УДК 519.163

Міністерство освіти та науки України

Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*
(повна назва)

Кафедра *Системотехніки*
(повна назва)

# АТЕСТАЦІЙНА РОБОТА

## Пояснювальна записка

*другий (магістерський)*
(рівень вищої освіти)

*ГЮИК.502520.015 ПЗ*
(позначення документа)

*Розробка методу інтелектуального аналізу даних клієнтів банківської сфери*
(тема)

Виконала:

студентка *2* курсу, групи *СПРм-19-1*

*Шемчук В.Н.*
(прізвище, ініціали)

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

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

Освітня програма *Системне проектування*
(повна назва освітньої програми)

Керівник *проф. Ситніков Д. Е*
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри системотехніки _____ *проф. Гребеннік І.В*
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*
(повна назва)

Кафедра *Системотехніки*
(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 *Системне проектування*
(повна назва)

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

# ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові *Шемчук Вікторії Насімівні*
(прізвище, ім'я, по батькові)

1. Тема роботи *Розробка методу інтелектуального аналізу даних клієнтів банківської сфери*
затверджена наказом по університету від *« 02 » 11* 20*20* р. № *1516 Ст*
2. Термін подання студентом роботи (проекту) *23 грудня 2020 р.*
3. Вихідні дані до роботи (проекту) *Науково-технічні публікації, дані статей, результати експериментальних досліджень. Перелік використаних програмних засобів: OS Windows 10, Jupiter Notebook, Anaconda Navigator , Python, MySQL, ERwin Data Modeler, веб-сервіс draw.io.*

4. Зміст пояснювальної записки (перелік питань, що потрібно розробити) *Вступ, Аналіз предметної області, Особливості ринку банківських продуктів та послуг, Формування профілю клієнта банку, основні методи аналізу даних та їх недоліки, Інтелектуальний аналіз даних та способи його застосування у банківській сфері, Організація та створення персоналізованих пропозицій для клієнтів банківської сфери, Постановка задачі, Вибір алгоритмів Data Mining для розробки комбінованого методу, Метод пошуку асоціативних правил. Особливості алгоритму Apriori, Метод логістичної регресії, Метод дерева рішень. Алгоритм Random*

*Forest, Побудова методу класифікації на основі комбінування обраних алгоритмів, Проектування системи для автоматизації створення персоналізованих пропозицій, Розробка функціональних вимог системи, Розробка діаграми прецедентів системи, Вибір архітектури для розробки додатку, Вибір СУБД, Логічне та фізичне моделюваня БД, Вибір мови програмування, Перевірка ефективності методу та оцінка Приклад використання навченої моделі, Висновки*

5. Перелік графічного матеріалу. Плакати на аркушах формату А4:

6. Консультанти розділів роботи (проекту)

| Найменування розділу | Консультант (посада, прізвище, ім'я, по батькові) | Позначка консультанта про виконання розділу | |
|---|---|---|---|
| | | підпис | дата |
| | | | |
| | | | |

7. Дата видачі завдання 02.11.2020

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів роботи (проекту) | Термін виконання етапів проекту (роботи) | Примітка |
|---|---|---|---|
| 1 | Отримання завдання на дипломну роботу | 02.11.20 | Виконано |
| 2 | Аналіз поставленої задачі | 06.11.20 | Виконано |
| 3 | Опрацювання літератури та аналіз об'єкту дослідження | 11.11.20 | Виконано |
| 4 | Дослідження сучасного стану проблеми створення персоналізованих пропозицій для клієнтів банку | 17.11.20 | Виконано |
| 6 | Розробка комбінованого методу аналізу заявок на отримання банківських кредитів | 24.11.20 | Виконано |
| 7 | Проектування системи автоматизації створення персоналізованих пропозицій з використанням комбінованого методу | 04.12.20 | Виконано |
| 8 | Перевірка ефективності комбінованого методу | 12.12.20 | Виконано |
| 9 | Оформлення пояснювальної записки | 16.12.20 | Виконано |
| 10 | Представлення на рецензування | 21.12.20 | Виконано |
| 12 | Представлення дипломної роботи | 24.12.20 | Виконано |

Студент _____ Шемчук В.Н.

Керівник роботи (проекту) Ситніков Д. Е

# РЕФЕРАТ

Пояснювальна записка: 76 с., 34 рис., 11 таб., 2 додатки, 42 джерела. Графічний матеріал атестаційної роботи містить 13 плакатів.

Об’єктом дослідження є процес створення персоналізованих пропозицій для клієнтів банку.

Предметом дослідження є методи аналізу даних клієнтів банку для генерації персоналізованих пропозицій банківських послуг.

Метою дослідження є підвищення ефективності маркетингової діяльності банку за рахунок використання системи автоматизації створення персоналізованих пропозицій на основі комбінованого методу Data Mining.

Методи дослідження – системний підхід, методи Data Mining і проектування.

Результати атестаційної роботи – метод інтелектуального аналізу даних, на основі комбінування алгоритмів Apriori, Random Forest і Logistic Regression та спроектована система, для автоматизації створення персоналізованих пропозицій для клієнтів банку.

Галузь застосування – банківська сфера, організації та підприємства, які використовують аналіз даних споживачів для створення таргетованої реклами.

БАНКІВСЬКІ ПОСЛУГИ, ПЕРСОНАЛІЗОВАНІ ПРОПОЗИЦІЇ, СУБД, ТАРГЕТОВАНА РЕКЛАМА, DATA MINING, BIG DATA, APRIORI, RANDOM FOREST, LOGISTIC REGRESSION, MYSQL

ABSTRACT

Explanatory note: 76 p., 34 pic., 11 tables, 2 applications, 42 sources. The graphic material of the attestation work contain 13 posters.

The object of research is the process of creating personalized offers for the bank's customers.

The subject of research is methods of data analysis of bank customers to generate personalized offers of banking services.

The purpose of the research is improving the efficiency of the bank's marketing activities through the use of automation system for creating personalized proposals based on a combined method of Data Mining.

Methods of research – system approach, Data Mining methods and design methods.

Results of attestation work – data mining method, based on a combination of Apriori, Random Forest and Logistic Regression algorithms and a designed system to automate the creation of personalized offers for the bank's customers.

Field of application – banking institutions, organizations and enterprises that use the analysis of costumers data to create targeted advertising.

BANKING SERVICES, PERSONALIZED OFFERS, DBMS, TARGETED ADVERTISING, DATA MINING, BIG DATA, APRIORI, RANDOM FOREST, LOGISTIC REGRESSION, MYSQL

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

ІАД – інтелектуальний аналіз даних;

СУБД – система управління базами даних;

API – application programming interface (програмний інтерфейс програми);

CASE – Computer Aided Software Engineering (сукупність методів та засобів автоматизованого проектування інформаційних систем систем);

CRISP-DM – Cross-Industry Standard Process for Data Mining (стандарт аналізу даних);

IDEF – Integrated DEFinition (методологія функціонального моделювання і графічного описання процесів);

JSON – JavaScript Object Notation (загальний формат для представлення значень і об'єктів);

REST – Representational State Transfer (стиль архітектури програмного забезпечення для розподілених систем);

SQL – Structured Query Language (мова програмування для взаємодії користувача з базами даних);

UML – Unified Modeling Language (уніфікована мова моделювання).

ЗМІСТ

ВСТУП ..... 9
1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ..... 11
1.1 Особливості ринку банківських продуктів та послуг ..... 11
1.2 Формування профілю клієнта банку, основні методи аналізу даних та їх недоліки ..... 14
1.3 Інтелектуальний аналіз даних та способи його застосування у банківській сфері ..... 17
1.4 Організація та створення персоналізованих пропозицій для клієнтів банківської сфери ..... 23
1.5 Постановка задачі ..... 26
2 ВИБІР АЛГОРИТМІВ DATA MINING ДЛЯ РОЗРОБКИ КОМБІНОВАНОГО МЕТОДУ ..... 28
2.1 Метод пошуку асоціативних правил. Особливості алгоритму Apriori ..... 28
2.2 Метод логістичної регресії ..... 33
2.3 Метод дерева рішень. Алгоритм Random Forest ..... 35
2.4 Побудова методу класифікації на основі комбінування обраних алгоритмів ..... 39
3 РОЗРОБКА ФУНКЦІОНАЛЬНИХ ВИМОГ СИСТЕМИ ..... 42
3.1 Характеристика основної функції системи ..... 42
3.2 Характеристика функції системи «Збір додаткових даних про клієнта» ..... 44
3.3 Характеристика функції «Обробка даних клієнта» ..... 46
3.4 Характеристика функції «Аналіз даних для створення персоналізованих пропозицій» ..... 47

4 ПРОЕКТУВАННЯ СИСТЕМИ ДЛЯ АВТОМАТИЗАЦІЇ СТВОРЕННЯ ПЕРСОНАЛІЗОВАНИХ ПРОПОЗИЦІЙ ..... 49
4.1 Розробка діаграми прецедентів системи ..... 49
4.2 Вибір архітектури для розробки додатку ..... 51
4.2.1 Характеристика 1-рівневої архітектури ..... 52
4.2.2 Характеристика 2-рівневої архітектури ..... 53
4.2.3 Характеристика 3-рівневої архітектури ..... 54
4.3 Вибір СУБД ..... 56
4.4 Логічне та фізичне моделюваня БД ..... 57
4.5 Вибір мови програмування ..... 60
5 ПЕРЕВІРКА ЕФЕКТИВНОСТІ МЕТОДУ ТА ОЦІНКА ОТРИМАНИХ РЕЗУЛЬТАТІВ ..... 62
5.1 Огляд вхідних даних ..... 62
5.2 Візуалізація даних ..... 65
5.3 Підготовка даних ..... 66
5.4 Навчання моделі ..... 68
5.5 Приклад використання навченої моделі ..... 69
ВИСНОВКИ ..... 71
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ ..... 72

# ВСТУП

На сучасному етапі розвитку суспільства інформація є одним із найголовніших ресурсів, а інформаційні технології охоплюють всі сфери нашого життя. Тому неможливо уявити функціонування банківської сфери, як однієї з основних у галузі економіки без цих технологій, так як запорукою успішних ділових відносин між клієнтом та банком є інформація [1] та її правильна обробка [2].

В Україні налічується 77 діючих банків [3], з чого випливає конкуренція за клієнтів, і для залучення нових користувачів, має бути проведена масштабна маркетингова політка. Саме аналіз даних клієнтів може допомогти у покращенні банківської реклами.

Як і у будь-якій сфері, банки також мають свої товари та послуги, які потрібні різним групам клієнтів або навпаки. Наприклад, торговий еквайринг навряд чи потрібний людині, у якої неї немає власного бізнесу або товарів для продажу. Вирішення проблеми маркетингу банківських послуг лежить в аналізі даних клієнтів банку та їх класифікації для створення кластерів споживачів і виділення цільової аудиторії [4]. Правильно обрана цільова аудиторія – важлива частина успішної маркетингової кампанії банку.

З цього виникає потреба у розробці методу інтелектуального аналізу даних клієнтів банків, який дозволить ефективно сформувати класи клієнтів та створити для них відповідні персоналізовані пропозицій.

Метою атестаційної роботи є підвищення ефективності маркетингової діяльності банку за рахунок використання системи автоматизації створення персоналізованих пропозицій на основі комбінованого методу Data Mining.

Наявність наукових праць, які висвітлюють питання розвитку технологій аналізу даних клієнтів та оптимізації маркетингової діяльності банківської сфери, а також наявність критичних точок зору фахівців-практиків свідчать про те, що окреслена проблема є актуальною і важливою для розвитку банківської сфери.

Деякі теоретичні, методологічні та практичні аспекти аналізу даних клієнтів банку з метою маркетингу знайшли відображення в роботах багатьох закордонних і вітчизняних вчених та дослідників, таких, як Д. В. Гайдунько, Л. В. Дубчак, Дж. Міланокіс, О. І. Лаврушина, Д.В.Завадська, Л.В.Жердецька [5-7].

Отже, у результаті даної атестаційної роботи повинен бути розроблений метод інтелектуального аналізу даних клієнтів банку для створення персоналізованих пропозицій та спроектована система, що використовує цей метод. Цільовою аудиторією проектованої системи є працівники банківської сфери.

Розроблений метод для підвищення ефективності маркетингової діяльності може бути застосований не лише у банкінгу, а й для для створення таргетованої реклами інтернет магазинів, ресторанного та готельного бізнесу

# 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

## 1.1 Особливості ринку банківських продуктів та послуг

Вагоме місце у функціонуванні ринку фінансових ресурсів займає ринок банківських продуктів та послуг, на якому функціонують банківські установи [8]. Однак, саме банківські установи, на даному ринку користуються значним попитом.

У державних законах щодо банківської сфери, помічається заміна поняття «банківська послуга» терміном «банківська операція». У статті 47 «Банківські операції» закону України «Про банки і банківську діяльність» серед можливих дій банку на ринку представлені поняття «операції» та «угоди», а термін послуги окремо не виділяється [9].

За формулюванням зарубіжного економіста Телінгера М. Ю. банківською послугою є складова діяльності банку, що грунтнується на отриманні вигоди від всього обсягу послуг, які банк здатний запропонувати клієнтові, починаючи від приймання вкладів та надання кредитів, котрі є основою банківської діяльності, закінчуючи новоствореними послугами [10]. Більшість науковців визначають банківську послугу як різновид банківського продукту.

Проаналізувавши різні точки зору, щодо визначення банківської послуги можна зробити висновок, що існує два загальних підходи щодо тлумачення даного поняття.

Сутність першого підходу полягає в тому, що банківська послуга належить до змісту банківських операцій, які є основним напрямом банківської діяльності. Прихильники іншого підходу, принципово розмежовують поняття «операції» та «послуги» банківських установ і не вважають їх тотожними.

Сучасні банки презентують обширну різновидність банківських послуг, як для корпоративних клієнтів, так і для індивідуальних. Конкуренція між банками, змушує їх вводити унікальні банківські послуги та удосконалювати старі. Тобто,

чисельність банківських послуг невпинно зростає, що зумовлює потребу у впорядкованій класифікації існуючих банківських послуг за різними ознаками, яка представлена у таблиці 1.1

Таблиця 1.1 – Класифікація банківських послуг

| Критерії класифікації | Тип банківських послуг |
|---|---|
| Відповідно до специфіки діяльності банку | – традиційні (депозитні, кредитні, розрахунково-касові);<br>– нетрадиційні (лізингові операції, трастові послуги, операції з дорогоцінними металами, посередницькі, консультаційні та інформаційні послуги, фінансовий інжиніринг). |
| Відповідно до типу клієнта | – послуги, запропоновані юридичним особам (факторинг, відкриття поточного рахунку, акредитив);<br>– послуги для фізичних клієнтів (депозити, відкриття рахунків, грошові перекази);<br>– послуги для іших банків (міжнародне фінансування, операції з векселями). |
| За призначенням | – депозити;<br>– кредити;<br>– інвестиції;<br>– розрахункові. |
| За частотою надання | – постійні (розрахунково-касові);<br>– перехідні (інформаційно-консультаційні). |
| За сутністю | – виключно послуги банку ( депозитний розрахунок, відкриття та обслуговування поточного рахунку);<br>– побічні (перевезення дорогоцінностей, конвертація валюти). |
| Залежно від обслуговування | – послуги за участю працівника банку (оформлення кредиту чи консультування);<br>– без участі банківського працівника (самообслуговування, оплата послуги у терміналі, проведення електронного платежу). |
| Залежно від ризиків | – чисті, що не представляють собою значних ризиків як для банку, так і для клієнтів (консультаційні послуги);<br>– сурогатні, що обумовлюють певний ризик для банку і можуть негативно вплинути на його діяльність (кредити, гарантійні послуги, підтверджені акредитиви тощо) . |

Продовження таблиці 1.1

| | |
|---|---|
| За спрямованістю на задоволення потреб клієнта | – прямі, що безпосередньо спрямовані на виконання конкретних запитів користувача (депозитні операції, кредитні, валютні);<br>– непрямі, що регулюють комплексність послуг клієнта (онлайн банкінг, інформаційні послуги, клірингові, видача пластикової картки на базі депозитного рахунку);<br>– додаткові, що приносять додатковий дохід. |
| Відповідно до виду банківського доходу від надання послуг | - комісійні, відповідно до яких банк отримує від клієнта платню у вигляді комісії (ведення реєстрів власників акцій тощо);<br>- гонорарні, відповідно до яких банк отримує від клієнта визначений наперед платіж (консультації, лізинг, трастові операції);<br>- спредові, відповідно до яких , чистий дохід визначається як різниця між комісією, сплаченою клієнтом та комісією від виконання конкретної послуги. |

Таблиця побудована за допомогою джерел [11-12].

Відповідно до Закону України «Про банк і банківську діяльність» клієнтом банку є будь-яка фізична або юридична особа, що має рахунок у банку, чи користується банківськими послугами [9]. Клієнти банківських пропозицій є основою розвитку та процвітання банку. Відповідно до типу відносин споживача послуг з банком, клієнтів можна поділити на такі групи:

а) звичні клієнти – це споживачі банківських послуг, що мають відкритий банківський рахунок та довгострокові взаємовідносини з банком;

б) тимчасові (разові), до них належать клієнти, що не потребують формування довгострокових стосунків з певним банком. Споживачі користуються його пропозиціями задля отримання разової або тимчасової послуги, чи в результаті непередбачуваних ситуацій;

в) нові (потенційні) – клієнти, які планують здобути вигоду від функціонування банку в прямій або в прихованій формі. Такі споживачі представляють собою потенціальну можливість росту і розвитку банку.

Тому задля збільшення обсягу продажу банківських продуктів і послуг, росту доходу та підвищення ефективності, банк повинен спрямувати свою увагу на залучення потенційних клієнтів та утримання постійних.

Основні принципи функціонування банківського ринку наведено на рис. 1.1 Нижче представлена схема побудована за даними джерел [13, 14].

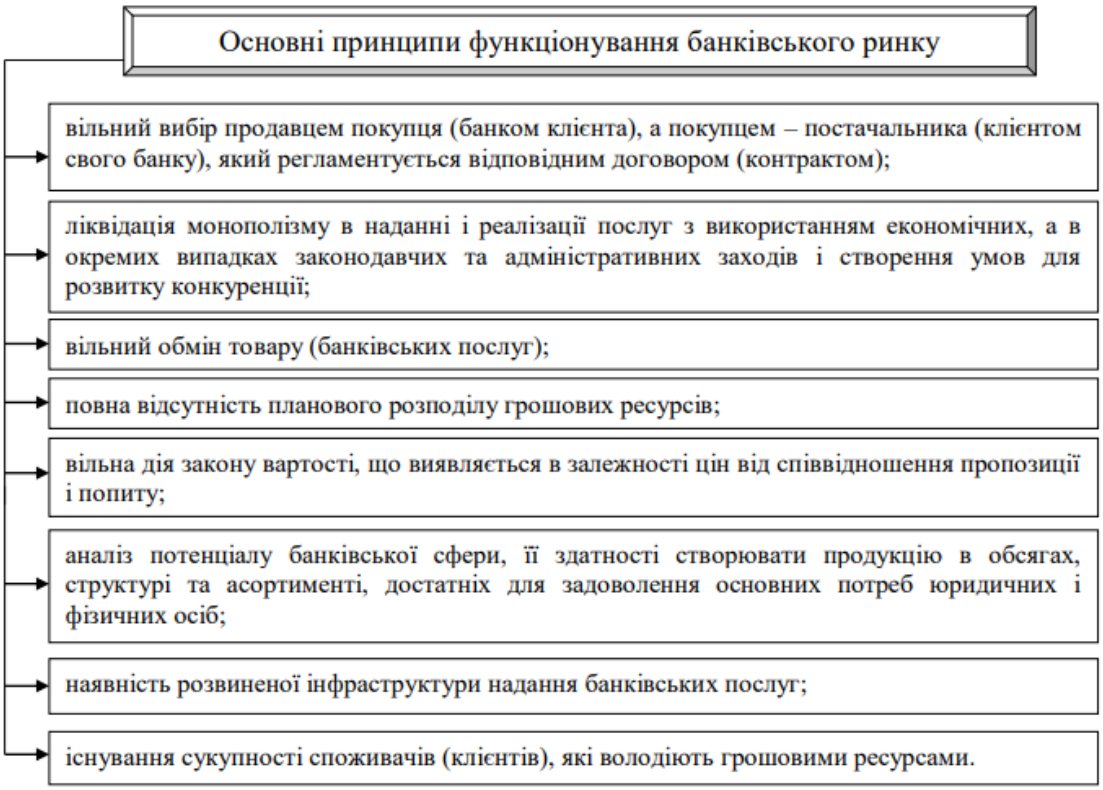


Рисунок 1.1 – Основні принципи функціонування банківського ринку

Важливою як для клієнтів, так і для банку є сегментація банківського ринку. Групування банків за рейтингом таким чином, надає споживачу можливість обрати достойний банк, щоб вигідно вкласти свої збереження або застосувати банківський продукт чи послугу.

