УДК 004.052.4:[330.322:37 **DOI:** 10.31673/2412-9070.2025.061205
Т. О. БАЖАН, ст. викл.;
ORCID: 0009-0007-6594-1695
В. Ф. КРИВОРУЧКО, аспірант,
ORCID: 0009-0004-4247-5722
Державний університет інформаційно-комунікаційних технологій, Київ

МЕТОДИ ОЧИЩЕННЯ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ ІНВЕСТИЦІЙ В ОСВІТУ

У статті розглядається актуальність очищення даних у контексті прогнозування інвестицій в освітню галузь, підкреслюючи, що якість вхідних даних є вирішальною для точності та надійності прогностичних моделей машинного навчання. Неякісні дані призводять до спотворення виявлених закономірностей та, як наслідок, до помилкових прогнозів інвестицій, що може негативно позначитися на розподілі фінансових ресурсів та розвитку освітньої системи. Специфіка освітніх даних, їхня різноманітність та схильність до помилок обумовлюють гостру потребу у ретельному очищенні.

На основі аналізу наявної літератури виявлено, що, хоча існує значний обсяг досліджень, присвячених загальним методам очищення даних та застосуванню машинного навчання в освіті, бракує цілеспрямованих робіт, що детально вивчають ефективність різних методів очищення даних саме для підвищення точності прогнозування інвестицій в освітню сферу. Це підкреслює наукову новизну та актуальність проведеного дослідження.

Метою дослідження є розробка та обґрунтування ефективного методу очищення даних, спрямованого на підвищення точності прогнозування інвестицій в освіту. Для досягнення цієї мети було поставлено низку завдань, включаючи аналіз існуючих методів, їхній порівняльний аналіз, виявлення найбільш придатних підходів, розробку можливих удосконалень, створення блок-схеми запропонованого методу та формування практичних рекомендацій.

У роботі детально розглянуто фундаментальний етап очищення даних у пайплайні машинного навчання, який передує створенню та навчанню моделей. Представлено порівняльний аналіз основних методів очищення даних, таких як обробка відсутніх значень (видалення рядків/стовпців, імпутація середнім/медіаною/модою, імпутація прогнозуванням), виявлення та обробка викидів (візуалізація, статистичні методи, алгоритми машинного навчання, перетворення викидів), видалення дублікатів, виправлення помилок та невідповідностей (перевірка правопису/формату, узгодження джерел, валідація за правилами), а також масштабування та нормалізація даних (Min-Max Scaling, StandardScaler) та перетворення типів даних.

Запропоновано виділення найкращих методів очищення для прогнозування інвестицій в освіту, враховуючи специфіку освітніх даних. До них віднесено комплексну обробку відсутніх значень, робастні методи виявлення та обробки викидів, ретельне виявлення та усунення дублікатів, строгу валідацію даних на основі правил та обмежень предметної області, а також узгодження форматів та перетворення типів даних. Обговорено можливості для вдосконалення методів очищення, зокрема розробку гібридних підходів, врахування контексту освітніх даних, автоматизацію процесу очищення з використанням машинного навчання, створення інтерактивних інструментів та оцінку впливу методів очищення на якість прогнозів.

Надано практичні рекомендації щодо використання методів очищення даних для прогнозів інвестицій в освіту, акцентуючи увагу на розумінні специфіки освітніх даних (їхнє походження, ієрархічна структура, часові залежності, категоріальні ознаки, чут-

ливість до змін у політиці), комплексній обробці відсутніх значень, робастному виявленні та обробці викидів, специфічних методах очищення для освітніх даних (стандартизація категоріальних ознак, контроль консистентності між рівнями, валідація за стандартами), інтеграції та узгодженні даних з різних джерел, а також на оцінці впливу очищення на якість прогнозів. Підкреслюється важливість залучення експертів з освітньої галузі на всіх етапах процесу.

У висновках зазначено, що якісне очищення даних є критично важливим для побудови надійних прогностичних моделей у сфері інвестицій в освіту. Запропонований комплексний підхід, що поєднує обробку відсутніх значень та робастні методи виявлення й обробки викидів, дозволяє значно покращити якість вхідних даних та підвищити точність прогнозів. Перспективи подальших досліджень включають апробацію методу на більших обсягах реальних даних та порівняння його ефективності з іншими підходами.

Ключові слова: очищення даних; якість даних; машинне навчання; прогнозування інвестицій; освітні дані; моделі прогнозування.

Вступ

Актуальність очищення даних у контексті прогнозування інвестицій в освіту є беззаперечною. Якість вхідних даних є фундаментом для побудови точних і надійних моделей машинного навчання для прогнозування. Навіть найскладніші алгоритми не зможуть виявити істинні закономірності та зробити обґрунтовані прогнози, якщо їх навчати на неякісних, так званих "брудних" даних. Такі дані неминуче призведуть до спотворення виявлених залежностей і, як наслідок, до помилкових прогнозів інвестицій. Специфіка освітньої галузі характеризується складністю та різноманітністю даних, що використовуються для аналізу та прогнозування. До них належать демографічні показники, економічні індикатори, статистичні дані про успішність учнів, інформація про освітню інфраструктуру та багато інших. Кожен з цих типів даних може бути схильний до неповноти, суперечливості або містити різноманітні помилки, що підкреслює необхідність ретельного очищення. Точні прогнози інвестицій в освіту мають вирішальне значення для прийняття обґрунтованих рішень щодо ефективного розподілу фінансових ресурсів, стратегічного планування освітньої політики та досягнення ключових цілей у цій важливій сфері. Неправильні прогнози можуть призвести до неефективного використання бюджетних коштів, уповільнення розвитку освітньої системи та, зрештою, негативно позначитися на якості освіти. Крім того, в епоху стрімкого розвитку інформаційних технологій та зростання обсягів даних, зокрема і в освітній сфері, проблема якісного очищення даних набуває ще більшої актуальності. Великі обсяги даних (Big Data) несуть у собі як значний потенціал для глибокого аналізу та точного прогнозування, так і підвищені ризики, пов'язані з наявністю помилок, дублікатів та інших недоліків, які можуть суттєво вплинути на результати моделювання. Таким чином, якісне очищення даних є невід'ємною передумовою для успішного застосування методів машинного навчання у прогнозуванні інвестицій в освіту.

