МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХЕРСОНСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

Факультет комп'ютерних наук, фізики та математики

Кафедра комп'ютерних наук та програмної інженерії

Автоматизоване машинне навчання для визначення цільової аудиторії вступників для закладів вищої освіти

Кваліфікаційна робота (проєкт)

на здобуття ступеня вищої освіти «магістр»

Виконав: студент 2 курсу 261М групи

Спеціальності

126 «Інформаційні системи та технології»

(шифр, назва)

Освітньо-професійної програми:

«Інформаційні системи та технології»

(назва)

Гулін Дмитро Вадимович

Керівник: доктор економічних наук, професор Кобець В.М.

Рецензент: Яцюта В.О.

Senior developer, team-lead, tech-lead, IT компанія DataArt

Івано Франківськ – 2024

ЗМІСТ

ВСТУП

У сучасному бізнес-середовищі, яке характеризується стрімким зростанням обсягів даних та необхідністю швидкого ухвалення обґрунтованих рішень, технології штучного інтелекту та машинного навчання стають важливим інструментом для підтримки процесу прийняття рішень. Бізнеси з різних галузей, від фінансових установ до ритейлу і виробництва, все частіше покладаються на ці технології для оптимізації процесів, підвищення ефективності та конкурентоспроможності. Завдяки автоматизації та машинному навчанню бізнес-аналітика здатна не лише вивчати історичні дані, але й будувати прогнози на майбутнє, що відкриває нові можливості для стратегічного планування та оперативного реагування на ринкові зміни.

Застосування автоматизованих систем машинного навчання у бізнес-аналітиці дозволяє ефективно аналізувати великі масиви даних, виявляти приховані закономірності, знаходити фактори, що впливають на динаміку показників, і прогнозувати події з високою точністю. Автоматизація аналізу даних зменшує залежність від ручної роботи, скорочує час на обробку інформації та мінімізує людські помилки, що підвищує точність і надійність аналітичних висновків. Це особливо актуально в умовах, коли дані надходять з різних джерел (наприклад, CRM-систем, маркетингових платформ, систем моніторингу соціальних медіа) та потребують комплексного підходу для аналізу.

Актуальність дослідження зумовлена зростаючою роллю автоматизованого машинного навчання у бізнес-аналітиці. Завдяки автоматизації процесів машинного навчання, компанії отримують не тільки перевагу у швидкості обробки інформації, але й забезпечують зменшення впливу людського фактора, що підвищує надійність отриманих даних. Окрім того, використання автоматизованих моделей дозволяє бізнесу фокусуватися на прийнятті стратегічних рішень, спираючись на точні прогнози та аналітичні дані.

Об’єктом дослідження є автоматизована підтримка прийняття рішень на основі даних.

Предметом дослідження є використання автоматизованих засобів машинного навчання для підтримки прийняття рішень керівників закладів вищої освіти на основі даних.

Мета дослідження полягає у розробці підходу до застосування автоматизованого машинного навчання в бізнес-аналітиці для підтримки прийняття рішень та оцінки його ефективності у контексті аналізу даних про вступників, що може слугувати базою для подальших управлінських рішень у сфері освіти.

Завдання дослідження:

1. Провести огляд літератури щодо ролі машинного навчання у прийнятті рішень та його застосувань у бізнес-аналітиці.
2. Дослідити основні параметричні і непараметричні методи машинного навчання в бізнес-аналітиці.
3. Розробити методологію аналізу даних вступників за допомогою автоматизованих методів машинного навчання для прогнозування контингенту майбутніх вступників.
4. Виконати експериментальний аналіз даних щодо вступу для визначення закономірностей та основних факторів впливу на рішення абітурієнтів.
5. Надати рекомендації щодо цільвої аудиторії вступників для освітніх програм ІТ спеціальностей на основі результатів автоматизованого аналізу даних вступників.

Методи дослідження включають використання алгоритмів автоматизованого машинного навчання, зокрема параметричних і непараметричних методів машинного навчання, а також статистичних методів для оцінки цільової аудиторії вступників. У дослідженні також використовувались алгоритми аналізу даних, що допомагають визначити ключові закономірності та тенденції у наборах даних про вступників.

Апробація. Апробацію магістерської роботи було проведено у вересні 2024 році на міжнародній конференції 19th International Conference on ICT in Education, Research, and Industrial Applications (ICTERI 2024), за результатами рецензування була відібрана стаття «Data-Driven Decision-Making to Identify the Target Audience of Higher Education Institutions Using Machine Learning Techniques» для публікації у виданні Springer, що індексується в Scopus.

Структура роботи. Дослідження містить вступ, три розділи, висновки і список використаних джерел. У першому розділі аналізується застосування ШІ і машинного навчання у прийнятті управлінських рішень у ЗВО. У другому розглядається методологія машинного навчання. Третій розділ присвячений прийняттю рішень на основі даних для визначення цільової аудиторії закладів вищої освіти з використанням методів машинного навчання.

РОЗДІЛ 1
АНАЛІЗ ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ І МАШИННОГО НАВЧАННЯ У ПРИЙНЯТТІ УПРАВЛІНСЬКИХ РІШЕНЬ У ЗВО

У статті Ляо [1] автори досліджують вплив навчання аудиторів з використання штучного інтелекту (ШІ) на швидкість підготовки аудиторських звітів. Важливим аспектом є те, що ШІ має потенціал прискорити багато рутинних процесів, які займають значну частину часу аудиторів. Це дозволяє фахівцям сконцентруватися на більш складних та стратегічних завданнях, таких як аналіз ризиків, перевірка фінансових даних або оцінка ефективності фінансового управління в організаціях [1].

Штучний інтелект здатний автоматизувати значну частину процесу аудиту, включаючи перевірку даних, пошук аномалій та виявлення потенційних шахрайств. Це робить аудиторські звіти більш точними та вивільняє час для аудиторів, що сприяє їх більш ефективній роботі. Крім того, навчання аудиторів використанню ШІ може мати важливий вплив на швидкість і своєчасність підготовки аудиторських звітів, що є ключовим фактором для сучасного бізнесу, оскільки прийняття рішень часто залежить від швидкості отримання фінансової інформації.

У дослідженні наголошується на важливості підвищення кваліфікації аудиторів шляхом навчання новітнім технологіям, що дозволяє їм використовувати переваги автоматизації. ШІ не тільки прискорює звітування, але й підвищує його точність. Це особливо важливо у випадках, коли аудитори працюють з великими обсягами фінансових даних, які можуть бути складними для перевірки вручну.

Таким чином, впровадження ШІ у сфері аудиту здатне суттєво скоротити час на підготовку звітів, що, в свою чергу, підвищує ефективність роботи аудиторських фірм та сприяє своєчасному наданню клієнтам фінансових звітів [1].

Дослідження Алкхаттані [2] присвячене ролі ШІ, природної мови та великих мовних моделей (LLM) у вищій освіті та наукових дослідженнях. Впровадження штучного інтелекту в навчальний процес створює нові можливості для студентів і викладачів. Моделі на основі природної мови, такі як ChatGPT, можуть автоматизувати рутинні завдання, наприклад, підготовку навчальних матеріалів, перевірку робіт або навіть проведення тестування знань [2].

Велика увага приділяється використанню таких моделей для персоналізації навчання. Застосування ШІ може суттєво вплинути на якість освітніх програм завдяки адаптації навчальних матеріалів під індивідуальні потреби студентів. Наприклад, штучний інтелект може аналізувати прогрес кожного студента, рекомендувати відповідні завдання або додаткові навчальні ресурси для поглибленого вивчення тем, які викликають труднощі.

Інше важливе застосування ШІ в освіті полягає у створенні віртуальних асистентів, здатних відповідати на запитання студентів у реальному часі. Це дозволяє значно розширити доступ студентів до інформації та забезпечити швидке вирішення навчальних проблем. Зокрема, віртуальні помічники можуть відповідати на запитання щодо навчальних програм, допомагати з підготовкою до іспитів або роз'яснювати складні теми [2].

Також слід зазначити, що великі мовні моделі можуть бути використані у наукових дослідженнях, де вони допомагають швидко аналізувати великі обсяги наукової інформації, генерувати ідеї для досліджень або навіть пропонувати нові підходи до вирішення наукових проблем. У дослідженні зазначається, що ШІ може полегшити роботу науковців шляхом автоматизації збору даних або обробки результатів досліджень, що робить науковий процес більш ефективним і прискорює отримання нових знань [2].

Загалом, використання ШІ та великих мовних моделей у вищій освіті та науці відкриває нові перспективи для поліпшення якості навчання та досліджень, надаючи студентам і викладачам нові інструменти для

персоналізованого підходу до навчання та автоматизації багатьох рутинних завдань.

У статті Таяна [3] обговорюються виклики та перспективи адаптації технологічних курсів у вищій освіті до впровадження великих мовних моделей, таких як ChatGPT. Автори зазначають, що хоча великі мовні моделі мають значний потенціал для підтримки освітнього процесу, існує також низка питань, які потребують вирішення для ефективного їх впровадження в навчальні програми [3].

Однією з головних проблем є необхідність навчання викладачів використанню ШІ та великих мовних моделей у навчальному процесі. Багато викладачів не мають достатнього рівня технічних знань для ефективного впровадження таких інструментів у свої курси. Це вимагає розробки спеціальних програм підвищення кваліфікації, які дозволять викладачам навчитися використовувати ШІ у повсякденній роботі.

Автори також наголошують на тому, що існує ризик надмірної залежності від ШІ, що може призвести до зниження якості викладання. Наприклад, якщо великі мовні моделі будуть використовуватися для автоматизації перевірки завдань або тестів, це може призвести до того, що студенти отримуватимуть менш індивідуальні відгуки на свою роботу, що може негативно вплинути на їх навчальний процес [3].

Крім того, важливо враховувати етичні питання, пов'язані з використанням ШІ в освіті. Наприклад, слід забезпечити прозорість алгоритмів, які використовуються для оцінки знань студентів, а також гарантувати, що дані студентів будуть захищені від несанкціонованого доступу або використання.

Водночас, великі мовні моделі можуть значно полегшити навчальний процес, особливо для студентів з обмеженими можливостями або тих, хто навчається на дистанційних програмах. Такі моделі можуть забезпечити доступ до навчальних матеріалів у будь-який час та допомогти студентам у вирішенні питань, які виникають у процесі навчання, без необхідності чекати відповіді від викладача [3].

Отже, хоча впровадження великих мовних моделей у навчальний процес має значний потенціал, для його успішної реалізації необхідно розв'язати низку проблем, зокрема підвищити кваліфікацію викладачів, забезпечити якісну оцінку знань та вирішити етичні питання.

Стаття Бубкера [4] присвячена дослідженню впливу інструментів штучного інтелекту на результати навчання студентів. Автор розглядає потенціал ШІ для покращення навчальних процесів та підвищення рівня знань у студентів. Особливу увагу приділено використанню адаптивних навчальних платформ, які підлаштовуються під рівень знань і темп навчання кожного студента, надаючи їм персоналізовані рекомендації для досягнення кращих результатів [4].

Інструменти ШІ, такі як навчальні помічники, можуть значно полегшити самонавчання, що дозволяє студентам глибше розуміти матеріал, не покладаючись виключно на традиційні методи навчання. Наприклад, використання віртуальних асистентів і автоматизованих систем для перевірки домашніх завдань дає студентам змогу отримувати миттєві відповіді та зворотний зв'язок щодо їхніх помилок. Це сприяє формуванню більшої відповідальності за власне навчання та зменшує час на очікування результатів [4].

Крім того, автор підкреслює, що інструменти ШІ здатні полегшити студентам доступ до додаткових освітніх ресурсів. Зокрема, за допомогою штучного інтелекту студенти можуть отримувати рекомендації щодо додаткової літератури, навчальних відео та інших матеріалів, які допомагають глибше зануритися у вивчення предмету. Це особливо корисно для студентів, які хочуть самостійно покращити свої знання поза межами традиційних занять [4].

Важливим аспектом також є те, що штучний інтелект дозволяє викладачам краще відстежувати прогрес студентів, виявляти слабкі місця в їхніх знаннях і адаптувати навчальні матеріали для кожного студента індивідуально. Це дозволяє забезпечити більш інклюзивний підхід до освіти,

особливо для студентів з різним рівнем підготовки або індивідуальними навчальними потребами [4].

Отже, інструменти штучного інтелекту мають великий потенціал для покращення результатів навчання, оскільки вони дозволяють студентам отримувати миттєві відповіді, доступ до додаткових ресурсів та персоналізовані рекомендації, що робить навчання більш ефективним і доступним для кожного.

В дослідженні Ванга [5] розглядається взаємодія штучного інтелекту та процесу прийняття рішень керівниками освітніх закладів на основі даних. Автор описує як ШІ може підтримувати освітніх лідерів у прийнятті обґрунтованих рішень, спираючись на аналіз великих обсягів даних, що дозволяє забезпечити більш точний контроль за освітнім процесом і підвищити його ефективність [5].