## 1.2 Формування профілю клієнта банку, основні методи аналізу даних та їх недоліки

Формування профілю клієнта є однією з основних передумов досягнення банками своєї основної цілі – отримання доходу. Крім того, організація клієнтської бази за рахунок упорядкування профілів підвищує стійкість та конкурентоспроможність банків.

Профіль клієнта – це набір ознак, що відображають відношення та ступінь прихильності клієнта до відповідних банківських продуктів та пропозицій, сприяють змозі коригування його ставлення до цих послуг з метою підвищення рівня роботи банківської структури [15].

Складання клієнтського профілю, відбувається згідно до визначених правил сегментації банківського ринку та використовує чотири основні ознаки споживача [16]: географічне положення, демографічний стан, персональні особливості (соціальний статус, стиль життя, особисті якості), поведінкові манери (стиль придбання послуг). Ступінь конкретизації кожної з ознак визначається відповідно до специфіки складових продуктового портфелю клієнта банківської установи.

При формуванні профілю здійснюють аналіз таких показників [13-14]:

– особливості регіону, у якому проживає клієнт;
– фізична особа чи корпоративний клієнт;
– професія клієнта;
– значимість споживача для банку;
– відношення до конкретних банківських послуг з обгрунтуванням;
– число банків, з якими співпрацює клієнт;
– ситуації припинення стосунків з клієнтом із боку банку;
– тривалість використання певних банківських продуктів і послуг;
– умови розпорядження розрахунковим рахунком та кредитом;
– користування сучасних засобів комунікації з банком;
– обізнаність щодо рекламованих пропозицій банку;
– інформаційні канали сповіщення про сезонні або нові послуги банку.

Клієнтський профіль дає банку можливість приймати правильні рішення щодо складових ознак продуктового кошику клієнта, що потребують змін.

Відповідно до мети створення клієнтського профілю, він може бути розширений даними з кредитних транзакцій клієнта, інформацією щодо виконання договірних зобов'язань. Для цього необхідно розраховувати такі коефіцієнти [16]:

- виконання клієнтом договірних зобов'язань;
- своєчасність розрахунків;
- мінливість стилю поведінки відповідно до конкретних банківських послуг;
- вподобання різних компонентів продуктового портфеля банку.

Враховуючи вище представлені коефіцієнти, банк може виділити додаткові структурні елементи клієнтської БД та цільові перспективні групи клієнтів, що підвищує ефективність розробки цільових стратегій маркетингу.

Найчастіше використовують при роботі з клієнтами застосовують такі методи опитування як: співбесіда, анкетування, телефонне опитування. Методами вивчення мотивів споживчої поведінки є: заглиблена співбесіда, групове інтерв’ю, асоціації, групове вивчення [1].

Головна задача аналізу клієнтського профілю в системі управління ресурсами полягає в тому, щоб максимально знизити невизначеність, в якій приймаються рішення щодо формування та використання ресурсів банку в цілому та поточних пасивів зокрема. Для цього активно використовуються: метод кластеризації, метод регресії [17]. А задля підвищення якості моделей застосовують деякі методи машинного навчання (нейронні мережі, Random Forest і Gradient Boosting тощо) [18].

Для прогнозування характеристики об’єкта використовують метод регресії. Наприклад, певний набір внутрішніх даних (транзакції за картками, обсяг заощаджень та ін.) по фізособі, яка не є зарплатним клієнтом банку, дозволяє прогнозувати його дохід. Знаючи оцінку доходу, банк може сформувати для клієнта адекватну продуктову пропозицію по кредиту або підібрати оптимальний інструмент заощадження коштів, які будуть відповідати його можливостям [7].

У маркетинговому аналізі клієнтського профілю можуть бути використані методи сценаріїв, експертних оцінок, порівняння, побудови систем показників і аналітичних таблиць. При цьому в залежності від конкретних маркетингових цілей аналізу до інформації можуть застосовуватися різні методи.

Для ранжування клієнтів використовують метод побудови рейтингів. Рейтингові показники можуть бути розраховані на основі різних методик [6]:

– рейтингової системи, що ґрунтується на поєднанні показників з присвоєнням кожному з поєднань певного місця в рейтингу;

– бальної системи, що ґрунтується на присвоєння показникам клієнта балів з визначенням підсумкового балу, на основі якого клієнта відносять до певної категорії;

– індексної системи, заснованої на розрахунку індексу для кожного з показників з подальшим визначенням зведених індексів на основі врахування вагових коефіцієнтів індексів кожного з показників.

Загальноприйняті методи, що використовуються у банківській сфері мають ряд недоліків. Рейтингову систему не завжди можливо застосувати через складнощі визначення допустимих поєднань значень показників.

Застосування методу комплексного аналізу також пов'язане з певними складнощами: бальні системи оцінювання потреб клієнта повинні бути статистично достовірними та вимагають систематичного оновлення даних, що часто буває коштовним для банку. Стандартний характер цих моделей не передбачає врахування персональних особливостей окремих споживачів послуг.

## 1.3 Інтелектуальний аналіз даних та способи його застосування у банківській сфері

У зв'язку з вдосконаленням технологій запису і зберігання даних на суспільство обрушилася колосальні потоки інформації в самих різних областях, що вимагає продуктивної переробки. Управління досвідом клієнтів стало ключовою темою в сучасній роздрібній банківській галузі за останні роки, оскільки банки намагаються вдосконалити свої традиційні бізнес-моделі, орієнтовані на продукти та транзакції. Зараз вони намагаються перейти на підхід, орієнтований на споживача, і вдосконалити перехресні продажі та збільшити частку гаманця.

Об’єми даних клієнта настільки обширні, що звичайна людина фізично не здатна опрацювати їх самостійно, хоча необхідність у проведенні такого аналізу є очевидною, так як саме в цих “сирих даних” міститься інформація, що може бути використана для прийняття рішень [19].

Традиційна математична статистика довгий час виконувала роль основного інструменту при аналізуванні даних, але з розвитком технологій не змогла повною мірою відповісти проблемам, що з’явились. Тому виникла необхіднсть у розвитку нових сучасних методологій обробки та аналізу даних. Такою новою методологією і став інтелектуальний аналіз даних (Data Mining).

Data Mining – це процес виділення в «сирих» даних ще невідомих до цього моменту доступних, практичних та корисних інтерпретацій знань, що є необхідними при прийнятті рішень у будь-якій сфері діяльності людини [20].

Загальновживані методи аналізу даних зазвичай спрямовані на перевірку завчасно сформульованих гіпотез і на "грубий" аналіз, що являє собою фундаментом своєчасної аналітичної обробки даних, тоді як одним з головних тверджень Data Mining є пошук неочевидних закономірностей.

Інструменти Data Mining здатні самостійно знаходити такі закономірності та самотужки формулювати гіпотези щодо взаємозв'язків. Так як саме формулювання гіпотези, щодо взаємозв'язків є одним з найскладніших завдань, переважання Data Mining у порівнянні з іншими методами аналізу є безсумнівним.

Data Mining допоможе перевірити величезні масиви і знайти важливу приховану  інформацію, що допоможе компвніям краще зрозуміти своїх клієнтів і передбачити їх поведінку [21].

Єдиного переліку того, які задачі варто включати до ІАД, немає. Та все ж, більшість авторитетних джерел виділяє такі основні задачі [19, 20] :

– Класифікація (Classification). Даний вид задач є найпростішим та найрозповсюдженішим в ІАД. У результаті розв’язання задач класифікації визначаються показники, що описують об'єкти набору даних, які досліджуються –

класи; враховуючи дані показники, новий об'єкт можна співвіднести до того чи іншого класу.

– Кластеризація (Clustering) є складнішою задачею, ніж класифікація. Основною рисою кластеризації є те, що класи об'єктів спочатку не визначені. Результатом даної задачі є розбиття об'єктів на групи. Метод кластеризації дозволяє по набору продуктових параметрів виділити клієнтські сегменти з унікальним поєднанням характеристик.

Наприклад, при кластеризації власників зарплатних карт банку ми бачимо значну різницю в моделі використання даного продукту: частина клієнтів орієнтована на дистанційні канали обслуговування та активно використовує супутні продукти й послуги; інші користуються карткою банку для покупок, здійснюючи при цьому частину операцій за допомогою готівки або за допомогою карт інших банків; ряд клієнтів знімають заробітну плату практично відразу після отримання.

Виявлення конкретних драйверів дозволяє оцінити цільовий портрет клієнта і вибудувати відповідну стратегію комунікації з метою максимізації доходу банку.

– Асоціація (Associations). Під час розв’язання задачі пошуку асоціативних правил, відбувається пошук закономірностей між взаємопов’язаними подіями у певному наборі даних. Різниця асоціації та вищепредставлених двох задач ІАД в тому, що процес знаходження звя’зків відбувається між подіями, що здійснюються одночасно.

– Прогнозування (Forecasting). У результаті вирішення задачі прогнозування, відбувається оцінювання пропущених чи майбутніх значень цільових числових показників за допомогою характеристик властивостей існуючих даних. Для знаходження рішення таких задач досить часто використовують методи математичної статистики, нейроні мережі тощо.

– Візуалізація (Visualization). Результатом задачі є представлення графічного образу даних, що аналізуються. Під час вирішення задач візуалізації застовуються методи графічної обробки, які показують наявність закономірностей у даних.

Структура методів Data Mining представлена на рис. 1.2. Рисунок зроблений автором за джерелами [23-24].

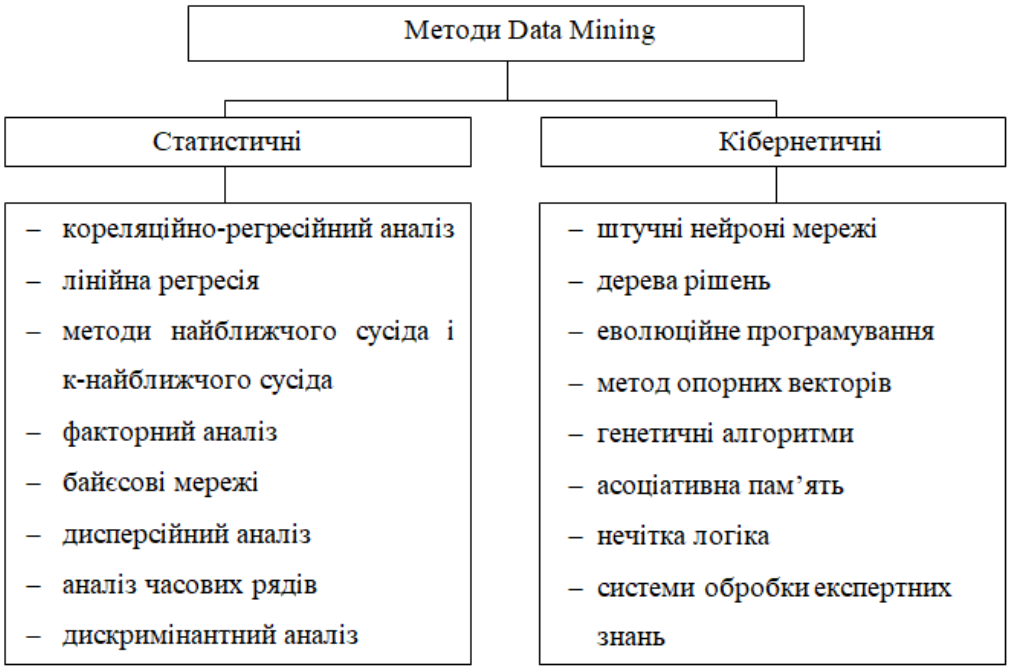


Рисунок 1.2 – Методи ІАД

Метод дерев рішень (decision trees) – один з найпопулярніших методів для розв'язання задач класифікації та прогнозування. Дерева рішень представляють правила в послідовній ієрархічній структурі, де кожному об'єкту відповідає єдиний вузол, що дає результат [23].

В інтелектуальному аналізі даних, дерева рішень використовуються як математичні та обчислювальні методи для опису, класифікації та узагальнення набору даних, які можуть виглядати таким чином (формула 1.1):

$$(x, Y) = (x_1, x_2, x_3 .. x_k, Y) \quad (1.1)$$

де $Y$ – цільова змінна.

До вектору $x$ належать вхідні змінні $x_1, x_2, x_3$ і т.д., які застосовані при виконанні цього завдання [25].

В Data Mining виділяють дерева рішень двох напрямків:

– класифікації, в яких результати представлені у вигляді класів, до яких належать дані;

– регресії, в яких передбачений результат розглядають як дійсне число (наприклад, вартсть автомобілю або тривалість участі велосипедиста у гонці).

Метод дерева рішень має певні переваги серед інших методів Data Mining. Модель класифікації, у вигляді дерева рішень, є інтуїтивно-зрозумілою, що полегшує процес розв’язання поставленої задачі. Результати алгоритмів побудови дерев рішень, на відміну від нейронних мереж, які представляють собою "чорні ящики", легко інтерпретується користувачем [22].

Переважаюча кількість алгоритмів навчання дерева рішень володіють здатністю обробки пропущених значень. Багато математичних методів, що вирішують завдання задачі класифікації, здатні працювати лише з числовими даними, в той час як дерева рішень працюють як з числовими, так і з категоріальними типами даних.

Тобто, дерева рішень можуть розв’язувати такі завдання Data Mining, у яких пропущена апріорна інформація про вид зв’язку між досліджуваними даними. Тому цей метод класифікації даних був використаний при розробці методу інтелектуального аналізу даних клієнтів банківської сфери.

Досягнення технології Data Mining використовуються у банківській справі для виконання таких завдань [14, 26]:

– виявлення шахрайства із кредитними картками. Аналізуючи минулі транзакції, які виявилися шахрайськими, банк виявляє "підозрілі стереотипи поведінки" в майбутніх транзакціях;

– видача кредитів. Дані клієнтів банку розбиваються на два класи (повернуті та не повернуті кредити); на основі груп клієнтів, що не повернули кредит, визначаються основні "риси" потенційного неплатника; при надходженні інформації про нового клієнта визначається його клас;

– залучення нових клієнтів банку. За допомогою інструментів Data Mining можливо провести класифікацію на "більш вигідних" та "менш вигідних" клієнтів. Після визначення найбільш вигідного сегменту клієнтів банку існує зміст, який проводить більш активну маркетингову політику за залученням клієнтів саме серед найвищої групи;

– сегментація клієнтів. У результаті використання Data Mining вирішується завдання сегментації клієнтів на основі їх прибутку. Аналіз виділяє ті сегменти покупців, які приносять найбільший прибуток. Сегментація також може здійснюватись на основі лояльності клієнтів.

У результаті сегментації всієї клієнтської бази буде здійснено поділ на певні сегменти з загальними характеристиками. Відповідно до цих характеристик компанія може індивідуально підбирати маркетингову політику для кожної групи клієнтів.

Під інтелектуальним аналізом даних клієнтів банку слід розуміти процес виокремлення та аналізу великої кількості неструктурованої інформації з використанням статистичних та математичних методів для отримання нових, структурованих даних щодо кредитоспроможності клієнтів банку [26].

Типові етапи, які супроводжують рішення задач інтелектуального аналізу даних у банку представлені на рис 1.3.

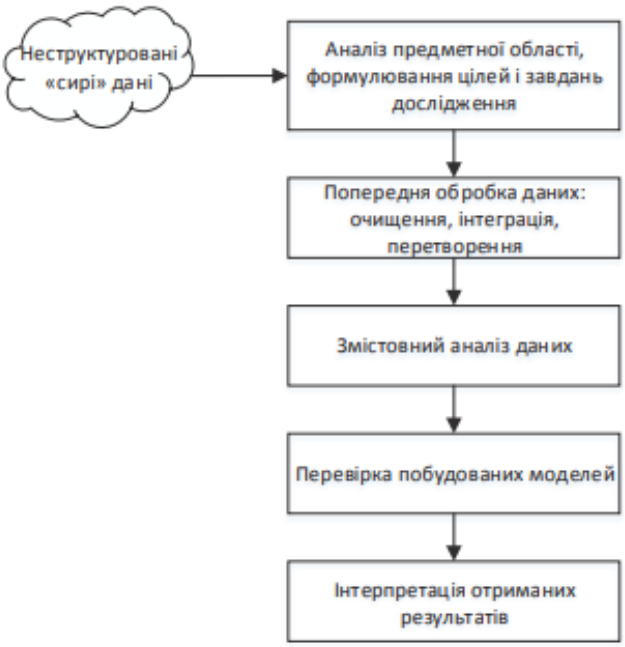


Рисунок 1.3 – Етапи інтелектуального аналізу даних клієнтів банку

Використання методів Data Mining під час обробки даних клієнів банківської сфери дозволить [27]:

– краще утримувати клієнтів і уникнути непотрібних дій;

– складати профілі клієнтів і розуміти їх поведінку;

– підтримувати і підвищувати рівень прибутку;

– скоротити витрати клієнтів при покупках;

– виходити на клієнтів з потрібними їм пропозиціями.

## 1.4 Організація та створення персоналізованих пропозицій для клієнтів банківської сфери

Практика показує, що фінансові установи не тільки можуть, але й зобов'язані змінювати механізми обслуговування споживачів. Тому в маркетингу була створена клієнт-орієнтована технологія, яка отримала назву «таргетована реклама» [5]. Сутність поняття полягає у створенні персоналізованих рекламних пропозицій, що максимально відповідають запитам і потребам користувачів.

Пошукові технології вже деякий час використовують обробку великих даних та машинне навчання для покращення результатів пошуку, але лише нещодавно банки зрозуміли, що ці самі методи можна використовувати для персоналізації.

Data Mining в банківському маркетингу – кращий спосіб для роботи з даними клієнтів. Перш за все, тому, що він може створювати унікальні персоналізовані пропозиції, що виходять за рамки звичайного продажу товару, і підтримувати купівельну лояльність протягом максимально тривалого періоду.

Користування даною технологією розширює межі банківської сфери, адже банк володіє основним ресурсом, якого немає у Google, Facebook і у тисячі Інтернет-магазинів — доступом до рахунків клієнта [28]. Банки не мають права відкривати іншим до такої конфіденційної інформації, як стан рахунку клієнта, але можуть користуватися нею для власних потреб і заробляти на співпраці з комерційними компаніями.

Одну з основних ролей при створенні персоналізованих пропозицій відіграють соціальні мережі. Вони надають банку додаткові ресурси, що при правильному застосуванні, можуть сприяти розширенню бази клієнтів та

удосконаленню власної діяльності, за рахунок дослідження відгуків про переваги та недоліки запропонованих банківських послуг [29].

Крім того, «лайкаючи» новини, фотографії, імпульсивно підписуючись на групи за інтересами, додаючи ту чи іншу пісню в плейлист, користувач розповідає про свої уподобання. Тому профіль клієнта банку в соціальній мережі може використовуватись комерційними банками для визначення його інтересів, а також рівня надійності та платоспроможності. Адже в такий спосіб можна отримати інформацію про те, скільки разів на рік людина їздить відпочивати, які ресторани відвідує, в якому колі спілкується, де живе, яким транспортом користується тощо.

Створення персоналізованих пропозицій означає відсутність масових пропозицій на загальних умовах для всіх одержувачів послуги. Наявність повноцінної інформації про клієнтів, як наявних, так і потенційних, дає змогу банку вибудовувати персоналізовані послуги, при цьому зберігаючи певну «конвеєрність» їх розповсюдження, що мінімізує витрати банку, а найголовніше – дає змогу не втрачати клієнтів.