Аналіз останніх досліджень

Аналізуючи наявну літературу, слід зазначити, що безпосередньо досліджень, які б фокусувалися виключно на очищенні даних для прогнозування інвестицій саме в освіту, наразі не так багато. Однак, існує значний обсяг робіт, присвячених загальним методам очищення даних у контексті машинного навчання та аналізу даних, а також досліджень, що розглядають застосування машинного навчання для прогнозування в освітній галузі, де питання якості даних також є важливим, хоча й не завжди центральним.

Так, у статті [1] детально розглядають різні методи очищення даних, включаючи обробку відсутніх значень, виявлення та усунення шумів і викидів, а також інтеграцію та трансформацію даних. Ці методи є універсальними і можуть бути застосовані до даних будь-якої предметної області, включаючи освітню.

У статті [2] систематизують основні проблеми, пов'язані з якістю даних, та пропонують загальні підходи до їхнього вирішення. Вони підкреслюють важливість розуміння джерел даних та характеру можливих помилок для вибору ефективних методів очищення.

Деякі дослідники, наприклад, у книзі [3], акцентують увагу на практичних аспектах очищення даних, пропонуючи конкретні методи та інструменти для підвищення якості даних на різних етапах їхньої обробки.

У контексті застосування машинного навчання в освіті, багато авторів підкреслюють важливість якісних даних для побудови ефективних прогностичних моделей. Наприклад, у роботах, присвячених прогнозуванню успішності студентів або відтоку учнів, часто згадується необхідність попередньої обробки та очищення даних від неточностей та пропусків. Однак, детальний аналіз конкретних методів очищення, адаптованих саме для підвищення точності прогнозування інвестицій, зазвичай не є основ-ним фокусом цих досліджень.

Щодо робіт українських науковців [4] і [5], то дослідження в галузі застосування машинного навчання в освіті також розвиваються. Проте, поки що не вдалося знайти значної кількості публікацій, які б безпосередньо розглядали питання очищення даних саме для цілей прогнозування інвестицій в освітню галузь. Більшість робіт зосереджені на застосуванні прогностичних моделей для аналізу успішності навчання, прогнозування потреб у певних освітніх послугах або аналізу ефективності освітніх програм. У цих роботах питання очищення даних часто розглядається як один із попередніх етапів, але без глибокого аналізу та порівняння різних методів саме для підвищення якості прогнозів інвестицій.

Таким чином, аналіз літератури показує, що хоча існують фундаментальні дослідження в галузі очищення даних та роботи, що застосовують машинне навчання для прогнозування в освіті, бракує цілеспрямованих досліджень, які б детально вивчали та порівнювали ефективність різних методів очищення даних спеціально для підвищення точності прогнозування інвестицій в освіту. Це підкреслює актуальність дослідження та потенційну наукову новизну розробки та обґрунтування методів очищення даних, адаптованих до специфіки освітніх даних та завдань прогнозування інвестицій.

Метою даної наукової статті є розробка та обґрунтування ефективного методу очищення даних, спрямованого на підвищення точності прогнозування інвестицій в освітню галузь.

Для досягнення поставленої мети передбачається вирішити наступні завдання:

1. Провести аналіз існуючих методів очищення даних, що застосовуються в контексті машинного навчання та прогнозування.

2. Здійснити порівняльний аналіз ефективності різних методів очищення даних на прикладі реальних або змодельованих даних, релевантних для прогнозування інвестицій в освіту.

3. Виявити найбільш придатні методи очищення даних, які забезпечують найвищу точність прогнозів інвестицій в освітній сфері.

4. Запропонувати можливі удосконалення до обраних методів очищення з урахуванням специфіки освітніх даних та завдань прогнозування.

5. Розробити наочну блок-схему запропонованого або модифікованого методу очищення даних для його практичного застосування.

6. Сформулювати практичні рекомендації щодо використання розробленого методу очищення даних для підвищення якості прогнозування інвестицій в освіту.

Вирішення зазначених завдань дозволить досягти поставленої мети та зробити внесок у розробку більш ефективних підходів до аналізу та прогнозування в освітній галузі.

Основна частина

Очищення даних – фундаментальний етап для якісного прогнозування інвестицій в освіту.

Якість даних є критично важливим фактором, що визначає успіх будь-якого проєкту машинного навчання, особливо в такій чутливій галузі, як прогнозування інвестицій в освіту. Використання неякісних, "брудних" даних неминуче призводить до побудови неефективних моделей, які генерують спотворені прогнози та можуть стати причиною прийняття помилкових управлінських рішень. Неточності, пропуски, викиди, дублікати та невідповідності в даних можуть приховати істинні закономірності, призвести до зміщення оцінок параметрів моделі та, як наслідок, до значних фінансових втрат та неефективного розподілу ресурсів в освіт-

ній сфері. Тому, ретельне очищення даних є не просто бажаним, а абсолютно необхідним етапом для забезпечення надійності та практичної цінності прогнозів.

