Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра _____Інформаційних управляючих систем_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____другий (магістерський)_____

Дослідження методів інтелектуального аналізу даних в задачах оцінювання можливості реалізації інформаційної системи
(тема)

Виконав:
студент __2__ курсу, групи ІУСТм-19-1_
_____Затолокіна Л. О._____
(прізвище, ініціали)

Спеціальність ____122 Комп'ютерні науки_____
(код і повна назва спеціальності)

Тип програми __освітньо-професійна_____
(освітньо-професійна або освітньо-наукова)

Освітня програма__Інформаційні управляючі системи та технології_____
(повна назва освітньої програми)

Керівник __проф. Чалий С.Ф._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ ___Петров К.Е.____
(підпис) (прізвище, ініціали)

2020р.
Харківський національний університет
радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Інформаційних управляючих систем
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Інформаційні управляючі системи та технології
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові Затолокіной Любові Олегівні
(прізвище, ім'я, по батькові)

1 Тема роботи Дослідження методів інтелектуального аналізу даних в задачах оцінювання можливості реалізації інформаційної системи

затверджена наказом по університету від 27 жовтня 2020 р. № 1455 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 17 12 2020 р.
3. Вихідні дані до роботи Література щодо існуючих моделей інтелектуального аналізу даних, керівництво з використання мов Python, CSS, HTML, Bootstrap, Flask, відомості про методи інтелектуального аналізу даних.

4. Перелік питань, що потрібно опрацювати в роботі Огляд процесу інтелектуального аналізу даних та його методів класифікації, створення технології оцінювання можливості реалізації інформаційної системи, вибір удосконаленого методу оцінювання можливості реалізації інформаційної системи, застосування удосконаленого методу оцінювання можливості реалізації інформаційної системи.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
	Отримання завдання на дипломну роботу	02.11.2020	
	Аналіз предметної галузі і постановка задачі	03.11.2020 - 10.11.2020	
	Створення теоретичного підходу для вирішення задачі	11.11.2020 - 12.11.2020	
	Створення технології удосконаленого методу для вирішення задачі	13.11.2020 - 20.11.2020	
	Застосування створеного удосконаленого методу	20.11.2020 - 23.11.2020	
	Експериментальна перевірка отриманого результату	23.11.2020 - 24.11.2020	
	Створення графічного інтерфейсу	25.11.2020 - 30.11.2020	
	Оформлення пояснювальної записки	01.12.2020-10.12.2020	
	Оформлення графічних матеріалів	10.12.2020 – 14.12.2020	
	Попередній захист	14.12.2020	
	Захист перед ЕК	17.12.2020	

Дата видачі завдання __02__ __11__ 2020 р.

Студент _____
(підпис)

Керівник роботи _____ _____проф. Чалий С.Ф._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до магістерської атестаційної роботи містить: 86 с., 4 розділи, 34 рис., 4 табл., 31 джерело.

АНСАМБЛІ КЛАСИФІКАТОРІВ, ВИЗНАННЯ ІМЕНОВАНОЇ СУТНОСТІ, ВІЗУАЛІЗАЦІЯ, ІНФОРМАЦІЙНІ СИСТЕМИ, МЕТОДИ КЛАСИФІКАЦІЇ, ТРАНСФОРМАЦІЯ ДАНИХ

У роботі виконано огляд процесу інтелектуального аналізу даних, його основних методів та підходів

Також було досліджено методи покращення моделювання та проаналізовано їх застосування. На підставі проведеного аналізу було побудовано удосконалену модель оцінки реалізації інформаційної системи (ІС).

В ході дослідження отримані такі результати: визначено стан розвитку сфери інтелектуального аналізу даних та дослідженні розв'язані задачі; проаналізовані основні методи побудування прогнозних моделей та засоби їх вдосконалення; побудовано основні моделі класифікації для вирішення задачі оцінки реалізації ІС; застосовано різноманітні методи поліпшення результатів роботи побудованих моделей класифікації; розроблено зручний інтерфейс користувача для використання поліпшеної прогнозної моделі.

ABSTRACT

Explanatory Note to master certification work contains 86 pages, 4 sections, 34 pictures, 4 tables, 31 sources.

CLASSIFICATION METHODS, ENSEMBLES OF CLASSIFIERS, INFORMATION SYSTEMS, VISUALIZATION, TRANSFORMATION, NAMED ENTITY RECOGNITION

The paper gives review of the data mining process, common methods and approaches. Methods of increasing the quality of modeling also were investigated. On the basis of the analysis, an improved model of assessing realization of the information system was made.

The study obtained the following results: identified the state of data mining development; explored solved data mining problems; analyzed main methods for building predictive models and ways for improving quality of the model; built common classification models for solving the problem of the information system realization assessment; applied different methods for increasing the quality of the model prediction; developed convenient graphical user interface for using built predictive model.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УСЛОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ ТА ТЕРМІНІВ

ІАД – Інтелектуальний аналіз даних;

ІС – Інформаційна система.

ВСТУП

На сьогоднішній день інтелектуальний аналіз даних займає дуже важливу роль в житті сучасного суспільства. За допомогою різноманітних методів інтелектуального аналізу даних (ІАД) з'являється можливість оптимізувати, покращити різноманітні сфери бізнесу. З використанням методів ІАД реалізується багато задач, пов'язаних з медициною, освітою, наукою тощо.

Не менш поширеним та важливим у сучасному суспільстві є інформаційні системи (ІС). Кожна організація має користь від їх впровадження. Інформаційні системи пропонують можливість організаціям будь-якої форми зберігати ключову інформацію, а потім використовувати її інформацію для впливу на важливі рішення. Інформаційні системи допомагають бізнесу в розробці більшої кількості систем доданої вартості в компанії. Щоб отримати максимальну вигоду від інформаційної системи треба використати всі її можливості. Реалізація ІС – невід'ємний процес у створені інформаційної системи. Багато компаній займаються реалізаціями ІС, але іноді витрачають на це багато фінансових ресурсів та часу. Саме через це, важливо оцінювати та розуміти результат реалізації ІС ще на початку робіт. Використовуючи систему оцінки реалізації ІС можна своєчасно дізнатися про можливі ризики щодо реалізації ІС та вплинути на подальший результат, саме це є актуальністю виконання роботи.

Тому важливою задачею є можливість оцінити успішність реалізації інформаційних систем, що дозволить багатьом компаніям заощадити багато часу та коштів. У сучасних ІТ-компаніях дану проблему вирішують за допомогою методів інтелектуального аналізу даних.

Однак по теперішній час, ця задача не розв'язана методами ІАД достатньою мірою. Існує безліч методів інтелектуального аналізу даних, які допоможуть у вирішенні поставленої задачі.

Для того, щоб процес інтелектуального аналізу даних проводився вдало, слід мати набір даних, на якому застосовуються різні методи та моделі цієї області. Також багато методів ІАД можуть бути модифіковані та удосконалені після проведення детального аналізу та дослідження можливої проблеми отриманого дослідження. Саме тому метою даної магістерської роботи є дослідження методів інтелектуального аналізу даних наборів даних щодо процесу реалізації ІС для оцінки можливості розробки та реалізації інформаційної системи при обмеженнях на строки виконання робіт.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Структуризація процесу оцінки можливості реалізації ІС

Інформаційна система (ІС) - система, призначена для зберігання, пошуку та обробки інформації, і відповідні організаційні ресурси, які забезпечують і поширюють інформацію. Інформаційна система використовується у багатьох сферах життя та застосовується для забезпечення користувачів різних систем необхідною інформацією. ІС є сукупністю методів и засобів інформаційних технологій.

Інформаційні технології є основою перебудови бізнес-процесів. Інформаційна система впливає на функціонування процесів і при правильному використанні призводить до багаторазового підвищення їх результативності. Структура інформаційної системи наведена на рисунку 1.1.

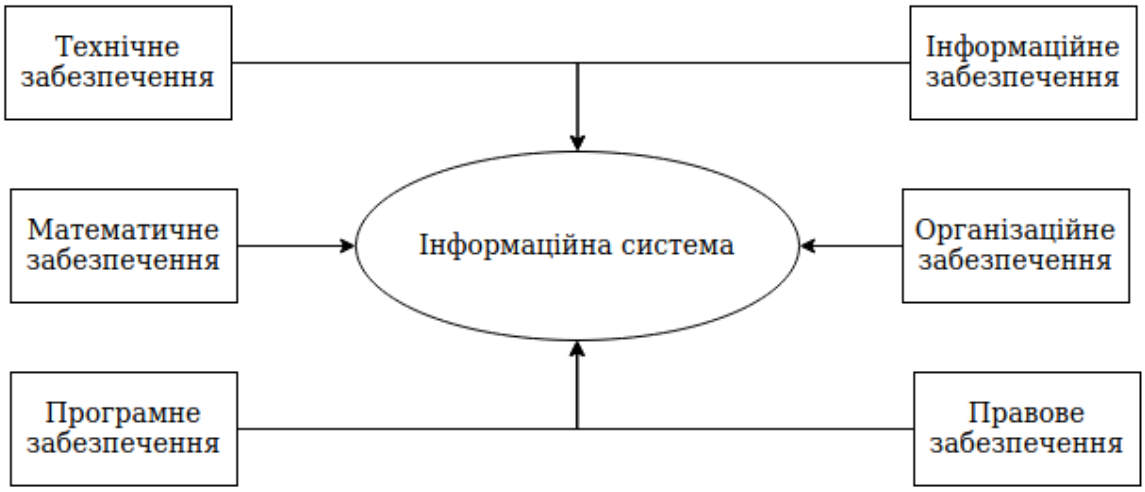


Рисунок 1.1 – Структура інформаційної системи

Створення інформаційної системи базується на шести основних стадіях: збір та аналіз вимог, логічне та фізичне проектування, реалізація та тестування (рисунок 1.2).

Рисунок 1.2 – Процес створення ІС

Процес, наведений на рисунку 1.2 починається з інформації, яка ініціює створення ІС. Першим етапом створення ІС є «Визначення вимог правовласників» в процесі якого визначаються основні характеристики та умови функціональних можливостей системи, результати відстеження зав’язків між вимогами правовласників та їх потребами. Саме таке інформація ініціює процес «Аналізу вимог». Вимоги до ІС є основною вхідною інформацією для проектування архітектури ІС, на основі якої виконуються роботи по розробці забезпечення систем.

На сьогоднішній день під вимогою розуміється:

– умова або можливість, необхідні користувачеві для вирішення проблеми або досягнення мети;

– умова або можливість, якими повинна володіти система або компонент системи, відповідні договору, стандарту, специфікації або іншого офіційного документа;

– документоване уявлення умови або можливості подібно описаного в перших двох визначеннях.

На виході стадії «Аналіз вимог» отримуються характеристики, властивості та функціональні вимоги та технічних рішень, а також деякі обмеження, які впливають на архітектурне проектування. Саме тому запропонований нижче процес оцінки можливості реалізації інформаційної системи повинен позитивно вплинути на подальший її розвиток (рисунок 1.3).

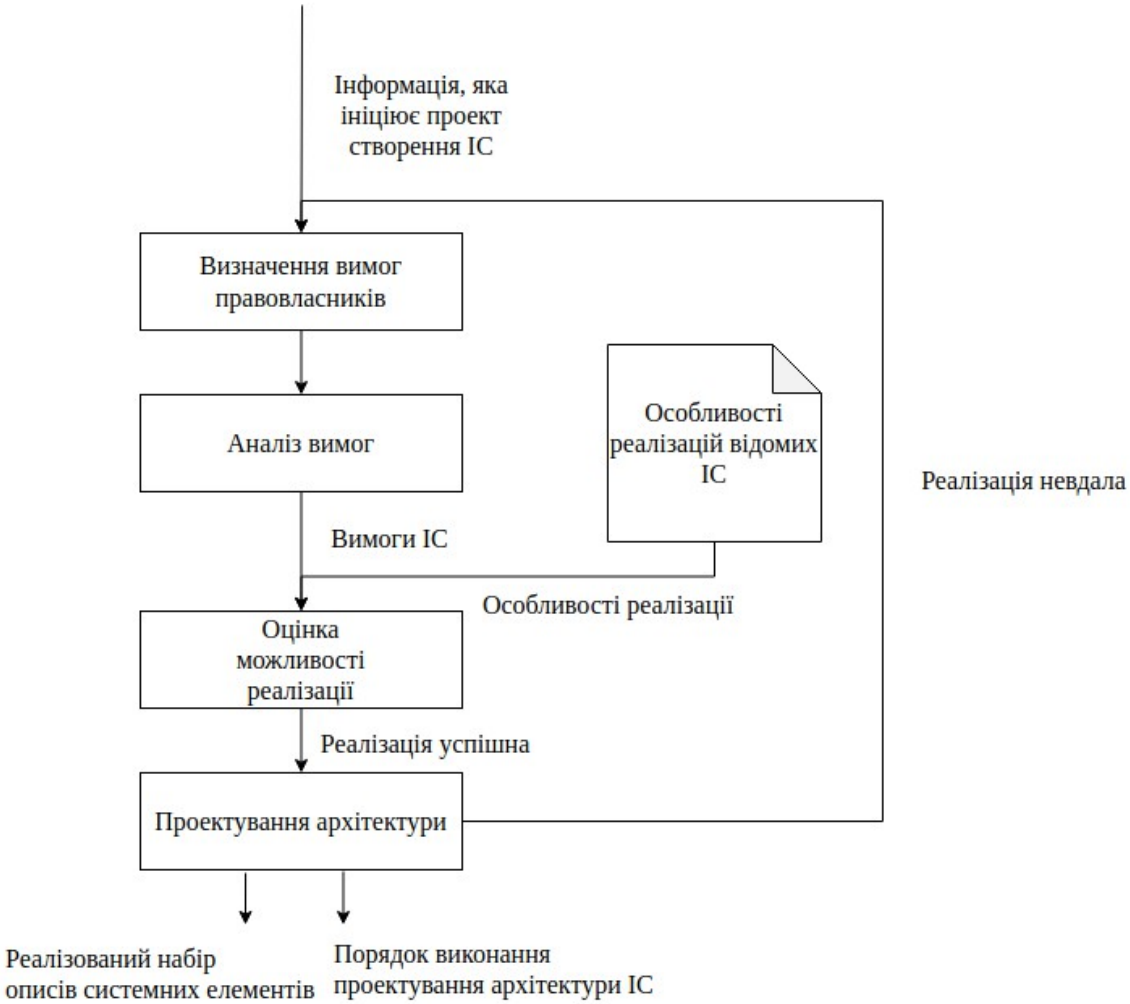


Рисунок 1.3 – Процес оцінки можливості реалізації ІС

Наведений на рисунку. 1.3. процес оцінки можливості реалізації, який використовує методи інтелектуального аналізу даних дозволяє заздалегідь отримати інформацію щодо можливих проблем з реалізацією ІС. Якщо на даному етапі буде визначено, що реалізація ІС може бути невдалою, то це

надає можливість переглянути деякі характеристики та особливості ІС. Даний етап відбувається з використанням інформації минулих років щодо особливостей реалізації відомих ІС. При отриманні успішної прогнозованої оцінки щодо можливості реалізації, відбувається етап «Проектування архітектури», виходом якого є реалізований набір описів системних елементів та порядок виконання архітектури ІС.

1.2 Дослідження методів інтелектуального аналізу даних

Інтелектуальний аналіз даних – це процес обробки початкових даних з метою вилучення нової корисної інформації. За допомогою методів інтелектуального аналізу даних можна знайти необхідні та корисні закономірності у значних об’ємах даних, які мають дуже важливе значення для подальшого їх аналізу та обробки.

Використовуючи різноманітні методи аналізу даних для пошуку шаблонів у наборах даних, підприємства можуть дізнатися більше про своїх клієнтів, що дасть їм змогу розробити ефективні маркетингові стратегії, збільшити продажі та зменшити витрати. ІАД дає можливість компаніям оптимізувати майбутнє, розуміючи минуле і сьогодення та роблячи точні прогнози щодо того, що, ймовірно, станеться далі. Також ІАД може сказати вам, які потенційні клієнти можуть стати прибутковими клієнтами на основі минулих профілів клієнтів, а які, найімовірніше, відгукнуться на конкретну пропозицію. Завдяки цим знанням можна збільшити рентабельність інвестицій, роблячи свою пропозицію лише тим потенційним клієнтам, які можуть відповісти і стати цінними клієнтами. Отже, ІАД можна використовувати для вирішення практично будь-якої бізнес-проблеми, яка включає дані, зокрема:

- збільшення доходу;
- розуміння сегментів та уподобань клієнтів;
- залучення нових клієнтів;

– покращення перехресних продажів та перепродажів;

– утримання клієнтів та підвищення лояльності;

– збільшення рентабельності інвестицій від маркетингових кампаній;

– виявлення кредитних ризиків;

– моніторинг експлуатаційних показників.

Оскільки широкомасштабні технології обробки даних, такі як машинне навчання та штучний інтелект, стають більш доступними, компанії тепер можуть перебирати терабайти даних за хвилини чи години, а не за дні чи тижні, допомагаючи їм впроваджувати інновації та швидше зростати.

Процес ІАД поділяють на кілька послідовних етапів, які зображені на рисунку 1.4.

Кожен наступний етап процесу ІАД залежить від результату попереднього, з цього можна зробити висновок про важливість розуміння мети та змісту кожного з етапів. Процес ІАД – це циклічний процес, тобто кожен з етапів може повторюватися по декілька разів, доки не буде отримано бажаного результату.

Процес ІАД стає неможливим без наявності даних. Дані – це факти та цифри, що передають щось конкретне, але які ніяк не організовані і не надають додаткової інформації щодо закономірностей, контексту. Для використання певних даних в процесі ІАД, треба щоб вони задовольняли вимогам початкової мети. Процес ІАД містить в собі п'ять основних етапів: визначення мети, аналіз даних, підготування даних, обробка даних та аналіз отриманого результату.

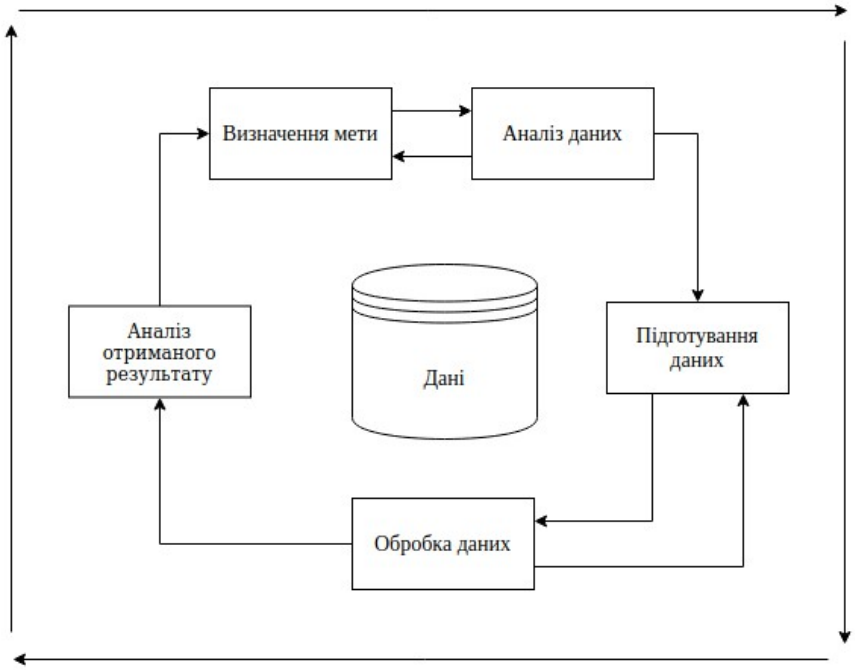


Рисунок 1.4 – Основні етапи процесу ІАД

Визначення мети – це етап на якому відбувається визначення ключових змінних, які аналіз повинен передбачити. Ці змінні позначаються як цілі моделі, і використовуються пов’язані з ними метрики для визначення успіху проекту, що розробляється. Мета проекту визначається з використанням уточнюючих запитань, які є актуальними, конкретними та однозначними. Також на цьому етапі визначаються показники успіху, які потім будуть використовуватися для аналізу отриманого результату.

