

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

Факультет комп'ютерних наук і технологій

(повне найменування факультету)

Кафедра програмних засобів

(повне найменування кафедри)

Пояснювальна записка

до дипломного проекту (роботи)

магістр

(ступінь вищої освіти)

на тему ДОСЛІДЖЕННЯ ТА РОЗРОБКА СИСТЕМИ ПРОГНОЗУВАННЯ
ЗВІЛЬНЕННЯ ПРАЦІВНИКІВ НА ОСНОВІ МОДЕЛЕЙ ШІ
RESEARCH AND DEVELOPMENT OF AN EMPLOYEE
TURNOVER PREDICTION SYSTEM BASED ON AI MODELS

Виконав: студент(ка) 2 курсу, групи КНТ-213м

Спеціальності 122 Комп'ютерні науки

(код і найменування спеціальності)

Освітня програма (спеціалізація)

Системи штучного інтелекту

ДОМАШЕНКО С.В.

(ПРИЗВИЩЕ та ініціали)

Керівник ГЛАДКОВА О.М.

(ПРИЗВИЩЕ та ініціали)

Рецензент ГОЛУБ Т.В.

(ПРИЗВИЩЕ та ініціали)

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

Факультет КНТ
Кафедра програмних засобів
Ступінь вищої освіти магістр
Спеціальність 122 Комп'ютерні науки
(код і найменування)
Освітня програма (спеціалізація) Системи штучного інтелекту
(назва освітньої програми (спеціалізації))

ЗАТВЕРДЖУЮ

Завідувач кафедри ПЗ, д.т.н, проф.
Сергій СУББОТІН
“ ” 2024 року

З А В Д А Н Н Я

НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТА(КИ)

ДОМАШЕНКО Світлани Володимирівни

(ПРИЗВИЩЕ, ім'я, по батькові)

1. Тема проєкту (роботи) Дослідження та розробка системи прогнозування звільнення працівників на основі моделей III. Research and Development of an Employee Turnover Prediction System Based on AI Models

керівник проєкту (роботи) к.т.н., ГЛАДКОВА Ольга Миколаївна,
(науковий ступінь, вчене звання, ПРИЗВИЩЕ, ім'я, по батькові)

затверджені наказом закладу вищої освіти від “ 07 ” жовтня 2024 року № 398

2. Строк подання студентом проєкту (роботи) 5 грудня 2024 року

3. Вихідні дані до проєкту (роботи) рекомендована література, технічне завдання

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) 1. Аналіз предметної області. 2. Комплексний аналіз набору даних. 3. Вибір моделі класифікації та методів попередньої обробки даних. 4. Розробка та тестування системи прогнозування. 5. Шляхи подальшого розвитку системи прогнозування плинності кадрів.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) Слайди презентації

6. Консультанти розділів проєкту (роботи)

Розділ	ПРИЗВИЩЕ, ініціали та посада консультанта	Підпис, дата	
		завдання видав	прийняв виконання завдання
1–5 Основна частина	ГЛАДКОВА О.М., доцент		
Нормоконтролер	БЄЛОВА А.В., асистент		

7. Дата видачі завдання “ 4 ” вересня 2024 року.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проєкту (роботи)	Строк виконання етапів проєкту (роботи)	Примітка
1	Постановка завдання роботи.	1 тиждень	Завдання, ТЗ
2	Аналіз предметної області.	2–3 тижні	Розділ 1
3	Розробка та удосконалення методів, моделей й алгоритмів вирішення задачі.	4–5 тижні	Розділ 2, 3
4	Розробка архітектури програми.	6 тиждень	Розділ 4
5	Розробка програми.	7–8 тижні	Розділ 4
6	Тестування програмного забезпечення.	9 тиждень	Розділ 5
7	Оформлення пояснювальної записки та документів до неї.	10–11 тижні	Додатки
8	Нормоконтроль та рецензування.	12 тиждень	
9	Захист роботи.	12 тиждень	

Студент(ка)

Світлана ДОМАШЕНКО
(Ім'я ПРІЗВИЩЕ)

Керівник проєкту (роботи)

Ольга ГЛАДКОВА
(Ім'я ПРИЗВИЩЕ))

РЕФЕРАТ

Пояснювальна записка до дипломної кваліфікаційної роботи магістра:
119 с., 4 табл., 40 рис., 2 дод., 24 джерела.

ПЛИННІСТЬ КАДРІВ, ЗВІЛЬНЕННЯ, ШТУЧНИЙ ІНТЕЛЕКТ,
МАШИННЕ НАВЧАННЯ, PYTHON, СИСТЕМА ПРОГНОЗУВАННЯ.

Об'єкт дослідження – процес розробки системи прогнозування звільнення працівників з використанням методів машинного навчання.

Предмет дослідження – методи та моделі машинного навчання для прогнозування звільнення працівників, а також підходи до розробки системи прогнозування на основі цих методів.

Мета роботи – розробка системи прогнозування звільнення працівників на основі методів машинного навчання для підвищення ефективності управління персоналом та зниження витрат, пов'язаних з непередбачуваною плинністю кадрів.

Матеріали, методи та технічні засоби: методи статистичного аналізу, машинного навчання та програмної інженерії; мова програмування Python; хмарне середовище Google Colab.

Результати. Розроблено систему прогнозування звільнення працівників з використанням методів машинного навчання.

Висновки. Розроблено систему прогнозування звільнення працівників з модульною архітектурою для своєчасного виявлення працівників з високим ризиком звільнення для прийняття обґрунтованих рішень в HR-менеджменті.

Галузь використання - HR-департаменти підприємств різних галузей.

ABSTRACT

Explanatory note to the diploma qualifying work of the master: 119 pages, 4 tables, 40 figures, 2 appendixes, 24 sources.

EMPLOYEE TURNOVER, RESIGNATION, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, PYTHON, PREDICTION SYSTEM.

Object of research is the process of developing an employee resignation prediction system using machine learning methods.

Subject of research is machine learning methods and models for predicting employee resignations, as well as approaches to developing a prediction system based on these methods.

The aim of the work is to develop an employee resignation prediction system based on machine learning methods to improve personnel management efficiency and reduce costs associated with unexpected employee turnover.

Materials, methods and technical tools: methods of statistical analysis, machine learning and software engineering; Python programming language; Google Colab cloud environment.

Results. An employee resignation prediction system has been developed using machine learning methods.

Conclusions. A modular architecture employee resignation prediction system has been developed for timely identification of employees with high resignation risk to make informed decisions in HR-management.

Field of application is HR-departments of enterprises across various industries.

ЗМІСТ

	C.
Перелік скорочень та умовних позначок.....	8
Вступ.....	9
1 Аналіз предметної області	11
1.1 Аналіз проблеми плинності кадрів та методів її прогнозування	11
1.2 Огляд сучасних AI-рішень для управління персоналом	15
1.3 Особливості використання методів машинного навчання для задач прогнозування плинності кадрів	18
1.4 Висновки за розділом 1	25
2 Комплексний аналіз набору даних	27
2.1 Аналіз структури та характеристик набору даних	27
2.2 Дослідження впливу факторів на ймовірність звільнення.....	40
2.3 Аналіз взаємозв'язків між змінними	50
2.4 Висновки за розділом 2.....	55
3 Вибір моделі класифікації та методів попередньої обробки даних	57
3.1 Обґрунтування вибору моделі машинного навчання.....	57
3.2 Підвищення ефективності моделей класифікації.....	62
3.3 Аналіз використання методів ансамблевого навчання	68
3.4 Обґрунтування вибору методу відбору ознак	72
3.5 Висновки за розділом 3	77
4 Розробка та тестування системи прогнозування	78
4.1 Архітектура та реалізація компонентів системи прогнозування плинності кадрів	78
4.2 Тестування роботи розробленої системи прогнозування	82
4.3 Висновки за розділом 4.....	90
5 Визначення шляхів подальшого розвитку розробленої системи прогнозування	91
5.1 Визначення переваг, недоліків та основних напрямків вдосконалення розробленої системи	91

	7
5.2 Подальший розвиток розробленої системи прогнозування.....	94
5.3 Висновки за розділом 5.....	101
Висновки.....	102
Перелік джерел посилання	103
Додаток А Фрагмент тексту програми.....	106
Додаток Б Слайди презентації.....	110

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАК

AI – Artificial Intelligence (штучний інтелект);

CRM – Customer Relationship Management (система управління взаємовідносинами з клієнтами);

ERP – Enterprise Resource Planning (система планування ресурсів підприємства);

FN – False Negative (хибно негативний);

FP – False Positive (хибно позитивний);

HR – Human Resources (управління персоналом);

IT – Information Technology (інформаційні технології);

KPI – Key Performance Indicators (ключові показники ефективності);

NLP – Natural Language Processing (обробка природної мови);

PCA – Principal Component Analysis (метод головних компонент);

RFECV – Recursive Feature Elimination with Cross-Validation (рекурсивне усунення ознак з перехресною перевіркою);

T-SNE – t-Distributed Stochastic Neighbor Embedding (t-розподілене стохастичне вкладення сусідів);

TN – True Negative (істинно негативний);

TP – True Positive (істинно позитивний).

ВСТУП

В умовах сучасного ринку праці проблема плинності кадрів набуває все більшої актуальності. Останні роки характеризуються значними змінами на ринку праці, викликаними як глобальною пандемією COVID-19, так і загальними трансформаціями у ставленні працівників до роботи. В умовах війни та повоєнного відновлення України актуальність цієї проблеми ще більше зростає. Масове переміщення населення, релокація бізнесу та знищення підприємств суттєво ускладнили стабільність кадрового потенціалу країни. Ці фактори ускладнюють повернення до нормальної роботи та збільшують навантаження на працівників, що залишилися.

У відповідь на виклики сучасності організації змушені переосмислювати свої підходи до управління персоналом та розробляти нові стратегії утримання працівників. Зростає потреба у впровадженні систем раннього прогнозування ризиків звільнення, що дозволяють своєчасно виявляти потенційні проблеми та вживати превентивних заходів. Тому розробка ефективних інструментів для прогнозування плинності кадрів та запобігання її негативним наслідкам є важливим завданням для роботодавців, держави та наукової спільноти.

Питаннями дослідження можливостей використання штучного інтелекту (ШІ) в управлінні персоналом займалися багато зарубіжних вчених, серед яких: Д. Вронтіс [1], П.Р. Палос–Санчес [2], Н.А. Ракха [3], Р. Менака [4], Ша Рі На [5] та інші. Аналіз наукових публікацій показує, що дослідники приділяють значну увагу питанням впровадження штучного інтелекту в HR–процеси. Зокрема, вивчаються аспекти автоматизації рекрутингу (Woods S.[6]), трансформації HR–функцій під впливом ШІ (Вронтіс Д. [1]), етичних аспектів застосування ШІ в управлінні персоналом (Денніс М. [7]). Дослідження впливу ШІ на управління персоналом та бізнес-процеси активно розвивається у вітчизняній науці. Кравчук О. І. з співавторами досліджували етичні аспекти та прозорість прийняття рішень при впровадженні ШІ в HR-процеси [8]-[9]. Бей Г.В. та Середа Г.В. зосередились на впливі технологічних змін на HR-фахівців та їх компетенції [10]. Чернишова О.О. з колегами вивчали роль ШІ в оптимізації

бізнес-процесів в кризових умовах, особливо через впровадження чат-ботів [11]. Черненко Н. дослідила практичні аспекти та ризики використання ІІІ в управлінні персоналом [12].

Водночас недостатньо дослідженими залишаються питання використання технологій ІІІ для прогнозування звільнень працівників. Зокрема, потребують глибшого вивчення методи предиктивної аналітики для виявлення ризиків звільнення ключових співробітників, алгоритми раннього попередження про можливу плинність кадрів, підходи до аналізу причин звільнень з використанням машинного навчання. Також бракує емпіричних досліджень ефективності систем ІІІ в прогнозуванні звільнень та їх впливу на утримання талантів в організації.

У дипломній кваліфікаційній роботі магістра розв'язується актуальне завдання розробки системи прогнозування звільнення працівників на основі моделей штучного інтелекту. Система повинна забезпечити високий рівень точності прогнозів, підтримувати прийняття управлінських рішень та сприяти зниженню рівня плинності кадрів в організаціях.

Для досягнення поставленої мети необхідно вирішити такі задачі:

- провести аналіз існуючих підходів до прогнозування плинності кадрів та методів машинного навчання, які застосовуються для вирішення подібних задач;
- визначити ключові фактори, які впливають на рішення працівників про звільнення;
- розробити ефективну методику обробки та аналізу даних для підготовки їх до моделювання;
- обрати та налаштувати алгоритми машинного навчання для побудови моделей прогнозування плинності кадрів;
- здійснити проєктування та реалізувати компоненти системи прогнозування звільнення працівників;
- виконати тестування розробленої системи прогнозування;
- визначити шляхи подальшого розвитку розробленої системи прогнозування.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Аналіз проблеми плинності кадрів та методів її прогнозування

Український ринок праці характеризується значною конкуренцією за висококваліфікованих фахівців та високим рівнем трудової міграції, що робить проблему плинності кадрів особливо актуальною. За даними Державної служби статистики України, середній рівень плинності кадрів в різних галузях становить 15–40%, що суттєво перевищує рекомендовані показники 5–10% [13].

Плинність кадрів негативно впливає на всі аспекти діяльності організацій, породжуючи як прямі, так і непрямі наслідки. Одним із найвідчутніших наслідків є зростання витрат. За даними Society for Human Resource Management (SHRM), заміна одного співробітника може обійтися організації від 6 до 9 місячних зарплат цього працівника [14]. Ці витрати включають пошук, підготовку та адаптацію нового персоналу. Водночас економічні наслідки не обмежуються лише прямими витратами. Втрата ключових співробітників знижує продуктивність компанії, адже нові кадри потребують часу для досягнення оптимального рівня ефективності. Для висококонкурентних галузей, таких як ІТ або консалтинг, це може означати втрату клієнтів чи зрив стратегічних проєктів.

Крім фінансових витрат, звільнення працівників впливають на психологічну атмосферу в колективі. Часті звільнення створюють атмосферу невизначеності, знижуючи мотивацію та залученість інших працівників. Це може призвести до "ефекту доміно", коли звільнення одного співробітника тягне за собою звільнення інших.

Соціальні наслідки плинності кадрів також вагомі. Постійна зміна місця роботи знижує лояльність працівників до роботодавців та загальне задоволення життям. Нестабільна зайнятість негативно впливає на добробут родин працівників, знижуючи довіру до роботодавців.

Репутаційні ризики компаній також посилюються через високий рівень плинності кадрів. Організації з поганою репутацією часто отримують негативні відгуки на платформах для пошуку роботи, таких як Glassdoor або українські

ресурси (наприклад, Work.ua та Rabota.ua). Ці відгуки швидко поширюються і можуть ускладнити залучення талановитих спеціалістів.

Зважаючи на зазначене, плинність кадрів створює значні економічні, соціальні та репутаційні виклики для організацій. У зв'язку з цим особливо важливо прогнозувати проблему, що стає можливим завдяки сучасним інструментам, які дозволяють ідентифікувати ризики та впроваджувати заходи для утримання кадрів.

Методи прогнозування залежать від рівня цифровізації компанії та її ресурсів. Традиційні підходи включають якісний аналіз даних, отриманих за допомогою опитувань, інтерв'ю та оцінок задоволеності працівників. Наприклад, регулярні анкетування дозволяють визначити загальні настрої у колективі. Однак ці методи мають суб'єктивний характер, вимагають значних часових витрат і не завжди є ефективними у великих організаціях. В Україні традиційні підходи поширені серед малих та середніх підприємств через їх доступність та невисокі витрати.

Поширеною практикою є Exit-інтерв'ю, яке проводять після звільнення працівника. Проте часто це здійснюється формально, оскільки працівники уникають прямої критики компанії. Серед методів оцінювання залученості працівників найпоширенішими є регулярні опитування, але їх ефективність обмежена низьким рівнем довіри до анонімності, формальним підходом до участі та культурними особливостями.

У середніх та малих компаніях часто взагалі відсутні формалізовані системи аналізу плинності кадрів. У таких організаціях рішення приймаються реактивно, після подачі заяви про звільнення. HR-менеджери часто покладаються на неформальні комунікації, моніторинг корпоративних чатів та настроїв у колективі.

Аналіз HR-метрик в організаціях зазвичай обмежується базовими показниками, такими як відвідуваність, виконання KPI, участь у заходах. Основними бар'єрами є недостатня культура збору даних, обмежені бюджети та брак спеціалістів. Це підкреслює потребу у сучасних аналітичних підходах до управління персоналом.

Сучасні аналітичні методи включають використання великих даних та алгоритмів машинного навчання. Ці технології дозволяють автоматизувати аналіз і враховувати численні фактори, що впливають на рішення про звільнення. Наприклад, компанії IBM та Google, використовують моделі прогнозування, що аналізують історичні та поведінкові дані працівників. В українському бізнесі сучасні інструменти прогнозування також стають популярними, особливо це помітно в IT–секторі, великих міжнародних корпораціях та фінансовій сфері, де високий рівень цифровізації [14].

Інноваційні технології кардинально змінюють роль HR–аналітики, перетворюючи її зі звичайного допоміжного інструменту в ключовий стратегічний ресурс для управління організацією [15].

Одним із найбільш вагомих досягнень у цій сфері є використання технологій аналізу великих даних (Big Data), які дозволяють компаніям вивчати поведінкові тренди працівників, прогнозувати їх активність і приймати стратегічні рішення на основі об'єктивних даних. Ці технології дають змогу враховувати не лише стандартні показники продуктивності, але й взаємодію в колективі, рівень задоволеності роботою та навіть поведінкові зміни, що можуть сигналізувати про ризик звільнення [15].

Невід'ємною частиною HR–аналітики стали алгоритми машинного навчання, адже вони здатні обробляти величезні масиви даних і знаходити складні закономірності, які залишаються непомітними при використанні традиційних підходів. Наприклад, такі алгоритми можуть аналізувати частоту відвідувань корпоративних заходів, зміни у стилі електронного листування або поведінку працівників у внутрішніх комунікаційних системах. Ці інструменти допомагають передбачати ризики плинності кадрів, ідентифікувати ключових співробітників, які можуть бути незадоволені своєю поточною посадою, та пропонувати своєчасні заходи для їх утримання.

Інтеграція HR–аналітики з корпоративними системами управління, такими як ERP (Enterprise Resource Planning) та CRM (Customer Relationship Management), забезпечує повноцінний доступ до даних у реальному часі. Цей підхід дозволяє керівникам отримувати комплексну картину стану персоналу,

що охоплює як операційні, так і стратегічні аспекти управління. У реальному часі можна оцінювати ефективність співробітників, рівень їх залученості та оперативно реагувати на будь-які негативні тенденції. В Україні великі підприємства, особливо у фінансовому секторі та виробничій галузі, активно впроваджують такі інтеграційні рішення для оптимізації роботи своїх HR-департаментів.

Проактивна HR-аналітика є одним із найбільш перспективних напрямів розвитку в управлінні персоналом. Вона дозволяє не лише аналізувати минулі події, а й прогнозувати майбутні ситуації, такі як ризик звільнення співробітників чи можливі проблеми у колективі. Використовуючи інструменти штучного інтелекту, проактивна аналітика допомагає виявляти приховані тренди та патерни, які можуть сигналізувати про потенційні проблеми. Наприклад, система може ідентифікувати групи працівників із подібними рівнями залученості або схожими викликами, з якими вони стикаються, і запропонувати індивідуальні рішення для кожної групи.

Персоналізований підхід, який є однією з основ сучасної HR-аналітики, дозволяє створювати індивідуальні програми розвитку та мотивації для кожного працівника. Наприклад, аналіз даних про продуктивність і залученість дозволяє розробити адаптовані тренінгові програми, які відповідають потребам конкретної особи. Такі системи також допомагають виявити співробітників із високим потенціалом для кар'єрного зростання і запропонувати їм відповідні можливості для розвитку.

Додатково, новітні технології, що використовують біометричні дані та аналіз емоцій, поступово впроваджуються у провідних міжнародних компаніях. Ці рішення дозволяють отримувати дані про стан працівників на основі аналізу міміки, тону голосу або поведінки під час комунікації. Наприклад, системи можуть визначити рівень стресу або втоми працівника та запропонувати відповідні заходи для їхнього зменшення, що сприяє створенню більш комфортного робочого середовища.

Таким чином, інтеграція інноваційних технологій у HR-аналітику відкриває нові можливості для організацій, дозволяючи не лише реагувати на

проблеми, але й запобігати їм на ранніх етапах. Це підвищує ефективність управління персоналом, знижує витрати, пов'язані з плинністю кадрів, та сприяє створенню стабільного та продуктивного робочого середовища.

Незважаючи на зростаючий інтерес до HR–аналітики в Україні, існують виклики, пов'язані з низьким рівнем цифровізації, браком фахівців у сфері Data Science і питаннями конфіденційності даних. Водночас зростання попиту на аналітичні рішення створює перспективи для їх подальшого впровадження.

1.2 Огляд сучасних AI–рішень для управління персоналом

Огляд інструментів штучного інтелекту для HR у 2024 році наведено у статті [14]. Авторами виділено переваги впровадження штучного інтелекту в HR, такі як підвищення ефективності навчання, автоматизація рутинних завдань, зменшення упередженості під час найму та моніторинг добробуту співробітників у режимі реального часу. Серед наведених систем, які можуть допомогти прогнозувати звільнення працівників, варто звертати увагу на рішення, що використовують аналітику ефективності та моніторинг залученості співробітників. Такі інструменти аналізують поведінкові та продуктивні показники, що дозволяє HR–менеджерам виявляти потенційні ризики звільнення та вживати превентивних заходів.

Серед численних рішень прогнозування плинності кадрів, представлених на ринку України, окремі системи мають свої унікальні особливості, що відповідають потребам різних компаній залежно від їх масштабу, рівня цифровізації та специфіки діяльності:

- PeopleForce, українська HRM–система, спеціалізується на наданні інструментів для аналізу залученості працівників та прогнозування плинності. Її перевагою є адаптованість до потреб місцевого ринку та підтримка українською мовою, що робить систему особливо привабливою для малих та середніх підприємств. Гнучка цінова політика дозволяє компаніям з обмеженими ресурсами інтегрувати сучасні інструменти HR–аналітики [16];

– інша українська система, Hurma System, також орієнтована на автоматизацію HR-процесів із додатковими можливостями прогнозування. Завдяки використанню алгоритмів штучного інтелекту, система дозволяє аналізувати поведінкові аспекти співробітників, ідентифікувати потенційні ризики звільнення та пропонувати заходи для їх запобігання. Локальна підтримка та доступність роблять цю систему привабливим вибором для компаній із середнім рівнем цифровізації [17];

– міжнародні рішення, такі як KISSFLOW HR Cloud, пропонують широкі можливості інтеграції з іншими бізнес-додатками та високий рівень функціональності, орієнтований на управління персоналом і автоматизацію процесів. Хоча система забезпечує базові інструменти для прогнозування плинності кадрів, її підтримка англійською мовою може стати бар'єром для локальних компаній, які прагнуть впровадити це рішення [18];

– Culture Amp, ще одна міжнародна платформа, вирізняється своєю спеціалізацією на вимірюванні залученості співробітників, зборі зворотного зв'язку та аналітиці для покращення корпоративної культури. Завдяки розширеним можливостям інтеграції з різними HR-системами, платформа дозволяє детально аналізувати дані про співробітників і впроваджувати персоналізовані заходи для їх утримання. Це робить Culture Amp особливо корисним для компаній, які зосереджуються на формуванні позитивної корпоративної культури та підвищенні залученості [19];

– BambooHR, пропонує широкий спектр функцій для управління персоналом, зокрема рекрутинг, прогнозування плинності та аналітику даних. Інтеграція з численними HR-додатками забезпечує високу гнучкість у використанні, що робить BambooHR ефективним рішенням для середніх і великих організацій. Водночас обмеження підтримки виключно англійською мовою може впливати на зручність використання цієї платформи в локальних умовах [20].

Порівняльний аналіз зазначених систем наведено в таблиці 1.1.

Таблиця 1.1 – Порівняльний аналіз HR–систем, представлених на ринку

України

Система	PeopleForce	Hurma System	KISSFLOW HR Cloud	Culture Amp	BambooHR
Локалізація	Українська розробка, адаптована під локальний ринок	Українська розробка, орієнтована на місцеві компанії	Міжнародне рішення, доступне в Україні	Міжнародне рішення, доступне в Україні	Міжнародне рішення, доступне в Україні
Функціонал	Управління персоналом, аналітика, прогнозування плинності	Автоматизація HR–процесів, аналітика, прогнозування плинності	Управління персоналом, автоматизація процесів, прогнозування плинності	Опитування залученості, зворотний зв'язок, аналітика, прогнозування плинності	Управління персоналом, рекрутинг, аналітика, прогнозування плинності
Інтеграції	Інтеграція з популярними сервісами та платформами	Підтримка інтеграцій з різними системами	Широкі можливості інтеграції з іншими бізнес–додатками	Інтеграція з HRIS, SFTP, Slack, SSO, MS Teams	Інтеграція з різними HR–додатками та платформами
Підтримка	Локальна підтримка українською мовою	Локальна підтримка українською мовою	Підтримка англійською мовою	Підтримка англійською мовою	Підтримка англійською мовою
Цінова політика	Гнучка цінова політика, орієнтована на український ринок	Доступні тарифи для малого та середнього бізнесу	Вартість залежить від обраного пакету та кількості користувачів	Вартість залежить від обраного пакету та кількості користувачів	Вартість залежить від обраного пакету та кількості користувачів

Таким чином, кожна з цих систем пропонує унікальні рішення для аналізу та управління плинністю кадрів, що дозволяє компаніям обирати платформу відповідно до потреб, бюджету та технічних можливостей. Це свідчить про зростаючу роль цифрових технологій у вдосконаленні процесів управління персоналом і формуванні стійких корпоративних структур.

1.3 Особливості використання методів машинного навчання для задач прогнозування плинності кадрів

Задача прогнозування звільнення працівників відноситься до задач бінарної класифікації, в якій кожен об'єкт $K_i (i = 1, \dots, N)$ характеризується m -мірним вектором ознак $(X_1 \dots X_m)$. Модель має передбачити приналежність об'єкта (працівника) до одного з двох класів: 1 – "звільниться" чи 0 – "залишиться". Найпоширенішими алгоритмами для таких задач є логістична регресія, метод опорних векторів, дерева рішень, нейронні мережі, байєсівський класифікатор та ін., які використовують різний математичний апарат та різні підходи при реалізації. Вибір конкретного методу залежить від багатьох факторів, включаючи обсяг та якість доступних даних, вимоги до інтерпретованості результатів, обчислювальні ресурси та специфіку конкретної організації. При цьому важливо розуміти, що жоден метод не є універсально найкращим, і часто оптимальним рішенням є комбінація різних підходів [21].

Для вирішення задачі прогнозування звільнення працівників доцільно використати ансамблювання моделей, оскільки цей підхід дозволяє поєднати переваги різних алгоритмів класифікації та підвищити точність прогнозування. Існує два основних підходи до комбінування моделей:

- гомогенний підхід передбачає ансамблювання моделей однакового типу. Наприклад, у методі Random Forest використовується множина дерев рішень, а в AdaBoost – послідовність простих класифікаторів одного типу. Перевагою такого підходу є можливість ефективної оптимізації параметрів однотипних моделей та простота реалізації;

- гетерогенний підхід базується на комбінуванні різнотипних моделей, що дозволяє використовувати їх взаємодоповнюючі сильні сторони. Наприклад, можна поєднати логістичну регресію, яка добре працює з лінійними залежностями, з нейронною мережею, здатною виявляти складні нелінійні патерни у даних. Такий підхід реалізується через метод Stacking, де мета-модель навчається оптимально комбінувати прогнози базових моделей.

Розглянемо основні методи ансамблювання [22].

Bagging (Bootstrap Aggregating) – метод паралельного навчання однотипних моделей на різних підвибірках даних з подальшим усередненням результатів (рис. 1.1). Наприклад, Random Forest використовує цей принцип, навчаючи множину дерев рішень, що особливо ефективно для зменшення дисперсії прогнозів та уникнення перенавчання.

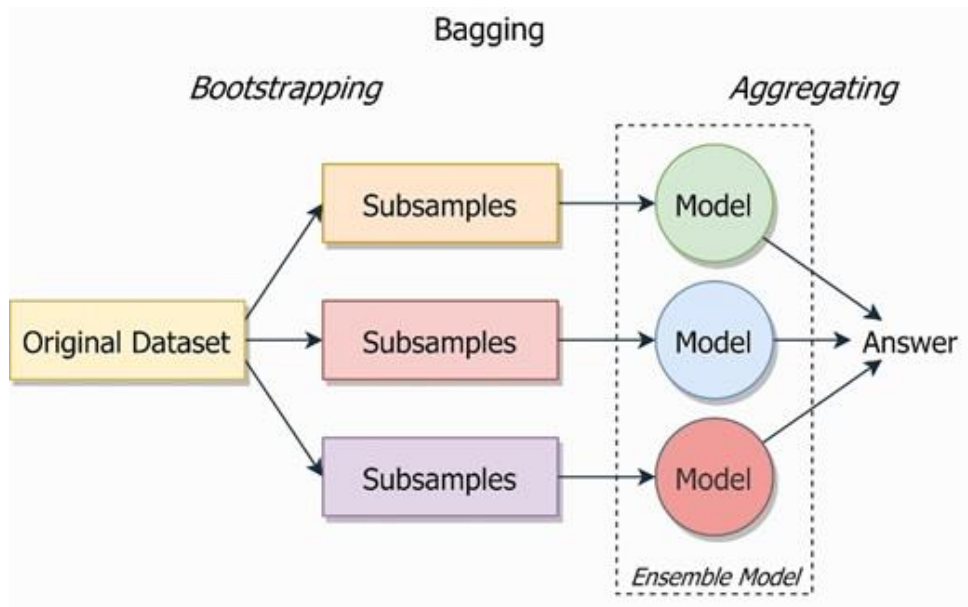


Рисунок 1.1 – Метод ансамблювання Bagging [23]

Boosting – послідовне навчання моделей з фокусом на помилках попередніх ітерацій (рис. 1.2). Алгоритми XGBoost та LightGBM, що базуються на цьому принципі, добре працюють з незбалансованими даними та дозволяють зменшити зміщення моделі. Це особливо актуально для задачі прогнозування звільнень, де класи часто бувають незбалансованими (звільнених працівників зазвичай менше, ніж тих, хто залишається).

Stacking – метод, що дозволяє комбінувати прогнози різнотипних моделей через навчання мета-моделі (рис. 1.3). Цей підхід дає можливість використовувати сильні сторони різних алгоритмів класифікації (наприклад, логістичної регресії, SVM, нейронних мереж) для отримання більш надійних прогнозів.

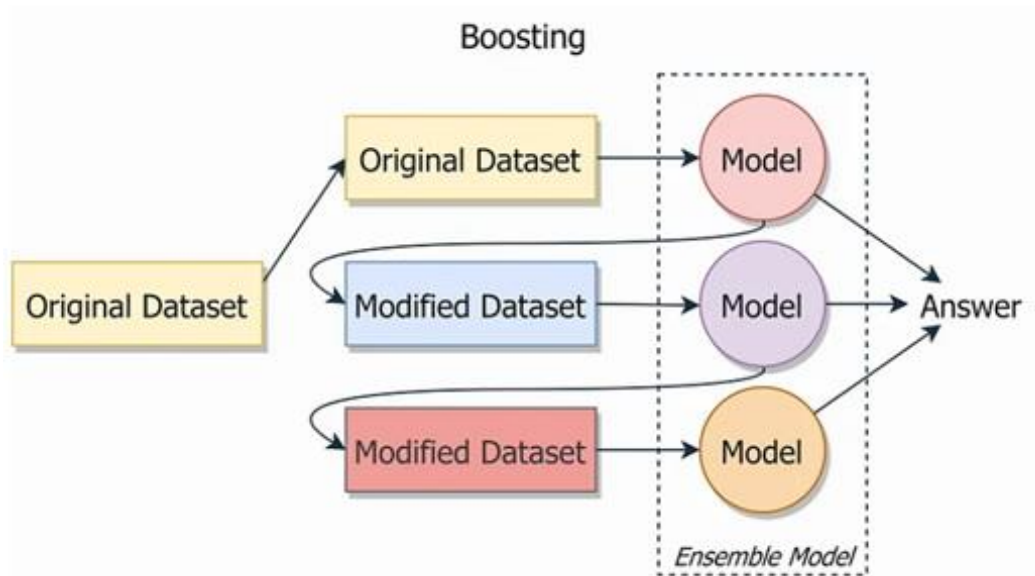


Рисунок 1.2 – Метод ансамблювання Boosting [23]

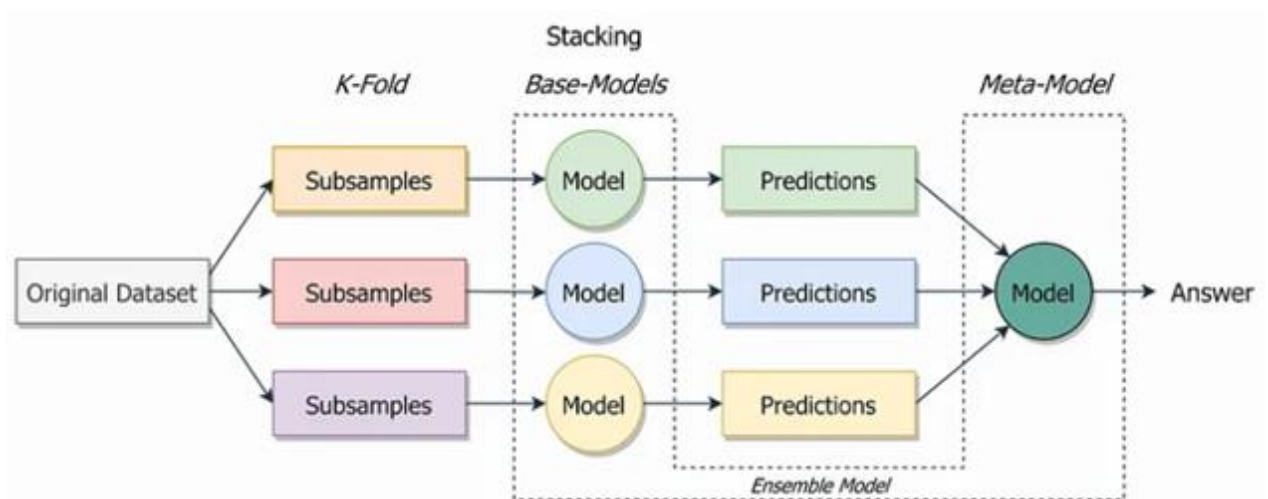


Рисунок 1.3 – Метод ансамблювання Stacking [23]

В контексті задачі прогнозування звільнень ансамблювання дозволить не тільки підвищити точність прогнозів, але й отримати більш стабільні результати навіть при наявності шуму в даних чи недостатній кількості прикладів звільнень у навчальній вибірці.

Ключовим етапом побудови моделей є ідентифікація факторів, які впливають на ймовірність звільнення. Це можуть бути: рівень зарплати, тривалість роботи на поточній посаді, зміна кар'єрного шляху, частота підвищень або конфлікти з керівництвом. Сучасні алгоритми дозволяють враховувати

десятки або навіть сотні таких факторів одночасно, створюючи персоналізовані прогнози для кожного працівника.

Рішення працівника про звільнення є багатофакторним процесом, на який впливають як внутрішні, так і зовнішні чинники. Внутрішні фактори найчастіше пов'язані з умовами роботи, рівнем задоволеності працівника та взаємовідносинами в колективі. Наприклад, низький рівень заробітної плати чи відсутність можливостей для кар'єрного зростання стають одними з головних причин звільнень. В Україні ця проблема є особливо актуальною через обмежену кількість програм професійного навчання у багатьох компаніях.

Важливу роль відіграють соціальні фактори, такі як атмосфера в колективі та стиль керівництва. Конфлікти з колегами чи керівниками, відсутність зворотного зв'язку або підтримки можуть суттєво знизити залученість працівника до роботи. Особливо це стосується великих організацій, де працівники можуть почуватися ізольованими або недооціненими.

Зовнішні фактори включають ситуацію на ринку праці, економічну стабільність та наявність альтернативних можливостей. Наприклад, зростання зарплат у конкурентних компаніях чи розширення можливостей для віддаленої роботи під час пандемії COVID-19 спонукало багатьох працівників до зміни роботодавця. Для України до зовнішніх факторів також відносяться міграційні процеси, викликані війною, які суттєво вплинули на мобільність робочої сили. За даними Міністерства економіки України, у 2023 році понад 5 мільйонів осіб змінили своє місце роботи через релокацію або економічні труднощі на попередньому місці роботи.

Необхідно також враховувати особисті характеристики працівника, такі як вік, стаж роботи, рівень освіти та сімейний статус. Молодші працівники частіше змінюють місце роботи, шукаючи кращі умови або перспективи, тоді як старші працівники більш схильні залишатися на стабільних посадах.

Таким чином, фактори, що впливають на рішення про звільнення, є складними та багатовимірними. Їх аналіз є критичним етапом для створення точних моделей прогнозування. Використання сучасних аналітичних методів

дозволяє виявляти неочевидні закономірності між різними змінними, що сприяє більш ефективному управлінню кадровими ресурсами.