Однак, поряд з перевагами, дослідження Ванга наголошує на обережності щодо надмірного використання ШІ у процесі прийняття рішень. ШІ може створювати ризики для приватності даних студентів та викладачів, а також впливати на суб'єктивність у прийнятті рішень. Наприклад, автоматизовані системи можуть видавати рішення, засновані на недостатньо перевірених алгоритмах, що може призвести до помилкових висновків або упереджень щодо окремих груп студентів [5].

Автор також підкреслює необхідність етичного підходу до використання штучного інтелекту в освітній сфері. Важливо не тільки забезпечити захист даних, але й використовувати ШІ таким чином, щоб рішення, прийняті на основі цих технологій, були справедливими, прозорими і відкритими для перевірки. Освітні установи повинні розробити чіткі правила щодо того, як і коли використовувати ШІ для прийняття рішень, щоб уникнути можливих зловживань та помилок [5].

Отже, хоча ШІ може бути корисним інструментом для освітніх лідерів, необхідно враховувати потенційні ризики та забезпечити етичне використання технологій для прийняття рішень у сфері освіти.

Дослідження Комлевої [6] стосується впровадження систем підтримки рішень у сфері управління якістю навчального процесу. Автори зазначають, що системи підтримки рішень, засновані на штучному інтелекті, можуть значно покращити процес моніторингу та оцінки якості освіти, оскільки дозволяють автоматизувати багато процесів, які раніше виконувалися вручну [6].

Завдяки використанню таких систем освітні заклади можуть більш ефективно відслідковувати прогрес студентів, оцінювати ефективність навчальних програм та вчасно виявляти проблеми, що виникають у процесі навчання. ШІ допомагає аналізувати великі обсяги даних і на основі цього надавати рекомендації для вдосконалення навчальних програм або поліпшення педагогічних підходів.

Комлева та ін. звертають увагу на те, що впровадження таких систем дозволяє суттєво зменшити адміністративне навантаження на викладачів та адміністрацію навчальних закладів, оскільки частина завдань з управління навчальним процесом автоматизується. Це звільняє час для викладачів, який вони можуть використовувати для більш творчої та інноваційної роботи зі студентами [6].

Також важливо відзначити, що системи підтримки рішень дозволяють краще враховувати індивідуальні потреби студентів. ШІ аналізує дані про кожного студента та пропонує оптимальні варіанти навчальних планів, завдань або додаткових матеріалів, що дозволяє забезпечити більш персоналізований підхід до навчання.

Отже, системи підтримки рішень є важливим інструментом для підвищення якості освіти, оскільки вони дозволяють автоматизувати рутинні процеси, забезпечити індивідуальний підхід до навчання та поліпшити моніторинг і оцінку результатів освітнього процесу.

У статті Саутворта [7] йдеться про необхідність розвитку штучного інтелекту в рамках освітніх програм та впровадження його у всі предмети навчання. Автори досліджують, як інтеграція ШІ у різні галузі знань може

змінити освітній ландшафт і підготувати студентів до викликів сучасного ринку праці [7].

Саутворт та його колеги зазначають, що сучасні освітні програми мають адаптуватися до швидких змін у технологіях і забезпечувати студентів навичками, необхідними для роботи з ШІ.

У статті розглядається необхідність створення інтегрованих освітніх програм, які охоплюють базові та розширені аспекти роботи зі штучним інтелектом, починаючи від програмування та алгоритмів, до етичних і правових питань використання цих технологій у суспільстві. Автори підкреслюють, що такі програми повинні бути частиною не лише технічних спеціальностей, але й гуманітарних наук, що дозволить студентам будь-яких спеціальностей краще розуміти можливості та виклики, які несе ШІ [7].

Одним із ключових аспектів інтеграції штучного інтелекту в навчальні програми є розвиток у студентів так званої "ШІ-грамотності" — навички ефективного використання інструментів ШІ для вирішення практичних завдань та розуміння принципів роботи алгоритмів. Це дозволить майбутнім спеціалістам краще адаптуватися до нових технологічних реалій і бути конкурентоспроможними на ринку праці.

Дослідження також наголошує на важливості міждисциплінарного підходу до викладання ШІ. Студенти мають отримувати знання не лише з технічних аспектів, а й з філософських, соціальних та економічних питань, пов'язаних з використанням ШІ. Це дозволить майбутнім фахівцям не лише створювати технології, але й аналізувати їх вплив на суспільство та враховувати етичні аспекти у своїй діяльності [7].

Отже, розвиток ШІ-грамотності у рамках навчальних програм є важливим кроком для підготовки студентів до майбутнього, де штучний інтелект відіграватиме центральну роль у багатьох професійних сферах. Важливо не лише забезпечити студентам технічні знання, але й розвинути їх критичне мислення щодо впровадження нових технологій у суспільстві.

У дослідженні Каролус [8] пропонується концептуальна модель цифрової взаємодії з штучним інтелектом, яка підкреслює важливість розвитку компетенцій для грамотного використання голосових систем ШІ. Автори досліджують різні аспекти взаємодії людини з системами на основі штучного інтелекту, зокрема з голосовими помічниками, які стають все більш поширеними у повсякденному житті [8].

Модель, розроблена авторами, включає кілька ключових компетенцій, які необхідно розвивати для ефективної взаємодії з голосовими системами штучного інтелекту. До них відносяться: технічна грамотність, критичне мислення щодо використання даних, знання про принципи роботи ШІ та навички ефективної комунікації з цифровими помічниками. Ці компетенції стають особливо важливими, оскільки голосові системи дедалі частіше використовуються не лише для простих завдань, таких як пошук інформації, але й для більш складних операцій, таких як управління пристроями або виконання бізнес-завдань [8].

Автори наголошують, що для успішної взаємодії з голосовими системами ШІ користувачі повинні розуміти обмеження та можливості цих систем, а також вміти критично оцінювати інформацію, яку вони надають. Це означає, що голосові помічники можуть бути корисними інструментами лише в тому випадку, якщо користувачі володіють необхідними знаннями для їх грамотного використання. Важливо також розвивати навички зворотного зв'язку з системами ШІ, оскільки голосові системи навчаються на основі взаємодії з користувачами [8].

Таким чином, модель цифрової взаємодії з голосовими системами ШІ, запропонована авторами, вказує на необхідність розвитку нових компетенцій для ефективної роботи з цими технологіями, що є важливим аспектом цифрової грамотності у сучасному світі.

У дослідженні Брессане [9] розглянуто роль штучного інтелекту у підтримці академічних досягнень студентів, зокрема у контексті розробки навчальних стратегій та підтримки студентів з навчальними труднощами.

Автори зосереджуються на тому, як інструменти штучного інтелекту можуть бути використані для індивідуалізації навчання та забезпечення спеціалізованої підтримки для тих, хто має труднощі з засвоєнням матеріалу [9].

Одним з ключових аспектів дослідження є аналіз того, як ШІ може допомогти в ідентифікації студентів з навчальними труднощами на ранніх етапах та запропонувати індивідуальні шляхи вирішення цих проблем. Наприклад, адаптивні навчальні платформи можуть використовувати дані про прогрес студентів, щоб визначити ті області, де вони потребують додаткової допомоги, та запропонувати відповідні навчальні матеріали або завдання. Це дозволяє не лише покращити академічні результати, але й підвищити мотивацію студентів до навчання, оскільки вони отримують персоналізовані рекомендації [9].

Дослідження також підкреслює важливість використання ШІ для підтримки студентів з особливими освітніми потребами. Інструменти ШІ можуть допомогти адаптувати навчальні матеріали для таких студентів, враховуючи їхні індивідуальні потреби. Це може включати надання додаткових пояснень, зміну формату завдань або навіть використання альтернативних методів навчання, таких як візуалізація або аудіо-супровід [9].

Загалом, ШІ має потенціал стати потужним інструментом для підтримки студентів з навчальними труднощами, забезпечуючи їм персоналізоване навчання та допомагаючи досягти кращих результатів у навчанні.

У дослідженні Ванга [10] розглянуто питання готовності викладачів до впровадження нових технологій, зокрема штучного інтелекту, у навчальний процес. Автори підкреслюють, що ефективне використання ШІ в освітньому процесі залежить не лише від технічних можливостей, але й від рівня підготовки викладачів до роботи з цими технологіями [10].

Дослідження показало, що більшість викладачів відчувають потребу в додатковій підготовці для ефективного використання ШІ у своїй роботі.

Автори підкреслюють важливість розвитку професійних навичок викладачів у сфері штучного інтелекту, включаючи базові знання про принципи

роботи алгоритмів та можливості ШІ для персоналізації навчання. Особливо наголошується на необхідності інтеграції технологічних інновацій у педагогічні практики для забезпечення кращих навчальних результатів студентів [10].

Згідно з дослідженням, готовність викладачів до впровадження штучного інтелекту в навчальний процес визначається кількома ключовими чинниками. Перш за все, це рівень їхньої технічної грамотності та обізнаності щодо можливостей ШІ. Викладачі, які мають знання про технології ШІ, більш схильні до їх використання в освітньому процесі. Крім того, важливу роль відіграє доступ до навчальних ресурсів та курсів підвищення кваліфікації, які можуть допомогти викладачам оволодіти необхідними навичками для ефективної роботи з новими технологіями [10].

Автори також відзначають, що одним із викликів для впровадження ШІ в освіту є страх викладачів перед змінами та їхнє занепокоєння щодо можливої автоматизації навчального процесу. Однак дослідження показує, що використання ШІ не замінює роль викладача, а лише доповнює його, дозволяючи краще організовувати навчальний процес і забезпечувати індивідуалізований підхід до кожного студента [10].

Таким чином, для успішної інтеграції штучного інтелекту в освіту необхідно підготувати викладачів, забезпечивши їм доступ до навчальних програм, спрямованих на розвиток технічної грамотності та навичок використання ШІ у викладанні. Це дозволить створити більш ефективну та персоналізовану систему освіти, орієнтовану на потреби студентів.

Тамікі [11] у своїй роботі акцентують увагу на розвитку освітніх програм, спрямованих на підготовку системних розробників та бізнес-продюсерів у майбутньому. Дослідження зосереджується на проекті "Future Strategy Design in Action", який проводився у рамках співпраці між університетами та промисловими підприємствами. Метою проекту було розроблення освітніх програм, що враховують потреби сучасної індустрії у кваліфікованих спеціалістах [11].

Автори зазначають, що підготовка системних розробників та бізнес-продюсерів вимагає особливого підходу, який поєднує технічні знання з навичками стратегічного мислення та управління. Освітні програми повинні включати як базові курси з програмування та системного аналізу, так і розширені дисципліни, спрямовані на розвиток навичок бізнес-аналізу, управління проектами та інноваційного мислення [11].

Проект "Future Strategy Design in Action" також акцентує увагу на важливості практичного навчання, яке передбачає участь студентів у реальних проектах у співпраці з компаніями. Це дозволяє майбутнім спеціалістам отримати практичний досвід роботи у своїй галузі, що є важливим чинником для успішного працевлаштування після закінчення навчання [11].

Автори підкреслюють, що співпраця між університетами та промисловими підприємствами є ключовим елементом успішного впровадження освітніх програм, оскільки вона дозволяє враховувати актуальні потреби ринку праці та забезпечувати студентам можливість отримати необхідні навички для подальшого розвитку у професійній сфері [11].

Таким чином, розробка освітніх програм для майбутніх системних розробників та бізнес-продюсерів повинна враховувати як технічні, так і управлінські аспекти, що дозволить підготувати спеціалістів, здатних ефективно працювати у швидко змінюваних умовах сучасної економіки.

У статті Кобца [12] розглядаються прогалини між компетенціями, які отримують студенти у навчальних закладах, та вимогами сучасного ринку праці до бізнес-аналітиків. Автори відзначають, що стандарти освіти у цій галузі не завжди відповідають потребам роботодавців, що може створювати труднощі для випускників при працевлаштуванні [12].

Дослідження показує, що сучасний ринок праці вимагає від бізнес-аналітиків не лише знання у сфері економіки та фінансів, але й навички роботи з великими обсягами даних, аналізу інформації за допомогою інструментів штучного інтелекту та вміння розробляти стратегії для покращення бізнес-процесів. Автори наголошують, що вищі навчальні заклади часто не

забезпечують студентів достатньо глибокими знаннями у цих галузях, що створює розрив між освітою та вимогами ринку праці [12].

Зокрема, дослідження вказує на необхідність перегляду освітніх програм з бізнес-аналітики для включення до них курсів з аналізу даних, роботи з інструментами ШІ, а також розвитку навичок критичного мислення та управління проектами. Автори пропонують створювати партнерські програми з бізнесом для розробки навчальних матеріалів, які відповідають сучасним потребам ринку [12].

Отже, для того щоб підготувати конкурентоспроможних бізнес-аналітиків, необхідно переглянути стандарти освіти у цій сфері, забезпечивши студентам доступ до сучасних інструментів аналізу даних та штучного інтелекту, а також можливість отримати практичний досвід під час навчання.