Етап очищення даних у пайплайні машинного навчання.

Процес машинного навчання зазвичай включає кілька послідовних етапів. Очищення даних є одним із ключових попередніх етапів, що слідує за збором та первинною обробкою даних (наприклад, об'єднанням різних джерел). Зазвичай етап очищення передує етапам feature engineering (створення нових ознак), feature selection (відбору важливих ознак) та безпосередньо навчанню моделі. Якісно очищені дані є фундаментом для всіх наступних кроків. Якщо на етапі очищення буде допущено помилки або не усунуто існуючі проблеми з даними, це неминуче позначиться на якості створеної моделі та її здатності до точного прогнозування. Інвестиції часу та зусиль в ретельне очищення даних окупаються за рахунок підвищення точності та надійності кінцевого прогнозу.

Розглянемо різні способи та виконаємо порівняльний аналіз цих методів.

Існує широкий спектр методів очищення даних, кожен з яких має свої особливості, переваги та недоліки. Нижче наведено порівняльну таблицю (табл. 1) основних методів, які застосовуються для обробки даних.

Порівняння основних методів очистки даних для обробки даних

Метод очищення	**Опис методу**	**Переваги**	**Недоліки**	**Випадки застосування (коли метод найбільш ефективний)**
Видалення рядків/ стовпців	Видалення записів або ознак, що містять значну кількість відсутніх значень.	Простота реалізації.	Може призвести до втрати цінної інформації, особливо якщо відсутні значення не є випадковими. Зменшує розмір вибірки.	Коли відсутніх значень дуже багато і їх розподіл є випадковим. Для видалення ознак з критично високим відсотком пропусків.
Імпутація середнім/ медіаною	Заповнення відсутніх числових значень середнім або медіанним значенням відповідної ознаки.	Простота реалізації. Зберігає розмір вибірки.	Може спотворити розподіл даних та зменшити дисперсію ознаки. Не враховує залежності між ознаками.	Коли відсутніх значень небагато і розподіл ознаки близький до нормального (середнє) або містить викиди (медіана).
Імпутація модою	Заповнення відсутніх категоріальних значень значенням (модою), яке найбільш часто зустрічається.	Простота реалізації. Зберігає розмір вибірки.	Може ввести упередження, особливо якщо мода не є репрезентативною .	Для категоріальних ознак з невеликою кількістю відсутніх значень.

Імпутація прогнозу-ванням	Використання моделей машинного навчання (наприклад, регресії або класифікації) для прогнозування відсутніх значень на основі інших наявних ознак.	Потенційно забезпечує більш точне заповнення, враховуючи залежності між ознаками.	Складність реалізації та вибору моделі для імпутації. Може ввести упередження, якщо модель для імпутації є неточною.	Коли існує сильна кореляція між ознаками і відсутні значення, ймовірно, залежать від інших ознак.

У наступній таблиці (табл. 2) розглянуто порівняння методів очистки даних для виявлення та обробки викидів. Встановлено переваги та недоліки кожного методу і розглянуто випадки, коли метод найбільш ефективний.

Порівняння основних методів очистки даних для виявлення та обробки викидів

Метод очищення	**Опис методу**	**Переваги**	**Недоліки**	**Випадки застосування (коли метод найбільш ефективний)**
Візуалізація	Використання графічних методів (діаграми розсіювання, box plots, гістограми) для візуального виявлення аномальних значень.	Дозволяє отримати інтуїтивне розуміння розподілу даних та виявити потенційні викиди.	Суб'єктивність інтерпретації. Може бути складно застосовувати для багатовимірних даних.	На початковому етапі аналізу даних для виявлення потенційних проблем. Для перевірки результатів статистичних методів.
Статистичні методи	Використання статистичних правил (наприклад, правило трьох сигм, міжквартильний розмах - IQR) для визначення викидів на основі розподілу даних.	Об'єктивність (при правильному виборі порогів). Простота застосування.	Припущення про нормальний розподіл (для правила трьох сигм). Може не виявляти складні комбінації викидів. Чутливість до екстремальних значень при розрахунку параметрів розподілу.	Для виявлення одиничних викидів у даних з відомим або близьким до нормального розподілом (правило трьох сигм) або в даних з довільним розподілом (IQR).
Алгоритми машинного навчання	Застосування спеціалізованих алгоритмів (наприклад, Isolation Forest, DBSCAN, One-Class SVM) для виявлення аномалій у багатовимірних даних.	Здатність виявляти складні та контекстуальні аномалії, які можуть бути непомітні при використанні статистичних методів.	Складність налаштування гіперпараметрів. Потребує достатньої кількості даних для навчання. Результати можуть бути менш інтерпретованими.	Для виявлення складних аномалій у багатовимірних даних, коли статистичні методи є неефективними.