Машинне навчання має значний потенціал для HR-аналітики, проте його застосування супроводжується певними особливостями. По-перше, HR-дані часто є неоднорідними та містять як числові, так і категорійні змінні, які вимагають попередньої обробки. Наприклад, інформація про вік, стаж роботи чи кількість підвищень потребує нормалізації, тоді як текстові дані, такі як відгуки з опитувань, можуть бути оброблені за допомогою методів обробки природної мови (NLP).

По-друге, забезпечення конфіденційності даних є критично важливим у HR. Законодавство вимагає, щоб дані працівників оброблялися відповідально та з дотриманням етичних норм. Це створює додаткові виклики для побудови моделей, які мають аналізувати дані без порушення приватності.

Дисбаланс даних є ще одним характерним викликом. У більшості компаній частка працівників, які звільняються, є значно меншою за тих, хто залишається. Це може призвести до зміщення моделей у бік більшості класу. Для вирішення цієї проблеми застосовуються такі підходи, як ресемплінг або зміна ваг у функціях втрат.

Для прогнозування звільнення працівників важливо правильно вибрати метрику для оцінювання моделей, оскільки різні метрики можуть мати різну вагу залежно від поставленої бізнес-мети. Важливо оцінювати не лише точність класифікації, але й те, наскільки добре модель передбачає позитивний клас (звільнення), який часто є менш поширеним у порівнянні з негативним класом (залишилися на роботі).

В задачах бінарної класифікації використовується матриця неточностей (confusion matrix), де об'єкти вибірки поділяються на два класи – позитивні (positive) та негативні (negative). Модель класифікації при цьому призначає об'єктам одну з двох міток: позитивну або негативну. У ході роботи моделі з тестовою вибіркою можуть виникати помилки, внаслідок чого класифіковані об'єкти розподіляються на чотири групи:

- істинно–позитивні (true positive, TP) об'єкти, які правильно класифіковано як позитивні;
- істинно–негативні (true negative, TN) об'єкти, які правильно класифіковано як негативні;
- хибно–позитивні (false positive, FP) – помилки, коли об'єкти помилково віднесено до позитивного класу (помилка 1–го роду);
- хибно–негативні (false negative, FN) – помилки, коли об'єкти помилково віднесено до негативного класу (помилка 2–го роду).

Матриця неточностей дає змогу оцінити роботу моделі, аналізуючи кількість коректних і некоректних передбачень для кожного класу. Ця оцінка є ключовою для розуміння якості роботи алгоритму класифікації (рис. 1.4).

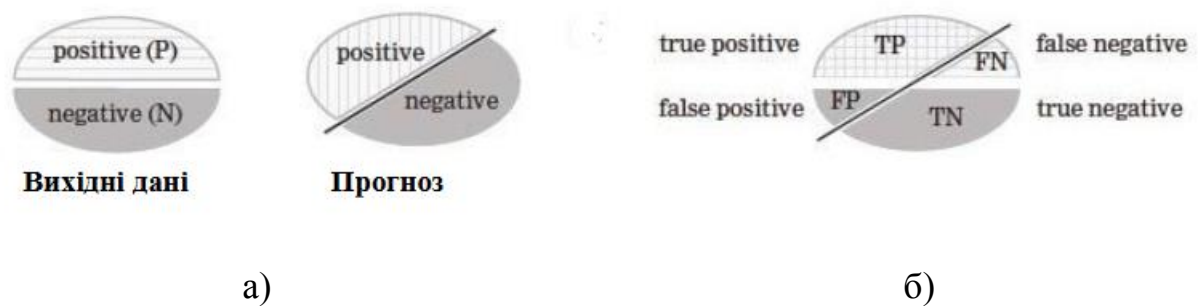


Рисунок 1.4 – Вихідні та прогнознi дані бiнарної класифікації (а) та матриця неточностей (б) [21]

Використовуючи матрицю неточностей, розглянемо метрики якості моделі для вибору релевантних в прогнозуванні звільнення працівників.

Recall (повнота) – частка працівників, що звільнилися, яких модель знайшла серед звільнених працівників, визначається за формулою (1.1):

$$\text{Recall} = \frac{TP}{TP+FN}, \quad (1.1)$$

де TP – працівники, які звільнилися та були правильно передбачені;

FN – працівники, які звільнилися, але модель передбачила, що вони залишаються.

Основна мета задачі – не пропустити працівників, які, ймовірно, звільняться. Високий Recall означає, що модель виявляє більшу кількість працівників, які планують звільнення.

Precision (точність) – частка дійсно звільнених працівників серед усіх, кого модель передбачила звільненими, визначається за формулою (1.2):

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (1.2)$$

де TP – працівники, які звільнилися та були правильно передбачені;

FP – працівники, які не планували звільнення, але модель помилково передбачила їх звільнення.

Низький показник точності означає, що модель часто помиляється, передбачаючи звільнення для працівників, які насправді залишаються.

F1-Score – гармонічне середнє між точністю (Precision) та повнотою (Recall), визначається за формулою (1.3):

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (1.3)$$

Ця метрика добре підходить у випадках, коли важливо досягти балансу між високою точністю і повнотою.

Accuracy (точність класифікації) – частка правильно передбачених класів (як залишилися, так і звільнилися), визначається за формулою (1.4):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}. \quad (1.4)$$

Цей показник є найпростішим для інтерпретації та розуміння, однак він працює, коли набір даних збалансований і відсутні перевага або особливий інтерес до прогнозування будь-якого з класів.

ROC AUC Score (Area Under Curve) – вимірює здатність моделі відрізняти між позитивним і негативним класом. Показує, наскільки добре модель розрізняє

працівників, які звільняються, від тих, що залишається. Площа під ROC–кривою, яка вимірює повноту (Recall) і специфічність на різних порогах. Високий ROC AUC Score вказує на те, що модель здатна добре розділяти класи.

Recall є найбільш критичною метрикою. Пропущений працівник, який планує звільнення, може бути більш критичним, ніж помилкове передбачення звільнення. Наприклад, якщо Recall високий, модель добре визначає всіх, хто схильний до звільнення, що дозволяє HR–команді швидко втрутитися та вжити запобіжних заходів. Треба також звернути увагу на Precision. Низька точність означає, що багато працівників, які не планують звільнення, будуть помилково позначені як схильні до цього, що може призвести до зайвих витрат на утримання. Щоб досягти балансу між Recall і Precision використовується F1–Score. Для оцінки загальної ефективності моделі використовується ROC AUC Score, особливо якщо є незбалансовані дані.

1.4 Висновки за розділом 1

Плинність кадрів є значною проблемою для українських підприємств, що призводить до суттєвих фінансових втрат, зниження продуктивності та погіршення психологічного клімату в колективі.

Традиційні методи прогнозування звільнень (опитування, інтерв'ю, аналіз базових HR–метрик) мають обмежену ефективність через суб'єктивність та реактивний характер. Натомість сучасні аналітичні методи, засновані на машинному навчанні та аналізі великих даних, дозволяють проактивно виявляти ризики звільнення працівників.

Представлений спектр HR–систем для управління персоналом на ринку України – від локальних рішень до міжнародних платформ – має свої особливості щодо функціоналу, локалізації та цінової політики, що дозволяє компаніям обирати оптимальне рішення відповідно до своїх потреб.

Для прогнозування звільнень доцільно використовувати ансамблювання моделей машинного навчання, яке дозволяє поєднати переваги різних алгоритмів класифікації.

При розробці моделей прогнозування важливо враховувати специфіку HR–даних: неоднорідність, необхідність забезпечення конфіденційності та проблему незбалансованості класів.

Важливим є правильний вибір метрик оцінювання моделей (Recall, Precision, F1–Score, ROC AUC Score) відповідно до бізнес–цілей та особливостей задачі.

2 КОМПЛЕКСНИЙ АНАЛІЗ НАБОРУ ДАНИХ

2.1 Аналіз структури та характеристик набору даних

Для розробки системи прогнозування було використано великий набір даних, що містить інформацію про 74 498 співробітників з джерела у відкритому доступі [24]. Набір даних охоплює 24 змінні, які характеризують різні аспекти професійного життя, особистих характеристик та умов роботи співробітників. Опис змінних, а також визначення підходів до їх кодування наведено в табл. 2.1.

Таблиця 2.1 – Опис та метод кодування ознак

Ознака	Опис
1	2
Цільова змінна, кодування за допомогою target mapping	
Attrition	містить дві категорії – не звільнився (Stayed), звільнився (Left)
Числові (кількісні) змінні, кодування не здійснюється	
Age	вік співробітників варіюється
Company Tenure	кількість років, протягом яких компанія існує
Distance from Home	відстань від місця мешкання до робочого місця
Employee ID	унікальний ідентифікаційний номер кожного співробітника
Monthly Income	щомісячна заробітна плата
Number of Dependents	кількість утриманців у працівника, варіюється від 0 до 6
Number of Promotions	кількість підвищень, які отримав співробітник протягом роботи, варіюється від 0 до 4
Years at Company	кількість років, які співробітник пропрацював у компанії
Бінарні змінні, кодування за допомогою LabelEncoder	
Gender	стать, містить дві категорії – Male (чоловік) та Female (жінка)
Innovation Opportunities	наявність можливостей для інновацій (Yes, No)
Remote Work	вказує, чи працює співробітник віддалено (Yes, No)
Leadership Opportunities	можливість виконання лідерських функцій (Yes, No)
Overtime	показує, чи працює співробітник понаднормово (Yes, No)

Кінець таблиці 2.1

1	2
Порядкові змінні, кодування за допомогою Ordinal Encoding	
Company Reputation	сприйняття працівником репутації компанії (Poor, Fair, Good, Excellent)
Company Size	розмір компанії (Small, Medium, Large)
Education Level	рівень освіти (High School, Associate Degree, Bachelor Degree, Master Degree, PhD)
Employee Recognition	рівень визнання співробітника (Low, Medium, High, Very High)
Job Level	рівень посади співробітника (Entry, Mid, Senior)
Job Satisfaction	задоволеність працівника своєю роботою (Low, Medium, High, Very High)
Performance Rating	оцінка продуктивності працівника (Low, Below Average, Average, High)
Work–Life Balance	задоволеність працівника балансом між роботою та особистим життям (Poor, Fair, Good, Excellent)
Номінативні змінні, кодування за допомогою One–Hot Encoding	
Job Role	напрямок, за яким працює співробітник (Education, Media, Healthcare, Technology, Finance)
Marital Status	сімейний стан працівника (Married, Divorced, Single)

На початковому етапі дослідження були проведені операції з імпорту, очищення та перетворення даних для створення якісної бази для подальшого моделювання. Для забезпечення цілісності даних було виконано попередній аналіз на наявність пропущених значень та дублікатів (рис. 2.1).

Результати показали, що в наборі даних:

- відсутні пропущені значення, жодна змінна не містить пропусків, що свідчить про повну цілісність даних;
- не було знайдено дубльованих рядків, що свідчить про унікальність записів у наборі даних.

Після обробки набору даних було видалено стовпець Employee ID, оскільки він не є інформативним для моделювання. Всі інші змінні були відповідним чином закодовані.

Аналіз основних характеристик колонок:

	Column	Non-null	Null	Unique	Data type
0	Attrition	74498	0	2	object
1	Age	74498	0	42	int64
2	Company Reputation	74498	0	4	object
3	Company Size	74498	0	3	object
4	Company Tenure	74498	0	127	int64
5	Distance from Home	74498	0	99	int64
6	Education Level	74498	0	5	object
7	Employee ID	74498	0	74498	int64
8	Employee Recognition	74498	0	4	object
9	Gender	74498	0	2	object
10	Innovation Opportunities	74498	0	2	object
11	Job Level	74498	0	3	object
12	Job Role	74498	0	5	object
13	Job Satisfaction	74498	0	4	object
14	Leadership Opportunities	74498	0	2	object
15	Marital Status	74498	0	3	object
16	Monthly Income	74498	0	9842	int64
17	Number of Dependents	74498	0	7	int64
18	Number of Promotions	74498	0	5	int64
19	Overtime	74498	0	2	object
20	Performance Rating	74498	0	4	object
21	Remote Work	74498	0	2	object
22	Work-Life Balance	74498	0	4	object
23	Years at Company	74498	0	51	int64

Стовпці з пропущеними значеннями: []

Кількість пропущених значень після заповнення: 0

Кількість дублюючих рядків: 0

Дублюючих рядків не виявлено.

=====

Кількість спостережень у наборі даних: 74498

=====

Рисунок 2.1 – Попередній аналіз набору даних

На рисунку 2.2 представлено розподіл цільової змінної Attrition, яка має два класи: Stayed (залишився в компанії) та Left (звільнився з компанії). Загалом, розподіл класів є відносно збалансованим. Кількість співробітників, які залишилися, незначно перевищує кількість тих, хто звільнився. Така структура даних є позитивною з точки зору моделювання, оскільки збалансовані дані зазвичай забезпечують більш точні результати та уникнення проблеми переобчислення для домінуючого класу.

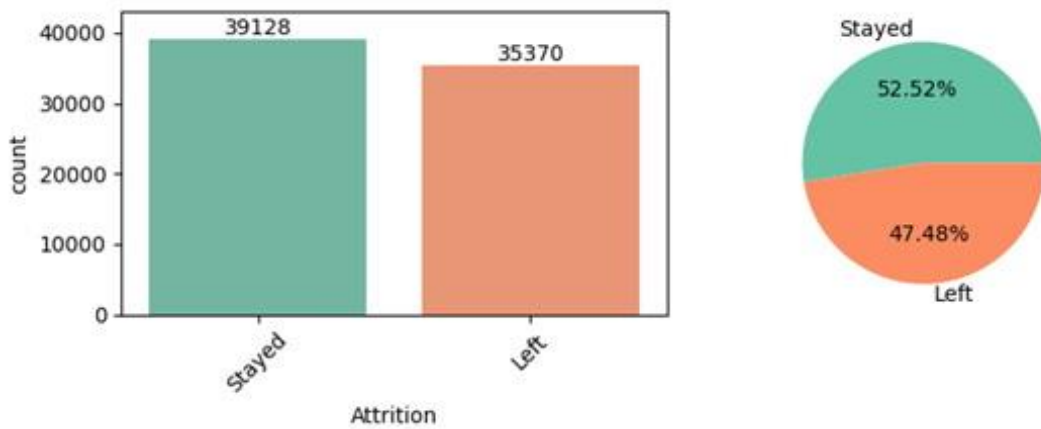


Рисунок 2.2 – Розподіл цільової змінної Attrition

З точки зору побудови моделі прогнозування звільнення, збалансованість класів є важливим фактором. У випадку, якщо класи були б сильно дисбалансовані (наприклад, 80% – Stayed і лише 20% – Left), моделі могли б надавати перевагу більш представленому класу, що призводить до зниження чутливості щодо меншого класу (Left). У поточному випадку, де розподіл є відносно збалансованим, ймовірність такої проблеми є мінімальною. Це означає, що моделі, побудовані на цьому наборі даних, повинні добре розпізнавати обидва стани.

З метою виявлення закономірностей та аномалій у даних було проведено детальний аналіз кількісних змінних. Розгляд змінних включав використання як описової статистики (середнє значення, медіана, інтерквартильний діапазон) (рис. 2.3), так і візуалізацій, таких як гістограми розподілу значень (рис. 2.4) та коробчасті діаграми (рис. 2.5) для ознак з кількістю унікальних значень, що перевищує 10.

З графіків видно, що розподіл вікових груп у наборі від 18 до 60 років є відносно рівномірним, є легке підвищення кількості працівників у середньому віковому діапазоні (30–40 років). Можливо, компанія активно залучає молодих фахівців і підтримує баланс у віковій структурі співробітників. Коробчаста діаграма не виявляє жодних аномалій, що свідчить про рівномірний розподіл працівників у межах основного вікового діапазону. Такий розподіл показує, що

вибірка є репрезентативною для різних вікових груп, від молодих фахівців до більш досвідчених співробітників.

Основні статистичні характеристики числових колонок:

	count	mean	std	min	25%	50%	75%	max
Employee ID	74498.0	37249.500000	21505.864514	1.0	18625.25	37249.5	55873.75	74498.0
Age	74498.0	38.529746	12.083456	18.0	28.00	39.0	49.00	59.0
Years at Company	74498.0	15.721603	11.223744	1.0	7.00	13.0	23.00	51.0
Monthly Income	74498.0	7299.379514	2152.508566	1226.0	5652.00	7348.0	8876.00	16149.0
Number of Promotions	74498.0	0.832935	0.995289	0.0	0.00	1.0	2.00	4.0
Distance from Home	74498.0	49.991584	28.513611	1.0	25.00	50.0	75.00	99.0
Number of Dependents	74498.0	1.650326	1.553633	0.0	0.00	1.0	3.00	6.0
Company Tenure	74498.0	55.727456	25.399349	2.0	36.00	56.0	76.00	128.0

Рисунок 2.3 – Описова статистика для числових ознак

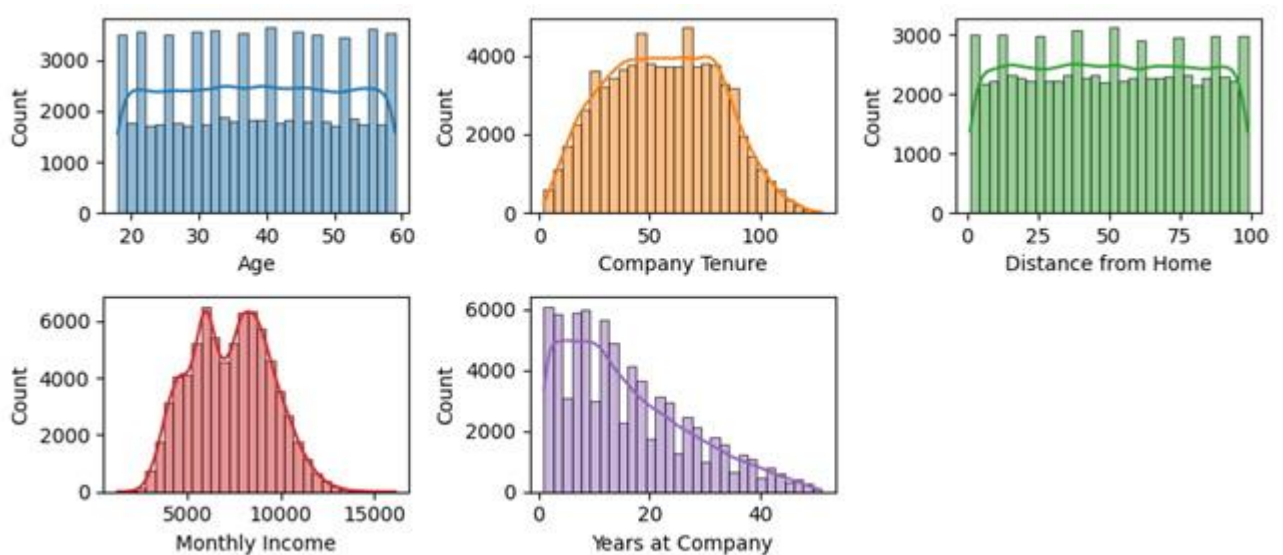


Рисунок 2.4 – Гістограми розподілу значень для числових ознак

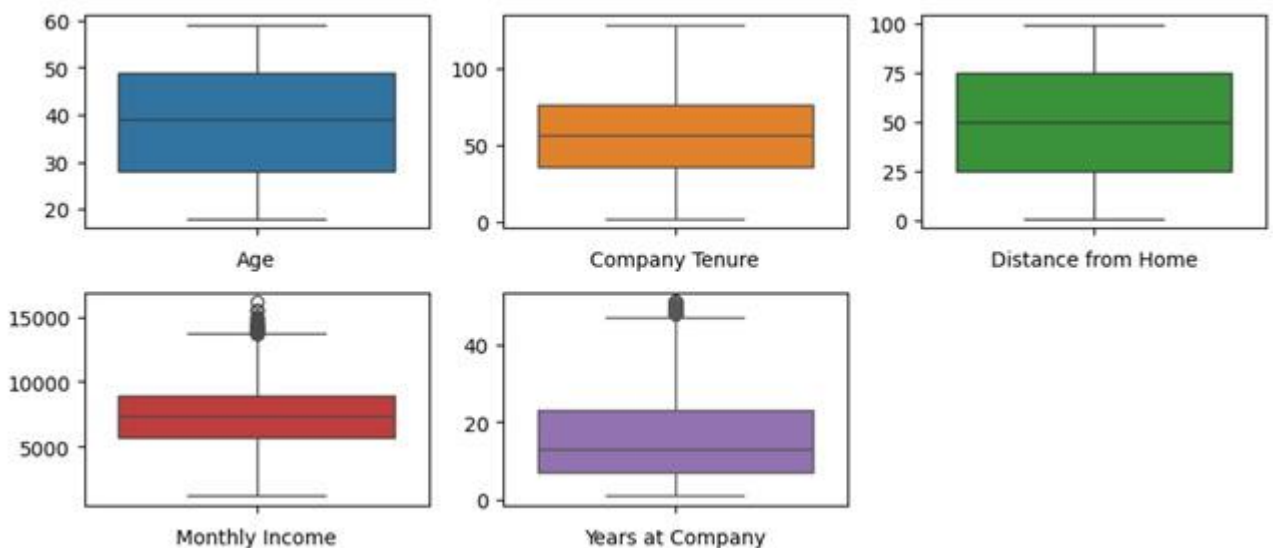


Рисунок 2.5 – Box-plot діаграми для числових ознак

Розподіл кількості років роботи у компанії значно зміщений ліворуч. Переважна кількість співробітників працюють у компанії менше ніж 10 років. Спостерігається різке зниження кількості працівників, які працюють понад 10 років, що може вказувати на високий рівень плинності кадрів або активний найм нових співробітників. Ті, хто працює понад 20 років, становлять лише невелику частину. Коробчаста діаграма показує наявність викидів для працівників, які працювали понад 40 років, що є нетиповим для більшості працівників. Такі викиди можуть вказувати на невелику групу довготривалих працівників, які мають значний досвід та лояльність до компанії.

Розподіл щомісячного доходу працівників має форму двогорбого розподілу. Розподіл зарплат працівників складається з двох основних груп: одна – з нижчими доходами (приблизно 2 500–7 500 \$), друга – з вищими доходами (7 500–12 500 \$). Це може відображати поділ працівників на молодших (менший дохід) і старших співробітників або розподіл між початковими і висококваліфікованими посадами. Коробчаста діаграма демонструє наявність значної кількості викидів для доходів понад 14 000 \$, що може вказувати на невелику групу висококваліфікованих працівників або керівників, які мають набагато вищі доходи порівняно з іншими працівниками.

Рівномірний розподіл відстань від місця проживання до роботи свідчить про те, що компанія має різноманітний склад працівників, які проживають на різних відстанях від робочого місця. Немає значного зростання чи зниження для певних груп, що може вказувати на відсутність географічного обмеження при наймі. Це може вказувати на те, що відстань до роботи не є критичним чинником, який призводить до значних аномалій у вибірці.

Розподіл кількості років існування компанії має форму нормального розподілу з піком у діапазоні 40–60 років. Це може означати, що більшість компаній у вибірці працюють на ринку досить тривалий час, але є також невелика кількість нових і дуже старих компаній. Широкий діапазон даних вказує на включення як молодих, так і зрілих компаній у вибірку, що є важливим для узагальнення висновків на різні типи організацій.

Аналіз показує, що для щомісячного доходу та кількості років роботи в компанії спостерігається наявність викидів, що може впливати на побудову моделей та результати прогнозування. Це свідчить про необхідність попередньої обробки цих змінних, зокрема усунення викидів або їх трансформації для мінімізації впливу аномальних значень.

Інші змінні не показують аномалій, що вказує на стабільний розподіл значень і відсутність потреби в їх додатковій обробці. Загалом, кількісні змінні показують широкий діапазон значень і можуть використовуватися для моделювання звільнення співробітників без значних перетворень, за винятком тих змінних, де були виявлені викиди.

Категорійні ознаки було поділено на кілька основних груп:

- соціально–демографічні характеристики: стать працівника; сімейний стан; кількість утриманців;
- професійні характеристики: відділ або посада працівника; рівень посади; рівень освіти; кількість підвищень; рівень визнання співробітника; рейтинг продуктивності працівника;
- робочі умови та навантаження: чи працює працівник віддалено; чи працює працівник понаднормово; можливості лідерства; можливості для інновацій; задоволеність роботою; баланс між роботою та особистим життям;
- організаційні фактори: розмір компанії; сприйняття працівником репутації компанії.

Аналіз соціально–демографічних характеристик працівників показав, що серед працівників більше чоловіків (54,8%), ніж жінок (45,2%). Даний розподіл вказує на певний дисбаланс за статтю, що може бути пов'язано з особливостями галузі або політикою найму. Для подальшого аналізу важливо перевірити, чи впливає стать на ймовірність звільнення, та чи існують гендерні відмінності в показниках лояльності та задоволення роботою (рис. 2.6).

Вибірка включає переважно одружених працівників (50,2%), що може впливати на стабільність та зобов'язання перед компанією. Розлучені працівники становлять найменшу групу (14,9%). Ці дані слід використовувати для подальшого аналізу залежності звільнення від сімейного стану, оскільки

працівники у шлюбі можуть мати нижчу ймовірність звільнення через фінансові та сімейні зобов'язання. Працівники, які не перебувають у шлюбі, навпаки, можуть виявляти вищу мобільність на ринку праці.

Більшість працівників мають не більше двох утриманців. Наявність утриманців може впливати на стабільність працевлаштування, оскільки працівники з більшою кількістю утриманців можуть мати більші фінансові зобов'язання, що може призвести до підвищеної лояльності до компанії.

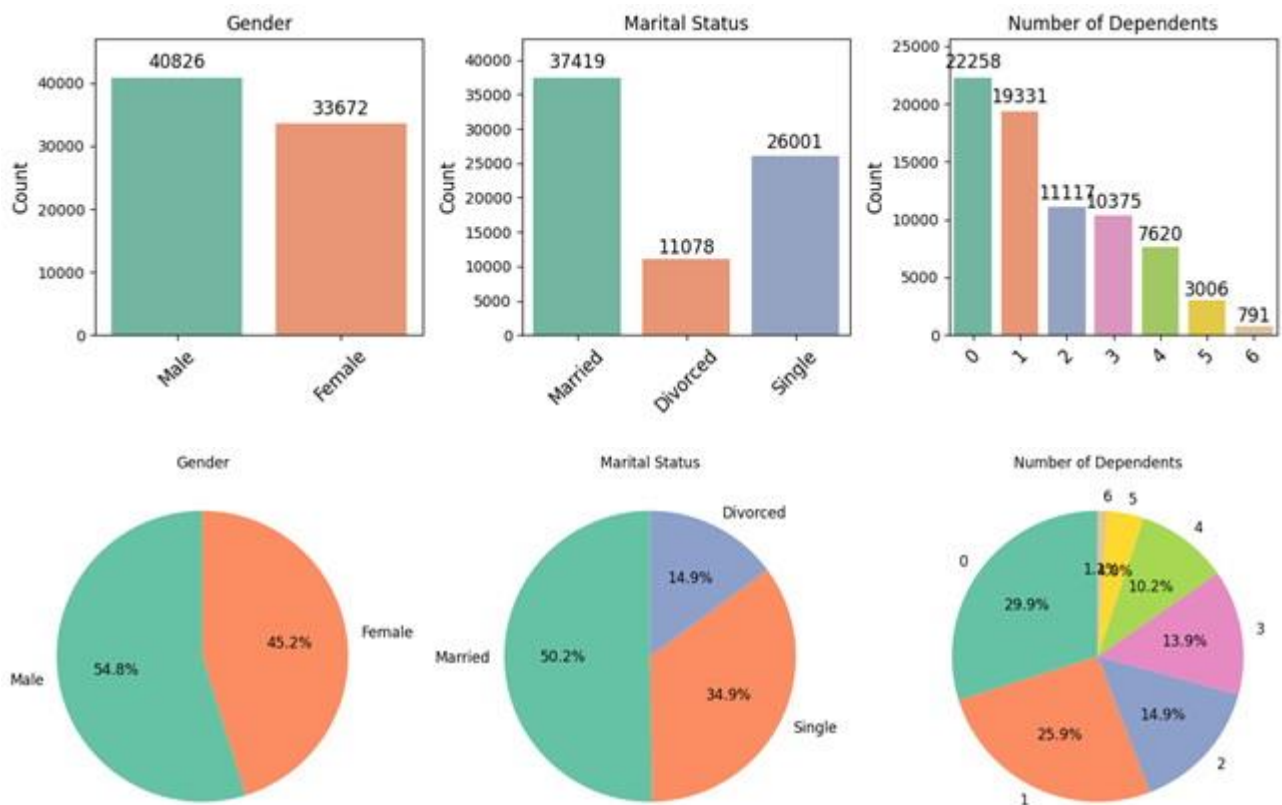


Рисунок 2.6 – Аналіз соціально–демографічних характеристик працівників

Важливим для розуміння кар'єрного шляху співробітників є аналіз їх професійних характеристик, досягнень та позиції в компанії (рис. 2.7).

Розподіл за рівнем освіти показує, що більшість працівників мають вищу освіту: 25,0% молодший спеціаліст, 30,0% з рівнем бакалавр та 20,2% магістр. Менші групи складають працівники з середньою освітою (19,7%) та з науковим ступенем (5.1%). Високий рівень освіти серед працівників вказує на те, що компанія має кваліфікований персонал.

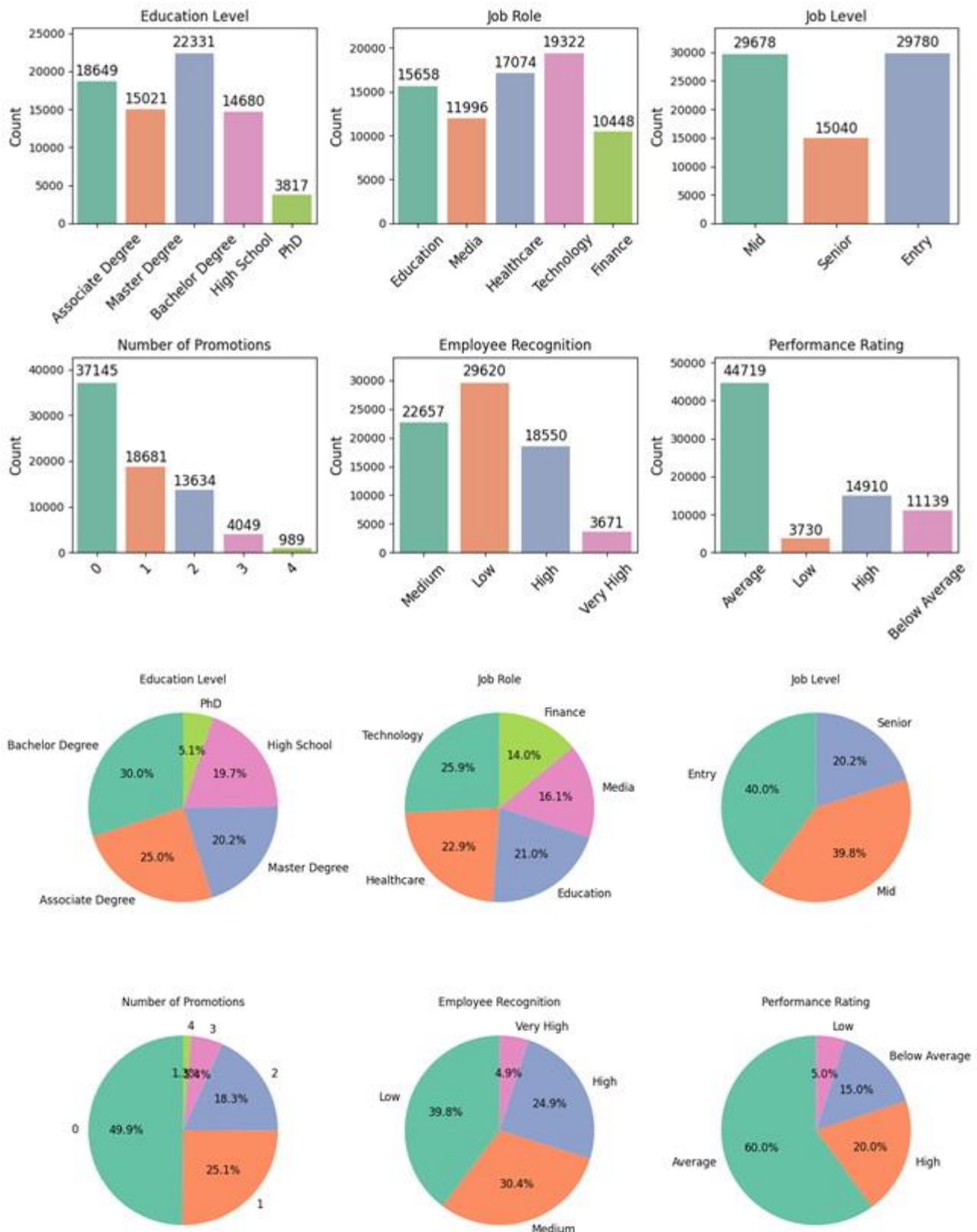


Рисунок 2.7 – Аналіз професійних характеристик працівників

Серед основних посад найбільша частка працівників зайнята у сферах технологій (25,9%), охорони здоров'я (22,9%) та освіти (21%). Менші групи

представлені працівниками у фінансах (14,0%) та у засобах масової інформації (16,1%). Розподіл за посадами вказує на галузеву специфіку компаній, де найбільше значення мають технологічні та освітні професії.

Працівники розподіляються між трьома основними рівнями посади: найбільша частка знаходиться на початковому (40%) та середньому рівнях (39,8%). Вищий рівень займає лише 20,2% працівників. Такий розподіл свідчить про те, що більшість працівників перебувають на нижніх кар'єрних сходинках, що може свідчити про обмежені можливості кар'єрного зростання.

Половина працівників (49,9%) не отримувала жодного підвищення, 25,1% отримали одне підвищення, і лише 4,8% працівників отримали три або більше підвищень. Відсутність кар'єрних змін може бути одним з основних факторів, що сприяють звільнення, особливо для тих працівників, які прагнуть кар'єрного розвитку.

Близько 39,8% працівників мають низький рівень визнання, тоді як тільки 4,9% мають дуже високий рівень визнання. Висока частка працівників із низьким рівнем визнання може бути сигналом про недостатню мотиваційну політику компанії. Недооціненість досягнень працівників може призводити до зниження їх мотивації та підвищення ймовірності звільнення.

Більшість працівників мають середню оцінку продуктивності (60%). Високу оцінку мають 20% працівників, тоді як 15% – нижчу від середньої. Найменшу групу становлять працівники з низькою оцінкою – 5%. Переважання середніх оцінок може вказувати на стандартизацію процесів оцінювання, яка не враховує індивідуальні досягнення. Працівники з нижчими оцінками можуть бути менш мотивованими та схильними до звільнення.

Аналіз робочих умов та навантаження дозволяє виявити фактори, які можуть сприяти утриманню або, навпаки, збільшувати ризик звільнення працівників. Дослідження показало, що лише 19,1% працівників мають можливість працювати віддалено (рис. 2.8), що може вказувати на обмеженість можливостей для дистанційної роботи. Це може знижувати гнучкість працівників та впливати на їх задоволеність роботою, особливо в умовах підвищеного попиту на гнучкі робочі графіки.

