

УДК 519.816

Міністерство освіти і науки України  
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук  
(повна назва)

Кафедра Системотехніки  
(повна назва)

## АТЕСТАЦІЙНА РОБОТА Пояснювальна записка

Рівень вищої освіти другий (магістерський)

ГЮИК 506413.009 ПЗ

Дослідження та розробка методів аналізу даних для класифікації заявок на

отримання банківських кредитів

(тема)

Виконав:

Студент 2 курсу, групи СПРм-19-1

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

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

Освітня програма Системне проектування

(повна назва освітньої програми)

Мизніков Р. І.

(прізвище, ініціали)

Керівник проф. Гребеннік І.В.

(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри

(підпис)

Гребеннік І. В.

(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 Мизнікову Руслану Ігоровичу  
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження та розробка методів аналізу даних для класифікації заявок на отримання банківських кредитів  
затверджена наказом по університету від « 02 » 11 2020 р. № 1516 Ст
2. Термін подання студентом роботи (проекту) 21 грудня 2020 р.
3. Вихідні дані до роботи (проекту) Науково-технічні публікації, дані статей, результати експериментальних досліджень. Перелік використаних програмних засобів: OS Windows 10, Anaconda Navigator, Jupiter Notebook, Python, Oracle, ERwin Data Modeler, веб-сервіс draw.io.
4. Зміст пояснювальної записки (перелік питань, що потрібно розробити) Вступ, Аналіз предметної області, Особливості діяльності банківського сектору, Огляд кредитних ризиків Аналіз існуючих рішень розрахунку кредитного ризику, Аналіз існуючих методів інтелектуального аналізу даних для оцінки кредитоспроможності клієнтів банку, Системи підтримки прийняття рішень, Комбінування методів аналізу даних, Основні положення методу CART, Основні положення методу SVM, Комбіноване використання методу CART та SVM, Визначення функціональних вимог системи, Проектування системи підтримки рішень для оцінки заявок на отримання кредиту, Обґрунтування архітектури системи, Обґрунтування вибору СУБД, Логічне та фізичне моделювання бази даних, Вибір мови програмування, Тестування та оцінка отриманих результатів методу, Висновки



## РЕФЕРАТ

Пояснювальна записка до атестаційної роботи: 77 с., 4 табл., 35 рис., 2 додатки, 36 джерел інформації. Графічний матеріал атестаційної роботи містить 10 плакатів.

Об'єктом досліджень є процес видачі кредитів клієнтам банку.

Предметом досліджень є методи аналізу даних клієнтів банку для прийняття рішень, щодо видачі банківського кредиту.

Метою дослідження є підвищення ефективності вирішення задачі класифікації заявок на отримання банківських кредитів за рахунок системи підтримки прийняття рішень на основі комбінованого методу інтелектуального аналізу даних.

Методи дослідження: системний підхід, методи інтелектуального аналізу і проектування.

У роботі розроблений комбінований метод інтелектуального аналізу даних на основі алгоритмів CART та SVM. Проведено тестування комбінованого методу на великому об'ємі тестових даних, отримані результати його роботи. У результаті були розроблені рекомендації, щодо підвищення продуктивності вирішення задачі класифікації заявок на отримання банківських кредитів. В роботі виконано проектування системи, для оптимізації роботи розробленого методу.

Галузь застосування – банківські установи, організації та компанії, що можуть надавати кредити.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, КРЕДИТНИЙ РИЗИК, СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, БАЗИ ДАНИХ, BIG DATA, CART, SVM, ORACLE

## ABSTRACT

Explanatory note to attestation work: 79p., 6 tables, 35 pic., 2 applications, 36 sources. The graphic part of the certification work is 10 posters.

The object of research is a process of issuing loans to bank customers.

The subject of research are methods of analyzing the bank's customer data for decisions making about issuance of a bank loan.

The purpose of the research is improving the efficiency of solving the problem of classification of applications to obtain bank loans due to the decision support system based on the combined method of data mining.

Methods of research: system approach, methods of intellectual analysis and design.

In the work was developed the combined method of data mining based on algorithms CART and SVM. The combined method was tested on a large extent of test data and were received the results of his work. As a result, were developed recommendations to increase productivity of solving problems of classification of applications to obtain bank loans. In the work was performed the system design to optimize the work of the developed method.

Field of application – banking institutions, organizations and companies that can provide loans.

DATA MINING, DECISION SUPPORT SYSTEMS, DATABASES, BIG DATA, CART, SVM, ORACLE

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

БД – база даних;

СУБД – система управління базами даних;

СППР – система підтримки прийняття рішень;

API – Application Programming Interface;

CASE – Computer Aided Software Engineering;

CART – Classification and Regression Tree;

JSON – JavaScript Object Notation;

HTTP – HyperText Transfer Protocol;

MVC – Model-View-Controller;

SVM – Support Vector Machine;

## ЗМІСТ

ВСТУП.....	9
1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ.....	10
1.1 Особливості діяльності банківського сектору .....	10
1.2 Огляд кредитних ризиків .....	13
1.3 Аналіз існуючих рішень розрахунку кредитного ризику .....	18
1.4 Аналіз існуючих методів інтелектуального аналізу даних для оцінки кредитоспроможності клієнтів банку .....	25
1.5 Системи підтримки прийняття рішень.....	28
1.6 Постановка задачі.....	30
2 КОМБІНУВАННЯ МЕТОДІВ АНАЛІЗУ ДАНИХ.....	32
2.1 Основні положення методу CART .....	32
2.2 Основні положення методу SVM.....	36
2.3 Комбіноване використання методу CART та SVM.....	38
3 ВИЗНАЧЕННЯ ФУНКЦІОНАЛЬНИХ ВИМОГ СИСТЕМИ.....	41
4 ПРОЕКТУВАННЯ СИСТЕМИ ПІДТРИМКИ РІШЕНЬ ДЛЯ ОЦІНКИ ЗАЯВОК НА ОТРИМАННЯ КРЕДИТУ .....	46
4.1 Діаграма варіантів використання системи.....	46
4.2 Розробка структурної схеми інформаційної системи підтримки прийняття рішень для аналізу заявок на отримання позики.....	48
4.3 Обґрунтування архітектури системи.....	49
4.3.1 Опис 2-шарової клієнт-серверної архітектури.....	50
4.3.2 Опис 3-шарової клієнт-серверної архітектури.....	50
4.3.3 Опис N-шарової клієнт-серверної архітектури .....	51
4.4 Обґрунтування вибору СУБД.....	53

4.5 Логічне та фізичне моделювання бази даних .....	53
4.6 Вибір мови програмування .....	57
5 ТЕСТУВАННЯ ТА ОЦІНКА ОТРИМАНИХ РЕЗУЛЬТАТІВ МЕТОДУ .....	59
ВИСНОВКИ .....	73
ПЕРЕЛІК ПОСИЛАНЬ .....	74

## ВСТУП

На сьогоднішній день, з розвитком інформаційних технологій, набирає оберти і розвиток банківської сфери. Користувач банківської системи все більше і більше відмовляється від готівкових грошей на користь безготівкових. Люди користуються банками постійно: отримують заробітну плату, обмінюють валюту, переводять гроші з картки на картку, проводять оплату товару за допомогою картки, беруть кредити чи кладуть гроші на депозит. У нашій країні нерідко у людей виникає проблема з грошима, для вирішення якої є можливість використання банківських кредитів. Таким чином, виникає потреба у методах аналізу даних для класифікації заявок на отримання банківських кредитів.

Банківський скоринг – це система, що використовується банками для оцінки кредитоспроможності клієнта, в основі якої закладені статистичні методи аналізу даних. Результатом банківського скорингу є числове значення, яке й використовують у прийнятті рішення про видачу кредиту [1].

Для кожного банку можуть використовуватися різні скорингові моделі, на основі різних методів аналізу даних. Щоб розрахувати скоринг, інформаційні системи використовують різні дані користувачів, починаючи з даних що вже знаходяться у системі, закінчуючи даними які можливо дістати з інших сервісів, наприклад: Бюро Кредитних історій, дані держстатистики, результати перепису населення та інші [2].

Метою дослідження є підвищення ефективності вирішення задачі класифікації заявок на отримання банківських кредитів за рахунок системи підтримки прийняття рішень на основі комбінованого методу інтелектуального аналізу даних.

Об'єктом досліджень є процес видачі кредитів клієнтам банку.

Предметом досліджень є методи аналізу даних клієнтів банку для прийняття рішень, щодо видачі банківського кредиту.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

## 1.1 Особливості діяльності банківського сектору

Сучасна банківська галузь - це мережа фінансових установ, ліцензованих державою на надання банківських послуг [3]. Основні пропоновані послуги стосуються зберігання, передачі, надання кредитів або управління ризиками, пов'язаними з утриманням різних форм багатства. Точний набір фінансових послуг, що пропонуються в будь-який момент часу, значно різнився в різних установах, у часі, різних юрисдикціях, розвитком економіки та досягненнями в галузі інформаційно-комунікаційних технологій.

Банки як фінансові посередники є учасниками переказу коштів від заощаджувача до кінцевого споживача коштів. Часто банки корисно змінюють умови договірної угоди, коли кошти рухаються в процесі переказу таким чином, щоб підтримувати та стимулювати економічну діяльність [4]. Виставляючи торгові вимоги (банківські депозити) проти себе, банк може додати гнучкості до обігових засобів обміну таким чином, щоб підвищити ефективність платіжної системи. Ці депозити можуть сприяти наданню особистих кредитів споживачам (роздрібне банківське обслуговування) або короткострокових кредитів нефінансовим підприємствам (комерційний банкінг). Якщо це так, банк сприяє управлінню ліквідністю, сприяючи тим самим споживанню домогосподарств та торгівлі.

Спеціалізуючись на оцінці ризику, банк може контролювати результати роботи позичальника; диверсифікуючи різні інвестиційні проекти, також мінімізує деякі види ризиків та сприяє розподілу коштів на ті заходи, що мають найбільший економічний потенціал [5]. Видаючи міжнародний торговий кредит (торгівельний банкінг), банк може сприяти міжнародній торгівлі та комерції.

Для виконання своїх різноманітних функцій банки всіх типів управляють портфелями фінансових активів та пасивів. Деякі з найбільш важливих питань для

банківської галузі та державних регуляторів зосереджуються на питаннях, як найкраще керувати портфелем депозитних банків, враховуючи життєво важливу роль цих банків у наданні комерційного кредиту та забезпеченні можливостей для здійснення платежів.

Деякі основні сучасні послуги, пропоновані банками зазначені у джерелах[6, 7]:

- надання позик;
- овердрафт;
- дисконтування векселів;
- збір та оплата кредитних інструментів;
- обмін іноземної валюти;
- консультація;
- банківська гарантія;
- переказ коштів;
- кредитні картки;
- послуги банкоматів;
- дебетові картки;
- домашній банкінг;
- інтернет-банкінг;
- мобільний банкінг;
- прийом депозиту;
- пріоритетні банківські справи.

Банки – це бізнес-організації, орієнтовані на прибуток [8]. Тому вони повинні надавати позику населенню та отримувати від них відсотки як прибуток. Утримуючи певні резерви готівки, банки надають короткострокові, середньострокові та довгострокові позики позичальникам.

Іноді банк надає своїм клієнтам можливість овердрафту, через який їм дозволяється зняти більше, ніж їх депозити [9]. За перевитрачену суму з клієнтів стягуються відсотки.

Банки мають справу з іноземними валютами. Відповідно до вимог клієнтів, вони обмінюють іноземну валюту з місцевою валютою, що є важливим для сплати збору в міжнародній торгівлі [10].

Сучасні комерційні банки - це великі організації. Вони можуть розширити свої функції до консалтингового бізнесу. За цією функцією банки наймають фінансових, юридичних та ринкових експертів, які надають консультації клієнтам щодо інвестицій, промисловості, торгівлі, доходів, податків тощо.

Підтримка кредитних карток – важлива складова банківського сектору. Кредитна картка – це картка, що дозволяє її власнику розпоряджатися коштами що надає постачальник у вигляді кредитного ліміту [3]. Умови такого кредиту обумовлюється договором, створеному при видачі картки.

Банкомати замінюють каси банків, які виконують функції надання таких банківських функцій, як депозити, зняття коштів, запити на рахунки. Ключові переваги банкоматів включають:

- цілодобова доступність;
- виключення вартості робочої сили;
- зручність розташування.

Дебетові картки використовуються для електронного зняття коштів безпосередньо з рахунків власників карток. Більшість дебетових карток вимагають персонального електронно-цифрового підпису (ЕЦП) для підтвердження транзакції.

Домашній банкінг - це процес завершення фінансової операції з власного будинку, на відміну від використання філії банку. Він включає такі дії, як запити на рахунки, переказ грошей, оплата рахунків, подання заявок на позики, спрямування депозитів [7].

Інтернет-банкінг – це послуга, яку пропонують банки, що дозволяє власникам рахунків отримувати доступ до даних своїх рахунків через Інтернет [8]. Інтернет-банкінг проводить операції через традиційні банки дозволяє клієнтам виконувати всі рутинні операції, такі як перекази рахунків, запити балансу, оплату рахунків та запити на зупинку платежів, а деякі навіть пропонують

онлайн-заявки на позики та кредитні картки. Інформація про обліковий запис доступна в будь-який час, вдень та вночі, і це можна зробити з будь-якого місця.

Мобільний банкінг (також відомий як M-Banking) - це термін, що використовується для перевірки балансу, операцій з рахунками, платежів, кредитних додатків та інших банківських операцій через мобільний пристрій [11].

Прийом депозитів від вкладників або власників рахунків є основною функцією банку. Банки приймають депозити від тих, хто може заощадити гроші, але не може використати їх у прибуткових секторах. Люди вважають, що краще зберігати свої заощадження у банку, оскільки цим вони заробляють відсотки.

Пріоритетні банківські послуги можуть включати низку різноманітних послуг, але деякі популярні включають безкоштовну перевірку, оплату рахунків через Інтернет, фінансові консультації та інформацію.

## 1.2 Огляд кредитних ризиків

В рамках магістерської атестаційної роботи, та мети дослідження, найбільшу увагу необхідно приділити кредитним ризикам.

Аналіз кредитного ризику можна розглядати як продовження процесу розподілу кредитів. Після того, як фізична особа чи бізнес звертається до банку чи фінансової установи щодо отримання позики, кредитна установа аналізує потенційні вигоди та витрати, пов'язані з позикою. Аналіз кредитного ризику використовується для оцінки витрат, пов'язаних із позикою [12].

Кредитний ризик або ризик дефолту - це тип ризику, з яким стикаються кредитори. Він виникає через те, що боржник завжди може відмовитись від сплати боргу. Комерційні банки, інвестиційні банки, компанії з управління активами, фонди прямих інвестицій, фонди венчурного капіталу та страхові компанії повинні аналізувати кредитні ризики, яким вони піддаються, щоб вигідно працювати на ринку.

В загальному, кредитний ризик можна представити формулою (1.1):

$$R = P(d) \times ex \times L \quad (1.1)$$

де  $R$  – кредитний ризик;

$P(d)$  – імовірність дефолту;

$ex$  – експозиція;

$L$  – рівень збитків.

Імовірність дефолту - це ймовірність відмови боржника від сплати боргу. Експозиція - це загальна сума, яку повинен отримати кредитор. У більшості випадків це просто сума, запозичена боржником плюс виплата відсотків. Рівень збитків також можна представити у вигляді формули (1.2):

$$L = 1 - k \quad (1.2)$$

де  $k$  - це частка загальної суми, яку можна повернути, якщо боржник не виконає заборгованість [13].

Наведені вище формули у загальному вигляді представляють розрахунок кредитного ризику, для більш точного представлення кредитного ризику, необхідно розглядати аналізувати показник імовірності дефолту.

Для кожної кредитної операції характерні свої особливості, що визначають ступінь ризику. Так, кредитний ризик може виникати через погіршення фінансового стану позичальника, відсутність належних організаційних здібностей у його керівництва, недостатню підготовку працівника, який приймає рішення про кредитування, та інші обставини. Слід зазначити, що кредитний ризик формується з кількох самостійно діючих видів ризику [14-15] (рис. 1.1).

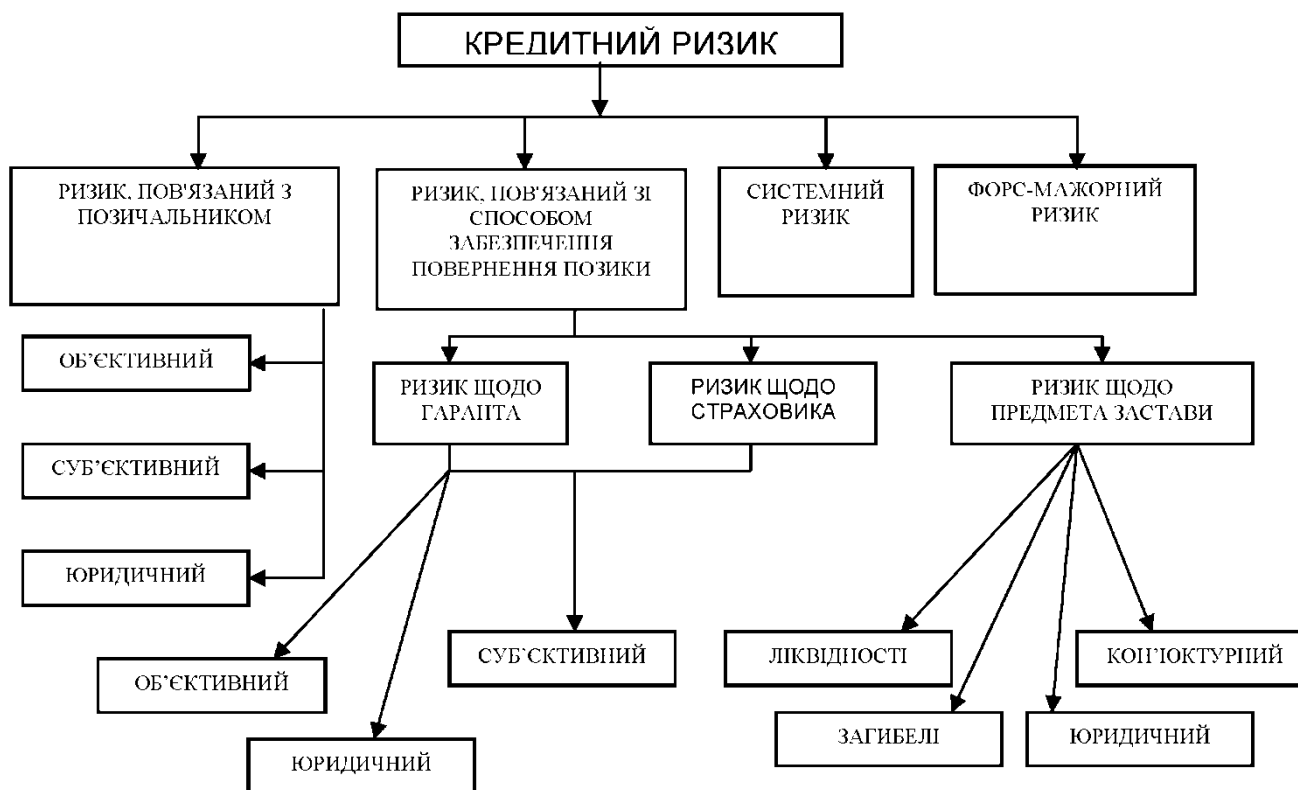


Рис. 1.1 – Структура кредитного ризику.