Перетворення викидів	Заміна екстремальних значень на менш екстремальні (наприклад, вінзоризація) або застосування математичних перетворень (наприклад, логарифмування) для зменшення їхнього впливу.	Дозволяє зберегти інформацію, що міститься у викидах, зменшуючи їхній вплив на модель.	Може спотворити справжній розподіл даних. Необхідність обґрунтування вибору методу перетворення та меж для вінзоризації.	Коли викиди є результатом помилок вимірювання або мають значний вплив на модель, але їх видалення може призвести до втрати цінної інформації.
Видалення дублікатів	Видалення повністю або частково ідентичних записів у наборі даних. Часткове видалення може ґрунтуватися на певних ключових ознаках.	Забезпечує унікальність записів, що може бути важливим для деяких моделей та запобігає спотворенню статистики.	Необхідність чіткого визначення критеріїв дублікації, особливо при частковому видаленні. Можлива втрата інформації, якщо дублікати містять різну цінну інформацію в неключових полях.	Для усунення випадкових повторень записів або стандартизації даних з різних джерел, де можуть бути дублікати.

Табл. 3 дає нам порівняння методів для виправлення помилок та невідповідностей. Тут розглянуто основні методи очистки даних, які використовуються, саме для виправлення помилок та невідповідностей.

Порівняння основних методів очистки даних для виправлення помилок та невідповідностей

Метод очищення	**Опис методу**	**Переваги**	**Недоліки**	**Випадки застосування (коли метод найбільш ефективний)**
Перевірка правопису/формату	Використання інструментів для перевірки орфографії, узгодження форматів даних (наприклад, дати, числа).	Покращує консистентність та читабельність даних. Зменшує кількість помилок, пов'язаних з людським фактором.	Може бути ресурсомістким для великих обсягів неструктурованих даних. Потребує налаштування правил для специфічних форматів.	Для текстових даних, що містять орфографічні помилки або даних з різними форматами представлення однієї й тієї ж інформації.

Узгодження джерел	Вирішення конфліктів та об'єднання інформації з різних джерел даних, що можуть містити суперечливі записи про одні й ті самі об'єкти.	Забезпечує цілісність та узгодженість даних, отриманих з різних систем.	Складність ідентифікації відповідних записів у різних джерелах. Потребує розробки стратегій вирішення конфліктів (якому джерелу довіряти більше, як агрегувати інформацію).	При інтеграції даних з кількох різних баз даних або систем.
Узгодження джерел	Вирішення конфліктів та об'єднання інформації з різних джерел даних, що можуть містити суперечливі записи про одні й ті самі об'єкти.	Забезпечує цілісність та узгодженість даних, отриманих з різних систем.	Складність ідентифікації відповідних записів у різних джерелах. Потребує розробки стратегій вирішення конфліктів (якому джерелу довіряти більше, як агрегувати інформацію).	При інтеграції даних з кількох різних баз даних або систем.
Валідація за правилами	Використання попередньо визначених правил та обмежень (наприклад, діапазони значень, логічні зв'язки між ознаками) для виявлення та виправлення некоректних даних.	Дозволяє виявити системні помилки та невідповідності, що порушують бізнес-логіку або обмеження предметної області.	Потребує глибокого розуміння предметної області та формалізації правил валідації.	Для даних, що підлягають певним бізнес-правилам або обмеженням, наприклад, вікові обмеження, логічні зв'язки між показниками.

І, нарешті, порівняння основних методів очистки даних розглянуто у табл. 4, які використовуються для масштабування та нормалізації даних.

Порівняння основних методів очистки даних для масштабування та нормалізації

Метод очищення	Опис методу	Переваги	Недоліки	Випадки застосування (коли метод найбільш ефективний)
Min-Max Scaling	Лінійне масштабування ознак до заданого діапазону (зазвичай [0, 1]).	Зберігає відносні співвідношення між значеннями. Корисний для алгоритмів, чутливих до масштабу ознак.	Чутливий до викидів.	Коли всі ознаки мають приблизно однаковий діапазон значень або коли алгоритм вимагає певного діапазону вхідних даних.
StandardScaler	Стандартизація ознак шляхом віднімання середнього значення та ділення на стандартне відхилення. Приводить ознаки до нульового середнього та одиничної дисперсії.	Менш чутливий до викидів, ніж Min-Max Scaling. Корисний для багатьох алгоритмів, особливо тих, що базуються на відстанях.	Може спотворити розподіл, якщо він суттєво відрізняється від нормального.	Коли розподіл ознак близький до нормального або коли алгоритм чутливий до дисперсії ознак.
Перетворення типів даних	Приведення даних до потрібного формату (наприклад, перетворення рядкових значень у числові або категоріальні).	Забезпечує можливість використання даних у моделях машинного навчання, які вимагають певних типів вхідних даних.	Можлива втрата інформації при перетворенні складних типів даних. Потребує розуміння семантики даних.	Коли дані представлені в невідповідному для аналізу форматі.
Видалення нерелевантних даних	Видалення ознак або записів, які не несуть корисної інформації для задачі прогнозування.	Зменшує розмірність даних, спрощує моделі, може покращити їхню продуктивність та запобігти перенавчанню.	Потребує ретельного аналізу значущості ознак. Видалення важливих ознак може призвести до зниження точності прогнозування.	На етапі попередньої обробки для видалення очевидно неінформативних ознак або викидів, що не несуть змістовного значення.

Виділення найкращих методів очищення для прогнозування інвестицій в освіту.

Враховуючи специфіку даних, що можуть використовуватися для прогнозування інвестицій в освіту (демографічні дані, економічні показники, статистика успішності учнів, дані про інфраструктуру тощо), а також потенційні проблеми, такі як пропуски у статистичних звітах, викиди, пов'язані з економічними кризами або регіональними особливостями та невідпо-

відності у форматах з різних джерел, можна виділити кілька найбільш перспективних методів очищення:

1. Комплексна обробка відсутніх значень.
2. Робастні методи виявлення та обробки викидів.
3. Ретельне виявлення та усунення дублікатів.
4. Строга валідація даних на основі правил та обмежень предметної області.
5. Узгодження форматів та перетворення типів даних.

