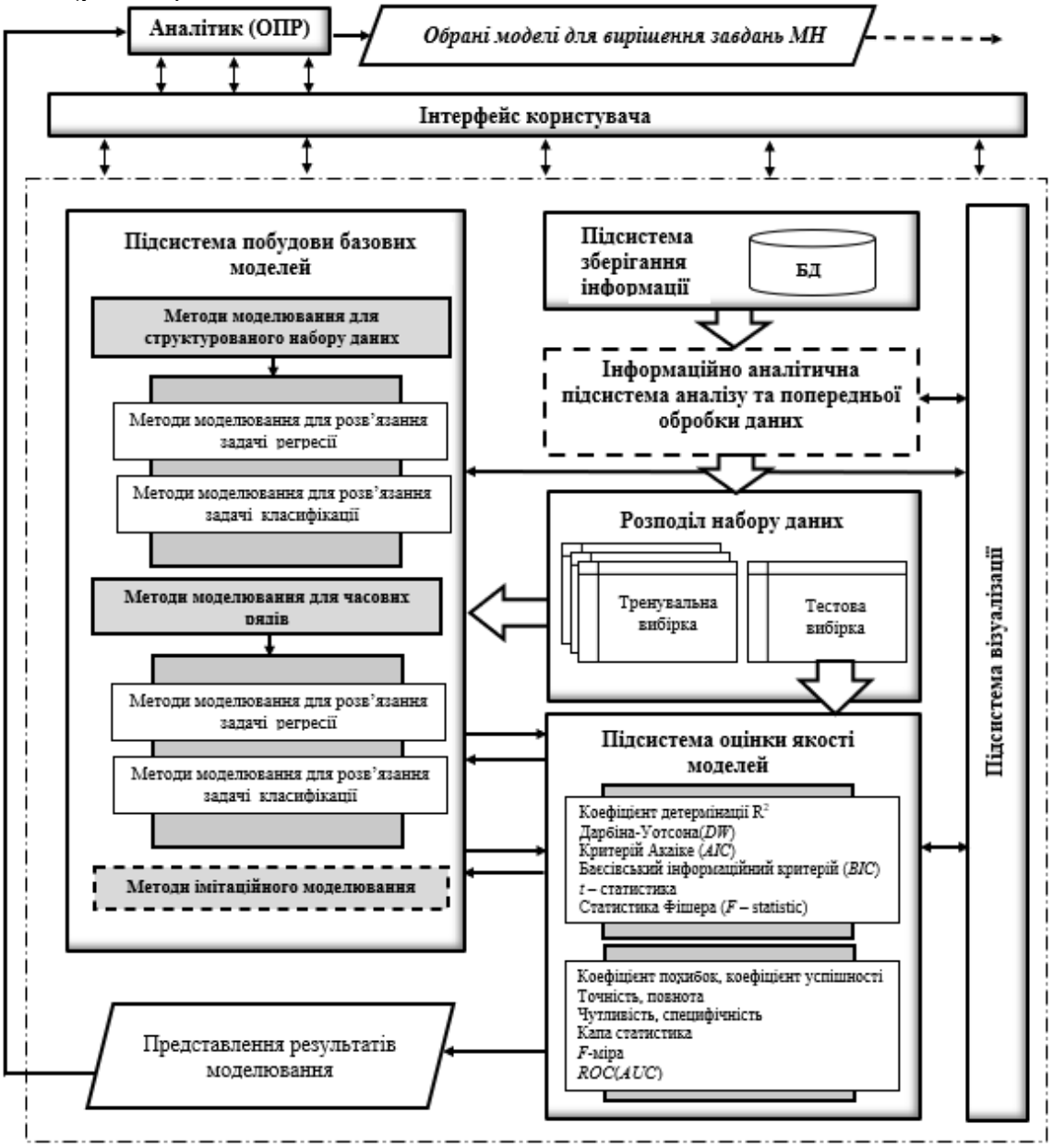

Рисунок 13 – Інформаційно аналітична система моделювання

У **п’ятому розділі** (*Створення інформаційних технологій прогнозування на основі ймовірнісно-статистичної обробки нелінійних та нестаціонарних даних в процедурах машинного навчання*). Вперше розроблено системний підхід до

розв’язання завдань моделювання та прогнозування на основі ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних в процедурах машинного навчання (рис. 14). Він об’єднує на системній основі методи та методології, направлені на вирішення задач: аналізу та попередньої обробки даних; побудови моделей та їх оцінки; побудови прогнозів та процедур їх оцінювання. Також він ґрунтується на аналізі досліджуваного процесу, встановленні типів наявних характерних невизначеностей, оцінюванні структури і параметрів моделі, та прогнозів.

Системний підхід є методологічної базою для побудови інформаційних

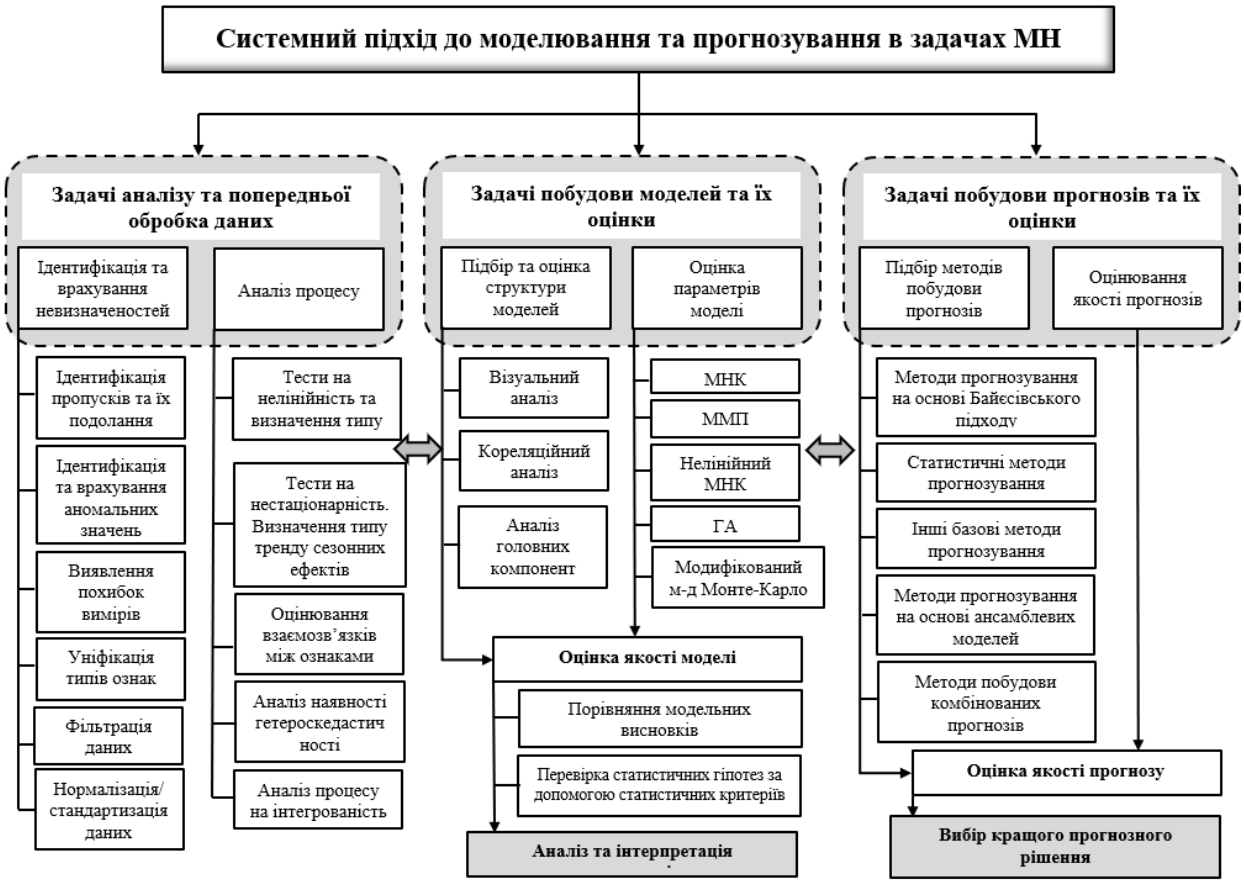

Рисунок 14 – Системний підхід до моделювання та прогнозування в задачах машинного навчання

технологій вирішення завдань ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних та вирішення окремих задач машинного навчання.

Проблема обліку та обробки нелінійностей та нестаціонарностей при прогнозуванні часових рядів є одним із головних завдань побудови адекватних прогнозних моделей досліджуваного процесу. Моделі даного типу повинні включати повне уявлення динаміки нелінійних і нестаціонарних систем на основі спостережуваних реальних даних. Одним із методів обробки нелінійностей та нестаціонарностей є метод байєсівських структурних часових рядів.

Як приклад його застосування розглянуто задача прогнозування цін акцій компанії *Amazon.* Базовим набором даних є *amzn_share*, який містить значення цін

акцій компанії. Дані є частиною багатовимірного часового ряду та взяті сайту *https://finance.yahoo.com/*. Після завантаження проведено аналіз структури та типів даних, виконано обробку пропущених значень за допомогою метода лінійної апроксимації. За допомогою набору статистичних тестів (ADF, KPSS, PP) вихідний ряд перевірено на стаціонарність. Проведено візуальний аналіз даних. Він дозволив визначитися із принципами моделювання.

Після розбиття вхідного набору даних на навчальну та тестову вибірки виконано адаптацію структурних моделей часових рядів з використанням фільтра Калмана та методу Монте-Карло за схемою марковських ланцюгів (MCMC). Проведено дослідження адекватності прогнозних моделей. Розраховано прогнозні значення для побудованих моделей з найкращими показниками якості.