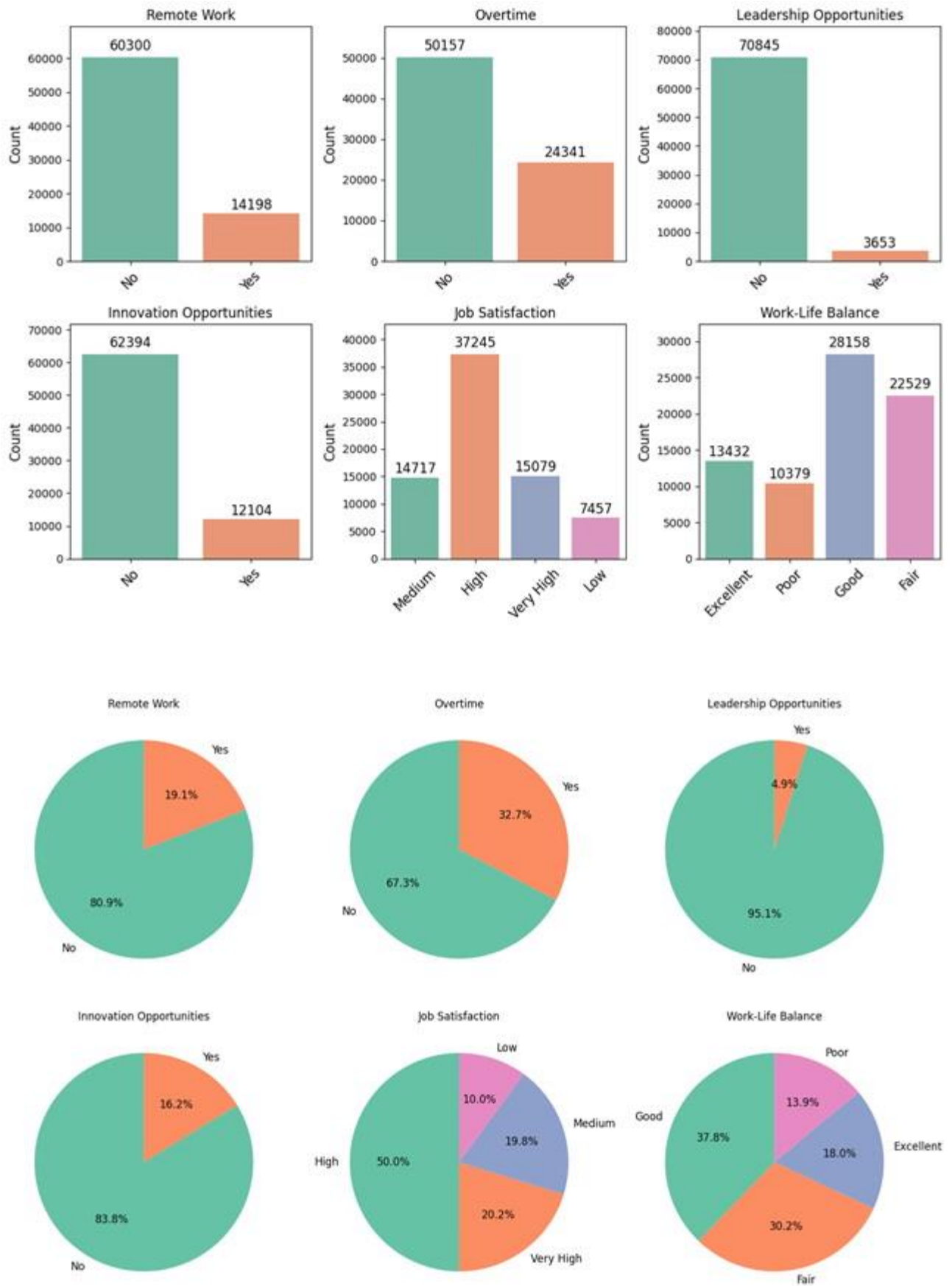


Рисунок 2.8 – Аналіз робочих умов та навантаження

Значна частка працівників (32,7%) працюють понаднормово. Наявність понаднормової роботи може свідчити про високе навантаження та вимоги до працівників, що може призводити до професійного вигорання. Понаднормова робота може бути фактором ризику звільнення, оскільки високі навантаження часто призводять до зниження задоволеності та підвищення рівня стресу серед працівників.

Лише 4,9% працівників мають лідерські можливості. Обмеженість у лідерських можливостях може призводити до зниження мотивації та обмеження кар'єрного розвитку, що негативно впливає на довгострокове утримання працівників у компанії.

Більшість працівників (83,8%) не мають можливостей для інновацій, що може призводити до зниження креативності та відчуття професійного розвитку. Це може бути фактором, що впливає на звільнення серед тих працівників, які прагнуть брати участь у нових проєктах та ініціативах.

Половина працівників (50,0%) висловлюють високий рівень задоволеності роботою, тоді як 10,0% – мають низький рівень. Задоволеність роботою є важливим показником утримання працівників. Високий рівень задоволення позитивно впливає на лояльність, тоді як низька задоволеність може бути основним фактором ризику звільнення.

Дослідження балансу між роботою та особистим життям показує, що найбільша група працівників (37,8%) вважає свій баланс між роботою та особистим життям задовільним. Інші групи розподіляються наступним чином: 30,2% вважають баланс прийнятним, 18,0% – чудовим, а 13,9% мають низький рівень балансу. Баланс між роботою та життям є одним із ключових факторів утримання працівників. Низький рівень балансу часто асоціюється з високим рівнем стресу та невдоволенням, що підвищує ризик звільнення.

Організаційні фактори, такі як розмір компанії та її репутація, відіграють важливу роль у визначенні задоволеності працівників та їх лояльності до організації. Ці змінні характеризують внутрішнє середовище компанії, її стабільність та привабливість для працівників. Аналіз цих змінних дозволяє зрозуміти їх вплив на утримання працівників (рис. 2.9).



Рисунок 2.9 – Аналіз організаційних факторів

Співробітники середніх компаній складають найбільшу групу (50,0%), що вказує на їх популярність серед працівників. Їх розмір дозволяє поєднувати стабільність із можливостями для розвитку.

Найбільша частка працівників оцінює репутацію своєї компанії як хорошу – 49,9% вибірки, поганою вважають 20,3% працівників, тоді як 19,8% – середньою. Лише 10,0% працівників оцінюють репутацію компанії як відмінну. Переважання позитивних оцінок репутації вказує на високий рівень задоволеності працівників іміджем своїх компаній. Однак варто звернути увагу на те, що лише 10% працівників вважають свою компанію дійсно відмінною, що може свідчити про наявність сфер для вдосконалення. Працівники, які негативно оцінюють репутацію компанії, можуть бути більш схильними до звільнення, оскільки імідж компанії є важливим фактором, що впливає на мотивацію та задоволеність працівників.

2.2 Дослідження впливу факторів на ймовірність звільнення

Наступним етапом є дослідження впливу характеристик на цільову змінну (Attrition) з метою визначення ключових змінних, що мають статистично значущий вплив на звільнення працівників.

Аналіз розподілу кількісних змінних за групами працівників дозволяє виявити відмінності та ідентифікувати ключові фактори, які можуть впливати на рішення працівників залишитися або звільнитися (рис. 2.10).

Розподіл віку працівників між двома групами має деякі відмінності. Працівники, які залишилися, демонструють більш рівномірний розподіл у всіх вікових категоріях. Натомість для працівників, які звільнилися, спостерігається певний пік у віковій категорії 20–30 років та спад у вікових категоріях після 50 років. Це може свідчити про те, що молоді працівники (20–30 років) більш схильні до звільнення, ймовірно через пошук кращих кар'єрних можливостей або невдоволеність поточними умовами роботи. Працівники старшого віку (50+ років) частіше залишаються у компанії, що вказує на їх лояльність або обмежені можливості для зміни роботи в цьому віці.

Розподіл стажу роботи у компанії показує, що працівники, які залишаються, мають значно більший стаж у компанії. Найбільша щільність для тих, хто залишився, спостерігається в діапазоні 10–20 років, тоді як для тих, хто звільнився, – у діапазоні до 10 років. Працівники з меншим стажем (менше 10 років) є більш схильними до звільнення, що може свідчити про недостатню адаптацію або відсутність кар'єрних перспектив. Працівники зі стажем понад 20 років мають низьку ймовірність звільнення, що вказує на лояльність та стабільність у роботі.

Розподіл тривалості існування компанії є майже однаковим для обох груп. Це свідчить про те, що цей фактор не є суттєвим детермінантом звільнення. Основна щільність розподілу знаходиться в діапазоні 50–80 років, що вказує на те, що більшість компаній у вибірці мають багаторічну історію. Відсутність відмінностей між групами вказує на те, що тривалість існування компанії не є основним фактором, який впливає на рішення працівників звільнитися.

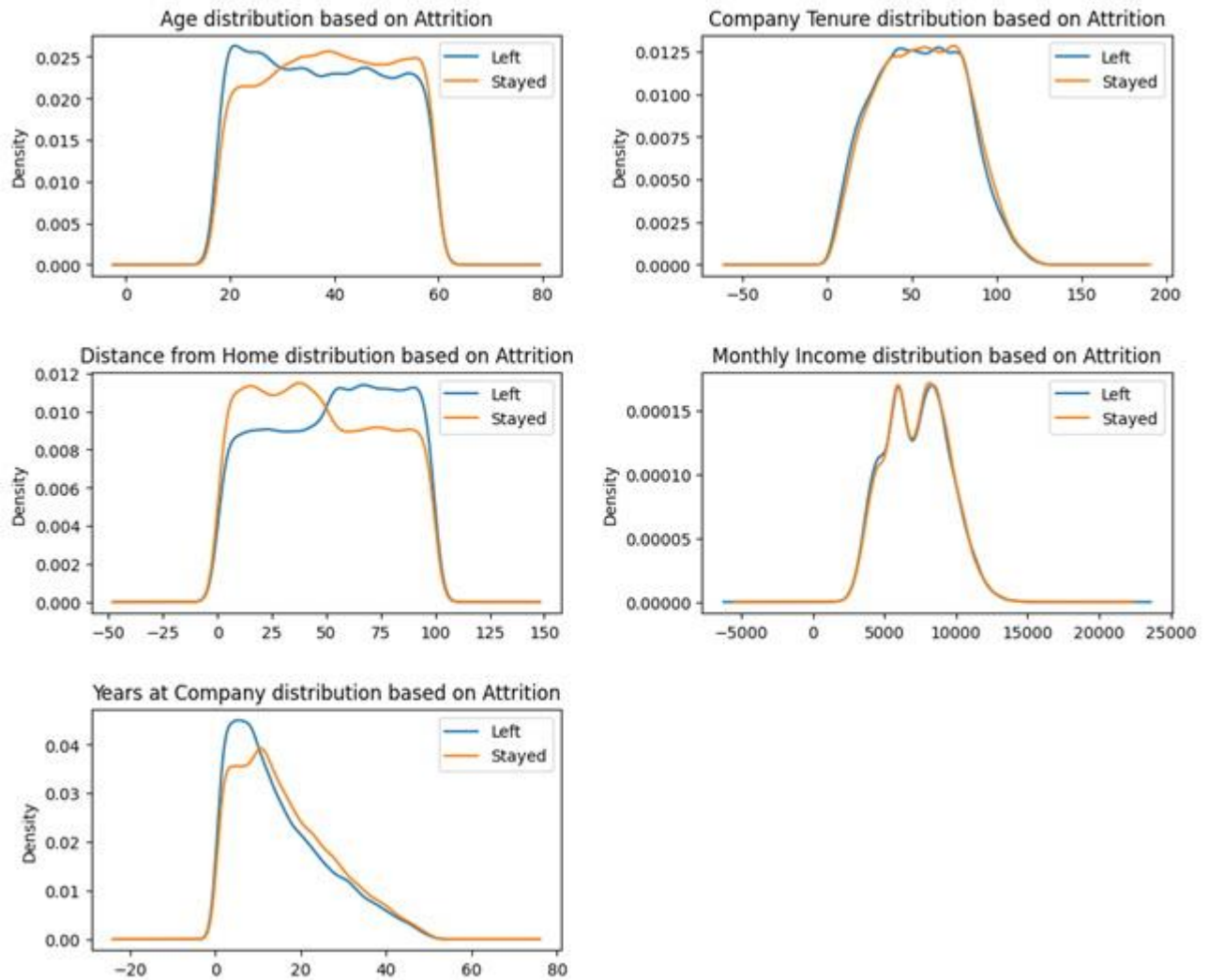


Рисунок 2.10 – Аналіз розподілу кількісних змінних за групами звільнення

Розподіл щомісячного доходу для обох груп дуже схожий, із піками у межах 5 000–10 000 \$. Немає суттєвих відмінностей між тими, хто залишився, та тими, хто звільнився. Це свідчить про те, що рівень заробітної плати не є основним фактором, що впливає на звільнення. Проте, для більш детального аналізу доцільно розглянути зв'язок доходу з іншими факторами, такими як задоволеність роботою або рівень визнання.

Дослідження впливу відстані до роботи демонструє, що працівники, які залишилися, демонструють рівномірний розподіл у всіх категоріях відстані. Натомість у працівників, які звільнилися, спостерігається дві щільності на початку та кінці діапазону, що вказує на тенденцію до звільнення як тих, хто живе близько, так і тих, хто живе дуже далеко від роботи. Відсутність чіткої

тенденції свідчить про те, що відстань до роботи може впливати на звільнення лише у поєднанні з іншими факторами, такими як робоче навантаження або гнучкість робочого графіка.

Таким чином вік та стаж у компанії є ключовими детермінантами звільнення. Молодші працівники та працівники з малим стажем більш схильні до звільнення, що вказує на необхідність додаткових заходів для утримання молодих фахівців та покращення адаптації нових співробітників. Щомісячний дохід та тривалість існування компанії не мають значного впливу на звільнення. Вплив відстані до роботи є фактором, який може бути суттєвим для окремих груп працівників залежно від їх робочого графіка та умов праці.

Розглянемо вплив соціально-демографічних характеристик на ймовірність звільнення працівників (рис. 2.11). За результатами аналізу жінки є більш схильними до звільнення (53,0%) порівняно з чоловіками. Це може бути пов'язано з тим, що жінки, ймовірно, частіше змінюють місце роботи через сімейні обов'язки або прагнення знайти гнучкіші умови праці.

Працівники, які перебувають у шлюбі, виявляють вищу стабільність і меншу схильність до звільнення (36%), ймовірно, через більші фінансові та сімейні зобов'язання. Натомість, співробітники, які не перебувають у шлюбі, можуть бути більш мобільними та відкритими до нових можливостей, що збільшує ймовірність їхнього звільнення (66.8%).

Працівники без утриманців та з 1–3 утриманнями демонструють майже однакові частки для обох класів. Зі збільшенням кількості утриманців (4 і більше) частка тих, хто залишився, знижується (лише 35–37% залишилися, решта – звільнилися). Працівники з невеликою кількістю утриманців (1–2) є більш стабільними та менш схильними до звільнення, що вказує на лояльність через фінансові зобов'язання. Однак працівники з більшою кількістю утриманців (4 і більше) можуть мати вищий ризик звільнення через необхідність знаходити кращі фінансові умови або більш гнучкий робочий графік.



Рисунок 2.11 – Вплив соціально-демографічних характеристик на ймовірність звільнення працівників

Аналіз впливу професійних чинників на ймовірність звільнення працівників (рис. 2.12) показав, що працівники з освітнім рівнем доктор філософії (PhD) мають найвищу стабільність: 75,6% залишилися в компанії. Вищий рівень освіти позитивно впливає на утримання працівників. Це може свідчити про те, що працівники з вищим ступенем мають більше можливостей для розвитку всередині компанії або більш стабільну кар'єру, що знижує ймовірність їх звільнення.

Працівники, зайняті у сфері фінансів та медіа, демонструють найвищу стабільність – 53% залишилися в компанії, тоді як працівники у сфері освіти, охорони здоров'я та технологій мають дещо нижчу стабільність (47–48% звільнилися). Висока стабільність у сфері фінансів може бути пов'язана з високими доходами та фінансовими зобов'язаннями, які стримують працівників від звільнення. Натомість, працівники у сфері технологій можуть бути більш мобільними на ринку праці, що призводить до підвищеного рівня звільнень.

Найвищу стабільність демонструють працівники на вищих посадах: 79,7% залишилися в компанії. Натомість працівники початкового рівня мають найвищий рівень звільнень – 63,2% звільнилися. Працівники на початкових посадах є більш схильними до звільнення, що може бути пов'язано з обмеженими кар'єрними можливостями. Старші працівники мають вищу

стабільність завдяки своїм досягненням, досвіду та більшій задоволеності кар'єрою.

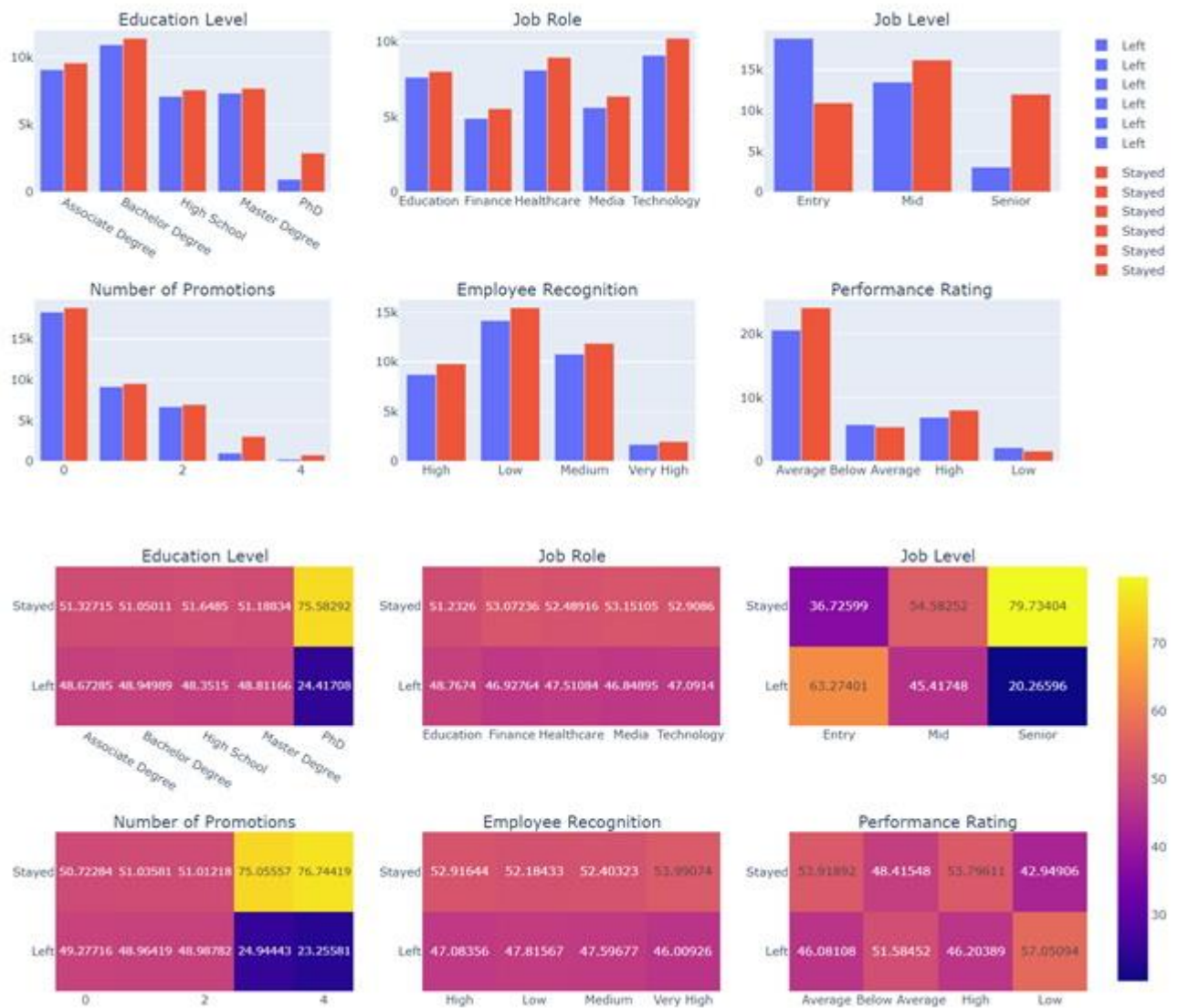


Рисунок 2.12 – Вплив професійних чинників на ймовірність звільнення

Працівники, які не отримували підвищень (0), мають майже рівні частки звільнених та тих, хто залишився. Працівники, які отримали хоча б одне підвищення, мають дещо вищу стабільність, зі збільшенням кількості підвищень до 3 – 4 стабільність найвища (75% залишилися). Наявність підвищень позитивно впливає на утримання працівників, оскільки вони відчувають перспективи кар'єрного росту. Відсутність підвищень є фактором ризику звільнення, особливо для молодих фахівців.

Працівники з високим та дуже високим рівнем визнання частіше залишаються у компанії – 52,9% та 54,0% відповідно. Працівники з низьким рівнем визнання мають вищу схильність до звільнення – 47,8% звільнилися. Відсутність визнання є значним демотивуючим фактором, що призводить до зниження лояльності працівників. Працівники з низьким рівнем визнання менш задоволені та більш схильні до пошуку нових можливостей.

Працівники з низьким рейтингом продуктивності демонструють найвищий рівень звільнення – 57,0% звільнилися. Низький рейтинг продуктивності може призводити до звільнення через невідповідність очікуванням компанії. Високий рейтинг продуктивності, навпаки, позитивно впливає на стабільність, проте його вплив може залежати від системи винагороди та визнання у компанії.

Таким чином, вищий рівень освіти та більша кількість підвищень позитивно впливають на утримання працівників. Це свідчить про необхідність інвестування у розвиток співробітників та створення умов для їхнього професійного росту. Посадовий рівень є значущим фактором: працівники на початкових позиціях є найбільш схильними до звільнення, що вимагає додаткових заходів для забезпечення мотивації та підтримки. Відсутність визнання та низький рейтинг продуктивності є основними демотивуючими чинниками, що призводять до звільнення.

Аналіз робочих умов показав, що віддалена робота позитивно впливає на утримання працівників, що може бути пов'язано із гнучкістю графіка та можливістю балансу між роботою та особистим життям (рис. 2.13). Відсутність можливості працювати віддалено може бути фактором ризику, особливо для молодих фахівців, які цінують гнучкі умови праці.

Працівники, які не працюють понаднормово, частіше залишаються в компанії – 45,5% звільнених. Працівники, які працюють понаднормово, мають майже однаковий рівень звільнення. Понаднормова робота не є основним фактором ризику, але може підвищувати стрес і впливати на мотивацію працівників. Залежно від поєднання з іншими факторами, цей аспект може як підвищувати, так і знижувати ризик звільнення.

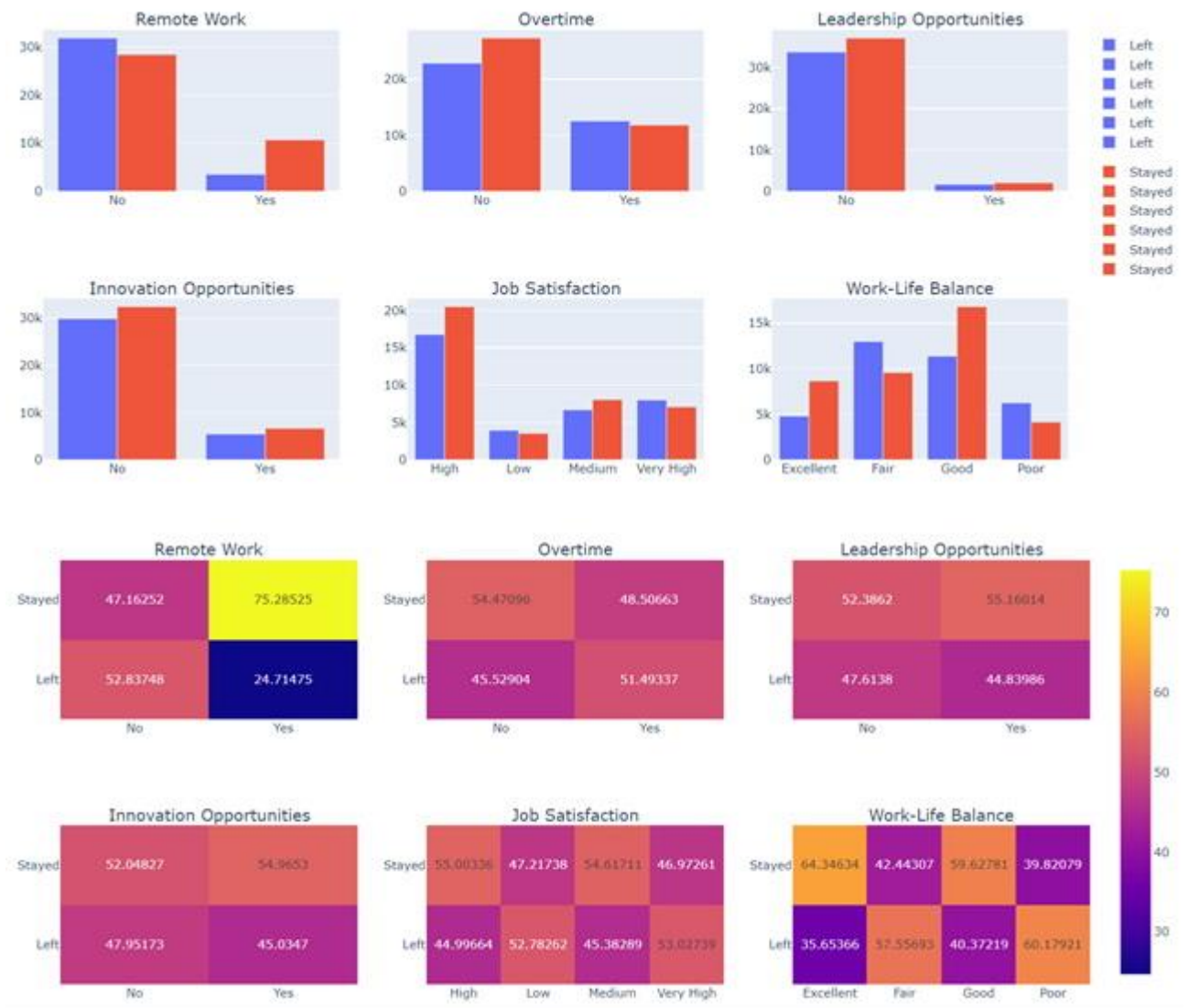


Рисунок 2.13 – Вплив робочих умов на ймовірність звільнення

Лише 4,9% працівників мають лідерські можливості, і серед них 55,2% залишилися в компанії. Відсутність лідерських можливостей є фактором ризику для працівників, які прагнуть кар'єрного розвитку. Працівники без таких можливостей можуть відчувати обмежені перспективи та шукати можливості в інших компаніях.

Працівники, які мають можливості для інновацій, частіше залишаються – 55,0%. Працівники без інноваційних можливостей демонструють майже однаковий рівень звільнення. Можливості для інновацій є мотиваційним фактором, що сприяє утриманню працівників. Відсутність таких можливостей може знижувати креативність та відчуття задоволення роботою, що підвищує ймовірність звільнення.

Найбільша частка звільнених спостерігається серед працівників із низьким рівнем задоволеності – 53,0% звільнилися та дуже високим рівнем – 53,0%. Низька задоволеність прямо впливає на ймовірність звільнення, тому покращення умов роботи та мотиваційних програм може знизити цей ризик.

Працівники, які оцінюють свій баланс між роботою та особистим життям як "відмінний", частіше залишаються. Працівники, які вважають свій баланс "поганим", демонструють найвищий рівень звільнення – 60,2% звільнилися. Поганий баланс між роботою та особистим життям є основним фактором ризику звільнення. Забезпечення гнучкості робочого графіка та можливості регулювати робочі навантаження може суттєво знизити ризик звільнення.

Отже, віддалена робота та можливості для інновацій позитивно впливають на утримання працівників. Це вказує на важливість створення гнучких умов праці та надання можливостей для розвитку. Відсутність лідерських можливостей та низький рівень задоволеності роботою є основними факторами ризику. Важливо створювати умови для професійного розвитку та визнання досягнень. Понаднормова робота та поганий баланс між роботою та життям підвищують ймовірність звільнення, що потребує перегляду політики робочих навантажень та управління часом.

Аналіз організаційних факторів є важливим для розуміння, як сприйняття організації та її структура можуть впливати на ймовірність звільнення працівників (рис. 2.14).

Працівники великих компаній демонструють дещо вищу стабільність: 53,4% залишилися. У середніх компаніях частка тих, хто залишився, майже рівна частці тих, хто звільнився. Натомість у малих компаніях спостерігається вищий рівень звільнень. Працівники великих компаній мають більш стабільну зайнятість, що може бути пов'язано з наявністю кращих можливостей для розвитку та більшою фінансовою стабільністю. Вищий рівень звільнень в малих компаніях може бути обумовлений обмеженими можливостями для кар'єрного росту та менш стабільним середовищем.



Рисунок 2.14 – Вплив організаційних факторів на ймовірність звільнення

Працівники, які оцінюють репутацію своєї компанії як "відмінну", демонструють високий рівень стабільності – 56,0% залишилися. Натомість працівники компаній із "поганою" репутацією частіше звільняються. Працівники з позитивним сприйняттям репутації компанії мають вищий рівень лояльності, тоді як погана репутація підвищує ймовірність звільнення. Це може бути обумовлено тим, що репутація компанії впливає на загальне задоволення роботою та рівень мотивації працівників.

Таким чином, розмір компанії є важливим фактором для утримання працівників. Великі компанії мають перевагу у стабільності зайнятості, тоді як малі компанії демонструють підвищений рівень звільнень, що може свідчити про обмежені ресурси для підтримки працівників. Репутація компанії є вагомим фактором для утримання працівників. Погана репутація знижує лояльність та підвищує ймовірність звільнення, тому покращення іміджу компанії має бути важливим завданням для управління.

Кореляційний аналіз факторів, що впливають на звільнення працівників, дозволяє виявити, які характеристики є найбільш значущими для утримання або звільнення працівників (рис. 2.15).

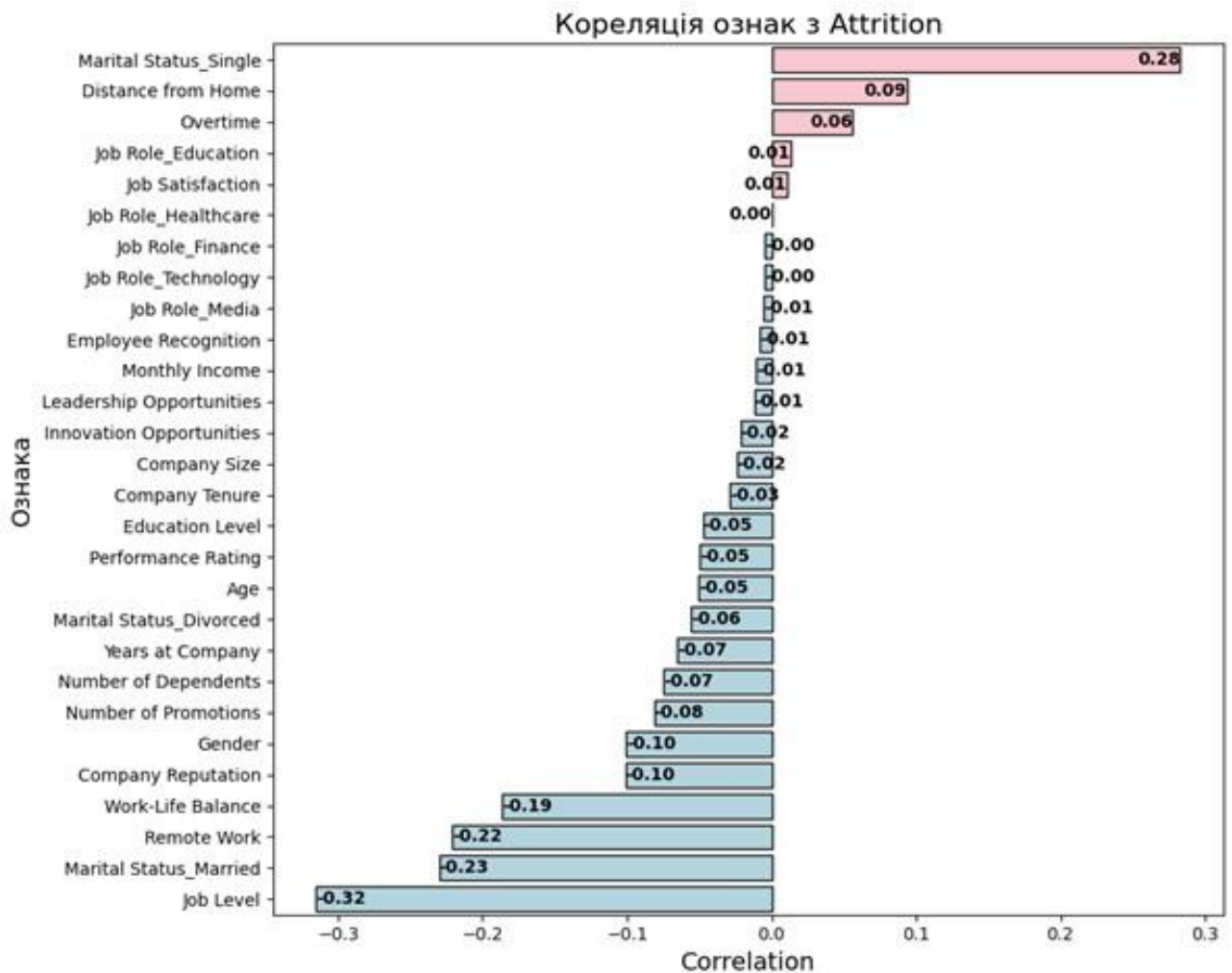


Рисунок 2.15 – Кореляція факторів з цільовою ознакою

Аналіз кореляційних зв'язків з цільовою змінною (Attrition) виявив різні за силою та напрямом впливу групи характеристик працівників.

Найсильніший вплив на звільнення демонструють такі фактори:

- рівень посади (Job Level) має найсильнішу негативну кореляцію ($-0,32$), що вказує на те, що працівники на вищих посадах менш схильні до звільнення;
- сімейний стан має значний вплив, при цьому неодружені працівники (Marital Status_Single) демонструють найсильнішу позитивну кореляцію ($0,28$), тоді як одружені (Marital Status_Married) – значну негативну ($-0,23$), що свідчить про більшу стабільність одружених працівників;
- можливість віддаленої роботи (Remote Work) має помітну негативну кореляцію ($-0,22$), вказуючи на те, що наявність такої можливості знижує ймовірність звільнення;

– баланс між роботою та особистим життям (Work–Life Balance) також має суттєву негативну кореляцію ($-0,19$), підтверджуючи важливість цього фактору для утримання працівників.

Помірний вплив на звільнення мають:

– репутація компанії (Company Reputation) та стать працівника (Gender) мають однакову негативну кореляцію ($-0,10$);

– кількість підвищень (Number of Promotions), кількість утриманців (Number of Dependents) та стаж у компанії (Years at Company) показують слабку негативну кореляцію (від $-0,07$ до $-0,08$);

– відстань до роботи (Distance from Home) та понаднормова робота (Overtime) показують незначну позитивну кореляцію ($0,09$ та $0,06$ відповідно);

– вік (Age), рівень освіти (Education Level), рейтинг продуктивності (Performance Rating) демонструють слабку негативну кореляцію ($-0,05$).

Інші ознаки демонструють мінімальний вплив на звільнення.

Такий розподіл кореляцій свідчить про те, що найбільший вплив на рішення про звільнення мають фактори, пов'язані з професійним статусом, особистим життям та умовами роботи, тоді як фінансові показники та специфіка професійної ролі мають менше значення. Це важливо враховувати при розробці стратегій утримання персоналу та при побудові прогностичних моделей.

2.3 Аналіз взаємозв'язків між змінними

Кореляційний аналіз змінних є важливим етапом дослідження, оскільки високий рівень взаємозв'язку між ознаками може призвести до мультиколінеарності у моделях прогнозування. В такому випадку змінні надають подібну інформацію, що може знижувати точність моделі та негативно впливати на роботу деяких моделей (наприклад, лінійних моделей). Для моделей, чутливих до мультиколінеарності (наприклад, лінійна регресія), варто обрати змінні з низькою взаємозалежністю. Однак, алгоритми на основі дерев (наприклад, Random Forest, XGBoost) можуть використовувати всі змінні без ризику деградації якості моделі.

Аналіз кореляційної матриці Пірсона показує, що більшість ознак мають слабку або відсутню кореляцію між собою (рис. 2.16). Це означає, що ризик мультиколінеарності є низьким для більшості змінних, і вони можуть бути використані в моделях машинного навчання без необхідності виключення чи об'єднання. Вік працівника та стаж роботи у компанії мають середню позитивну кореляцію (0,5), що вказує на те, що чим старший працівник, тим більше часу він працює у компанії. Стаж роботи працівника у компанії має помірний позитивний зв'язок із загальною тривалістю існування компанії (0,4). Кореляція між віком та стажем роботи в компанії є очікуваною, оскільки збільшення віку зазвичай супроводжується збільшенням тривалості роботи. Це може бути використано для створення похідних змінних, які будуть більш інформативними для моделей прогнозування.

Дослідження показало, що є кілька груп ключових взаємозв'язків (рис. 2.17).

Група 1 включає такі ознаки, як "вік", "досвід роботи в компанії" та "роки на посаді". Кореляція між цими показниками вказує на їх спільний внесок у формування профілю досвіду працівника, що робить доцільним застосування методу головних компонент (PCA) для їх об'єднання в інтегральну ознаку, яку можна позначити як "Experience_Score". Це дозволяє зменшити кількість змінних у моделі, зберігаючи суттєву інформацію щодо досвіду працівника.

Друга група містить ознаки, пов'язані з посадою та доходом. З огляду на те, що окремі посади можуть впливати на рівень доходу в різних галузях, доцільно ввести нові змінні, які відображають співвідношення доходу до конкретної посади у відповідних галузях. Це дозволяє відобразити відмінності в доходах серед різних посад та забезпечити більш гнучкий підхід до оцінки ролі цих змінних у моделі.