Можливості для вдосконалення методів очищення даних для прогнозування інвестицій в освіту.

Для підвищення ефективності очищення даних, специфічно орієнтованого на прогнозування інвестицій в освіту, можна розглянути кілька важливих напрямків. По-перше, це розробка гібридних методів очищення (рис. 1). Комбінування різних підходів може дозволити скористатися перевагами кожного з них та компенсувати їхні недоліки. Наприклад, спершу можна використовувати статистичні методи для первинного виявлення викидів, а потім застосовувати алгоритми машинного навчання для ідентифікації більш складних та неочевидних аномалій у даних.

По-друге, необхідно враховувати контекст освітніх даних. Розробка методів очищення повинна базуватися на специфічних характеристиках освітніх даних, таких як ієрархічна структура (наприклад, школа – район – область – країна), часові залежності між показниками та взаємозв'язок між різними освітніми індикаторами. Такий підхід забезпечить більш точне та релевантне очищення.

Третій напрямок полягає в автоматизації процесу очищення з використанням машинного навчання. Застосування алгоритмів машинного навчання не обмежується лише імпутацією відсутніх значень або виявленням викидів. Вони також можуть бути використані для автоматичного виправлення помилок на основі виявлених закономірностей у даних, що значно пришвидшить та спростить процес.

Четвертий аспект - це розробка інтерактивних інструментів для очищення даних. Створення зручних та інтуїтивно зрозумілих інструментів дозволить експертам з предметної області (освіти) активно долучатися до процесу очищення даних. Використання їхніх знань та інтуїції є безцінним для виявлення та виправлення помилок, які можуть бути неочевидними для автоматичних алгоритмів.

Нарешті, надзвичайно важливою є оцінка впливу методів очищення на якість прогнозів. Необхідно проводити експерименти для кількісної оцінки того, як застосування різних методів очищення даних впливає на точність моделей прогнозування інвестицій в освіту. Це дозволить емпірично визначити та обґрунтувати найбільш ефективні підходи, забезпечуючи практичну цінність та надійність отриманих прогнозів.

Проведення експериментів для кількісної оцінки того, як застосування різних методів очищення даних впливає на точність моделей прогнозування інвестицій в освіту. Це дозволить емпірично визначити найбільш ефективні підходи.

Реалізація цих ідей може значно підвищити якість даних, що використовуються для прогнозування інвестицій в освіту, а отже, і точність та надійність самих прогнозів. Впровадження адаптованих та комбінованих методів очищення, які враховують специфіку освітньої галузі, є важливим кроком на шляху до прийняття більш обґрунтованих та ефективних рішень у сфері освітньої політики та розподілу фінансових ресурсів.