Досліджено особливості побудови прогнозів на основі нейронних мереж прямого розповсюдження та проведено оцінка їх ефективності. Розглянуто прогнозування на основі нейронних мереж для структурованого набору даних та для часових рядів. Розроблено метод підготовки вхідних даних для моделювання та прогнозування на основі нейронних мереж. Розроблено алгоритм побудови прогнозних моделей для часових рядів на основі нейронних мереж прямого розповсюдження.

Для підвищення ефективності розв’язання задачі прогнозування було удосконалено метод побудови і використання комбінованих прогнозів за рахунок ітеративного оцінювання різних схем комбінування. Структурна схема методики побудови комбінованих прогнозів для часових рядів представлено на рисунку 15. Було розглянуто моделі простого усереднення, зваженого усереднення та регресії. За рахунок зменшення дисперсії доведено ефективність такого підходу.

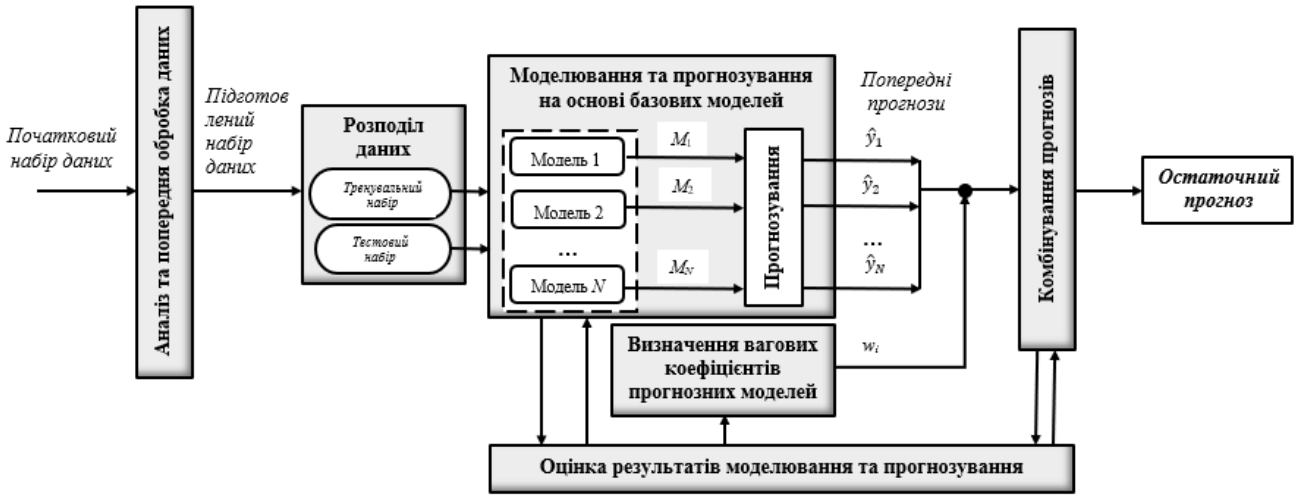

Рисунок 15– Структурна схема методу побудови комбінованих прогнозів на основі часових рядів.

Першим етапом методу є процес аналізу та попередньої обробки набору даних. На цьому етапі реалізуються процедури: виявлення та обробки пропусків в наборі даних, виявлення аномалій, перевірки на наявність нелінійності, нестаціонарності та їх врахування, фільтрації та згладжування даних тощо. Після

цього етапу первинний набір даних повністю підготовлений до процесу моделювання. На другому етапі відбувається поділення набору даних на дві частини: тренувальну та тестову. Наступним етапом є моделювання та прогнозування на основі базових моделей. Базові моделі будуються на основі обраних методів. Вони перевіряються на адекватність за допомогою метрик якості, значення яких передаються в блок результатів оцінювання моделей. Із базових моделей формуються попередні прогнози. Оцінки якості моделей є основою для формування вагових коефіцієнтів при комбінуванні прогнозів. Заключним етапом методу є етап комбінування, на якому визначається спосіб комбінування та оцінюється його ефективність. Якщо покращення точності прогнозу не виявлено, то необхідно повернутись на етап формування базових моделей, або змінити їх кількість та тип комбінування. Запропоновано архітектуру інформаційно-аналітичної системи прогнозування на основі комбінованих прогнозів. Ефективність використання підтверджена прикладом.

Іншим підходом до підвищення якості прогнозних значень є підхід до прогнозування на основі ансамблевих методів. Він дозволяє зменшувати похибки прогнозів за рахунок одночасного зменшення *зміщення* та *дисперсії* на основі використання багаторівневих гетерогенних ансамблів прогнозних моделей.

Для будь-якого перевірочного спостереження $\boldsymbol{x_0}$ математичне сподівання середньоквадратичної помилки його прогнозу можна розкласти на суму трьох величин: дисперсії $\boldsymbol{f(x_0)}$, квадрата зміщення $\boldsymbol{f(x_0)}$ та дисперсії залишків ε:

$$E[y_0 - f(\boldsymbol{x_0})]^2 = Var[f(\boldsymbol{x_0})] + [Bias\big(f(\boldsymbol{x_0})\big)]^2 + Var(\boldsymbol{\varepsilon}),$$

де $Bias$ означає зміщення, а Var – дисперсію.

При підборі оптимальної прогнозної моделі необхідно враховувати компроміс між *зміщенням* та *дисперсією*. Техніка ансамблювання, тобто агрегування прогнозних значень різних базових моделей для створення однієї «оптимальної», є прийомом, що дозволяє пом'якшити компроміс між зміщенням та дисперсією.

Застосування ансамбля моделей це процес, у якому різні та незалежні моделі поєднуються для отримання кращого результату. При побудові ансамблю використовується кілька технік агрегування результатів базових моделей, кожен з яких забезпечує різну точність діагностики. Розглядались три, найбільш поширені, методи агрегування прогнозних значень: *bagging*, *boosting* та *staking*. У таблиці 4 наведено обґрунтування вибору типу моделей для створення гетерогенної багаторівневої ансамблевої структури.

Після експериментального підбору типу ансамблю розроблено дворівневу архітектуру системи класифікації на основі методів *Staking* та *Bagging* (рис. 16). Дворівневий класифікатор на обох рівнях агрегації використовує зважену суму. Ваги розраховуються на основі значення *F*-міри кожного з базових методів. Розроблено алгоритм побудови багаторівневого ансамблю моделей на основі різних базових методів, він складається із 7 кроків:

Крок 1. Розділення підготовленого до моделювання набору даних на дві вибірки (навчальну і тестову).

Крок 2. Навчання *n* різнорідних базових моделей на визначеній навчальній вибірці. Оцінка якості моделей. Підбір оптимальних параметрів для моделей кожного типу.

Крок 3. Розрахунок для кожної з *n* моделей прогнозних значень. Оцінка якості прогнозів за допомогою тестової вибірки.

Крок 4. Обчислення (оцінка) дисперсії та зміщення для кожної з *n* базових моделей.

Крок 5. Розділення *n* базових моделей на дві групи (з високим зміщенням – недонавчені та з високою дисперсією – перенавчені).

Крок 6. Формування на основі *m* ($m<n$) моделей першої групи 1-го рівня ансамблевої структури – *Stacking*.

Крок 7. Формування на основі *l* ($l \le n-m$) моделей другої групи + результат метода *Stacking* 2-го рівня ансамблевої структури – *Bagging*. Результат вважати остаточним прогнозним значенням.

Таблиця 4 – Обґрунтування вибору типу моделей для створення ансамблю

Тип ансамблю	Тип базових моделей	Характер помилки моделі	Результат агрегації
Bagging (модельне усереднення)	однорідні моделі	моделі з низьким зміщенням та високою дисперсією (перенавчені моделі)	зменшує дисперсію
Boosting (модельне підсилювання)	однорідні моделі	моделі з низькою дисперсією та високим зміщенням (недонавчені моделі)	зменшує зміщення
Stacking (модельне накладення)	різнорідні моделі	моделі з низькою дисперсією та високим зміщенням (недонавчені моделі)	зменшує зміщення, але залежно від вибору метамоделі може зменшувати дисперсію

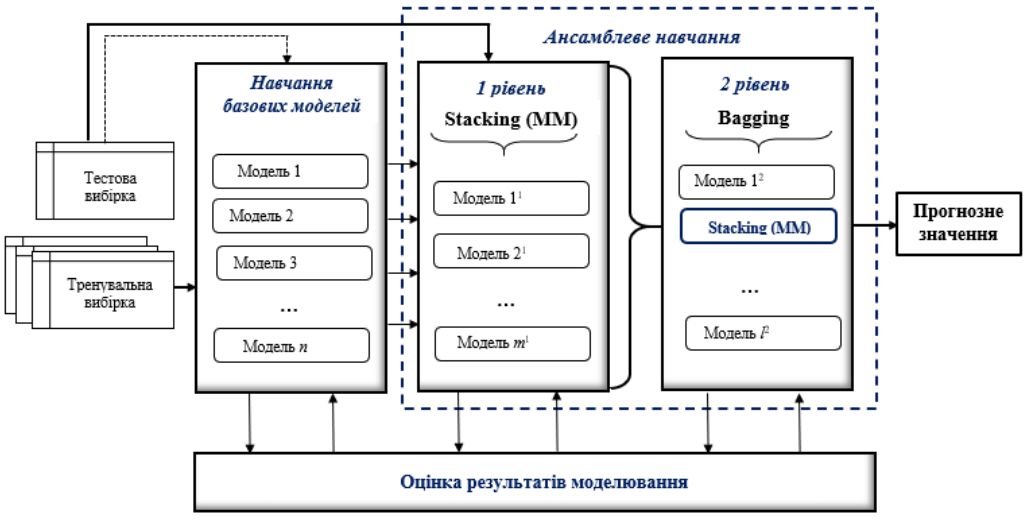

Рисунок 16 – Дворівнева архітектура системи прогнозування на основі *Staking* та *Bagging*

Запропоновано архітектуру інтелектуальної системи класифікації на основі дворівневого гетерогенного ансамбля моделей. Ефективність використання підтверджена прикладами.

