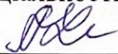


Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації
Кафедра системного аналізу та інформаційних технологій

Бакалаврська кваліфікаційна робота на тему:
**«СИСТЕМНИЙ АНАЛІЗ СОЦІАЛЬНОЇ ВЗАЄМОДІЇ ВІННИЦЬКОЇ МІСЬКОЇ
ТЕРИТОРІАЛЬНОЇ ГРОМАДИ ЧЕРЕЗ ЦИФРОВІ КАНАЛИ»**

Виконала: студентка 4 курсу, групи СА–216
спеціальності 124 – «Системний аналіз»

 Анна БОЯНОВСЬКА

Керівник: Ph.D., ст. викладач каф. САІТ

 Ольга ВОЙЦЕХОВСЬКА

« 08 » 06 2025 р.

Рецензент: к.т.н., доц. каф. КН

 Володимир ОЗЕРАНСЬКИЙ

« 11 » 06 2025 р.

Допущено до захисту

Завідувач кафедри САІТ

 д.т.н., проф. Віталій МОКІН

« 03 » 06 2025 р.

Вінниця ВНТУ – 2025 рік

Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації
Кафедра системного аналізу та інформаційних технологій
Рівень вищої освіти – перший (бакалаврський)
Галузь знань – 12 Інформаційні технології
Спеціальність 124 – «Системний аналіз»
Освітньо-професійна програма – Системний аналіз

ЗАТВЕРДЖУЮ

Завідувач кафедри САІТ

 д.т.н., проф. Віталій МОКІН

“18” 05 2025 року

ЗАВДАННЯ НА БАКАЛАВРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

Бояновській Анні Ігорівні

1. Тема роботи: «Системний аналіз соціальної взаємодії Вінницької міської територіальної громади через цифрові канали»
керівник роботи: Войцеховська О. О., Ph.D., ст. викладач каф. САІТ
затверджені наказом ВНТУ від «20» 05 2025 року № 37
2. Термін подання студентом роботи 31.05.25
3. Вихідні дані до роботи:
 - 1) Дані соціологічного опитування серед мешканців Вінницької міської територіальної громади, проведеного Благодійною організацією «Благодійний фонд «Фонд громади Подільська громада» у січні-лютому 2025 року.
4. Зміст текстової частини:
 - 1) Загальна характеристика об'єкта дослідження.
 - 2) Вибір оптимального рішення та налаштувань для розв'язання поставленої задачі.
 - 3) Результати передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали.
5. Перелік ілюстративного матеріалу:
 - 1) Інформація про дані;
 - 2) Кореляційна матриця числових даних;
 - 3) Графіки важливості ознак;
 - 4) Графіки розподілу за різними ознаками;
 - 5) Оцінка точності моделей.

6. Консультанти розділів роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 28.05

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва та зміст етапу	Термін виконання		Примітка
		початок	закінчення	
1	Аналіз предметної області	01.04	15.04	вм
2	Аналіз наукових робіт дослідження соціальної взаємодії територіальних громад через цифрові канали	15.04	30.04	вм
3	Вибір оптимального рішення та налаштувань для розв'язання поставленої задачі.	01.05	10.05	вм
4	Результати передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали.	10.05	20.05	вм
5	Оформлення матеріалів до захисту БКР	20.05	30.05	вм

Студентка

Анна БОЯНОВСЬКА

Керівник роботи

Ольга ВОЙЦЕХОВСЬКА

АНОТАЦІЯ

Бакалаврська кваліфікаційна робота складається зі 106 сторінок формату А4, на яких є 95 рисунків, список використаних джерел містить 28 найменувань.

Метою роботи є системний аналіз соціальної взаємодії Вінницької міської територіальної громади та підвищення точності передбачення використання цифрових каналів.

Проведено аналіз предметної області, здійснено попередню обробку даних та відбір ознак, побудовано моделі машинного навчання на основі Logistic Regression, Random Forest і XGBoost. Реалізація проведена в середовищі Google Colab з використанням мови Python. Розроблена система дає змогу передбачати, чи користуються респонденти певним цифровим каналам для соціальної взаємодії на основі їхньої соціально демографічної характеристики. Аналіз результатів дозволив визначити найбільш впливові чинники цифрової взаємодії та сформулювати практичні рекомендації щодо підвищення ефективності комунікаційної політики громади.

Ключові слова: системний аналіз, соціальна взаємодія, цифрові канали, передбачення даних, машинне навчання, Python, Google Colab.

ABSTRACT

The bachelor's qualification consists of 106 A4-format pages, on which there are 95 figures, the list of sources used contains 28 names.

The aim of the work is a systematic analysis of the social interaction of the Vinnytsia city territorial community and increasing the accuracy of predicting the use of digital channels.

An analysis of the subject area of data was conducted, preliminary processing and feature selection were performed, and machine learning models based on Logistic Regression, Random Forest, and XGBoost were built. The implementation was carried out in the Google Colab environment using the Python language. The developed system allows to predict whether respondents use certain digital channels for social interaction based on their socio-demographic characteristics. Analysis of the results allowed to determine the most influential factors of digital interaction and to form practical recommendations for improving the effectiveness of communication policy of the community.

Keywords: systems analysis, social interaction, digital channels, data prediction, machine learning, Python, Google Colab.