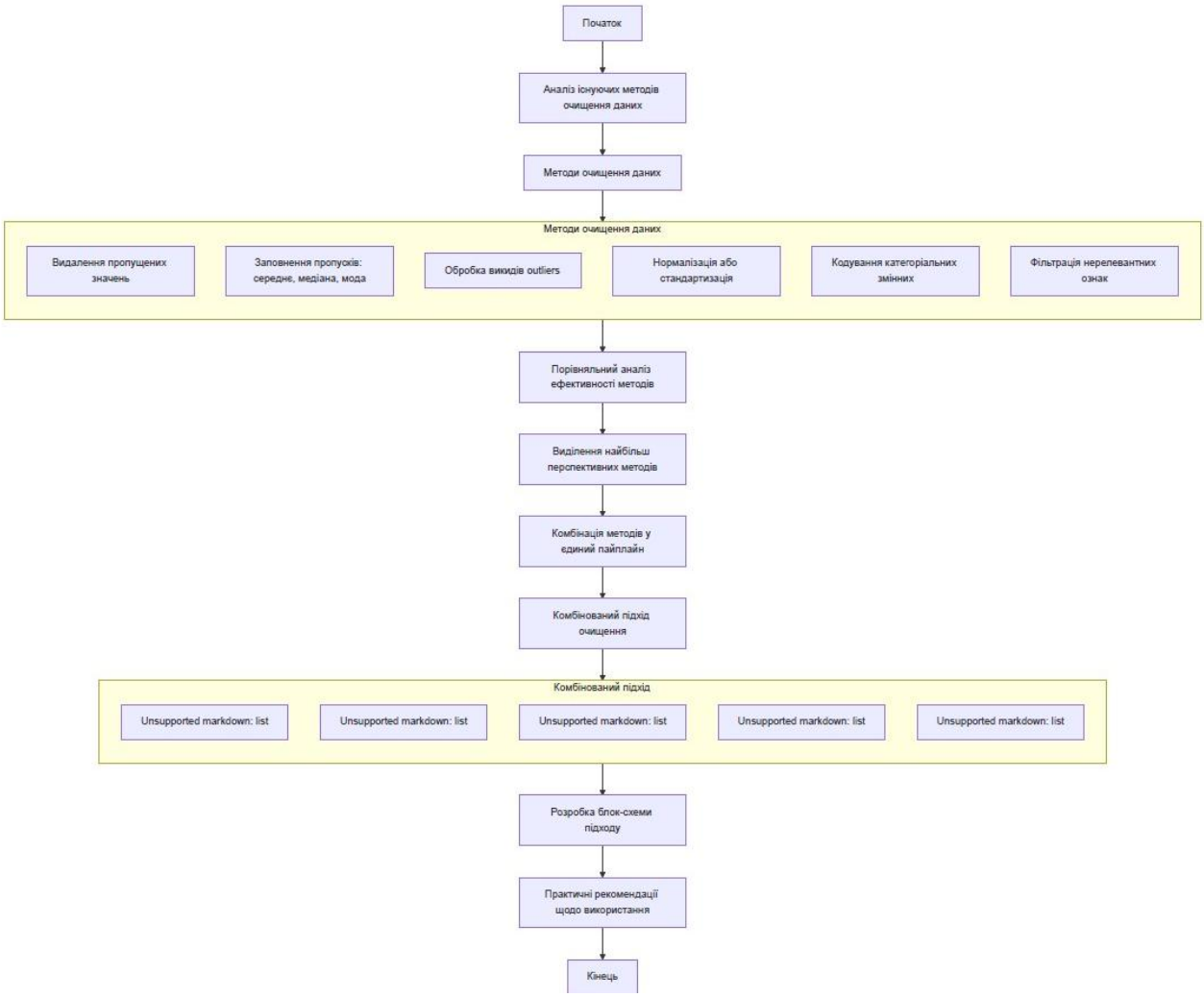

Рис. 1. Блок-схема модифікованого методу очищення даних

Ефективне застосування методів очищення даних є ключовим для отримання надійних прогнозів інвестицій в освіту, тому варто надати кілька практичних порад щодо їхнього використання. Запропонований комплексний підхід, який поєднує обробку відсутніх значень та робастні методи виявлення й обробки викидів є найбільш придатним для широкого спектра даних, що використовуються для прогнозування інвестицій в освіту. До таких даних належать статистичні відомості про освіту, фінансові показники, демографічні дані та соціально-економічні індикатори. Важливо зазначити, що цей метод є гнучким і може бути адаптований до різних типів даних, включаючи числові, категоріальні та часові ряди, причому для останніх слід враховувати специфічні підходи до обробки пропусків та аномалій, зумовлених часовою залежністю.

Для успішного застосування цих методів очищення даних рекомендується використовувати потужні інструменти та програмне забезпечення, такі як мови програмування Python або R з відповідними бібліотеками для аналізу даних та машинного навчання, середовища розробки на кшталт Jupyter Notebook або RStudio, інструменти візуалізації, такі як Matplotlib або ggplot2, а також системи керува-ння базами даних, як SQL для роботи з даними з різних джерел. Вибір конкретних інструментів визна-чається наявними навичками, обсягом даних та складністю аналітичних завдань.

Процес очищення даних може супроводжуватися певними складнощами. Інтеграція та очищення великих обсягів різнорідних даних може бути складним завданням, що потребує розробки чітких про-токолів та використання автоматизованих інструментів. Пропуски, які виникають через системні фак-тори, вимагають ретельного аналізу їхніх причин та використання методів імпутації, що враховують залежності між ознаками, а також залучення експертів

для розуміння контексту. Викиди, які можуть відображати реальні події, потребують обережного підходу до їхньої обробки та врахування контексту при інтерпретації прогнозів. Крім того, визначення обґрунтованих порогів для виявлення викидів може бути суб'єктивним, тому рекомендується експериментувати з різними значеннями та оцінювати їхній вплив на дані та якість прогнозів, консультуючись з експертами.

Оцінка якості очищених даних є важливим етапом, що включає перевірку на зменшення кількості помилок, повноту даних, збереження їхнього розподілу, зменшення впливу викидів та підвищення кон-систентності. Кінцевим критерієм є покращення продуктивності та точності моделей прогнозування.

Нарешті, надзвичайно важливим є залучення експертів з освітньої галузі на всіх етапах очищення даних. Їхні знання специфіки освітньої системи, джерел даних та потенційних проблем можуть значно підвищити якість процесу, допомогти у виборі методів, виявленні неправдоподібних значень та оцінці очищених даних з точки зору їхньої змістовності. Співпраця між аналітиками даних та експертами з освіти є запорукою успішного очищення даних та побудови надійних прогнозів інвестицій у цю важ-ливу сферу.

Рекомендації щодо використання методів очищення даних для прогнозів інвестицій в освіту

Прогнозування інвестицій в освіту вимагає особливої уваги до якості вхідних даних. Ось специфічні рекомендації щодо використання методів очищення даних у цьому контексті:

1. Розуміння специфіки освітніх даних. Дані можуть надходити з різних джерел: державні статистичні служби, міністерства освіти, місцеві органи управління освітою, дані навчальних закладів, соціологічні опитування, міжнародні бази даних. Кожне джерело може мати свої особливості формату, точності та повноти. Освітні дані часто мають ієрархічну структуру (учень – клас – школа – район – область – країна). При очищенні необхідно враховувати цю ієрархію та можливі невідповідності на різних рівнях. Багато освітніх показників є часовими рядами, що вимагає врахування часових залежностей при обробці відсутніх значень та викидів. Велика кількість важливих ознак може бути категоріальною (тип навчального закладу, рівень освіти, спеціальність тощо). Методи очищення для категоріальних даних (наприклад, стандартизація назв, обробка рідкісних категорій) є важливими. Дані можуть бути чутливими до змін у державній політиці, демографічних тенденціях та соціальних пріоритетах, що може призводити до різких змін або аномалій.