У дослідженні Хаббала [13] розглянуто новий підхід до управління ризиками, довірою та безпекою у сфері штучного інтелекту — AI TRiSM (Artificial Intelligence Trust, Risk, and Security Management). Автори пропонують концептуальну модель, яка дозволяє впроваджувати штучний інтелект, враховуючи виклики, пов'язані з безпекою та довірою користувачів до цих технологій [13].

AI TRiSM включає кілька ключових елементів, які дозволяють забезпечити безпечне та етичне впровадження ШІ у різні сфери діяльності.

Модель передбачає виявлення ризиків, пов’язаних із застосуванням штучного інтелекту, та розробку стратегій їх мінімізації. Це включає оцінку впливу технологій на безпеку даних, конфіденційність користувачів та загальний соціальний вплив ШІ. Автори наголошують, що недостатня довіра до технологій ШІ може стримувати їхнє впровадження, тому важливо розробляти механізми для підвищення рівня довіри користувачів [13].

Одним із ключових аспектів AI TRiSM є забезпечення прозорості алгоритмів, які використовуються в ШІ. Це передбачає розробку стандартів для документування та пояснення рішень, які приймаються алгоритмами, щоб користувачі могли зрозуміти, як та чому були ухвалені певні рішення.

Важливими також є етичні норми, які мають бути впроваджені в практику для забезпечення справедливості, ненупередженості та етичного використання штучного інтелекту [13].

Крім того, автори акцентують увагу на важливості навчання всіх учасників процесу – від розробників до кінцевих користувачів – про ризики та виклики, пов'язані зі штучним інтелектом. Це дозволяє створити більш свідоме та обізнане суспільство, яке здатне відповідально користуватися новими технологіями [13].

Таким чином, впровадження концепції AI TRiSM є важливим кроком до безпечного і етичного використання штучного інтелекту, яке може допомогти підвищити довіру до цих технологій серед користувачів і забезпечити їхнє успішне інтегрування в різні сфери діяльності.

У статті Кравцова та Кобца [14] обговорюється процес впровадження інновацій у навчальні програми інформаційно-комунікаційних технологій (ІКТ) відповідно до вимог різних зацікавлених сторін. Автори зазначають, що інтеграція нових технологій у навчальний процес є складним завданням, яке вимагає врахування інтересів студентів, викладачів, роботодавців та інших учасників освітнього процесу [14].

Дослідження вказує на те, що для успішного впровадження інноваційних підходів у навчання необхідно провести детальний аналіз потреб усіх учасників. Це може включати опитування, фокус-групи та інші методи збору інформації, щоб виявити, які знання та навички є найбільш важливими для ринку праці. Автори стверджують, що тільки шляхом тісної співпраці з бізнесом та індустрією можна створити навчальні програми, які відповідають сучасним вимогам [14].

Крім того, важливим аспектом є забезпечення підготовки викладачів до роботи з новими технологіями. Автори наголошують на необхідності професійного розвитку викладачів, які повинні мати змогу постійно оновлювати свої знання у сфері ІКТ та використовувати нові методи навчання для залучення студентів [14].

Таким чином, успішне впровадження інновацій у навчальні програми ІКТ вимагає системного підходу, який включає активну участь усіх зацікавлених сторін, постійний моніторинг змін на ринку праці та адаптацію освітніх програм відповідно до нових викликів і можливостей, що виникають у технологічному середовищі.

У статті Кобца та Осипової [16] розглядаються фактори, які впливають на забезпечення якості освіти у вищих навчальних закладах, зокрема у контексті академічної успішності студентів. Автори акцентують увагу на важливості створення сприятливого навчального середовища, яке включає не лише академічні, але й соціальні та емоційні аспекти [16].

Дослідження показує, що академічна успішність студентів значно залежить від їхнього рівня мотивації, здатності до самоорганізації та підтримки з боку викладачів та однокурсників. Важливим є також доступ до навчальних ресурсів та інструментів, які допомагають студентам у процесі навчання. Автори підкреслюють, що впровадження нових технологій, таких як штучний інтелект, може значно полегшити навчальний процес і допомогти студентам у досягненні кращих результатів [16].

Крім того, важливим аспектом є врахування індивідуальних особливостей студентів, що дозволяє адаптувати навчальні програми до їхніх потреб. Використання інструментів ШІ для аналізу даних про навчання може допомогти викладачам виявляти тих студентів, які потребують додаткової підтримки, і розробляти для них індивідуальні плани навчання [16].

Отже, забезпечення якості освіти є комплексним завданням, яке потребує інтеграції академічних, соціальних та технологічних чинників. Лише спільними зусиллями усіх учасників освітнього процесу можна створити умови для досягнення високих результатів у навчанні та розвитку студентів.

РОЗДІЛ 2
МЕТОДОЛОГІЯ МАШИННОГО НАВЧАННЯ

2.1 Машинне навчання

Машинне навчання (ML) — це напрямок в області штучного інтелекту (ШІ), який дає комп'ютерам здатність «навчатися» на даних і поліпшуватися в задачах без явного програмування. Основна ідея полягає в тому, що замість написання конкретних алгоритмів для виконання завдань, ми забезпечуємо систему великою кількістю даних та загальними правилами для аналізу і обробки, після чого вона самостійно визначає найбільш ефективні способи вирішення завдань.

Основні концепції машинного навчання:

1. **Алгоритм** — це послідовність математичних і статистичних операцій, яка дозволяє системі аналізувати дані, розпізнавати закономірності і створювати прогнозну модель. Алгоритми — це «серце» машинного навчання, саме вони виконують обчислення та узагальнюють дані в моделі.

2. **Модель** — це результат навчання алгоритму на певних даних, який зберігає узагальнені знання для використання в майбутньому. Модель є ключовим компонентом, що дозволяє передбачати, класифікувати або приймати рішення на основі нових даних.

3. **Навчальні дані (Тренувальний набір)** — це дані, на яких навчається алгоритм, збагачуючи модель знаннями для подальших завдань. Тренувальні дані забезпечують базу для розпізнавання закономірностей, що використовуються для прийняття рішень або прогнозування.

4. **Навчання** — це процес, під час якого модель обробляє навчальні дані, щоб розпізнати зв'язки та залежності між вхідними даними (ознаками) та результатом. Зазвичай під час навчання модель коригує свої параметри на основі зворотного зв’язку або оцінки помилок, з метою підвищення точності прогнозів або класифікацій.

5. **Оцінка якості** — це процес вимірювання точності та надійності моделі за допомогою тестових даних або окремих метрик, таких як точність,

корінь середньоквадратичної помилки (RMSE) тощо. Оцінка якості є важливим етапом, оскільки вона дозволяє зрозуміти, наскільки добре модель справляється із завданням, для якого вона була створена.

2.1.1 Методи машинного навчання з вчителем

Методи навчання з вчителем орієнтовані на роботу з позначеними даними, тобто такими, де для кожного прикладу відомі і вхідні значення (ознаки), і правильний результат або мітка. Ця мітка діє як «вчитель», який направляє модель, дозволяючи їй зрозуміти зв’язок між ознаками та результатом. Модель навчається на основі цих даних, спочатку аналізуючи приклади і формуючи прогноз, який потім порівнюється з фактичним результатом. Якщо прогноз і мітка не збігаються, модель коригує свої параметри, поступово підвищуючи точність. Завдяки такому підходу, модель не просто «запам'ятовує» дані, а й вчиться узагальнювати закономірності, що дозволяє робити більш точні передбачення на нових даних.

Лінійна регресія

Лінійна регресія є одним із найпростіших і водночас потужних методів для прогнозування числових значень та аналізу залежностей між змінними. Суть лінійної регресії полягає в тому, щоб знайти лінію, яка найкраще описує залежність між двома змінними, де одна є незалежною (ознакою), а інша — залежною (результатом). Формула для лінійної регресії виглядає як $y=mx+by = mx + by=mx+b$, де mmm визначає нахил лінії, що показує, як зміна ознаки впливає на результат, а b — це перетин лінії з віссю y. Для знаходження параметрів mmm і b застосовується метод найменших квадратів, що дозволяє мінімізувати суму квадратів відхилень між передбаченими та фактичними значеннями. Наприклад, лінійна регресія може передбачити ціну нерухомості, якщо є дані про площу, кількість кімнат та інші параметри житла.

Метод k-найближчих сусідів (KNN)

Метод k-найближчих сусідів (KNN) є простим та інтуїтивним методом, який використовується як для класифікації, так і для регресії. Його суть полягає

в тому, що для нового зразка визначається k найближчих сусідів у тренувальних даних, а новий зразок отримує клас або значення на основі цих сусідів. Вибір значення k є ключовим: маленьке значення k може зробити модель надмірно чутливою до шуму (перенавчання), а велике — може знизити її адаптивність до дрібних змін. Наприклад, якщо банк хоче визначити категорію клієнта (наприклад, потенційний позичальник або вкладник), KNN може порівняти його з іншими клієнтами зі схожими характеристиками і на основі класів найближчих сусідів визначити його категорію.

Випадковий ліс (Random Forest)

Випадковий ліс є ансамблевим методом, що комбінує кілька дерев рішень для покращення точності й надійності передбачень. Кожне дерево тренується на випадковій підмножині даних і використовує випадковий набір ознак, що дозволяє моделі виявляти різні аспекти даних і уникати перенавчання. У випадку класифікації кожне дерево «голосує» за певний клас, а результатом є клас із найбільшою кількістю голосів. Для задач регресії обчислюється середнє значення передбачень усіх дерев. Наприклад, випадковий ліс може бути використаний для прогнозування кредитоспроможності клієнтів банку, об'єднуючи результати різних дерев для точнішого результату.

Дерева рішень

Дерева рішень — це ієрархічні моделі, що працюють за принципом набору правил «якщо-то». Кожен вузол дерева представляє певну ознаку, а його гілки — можливі значення цієї ознаки. Залежно від умов на кожному вузлі, процес переходить по гілках дерева, доки не досягне листа, який і визначає кінцевий результат. Такий підхід є простим і зрозумілим для інтерпретації, тому дерева рішень часто використовуються в задачах, де важливо пояснити процес ухвалення рішень. Наприклад, інтернет-магазин може класифікувати своїх клієнтів на основі параметрів, таких як кількість покупок і візитів на сайт, прогнозуючи ймовірність повторної покупки.

Метод опорних векторів (SVM)

Метод опорних векторів (SVM) використовується як для задач класифікації, так і для регресії. Його основна ідея полягає у пошуку оптимальної гіперплощини, яка ділить простір ознак таким чином, щоб відстань між найближчими точками обох класів (опорними векторами) та цією гіперплощиною була максимальною. Чим більша ця відстань, тим надійніше відбувається класифікація. Якщо дані не можна поділити лінійно, застосовуються ядрові методи, такі як ядро Радіальної базисної функції (RBF), що перетворюють дані у простір вищої розмірності, де їх вже можна розділити лінійно. Наприклад, у задачах класифікації електронної пошти SVM може ефективно відокремити спам від легітимних повідомлень, формуючи чітку межу між класами.

2.1.2 Методи машинного навчання без вчителя

Методи машинного навчання без вчителя використовуються для роботи з неструктурованими даними, де немає позначок або правильних відповідей, що дозволяє моделі самостійно аналізувати та структурувати інформацію. Основна мета таких методів — виявити приховані шаблони, зв’язки або групи, щоб створити представлення даних, що дозволяє системі працювати з великим обсягом інформації. Методи без учителя ефективні в задачах, де потрібно виявити структуру в даних без попереднього знання їхньої організації.

Алгоритм K-середніх (K-Means)

Алгоритм K-середніх (K-Means) — це один із найпоширеніших методів кластеризації. Його мета полягає у поділі даних на KKK кластерів, кількість яких задається заздалегідь. Алгоритм обчислює центр кожного кластера, до якого потім призначаються точки, найближчі до цього центру. Коли точки перестають переходити між кластерами, процес завершується. K-середніх ефективний для даних, що мають чітко розділені кластери з однаковою кількістю точок, але менш ефективний для складних або нерівномірних кластерів.

Аналіз головних компонент (PCA)

Аналіз головних компонент (PCA) — це метод зниження розмірності, який дозволяє зберегти максимум інформації, зменшивши кількість ознак. PCA обчислює нові координати (головні компоненти) на основі варіацій у даних, а перші кілька компонент зберігають найбільшу варіативність, дозволяючи спростити дані без значних втрат. PCA часто застосовується для аналізу зображень, звуку та тексту.

Гаусові змішані моделі (GMM)

Гаусові змішані моделі (GMM) є розширенням кластеризації, які припускають, що дані можуть бути розподілені за кількома нормальними розподілами. Модель оцінює ймовірності належності кожної точки до кожного з кластерів, що дозволяє створювати більш гнучкі ймовірнісні структури кластерів. GMM підходить для даних, що перетинаються між кластерами або мають складну форму.

Ієрархічна кластеризація

Ієрархічна кластеризація будує дерево або дендрограму, що представляє багаторівневу структуру кластерів. На початковому етапі кожен об'єкт є окремим кластером, потім об’єкти об’єднуються на основі подібності. Ієрархічна кластеризація не потребує заздалегідь визначення кількості кластерів, що робить її гнучкішою для задач, де структура даних невідома. Наприклад, компанія може використовувати ієрархічну кластеризацію для сегментації клієнтів за звичками покупок, щоб виділити різні типи клієнтів.