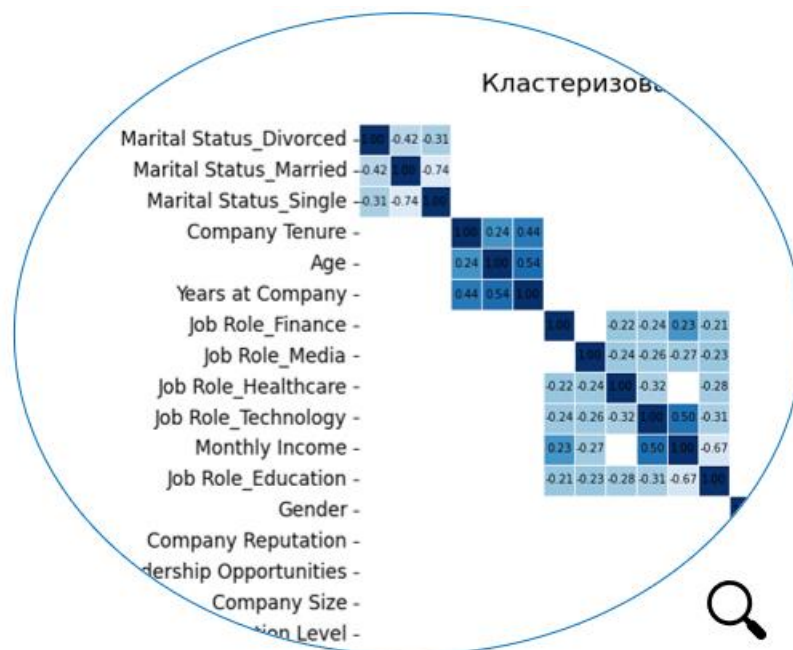
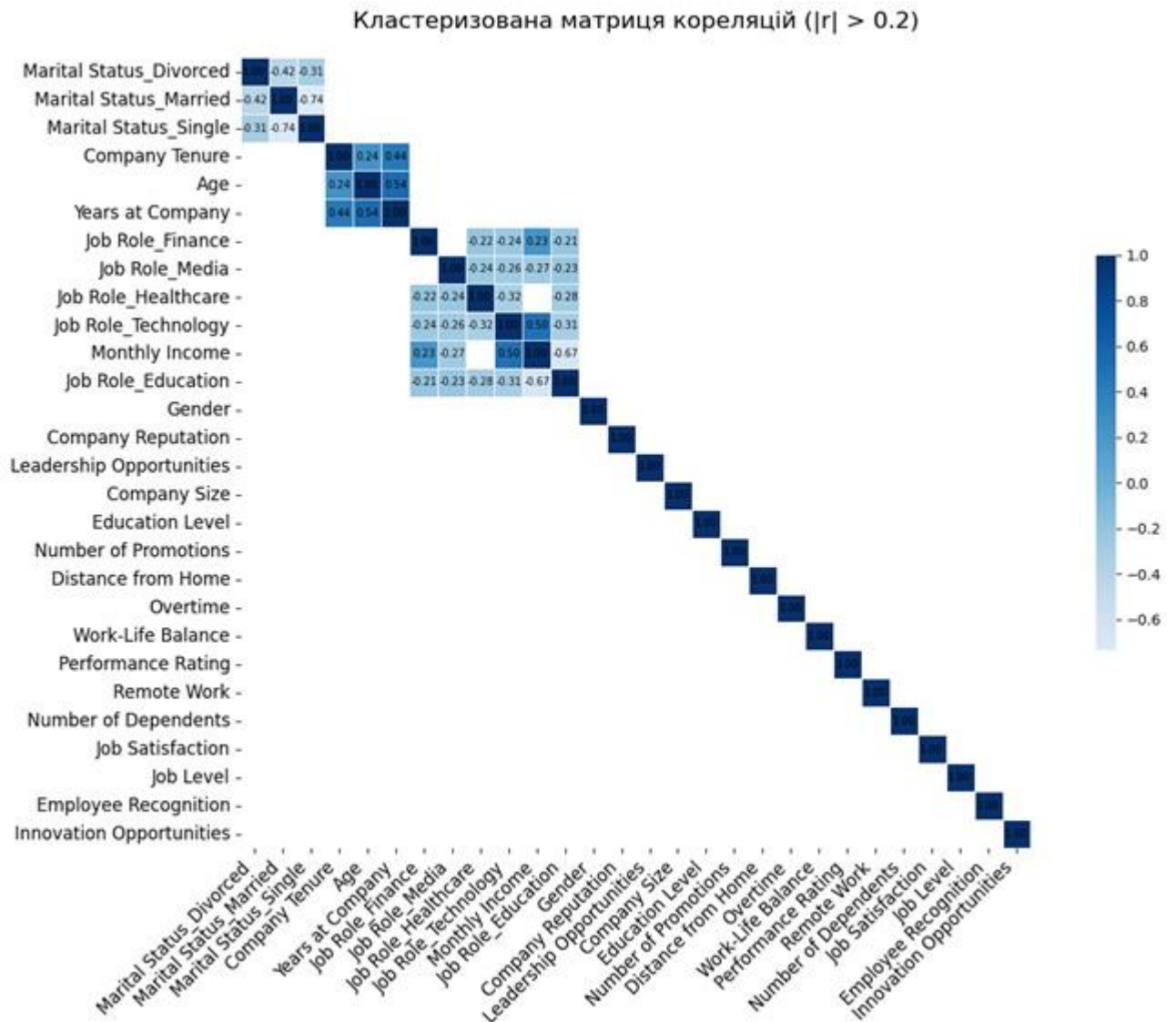


Рисунок 2.16 – Кореляційна матриця Пірсона

Найбільш корельовані пари ознак:

	Feature 1	Feature 2	Correlation
8	Marital Status_Married	Marital Status_Single	-0.735563
2	Monthly Income	Job Role_Education	-0.671651
0	Age	Years at Company	0.537318
3	Monthly Income	Job Role_Technology	0.497799
1	Company Tenure	Years at Company	0.442580
6	Marital Status_Divorced	Marital Status_Married	-0.419855
5	Job Role_Healthcare	Job Role_Technology	-0.322680
7	Marital Status_Divorced	Marital Status_Single	-0.306024
4	Job Role_Education	Job Role_Technology	-0.305269

Групи корельованих ознак:

Група 1:

['Age', 'Company Tenure', 'Years at Company']

Група 2:

['Job Role_Education', 'Job Role_Technology', 'Monthly Income']

Група 3:

['Marital Status_Divorced', 'Marital Status_Single', 'Marital Status_Married']

Рисунок 2.17 – Найбільш корельовані ознаки

Третя група об'єднує показники сімейного стану. Оскільки кореляції між різними статусами вказують на взаємозв'язок цих категорій, пропонується зберегти тільки індикатор для статусу "неодружений", що зумовлено його найбільш значущим кореляційним зв'язком із показником "Attrition" та вказує на підвищену ймовірність звільнення серед співробітників, які не перебувають у шлюбі (рис. 2.15). Таким чином, включення індикатора "неодружений" до моделі дозволяє більш точно передбачати ризик звільнення. Інші категорії сімейного стану, хоч і корелюють між собою, мають меншу силу кореляції з Attrition, що робить їх менш значущими для прогнозування.

Такий підхід спрощує структуру даних, забезпечуючи при цьому достатній рівень деталізації для моделі прогнозування.

Для визначення характеру розподілу даних та виявлення прихованих патернів у поведінці працівників щодо звільнення було застосовано метод T-SNE (T-Distributed Stochastic Neighbor Embedding). Цей метод дозволяє ефективно візуалізувати багатовимірні дані у двовимірному просторі, зберігаючи при цьому топологічну структуру даних та відносні відстані між точками. T-SNE особливо ефективний для візуалізації кластерів у даних,

оскільки він намагається зберегти локальну структуру, розміщуючи близькі точки в оригінальному просторі поряд у проекції. Це дозволяє виявити природні групування в даних та взаємозв'язки між різними характеристиками працівників.

Візуалізація результатів T–SNE демонструє складну кластерну структуру даних, де кожна точка представляє окремого працівника, а кольорове кодування відображає статус працівника: фіолетовим позначено працівників, які залишились в компанії, жовтим – тих, хто звільнився. На графіку можна спостерігати формування приблизно 20–25 окремих кластерів різного розміру та щільності (рис. 2.18).

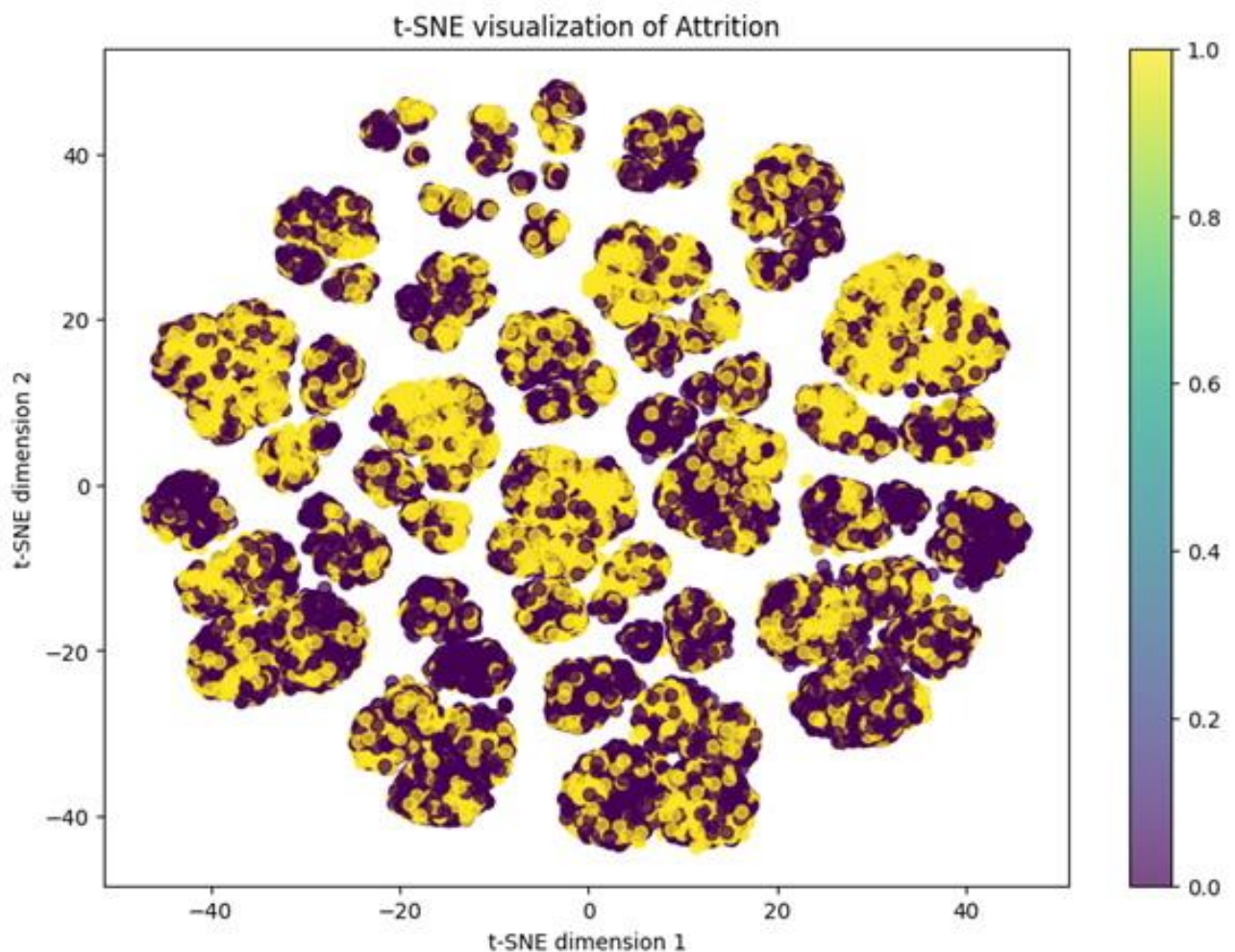


Рисунок 2.18 – Результати кластерного аналізу за методом T–SNE

Аналіз отриманої візуалізації виявляє кілька важливих особливостей:

- спостерігається значне перекриття класів, у більшості кластерів спостерігається змішування фіолетових та жовтих точок, що вказує на суттєве

перекриття характеристик між працівниками, які звільняються, та тими, хто залишається, що свідчить про складність задачі прогнозування;

- присутня неоднорідність розподілу, на графіку присутні як щільні, так і розріджені області, що вказує на існування як чітко визначених груп працівників зі схожими характеристиками, так і перехідних станів між цими групами;

- наявні як великі кластери, так і менші групи точок, що може відображати різні професійні та соціальні групи працівників;

- кластери розташовані нерівномірно у просторі ознак, формуючи складну топологічну структуру, що свідчить про багатофакторність впливу на рішення про звільнення.

Наявність як щільних, так і розріджених областей у візуалізації вказує на нерівномірний розподіл працівників у просторі ознак. Деякі кластери демонструють більшу гомогенність щодо рішення про звільнення, тоді як інші характеризуються значною гетерогенністю. Така структура даних свідчить про складність задачі прогнозування звільнень.

2.4 Висновки за розділом 2

Проведений аналіз набору даних показав його цілісність та відсутність пропущених значень. У наборі даних представлено характеристики працівників, що охоплюють соціально-демографічні, професійні, організаційні фактори та показники задоволеності роботою. Розподіл цільової змінної є відносно збалансованим, що створює сприятливі умови для побудови моделей прогнозування.

Виявлено найбільш значущі фактори впливу на звільнення: рівень посади ($-0,32$), сімейний стан ($0,28$ для неодружених), можливість віддаленої роботи ($-0,22$) та баланс між роботою і особистим життям ($-0,19$). При цьому фінансові показники та специфіка професії демонструють мінімальний вплив на рішення про звільнення, що важливо враховувати при розробці стратегій утримання персоналу.

Кореляційний аналіз встановив помірні взаємозв'язки між деякими характеристиками працівників. Для подальшого моделювання доцільно застосувати РСА для створення інтегральної ознаки "Experience_Score" для ознак "вік", "досвід роботи в компанії" та "роки на посаді", додати змінні, що відображають співвідношення доходу до посади. Також варто зберегти індикатор "неодружений" для підвищення точності прогнозування ризику звільнення.

3 ВИБІР МОДЕЛІ КЛАСИФІКАЦІЇ ТА МЕТОДІВ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ

3.1 Обґрунтування вибору моделі машинного навчання

Для вирішення основної задачі дослідження – розробки системи, здатної ефективно і точно прогнозувати ймовірність звільнення на основі наявних даних, було проведено комплексний аналіз методів машинного навчання для бінарної класифікації. Враховуючи значний обсяг даних та різноманітність ознак, доцільно розглянути широкий спектр алгоритмів, від простих лінійних моделей до складних ансамблевих методів (рис. 3.1). Цей підхід дозволяє всебічно оцінити ефективність різних методів та обрати оптимальне рішення для задачі.

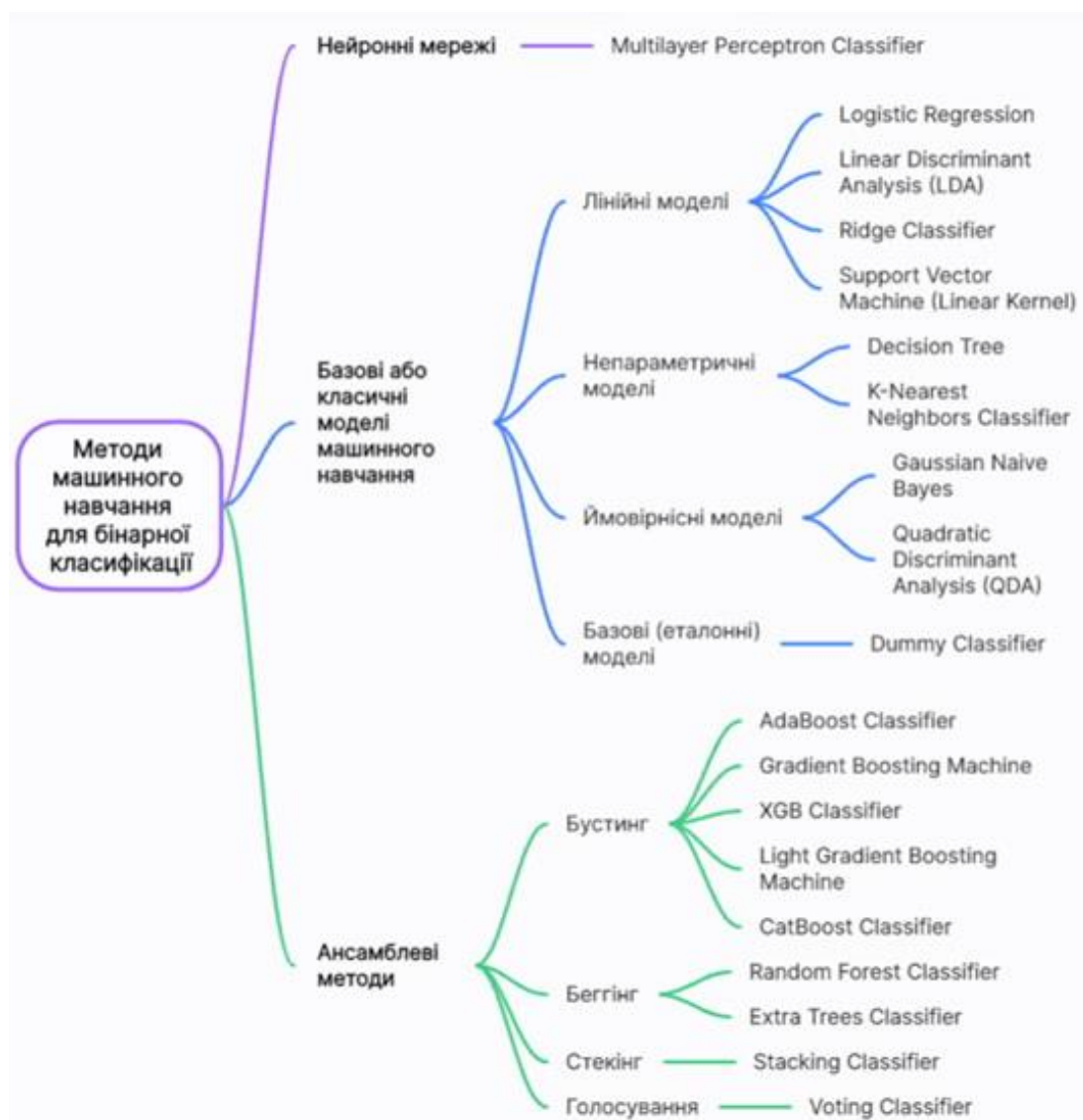


Рисунок 3.1 – Класифікація досліджених методів машинного навчання

Лінійні моделі, такі як логістична регресія, лінійний дискримінантний аналіз (LDA) та SVM з лінійним ядром, обрані через їх інтерпретованість та ефективність для лінійно розділених даних, що робить їх корисними для розуміння основних закономірностей у даних.

Непараметричні моделі, зокрема дерева рішень та k -найближчих сусідів, включені до дослідження завдяки їх здатності обробляти нелінійні залежності та різнотипні дані. Ці методи особливо корисні для виявлення складних взаємозв'язків між ознаками.

Ймовірнісні моделі, такі як наївний баєсівський класифікатор та квадратичний дискримінантний аналіз (QDA), обрані завдяки їх здатності працювати з імовірнісними розподілами та потенційної ефективності при наявності категоріальних ознак.

Особлива увага приділяється ансамблевим методам, які об'єднують результати кількох базових моделей для підвищення загальної ефективності класифікації. Бустингові алгоритми, такі як AdaBoost, Gradient Boosting та його варіації (XGBoost, LGBM, CatBoost), ітеративно покращують модель, фокусуючись на складних для класифікації прикладах. Беггінг, представлений Random Forest та Extra Trees класифікаторами, використовує ідею усереднення результатів багатьох незалежних моделей. Це дозволяє знизити варіативність прогнозів та підвищити стійкість моделі до перенавчання. Стекінг, як метод ансамблювання вищого рівня, дозволяє комбінувати прогнози різних моделей, потенційно використовуючи сильні сторони кожного алгоритму. Це може бути особливо корисним для складних задач, де жоден окремий алгоритм не дає оптимальних результатів.

Нейронні мережі, представлені MLP Classifier, включені до дослідження як потужний інструмент для виявлення складних нелінійних залежностей. Їх використання обґрунтоване великим обсягом даних та потенційною наявністю прихованих паттернів, які можуть бути не виявлені більш простими методами.

Кожен метод має свої переваги та обмеження, і їх порівнювальний аналіз допоможе визначити оптимальну стратегію моделювання для даної задачі. Експериментальне порівняння різних алгоритмів на конкретному наборі даних

дозволить визначити оптимальний підхід для прогнозування звільнення працівників, забезпечуючи баланс між точністю, ефективністю та інтерпретованістю результатів.

Для надійної оцінки продуктивності моделей машинного навчання та вибору найкращої з них була використана техніка крос-валідації, яка дозволяє розподілити наявні дані на кілька частин, де кожна частина по черзі використовується як тестова, а решта – для навчання моделі. Такий підхід забезпечує оцінку якості моделей на різних підмножинах даних, що знижує ризик отримання завищених оцінок продуктивності внаслідок перенавчання.

Для даного дослідження було обрано стратегію крос-валідації з п'ятьма розділеннями (5-fold), що дозволило отримати стабільні оцінки та зменшити ймовірність випадкового розподілу результатів. Як метрика для оцінки моделей використовувався ROC AUC, що є стандартним показником для задач класифікації та дозволяє оцінити здатність моделі розпізнавати позитивні і негативні класи, де значення, близькі до 1, вказують на високу якість розділення.

Аналіз застосування різних моделей при базових налаштуваннях виявив результати щодо ефективності різних моделей, наведені на рис. 3.2.

Результати класифікаторів за ROC AUC Score:							
	Classifier	Accuracy	ROC AUC Score	Recall	Precision	Training Time (s)	Prediction Time (s)
0	Ada Boost Classifier	0.756003	0.846545	0.738353	0.745841	5.649492	0.126470
2	Gradient Boosting Classifier	0.753691	0.843832	0.735529	0.743578	18.576224	0.048458
15	CatBoostClassifier	0.754661	0.843257	0.742431	0.741733	20.481305	0.021651
1	Light Gradient Boosting Machine	0.752573	0.843129	0.741804	0.738791	1.788337	0.144223
3	Extreme Gradient Boosting	0.748024	0.835317	0.736314	0.734355	3.689390	0.047465
16	SVM Linear Kernel	0.747129	0.834271	0.760471	0.722288	2333.808681	34.210572
5	Random Forest Classifier	0.738404	0.828053	0.712941	0.730355	13.718327	0.456740
7	Logistic Regression	0.739672	0.825270	0.714667	0.731535	0.378310	0.006918
9	Linear Discriminant Analysis	0.739299	0.825182	0.717333	0.729695	0.194932	0.010017
6	Extra Trees Classifier	0.740045	0.824227	0.714667	0.732123	11.767069	0.684935
10	Quadratic Discriminant Analysis	0.733930	0.816728	0.745412	0.709571	0.159781	0.019244
4	MLP Classifier	0.727218	0.814034	0.721725	0.709483	103.562641	0.059652
11	Naive Bayes	0.730425	0.811519	0.745725	0.704505	0.042488	0.008391
12	K Neighbors Classifier	0.665548	0.714301	0.646118	0.648866	0.017272	6.734551
13	Decision Tree Classifier	0.668531	0.667575	0.648157	0.652352	0.641625	0.008494
14	Dummy Classifier	0.524609	0.500000	0.000000	0.000000	0.002624	0.000309
8	Ridge Classifier	0.739299	NaN	0.717333	0.729695	0.087045	0.010715

Чотири найкращі класифікатори: ['Ada Boost Classifier' 'Gradient Boosting Classifier' 'CatBoostClassifier' 'Light Gradient Boosting Machine']

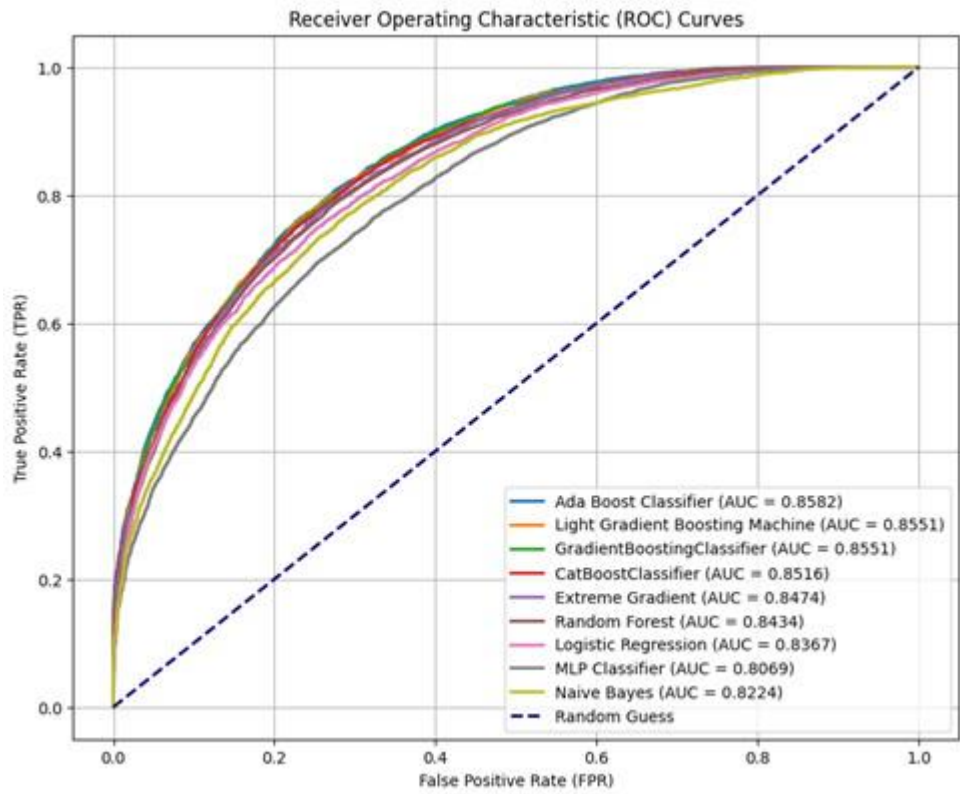
Рисунок 3.2 – Порівняння ефективності класифікаторів при базових налаштуваннях

Найвищу ефективність продемонстрували ансамблеві методи, які зайняли перші місця за всіма метриками. Ці моделі стабільно досягають показників Recall та Precision на рівні 0,74 та ROC AUC Score понад 0,84, що вказує на високу прогностичну здатність та збалансовану продуктивність між класами.

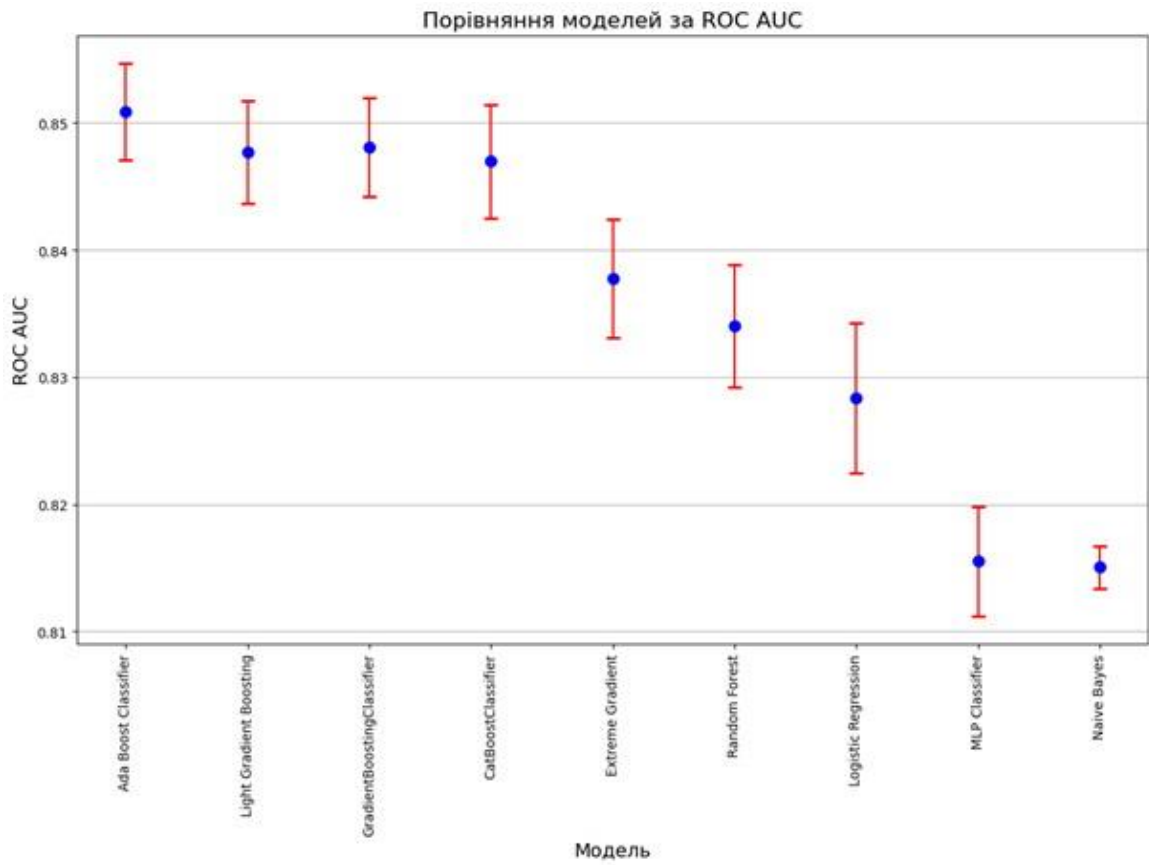
Серед класичних методів найкращі результати продемонстрував метод опорних векторів (SVM) з лінійним ядром. Особливу увагу слід звернути на той факт, що час роботи моделі був значно більшим порівняно з іншими алгоритмами. Така тривалість обробки пояснюється високою обчислювальною складністю методу опорних векторів при великій кількості ознак та спостережень, що робить його малоефективним для задачі з таким обсягом даних. Порівняно з результатами інших класифікаторів, Support Vector Classifier не показав переваг у контексті цієї задачі. Відповідно, доцільність використання даного алгоритму для прогнозування на даному етапі є сумнівною, враховуючи тривалість обробки.

Хоча моделі бустингу продемонстрували найкращі результати за метрикою ROC AUC, інші алгоритми також заслуговують на увагу через їх особливу архітектуру, яка може дати додаткові переваги для інтерпретації або при Stacking ансамблюванні. Логістична регресія, Random Forest, Naive Bayes і MLP Classifier є важливими представниками різних класів моделей: лінійних, ансамблевих, ймовірнісних і нейронних мереж відповідно. Використання цих моделей у подальших дослідженнях доцільне для порівняння їх сильних сторін у різних сценаріях. Логістична регресія забезпечує простоту та інтерпретованість, Random Forest – гнучкість і стійкість, Naive Bayes – здатність працювати з імовірнісними розподілами, а MLP – здатність до моделювання складних нелінійних залежностей.

Детальний аналіз ROC-кривих та діаграми порівняння моделей за метрикою ROC AUC показує чітку ієрархію ефективності різних алгоритмів (рис. 3.3).



а)



б)

Рисунок 3.3 – Порівняння ROC-кривих (а) та діаграми за метрикою ROC AUC (б) вибраних моделей при базових налаштуваннях

Найкращі результати демонструє група бустингових алгоритмів, де Ada Boost Classifier досягає найвищого значення AUC, а Light Gradient Boosting Machine та GradientBoostingClassifier показують майже ідентичні результати. CatBoostClassifier, хоча й дещо поступається лідерам, все ще демонструє високу ефективність.

У середній групі за ефективністю знаходяться Extreme Gradient та Random Forest, які показують стабільні результати з меншим розкидом значень на діаграмі розмаху, що свідчить про їх надійність при різних конфігураціях даних.

Нижчі показники демонструє Logistic Regression, та найнижчі результати показують MLP Classifier та Naive Bayes.

Важливо відзначити, що діаграма розмаху демонструє відносно невеликий розкид значень для всіх моделей, особливо для бустингових алгоритмів, що свідчить про стабільність їх роботи. Це особливо важливо для практичного застосування, оскільки вказує на надійність прогнозів при різних конфігураціях вхідних даних.

3.2 Підвищення ефективності моделей класифікації

Для підвищення ефективності моделей прогнозування було застосовано техніку налаштування гіперпараметрів за допомогою RandomizedSearchCV. Цей підхід вибрано з огляду на його ефективність у пошуку оптимальної конфігурації параметрів моделей, що дозволяє підвищити продуктивність без надмірного обчислювального навантаження, яке характерне для повного перебору варіантів.

Після налаштування гіперпараметрів спостерігається покращення метрик для більшості моделей (рис. 3.4), що підкреслює важливість точного налаштування параметрів для підвищення ефективності моделі.

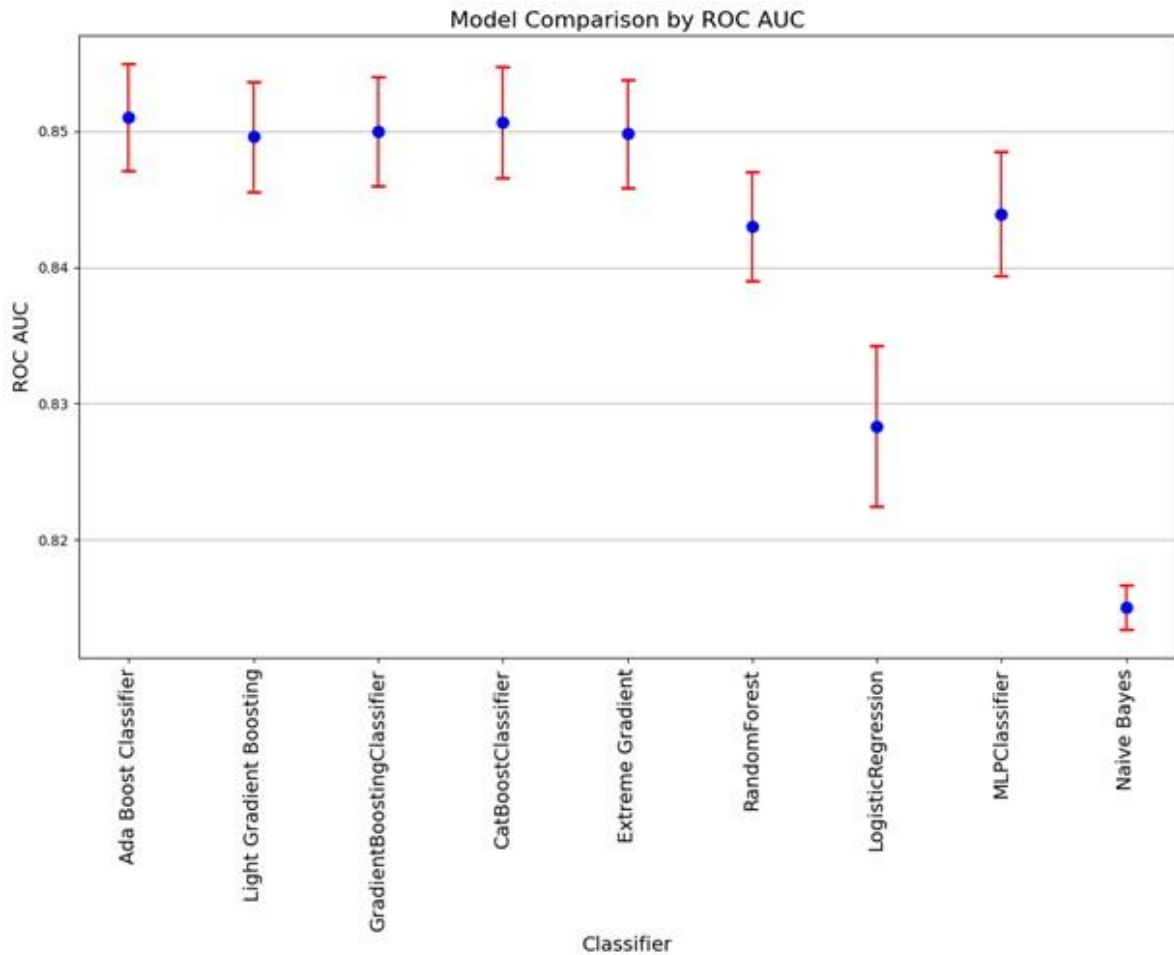
Важливо відзначити, що не всі моделі показали покращення. Наприклад, оптимізація Logistic Regression призвела лише до незначного підвищення показників, що може вказувати на обмеження лінійних моделей у вирішенні даної задачі або на те, що початкові параметри вже були близькими до оптимальних.

Результати класифікаторів за ROC AUC Score:

	Classifier	Accuracy	ROC AUC Score	Recall	Precision	Training Time (s)	Prediction Time (s)
0	Ada Boost Classifier	0.754810	0.846751	0.737569	0.744341	12.167056	0.334641
7	CatBoostClassifier	0.754213	0.845913	0.742118	0.741188	4.506820	0.009734
2	Gradient Boosting Classifier	0.755406	0.845746	0.741333	0.743432	28.377138	0.221525
1	Light Gradient Boosting Machine	0.755332	0.845617	0.741490	0.743239	1.469117	0.118492
3	Extreme Gradient Boosting	0.754362	0.845353	0.740235	0.742331	1.512899	0.047304
4	MLP Classifier	0.746905	0.839995	0.741176	0.730407	36.301441	0.055727
5	Random Forest Classifier	0.751081	0.838283	0.721255	0.746550	18.081788	0.523300
6	Logistic Regression	0.739746	0.825244	0.714824	0.731578	2.633925	0.024252
8	Naive Bayes	0.730425	0.811519	0.745725	0.704505	0.030159	0.006219

Чотири найкращі класифікатори: ['Ada Boost Classifier' 'CatBoostClassifier' 'Gradient Boosting Classifier' 'Light Gradient Boosting Machine']

a)



б)

Рисунок 3.4 – Порівняння ефективності вибраних класифікаторів після налаштування гіперпараметрів (а) та діаграми за метрикою ROC AUC, отриманою методом 5-fold крос-валідації (б)

Основою для висновків про надійність та природу кожного з алгоритмів може слугувати аналіз розподілу ймовірностей. На рисунку 3.5 наведено розподіл ймовірності окремо для класів «0» та «1» (лівий графік) та загальний розподіл ймовірностей звільнень для всієї вибірки (правий графік).