2. Комплексна обробка відсутніх значень в освітніх даних. Це включає аналіз причин пропусків та використання ієрархічної інформації для імпутації. Також варто застосовувати часові методи імпутації. Обережне видалення навчальних закладів або регіонів з великою кількістю пропусків є крайнім заходом, оскільки це може призвести до втрати важливої регіональної специфіки.

3. Робастне виявлення та обробка викидів в освітніх даних. При виявленні викидів необхідно враховувати можливий вплив економічних криз, реформ освіти або демографічних змін на інвестиційні показники. Метод міжквартильного розмаху (IQR) може бути корисним для виявлення нетипових рівнів фінансування або кількості учнів, які не відповідають загальній тенденції. Також ефективним є застосування алгоритмів машинного навчання для виявлення аномалій, зокрема Isolation Forest або DBSCAN, що можуть допомогти виявити складніші аномалії у багатовимірних даних (наприклад, нетипове співвідношення між різними освітніми показниками). Важливо проводити аналіз причин викидів: перш ніж обробляти їх, спробуйте зрозуміти їхню природу, адже різке зростання інвестицій може бути пов'язане з реалізацією великого освітнього проєкту.

4. Специфічні методи очищення для освітніх даних. Це включає стандартизацію категоріальних ознак для узгодження назв типів навчальних закладів, рівнів освіти та спеціальностей, щоб уникнути дублювання інформації через різне написання. Необхідний також контроль консистентності між різними рівнями, тобто перевірка узгодженості даних між рівнями ієрархії (наприклад, загальна кількість учнів в області має відповідати сумі кількості учнів у всіх

районах). Крім того, важлива валідація на основі освітніх стандартів та нормативів, що дозволяє використовувати знання про освітні стандарти та нормативи для виявлення неправдоподібних значень.

5. Інтеграція та узгодження даних з різних джерел. Важливо використовувати унікальні ідентифікатори та ефективно вирішувати конфлікти при об'єднанні даних.

6. Оцінка впливу очищення на прогнози інвестицій в освіту. Необхідно порівнювати точність прогнозних моделей, побудованих на неочищених та очищених даних. Для цього слід використовувати метрики, релевантні для задачі прогнозування інвестицій (наприклад, MAE, RMSE). Також вкрай важливо залучати експертів з освітньої галузі для оцінки правдоподібності прогнозів на основі очищених даних.

Застосування цих рекомендацій допоможе забезпечити високу якість даних, що використовуються для прогнозування інвестицій в освіту, що, в свою чергу, сприятиме прийняттю більш обґрунтованих та ефективних управлінських рішень у цій важливій сфері.

Висновки та перспективи

Проведене дослідження було спрямоване на розробку та обґрунтування ефективного підходу до очищення даних для підвищення точності прогнозування інвестицій в освіту. В рамках роботи було здійснено аналіз існуючих методів очищення даних, проведено порівняльний аналіз їхньої ефективності в контексті освітніх даних, виділено найбільш перспективні методи та запропоновано їхню комбінацію для комплексного очищення. Розроблено наочну блок-схему запропонованого підходу та надано практичні рекомендації щодо його використання, враховуючи специфіку даних освітньої галузі.

Результати дослідження підкреслюють критичну важливість якісного очищення даних як попереднього етапу для побудови надійних прогнозних моделей у сфері інвестицій в освіту. Застосування комплексного підходу, що поєднує обробку відсутніх значень та робастні методи виявлення та обробки викидів, дозволяє значно покращити якість вхідних даних, зменшити вплив шумів та аномалій та, як наслідок, підвищити точність майбутніх прогнозів.

Незважаючи на отримані результати, існують перспективи для подальших досліджень у цій галузі. Одним із напрямків є застосування розробленого методу на більших обсягах реальних даних про інвестиції в освіту для оцінки його практичної ефективності та масштабованості. Також важливим є порівняння ефективності запропонованого комплексного методу з іншими існуючими підходами до очищення даних, що використовуються в задачах прогнозування.

Подальші дослідження можуть бути спрямовані на врахування нових джерел даних, які стають доступними у сфері освіти, таких як дані з освітніх платформ, результати онлайн-навчання, дані про працевлаштування випускників тощо. Інтеграція та очищення цих нових типів даних може відкрити нові можливості для більш глибокого та точного прогнозування інвестиційних потреб.

Актуальним напрямком є розробка інструментів для автоматизації процесу очищення освітніх даних, що дозволить зменшити ручну працю, підвищити швидкість та консистентність обробки даних. У довгостроковій перспективі заслуговує на увагу розробка спеціалізованих інструментів, орієнтованих саме на очищення даних освітньої галузі, враховуючи її специфічні характеристики та потенційні проблеми з якістю даних.

Також актуальним продовженням даного дослідження є обрання найкращого методу машинного навчання для використання даного алгоритму очистки даних для прогнозування інвестицій.

Таким чином, проведене дослідження є важливим кроком у напрямку покращення якості прогнозування інвестицій в освіту через застосування ефективних методів очищення даних. Подальші дослідження та розробки в цій галузі мають значний потенціал для підвищення обґрунтованості прийняття стратегічних рішень у сфері освітньої політики та розподілу ресурсів.

Список літератури

1. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Amsterdam : Morgan Kaufmann, 2012. 744 p.

2. Rahm E., Do H.-H. Data cleaning: Problems and current approaches // IEEE Data Engineering Bulletin. 2000. Vol. 23, No. 4. P. 3–13.

3. Dasu T., Johnson T., Muthukrishnan S., Shiyi C. Data Quality and Data Cleansing: Practical Approaches. New York : Hewlett-Packard Labs, 2002. 34 p. (Technical Report HPL-2002-98).