На основі системного використання методів та підходів до моделювання і прогнозування та представлених методів розроблено інформаційну технологію прогнозування (рис. 17). Інформаційна технологія прогнозування для задач машинного навчання ґрунтується на запропонованих методах прогнозування по базовим моделям, на статистичних методах прогнозування часових рядів, на байєсівських методах прогнозування, на методах прогнозування на основі

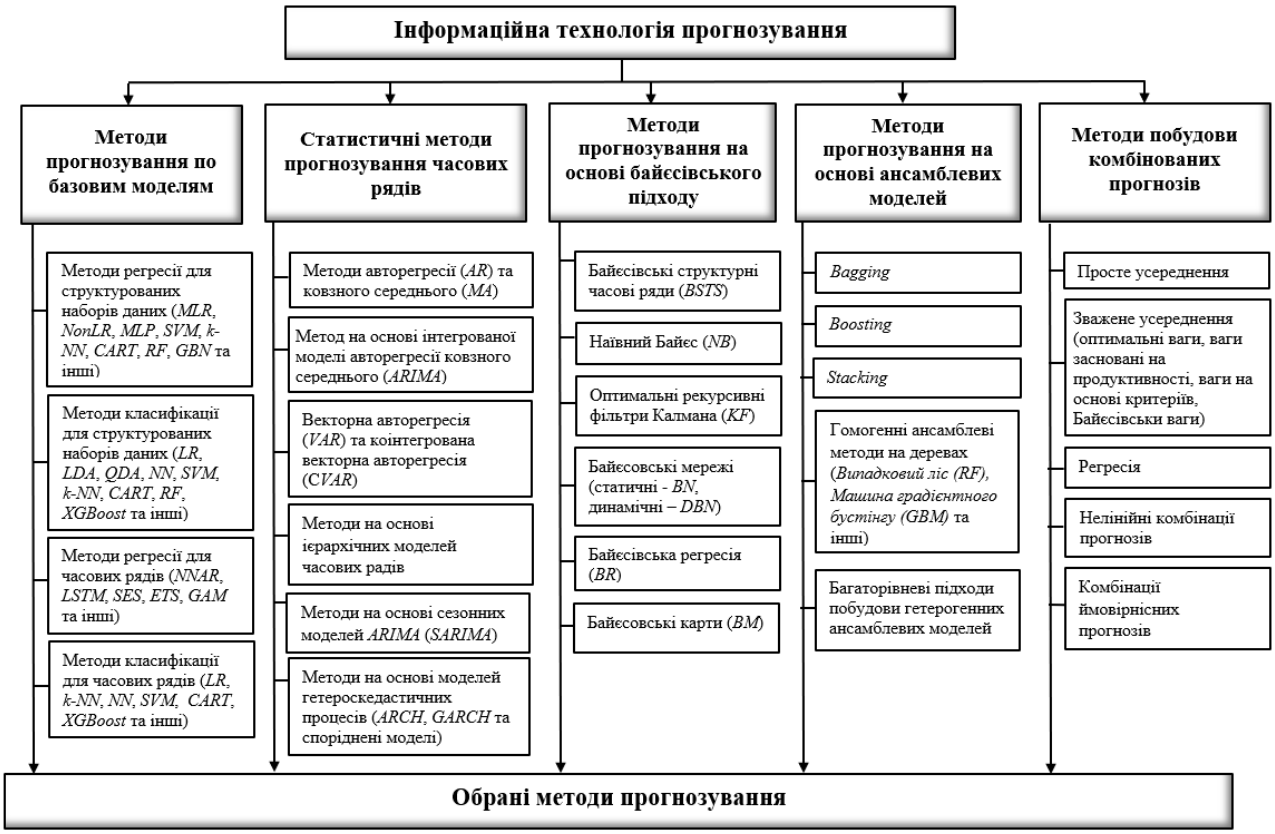

Рисунок 17 – Інформаційна технологія прогнозування

комбінованих прогнозів та методах прогнозування на основі ансамблевих моделей.

На основі інформаційної технології прогнозування розроблено архітектуру інформаційно-аналітичної системи прогнозування для вирішення задач машинного навчання (рис. 18).

У **шостому** розділі (*Вирішення прикладних задач машинного навчання на основі інформаційних технологій ймовірнісно-статистичного аналізу, моделювання та прогнозування нелінійних нестаціонарних даних*) розглянуто на прикладах практична ефективність розроблених підходів та методів до розв’язання прикладних задач машинного навчання для нелінійних нестаціонарних даних різної природи.

Першою розглянуто ***задачу прогнозування та моделювання поведінки валютних котирувань*** на валютному ринку. В якості вхідних даних використано

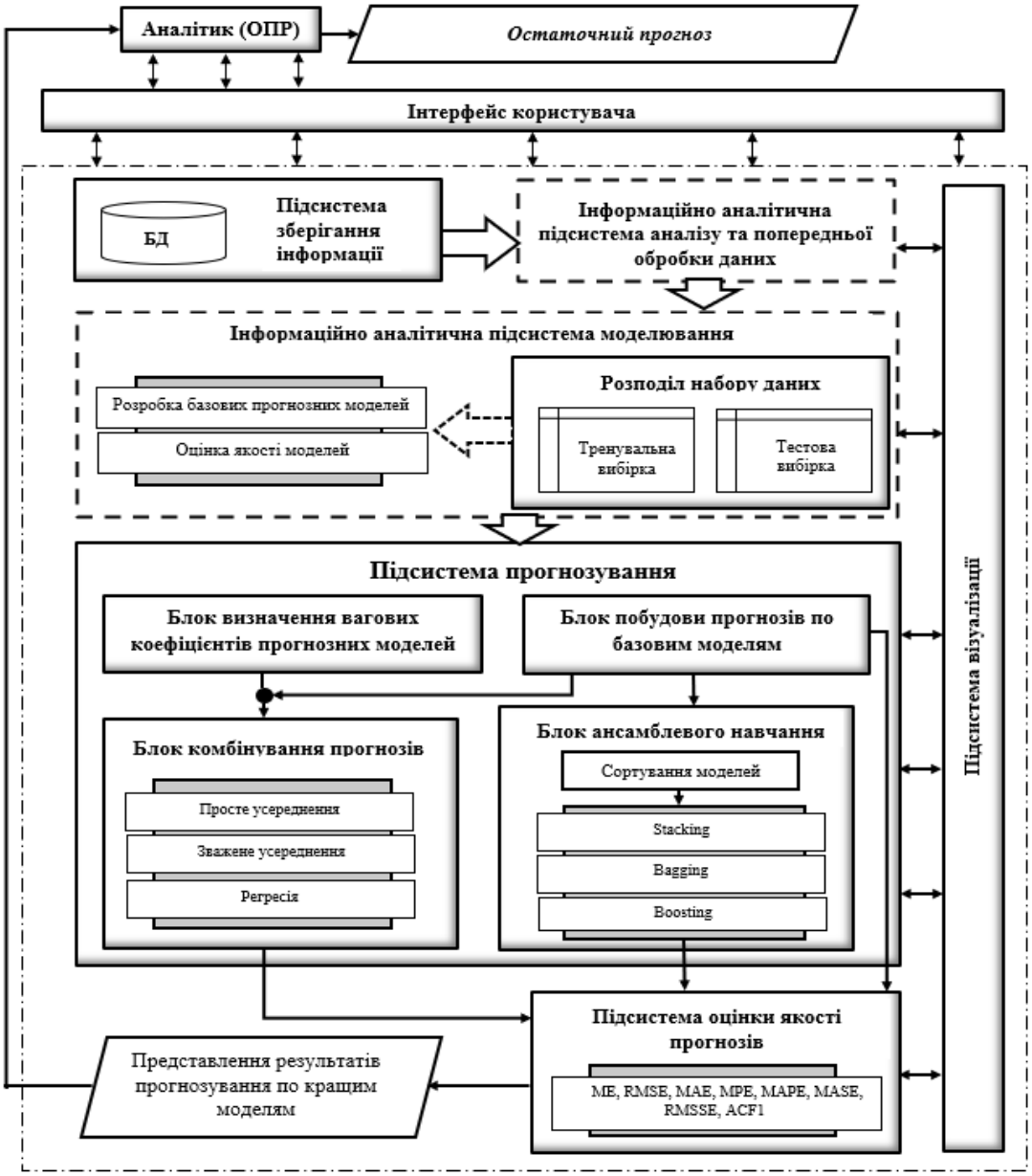

Рисунок 18 – Інформаційно аналітична система прогнозування

історичні дані 6 валютних пар: AUD/USD, CHF/JPY, EUR/USD, GBP/USD, NZD/USD, USD/JPY з періоду 1990 по 2020 роки з інтервалом 1 день. Джерелом даних став сервіс *Investing.com.* Вирішення завдання моделювання та короткострокового прогнозування поведінки котирувань на валютному ринку складається з наступних *етапів*: попередній аналіз та обробка даних; кластеризація та агрегація; створення моделей множинної класифікації; прогнозування.

На етапі попереднього аналізу та обробки даних здійснено ідентифікація та заповнення пропусків в даних, виявлення аномальних значень та їх обробка. Використано методи фільтрації. Завершує етап підготовки даних до процесу моделювання нормалізація даних.

Кластеризацію в даній задачі використано для пошуку цінових шаблонів та маркування часових рядів за схожістю до цих шаблонів. Часовий ряд по кожній

валютній парі представлено фрагментами довжиною 15 днів (3 тижні без вихідних) для подальшого маркування даних. Для кластеризації часових рядів використано метод K-means з метрикою якості DTW. Таке поєднання методів дозволило з максимальною точністю групувати часові ряди в кластери. Дані поділено на 100 кластерів в зв'язку з великим обсягом даних та для більшої точності майбутнього прогнозу.