Чим більше інформації про клієнта є у банку, тим легше буде розробити персоналізовану пропозицію. Так,наприклад, на основі даних клієнта, у якого професійною діяльністю є IT-підприємництво можна створити персоналізовану пропозицію – картку для IT-підприємців. За фіксовану плату банк бере на себе всі бухгалтерські турботи: починаючи від відкриття ФОПу й завершуючи повним аутсорсингом бухобліку. Або банк бачить, що машині клієнта вже більше трьох років, і пропонує йому пільговий автокредит.

Важливо, щоб пропозиції банку приходили через найзручніший для клієнта канал комунікацій. Це можуть бути розсилки, SMS, push-повідомлення, повідомлення в месенджерах.

Можна виділити п'ять типових рівнів наявності і використання аналітичних даних про клієнтів для персоналізації пропозицій [30]:

– «Килимове бомбардування». Всі клієнти отримують практично ідентичні пропозиції через один або кілька каналів. Обсяг даних про клієнтів у

банку на даному етапі обмежений, і лише незначна частина даних використовується для розробки конкретних дій;

– «Широкими мазками». Пропозиція і стратегія взаємодії диференціюються з урахуванням сегментації (як правило, за критеріями віку і рівня доходу) та наявності у клієнта тих чи інших продуктів. Інформаційна взаємодія в основному здійснюється через активні канали продажів - колл-центр, відділення, що виходить розсилку листів, електронну пошту і SMS.

– «Мікросегментація». На цьому етапі банк виділяє десятки сегментів клієнтів з урахуванням їх прибутковості, переваг і поведінкових стереотипів. Пропозиції для кожного мікросегменту персоналізуються і пріорітизуються методами математичного моделювання.

– «Персональний підхід». Від аналізу сегментів банк переходить до аналізу кожного конкретного клієнта. Замість активних продажів банк прагне конвертувати обслуговування в продажах. У будь-який момент часу для кожного клієнта в кожному з каналів є спеціально адаптована індивідуальна пропозицію.

– «Динамічна персоналізація». Для аналізу використовуються не тільки традиційні дані - демографічні, транзакційні, продуктові, поведінкові, - а й інформація з соціальних мереж. Уподобання клієнтів і схильність до покупки певних продуктів враховуються при будь-якій взаємодії з клієнтом по всіх каналах. Пропозиції динамічно коригуються з урахуванням формату та історії взаємодії з клієнтом.

Більшість українських фінансових інститутів знаходяться на другому рівні, і лише деякі з них переходять на третій.

У банківській сфері можна виділити такі види таргетованої реклами, що використовуються для продажу персоналізованих пропозицій [31]:

– радіусний таргетинг. Оголошення пропонуються тим клієнтам, які перебувають у встановленому певному радіусі поблизу банківських відділень. Цей вид таргетингу надає можливість знайти клієнтів, які в даний момент знаходяться в пошуках банківських послуг або перебувають біля відділення.

– таргетинг за періодом розміщення рекламних повідомлень (ранок чи вечір – будні або вихідні) скорочує число нерезультативних оголошень. Для банківських установ оптимально виставляти рекламу під час робочих годин відділень у будні дні.

– поведінковий таргетинг є найбільш результативним. Передбачає відбір інформації про дії користувача, використовуючи cookie-файли: особисті дані, хобі, інтереси, покупки, пошукові запити тощо. Усі ці дані дозволяють сформувати персоналізовані пропозиція для користувача.

Таким чином, розробка нових методів обробки та аналізу персональної інформації стають основним джерелом підвищення успішності маркетингової діяльності банків. В умовах диджиталізації суспільства, Інтернет та інші цифрові канали змінюють форми і методи маркетингової діяльності в банківській сфері, стають поштовхом для розвитку ефективної форми маркетингової діяльності – персоналізованого маркетингу.

## 1.5 Постановка задачі

Для досягнення мети атестаційної роботи, необхідно виконати наступний ряд завдань:

- проаналізувати ринок банківських продуктів і послуг;
- проаналізувати методи ІАД;
- проаналізувати методи створення персоналізованих пропозицій;
- проаналізувати джерела отримання інформації про клієнта;
- обрати методи ІАД для комбінування;
- розробити комбінований метод на основі обраних алгоритмів Data Mining;
- спроектувати автоматизовану інформаційну систему, що використовує розроблений метод;
- провести тестування та оцінку результатів розробленого методу.

Для розробки методу інтелектуального аналізу даних щодо створення персоналізованих пропозицій необхідно [32]:

- обрати критерії та коефіцієнти для побудови найбільш правдивої моделі класифікації даних;
- проаналізувати та вибрати інструменти для розробки методу;
- обрати методику на основі комбінування алгоритмів Data Mining.

Також для ефективного використання даного методу необхідно спроектувати та розробити систему, яка зменшить поріг входу для використання методу. Для проектування системи необхідно:

- розробити системні та функціональні вимоги до системи, що проектується;
- розробити діаграму варіантів використання;
- провести обґрунтування мови програмування, що буде використовуватися для створення інформаційної системи та обґрунтування СУБД;
- провести логічне та фізичне моделювання даних з реалізацією бази даних;
- визначити інструменти розробки системи, що проектується,
- визначити архітектуру системи.

# 2 ВИБІР АЛГОРИТМІВ DATA MINING ДЛЯ РОЗРОБКИ КОМБІНОВАНОГО МЕТОДУ

## 2.1 Метод пошуку асоціативних правил. Особливості алгоритму Apriori

Асоціативні правила являють собою механізм знаходження логічних закономірностей між пов'язаними елементами (подіями або об'єктами) у великих базах даних. Вперше цей метод було запропоновано для пошуку типових шаблонів покупок, що здійснюються в супермаркетах.

Так як зазвичай людина купує не один товар, а набор товарів, виникає можливість створення персоналізованих рекламних пропозицій на основі аналізу її продуктового кошику. Під аналізом кошику розуміють процес аналізу наборів даних, ціллю якого є визначення комбінацій продуктів, що пов'язані один з одним. Тобто, проводиться пошук продуктів, наявність яких у відповідній транзакції впливає на можливість присутності у ній інших продуктів або їх комбінацій.

Транзакція – це певна множина випадків, які розвиваються спільно [32]. Типовою транзакцією є придбання клієнтом банківського продукту або послуги.

Асоціативні правила складаються із двох необхідних складових частин, які називаються antecedent (умова), що позначено як X, та consequent (наслідок), позначений як Y.

Нехай даний набір $I = \{i_1, i_2, \dots i_n\}$ із $n$-двійкових атрибутів, що називаються об’єктами та набір $D = \{t_1, t_2, \dots t_n\}$ транзакцій, який є базою даних. Тоді асоціативне правило буде мати вигляд: $X \rightarrow Y$, де $X, Y \subseteq I$.

Основними показниками, які описують асоціативні правила є support (підтримка), confidence (довіра) та improvement (покращення).

Support (Sup) – це показник того, як часто елемент або набір елементів з’являється у наборі даних [33]. Наприклад, support для асоціації A → B можна записати таким чином (формула 2.1):

$$Sup(A \rightarrow B) = P(A \cap B) = \frac{n(\{A; B\} \in d_i)}{N} \quad (2.1)$$

де $P(A \cap B)$ – ймовірність події, де продукти A і B будуть придбані разом.

Confidence (Conf) є показником того, як часто правило виявляється істинним і позначається через відношення кількості транзакцій, що містять умову і наслідок, до числа транзакцій, у яких міститься тільки умова (формула 2.2):

$$Conf(A \rightarrow B) = P(A|B) = \frac{n1(\{A; B\} \in d_i)}{n1(\{A\} \in d_i)} \quad (2.2)$$

Improvement (Imp)– відношення ймовірності правила до ймовірності результата [23]. Тобто воно показує, у скільки разів частіше від випадкового вгадування, правило, що розглядається, видає правильний прогноз (формула 2.3).

$$Imp(A \rightarrow B) = \frac{P_A(B)}{P(B)} = \frac{C(A \rightarrow B)}{S(B)} \quad (2.3)$$

Якщо L >1, це дає нам зрозуміти, в якій мірі ці два випадки залежать одне від одного, дозволяє використати ці правила для прогнозування наслідків у наступних наборах даних.

У випадку якщо L < 1, стає зрозуміло, що елементи заміщують одне одного. Іншими словами, існування одного предмета негативно позначається на наявності іншого товару і навпаки.

Для розширення показників асоціації можна використати таку характеристику, як інформативність асоціації (формула 2.4), що запропонована у роботі [34] і розраховується на основі рівнів support, confidence та improvement:

$$
\begin{aligned}
I_{A\leftrightarrow B} = {} & Sup(A \to B) \cdot log\big(Imp(A \to B)\big) + \\
& + Sup(A \to \bar{B}) \cdot log\big(Imp(A \to \bar{B})\big) + \\
& + Sup(\bar{A} \to B) \cdot log\big(Imp(\bar{A} \to B)\big) + \\
& + Sup(\bar{A} \to \bar{B}) \cdot log\big(Imp(\bar{A} \to \bar{B})\big)
\end{aligned}
\tag{2.4}
$$

Інформативність правила (повна взаємна інформація) може використовуватися в якості обмеження під час генерації асоціативних залежностей при виділенні знань із даних.

Алгоритм Apriori використовує часті набори компонентів для генерації правил асоціації і призначений для роботи з БД, у яких містяться транзакції. Даний алгоритм визначає компоненти, які найчастіше трапляються у базі й збільшує їх перелік до ще більших множин, дотримуючись правила достатньої частотності.

Apriori застосовує підхід "знизу – вгору", коли часті підмножини розширюються по одному компоненту (етап, під назвою «генерація кандидатів»), а групи кандидатів контролюються на відповідність даним [19]. Процес роботи алгоритму припиняється, коли можливості для наступних ефнктивних розширень не знайдено.

На табл. 2.1 представлені переваги та недоліки використання алгоритму для системи маркетингового аналізу у банківській сфері.

Таблиця 2.1 – Переваги та недоліки алгоритму Apriori

| Переваги | Недоліки |
|---|---|
| – простота | – багаторазове сканування базового набору |
| – швидке зменшення числа згенерованих кандидатів, при установці високої мінімальної підтримки або відносно розрідженому базовому наборі | – велика кількість згенерованих кандидатів, при занадто великому наборі даних або при дуже низькій підтримці |

Генерація правил асоціації за алгоритмом Apriori відбувається у чотири кроки і зображена на рис. 2.1

Рисунок 2.1 – Кроки алгоритму Apriori для генерації асоціативних правил

Робота алгоритму, розглянута на прикладі бази даних представлена на рис 2.2. Мінімальний рівень support дорівнює 3. TID є унікальним ідентифікатором, який позначає відповідну транзакцію. Itemset – набір елментів.

На початку формуються одноелементні кандидати. Потім алгоритм підраховує support одноелементних наборів. Набори з показниками support менше встановленого, тобто менше 3, відсіюються. У даному прикладі це набори e і f, у яких support рівний 1. Набори товарів, що залишилися вважаються одноелементними наборами товарів, що часто зустрічаються: це набори a, b, c, d.

Далі відбувається формування двохелементних кандидатів, підрахунок їх support та відсуювання наборів з показниками support, меншими 3.

Двоелементні набори товарів, що залишилися, вважаються двоелементними наборами, які часто зустрічаються, тобто ab, ac, bd приймають участь у подальшій роботі алгоритму.

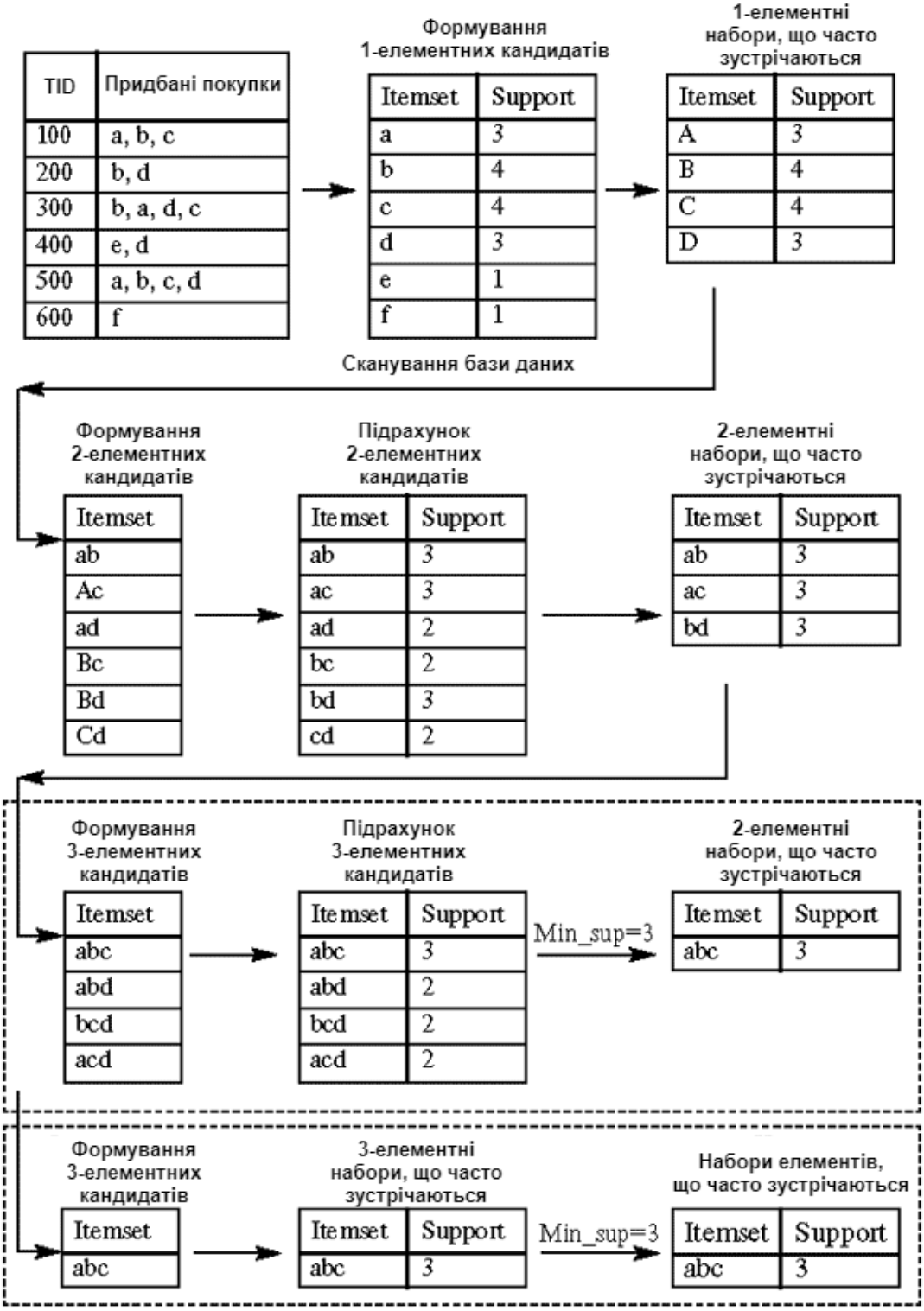


Рисунок 2.2 – Алгоритм Apriori

Якщо дивитися на роботу алгоритму прямолінійно, на останньому етапі алгоритм формує триелементні набори товарів: abc, abd, bcd, acd, підраховує їх support та відсіює набори з рівнем support менше 3. Набір товарів abc може бути названий набором, що часто зустрічається.

Однак алгоритм Apriori зменшує кількість кандидатів, апріорі відсіюючи тих, які явно не можуть стати тими, що часто зустрічаються, на основі інформації про відсіяних кандидатів на попередніх етапах роботи алгоритму.

Відсіювання кандидатів відбувається на основі припущення про те, що у набору товарів, який часто зустрічається, всі підмножини повинні бути такими, що часто зустрічаються. Якщо в наборі наявна підмножина, що на попередньому кроці визначилась як та, що рідко трапляється, то цей кандидат вже не включається у формування і підрахунок кандидатів.

Тому набори товарів ad, bc, cd були відсіяні як ті, що зустрічаються нечасто і алгоритм не розглядав набір товарів abd, bcd, acd. За умови розгляду цих наборів, формування триелементних кандидатів відбувалося б за схемою, наведеною в верхньому пунктирному прямокутнику. Оскільки алгоритм апріорі відкинув, які завідомо нечасто зустрічаються, останній етап алгоритму визначив набір abc як єдиний триелементний набір, що часто зустрічається (етап наведено в нижньому пунктирному прямокутнику).

## 2.2 Метод логістичної регресії

Логістична регресія – це статистичний метод, що використовується для моделювання двійкової залежної змінної за допомогою логістичної функції [24]. Вона використовується, коли залежна змінна має двійковий характер (0/1, True / False, Так / Ні) .

Логістична регресія також відома як біноміальна логістична регресія. За допомогою логістичної регресії можна оцінити ймовірність настання або ненастання події: пацієнт хворий чи здоровий, позичальник повернув позику або дозволена затримка тощо. Завдяки цьому логістичну регресію можна розглядати як потужний інструмент підтримки прийняття рішень.

Модель логістичної регресії використовує логістичну функцію для моделювання залежності вихідної змінної від набору вхідних у випадку, коли

вихідна змінна є бінарною [26]. Статистична модель логістичної регресії відображена у формулі (2.5):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \qquad (2.5)$$

де $p$ – це біноміальна пропорція;

$x$ – пояснювальна змінна;

$\beta_0$ та $\beta_1$ – параметри моделі логістичної регресії.

На рис. 2.3 представлені графіки залежності $p$ від $x$ для різних значень $\beta_0$ і $\beta_1$.

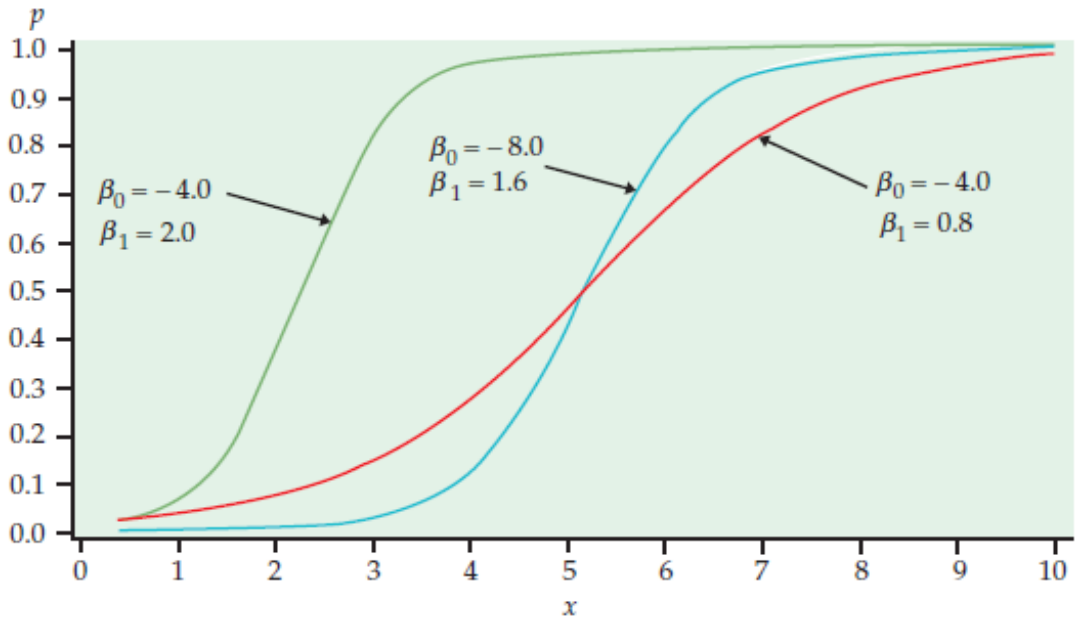


Рисунок 2.3 – Залежність $p$ від $x$ для різних моделей логістичної регресії

Переваги та недоліки методу біномінальної логістичної регресії зображені у табл. 2.2

Таблиця 2.2 – Переваги та недоліки методу логістичної регресії

| Переваги | Недоліки |
|---|---|
| – легка у впорядкуванні, інтерпретації та навчанні | – не дуже добре опрацьовує велику кількість категорійних змінних |
| – не вимагає занадто великої кількості обчислювальних ресурсів | – необхідне перетворення нелінійних функцій |

Продовження таблиці 2.2

| | |
|---|---|
| – не тільки передбачає міру доречності предиктора (величина коефіцієнта), а й напрямок його асоціації (позитивний чи негативний) | – без правильної ідентифікації незалежних змінних логістична регресія не може працювати правильно |
| – не вимагає масштабування вхідних функцій | – недостатньо гнучка, щоб природно охоплювати більш складні відношення |
| – через свою простоту є хорошою базою для вимірювання продуктивності інших більш складних алгоритмів | – може використовуватися лише для прогнозування дискретних функцій. Тому, залежна змінна логістичної регресії прив’язана до дискретного набору чисел |
| – має хорошу точність для багатьох простих наборів даних | – не вирішує нелінійні задачі, так як її поверхня прийняття рішень є лінійною |
| – логістичні моделі можуть бути легко оновлені новими даними з використанням стохастичного градієнтного спуску | – не буде добре працювати з незалежними змінними, які не пов'язані з цільовою змінною і дуже схожі або пов'язані один з одним |

## 2.3 Метод дерева рішень. Алгоритм Random Forest

Дерева рішень є одним з найбільш ефективних методів інтелектуального аналізу даних. Переваги та недоліки цього засобу підтримки прийняття рішень представлено у таб. 2.3

Таблиця 2.3 – Переваги та недоліки методу дерева рішень

| Переваги | Недоліки |
|---|---|
| – не потребує попередньої обробки даних (їх нормалізації та масштабування) | – невелика зміна у даних може спричинити значні зміни в структурі дерева рішень |
| – Висока надійність, точність та ефективність | – проблематичність побудови оптимального дерева рішень |
| – відсутність змінних у даних суттєво не впливають на процес побудови дерева рішень | – навчання дерева рішень є відносно дорогим, оскільки складність та час забирають більше ресурсів |
| – є можливість інтуїтивного розуміння та інтерпретації | – вірогідність створення занадто складного дерева |

Структура дерева рішень складається з компонентів двох типів: вузлів (node) і листя (leaf). У вузлах містяться вирішальні правила і проводиться

перевірка відповідності прикладів цьому правилу з якого-небудь атрибуту навчальної множини [20]. У найпростішому варіанті, в результаті перевірки, множина прикладів, що потрапила до вузла, ділиться на дві підмножини, до одної потрапляють приклади, що задовольняють правило, а до іншої – що не задовольняють.