4. Гончаренко С. У., Полякова О. В. Використання моделей машинного навчання у прогнозуванні результатів навчальної діяльності студентів // Інформаційні технології і засоби навчання. 2020. № 4 (78). С. 107–120. DOI: https://doi.org/10.33407/itlt.v78i4.3732.

5. Іващенко Г. М., Злочевська О. В. Прогнозування успішності навчання студентів з використанням методів машинного навчання // Освітній вимір. 2021. № 1. С. 25–33.

6. Chapman, W. T., Pignatiello, A. Data Quality in Educational Research: Challenges and Solutions. Journal of Educational Data Mining. 2019. Vol. 11, No. 2. P. 1–20.

7. Chen, Y., Liu, Y. A Comprehensive Review of Data Preprocessing Techniques for Educational Data Mining. International Journal of Machine Learning and Computing. 2020. Vol. 10, No. 4. P. 502–508.

8. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine. 1996. Vol. 17, No. 3. P. 37–54.

9. Kotsiantis, S. B., Kanellopoulos, D., Pintelas, P. E. Data Preprocessing for Educational Data Mining. Studies in Computational Intelligence. Berlin ; Heidelberg : Springer, 2006. Vol. 45. P. 1–21.

10. Malik, S., Hussain, M. A Survey of Data Cleaning Techniques for Big Data. Journal of Big Data. 2018. Vol. 5, No. 1. P. 1–24.

11. Pyle, D. Data Preparation for Data Mining. San Francisco : Morgan Kaufmann, 1999. 477 p.

12. Srivastava, P., Singh, R. Impact of Data Preprocessing on Machine Learning Model Performance: A Review. International Journal of Engineering Research & Technology. 2021. Vol. 10, No. 01. P. 1–6.

13. Wang, R., Strong, D. M. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996. Vol. 12, No. 4. P. 5–33.

14. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Steinberg, D. Top 10 Algorithms in Data Mining. Boca Raton : Chapman and Hall/CRC, 2008. 202 p.

15. Xu, S., Li, D. Big Data Analytics in Education: Challenges and Opportunities. In Proceedings of the International Conference on Big Data and Education (ICBDE). 2022. P. 100–105.

T. Bazhan, V. Kryvoruchko

METHODS OF DATA CLEANING FOR FORECASTING INVESTMENTS IN EDUCATION

This article addresses the relevance of data cleaning in the context of forecasting investments in the educational sector, emphasizing that the quality of input data is crucial for the accuracy and reliability of machine learning prognostic models. Poor quality data inevitably leads to distorted patterns and, consequently, to erroneous investment forecasts, which can negatively impact the allocation of financial resources and the development of the educational system. The specificity of educational data, its diversity, and susceptibility to errors highlight the urgent need for thorough cleaning.

Based on an analysis of existing literature, it was found that while there is a significant body of research on general data cleaning methods and the application of machine learning in education, there is a lack of focused studies that specifically investigate the effectiveness of various data cleaning methods for improving the accuracy of investment forecasting in the educational field. This underscores the scientific novelty and relevance of the conducted research.

The aim of the study is to develop and substantiate an effective data cleaning method aimed at improving the accuracy of forecasting investments in education. To achieve this goal, a number of tasks were set, including analyzing existing methods, conducting a comparative analysis of their effectiveness, identifying the most suitable approaches, developing possible enhancements, creating a block diagram of the proposed method, and formulating practical recommendations.

The paper thoroughly examines the fundamental stage of data cleaning in the machine learning pipeline, which precedes the creation and training of models. A comparative analysis of key data cleaning methods is presented, including handling missing values (row/column deletion, mean/median/mode imputation, predictive imputation), outlier detection and treatment (visualization, statistical methods, machine learning algorithms, outlier transformation), duplicate removal, error and inconsistency correction (spell check/format validation, source reconciliation, rule-based validation), as well as data scaling and normalization (Min-Max Scaling, StandardScaler) and data type conversion.

Selection of the best cleaning methods for forecasting investments in education is proposed, considering the specificity of educational data. These include comprehensive missing value handling, robust outlier detection and treatment, thorough duplicate detection and elimination, strict rule-based data validation using domain knowledge, and format harmonization and data type conversion. Opportunities for improving data cleaning methods are discussed, specifically the development of hybrid approaches, consideration of the context of educational data, automation of the cleaning process using machine learning, creation of interactive tools, and evaluation of the impact of cleaning methods on forecast quality.

Practical recommendations for using data cleaning methods for educational investment forecasts are provided, with an emphasis on understanding the specific characteristics of educational data (its origin, hierarchical structure, temporal dependencies, categorical features, sensitivity to policy changes), comprehensive handling of missing values, robust outlier detection and treatment, specific cleaning methods for educational data (standardization of categorical features, consistency control between levels, validation based on standards), integration and reconciliation of data from different sources, and evaluation of the impact of cleaning on forecast accuracy. The importance of involving experts from the educational sector at all stages of the process is highlighted.

In the conclusions, it is stated that high-quality data cleaning is critically important for building reliable prognostic models in the field of educational investment. The proposed comprehensive approach, combining missing value handling and robust outlier detection and treatment methods, significantly improves the quality of input data and enhances forecast accuracy. Prospects for further research include testing the method on larger volumes of real-world data and comparing its effectiveness with other existing data cleaning approaches.

Keywords: data cleaning; data quality; machine learning; investment forecasting; educational data; predictive models.