Метод Apriori

Метод Apriori застосовується для пошуку частих комбінацій об'єктів у даних. Він використовується для аналізу моделей купівлі в ритейлі, наприклад, для виявлення асоціацій, коли покупка одного продукту підвищує ймовірність придбання іншого.

2.1.3 Оцінка якості методів машинного навчання

Якість методів машинного навчання оцінюється за допомогою різних метрик залежно від типу задачі (регресія чи класифікація), що дозволяє зрозуміти точність передбачень або класифікацій.

Метрики для задач регресії

Основні метрики для регресії — це корінь середньоквадратичної помилки (RMSE) та коефіцієнт детермінації (R^2). RMSE вимірює середню різницю між прогнозом і фактичним значенням, а R^2 показує, наскільки добре модель описує варіативність. Це дозволяє зрозуміти, наскільки точно модель прогнозує значення, враховуючи всі зміни в даних.

Метрики для задач класифікації

Класифікаційні задачі оцінюються через матрицю неточностей, точність, повноту, точність (Precision) і F1-міру. Матриця неточностей дозволяє оцінити успішність класифікації, враховуючи справжні та помилкові прогнози, а F1-міра балансує між точністю і повнотою, забезпечуючи комплексну оцінку.

Перехресна валідація

Перехресна валідація — це метод розбиття даних на кілька підвибірок для об’єктивної перевірки моделі. Це дозволяє краще оцінити модель і уникнути надмірної залежності від особливостей одного набору даних.

2.2 Бізнес аналітика та технічні аспекти впровадження автоматизованого машинного навчання в бізнес аналітику

Бізнес-аналітика (BA) — це дисципліна, що спрямована на збір, обробку і аналіз даних з метою підтримки прийняття обґрунтованих бізнес-рішень. Основна мета бізнес-аналітики полягає у виявленні закономірностей, трендів і причинно-наслідкових зв’язків у великих масивах даних, які допомагають у прогнозуванні майбутніх подій, покращенні ефективності та оптимізації ресурсів. Сучасна бізнес-аналітика охоплює кілька ключових напрямів, включаючи описову аналітику (опис поточного стану речей), діагностичну

аналітику (аналіз причин подій), прогнозну аналітику (передбачення майбутніх результатів) та прескриптивну аналітику (рекомендації щодо подальших дій).

Бізнес-аналітика використовує численні методи та інструменти, зокрема статистику, моделювання, обробку великих даних (Big Data) та машинне навчання. Впровадження автоматизованого ML стає важливим інструментом для розширення можливостей бізнес-аналітики, дозволяючи отримувати більш точні прогнози та автоматизувати процеси аналізу.

2.2.1 Автоматизоване машинне навчання в бізнес-аналітиці

Автоматизоване машинне навчання (AutoML) — це технологія, що автоматизує основні етапи побудови ML-моделей: підготовку даних, вибір алгоритму, тренування моделі, оптимізацію гіперпараметрів, оцінку якості та розгортання. Використання AutoML дає змогу прискорити процес створення моделей і робить технології машинного навчання доступними для ширшого кола спеціалістів, навіть без глибоких знань у галузі ML.

Для ефективного впровадження AutoML у бізнес-аналітику важливим є створення належної інфраструктури для збору, зберігання та обробки великих обсягів даних. Хмарні платформи, такі як Amazon Web Services (AWS), Google Cloud Platform (GCP) і Microsoft Azure, надають необхідні засоби для підтримки AutoML. Ці платформи пропонують інтегровані рішення для збору даних, створення моделей та розгортання їх в продуктивному середовищі. Наприклад, сервіс Amazon SageMaker забезпечує повний цикл розробки AutoML, включаючи засоби для автоматизованого тренування, оцінки та масштабування моделей (рис. 2.1).

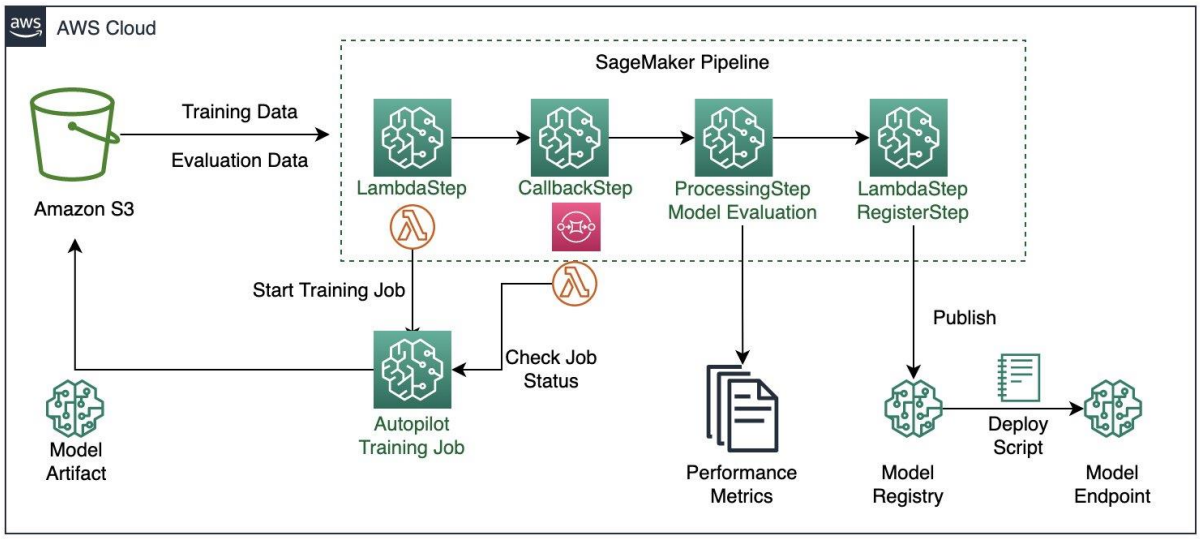

Рис. 2.1. Архітектура використання Amazon SageMaker для автоматизації ML-процесів

2.2.2 Технічні аспекти впровадження автоматизованого машинного навчання.

Впровадження AutoML у бізнес-аналітику включає кілька основних технічних етапів, кожен з яких має свої особливості та вимоги до інфраструктури.

Підготовка та очищення даних. Перший етап полягає у підготовці даних, включаючи їх очищення, нормалізацію та обробку. Якість даних безпосередньо впливає на точність моделей, тому їх підготовка є ключовим кроком для успішного функціонування автоматизованих моделей. Наприклад, дані, що містять аномалії, пропущені значення або помилки, можуть призвести до неточних прогнозів та підвищити ризик перенавчання. Один з методів нормалізації даних, що часто використовується у бізнес-аналітиці, є масштабування значень у межах від 0 до 1 за допомогою формули(1):

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (2.1)$$

де x — значення ознаки, min(x)— мінімальне значення ознаки, max(x) — максимальне значення ознаки. Цей процес забезпечує стабільність роботи алгоритму та дозволяє уникнути домінування одних ознак над іншими.

Вибір та налаштування моделей. На наступному етапі вибирається алгоритм машинного навчання, який найбільше підходить для задачі.

Автоматизовані ML-платформи пропонують широкий вибір алгоритмів для різних типів задач, таких як лінійна регресія для прогнозування числових значень або метод опорних векторів для задач класифікації. Наприклад, лінійна регресія дозволяє моделювати залежності між змінними за допомогою лінійного рівняння:

$$y = mx + b \qquad (2.2)$$

де y — залежна змінна, x — незалежна змінна, m — коефіцієнт нахилу (градієнт), а b — вільний член. У бізнес-аналітиці лінійна регресія широко використовується для прогнозування витрат, доходів, цін або інших важливих показників на основі історичних даних.

Тренування моделі та оптимізація гіперпараметрів. Під час тренування модель навчається на основі тренувальних даних, а також коригує свої параметри для підвищення точності прогнозів. Паралельно з тренуванням відбувається автоматична оптимізація гіперпараметрів, що є важливим процесом для налаштування моделі. Оптимізація гіперпараметрів дозволяє знайти найкращі значення параметрів, які впливають на продуктивність моделі. Багато AutoML-платформ використовують методи пошуку, такі як випадковий пошук або байєсівський оптимізаційний підхід, для визначення оптимальних значень гіперпараметрів.

Оцінка та перевірка моделі. Оцінка якості моделі є важливим етапом, що дозволяє зрозуміти, наскільки точно модель справляється з поставленим завданням. Для задач регресії часто використовується метрика кореня середньоквадратичної помилки (RMSE), яка вимірює середнє відхилення між фактичними та прогнозованими значеннями. Формула для обчислення RMSE має вигляд:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2.3)$$

де n — кількість спостережень, yi — фактичне значення, yi^ — прогнозоване значення. Менше значення RMSE свідчить про вищу точність моделі. Для задач класифікації використовуються інші метрики, такі як точність (Accuracy), повнота (Recall) та F1-міра, що дозволяють оцінити здатність моделі правильно класифікувати об’єкти.

Розгортання моделі та інтеграція з бізнес-процесами. Після тренування та оцінки модель інтегрується в бізнес-процеси, де вона може безперервно використовуватись для аналізу поточних даних і підтримки прийняття рішень. Сучасні AutoML-платформи забезпечують можливості розгортання моделей у хмарі, на локальних серверах або як сервіси з доступом через API. Це дозволяє інтегрувати ML-модель в інші інформаційні системи компанії, такі як ERP чи CRM, і забезпечувати доступ до актуальних прогнозів у реальному часі.

Моніторинг та оновлення моделі. Моніторинг продуктивності моделі є обов’язковим етапом, оскільки в умовах реального світу модель може втрачати свою актуальність через зміни у даних. Автоматизовані платформи для машинного навчання, такі як DataRobot та H2O.ai, надають можливості для постійного моніторингу показників ефективності моделей та їх оновлення в разі погіршення точності. На рис. 2.2 зображена архітектура моніторингу моделі, що включає процес перевірки якості, автоматичне оновлення параметрів та підтримку історії змін для аналізу ефективності моделі.

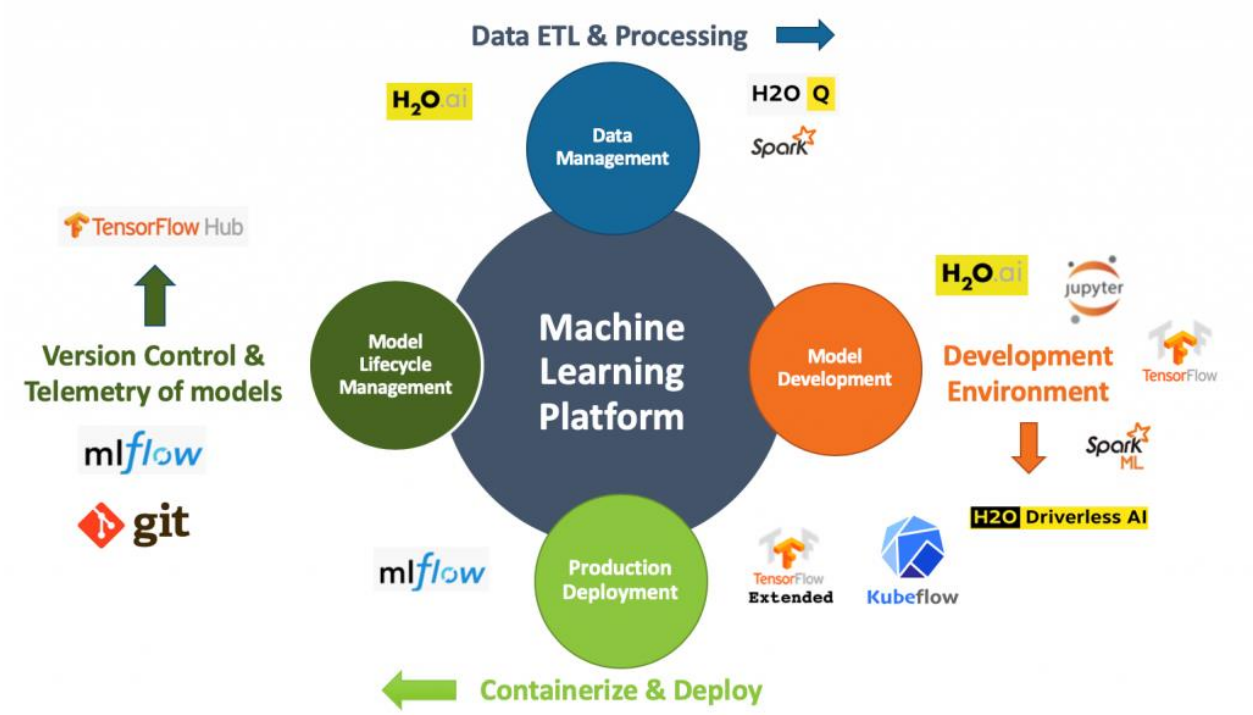

Рис. 2.2 - Архітектура моніторингу моделей машинного навчання

Виклики впровадження автоматизованого ML в бізнес-аналітику. Основні виклики впровадження AutoML включають проблеми якості та обсягу даних, потребу в інтеграції з наявними системами, необхідність кваліфікованого персоналу та відповідність вимогам щодо захисту персональних даних. Недостатня кількість якісних даних може знизити ефективність моделі, а недотримання стандартів конфіденційності може призвести до юридичних проблем. Крім того, для ефективного впровадження та підтримки AutoML необхідні фахівці, зокрема інженери даних та аналітики.