До основних і найбільш поширених методів мінімізації кредитного ризику належать [10]:

- диверсифікація капіталу;
- оцінювання юридичної і фінансової кредитоспроможності позичальника;
- регулювання обсягів кредитних вкладень;
- використання ефективних форм забезпечення повноти та своєчасності повернення кредиту;
- створення комерційними банками спеціальних резервів для відшкодування втрат від кредитної діяльності.

Надійним способом вирішення проблеми мінімізації кредитного ризику є розробка і використання обґрунтованої методики, що дозволить банку реально оцінювати кредитоспроможність позичальників. Проте слід зазначити, що не існує ні єдиної методики, ні єдиного підходу до її побудови. Більше того, не існує навіть єдиного обґрунтованого визначення поняття кредитоспроможності, про що,

зокрема, свідчать наведені нижче різноманітні тлумачення поняття кредитоспроможності [16]:

- кредитоспроможність позичальника – це його спроможність повністю і своєчасно розрахуватися за своїми борговими зобов'язаннями;
- кредитоспроможність – система умов, що визначають спроможність підприємства залучати позиковий капітал і повертати його в повному обсязі у передбачені терміни”;
- кредитоспроможність – спроможність компанії або приватної особи залучати позиковий капітал і в майбутньому належним чином обслуговувати свій борг;
- кредитоспроможність – це здатність позичальника в повному обсязі і у визначений кредитною угодою термін розрахуватися за своїми борговими зобов'язаннями”

Головна мета процесу управління кредитним ризиком – управління кредитним портфелем банку, що полягає в забезпеченні максимальної доходу за певного рівня ризику. Рівень доходу кредитного портфеля залежить від структури й обсягу портфеля, а також від рівня відсоткових ставок за кредитами [13].

Основними причинами виникнення кредитного ризику на рівні окремої позики є:

- нездатність позичальника до створення адекватного грошового потоку;
- ризик ліквідності застави;
- моральні та етичні характеристики позичальника.

До факторів, які збільшують ризик кредитного портфелю банку, належать:

- надмірна концентрація - зосередження кредитів в одному із секторів економіки;
- надмірна диверсифікація, яка призводить до погіршення якості управління за відсутності достатньої кількості висококваліфікованих фахівців зі знаннями особливостей багатьох галузей економіки;
- валютний ризик кредитного портфеля;

- структура портфеля, якщо він сформований лише з урахуванням потреб клієнтів, а не самого банку;
- рівень кваліфікації персоналу банку.

Кредитна заявка - це прохання про продовження кредиту. Заявки на кредит можна подавати в усній або письмовій формі, як правило, за допомогою електронної системи. Незалежно від того, чи робиться це особисто або за допомогою співробітника банку, заявка повинна юридично містити всю відповідну інформацію, що стосується вартості кредиту для позичальника, включаючи річний процентний дохід та всі пов'язані з ним збори [2].

Система відбору суб'єктів, що залишають свої заявки на отримання кредиту, відповідно до якої працюють сучасні банки має ряд недоліків:

- оцінювання зазвичай має суб'єктивний характер, через це точність менша;
- велика частина експертних оцінок базуються на дистанційному аналізі, який часто призводить до хибних висновків;
- результативність аналітики залежить від кількості разів її проведення.
- формування оцінки кредитоспроможності клієнта має зазвичай формальний, ситуативний характер;
- схвалення заявки, цілком залежить від рішення аналітика банку. Його знання та персональні стан, під час оцінювання заявки значно впливають на долю клієнта;
- через високу вартість експертизи встановлюють обмеження на розмір кредитної заявки;
- кількість заявок, що може обробити аналітик обмежена його робочим днем.

Розглядаючи ключові моменти кредитних ризиків в банківській сфері та проаналізувавши основні фактори, які впливають на величину ризику, можна зробити висновок, що формування рішення щодо надання, чи не надання кредиту є достатньо складним і відповідальним.

### 1.3 Аналіз існуючих рішень розрахунку кредитного ризику

Перелік ознак кредитоспроможності клієнта залежить від мети аналізу, типів кредиту, кредитної історії окремого клієнта, результатів грошової діяльності позичальника, а також від поточних коштів. У кожного банку розробляється своя система оцінок, виходячи з важливих аспектів кредитної політики, технологічних можливостей, направленості банку, поточних умов договору, конкурентоспроможності, стану взаємовідносин із клієнтами, рівня економічної та політичної стабільності в державі тощо [17].

При визначенні кредитоспроможності позичальника, як правило, треба брати до уваги такі фактори:

- діє- і правоздатність позичальника для здійснення кредитної угоди;
- його моральний вигляд, репутація;
- наявність забезпечення позики;
- здатність позичальника отримувати дохід.

Методики аналізу і система показників кредитоспроможності потенційного позичальника, що використовуються комерційними банками, досить різноманітні. Умовно їх можна розбити на класифікаційні (статистичні) методи оцінки та комплексний аналіз кредитоспроможності [14] (рис. 1.2).

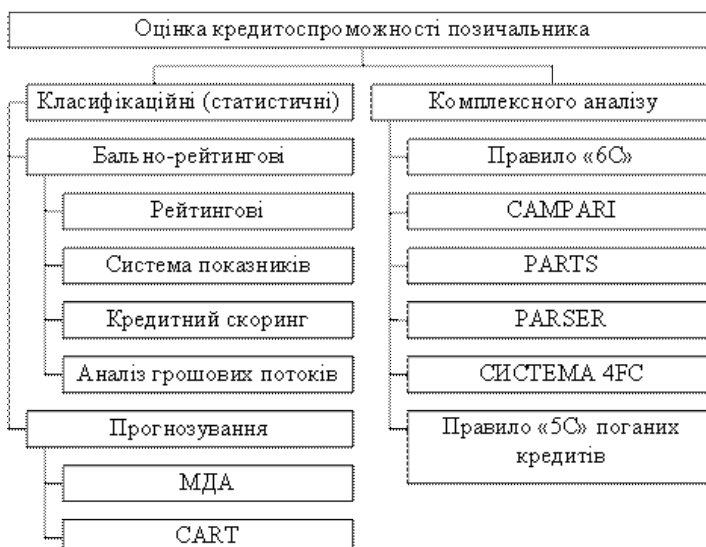


Рис.1.2 – Методи оцінки кредитоспроможності позичальника

До методів комплексного аналізу належать: Правило «6С», PARSER, CAMPARI, PARTS, MEMO RISK, Система 4FC.

Кожен з методів, має свої переваги та недоліки (табл. 1.1).

Табл. 1.1 – Переваги та недоліки методик оцінки кредитоспроможності.

Назва методики	Сутність методики	Переваги	Недоліки
Класифікаційні (статистичні методики)	Приймаються до уваги показники, які описують кредитоспроможність клієнта. Вони можуть бути розглянуті, як в будь-який момент часу.	Велика частота використання та об'єктивна оцінка.	Необхідно сформувати велику кількість вихідних даних. Методика не бере до уваги слабоформалізовані якісні ознаки кредитоспроможності та умови кредитування. В умовах економіки України – неможливо застосувати.
Комплексного аналізу	Проставляються коефіцієнти важливості в балах, щодо кожної з ознак кредитоспроможності клієнта	Зменшує кількість витрат праці, збільшує кількість проаналізованих заявок, що забезпечує зменшення банківських витрат.	Висока вартість підготовки. Невеликі банки не можуть собі дозволити. Потребує багато часу для підготовки моделі, а також постійного оновлення ознак

До бально-рейтингових моделей статистичних методів включають кредитний скоринг. Кредитний скоринг – це метод оцінки кредитоспроможності людини, який спираючись на статистичні дані по кредитній історії банку проводить оцінку заявки на кредит потенційного клієнта, аналізуючи його соціально-демографічні ознаки, такі як вік, стать, освіта, посада, трудовий стаж, термін проживання в регіоні тощо [18].

Маючи базу даних випланих (успішних) і не випланих (провальних) кредитів, банківська установа за допомогою статистичного аналізу має змогу виявити основні фактори, що впливають на можливість позичальника повернути борг. Загально прийнято, що існує кореляція між певними соціальними даними і надійністю позичальника. З огляду на специфіку наявної інформації існують чотири види кредитного скорингу [19] (рис. 1.3):

- аплікаційний – оцінка кредитної спроможності клієнта, для прийняття рішення щодо можливості видачі йому кредиту;
- поведінковий – оцінка ймовірності повернення вже наданих кредитів. Скорингові моделі дозволяють спрогнозувати зміну кредитоспроможності клієнта, завдяки чому можна контролювати оптимальний ліміт по кредитній карті;
- колекторський – оцінка можливості повернення кредиту (повного або часткового) позичальником при порушенні ним термінів погашення заборгованості;
- антишахрайський скоринг – оцінка ймовірності шахрайських дій з боку потенційного позичальника.



Рис. 1.3 – Види кредитного скорингу.

Важливим при оцінюванні заявки позичальника є рейтинг клієнта, який повинен являти собою банківську оцінку можливості і готовності позичальника виконати контрактні зобов'язання незалежно від несприятливих економічних умов або настання неочікуваних подій [20].

Методика визначення рейтингу кредитоспроможності (як і в процесі визначення класу позичальника) базується на аналізі формалізованих і неформалізованих показників (критеріїв кредитоспроможності), які впливають на можливість підприємства виконувати свої фінансові зобов'язання. Всі показники об'єднуються в групи факторів, аналіз яких дозволяє оцінити ризики, пов'язані з певним аспектом діяльності підприємства [17]. Порядок визначення класу позичальника на основі внутрішнього кредитного рейтингу наведено на рис. 1.4.

Шкала внутрішнього кредитного рейтингу складається з таких груп: А, В, С, D та інгредієнтів: “+”, “нейтральний”, “–”. Рейтингова група А відповідає

нижчому рівню ризику, група D – найвищому відповідно, причому дана група має тільки два інгредієнти: “нейтральний” та “–”.

Бальна оцінка об’єктивних факторів ризику здійснюється шляхом аналізу фінансових коефіцієнтів та окремих показників, що характеризують фінансово-економічний стан контрагента [15]. На підставі суми балів, отриманих за результатами аналізу об’єктивних факторів ризику, визначається група рейтингу, до якої належатиме клієнт: група А, група В, група С або група D.

Бальна оцінка додаткових факторів ризику проводиться з метою корекції групи рейтингу. Результатом проведення оцінки об’єктивних та додаткових факторів ризику є визначення внутрішнього рейтингу клієнта у вигляді “А+”, “А нейтральний”, “А–”, “В+”, “В нейтральний”, “В–”, “С+”, “С нейтральний”, “С–”, “D нейтральний”, “D –”.

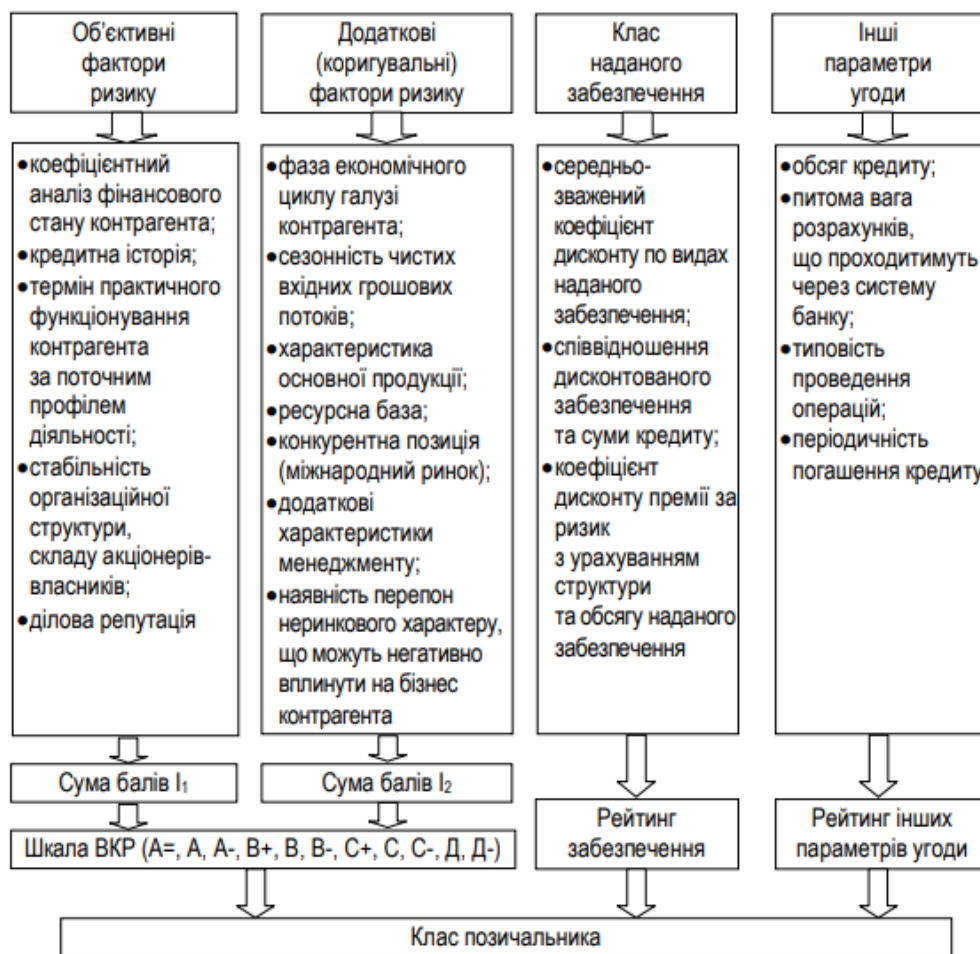


Рис 1.4. Процес визначення класу позичальника на основі внутрішнього кредитного рейтинга

Приймаючи до уваги недоліки українського банківського сектору, та крім цього значення кредитоспроможності клієнтів як головного фактору зменшення ризику у кредитуванні, покращення цієї сфери аналізу може розвиватися у двох напрямках: перший це – збільшення кількості базових показників до тих пір, поки б вони не описали діяльність клієнта максимально, а другий – це перевага у аналізі грошового потоку організації, що надає можливість оцінити кількість транзакцій коштів позичальника.

Аналіз грошового потоку — методика аналізу кредитоспроможності клієнта, що базується на використанні критеріїв, що описують обіг грошей клієнта за певний період.

Аналіз грошового потоку включає в себе порівняння притоку та відтоку грошей протягом періоду надання позики, та надає можливість проаналізувати кредитоспроможність клієнта. Якщо сума притоку грошей більша за суму відтоку, то клієнт здатний на погашення боргу у процесі кругообігу фондів.

Використання методів класифікації позичальників необхідний для виведення стандартних практик для об'єктивного аналізу клієнтів, виділення кількісних характеристик для визначення класу клієнтів на основі наданих ними інформації, на надійних і ненадійних.

Бально-рейтингова оцінка забезпечує прогнозування своєчасності здійснення виплат, оцінювання фінансового стану підприємства і його стійкість, а також дає можливість визначити межі зниження обсягу прибутку, в яких здійснюється погашення частки фіксованих платежів.

Перевагами бально-рейтингових моделей є простота, можливість розрахунку оптимальних значень за окремими показниками, здатність ранжирування підприємств за результатами, комплексний підхід до оцінки кредитоспроможності (використовуються показники, що відображають різні аспекти діяльності позичальника) [19].

Правило шести «С», в рамках якого по кожному позичальнику оцінюються:

- Характер (Character) - показник відповідальності позичальника.
- Платоспроможність (Capacity) - аналіз фінансового стану клієнта.

- Капітал (Capital) - оцінка грошового капіталу і майна позичальника.
- Забезпечення (Collateral) - оцінка забезпечення, яке може бути надано за кредитом і його ліквідності.
- Умови (Conditions) - загальні економічні умови в країні, світі, галузі, в якій діє позичальник, і прогноз їх змін на термін кредиту, тенденції розвитку ринку і зміни рівня попиту на продукцію позичальника.
- Грошові кошти (Currency) -Прогнозування потоку грошових коштів позичальника на весь період кредитування

Також, у світовій практиці для оцінки кредитоспроможності позичальника використовують такі підходи, як PARSEER і CAMPARI.

Таблиця 1.3 – Розшифрування методик PARSEER та CAMPARI

PARSEER	CAMPARI
P (person) – дані потенційного позичальника ; A (amount) –сума кредиту; R (repayment) – здатність погашення; S (security) – оцінка забезпечення; E (expediency) – доцільність кредиту; R (remuneration) – вигода банку (відсоток) за ризик надання кредиту.	C (character) – характеристика клієнта; A (ability) – можливість клієнта виплачувати кредит; M (margin) –дохід клієнта; P (purpose) – ціль позики; A (amount) – об’єм кредиту; R (repayment) – умови погашення позики; I (insurance) – страхування ризику непогашення позики.

Методика CAMPARI полягає у почерговому виділенні з кредитної заявки і документів, що додаються, найважливіших показників, що визначають діяльність клієнта в їх оцінці та уточненні після особистої зустрічі з клієнтом. Дана методика включає найбільшу кількість параметрів з розглянутих, що дозволяє найбільш точно визначати кредитоспроможність позичальника. Однак частина параметрів не може бути використана при оцінці кредитоспроможності українських організацій через відсутності даних показників. Наприклад, в Україні відсутні оцінки репутації позичальника, або такий критерій використовуються лише великими корпораціями, які мають іноземний капітал в структурі статутного капіталу [20].

Для оцінювання здатності виплати кредиту позичальником, ризикованості та класифікації кредитів часто застосовують модель CART. Це непараметрична модель, основними перевагами якої є можливість широкого застосування, доступність для розуміння і незначна трудомісткість розрахунків, хоча при побудові таких моделей застосовують складні статистичні методи.

Методика «PARTS» застосовується банками Великобританії і розшифровується:

P (Purpose) - мета отримання кредиту;

A (Amount) - обсяг кредиту;

R (Repayment) - сплата боргу та відсотків від нього;

T (Term) - термін надання кредиту;

S (Security) - гарантія сплати кредиту.

Комплексні методики оцінки кредитоспроможності позичальника застосовуються багатьма комерційними банками, однак ці методики недостатньо теоретично опрацьовані і в них мало використаний математичний апарат [17].

У світовій практиці досить широко застосовуються така система аналізу кредитоспроможності як MEMO RISK:

- Management – якість обслуговування;
- Experience – досвід;
- Market – загальні обставини роботи клієнта;
- Operations – оцінка професії позичальника;
- Repayment – оцінка здатності погашення позики;
- Interest – розмір відсоткової ставки;
- Security – безпека;
- Kontrol – контроль.

Головним недоліком даних комплексних методик є їх орієнтація здебільшого на якісні чинники, а також треба враховувати той факт, що дані моделі побудовані на основі експертних висновків та в окремих випадках можуть мати суб'єктивний характер.

#### 1.4 Аналіз існуючих методів інтелектуального аналізу даних для оцінки кредитоспроможності клієнтів банку

За останні роки, було проведено багато досліджень на основі аналізу даних у галузі фінансів та банківської справи. В рамках атестаційної роботи розглянуті деякі з них.