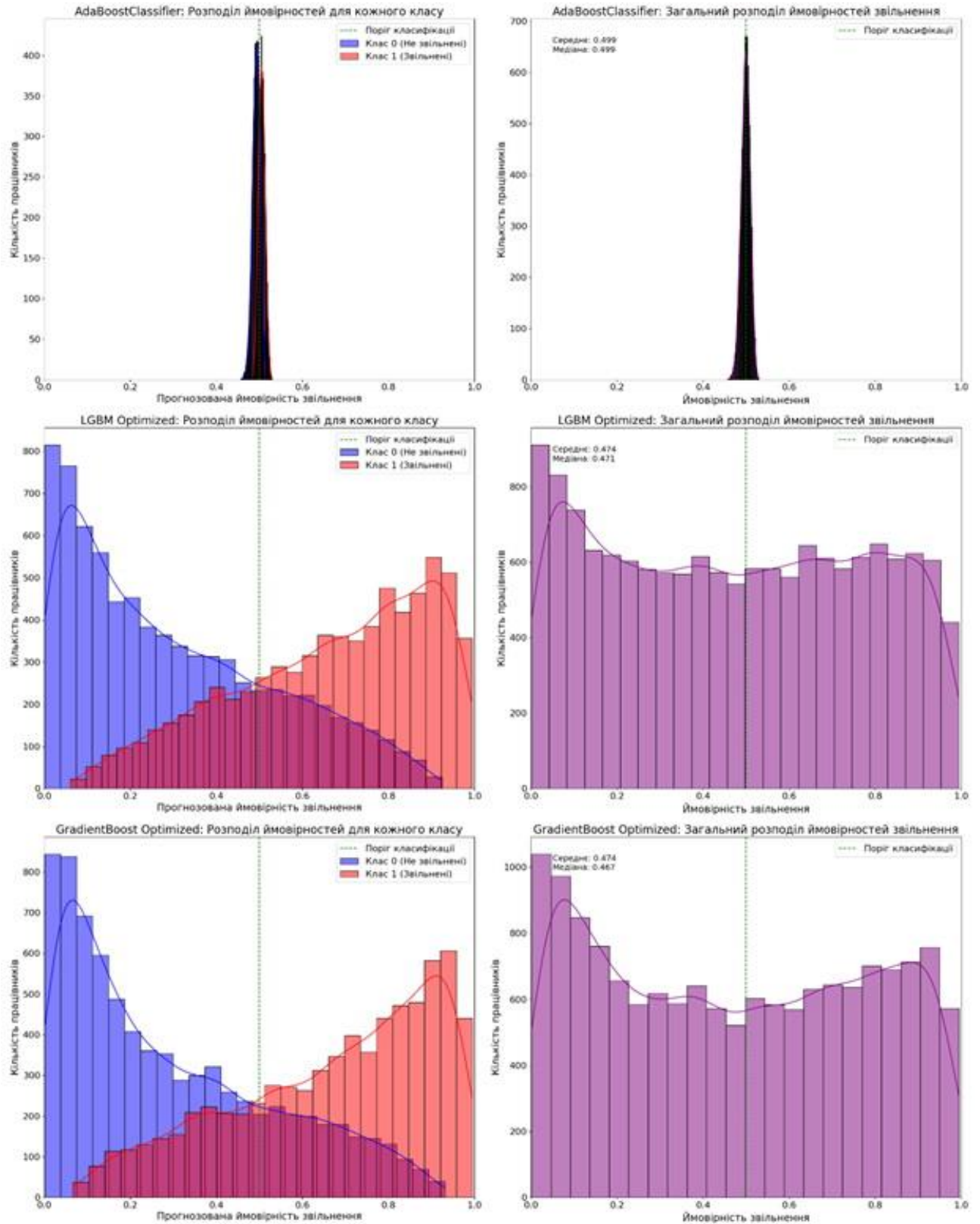


Рисунок 3.5 – Порівняння розподілів ймовірностей для різних моделей класифікації

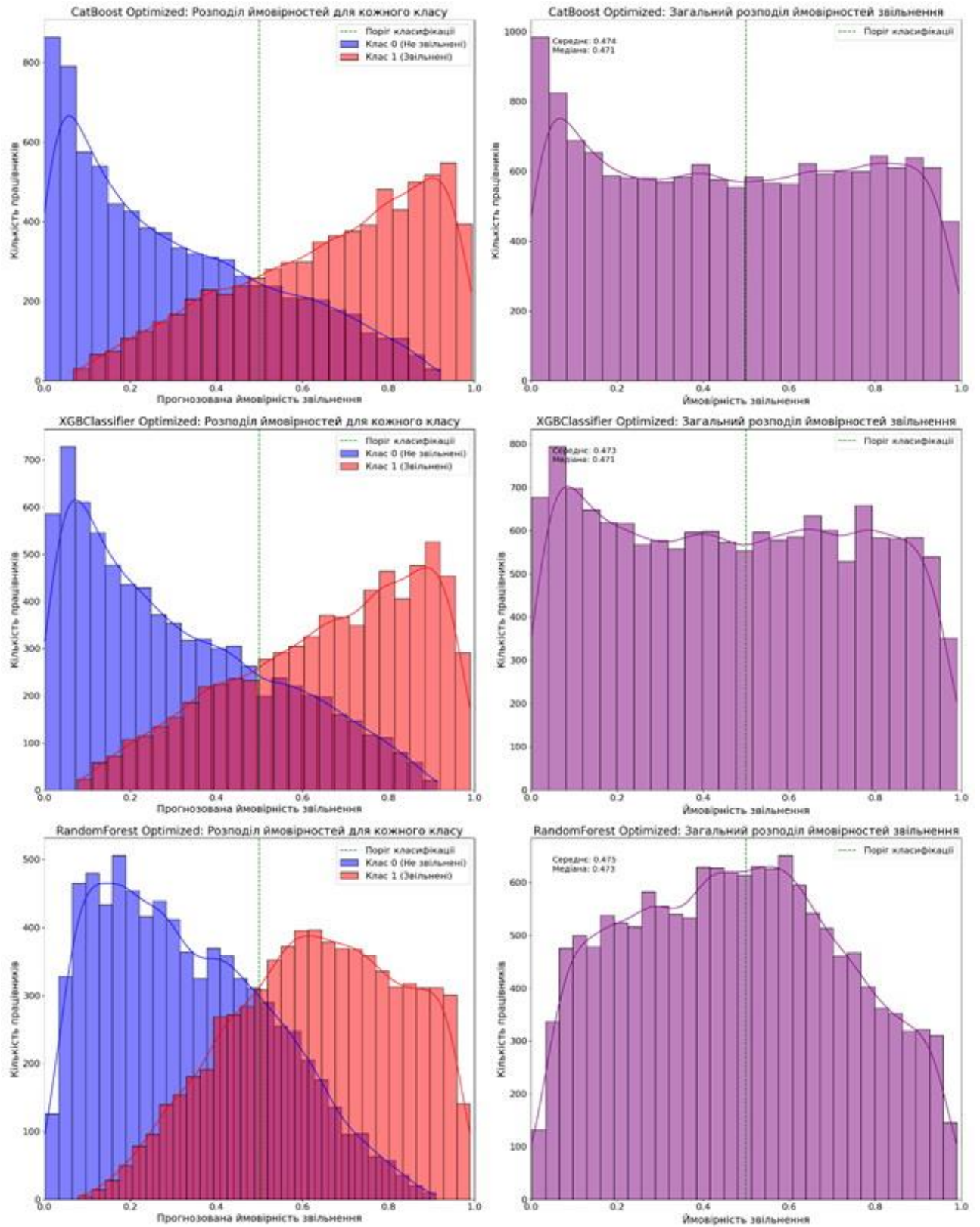


Рисунок 3.5, аркуш 2

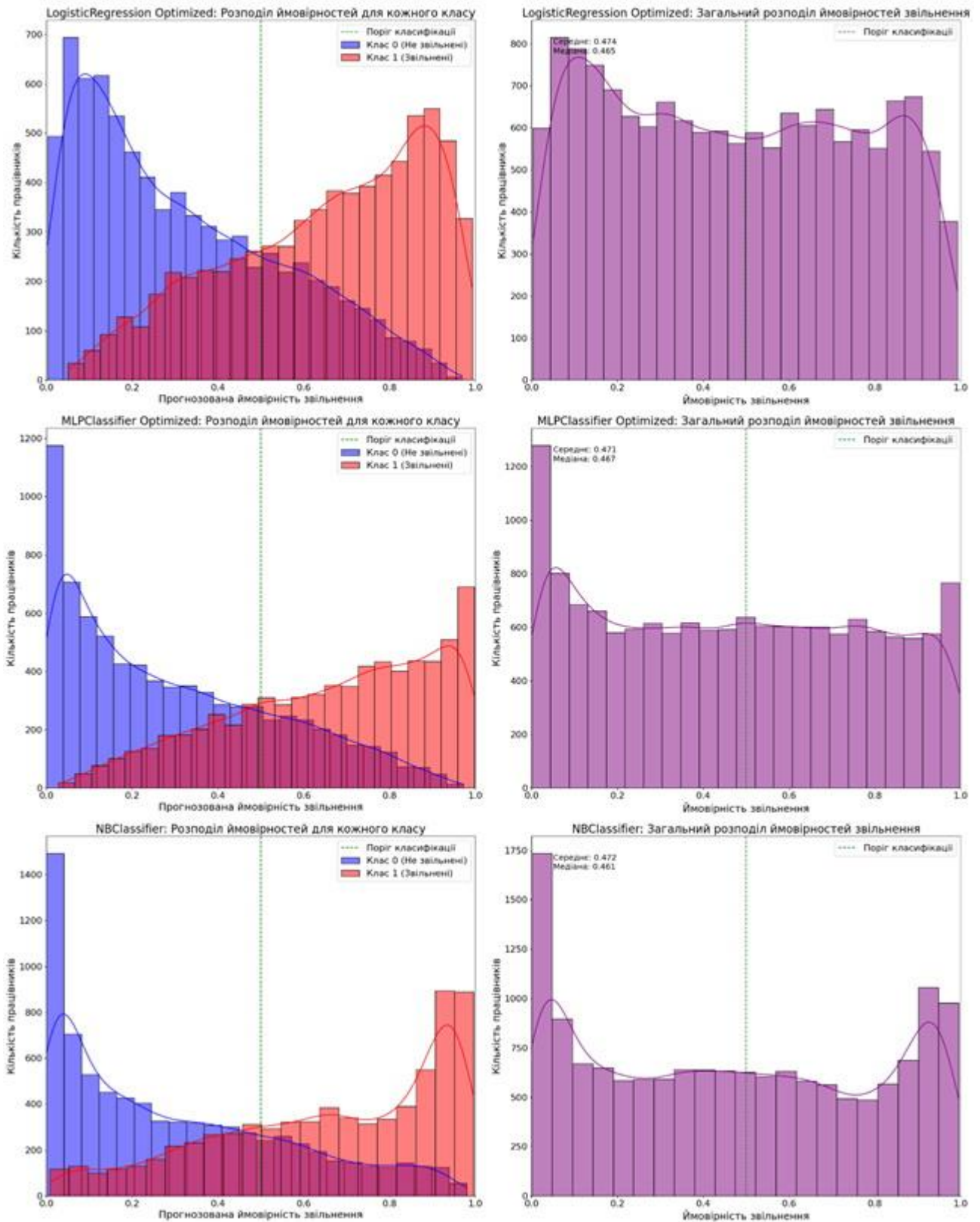


Рисунок 3.5, аркуш 3

Перший графік допомагає зрозуміти, наскільки добре модель передбачає ймовірності для дійсно позитивних випадків. Менше перекриття свідчить про

кращу здатність моделі розрізняти класи. Графік допомагає визначити оптимальний поріг класифікації, який може відрізнитися від стандартного значення 0,5, особливо у випадках з незбалансованими класами.

Другий дає загальну картину розподілу ймовірностей для всього набору даних. Він дозволяє оцінити загальну тенденцію прогнозів моделі та її схильність до певних типів прогнозів (наприклад, чи схильна модель давати крайні значення ймовірностей). Ширина та форма розподілу дають уявлення про рівень невизначеності в прогнозах моделі. Вузкий пік вказує на високу впевненість моделі, тоді як широкий розподіл свідчить про більшу невизначеність.

Аналіз показує, що градієнтні моделі LightGBM, GradientBoosting, CatBoost, XGBoost забезпечують чіткий поділ між класами та рівномірний розподіл ймовірностей з мінімальним перекриттям між класами, що робить їх ідеальними для задачі прогнозування звільнення працівників та ідентифікації груп ризику. Це забезпечує можливість формування чітких груп ризику: низький ризик (ймовірності близькі до 0), високий ризик (ймовірності близькі до 1), та середній ризик (перекриття ймовірностей).

Моделі Naive Bayes та MLP демонструють неефективний розподіл ймовірностей, з сильно вираженим скупченням значень з боків (на краях 0 та 1), що призводить до дуже різкого розподілу та не дає надійної можливості визначити проміжні групи ризику. Моделі надто впевнено класифікують випадки як "залишилися" або "звільнилися", що не залишає місця для визначення середніх ризиків. Це робить їх менш придатними для задачі прогнозування звільнення з формуванням груп ризику.

Модель, яку можна не розглядати, це AdaBoost. На графіку розподілу ймовірностей видно, що цей класифікатор генерує ймовірності, зосереджені навколо середини (приблизно 0,5), і не розділяє класи чітко. Це ускладнює завдання виявлення працівників з високим ризиком, оскільки модель фактично не забезпечує чіткого поділу між тими, хто залишився, та тими, хто звільнився.

Моделі RandomForest та Logistic Regression можуть бути використані для задачі прогнозування звільнення, проте вони мають певні обмеження і можуть

не забезпечити таку ж точність або чіткість розподілу, як градієнтні бустингові моделі.

Logistic Regression забезпечує поступовий розподіл ймовірностей між класами. Це дозволяє ідентифікувати працівників з високим і низьким ризиком, хоча й з меншим поділом у середині, ніж у градієнтних бустингових моделей. Модель може бути корисною, якщо потрібна проста та інтерпретована модель.

RandomForest забезпечує помірний розподіл ймовірностей, де класи все ж можна розрізнити, але з певним перекриттям у середніх значеннях ймовірностей. Вона є стабільною та інтерпретованою, однак поступається градієнтним бустинговим моделям у точності.

3.3 Аналіз використання методів ансамблевого навчання

В роботі було досліджено методологію ансамблевого навчання з метою подолання обмежень окремих моделей та підвищення загальної точності прогнозів. Ансамблеві методи, зокрема VotingClassifier та StackingClassifier, були вибрані з огляду на їх здатність комбінувати сильні сторони різних алгоритмів, одночасно зменшуючи вплив індивідуальних недоліків кожної моделі. Для створення ансамблю було використано класифікатори з першої групи «найкраще підходять» та класифікатори з групи «середній варіант»:

- градієнтні бустингові моделі LightGBM, CatBoost та GradientBoosting;
- модель RandomForest;
- модель Logistic Regression включена в ансамбль VotingClassifier як базова модель, та в ансамблі StackingClassifier включена як метамодель.

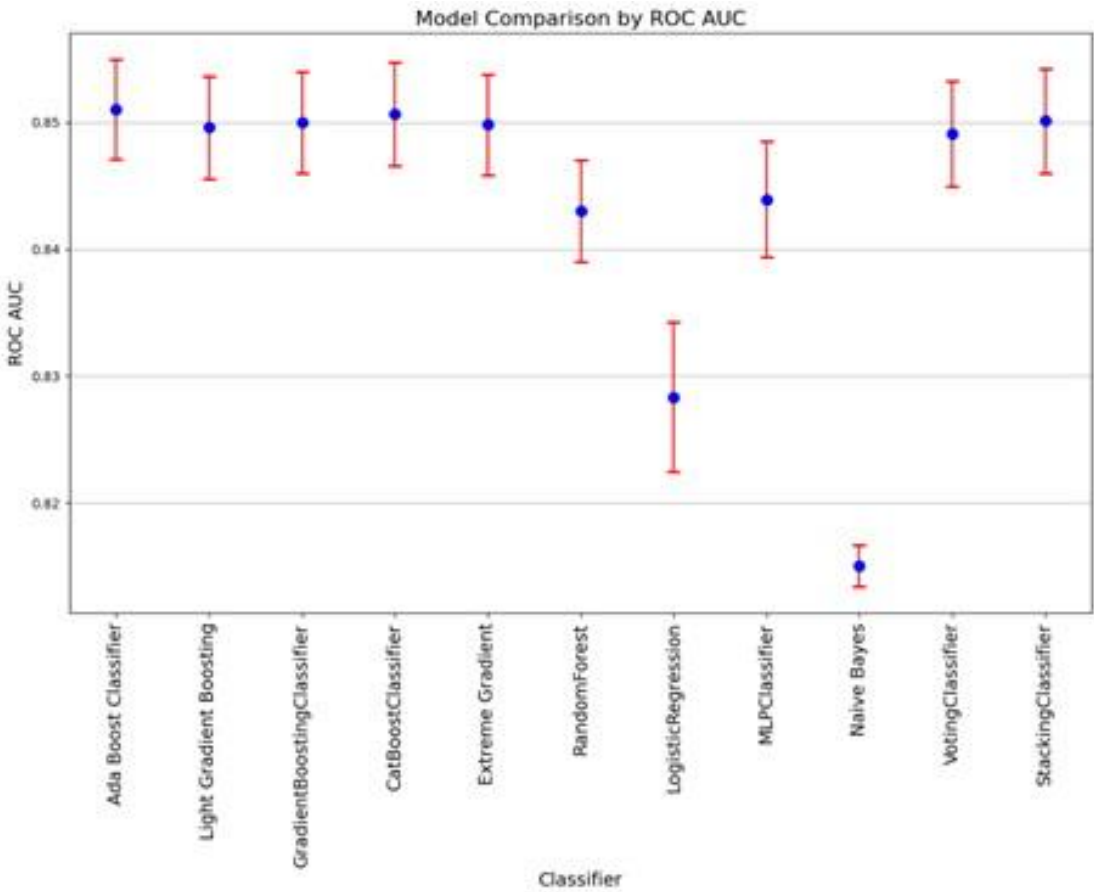
Така комбінація може бути ефективною, оскільки бустингові моделі забезпечують високу точність і гарний розподіл ймовірностей для визначення працівників з різними рівнями ризику звільнення, RandomForest допомагає збалансувати ансамбль, додаючи елементи інтерпретованості та стабільності, метамодель Logistic Regression агрегує прогнози базових моделей, що дозволяє підвищити точність і знизити варіативність у прогнозах. Такий підхід дозволяє досягти більш робастних та надійних результатів у порівнянні з використанням

окремих класифікаторів. Порівняння ансамблевих моделей з індивідуальними класифікаторами підтвердило ефективність ансамблевого підходу (рис. 3.6).

Порівняльна таблиця результатів для обраних моделей:

	Classifier	Accuracy	ROC AUC Score	Recall	Precision	F1 Score
0	Ada Boost Classifier	0.7548	0.8468	0.7376	0.7443	0.7409
1	Light Gradient Boosting	0.7553	0.8456	0.7415	0.7432	0.7424
2	GradientBoostingClassifier	0.7554	0.8457	0.7413	0.7434	0.7424
3	CatBoostClassifier	0.7542	0.8459	0.7421	0.7412	0.7417
4	Extreme Gradient	0.7544	0.8454	0.7402	0.7423	0.7413
5	RandomForest	0.7511	0.8383	0.7213	0.7465	0.7337
6	LogisticRegression	0.7397	0.8252	0.7148	0.7316	0.7231
7	MLPClassifier	0.7469	0.8400	0.7412	0.7304	0.7358
8	Naive Bayes	0.7304	0.8115	0.7457	0.7045	0.7245
9	VotingClassifier	0.7532	0.8449	0.7363	0.7425	0.7394
10	StackingClassifier	0.7533	0.8453	0.7376	0.7420	0.7398

а)



б)

Рисунок 3.6 – Ефективність індивідуальних класифікаторів (а) та діаграми за метрикою ROC AUC, отриманою методом 5-fold крос-валідації (б) після налаштування гіперпараметрів

Варто зазначити, що оптимізація гіперпараметрів в ансамблях не була проведена в рамках цього експерименту. Ансамблі були складені з моделей на базових налаштуваннях, що дозволило отримати попередні результати та оцінити ефективність підходу без оптимізації. Зокрема, `StackingClassifier` продемонстрував високу медіану та менший розкид значень, що свідчить про його стабільність та ефективність у порівнянні з іншими моделями. Це підтверджує, що комбінування різних моделей дозволяє отримати більш надійні та точні прогнози, використовуючи сильні сторони кожного класифікатора та компенсуючи їх слабкості.

Ансамблі `VotingClassifier` і `StackingClassifier` показують достатньо чіткий розподіл між класами, схожий на градієнтні методи (`GradientBoosting`, `CatBoost` тощо), що свідчить про стабільний прогноз для визначення груп ризику. Особливо `StackingClassifier` показує хороший поділ між класами та розподіл ймовірностей, що можна використати для більш надійного прогнозування (рис. 3.7).

На графіку порівняння за метрикою ROC AUC видно, що більшість моделей демонструють значення, які перевищують 0,84, що свідчить про високу якість підготовки даних. Ретельна попередня обробка, включаючи масштабування, значно покращує стабільність та продуктивність моделей, забезпечуючи високий рівень прогнозування для різних класифікаторів.

Хоча ансамблі `StackingClassifier` та `VotingClassifier` на базових налаштуваннях за метрикою ROC AUC не перевершили найкращі індивідуальні моделі з налаштованими гіперпараметрами, їх оптимізація за допомогою методів, таких як `RandomizedSearchCV` або `GridSearchCV`, може дозволити розкрити повний потенціал окремих моделей і посилити їх взаємодію в ансамблях. При цьому треба врахувати, що оптимізація гіперпараметрів кожного класифікатора перед додаванням їх у стекінг значно збільшує загальний час роботи, оскільки час оптимізації для кожного класифікатора додається до загального часу навчання.

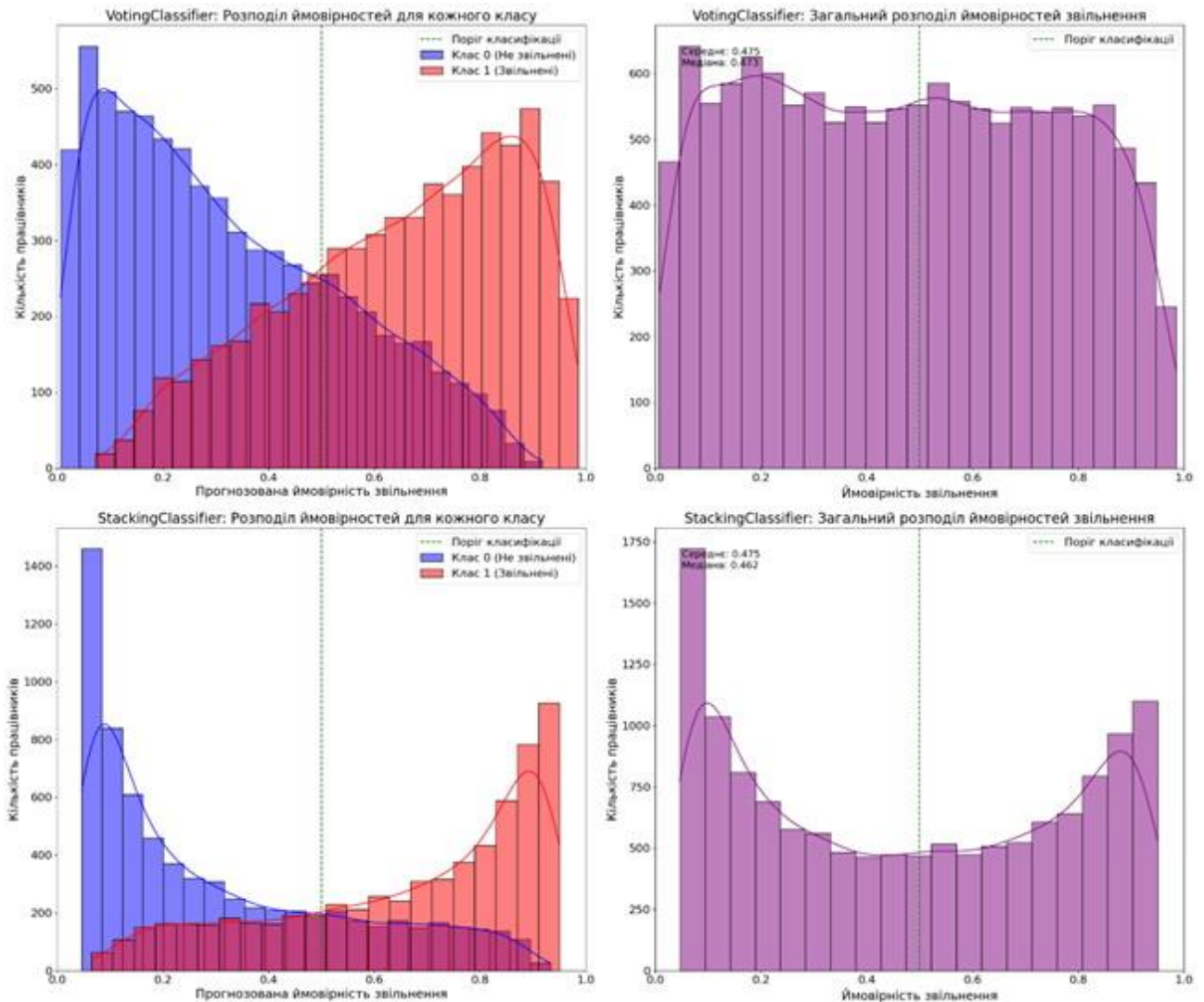


Рисунок 3.7 – Порівняння розподілів ймовірностей ансамблевих методів VotingClassifier та StackingClassifier

Таким чином, для задачі прогнозування звільнення працівників та формування груп ризику краще використовувати LGBM, GradientBoosting, CatBoost, XGBoost або StackingClassifier. Ці моделі забезпечують чіткий розподіл ймовірностей між класами, що дозволяє більш точно визначити працівників з високим ризиком звільнення. AdaBoost можна виключити, оскільки її розподіл ймовірностей не дозволяє ефективно визначити працівників з групи ризику.

3.4 Обґрунтування вибору методу відбору ознак

Під час підготовки до моделювання значну увагу було приділено етапу відбору ознак. Цей крок дозволяє зосередитися на найбільш інформативних характеристиках, зменшуючи шум та надмірність даних. Відбір ознак сприяє кращій інтерпретованості, що особливо важливо в контексті управління персоналом.

Для відбору ознак було застосовано кілька методів, кожен із яких базується на різних принципах та критеріях. Кореляційний аналіз використовував поріг 0,05, що дозволяло відсіяти ознаки зі слабким кореляційним зв'язком із цільовою змінною. Метод головних компонент (PCA) був застосований із параметром, що зберігає 95% варіації даних та зменшує кількість змінних, залишаючи лише ті, що найбільше пояснюють розподіл ознак.

Lasso регресія, як метод із регуляризацією, відібрала ознаки на основі ненульових коефіцієнтів у моделі, що дозволило залишити лише найважливіші ознаки. У випадку методу Random Forest використовувався поріг важливості ознак 0,05, завдяки якому відібрали ознаки з важливістю, вищою за задане значення. Рекурсивне виключення ознак із крос-валідацією (RFECV) поступово видаляло найменш значущі ознаки, забезпечуючи оптимальний набір для моделі за допомогою внутрішньої крос-валідації. Генетичний алгоритм (GA) здійснював пошук комбінацій ознак, оцінюючи продуктивність моделі на кожному етапі еволюційного процесу, щоб знайти найбільш значущі ознаки.

Таким чином, кожен із методів спрямований на оптимізацію набору ознак за своїм принципом, дозволяючи виділити ті змінні, які найбільше сприяють підвищенню точності моделі та зменшенню її складності.

Для оцінки ефективності різних методів відбору ознак було проведено порівняльний аналіз з використанням класифікатора LGBM Classifier, що дозволяє усунути вплив різних моделей на результати. Метою було визначити, який з цих методів забезпечує найвищу якість класифікації.

Аналіз результатів, отриманих за різними методами відбору ознак, показав, що методи відбору суттєво впливають на якість моделі та характеристику розподілу ймовірностей (рис. 3.8).

Високу продуктивність, зокрема, високий ROC AUC Score, забезпечив метод рекурсивного виключення ознак з крос-валідацією (RFECV), що свідчить про значну силу цього підходу у виборі ознак. Відзначається також метод Lasso який надав близькі результати, що вказує на його надійність і відповідність критеріям для прогнозування.

Порівняльна таблиця результатів:

	Classifier	Accuracy	ROC AUC Score	Recall	Precision	F1 Score	Selected Features
0	LGBMClassifier	0.7556	0.8456	0.7423	0.7432	0.7427	23.0000
1	Correlation	0.7456	0.8343	0.7311	0.7331	0.7321	11.0000
2	PCA	0.7184	0.8015	0.7107	0.7011	0.7059	11.0000
3	Lasso	0.7534	0.8421	0.7399	0.7410	0.7404	18.0000
4	RandomForest	0.7030	0.7789	0.6739	0.6929	0.6833	5.0000
5	RFECV	0.7548	0.8451	0.7413	0.7425	0.7419	18.0000
6	GA	0.7321	0.8179	0.7250	0.7154	0.7202	7.0000

Рисунок 3.8 – Порівняння ефективності класифікатора при використанні різних методів відбору ознак

Метод головних компонент (PCA) забезпечив ефективне зниження кількості ознак, зберігаючи значну частину варіації в даних, але показав менш значущі результати в термінах точності. Генетичний алгоритм (GA) також заслуговує уваги, хоча він обрав менше ознак, але показав помітне зниження продуктивності порівняно з іншими методами. Це може свідчити про обмежену адаптацію GA до специфіки задачі прогнозування. Методи кореляційного аналізу та випадкового лісу (RandomForest) показали помірні результати, але виявилися менш ефективними порівняно з Lasso та RFECV.

Аналіз розподілу ймовірностей для кожного методу відбору ознак виявив особливості кожного підходу у прогнозуванні ймовірностей звільнення (рис. 3.9). Методи RFECV та Lasso продемонстрували чіткий розподіл для обох

класів, що свідчить про здатність відбирати релевантні ознаки для прогнозування. Вони забезпечують гарну диференціацію між класами і загалом створили стабільний розподіл ймовірностей навколо порогу класифікації.

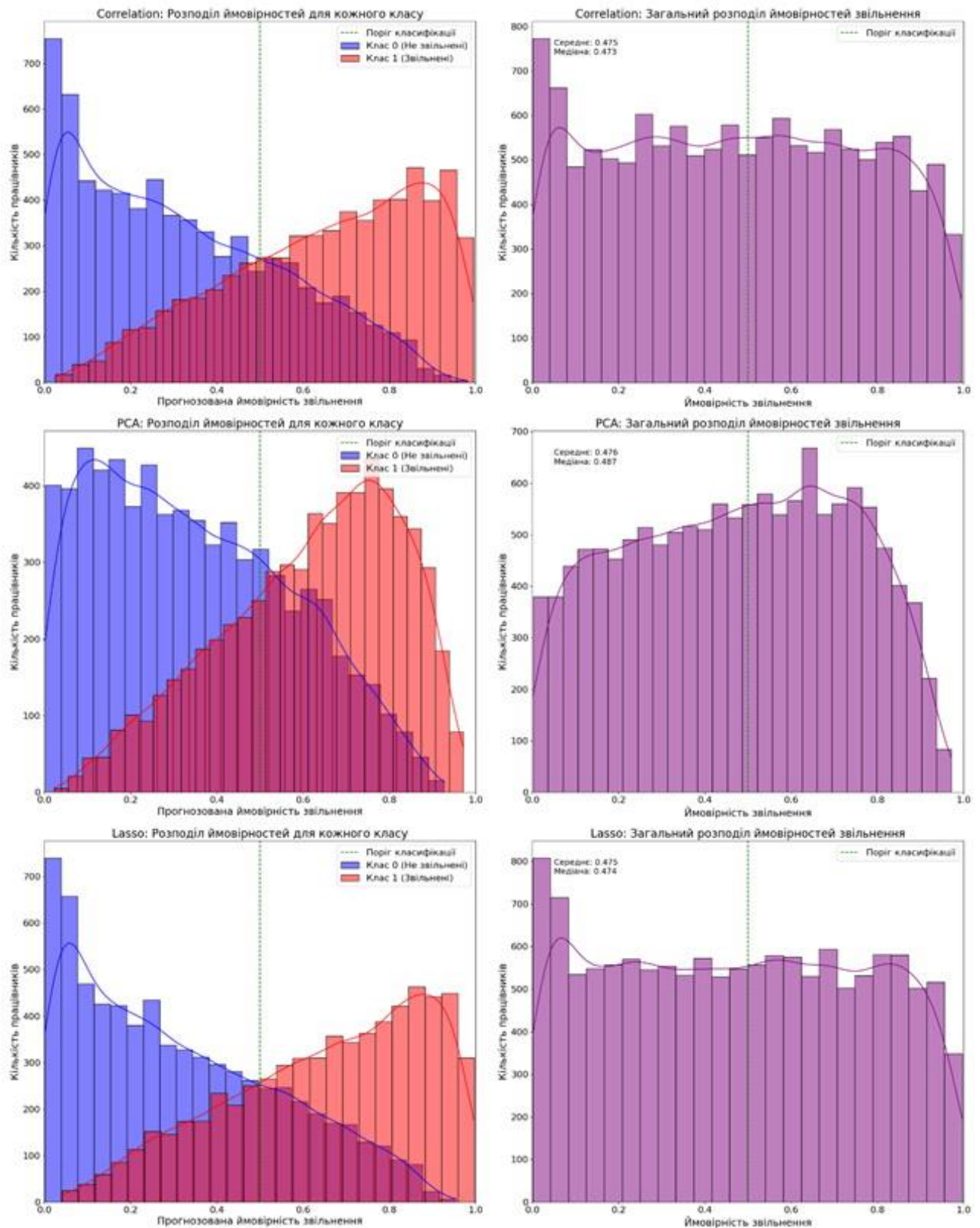


Рисунок 3.9 – Розподіл ймовірностей для різних методів відбору ознак

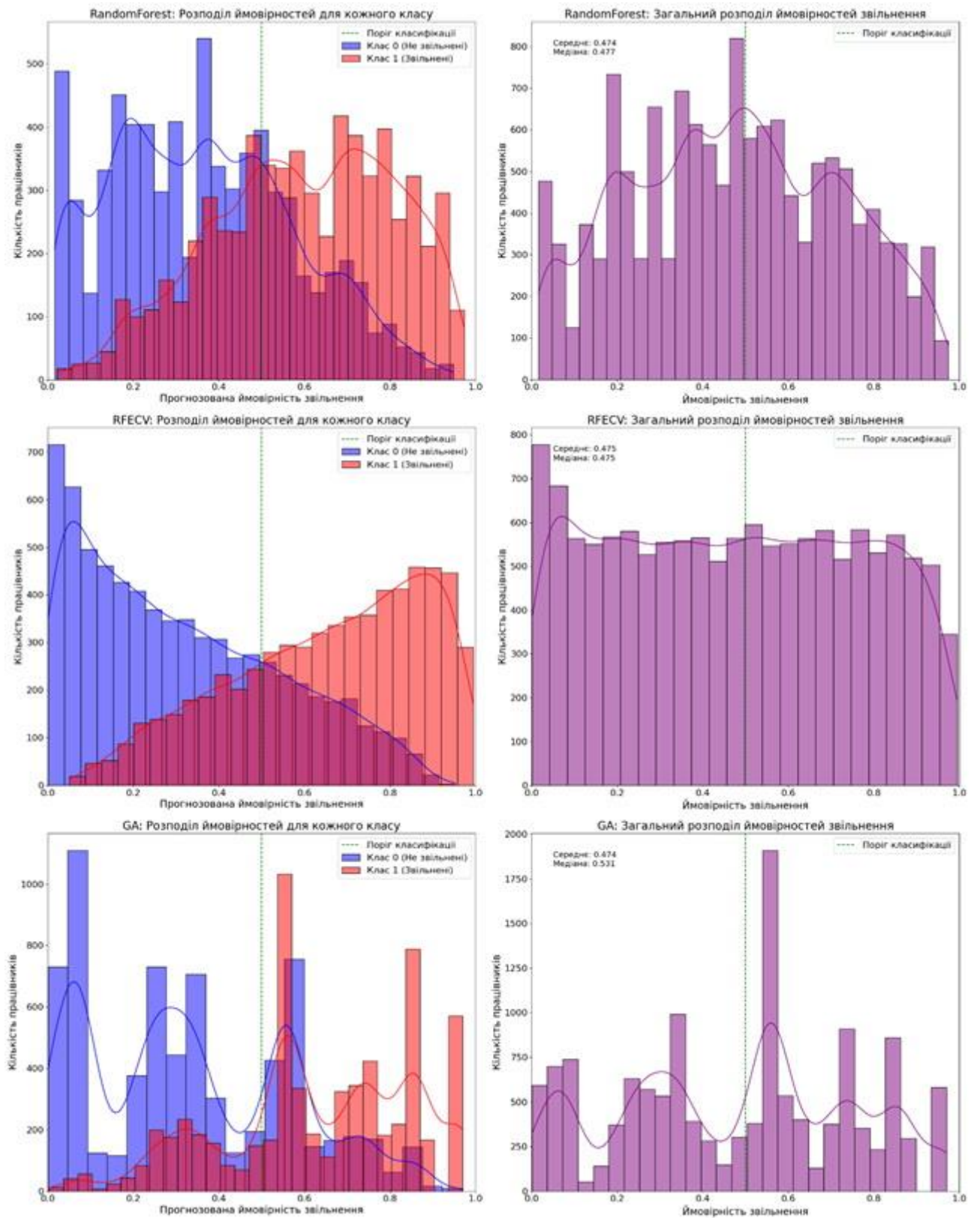


Рисунок 3.9, аркуш 2

Кореляційний аналіз продемонстрував подібний розподіл, хоча щільність розподілу для класу "залишилися" виглядає менш однорідною. Це може свідчити про деякі обмеження в здатності кореляційного аналізу охоплювати всі важливі ознаки, що впливають на ймовірність звільнення.