Переваги автоматизації машинного навчання в бізнес-аналітиці. Автоматизоване ML дозволяє бізнесу швидко адаптуватися до ринкових змін, знижує витрати на розробку моделей, усуває людські помилки, масштабовано забезпечує аналітичні потреби та підвищує точність прогнозів і обґрунтованість рішень.

РОЗДІЛ 3
ПРИЙНЯТТЯ РІШЕНЬ НА ОСНОВІ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ ЦІЛЬОВОЇ АУДИТОРІЇ ЗАКЛАДІВ ВИЩОЇ ОСВІТИ З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ

3.1 Обґрунтування вибору методології проекту

Згідно з очікуваннями, до 2030 року штучний інтелект автоматизує 40% завдань вчителів початкової школи. Для забезпечення якісного навчання будь-яка система штучного інтелекту залежить від доступу до великого обсягу даних. Чим більше даних подається в ШІ, тим точнішими стають ці системи. Даних повинно бути багато і вони мають бути різноманітними, що надає велику перевагу великим вищим навчальним закладам.

Зі зростанням поширення штучного інтелекту на робочих місцях ті, хто розуміє та взаємодіє з ним, матимуть чітку перевагу над тими, хто має менше розвинені навички роботи з ШІ. Лише 33% опитаних користувачів ІТ-послуг заявили, що використовували ШІ для виконання конкретного завдання, хоча 77% пристроїв, що використовуються, вже мають певну функціональність ШІ. Найпоширеніші з таких завдань: 1) покупки, 2) пошук інформації, та 3) дослідження.

Метою вищого навчального закладу є визначення цільової аудиторії вступників для ефективного розподілу фінансових, людських і часових ресурсів для проведення профорієнтаційної роботи та маркетингових заходів з метою залучення абітурієнтів до академічних програм. Одним з інструментів для досягнення цієї мети є вивчення характеристик абітурієнтів за попередні роки навчання за допомогою методів машинного навчання із використанням алгоритмів класифікації. Щоб використовувати цей інструмент, керівники структурних підрозділів ВНЗ повинні мати навички його застосування та інтерпретації результатів. Згідно з останнім опитуванням, 91% бізнес-лідерів наймають працівників з досвідом роботи з ChatGPT для 1) розробки програмного забезпечення, 2) обслуговування клієнтів, 3) відділу кадрів (HR) та

4) відділів маркетингу. Водночас у сфері освіти навички отримання інсайтів з наявної інформації про абітурієнтів обмежуються розрахунком описової статистики та динаміки вступу. Було б недоцільно для ВНЗ ігнорувати такі інструменти, оскільки прийняття рішень на основі даних є конкурентною перевагою, яка дозволяє зосередитися на тих абітурієнтах, які мають ймовірність вступу понад 50%.

Метою цього дослідження є розробка моделей на основі машинного навчання, які можуть точно прогнозувати ймовірність вступу абітурієнтів до вищих навчальних закладів за допомогою прийняття рішень на основі даних.

Структура статті організована наступним чином. Короткий огляд пов'язаних робіт представлено у розділі 2. Методологія прогнозування вступу абітурієнтів описана у розділі 3. Результати обговорено у розділі 4. Нарешті, висновки надано у розділі 4.4.6.

3.2 Схожі проекти

Передумовами для впровадження інструментів ШІ/МН у вищих навчальних закладах [5] є: 1) сформована команда з культурою використання ШІ; 2) розроблені стратегії навчання ШІ; 3) встановлена співпраця з постачальниками послуг ШІ та програмного забезпечення; 4) оцифровані дані вищих навчальних закладів для обробки ШІ; 5) наявна інфраструктура даних для ШІ. Вимоги до даних для якісної обробки включають: 1) доступність та якість даних (текст, зображення, відео, аудіо, хештеги в соціальних мережах, пости, коментарі, вподобання та ретвіти); 2) підтримка керівництва та залучення зацікавлених сторін (відкриті дані, людські ресурси, бюджет та фінансова інформація) [6]; 3) знання та навички окремих осіб і команд у роботі з даними.

Грамотність у сфері ШІ [7] — це здатність розуміти (навчальна та тестова вибірка), використовувати (інструменти ШІ для виконання завдань), оцінювати (якість і надійність ШІ) та етично керувати ШІ (моральні та етичні

наслідки ШІ, справедливість, прозорість, відповідальність перед суспільством і окремими людьми).

Грамотність у сфері ШІ [8] може бути впроваджена в рамках програм перекваліфікації та підвищення кваліфікації керівного складу вищих навчальних закладів із навчальними результатами програм (НРП), як це показано в таблиці 3.1, де вміст ШІ — це частка навчальних матеріалів про ШІ у відповідній категорії.

Таблиця 3.1

Класифікація систем підтримки прийняття рішень

Категорії грамотності ШІ	**Опис**	**Вміст ШІ**	**Результати навчання**
Впровадження ШІ	Підтримка ШІ відповідними знаннями та (програмування, статистика)	10–49%	PLO1. Визначити, описати та пояснити компоненти, вимоги та/або характеристики ШІ. PLO2. Розпізнавати, ідентифікувати, описувати, визначати та/або пояснювати застосування ШІ в різних сферах.
Знання та розуміння ШІ	Знати основні функції ШІ та використовувати програми ШІ	> 50%	
Використання та застосування ШІ	Застосування знань, концепцій і програм ШІ в різних сценаріях	> 50%	PLO3. Вибирайте та/або використовуйте інструменти та методи штучного інтелекту, які відповідають конкретному контексту та застосуванню (критичне мислення та знання змісту).
Оцінка та розвиток ШІ	Навички мислення вищого рівня (наприклад, оцінювання, прогнозування, проектування) із застосуванням ШІ	> 50%	PLO4. Оцінка контекстної цінності або якості інструментів і програм AI (критичне мислення) PLO5. Концептуалізація та/або розробка інструментів, обладнання, даних та/або алгоритмів, що використовуються в рішеннях ШІ (критичне мислення)
Етика ШІ	Людиноорієнтовані міркування (справедливість, підзвітність, прозорість, етика, безпека)	> 50%	PLO6. Розробляти, застосовувати та/або оцінювати контекст етичних рамок для використання в усіх аспектах ШІ.

Педагогіка ШІ надасть можливість застосовувати прийняття рішень за допомогою ШІ та ML.

Методи підтримки прийняття рішень із використанням ШІ включають [5]:

- експертні системи, засновані на правилах (якщо X, то Y)
- машинне навчання («навчання» на основі даних без явного програмування)
- нейронні мережі (виявлення шаблонів або закономірностей у даних)
- глибоке навчання (обробка даних на декількох рівнях)

Штучна нейронна мережа, заснована на архітектурі глибокого навчання та вхідних даних, допомагає вибрати найкращу стратегію для досягнення цілей ЗВО (Рис. 3.1).

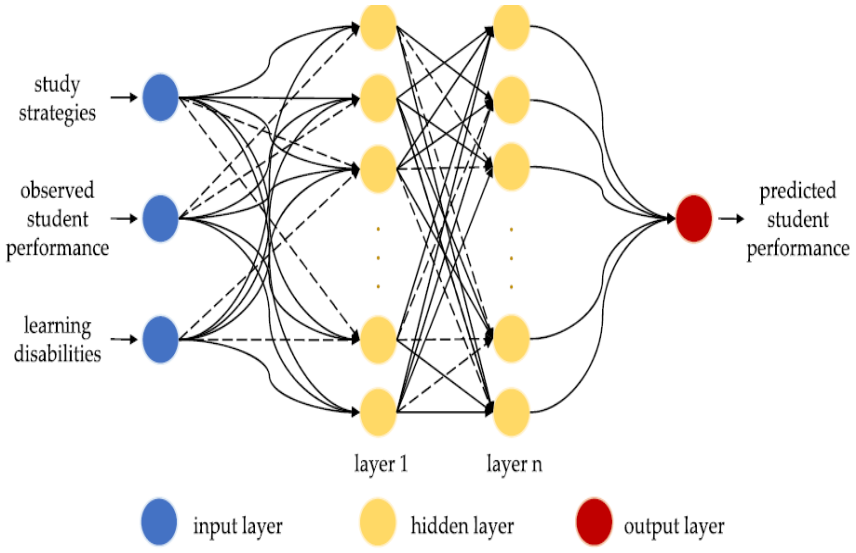

Рис. 3.1. Штучна нейронна мережа, заснована на архітектурі глибокого навчання [9]

Менеджерські знання та навички для інтерпретації інформації з ШІ для прийняття рішень охоплюють [5]:

- сучасне статистичне моделювання,
- аналітичні методи текстового аналізу,
- аналіз настроїв,
- аналіз мереж та згорткових мереж.

За відсутності цих аналітичних знань і навичок керівництво ВНЗ може приймати менш обґрунтовані рішення. Сфери застосування ШІ для підтримки прийняття рішень адміністрацією університету описані в таблиці 3.2.

Таблиця 3.2

Класифікація систем підтримки прийняття рішень

Сфери застосування	**Опис**
Прогнозування кількості претендентів	Моделі ШІ можуть аналізувати історичні дані та тенденції, щоб точніше прогнозувати кількість і профілі майбутніх студентів. Це допомагає планувати майбутнє
Персоналізація навчання студентів	Аналізуючи дані, ШІ може визначити сильні та слабкі сторони кожного студента та розробити індивідуальні плани навчання (введення спеціальності, а не спеціальності, реагування на емоції студента, допомога абітурієнтам з особливими потребами)
Запобігання відрахуванню студентів	Аналітика штучного інтелекту дозволяє виявляти студентів із високим ризиком відсіву та вживати профілактичних заходів.
Підвищення ефективності	ШІ може оптимізувати планування, управління гуртожитком, розподіл ресурсів, дослідження, маркетинг тощо.
Набір викладачів	На основі попередніх результатів викладання ШІ може рекомендувати найкращих кандидатів на вакантні посади.
Автоматизація рутинних завдань	За допомогою МН ШІ може приймати рішення щодо простих повторюваних питань без втручання людини.

ШІ може значно покращити аналітику, планування, персоналізацію та ефективність в університетах. Однак людська участь все ще є важливою. ШІ не може замінити людський дотик у незвичних і чутливих сферах (табл. 3.3).

Таблиця 3.3

Класифікація систем підтримки прийняття рішень

Сфери застосування	**Опис**
Морально-етичні рішення	Рішення про виключення студентів, звільнення співробітників і розслідування скарг вимагають етичної оцінки [10], яку може надати лише людина. Моральні цінності (справедливість, чесність, відсутність шкоди) можуть суперечити використанню ШІ для прийняття рішень на основі даних.
Невизначені та нестандартні ситуації	У складних, непередбачуваних випадках може допомогти лише людський життєвий досвід, критичне мислення та креативність.
Стратегічні рішення	Формування довгострокових цілей і стратегій розвитку вимагає суто людського бачення. AI може служити засобом реалізації, але не визначення стратегії [11].
Кризові ситуації	У часи кризи, коли ставки дуже високі, люди часто не довіряють автоматизованим системам і радше покладаються на людську мудрість і досвід.
Креативність та інноваційність	Розвиток нових ідей, шляхів розвитку, освітніх підходів тощо базується на людській уяві та творчості.

У багатьох випадках адміністраторам університетів доводиться покладатися на людський фактор [12], а не лише на рекомендації ШІ. Це означає, що оптимальним є комплексний підхід. Обмеження та ризики рішень на основі даних за допомогою ШІ наведено в таблиці 3.4.

Таблиця 3.4

Обмеження та ризики рішень на за допомогою ШІ [5, 13]

Обмеження	**Ризики**
Робота освітніх керівників полягає не в тому, щоб реалізовувати рекомендації, створені штучним інтелектом, а підтримувати процес прийняття рішень, забезпечуючи контекст для рекомендацій штучного інтелекту.	Підвищене упередження (за статтю, расою, багатством)
Лідери повинні протистояти спокусі шукати швидких рішень, не розглядаючи деталі та не підтримуючи соціальні стосунки.	Морально-етичне прийняття рішень (люди не можуть бути зведені до даних)
Керівники освітніх закладів повинні розглядати ШІ як радника, а не як орган, який приймає рішення. Дані можуть лише інформувати, але «ніколи повністю не керують рішеннями».	Питання безпеки та конфіденційності (витік конфіденційної інформації для учасників ВНЗ)

Усвідомлення прихованих упереджень у ШІ є першим кроком для керівників освіти, щоб розглянути можливості використання ШІ в прийнятті рішень на основі даних для протидії упередженням, а не їх підсилення.