Процедура агрегації реалізовано з метою об'єднання часових рядів, що знаходяться в одному кластері в єдиний шаблон, який демонструє динаміку змін ціни. Агрегація виконано шляхом пошуку *барицентру*. Для знаходження оптимального барицентру реалізовано декілька алгоритмів пошуку барицентрів. В результаті порівняння середньої відстані DTW для агрегування обрано метод *Soft DTW* зі значенням параметра $\gamma = 0.01$.

Створення класифікаційних моделей фрагментів часового ряду для виявлення належності їх до відповідного шаблону починалось з розділення початкового набору даних на навчальний і тестовий. Навчальний набір – 7437 днів, а тестовий набір – 281 день. На наступному кроці дані об'єднано в часові ряди довжиною 15 днів для маркування даних DTW-K-means і 10 днів для прогнозування. Для виявлення найкращого методу класифікації реалізовано декілька алгоритмів. Результати роботи методів класифікації представлено в таблиці 5. У результаті дослідження методів класифікації часових рядів обрано метод *SVM*, оскільки він показав кращу точність класифікації порівняно з іншими методами. Для агрегування кластера *K-means* обрано метод *Soft-DTW*.

Таблиця 5 – Результати роботи методів класифікації

Алгоритми	Точність класифікації шаблону, %	Точність класифікації напрямку ціни,%	Область ROCAUC
Наївний Баєс (NB)	37.03	70.37	0.71
Логістична регресія (LR)	25.93	74.07	0.75
Багатошаровий перцептрон (ANN)	35.19	72.20	0.73
Машина опорних векторів (SVM)	38.89	75.93	0.76
Дерево рішень (DT)	22.22	68.52	0.69
Градієнтний бустинг (XGBoost)	27.78	64.81	0.65

На етапі прогнозування реалізовано *алгоритм*, що складається з 4-х послідовних кроків:

Крок 1. На вхід *SVM*-моделі подається фрагмент часового ряду, що містить інформацію про останні 10 днів спостережень.

Крок 2. Часовий ряд класифікується та отримується номер його цінового шаблону.

Крок 3. Ціновий шаблон експортується з *K-means* -моделі у вигляді часового ряду довжиною 15 днів.

Крок 4. Класифікується ціновий рух (зміни ціни) на основі останніх 5 днів цінового шаблону.

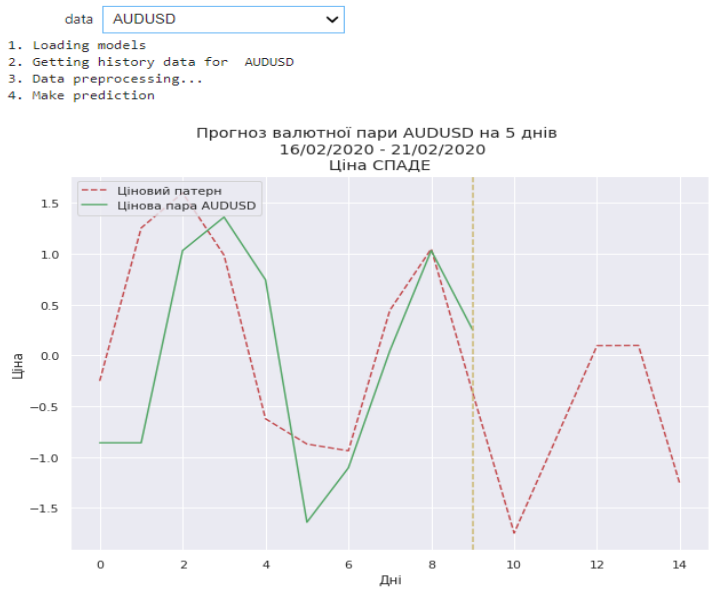

Рисунок 19 – Результати прогнозу AUD/USD на 5 днів

Аналіз показників якості прогнозних моделей показав, що запропонований підхід до прогнозування – ефективний. Інтерфейс додатку з результатами прогнозу по валютній парі австралійський долар /американський долар на 5 днів представлено на рисунку 19. Це дозволяє приймати рішення аналітикам щодо поточної ситуації на валютному ринку.

Наступною розглянуто ***задачу прогнозування показників виробництва електроенергії гібридними електростанціями*** на основі характеристик навколишнього середовища. Експериментальний набір даних *Combined Cycle Power Plant* складається з 9568 спостережень, зібраних коли електростанція з комбінованим циклом працювала 674 днів протягом 6 років. Розроблено методологію вирішення задачі прогнозування на основі методів машинного навчання. Реалізовано декілька альтернативних прогнозних моделей (табл. 6). Найкращі результати продемонстрував *boosting*. Застосування методології та використання методики покращення якості моделей значно підвищило якість прогнозування.

Таблиця 6 – Порівняльна таблиця показників якості прогнозу кращих прогностичних моделей

Тип моделі	MAE	MSE	RMSE	Cor
Удосконалена регресійна модель	3,61	21,39	4,62	0,962
Boosting -модель	2,28	11,02	3,32	0,980
ANN модель	4,66	27,67	5,26	0,967
DNN модель	3,34	18,18	4,26	0,978

Розглянуто використання методів класифікації для вирішення ***задачі прогнозування аеродинамічних властивостей матеріалів*** на основі дворівневого гетерогенного ансамбля моделей. Для вирішення задачі класифікації використано набір даних *Airfoil Self-Noise Data Set*. Набір даних NASA включає аеродинамічні профілі NACA 0012 різних розмірів з різними швидкостями та кутами атаки в аеродинамічній трубі. Набір складається з 1506 спостережень і 6 атрибутів: частота в герцах; кут атаки в градусах; довжина хорди в метрах; швидкість напливного потоку в метрах за секунду; і товщина витіснення на стороні всмоктування в метрах. Результуюче значення є масштабованим рівнем звукового тиску в децибелах.

На етапі попереднього аналізу даних виявлено нечислові та відсутні значення, виявлено та оброблено аномальні значення, ознаки перевірянні на

унікальність та нормовані на основі розробленого алгоритму «Очищення даних». Для обробки викидів, розроблено алгоритм, що демонструє розрахунок порогів відсікання. Відбір ознак для побудови ML-моделі виконано на підставі розробленого алгоритму в основі якого покладено метод *Backward Elimination*.

На етапі моделювання дані підготовлено до навчання моделей. Набір даних поділено на дві вибірки: навчальну і тестову, в пропорції 70/30. Спостереження відібрано за допомогою генератора випадкових чисел. Реалізовано 8 базових моделей класифікації (знайдено оптимальну структуру та параметри). Для виявлення дисперсії і зміщення кожної моделі проаналізовано частоту помилок класифікації на навчальній і контрольній вибірках (табл. 7).

На основі розробленого алгоритму побудови дворівневого ансамблю базові моделі розподілене на два рівні. Показники ефективності моделей класифікації представлено в таблиці 8.

Таблиця 7 – Частота помилок класифікації для моделей набору даних

Тип моделі	Частота похибок	
	на навчальній вибірці, %	на контрольній вибірці, %
Decision Tree	22,2	26,4
NB	25,5	29,8
QDA	14,8	23,2
LR	24,4	28,9
SVM	23,7	25,5
KNN	20,5	21,1
ANN	8,3	22,2
RF	20,8	22,1

Таблиця 8 – Показники ефективності моделей класифікації

Тип моделі	Коефіцієнт успішності (accuracy)	Каппа-статистика (Kappa)	Чутливість (sensitivity)	Специфічність (specificity)	Точність (precision)	Повнота (recall)	Fміра (F-measure)
Decision Tree	0,736	0,475	0,600	0,878	0,678	0,878	0,765
NB	0,711	0,419	0,800	0,617	0,747	0,647	0,676
QDA	0,745	0,488	0,808	0,678	0,772	0,678	0,722
LR	0,770	0,540	0,775	0,765	0,765	0,765	0,765
KNN	0,774	0,549	0,791	0,756	0,777	0,757	0,767
ANN	0,240	0,517	0.275	0,209	0,216	0,209	0,212
Stacking (LR)	0,770	0,542	0,708	0,835	0,733	0,835	0,780
RF	0,813	0,625	0,817	0,809	0,809	0,809	0,809
SVM	0,787	0,575	0,742	0,835	0,756	0,835	0,793
Bagging	0,817	0,634	0,825	0,809	0,816	0,809	0,812

Метрики оцінки якості роботи базових класифікаторів показали результати, що потребують покращення**.** Використовуючи *stacking* для комбінування шести не високих результатів базових класифікаторів, покращено загальний результат. Використовуючи *bagging* для комбінування трьох не високих результатів базових класифікаторів та результату роботи моделі першого рівня – *stacking*, значно покращено загальний результат по всім метрикам оцінювання. Доведено, що використання дворівневого ансамблю підвищує ефективність класифікаційних моделей.

Розглянуто ***задачу побудови прогнозних моделей з предикторами вартості комерційних компаній*** на основі байєсівських структурних часових рядів. Прогнозні моделі з предикторами побудовано на базі трьох наборів даних, в яких відображено вартість акцій компаній *Amazon*, *Facebook* і *Google*. Вони є частиною багатовимірного часового ряду (*https://finance.yahoo.com/*).

При розв'язанні задачі прогнозування на часовому ряді за результатами попереднього аналізу та обробки даних складається декілька альтернативних моделей BSTS. Кожна з моделей доповнюється компонентами, які можуть відображати характер змін даних. Щоб побудувати BSTS-модель для прогнозування вартості акцій компанії *Amazon*, використано, як предиктори, ціни акцій інших компаній на момент закриття торгів у період з 1 січня 2016 року по 26 травня 2019 року (*Facebook* і *Google*). Набір компонентів найкращої моделі без предикторів є основою для комплектування компонентів моделей часових рядів предикторів. Для вирішення задачі побудови прогнозних моделей розроблено алгоритм побудови моделі BSTS з предикторами. В байєсівські структурні моделі часових рядів можна включати велику кількість предикторів без ризику перенавчання.