Метод RandomForest виявив обмежену здатність до розрізнення класів, що відображено в менш чітко вираженому розподілі для кожного класу. Нарешті, генетичний алгоритм (GA) продемонстрував найбільш нерівномірний розподіл ймовірностей. Незважаючи на можливість виділення корисних ознак, він не зміг забезпечити стабільний розподіл для кожного класу, що свідчить про його потенційну недостатність для цієї задачі.

Виходячи з результатів, методи Lasso та RFECV показали найвищу продуктивність та можуть бути використані для моделювання. Вони забезпечують високий рівень точності та стабільний розподіл ймовірностей, що є важливими критеріями для побудови надійної прогностичної моделі.

Порівнюючи результати моделей можна помітити, що найкращий результат серед методів відбору ознак не перевершує показників найкращих оптимізованих моделей.

Однак, важливо враховувати кілька важливих аспектів.

1. Хоча методи відбору ознак не перевершили оптимізовані моделі за кількісними показниками, вони можуть значно підвищити інтерпретованість моделі. Зменшення кількості вхідних параметрів дозволяє краще зрозуміти, які фактори найбільше впливають на рішення співробітників про звільнення.

2. Модель з меншою кількістю ознак може працювати швидше та ефективніше, що особливо важливо при роботі з великими обсягами даних або при необхідності частого оновлення прогнозів.

3. Відбір найбільш релевантних ознак може допомогти зменшити ризик перенавчання моделі, особливо при роботі з обмеженими наборами даних.

4. У контексті прогнозування звільнень співробітників розуміння ключових факторів, що впливають на рішення про звільнення, може бути не менш важливим, ніж досягнення максимальної точності прогнозу.

Враховуючи ці фактори, хоча прямий відбір ознак не призвів до підвищення кількісних показників ефективності моделі, його застосування може бути корисним у контексті даної задачі.

Результати дослідження підкреслюють важливість комплексного підходу до побудови моделі, який включає як оптимізацію гіперпараметрів так і відбір ознак. Оптимізація дозволила не лише підвищити загальну точність прогнозів, але й покращити здатність моделей до розрізнення складних випадків, що особливо важливо в контексті прогнозування звільнень співробітників. Модель на відібраному наборі ознак після оптимізації також демонструє високу ефективність, пропонуючи компроміс між точністю та складністю моделі.

3.5 Висновки за розділом 3

Встановлено, що для даної задачі оптимальним є використання метрик ROC AUC та F1-Score для оцінки якості моделей, враховуючи необхідність балансу між точністю виявлення працівників, схильних до звільнення, та мінімізацією помилкових спрацювань.

Показано, що серед різних методів машинного навчання найвищу ефективність для даного набору даних мають ансамблеві методи, зокрема різновидів градієнтного бустингу.

Встановлено, що оптимізація гіперпараметрів за допомогою RandomizedSearchCV дозволила покращити метрики для більшості моделей. Аналіз розподілу ймовірностей показав, що градієнтні моделі (LightGBM, GradientBoosting, CatBoost, XGBoost) забезпечують найкращий поділ між класами та рівномірний розподіл ймовірностей, що робить їх оптимальними для задачі прогнозування звільнення працівників.

Аналіз методів відбору важливих ознак показав, методи Lasso та RFECV забезпечують високий рівень точності та стабільний розподіл ймовірностей, що є важливими критеріями для побудови надійної прогностичної моделі.

4 РОЗРОБКА ТА ТЕСТУВАННЯ СИСТЕМИ ПРОГНОЗУВАННЯ

4.1 Архітектура та реалізація компонентів системи прогнозування плинності кадрів

За результатами проведеного дослідження було розроблено комплексну систему прогнозування звільнення працівників, що базується на методах машинного навчання. Система має модульну структуру, що забезпечує чітке розділення функціоналу та гнучкість у використанні (рис. 4.1).

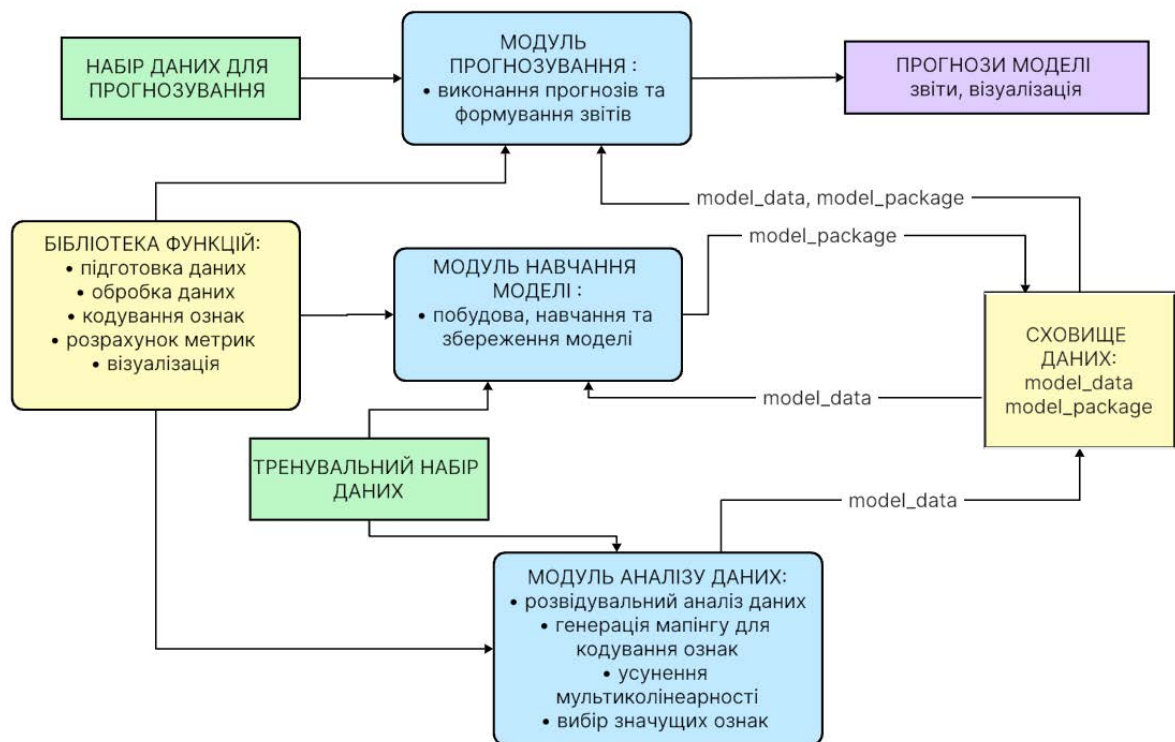


Рисунок 4.1 – Структура розробленої системи прогнозування звільнення працівників

Розроблена система працює з двома типами вхідних даних: тренувальним набором даних, що містить інформацію про працівників (включаючи такі характеристики як стаж, рівень задоволення роботою, баланс між роботою та життям, роль, рівень освіти та інше), та набором даних для прогнозування, який відповідає структурі тренувального набору.

Архітектура системи складається з чотирьох основних функціональних модулів.

Модуль аналізу даних (`data_analysis.py`) відповідає за попередню обробку та комплексний аналіз даних, що є важливим етапом для забезпечення якості прогнозування. Модуль виконує наступні ключові функції:

а) статистичний аналіз характеристик даних:

- 1) розрахунок базових статистичних метрик;
- 2) виявлення викидів та аномалій;
- 3) аналіз розподілів змінних;
- 4) перевірка якості та повноти даних;

б) аналіз та візуалізація розподілу класів:

- 1) оцінка збалансованості класів;
- 2) створення візуалізацій для кращого розуміння структури даних;

в) обробка категоріальних змінних:

- 1) аналіз унікальних значень;
- 2) генерація оптимальних схем кодування;
- 3) створення мапінгів для перетворення категоріальних змінних;

г) аналіз взаємозв'язків між змінними:

- 1) побудова кореляційних матриць;
- 2) виявлення мультиколінеарності;
- 3) відбір оптимального набору предикторів;

д) застосування відбору значущих ознак:

- 1) оцінка важливості різних характеристик;
- 2) оптимізація набору предикторів;
- 3) зменшення розмірності даних.

Результатом роботи цього модуля є створення файлу `model_data`, який містить всі необхідні налаштування та параметри для подальшої узгодженої обробки даних, включаючи:

- мапінги для кодування категоріальних змінних;
- перелік колонок для створення нових ознак;
- список відібраних значущих ознак;

– перелік змінних для застосування методу PCA.

Модуль навчання моделі (`model.py`) відповідає за побудову, навчання та збереження моделі машинного навчання.

Основні функції модуля включають:

а) підготовка даних для навчання:

1) кодування категоріальних змінних згідно з визначеними мапінгами;

2) створення нових ознак за визначеними колонками;

3) застосування методу PCA за визначеним переліком змінних;

4) відбір значущих ознак за визначеним списком;

5) масштабування числових ознак за допомогою `scaler`;

6) балансування навчального набору даних;

б) навчання моделі:

1) розділення даних на навчальну та валідаційну вибірки;

2) налаштування параметрів базової моделі;

3) навчання моделі на збалансованому наборі даних;

4) оцінка якості моделі на валідаційній вибірці.

Результатом роботи цього модуля є створення файлу `model_package`, що містить: натреновану модель, об'єкт `scaler` для масштабування даних, компонент PCA для обробки ознак.

Модуль прогнозування (`prognosz.py`) призначений для роботи з новими даними та формування прогнозів.

Він забезпечує:

а) підготовка даних для прогнозування:

1) кодування категоріальних змінних згідно з визначеними мапінгами;

2) створення нових ознак за визначеними колонками;

3) застосування методу PCA за визначеним переліком змінних;

4) відбір значущих ознак за визначеним списком;

5) масштабування числових ознак за допомогою `scaler`;

б) формування прогнозів:

- 1) розрахунок ймовірностей звільнення для кожного працівника;
- 2) визначення груп ризику;
- 3) генерація пояснень для прогнозів;

в) створення звітів:

- 1) формування детальних звітів по групах ризику;
- 2) створення візуалізацій розподілу ймовірностей;
- 3) збереження результатів у форматі CSV.

Бібліотека функцій (`data_processing.py`) надає набір допоміжних функцій які використовуються іншими модулями для обробки даних.

Особливістю даного модуля є те, що він не працює з даними безпосередньо, а надає інструментарій для їх обробки. Такий підхід забезпечує:

- стандартизацію процесів обробки даних;
- зменшення дублювання коду;
- спрощення процесу внесення змін та оновлень;
- підвищення надійності системи через централізоване тестування функцій.

Взаємодія між компонентами системи організована через проміжні файли `model_data` та `model_package`, що забезпечує узгодженість у підході до обробки даних та прогнозування. Система має три основні потоки даних:

а) створення конфігурацій обробки даних:

- 1) завантаження та аналіз тренувальних даних;
- 2) формування файлу `model_data` з налаштуваннями.

б) навчання моделі:

- 1) завантаження тренувальних даних та налаштувань з `model_data`;
- 2) обробка даних;
- 3) навчання моделі;
- 4) збереження компонентів у `model_package`.

в) прогнозування:

- 1) завантаження нових даних, налаштувань з `model_data` та компонентів з `model_package`;
- 2) обробка нових даних;

3) формування прогнозів та звітів.

Всі компоненти системи реалізовані з використанням мови програмування Python та сучасних бібліотек для аналізу даних та машинного навчання.

4.2 Тестування роботи розробленої системи прогнозування

Для перевірки ефективності розробленої системи прогнозування звільнення працівників було проведено комплексне тестування всіх її компонентів. Початковий набір даних був розділений на дві частини: 90% для навчання моделі і 10% для прогнозування. При цьому розподілі було збережено початкове співвідношення класів, що забезпечило репрезентативність наборів.

Першим етапом тестування стала перевірка роботи модуля `data_analysis.py`, який відповідає за попередню обробку та аналіз даних. У результаті роботи модуля було створено конфігураційний файл `model_data.pkl`, який містить:

- мапінг для початкового кодування категоріальних змінних;
- трансформери для створення нових ознак;
- перелік відібраних ознак;
- список ознак для застосування методу PCA (рис. 4.2).

Важливим результатом роботи модуля стала візуалізація значущості ознак на основі коефіцієнтів RFECV. Аналіз важливості ознак дозволив виявити найбільш впливові фактори, що визначають ймовірність звільнення працівника, та надає цінну інформацію для розробки стратегій утримання персоналу. Аналіз графіку коефіцієнтів показує значний вплив окремих ознак на ймовірність звільнення (рис. 4.3).

Однією з найсильніших ознак, що підвищує цей ризик, є сімейний статус. Працівники, які не перебувають у шлюбі, мають вищу ймовірність звільнення, що вказує на можливий зв'язок між стабільністю сімейного стану та прихильністю до роботи. Також значно підвищує ризик звільнення

понаднормова робота. Це може свідчити про можливу незадоволеність роботою чи стрес через надмірне навантаження.

Перевірка вмісту `model_data`:

Мапінг для початкового кодування (mapping):

```
{'target_mapping': {'Stayed': 0, 'Left': 1}, 'class_weights': {'Stayed': 0.5255637752058226, 'Left': 0.4744362247941773}, 'ordinal_mappings': {'Work-Life Balance': {'Poor': 0, 'Fair': 1, 'Good': 2, 'Excellent': 3}, 'Job Satisfaction': {'Low': 0, 'Medium': 1, 'High': 2, 'Very High': 3}, 'Performance Rating': {'Low': 0, 'Below Average': 1, 'Average': 2, 'High': 3}, 'Education Level': {'High School': 0, 'Associate Degree': 1, 'Bachelor Degree': 2, 'Master Degree': 3, 'PhD': 4}, 'Job Level': {'Entry': 0, 'Mid': 1, 'Senior': 2}, 'Company Reputation': {'Poor': 0, 'Fair': 1, 'Good': 2, 'Excellent': 3}, 'Employee Recognition': {'Low': 0, 'Medium': 1, 'High': 2, 'Very High': 3}, 'Company Size': {'Small': 0, 'Medium': 1, 'Large': 2}}, 'nominal_columns': ['Job Role', 'Marital Status'], 'label_encoders': {'Attrition': LabelEncoder(), 'Gender': LabelEncoder(), 'Innovation Opportunities': LabelEncoder(), 'Leadership Opportunities': LabelEncoder(), 'Overtime': LabelEncoder(), 'Remote Work': LabelEncoder()}}
```

Трансформери для обробки мультиколінеарності (transformers):

```
{'feature_groups': {'job_roles': ['Media', 'Technology', 'Finance', 'Healthcare', 'Education'], 'monthly': ['Monthly Income'], 'marital_features': ['Married', 'Divorced', 'Single']}, 'columns_to_drop': ['Job Role_Media', 'Job Role_Technology', 'Job Role_Finance', 'Job Role_Healthcare', 'Job Role_Education', 'Monthly Income', 'Marital Status_Married', 'Marital Status_Divorced']}
```

Відібрані ознаки (selected_features):

```
['Marital Status_Single', 'Remote Work', 'Job Level', 'Work-Life Balance', 'Gender', 'Overtime', 'Company Reputation', 'Number of Promotions', 'Leadership Opportunities', 'Performance Rating', 'Number of Dependents', 'Innovation Opportunities', 'Education Level', 'Company Size', 'Job Satisfaction', 'Employee Recognition', 'Distance from Home', 'Education_Income_Ratio']
```

Ознаки для PCA (experience_features):

```
['Years at Company', 'Age', 'Company Tenure']
```

Рисунок 4.2 – Конфігураційний файл `model_data.pkl`

Інші ознаки, такі як задоволеність роботою та відстань до офісу, мають менший, але все ж таки позитивний вплив на ймовірність звільнення. Низький рівень задоволеності роботою та значна віддаленість можуть викликати в працівників бажання змінити місце роботи, зокрема у випадках, коли витрати часу на дорогу негативно позначаються на життєвому балансі.

Окрім того, є фактори, що суттєво знижують ризик звільнення. Серед них найбільш виразний вплив має можливість працювати віддалено. Працівники, які мають можливість виконувати роботу з дому, демонструють нижчий рівень звільнень, що вказує на важливість гнучких умов праці. Також високий рівень посади сприяє утриманню працівників, оскільки вони, ймовірно, мають більше можливостей для розвитку та вищий рівень задоволення.

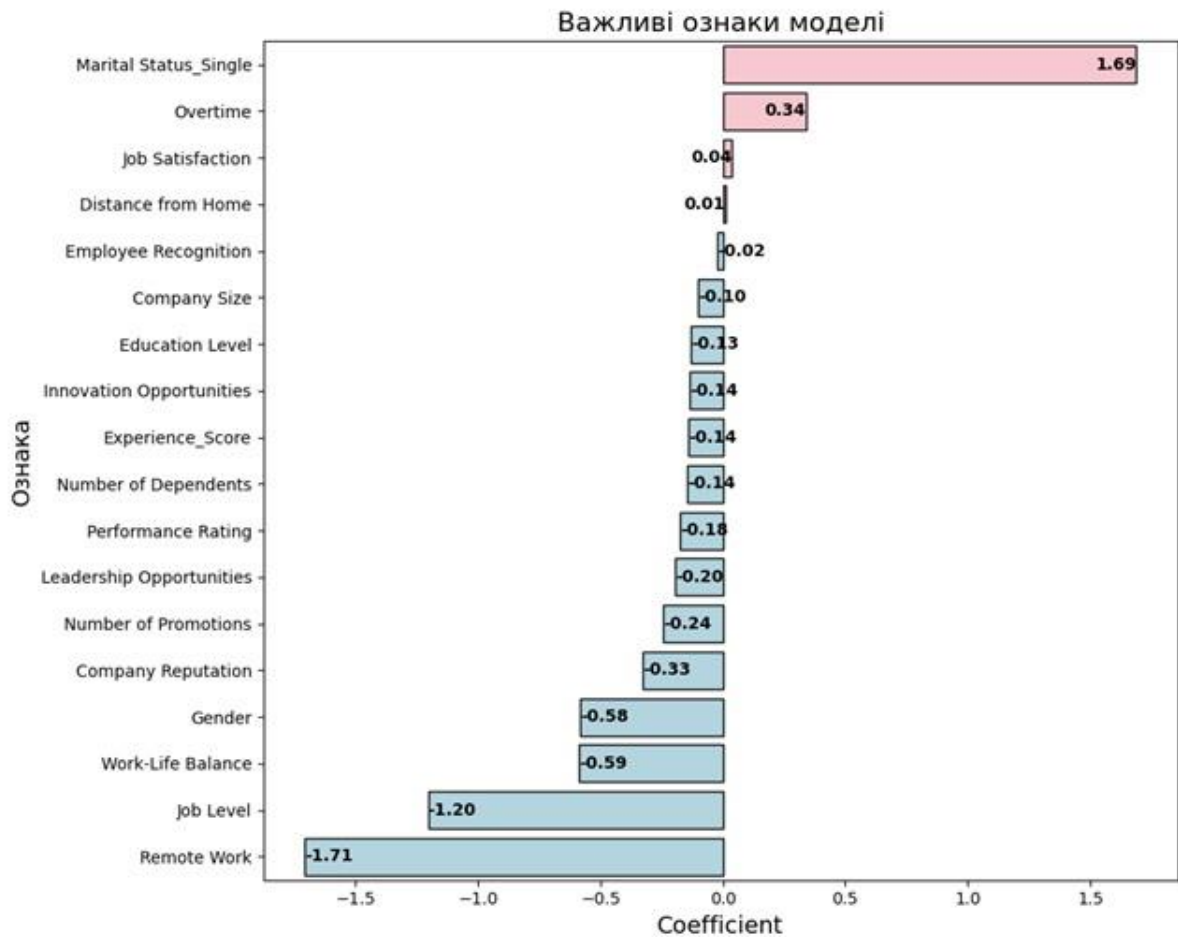


Рисунок 4.3 – Важливість ознак на основі коефіцієнтів RFECV

Баланс між роботою та особистим життям також є вагомим чинником, який знижує ризик звільнення. Працівники, що відчують підтримку цього балансу, демонструють більшу стабільність у своїй роботі. Репутація компанії та оцінка ефективності працівника мають додатковий, хоч і менш виражений, позитивний вплив, який також сприяє утриманню працівників.

Результати аналізу показують, що для зниження ризику звільнення компаніям варто звернути увагу на умови праці, зокрема на можливість роботи віддалено, підтримку балансу між роботою та життям, а також уникати надмірного навантаження через понаднормову роботу. Такі фактори, як репутація компанії, можливості для розвитку та підтримка позитивної робочої атмосфери, також є важливими для забезпечення стабільності працівників.

Інші фактори мають відносно невеликий вплив на рішення про звільнення згідно з даною моделлю. Ці результати надають важливі інсайти для HR-відділів

та керівництва компаній. Вони вказують на необхідність фокусування на розвитку кар'єрних можливостей, забезпеченні гнучких умов праці, покращенні балансу між роботою та особистим життям, а також на підтримці позитивної репутації компанії. Особливу увагу слід приділити розробці стратегій утримання одиноких співробітників, які, згідно з моделлю, мають вищий ризик звільнення.

Наступним етапом стало тестування модуля `model.py`, який відповідає за побудову та навчання моделі машинного навчання. Модуль використовує налаштування з файлу `model_data.pkl` для узгодженої обробки даних. Навчання моделі проводилося з використанням ансамблевого методу `Stacking Classifier`, складеного з моделей на базових налаштуваннях. Результатом роботи модуля є:

- навчена модель машинного навчання;
- об'єкт масштабування даних (`scaler`);
- компонент зниження розмірності даних (`PCA`).

Всі ці компоненти об'єднуються в єдиний файл `model_package.pkl`, який використовується модулем прогнозування для обробки нових даних.

Наступним етапом стало тестування модуля `prognosz.py`, який призначений для роботи з новими даними та формування прогнозів.

Порівняльний аналіз результатів тестування на валідаційному та тестовому наборах даних показав стабільну ефективність класифікатора (табл. 4.1). На валідаційному наборі модель продемонструвала загальну точність (`Accuracy`) 0,7545, показник `ROC AUC Score` склав 0,8453, що свідчить про хорошу здатність моделі розрізняти класи. Метрики `Precision` (0,7431) та `Recall` (0,7391) мають близькі значення, що вказує на збалансованість моделі щодо помилок першого та другого роду.

На тестовому наборі спостерігається незначне підвищення деяких показників, зокрема загальна точність зросла до 0,7630, а показник `ROC AUC Score` склав 0,8563. Аналіз розподілу ймовірностей звільнення показує схожі патерни на обох наборах даних – бімодальний розподіл з піками в областях низьких (0.1–0.2) та високих (0.8–0.9) ймовірностей. Це свідчить про стабільну здатність моделі чітко розділяти працівників на групи з високим та низьким ризиком звільнення (табл. 4.1).

Таблиця 4.1 – Результати тестування розробленої системи прогнозування

	Stacking Classifier	Розподіл ймовірності звільнення
Ефективність моделі	Accuracy: 0.7545 Balanced Accuracy: 0.7538 ROC AUC Score: 0.8453 Average Precision Score: 0.8359 F1 Score: 0.7411 Recall (Sensitivity): 0.7391 Specificity: 0.7684 G-mean: 0.7536 Precision: 0.7431 Confusion Matrix: [[5406 1629] [1663 4712]]	
Тестовий набір	Accuracy: 0.7630 Balanced Accuracy: 0.7618 ROC AUC Score: 0.8563 Average Precision Score: 0.8447 F1 Score: 0.7477 Recall (Sensitivity): 0.7399 Specificity: 0.7838 G-mean: 0.7615 Precision: 0.7557 Confusion Matrix: [[3067 846] [920 2617]]	
Набір Stayed	Accuracy: 0.7838 Balanced Accuracy: 0.3919 ROC AUC Score: Недоступно Average Precision Score: Недоступно F1 Score: 0.0000 Recall (Sensitivity): 0 Specificity: 0.7838 G-mean: 0.0000 Precision: 0.0000 Confusion Matrix: [[3067 846] [0 0]]	
Набір Left	Accuracy: 0.7399 Balanced Accuracy: 0.3699 ROC AUC Score: Недоступно Average Precision Score: Недоступно F1 Score: 0.8505 Recall (Sensitivity): 0.7399 Specificity: 0 G-mean: 0.0000 Precision: 1.0000 Confusion Matrix: [[0 0] [920 2617]]	

Загалом, близькі значення метрик на обох наборах даних та схожі розподіли ймовірностей вказують на стабільність та надійність розробленої моделі, її здатність ефективно працювати з новими даними без суттєвої втрати якості прогнозування.

Для подальшої перевірки робастності моделі було проведено додаткове тестування на двох спеціально підготовлених наборах даних. Набір тестових даних був розділений на дві частини: набір Stayed, що містить лише клас 0 (працівники, які залишилися), і набір Left, що містить лише клас 1 (працівники, які звільнилися).

Результати тестування на наборі Stayed показали, що модель не змогла правильно класифікувати жодного випадку (Recall, Precision, F1 Score дорівнюють 0). Це очікуваний результат, оскільки в цьому наборі відсутні позитивні приклади (клас 1), на яких модель була навчена. Проте, варто відзначити, що модель все ж таки визначила 846 випадки як потенційні звільнення (фальшиві позитивні результати), що може бути корисним для виявлення працівників з високим ризиком звільнення.

На наборі Left модель правильно виявила близько 74% всіх випадків звільнення. F1 Score дорівнює 0,8505, що підтверджує високу ефективність моделі на цьому наборі даних.

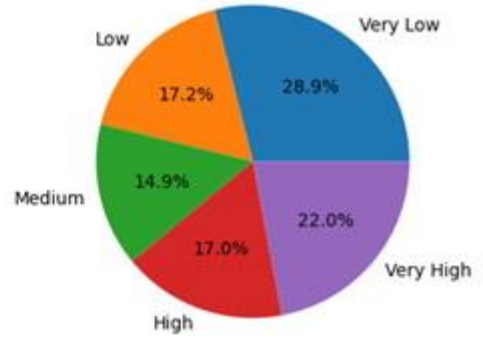
Особливу увагу варто звернути на результати для випадків з ймовірністю звільнення більше 0,8 (група Very High). На тестовому наборі модель виявила 1641 працівника з групи ризику. У наборі Stayed було виявлено лише 180 таких випадків, тоді як у наборі Left – 1461. Це свідчить про високу впевненість моделі у своїх прогнозах щодо звільнення працівників (рис. 4.4).

Загалом, результати тестування демонструють, що розроблена модель має високу точність у прогнозуванні звільнення працівників, особливо для випадків з високою ймовірністю звільнення. Проте, модель також показує певну схильність до фальшивих позитивних результатів, що може бути як перевагою (виявлення потенційних ризиків), так і недоліком (можливі помилкові тривоги) залежно від конкретного застосування.



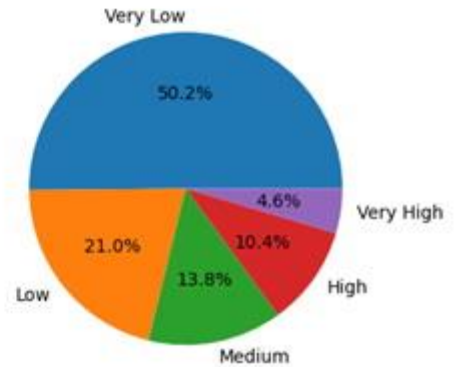
а)

Відсотковий розподіл груп ризику



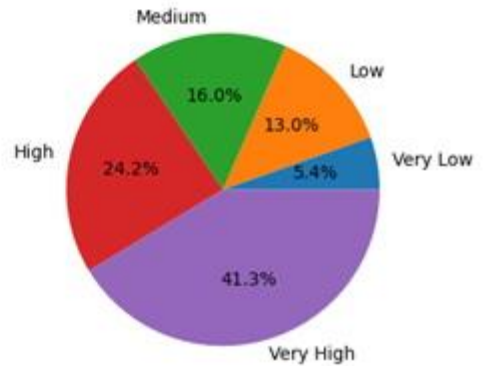
б)

Відсотковий розподіл груп ризику



в)

Відсотковий розподіл груп ризику



а) тестовий набір; б) набір Stayed; в) набір Left

Рисунок 4.4 – Розподіл ймовірності звільнення за групами ризику

Модуль прогнозування генерує комплексну систему звітності, яка включає як деталізовані дані про кожного працівника, так і агреговані звіти за групами ризику (рис. 4.5). В основному звіті full_report.csv для кожного працівника відображається не лише прогнозована ймовірність звільнення та відповідний рівень ризику, але й значення найбільш впливових характеристик (рис. 4.6). Така деталізація дозволяє HR-фахівцям глибше розуміти індивідуальну ситуацію кожного працівника та фактори, що впливають на ризик його звільнення.

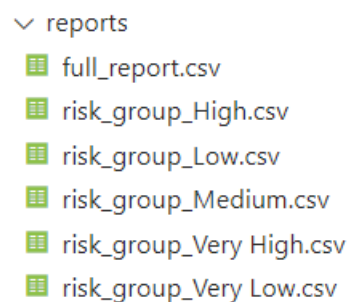


Рисунок 4.5 – Перелік звітів про розподіл працівників за групами ризику на основі ймовірності звільнення

	A	B	C	D	E	F	G	H	I
1	Employee ID	Prediction	Probability	Risk Level	Remote Work	Marital Status_Single	Job Level	Work-Life Balance	Gender
2	49797	0	0.5	Medium	No	ЛОЖЬ	Mid	Poor	Male
3	16721	0	0.4	Medium	No	ЛОЖЬ	Entry	Excellent	Male
4	18956	0	0.15	Very Low	No	ЛОЖЬ	Mid	Good	Male
5	23781	0	0.18	Very Low	Yes	ИСТИНА	Entry	Excellent	Male
6	19694	0	0.08	Very Low	Yes	ЛОЖЬ	Mid	Good	Female
7	46549	1	0.92	Very High	No	ИСТИНА	Entry	Excellent	Female
8	6409	1	0.73	High	No	ЛОЖЬ	Mid	Fair	Male
9	11806	0	0.27	Low	No	ИСТИНА	Senior	Excellent	Male
10	13051	1	0.81	Very High	No	ИСТИНА	Mid	Good	Male
11	48186	1	0.94	Very High	No	ИСТИНА	Entry	Poor	Female
12	22117	1	0.51	Medium	Yes	ЛОЖЬ	Entry	Poor	Female

Рисунок 4.6 – Структура звіту про розподіл працівників за групами ризику на основі ймовірності звільнення

Окремі звіти за групами ризику (Very Low, Low, Medium, High, Very High) допомагають HR-менеджерам ефективно працювати з різними категоріями працівників, розробляючи відповідні програми утримання та розвитку персоналу. Це дозволяє забезпечити чітку класифікацію працівників за рівнем

ризиків, а також гнучко працювати з кожною групою та аналізувати специфічні тенденції в рамках різних ризикових категорій.

Доповнюють звітність візуалізації у вигляді діаграм, які забезпечують наочне представлення розподілу працівників за групами ризику та допомагають оцінювати загальну ситуацію в колективі (рис.4.3). Створюється стовпчаста діаграма, що відображає кількість працівників у кожній з п'яти ризикових груп, та кругова діаграма, яка демонструє відсотковий розподіл працівників за ризиковими групами. Такі візуалізації дозволяють швидко оцінити загальний стан кадрів з точки зору ризику звільнення, а також ідентифікувати найбільш вразливі групи.

Така структурована система звітності забезпечує HR-відділ необхідними інструментами для ефективного управління персоналом та своєчасного реагування на потенційні ризики звільнення працівників.

4.3 Висновки за розділом 4

Розроблено систему прогнозування звільнення працівників з модульною архітектурою, що складається з чотирьох основних компонентів: бібліотеки функцій (`data_processing.py`), модуля аналізу даних (`data_analysis.py`), модуля навчання моделі (`model.py`) та модуля прогнозування (`prognoz.py`). Взаємодія між компонентами організована через проміжні файли `model_data` та `model_package`, що забезпечує узгодженість у підході до обробки даних та прогнозування.

Тестування роботи системи на різних наборах даних (тестовому, `Stayed`, `Left`) підтвердило її високу ефективність. Аналіз важливості ознак за допомогою RFECV регресії виявив ключові фактори впливу на звільнення, такі як сімейний стан, можливість віддаленої роботи та рівень посади, що надає цінну інформацію для HR-відділів.

5 ВИЗНАЧЕННЯ ШЛЯХІВ ПОДАЛЬШОГО РОЗВИТКУ РОЗРОБЛЕНОЇ СИСТЕМИ ПРОГНОЗУВАННЯ

5.1 Визначення переваг, недоліків та основних напрямків вдосконалення розробленої системи

Розроблена система прогнозування звільнення працівників, продемонструвала високу ефективність під час тестування. Для забезпечення подальшого розвитку та вдосконалення системи важливо провести комплексний аналіз її поточного стану, визначити сильні та слабкі сторони реалізації, а також оцінити потенційні можливості та загрози.

Розроблена система характеризується наступними ключовими особливостями.

Модульність:

- чітке розділення функціоналу між компонентами;
- простота тестування та оновлення окремих модулів;
- можливість повторного використання компонентів.

Гнучкість обробки даних:

- централізоване зберігання функцій обробки;
- можливість легкої адаптації до нових вимог;
- масштабованість рішення.

Ефективне управління даними:

- збереження проміжних об'єктів;
- узгоджений підхід до обробки даних;
- чітке розділення навчальних та прогнозних даних.

Практична корисність:

- формування зрозумілих звітів;
- визначення груп ризику;
- можливість пояснення прогнозів.

Для оцінки поточного стану системи, визначення шляхів розвитку, передбачення можливих проблем та планування вдосконалення в роботі було проведено SWOT-аналіз (рис. 5.1).

Сильні сторони (Strengths):	Слабкі сторони (Weaknesses):
1. Модульна архітектура: – чітке розділення функціоналу; – легкість підтримки та модифікації; – можливість повторного використання компонентів. 2. Централізація допоміжних функцій: – єдине місце для базових операцій; – уникнення дублювання коду; – спрощене оновлення функціоналу. 3. Послідовна обробка даних: – чіткий процес від аналізу до прогнозування; – збереження проміжних результатів; – можливість контролю на кожному етапі. 4. Незалежність модулів: – можливість паралельної розробки; – простота тестування; – легка заміна окремих компонентів.	1. Залежність від проміжних файлів: – необхідність зберігання додаткових файлів; – можливі проблеми з версійністю файлів; – ризик пошкодження/втрати файлів. 2. Відсутність системи логування: – складність відстеження помилок; – обмежені можливості моніторингу; – складність діагностики проблем. 3. Обмежена валідація даних: – відсутність централізованої перевірки вхідних даних; – можливі проблеми з якістю даних; – ризик некоректної обробки. 4. Відсутність паралельної обробки: – послідовне виконання операцій; – обмежена продуктивність; – неефективне використання ресурсів.
Можливості (Opportunities):	Загрози (Threats):
1. Розширення функціоналу: – додавання нових методів аналізу; – впровадження додаткових моделей; – розширення можливостей візуалізації. 2. Оптимізація процесів: – впровадження паралельної обробки; – покращення алгоритмів; – оптимізація використання пам'яті. 3. Покращення взаємодії: – додавання API; – інтеграція з іншими системами; – створення веб-інтерфейсу. 4. Вдосконалення безпеки: – впровадження системи авторизації; – шифрування даних; – аудит доступу.	1. Технічні ризики: – проблеми сумісності версій Python/бібліотек; – обмеження масштабованості; – технічний борг. 2. Якість даних: – неповні або некоректні вхідні дані; – зміни в форматі вхідних даних; – проблеми з кодуванням. 3. Безпека: – доступ до конфіденційних даних; – можливі витіки інформації; – вразливості системи. 4. Обслуговування: – складність підтримки при розширенні; – залежність від специфічних знань; – потреба в регулярному оновленні.