Популярним підходом, щодо вирішення завдання оцінки кредитоспроможності фізичних осіб є застосування алгоритмів, що вирішують задачу класифікації, а саме віднесення будь-якого об'єкта (потенційного позичальника) до одного із заздалегідь відомих класів («хороший»/«поганий»). Такого роду завдання можуть вирішуватися за допомогою дерев рішень.

Метод дерев рішень дозволяє автоматично аналізувати величезні масиви даних. Автор Судхакар М. у своїй роботі , зосередив свою увагу на методах, що використовують кілька алгоритмів для аналізу даних, таких як Decision Tree та штучні мережі Radial Basis Neural Networks [22]. У своїй роботі він розробив вигляд заявки, яку необхідно подавати для аналізу.

Отримана в результаті використання цього методу модель – це зручний засіб представлення правил у вигляді ієрархічної, послідовної структури, де кожний об'єкт потрапляє в єдиний для нього кінцевий вузол. Під правилом розуміється логічна конструкція, представлена у формі «якщо..., то...». Менеджери найчастіше дають перевагу даному підходу за його наочність та інтуїтивну зрозумілість процесу прийняття рішення у вигляді простих та зрозумілих правил (рис. 1.5).

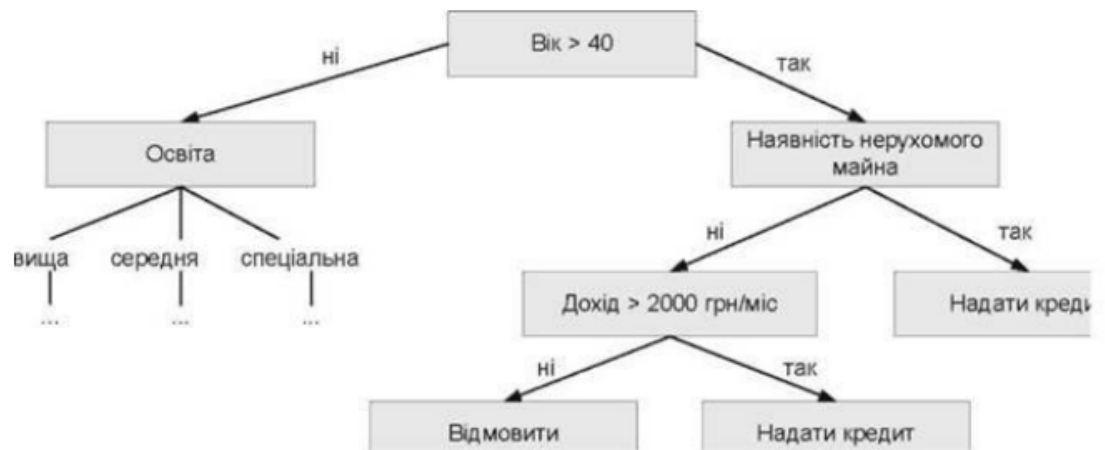


Рис.1.5 – Приклад побудованого дерева рішень

Розглянемо сутність даного підходу. Дерево будується на основі даних з минулих періодів, завдяки чому заздалегідь відомо до якого класу відноситься кожна з ситуацій. Тобто має бути відомо, чи був погашений кредит. Спочатку всі ситуації з навчальної вибірки потрапляють в верхній (перший) вузол, а потім вони розподіляються по нижніх вузлах, що теж можуть бути розподілені на дочірні вузли. Критерієм розбиття є певні значення одного з вхідних параметрів. Для визначення поля, по якому буде відбуватися розбиття, використовується ентропія (міра невизначеності). Обирається те поле, при розбитті по якому вдасться найбільше позбавитися невизначеності. Чим більше об'єктів, що відносяться до різних класів (домішок) знаходиться в одному вузлі, тим більшою є невизначеність. Якщо у вузлі знаходяться об'єкти, що відносяться до одного класу, то ентропія повинна дорівнювати нулю. Після побудови моделі на навчальній вибірці, отриману модель можна використовувати для визначення класу («хороший»/«поганий») нових ситуацій, тобто коли потенційний клієнт хоче отримати кредит. Якщо на ринку істотно змінюється ситуація, то модель дерева можна перебудувати, адаптувавши до існуючої обстановки.

У своїй роботі Джозеф Зурада та співавтори, описали як за допомогою дерев рішень можливо проводити оцінку кредитних ризиків [23]. Вони направили свої зусилля на розробку архітектури штучної мережі для оцінки заявок на комерційні

позики. Розроблена модель мережі змогла класифікувати близько 75% заявок на кредити.

На сьогоднішній день відомо багато алгоритмів, що дозволяють побудувати дерева рішень: QUEST, CART, CHAID, C4.5, CN2, C&R тощо [19].

Метод CHAID (Chi-Square Automatic Interaction Detection – автоматичний детектор взаємозв'язків на основі критерію  $\chi^2$ ) розроблений в 1980 р. Є найбільш вживаним і швидкодіючим багатовимірним статистичним методом побудови дерева рішень, що базується на використанні критерію зв'язку  $\chi^2$  для пошуку оптимального розбиття між категоріальними змінними. При необхідності кожна вершина дерева може бути поділена більше ніж на дві вершини наступного рівня. У випадку інтервальної залежної змінної в якості критерія оптимізації використовується F-критерій Фішера. Якщо незалежні змінні є інтервальними, то вони автоматично перетворюються в категоріальні, окрім того кількість категорій можна змінювати [24].

В 1991р. був розроблений метод Вичерпний (Exhaustive) CHAID, що є модифікацією методу CHAID. Перевага даного методу полягає в тому, що при побудові дерева відбувається попередній аналіз всіх можливих розбиттів на наступних етапах алгоритму. Звичайно такі аналіз потребує більше ресурсів та часу.

Метод QUEST (Quick, Unbiased, Efficient Statistical Trees) – швидкі, незсунені, результативні статистичні дерева) був розроблений для усунення недоліків попереднього методу. Цей статистичний метод призначений для швидкої і ефективно побудови бінарних дерев був розроблений в 1997 р. Однак, цей метод застосовується, тільки в тому випадку якщо цільова змінна є номінальною. В залежності від типу тої чи іншої незалежної змінною використовуються різні статистичні критерії, а тому незалежні змінні можуть бути будь-якими.

У роботі Чаван П. М. проводиться опис різних алгоритмів аналізу даних, та їх поєднання за допомогою техніки Bagging [24]. Навчання методів проводиться на банківських даних, проте на вибірках різних країн. Результати показали, що на

різних даних деякі алгоритми працюють краще, деякі гірше. Саме в роботі найкраще показав себе Random Forest, хоча прогнози на основі техніки Bagging показали гарні результати.

Лінійна імовірнісна модель (ЛІМ) – це модель у формі лінійної регресії, залежна змінна якої набуває значення 0 або 1 залежно від того, яким є результат повернення особою наданого споживчого кредиту.

Логістична регресія – є корисним класичним інструментом для вирішення завдань регресії та класифікації. В останні роки логістична регресія набула поширення в скорингу для розрахунку рейтингу позичальників і управління кредитними ризиками. Логістична регресія є різновидом множинної регресії. Призначення логістичної регресії – аналіз зв'язку між цільовою змінною та незалежними змінними, або, як їх ще називають, регресорами чи предикторами. Якщо цільова змінна приймає тільки два значення (наприклад, 0 або 1), тобто є бінарною, то в такому випадку застосовується бінарна логістична регресія [23]. Логістична регресія дозволяє оцінити ймовірність того, що відбудеться якась конкретна подія для певного випробування (наприклад, повернення кредиту).

Логістична регресія фактично служить не для передбачення значень цільової змінної, а скоріше для оцінки ймовірності того, що залежна змінна прийме задане значення.

### 1.5 Системи підтримки прийняття рішень

Система підтримки прийняття рішень (СППР) – це комп'ютеризована автоматизована система, метою якої є допомога людям, які приймають рішення в складних умовах для повного і об'єктивного аналізу предметної області. СППР, на основі вхідних даних, видає інформацію, що допомагає людині швидко і точно оцінити ситуацію та прийняти рішення. Особливу увагу потрібно надати надзвичайним ситуаціям, в моменти яких, людині складно оцінювати проблему та її ознаки об'єктивно, та приймати зважені рішення, у чому можуть зіграти ключову роль СППР [26, 27].

СППР використовують різні методи для аналізу та видачі інформації. Найпоширеніші з методів аналізу даних у СППР:

- інформаційний пошук;
- інтелектуальний аналіз даних;
- пошук знань у базі даних;
- висновки на основі прецедентів;
- імітаційне моделювання;
- еволюційні обчислення;
- генетичні алгоритми;
- нейронні мережі.

Концептуальна модель СППР зображена на рисунку 1.6.



Рис. 1.6 – Концептуальна модель СППР

Розробка СППР, як правило є результатом дослідження проведеного не тільки в рамках комп'ютерних наук, а й в рамках предметної області та методів її аналізу. Існує декілька видів СППР на рівні взаємодії з користувачем, а саме активні, пасивні та кооперативні СППР. Пасивна СППР характеризується лише допомогою процесу прийняття рішень, та не може видати прогноз, яке саме рішення потрібно прийняти. Активна СППР може зробити прогноз щодо прийняття рішення. Кооперативна СППР дозволяє людині, що приймає рішення змінювати, додавати і покращувати рішення, які прогнозуються системою. Процес продовжується до отримання однакового рішення.

На концептуальному рівні виділяються наступні типи СППР:

- на основі управління повідомленнями;

- на основі управління даними;
- на основі управління документами;
- на основі управління моделями.

Кожен тип, що приведений вище виконує кінцеву задачу у підтримці прийняття рішення за допомогою окремих особливостей.

## 1.6 Постановка задачі

У рамках атестаційної роботи для досягнення мети, щодо класифікації банківських кредитів потрібно вирішити наступні задачі:

- a) провести аналіз досліджуваної предметної області, а саме:
  - провести аналіз та структурування кредитних ризиків;
  - провести дослідження СППР та їх розробки;
  - провести аналіз існуючих методів інтелектуального аналізу даних;
- b) вибір методу оцінювання параметрів моделей;
- c) розробити комбінований метод аналізу даних на основі алгоритмів CART та SVM.
- d) спроектувати СППР, що використовує комбінований метод;
- e) провести тестування комбінованого методу та провести аналіз його точності.

Для проектування системи необхідно виконати наступні задачі:

- визначити сферу застосування системи, що розроблюється;
- розробити системні та функціональні вимоги до системи, що проектується;
- побудувати діаграму варіантів використання;
- провести обґрунтування вибору СУБД;
- провести логічне та фізичне моделювання даних з реалізацією бази даних;

– видати рекомендації для використання СППР на основі комбінованого методу аналізу даних, щодо класифікації заявок клієнтів на банківські кредити.

## 2 КОМБІНУВАННЯ МЕТОДІВ АНАЛІЗУ ДАНИХ

### 2.1 Основні положення методу CART

Основне призначення методу – це визначення класу, поданого на вхід об'єкту. Кожне бінарне дерево складається з вузлів, кожен вузол має лише два нащадки, що розділяє об'єкти на виділені групи. Ці нащадки мають назви лівий «left» і правий «right». Для розбиття нащадків на виділені групи виконується перевірка правила, що формується на кожному вузлі дерева. Якщо умова виконується, то ці нащадки відносять до правої групи, якщо ні, то до лівої. На рисунку 2.1 представлений приклад дерева рішень на основі даних для апроксимації синусоїдальної кривої.

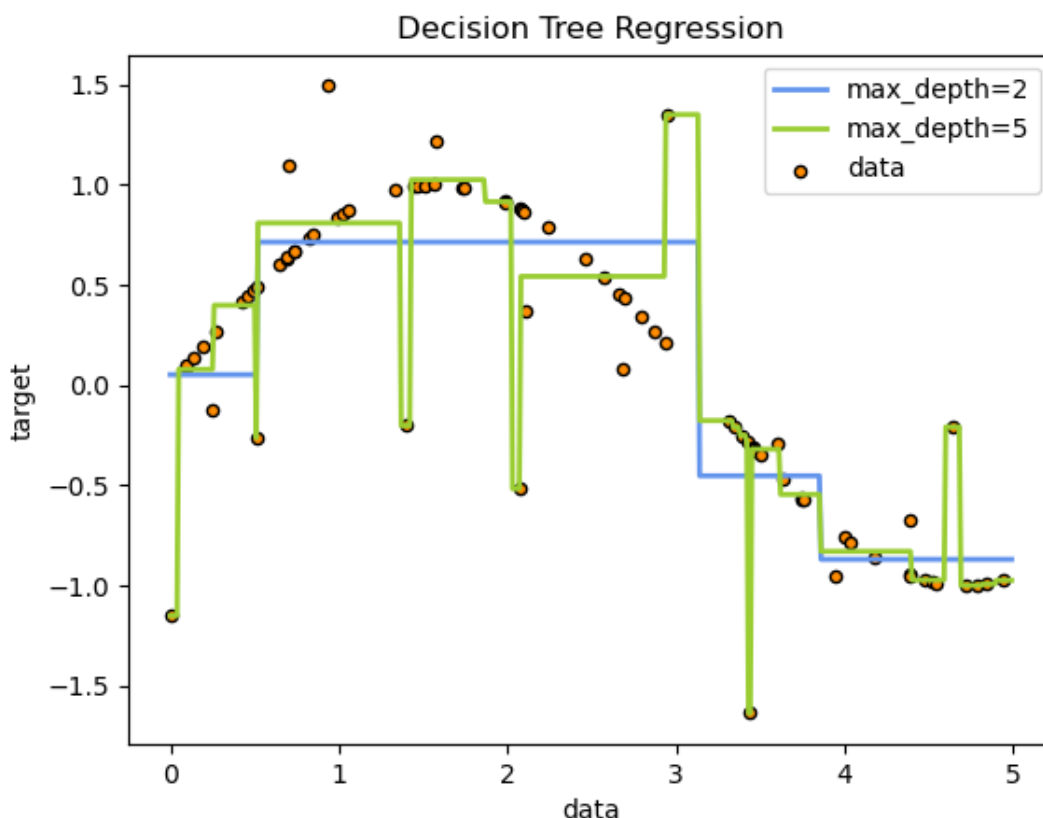


Рис. 2.1 - Приклад дерева рішень на основі даних для апроксимації синусоїдальної кривої.

Метод CART застосовується для номінальних (зазвичай дворівневих) і порядкових змінних. У цьому методі перебираються всі можливі варіанти розгалуження для кожного вузла, і вибирається та змінна, при якій оцінна функція дає найкращий показник [28]. Для вирішення задач атестаційної роботи, було обрано застосовувати метод на номінальних змінних, для вирішення задачі класифікації заявок на отримання банківських кредитів.

Метод CART має свої переваги та недоліки, які представлені на приведеній нижче таблиці 2.1.

Табл. 2.1 – Переваги та недоліки методу CART.

Переваги	Недоліки
Простий для розуміння та тлумачення. Дерева можна візуалізувати.	Навчені дерева прийняття рішень можуть створювати надскладні дерева, які погано узагальнюють дані. Це називається перенавчанням. Такі механізми, як обрізка, встановлення мінімальної кількості зразків, необхідних у листовому вузлі, або встановлення максимальної глибини дерева необхідні, щоб уникнути цієї проблеми.
Потрібна невелика підготовка даних. Інші методи часто вимагають нормалізації даних, потрібно створити фіктивні змінні та видалити порожні значення. Однак зауважте, що цей модуль не підтримує відсутні значення.	Дерева рішень можуть бути нестабільними, оскільки невеликі варіації даних можуть призвести до генерування зовсім іншого дерева. Цю проблему пом'якшує використання дерев рішень у межах групи.
Вартість використання дерева (тобто прогнозування даних) є логарифмічною за кількістю точок даних, що використовуються для навчання дерева.	Практичні алгоритми навчання на дереві рішень засновані на евристичних алгоритмах, таких як жадібний алгоритм, де локально оптимальні рішення приймаються на кожному вузлі. Такі алгоритми не можуть гарантувати повернення глобально оптимального дерева рішень. Це можна пом'якшити шляхом навчання декількох дерев у групі, яка навчається, де функції та вибірки будуються із заміною.
Здатний обробляти як числові, так і категоріальні дані. Інші методи, як правило, спеціалізуються на аналізі наборів даних, які мають лише один тип змінних. Дивіться алгоритми для отримання додаткової інформації.	Навчені дерева прийняття рішень створюють упереджені дерева, якщо деякі класи домінують. Тому рекомендується збалансувати набір даних перед узгодженням з деревом рішень.
Здатний вирішувати проблеми з кількома виходами.	
Використовує модель білої коробки. Якщо	

дана ситуація спостерігається в моделі, пояснення умови легко пояснюється булевою логікою. Навпаки, у моделі чорного ящика (наприклад, у штучній нейронній мережі) результати можуть бути складнішими для інтерпретації.	
Можливо перевірити модель за допомогою статистичних тестів. Це дає змогу враховувати надійність моделі.	
Добре працює, навіть якщо його припущення дещо порушені справжньою моделлю, з якої були створені дані.	

Дано вектор для навчання  $x_i \in R^n$ ,  $i = 1, \dots, l$  та вектор мітки  $y \in R^l$ , дерево рішень рекурсивно розділяє простір так, що зразки з однаковими мітками групуються разом. Нехай дані на вузлі  $m$  будуть представлені як  $Q$ . На кожного кандидата розділити  $\theta = (j, t_m)$  що складається з ознаки  $j$  і поріг  $t_m$ , розділити дані на  $Q_{left}(\theta)$  і  $Q_{right}(\theta)$  підмножини (формула 2.1).

$$\begin{aligned} Q_{left}(\theta) &= (x, y) | x_j \leq t_m \\ Q_{right}(\theta) &= Q \setminus Q_{left}(\theta) \end{aligned} \quad (2.1)$$

Домішка при  $m$  обчислюється за допомогою домішкової функції  $H()$ , вибір якого залежить від вирішуваного завдання (формула 2.2).

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \quad (2.2)$$

Для мінімізації домішок використовують критерій (формула 2.3)

$$\theta^* = \operatorname{argmin}_{\theta} G(Q, \theta) \quad (2.3)$$

Рекурсивно застосувати для підмножин  $Q_{left}(\theta^*)$  та  $Q_{right}(\theta^*)$  до тих пір, поки не буде досягнена максимально дозволена глибина.

Якщо ціль - це результат класифікації, який приймає значення  $0, 1, \dots, K-1$  для вузла, що представляє регіон  $R_m$  з  $N_m$  спостереження, нехай (формула 2.4)

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i = k) \quad (2.4)$$

Звичайно для оцінки якості моделі використовують індекс Gini (формула 2.5), також ентропію (формула 2.6) та неправильну класифікацію (формула 2.7)

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \quad (2.5)$$

$$H(X_m) = - \sum_k p_{mk} \log(p_{mk}) \quad (2.6)$$

$$H(X_m) = 1 - \max(p_{mk}) \quad (2.7)$$

Де  $X_m$  – це дані навчання у вузлі  $m$ .

В результаті, на виході можна візуалізувати наше дерево, для кращого сприйняття та аналізу. На рисунку 2.2 представлений приклад візуалізації дерева рішень для вирішення задачі класифікації, навчений на тестових даних про рослини.