Рекомендації для прийняття рішень на основі даних із використанням ШІ:

1. Громадський контроль над ШІ з боку зацікавлених сторін [14]: люди, які добре знають дані; люди, на яких вплинуть рішення (наприклад, вчителі, батьки, студенти, громада); люди, які розробляють алгоритми ШІ.

2. Ставлення до ШІ як до підтримки прийняття рішень, а не як до заміни людських рішень.

3. Вища освіта повинна інтегрувати теорію та практику ШІ в усі сфери освітнього процесу, а не розглядати їх як "додаткову" вимогу.

3.3 **Методологія.**

Для дослідження впливу предикторів на зарахування вступників будуть використані такі моделі МН: лінійна регресія, логістична регресія, метод K-найближчих сусідів, дерево рішень та випадковий ліс. Адміністрація вищих навчальних закладів може ухвалювати рішення щодо своєї цільової аудиторії та маркетингових зусиль для конкретних закладів, враховуючи характеристики вступників, які стали студентами ВНЗ, на основі результатів моделей. Набір даних складається зі 122 заявок абітурієнтів на спеціальності галузі знань 12 "Інформаційні технології" для денної форми навчання на бакалавраті Херсонського державного університету (ХДУ) у 2023 році [15]. Описова статистика даних представлена у таблиці 3.5. Ми можемо перетворити якісні показники [16] (стать, регіон, відзнаки) на кількісні, використовуючи бінарні або дискретні індикатори, щоб розширити кількість предикторів у моделях МН.

Таблиця 3.5

Описова статистика абітурієнтів

Predictors	**Mean**	**Standard deviation**	**Min**	**Max**
Join (Y)	0.18	0.4	0	1
Age (X1)	17.3	0.7829	17	20
Sex (X2)	0.8	0.4	0	1
Score (X3)	149.69	13.967	0	186.5
Region (X4)	0.607	0.491	0	1
Category (X5)	0.5983	0.4922	0	1
Budget (X6)	0.787	0.411	0	1
Honor (X7)	0.082	0.275	0	1
Priority (X8)	2.1721	1.6599	0	5
Speciality (X9)	1.7459	0.8582	0	3

Наступні предиктори вступників використовуються в ML-моделях для визначення їхнього впливу на зарахування: вік, стать, середній бал при вступі

(Score), регіон подання документів до ХДУ (Region: 1 – Херсонська область, 0 – інші), чи був вступник зарахований (Join: 1 – зарахований, 0 – не зарахований), пільгові категорії (Category: 1 – має пільги, 0 – не має пільг), заявка на бюджет (Budget: 1 – подає на бюджет, 0 – не подає), нагорода за останнє місце навчання, наприклад, золота/срібна медаль (Honor: 1 – наявна, 0 – відсутня), пріоритет, встановлений вступником при поданні документів (Priority: 0 – контракт, 1, 2, 3, 4, 5 – пріоритет вступу на бюджет) і спеціальність, на яку подають документи (Speciality: 1 – Середня освіта (інформатика), 2 – Програмна інженерія, 3 – Комп'ютерні науки, 4 – Інформаційні системи та технології). Аналіз даних для пояснення виявляє рівень кореляції між усіма факторами (Рис. 3.2) і розподіл балів вступників (Рис. 3.3).

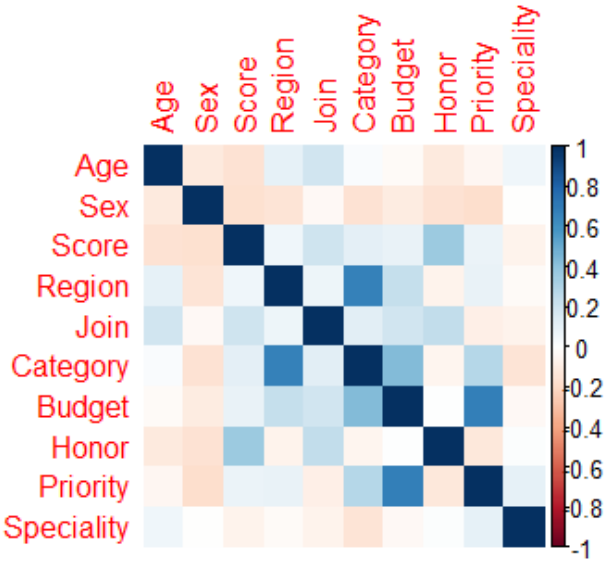

Рис. 3.2. Кореляція всіх факторів (предиктори та залежна змінна)

Ми розглядаємо лінійну регресію, логістичну регресію, метод K найближчих сусідів, дерево рішень і випадковий ліс як методи машинного навчання для оцінки їхнього потенціалу в прогнозуванні зарахування абітурієнтів. Точність результатів моделей машинного навчання визначається середньоквадратичною помилкою (RMSE) та наступною формулою з матриці помилок:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3.1)$$

де TP, FP, TN, FN позначають істинно позитивні, хибно позитивні, істинно негативні та хибно негативні передбачення відповідно.

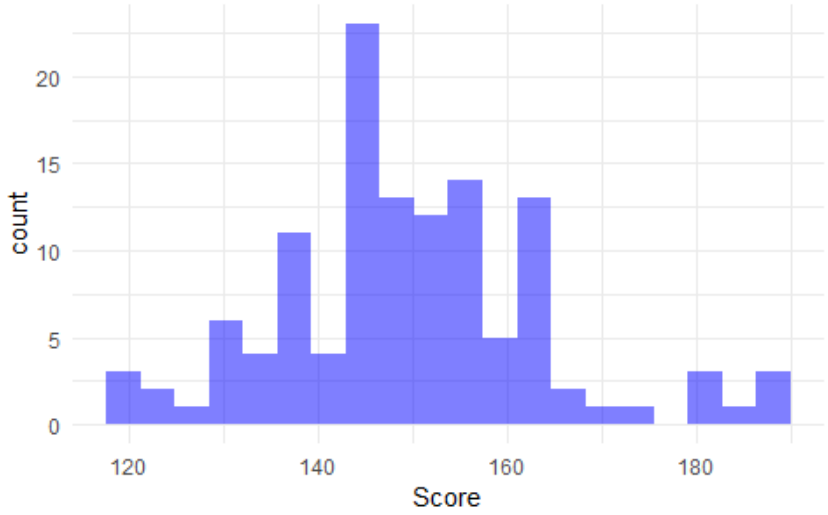

Рис. 3.3. Гістограма для змінної Score

3.4 Результати

3.4.1 Лінійна регресія

Спочатку розглянемо результати прогнозування за допомогою моделі лінійної регресії при визначенні її параметрів на основі навчальної вибірки (Рис. 3.4). Код доступний у [15].

```
Call:
lm(formula = Join ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65245 -0.18375 -0.07080  0.04085  0.87466

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.426396   1.028235  -2.360   0.0209 *
Age          0.117479   0.047907   2.452   0.0165 *
Sex          0.125895   0.101585   1.239   0.2190
Score        0.002337   0.003144   0.743   0.4596
Region      -0.107029   0.101647  -1.053   0.2957
Category     0.116431   0.109817   1.060   0.2924
Budget       0.325127   0.134605   2.415   0.0181 *
Honor        0.352397   0.155166   2.271   0.0260 *
Priority    -0.076803   0.031859  -2.411   0.0183 *
Speciality  -0.009196   0.043477  -0.212   0.8331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.341 on 76 degrees of freedom
Multiple R-squared:  0.2461,    Adjusted R-squared:  0.1568
F-statistic: 2.756 on 9 and 76 DF,  p-value: 0.007631
```

Рис. 3.4. Результати оцінки моделі лінійної регресії з використанням навчальної вибірки

Параметри моделі: Кожен з коефіцієнтів показує, як зміна на одиницю відповідного предиктора (наприклад, вік, стать, бюджет тощо) впливає на залежну змінну Join (Рис. 4). Наприклад, коефіцієнт для віку вказує на те, що збільшення віку абітурієнта на один рік, як очікується, збільшить ймовірність зарахування на 11,7% (приймається більш свідоме та обґрунтоване рішення). Якщо абітурієнт зарахований на бюджетну програму, ймовірність зарахування зростає на 32,5%. Якщо абітурієнт має нагороду за успіхи в навчанні, ймовірність зарахування зростає на 35,2%. Негативний коефіцієнт для пріоритету означає, що збільшення пріоритету на 1 (що означає менш привабливу альтернативу для абітурієнта) зменшує ймовірність зарахування на 7,7%.

Рівень значущості: Значення "Pr(>|t|)" показує статистичну значущість параметра, тобто ймовірність того, що предиктор не впливає на залежну змінну (заходження). Якщо ця ймовірність перевищує 0,1 (10%), предиктор вважається статистично незначущим і не враховується при аналізі впливу предикторів на зарахування (нульова гіпотеза підтверджується: параметр дорівнює нулю). Якщо ця ймовірність не менше 0,1 (10%), наприклад, 0,05 (5%), тоді альтернативна гіпотеза про статистичну значущість предиктора, що відповідає цьому параметру, підтверджується.

Коефіцієнт детермінації (R-квадрат) і RMSE: R-квадрат показує, яку частину варіації в залежній змінній пояснює модель (набір предикторів). У нашому випадку R-квадрат=0,2461, що означає, що модель пояснює приблизно 24,61% варіації в залежній змінній. Метрика RMSE=0,381 означає, що модель відхиляється на 0,381 одиниці в середньому при прогнозуванні залежної змінної. Чим менше значення RMSE, тим краще, оскільки це вказує на більш точні прогнози.

p-значення F-статистики: показує ймовірність того, що всі параметри регресії дорівнюють нулю, за умови, що модель правильно побудована, тобто лінійна регресія відображає існуючий лінійний зв’язок. p-значення=0,007631 близьке до нуля і вказує на те, що модель є статистично значущою, тобто в цій

моделі є принаймні один статистично значущий предиктор.

Мультиколінеарність: предиктори з високою кореляцією між собою можуть призводити до мультиколінеарності, що ускладнює інтерпретацію коефіцієнтів. У нашому випадку, наприклад, можна побачити, що існує висока кореляція між регіоном і категорією (0,675), що може вплинути на інтерпретацію та точність коефіцієнтів. Це означає, що, ймовірно, лише один з цих факторів можна залишити в лінійній моделі, щоб уникнути мультиколінеарності. Давайте обчислимо фактори інфляції дисперсії (VIF) для кожної змінної, щоб перевірити мультиколінеарність (Рис. 3.5):

```
> vif_values <- vif(model)
> print(vif_values)
      Age       Sex     Score    Region  Category    Budget     Honor
1.083354  1.099111  1.126686  1.826847  2.152838  2.125456  1.155766
Priority Speciality
2.098383  1.094168
```

Рис. 3.5. Фактори інфляції дисперсії (VIF) предикторів

За результатами розрахунків всі предиктори мають VIF<10, тобто мультиколінеарність між предикторами в моделі відсутня.

Також розрахуємо важливісить факторів інфляції дисперсії (VIF) предикторів, для моделі лінійної регресії (Рис. 3.6):

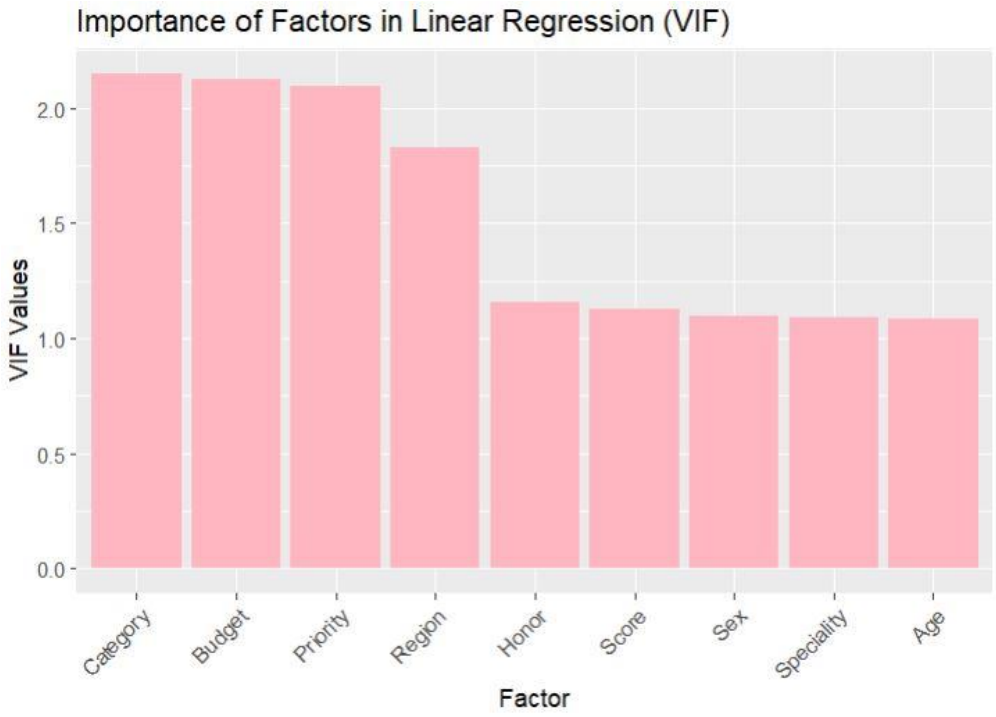

Рис. 3.6. Важливість факторів для лінійної регресії.