Рисунок 5.1 – SWOT-аналіз розробленої системи прогнозування

Проведений аналіз виявив ряд суттєвих переваг у її архітектурі. Основною перевагою є модульна структура, яка забезпечує чітке розділення функціоналу та полегшує підтримку системи. Централізація допоміжних функцій в окремому модулі `data_processing.py` дозволяє уникнути дублювання коду та спрощує процес оновлення базового функціоналу. Важливою перевагою також є послідовна обробка даних з можливістю контролю на кожному етапі та збереженням проміжних результатів. Незалежність модулів системи забезпечує можливість паралельної розробки та спрощує процес тестування окремих компонентів.

Однак, дослідження також виявило певні недоліки системи. Значна залежність від проміжних файлів створює ризики, пов'язані з їх зберіганням та версійністю. Відсутність системи логування ускладнює процес відстеження помилок та діагностики проблем у роботі системи. Суттєвим недоліком є обмежена валідація вхідних даних, що може призвести до проблем з якістю обробки інформації. Послідовне виконання операцій без можливості паралельної обробки обмежує продуктивність системи та ефективність використання обчислювальних ресурсів.

За результатами проведеного аналізу було визначено шляхи вдосконалення системи. Пріоритетним є впровадження інтегрованого рішення для моніторингу та логування, яке дасть змогу відстежувати функціонування системи та своєчасно усувати можливі проблеми. Не менш суттєвим є створення комплексної перевірки даних та запровадження системи версійності проміжних файлів. Для оптимізації роботи доцільно реалізувати механізми паралельної обробки даних.

Перспективним напрямком розвитку є створення програмного інтерфейсу (API) для інтеграції з іншими системами та розробка веб-інтерфейсу для зручності користування. Особливу увагу слід приділити питанням безпеки: впровадження системи авторизації, шифрування даних та аудиту доступу. Для забезпечення надійності роботи необхідно розробити систему резервного копіювання та відновлення даних.

Реалізація запропонованих вдосконалень дозволить підвищити надійність, безпеку та ефективність системи, розширити її функціональні можливості та забезпечити більш зручне використання. Модульна архітектура системи створює хорошу основу для впровадження цих покращень без необхідності кардинальної перебудови існуючої структури.

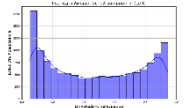
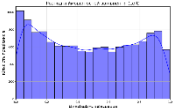
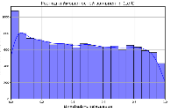
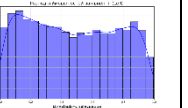
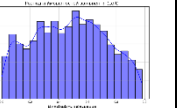
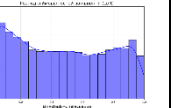
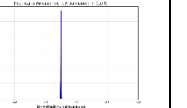
5.2 Подальший розвиток розробленої системи прогнозування

Для реалізації одного з напрямів подальшого розвитку розробленої системи прогнозування було проведено порівняльне дослідження з використанням різних класифікаторів. Такий підхід дозволяє не лише визначити найбільш підходящий алгоритм для конкретної задачі, але й оцінити стійкість результатів та виявити потенційні слабкі місця в обробці даних чи моделюванні.

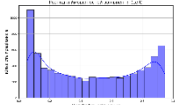
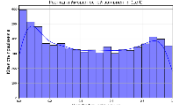
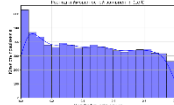
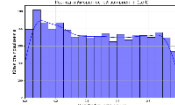
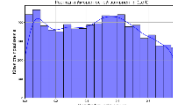
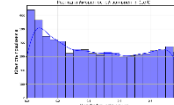
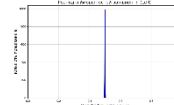
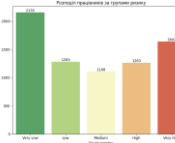
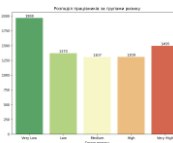
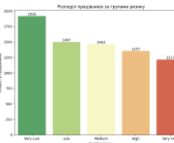
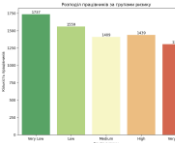
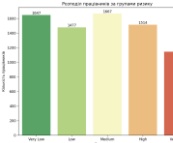
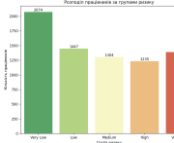
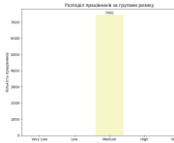
Порівняльний аналіз результатів прогнозування на валідаційному наборі даних показав, що всі досліджені моделі продемонстрували достатньо високу та стабільну якість прогнозування, з показниками ROC AUC Score 0,83 – 0,85, що свідчить про надійність обраного підходу до обробки даних та формування ознак (табл. 5.1). Аналіз розподілу ймовірностей звільнення та матриць помилок показує, що всі моделі демонструють схожі патерни в прогнозуванні, що підтверджується близькими значеннями метрик Precision та Recall. Це свідчить про те, що основні закономірності в даних успішно виявляються різними алгоритмами.

Дослідження показало, що особливу увагу при виборі класифікатора варто приділяти аналізу розподілу прогнозованих ймовірностей звільнення. Яскравим прикладом важливості такого аналізу є результати AdaBoostClassifier. Незважаючи на високі показники точності ($\text{ROC AUC} = 0,85$), цей класифікатор виявив суттєвий недолік – нездатність ідентифікувати випадки з високою ймовірністю звільнення (більше 0,8), що обмежує можливості системи щодо виявлення працівників з найвищим ризиком звільнення. Це робить його непридатним для практичного використання, оскільки обмежує можливості системи щодо виявлення працівників з найвищим ризиком звільнення.

Таблиця 5.1 – Результати тестування розробленої системи прогнозування на різних класифікаторах

Класифікатор	Stacking (оптиміз.)	Gradient Boost	CatBoost	LGBM	Logistic Regression	Random Forest	MLP	AdaBoost
Ефективність моделі на валідаційному наборі даних								
Час оптимізації Randomized SearchCV	2613,8545	852,1347	1054,6521	217,0202	27,5738	561,6284	2916,1577	584,0277
Час оптимізації GridSearchCV	208,0285	26,0182	88,7411	4,6450	1,3690	93,3756	213,9620	40,1548
Час навчання моделі	305,6151	11,1144	3,3504	1,2232	0,5179	44,1696	75,2652	20,7820
Accuracy:	0,7538	0,7556	0,7540	0,7558	0,7400	0,7500	0,7477	0,7558
Balanced Accuracy:	0,7531	0,7549	0,7521	0,7551	0,7388	0,7492	0,7458	0,7550
ROC AUC Score	0,8451	0,8458	0,8460	0,8452	0,8253	0,8381	0,8400	0,8469
Average Precision Score:	0,8361	0,8368	0,8374	0,8365	0,8142	0,8281	0,8300	0,8371
F1 Score:	0,7406	0,7422	0,7339	0,7428	0,7236	0,7360	0,7270	0,7419
Recall (Sensitivity):	0,7395	0,7401	0,7137	0,7420	0,7161	0,7332	0,7065	0,7382
Precision:	0,7418	0,7444	0,7553	0,7437	0,7313	0,7389	0,7487	0,7456
Confusion Matrix:	[[5394 1641] [1661 4714]]	[[5415 1620] [1657 4718]]	[[5561 1474] [1825 4550]]	[[5405 1630] [1645 4730]]	[[5358 1677] [1810 4565]]	[[5383 1652] [1701 4674]]	[[5523 1512] [1871 4504]]	[[5429 1606] [1669 4706]]
Середній Average Precision на крос-валідації	0,8393 (+/- 0,0071)	0,8393 (+/- 0,0075)	0,8400 (+/- 0,0073)	0,8385 (+/- 0,0078)	0,8147 (+/- 0,0115)	0,8312 (+/- 0,0068)	0,8314 (+/- 0,0101)	0,8390 (+/- 0,0052)
Розподіл ймовірності звільнення								

Кінець таблиці 5.1

Класифікатор	Stacking	Gradient Boost	CatBoost	LGBM	Logistic Regression	Random Forest	MLP	AdaBoost
Ефективність моделі на тестовому наборі								
Accuracy	0,7779	0,7634	0,7636	0,7660	0,7470	0,8223	0,7608	0,7617
Balanced Accuracy	0,7769	0,7622	0,7614	0,7650	0,7457	0,8218	0,7585	0,7606
ROC AUC Score	0,8726	0,8592	0,8578	0,8587	0,8311	0,9118	0,8536	0,8560
Average Precision Score	0,8632	0,8482	0,8463	0,8474	0,8160	0,9035	0,8411	0,8449
F1 Score	0,7641	0,7478	0,7421	0,7513	0,7301	0,8127	0,7390	0,7465
Recall (Sensitivity)	0,7577	0,7390	0,7164	0,7444	0,7207	0,8123	0,7133	0,7388
Precision	0,7706	0,7568	0,7697	0,7584	0,7397	0,8132	0,7666	0,7543
Confusion Matrix	[[3115 798] [857 2680]]	[[3073 840] [923 2614]]	[[3155 758] [1003 2534]]	[[3074 839] [904 2633]]	[[3016 897] [988 2549]]	[[3253 660] [664 2873]]	[[3145 768] [1014 2523]]	[[3062 851] [924 2613]]
Розподіл ймовірності звільнення								
Розподіл за групами ризику (Count/Percentage)								
Very Low	2156/ 28,94	1969/ 26,43	1916/ 25,72	1925/ 25,84	1737/ 23,32	1647/ 22,11	2074/ 27,84	0/ 0,00
Low	1284/ 17,23	1370/ 18,39	1497/ 20,09	1365/ 18,32	1559/ 20,93	1477/ 19,83	1447/ 19,42	0/ 0,00
Medium	1092/ 14,66	1307/ 17,54	1463/ 19,64	1376/ 18,47	1409/ 18,91	1667/ 22,38	1304/ 17,50	7450/ 100,00
High	1255/ 16,85	1309/ 17,57	1357/ 18,21	1345/ 18,05	1439/ 19,32	1514/ 20,32	1235/ 16,58	0/ 0,00
Very High	1663/ 22,32	1495/ 20,07	1217/ 16,34	1439/ 19,32	1306/ 17,53	1145/ 15,37	1390/ 18,66	0/ 0,00
Розподіл за групами ризику								

Особливу увагу варто звернути на часові характеристики навчання та оптимізації моделей. При використанні RandomizedSearchCV найбільші часові витрати показали MLP Classifier (2916 секунд) та CatBoost (1055 секунд), тоді як Light Gradient Boosting Machine (LGBM) та логістична регресія виявилися найшвидшими (217 та 28 секунд відповідно). При застосуванні GridSearchCV спостерігається аналогічна тенденція, але з суттєво меншими часовими витратами: від 0,5 секунди для логістичної регресії до 214 секунд для MLP.

Для оцінки ефективності стекінгового класифікатора було проведено порівняльне тестування моделі з базовими налаштуваннями та після оптимізації гіперпараметрів. Модель з базовими налаштуваннями потребувала 52,44 секунди на навчання та продемонструвала достатньо високі показники ефективності на тестовому наборі: ROC AUC Score склав 0,8453, Accuracy – 0,7547. На крос-валідації модель показала стабільні результати з середнім значенням Average Precision 0,8380 ($\pm 0,0068$).

Застосування методів оптимізації гіперпараметрів значно збільшило час навчання моделі: RandomizedSearchCV потребував 2613,85 секунд, GridSearchCV – 208,03 секунд, а час навчання оптимізованої моделі склав 305,62 секунди. При цьому показники ефективності залишились практично незмінними: ROC AUC Score – 0,8451, Accuracy – 0,7538. Середнє значення Average Precision на крос-валідації незначно покращилось до 0,8393 ($\pm 0,0071$).

Такі результати свідчать про те, що базова конфігурація стекінгового класифікатора вже забезпечує достатньо високу ефективність, а оптимізація гіперпараметрів, незважаючи на значне збільшення обчислювальних витрат, не призводить до суттєвого покращення якості прогнозування. Це може вказувати на те, що для даної задачі базова конфігурація є оптимальним вибором з точки зору балансу між ефективністю та обчислювальними витратами.

При тестуванні різних класифікаторів на валідаційному та тестовому наборах даних було проведено порівняльний аналіз їх ефективності. На валідаційному наборі всі моделі продемонстрували досить стабільні результати. Показник ROC AUC Score для різних класифікаторів знаходиться в діапазоні 0,82–0,85, загальна точність (Accuracy) коливається близько 0,75, а метрики

Precision та Recall демонструють збалансовані значення в межах 0,72–0,75. При цьому найбільш стабільні результати за всіма метриками показали градієнтні методи (Gradient Boost, LGBM) та Stacking класифікатор (рис. 5.2).

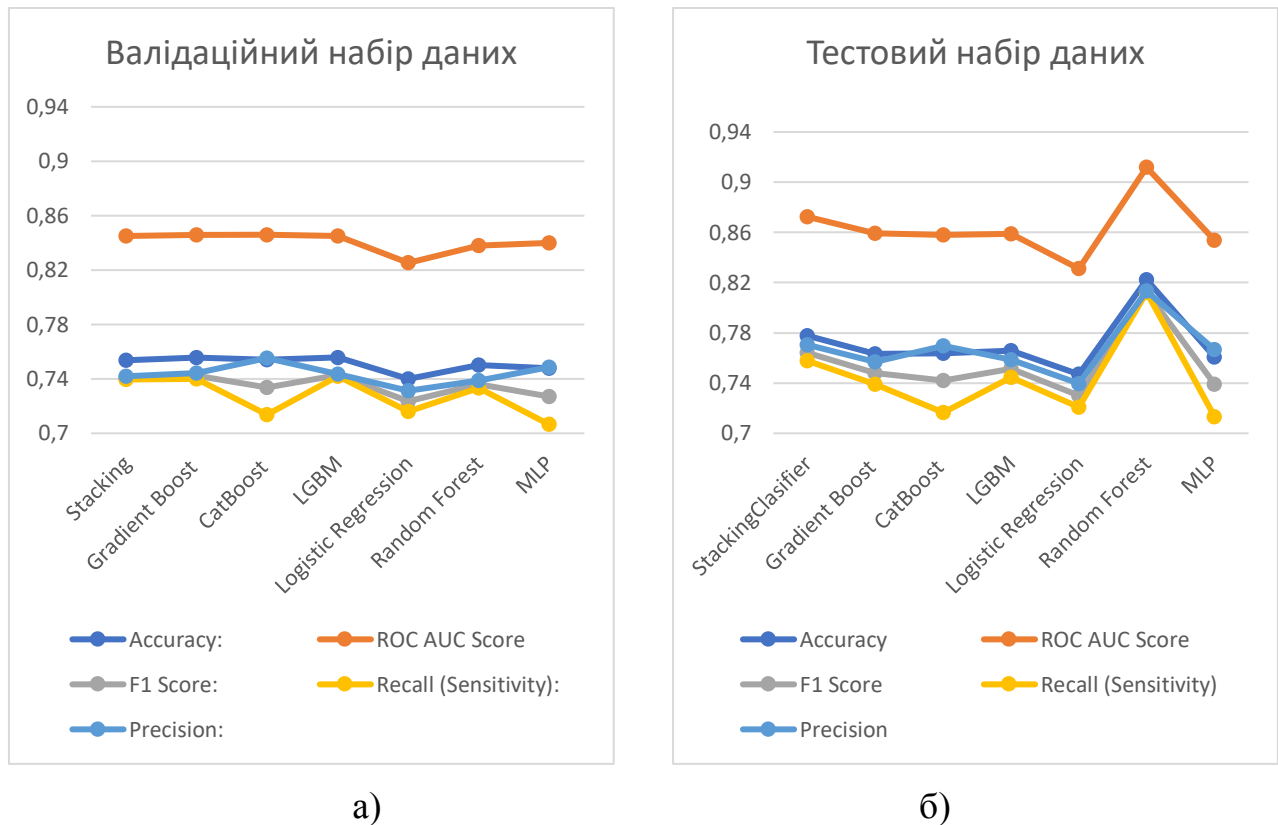


Рисунок 5.2 – Порівняльний аналіз ефективності класифікаторів на валідаційному (а) та тестовому (б) наборах даних

Результати тестування на тестовому наборі даних виявили більш виражені відмінності між моделями. Random Forest продемонстрував значне покращення всіх метрик, досягнувши ROC AUC Score 0,91 та найвищих показників Ассурасу (0,82) і F1 Score (0,82). Градієнтні методи зберегли стабільність показників з невеликим покращенням ROC AUC Score до 0,86. Найнижчі результати на обох наборах даних показала логістична регресія.

Спостережувана різниця в результатах тестування, особливо значне покращення показників Random Forest на тестовому наборі, може бути пов'язана з особливостями розподілу даних у тестовій вибірці. Проте з точки зору практичного застосування більш надійним є орієнтування на результати валідаційного набору, де моделі демонструють більш реалістичні та стабільні

показники. Враховуючи це, градієнтні методи та Stacking класифікатор можна вважати найбільш надійним вибором для впровадження в систему прогнозування звільнення працівників.

Враховуючи баланс між якістю прогнозування та обчислювальною ефективністю, найбільш оптимальним вибором для практичного застосування виглядає Light Gradient Boosting Machine (LGBM). Цей алгоритм забезпечує високу точність прогнозування при порівняно невеликих часових витратах на оптимізацію та навчання.

Для підтвердження стабільності роботи класифікаторів було проведено оцінку на крос-валідації, зокрема проаналізовано середнє значення Average Precision та його варіативність (рис. 5.3).

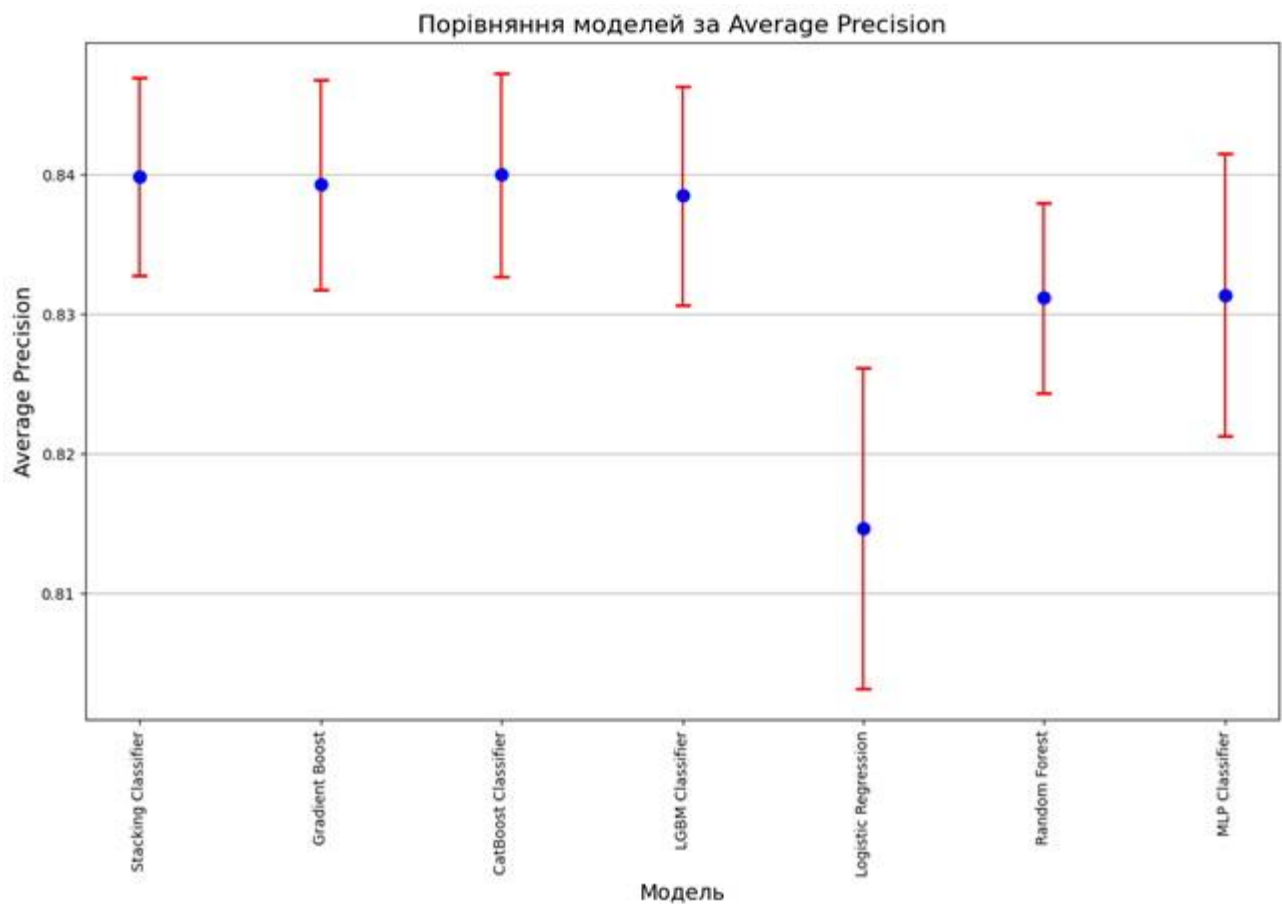


Рисунок 5.3 – Результати на крос-валідації за середнім значенням Average Precision

Всі градієнтні методи показують високу стабільність на крос–валідації. Stacking класифікатор демонструє схожу стабільність, тоді як Logistic Regression та MLP показують більшу варіативність результатів, що може свідчити про меншу надійність при роботі з різними наборами даних.

Стабільні результати крос–валідації LGBM з середнім значенням Average Precision 0.8385 (± 0.0078) свідчить про надійність моделі, що додатково підтверджує її перевагу як оптимального вибору для практичного застосування. Класифікатор забезпечує не лише найвищу точність, але й стабільність прогнозів при роботі з різними даними.

Таким чином, враховуючи всі критерії – високу точність прогнозування, здатність ефективно розподіляти працівників за групами ризику та прийнятний час налаштування гіперпараметрів та навчання, Light Gradient Boosting Machine є найбільш оптимальним вибором для практичного застосування в системі прогнозування звільнення працівників.

В результаті проведеного дослідження було розроблено модифікований модуль побудови та навчання моделі на базі класифікатора Light Gradient Boosting Machine. Модифікований модуль, який містить реалізацію двох підходів до оптимізації гіперпараметрів – RandomizedSearchCV та GridSearchCV, представлений в додатку А.

Модульна архітектура розробленої системи, зокрема виділення класифікатора в окремий модуль, є її важливою особливістю. Така організація системи забезпечує гнучкість у виборі та заміні алгоритму машинного навчання без необхідності змін в інших компонентах системи. Це дозволяє адаптувати систему під потреби організації, наприклад, обираючи більш швидкий алгоритм для обробки великих обсягів даних або більш точний – для критично важливих прогнозів.

Проведене дослідження різних класифікаторів підтверджує ефективність такого підходу, демонструючи можливість досягнення високої якості прогнозування з використанням різних алгоритмів при збереженні загальної структури системи.

5.3 Висновки за розділом 5

Проведений SWOT-аналіз системи виявив її ключові переваги (модульність, централізація функцій, послідовна обробка даних) та визначив основні напрямки вдосконалення, включаючи впровадження системи логування, покращення валідації даних, розробку механізмів паралельної обробки та створення API для інтеграції.

Порівняльне дослідження різних класифікаторів показало, що Light Gradient Boosting Machine забезпечує оптимальний баланс між швидкістю та точністю прогнозування. На основі цього дослідження розроблено модифікований модуль побудови та навчання моделі, що дозволяє гнучко обирати метод оптимізації гіперпараметрів залежно від вимог до швидкодії системи.

ВИСНОВКИ

У роботі вирішено актуальне завдання розробки системи прогнозування звільнення працівників на основі методів штучного інтелекту. Основні результати роботи полягають у наступному.

Проведений аналіз існуючих підходів до прогнозування звільнення працівників виявив необхідність розробки комплексного рішення, яке б враховувало широкий спектр характеристик працівників та забезпечувало високу точність прогнозування.

Розроблено систему прогнозування з модульною архітектурою, що включає чотири основні компоненти: бібліотеку функцій, модулі аналізу даних, навчання моделі та прогнозування. Взаємодія між компонентами реалізована через проміжні файли, що забезпечує узгодженість обробки даних та можливість контролю на кожному етапі.

Тестування системи на різних наборах даних підтвердило її високу ефективність. Проведений SWOT-аналіз виявив ключові переваги системи (модульність, централізація функцій, послідовна обробка даних) та визначив перспективні напрями вдосконалення.

На основі порівняльного аналізу різних методів машинного навчання встановлено, що найвищу ефективність демонструють ансамблеві методи, зокрема градієнтний бустинг. Розроблено модифікований модуль навчання моделі на базі Light Gradient Boosting Machine, який забезпечує оптимальний баланс між швидкодією та точністю прогнозування.

Практичне значення отриманих результатів полягає в можливості використання розробленої системи для ефективного прогнозування ризику звільнення працівників та прийняття обґрунтованих рішень в HR-менеджменті. Модульна архітектура системи забезпечує її гнучку адаптацію під потреби різних організацій.

Подальші дослідження можуть бути спрямовані на розширення функціональності системи через впровадження API, розробку веб-інтерфейсу та вдосконалення механізмів валідації даних і системи логування.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Vrontis, D. Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review / D. Vrontis, M. Christofi, V. Pereira et al. // *The International Journal of Human Resource Management*. – 2021. – P. 1–30.
2. Palos–Sánchez, P. R. Artificial Intelligence and Human Resources Management: A Bibliometric Analysis [Electronic resource] / P. R. Palos–Sánchez, P. Baena–Luna, A. Badicu, et al. // *Applied Artificial Intelligence*. – 2022. – № 36 (1). – Access mode: <https://doi.org/10.1080/08839514.2022.2145631>.
3. Rakha, A. N. Artificial Intelligence and HR Management [Electronic resource] / A. N. Rakha // *International Journal of Management and Finance*. – 2023. – № 1(1). – Access mode: <https://irshadjournals.com/index.php/ijmf/article/view/25>.
4. Menaka, R. Role of Artificial Intelligence (AI) in Human Resource Management (HRM) in Recent Era / R. Menaka // *Shanlax International Journal of Management*. – 2023. – № 2(11). – P. 32–38.
5. Sha, Ri Na. Application of Artificial Intelligence in Recruitment and Selection / Sha Ri Na // *Academic Journal of Science and Technology*. – 2024. – № 9(2). – P. 56–58.
6. Woods, S. A. Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges / S. A. Woods, S. Ahmed, I. Nikolaou et al. // *European Journal of Work and Organizational Psychology*. – 2019. – № 29(1). – P. 64–77.
7. Dennis, M. J. The Ethics of AI in Human Resources / M. J. Dennis, E. Aizenberg // *Business Horizons*. – 2022. – № 65(4). – P. 567–578.
8. Кравчук, О. І. Трансформація HR менеджменту через призму штучного інтелекту: комплексний аналіз імплементації, викликів та етичних аспектів [Електронний ресурс] / О. І. Кравчук, І. О. Варіс, Н. Г. Каленська // *Проблеми сучасних трансформацій. Серія: економіка та управління*. – 2024. – № 15. – Режим доступу: https://reicst.com.ua/pmt/issue/view/issue_15_2024.

9. Кравчук, О. І. Цифровізація менеджменту персоналу: концептуальні аспекти та тенденції [Електронний ресурс] / О. І. Кравчук, І. О. Варіс, К. О. Рубель // Проблеми сучасних трансформацій. Серія: економіка та управління. – 2024. – № 12. – Режим доступу: <https://reicst.com.ua/pmt/article/view/2024-12-07-04>.
10. Бей, Г. В. Трансформація HR–технологій під впливом цифровізації бізнес–процесів / Г. В. Бей, Г. В. Середа // Економіка і організація управління. – 2019. – №2. – С. 93–101.
11. Чернишова, О. О. Вплив штучного інтелекту на бізнес-процеси з метою оптимізації та покращення ефективності роботи організації / О. О. Чернишова, С. В. Домашенко, Д. Г. Домашенко // Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки. – 2024. – 35(74). – №2. – С. 196–204.
12. Черненко Н. Штучний інтелект в управлінні персоналом / Н. Черненко // Таврійський науковий вісник. Серія: Економіка. – 2022. – № 12. – С.76–83.
13. Державна служба статистики України [Електронний ресурс]. – Режим доступу: <https://www.ukrstat.gov.ua/>.
14. Найкращі інструменти штучного інтелекту для HR у 2024 році: повне керівництво [Електронний ресурс]. – Режим доступу: <https://ukr.pritula.academy/tpost/frm9ju8111-naikrasch-nstrumenti-shtuchnogo-ntelektu>.
15. AI in HR (2022). HR encyclopedia Eddy [Electronic resource]. – Access mode: <https://eddy.com/hr-encyclopedia/ai-in-hr/>.
16. PeopleForce [Electronic resource]. – Access mode: <https://peopleforce.io/uk>.
17. Hurma System [Electronic resource]. – Access mode: <https://hurma.work/>.
18. KISSFLOW HR Cloud [Electronic resource]. – Access mode: <https://kissflow.com/hr>.

19. Culture Amp [Electronic resource]. – Access mode: <https://www.cultureamp.com/>.
20. BambooHR [Electronic resource]. – Access mode: <https://www.bamboohr.com/>.
21. Піскун, О. В. Аналіз алгоритмів машинного навчання для задачі бінарної класифікації / О. В. Піскун // Вісник Черкаського національного університету імені Богдана Хмельницького. Серія «Прикладна математика. Інформатика». – 2019. – Вип. № 2. – С. 86–95.
22. Bagging, Boosting, and Stacking in Machine Learning [Електронний ресурс]. – Access mode: <https://www.baeldung.com/cs/bagging-boosting-stacking-ml-ensemble-models>.
23. Supervised Learning algorithms cheat sheet [Електронний ресурс]. – Access mode: <https://towardsdatascience.com/supervised-learning-algorithms-cheat-sheet-40009e7f29f5>.
24. Employee Attrition Classification Dataset [Електронний ресурс]. – Access mode: <https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data>

ДОДАТОК А**Фрагмент тексту програми**

```

# model.py
import pandas as pd
import numpy as np
import joblib
import timeit
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler
from lightgbm import LGBMClassifier
from sklearn.model_selection import RandomizedSearchCV, GridSearchCV
from sklearn.decomposition import PCA
from data_processing import (
    preprocess_data,
    evaluate_model_performance,
    plot_probability_distribution,
    balance_dataset,
    prepare_features,
    plot_roc_curve,
    plot_precision_recall_curve,
)
import warnings
warnings.filterwarnings('ignore')

print("Завантаження та попередня обробка даних...")

# Завантаження набору даних
hr_data = pd.read_csv('train_data.csv')
columns_sorted = ['Attrition'] + sorted([col for col in hr_data.columns if col != 'Attrition'])
hr_data = hr_data[columns_sorted]

# Попередня обробка набору даних
hr_data = preprocess_data(hr_data)
emp_data = hr_data.copy()

# Поділ на тренувальний і тестовий набори
train, test = train_test_split(emp_data, test_size=0.2, random_state=42)

# Завантаження та застосування моделі обробки даних
model_data = joblib.load('model_data.pkl')
train_processed = prepare_features(train, model_data)
test_processed = prepare_features(test, model_data)
print("Попередня обробка даних завершена")
print('=' * 30)

# Підготовка даних для навчання
X_train = train_processed.drop(columns=['Attrition'])
y_train = train_processed['Attrition']
X_test = test_processed.drop(columns=['Attrition'])
y_test = test_processed['Attrition']

# Балансування тренувального набору ('smote', 'undersample', 'combined', 'not')
print("\nБалансування тренувального набору даних...")
X_train_balanced, y_train_balanced = balance_dataset(X_train, y_train, method='not')

# Масштабування числових колонок
numeric_features = X_train_balanced.select_dtypes(include=['float64', 'int64']).columns
numeric_features = [col for col in numeric_features if X_train_balanced[col].nunique() > 10]
if numeric_features:
    scaler = StandardScaler()
    X_train_balanced[numeric_features] = scaler.fit_transform(X_train_balanced[numeric_features])
    X_test[numeric_features] = scaler.transform(X_test[numeric_features])

# Створення ознаки PCA
experience_features = model_data['experience_features']

```

```

pca_experience = PCA(n_components=1)

# Застосування PCA для тренувального та тестового наборів
X_train_balanced['Experience_Score'] = pca_experience.fit_transform(X_train_balanced[experience_features])
X_test['Experience_Score'] = pca_experience.transform(X_test[experience_features])

# Видалення початкових ознак з обох наборів
X_train_balanced = X_train_balanced.drop(columns=experience_features)
X_test = X_test.drop(columns=experience_features)

# Оптимізація гіперпараметрів з RandomizedSearchCV
param_dist = {
    'num_leaves': np.arange(10, 100, 10),
    'max_depth': np.arange(3, 10),
    'learning_rate': np.linspace(0.01, 0.3, 10),
    'n_estimators': np.arange(100, 1000, 100),
    'subsample': np.linspace(0.6, 1.0, 5),
    'colsample_bytree': np.linspace(0.6, 1.0, 5),
    'reg_alpha': np.linspace(0, 1, 5),
    'reg_lambda': np.linspace(0, 1, 5)}

base_model = LGBMClassifier(random_state=42, verbose=-1)

# RandomizedSearchCV
print("\nОптимізація гіперпараметрів за допомогою RandomizedSearchCV...")
random_search = RandomizedSearchCV(
    estimator=base_model,
    param_distributions=param_dist,
    n_iter=50,
    cv=StratifiedKFold(n_splits=5, shuffle=True, random_state=42),
    scoring='average_precision',
    random_state=42,
    n_jobs=-1)

start_time = timeit.default_timer()
random_search.fit(X_train_balanced, y_train_balanced)
work_time_random_search = timeit.default_timer() - start_time

print(f"\nНайкращі параметри після RandomizedSearchCV:")
for param, value in random_search.best_params_.items():
    print(f'{param}: {value}')
print(f"\nЧас оптимізації RandomizedSearchCV: {work_time_random_search:.4f} секунд")

# Додаткова оптимізація за допомогою GridSearchCV
param_grid = {
    'num_leaves': [random_search.best_params_['num_leaves']],
    'max_depth': [random_search.best_params_['max_depth']],
    'learning_rate': [random_search.best_params_['learning_rate']],
    'n_estimators': [random_search.best_params_['n_estimators']],
    'subsample': [random_search.best_params_['subsample']],
    'colsample_bytree': [random_search.best_params_['colsample_bytree']],
    'reg_alpha': [random_search.best_params_['reg_alpha']],
    'reg_lambda': [random_search.best_params_['reg_lambda']]

print("\nОптимізація гіперпараметрів за допомогою GridSearchCV...")
grid_search = GridSearchCV(
    estimator=base_model,
    param_grid=param_grid,
    cv=StratifiedKFold(n_splits=5, shuffle=True, random_state=42),
    scoring='average_precision',
    n_jobs=-1)

start_time = timeit.default_timer()
grid_search.fit(X_train_balanced, y_train_balanced)

```

```

work_time_grid_search = timeit.default_timer() - start_time

print(f"\nНайкращі параметри після GridSearchCV:")
for param, value in grid_search.best_params_.items():
    print(f"{param}: {value}")
print(f"\nЧас оптимізації GridSearchCV: {work_time_grid_search:.4f} секунд")

# Навчання моделі з оптимізованими параметрами
model = LGBMClassifier(**grid_search.best_params_, random_state=42)

start_time = timeit.default_timer()
model.fit(X_train_balanced, y_train_balanced)
work_time_model = timeit.default_timer() - start_time

print(f"\nЧас навчання моделі: {work_time_model:.4f} секунд")

# Об'єднання scaler, PCA та моделі в один об'єкт
model_package = {
    'scaler': scaler,
    'pca': pca_experience,
    'model': model
}

# Збереження об'єданого об'єкта в один файл
joblib.dump(model_package, 'model_with_scaler_and_pca.pkl')
print(f"\nМодель збережено у файл: model_with_scaler_and_pca.pkl")

# Оцінка якості моделі
predictions = model.predict(X_test)
probabilities = model.predict_proba(X_test)
probability_left = probabilities[:, 1]

# Розрахунок та виведення метрик
print("\nОцінка якості моделі на тестовому наборі:")
metrics = evaluate_model_performance(y_test, predictions, probability_left)

# Крос-валідація моделі
print("\nПроведення крос-валідації...")
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cv_scores = cross_val_score(model, X_train_balanced, y_train_balanced, cv=cv, scoring='average_precision')
print(f"\nСередній Average Precision на крос-валідації: {cv_scores.mean():.4f} (+/- {cv_scores.std() * 2:.4f})")

# Візуалізація результатів
plot_probability_distribution(probability_left)

# Візуалізація ROC-кривої
plot_roc_curve(y_test, probability_left, title='ROC Curve for Model')

# Візуалізація Precision-Recall кривої
plot_precision_recall_curve(y_test, probability_left, title='Precision-Recall Curve')