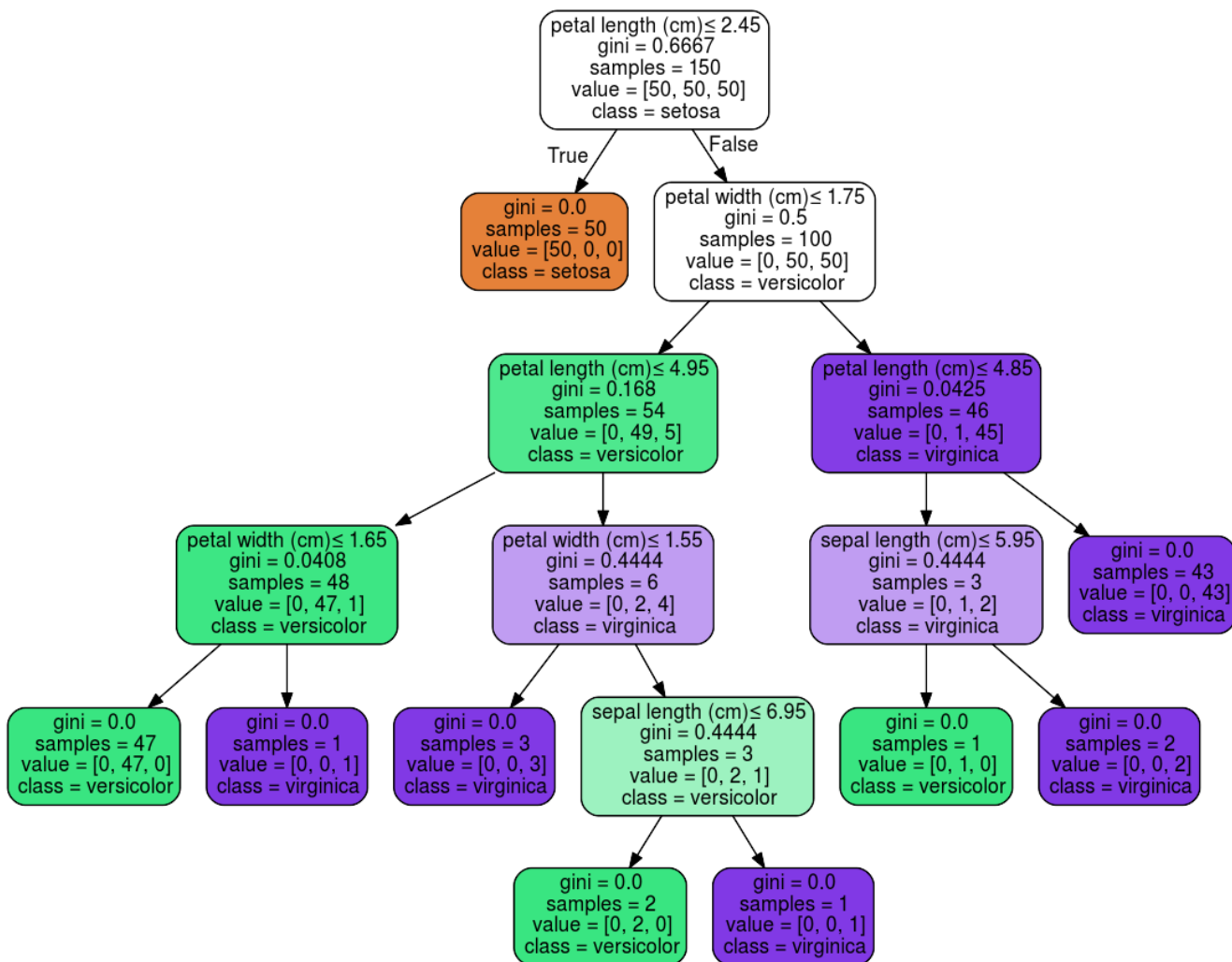


Рис. 2.2 – Приклад візуалізації дерева рішень, для вирішення задачі класифікації.

## 2.2 Основні положення методу SVM

Основне призначення методу – вирішення задачі бінарної класифікації. Бінарна класифікація – це тоді коли дані необхідно розділити лише на два класи. Спочатку алгоритм тренується на об'єктах з навчальної вибірки, для яких заздалегідь відомі мітки класів. Далі вже навчений алгоритм пророкує мітку класу для кожного об'єкта з тестової вибірки. Мітки класів можуть набувати значень  $Y = \{-1, +1\}$ . Об'єкт - вектор з  $N$  ознаками  $x = (x_1, x_2, \dots, x_n)$  в просторі  $R^n$ . При навчанні алгоритм повинен побудувати функцію  $F(x) = y$ , яка приймає в себе аргумент  $x$  - об'єкт з простору  $R^n$  і видає мітку класу  $y$  [29].

Головна мета SVM як класифікатора - знайти рівняння розділяє гіперплощини  $w_1x_1 + w_2x_2 + \dots + w_nx_n + w_0 = 0$  в просторі  $R^n$ , яке б розділила два класи якимось оптимальним чином [30].

Загальний вигляд перетворення  $F$  об'єкта  $x$  в мітку класу  $Y: F(x) = \text{sign}(w^T x - b)$ . Якщо  $w = (w_1, w_2, \dots, w_n)$ , то  $b = -w_0$ . Після налаштування вагових коефіцієнтів алгоритму  $w$  і  $b$  для навчання, всі об'єкти, які потрапляють по одну сторону від побудованої гіперплощини, будуть позначені як перший клас, а об'єкти, що потрапляють по іншу сторону – другий клас. Функція  $\text{sign}()$  використовує лінійну комбінацію ознак об'єкта, з ваговими коефіцієнтами алгоритму, саме тому алгоритм SVM відноситься до лінійних алгоритмів Data Mining.

Гіперплощина може будуватися різними способами, але у методі SVM вагові коефіцієнти  $w$  і  $b$  налаштовуються таким чином, щоб об'єкти класів лежали якнайдалі від поділяючої гіперплощини. Іншими словами, алгоритм максимізує відстань між об'єктами різних класів та гіперплощиною, які розташовані до неї найближче (рисунок 2.3).

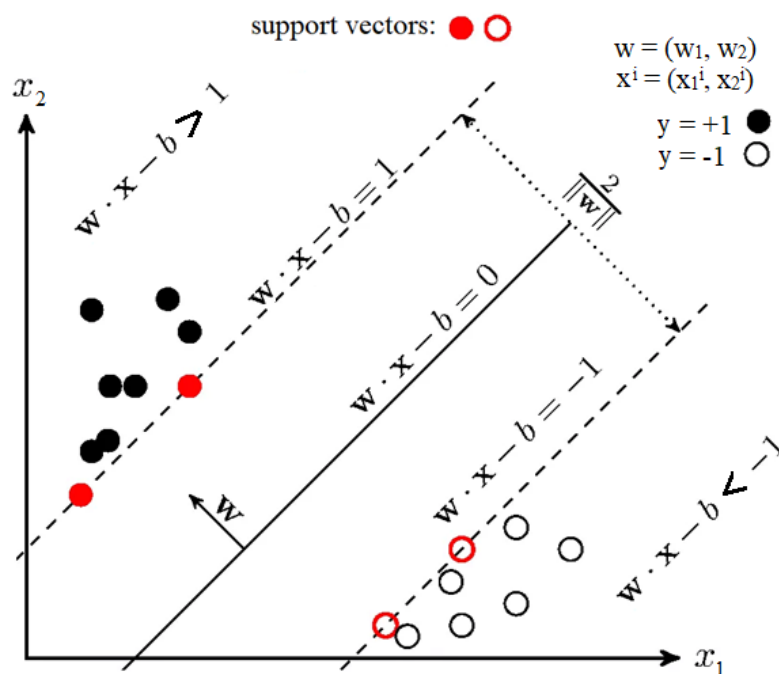


Рис. 2.3 – Графічне представлення роботи методу SVM (опорних векторів)

Для налаштування вагових коефіцієнтів в SVM, для того щоб гіперплощина знаходилась якнайдалі від точок вибірки ширина смуги має бути максимальною. Вектор  $w$  – вектор нормалі до поділяючої гіперплощини.

В рамках атестаційної роботи, SVM використовувався для вирішення задачі бінарної класифікації заявок на отримання банківських кредитів.

### 2.3 Комбіноване використання методу CART та SVM

У разі комбінованого використання методів CART та SVM на початковому етапі було вирішено використовувати метод на основі дерева рішень. За допомогою нього були проставлені вагові коефіцієнти для кожної з ознак, що дозволило виділити найбільш та найменш вагомі параметри моделі.

Для початку, необхідно було провести навчання методу. Відбувається ініціалізація параметрів користувача, зокрема  $W_{min}$  який служить для відсіювання ознак, що мають значення важливості  $W < W_{min}$ .

На вхід до методу подаються дані, проходить поверхневий аналіз даних, де метод перевіряє, чи є пропущені значення у даних, що поступили. Якщо пропущені значення є, то відбувається їх заповнення відповідно до медіанного значення цієї ознаки, якщо медіанне значення не дає результат, що близький до справжнього, то метод відсіює ознаку повністю, для того щоб покращити прогнози після навчання. Після того як пусті значення заповнені, алгоритм проводить нормалізацію даних, для подальшого аналізу.

Детальніше розглянемо на прикладі. Існує декілька типів клієнтів:

- клієнти, що виплачували кредити вчасно (тип С);
- клієнти, що не брали кредити (тип Х);
- клієнти, що брали позику, та мають заборгованість протягом від 1 до 29 днів (тип 0);
- клієнти, що брали позику, та мають заборгованість протягом від 30 до 59 днів (тип 1);

- клієнти, що брали позику, та мають заборгованість протягом від 60 до 89 днів (тип 2);
- клієнти, що брали позику, та мають заборгованість протягом від 90 до 119 днів (тип 3);
- клієнти, що брали позику, та мають заборгованість протягом від 120 до 149 днів (тип 4);
- клієнти, що брали позику, та мають заборгованість протягом від 150 днів, та вважаються безнадійними (тип 5).

Для побудови найбільш безпечної моделі, було прийнято рішення визначити типи С та Х, як надійних, а типи 0, 1, 2, 3, 4, 5 вважати боржниками. На рисунку 2.4 представлено, як позначені ці класи у методі. Тобто на виході методу буде клас, чи клієнт надійний, чи клієнт боржник.

	ID	MONTHS_BALANCE	STATUS	Label
0	5001711	0	X	0
1	5001711	-1	0	1
2	5001711	-2	0	1
3	5001711	-3	0	1
4	5001712	0	C	0
5	5001712	-1	C	0
6	5001712	-2	C	0
7	5001712	-3	C	0
8	5001712	-4	C	0
9	5001712	-5	C	0

Рис. 2.4 – Відношення типу клієнта, до його класу, де 0 – надійний, 1 – боржник.

На прикладі показано як відбувається нормалізація даних для типу клієнта, таким же чином відбувається нормалізація і для інших ознак.

Далі дані розбиваються на навчальну та тестову вибірки, і йдуть на вхід до класифікатора CART, який проводить класифікацію, та розставляє  $W$  для кожної з ознак. Після отримання вагових коефіцієнтів, ознаки проходять відсіювання, при  $W < W_{min}$ . Дані остаточно оброблені, після чого ставиться задача бінарної класифікації, для вирішення якої було обрано використовувати метод SVM.

Метод опорних векторів класифікацію та видає остаточний результат, чи можна схвалювати заявку на кредит. На рисунку 2.5 представлений алгоритм, використаний для навчання методу, який комбінує в собі CART та SVM.

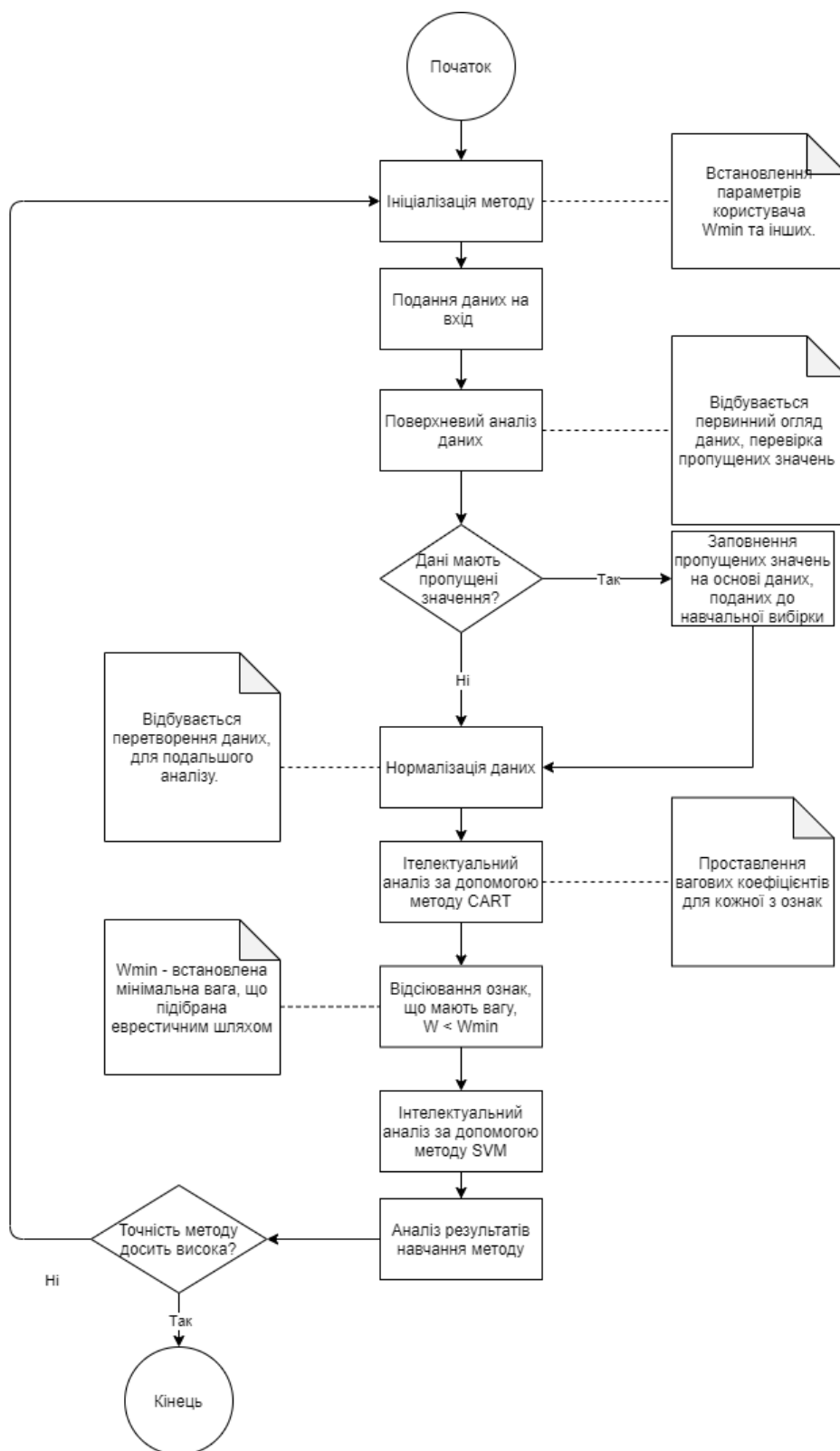


Рис. 2.5 – Алгоритм навчання комбінованого методу на основі CART та SVM

### 3 ВИЗНАЧЕННЯ ФУНКЦІОНАЛЬНИХ ВИМОГ СИСТЕМИ

Для проектування СППР необхідно виконати наступні задачі:

- визначити сферу застосування системи, що розроблюється;
- розробити системні та функціональні вимоги до системи, що проектується;
- побудувати діаграму варіантів використання;
- провести обґрунтування вибору СУБД;
- провести логічне та фізичне моделювання даних з реалізацією бази даних;
- видати рекомендації для використання СППР на основі комбінованого методу аналізу даних, щодо класифікації заявок клієнтів на банківські кредити.

В рамках атестаційної роботи було вирішено, що система підтримки прийняття рішень, щодо видачі банківських кредитів, представляє собою веб-орієнтований додаток, спроектований на клієнт-серверній архітектурі. Мається на увазі, що розроблену систему будуть використовувати співробітники банку, та вносити клієнтські дані для аналізу заявок на отримання банківських кредитів.

Для визначення функціональних вимог системи, було обрано стандарт методології IDEF0 (I-CAM DEFinition) [31]. Модель IDEF представляє собою ієрархічну модель опису бізнес-процесів. Процеси моделі можна декомпонувати до тих пір, поки не буде досягнута необхідна кількість деталей. На рисунку 3.1 представлена концептуальна діаграма функції системи.

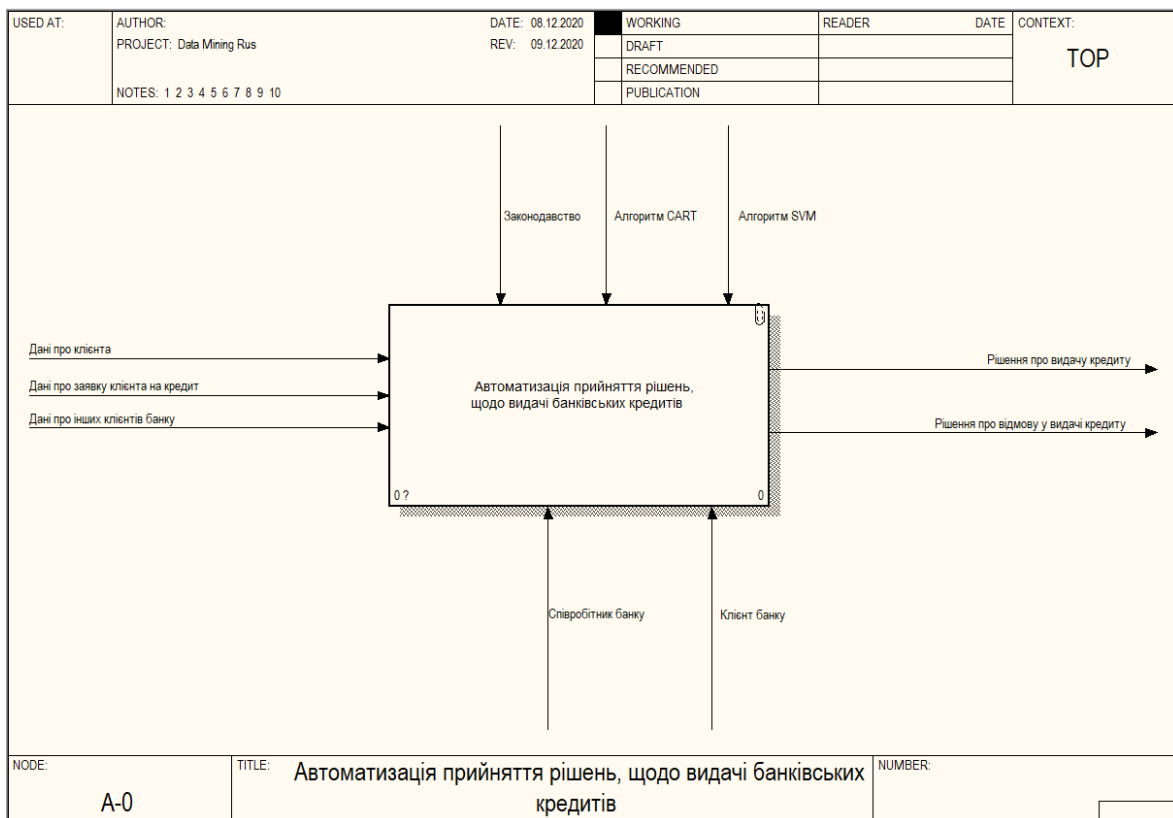


Рисунок 3.1 – Концептуальна діаграма функції «Автоматизація прийняття рішень, щодо видачі банківських кредитів» з точки зору користувача

Для продуктивної обробки та аналізу даних, необхідно зібрати дані, які будуть корисні для отримання ефективного прогнозу. Під корисними даними, передбачається така інформація, яка має вагу щодо вирішення надання чи відмови у позиції. Загалом, кожна людина, що приходить до банку, в очікуванні отримати кредит, надає первинні дані, які потрапляють до системи. Ця інформація, що необхідна для заповнення передбачена системою і вважається корисною. Подається для заповнення у вигляді форми, дані вводяться за допомогою співробітника банку. На рисунку 3.2 представлені декомпозиція основної функції системи.