Визначено дев'ять прогнозних моделей. Оскільки моделі BSTS передбачають велику кількість можливих реалізацій майбутніх значень залежної змінної, для розрахунку показників використовувалися медіанні значення можливих реалізацій (табл. 9).

Таблиця 9 – Оцінка якості прогнозних моделей

Найменування моделі	Метрики якості прогнозних моделей			
	MAPE	MAE	RMSE	U-статистика
Model 2	3,740	68,6478	59,6242	0,0158
Model 3	4,387	80,5957	68,7171	0,1822
Model 4	4,164	76,3945	67,0756	0,0178
Model 5	3,449	63,4271	52,3890	0,0139
Model 7	3,395	62,3760	52,5126	0,0140
Model 12	2,323	43,1211	36,1843	0,0097

Після реалізації кроків алгоритму результати показують, що модель BSTS з предикторами в порівнянні з іншими моделями є найкращою та може бути ефективно використана при побудові прогнозів складних даних, представлених у вигляді байєсівських структурних часових рядів.

Розглянуті варіанти розв'язання практичних задач демонструють ефективність застосування методів машинного навчання для вирішення завдань ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних. Доцільність їх використання пояснюється їх високою гнучкістю, здатністю здійснювати аналіз складних даних, будувати моделі та прогнозувати, для вирішення завдань машинного навчання.

ОСНОВНІ РЕЗУЛЬТАТИ РОБОТИ І ВИСНОВКИ

На основі виконаних теоретичних та експериментальних досліджень вирішено важливу науково-прикладну проблему в галузі інформаційних технологій – підвищення ефективності ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в задачах машинного навчання засобами сучасних інформаційних технологій з урахуванням нелінійності і нестаціонарності даних, а також можливих невизначеностей, що є характерними для них.

У результаті виконання цієї роботи одержані наступні результати:

1. Виконано аналіз сучасного стану досліджень в області існуючих методів, моделей і алгоритмів ймовірнісно-статистичного аналізу, моделювання та прогнозування, з урахуванням нелінійностей та нестаціонарностей даних і можливих супутніх невизначеностей, для розв’язання завдань машинного навчання.
2. Вперше розроблено метод синтезу нових інформаційних технологій для розв’язування завдань машинного навчання, який ґрунтується на системному використанні методів ймовірнісно-статистичного аналізу даних, математичного моделювання, методів прогнозування, що підвищує ефективність процесу машинного навчання в умовах наявності нелінійних нестаціонарних процесів, які досліджувались, та різних типів невизначеностей даних.
3. Вперше розроблено метод обробки пропусків в наборах даних на основі системного використання методів пошуку закономірностей появи відсутніх значень та методів аналізу наборів даних без пропусків, що дало змогу значно підвищити ефективність попередньої обробки даних.
4. Вперше розроблено метод пошуку викидів та аномалій в нелінійних нестаціонарних даних на основі системного використання методів виявлення аномальних значень, аналізу причин та методів обробки викидів, що дало змогу підвищити точність ідентифікації аномальних значень в наборах даних різного типу.
5. Отримав подальший розвиток метод фільтрації статистичних даних завдяки системному використанню різних типів фільтрів за рахунок модифікації алгоритмів гранулярної фільтрації, що дало можливість в процесі попередньої обробки врахувати нелінійність та нестаціонарність (гетероскедастичність) даних, і описати сам процес та динаміку його дисперсії.
6. Отримав подальший розвиток підхід до нормалізації та стандартизації даних на основі системного поєднання методів перетворення даних та особливостей вирішення завдань машинного навчання, що дало змогу спростити і прискорити процедури нормалізації складних наборів даних.
7. Вперше розроблено метод побудови прогнозних моделей на основі

байєсівських структурних часових рядів, якій базується на процесі навчання структурних моделей часових рядів та на алгоритмі побудові BSTS-моделі (з предикторами та без предикторів) у середньому на 6-27% в порівнянні зі статистичними моделями ARIMA, що дало можливість враховувати нелінійність та нестаціонарність процесів при аналізі часових рядів та підвищити точність прогнозування.

8. Вперше розроблено метод побудови імітаційних моделей систем зі складним стохастичним процесом обробки даних на основі колірних мереж Петрі для розв'язування завдань аналізу та моделювання нелінійних та нестаціонарних процесів, який базується на системному використанні стохастичних, часових, ієрархічних мереж Петрі для імітаційного моделювання динамічних процесів, що дало можливість значно підвищити ефективність і точність побудови математичних моделей для аналізу складних нелінійних нестаціонарних процесів.

9. Отримав подальший розвиток метод синтезу параметрів нелінійної прогнозної моделі на основі використання генетичного алгоритму, що підвищило ефективність підбору параметрів прогнозних моделей.

10. Вперше розроблено системний підхід до моделювання, прогнозування та підтримки прийняття рішень в задачах машинного навчання, що враховує можливі невизначеності при аналізі процесу, нелінійності, нестаціонарності даних, особливості підбору й оцінки структури та параметрів прогнозних моделей та їх оцінювання, що дало змогу підвищити точність оцінювання якості прогнозу та визначити краще прогнозне рішення.

11. Удосконалено метод розв'язання задач прогнозування за рахунок використання багатошарових нейронних мереж прямого розповсюдження, яка забезпечує підвищення точності прогнозування на часових рядах до 25% (по показнику RMSE) в порівнянні з кращою зі статистичних моделей ARIMA.

12. Удосконалено метод побудови і використання комбінованих прогнозів за рахунок ітеративного оцінювання різних схем комбінування, що дає змогу підвищити ефективність та точність прогнозних рішень у середньому на 22%.

13. Вперше розроблено метод до зменшення похибки прогнозу за рахунок одночасного зменшення зміщення та дисперсії на основі використання багаторівневих гетерогенних ансамблів прогнозних моделей, що дало змогу підвищити якість та точність класифікації до 31%, по показнику F-міри до 29% в порівнянні з кращою з альтернативних моделей.

14. На основі розроблених методів, моделей ймовірнісно-статистичного аналізу, моделювання та прогнозування, з урахуванням нелінійностей та нестаціонарностей даних і можливих супутніх невизначеностей, розроблено інформаційні технології (ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання і прогнозування) та інформаційно-аналітичні системи для вирішення завдань машинного навчання.

15. Достовірність наукових та практичних результатів підтверджується відповідними матеріалами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації

Статті у періодичних виданнях, індексованих у наукометричних базах Scopus та Web of Science

1. Babichev S., Liakh I., Kalinina I. Applying a Recurrent Neural Network-Based Deep Learning Model for Gene Expression Data Classification. *Applied sciences an Open Access Journal by MDPI*. 2023. Vol. 13, Issue 21, No. 11823. https://doi.org/10.3390/app132111823. (*Scopus*, ***Q2***)
2. Danilov V., Gozhyj O., Kalinina I., Bidyuk P., Jirov O. Adaptive forecasting and financial risk estimation. *System Research and Information Technologies*, 2020. Vol. 1, pp. 34-53. DOI: 10.20535/SRIT.2308-8893.2020.1.04. (*Scopus*).
3. Kalinina I., Gozhyj A. Methodology for Solving Forecasting Problems Based on Machine Learning Methods. *Lecture Notes on Data Engineering and Communications Technologies* (Switzeland). 2023. Vol. 149, pp. 105-125. (ISSN: 2367 – 4512). (*серійне закордонне видання*) (*bookchapter*, *Scopus*).
4. Bidyuk P., Kalinina I., Gozhyj A. An Approach to Identifying and Filling Data Gaps in Machine Learning Procedures. *Lecture Notes on Data Engineering and Communications Technologies* (Switzeland). 2022. Vol. 77, pp. 164-176. (ISSN: 2367-4512). (*серійне закордонне видання*) (*bookchapter*, *Scopus*).
5. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. Web Resources Management Method Based on Intelligent Technologies. *Advances in Intelligent Systems and Computing* III. Vol. 871. Springer. Pp. 206-221. DOI: 10.1007/978-3-030-01069-0_15. (*Scopus*, ***Q3***).
6. Bidyuk P., Gozhyj A., Kalinina I., Gozhyj V. Analysis of uncertainty types for model building and forecasting dynamic processes. Conference on Computer Science and Information Technologies, CSIT 2017: *Advances in Intelligent Systems and Computing* II. 2017. Vol. 689. Springer-Verlag, pp. 66-78. DOI: 10.1007/978-3-319-70581-1_5(*Scopus*, ***Q3***).