```

ДОДАТОК Б
Слайди презентації

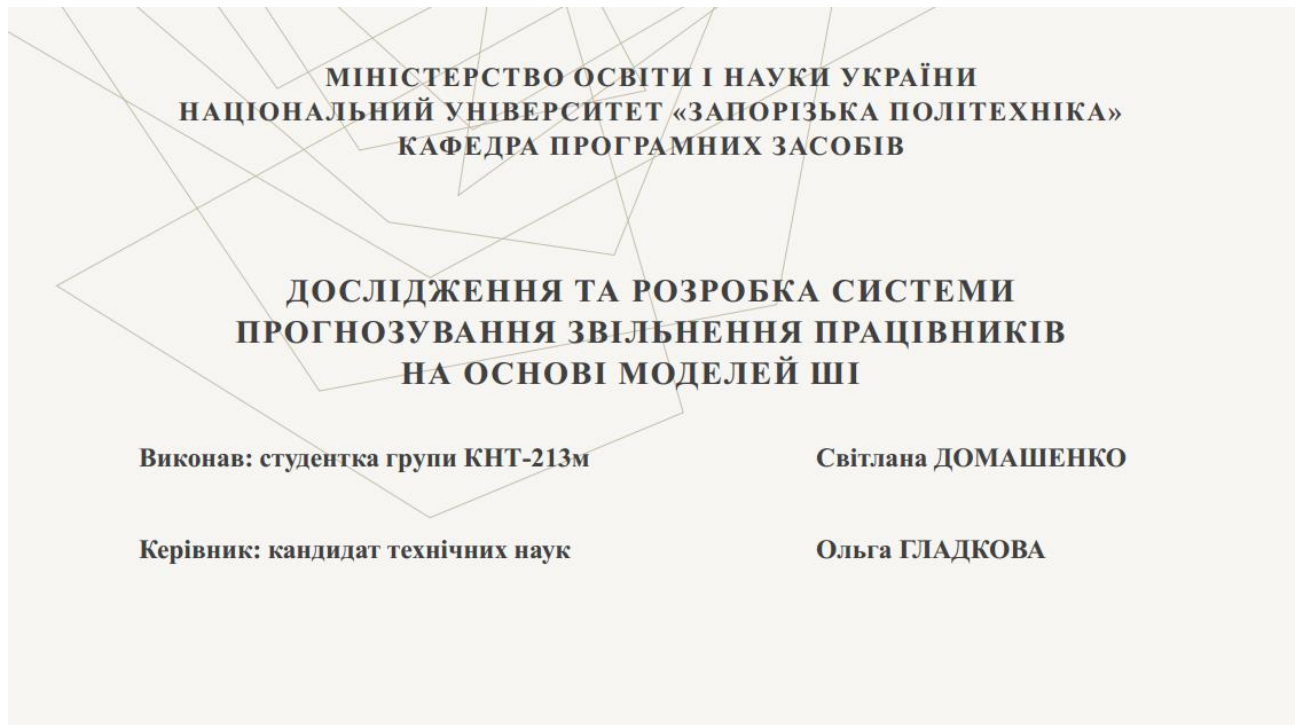


Рисунок Б.1 – Слайд 1

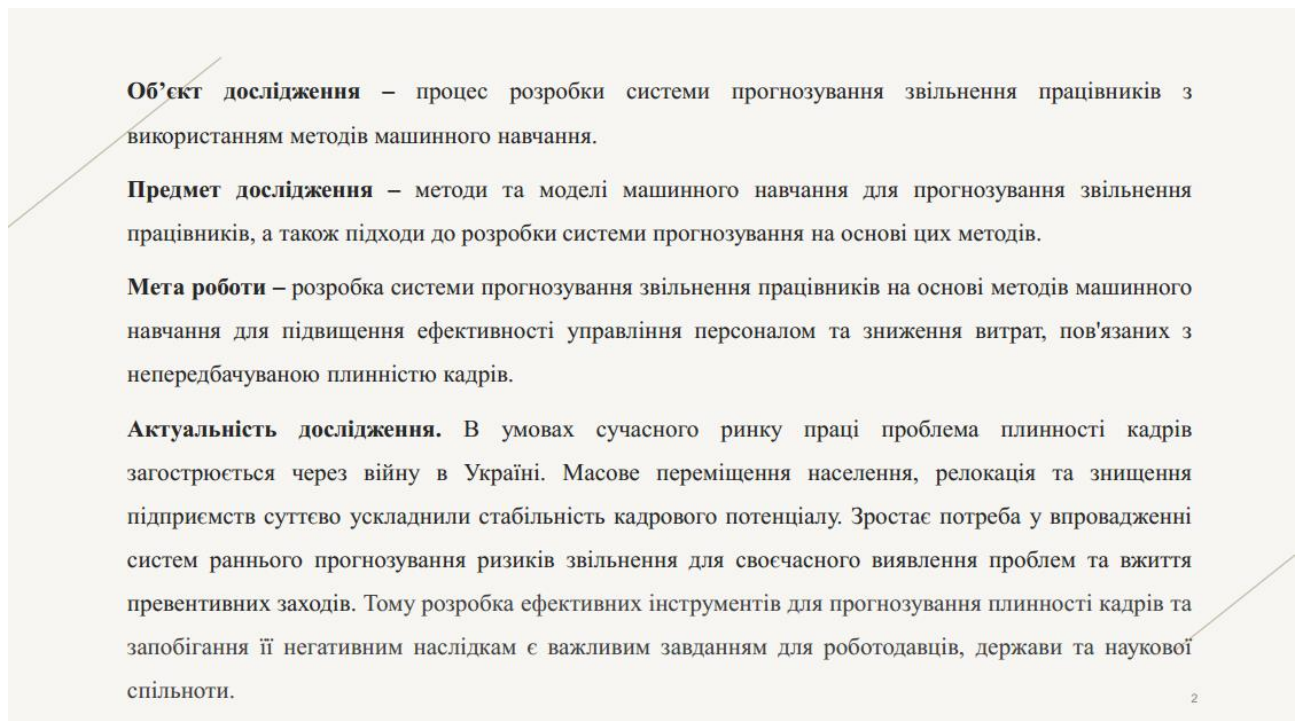


Рисунок Б.2 – Слайд 2

ЗАВДАННЯ ДОСЛІДЖЕННЯ:

- провести аналіз існуючих підходів до прогнозування плинності кадрів та методів машинного навчання, які застосовуються для вирішення подібних задач;
- визначити ключові фактори, які впливають на рішення працівників про звільнення;
- розробити ефективну методику обробки та аналізу даних для підготовки їх до моделювання;
- обрати та налаштувати алгоритми машинного навчання для побудови моделей прогнозування плинності кадрів;
- здійснити проєктування та реалізувати компоненти системи прогнозування звільнення працівників;
- виконати тестування розробленої системи прогнозування;
- визначити шляхи подальшого розвитку розробленої системи прогнозування.

3

Рисунок Б.3 – Слайд 3

ПОРІВНЯЛЬНИЙ АНАЛІЗ HR-СИСТЕМ З ПІДТРИМКОЮ AI, ПРЕДСТАВЛЕНИХ НА РИНКУ УКРАЇНИ

Система	PeopleForce	Hurma System	KISSFLOW HR Cloud	Culture Amp	BambooHR
Локалізація	Українська розробка, адаптована під локальний ринок	Українська розробка, орієнтована на місцеві компанії	Міжнародне рішення, доступне в Україні	Міжнародне рішення, доступне в Україні	Міжнародне рішення, доступне в Україні
Функціонал	Управління персоналом, аналітика, прогнозування плинності	Автоматизація HR-процесів, аналітика, прогнозування плинності	Управління персоналом, автоматизація процесів, прогнозування плинності	Опитування залученості, зворотний зв'язок, аналітика, прогнозування плинності	Управління персоналом, рекрутинг, аналітика, прогнозування плинності
Інтеграція	Інтеграція з популярними сервісами та платформами	Підтримка інтеграцій з різними системами	Широкі можливості інтеграції з іншими бізнес-додатками	Інтеграція з HRIS, SFTP, Slack, SSO, MS Teams	Інтеграція з різними HR-додатками та платформами
Підтримка	Локальна підтримка українською мовою	Локальна підтримка українською мовою	Підтримка англійською мовою	Підтримка англійською мовою	Підтримка англійською мовою
Цінова політика	Гнучка цінова політика, орієнтована на український ринок	Доступні тарифи для малого та середнього бізнесу	Вартість залежить від обраного пакету та кількості користувачів	Вартість залежить від обраного пакету та кількості користувачів	Вартість залежить від обраного пакету та кількості користувачів

4

Рисунок Б.4 – Слайд 4

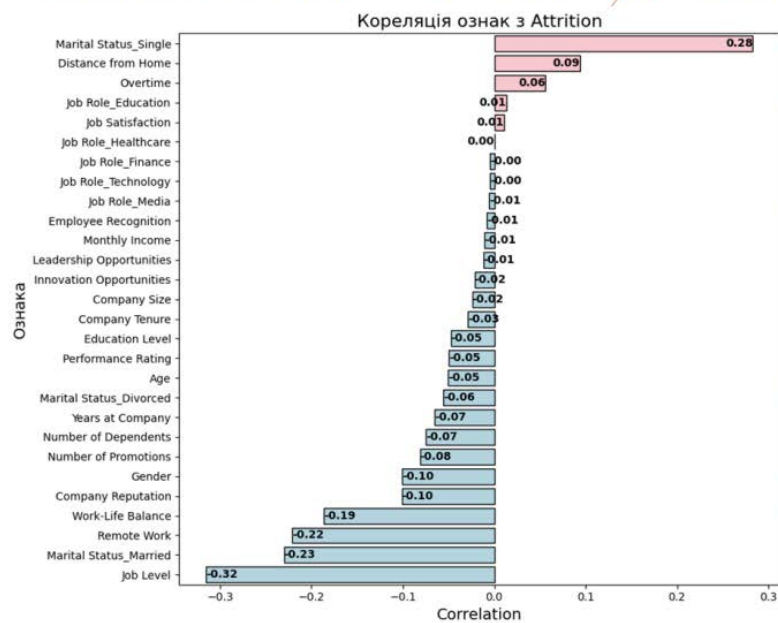
АНАЛІЗ НАБОРУ ДАНИХ

	Column	Non-null	Null	Data type	Number of unique	Unique values
0	Attrition	74498	0	object	2	Stayed, Left
1	Age	74498	0	int64	42	31, 59, 24, 36, 56, 38, 47 ...
2	Company Reputation	74498	0	object	4	Excellent, Fair, Poor, Good
3	Company Size	74498	0	object	3	Medium, Small, Large
4	Company Tenure	74498	0	int64	127	89, 21, 74, 50, 68, 47, 93 ...
5	Distance from Home	74498	0	int64	99	22, 21, 11, 27, 71, 37, 75 ...
6	Education Level	74498	0	object	5	Associate Degree, Master Degree, Bachelor Degree, High School, PhD
7	Employee ID	74498	0	int64	74498	8410, 64756, 30257, 65791, 65026, 24368, 64970 ...
8	Employee Recognition	74498	0	object	4	Medium, Low, High, Very High
9	Gender	74498	0	object	2	Male, Female
10	Innovation Opportunities	74498	0	object	2	No, Yes
11	Job Level	74498	0	object	3	Mid, Senior, Entry
12	Job Role	74498	0	object	5	Education, Media, Healthcare, Technology, Finance
13	Job Satisfaction	74498	0	object	4	Medium, High, Very High, Low
14	Leadership Opportunities	74498	0	object	2	No, Yes
15	Marital Status	74498	0	object	3	Married, Divorced, Single
16	Monthly Income	74498	0	int64	9842	5390, 5534, 8159, 3989, 4821, 9977, 3681 ...
17	Number of Dependents	74498	0	int64	7	0, 3, 2, 4, 1, 5, 6
18	Number of Promotions	74498	0	int64	5	2, 3, 0, 1, 4
19	Overtime	74498	0	object	2	No, Yes
20	Performance Rating	74498	0	object	4	Average, Low, High, Below Average
21	Remote Work	74498	0	object	2	No, Yes
22	Work-Life Balance	74498	0	object	4	Excellent, Poor, Good, Fair
23	Years at Company	74498	0	int64	51	19, 4, 10, 7, 41, 3, 23 ...

5

Рисунок Б.5 – Слайд 5

АНАЛІЗ ФАКТОРІВ ВПЛИВУ НА ЗВІЛЬНЕННЯ



6

Рисунок Б.6 – Слайд 6

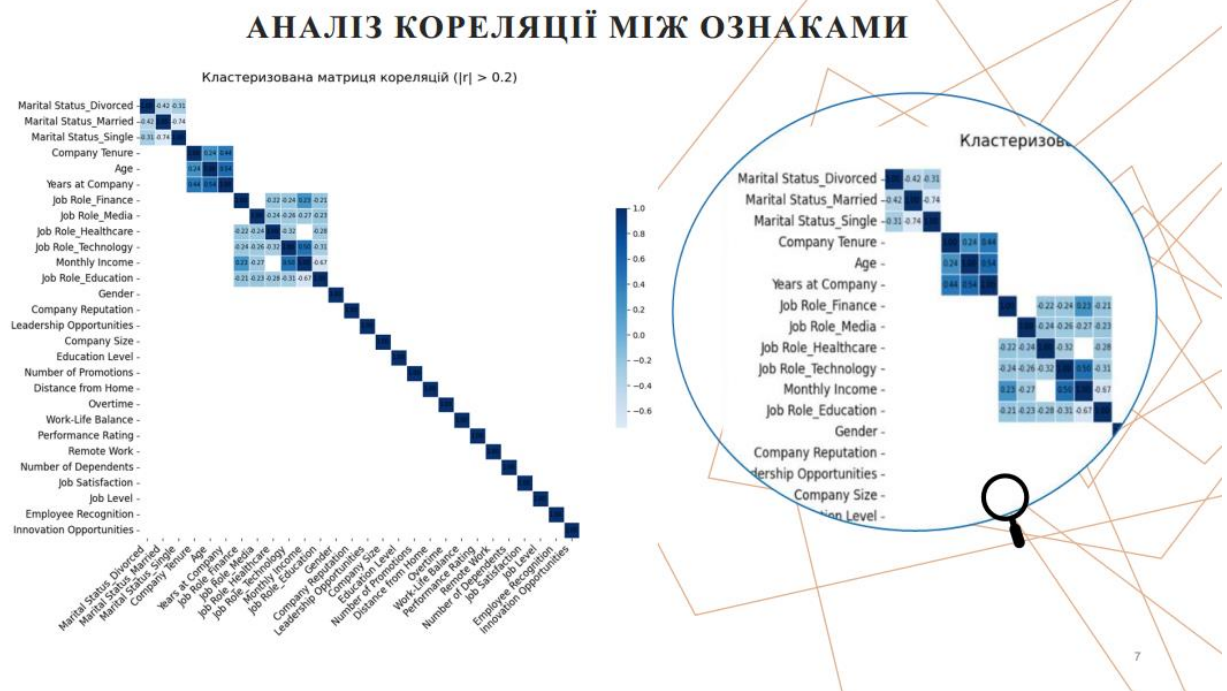


Рисунок Б.7 – Слайд 7

ПОРІВНЯННЯ ЕФЕКТИВНОСТІ КЛАСИФІКАТОРІВ ПРИ БАЗОВИХ НАЛАШТУВАННЯХ

Результати класифікаторів за ROC AUC Score:							
	Classifier	Accuracy	ROC AUC Score	Recall	Precision	Training Time (s)	Prediction Time (s)
0	Ada Boost Classifier	0.756003	0.846545	0.738353	0.745841	5.649492	0.126470
2	Gradient Boosting Classifier	0.753691	0.843832	0.735529	0.743578	18.576224	0.048458
15	CatBoostClassifier	0.754661	0.843257	0.742431	0.741733	20.481305	0.021651
1	Light Gradient Boosting Machine	0.752573	0.843129	0.741804	0.738791	1.788337	0.144223
3	Extreme Gradient Boosting	0.748024	0.835317	0.736314	0.734355	3.689390	0.047465
16	SVM Linear Kernel	0.747129	0.834271	0.760471	0.722288	2333.808681	34.210572
5	Random Forest Classifier	0.738404	0.828053	0.712941	0.730355	13.718327	0.456740
7	Logistic Regression	0.739672	0.825270	0.714667	0.731535	0.378310	0.006918
9	Linear Discriminant Analysis	0.739299	0.825182	0.717333	0.729695	0.194932	0.010017
6	Extra Trees Classifier	0.740045	0.824227	0.714667	0.732123	11.767069	0.684935
10	Quadratic Discriminant Analysis	0.733930	0.816728	0.745412	0.709571	0.159781	0.019244
4	MLP Classifier	0.727218	0.814034	0.721725	0.709483	103.562641	0.059652
11	Naive Bayes	0.730425	0.811519	0.745725	0.704505	0.042488	0.008391
12	K Neighbors Classifier	0.665548	0.714301	0.646118	0.648866	0.017272	6.734551
13	Decision Tree Classifier	0.668531	0.667575	0.648157	0.652352	0.641625	0.008494
14	Dummy Classifier	0.524609	0.500000	0.000000	0.000000	0.002624	0.000309
8	Ridge Classifier	0.739299	NaN	0.717333	0.729695	0.087045	0.010715

Чотири найкращі класифікатори: ['Ada Boost Classifier' 'Gradient Boosting Classifier' 'CatBoostClassifier' 'Light Gradient Boosting Machine']

Рисунок Б.8 – Слайд 8

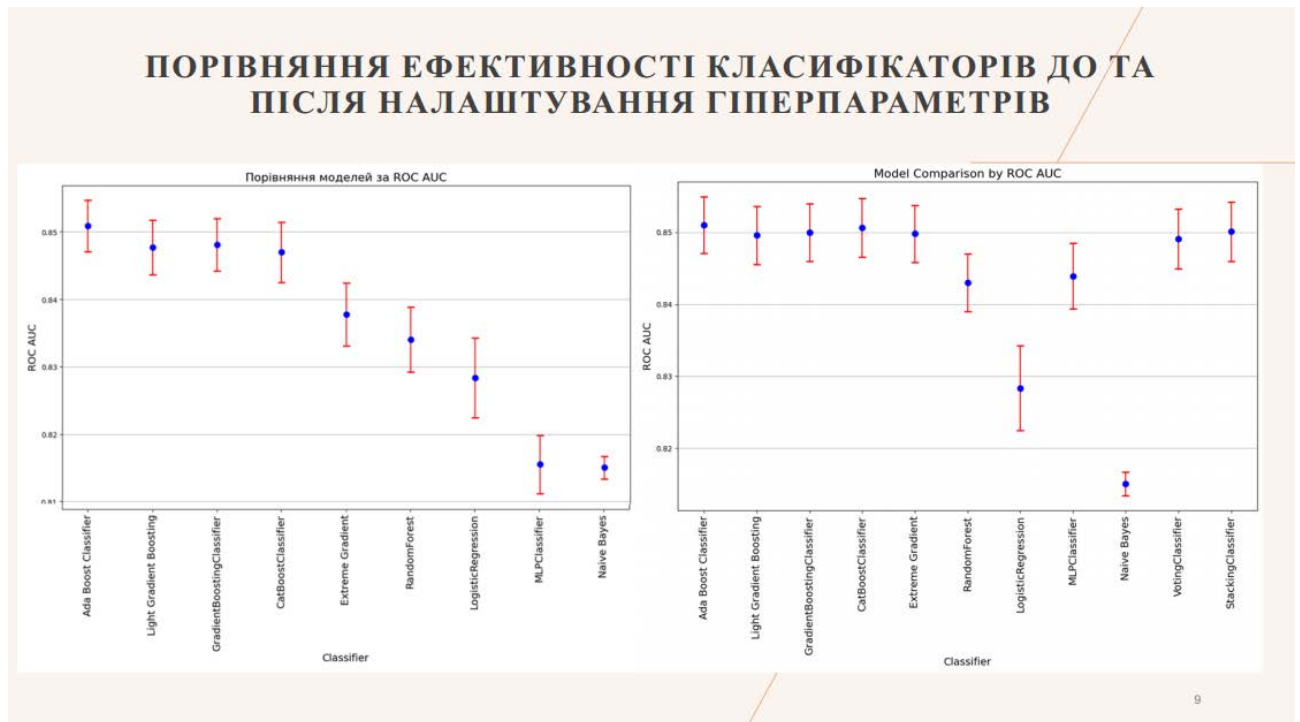


Рисунок Б.9 – Слайд 9

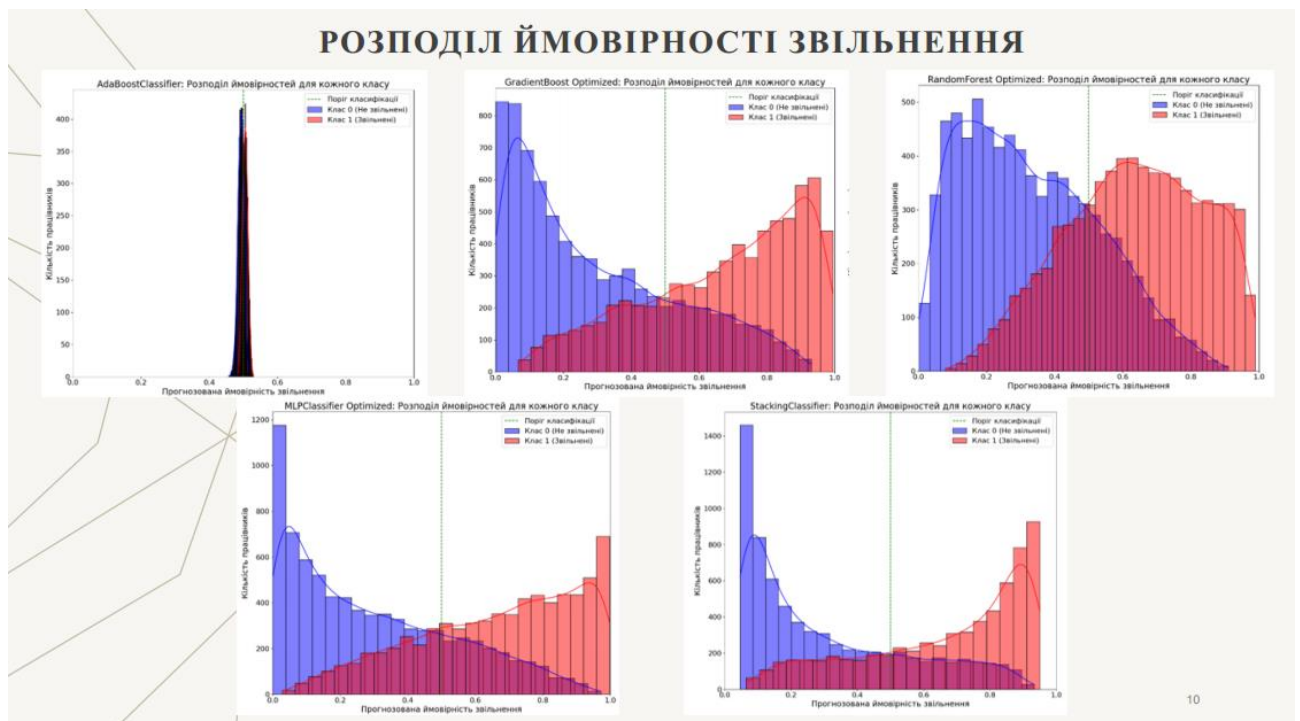


Рисунок Б.10 – Слайд 10

ПОРІВНЯННЯ МЕТОДІВ ВІДБОРУ ОЗНАК

Порівняльна таблиця результатів:

	Classifier	Accuracy	ROC AUC Score	Recall	Precision	F1 Score	Selected Features
0	LGBMClassifier	0.7556	0.8456	0.7423	0.7432	0.7427	23.0000
1	Correlation	0.7456	0.8343	0.7311	0.7331	0.7321	11.0000
2	PCA	0.7184	0.8015	0.7107	0.7011	0.7059	11.0000
3	Lasso	0.7534	0.8421	0.7399	0.7410	0.7404	18.0000
4	RandomForest	0.7030	0.7789	0.6739	0.6929	0.6833	5.0000
5	RFECV	0.7548	0.8451	0.7413	0.7425	0.7419	18.0000
6	GA	0.7321	0.8179	0.7250	0.7154	0.7202	7.0000

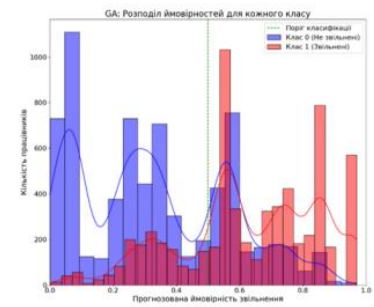
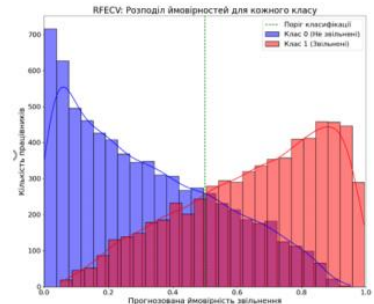
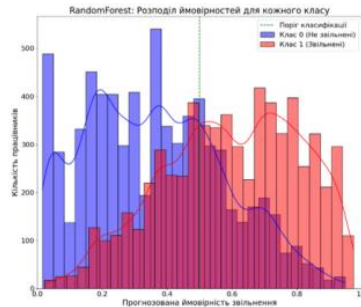
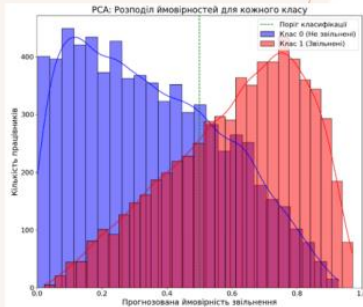


Рисунок Б.11 – Слайд 11

СТРУКТУРА СИСТЕМИ ПРОГНОЗУВАННЯ ЗВІЛЬНЕННЯ ПРАЦІВНИКІВ

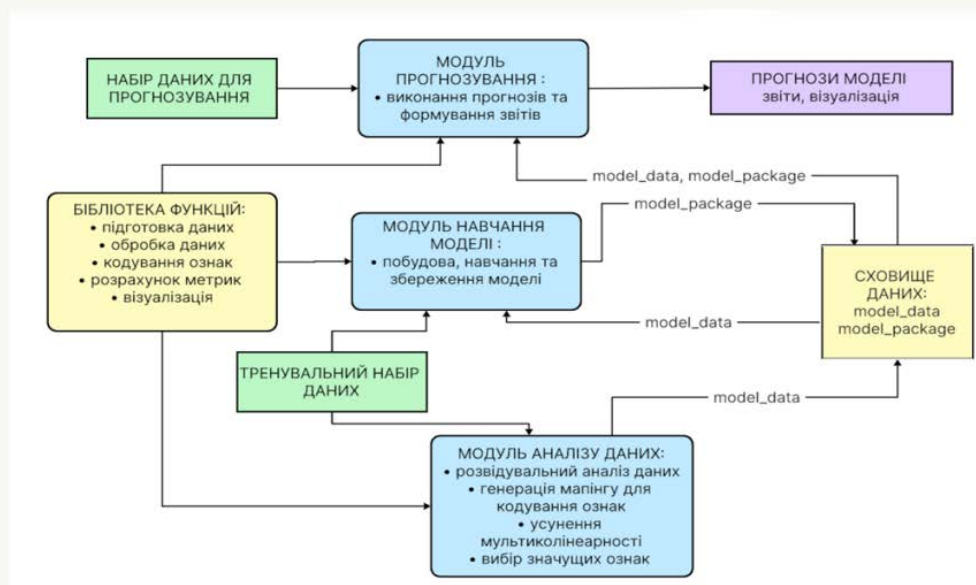


Рисунок Б.12 – Слайд 12

РЕЗУЛЬТАТИ ТЕСТУВАННЯ STACKING CLASSIFIER

	Метрики	Розподіл ймовірності звільнення	Метрики	Розподіл ймовірності звільнення
Ефективність моделі	Accuracy: 0.7545 Balanced Accuracy: 0.7538 ROC AUC Score: 0.8453 Average Precision Score: 0.8359 F1 Score: 0.7411 Recall (Sensitivity): 0.7391 Specificity: 0.7684 G-mean: 0.7536 Precision: 0.7431 Confusion Matrix: [[5406 1629] [1663 4712]]		Accuracy: 0.7838 Balanced Accuracy: 0.3919 ROC AUC Score: Недоступно Average Precision Score: Недоступно F1 Score: 0.0000 Recall (Sensitivity): 0 Specificity: 0.7838 G-mean: 0.0000 Precision: 0.0000 Confusion Matrix: [[3067 846] [0 0]]	
Тестовий набір	Accuracy: 0.7630 Balanced Accuracy: 0.7618 ROC AUC Score: 0.8563 Average Precision Score: 0.8447 F1 Score: 0.7477 Recall (Sensitivity): 0.7399 Specificity: 0.7838 G-mean: 0.7615 Precision: 0.7557 Confusion Matrix: [[3067 846] [920 2617]]		Accuracy: 0.7399 Balanced Accuracy: 0.3699 ROC AUC Score: Недоступно Average Precision Score: Недоступно F1 Score: 0.8505 Recall (Sensitivity): 0.7399 Specificity: 0 G-mean: 0.0000 Precision: 1.0000 Confusion Matrix: [[0 0] [920 2617]]	

13

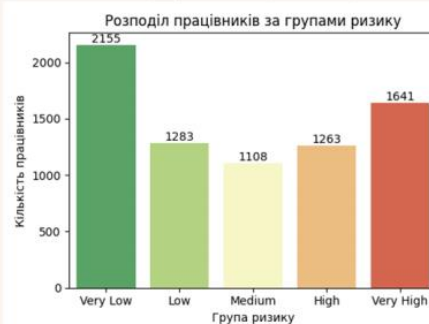
Рисунок Б.13– Слайд 13

РЕЗУЛЬТАТИ ПРОГНОЗУВАННЯ НА
ТЕСТОВОМУ НАБОРІ ДАНИХ

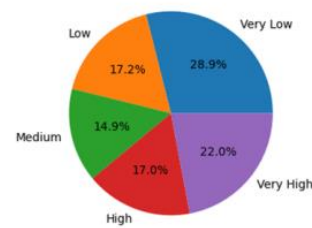
▼ reports

- full_report.csv
- risk_group_High.csv
- risk_group_Low.csv
- risk_group_Medium.csv
- risk_group_Very High.csv
- risk_group_Very Low.csv

	A	B	C	D	E	F	G	H	I
1	Employee ID	Prediction	Probability	Risk Level	Remote Work	Marital Status_Single	Job Level	Work-Life Balance	Gender
2	49797	0	0.5	Medium	No	ЛОЖЬ	Mid	Poor	Male
3	16721	0	0.4	Medium	No	ЛОЖЬ	Entry	Excellent	Male
4	18956	0	0.15	Very Low	No	ЛОЖЬ	Mid	Good	Male
5	23781	0	0.18	Very Low	Yes	ИСТИНА	Entry	Excellent	Male
6	19694	0	0.08	Very Low	Yes	ЛОЖЬ	Mid	Good	Female
7	46549	1	0.92	Very High	No	ИСТИНА	Entry	Excellent	Female
8	6409	1	0.73	High	No	ЛОЖЬ	Mid	Fair	Male
9	11806	0	0.27	Low	No	ИСТИНА	Senior	Excellent	Male
10	13051	1	0.81	Very High	No	ИСТИНА	Mid	Good	Male
11	48186	1	0.94	Very High	No	ИСТИНА	Entry	Poor	Female
12	22117	1	0.51	Medium	Yes	ЛОЖЬ	Entry	Poor	Female



Відсотковий розподіл груп ризику



14

Рисунок Б.14 – Слайд 14

SWOT-АНАЛІЗ

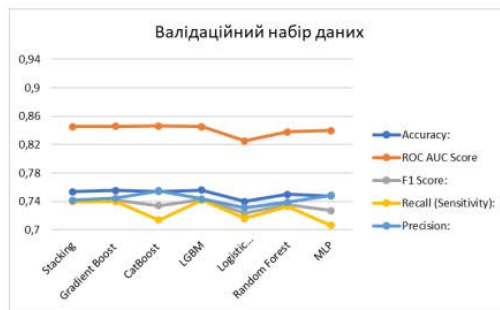
СИЛЬНІ СТОРОНИ (STRENGTHS):		СЛАБКІ СТОРОНИ (WEAKNESSES):	
1. Модульна архітектура: – чітке розділення функціоналу; – легкість підтримки та модифікації; – можливість повторного використання компонентів.	3. Послідовна обробка даних: – чіткий процес від аналізу до прогнозування; – збереження проміжних результатів; – можливість контролю на кожному етапі.	1. Залежність від проміжних файлів: – необхідність зберігання додаткових файлів; – можливі проблеми з версійністю файлів; – ризик пошкодження/втрати файлів.	3. Обмежена валідація даних: – відсутність централізованої перевірки вхідних даних; – можливі проблеми з якістю даних; – ризик некоректної обробки.
2. Централізація допоміжних функцій: – єдине місце для базових операцій; – уникнення дублювання коду; – спрощене оновлення функціоналу.	4. Незалежність модулів: – можливість паралельної розробки; – простота тестування; – легка заміна окремих компонентів.	2. Відсутність системи логування: – складність відстеження помилок; – обмежені можливості моніторингу; – складність діагностики проблем.	4. Відсутність паралельної обробки: – послідовне виконання операцій; – обмежена продуктивність; – неефективне використання ресурсів.
МОЖЛИВОСТІ (OPPORTUNITIES):		ЗАГРОЗИ (THREATS):	
1. Розширення функціоналу: – додавання нових методів аналізу; – впровадження додаткових моделей; – розширення можливостей візуалізації.	3. Покращення взаємодії: – додавання API; – інтеграція з іншими системами; – створення веб-інтерфейсу.	1. Технічні ризики: – проблеми сумісності версій Python/бібліотек; – обмеження масштабованості; – технічний борг.	3. Безпека: – доступ до конфіденційних даних; – можливі витoki інформації; – вразливості системи.
2. Оптимізація процесів: – впровадження паралельної обробки; – покращення алгоритмів; – оптимізація використання пам'яті.	4. Вдосконалення безпеки: – впровадження системи авторизації; – шифрування даних; – аудит доступу.	2. Якість даних: – неповні або некоректні вхідні дані; – зміни в форматі вхідних даних; – проблеми з кодуванням.	4. Обслуговування: – складність підтримки при розширенні; – залежність від специфічних знань; – потреба в регулярному оновленні.

15

Рисунок Б.15 – Слайд 15

РЕЗУЛЬТАТИ ТЕСТУВАННЯ СИСТЕМИ З РІЗНИМИ КЛАСИФІКАТОРАМИ

Класифікатор	Stacking	Gradient Boost	CatBoost	LGBM	Logistic Regression	Random Forest	MLP
Час оптимізації Randomized SearchCV	2613,8545	852,1347	1054,6521	217,0202	27,5738	561,6284	2916,1577
Час оптимізації GridSearchCV	208,0285	26,0182	88,7411	4,6450	1,3690	93,3756	213,9620
Час навчання моделі	305,6151	11,1144	3,3504	1,2232	0,5179	44,1696	75,2652



16

Рисунок Б.16 – Слайд 16

ВИСНОВКИ

У роботі вирішено актуальне завдання розробки системи прогнозування звільнення працівників на основі методів штучного інтелекту. Основні результати роботи полягають у наступному.

Проведений аналіз існуючих підходів до прогнозування звільнення працівників виявив необхідність розробки комплексного рішення, яке б враховувало широкий спектр характеристик працівників та забезпечувало високу точність прогнозування.

Розроблено систему прогнозування з модульною архітектурою, що включає чотири основні компоненти: бібліотеку функцій, модулі аналізу даних, навчання моделі та прогнозування. Взаємодія між компонентами реалізована через проміжні файли, що забезпечує узгодженість обробки даних та можливість контролю на кожному етапі.

Тестування системи на різних наборах даних підтвердило її високу ефективність. Проведений SWOT-аналіз виявив ключові переваги системи (модульність, централізація функцій, послідовна обробка даних) та визначив перспективні напрями вдосконалення.

На основі порівняльного аналізу різних методів машинного навчання встановлено, що найвищу ефективність демонструють ансамблеві методи, зокрема градієнтний бустинг. Розроблено модифікований модуль навчання моделі на базі Light Gradient Boosting Machine, який забезпечує оптимальний баланс між швидкістю та точністю прогнозування.

Практичне значення отриманих результатів полягає в можливості використання розробленої системи для ефективного прогнозування ризику звільнення працівників та прийняття обґрунтованих рішень в HR-менеджменті. Модульна архітектура системи забезпечує її гнучку адаптацію під потреби різних організацій.

Подальші дослідження можуть бути спрямовані на розширення функціональності системи через впровадження API, розробку веб-інтерфейсу та вдосконалення механізмів валідації даних і системи логуювання.

17

Рисунок Б.17 – Слайд 17

ПЕРЕЛІК ПУБЛІКАЦІЙ

- Domashenko, S. Intelligent employee turnover prediction system as a tool for optimizing HR processes / S. Domashenko // Science and Global Challenges in the Modern World: Proceedings of the International Scientific Conference, 4 November 2024. – Leicester, UK: Bookmundo, 2024. - P. 43 – 46.
- Домашенко, С. В. Штучний інтелект у HR-аналітиці: рішення для прогнозування плинності кадрів / С. В. Домашенко, Д. М. Морозов, О. В. Гнездовський // Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій: матеріали XII Міжнародній науково-практичній конференції, 10–12 грудня 2024р. - Запоріжжя : НУ «Запорізька політехніка», 2024.

ДЯКУЮ ЗА УВАГУ!

18

Рисунок Б.18 – Слайд 18