На етапі аналізу даних відбувається ознайомлення з отриманими даними, виявлення неявних закономірностей, перевірка якості даних та можливість використання їх для вирішення поставленої мети. Цей етап є дуже важливим тому що при проведенні неповного аналізу можна зовсім по іншому зрозуміти мету проекту або почати наступний етап у помилковому напрямку. Саме тому етапи аналізу даних та визначення мети залежить один від одного, тобто визначення мети буде відбуватися доки поставлена мета не буде узгоджена на наступному етапі.

На етапі підготування даних виконуються кілька послідовних дій, які у майбутньому допоможуть отримати кращий результат. Зазвичай цей етап необхідний для того, щоб перетворити початкові дані у формат, відповідний для вирішення поставленої мети. Для цієї задачі використовуються різні методи роботи з даними, головною метою яких є трансформація.

Після етапу підготування даних виконується обробка даних, яка полягає у застосуванні різних методів моделювання до перетворених даних. Методи моделювання обираються згідно з зазначеною метою. До відповідного методу моделювання обирається метрика оцінки якості результату. Використовуючи обрану метрику, можна аналізувати та вдосконалювати відповідний метод моделювання.

Аналіз отриманого результату – це етап, на якому зіставляється очікуваний та отриманий результати. У разі розбіжності двох результатів, весь ІАД процес відбувається ще раз з початку. Цикл ІАД процесу зупиняється при досягненні узгодженості двох результатів.

Якість процесу інтелектуального аналізу даних залежить від наступних факторів:

– Правильного збору даних. Дані повинні повністю відповідати зазначеної мети та містити усі необхідну інформацію задля її досягнення. Дані не повинні містити пропущених значень або значень, які сильно відрізняються від інших. При досягненні усіх перелічених вимог, вірогідність отримання гарного результату стає вище.

– Надійного зберігання. З плином часу даних стає ще більше, отже об'єм займаної пам'яті зростає. Тому дуже важливо, щоб дані зберігалися без перебоїв та рівень надійності місця зберігання був високий.

– Якісної комп'ютерної обробки. Основною метою процесу ІАД є узгодження очікуваного і отриманого результату, задля досягнення цієї мети необхідно добре підготувати та обробити отримані дані. При недосягненні бажаного результату, треба вдосконалювати обробку всіма можливими методами.

ІАД містить у собі багато різних методів, які використовуються для різних потреб та сфер життя. Ці методи можна поділити на кілька узагальнених категорій, наведених на рисунку 1.5.

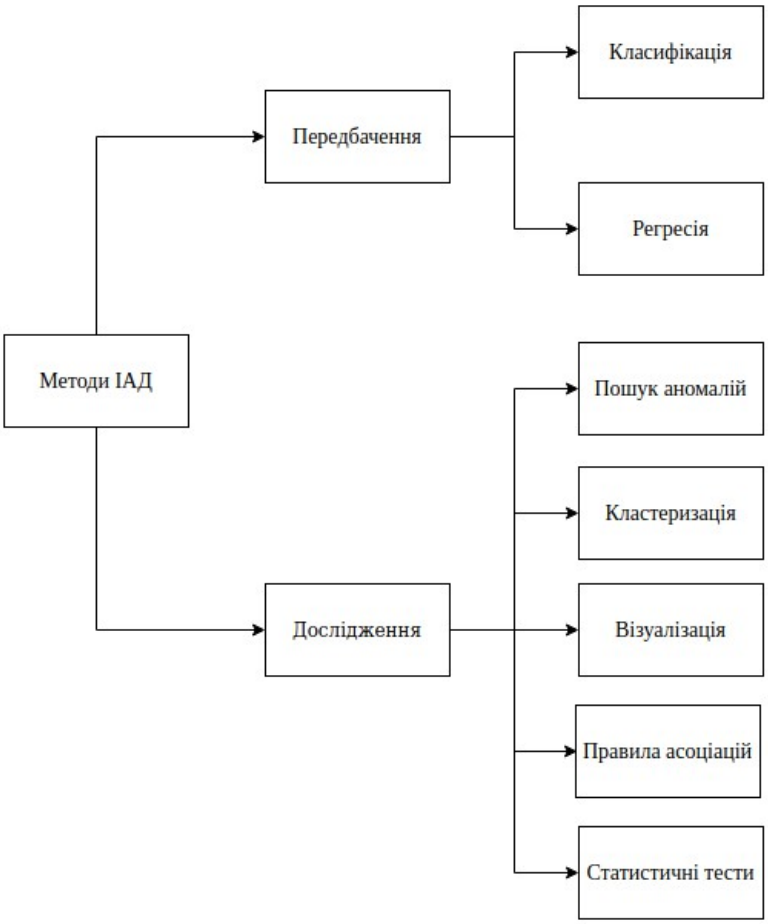


Рисунок. 1.5 – Методи ІАД

Загалом методи ІАД поділяють на дві основні категорії: передбачення та дослідження. До першої категорії належать методи після використання яких можна отримати прогноз на майбутнє, маючи попередні дані. Ці методи використовують для побудови різних типів моделей прогнозування. До категорії дослідження відносяться методи, які допомагають отримати нову, додаткову інформацію про існуючі дані, що неможливо при застосуванні усіх інших методів.

Класифікація в машинному навчанні, статистиці та ІАД – це підхід, при якому комп’ютерна програма вивчає дані, що їй передаються, та робить нові спостереження чи класифікації, використовуючи ці дані. Це процес класифікації даного набору даних за класами, він може виконуватися як на структурованих, так і на неструктурованих даних. Процес починається з визначення класу вхідних даних. Класи часто називають цільовими, мітками або категоріями. Класифікаційне моделювання – наближення функції відображення від вхідних змінних до дискретних вихідних змінних. Головною метою класифікації є визначення класу або категорії нових даних.

Поширеним прикладом класифікації є виявлення спаму. Для того, щоб написати програму для фільтрації електронної пошти зі спамом, можна навчити алгоритм машинного навчання із набором електронних листів, подібних до спаму, позначених як спам, та звичайних листів, позначених як не спам. Ідея полягає в тому, щоб створити алгоритм, який знайде характеристики спам-повідомлень з цього навчального набору, та потім виявити спам-повідомлення з усього набору електронних листів. Класифікація є важливим інструментом у сучасному світі, де великі дані використовуються для прийняття різного роду рішень в уряді, економіці, медицині тощо. Дослідники мають доступ до величезних обсягів даних, і класифікація є одним із інструментів, який допомагає їм зрозуміти дані та знайти закономірності. Хоча класифікація іноді вимагає використання складних алгоритмів, класифікація – це те, що люди роблять природним шляхом щодня – це просто групування речей за подібними ознаками та ознаками

Регресійний аналіз – це форма прогнозного моделювання, яка досліджує взаємозв'язок між залежною (цільовою) та незалежною змінними. Ця методика використовується для прогнозування, моделювання часових рядів та пошуку причинно-наслідкового зв’язку між змінними. Наприклад, взаємозв'язок між несвідомою їздою та кількістю дорожньо-транспортних

пригод водієм найкраще вивчити шляхом регресії. Регресійний аналіз є важливим інструментом для моделювання та аналізу даних.

Використання регресійного аналізу має наступні переваги:

– вказує на значні взаємозв'язки між залежною змінною та незалежною змінною;

– вказує на силу впливу декількох незалежних змінних на залежну змінну.

Регресійний аналіз також дозволяє порівняти вплив зовсім різних змінних, таких як зміна ціни та кількість рекламних заходів. Ці переваги допомагають дослідникам ринку, аналітикам даних, науковцям обрати найкращий набір змінних, які будуть використовуватися для побудови прогнозних моделей.

Пошук аномалій – це метод у ІАД, який визначає точки даних, події або спостереження, що відхиляються від звичайної поведінки набору даних. Аномальні дані можуть свідчити про критичні випадки, такі як технічний збій або потенційні можливості, наприклад, про зміну поведінки споживачів.

Метод правила асоціацій знаходить цікаві асоціації та взаємозв'язки між великими наборами елементів даних. Це правило показує, як часто набір елементів трапляється в транзакції. Типовим прикладом є аналіз ринку. Аналіз ринку – це один з ключових методів, який використовують відносини для демонстрації асоціацій між товарами. Враховуючи набір транзакцій, можна знайти правила, які передбачатимуть появу предмета на основі випадків інших елементів у транзакції.

Статистичні тести допомагають проаналізувати важливість вхідних ознак. Використовуючи різні типи тестів, можна отримати рівень значущості (альфа значення) для кожної ознаки. Проаналізувавши отримане альфа значення, можна зробити висновок щодо важливості відповідної ознаки та відредагувати початковий перелік ознак. Ознаки з низьким рівнем значущості краще видалити з початкових даних, тому що краще мати менше ознак, але дійсно важливих, ніж багато беззмістовних.

1.3 Дослідження підходів до реалізації процесу інтелектуального аналізу даних

Отримавши дані у своєму початковому вигляді, їх треба проаналізувати та зробити висновок щодо вигляду вихідних даних. На цьому етапі перш за все треба визначити, які з ознак є важливими та добре впливають на результат, та які з ознак не мають жодного сенсу та після видалення яких кінцевий результат ніяк не зміниться. Також треба проаналізувати, які з ознак мають багато аномалій в даних, що може дуже погано позначитися на результаті. Саме тому іноді ознаки з аномаліями краще видалити с усього набору даних або трансформувати у будь-який відомий спосіб. Процес аналізу даних за допомогою методів ІАД схематично зображено на рисунку 1.6.

Основні методи ІАД, які використовуються на етапі аналізу даних – це візуалізація та статистичні тести.

З використанням візуалізації можна краще зрозуміти дані, які надійшли, ознайомитися з переліком ознак, дослідити значення та особливості кожної з ознак. Із застосуванням різних типів графіків та методів побудова діаграм, можна дослідити які проблеми має та чи інша ознака. Після вживання методів візуалізації, користувач отримує набір графіків та діаграм, аналізуючи які можна знати аномалії в даних та відібрати значущі ознаки. Ознаки можна відібрати, використовуючи інший метод – статистичні тести. Після використання цього метода користувач отримує рівень значущості для кожної з ознак, аналізуючи які обирає для себе перелік вихідних ознак.

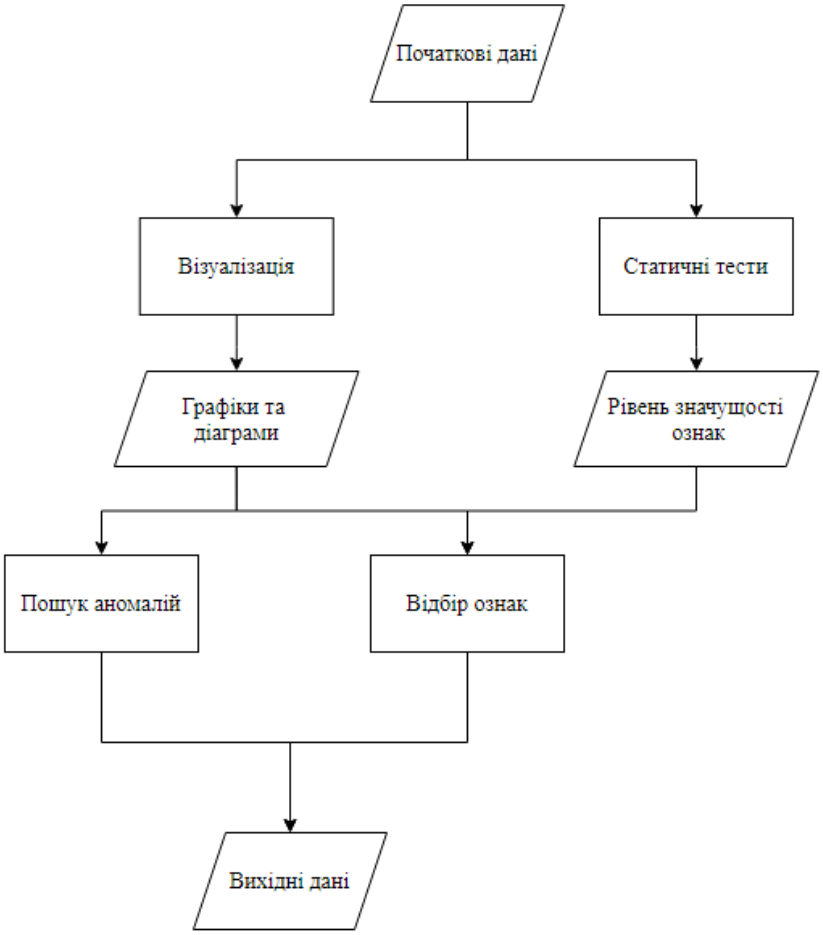


Рисунок 1.6 – Процес аналізу даних з використанням методів ІАД

Після отримання оновленого набору даних, необхідно зробити його обробку, результат якої повинен узгоджуватися з первинною метою. Користувач повинен визначитися з поставленим завданням: регресії або класифікації. А потім діяти згідно за схемою на рисунку 1.7, на якому зображено процес обробки даних з використанням методів ІАД.

Після визначення вхідної задачі, відбувається процес побудови відповідної моделі. На виході користувач отримує регресійну або класифікаційну моделі. Побудована модель аналізується згідно з поставленою метою. У випадку, коли поставлена мета не узгоджується з побудованою моделлю – процес побудови повторюється. Цикл побудови моделі зупиняється, коли отриманий результат стає узгодженим с первинною

метою. Тоді користувач отримує на виході побудовану модель, яка може прогнозувати, визначати клас нових даних.

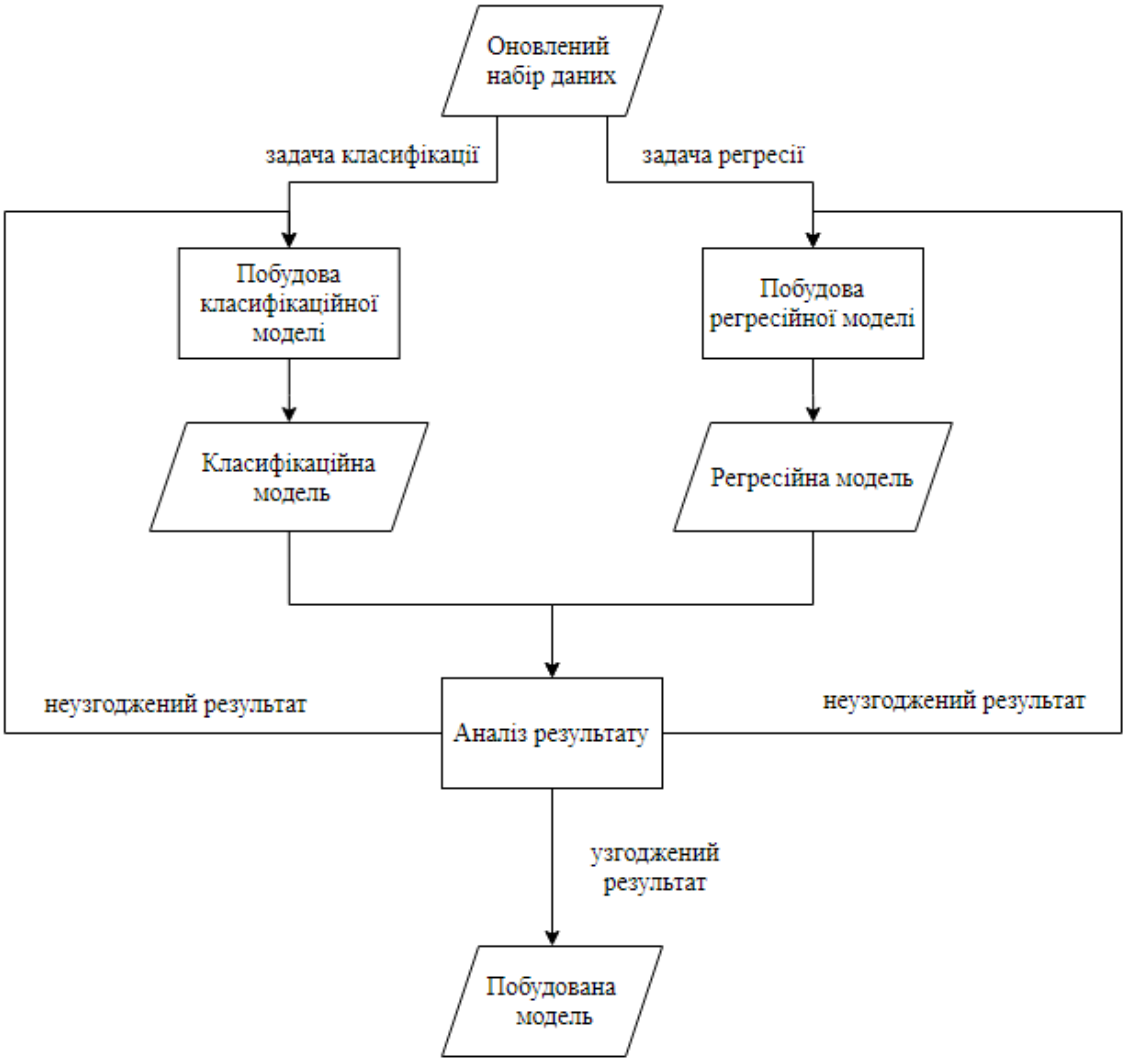


Рисунок 1.7 – Процес обробки даних з використанням методів ІАД

В окремих випадках, задачу класифікації можна вирішувати за допомоги регресійної моделі. Однак регресійна модель на виході прогнозує безперервне число, яке треба округляти, щоб отримати дискретне значення (клас об'єкту). В загалом за погано впливає на результат, але іноді регресійна модель показує кращий результат, ніж класифікаційна модель.

1.4 Постановка задачі

З усього вищезазначеного можна визначити об’єкт, предмет та мету дослідження.

Об’єктом дослідження є процес оцінки можливості реалізації інформаційної системи.

Предметом дослідження є методи інтелектуального аналізу даних в задачах оцінки можливості реалізації інформаційної системи.