Статті у наукових фахових виданнях України

7. Kalinina I.A., Gozhyj A.P. Modeling and forecasting of nonlinear nonstationary processes based on the Bayesian structural time series. *Applied Aspects of Information Technology*. 2022. Vol. 5, no. 3, pp. 240-255. DOI: https://doi.org/10.15276/aait.05.2022.17.
8. Gozhyj A.P., Kalinina I.A., Gozhyj V.A. Method for developing and modelling composite web-services. *Herald of Advanced Information Technology*. 2022. Vol. 5, no. 3, pp. 185-197. DOI: https://doi.org/10.15276/aait.05.2022.14.
9. Калініна І. О., Гожий О. П. Дослідження ефективності методів класифікації при прогнозуванні в задачах машинного навчання. *Управління розвитком складних систем. Управління технологічними процесами*. 2021. Вип. 46. С. 173-180. DOI: 10.32347/2412-9933.2021.46.173-180.
10. Bidyuk P.I., Gozhyj O.P., Kalinina I.O., Danilov V.J., Jirov O.L. Adaptive modeling and forecasting economic and financial. *Informatics and Mathematical*

Methods in Simulation. 2019. Vol. 9, no. 4, pp. 231-250. DOI: 10.15276/imms.v9.no4.231.
11. Гожий О.П., Жебко О.О., Калініна І.О., Ганніченко Т.А. Інтелектуальна система класифікації на основі ансамблевих методів. *Регіональний міжвузівський збірник наукових праць* «*Системні технології*», Дніпро. 2023. Вип. 3, № 146. С. 61-75. DOI 10.34185/1562-9945-3-146-2023-07.
12. Калініна І.О., Гожий О.П., Нечахін О.П., Шиян С.І. Синтез параметрів нелінійної прогнозної моделі за допомогою генетичного алгоритму. *Регіональний міжвузівський збірник наукових праць* «*Системні технології*», Дніпро. 2023. Вип. 2, № 145. С. 66-75. DOI: 10.34185/1562-9945-2-145-2023-07.
13. Калініна І.О., Гожий О.П., Нечахін В.В., Шиян С.І. Імітаційне моделювання систем зі складним стохастичним процесом обробки даних за допомогою кольорових мереж Петрі. *Регіональний міжвузівський збірник наукових праць* «*Системні технології*», Дніпро. 2022. Вип. 6, № 143. С. 42-56. DOI: https://doi.org/10.34185/1562-9945-6-143-2022-04.
14. Гожий О. П., Калініна І. О., Андрєєва Н. Ю. Динамічне планування розподілу ресурсів в автономній енергосистемі. *Науково-технічний журнал* «*Авиационно-космическая техника и технология*». *Информационные технологии*. 2014. Вип. № 10 (117). С. 131-134.
15. Гожий О. П., Калініна І. О., Гожий В.О. Побудова динамічних прогнозів в задачах планування. *Регіональний міжвузівський збірник наукових праць* «*Системні технології*», Дніпро. 2015. Вип. 2, № 97. С. 13-24. DOI: http://nbuv.gov.ua/UJRN/st_2015_2_5.
16. Bidyuk P.I., Korshevnyuk L.O., Gozhyj O.P., Kalinina I.O., Prosyankina-Zharova T.I., Terentiev O.M. Modelling and forecasting financial and economic processes with decision support system. *Інформаційні технології, системний аналіз та керування. Science News KPI*. 2019. №5-6. Pp. 7-17. DOI: 10.20535/kpi-sn.2019.5-6.176835.
17. Гожий В.О., Калініна І.О. Використання ієрархічних часових мереж Петрі для моделювання web-сервісів. *Наукові праці ЧДУ ім. Петра Могили*: Науково-методичний журнал, серія: Комп'ютерні технології. 2018. Вип. 305, т. 317. Миколаїв: Вид-во ЧНУ ім. П. Могили. С. 30-35.
18. Гожий О.П., Калініна, І.О. Особливості використання нечітких ситуаційних мереж для вирішення задач прийняття рішень. *Наукові праці ЧДУ ім. Петра Могили*. Науково-методичний журнал, серія: Комп'ютерні технологіїю 2014. Вип. 225, т.237, Миколаїв, вид-во ЧДУ ім. Петра Могили. С. 19-24.
19. Калініна І.О. Особливості застосування генетичних алгоритмів в задачах прогнозування. *Наукові праці ЧДУ ім. Петра Могили*. Науково-методичний журнал, серія: Комп'ютерні технології. 2010. Вип. 121, т. 134, Миколаїв, вид-во ЧДУ ім. Петра Могили. С. 137-141.
20. Калініна І.О. Дослідження алгоритмів навчання нейронних мереж у задачах прогнозування. *Наукові праці ЧДУ ім. Петра Могили*. Науково-методичний журнал, серія: Комп'ютерні технології. 2009. Вип. 104, т. 117, Миколаїв, вид-во ЧДУ ім. Петра Могили. С. 160-171.

21. Калініна І.О. Використання нейромережевих методів в задачах прогнозування. *Наукові праці ЧДУ ім. Петра Могили*. Науково-методичний журнал, серія: Комп'ютерні технології. 2009. Вип. 93, т. 106, Миколаїв, вид-во ЧДУ ім. Петра Могили. С. 132-138.
22. Калініна І.О. Використання нейромережевих методів у задачах фінансового менеджменту. *Наукові праці МДГУ ім. П. Могили*. Серія: Комп'ютерні технології. 2008. Вип. 77, т. 90, Миколаїв, вид-во ЧДУ ім. Петра Могили. С. 160-167.

Монографії

23. Гожий О.П., Калініна І.О., Нечахін В.В. Інтелектуальні технології в керуванні гібридними енергетичними системами: монографія. Херсон: Книжкове видавництво ФОП Вишемирський В.С., 2021. 200 с. ISBN 978-617-7941-56-8.
24. Бідюк П.І., Калініна І.О., Гожий О.П. Байєсівський аналіз даних: монографія. Херсон: Книжкове видавництво ФОП Вишемірський В.С., 2021. 208 с. ISBN 978-617-7941-52-0.

Наукові праці, які засвідчують апробацію матеріалів дисертації

Статті у матеріалах міжнародних конференцій, які індексуються у наукометричних базах Scopus та Web of Science

25. Bidyuk P., Kalinina I., Zhebko O., Gozhyj A., Hannichenko T. Classification System Based on Ensemble Methods for Solving Machine Learning Tasks. *CEUR- WS*. 2023. Vol. 3426. Pp. 1-11. CEUR-WS.org/Vol-3426/paper5.pdf. (ISSN 1613-0073). (*Scopus*)
26. Kalinina I., Bidyuk P., Gozhyj A., Malchenko P. Combining Forecasts Based on Time Series Models in Machine Learning Tasks. *CEUR-WS*. 2023. Vol. 3426. Pp. 25-35. CEUR-WS.org/Vol-3426/paper2.pdf. (ISSN 1613-0073). (*Scopus*)
27. Kalinina I., Bidyuk P., Gozhyj A. Construction of Forecast Models based on Bayesian Structural Time Series. *International Scientific and Technical Conference on Computer Sciences and Information Technologies*. CSIT_2022. 2022. Pp. 180-184. DOI 10.1109/CSIT56902.2022.10000484. (*Scopus*)
28. Kalinina I., Gozhyj A., Gozhyj V. Modeling a Pharmaceutical Web Service Using Colored Petri Nets. *International Scientific and Technical Conference on Computer Sciences and Information Technologie*s. CSIT_2022. 2022. Pp. 345-348. DOI 10.1109/CSIT56902.2022.10000824. (*Scopus*)
29. Gozhyj A., Kalinina I., Nechakhin V., Gozhyj V., Vysotska V. Modeling an Intelligent Solar Power Plant Control System Using Colored Petri Nets. Proceedings of the 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: *Technology and Applications*. 2021. Cracow, Poland. Vol. 2, pp. 626–631. (Electronic ISSN: 2770-4254). DOI: 10.1109/IDAACS53288.2021.9660860. (*WoS/Scopus*)
30. Trofymchuk O., Bidyuk P., Kalinina I., Gozhyj A. Financial Risk Estimation in Conditions of Stochastic Uncertainties. *Lecture Notes in Computational Intelligence and Decision Making*. *Lecture Notes on Data Engineering and Communications Technologies.* 2022. Vol 77. Pp. 3-24, Springer, Cham. https://doi.org/10.1007/978-3-030-82014-5_1. (*bookchapter, Scopus*)

31. Aksonov D., Gozhyj A., Kalinina I., Vysotska V. Question-Answering Systems Development Based on Big Data Analysis. IEEE 16th *International Conference on Computer Sciences and Information Technologies* (CSIT). 2021. Vol.1, pp. 113–118. DOI: 10.1109/CSIT52700.2021.9648631. (*Scopus*)
32. Bidyuk P., Gozhyj A., Matsuki Y., Kuznetsova N., Kalinina I. Modeling and Forecasting Economic and Financial Processes Using Combined Adaptive Models. International Scientific Conference "Intellectual Systems of Decision Making and Problem of Computational Intelligence", ISDMCI 2020: *Lecture Notes in Computational Intelligence and Decision Making, Part of the Advances in Intelligent Systems and Computing book series* (AISC, vol. 1246). Pp. 395-408, (ISSN 2194-5357). http://doi.org/10.1007/978-3-030-54215-3_25. (*Scopus*)
33. Bidyuk P., Gozhyj A., Kalinina I. Probabilistic Inference Based on LS-Method Modifications in Decision Making Problems. *Lecture Notes in Computational Intelligence and Decision Making. Advances in Intelligent Systems and Computing*. 2020. Vol. 1020, pp. 422-433. Springer, Cham. (ISSN 2194-5357). DOI: 10.1007/978-3-030-26474-1_30. (*Scopus*)
34. Bidyuk P., Matsuki Y., Gozhyj A., Beglytsia V., Kalinina I. Features of Application of Monte Carlo Method with Markov Chain Algorithms in Bayesian Data Analysis. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1080 AISC, pp. 361-376. Springer, Cham. DOI: 10.1007/978-3-030-33695-0_25. (*Scopus*)
35. Bidyuk P., Gozhyj A., Kalinina I., Vysotska V. Methods for Forecasting Nonlinear Non-Stationary Processes in Machine Learning. *Communications in Computer and Information Science*. 2020. Vol. 1158. Springer, Cham. Pp. 470-485. DOI: https://doi.org/10.1007/978-3-030-61656-4_32. (*Scopus*)
36. Gozhyj A., Kalinina I., Gozhyj V., Danilov V. Approach for Modeling Search Web-Services Based on Color Petri Nets. *Communications in Computer and Information Science*. 2020. Vol. 1158. Springer, Cham. Pp. 525-538. (ISSN 1865-0929). DOI: 10.1007/978-3-030-61656-4_35. (*Scopus*)
37. Bidyuk P., Kalinina I., Gozhyj A. Methodology of Constructing Statistical Models for Nonlinear Non-stationary Processes in Medical Diagnostic Systems. *CEUR-WS*. 2020, vol. 2753, pp. 36-45. (ISSN 1613-0073). http://ceur-ws.org/ Vol-2753/paper4.pdf. (*Scopus*)
38. Bidyuk P., Trofymchuk P., Kalinina I., Gozhyj A. Modeling Risk Factor Interaction Using Copula Functions. Proceedings of the 1st International Workshop on Computational & Information Technologies for Risk-Informed Systems (CITRisk-2020). Kherson. *CEUR-WS*. 2020. Pp. 87-101. (ISSN 1613-0073). CEUR-WS.org/Vol-2805/paper7.pdf. (*Scopus*)
39. Gozhyj A., Nechachin V., Kalinina I. Solar Power Control System based on Machine Learning Methods. *International Scientific and Technical Conference on Computer Sciences and Information Technologies*. IEEE CSIT 2020. Lviv. 2020. Vol. 1. Pp. 24-27. DOI: 10.1109/CSIT49958.2020.9321953. (*Scopus*)
40. Bidyuk, P., Gozhyj, A., Kalinina I., Vasilev, M., Malets, R. Forecasting nonlinear nonstationary processes in machine learning task. Proceedings of the 2020 IEEE 3rd