Діаграма показує значення **VIF (Variance Inflation Factor)** для факторів у моделі лінійної регресії, що відображає рівень мультиколінеарності між ними.

- Високі значення VIF у факторів *Category*, *Budget*, і *Priority* вказують на сильну кореляцію з іншими змінними, що може впливати на стабільність моделі.
- Низькі значення VIF для факторів *Honor*, *Score*, *Sex*, *Speciality*, і *Age* свідчать про їхню низьку кореляцію з іншими ознаками.

Фактори з високими VIF можуть потребувати перегляду, щоб зменшити мультиколінеарність і покращити точність моделі.

3.4.2 Логістична регресія

Результати моделі логістичної регресії показані на рис. 3.7 та в [15].

```
Call:
glm(formula = Join ~ ., family = binomial(link = "logit"), data = df2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -27.00612    8.33207  -3.241  0.00119 **
Age           0.99783    0.34768   2.870  0.00411 **
Sex           0.74396    0.82031   0.907  0.36445
Score         0.03617    0.02549   1.419  0.15584
Region       -0.73872    0.84079  -0.879  0.37962
Category      0.75853    0.91653   0.828  0.40789
Budget        3.77545    1.37634   2.743  0.00609 **
Honor         1.52154    1.04438   1.457  0.14515
Priority     -0.64450    0.25603  -2.517  0.01183 *
Speciality   -0.06585    0.34501  -0.191  0.84864
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 115.141  on 121  degrees of freedom
Residual deviance:  84.729  on 112  degrees of freedom
AIC: 104.73
```

Рис. 3.7. Результати моделі логістичної регресії для всього набору даних

У логістичній моделі статистично значущими предикторами є вік, бюджет і пріоритет, де пріоритет має негативний вплив, а інші предиктори мають позитивний вплив. При поділі набору даних у співвідношенні 70:30 для навчального набору та тестового набору для логістичної моделі ми отримуємо наступні результати точності для матриці плутанини (рис. 3.8):

```
> table(final.test$Join, fitted.probabilities > 0.5)

    FALSE TRUE
  0    29    1
  1     6    1
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.810810810810811"
> rmse
[1] 0.4349588
```

Рис. 3.8. Матриця похибок для логістичної моделі

Результати моделі: Результати моделі логістичної регресії представляють значний прогрес у прогнозуванні прийому абітурієнтів. Модель демонструє високу точність прогнозування, досягаючи 81,08%, що вказує на її здатність ефективно класифікувати дані та робити більш точні прогнози, ніж лінійна модель.

RMSE (середньоквадратична помилка): RMSE=0,435 для моделі логістичної регресії, що вказує на те, що середнє квадратів відхилень між фактичними та прогнозованими значеннями є досить малим. Це означає, що модель добре адаптується до даних і має невелику помилку прогнозування.

Обмеження: Неврівноважені класи даних можуть викликати спотворення в оцінках якості моделі, оскільки пропорції зарахованих і не зарахованих значно відрізняються. Крім того, виявилося, що багато ознак є статистично незначущими, що вимагає додаткового аналізу щодо того, чи слід їх виключити з моделі.

На діаграмі (Рис. 3.9) показано нормалізовану важливість факторів у моделі логістичної регресії:

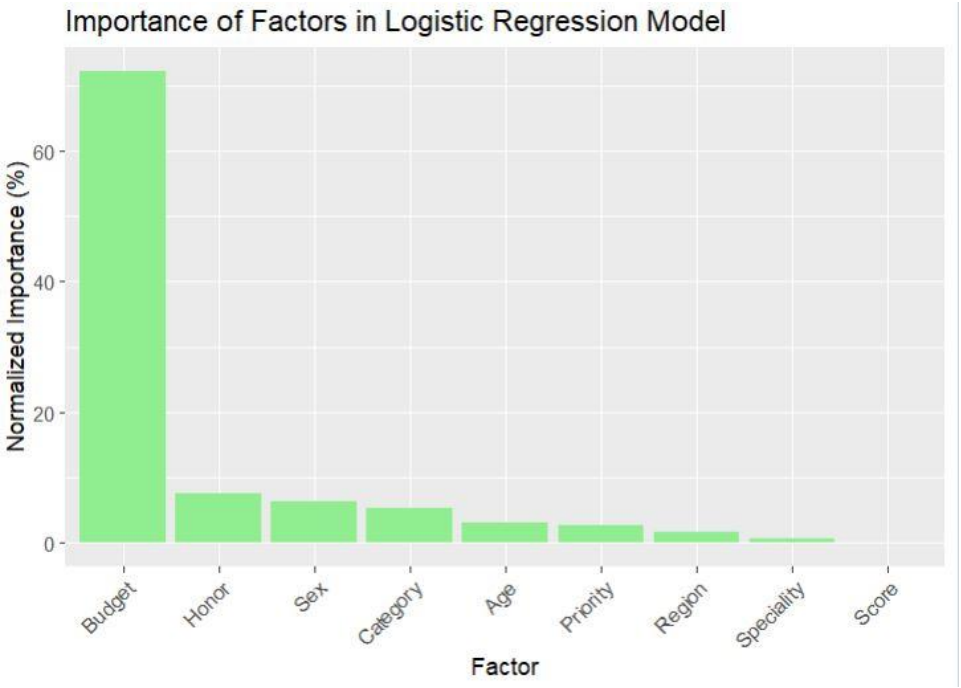

Рис. 3.9. Важливість факторів для моделі логістичної регресії.

Фактор *"Budget"* має найвищу важливість і домінує над іншими факторами, показуючи найбільший вплив на результат моделі.

Інші фактори, такі як *Honor, Sex, Category*, та *Age*, мають значно нижчу важливість, що вказує на те, що вони практично не впливають на модель.

3.4.3 К найближчих сусідів

Результати алгоритму KNN показані на рис. 3.10 і в [15].

```
#KNN

set.seed(101)
split <- sample.split(df$Join, SplitRatio = 0.70)
train <- subset(df, split == TRUE)
test <- subset(df, split == FALSE)

k <- 3
knn_model <- knn(train[, -1], test[, -1], train$Join, k)

knn_model_numeric <- as.numeric(as.character(knn_model))

misClasificError <- mean(knn_model_numeric != test$Join)
accuracy <- 1 - misClasificError
print(paste('Accuracy:', accuracy))

rmse <- sqrt(mean((knn_model_numeric - test$Join)^2))
print(paste('RMSE:', rmse))
```

Рис. 3.10 Результати алгоритму найближчих сусідів

Accuracy: Модель KNN продемонструвала вищу точність порівняно з

лінійною та логістичною регресією, досягнувши 83,78%.

RMSE: 0,4027, що вказує на середню відстань між фактичними та прогнозованими значеннями. Нефункціональний метод класифікації KNN показав вищу якість прогнозування порівняно з функціональними методами лінійної та логістичної регресій, що дозволяє KNN приймати кращі рішення для прогнозування вступу абітурієнтів. Модель KNN має певні обмеження, наприклад, незбалансовані класи даних можуть призвести до спотворень у оцінках якості моделі, неможливість інтерпретувати вплив предикторів і необхідність ручного вибору параметра k. У цій моделі коефіцієнт k=3 дав найкращий результат.

На діаграмі (рис. 3.11) зображено нормалізовану важливість факторів у моделі К-найближчих сусідів (KNN).

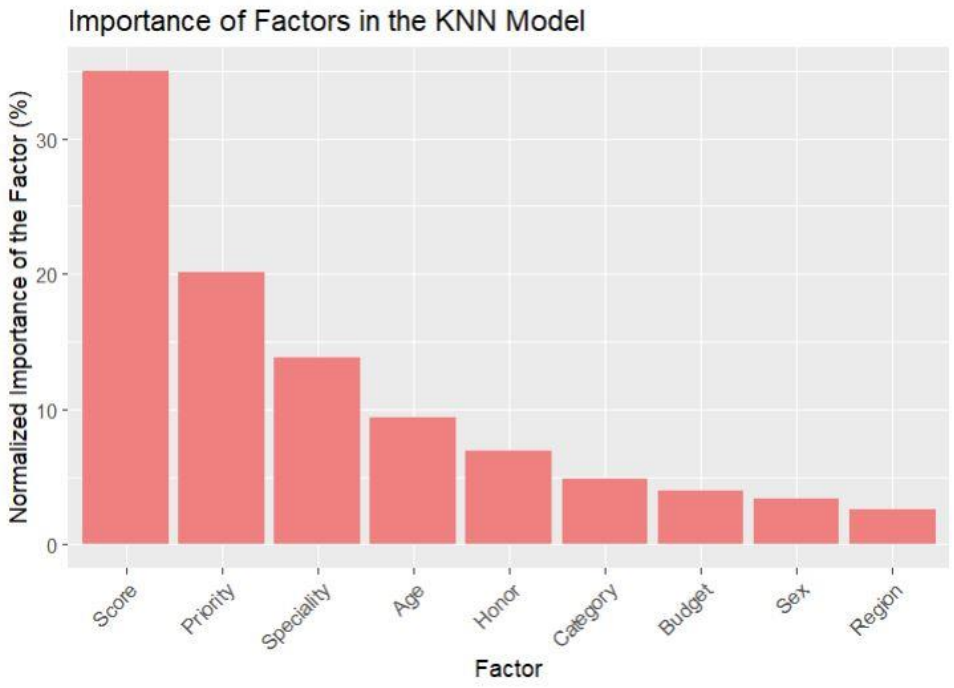

Рис. 3.11. Важливість факторів для моделі К-найближчих сусідів.

Фактор "*Score*" є найбільш важливим і має найбільший вплив на результати моделі KNN.

Другий за значимістю — *Priority*, тоді як *Speciality* та *Age* також мають помітний вплив, але значно менший.

Фактори *Budget, Sex,* та *Region* мають низьку важливість, що вказує на мінімальний вплив на модель KNN.

Таким чином, *"Score"* є головним фактором для цієї моделі, а деякі інші фактори мають лише незначний вплив на результати.

3.4.4 Випадковий ліс

Результати моделі випадкового лісу показані на рис. 3.12 [15].

```
[1] "RMSE: 0.464990554975277"
> conf_matrix <- confusionMatrix(predictions, test$Join)
> print(conf_matrix$overall['Accuracy'])
 Accuracy
0.7837838
```

Рис. 3.12. Результати моделі випадковий ліс

Accuracy: 0.7838 означає, що модель правильно класифікує близько 78.38% випадків, що нижче за середню точність серед попередніх методів.

RMSE: 0.465 показує, що середня помилка прогнозу є вищою за середню помилку попередніх методів.

Діаграма (рис. 3.13) показує нормалізовану важливість факторів у моделі випадкового лісу (Random Forest).

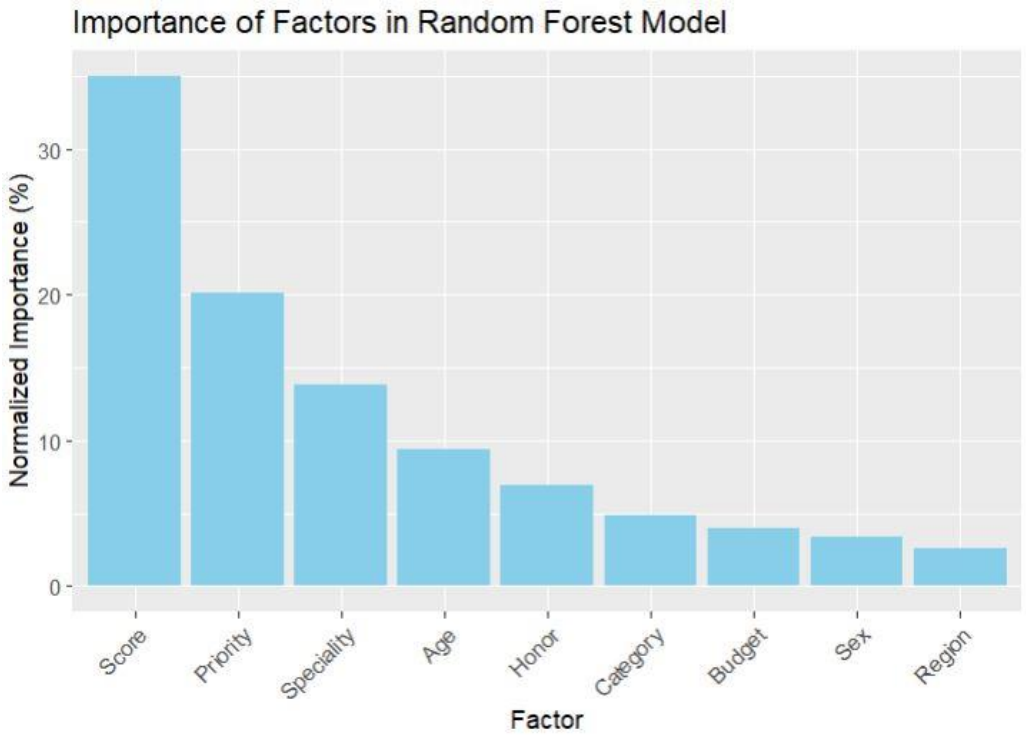

Рис. 3.13. Важливість факторів для моделі випадкового лісу.