Метою роботи є дослідження методів інтелектуального аналізу даних наборів даних щодо процесу реалізації ІС для оцінки можливості розробки та реалізації інформаційної системи при обмеженнях на строки виконання робіт про проблему, об'єкт та предмет дослідження

З усього вищезазначеного можна зробити висновок, що для вирішення поставленої задачі необхідно:

– дослідити процес створення ІС та структурувати процес оцінки можливості ІС;

– проаналізувати область та методи інтелектуального аналізу даних;

– визначити особливості застосування досліджених методів ІАД;

– проаналізувати підходи до підвищення ефективності методу оцінювання можливості реалізації ІС;

– запропонувати удосконалений метод оцінювання можливості реалізації ІС;

– визначити технологію оцінювання можливості реалізації ІС;

– практично реалізувати запропонований удосконалений підхід;

– провести експериментальну перевірку, отриманого удосконаленого методу.

2. ДОСЛІДЖЕННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ ОЦІНКИ МОЖЛИВОСТІ РЕАЛІЗАЦІЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ

2.1 Трансформація даних для оцінювання можливості реалізації ІС методами інтелектуального аналізу даних

Вхідні дані можуть бути наведені у будь-якому вигляді та містити у собі різні формати представлення даних. Моделі класифікації та регресії можна побудувати маючи тільки чисельний набір даних, але реальні дані переважно мають різні типи ознак: категоріальні значення, чисельні значення, текстові повідомлення. Саме тому важливо вміти перетворити вхідний набір даних до вигляду, прийнятного для побудови моделі.

Якщо ознака в наборі дані має чисельний тип даних, моделі можна побудувати, використовуючи первинне значення ознаки. Задля покращення результатів роботи моделі, можна застосовувати наступні ІАД методи: масштабування даних (data scaling) та перетворення даних (data transformation). Саме це допоможе змінити формату, структури або значень даних.

Перелічені методи застосовують для того, щоб:

- зробити дані краще організованими, простішими як для людей, так і для комп’ютерів;
- покращити якість даних та захистити програми від потенційних проблем, таких як нульові значення, несподівані дублікати, неправильна індексація та несумісні формати;
- полегшити сумісність між програмами, системами та типами даних: дані, що використовуються для багатьох цілей, можуть потребувати перетворення різними способами.

Якщо ознака має категоріальний тип даних, то можна застосувати декілька методів ІАД для перетворення її до чисельного вигляду:

– призначення відповідних номерів для кожної з категорій. Цей метод використовуються у тих випадках, коли кількість унікальних значень категоріальної змінної невелика. Наприклад, ознака містить інформації про колір та має такі значення: червоний, жовтий та зелений. У цьому випадку кількість унікальних значень ознаки дорівнює трьом. Зазначену ознаку можна перетворити на чисельну ознаку зі значеннями: 1, 2 та 3;

– одноразове кодування категоріальних ознак (one-hot encoding). У цьому методі видаляється первісна змінна та замість неї додається одна нова двійкова змінна для кожного унікального цілого значення змінної. У прикладі вище є три категорії, а отже, потрібні три бінарні змінні. Значення 1 розміщується у двійковій змінній відповідного кольору, а 0 – для інших кольорів;

– фіктивне копіювання змінних (dummy variable encoding). Одноразове кодування створює одну двійкову змінну для кожної категорії. Проблема полягає в тому, що це подання включає надмірність. Наприклад, якщо відомо, що [1, 0, 0] являє собою синій колір, а [0, 1, 0] означає зелений, то не обов'язково створювати двійкова змінну для червоного кольору. Замість цього можна використовувати [0, 0]. Фіктивне кодування змінних завжди представляє N категорій з двійковими змінними N-1.

Якщо ознака має текстовий тип даних, то можна застосувати декілька методів, які допоможуть вилучити додаткову інформацію з тексту. Отриману інформацію надалі використовувати, як новоутворені ознаки.

Методи, які допоможуть у вилученні ознак з текстових даних наступні:

– збір статистики тексту (кількість літер, слів, слів з великої літери, слів англійською мовою та інші). Цю інформацію зібрати дуже легко, однак іноді вона може являти собою дуже корисні факти. Наприклад, інформаційні системи, які у базі даних представлені англійською мовою, можливо, мають кращі показники реалізуємості. У цьому випадку ознака «кількість слів англійською мовою» допоможе у побудові моделі.

– визнання іменованої сутності (named entity recognition) – це завдання виявлення та категоризації ключової інформації (сутностей) у тексті. Суб'єктом може бути будь-яке слово або серія слів, які послідовно посилаються на одне і те ж. Кожна виявлена сутність класифікується на заздалегідь визначену категорію, наприклад: людина, організація, час, розташування та інші.

– аналіз настрою (sentiment analysis) – це інтерпретація та класифікація емоцій (позитивних, негативних та нейтральних) у текстових даних за допомогою методів текстового аналізу. Інструменти аналізу настрою дозволяють за допомогою ІАД визначати настрої користувачів, які писали відповідний текст.

2.2 Методи оцінювання процесу реалізації ІС засобами інтелектуального аналізу даних

Аналіз процесу реалізації ІС – це задача класифікації, метою якої є передбачення реалізується відповідна система або ні. Саме тому після того, як дані були підготовлені до обробки, слід почати будувати моделі класифікації. На рисунку 2.1 представлені найбільш популярні моделі класифікації.

Метод опорних векторів (Support Vector Machine, SVM) – це лінійний метод для задач класифікації та регресії. Він може вирішувати лінійні та нелінійні задачі та добре працювати для багатьох практичних задачах. Ідея SVM полягає у знаходженні лінії або гіперплощини, яка може поділити класи між собою. Цей метод виконує мінімізацію структурних ризиків, що покращує складність класифікатора з метою досягнення чудових показників при роботі.

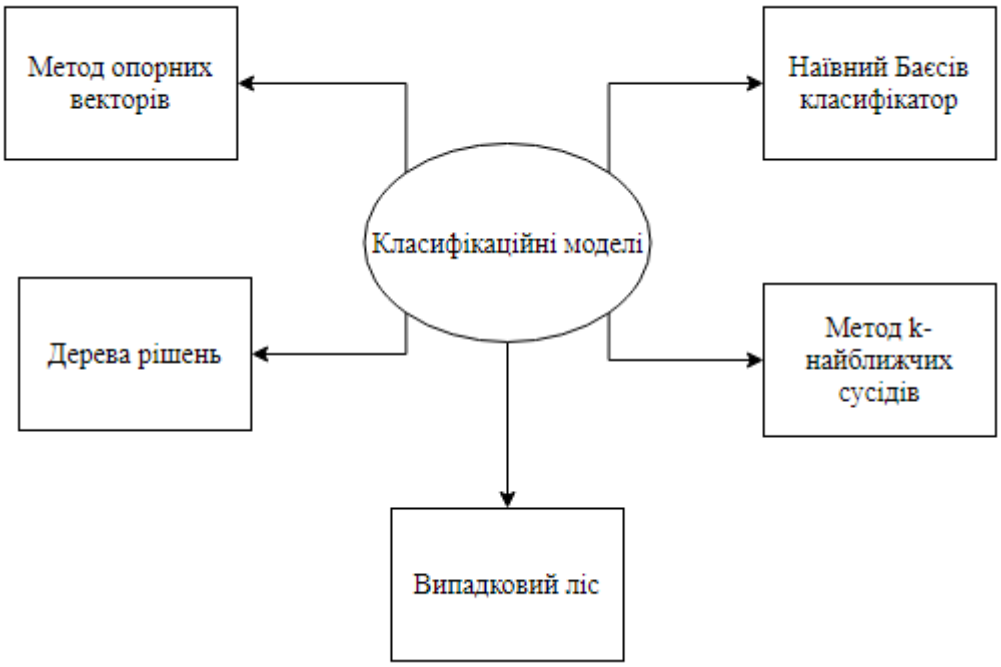


Рисунок 2.1 – Моделі класифікації

Наївний Байєсів алгоритм (Naive Bayes) – це імовірнісний алгоритм машинного навчання, заснований на теоремі Байєса, що використовується в широкому спектрі завдань класифікації. Основне припущення Наїва Байєса полягає в тому, що кожна особливість робить незалежний та рівний внесок у результат. Наївні алгоритми Байєса часто використовуються в аналізі настроїв, фільтрації спаму, системах рекомендацій тощо. Вони швидко та легко впроваджуються, але найбільшим їх недоліком є вимога до незалежності предикторів.

Алгоритм k-найближчих сусідів (k-nearest neighbors algorithm, KNN) – це простий у реалізації керований алгоритм машинного навчання, який може бути використаний для вирішення як задач класифікації, так і регресії. Алгоритм KNN передбачає, що подібні речі існують у безпосередній близькості. Основним недоліком KNN є дуже низька швидкість обробки із збільшенням обсягу даних. В загалом прогнозування треба робити швидко, саме тому KNN не є популярним методом.

Дерево рішень (Decision tree) є найпотужнішим та найпопулярнішим інструментом класифікації та прогнозування. Дерево рішень – це блок-схема, схожа на деревоподібну структуру, де кожен внутрішній вузол позначає тест на атрибут, кожна гілка являє собою результат тесту, а кожен листовий вузол (кінцевий вузол) містить мітку класу. Дерева рішень здатні генерувати зрозумілі правила, дають чітке розуміння того, які поля є найбільш важливими для прогнозування або класифікації. З іншого боку, дерева рішень схильні до помилок у проблемах класифікації з багатьма класами та відносно невеликою кількістю прикладів навчання.

Випадковий ліс (Random Forest) – це гнучкий, простий у використанні алгоритм машинного навчання, який, навіть без налаштування гіперпараметрів, більшу в загалом показує чудові результати. Це також один з найбільш часто використовуваних алгоритмів, завдяки своїй простоті та різноманітності. Також його можна використовувати як для класифікації, так і для регресії. Основне обмеження випадкового лісу полягає в тому, що велика кількість дерев може зробити алгоритм занадто повільним та неефективним для прогнозів у реальному часі. Для більш точного прогнозування потрібно більше дерев, що призводить до уповільнення моделі. У теперішній реалізації методу оцінки реалізації інформаційної системи використовується саме ця класифікаційна модель.

2.3 Підходи до підвищення ефективності процесу оцінювання можливості реалізації ІС

Підвищення продуктивності моделі часом може бути складним завданням. Іноді навіть, застосовуючи різні техніки для покращення результату, кращий результат отримати не вдається. Однак, можна виділити декілька технік, які можуть добре вплинути на кінцевий результат: розширити первинний набір даних, змінити спосіб обробки проблемних даних, застосувати методи трансформації ознак, проаналізувати перелік

ознак, побудувати іншу модель, застосувати перехресну перевірку або використати ансамблі класифікаторів.

Розширити первинний набір даних. В деяких випадках побудована модель може показувати поганий результат через те, що набір даних маленький та моделі було складно знайти якісь закономірності для того, щоб навчитися прогнозувати нові дані. Саме тому в деяких випадках збільшення набору даних призводить до кращої роботи моделі. Також, якщо розширити набір даних неможливо, то можна замінити поточний набір іншим, але більшим.

Змінити спосіб обробки проблемних даних (пропущені значення, дублікати, аномальні значення тощо). Існує декілька способів обробки проблемних даних: заміна даних середнім значенням, медіаною, модою, нульовими значеннями, видалення з набору даних та інші. На етапі підготування даних користувач обирає спосіб обробки та виконує його. Іноді цей спосіб може бути обраний помилково та застосування іншого методу покращить результат роботи відповідної моделі.

Застосувати методи трансформації ознак. Набір даних містить у собі перелік ознак, ці ознаки можуть бути схожі один з одним, але також можуть мати зовсім інші значення або діапазон значень. У цьому випадку застосовують методи трансформації або масштабування, які призводять усі ознаки в наборі даних до схожого між собою вигляду.

Проаналізувати перелік ознак. Це можна зробити за допомогою перелічених методів ІАД: візуалізація, статистичні тести тощо. Після застосування цих методів, користувач обирає ознаки, які краще видалити, які треба залишити. Але не завжди користувач приймає вірне рішення, саме тому після отримання результату моделювання, слід переглянути перелік ознак та спробувати побудувати модель ще раз. Цей метод може допомогти зменшити набір даних, що призведе до більш швидкої побудови моделі та кращої точності.

Спробувати побудувати іншу модель. Деякі алгоритми краще підходять для певного типу наборів даних, ніж інші. Отже, необхідно застосувати всі відповідні моделі та перевірити ефективність кожної. Доцільно обирати модель з кращим результатом та потім застосувати різні методи для її покращення.

Налаштувати відповідну модель. Алгоритми машинного навчання керуються параметрами. Ці параметри значною мірою впливають на результат навчального процесу. Завданням налаштування параметрів є пошук оптимального значення для кожного параметра для підвищення результату роботи моделі. Щоб налаштувати ці параметри, треба добре розуміти це значення та їх індивідуальний вплив на модель. Якісне налаштування параметрів призведе до кращого результату роботи моделі.

Перехресна перевірка (cross validation) – це техніка, яка передбачає використання певного зразка набору даних, на якому буде тестуватися побудована модель. Спочатку треба зарезервувати зразок набору даних, потім навчити модель на залишку. Побудовану модель можна перевірити за допомогою зразка, це допоможе вам оцінити ефективність роботи моделі.

Застосування ансамблів класифікаторів – це найпоширенішим метод. Головною ідеєю його є поєднання результатів багатьох класифікаційних моделей та обрання моделі з кращою точністю. Найбільш поширені ансамблеві методи:

– bagging. Головна ідея полягає підборі кілька незалежних моделей та усередненні їх прогнозів для того, щоб отримати модель з меншою дисперсією. Принцип роботи bagging методу зображений на рисунку 2.2.

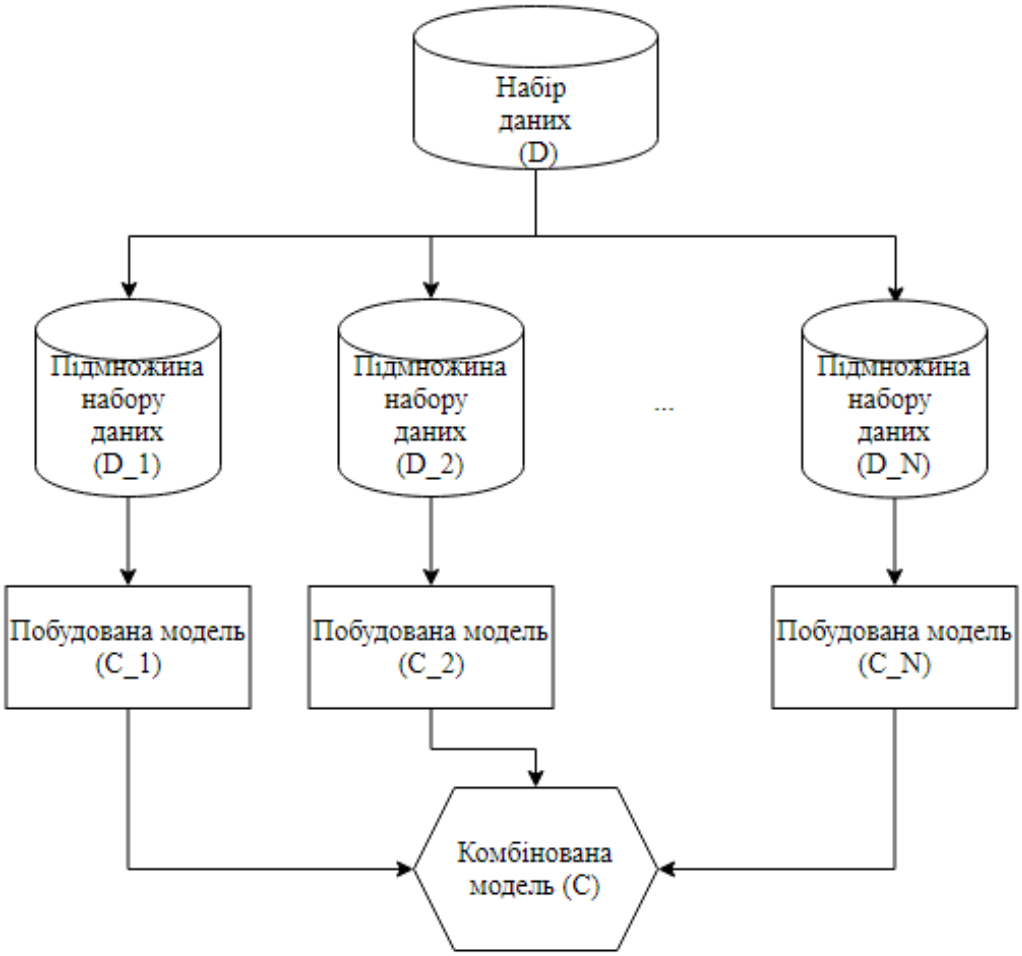


Рисунок 2.2 – Принцип роботи bagging методу

– boosting – це ітераційний прийом, який регулює вагу спостереження на основі останньої класифікації. Якщо спостереження було класифіковано неправильно, воно намагається збільшити вагу цього спостереження і навпаки. Boosting допомагає створювати потужні прогнозні моделі, однак іноді може виникнути проблема перенавчання. Принцип роботи bagging методу зображений на рисунку 2.3.

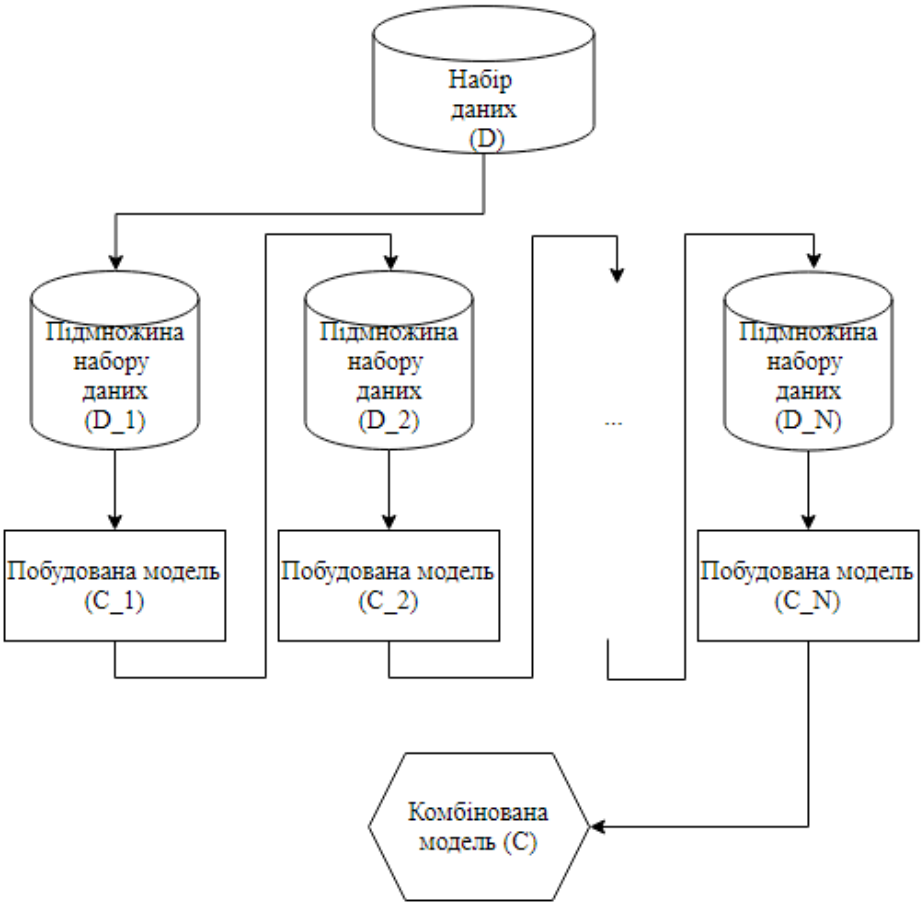


Рисунок 2.3 – Принцип роботи boosting методу

– stacking. Головна ідея складання полягає в тому, щоб навчити декількох різних класифікаторів та поєднати їх, навчаючи мета-моделі для виведення прогнозів на основі безлічі передбачень, повернутих цими класифікаторами. Наприклад, для проблеми класифікації можна обрати класифікатор KNN та SVM, і вирішити вивчити нейронну мережу як метамодель. Потім нейронна мережа прийматиме як вхідні дані результати двох класифікаторів та вчиться повертати остаточні прогнози. Принцип роботи stacking методу зображений на рисунок 2.4.