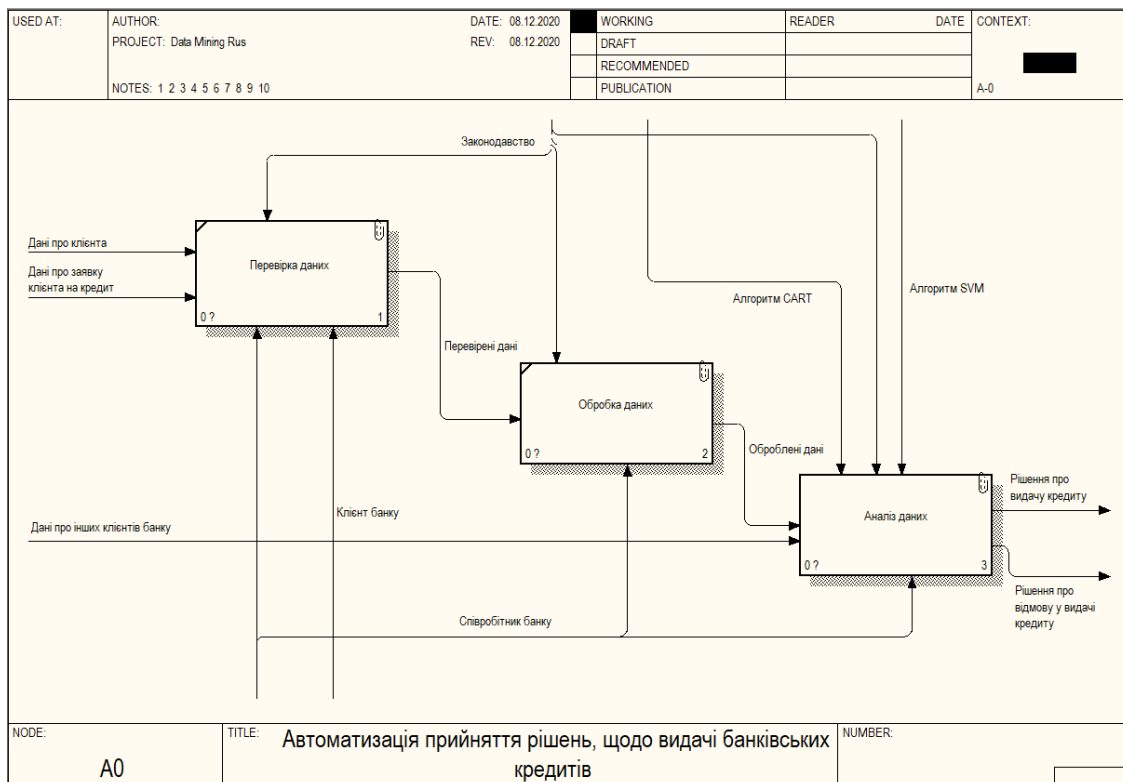


Рисунок 3.2 – Діаграма декомпозиції, що представляє функції системи

Наступним кроком є взаємодія з блоком обробки даних, де проходить перевірка інформації та внесення до системи. Система передбачає використання інформації з різних джерел. Далі проводиться пошук інформації про клієнта на різних сервісах. Після видобутку всієї доступної інформації, збір даних завершено і вони передаються у блок обробки.

Будь-яка інформація, що потрапляє до системи є необробленою, тобто кожна ознака клієнта, яка може надходити до системи з інших систем чи підсистем, має проходити перевірку, перетворення, нормалізацію.

Для того, щоб метод аналізу даних для класифікації заявок на кредити відпрацьовував найбільш ефективно, система виконує перевірку даних. Наприклад, користувач залишив номер телефона, відбувається перевірка номера, чи дійсно він належить цьому користувачеві. Перевірка паспортних даних, перевірка нерухомості та ін. Кожна з ознак, що пройшла, або не пройшла перевірку подається до системі у вигляді булевої ознаки.

Деякі дані, що надає користувач, мають лише інформативний характер і при аналізі можуть дати слабкий результат, але якщо їх трохи перетворити то ситуація зміниться. Наприклад, користувач вказав свій робочий телефон, система обробить ці дані та використає у якості ознаки, не сам номер, а ознаку його належності у вигляді булевого показника.

Всі дані зберігаються у базу даних, на основі СУБД Oracle, тож перед внесенням вони проходять нормалізацію до 3 нормальної форми.

Блок аналізу даних є основою для прийняття рішення, щодо підтвердження чи відхилення заявки клієнта на отримання кредиту. Дані, що потрапляють до цього блоку є обробленими, перевіреними та нормалізованими. Для аналізу даних, використовується метод на основі комбінування алгоритмів CART та SVM, який поєднує в собі деякі переваги обох методів. На рисунку 3.3 представлена декомпозиція функції «Аналіз даних». Прогнозування проходить за допомогою комбінованого методу, до якого подаються детальні дані клієнта.

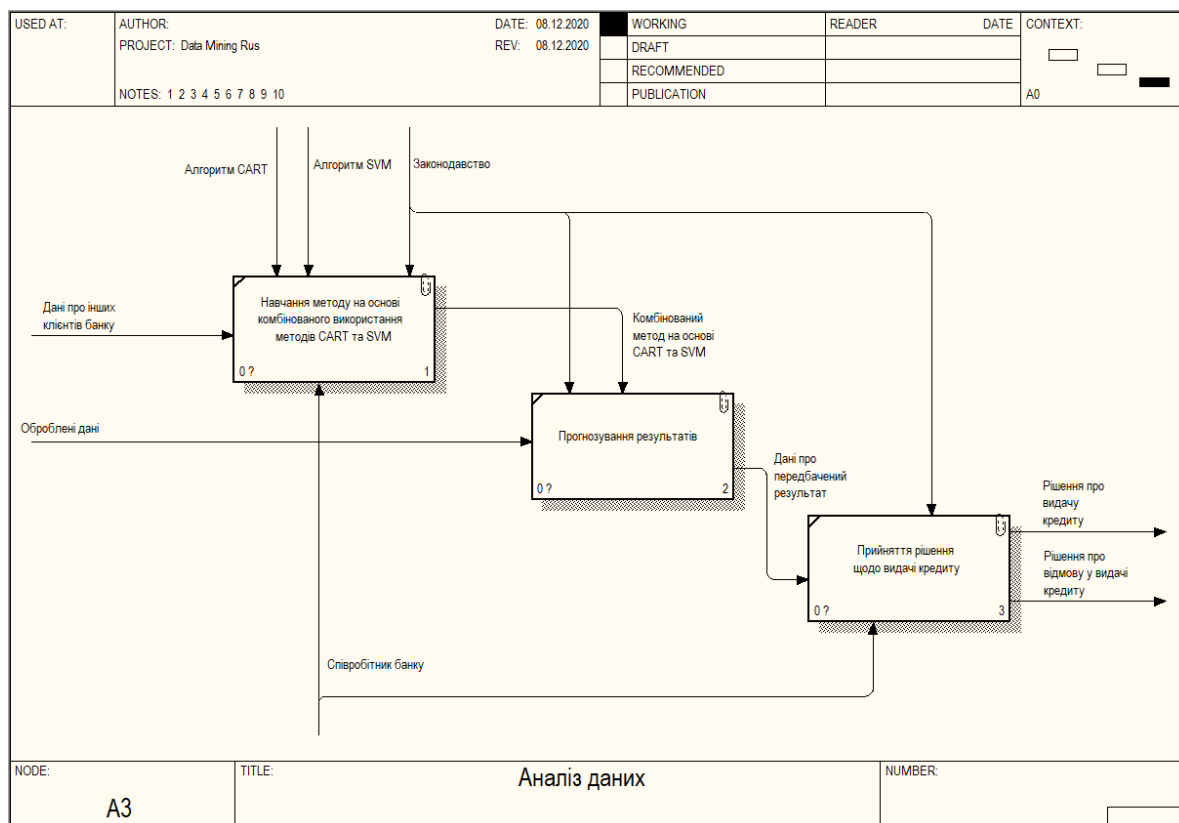


Рисунок 3.3 – Діаграма декомпозиції функції «Аналіз даних»

Спочатку метод проходить навчання на вибірці всіх клієнтів, де за допомогою дерева рішень визначаються вагові коефіцієнти для кожної з ознак, далі відсіюються неважливі ознаки та дані передаються до алгоритму SVM. Після навчання, система може приступити до прогнозування, щодо видачі кредиту клієнтові. Система надає своє рішення працівнику, з додатковими метриками точності, співробітник банку приймає рішення на основі отриманих результатів.

Розроблена модель функціональних вимог, з усіма описаними функціями використана у проектуванні та розробці системи. Було виділено концептуальний зміст системи, основні функції та важливі моменти головної функції, що виконує аналіз даних.

## 4 ПРОЕКТУВАННЯ СИСТЕМИ ПІДТРИМКИ РІШЕНЬ ДЛЯ ОЦІНКИ ЗАЯВОК НА ОТРИМАННЯ КРЕДИТУ

### 4.1 Діаграма варіантів використання системи

При проектуванні системи важливою складовою є виділення ролей та варіантів використання системи. Діаграма варіантів використання (Use case) описує взаємовідносини і залежності між групами варіантів використання і дійовими особами, які беруть участь в процесі [32]. На рисунку 4.1 представлена Use Case діаграма системи.



Рис. 4.1 – Use Case діаграма системи

Відповідно до діаграми, наведеної вище, було виділено дві основні ролі у системі: співробітник банку, клієнт банку. Співробітник банку може використовувати наступні функції системи:

- реєстрація;
- авторизація;
- створення заявки на кредит;
- керування інформацією про клієнта, що включає в себе:
  - внесення інформації про клієнта;
  - перегляд інформації про клієнта;
- проведення аналізу заявки клієнта на кредит, що включає в себе:
  - перегляд результатів після аналізу;
  - експорт даних, отриманих після аналізу;
- керування рішенням про видачу позики, що включає в себе:
  - підтвердження заявки на кредит;
  - відмова у кредитуванні;
  - формування документу про результат.

Клієнт банку у свою чергу може користуватися наступними функціями системи:

- створення заявки на кредит;
- перегляд заявки на кредит;
- авторизація;
- перегляд рішення, щодо його заявки на кредит;

Таким чином, співробітник банку, разом з клієнтом банку, заповнюють заявку на отримання кредиту, вносять необхідну інформацію про клієнта до системи. Співробітник переглядає передбачення та необхідні метрики, сформовані системою, на основі чого, виносить рішення, підтвердити чи відмовити у кредитуванні. Клієнт, може зайти у систему та перевірити дані, що містяться про нього у системі, а також дізнатися рішення щодо отримання одержання позики.

## 4.2 Розробка структурної схеми інформаційної системи підтримки прийняття рішень для аналізу заявок на отримання позики

Завдяки тому, що функціональні вимоги визначені, ролі та прецеденти системи з'ясовано, веб-системи, що подібні до розроблюваної проаналізовано, було розроблену структурну схему, на якій зображені основні компоненти системи (рисунок 4.2).

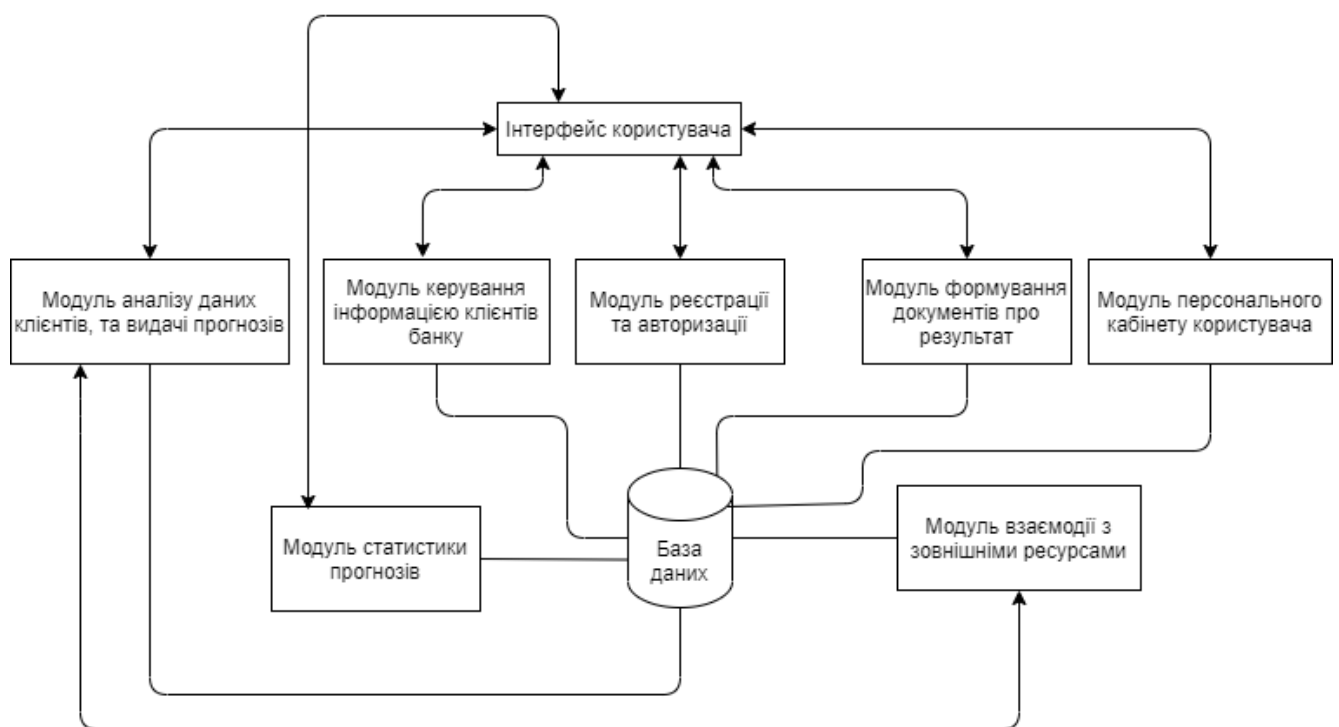


Рис. 4.2 – Структурна схема компонентів системи, та їх взаємодії

На структурній схемі представлені наступні компоненти проектованої системи:

- інтерфейс користувача;
- модуль аналізу даних клієнтів, та видачі прогнозів;
- модуль керування інформацією клієнтів банку;
- модуль реєстрації та авторизації;
- модуль формування документів про результат;
- модуль персонального кабінету користувача;

- модуль взаємодії з зовнішніми ресурсами;
- база даних.

Всі компоненти мають свої зв'язки з іншими компонентами системи, деякі можуть бути використані через інтерфейс користувача, а деякі лише іншими компонентами. Структурна схема дозволяє більш детально розібратися у зв'язках модулів між собою.

### 4.3 Обґрунтування архітектури системи

Для розробки системи, було прийнято рішення використовувати клієнт-серверну архітектуру. Така архітектура дозволяє розділити навантаження між клієнтом і сервером. Клієнт-серверна архітектура описує зв'язок між двома комп'ютерними додатками, клієнтський додаток виконує запит на обслуговування серверному додатку. Ця особливість дозволяє декільком користувачам мати доступ до однієї і тієї ж бази даних одночасно, а також зберігати та накопичувати великі об'єми даних [33].

Клієнт розпочинає спілкування та використовується для запускання функцій та доступу до даних, що зберігаються на сервері. Зазвичай, клієнтські додатки менш складні ніж серверні, адже вони, та у більшості випадків, не потребують спеціальних системних дозволів.

Сервер оброблює запрошені запити на сервіси з клієнту. Іноді для цього необхідні додаткові можливості, до яких клієнтський додаток не має доступу. Серверний додаток також використовується як інтерфейс доступу до бази даних.

Існують наступні типи клієнт серверної архітектури:

- 2-шарова архітектура;
- 3-шарова архітектура;
- N-шарова архітектура.

Кожен тип архітектури використовують для різних цілей. Розглянемо детальніше кожен з типів.

#### 4.3.1 Опис 2-шарової клієнт-серверної архітектури

У 2-шаровій клієнт-серверній архітектурі робоче навантаження розділено між клієнтським додатком, на якому розміщений інтерфейс користувача, та серверним додатком, на якому знаходяться основні сервіси та база даних. В дійсності ж, обидва додатки можуть бути розміщені на одному комп'ютері. На рис. 4.3 зображена схема 2-шарової архітектури.

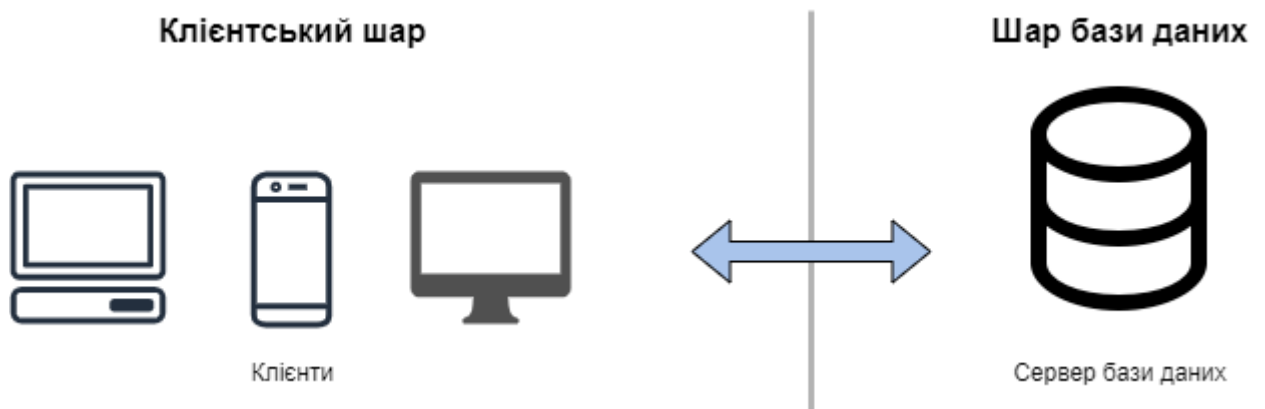


Рис. 4.3 – Схема 2-шарової клієнт-серверної архітектури

У такому типі архітектури, сервер та база даних знаходяться у одному шарі. Вибір архітектури 2-шарового типу, може підійти для систем, що будуть витримувати навантаження від невеликої кількості користувачів. Також, можна використовувати 2-шаровий тип, для створення прототипів, та якщо додаток буде працювати з невеликою кількістю даних. Рекомендується використовувати системи на такій архітектурі, при обмеженні, що кількість одночасних підключень користувачів буде меншою за 100.