International Conference on Data Stream Mining and Processing, DSMP 2020. Pp. 28-32. DOI: 10.1109/DSMP47368.2020.9204077. (*Scopus*)
41. Bidyuk, P., Beglytsia, V., Gozhyj, A., Kalinina, I. Using the Metropolis-Hastings algorithm in Bayesian data analysis procedures. 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT). 2019. Vol. 2. Pp. 98-101. DOI: 10.1109/STC-CSIT.2019.8929797. (*Scopus*)
42. Gozhyj, A., Kalinina, I., Gozhyj, V., Vysotska V. Web service interaction modeling with colored petri nets. *Journal "Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, IDAACS 2019. 2019. Vol. 1, pp. 319-323. DOI: 1109/IDAACS.2019.8924400. (*WoS/Scopus*)
43. Bidyuk P., Gozhyj A., Szymanski, Z., Kalinina, I., Beglytsia, V. The Methods Bayesian Analysis of the Threshold Stochastic Volatility Model. *Journal 2018 IEEE 2nd International Conference on Data Stream Mining and Processing*, DSMP 2018, Lviv. 2018. Pp. 70-74. DOI: 10.1109/DSMP.2018.8478474. (*Scopus*)
44. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. The method of web-resources management under conditions of uncertainty based on fuzzy logic. *Journal "2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies*, (CSIT 2018). Lviv: "Vega and Ko". 2018. Vol. 1. Pp. 343-346. DOI: 10.1109/STC-CSIT.2018.8526761.(*Scopus*)
45. Bidyuk P., Gozhyj A., Kalinina I. Modeling military conflicts using Bayesian networks. *Journal "2018 IEEE 1st International Conference on System Analysis and Intelligent Computing*, (SAIC 2018). Pp. 155-160. DOI: 10.1109/SAIC.2018.8516861. (*Scopus*)
46. Gozhyj A., Vysotska V., Yevseyeva I., Kalinina I., Gozhyj V. Web Resources Management Method Based on Intelligent Technologies. CSIT 2018: *Advances in Intelligent Systems and Computing* III. 2019. Vol. 871. Pp. 206-221. DOI: 10.1007/978-3-030-01069-0_15. (*Scopus*)
47. Bidyuk P., Kalinina I., Gozhyj A., Gozhyj V. Methods for processing uncertainties in solving dynamic planning problems. *Journal 'Proceedings of the 2017 12th international scientific and technical conference on computer sciences and information technologies* (CSIT 2017). Lviv: Publishing Lviv Polytechnic. 2017. Vol. 1. Pp. 151-155. (*Scopus*)
48. Gozhyj A., Kalinina I., Gozhyj V. Fuzzy cognitive analysis and modeling of water quality. 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: *Technology and Applications* (IDAACS). Bucharest, Romania, 21-22 September 2017. Pp. 289-294. DOI: 10.1109/IDAACS.2017.8095092. (*WoS/Scopus*)

Статті та тези доповідей у збірниках праць конференцій

49. Калініна І.О. Особенности применения нейронных сетей для решения задач прогнозирования. Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: Матеріали міжнародної науково-практичної

конференції. Том 3. (ч. 1). Теоретичні і прикладні аспекти систем прийняття рішень. Євпаторія, Крим, Україна, 2008. С. 138-140.
50. Калініна І.О. Використання нейромережевих методів у задачах фінансового менеджменту. «Ольвійський форум – 2008: Стратегії України в геополітичному просторі»: Матеріали міжнародної науково-практичної конференції. Частина 2. Ялта, Крим, Україна, 2008. С. 97-98.
51. Калініна І.О. Особливості використання нейромережевих методів в прогнозуванні фінансових показників. «Ольвійський форум – 2009: Стратегії України в геополітичному просторі»: Матеріали міжнародної науково-практичної конференції. Тези доповідей. Ялта, Крим, Україна, том 3, 2009. С. 221-224.
52. Гожий А.П., Калініна И.А., Чирун Л.Б. Построение процедур принятия решений на основе непараметрических методов. VIII міжнародна школа семінар Теорія прийняття рішень. 2016. Ужгород, УНУ-2016. С. 88-89.
53. Гожий О.П., Калініна І.О. Метод оцінювання ризиків і невизначеностей в задачах ситуаційного моделювання і планування. Міжнародна наукова конференція: Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту. Section „Analysis and modeling of complex systems and processes" ISDMCI'2016. *Збірка наукових праць*. Херсон, ПП Вишемирський В.С. 2017. С. 48-50.
54. Калініна І.О., Гожий О.П. Динамічне моделювання складних систем за допомогою кольорових мереж Петрі. «Ольвійський форум – 2017: стратегії країн Причорноморського регіону в геополітичному просторі»: Матеріали міжнародної науково-практичної конференції. Секція: Інформаційні технології у розвитку суспільства. Тези доповідей. Миколаїв : Вид-во ЧНУ ім. Петра Могили, 2017. С. 31-33.
55. Калініна І.О. Алгоритм побудови архітектури нейронної мережі за допомогою генетичного алгоритму. «Могилянські читання – 2017. Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти». Всеукр. наук.-метод. конф. Тези доповідей. Комп'ютерні науки. Технічні науки, Миколаїв. ЧНУ ім. Петра Могили, 2017. С. 125-126.
56. Гожий О.П., Калініна І.О. Аналіз моделей стохастичної волатильності. Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту». ISDMCI'2018. Залізний порт: ХНТУ. 2018. С.47-49.
57. Гожий О.П., Калініна І.О. Підхід до представлення порогової моделі стохастичної волатильності. «Могилянські читання – 2018. Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти». Всеукр. наук.-метод. конф. Тези доповідей. Комп'ютерні науки. Технічні науки. Миколаїв. ЧНУ ім. Петра Могили, 2018. С. 18-20.
58. Гожий О.П., Калініна І.О. Розробка Інтелектуальної системи керування автономною гібридною енергетичною системою. Міжнародний науковий симпозіум «Інтелектуальні рішення». Обчислювальний інтелект (результати, проблеми, перспективи): праці між нар.наук.-практ. конф., 2019. Ужгород. ДВНЗ «Ужгородський національний університет». С.186-187.

59. Калініна І.О., Гожий О.П., Особливості використання алгоритму Метрополіса-Хастінгса в процедурах машинного навчання. «Ольвійський форум – 2019. Стратегії країн Причорноморського регіону в геополітичному просторі». Матеріали міжнародної науково-практичної конференції. Секція: Автоматизація та комп'ютерно-інженерні технології. АСУ, CASE – засоби та програмна інженерія. Інтелектуальні інформаційні системи. Комп'ютерна інженерія. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2019. С. 64-66.
60. Калініна І. О. Етапи побудови та верифікації статистичних моделей в задачах машинного навчання. «Ольвійський форум – 2020. Стратегії країн Причорноморського регіону в геополітичному просторі». Матеріали міжнародної науково-практичної конференції. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2020. С. 21-22.
61. Калініна І. О. Підхід до ймовірнісного моделювання взаємодії факторів ризику. «Могилянські читання – 2020. Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти». Всеукр. наук.-метод. конф. Тези доповідей. Комп'ютерні науки. Технічні науки. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2020. С. 9-11.
62. Калініна І. О. Особливості генерування вибірки за Гіббсом в процедурах Байєсівського аналізу даних. «Могилянські читання-2021. Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти». Всеукр. наук.-метод. конф. Тези доповідей. Комп'ютерні науки. Технічні науки. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2021. С. 11-13.
63. Калініна І.О. Побудова байєсівських динамічних моделей в просторі станів. «Ольвійський форум – 2021. Стратегії країн Причорноморського регіону в геополітичному просторі». Матеріали міжнародної науково-практичної конференції. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2021. С. 41-43.
64. Гожий О.П., Нечахін В.В., Калініна І. О. Застосування нейромережевої архітектури LSTM в системі керування сонячною електростанцією. Міжнародний науковий симпозіум «Інтелектуальні рішення». Обчислювальний інтелект (результати, проблеми, перспективи): праці між нар.наук. симпозіуму, 2021. Ужгород, ДВНЗ «Ужгородський національний університет». С. 38-39.
65. Калініна І. О., Скубак О. Д., Петроченко О.О. Аналіз часових рядів за допомогою машинного навчання. Інформаційні технології та інженерія: Всеукраїнська науково-практична конференція молодих вчених, аспірантів і студентів: тези доп., 2022. ЧНУ імені Петра Могили. Миколаїв, 2022. С. 67-69.
66. Калініна І. О. Порівняльний аналіз прогностичних моделей на основі дерев рішень. «Ольвійський форум – 2022. Стратегії країн При-чорноморського регіону в геополітичному просторі». Матеріали міжнародної науково-практичної конференції. м. Миколаїв: Вид-во ЧНУ ім. Петра Могили. 2022. С. 47-52.
67. Калініна І. О., Мальченко П. О. Прогнозування вартості комерційних компаній на основі модифікованого методу ARIMA. «Могилянські читання-2022. Досвід та тенденції розвитку суспільства в Україні: глобальний, національний та регіональний аспекти». Всеукр. наук.-метод. конф. Тези доповідей. Комп'ютерні науки. Технічні науки. Миколаїв: Вид-во ЧНУ ім. Петра Могили, 2022. С. 26-28.