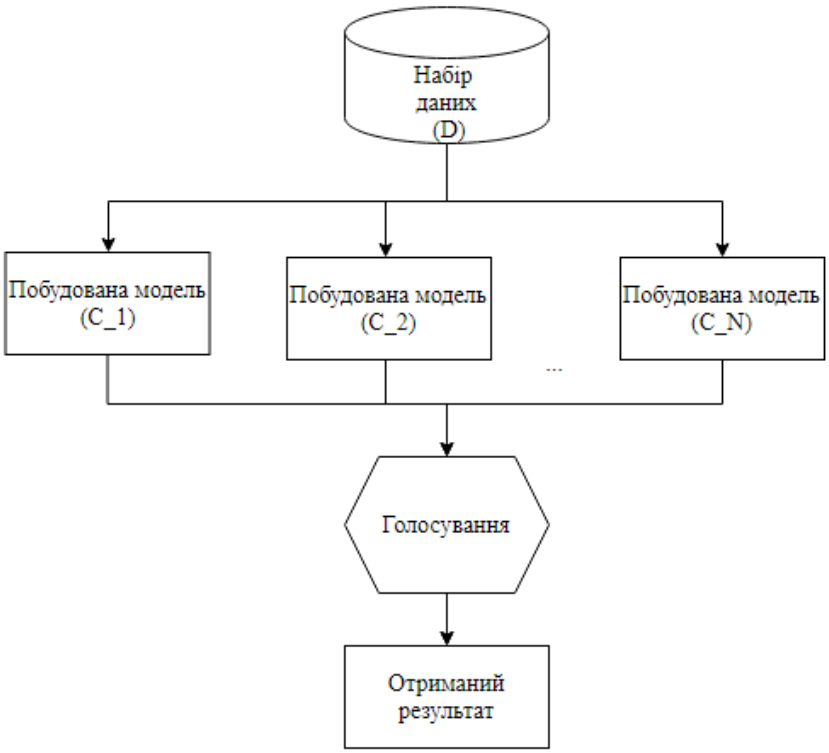


Рисунок 2.4 – Принцип роботи stacking методу

Усі перелічені вище техніки можна використовувати для підвищення результату роботи моделі. Іноді навіть застосовуючи ці техніки, поліпшити результат не вдається, але у багатьох випадках одна із технік обов'язково має допомогти у вирішення заданої проблеми.

2.4 Показники оцінювання результатів класифікації

Покращення роботи моделі розуміє під собою поліпшення показника (метрики) класифікації. В ІАД використовуються декілька різних метрик, кожна з якої має свої переваги та недоліки. Найбільш поширеною метрикою в розв'язанні задач класифікації є точність. Точність має великий недолік – це залежність від кількості екземплярів різних класів, тобто якщо розподілення різних класів не схоже (екземплярів одного класу набагато більше, ніж іншого), то точність не може дати об'єктивну оцінку результату. У цьому

випадку бажано використовувати метрику, яка добре справляється з проблемами такого роду – F1 оцінку.

Оцінка F1 поєднує в собі точність (precision) і повноту (recall) класифікатора в єдину метрику, беручи середнє гармонічне значення. Він в основному використовується для порівняння результатів роботи двох класифікаторів. Припустимо, що класифікатор А має кращу повноту, а класифікатор В – вищу точність. У цьому випадку оцінки F1 для обох класифікаторів можуть бути використані для визначення того, який з них дає кращі результати.

Оцінка F1 класифікаційної моделі обчислюється за формулою (2.1):

, (2.1)

де Precision – точність моделі,

Recall – повнота моделі.

Точність в основному позначає, що серед результатів, класифікованих моделлю як позитивні, скільки насправді було позитивними. Точність обчислюється за формулою (2.2):

, (2.2)

де True Positives – це значення, які насправді є позитивними і які модель також передбачає як позитивні,

False Positives – це значення, які насправді є негативними, але модель передбачає як позитивні.

Повнота позначає, скільки справжніх позитивних значень було знайдено побудованою моделлю, та обчислюється за формулою (2.3):

, (2.3)

де True Positives – це значення, які насправді є позитивними і які модель також передбачає як позитивні,

False Negatives – це значення, які насправді є позитивними, але модель також прогнозує негативні.

2.5 Удосконалений метод оцінювання можливості реалізації ІС

Проаналізувавши методи підвищення ефективності методів оцінювання можливості реалізації ІС, було обрано три основних етапи застосування методів ІАД, а саме: формування набору даних на основі аналізу вимог та опису ІС, перетворення набору даних з урахуванням різних масштабів значень та застосування ансамблів моделей (рисунок 2.5).

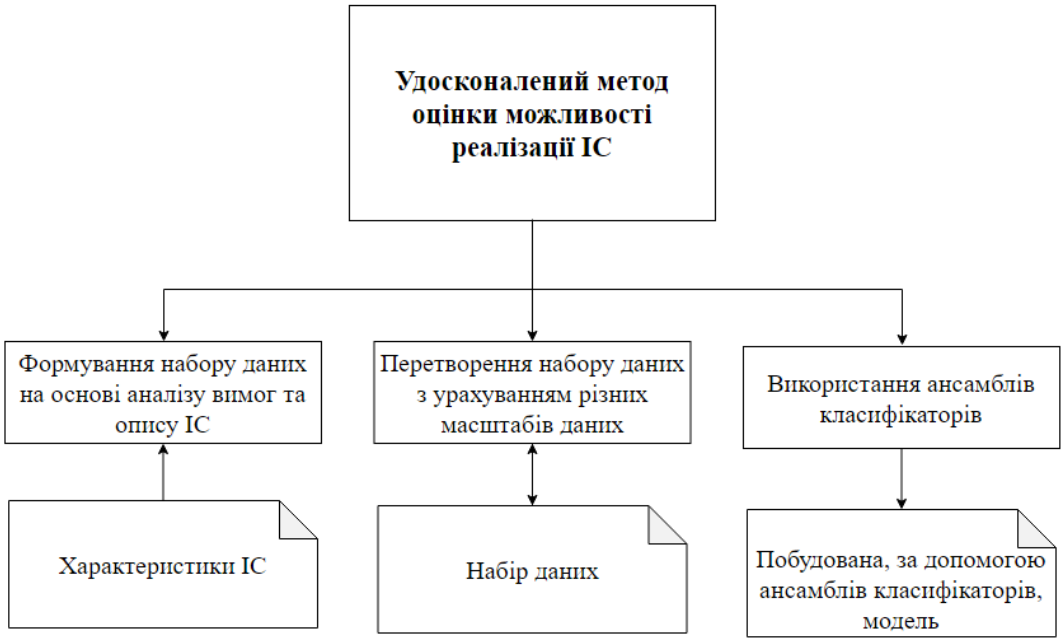


Рисунок 2.5 – Удосконалений метод оцінки можливості реалізації

Загалом запропонований метод можна поділити на шість основних етапів.

Етап 1. Доповнення набору даних на основі аналізу вимог та опису ІС з використанням методу інтелектуального аналізу даних «Визнання іменованої

сутності». Цей метод застосовується на первинного набору даних, в якому існують текстові ознаки. Для кожної характеристики ІС обчислюється кількість її входжень в певному тексті за формулою (2.4).

(2.4)

де – кількість співпадінь поточного слова в описі або вимогах ІС з характеристикою ІС;

N – кількість слів у тексті;

m – кількість характеристик ІС.

Етап 2. Перетворення первинного набору даних з урахуванням різних масштабів даних за допомогою методу інтелектуального аналізу «Трансформування Йео-Джонсона». Для кожної точки набору даних робиться логарифмічне перетворення, щоб усі дані були представлені в одному діапазон значень (2.5).

(2.5)

де – початкове значення екземпляру в наборі даних;

– отримане трансформоване значення.

Етап 3. Прогнозування можливості реалізації ІС за допомогою методів класифікації Random Forest, KNN, SVM, Decision tree. Отримання їх передбачень та оцінок .

Етап 4. Розрахунок вагових коефіцієнтів методів на основі F1 оцінок побудованих моделей. Для кожної побудованої моделі слід проаналізувати її важливість згідно до отриманих F1 оцінок. Моделі, які мають найбільшу F1 оцінку, будуть мати найбільший ваговий коефіцієнт та впливати на кінцевий

результат більше всього. Ваговий коефіцієнт обчислюється за формулою (2.6).

(2.6)

де – отримана оцінка прогнозування відповідної моделі.

Етап 5 поділяється на дві наступні фази:

– крок 5.1. Побудова ансамблів класифікаторів, який використовує отримані вагові коефіцієнти та передбачення відповідної моделі (2.7).

(2.7)

де – нормований коефіцієнт згідно до отриманих F1 оцінок,

– результат прогнозування відповідної моделі,

N – кількість моделей прогнозування.

– крок 5.2. Оцінювання можливості реалізації за допомоги ансамблів класифікаторів. Це процес, на вхід якого подаються ознаки з набору даних, а виходом якого є число 0 або 1. Число 1 позначає, що побудована модель передбачила, що реалізація ІС буде успішною. З іншого боку, 0 повідомляє про невдалий результат (2.8).

–

(2.8)

де 0 – прогноз моделі, який позначає, що реалізація ІС буде невдалою;

1 – прогноз моделі, який позначає, що реалізація ІС буде успішною.

3 ДОСЛІДЖЕННЯ ОТРИМАНИХ НАУКОВИХ РЕЗУЛЬТАТІВ

3.1 Вхідні дані для вирішення задачі оцінювання реалізації ІС

Для реалізації створеного теоретичного підходу для оцінювання можливості реалізації ІС, перш за все необхідно знайти відповідний набір даних, за допомогою якого будуть будуватися моделі класифікації. Знайдений набір даних містить у собі 3309 екземплярів даних у форматі, який представлено на рисунку 3.1.

is_id	responsible_person	deputy_person	closing_date	created_time	stage	type_1	type_2	amount	parsed_country	is_corp_mail	Potential_text	Notes_text
)005022011	Ivan Ivanov	None	2018-02-28 00:00:00	2016-07-15 18:14:00	5. Objections and offer adaptation	automatic	Information retrieval	$ 368,190.00	us	True	hi alan safely back home overcame jet lag serv...	hi alan hope well victor done investigation wo...
014825009	Ivan Ivanov	Alexander Kravchenko	2018-09-13 00:00:00	2018-05-24 18:26:00	7. Closed Won	automatic	Information Critical	$ 41,184.00	us	True	mobile app partially built jon partner needs e...	hi alan pleasure mine thank time support get b...

Рисунок 3.1 – Первинний набір даних

Первинний набір даних містить у собі інформації щодо інформаційних систем, які було реалізовано та які реалізувати не вдалося. Набір даних містить у собі таку інформацію:

- is_id – унікальний ідентифікатор інформаційної системи;
- responsible_person – людина, яка відповідала за реалізацію ІС;
- deputy_person – заступник відповідальної людини;
- closing_date та created_time містять інформацію щодо початку та кінця реалізації ІС;
- stage – стадія реалізації ІС. Якщо значення цієї ознаки дорівнює «Closed Won», то реалізація ІС пройшла успішно. При іншому значенні ІС реалізувати не вдалося;
- type_1 та type_2 – характеристики (типи) ІС;
- amount – приблизна вартість реалізації;

– parsed_country – країна замовника ІС;

– is_corp_email – наявність корпоративної електронної пошти у заявника;

– Potential_text та Notes_text – листування із замовником, опис ІС тощо.

Для застосування даного набору даних у реалізації поставленої задачі необхідно його перетворити до іншого формату. Для побудови класифікаційної моделі кожен набір даних повинен мати цільову функцію, тобто клас чи категорію об'єкту. Оцінювання можливості реалізації ІС – це теж саме, що визначення кінцевої стадії ІС, ця інформація міститься в ознаці stage. Класифікаційна модель має передбачати клас об'єкту, саме тому замінимо значення «Closed Won» на одиницю, усі інші значення на нуль.

Метою даної задачі є оцінка реалізації ІС, тобто визначення значення в ознаці stage. Клас 1 (Closed Won) буде позначати, що ІС була реалізована, в іншому випадку ознака дорівнюватиме нулю (Closed Lost). Оцінка F1 повинна бути більшою, ніж у теперішній реалізації. Зараз створена модель має оцінку 91. При значеннях оцінки вище 90, кожен процент є дуже важливим в оцінці роботи моделі.

На етапі аналізу даних необхідно проаналізувати важливість кожної з ознаки методами ІАД. Важливість ознаки amount можна оцінити за допомогою ядрової оцінки густини (KDE plot). Результат побудови графіка з використанням Python бібліотеки Matplotlib можна побачити на рисунку 3.2.

Згідно з графіком вище, можна сказати, що ознака amount є важливою, тому що має різні розподіли для кожного з класів. Також можна побачити, що ця ознака містить у собі багато нульових значень, це може погано вплинути на результат прогнозування.

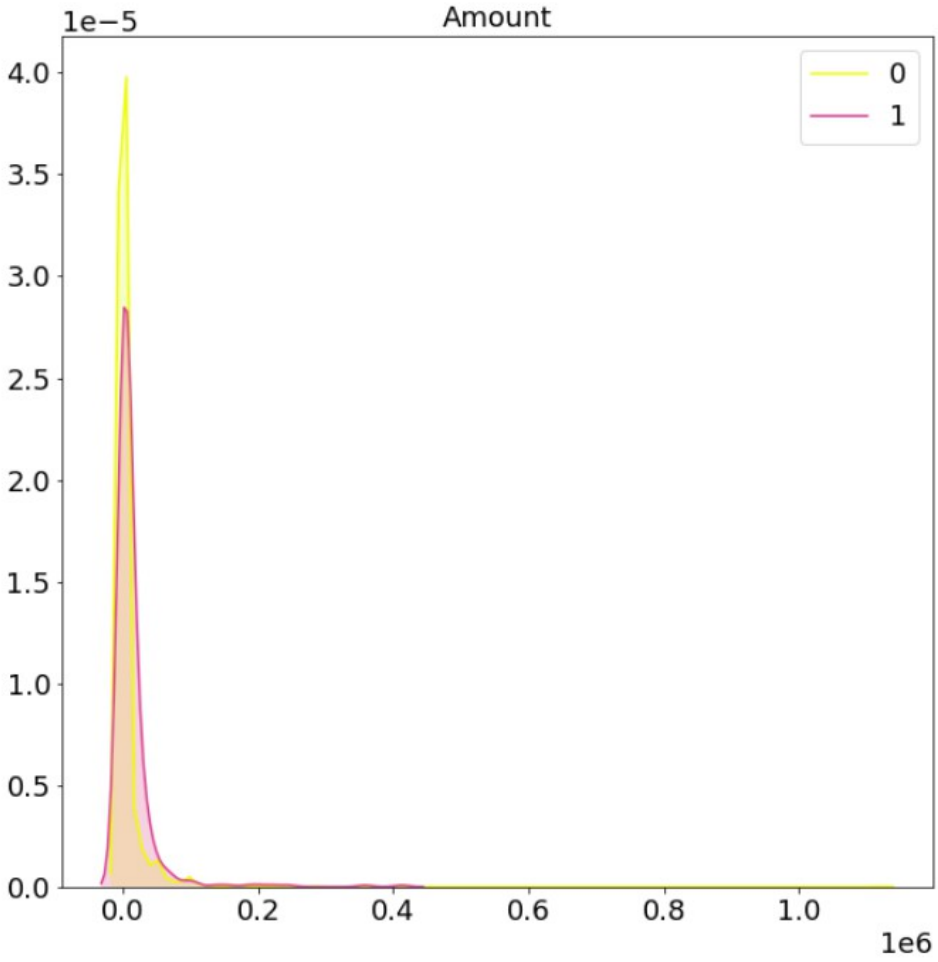


Рисунок 3.2 – Розподіл ознаки amount

Ознаки closing_data та created_time представлені у форматі часу та не можуть бути використовуватися для побудови моделі. Для того, щоб застосувати інформацію, надану в цих ознаках, можна створити нову ознаку – тривалість реалізації (duration_of_implementation). KDE графік розподілу представлено на рисунку 3.3.

Ознака parsed_country має дуже багато унікальних значень, саме тому треба добре зрозуміти її ефективність. На рисунку 3.4 зображена карта світу, яка була побудована з використанням бібліотеки Plotly, на якій мітками позначені країни замовника, в яких реалізація ІС була не успішною.