#### 4.3.2 Опис 3-шарової клієнт-серверної архітектури

Для того, щоб покрити деякі обмеження 2-шарової архітектури, була розроблена 3-шарова архітектура, яка дозволяє розгортувати клієнтський додаток, серверний додаток та сервер бази даних на різних ізольованих платформах. Звісно, якщо необхідно, можна все розгорнути на одному комп'ютері, проте це

рекомендовано лише для тестування чи розробки. Схема 3-шарової архітектури зображена на рисунку 4.4.

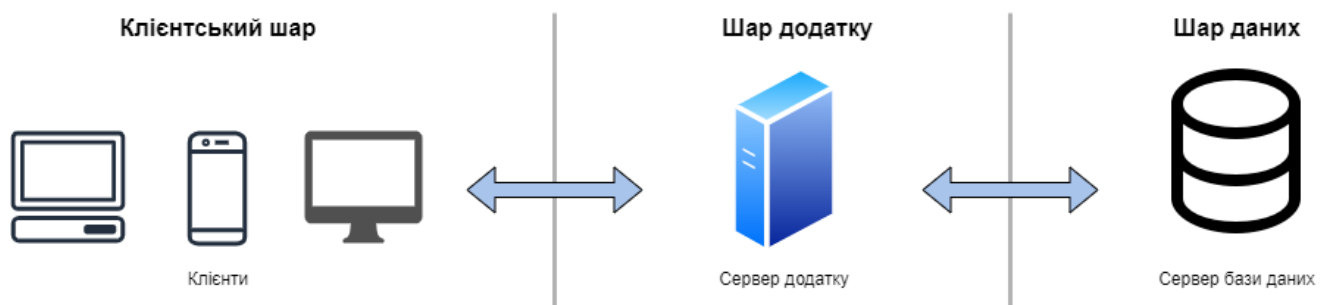


Рис. 4.4 – Схема 3-шарової архітектури

Кожен з рівнів такої архітектури виконує свої обов'язки. Клієнтський шар у свою чергу виконує запити до серверу додатку, який у свою чергу викликає потрібні сервіси і, якщо необхідно, звертається до серверу бази даних. Інформація з шару даних повертається до серверу додатку, оброблюється та повертається до клієнта. Для розробки додатків, на такому типі архітектури, необхідно витратити більше часу на розробку, а це означає і більше коштів. Проте, коли річ йде про великі об'єми даних та про безпеку цих даних, то 3-шарова архітектура дозволяє досягти бажаного результату.

#### 4.3.3 Опис N-шарової клієнт-серверної архітектури

Архітектура додатку вважається N-шаровою тоді, коли в ній присутні всі шари, що описані у 3-шаровій, та додаткові. Архітектури такого типу, використовуються для розробки складних високонавантажених систем. Звісно, такими системами легко керувати, вони гнучкі та гарно масштабуються. Проте, найголовнішим мінусом є ціна, кожен шар має розроблюватися різними командами, необхідні аналітики та менеджери. На рисунку 4.5 зображено схему N-шарової архітектури.

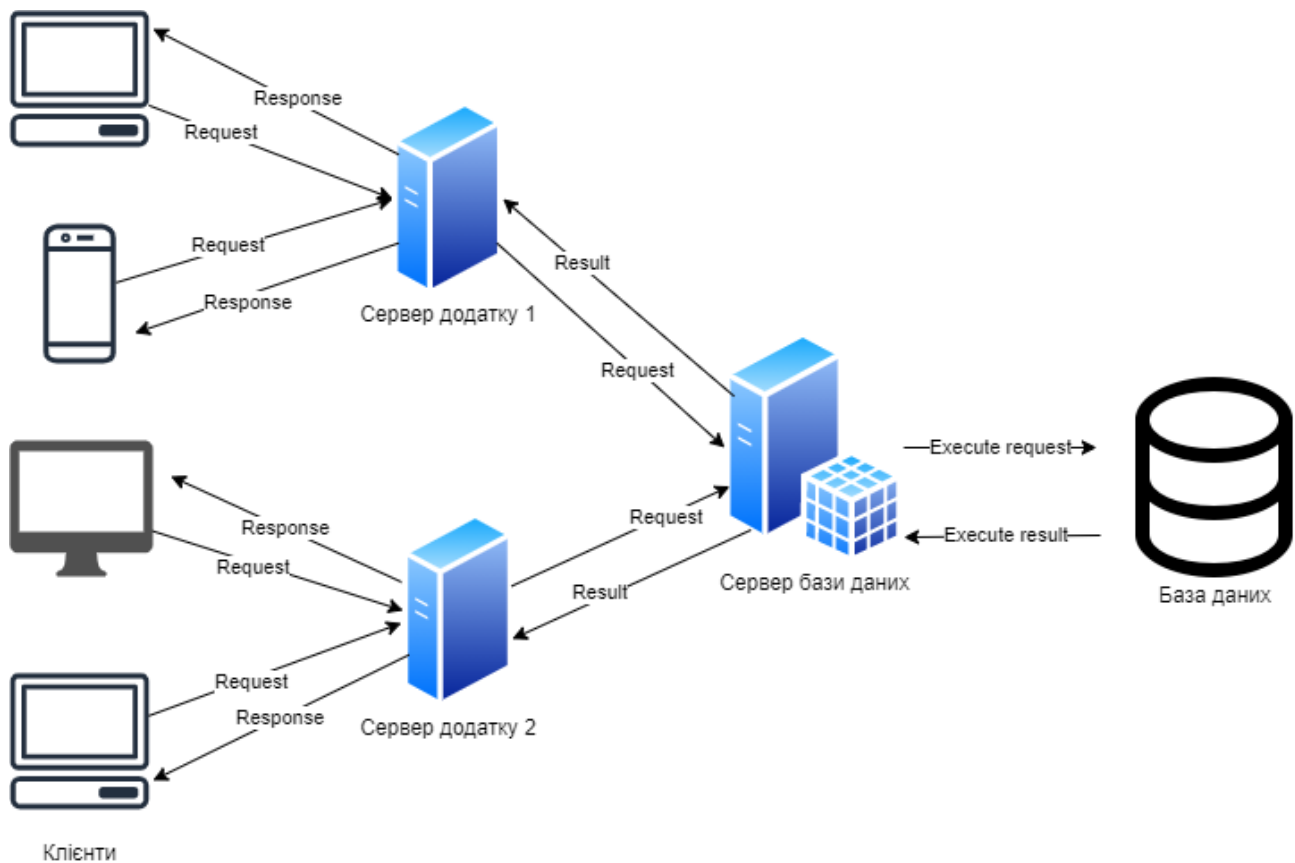


Рис. 4.5 – Схема N-шарової архітектури

При порівнянні архітектурних рішень для проектованої системи підтримки рішень, щодо видачі банківських кредитів, було обрано 3-шарову клієнт-серверну архітектуру. Зважаючи на ціну та час розробки, архітектура N-шарового типу – не підходить. Оцінюючи навантаження даних та силу виділених обчислювальних ресурсів, стало зрозуміло, що 2-шарова архітектура не зможе витримати навантаження.

В рамках атестаційної роботи, проектується прототип системи, отже була обрана 3-шарова архітектура. Для подальшого розвитку системи, можливе додавання додаткових шарів та переходу до N-шарового типу.

#### 4.4 Обґрунтування вибору СУБД

Для розробки системи, було обрано СУБД Oracle. Ця СУБД має велику кількість переваг, що і вплинуло на цей вибір, деякий перелік яких приведений нижче:

- Oracle підтримує найбільші бази даних;
- забезпечує підтримку великої кількості користувачів;
- підтримка великої кількості одночасних запитів;
- забезпечує якісну обробку транзакцій;
- велика швидкість транзакцій;
- велика швидкість роботи з даними;
- передбачена робота з критичними ситуаціями;
- відкат БД чи системний збій не приведе до припинення роботи БД;
- легко встановлювати на будь-яку ОС.

СУБД Oracle дає можливість скористатися якісним ПО для розробки системи. Також вважається, що дана СУБД гарно підходить для роботи Big Data, що у випадку з нашою системою є великим плюсом [34].

#### 4.5 Логічне та фізичне моделювання бази даних

Для моделювання схеми бази даних був використаний CASE-засіб «Allfusion ErWin Data Modeler». Даний засіб дає можливість змоделювати повноцінні логічну та фізичні моделі даних системи, розділити дані на сутності, визначити зв'язки та обрати структуру сховища даних.

Для початку було розроблено логічну модель, а потім на її основі фізичну. На рисунку 4.6 представлена логічна модель бази даних.

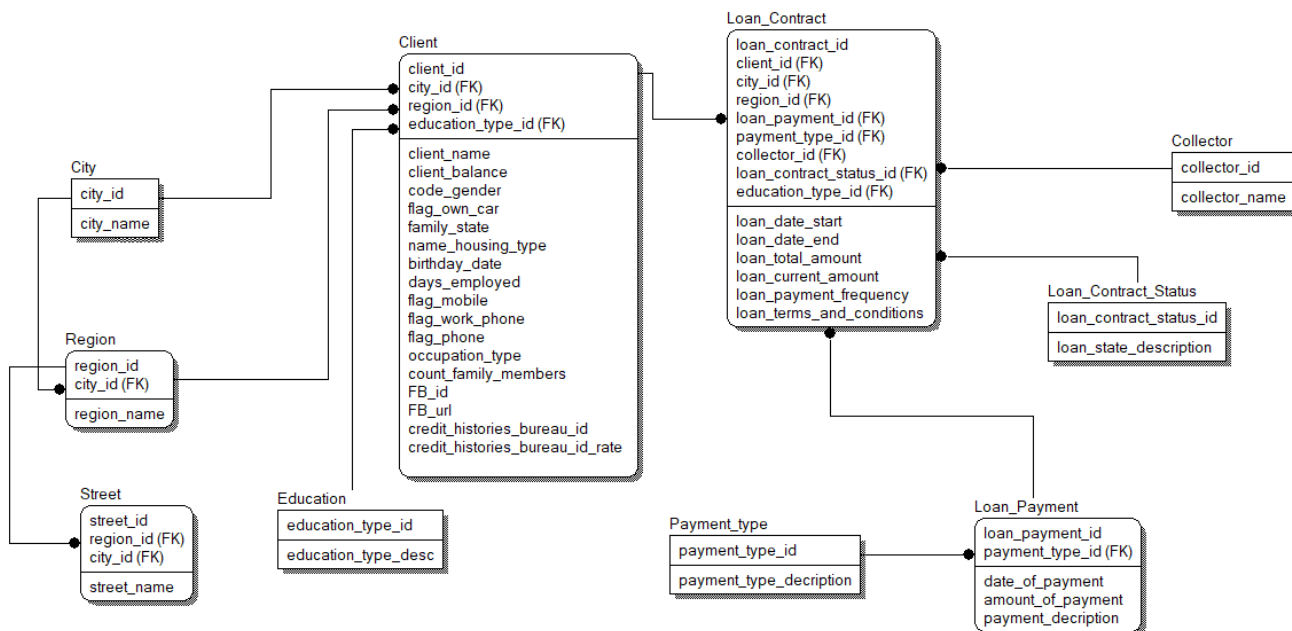


Рис. 4.6 – Логічне моделювання бази даних системи

Сутності, що увійшли до моделі та відповідають предметній області, описані у таблиці 4.1.

Таблиця 3.1 – Сутності БД та опис їх атрибутів

№	Назва сутності	Назва атрибуту	Опис	Тип
1	Client	client_id	Ідентифікатор клієнта	PK
		city_id	Ідентифікатор міста	FK
		region_id	Ідентифікатор району	FK
		education_type_id	Ідентифікатор типу освіти	FK
		client_name	Ім'я клієнта	
		client_balance	Баланс клієнта	
		code_gender	Стать клієнта	
		flag_own_car	Флаг наявності авто	
		family_state	Сімейний статус	
		name_housing_type	Тип проживання	
		birthday_date	Дата народження	
		days_employed	Днів праці	
		flag_mobile	Флаг наявності мобільного	
flag_work_phone	Флаг наявності			

			робочого телефону	
		flag_phone	Флаг наявності телефону	
		occupation_type	Тип професії	
		count_family_members	Кількість членів родини	
		fb_id	Facebook ідентифікатор	
		fb_url	Facebook посилання	
		credit_histories_bureau_id	Ідентифікатор користувача в Бюро Кредитних Історій	
		credit_histories_bureau_rate	Рейтинг користувача в Бюро кредитних історій	
2	Loan_Contract	loan_contract_id	Ідентифікатор контракту позики	PK
		client_id	Ідентифікатор клієнта	FK
		city_id	Ідентифікатор міста	FK
		region_id	Ідентифікатор району	FK
		loan_payment_id	Ідентифікатор оплати	FK
		payment_type_id	Ідентифікатор типу оплати	FK
		collector_id	Ідентифікатор колектору	FK
		loan_contract_status_id	Ідентифікатор статусу контракту	FK
		education_type_id	Ідентифікатор типу освіти	FK
		loan_date_start	Дата укладення контракту	
		loan_date_end	Дата закінчення дії контракту	
		loan_total_amount	Повна сума позики	
		loan_current_amount	Сума, що залишилось повернути	
		loan_payment_frequency	Термін	

			повторності платежів	
		loan_terms_and_conditions	Інформація та умови кредитування	
3	Loan_Contract_Status	loan_contract_status_id	Ідентифікатор статусу контракту	PK
		loan_state_description	Опис статусу	
4	Loan_Payment	loan_payment_id	Ідентифікатор платежу	PK
		payment_type_id	Ідентифікатор типу платежу	FK
		amount_of_payment	Сума платежу	
		payment_decription	Інформація про платіж	
		date_of_payment	Дата платежу	
5	Payment_type	payment_type_id	Ідентифікатор типу платежу	PK
		payment_type_decription	Опис типу платежу	
6	Collector	collector_id	Ідентифікатор колектору	PK
		collector_name	Ім'я колектору	
7	Education	education_type_id	Ідентифікатор типу освіти	
		education_type_desc	Опис типу освіти	
8	City	city_id	Ідентифікатор міста	PK
		city_name	Назва міста	
9	Region	region_id	Ідентифікатор району	PK
		city_id	Ідентифікатор міста	FK
		region_name	Назва району	
10	Street	street_id	Ідентифікатор вулиці	PK
		region_id	Ідентифікатор району	FK
		city_id	Ідентифікатор міста	FK
		street_name	Назва вулиці	

Також за допомогою CASE засобу була згенерована фізична модель бази даних, на основі логічної, що представлена на рисунку 4.7.



частині додатку, та оброблятися асинхронно. За допомогою фреймворку Angular 2+, час розробки клієнтської частини буде менший ніж, якби розробка велась на чистому JavaScript [35].

Для розробки серверної частини системи, було обрано мову Node.js, з огляду її переваг та недоліків. Node.js підтримує асинхронну обробку запитів, двигун V8 та використовує модель подій, яка підвищує швидкість обробки запитів у реальному часі. Однією з найбільших переваг цієї мови є масштабованість, що дозволяє розгортувати додаткові мікросервіси, для того щоб розвивати архітектуру і робити систему більш гнучкою. Також однією з переваг можна вважати, що мова JavaScript використовується для написання і клієнтської частини і серверної, що зменшує поріг входження у розробку системи.

Також потрібно взяти до уваги метод для аналізу клієнтських заявок на отримання банківського кредиту, який найкраще розроблювати на мові Python, та використовувати його найпопулярніші пакети такі як: pandas, numpy та інші. Також мають велике значення алгоритми класифікації, що були обрані для розробки методу, та які мають імплементацію у пакеті sklearn. Отже для розробки методу було обрано мову Python [36].

## 5 ТЕСТУВАННЯ ТА ОЦІНКА ОТРИМАНИХ РЕЗУЛЬТАТІВ МЕТОДУ

Основним завданням методу є вирішення задачі класифікації за допомогою аналізу даних. У контексті предметної області завдання методу – це класифікація заявок клієнтів банку на отримання банківських кредитів.

Аналізуючи предметну область, та аналізуючи методи інтелектуального аналізу даних, в рамках атестаційної роботи було прийнято рішення застосувати поєднати 2 методи інтелектуального аналізу, а саме CART та SVM для створення нового ансамблевого методу аналізу.

Можливості перевірити роботу методу на справжніх даних не було, тож набір даних був взятий з популярного сервісу Kaggle для Data Mining [37]. Набір даних призначений для аналізу та класифікації, щодо отримання банківських кредитів, який повністю підходить під предметну область у атестаційній роботі.

Для початку необхідно подати дані до методу, програмний код отримання вхідних даних знаходиться у лістингу 5.1, та зразок даних з набору на рисунку 5.1.

### Лістинг 5.1 – Програмний код отримання вхідних даних

```
import pandas as pd
application_record =
pd.read_csv('/kaggle/input/credit-card-approvalprediction/application_record.csv')
application_record.head()
```

	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH
0	5008804	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005
1	5008805	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005
2	5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special	Married	House / apartment	-21474
3	5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110
4	5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110

Рис. 5.1 – Приклад даних з набору

Далі проаналізуємо масштаби даних, рисунок 5.2 показує кількість записів та ознак в цілому.

(438557, 18)

Рис. 5.2 – Кількість записів та ознак у таблиці

Отже метод буде працювати з 483 557 записами, що є заявками на кредити, а також з 18 ознаками. Обробимо дані, та відсічемо не унікальні дані (рисунок 5.3).

438510

Рис. 5.3 – Кількість унікальних заявок

Також до тестового набору даних йде додаткова таблиця з результатами кредитування по заявці. Завантажимо ці дані та розглянемо їх (лістинг 5.2, рисунок 5.4).

Лістинг 5.2 – Програмний код отримання результуючих даних

```
credit_record =
pd.read_csv('/kaggle/input/credit-card-approval-prediction/credit_record.csv')
credit_record.head()
```

	ID	MONTHS_BALANCE	STATUS
0	5001711	0	X
1	5001711	-1	0
2	5001711	-2	0
3	5001711	-3	0
4	5001712	0	C

Рис. 5.4 – Приклад результуючих даних

Детально про сприйняття даних цієї таблиці проводився огляд у пункті 2.3 даної атестаційної роботи. Переглянемо загальну кількість заявок, та скільки унікальних клієнтів всього (рисунки 5.5-5.6).

(1048575, 3)

Рис. 5.5 – Кількість заявок клієнтів, що мають статуси

45985

Рис. 5.6 – Кількість унікальних клієнтів, у результуючій вибірці

Змінимо осі представлення для відображення всіх ознак заявки на кредит (рисунок 5.7).