68. Калініна І. О, Гожий О. П. Прогнозування нелінійних нестаціонарних процесів на основі байєсівського підходу. «Ольвійський форум – 2023. Стратегії країн Причорноморського регіону в геополітичному просторі». Технічні науки. Матеріали міжнародної науково-практичної конференції. Миколаїв: Вид-во ЧНУ ім. Петра Могили. 2023. С. 170-176.

Наукові праці, які додатково відображають наукові результати дисертації:

Навчальний посібник

69. Калініна І.О., Гожий О.П. Моделювання складних систем на основі кольорових мереж Петрі. Навчальний посібник [Текст]. Херсон: Книжкове видавництво ФОП Вишемірський В.С., 2021. 58 с.

70. Matsuki Y., Gozhyj A., Kalinina I., Bidyuk P. Method to Find the Original Source of COVID-19 by Genome Sequence and Probability of Electron Capture. *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*, 2022 (2023), vol. 149, pp. 214–230. Springer, Cham. DOI: 10.1007/978-3-031-16203-9_13. (*bookchapter*, *Scopus*)

71. Demchuk A., Rusyn B., Pohreliuk L., Gozhyj A., Kalinina I., Chyrun L., Antonyuk N. Commercial content distribution system based on neural network and machine learning. *CEUR-WS*. 2019, 2516, pp. 40-57. CEUR-WS.org/Vol-2516/paper3.pdf. (*Scopus*).

АНОТАЦІЯ

Калініна І.О. Моделі та інформаційні технології ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів в задачах машинного навчання. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Чорноморський національний університет імені Петра Могили, Миколаїв, Українська академія друкарства, Міністерство освіти і науки України, Львів, 2023.

У дисертаційній роботі вирішено актуальну науково-практичну проблему: підвищення ефективності ймовірнісно-статистичного аналізу даних, моделювання та прогнозування в задачах машинного навчання засобами сучасних інформаційних технологій з урахуванням нелінійності і нестаціонарності даних, а також можливих невизначеностей, що є характерними для них.

У роботі проаналізовано стан досліджень в області вирішення задач ймовірнісно-статистичного аналізу нелінійних нестаціонарних процесів, показані певні напрямки досліджень по створенню інформаційних технологій аналізу в задачах машинного навчання. Доведено, що проблема ефективного вирішення задач ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних та використання методів, моделей та сучасних інформаційних технологій для вирішення задач машинного навчання є актуальною. Визначено головні аспекти побудови інформаційних технологій ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних в задачах машинного навчання. Розроблено

структуру процесу обробки інформації при розв'язанні задач машинного навчання. Сформульовано математичну постановку задачі ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних при вирішенні задач машинного навчання. Досліджено особливості байєсівського підходу для вирішення задач ймовірнісно-статистичного аналізу даних у задачах машинного навчання. Розроблено метод синтезу інформаційних технологій для розв'язування задач машинного навчання з врахуванням нелінійностей та нестаціонарностей даних.

Розроблено метод обробки пропусків в даних, який ідентифікує пропуски в даних, виявляє закономірності їх появи та формує набори даних без пропусків. Розроблено метод виявлення та обробки аномальних значень в наборах даних, який ідентифікує екстремальні дані, аналізує причини їх появи, та здійснює їх обробку. Розвинуто метод для вирішення завдань фільтрації даних на основі байєсівського підходу та методу гранулярної фільтрації. Розвинуто метод нормалізації та стандартизації даних на основі системного поєднання методів перетворення даних та особливостей вирішення завдань машинного навчання.

Виконано класифікацію задач і методів моделювання в завданнях машинного навчання, яка надає можливість вибрати метод побудови моделі залежно від конкретної постановки задачі. Розроблено метод до побудови моделей на основі байєсівських часових рядів для вирішення завдань машинного навчання. Розроблено метод побудови моделей аналізу та моделювання нелінійних нестаціонарних процесів на основі колірних мереж Петрі. Розроблено метод синтезу параметрів нелінійної прогнозної моделі за допомогою генетичного алгоритму.

Розроблено системний підхід до розв'язання задач моделювання та прогнозування на основі ймовірнісно-статистичного аналізу нелінійних нестаціонарних даних в процедурах машинного навчання. Він об'єднує на системній основі методи та методології, спрямовані на вирішення таких задач: аналізу та попередньої обробки даних; побудови моделей та їх оцінки; побудови прогнозів та процедур їх оцінювання. Розглянуто вирішення задачі прогнозування на основі нейронних мереж для структурованого набору даних та для часових рядів. Розроблено метод побудови комбінованих прогнозів на основі часових рядів. Запропоновано архітектуру інформаційно-аналітичної системи прогнозування на основі комбінування прогнозів. Для вирішення задач класифікації розроблено алгоритм побудови багаторівневого гетерогенного ансамблю моделей. Розроблено дворівневу архітектуру системи прогнозування на основі методів *Staking* та *Bagging*. Запропоновано архітектуру інтелектуальної системи класифікації на основі дворівневого гетерогенного ансамбля моделей.

Розроблено структурні моделі інформаційних технологій ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання та прогнозування. Розроблено архітектури інформаційно-аналітичних систем для вирішення завдань ймовірнісно-статистичного аналізу та попередньої обробки даних, моделювання та прогнозування нелінійних нестаціонарних процесів, характерних для досліджуваних галузей.

Здійснено практичну реалізацію розроблених моделей, методів та інформаційних технологій при вирішенні практичних задач машинного навчання.

Ключові слова: ймовірнісно-статистичний аналіз, машинне навчання, нелінійність, нестаціонарність, невизначеності, синтез інформаційних технологій, аналіз та попередня обробка даних, моделювання, прогнозування, колірні мережі Петрі, ансамблі гетерогенних моделей, байєсівські структурні часові ряди, комбіновані прогнози.

ANNOTATION

Kalinina I.O. Models and information technologies of probabilistic-statistical analysis of non-linear non-stationary processes in machine learning problems. – On the rights of the manuscript.

Dissertation for obtaining the Doctor of Technical Sciences degreein the speciality 05.13.06 –Information Technologies. – Black Sea National University named after Petro Mohyla, Mykolaiv, Ukrainian Academy of Printing, Ministry of Education and Science of Ukraine, Lviv, 2023.

The dissertation solves an actual scientific and practical problem: increasing the efficiency of probabilistic and statistical data analysis, modelling and forecasting in machine learning tasks by means of modern information technologies, taking into account the nonlinearity and non-stationarity of data, as well as possible uncertainties that are characteristic of them.

The work analyses the state of research in the field of solving problems of probabilistic and statistical analysis of non-linear non-stationary processes, shows certain directions of research on the development of information technologies for analysis in machine learning problems. It has been proven that the problem of effectively solving problems of probabilistic statistical analysis of non-linear non-stationary data and the use of methods, models and modern information technologies for solving machine learning problems is relevant. The main aspects of the construction of information technologies for probabilistic and statistical analysis of non-linear non-stationary data in machine learning problems are defined. The structure of the information processing process for solving machine learning problems has been developed. The mathematical formulation of the probabilistic statistical analysis of non-linear non-stationary data in solving machine learning problems is formulated. Peculiarities of the Bayesian approach to solving probabilistic-statistical data analysis problems in machine learning problems are studied. A method of synthesis of information technologies has been developed for solving machine learning problems taking into account non-linearities and non-stationarity of data.

A method for processing gaps in data has been developed, which identifies gaps in data, reveals patterns of their occurrence, and forms data sets without gaps. A method of detecting and processing anomalous values in data sets has been developed, which identifies extreme data, analyses the reasons for their appearance, and performs their processing. A method for solving data filtering problems based on the Bayesian approach and the granular filtering method has been developed. A method of data normalization and standardization has been developed based on a systematic

combination of data transformation methods and features of solving machine learning tasks.

The classification of tasks and modelling methods in machine learning tasks has been carried out, which provides an opportunity to choose a method of building a model depending on the specific formulation of the task. A method for building models based on Bayesian time series for solving machine learning problems has been developed. A method of building models of analysis and simulation of non-linear non-stationary processes based on colour Petri nets has been developed. A method of synthesizing the parameters of a nonlinear predictive model using a genetic algorithm has been developed.

A systematic approach to solving modelling and forecasting problems based on probabilistic-statistical analysis of non-linear non-stationary data in machine learning procedures has been developed. It combines on a systematic basis methods and methodologies aimed at solving the following problems: data analysis and pre-processing; construction of models and their evaluation; construction of forecasts and their evaluation procedures. Forecasting based on neural networks for structured datasets and for time series is considered. A method of building combined forecasts based on time series has been developed. The architecture of the information and analytical forecasting system based on combined forecasts is proposed. An algorithm for building a multi-level heterogeneous ensemble was developed to solve classification problems. A two-level forecasting system architecture based on Staking and Bagging methods has been developed. The architecture of an intelligent classification system based on a two-level heterogeneous ensemble of models is proposed.

Structural models of information technologies of probabilistic statistical analysis and data pre-processing, modelling and forecasting have been developed. Architectures of information-analytical systems have been developed to solve the problems of probabilistic-statistical analysis and data pre-processing, modelling and forecasting of non-linear non-stationary processes characteristic of the studied industries.

Practical implementation of the developed models, methods and information technologies in solving practical problems of machine learning was carried out.

Keywords: probabilistic statistical analysis, machine learning. nonlinearity, non-stationarity, uncertainties, synthesis of information technologies, data analysis and pre-processing, modelling, forecasting, colour Petri nets, ensembles of heterogeneous models, Bayesian structural time series, combined forecasts.