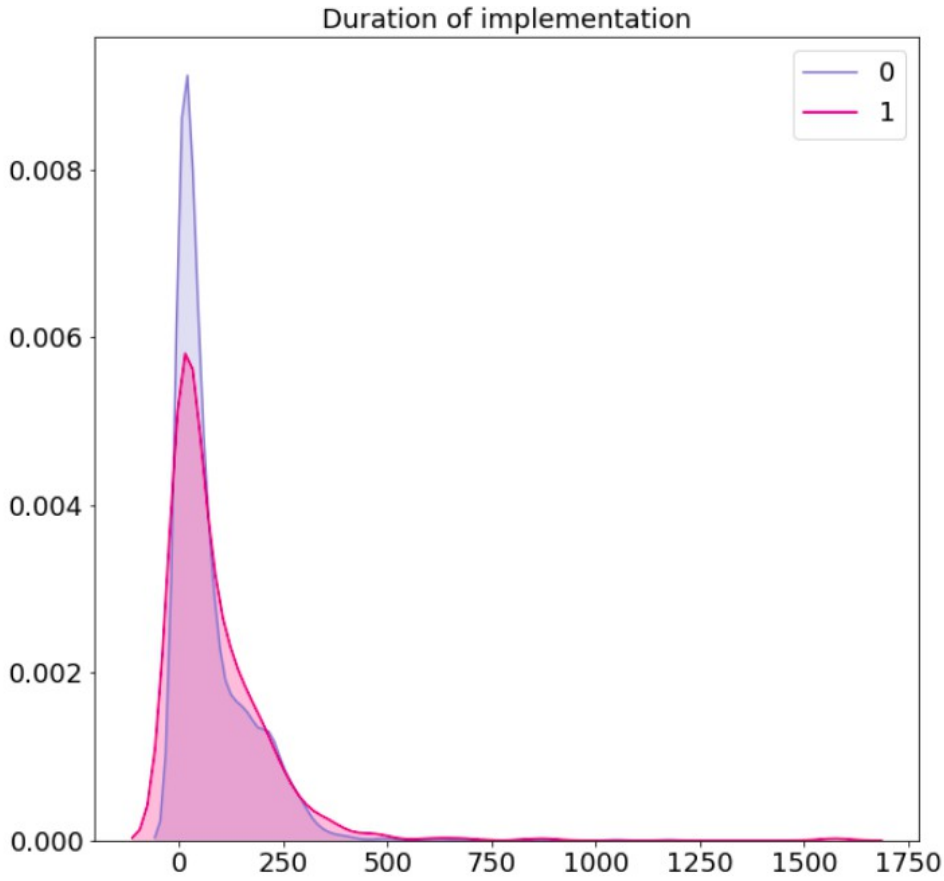


Рисунок 3.3 – Графік розподілу ознаки duration_of_implementation

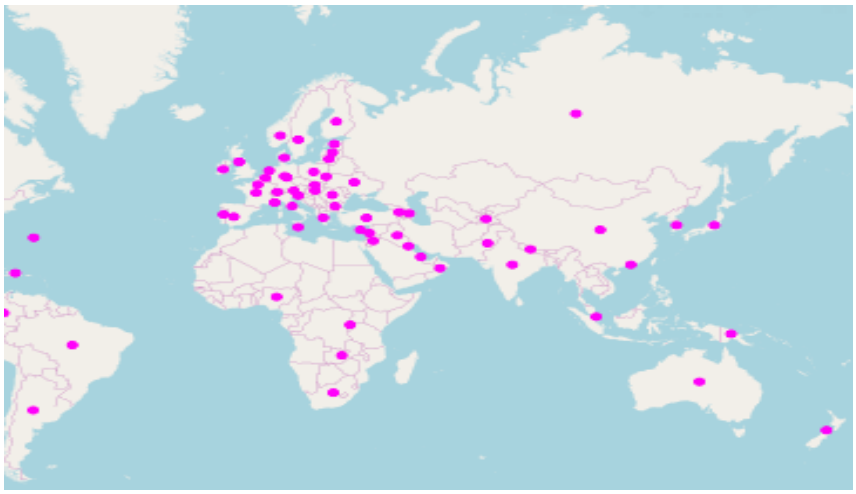


Рисунок 3.4 – Країни з неуспішними реалізаціями ІС

З іншого боку, рисунок 3.5 містить у собі інформацію щодо країн з вдалими реалізаціями ІС. Виходячи з рисунків 3.4 та 3.5 можна зробити

висновок, що ознака parsed_country є важливою для класифікаційної моделі. Про це свідчить те, що країн з успішними реалізаціями значно менше, саме тому країна може значно вплинути на оцінку реалізації.

Рисунок 3.5 – Країни з вдалими реалізаціями ІС

Наступна гістограма, яка зображена на рисунку 3.6, – це розподіл ознаки is_corp_email. Ця гістограма вказує на те, що замовники з корпоративною електронною поштою, мають більше успішних ІС та менше неуспішних, на відміну від інших замовників. Цей факт говорить про те, що дана ознака є важливою та її варто залишити в первинному наборі даних.

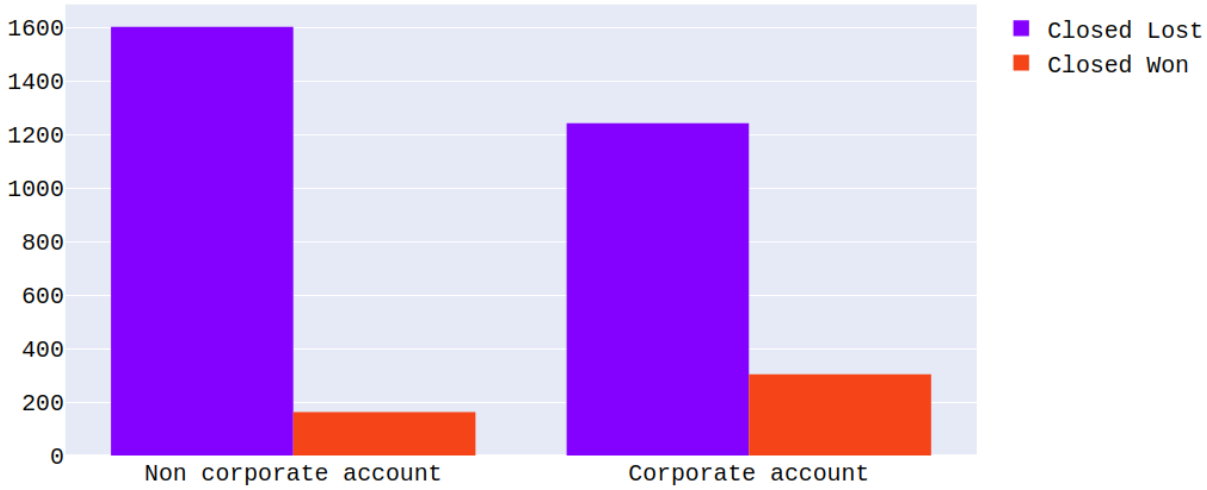


Рисунок 3.6 – Розподіл ознаки is_corp_email

Застосування техніки визнання іменованої сутності може здійснити за допомогою Python бібліотеки Spacy. Ця бібліотека здатна розпізнавати деякі сутності із тексту, наприклад: person (людей), gpe (геополітичні назви), org (організації), cardinal (чисельні значення) та інші. Використання цієї технології з метою створення додаткових ознак, може бути реалізовано підрахунком кількості знайдених сутностей у тексті. Також, для того, щоб зробити цей процес трохи ближчим до первинних даних, було проаналізовано тексти повідомлень та опису ІС, та виявлено шість нових сутностей. Ці сутності допомагають надати більш детальний опис до ІС, тобто які технології або мову програмування вона буде використовувати наведено на рисунку 3.7.

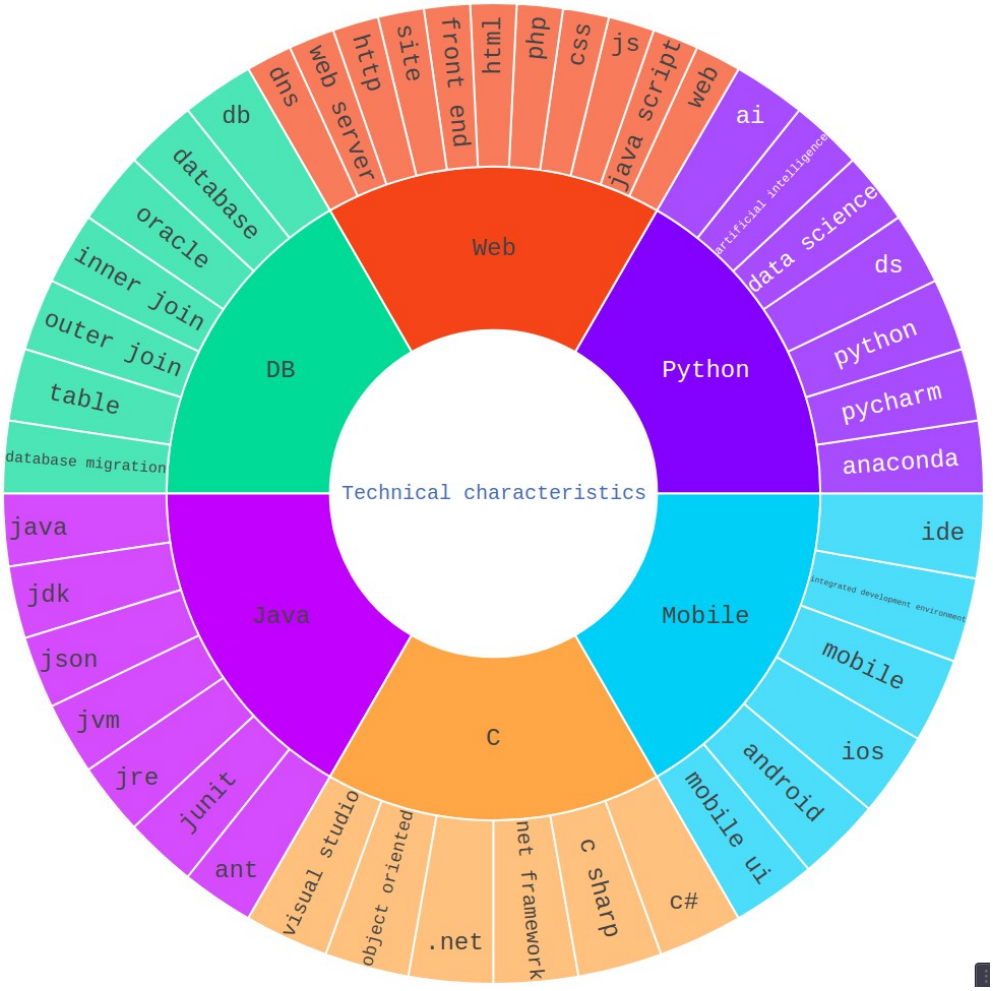


Рисунок 3.7 – Додатково знайдені технічні характеристики

Як показано на рисунку 3.7 усього додаткових сутностей було знайдено шість: застосування мови програмування Java, Python або C подібних мов, використання баз даних, додатки с інтерфейсом користувача, мобільні додатки. Кожна із додаткових сутностей описується уточнюючими словами, які допомагають техніці визнання іменованої сутності розпізнавати їх у тексті. На наступному рисунку можна побачити, що ці сутності є важливі для побудови моделі. Тому що, наприклад, для ІС з використанням мови програмування Java та баз даних, ймовірність успішної реалізації стає вище згідно рисунку 3.8.

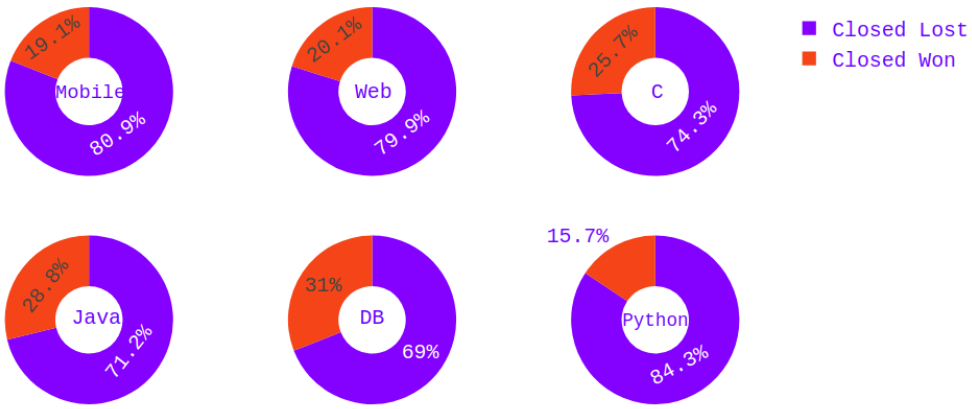


Рисунок 3.8 – Розподіл додатково знайдених сутностей

Усі категоріальні змінні було перетворено за дорогою методу фіктивного копіювання (dummy encoding). Це було зроблено через те, що майже усі вони мають багато унікальних значень, тому усі інші методи застосувати не можна. На рисунку 3.9 представлено приклад фіктивного копіювання для ознаки type.

type_manual	type_automatic	type_No info
0	0	1
0	1	0
0	1	0

Рисунок 3.9 – Результат фіктивного копіювання для ознаки type

Для вирішення проблеми різноманітності значень, було застосовано метод трансформування даних Йео-Джонсона.

Це було зроблено за допомогою Python бібліотеки sklearn. Основною метою цього методу є приведення значень ознаки до однакого діапазону. Результат роботи Йео-Джонсона методу приведено на рисунку 3.10.

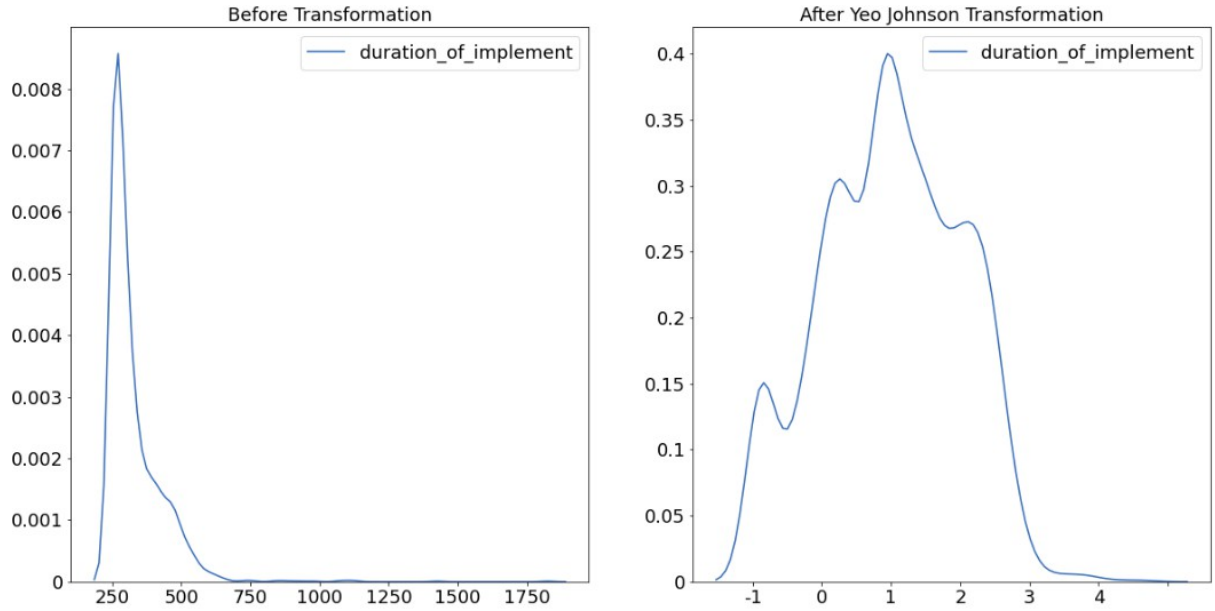


Рисунок 3.10 – Результат роботи методи Йео-Джонсона

На рисунку 3.10 зліва зображено початковий розподіл ознаки duration_of_implementation, справа результат трансформування. Можна побачити, що ознака значно змінила діапазон значень, що вплине на результат роботи моделі.

3.2 Технологія оцінювання можливості реалізації ІС

Для розробки моделі бізнес-процесу використовується методологія SADT. SADT – це техніка, яка корисна для планування системи, аналізу вимог та проектування. Вона була розроблена для розробки

дисциплінованого підходу, метою якого є досягнення розуміння потреб користувачів. SADT – це інструмент для структурування інформації та надання кращого розуміння користувачам проблеми, яку потрібно вирішити. Тому SADT в основному використовується для налаштування і структурування фаз аналізу та проектування.

Бізнес-процес – це система послідовних та цілеспрямованих видів діяльності, в якій за допомогою деяких керуючих факторів і за допомогою відповідних ресурсів входи процесу перетворюються у результат процесу, званий виходом. Для відображення потоку робіт робиться моделювання бізнес-процесу, метою якою є структурування відомостей про ІС у графічному вигляді. Це дає змогу у майбутньому легко аналізувати та поліпшувати існуючу ІС.

Найбільш популярні методі моделювання бізнес-процесу наступні:

- метод функціонального моделювання SADT (IDEF0);
- метод моделювання потоків даних DFD;
- уніфікована мова моделювання UML;
- нотація бізнес-процесів BPMN;
- діаграма Ґанта.

Для побудування моделі бізнес-процесу буде використано метод функціонального моделювання IDEF0. IDEF0 можна використовувати як інструмент для проектування процесів. Він містить багато важливої інформації та з'єднує різні частини між собою, щоб показати їх вплив на процес. IDEF0 дозволяє побачити значення вихідних даних процесу, щоб переконатися, в ефективності бізнес-процесу. Розроблена модель бізнес-процесу представлена на рисунку 3.11.

Технологія використання методів ІАД для оцінювання можливості реалізації ІС містить у собі п'ять основних етапів. Процес починається з надходженням набору даних на вхід. Після отримання набору даних, відбувається основні етапи процесу ІАД: формування набору даних, перетворення, побудова класифікаційної моделі та аналіз отриманого

результату. Вихід кожного з етапів одержується згідно до мети кожного з процесів ІАД (рисунок 3.11).

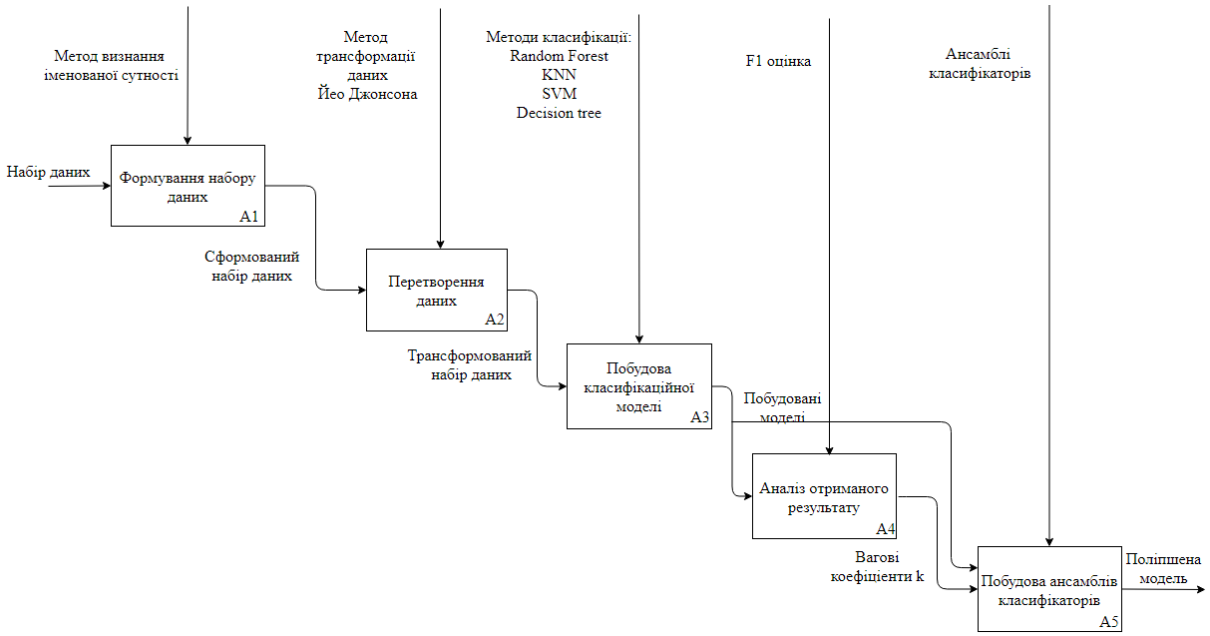


Рисунок 3.11 – Технологія використання методів ІАД для оцінювання можливості реалізації ІС

На виході процесу може одержується поліпшена різними методами ІАД модель прогнозування можливості реалізації ІС.