	0	1	2	3	4
ID	5008804	5008805	5008806	5008808	5008809
CODE_GENDER	M	M	M	F	F
FLAG_OWN_CAR	Y	Y	Y	N	N
FLAG_OWN_REALTY	Y	Y	Y	Y	Y
CNT_CHILDREN	0	0	0	0	0
AMT_INCOME_TOTAL	427500	427500	112500	270000	270000
NAME_INCOME_TYPE	Working	Working	Working	Commercial associate	Commercial associate
NAME_EDUCATION_TYPE	Higher education	Higher education	Secondary / secondary special	Secondary / secondary special	Secondary / secondary special
NAME_FAMILY_STATUS	Civil marriage	Civil marriage	Married	Single / not married	Single / not married
NAME_HOUSING_TYPE	Rented apartment	Rented apartment	House / apartment	House / apartment	House / apartment
DAYS_BIRTH	-12005	-12005	-21474	-19110	-19110
DAYS_EMPLOYED	-4542	-4542	-1134	-3051	-3051
FLAG_MOBIL	1	1	1	1	1
FLAG_WORK_PHONE	1	1	0	0	0
FLAG_PHONE	0	0	0	1	1
FLAG_EMAIL	0	0	0	1	1
OCCUPATION_TYPE	NaN	NaN	Security staff	Sales staff	Sales staff
CNT_FAM_MEMBERS	2	2	2	1	1

Рис. 5.7 – Приклад даних з відображенням усіх ознак

Далі слідуючи алгоритму навчання роботи методу, розробленому у пункті 2.3 даної атестаційної роботи, необхідно перевірити дані на наявність пропущених значень (рисунок 5.7).



ID	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
FLAG_MOBIL	0
FLAG_WORK_PHONE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	134203
CNT_FAM_MEMBERS	0
dtype: int64	

Рис. 5.7 – Кількість пропущених даних кожної ознаки

Виявилось що тип професії не вказаний у 134203 заявках, якщо підрахувати відсоткову кількість від всіх заявок, це близько 30%. Проаналізуємо цю ознаку, щоб дізнатися її корисність та необхідність у даній вибірці (рисунок 5.8). Виявилось, що в ознаці використовуються дуже різні дані, та якщо заповнити їх середнім значенням то вибірка може втратити свою репрезентативність, тож щоб не зіпсувати кінцевий результат, краще видалити ознаку (лістинг 5.3).

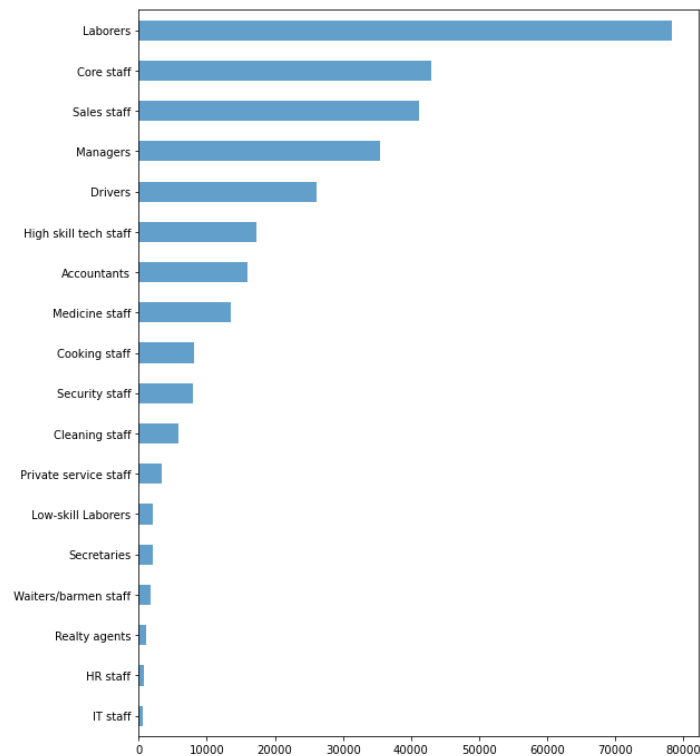


Рис. 5.8 – Діаграма співвідношення ознаки «тип професії» до її кількості

Лістинг 5.3 – Програмний код видалення стовпця з вибірки  
`application_record.drop('OCCUPATION_TYPE', axis=1, inplace=True)`

Таким чином, вибірка залишилась без пропущених значень, а отже можна переходити до нормалізації даних. Скористаймося рішенням прийнятим у пункті 2.3, та розіб'ємо результуючі заявки за 2 класами, де 0 – надійний клієнт, 1 – боржника боржник (лістинг 5.4, рисунок 5.9).

Лістинг 5.4 – Програмний код для розбиття заявок на класи  
`defaulter_codes= ['0', '1', '2', '3', '4', '5']`  
`credit_record['Label']= np.where(credit_record.STATUS.isin(defaulter_codes), 1, 0)`  
`credit_record.head(10)`

	ID	MONTHS_BALANCE	STATUS	Label
0	5001711	0	X	0
1	5001711	-1	0	1
2	5001711	-2	0	1
3	5001711	-3	0	1
4	5001712	0	C	0
5	5001712	-1	C	0
6	5001712	-2	C	0
7	5001712	-3	C	0
8	5001712	-4	C	0
9	5001712	-5	C	0

Рис. 5.9 – Результуючі дані з поділом на класи

Тепер ознака STATUS у результуючій вибірці нам не потрібна, її можна видалити, а також проаналізуємо дані (лістинг 5.5, рисунок 5.10).

Лістинг 5.5 – Програмний код видалення ознаки та виводу даних

```
credit_record.drop('STATUS', axis=1, inplace=True)
credit_record['Label'].value_counts()
```

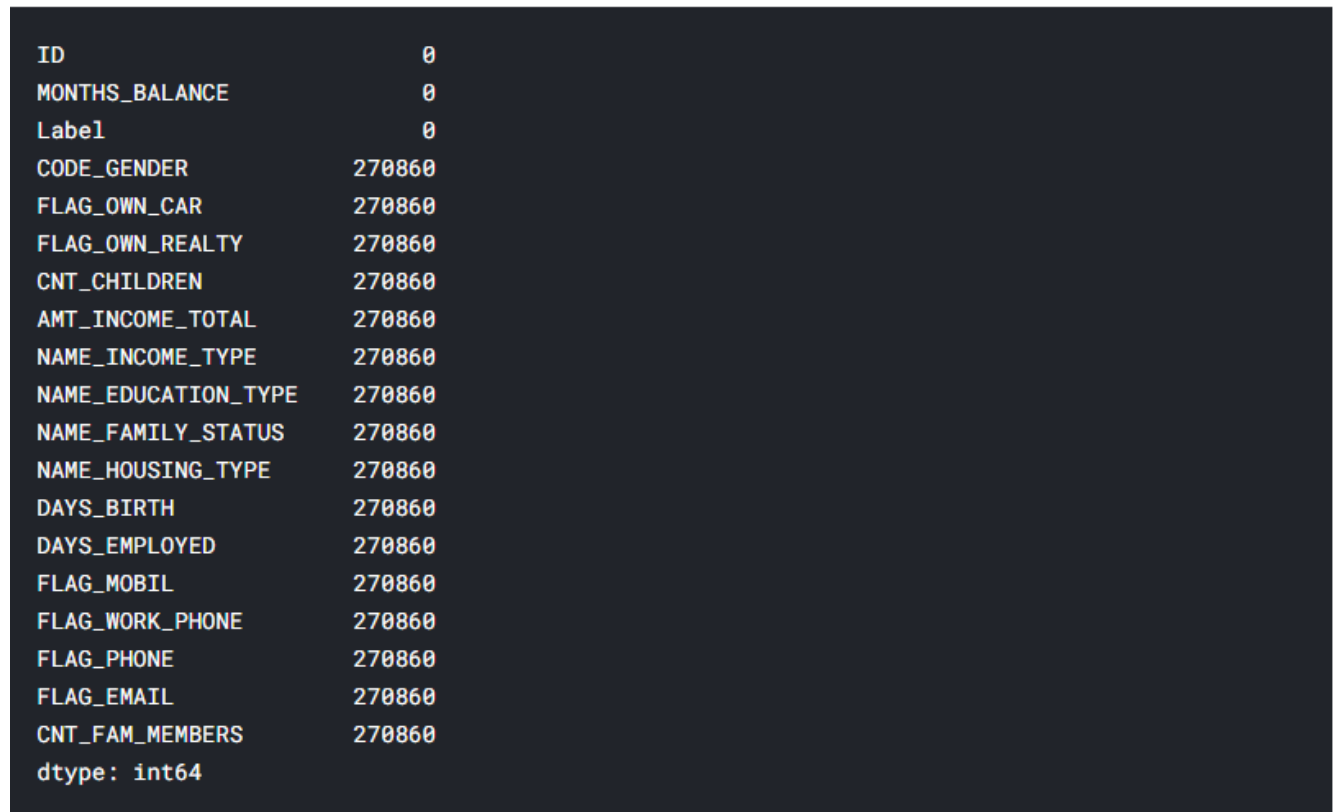
```
0    651261
1    397314
Name: Label, dtype: int64
```

Рис. 5.10 – Кількість заявок кожного класу

Після проведених перетворень видно, що успішних заявок більше ніж неуспішних, проте близько 38% неуспішних заявок – доволі багато. Отже провівши перетворення, відсіявши пропущені значення, виділивши класи, в рамках тестових даних тепер необхідно з'єднати дві таблиці, по ознаці ID, а також виведемо дані і перевіримо, чи є пропущені значення після злиття таблиць (лістинг 5.6, рисунок 5.11).

## Лістинг 5.6 – Програмний код з'єднання таблиць та виведення даних

```
record = pd.merge(credit_record, application_record, on='ID', how='left')
record.head()
record.isnull().sum()
```



```
ID          0
MONTHS_BALANCE  0
Label      0
CODE_GENDER  270860
FLAG_OWN_CAR  270860
FLAG_OWN_REALTY  270860
CNT_CHILDREN  270860
AMT_INCOME_TOTAL  270860
NAME_INCOME_TYPE  270860
NAME_EDUCATION_TYPE  270860
NAME_FAMILY_STATUS  270860
NAME_HOUSING_TYPE  270860
DAYS_BIRTH  270860
DAYS_EMPLOYED  270860
FLAG_MOBIL  270860
FLAG_WORK_PHONE  270860
FLAG_PHONE  270860
FLAG_EMAIL  270860
CNT_FAM_MEMBERS  270860
dtype: int64
```

Рис. 5.11 – Кількість пропущених даних після злиття таблиць для кожної ознаки

Виявилось, що дані з 2 таблиць співпадають не до кінця, і 270860 заявок на кредити залишились без статусу. Для навчання моделі, необхідно щоб всі заявки мали свої класи, тож видалимо рядки, що не мають класу і виведемо кількість заявок, що мають всі ознаки і приналежність до класу (лістинг 5.7, рисунок 5.12).

## Лістинг 5.7 – Програмний код видалення рядків з пустими значеннями і виведення даних

```
record.dropna(inplace=True)
record['Label'].value_counts()
```

```

0    475486
1    302229
Name: Label, dtype: int64

```

Рис. 5.12 – Кількість заявок кожного класу, з описом всіх ознак

Оглянувши дані, можна зробити висновок, що після видалення пустих рядків статистична ситуація не змінилась відношення і залишилось приблизно 62:38. Представимо це графіком (рисунок 5.12)

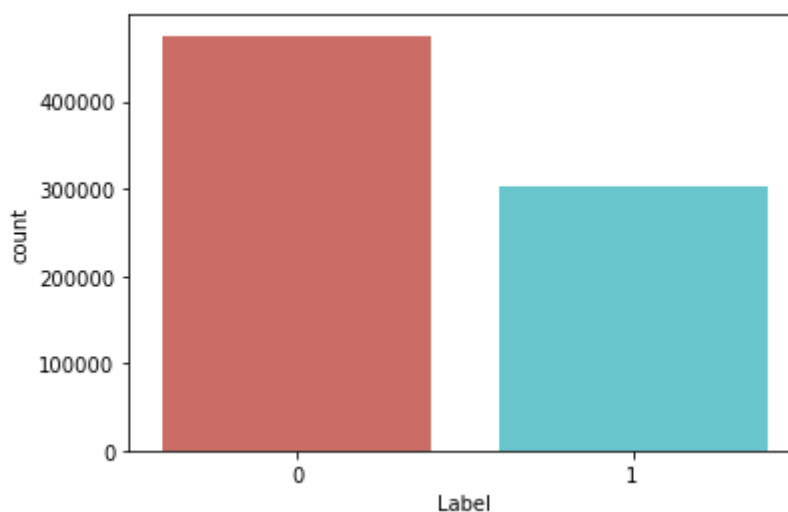


Рис. 5.13 – Співвідношення вдалих класів заявок до кількості

Далі, змінимо дані таким чином, щоб алгоритм використав їх правильно. Загалом, з усіх ознак виділимо нові, для того щоб значення ознак змінились з номінальних на порядкові або індикаторні (лістинг 5.8, рисунок 5.14).

Лістинг 5.9 – Програмний код спрощення номінальних даних

```

record = pd.get_dummies(record,
columns=['CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE',
'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE'], drop_first =
False)

```

	92938	92939	92940	92941	92942
ID	5008804_0	5008804_-1	5008804_-2	5008804_-3	5008804_-4
MONTHS_BALANCE	0	-1	-2	-3	-4
Label	0	0	0	0	0
CNT_CHILDREN	0	0	0	0	0
AMT_INCOME_TOTAL	427500	427500	427500	427500	427500
DAYS_BIRTH	-12005	-11975	-11945	-11915	-11885
DAYS_EMPLOYED	-4542	-4512	-4482	-4452	-4422
FLAG_MOBIL	1	1	1	1	1
FLAG_WORK_PHONE	1	1	1	1	1
FLAG_PHONE	0	0	0	0	0
FLAG_EMAIL	0	0	0	0	0
CNT_FAM_MEMBERS	2	2	2	2	2
INCOME_PER_MEMBER	213750	213750	213750	213750	213750
CODE_GENDER_F	0	0	0	0	0
CODE_GENDER_M	1	1	1	1	1

Рис. 5.14 – Спрощення номінальних даних

Після того як дані були оброблені, відсіяні пропущені значення, таблиці об'єднані, виділення класів було здійснено тепер є можливість проводити навчання моделей. Для початку проведемо навчання простої моделі дерева рішень CART (лістинг 5.10).

#### Лістинг 5.10 – Програмний код навчання моделі CART

```
from sklearn.tree import DecisionTreeClassifier

tree = DecisionTreeClassifier(random_state=0, max_depth=25)
tree.fit(features_train, label_train)

# predict train set
pred_train=tree.predict(features_train)
# predict test set
pred_test=tree.predict(features_test)

from sklearn.metrics import accuracy_score
accuracy_train = accuracy_score(pred_train,label_train)
accuracy_test = accuracy_score(pred_test,label_test)
```

Тепер подивимось результати класифікації, та проаналізуємо їх (рисунок 5.15).

```

Accuracy Score
0.6444333333333333

Precision Score
[0.68054713 0.5542604 ]

Confusion Matrix
          Non Defaulter  Defaulter
Non Defaulter          14578      3824
Defaulter              6843      4755

Classification Report
          precision    recall  f1-score   support

     0           0.68      0.79      0.73     18402
     1           0.55      0.41      0.47     11598

 accuracy                   0.64     30000
 macro avg                  0.62     30000
 weighted avg               0.63     30000

```

Рис 5.15 – Класифікаційний звіт роботи методу CART

Перевіряючи на тестових даних, виявилось що точність моделі дорівнює близько 64%. Користуючись алгоритмом, що описаний у пункті 2.3 розглянемо як модель зробила оцінку ознак (рисунок 5.16). Після цього, необхідно підібрати мінімальну оцінку ознак  $W_{min}$  для навчання наступної моделі на основі алгоритму SVM.

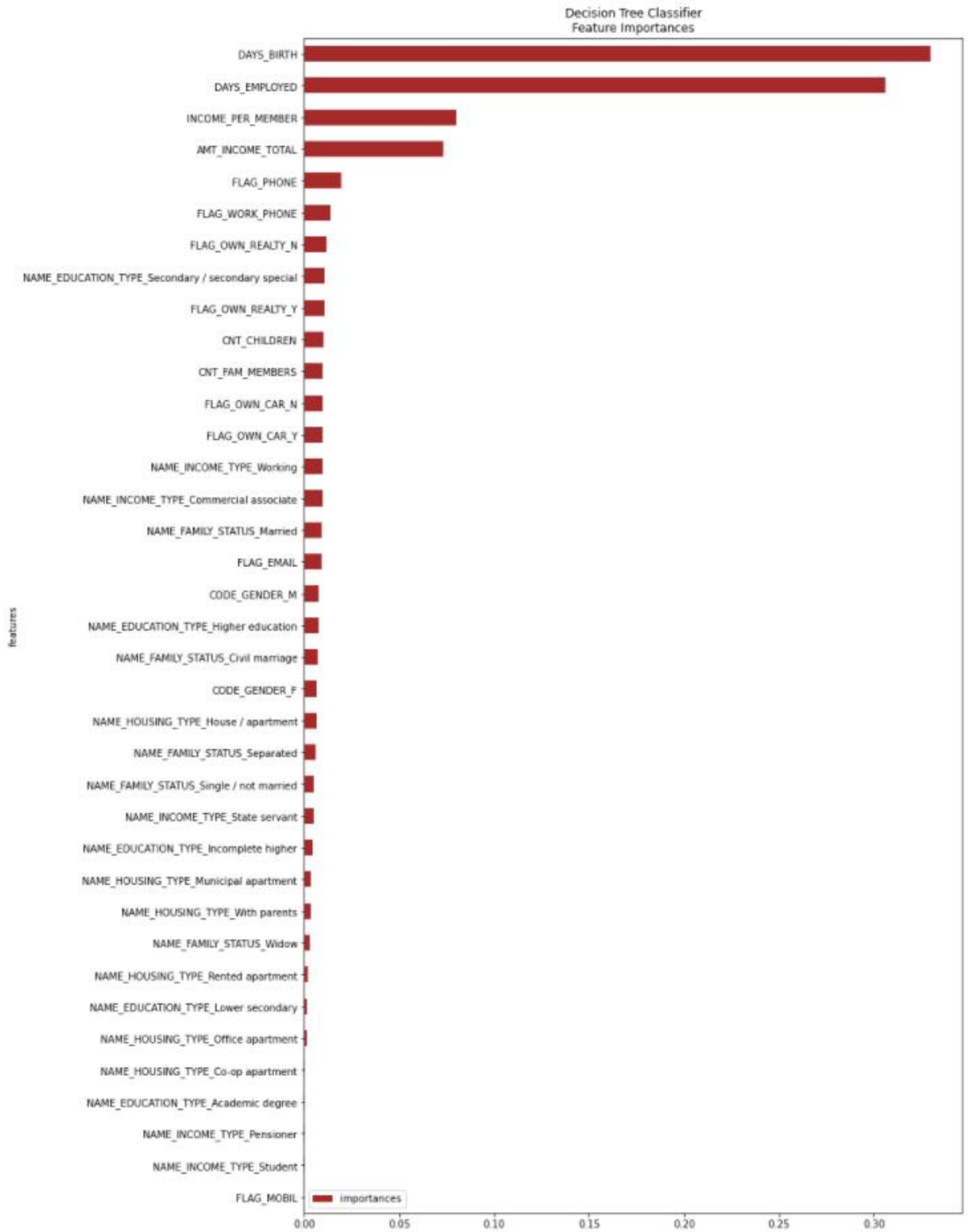


Рис. 5.16 – Оцінка ознак за допомогою моделі CART

Далі проаналізуємо цифри, для того щоб підібрати  $W_{min}$  (рисунок 5.17).

```
[('CNT_CHILDREN', 0.007967375218599484),
 ('AMT_INCOME_TOTAL', 0.00808580903971397),
 ('DAYS_BIRTH', 0.33266993169584147),
 ('DAYS_EMPLOYED', 0.2839949256538836),
 ('FLAG_MOBIL', 0.0),
 ('FLAG_WORK_PHONE', 0.012401439853414136),
 ('FLAG_PHONE', 0.018887320457038774),
 ('FLAG_EMAIL', 0.007433783150857225),
 ('CNT_FAM_MEMBERS', 0.009597284517581647),
 ('INCOME_PER_MEMBER', 0.08281453937002511),
 ('CODE_GENDER_F', 0.00893010076163038),
 ('CODE_GENDER_M', 0.008887964447149439),
 ('FLAG_OWN_CAR_N', 0.01069594932013490),
 ('FLAG_OWN_CAR_Y', 0.011358167930883174),
 ('FLAG_OWN_REALTY_N', 0.010152471696322666),
 ('FLAG_OWN_REALTY_Y', 0.010594249169580218),
 ('NAME_INCOME_TYPE_Commercial associate', 0.009638284055174088),
 ('NAME_INCOME_TYPE_Pensioner', 0.0001745574257301678),
 ('NAME_INCOME_TYPE_State servant', 0.006967483116164603),
 ('NAME_INCOME_TYPE_Student', 0.00846270782726417),
 ('NAME_INCOME_TYPE_Working', 0.009519713460739364),
 ('NAME_EDUCATION_TYPE_Academic degree', 0.008456593598610444),
 ('NAME_EDUCATION_TYPE_Higher education', 0.010082935472923764),
 ('NAME_EDUCATION_TYPE_Incomplete higher', 0.004452044529061358),
 ('NAME_EDUCATION_TYPE_Lower secondary', 0.0019456418856648673),
 ('NAME_EDUCATION_TYPE_Secondary / secondary special', 0.007914529277420555),
 ('NAME_FAMILY_STATUS_Civil marriage', 0.00806443551251236),
 ('NAME_FAMILY_STATUS_Married', 0.008801926873362901),
 ('NAME_FAMILY_STATUS_Separated', 0.0056401907561322755),
 ('NAME_FAMILY_STATUS_Single / not married', 0.006284107515060277),
 ('NAME_FAMILY_STATUS_Widow', 0.004224799328582286),
 ('NAME_HOUSING_TYPE_Co-op apartment', 0.000469568663112154),
 ('NAME_HOUSING_TYPE_House / apartment', 0.007390322142520077),
 ('NAME_HOUSING_TYPE_Municipal apartment', 0.004045750228821025),
 ('NAME_HOUSING_TYPE_Office apartment', 0.0007056790166570191),
 ('NAME_HOUSING_TYPE_Rented apartment', 0.0012813819357129243),
 ('NAME_HOUSING_TYPE_With parents', 0.004608721462116798)]
```