Фактор "*Score*" має найвищу значимість і найбільше впливає на результати моделі.

Priority і *Speciality* також значущі, але їхній вплив поступається "Score".

Age і *Honor* мають середню важливість, тоді як *Category*, *Budget*, *Sex* та *Region* демонструють мінімальний вплив.

Таким чином, "*Score*" є ключовим предиктором у моделі, тоді як інші фактори мають другорядне значення.

3.4.5 Дерево рішень

Результати моделі дерева рішень (DT) показані на рис. 3.14 [15].

```
> print(paste('Accuracy:', accuracy))
[1] "Accuracy: 0.777777777777778"
> rmse <- sqrt(mean((as.numeric(as.character(predictions)) - as.numeric(as.character
(test$Join)))^2))
> print(paste('RMSE:', rmse))
[1] "RMSE: 0.471404520791032"
```

Рис. 3.14. Результати моделі дерева рішень

Accuracy=77.78%. Висока точність означає, що модель дерева рішень працює добре і може надійно класифікувати нові приклади.

RMSE=0,4714 вказує на те, що середньоквадратична помилка моделі дерева рішень є помірною порівняно з попередніми моделями. Загалом, враховуючи точність і RMSE, можна зробити висновок, що модель дерева рішень показує прийнятні результати, але є можливості для покращення у вигляді більш складних моделей або налаштування гіперпараметрів.

На діаграмі (Рис. 3.15) показано нормалізовану важливість факторів у моделі дерева рішень (Decision Tree).

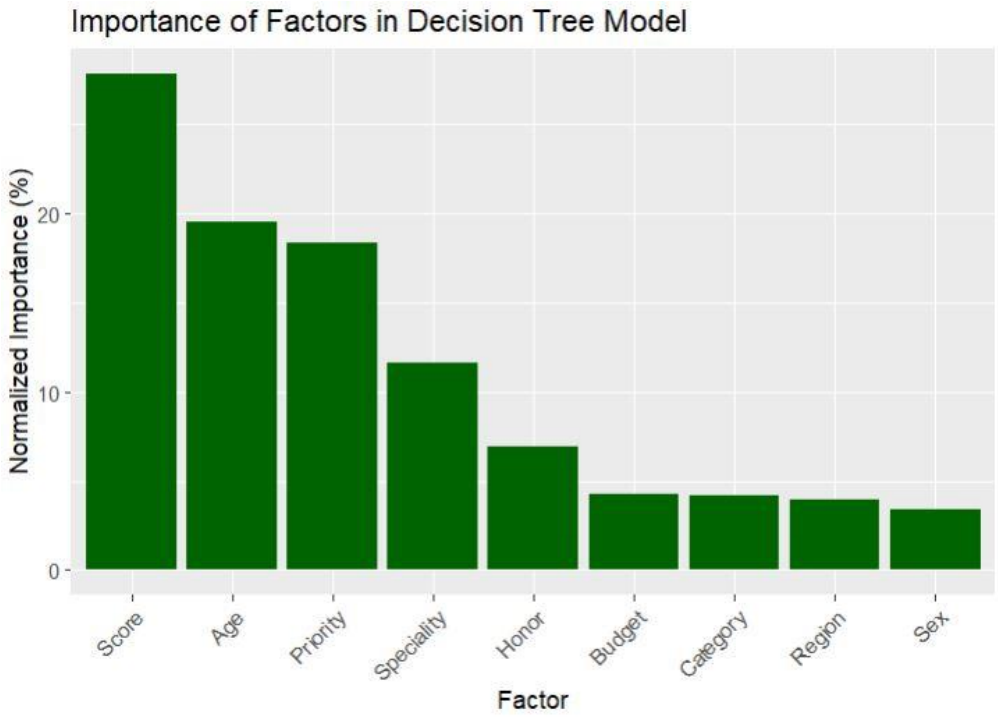

Рис. 3.15. Важливість факторів для моделі дерева рішень.

Фактор "*Score*" має найвищу значимість, що свідчить про його ключову роль у прийнятті рішень моделі. "*Age*" і "*Priority*" також мають значний вплив, хоча вони менш важливі порівняно з "*Score*". Фактори "*Speciality*" та "*Honor*" впливають на результати моделі, але в меншій мірі. Натомість "*Budget*", "*Category*", "*Region*" і "*Sex*" мають найнижчу значимість, отже, їхній вплив на рішення моделі є мінімальним.

Загалом, "*Score*" є основним предиктором, тоді як інші фактори мають помірний або незначний вплив у цій моделі дерева рішень.

3.5 Результати

На основі аналізу значень RMSE і метрик точності (Таблиця 6) можна зробити висновок, що модель лінійної регресії демонструє найнижче RMSE, рівне 0,381, що вказує на найкращу загальну прогностичну ефективність серед порівнюваних методів.

Однак лінійна регресія більше підходить для завдань прогнозування, ніж для класифікації. Серед методів класифікації алгоритм KNN показує наступну найкращу ефективність із RMSE 0,4027, за яким йде логістична регресія зі значенням 0,435. Моделі дерева рішень і випадкового лісу демонструють трохи вищі значення RMSE, що свідчить про те, що вони можуть бути менш точними у цьому конкретному застосуванні.

Таблиця 3.6

RMSE та точність для різних моделей машинного навчання

Model	**RMSE**	**Accuracy**
Linear Regression	**0.381**	-
Logistic Regression	0.435	**81.08%**
K-Nearest Neighbors	**0.4027**	**83.78%**
Random Forest	0.465	78.38%
Decision Tree	0.4714	77.78%

При оцінюванні точності алгоритм KNN показав найкращий результат у передбаченні прийняття кандидатів з точністю 0.8378, за ним слідує модель логістичної регресії з точністю 0.8108. Якщо точність і RMSE є ключовими метриками, тоді KNN або логістична регресія можуть бути найкращими варіантами. Крім того, логістична регресія дозволяє аналізувати кожен предиктор окремо для нелінійних зв'язків і виявляти статистично значущі пояснювальні змінні. У майбутньому прогнозування ймовірності прийняття кандидата може бути розширене аналізом його настрою в соціальних мережах та інших неструктурованих даних [17].

ВИСНОВКИ

Результати дослідження дозволили розробити та випробувати методологію застосування автоматизованого машинного навчання для підтримки ухвалення рішень у бізнес-аналітиці, зокрема на основі даних про вступників. Застосування алгоритмів кластеризації та прогнозування сприяло виявленню ключових закономірностей та факторів, що впливають на рішення щодо вибору навчального закладу, а також створило основу для прогнозування попиту на освітні послуги в майбутньому.

Отримані результати можуть бути використані для оптимізації освітніх програм та підвищення їх відповідності потребам ринку, що, у свою чергу, сприятиме підвищенню конкурентоспроможності випускників. Запропонована методологія автоматизованого машинного навчання є універсальною та може застосовуватись як у сфері освіти, так і в інших галузях, де прийняття рішень базується на великих обсягах даних, що забезпечує її широке практичне застосування.

Серед усіх методів алгоритм KNN продемонстрував найкращий результат у передбаченні прийняття кандидатів з точністю 0.8378 і RMSE=0.4027. Модель логістичної регресії також має високу точність 0.8108. Лінійна регресія показала найкращий результат у частині RMSE=0.381, але вона більше підходить для задач прогнозування, а не для класифікації. Таким чином, модель KNN (RMSE=0.4027) є найкращою за показниками RMSE і точності.

При виборі найкращої моделі необхідно також враховувати контекст завдання та інші фактори. Наприклад, якщо точність є ключовою метрикою, то метод випадкового лісу може бути найкращим варіантом, але якщо важлива точність прогнозу для лінійних залежностей, тоді може бути кращою лінійна регресія. Модель логістичної регресії дозволяє аналізувати кожен предиктор окремо для нелінійних зв'язків і виявляти статистично значущі предиктори.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Liao, F., Zhang, C., Zhang, J., Yan, Xiang, Chen, T.: Hyperbole or reality? The effect of auditors' AI education on audit report timeliness. International Review of Financial Analysis 91, 103050 (2024). [https://doi.org/10.1016/j.irfa.2023.103050]
2. Alqahtani, T., Badreldin, H.A., Alrashed, M., Alshaya, A.I., Alghamdi S.S., Saleh, K., Alowais, S.A., Alshaya, O.A., Rahman, I., Yami, M.S.A., Albekairy A.M.: The emergent role of AI, natural learning processing, and large language models in higher education and research. Research in Social and Administrative Pharmacy 19, 1236–1242 (2023). [https://doi.org/10.1016/j.sapharm.2023.05.016]
3. Tayan, O., Hassan, A., Khankan, K., Askool, S.: Considerations for adapting higher education technology courses for AI large language models: A critical review of the impact of ChatGPT. Machine Learning with Applications 15, 100513 (2024). [https://doi.org/10.1016/j.mlwa.2023.100513]
4. Boubker, O.: From chatting to self-educating: Can AI tools boost student learning outcomes? Expert Systems With Applications 238, 121820 (2024). [https://doi.org/10.1016/j.eswa.2023.121820]
5. Wang, Y.: When artificial intelligence meets educational leaders' data-informed decision-making: A cautionary tale. Studies in Educational Evaluation 69, 100872 (2021). [https://doi.org/10.1016/j.stueduc.2020.100872]
6. Komleva, N., Liubchenko, V., Zinovatna, S., Kobets, V.: Decision support system for quality management in learning process. In: Proceedings of 9th International Conference Information Control Systems and Technologies, pp. 430–442. CEUR-WS 2711, AU (2020)
7. Southworth, J., Migliaccio, K., Glover, J., Glover, Ja'Net, Reed, D., McCarty, C., Brendemuhl, J., Thomas, A.: Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in

AI literacy. Computers and Education: Artificial Intelligence 4, 100127 (2023). [https://doi.org/10.1016/j.caeai.2023.100127]

8. Carolus, A., Augustin, Y., Markus, A., Wienrich, C.: Digital interaction literacy model – Conceptualizing competencies for literate interactions with voice-based AI systems. Computers and Education: Artificial Intelligence 4, 100114 (2023). [https://doi.org/10.1016/j.caeai.2022.100114]

9. Bressane, A., Zwirn, D., Essiptchouk, A., Saraiva, A.C.V., Carvalho, F.L.C., Formiga, J.K.S., Medeiros, L.C.C., Negri, R.G.: Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: A proposal using AI. Computers and Education: Artificial Intelligence 6, 100196 (2024). [https://doi.org/10.1016/j.caeai.2023.100196]

10. Wang, X., Li, L., Tan, S.C., Yang, L., Lei, J.: Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness. Computers in Human Behavior 146, 107798 (2023). [https://doi.org/10.1016/j.chb.2023.107798]

11. Tamakia, K., Arakawab, M., Aramec, M., Onod, Y.: Development of Educational Programs for System Creators and Business Producers in Future Strategy Design in Action Project Group Activities Through Industry-University Cooperation. Procedia Manufacturing 39, 1377-1382 (2019). [https://doi.org/10.1016/j.promfg.2020.01.319]

12. Kobets, V., Yatsenko, V., Buiak, L.: Bridging Business Analysts Competence Gaps: Labor Market Needs Versus Education Standards. In: ICTERI 2020. Communications in Computer and Information Science, vol. 1308, pp. 22-45 (2021). [https://doi.org/10.1007/978-3-030-77592-6_2]

13. Habbal, A., Ali, M.K., Abuzaraida, M.A.: Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. Expert Systems With Applications 240, 122442 (2024). [https://doi.org/10.1016/j.eswa.2023.122442]

14. Kravtsov, H., Kobets, V.: Implementation of stakeholders' requirements and innovations for ICT curriculum through relevant competences. In: Proceedings of 13th International Conference on ICTERI 2017, pp. 414–427. CEUR-WS 1844, Aachen University (2017)

15. Code in RStudio: [https://github.com/MatrimFox/R-studio/tree/main], last accessed 2024/08/03

16. Kobets, V., Osypova, N.V.: Identification of factors for providing the higher education quality assurance for students. International Journal for Quality Research, 17(1), 195–208 (2023). [https://doi.org/10.24874/ijqr17.01-12]

17. Derbentsev, V., Bezkorovainyi, V., Akhmedov, R.: Machine learning approach of analysis of emotional polarity of electronic social media. Neuro-Fuzzy Modeling Techniques in Economics 9, 95-137 (2020). [http://doi.org/10.33111/nfmte.2020.095]