Приклад структури дерева рішень при класифікації заявки клієнта, щодо надання кредиту зображений на рис 2.4

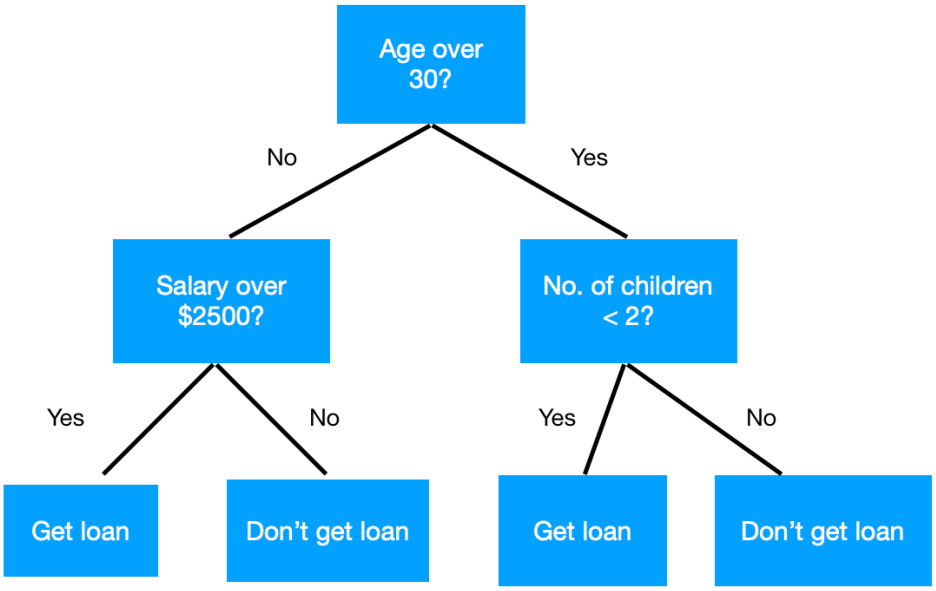


Рисунок 2.4 – Приклад структури дерева рішень

Потім до кожної підмножини знову застосовується правило і процедура рекурсивно повторюється поки не буде досягнуто деякої умови зупинки алгоритму. В кінцевому випадку, в останньому вузлі перевірка та розбиття не проводяться і його оголошують листом. Лист визначає рішення для кожного, потрапившого до нього прикладу.

При побудові дерева класифікації, для створення розгалуження в черговому вузлі, необхідно вибрати атрибут, за яким це буде зроблено. Для цього застосовують критерії оцінки способу розгалуження (вибору ознаки для

розгалуження і вибору порога для цієї ознаки). Найбільш популярними критеріями є інформаційна ентропія (формула 2.6) та статистичний індекс Gini.

Під ентропією розуміють "ступінь невизначеності" або "ступінь хаосу" в системі [22]. Завдання дерева рішень –ввести порядок в набір значень цільової змінної. Отже, їх метою буде максимальне зменшення ентропії.

$$\mathrm{H} = -\sum_{i=1}^{n} \frac{N_i}{N} \log\left(\frac{N_i}{N}\right) \tag{2.6}$$

де $n$ – число класів у вихідній підмножині;

$N_i$ – число прикладів і-го класу;

$N$ – загальна кількість прикладів у підмножині.

Коли класи представлені в рівних частинах і невизначеність класифікації найбільша, ентропія також максимальна. Якщо всі приклади у вузлі відносяться до одного класу, тобто $N = N_i$, логарифм від одиниці перетворює ентропію в нуль.

Індекс Gini може бути розрахований за формулою (2.7). Сенс даного показника в тому, що він показує - наскільки часто випадково обраний приклад навчальної вибірки буде розпізнано неправильно, за умови, що цільові значення в цій вибірці були взяті з певного статистичного розподілу.

Таким чином індекс Gini фактично показує відстань між двома розподілами - розподілом цільових значень і розподілом передбачень моделі [21].

$$\mathrm{Gini}(Q) = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.7}$$

де $Q$ – множина результатів;

$n$ – число класів у ньому;

$p_i$ – ймовірність і-того класу (виражена як відносна частота прикладів відповідного класу).

На сьогодні, існує значна кількість алгоритмів навчання дерева рішень: ID3, CART, C4.5, C5.0, NewId, ITrule, CHAID, CN2 і т.д. У даній атестаційній роботі

для вирішення задач класифікації був обраний алгоритм Random Forest. Його переваги та недоліки представлені у таб. 2.4

Таблиця 2.4 – Переваги та недоліки алгоритму Random Forest

| Переваги | Недоліки |
|---|---|
| – добре працюють з великою кількістю елементів даних | – для реалізації алгоритму Random Forest потрібно більше обчислювальних ресурсів |
| – масштабування даних не вимагає в алгоритмі випадкового лісу. Він зберігає високу точність навіть після надання даних без масштабування | – процес прогнозування більш трудомісткий у порівнянні з іншими алгоритмами |
| – мають меншу дисперсію, ніж одне дерево рішень | – менш інтуїтивно зрозуміло в разі, коли наявна є велика колекція дерев рішень |
| – дуже гнучкі і мають дуже високу точність | – вірогідність побудови занадто складного дерева |
| – підтримують високу точність навіть при відсутності значної частини даних | |

Random Forest складається з великої кількості окремих дерев рішень, які діють як ансамбль. Даний алгоритм створює дерева рішень на зразках даних, а потім отримує прогноз від кожного з них і, нарешті, вибирає найкраще рішення за допомогою голосування. Кожне дерево в класифікаціях бере вхідні дані із зразків початкового набору даних. Потім випадковим чином вибираються ознаки, які використовуються для створення дерева на кожному вузлі.

Процес навчання алгоритму Random Forest представлений в наступних кроках та зображений на рис 2.5 :

- крок 1 – обираються випадкові вибірки із заданого набору даних;
- крок 2 – алгоритм будує дерева рішень для кожного зразка та отримує результати прогнозування з кожного дерева рішень;
- крок 3 – голосування проводиться для кожного прогнозованого результату;
- крок 4 – обирається результат з найбільшою кількістю голосів як підсумковий результат голосування.

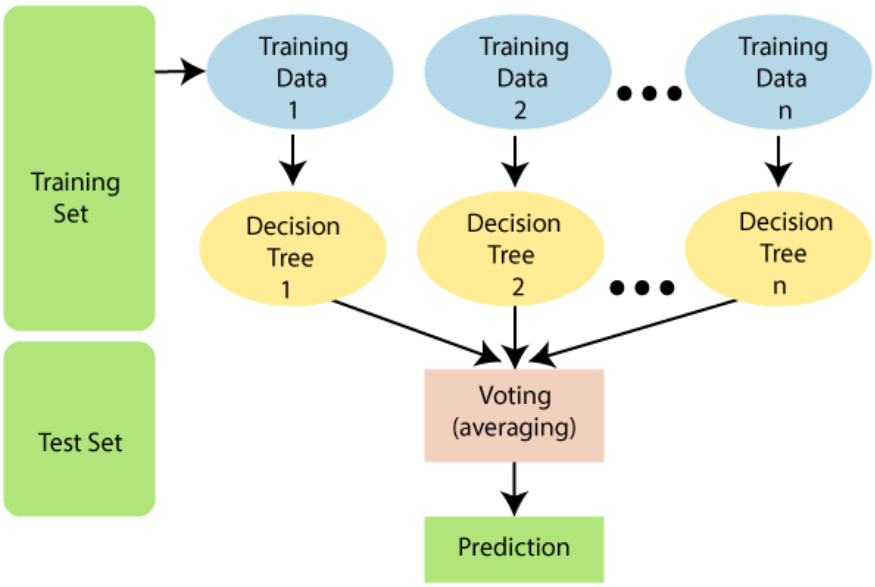


Рисунок 2.5 – Навчання алгоритму Random Forest

Після проходження алгоритмом навчання він готовий для обробки нових даних і створення нових прогнозів.

## 2.4 Побудова методу класифікації на основі комбінування обраних алгоритмів

У випадку комбінованого використання методів, було проведено дослідження щодо маркетингових політик банку та загалом. В деяких випадках, наприклад, коли клієнт новий і в нього ще немає ніяких покупок щодо банківських продуктів та послуг, можна покластися на класифікатор клієнтів, та віднести клієнта за його ознаками. З цим можуть допомогти методи Random Forest та Logistic regression. Якщо брати до уваги клієнтів, що вже робили покупки продуктів та послуг, то з ними можна провести більш детальний аналіз даних. Наприклад проаналізувати продукти та послуги, які були придбані цими клієнтами, за допомогою методу Apriori.

На рис. 2.6 представлений алгоритм навчання методу аналізу даних клієнтів банку.

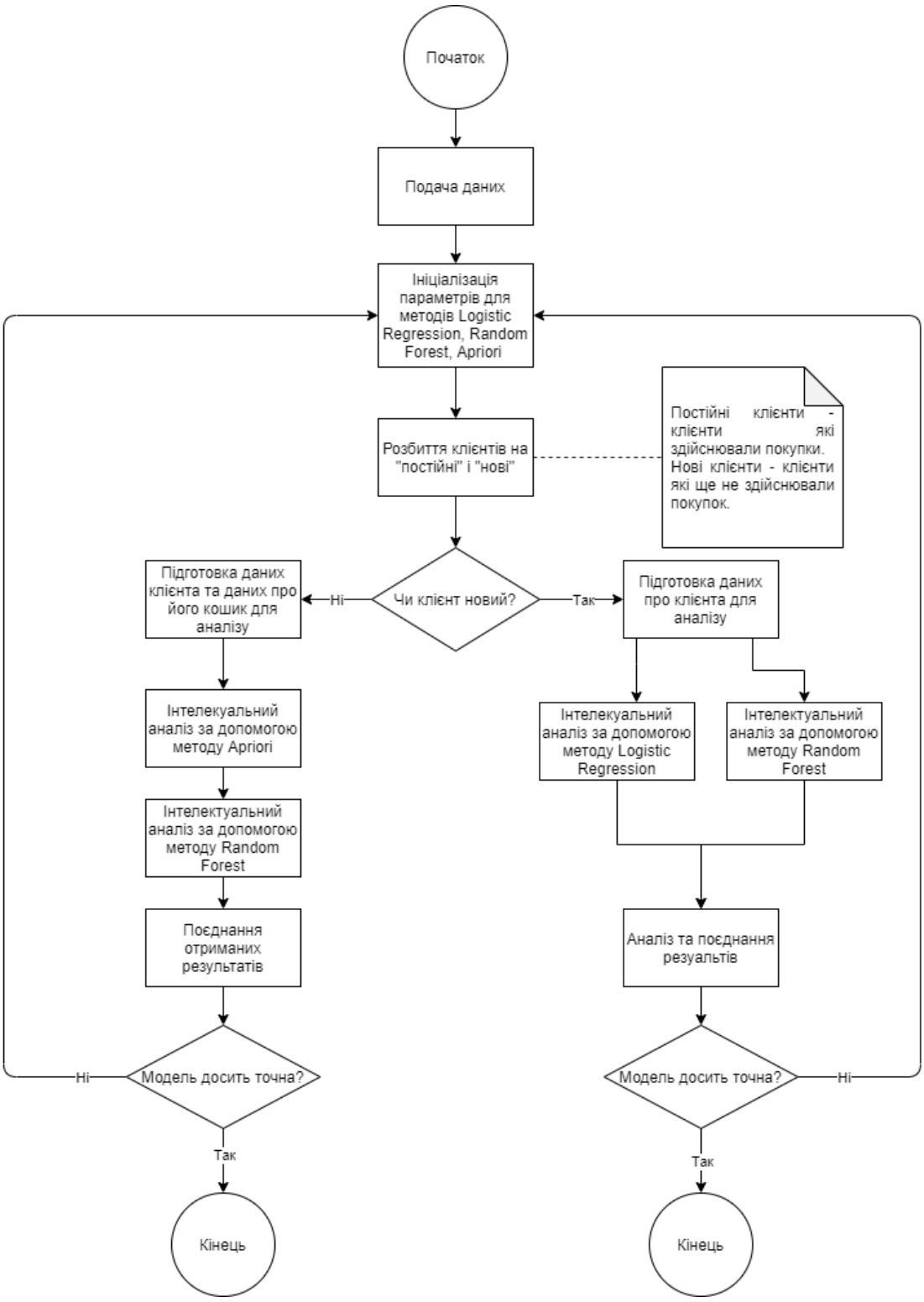


Рисунок 2.6 – Алгоритм навчання методу на основі комбінування методів Apriori, Random Forest та логістичної регресії

Для початку, проходить подача даних про клієнтів до методу. Далі здійснюється огляд даних на предмет пропусків, неможливих значень, після чого

проводиться оцінка даних та ініціалізація параметрів для комбінованого використання методів Random Forest, Logistic regression та методу Apriori.

В рамках поєднання методів, було вирішено, що необхідно розділити логіку навчання для різних видів клієнтів. Було виділено два види:

- постійні клієнти – клієнти які здійснювали покупки;
- нові клієнти – клієнти які ще не здійснювали покупок.

Розділення необхідно, для того щоб долучити до аналізу метод Apriori, що спеціалізується на аналізі кошиків клієнтів. Так як відомо, що у «постійних» клієнтів вже є покупки, то можна на основі цих даних сформувати пропозиції, які можу знадобитися клієнтові у майбутньому. Крім цього, його дані також будуть проаналізовані за допомогою іншого методу, а саме Random Forest. Тож максимально проаналізувавши «постійного» клієнта, ми отримали рекламну групу, яку виділив Apriori та рекламну групу яку виділив Random Forest, об’єднавши їх, отримуємо пакет продуктів які можна запропонувати клієнту.

Щодо «нових» клієнтів, до моделі не надходять дані, щодо його минулих покупок, тож необхідно опиратися на інші ознаки. Для того, щоб видати найбільш рекламу більш точніше, було вирішено використати схему під назвою Stacking [35], та проаналізувати дані за допомогою методів Random Forest та Logistic regression. На вхід до кожного класифікатора подаються однакові дані, далі проходить навчання обох методів. Після того, як дані з тестової вибірки були проаналізовані обома методами, дані передаються далі, де вони аналізуються та поєднуються в одне рішення, за допомогою методу найбільшої кількості голосів.

# 3 РОЗРОБКА ФУНКЦІОНАЛЬНИХ ВИМОГ СИСТЕМИ

## 3.1 Характеристика основної функції системи

Для розробки функціональних вимог системи, було обрано стандарт методології IDEF0 [36]. Стандарт IDEF успішно використовується для представлення функціонального каркасу системи. Розробка функціональних вимог дозволяє полегшити розуміння системи для замовника, та виділити основні функціії системи.

В рамках атестаційної роботи були виділені основні функції системи, та представлені у вигляді концептуальної діаграми (рисунок 3.1) та декомпозиціії першого рівня (рисунок 3.2).

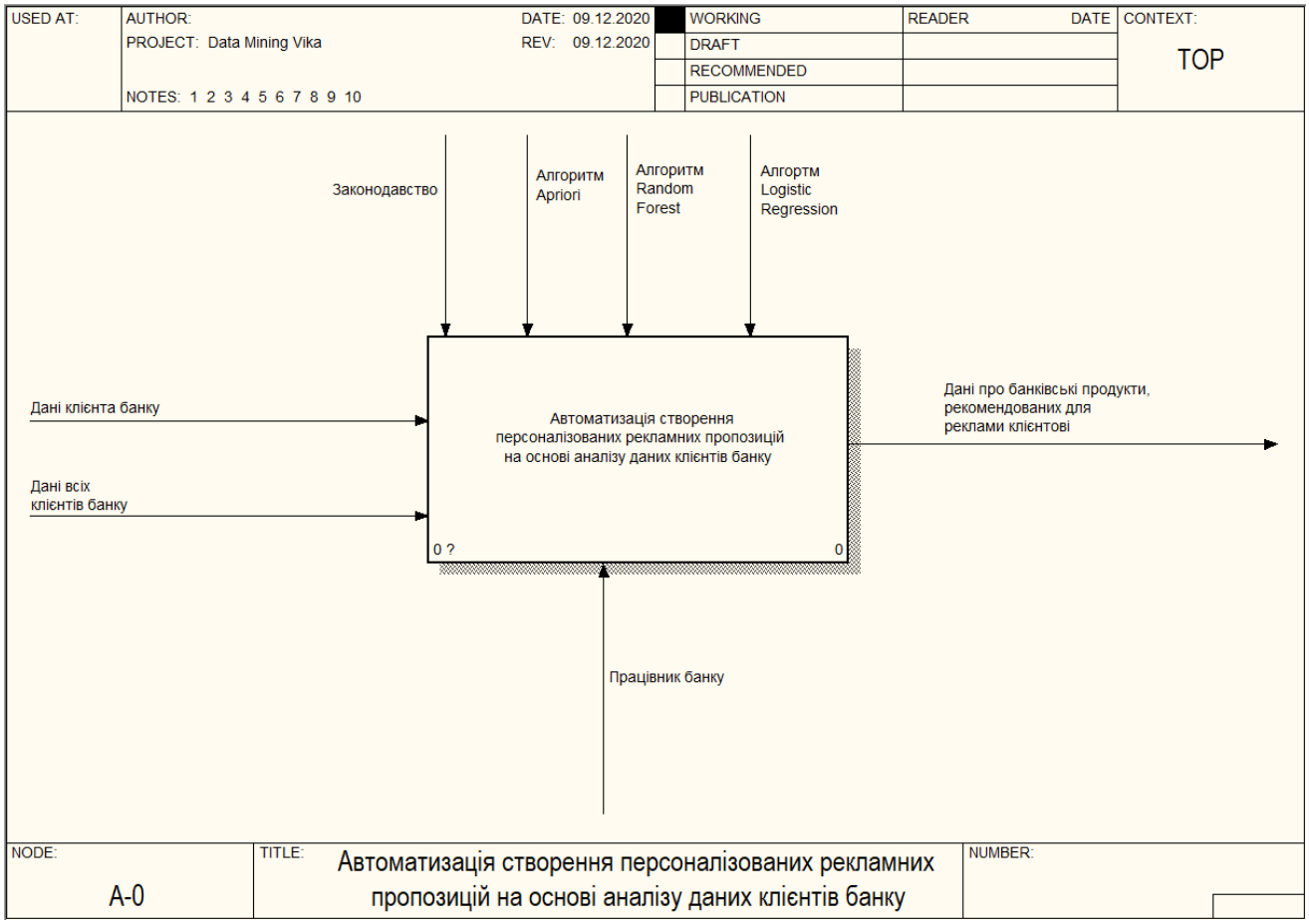


Рисунок 3.1 – Концептуальна діаграма функції «Автоматизація створення персоналізованих пропозицій на основі аналізу даних клієнтів банку»

Основна функція проектованої системи – «автоматизація створення персоналізованих пропозицій на основі аналізу даних клієнтів банку». На вхід до системи подаються дані про клієнта банку та дані інших клієнтів банку. Система керується законодавством та використовує в собі алгоритми: Apriori, Random Forest, Logistic Regression. Для внесення даних до системи, було обрано співробітника банку. На виході система представляє дані про банківські продукти, що рекомендовані для таргетованої реклами.