Рис. 5.17 – Важливість ознак

Отже можна побачити, що деякі ознаки мають більший вплив на те, чи буде виданий кредит чи ні, а деякі майже не грають ролі. Тож, відсіємо зовсім неважливі ознаки, задавши значення  $W_{min} = 0.00125$  (рисунок 5.18).

```
[('CNT_CHILDREN', 0.007967375218599484),
 ('AMT_INCOME_TOTAL', 0.00808580903971397),
 ('DAYS_BIRTH', 0.33266993169584147),
 ('DAYS_EMPLOYED', 0.2839949256538836),
 ('FLAG_MOBIL', 0.012401439853414136),
 ('FLAG_WORK_PHONE', 0.018887320457038774),
 ('FLAG_PHONE', 0.007433783150857225),
 ('FLAG_EMAIL', 0.009597284517581647),
 ('CNT_FAM_MEMBERS', 0.08281453937002511),
 ('INCOME_PER_MEMBER', 0.00893010076163038),
 ('CODE_GENDER_F', 0.008887964447149439),
 ('CODE_GENDER_M', 0.01069594932013490),
 ('FLAG_OWN_CAR_N', 0.011358167930883174),
 ('FLAG_OWN_CAR_Y', 0.010152471696322666),
 ('FLAG_OWN_REALTY_N', 0.010594249169580218),
 ('FLAG_OWN_REALTY_Y', 0.009638284055174088),
 ('NAME_INCOME_TYPE_Commercial associate', 0.006967483116164603),
 ('NAME_INCOME_TYPE_Pensioner', 0.009519713460739364),
 ('NAME_INCOME_TYPE_State servant', 0.010082935472923764),
 ('NAME_INCOME_TYPE_Student', 0.004452044529061358),
 ('NAME_INCOME_TYPE_Working', 0.0019456418856648673),
 ('NAME_EDUCATION_TYPE_Academic degree', 0.007914529277420555),
 ('NAME_EDUCATION_TYPE_Higher education', 0.00806443551251236),
 ('NAME_EDUCATION_TYPE_Incomplete higher', 0.008801926873362901),
 ('NAME_EDUCATION_TYPE_Lower secondary', 0.0056401907561322755),
 ('NAME_EDUCATION_TYPE_Secondary / secondary special', 0.006284107515060277),
 ('NAME_FAMILY_STATUS_Civil marriage', 0.004224799328582286),
 ('NAME_FAMILY_STATUS_Married', 0.00790322142520077),
 ('NAME_FAMILY_STATUS_Separated', 0.004045750228821025),
 ('NAME_FAMILY_STATUS_Single / not married', 0.0012813819357129243),
 ('NAME_FAMILY_STATUS_Widow', 0.004608721462116798)]
```

Рис. 5.18 – Ознаки, що мають  $W > W_{min}$

Далі відбувається фільтрація стовпців вибірки, та дані подаються для навчання моделі SVM (лістинг 5.11).

## Лістинг 5.11 – Подання даних для навчання алгоритму SVM

```

from sklearn import svm
classification_svm = svm.SVC(kernel='linear')
classification_svm.fit(features_train, label_train)
pred_train = classification_svm.predict(features_train)
pred_test=classification_svm.predict(features_test)
from sklearn.metrics import accuracy_score
accuracy_train = accuracy_score(pred_train,label_train)
accuracy_test = accuracy_score(pred_test,label_test)

```

Тепер проаналізуємо результат роботи комбінованого методу, за допомогою класифікаційного звіту (рисунок 5.19).

```

Accuracy Score
0.7208333333333333

Precision Score
[0.73903405 0.71823157]

Confusion Matrix

```

	Non Defaulter	Defaulter
Non Defaulter	14001	4331
Defaulter	4944	6724

```

Classification Report

```

	precision	recall	f1-score	support
0	0.74	0.76	0.75	18332
1	0.69	0.65	0.59	11668
accuracy			0.69	30000
macro avg	0.67	0.67	0.67	30000
weighted avg	0.69	0.69	0.69	30000

Рис. 5.19 – Класифікаційний звіт для комбінованого методу

Отже, класифікаційний звіт показує оцінку точності – 72%. В порівнянні з використанням чистого методу CART різниця майже у 8%. Чи можна сказати, що даний метод на будь-яких даних буде показувати кращі результати? Ні, не можна, дуже часто результати залежать від самих даних, та методик їх обробки, нормалізації. Щодо предметної області, та проектованої системи 72% – і досить непоганий результат, проте недосконалий. Для покращення методу, можливо, потрібно використати інші методи в ансамблі, спробувати техніки Stacking, Boosting.

Загалом, таку модель можна використовувати у реальній системі, так як комбінований метод буде виконувати роль лише підтримки рішення співробітника банку, а видавати, чи не видавати кредит буде його рішенням.

Для того, щоб отримати передбачення щодо клієнтської заявки на отримання банківського кредиту необхідно подати дані про клієнта на вхід до навченої моделі (лістинг 5.12, рисунок 5.20)

Лістинг 5.12 – Програмний код подання даних про заявку до навченої моделі

```
predict = classification_svm.predict(clientData)
print("Predicted=%s" % ("defaulter" if predict == 0 else "non_defaulter"));
```

```
Predicted=defaulter
```

Рис. 5.20 – Приклад прогнозування методу

Отже, подавши дані на вхід до методу, був отриманий прогноз щодо заявки, та клас заявки – боржник.

## ВИСНОВКИ

В атестаційній роботі проведено дослідження систем методів інтелектуального аналізу даних, а також спроектовано систему підтримки прийняття рішень на основі комбінування обраних методів.

Під час аналізу предметної області та проблеми оцінки кредитоспроможності клієнтів банку були розглянуті основні існуючі рішення розрахунку кредитного ризику, досліджені методи інтелектуального аналізу даних, а також системи підтримки прийняття рішень. В результаті аналізу предметної області були виділені основні задачі, що були розглянуті та виконані в рамках атестаційної магістерської роботи.

В атестаційній роботі вирішувалась проблема підвищення ефективності вирішення задачі класифікації заявок на отримання банківських кредитів за рахунок СППР на основі комбінованого методу інтелектуального аналізу даних.

Було проведено проектування СППР, а саме:

- визначено функціональні вимоги до системи;
- розроблена діаграма прецедентів;
- обрано архітектуру системи;
- обрано СУБД;
- проведено логічне та фізичне моделювання бази даних;
- обрано мову програмування для кожного шару архітектури.

У роботі детально описаний метод комбінування існуючих алгоритмів таких як CART та SVM. В ході тестування описаного методу, використовувалися дані з публічного доступу, схожі до реальних. В результаті тестування, було проведено аналіз результатів роботи методу за допомогою класифікаційного звіту, наданого пакетами мови Python. Він дозволив дізнатися точність навченої моделі, а також виконати порівняльний аналіз моделі, навченої за допомогою алгоритму CART та моделі на основі комбінування. Виявилося, що розроблений метод в атестаційній роботі показав кращі результати, відносно моделі на основі алгоритму CART та його можна розгортувати до спроектованої в рамках роботи системи СППР.

## ПЕРЕЛІК ПОСИЛАНЬ

1. Косова Т. Д. Аналіз банківської діяльності: навч. посіб. / Т. Д. Косова. – К.: Центр учбової літератури, 2010. – 486 с.
2. Лагутін В. Д. Кредитування: теорія і практика: навч. посібн. / В. Д. Лагутін. – К.: ТОВ "Знання", КОО, 2010. – 215 с.
3. Г. Н. Белоглазова Банковское дело. Организация деятельности коммерческого банка. / Белоглазова Г. Н., Кроливецкая Л. П. – М.: Высшее образование, 2011. – 422 с.
4. Пшик М.І. Етапи становлення і розвитку фінансово-кредитних відносин в Україні // Проблеми і перспективи розвитку банківської системи України: Зб. тез доп. ІХ Всеукр. наук.-практ. конф. – Суми: УАБС НБУ, 2011. – 188 с.
5. Мельник П. В. Банківські системи зарубіжних країн – банківські системи зарубіжних країн : підручник / П. В Мельник. – К. : Вид-во ЦУЛ ; Алерта. – 2010. – 574 с.
6. Холодна Ю. Є. Банківська система : навчальний посібник / Ю. Є. Холодна, О. М. Рац. – Х. : Вид. ХНЕУ, 2013. – 316 с. (Укр. мов.)
7. Варцаба В.І. Сучасне банківництво: теорія і практика: Навч. посібник. / В.І. Варцаба, О.І. Заславська— Ужгород: Видавництво УжНУ «Говерла», 2018. — 364 с.
8. Копилюк О.І. Банківські операції : навч. посібник / О.І. Копилюк, Г.В. Боднарчук Г. – К.: Знання, 2010. – 447 с.
9. Маркетинг у банку : навчальний посібник / [Т. А. Васильєва, С. М. Козьменко, І. О. Школьник та ін.] ; за заг. ред. Т. А. Васильєвої ; Державний вищий навчальний заклад “Українська академія банківської справи Національного банку України”. – Суми : ДВНЗ “УАБС НБУ”, 2014. – 353 с.
10. Економічний ризик: методи оцінки та управління / За заг. ред..Васильєвої Т.А., Кривич Я.М. Суми :ДВНЗ «УАБС НБУ», 2015. – 208 с.
11. Винник Т. Переваги та вигоди мобільноґо банкінґу / Т. Винник, О. Пастущин // Матеґ ріали міжнародноґо науковоґпрактичноґо конфеґ ренціґ

студентів і молодих учених "Соціально-економічні аспекти розвитку економіки", 27—28 квітня 2017 року. — Т.: ТНТУ. — 2017. — С. 46—47.

12. Азаренкова Г. М. Аналіз моделювання і управління ризиком (у схемах та прикладах) : навч. посібн. / Г. М. Азаренкова. — Львів : Новий Світ — 2010. — 240 с.

13. Кожухівська О. А. Прогнозування ризиків кредитування фізичних осіб за математичними моделями [Електронний ресурс] / О. А. Кожухівська // Вісник Національного університету «Львівська політехніка». Інформаційні системи та мережі. — 2013. — № 770. — С. 177-185. — Режим доступу: [http://nbuv.gov.ua/UJRN/VNULPICM\\_2013\\_770\\_23](http://nbuv.gov.ua/UJRN/VNULPICM_2013_770_23).

14. Ильясов С. М. Об оценке кредитоспособности банковского заемщика / С. М. Ильясов // Деньги и кредит. — 2015. — № 9. — С. 28—34.

15. Кузнєцова Н. В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / Н. В. Кузнєцова, П. І. Бідюк // Наукові вісті НТУУ «КПІ». — 2010. — №1. — С. 42-53.

16. Васильчак С. В. Оцінка кредитоспроможності позичальника як один із методів економічної безпеки / С. В. Васильчак, Л. Р. Демус // Науковий вісник НЛТУ України. - 2012. - Вип. 22.1. - С. 154-161.

17. Бондаренко Ю.В. Порівняльний аналіз підходів щодо оцінки кредитоспроможності позичальника – юридичної особи / Ю.В. Бондаренко // Молодіжний науковий вісник УАБС НБУ. Серія: Економічні науки. — 2012. - № 1. — С. 18-25.

18. Бучко І. Є. Скоринг як метод зниження кредитного ризику банку [Електронний ресурс] / І. Є. Бучко // Вісник Університету банківської справи Національного банку України. — 2013. — № 2. — Режим доступу: [http://nbuv.gov.ua/UJRN/VUbsNbU\\_2013\\_2\\_37](http://nbuv.gov.ua/UJRN/VUbsNbU_2013_2_37).

19. Самойлова С. С. Скоринговые модели оценки кредитного риска [Электронный ресурс] / С. С. Самойлова, М. А. Курочка // Социальноэкономические явления и процессы. — 2014. — № 3 (61). — Режим доступа: <http://cyberleninka.ru/article/n/skoringovye-modeli-otsenki-kreditnogo-riska>.

20. Виговський В.Г. Методичні підходи до оцінки кредитоспроможності суб'єктів господарювання комерційними банками [Текст] / В.Г. Виговський // Облік і фінанси АПК. – 2012. – №3. – С.125-128.

21. Кожухівська О. А. Прогнозування ризиків кредитування фізичних осіб за математичними моделями [Електронний ресурс] / О. А. Кожухівська // Вісник Національного університету «Львівська політехніка». Інформаційні системи та мережі. – 2013. – № 770. – С. 177-185. – Режим доступу: [http://nbuv.gov.ua/UJRN/VNULPICM\\_2013\\_770\\_23](http://nbuv.gov.ua/UJRN/VNULPICM_2013_770_23).

22. Sudhakar M and Dr. C. V. Krishna Reddy, "CREDIT EVALUATION MODEL OF LOAN PROPOSALS FOR BANKS USING DATA MINING," International Journal of Latest Research in Science and Technology, pp. 126-131, July 2014.

23. Zurada, Jozef, and Martin Zurada. "How Secure Are “Good Loans”: Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans." Review of Business Information Systems (RBIS) 6.3 (2011): 65-84.

24. Christopher Pal, Mark Hall, Eibe Frank, Ian Witten. Data Mining: Practical Machine Learning Tools and Techniques, 4rd ed. / Morgan Kaufmann, 2016.

25. Abhijit A. Sawant and P. M. Chawan, "Comparison of Data Mining Techniques used for Financial Data Analysis," International Journal of Emerging Technology and Advanced Engineering, June 2013

26. Grebennik I., Khriapkin O., Ovezgeldyyev A., Pisklakova V., Urniaieva I. (2019) The Concept of a Regional Information-Analytical System for Emergency Situations. In: Murayama Y., Velev D., Zlateva P. (eds) Information Technology in Disaster Risk Reduction. ITDRR 2017. IFIP Advances in Information and Communication Technology, vol 516. Springer, Cham Scopus

27. Grebennik I., Reshetnik V., Ovezgeldyyev A., Ivanov V., Urniaieva I. (2019) Strategy of Effective Decision-Making in Planning and Elimination of Consequences of Emergency Situations In: Murayama Y., Velev D., Zlateva P. (eds) Information Technology in Disaster Risk Reduction. ITDRR 2018. IFIP Advances in Information and Communication Technology. Springer, Cham Scopus

28. Моделі й методи прийняття рішень: навч. посіб. / С.А. Ус, Л.С. Коряшкіна; М-во освіти і науки України, Нац. гірн. ун-т. – Д. : НГУ, 2014. – 300 с

29. Тези Міжнародної науково-технічної конференції «Інформаційно-комп'ютерні технології – 2018» (20–21 квітня 2018 р.). – Житомир : Вид. О. О. Євенок, 2018. – 296 с.

30. Вьюгин В.В. «Математические основы теории машинного обучения и прогнозирования» : учебник / В.В.. Вьюгин. – М. : Знания, 2013. - 387 с.

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

32. Бевз О. М. Проектування програмних засобів систем управління : навч. посіб. Ч. 1. Основи об'єктно-орієнтованого проектування / О. М. Бевз, В. М. Папінов, Ю. А. Скидан; Вінниц. нац. техн. ун-т. - Вінниця, 2010. - 124 с.

33. Мизніков, Р., Гребеннік, І. 2020. Дослідження та розробка методів аналізу даних для класифікації заявок на отримання банківських кредитів. Матеріали конференцій МЦНД, 2 жовтня 2020, Харків, Україна, МЦНД Науковий простір: актуальні питання, досягнення та інновації, Том 1, с 65-66.

34. Іванець О. Б. Побудова моделей прогнозування реляційних даних / О. Б. Іванець, І. В. Морозова, В. В. Нечипорук // Електроніка та системи упр. - 2011. - № 3. - С. 124-127

35. Пономаренко, Николай Николаевич. ВЕБ-программирование: HTML, CSS, JavaScript, jQuery, PHP, MySQL [Текст] : учеб. пособие / Н. Н. Пономаренко ; Нац. аэрокосм. ун-т им. Н. Е. Жуковского "Харьк. авиац. ин-т". - Харьков : ХАИ, 2014. - 143 с

36. Програмування числових методів мовою Python [Текст] : навч. посіб. для студентів ВНЗ / [А. Ю. Дорошенко та ін.] ; за ред. чл.-кор. НАН України А. В. Анісімова ; Київ. нац. ун-т ім. Тараса Шевченка. - Київ : Київський університет, 2013. - 463 с.

37. Kaggle [Электронный ресурс]. – Режим доступа: [www/ URL: https://www.kaggle.com/](http://www.kaggle.com/).