4 ПРАКТИЧНЕ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

4.1 Обґрунтування вибору інструментальних програмних засобів

Для вирішення задач ІАД найбільш поширеним та використовуваним є мова програмування Python. В ході недавнього всесвітнього опитування було виявлено, що 83% з майже 24 000 спеціалістів з обробки даних використовували Python. Ця мова є дуже легкою у вивченні та має читабельний синтаксис. За допомогою Python, можна будувати великі наукові системи, не турбуючись про проблеми сумісності чи взаємодії. Використовувати Python можна з незначними витратами на обслуговування, що є значною перевагою цієї мови програмування. Також завдяки інтеграційній підтримці з усіх основних API ML та глибокого навчання з'являється можливість застосовувати усі методи ІАД дуже швидко та ефективно.

Для використання мови програмування Python було застосовано додаток Anaconda Navigator. Anaconda Navigator – графічний користувальницький інтерфейс, що дозволяє запускати програми та легко керувати різними пакетами, середовищами та каналами без необхідності використання команд командного рядка.

Використання Anaconda Navigator має кілька переваг:

- містить у собі понад 1500 пакетів для роботи з задачами ІАД;
- його застосування спрощує управління та розгортання пакетів;
- має у собі інструменти для легкого збору даних із джерел за допомогою ІАД;
- створює середовище, яким легко керувати для розгортання будь-якого проекту.

Реалізація створеного теоретичного підходу буде відбуватися засобами наступних Python шести бібліотек.

Pandas – це швидкий, гнучкий та простий у використанні інструмент аналізу та маніпулювання даними, який в даній роботі використовується для роботи з таблицями даних: відображення, зміна тощо.

Numpy – Python бібліотека, особливістю якої є підтримка високорівневих математичних функцій, в даній роботи застосовувалася з метою вилучення нової інформації з даних.

Sklearn – інструмент для прогнозного аналізу даних, який був використаний для побудови моделей класифікації та підготування даних.

Plotly – бібліотека, метою якою є створення інтерактивних графіків та гістограм.

Matplotlib призначена для візуалізації даних у двомірному просторі.

Seaborn – бібліотека візуалізації даних Python, заснована на Matplotlib, яка забезпечує високорівневий інтерфейс для малювання інформативної статистичної графіки.

Побудована модель ансамблів класифікаторів може прогнозувати успішність ІС згідно з її параметрів. Для можливості застосування цієї моделі треба створити графічний інтерфейс з її використанням. За допомогою Python бібліотеки sklearn та методу joblib, побудована модель ансамблів класифікаторів була збережена до окремого файлу для подальшого застосування. Також в ході дослідження кращої моделі та методів для оцінки реалізації ІС, був використаний трансформатор Йео-Джонсона, який також було збережено до окремого файлу.

Графічний інтерфейс користувача та його інтеграція з мовою Python було реалізовано за допомогою таких технологій: Flask, Bootstrap з використанням HTML та CSS.

Flask – це веб-фреймворк, який надає інструменти, бібліотеки та технології для створення веб-додатків. Flask є досить є досить легким для розуміння та швидким фреймворком існує невелика залежність для оновлення та стеження за помилками. Bootstrap – це набір інструментів, які застосовують для створення сайтів і веб-додатків. Він містить у собі HTML

та CSS-шаблони оформлення для веб-форм, кнопок, міток, блоків навігації та інших компонентів веб-інтерфейсу, включаючи JavaScript-розширення. HTML та CSS були застосовані з метою створення зовнішнього вигляду відповідної веб-сторінки.

4.2 Розробка програмного модулю оцінювання можливості реалізації ІС

В ході розробки графічного інтерфейсу користувача, було використано наступні елементи: форми, текстові поля, кнопки, радіо-кнопки, блоки попередження (alert), часові поля, опціонні блоки та інші. Результат створення графічного інтерфейсу можна побачити нижче на рисунку 4.1.

Розроблений інтерфейс згідно з рисунком 4.1. містить у собі 11 обов'язкових для заповнення полей різного типу даних. Для простоти сприйняття інформації користувачем, інтерфейс був розроблений із використанням української мови.

Перш за все користувач має визначитися щодо відповідальної особи та заступника відповідальної особи. Дані поля представлені у вигляді списку кандидатів. Список кандидатів був зроблений відповідно до кандидатів в набору даних. Тобто, якщо користувач обирає зі списку якусь конкретну особу, то в наборі даних ознака обраної особи дорівнюватиме одиниці. Якщо користувач бажаної особи у списку не знайшов, то значення усіх ознак, які пов'язані з відповідальною особою або її заступником, будуть дорівнюватиме нулю. Для легшого пошуку бажаної особи можна скористатися відповідним полем, яке забезпечує швидкий пошук за списком існуючих осіб згідно за рисунком 4.2.

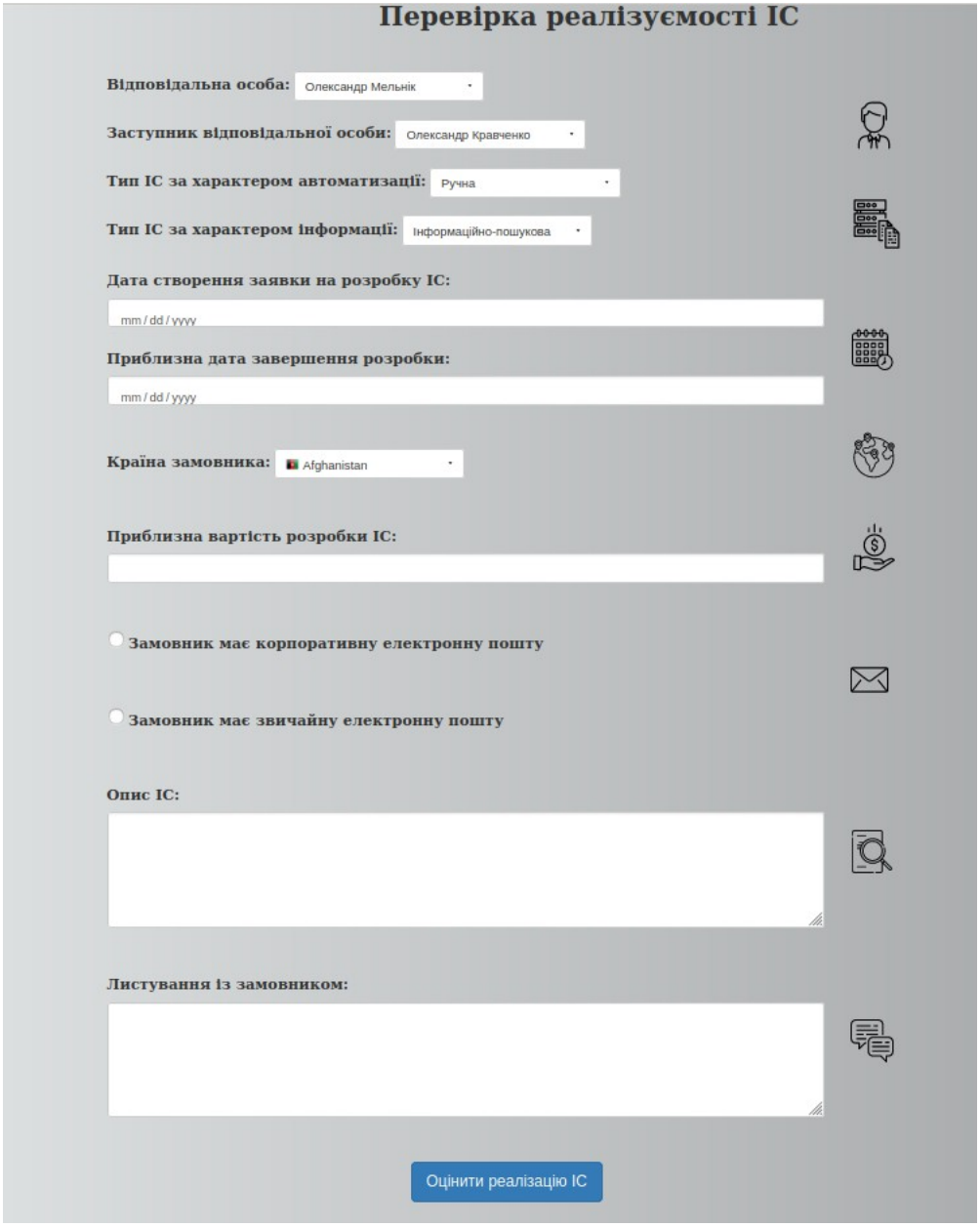


Рисунок 4.1 – Розроблений графічний інтерфейс користувача

Наступною опцією, є обрання типів ІС за характером автоматизації та інформації. Дані поля містять у собі декілька можливих варіантів типу ІС. У випадку, коли користувач знайшов бажаний тип ІС з представленого списку, то обирає відповідний пункт. У іншому випадку, коли бажаного типу ІС у списку немає або користувач однозначно не визначився із типом, то він має обрати пункт «Інша» (рисунок 4.3).

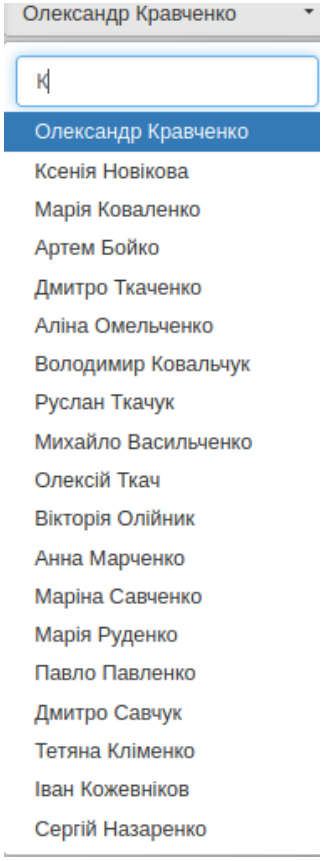


Рисунок 4.2 – Формат полів «Відповідальна особа» та «Заступник відповідальної особи»

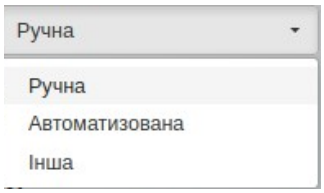


Рисунок 4.3 – Формат поля «Тип ІС за характером автоматизації»

Для простоти визначення значення ознаки duration_of_implementation користувач має обрати лише дві дати: дату створення заявки на розробки та приблизну дату завершення розробки. Ці дві дати користувач може обрати з використанням календаря, відповідних кнопок перемикання або ввести дату власноруч у форматі «MM/DD/YYYY». Після коректного заповнення цих

двох полей, розраховується різниця між двома датами та автоматично заповнюється значення відповідної ознаки. Приклад заповнення та використання часових полей можна побачити на рисунку 4.4.

Рисунок 4.4 – Формат полів «Дата створення заявки на розробку ІС» та «Приблизна дата завершення розробки»

Наступним полем для заповнення є країна замовника. Для його заповнення треба обрати країну, використовуючи назву або відповідний прапор країни. Для визначення значення ознаки, приналежної до країни, треба застосувати абревіатуру країни та встановити відповідну ознаку рівною одиниці. Формат та приклад заповнення поля «Країна замовника» представлено на рисунку 4.5.

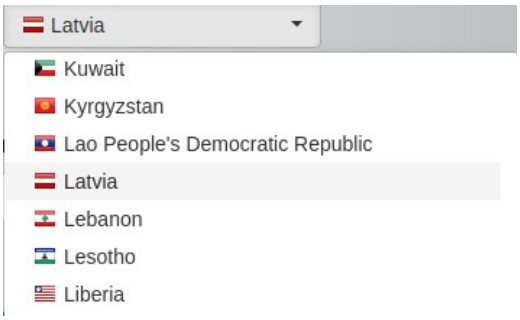


Рисунок 4.5 – Формат поля «Країна замовника»

Ознака amount має чисельний тип даних, саме тому задля її заповнення можна використати звичайне текстове поле, в яке користувач має вписати

вартість розробки. Вартість розробки необхідно вказати у доларах США, для отримання більш точного результату прогнозування. Це пов'язано з тим, що початковий набір даних містив у собі інформацію щодо вартості розробки саме у такій грошовій одиниці. Приклад заповнення поля вартості розробки представлено на рисунку 4.6.

Рисунок 4.6 – Формат поля «Приблизна вартість розробки»

Для визначення типу електронної пошти замовника (звичайна або корпоративна) були застосовані радіо-кнопки. При наявності корпоративної пошти у замовника, користувач має обрати першу опцію. У іншому випадку, обрати кнопку із назвою «Користувач має звичайну електронну пошту». Тоді значення відповідної ознаки буде дорівнюватиме нулю. Нижче представлено формат полей, призначених для визначення наявності електронної пошти у замовника (рисунок 4.7).

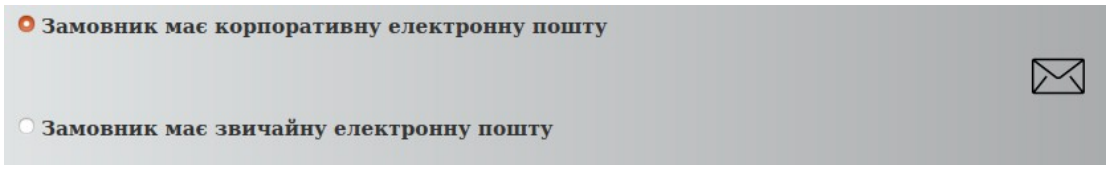
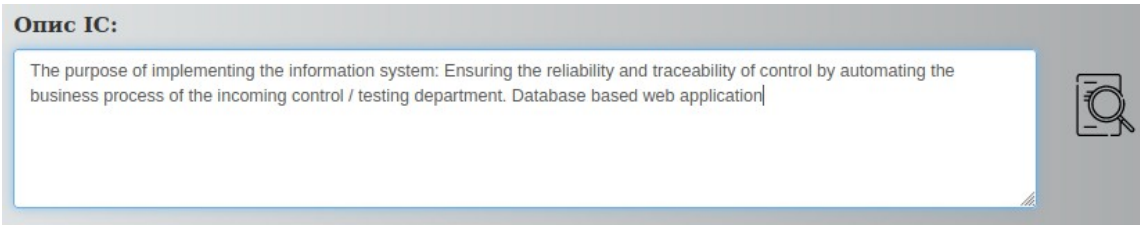


Рисунок 4.7 – Виявлення типу електронної пошти замовника

Поля «Опис ІС» та «Листування із замовником» реалізовані за допомогою текстових полів. Опис ІС містить інформацію щодо цілей, основних принципів, етапів ІС тощо. Листування із замовником – це усі текстові повідомлення у електронній пошті. Дані поля заповнюються англійською мовою через декілька причин. По-перше, у більшості випадків

листування із замовником та уточнення вимог відбувається англійською мовою, тому що вона є інтернаціональною. По-друге, у наборі даних усі ознаки містять інформацію щодо опису та листування теж англійською мовою. При використанні Python перекладача, з'являється можливість отримати гірший результат прогнозування через низьку якість перекладу. Саме через це, краще заповнювати ці поля на оригінальній (англійській мові) або в окремих випадках переводити тексти власноруч. На рисунку 4.8. зображено приклад заповнення опису ІС.



Рисунку 4.8 – Формат поля «Опис ІС»

У разі заповнення усіх перелічених вище полів, користувач має натиснути на кнопки «Оцінити реалізацію ІС» для того, щоб отримати результати прогнозної моделі. Уся введена інформація перетворюється у перелік ознак, створюючи об'єкт для класифікації. Об'єкт подається на вхід до збереженої моделі, яка виводить 0 чи 1. У разі отримання одиниці на виході моделі, користувач отримує повідомлення щодо успішності реалізації відповідної ІС (рисунок 4.9)

Рисунок 4.9 – Повідомлення про успішну реалізацію ІС

В іншому випадку, при прогнозі моделі значення 0, користувач отримує повідомлення про те, що реалізація ІС буде невдалою (рисунок 4.10)

Рисунок 4.10 – Повідомлення про невдалу реалізацію ІС

У випадку отримання попередження щодо невдалої реалізації, користувач має змогу змінювати деякі параметри, доки не отримає бажаний результат. З усього вищезазначеного можна зробити висновок, що розроблена модель з відповідним інтерфейсом допомагають на етапі початку розробки ІС запобігти невдалого результату, або скорегувати декілька характеристик ІС задля забезпечення успішності її реалізації. Розроблений інтерфейс має зручний та легкий для розуміння інтерфейс. Містить у собі поля, для збору усієї необхідної інформації.

4.3 Експериментальна перевірка удосконаленого методу оцінювання можливості реалізації ІС

Для аналізу результату побудованої моделі треба поділити існуючий набір даних на дві частини: навчальний та тестовий набори даних. Застосовувати навчальний набір даних необхідно для того, щоб навчити модель робити прогнози на нових даних. Метою тестового набору є перевірка побудованої моделі. Дуже важливо, щоб тренувальний та тестовий набори даних мали однакову пропорцію класів, це збільшить вірогідність отримання гарного результату. Зробити розбиття початкового набору на дві частини можна за допомогою Python бібліотеки Sklearn. Результат розбиття можна побачити на рисунку 4.11.

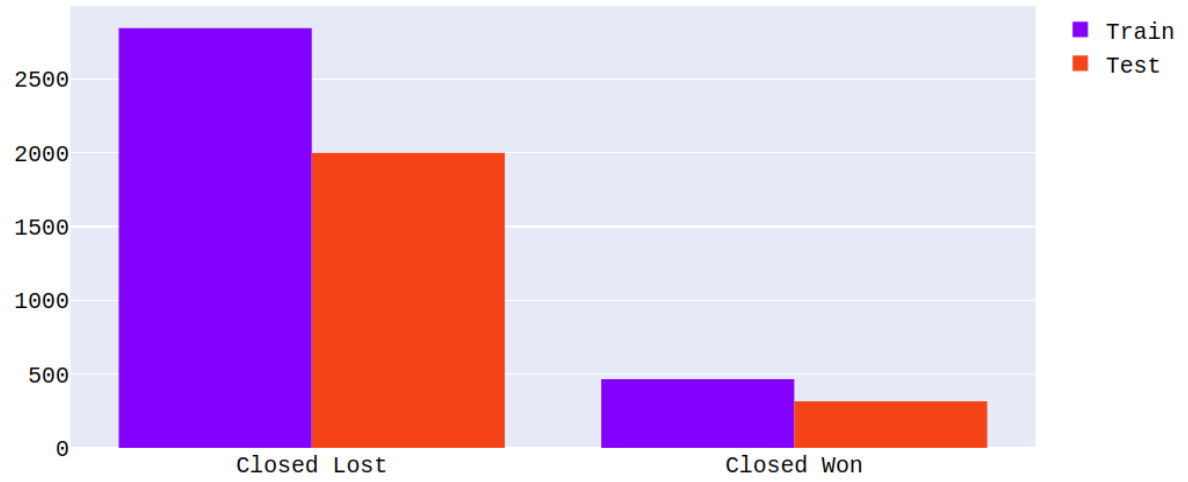


Рисунок 4.11 – Результат розбиття набору даних на начальний та тестовий

Після завершення усіх підготовчих етапів, можна почати процес обробки даних, тобто побудови класифікаційної моделі. Для побудови моделі необхідно на вхід подати набір даних. В ході дослідження було отримано два набори: початковий та трансформований за допомогою Йео-Джонсона. Для того, щоб визначитися з набором даних, який буде використовуватися у наступних дослідженнях, слід побудувати моделі із застосуванням обох та обрати найліпший набір з двох. У першому досліді для побудови моделей були використані параметри за замовчуванням та отримані результати, представлені у табл. 4.1. Згідно з табл. 4.1 отримана оцінка F1 не змінила своє значення для моделей Decision tree та Random Forest, але збільшилася для моделей SVM та KNN. З іншого боку, модель Naive Bayes має значно гірший результат на трансформованому наборі даних, ніж на початковому. Це сталося через те ця модель повинна працювати із початковими даними, відповідно до алгоритму Naive Bayes. Саме тому, для наступних досліджень було обрано трансформований за допомогою Йео-Джонсона набір даних, який має F1 оцінку близьку до 90 для трьох моделей із п'яти. Отже, при першій спробі побудови моделі класифікації, найкраща оцінка F1, яка була отримана – це 93. Цей результат не узгоджується з

первинною метою, мінімальна оцінка F1 якої є 94. Тому необхідно переходити до наступного етапу - поліпшення моделювання.