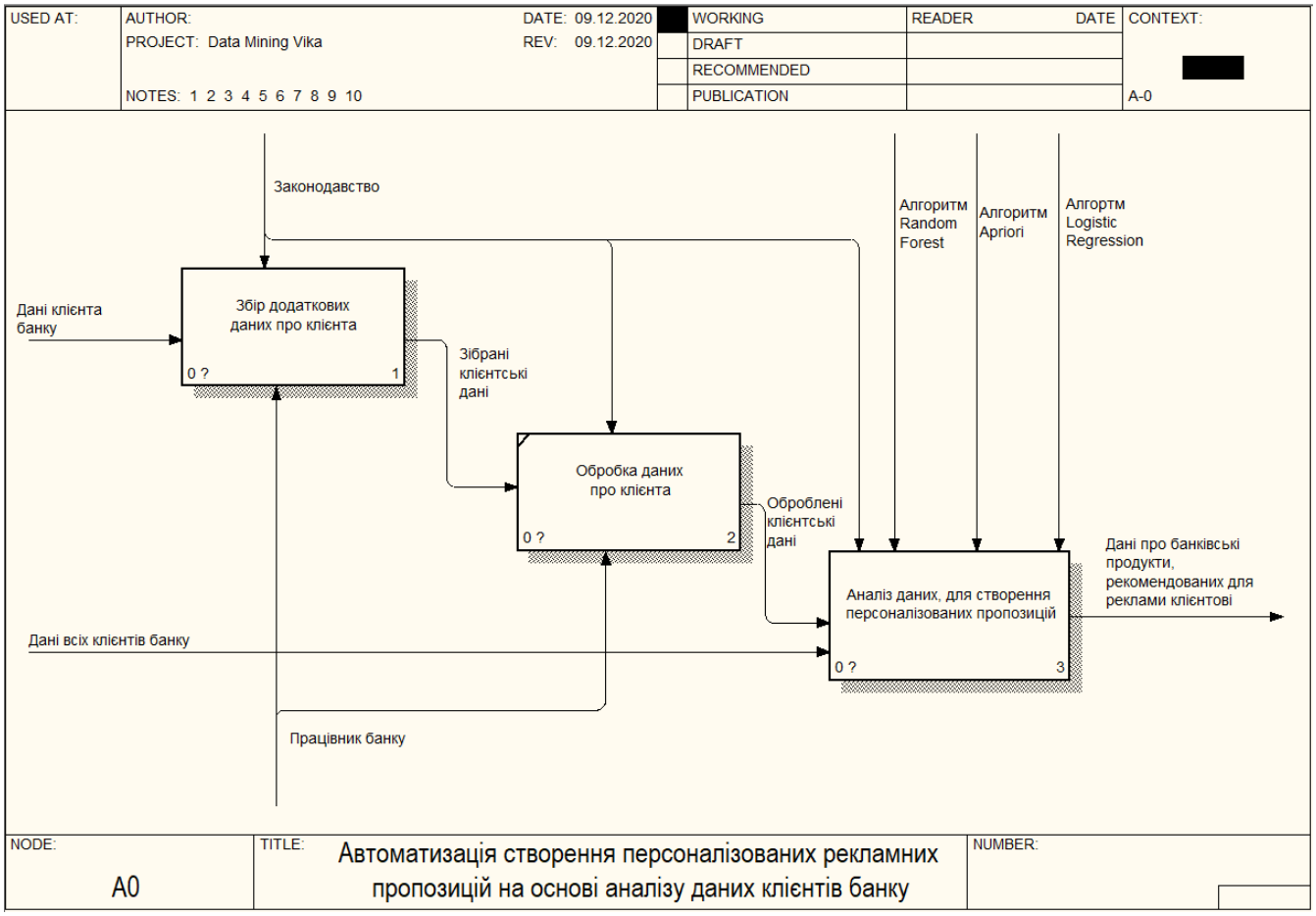


Рисунок 3.2 – Діаграма декомпозиції, що представляє основні функції системи

Основна функціональна сутність декомпозується на 3 функції: «Збір додаткових даних про клієнта», «Обробка даних клієнта» та «Аналіз даних для створення персоналізованих пропозицій».

## 3.2 Характеристика функції системи «Збір додаткових даних про клієнта»

Спочатку система отримує на вхід початкові дані про клієнта, що були надані у ході його реєстрації, а саме ідентифікаційні та контактні дані. Далі до системи вносяться дані за допомогою співробітника банку, та виконується збір даних на додаткових сервісах, наприклад у соціальних мережах, маркетингових мережах та інших. На рисунку 3.3 представлена декомпозиція функції «Збір додаткових даних про клієнта».

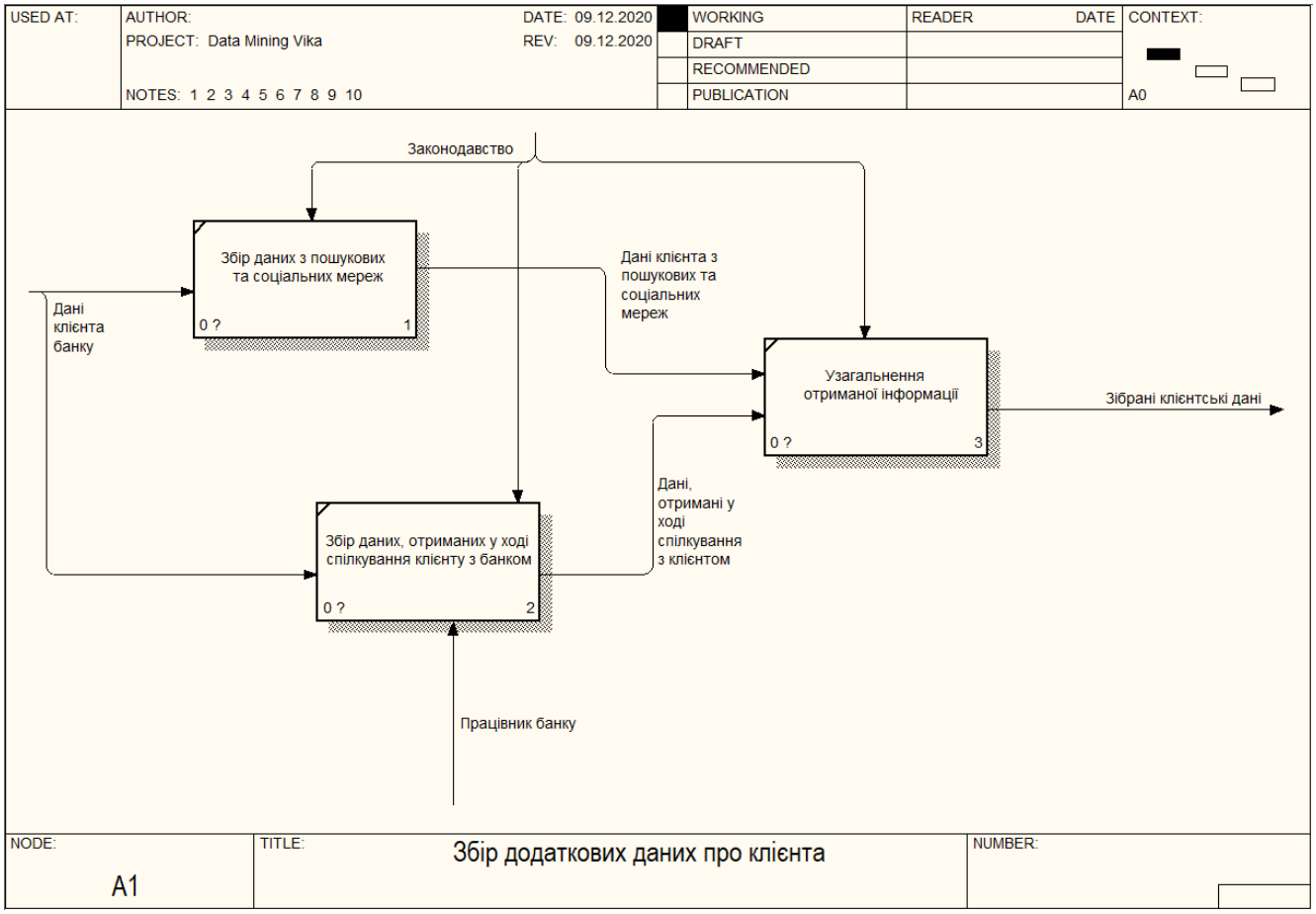


Рисунок 3.3 – Діаграма декомпозиції функції «Збір додаткових даних про клієнта»

Функція «Збір додаткових даних про клієнта» декомпозується на наступні функції: «Збір даних з пошукових та соціальних мереж», «Збір даних, отриманих у ході спілкування з клієнтом» та «узагальнення отриманої інформації».

Нижче представлений перелік інформації, яка необхідна для корректної роботи системи, та яку збирають функції описані вище:

- ідентифікаційні та контактні дані;
- фінансова інформація;
- інформація щодо освіти та сімейного стану;
- інформація щодо професійної діяльності клієнта;
- договірні зобов’язання в банку;
- інформація щодо отримуваних послуг;
- інформація про участь у акціях, розіграшах, лотереях банку;
- інформація, отримана у ході спілкування клієнту з банком;
- інформація від пошукових та соціальних мереж;
- дані щодо геолокації;
- інформація, що міститься в документах, представлених банку тощо.

На рисунку 3.4 показаний блок збору інформації, та дані які використовує система.

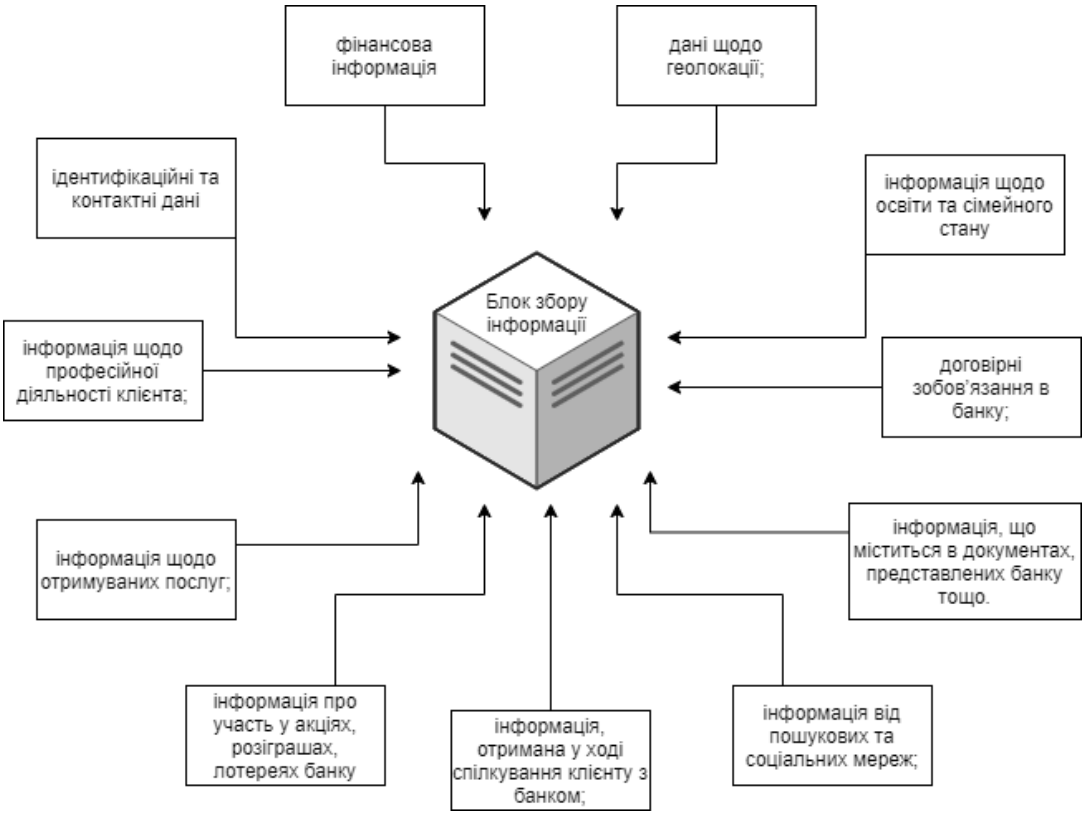


Рисунок 3.4 – Блок збору інформації

## 3.3 Характеристика функції «Обробка даних клієнта»

Одним з основних етапів побудови моделі є попередня обробка необроблених даних. Кожна ознаку, яку алгоритм аналізу має опрацювати, повинна бути представлена у правильному вигляді. Дана функція системи оброблює дані про клієнта що подаються на вхід. Підготовка даних – один з найважливіших етапів в Data Mining за методолгією CRISP-DM (Cross-Industry Standard Process for Data Mining) [37], концепція якої зображена на рисунку 3.5

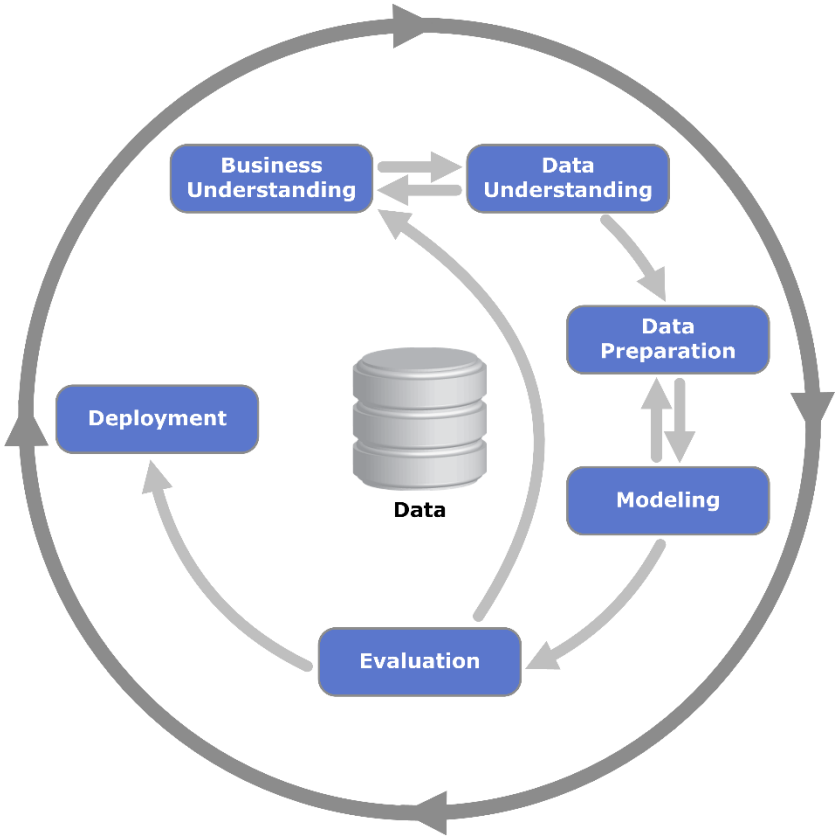


Рисунок 3.5 – Концепція методології CRISP-DM

Етапи Data Preparation та Modeling передбачено виконувати у зв’язку, та паралельно. При оцінці моделі, що побудована на оброблених даних, може виникнути необхідність повторного моделювання, адже саме після оцінки моделювання виявляються нові вимоги для обробки даних

## 3.4 Характеристика функції «Аналіз даних для створення персоналізованих пропозицій»

На вхід до функції подаються оброблені досліджуваного дані клієнта, та дані всіх клієнтів банку. Далі, керуючись алгоритмами Random Forest, Apriori та Logistic Regression, виконується побува моделі, за допомогою підходу Stacking. Для більше детального розуміння, на рисунку 3.6 представлена декомпозиція даної функції системи.

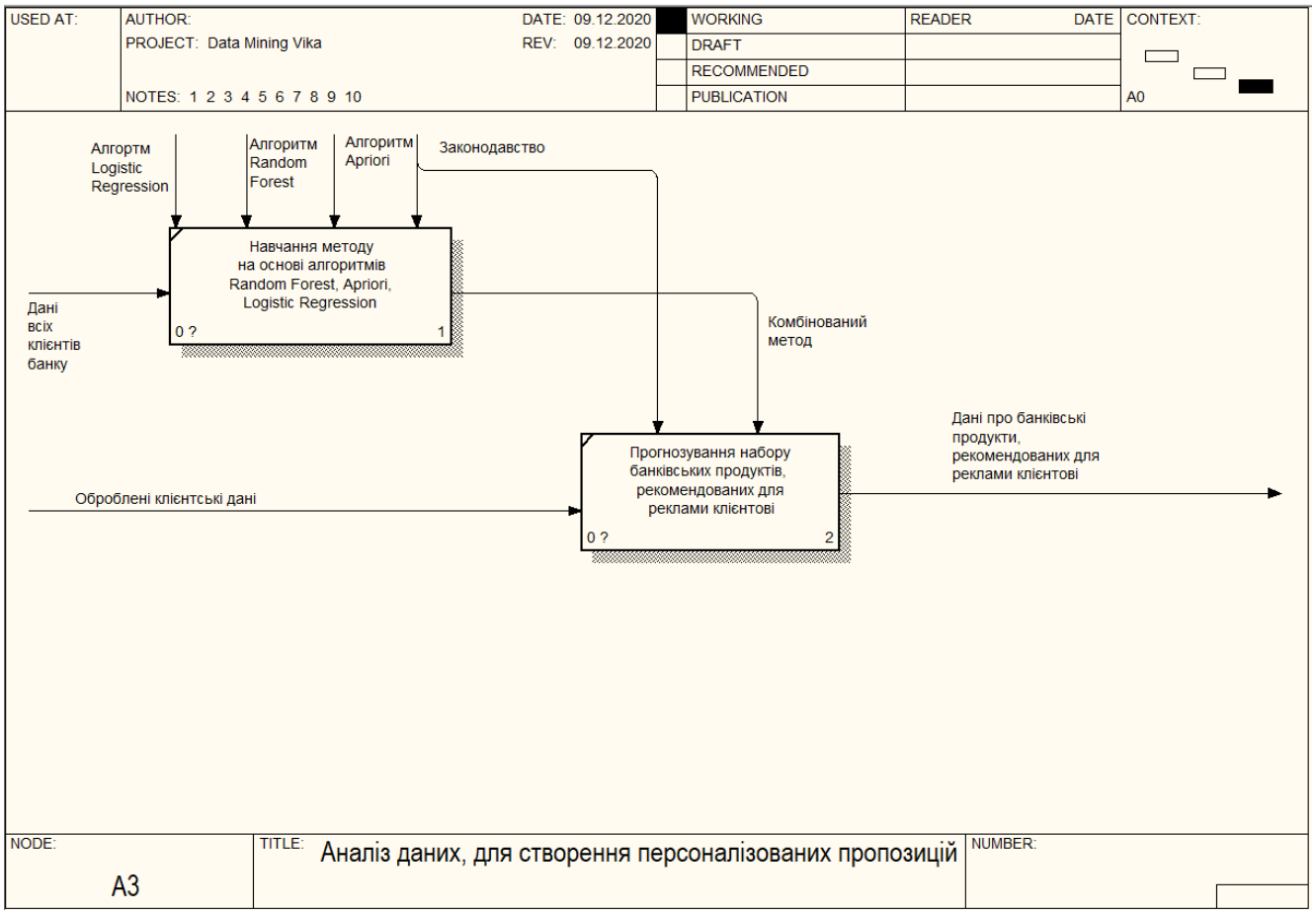


Рисунок 3.6 – Декомпозиція функції «Аналіз даних для створення персоналізованих пропозицій»

При декомпозиції, було виділено 2 основні функції: «Навчання методу на основі алгоритмів Random Forest, Apriori, Logistic Regression» та «Прогнозування набору банківських продуктів, рекомендованих для реклами клієнтові». Перша

функція, необхідна саме для навчання методу, та виконує комбінування трьох алгоритмів для їхньої сумісної роботи. Функція навчання методу, буде часто використовуватися у системі, так як з плином часу дані змінюються і точність методу пронозування буде змінюватися. На виході з цієї функції знаходиться навчений метод на основі комбінаціїї трьох, вище названих, алгоритмів, саме за допомогою нього і буде керуватися прогнозування набору банківських продуктів, рекомендованих для реклами клієнтові.