Таблиця 4.1 – Порівняння результатів побудови моделей на різних наборах даних

	F1 оцінка	
	Початковий набір даних	Набір даних, трансформований за допомогою Йео-Джонсона
SVM	78	88
Naive Bayes	82	11
KNN	78	85
Decision tree	90	90
Random Forest	93	93

Одним із методів поліпшення моделювання – налаштування параметрів моделі. Налаштувати параметри моделі можна за допомогою Python методу Grid Search. Принцип роботи цього методу полягає у тому, щоб методом перебору обрати кращі параметри моделі із запропонованих користувачем. Також користувач обирає відповідну метрику, за якою обирається краща модель в ході перебору. Результат налаштування параметрів моделі методом Grid Search наведено у табл. 4.2.

Відповідно до табл. 4.2 метод налаштування параметрів значно підвищив оцінку F1 для більшості моделей. Саме тому для подальших досліджень буде застосовано моделі із вищою оцінкою, тобто з налаштованими параметрами.

Таблиця 4.2 – Результат налаштування параметрів моделі для поліпшення моделювання

	F1 оцінка	
	Параметри моделі за замовчуванням	Налаштовані параметри за допомогою Grid Search

SVM	88	91
KNN	85	86
Decision tree	90	92
Random Forest	93	93

Наступним методом поліпшення моделювання є метод ансамблів. Реалізація даного методу була виконана із застосуванням Python бібліотеки sklearn. Для bagging методу було обрано модель Decision tree, яка при ітеративному підході показала результати кращі, ніж усі інші моделі. Boosting метод було виконано із використанням Gradient Boosting Classifier, який дозволяє оптимізувати довільні диференційовані функції втрат. Реалізація stacking методу відбулася із застосуванням таких моделей: Random Forest, KNN, SVM, але з фінальне рішення відповідала модель Decision tree. Саме при такому наборі моделей stacking метод показав найкращий результат. Ще одним методом імплементації ансамблів є voting метод, основний принцип якого базується на наступному: приймати фінальне рішення щодо класу об'єкту за допомогою голосування всіх перелічених моделей. Voting метод було реалізовано із використанням моделей Random Forest, Decision tree, KNN, SVM, голоси яких враховувалися у такій пропорції: Random Forest – 2, Decision tree – 2, SVM – 2, KNN – 1. Саме у таких пропорціях voting метод показав найліпший результат. KNN модель має низьку вагу при голосуванні через найменшу оцінку F1. Результати застосування усіх перелічених вище методів наведено у табл. 4.3.

Таблиця 4.3 – Результат застосування ансамблів для поліпшення моделювання

Метод	Використані моделі	F1 оцінка
Bagging	Decision tree	90
Boosting	Gradient Boosting Classifier	91

Stacking	Random Forest, KNN, SVM, Decision tree	91
Voting		94

Таблиця 4.4 – Результат експериментального дослідження

Методи поліпшення	Моделі класифікації					
		Naive Bayes	kNN	SVM	Decision tree	Random Forest
	Існуюча модель	75	62	82	87	91
	Визначення іменованої сутності	82	78	78	90	93
	Трансформування даних	11	85	88	90	93
	Ансамблі класифікаторів	-	94			

Відповідно до табл. 4.3 можна побачити, що один із методів імплементації ансамблів допоміг підвищити F1 оцінку. При застосуванні voting методу ансамблів та трансформованого набору даних, F1 оцінка дорівнюватиме 94. Виходячи з результатів експериментального дослідження, наведених у таблиці 4.4 можна зробити висновок, що отриманий удосконалений метод покращив F1 оцінку на 3% у порівнянні з попереднім початковим методом оцінювання можливості ІС. Таким результат є значним, тому що з використанням удосконаленого методу оцінювання можливості реалізації ІС, кількість хибних передбачень стане менше.

ВИСНОВКИ

Магістерська робота присвячена вирішенню проблеми оцінювання можливості реалізації ІС з урахуванням вимог, особливостей її реалізації та наявних обмежень у часі та ресурсах.

Проведено аналіз існуючих рішень поставленої задачі, який не враховує технічні та організаційні її характеристики.

Удосконалено метод шляхом використання згортки оцінок, отриманих від ансамблів класифікаторів, що враховують технічні та організаційні характеристики інформаційної системи.

Виконано експериментальну перевірку методу. Експериментальна перевірка показала, що використання запропонованого удосконаленого методу дозволяє підвищити F1 оцінку з 91% до 94%.

Для використання удосконаленого методу, за допомогою фреймворку Flask був розроблений зручний графічний інтерфейс користувача. Завдяки своїй простоті та зрозумілості, користувач легко може отримати оцінку реалізації бажаної ІС згідно з її параметрів.

Розроблений веб-додаток може бути використано у будь-яких компаніях, пов’язаних з розробкою ІС. Його застосування допоможе зменшити ризики невдалої реалізації ІС, що позитивно відобразиться на фінансовому стані відповідної компанії.

Створена система оцінки реалізації ІС має наступні перспективи розвитку: покращення прогнозної моделі, використовуючи збільшений набір даних; розширення списку параметрів ІС; створення більш доступного та орієнтованого на користувача графічного інтерфейсу. Усі перелічені перспективи розвитку допоможуть створити потужну систему оцінки реалізації ІС, яка допоможе у запобіганні фінансових та економічних кризах.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Методичні вказівки щодо розробки та оформлення магістерської атестаційної роботи за спеціальністю 122 Комп'ютерні науки (освітня програма «Інформаційні управляючі системи та технології» магістр / Упоряд.: Петров К.Е., Левикін В.М., Чалий С.Ф., Євланов М.В., Саєнко В.І., Міхнов Д.К., Міхнова А.В., Чала О.В. – Харків: ХНУРЕ, 2019. – 24 с.

2. ДСТУ 3008:2015. Інформація та документація. Звіти у сфері науки і техніки. Структура і правила оформлювання. [Чинний від 2015-06-22]. Вид. офіц. Київ, 2016. 31 с.

3. Олійник А. О., Субботін С. О., Олійник О. О. Інтелектуальний аналіз даних : навчальний посібник. Запоріжжя : ЗНТУ, 2012. 278 с.

4. Чалий С. Ф., Чередниченко А. А. Дослідження методів аналізу соціальних мереж для визначення груп користувачів програмного продукту. *АСУ та прилади автоматики*. Харків: ХНУРЕ, 2013. № 165. С. 107-111.

5. Chalyi S., Leshchynskyi V. Knowledge Representation in the Recommendation System Based on the White Box Principle Сучасні інформаційні системи, 2019. Т. 3, № 3. С. 82-86.

6. Чалий С. Ф., Альшейх А. Д. Формалізація бізнес-правил класифікації об'єктів на основі анали за журналу реєстрації подій бізнес-процесу. *Інформаційні технології та в навігації и управлінні: стан та перспективи розвитку*. 2011. С.47.

7. Євланов, М. В. Моделі, методи та інформаційна технологія розробки архітектури складних інформаційних систем на основі функціональних вимог : автореф. дис. . д-ра техн. наук : 05.13.06 / М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. Харків, 2017. 39 с.

8. Люгер Дж. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. М. : Вильямс, 2005. 864 с.

9. Ясницкий, Л. Н. Введение в искусственный интеллект : учеб. пособие для студ. высш. учеб. заведений / Л.Н.Ясницкий. – 3-е изд., стер. -. М. : Издательский центр, 2010. – 176 с.

10. Чубукова И.А. Data Mining: учебн. пособ. – М.: Интернет-университет информационных технологий БИНОМ: Лаборатория знаний, 2006. 382 с.

11. Дюк В., Самойленко А. Data Mining: учеб. курс (+CD); СПб.: Изд-во Питер, 2001. 368 с.

12. Барсегян Ф., Куприянов М., Степанеенко В.. Методы и модели анализа данных OLAP и DataMining / И. Холод. – СПб.: БХВ-Петербург, 2008. 354 с.

13. Ілляшенко К. Інформаційні методи інтелектуального аналізу даних / Економічний аналіз, 2010. С. 390-392.

14. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень: Навчальний посібник / Запоріжжя : ЗНТУ, 2008. 341 с.

15. Паклин Н., Орешков И. Бизнес-аналитика: от данных к знаниям / СПб. : Питер , 2009. 624 с.

16. Provost F. and Fawcett T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, 2013. 414 p.

17. Jensen C. Data Mining: Beginners' Analytics Guide for Business and Science. Amazon Digital Services LLC, 2017. 55 p.

18. Park A. Data Science for Beginners: 4 Books in 1: Python Programming, Data Analysis, Machine Learning. A Complete Overview for Beginners to Master The Art of Data Science From Scratch Using Python. Amazon.com Services LLC, 2020. 513 p.

19. Swamynathan M. Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. Apress, 2017. 358 p.

20. Dong G. and Liu H.Feature Engineering for Machine Learning and Data Analytics. CRC Press, 2018. 419 p.

21. Bhatnagar V. Data Mining and Analysis in the Engineering Field. IGI Global, 2014. 405 p.

22. Singh A. Machine Learning With Python, Independently published, 2019. 137 p.

23. Zinoviev D. Data Science Essentials in Python: Collect – Organize – Explore – Predict – Value. Pragmatic Bookshelf, 2016. 200 p.

24. Karahoca A. Advances in Data Mining Knowledge Discovery and Applications. InTech, 2012. 400 p.

25. Booth T. Python Data Science: Hands on Learning for Beginners. Independently published, 2019. 151 p.

26. Bruce P. and Bruce A. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media, 2016. 250 p.

27. Alfaro E., Gámez M., García N. Ensemble Classification Methods with Applications in R. Wiley, 2018. 224 p.

28. Marca D. A and McGowan C. L. IDEF0 and SADT: A Modeler's Guide. OpenProcess, 2005. 392 p.

29. Stouffer J. Mastering Flask. Packt Publishing, 2015. 288 p.

30. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – Tandem Computers Inc., 1996. 253 p.

31. Usama Fayyad. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? – Tandem Computers Inc.,1996. 54 p.

ДОДАТОК А

Цілі та результати досліджень

Об’єкт дослідження	Процес оцінки можливості реалізації інформаційної системи
Предмет дослідження	Методи інтелектуального аналізу даних в задачах оцінки можливості реалізації інформаційної системи.
Мета роботи	Дослідження методів інтелектуального аналізу даних наборів даних щодо процесу реалізації ІС для оцінки можливості розробки та реалізації інформаційної системи при обмеженнях на строки виконання робіт.
Научна новизна	Удосконалено методи оцінки можливості реалізації інформаційної системи шляхом використання згортки оцінок, отриманих від ансамблів класифікаторів, що враховують технічні та організаційні характеристики інформаційної системи.
Практичні результати	Отримана модель, яка дозволяє прогнозувати можливість реалізації інформаційної системи та заздалегідь коригувати її характеристики з заданими обмеженнями по строкам та вартості. Підвищення точності оцінки отриманої моделі за рахунок комбінації різних способів класифікації, а також врахування додаткових характеристик процесу розробки інформаційної системи.

Задачі дослідження

- аналіз процесу створення інформаційної системи;

- огляд методів інтелектуального аналізу даних;

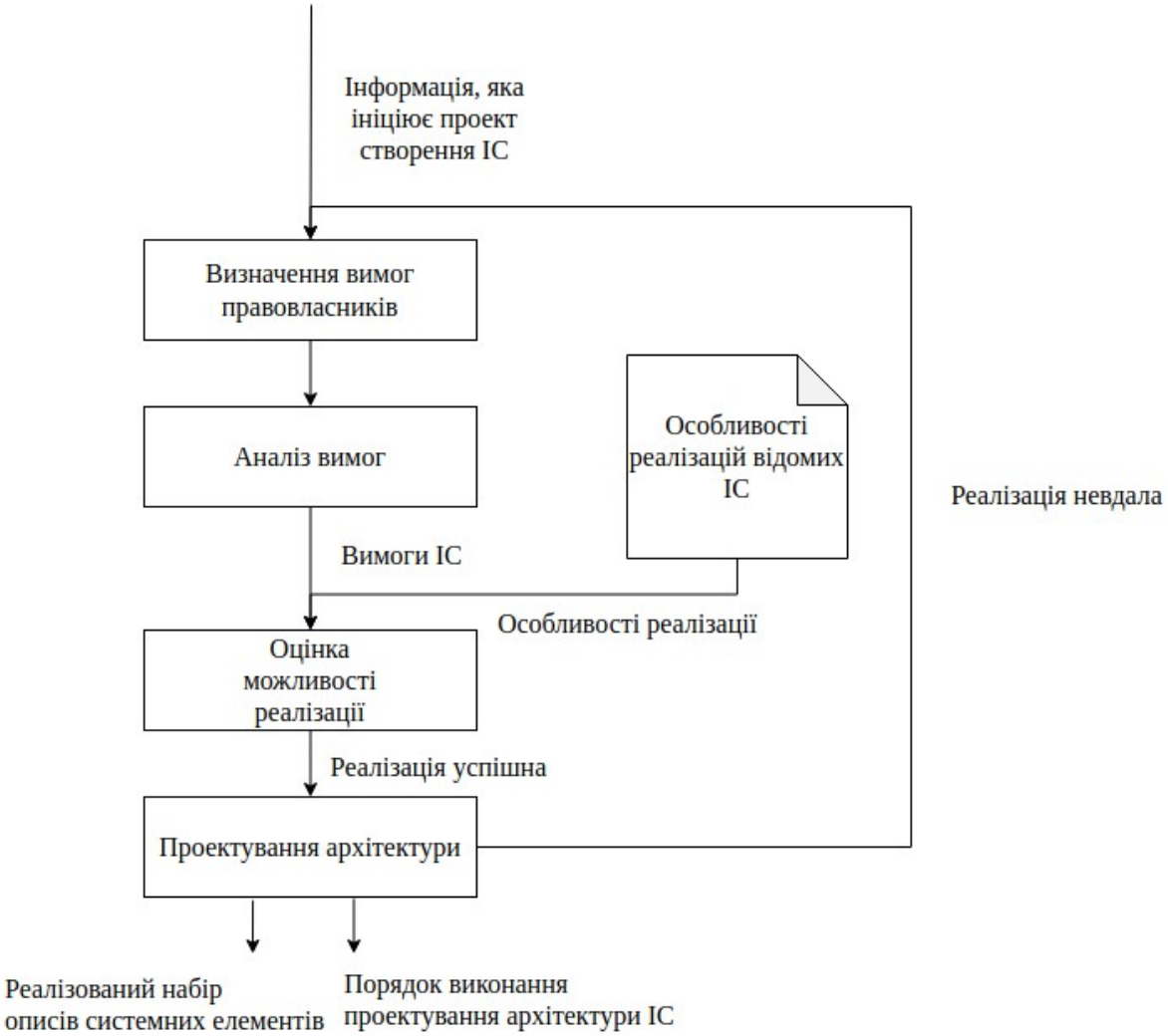
- дослідження проблеми існуючих рішень поставленої задачі;

- створення удосконаленого методу вирішення поставленої задачі;

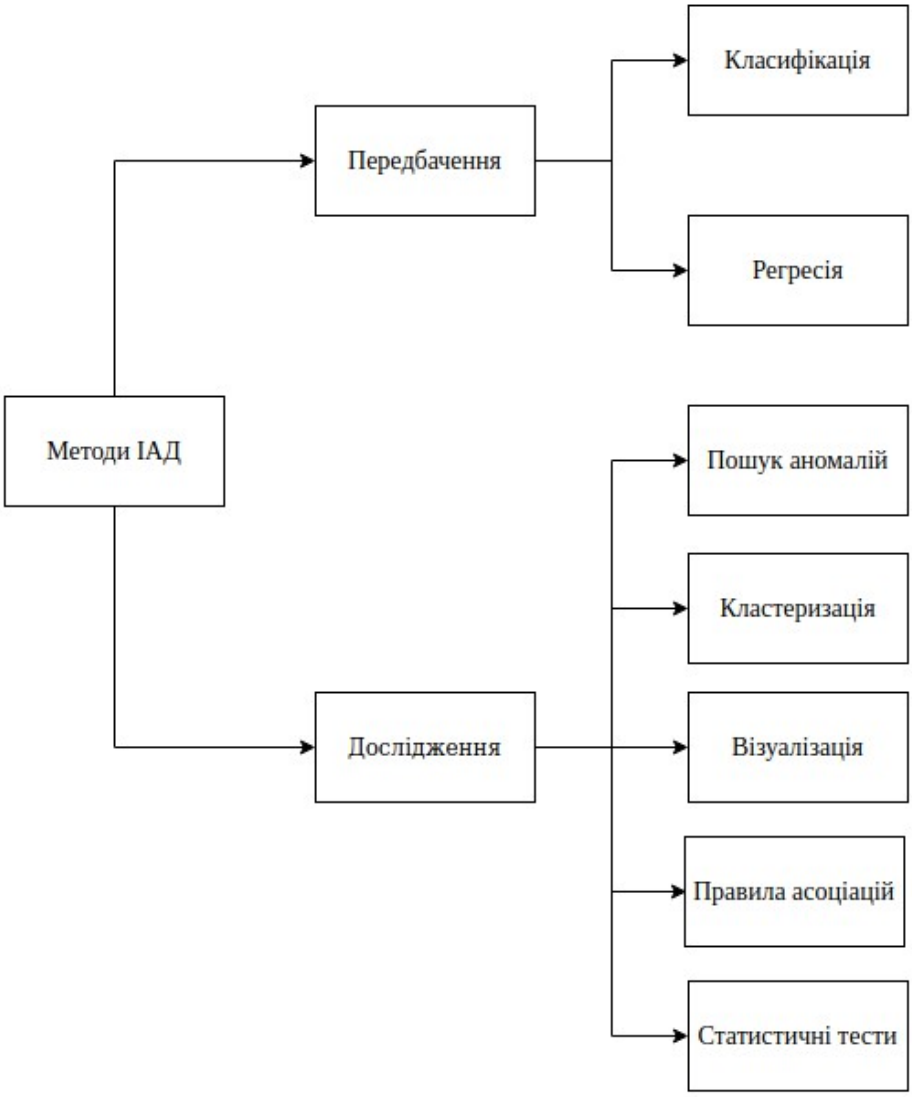
- експериментальна перевірка отриманого результату.