Інша функція, проводить прогнозування, на основі оброблених клієнтських даних, що подаються на вхід, та креючись кобінованим методом. Саме за допомогою цієї функції система і дасть відповідь, на питання: «чи необхідна ця банківська послуга клієнтові», «які продукти, можуть зацікавити клієнта».

# 4 ПРОЕКТУВАННЯ СИСТЕМИ ДЛЯ АВТОМАТИЗАЦІЇ СТВОРЕННЯ ПЕРСОНАЛІЗОВАНИХ ПРОПОЗИЦІЙ

## 4.1 Розробка діаграми прецедентів системи

Діаграма прецедентів, або Use Case діаграма, важлива складова на всіх етапах життєвого циклу проектування системи [38]. Дана діаграма служить для описання послуг, як система може надавати користувачеві. Діаграма прецедентів представляється у вигляді графа, що відображає різні відношення між множиною акторів та множиною прецедентів або також варіантів використання. Прецедент використовують для характеристики функцій та послуг, що надає система множині акторів. Use Case діаграма не описує реалізацію та яким чином актор буде взаємодіяти з прецедентами системи.

Мова UML (Unified Modeling Language) дозволяє використовувати наступні відношення:

- включення;
- асоціації;
- узагальнення;
- розширення.

Кожен тип відношення, що наведений вище, дозволяє уточнити вимоги до варіантів використання системи. На рисунку 4.1 представлена діаграма прецедентів системи.

Відношення асоціації – одне з загальновживаних у мові моделювання UML та використовується для того щоб, показати відношення між прецедентом та актором.

Відношення включення у мові моделювання UML використовується для створення залежності між базовим варіантом прецеденту та його окремим випадком.

Відношення розширення показує зв’язок між базовим варіантом використання та прецедентом, що може бути лише при виконанні додаткових умов.

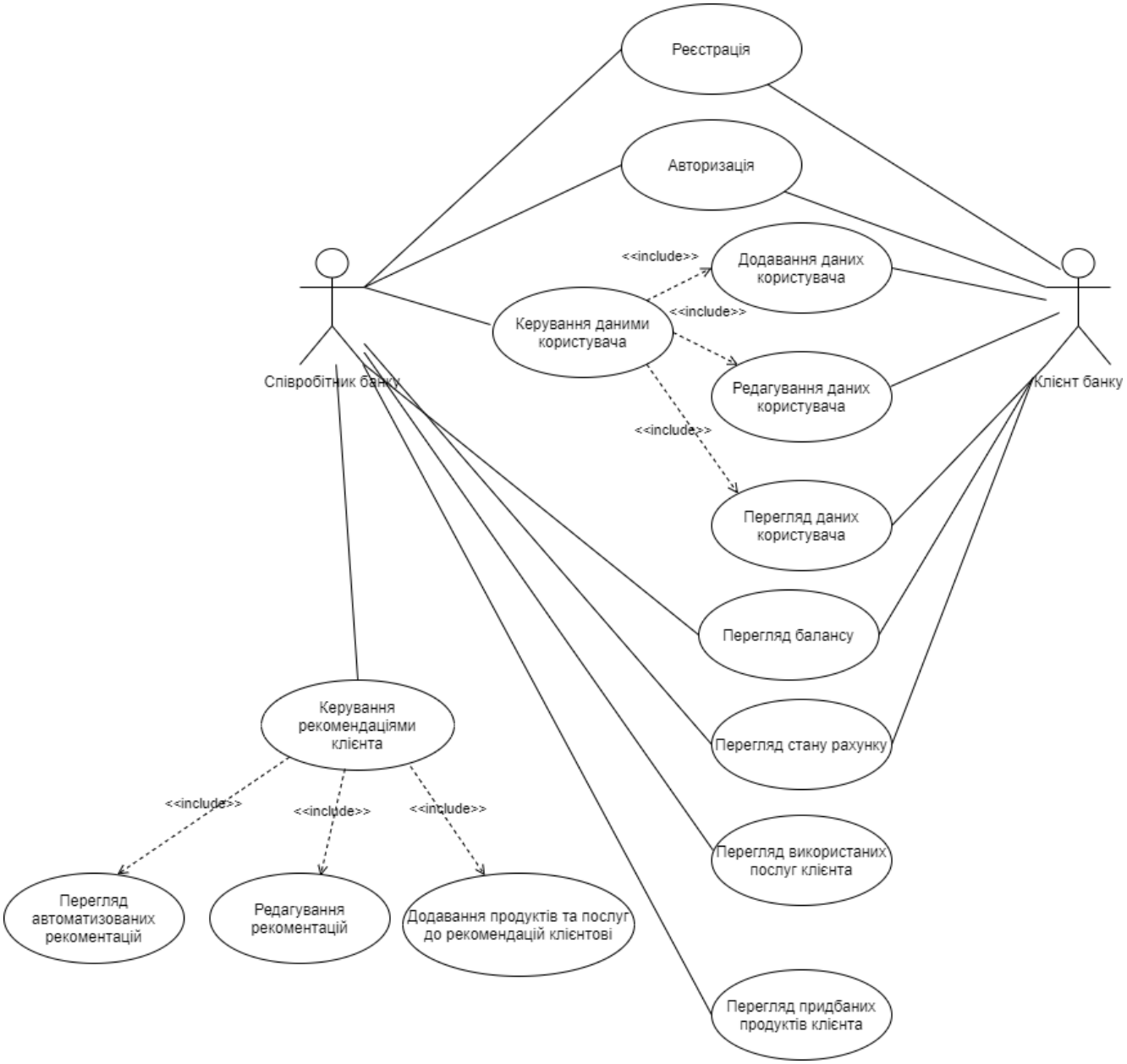


Рисунок 4.1 – Діаграма прецедентів системи

На діаграмі, що зображена вище, виділено 2 типи акторів:

– співробітник банку;
– клієнт банку.

Також виделені наступні варіанти використання системи:

a) реєстрація;
b) авторизація;
c) керування даними користувача, що включає в себе:
    - додавання даних користувача;
    - редагування даних користувача;
    - перегляд даних користувача;
d) перегляд балансу;
e) перегляд стану рахунку;
f) перегляд послуг, якими скористався клієнт;
g) перегляд продуктів, що придбав клієнт;
h) керування рекомендаціями клієнта, що включає в себе:
    - перегляд автоматизованих рекомендацій;
    - редагування рекомендацій;
    - додавання продуктів та послуг до рекомендацій клієнтові.

Отже Use Case діаграма, дозволила виділити основні вимоги до системи, щодо ролей, акторів та варіантів використання. У системі передбачені прецеденти, що описані вище, проте проектована система знаходиться у варіанті прототипу та можливість додавати нові прецеденти, та нових акторів залишається.

## 4.2 Вибір архітектури для розробки додатку

На сьогодні, існує непоганий вибір архітектурних альтернатив для розробки Web-системи. Кожна з них має свої переваги та недоліки, а також область для використання. При дослідженні, були виділені наступні архітектурні паттерни, які в тій чи іншій мірі можуть задовольнити потреби системи:

- 1-рівнева архітектура;
- 2-рівнева архітектура;

– 3-рівнева архітектура.

Приведені вище архітектурні шаблони призначені для вирішення окремих проблем при розробці веб-системи. У рамках атестаційної роботи, було прийнято рішення, що веб-система буде розгортуватися як клієнт-серверний додаток, на рисунку 4.2 зображена схема роботи таких додатків.

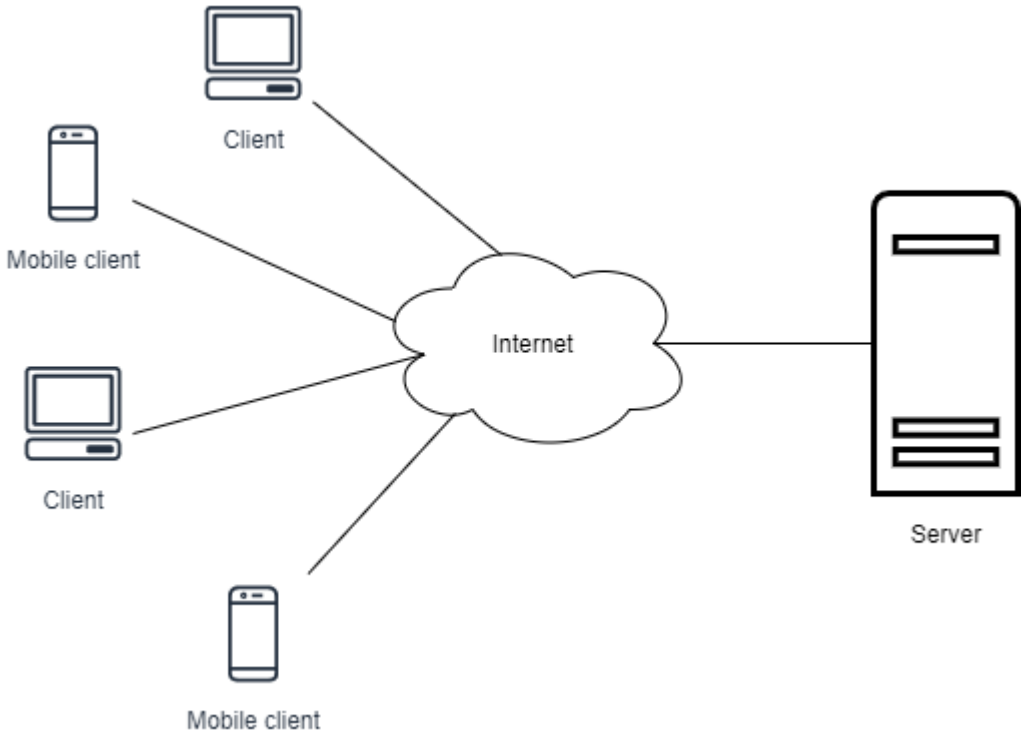


Рисунок 4.2 – Схема роботи клієнт-серверного додатку

Розглянемо приведені вище архітектурні шаблони, для прийняття рішення, щоб зрозуміти, який з них краще підходить для системи автоматизації створення персоналізваних пропозицій на основі нтелектуального аналізу даних клієнта.

### 4.2.1 Характеристика 1-рівневої архітектури

Система, що має однорівневу архітектуру представляє собою тришаровий монолітний додаток [39]. Схему 1-рівневої архітектури представлено на рис. 4.3.

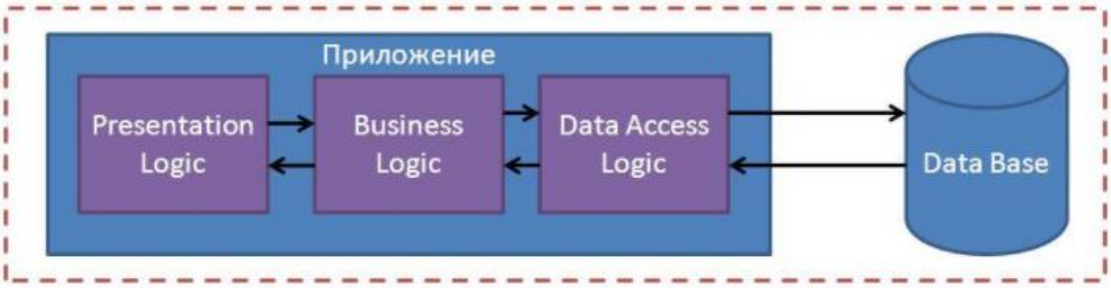


Рисунок 4.3 – Схема 1-рівневої архітектури

Можна виділити наступні основні переваги та недоліки даної архітектури у табл. 4.1

Таблиця. 4.1 – Переваги та недоліки 1-рівневої архітектури

| Переваги | Недоліки |
|---|---|
| - увесь код і всі процеси виконуються на одній машині. | - на одному ПК можна запустити обмежену кількість процесів;<br>- для розгортання на новій машині необхідно переписувати весь код;<br>- при зміні одного шару, необхідно змінювати інші. |

Проаналізувавши переваги та недоліки методу, було прийнято рішення відмовитися від 1-рівневого архітектурного шаблону, через його погану підтримку у ході розробки.

### 4.2.2 Характеристика 2-рівневої архітектури

Дворівнева архітектура включає в себе розбиття основних шарів додатку, таких як PL (Presentation Logic), BL (Business Logic), та Data Access Logic, на 2 рівні: клієнтський та серверний [39]. Схема архітектури зображена на рис. 4.4

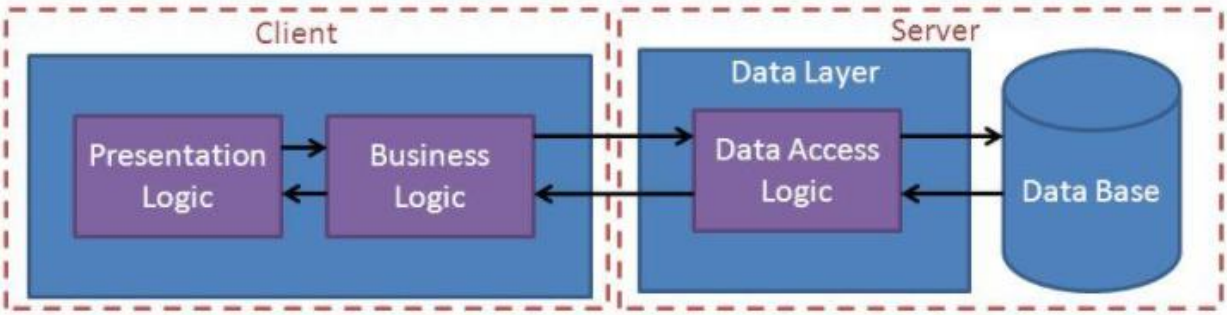


Рисунок 4.4 – Схема 2-рівневої архітектури

Такий архітектурний підхід має свої переваги та недоліки, що описані у таблиці 4.2.

Таблиця 4.2 – Переваги та недоліки 2-рівневої архітектури

| Переваги | Недоліки |
|---|---|
| - дані знаходяться на окремому від клієнта сервері;<br>- клієнтський додаток можна налаштувати на інший сервер. | - високе навантаження на сервер;<br>- PL та BL залишаються сильно зв'язаними<br>- при відключенні сервера, клієнт працювати не буде. |

Отже, можна зробити висновок, що для розробки системи, що проектується у даній атестаційній роботі даний підхід краще не використовувати, так як система розрахована на роботу під великим навантаження, що є недоліком розглянутого підходу.

### 4.2.3 Характеристика 3-рівневої архітектури

Трирівневий архітектурний підхід включає в себе розбиття системи на 3 рівні:

- клієнтський рівень;
- серверний рівень;
- рівень серверу бази даних.

Клієнтський рівень включає в себе PL, серверний – BL та Data Access Logic, а рівень серверу БД включає базу даних [39], схема зображена на рис. 4.5

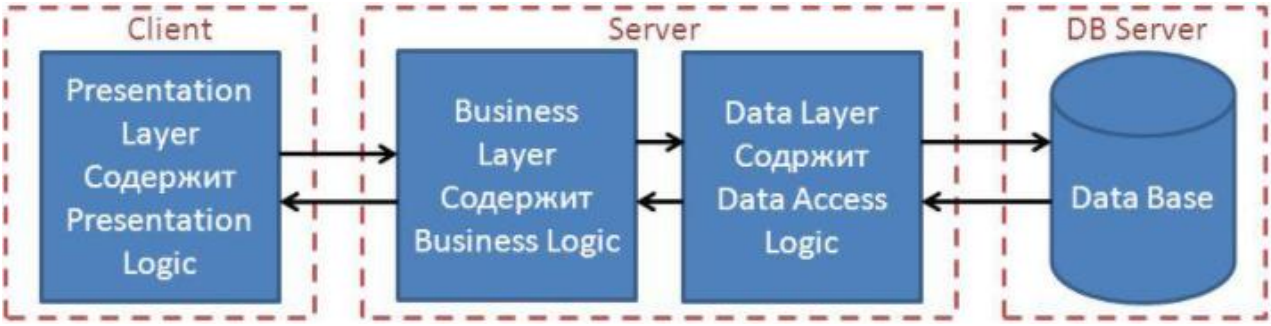


Рисунок 4.5 – Схема 3-рівневої архітектури

Такий архітектурний шаблон має свої переваги та недоліки, що описані далі у табл. 4.3.

Табл. 4.3 – Переваги та недоліки 3-рівневої архітектури

| Переваги | Недоліки |
|---|---|
| - потенційно, кожен шар додатку може бути розгорнутий на окремій машині;<br>- логіка знаходиться окремо від представлення та даних;<br>- може працювати з великими об’ємами даних.<br>- не з’єднані шари не контактують зовсім. | - вартість, розробка кожного шару потребує більше часу. |

Таким чином, для розробки системи у даній атестаційній роботі було обрано 3-рівневу архітектуру, заважаючи на її переваги. Кожна з переваг дозволить у майбутньому наладити гарну підтримку додатку та додасть гнучкості. Також закладається можливість, після реалізації додатку, рухатися у сторону мікросервісної архітектури, що набагато складніше зроботи у випадках 2-рівневох та 3-рівневої архітектури.

## 4.3 Вибір СУБД

Для розробки рівня даних проектованої системи, що включає в себе базу даних, було обрано СУБД MySQL. СУБД (Система Управління Базами Даних) – це програмний додаток, що забезпечує взаємодію з користувачем, додатком і самою базою даних та використовується для збору та обробки даних [40].

MySQL – являє собою реляційну СУБД та використовує структури, що дозволяють ідентифікувати і отримати дані з інших структур цієї БД.

Реляційні бази даних використовують мову структурованих запитів SQL для доступу до даних, а також для управління ними. MySQL є продуктом з відкритим вихідним кодом для управління реляційними базами даних [41], що працює та підтримується на багатьох платформах.

СУБД MySQL підтримує 2 типи двигунів: MyISAM та InnoDB. Кожен з них використовується для різних ситуацій, перш ніж обрати двигун, необхідно розглянути різницю між ними, що представлена у таблиці 4.4.

Таблиця 4.4 – Ключові відміннтості двигунів MyISAM та InnoDB

| MyISAM | InnoDB |
|---|---|
| - підтримує зжимання таблиць<br>- підтримує повнотекстовий пошук | - підтримує транзакції;<br>- підтримує блокування на рівні рядка;<br>- підримує обмеження зовнішніх ключів;<br>- має більшу надійність при великиї об’ємах даних;<br>- має більшу швидкість. |

Отже, розглянувши особливості двигунів СУБД MySQL, було прийнято рішення використовувати двигун InnoDB, через роботу з великою кількістю даних.

## 4.4 Логічне та фізичне моделюваня БД

Важливим етапом проектування системи є моделювання бази даних. Для вирішення цієї задачі було використано CASE засіб «Allfusion ErWin Data Modeler», що дозволив розробити логічну та фізичну моделі БД саме для СУБД MySQL. На рисунку 4.6 зображена логічна модель бази даних.