Визначення вимог до ІС

Процес оцінки можливості реалізації інформаційної системи



Методи інтелектуального аналізу даних



Методи класифікації в задачах інтелектуального аналізу даних

Метод опорних векторів (Support Vector Machine)	основна ідея алгоритму полягає у знаходженні лінії або гіперплощини, яка може поділити класи між собою
Наївний Баєсів класифікатор (Naive Bayes Classifier)	імовірнісний алгоритм машинного навчання, який використовує теорему Байєса для побудови класифікаційної моделі
Метод k-найближчих сусідів (kNN neighbors)	алгоритм визначає клас нового екземпляру згідно до класу його k схожих сусідів
Дерева рішень (Decision Tree)	уявлення вирішальних правил в ієрархічній структурі, що складається з елементів двох типів - вузлів і листя, використовуючи які приймається рішення щодо класу екземпляру
Випадковий ліс (Random Forest)	сукупність дерев рішень

Оцінювання методів класифікації

,

де True Positives - це значення, які насправді є позитивними і які модель також передбачає як позитивні,

False Positives - це значення, які насправді є негативними, але модель передбачає як позитивні,

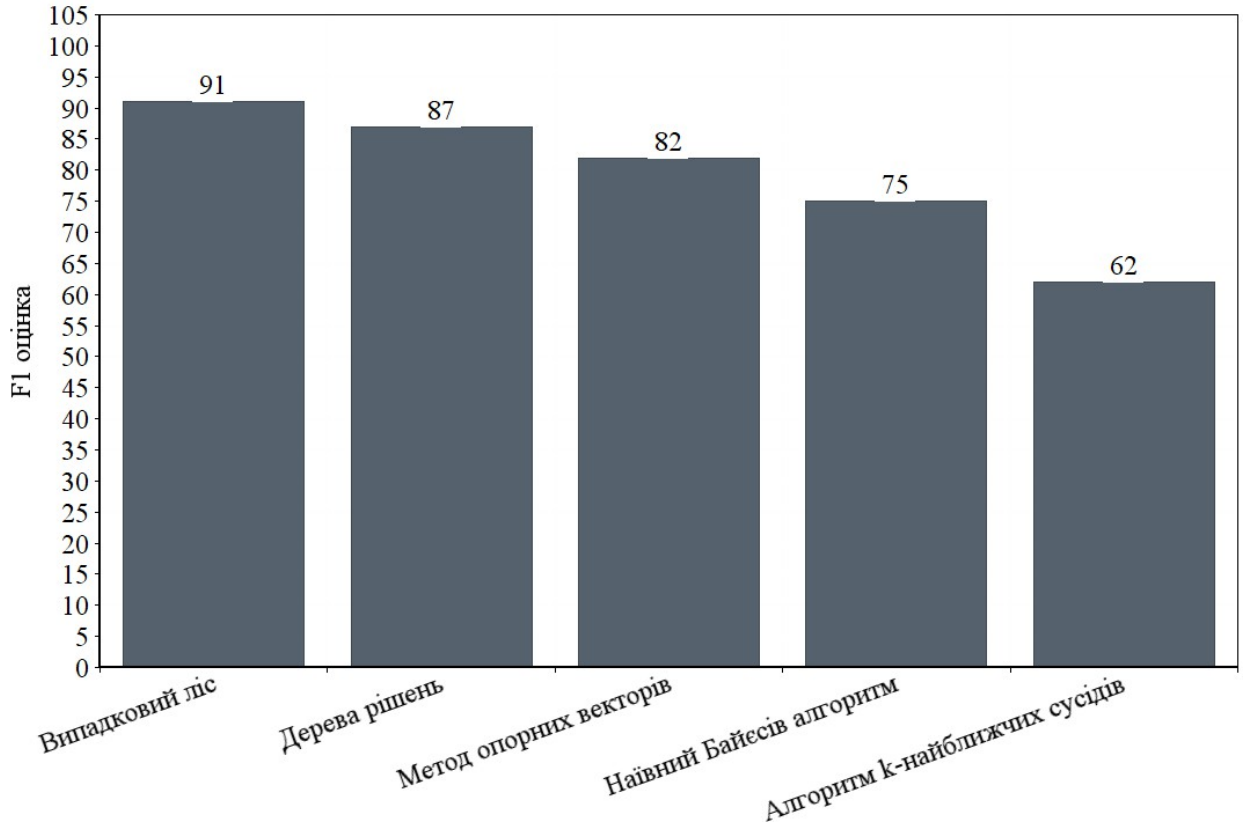
False Negatives - це значення, які насправді є позитивними, але модель також прогнозує негативні.

,

де Precision - точність моделі,

Recall - повнота моделі.

Порівняльна характеристика існуючих підходів



Проблема дослідження

1. Проведений порівняльний аналіз показав, що використання окремих методів класифікації не дозволяє отримати значення оцінки F1 більше 91%, що недостатньо для вирішення задачі оцінювання можливості реалізації інформаційної системи.
2. Для підвищення ефективності оцінювання можливості реалізації інформаційної системи доцільно дослідити можливості використання ансамблів методів класифікації.

Запропонований удосконалений метод

Вхідні дані методу

Набір даних містить у собі наступну інформацію:

- строки створення ІС;
- кінцевий результат створення ІС (вдала реалізація чи невдала);
- типи ІС;
- приблизна вартість реалізації;
- країна замовника ІС;
- інформацію про людину, яка відповідала за реалізацію ІС та її заступника;
- листування із замовником та опис ІС.

Приклад:

is_id	responsible_person	deputy_person	closing_date	created_time	stage	type_1	type_2	amount	parsed_country	is_corp_mail	Potential_text	Notes_text
005022011	Ivan Ivanov	None	2018-02-28 00:00:00	2016-07-15 18:14:00	5. Objections and offer adaptation	automatic	Information retrieval	$ 368,190.00	us	True	hi alan safely back home overcame jet lag serv...	hi alan hope well victor done investigation wo...
014825009	Ivan Ivanov	Alexander Kravchenko	2018-09-13 00:00:00	2018-05-24 18:26:00	7. Closed Won	automatic	Information Critical	$ 41,184.00	us	True	mobile app partially built jon partner needs e...	hi alan pleasure mine thank time support get b...

Основні етапи удосконаленого методу

Етап 1. Доповнення набору даних на основі аналізу вимог та опису ІС з використанням методу інтелектуального аналізу даних «Визнання іменованої сутності».

(1)

де – кількість співпадінь поточного слова в описі або вимогах ІС з характеристикою ІС;

N – кількість слів у тексті;

m – кількість характеристик ІС.

Основні етапи удосконаленого методу

Етап 2. Перетворення первинного набору даних з урахуванням різних масштабів даних за допомогою методу інтелектуального аналізу «Трансформування Йео-Джонсона».

де – початкове значення екземпляру в наборі даних;

– отримане трансформоване значення.

Основні етапи удосконаленого методу

Етап 3. Прогнозування можливості реалізації ІС за допомогою методів класифікації Random Forest, KNN, SVM, Decision tree. Отримання їх передбачень та оцінок .

Етап 4 Розрахунок вагових коефіцієнтів методів на основі F1 оцінок побудованих моделей.

де – отримана оцінка прогнозування відповідної моделі.

Основні етапи удосконаленого методу

Етап 5 поділяється на дві наступні фази:

– крок 5.1. Побудова ансамблів класифікаторів.

(**1**)

де – нормований коефіцієнт згідно до отриманих F1 оцінок,

– результат прогнозування відповідної моделі,

N – кількість моделей прогнозування.

– крок 5.2. Оцінювання можливості реалізації за допомоги ансамблів класифікаторів.

–

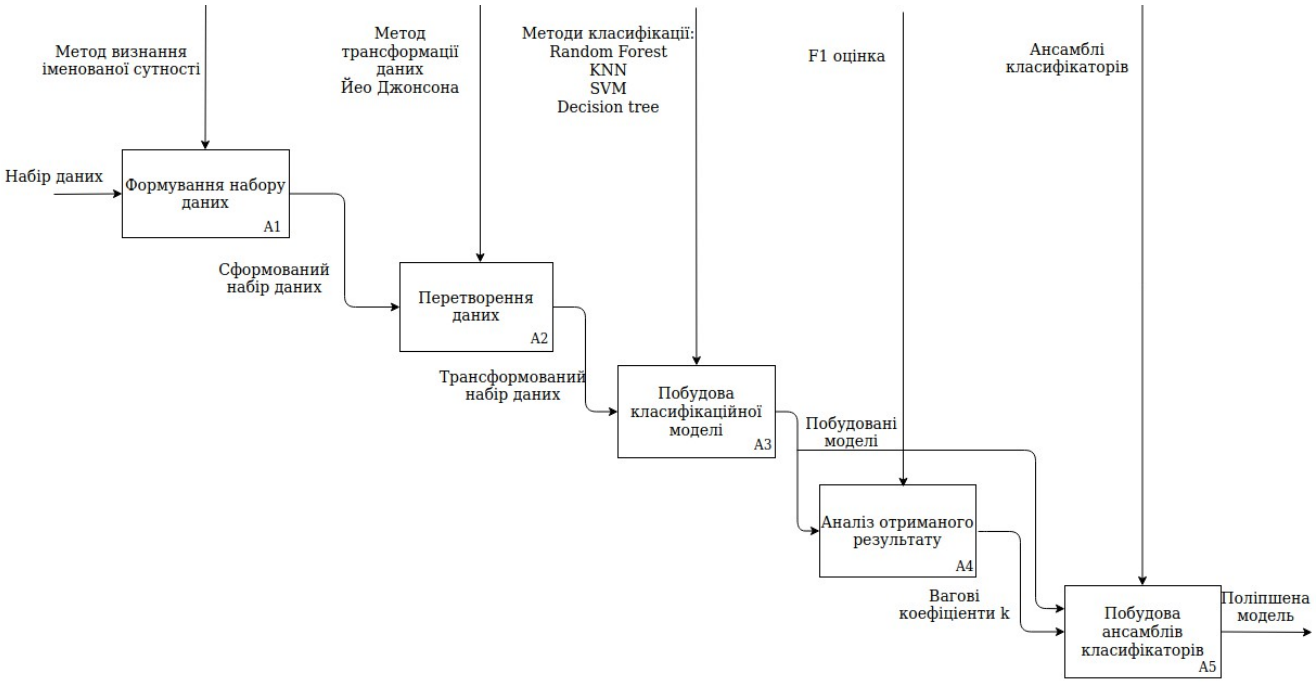
(2)

де 0 – прогноз моделі, який позначає, що реалізація ІС буде невдалою;

1 – прогноз моделі, який позначає, що реалізація ІС буде успішною.

Технологія оцінювання можливості

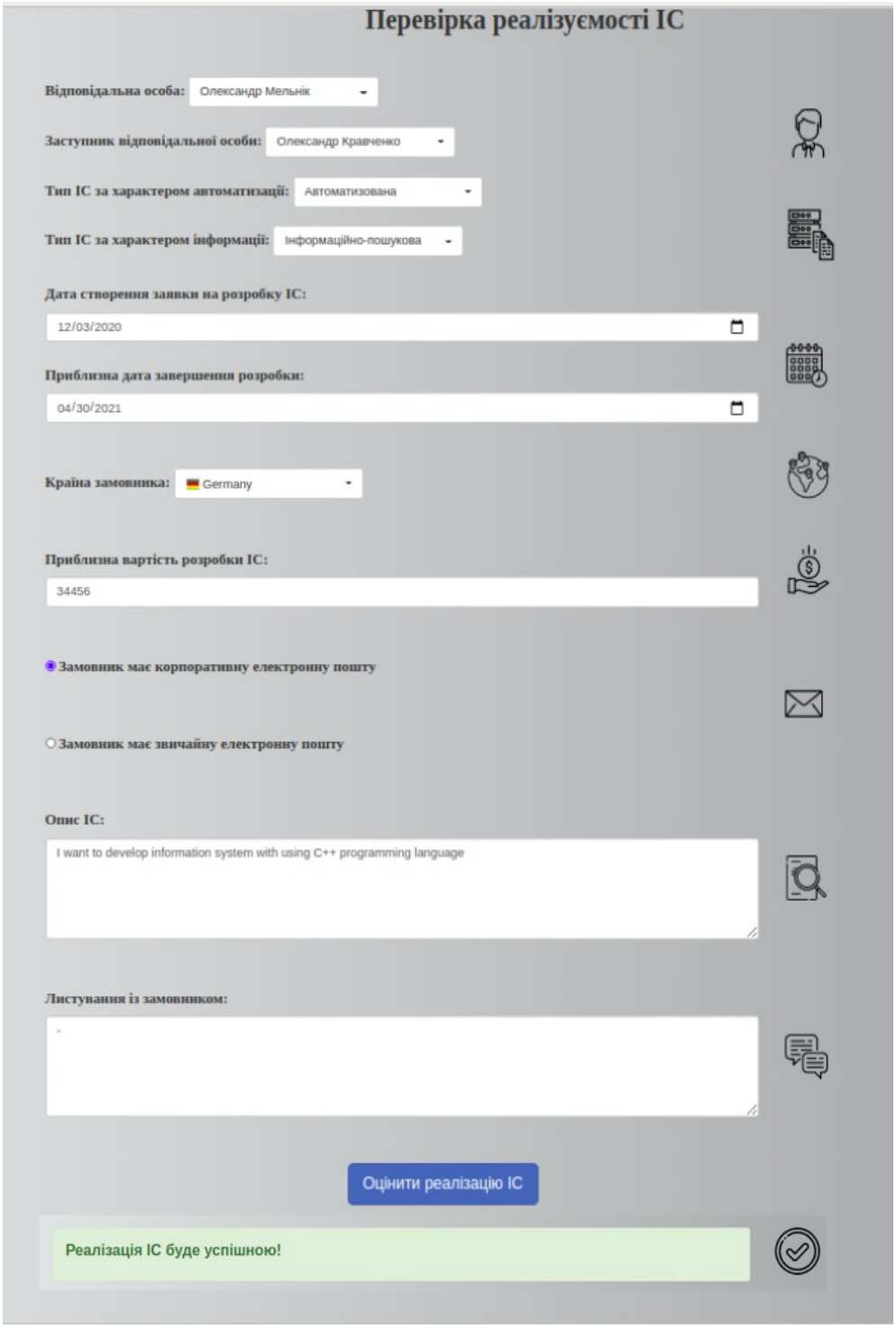
реалізації ІС на основі удосконаленого методу



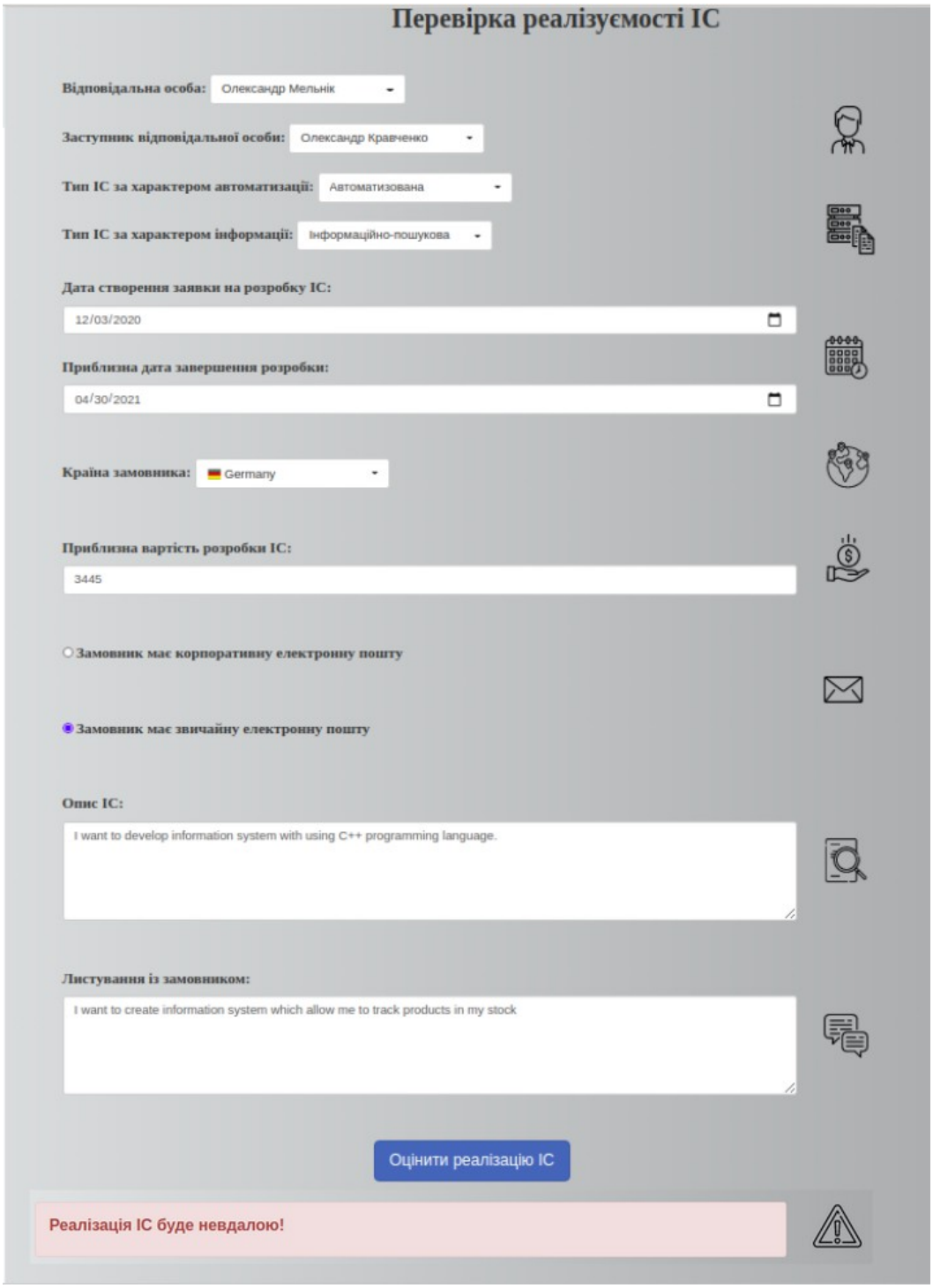
Експериментальна перевірка методу

	Моделі класифікації					
		Naive Bayes	**kNN**	**SVM**	**Decision tree**	**Random Forest**
Методи поліпшення	Існуюча модель	75	62	82	87	91
	Визначення іменованої сутності	82	78	78	90	93
	Трансформування даних	11	85	88	90	93
	Ансамблі класифікаторів	-	**94**			

Застосування удосконаленого методу



Застосування удосконаленого методу



Висновки

1. Магістерська робота присвячена дослідженню проблеми оцінювання можливості реалізації ІС з урахуванням вимог щодо особливостей її реалізації та наявних обмежень у часі та ресурсах.
2. Проведено аналіз існуючих рішень поставленої задачі, які не враховують технічні та організаційні її характеристики.
3. Удосконалено метод оцінювання можливості реалізації ІС шляхом використання згортки оцінок, отриманих від ансамблів класифікаторів, що враховують технічні та організаційні характеристики інформаційної системи.
4. Виконано експериментальну перевірку методу. Експериментальна перевірка показала, що використання запропонованого удосконаленого методу дозволяє підвищити F1 оцінку з 91% до 94%.