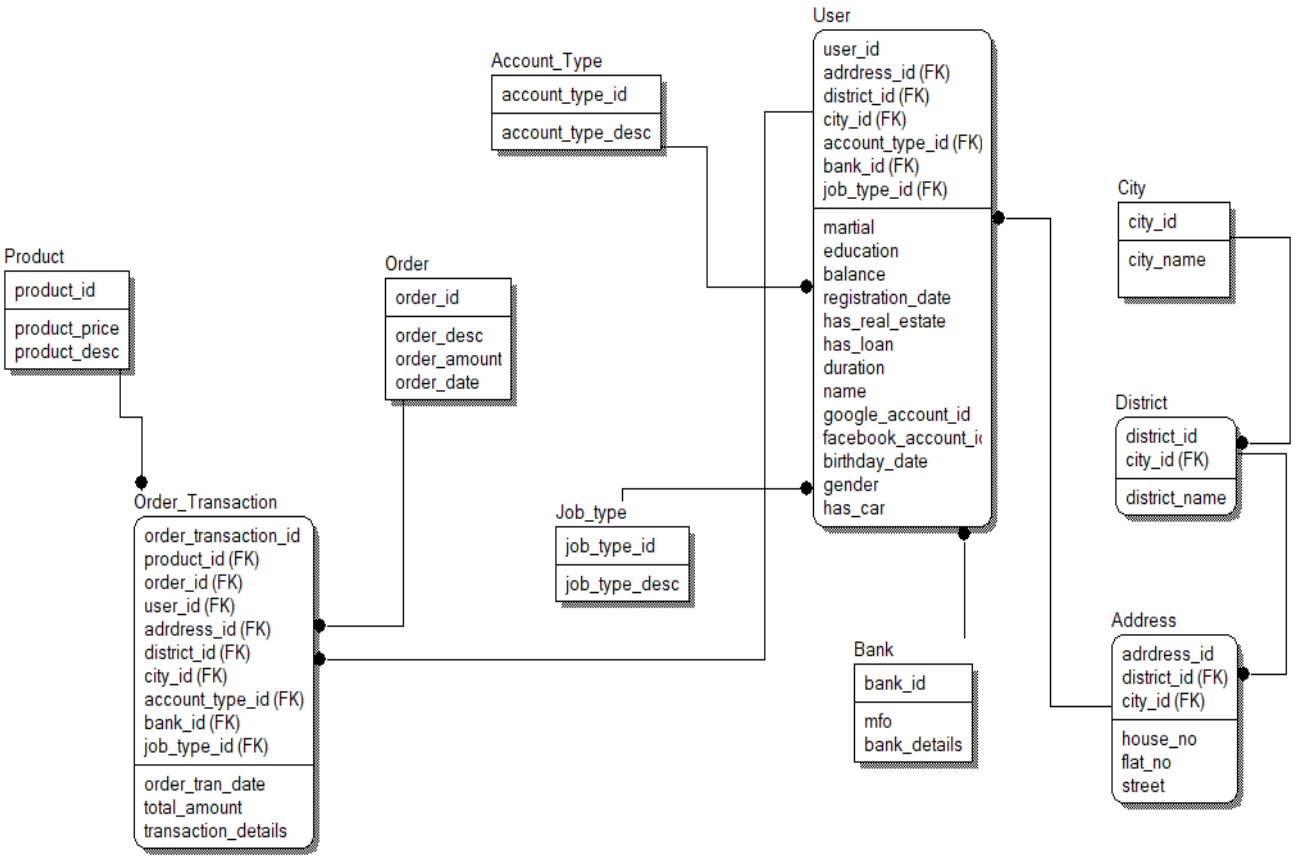


Рисунок 4.6 – Логічна модель бази даних

На логічній моделі бази даних можна виділити 2 найбільші та найважливіші таблиці, це сутності User та Order_Transaction. Сутність User – описує дані клієнта банку, та використовує допоміжні таблиці для конкретизації даних, а саме сутності Account_Type, Hob_Type, Address, City, District, Bank.

Інша таблиця Order_Transactions – описує транзакції та товари, що були придбані клієнтом банку. У таблиці 4.5 знаходиться опис сутностей моделі БД.

Таблиця 4.5 – Опис сутностей моделі БД

| Сутність | Опис |
|---|---|
| User | Інформація про клієнта банку |
| City | Список міст |
| District | Список районів міста |
| Address | Детальна інформація про адресу |
| Bank | Список банків |
| Job_type | Тип професійної діяльності |
| Order | Інформація про замовлення продукту |
| Product | Інформація про продукт |
| Account_type | Тип профілю |
| Order_Transaction | Транзакція по замовленню продукту |

Спроектована модель БД знаходиться у 3-ій нормальній форма, та сутності використовують зв’язки типу «один до багатьох». Опис атрибутів сутностей приведений у таблиці 4.6.

Таблиця 4.6 – Опис атрибутів сутностей

| Сутність | Атрибут | Опис | Обмеження |
|---|---|---|---|
| User | user_id | ID користувача | PK |
| | adrdress_id | ID адреси | FK |
| | district_id | ID регіону | FK |
| | city_id | ID міста | FK |
| | account_type_id | ID типу профіля | FK |
| | bank_id | ID банку | FK |
| | job_type_id | ID типу професії | FK |
| | martial | Сімейний статус | |
| | education | Освіта | |
| | balance | Поточний баланс | |
| | registration_date | Дата реєстрації | |
| | has_real_estate | Наявність своєї нерухомості | |
| | has_loan | Наявність кредитів | |
| | duration | Різниця між тривалістю від активів і зобов'язань | |
| | name | Ім’я | |
| | google_account_id | ID профілю Google | |
| | facebook_account_id | ID профілю Facebook | |
| | birthday_date | Дата народження | |
| | gender | Стать | |
| | has_car | Наявність машини | |

Продовження табл. 4.6

| | | | |
|---|---|---|---|
| Order_Transaction | order_transaction_id | ID транзакції замовлення | PK |
| | product_id | ID продукту/послуги | FK |
| | order_id | ID замовлення | FK |
| | user_id | ID користувача | FK |
| | adrdress_id | ID адреси | FK |
| | district_id | ID регіону | FK |
| | city_id | ID міста | FK |
| | account_type_id | ID типу профіля | FK |
| | bank_id | ID банку | FK |
| | job_type_id | ID типу професії | FK |
| | order_tran_date | Дата транзакції по замовленню | |
| | total_amount | Кінцева вартість | |
| | transaction_details | Деталі транзакції | |
| Account_Type | account_type_id | ID типу профілю | PK |
| | account_type_desc | Опис типу профілю | |
| Address | adrdress_id | ID Адреси | PK |
| | district_id | ID регіону | FK |
| | city_id | ID міста | FK |
| | house_no | Номер будинку | |
| | flat_no | Номер квартири | |
| | street | Назва вулиці | |
| Bank | bank_id | ID банку | PK |
| | mfo | МФО банку | |
| | bank_details | Деталі банку | |
| City | city_id | ID міста | PK |
| | city_name | Назва міста | |
| District | district_id | ID регіону | PK |
| | city_id | ID міста | FK |
| | district_name | Назва регіону | |
| Job_type | job_type_id | ID типу професії | PK |
| | job_type_desc | Опис типу професії | |
| Order | order_id | ID замовлення | PK |
| | order_desc | Опис замовлення | |
| | order_amount | Ціна замовлення | |
| Order | order_date | Час замовлення | |
| Product | product_id | ID продукту/послуги | PK |
| | product_price | Ціна продукту | |
| | product_desc | Опис продукту | |

Також було спроектовано і фізичну модель БД для системи, що буде створювати персоналізовані пропозиції на основі даних клієнтів банку. Фізична модель представлена на рисунку 4.7.

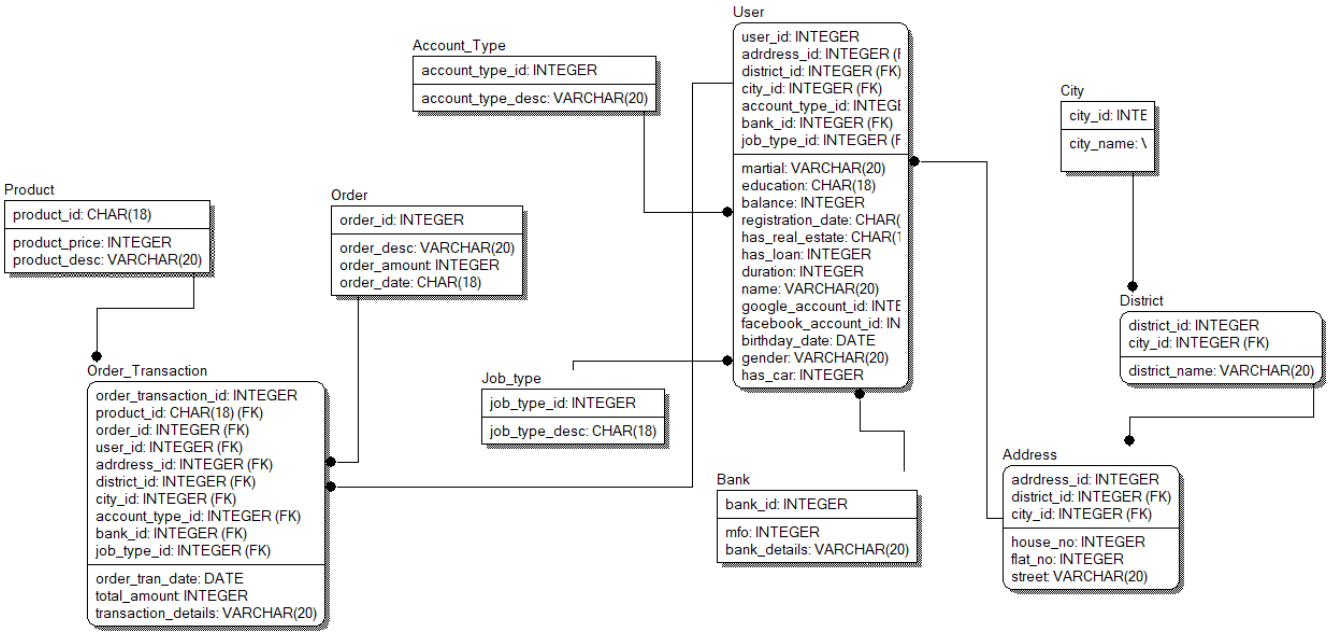


Рисунок 4.7 – Фізична модель БД

Отже, при проектуванні бази даних, були розроблені фізична та логічна модель бази даних, за допомого CASE засобу «Allfusion ErWin Data Modeler». Обраний засіб проектування дозволяє не тільки спроектувати БД, а й згенерувати її для обраної СУБД MySQL.

## 4.5 Вибір мови програмування

Під час проектування архітектури системи було визначено, що система буде розгортуватися як клієнт-серверний додаток на основі 3-рівневої архітектури, а це означає що BL та PL будуть незалежними шарами і розгортуватися на різних машинах.

Клієнтський рівень, що також називають Front-end, буде розроблюватися на мові JavaScript. JavaScript – це динамічна, об’єктно-орієнтована мова програмування, найчастіше знайшла своє застосування у клієнтських додатках.

Прямим застосування даної мови програмування є створення динамічного інтерфейсу користувача а також валідація веб форм.

На сьогоднішній день JavaScript дуже розвинувся та удосконалились фреймворки для усунення рутинної роботи Front-end розробників. В рамках розробки системи для атестаційної роботи, у якості фреймворку для мови JavaScript було обрано Angular 9.

За допомогою мови JavaScript будуть виконуватися динамічні операції, валідація форм, анімація, побудова графіків, та асинхронний обмін даними з серверною частиною системи за допомогою REST API, та форматом JSON. Також ставиться задача, мінімізувати залежність роботи системи від інтернету так, щоб користувач міг скористатися даними з кешу і незалежні функції системи працювали.

У якості розробки серверної частини додатку була обрана мова програмування Java. Java – це об'єктно-орієнтована мова програмування на основі класів, яка розроблена з якомога меншою залежністю реалізації. Обумовлений вибір тим, що мова підтримує строгу типізацію та кросплатформеність. У якості фреймворку було обрано Spring Boot, що дає можливість використовувати багато додаткових компонентів, що полегшує написання коду.

# 5 ПЕРЕВІРКА ЕФЕКТИВНОСТІ МЕТОДУ ТА ОЦІНКА ОТРИМАНИХ РЕЗУЛЬТАТІВ

В даній атестаційній роботі для перевірки ефективності та можливості роботи методу використовувалося середовище JupyterLab та мова програмування Python. Концептуально, метод аналізу даних клієнтів банку необхідний для створення персоналізованих пропозицій для збільшення продажів банківських продуктів та послуг. В даній роботі розроблений метод класифікації клієнтів, який пропонується для використання у автоматизації створення маркетингових пропозицій.

## 5.1 Огляд вхідних даних

Для перевірки роботи методу, в якості ресурсу клієнтських даних була використана вибірка даних з сервісу Kaggle [42], в якій міститься близько 45 тисяч записів про клієнтів банку, що користувались або не користувались строковими депозитами. Користуючись стандартами CRISP-DM для Data Mining, після проведення бізнес аналізу, необхідно провести огляд даних, з якими ведеться робота.

Подання даних на вхід та приклад виведення даних представлено у лістингу 5.1 та на рисунку 5.1.

Лістинг 5.1 – Подання даних на вхід та виведення на екран

```
#Reading training and testing data as dataframes
train_df = pd.read_csv("train.csv",sep =";")
test_df = pd.read_csv("test.csv", sep =";")
train_df.head()
```

| | age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 58 | management | married | tertiary | no | 2143 | yes | no | unknown | 5 | may | 261 | 1 | -1 | 0 | unknown | no |
| 1 | 44 | technician | single | secondary | no | 29 | yes | no | unknown | 5 | may | 151 | 1 | -1 | 0 | unknown | no |
| 2 | 33 | entrepreneur | married | secondary | no | 2 | yes | yes | unknown | 5 | may | 76 | 1 | -1 | 0 | unknown | no |
| 3 | 47 | blue-collar | married | unknown | no | 1506 | yes | no | unknown | 5 | may | 92 | 1 | -1 | 0 | unknown | no |
| 4 | 33 | unknown | single | unknown | no | 1 | no | no | unknown | 5 | may | 198 | 1 | -1 | 0 | unknown | no |

Рис. 5.1 – Приклад клієнтських даних

Переглянемо кількість записів, та ознак кожного клієнта, а також проведемо огляд типів даних (лістинг 5.2, рисунки 5.2 – 5.3).

Лістинг 5.2 – Розрядність та інформація про вхідні дані

```
#Dimensions of dataset
train_df.shape

#Information about the data types of features
train_df.info()
```

```
[30]: (45211, 17)
```

Рис. 5.2 – Кількість записів, ознак

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Рис. 5.3 – Опис типів даних ознак та їх кількість

Переглянемо статистичні підсумки аналізу вибірки (лістинг 5.3, рисунок 5.4)

Лістинг 5.3 – Виклик функції для формування статистичного звіту

```
#Statistical summary of training dataset
train_df.describe(include ='all')
```

| | age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 45211.000000 | 45211 | 45211 | 45211 | 45211 | 45211.000000 | 45211 | 45211 | 45211 | 45211.000000 | 45211 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211 | 45211 |
| **unique** | NaN | 12 | 3 | 4 | 2 | NaN | 2 | 2 | 3 | NaN | 12 | NaN | NaN | NaN | NaN | 4 | 2 |
| **top** | NaN | blue-collar | married | secondary | no | NaN | yes | no | cellular | NaN | may | NaN | NaN | NaN | NaN | unknown | no |
| **freq** | NaN | 9732 | 27214 | 23202 | 44396 | NaN | 25130 | 37967 | 29285 | NaN | 13766 | NaN | NaN | NaN | NaN | 36959 | 39922 |
| **mean** | 40.936210 | NaN | NaN | NaN | NaN | 1362.272058 | NaN | NaN | NaN | 15.806419 | NaN | 258.163080 | 2.763841 | 40.197828 | 0.580323 | NaN | NaN |
| **std** | 10.618762 | NaN | NaN | NaN | NaN | 3044.765829 | NaN | NaN | NaN | 8.322476 | NaN | 257.527812 | 3.098021 | 100.128746 | 2.303441 | NaN | NaN |
| **min** | 18.000000 | NaN | NaN | NaN | NaN | -8019.000000 | NaN | NaN | NaN | 1.000000 | NaN | 0.000000 | 1.000000 | -1.000000 | 0.000000 | NaN | NaN |
| **25%** | 33.000000 | NaN | NaN | NaN | NaN | 72.000000 | NaN | NaN | NaN | 8.000000 | NaN | 103.000000 | 1.000000 | -1.000000 | 0.000000 | NaN | NaN |
| **50%** | 39.000000 | NaN | NaN | NaN | NaN | 448.000000 | NaN | NaN | NaN | 16.000000 | NaN | 180.000000 | 2.000000 | -1.000000 | 0.000000 | NaN | NaN |
| **75%** | 48.000000 | NaN | NaN | NaN | NaN | 1428.000000 | NaN | NaN | NaN | 21.000000 | NaN | 319.000000 | 3.000000 | -1.000000 | 0.000000 | NaN | NaN |
| **max** | 95.000000 | NaN | NaN | NaN | NaN | 102127.000000 | NaN | NaN | NaN | 31.000000 | NaN | 4918.000000 | 63.000000 | 871.000000 | 275.000000 | NaN | NaN |

Рис. 5.4 – Результати статистичного звіту

Переглянувши статистичну інформацію, можна помітити, що мінімальний вік клієнта банку становить 18 років, також, що в строкові депозити вкладаються навіть, коли від’ємний баланс і те, що середня тривалість контакту банку з клієнтом близько 258 секунд. Все, що описано вище свідчить про те, що дані близькі до реальних і з ними можна працювати.

Перевіримо дані на наявність пропусків (лістинг 5.4, рисунок 5.5).

Лістинг 5.4 – Перевірка на наявність пропуску

```
#Checking if there are any missing values
train_df.isnull().sum()
```

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
```

Рис. 5.5 – Кількість пропущених значень кожної ознаки

Пропущених значень немає, а отже заповнювати їх або видаляти ознаки не потрібно.

## 5.2 Візуалізація даних

Для кращого розуміння відношення деяких ознак до результату прогнозування моделі, проведемо візуалізацію залежностей. Розглянемо, як освіта, сімейний статус, тип роботи, наявність кредиту, наявність нерухомості, впливають на рішення використання строкового депозиту (лістинг 5.5, рисунки 5.6-5.7).

Лістинг 5.5 – Відображення діаграм залежностей

```
sns.countplot(x="education",data=train_df, hue = "y")
plt.title("Education type vs Count")
sns.countplot(x="marital", data = train_df)
plt.title("Martial Status vs Count")
sns.countplot(x="job", data = train_df, hue ="y")
plt.title("Job vs Count")
plt.xticks(rotation=90)
sns.countplot(x="loan", data = train_df, hue ="y")
plt.title("personal loan vs Count")
```

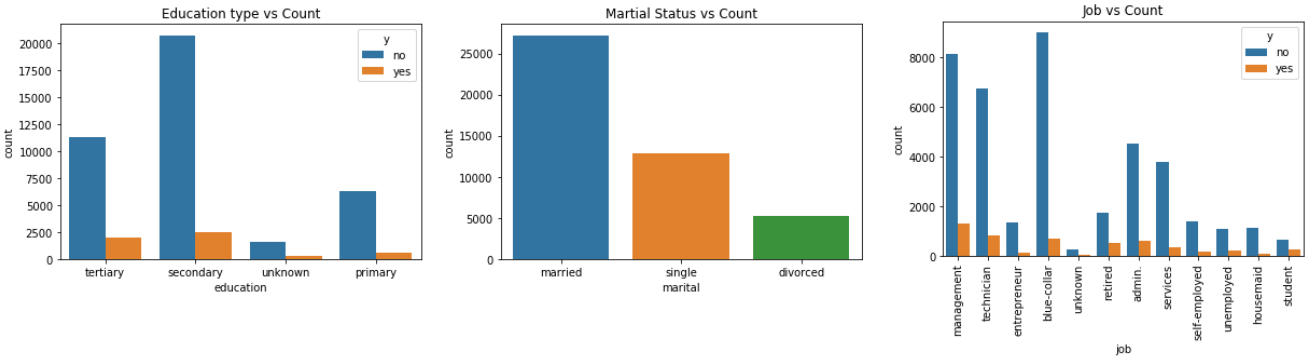


Рис. 5.6 – Діаграми залежностей для ознак: освіта, сімейний статус, тип роботи

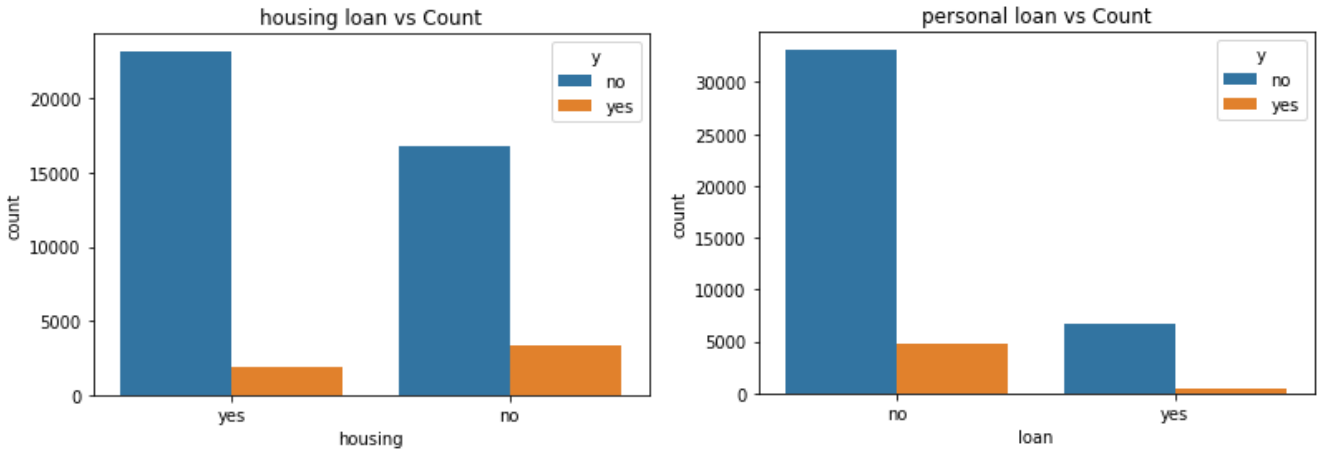


Рис. 5.7 - Діаграми залежностей для ознак: наявність нерухомості, наявність кредиту

Діаграми свідчать, що строкові депозити краще беруть люди з середньою освітою, одружені, приналежать до робочого класу, не мають кредитів, а також мають нерухомість.

## 5.3 Підготовка даних

Перед тим, як провести навчання моделей, необхідно підготувати дані для обраних класифікаторів. Для початку необхідно позбутися категоріальних даних (лістинг 5.6, рисунок 5.8).

Лістинг 5.6 – Програмний код для обробки категоріальних даних

```
df = pd.get_dummies(df,columns =
['job','marital','education','default','housing','month','loan','contact','poutcome'],
drop_first = True)
df.head()
```

| | age | balance | day | duration | campaign | pdays | previous | y | job_blue-collar | job_entrepreneur | ... | month_may | month_nov | month_oct | month_sep | loan_yes | contact_telephone | contact_unknown | poutcome_other |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 58 | 2143 | 5 | 261 | 1 | -1 | 0 | no | 0 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **1** | 44 | 29 | 5 | 151 | 1 | -1 | 0 | no | 0 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **2** | 33 | 2 | 5 | 76 | 1 | -1 | 0 | no | 0 | 1 | ... | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| **3** | 47 | 1506 | 5 | 92 | 1 | -1 | 0 | no | 1 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4** | 33 | 1 | 5 | 198 | 1 | -1 | 0 | no | 0 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

5 rows × 43 columns

Рис. 5.8 – Результат обробки категоріальних даних

Після обробки категоріальних даних, кількість ознак клієнта збільшилось до 43, проте це дозволить збільшити якість моделі.

Проведемо масштабування даних (лістинг 5.7, рисунок 5.9).

Лістинг 5.7 – Масштабування даних

```
target = df['y']
df = df.drop('y',axis = 1)
columns = df.columns
scaler = MinMaxScaler()
df = scaler.fit_transform(df)
df = pd.DataFrame(df,columns=[columns])
df.head()
```

| | age | balance | day | duration | campaign | pdays | previous | job_blue-collar | job_entrepreneur | job_housemaid | ... | month_may | month_nov | month_oct | month_sep | loan_yes | contact_telephone | contact_unknown |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.519481 | 0.092259 | 0.133333 | 0.053070 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| **1** | 0.337662 | 0.073067 | 0.133333 | 0.030704 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| **2** | 0.194805 | 0.072822 | 0.133333 | 0.015453 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 |
| **3** | 0.376623 | 0.086476 | 0.133333 | 0.018707 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| **4** | 0.194805 | 0.072812 | 0.133333 | 0.040260 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

Рис. 5.9 – Результат масштабування даних

На даному етапі було проведено обробку даних, а саме перетворення категоріальних даних та масштабування даних. Так як пропущених значень немає, дані тепер готові до подання на вхід до навчання моделі.

## 5.4 Навчання моделі

За алгоритмом, що описано у пункті 2.4, було вирішено, що модель буде навчатися на основі алгоритмів Logistic Regression та Random Forest. Класифікація буде проходити за методикою найбільшої кількості голосів.

Отже для початку потрібно навчити моделі за допомогою алгоритмів Logistic Regression та Random Forest (лістинг 5.8-5.9).

Лістинг 5.8 – Навчання моделі за допомого алгоритму Logistic Regression

```
y = np.array(target)
X = df
#Splitting the data into train and test data
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.25, random_state = 20)
#Initializing and fitting the logistic regression model
lr_model = LogisticRegression(max_iter=125)
lr_model.fit(X_train,y_train)
y_pred = lr_model.predict(X_test)
```

Лістинг 5.9 – Навчання моделі за допомого алгоритму Random Forest

```
from sklearn.ensemble import RandomForestClassifier
RFC =  RandomForestClassifier(n_estimators = 500, n_jobs = -1)
RFC.fit(X_train,y_train)
y_pred = RFC.predict(X_test)
print(confusion_matrix(y_test, y_pred))
```

Провівши навчання моделей, подаємо їх до вирішального класифікатору (лістинг 5.10).

Лістинг 5.10 – Навчання моделі на основі комбінування алгоритмів Logistic Regression та Random Forest методом голосування

```
from sklearn.ensemble import VotingClassifier
VC = VotingClassifier(estimators=[('lr', lr_model), ('rf', RFC)], voting='hard')
VC.fit(X_train,y_train)
y_pred = VC.predict(X_test)
print(confusion_matrix(y_test, y_pred))
```

Для оцінки результуючої моделі, порівняймо результати за допомогою класифікаційного звіту (рисунок 5.10).

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.98 | 0.94 | 10938 |
| 1 | 0.66 | 0.33 | 0.44 | 1495 |
| accuracy | | | 0.90 | 12433 |
| macro avg | 0.79 | 0.65 | 0.69 | 12433 |
| weighted avg | 0.88 | 0.90 | 0.88 | 12433 |

Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.98 | 0.96 | 10938 |
| 1 | 0.77 | 0.49 | 0.60 | 1495 |
| accuracy | | | 0.92 | 12433 |
| macro avg | 0.85 | 0.74 | 0.78 | 12433 |
| weighted avg | 0.91 | 0.92 | 0.91 | 12433 |

Combined model

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.98 | 0.95 | 10938 |
| 1 | 0.73 | 0.40 | 0.51 | 1495 |
| accuracy | | | 0.91 | 12433 |
| macro avg | 0.83 | 0.69 | 0.73 | 12433 |
| weighted avg | 0.90 | 0.91 | 0.90 | 12433 |

Logistic Regression

Рис. 5.10 – Порівняння результатів роботи різних моделей

Отже, при перевірці прогнозів за допомогою тестової вибірки, моделі отримали різні результати: Logistic Regression – 87%, Random Forest – 89%, там модель на основі комбінування – 93%. Можна зробити висновок, що комбінування методів підвищує продуктивність маркетингової компанії. Звісно в реальному житті майже не можливо досягти маркетингової кампанії такої точності як 93%, все залежить від даних і наскільки вони правильно відображають ситуацію в реальному житті.

## 5.5 Приклад використання навченої моделі

Після проведення навчання моделі, за стандартом CRISP-DM, якщо отримані результати точності влаштовують, метод можна впроваджувати в рамках системи. Приклад прогнозування можливості вкладу клієнта коштів на строковий депозит представлений у лістингах 5.11-5.12 та на рисунках 5.11 – 5.12.

Лістинг 5.11 – Програмний код відображення даних про нового клієнта

```
print(newClient)
```

| | age | balance | day | duration | campaign | pdays | previous | job_blue-collar | job_entrepreneur | job_housemaid | ... | month_may | month_nov | month_oct | month_sep | loan_yes | contact_telephone | contact_unknown |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.519481 | 0.115474 | 0.5 | 0.01891 | 0.080645 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |

Рис. 5.11 – Відображення ознак нового клієнта

Лістинг 5.12 – Програмний код для прогнозування вкладу коштів на строковий депозит

```
VC.predict(newClient)
```

```
array([0], dtype=int64)
```

Рис. 5.12 – Результат прогнозування методу

Отже було змоделювано ситуацію, де до методу на вхід надійшли дані про нового клієнта, та він видав його клас – 0, що означає клієнт скоріш за все відмовиться від запропонованого йому продукту.

# ВИСНОВКИ

У результаті виконання атестаційної роботи було розроблено метод інтелектуального аналізу даних клієнтів банківської сфери, на основі комбінування алгоритмів Data Mining і спроектовано систему, яка використовує даний метод.

Під час аналізу предметної області було розглянуто сучасний стан проблеми створення персоналізованих пропозицій для клієнтів банківської сфери, досліджено традиційні методи, які на сьогодні використовують у банках та виявлені їх недоліки, розглянуто методи ІАД, а також розібрано існуючі маркетингові стратегії з продажу банківських продуктів. У підсумку до розділу було сформульовано основні задачі, які вирішувались у рамках виконання атестаційної роботи.

Для вирішення поставленої мети було спроектовано систему, що дозволить співробітникам банку створювати персоналізовані пропозиції та підвищить ефективність маркетингової кампанії банку. При проектуванні системи було обрано архітектуру системи, визначено функціональні вимоги, сформовані ролі та прецеденти, побудовані логічна та фізична моделі БД, а також обрано мову програмування.

Для розробки методу було обрано алгоритми Apriori, Random Forest і Logistic Regression  та стратегії їх комбінування. Для перевірки роботи методу було використано тестовий набір даних, що знаходився у відкритому доступі. Результати роботи методу оцінювалися за допомогою класифікаційного звіту. З метою визначення конкурентоспроможності комбінованого методу був проведений порівняльний аналіз між самостійними моделями на основі алгоритмів Random Forest, Logistic Regression та моделлю на основі розробленого методу.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ


1. Шелудько В. М. Фінансовий ринок : підручн. / В. М. Шелудько. – 3-тє вид., стер. – К. : Знання, 2015. – 535 с.
2. Лютий І. О. Банківський маркетинг : підручн. [для студ. вищ. навч. закл.] / І. О. Лютий, О. О. Солодка – К. : Центр учбової літератури, 2010. – 776 с.
3. Перелік файлів з показниками статистичної звітності, що подаються банками відповідно до додатка 2 до Правил № 120, згруповані за напрямками звітної інформації [Электронный ресурс] / Національний банк України: [офіц. веб-сайт]. – Електронні дані. – Київ, 2020. – Режим доступу : https://bank.gov.ua/ua/statistic/nbureport/regindicator-statreport/files-21 – Дата перегляду: 02.09.2020. – Назва з екрану.
4. Стремительный рост monobank: неожиданный успех или продуманная до мелочей виральность [Електронний ресурс] – Режим доступу : https://lafounder.com/case/monobank#note-2 – Дата перегляду: 02.09.2020. – Назва з екрану.
5. Гайдунько В. Специфика банковского маркетинга в современных условиях. / В. Гайдунько // Банковские услуги. – 2010. - №12. – с . 13–15.
6. Дубчак Л. В. Інформаційні системи і технології в банківській діяльності : навч. посіб. / Л. В. Дубчак, Л. А. Клюско, В. Ю. Свириденко. – Ірпінь: Видавництво Національного університету державної податкової служби України, 2016. – 248 с.
7. Завадська Д. В. Маркетинг у банку: теорія та методологія / Д. В. Завадська, Л. В. Жердецька, Л. В. Кузнєцова. – Одеса: Атлант, 2010. – 717 с.
8. Угнівенко Р. Р. Теоретичні аспекти сутності ринку банківських продуктів та послуг / Р. Р. Угнівенко // Вісник Львівської комерційної академії. Серія економічна. – 2014. – № 45. – С. 136-140.
9. Про банки та банківську діяльність [Електронний ресурс] : Закон України від 07.12.2000 р. № 2121-111 (зі змінами та доповненнями) / Верховна

Рада України. – Режим доступу: http://zakon3.rada.gov.ua/laws/show/2121-14, вільний. – Назва з екрана.

10. Телінгер М. Ю. Сутність та особливості визначення поняття «банківська послуга» / М. Ю. Телінгер // Прометей. – 2011. – № 1 (34). – С. 233-235.

11. Аналіз банківської діяльності : Підручник / [А. М. Герасимович, М. Д. Алексеєнко, І. М. Парасій-Вергуненко, А. М. Арістова, В. С. Білошапка та ін.]; за ред. А. М. Герасимовича. – К. : КНЕУ, 2013. — 599 с

12. Чумак Н. К. Перспективы банков 2013–2015: стратегии, маркетинг, продукты и каналы // Банковский ритейл. – № 1 (29). – 2013. – С. 8-19.

13. Копилюк О. І. Аналіз банківської діяльності : навчальний посібник / О. І. Копилюк, О. М. Музичка. – Львів : Львівський торгівельно-економічний університет, 2016. – 416 с.

14. Бондаренко Л. А. Ринок банківських продуктів і послуг : поняття та сегментація / Л. А. Бондаренко, І. О. Жук // Ефективна економіка : електронне наукове фахове видання. – 2013. – № 6. – [Електронний ресурс]. – Режим доступу : http://www.economy.nayka.com.ua/?op=1&z=2121.

15. Chen H. Business intelligence and analytics: From big data to big impact / H. Chen, R. H. Chiang, V. C. Storey // MIS quarterly.– 2012. – № 36(4).

16. Варцаба В. І. Сучасне банківництво: теорія і практика : навч. посібник. / В. І. Варцаба, О. І Заславська. – Ужгород: УжНУ «Говерла», 2018. — 364 с.

17. Ткачук В. О. Маркетинг у банку : навч. посіб. / В. О. Ткачук. – Тернопіль : «ТАЙП», 2010. – 270 с.

18. Субботін С. О. Нейронні мережі: навч. посіб. / С. О. Субботін, А. О. Олійник; за ред. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 132 с.

19. Сергеєв-Горчинський О. О. Інтелектуальний аналіз даних: Комп’ютерний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 122 «Комп’ютерні науки та інформаційні технології», спеціалізацій «Інформаційні системи та технології проектування», «Системне проектування сервісів» / О. О. Сергеєв-Горчинський, Г. В. Іщенко ; КПІ ім. Ігоря Сікорського. –

Електронні текстові данні (1 файл: 1,72 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2018. – 73 с.

20. Черняк О. І. Інтелектуальний аналіз даних: Підручник / О. І. Черняк, П. В. Захарченко – Київ : Знання, 2014. — 599 с.

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

22. Інформаційні технології в моделюванні: Матеріали V-ої всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених (19-20 березня 2020 р., м. Одеса). – Одеса: ОНПУ, 2020. – 164 с.

23. Методи інтелектуального аналізу даних : конспект лекцій для здобувачів другого (магістерського) рівня освітньо-професійної програми «Інженерія програмного забезпечення» галузі знань 12 Інформаційні технології спеціальності 121 Інженерія програмного забезпечення денної та заочної форм навчання / уклад. Н.М. Ліщина. – Луцьк : Луцький НТУ, 2016. – 112 с.

24. Методи інтелектуального аналізу даних [Електронний ресурс]. –Режим доступу : http://buklib.net/books/24506/.

25. Лесная Н. С. Методы и модели Data Mining в построении оценочной модели предпочтений пользователя / Н. С. Лесная, К. К. Дикарева // Збірник наукових праць Харківського університету Повітряних Сил. – 2013 – № 2(35) . – с. 27 – 31.

26. Задворная И. А Применение алгоритмов Data Mining для анализа данных в сфере кредитования. Математическое и компьютерное моделирование в экономике, страховании и управлении рисками: Материалы VII Международной молодежной научно-практической конференции. / И. А Задворная, О. М. Ромакина. – Саратов : ООО Изд-во "Научная книга", 2018. с.61-66.

27. Радіоелектроніка та молодь у XXI столітті. Т. 11 : Конференція "Україна в історико-культурному та мовному просторі" : матеріали 24-го Міжнар. молодіж. форуму, 7 – 9 квітня 2020 р. / М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. – Харків : ХНУРЕ, 2020. – 68 с.

28. Косар Н. С. Особливості маркетингових стратегій банків у сучасних умовах / Н. С. Косар, Є. О. Угольков, А .Б. Гайдук // Вісник Хмельницького національного університету. Науковий журнал. – 2010. – № 6. – С.184–187.

29. Смородников О. Маркетинг банковских услуг для частных клиентов. / О. Смородников // Банковские технологии. – 2012. – №2. – с.26–30.

30. Васильєва Т. А. Маркетинг у банку : навчальний посібник / за заг. ред. проф. Т. А. Васильєвої ; Т. А. Васильєва, С. М. Козьменко, І. О. Школьник та ін. – Суми : ДВНЗ «УАБС НБУ», 2014. – 353 с.

31. Баранов П. А. Банковский маркетинг : особенности рекламы в коммерческих банках. / П. А. Баранов // Финансы и Кредит. – 2013. – № 7. – С. 12-13.

32. Шемчук В. Н. Розробка методу інтелектуального аналізу даних клієнтів банківської сфери / В. Н. Шемчук // Науковий простір: актуальні питання, досягнення та інновації : матеріали міжнародної наук. конф., Харків, 2 жов. 2020 р. / МЦНД. – Харків, 2020. – С. 72 – 73.

33. Sitnikov D. Assessment of extended aggregated association rules / Sitnikov D.,Ryabov O., Titova O., Kovalenko A. // Proceedings of 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies. – Kyiv , 2018. – P. 95 – 99.

34. Sitnikov, D. Informativity of Association Rules from the Viewpoint of Information Theory / Sitnikov, D., Titova, O., Minukhin, S., Kovalenko, A., Titov, S. // International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, Proceedings. – 2019. – P. 595–598.

35. J. Rocca Ensemble methods: bagging, boosting and stacking// Towards data science. – 2019. – [Електронний ресурс]. – Режим доступу : https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

36. Новикова Т.Б. МЕТОДОЛОГИИ IDEF0: ТИПЫ СВЯЗЕЙ, ТУННЕЛИРОВАНИЕ СТРЕЛОК // Международный журнал экспериментального образования. – 2016. – № 12-3. – С. 377-383.

37. Афанасьева Т. В. Введение в проектирование систем интеллектуального анализа данных: учебное пособие / Т. В. Афанасьева, А. Н. Афанасьев. – Ульяновск : УлГТУ, 2017. – 64 с.

38. Табунщик Г. В. Проектування та моделювання програмного забезпечення сучасних інформаційних систем / Г. В. Табунщик, Т.І. Каплієнко, О.А. Петрова − Запоріжжя : Дике Поле, 2016. – 250 с.

39. Архітектура комп'ютера: навчальний посібник / Б. І. Мороз, В. П. Дюбко, С. В. Клименко, В. О. Яковенко, В. В. Поліщук. – Дніпропетровськ: «Академія митної служби України», 2012. – 169 с.

40. Трофименко О. Г. Організація баз даних : навч. посібник / О. Г. Трофименко, Ю. В. Прокоп, Н. І. Логінова, І. М. Копитчук. 2-ге вид. виправ. і доповн. – Одеса : Фенікс, 2019. – 246 с.

41. Мулеса О.Ю. Інформаційні системи та реляційні бази даних. Навч.посібник. – Електронне видання, 2018. – 118 с.

42. Kaggle – data science community with tools and resources [Електронний ресурс] – Режим доступу до ресурсу: https://www.kaggle.com/