ЗМІСТ

ВСТУП.....	3
1 ЗАГАЛЬНА ХАРАКТЕРИСТИКА ОБ’ЄКТУ ДОСЛІДЖЕНЬ.....	5
1.1 Аналіз предметної області	5
1.2 Аналіз наукових робіт дослідження соціальної взаємодії територіальних громад через цифрові канали	5
1.3 Аналіз середовища розробки та мови програмування для задачі передбачення взаємодії Вінницької міської територіальної громади через цифрові канали....	10
1.4 Висновки.....	12
2 ВИБІР ОПТИМАЛЬНОГО РІШЕННЯ ТА НАЛАШТУВАНЬ ДЛЯ РОЗВ’ЯЗАННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ.....	13
2.1 Розвідувальний аналіз даних для задачі передбачення соціальної взаємодії через цифрові канали	13
2.2 Відбір ознак для передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали	20
2.3 Тренування моделей передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали	32
2.4 Висновки.....	40
3 РЕЗУЛЬТАТИ ПЕРЕДБАЧЕННЯ СОЦІАЛЬНОЇ ВЗАЄМОДІЇ ВІННИЦЬКОЇ МІСЬКОЇ ТЕРИТОРІАЛЬНОЇ ГРОМАДИ ЧЕРЕЗ ЦИФРОВІ КАНАЛИ	42
3.1 Передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали	42
3.2 Висновки.....	73
ВИСНОВКИ	75
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	78
Додаток А Технічне завдання.....	83
Додаток Б Протокол перевірки кваліфікаційної роботи.....	85
Додаток В Фрагмент лістингу програми.....	86
Додаток Г Ілюстративна частина	92

ВСТУП

Актуальність теми. У сучасному інформаційному суспільстві цифрові канали комунікації відіграють ключову роль у забезпеченні ефективної взаємодії між органами місцевого самоврядування та мешканцями територіальних громад. З огляду на стрімкий розвиток технологій та зростання ролі інтернет–ресурсів, соціальних мереж і мобільних застосунків, важливим завданням є дослідження цифрової активності громадян та аналіз чинників, які впливають на вибір каналів комунікації.

Особливої актуальності тема набуває в умовах децентралізації та цифрової трансформації муніципального управління. Для міських рад, громадських організацій та інших зацікавлених сторін стає важливо не лише поширювати інформацію, але й робити це через ті канали, які реально використовуються мешканцями. Розуміння цифрової поведінки різних соціально–демографічних груп дозволяє ефективніше будувати інформаційні кампанії, підвищувати рівень залученості та покращувати якість обслуговування населення.

Системний аналіз у поєднанні з методами машинного навчання відкриває нові можливості для глибокого дослідження структури соціальної взаємодії, а також для прогнозування моделей поведінки користувачів цифрових каналів. Це дозволяє виявляти приховані закономірності, адаптувати канали зв'язку під потреби окремих категорій населення та приймати управлінські рішення на основі даних.

Мета і задачі дослідження. Метою дослідження є системний аналіз соціальної взаємодії Вінницької міської територіальної громади та підвищення точності передбачення використання цифрових каналів.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

- здійснити аналіз предметної області;
- здійснити вибір оптимальних технологій машинного навчання для передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали;

– застосувати обрану модель машинного навчання для передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали.

Об’єктом дослідження є соціальна взаємодія Вінницької міської територіальної громади через цифрові канали.

Предметом дослідження є моделі машинного навчання для передбачення використання цифрових каналів взаємодії Вінницької міської територіальної громади.

1 ЗАГАЛЬНА ХАРАКТЕРИСТИКА ОБ'ЄКТУ ДОСЛІДЖЕНЬ

1.1 Аналіз предметної області

Системний аналіз соціальної взаємодії Вінницької міської територіальної громади – це дослідження складної соціально–цифрової системи, яка включає численні підсистеми: демографічну групу населення, просторово–локаційну, соціально–економічну та комунікаційну. Проблематика прийняття рішень у цій системі (наприклад, вибір каналів інформування для певної категорії населення) залежить від численних взаємозалежних факторів – віку, статі, освіти, місця проживання, працевлаштування тощо [1].

Складність системи стає помітною при наявності великої кількості взаємопов'язаних соціально–демографічних ознак, які впливають на громадянина під час вибору цифрового каналу. За допомогою методів машинного навчання пропонується здійснити аналіз цих ознак та виявити ті, що мають найбільший вплив на вибір респондента. Результати аналізу дадуть змогу ранжувати ознаки за силою впливу та встановити, як цільові групи взаємодіють з різними цифровими каналами.

Розуміння структури взаємодії в межах складної системи дозволяє зацікавленим організаціям (органам влади, місцевого самоврядування, громадським організаціям, бізнесу тощо) створювати та проводити ефективніші комунікаційні кампанії й легше знаходити цільову аудиторію. Оптимізація відбувається через адаптацію каналів комунікації до характеристик цільових груп, що підвищує точність інформування й обґрунтованість прийнятих рішень.

1.2 Аналіз наукових робіт дослідження соціальної взаємодії територіальних громад через цифрові канали

Покращення цифровізації комунікативного простору Вінницької міської територіальної громади безпосередньо сприяє підвищенню ефективності

комунікації та інформування населення. Широке використання соціальних мереж, месенджерів, офіційних сайтів та інших електронних засобів зв'язку формує комунікативне середовище, у якому громадяни мають розширені можливості для отримання інформації, комунікації, доступу до публічних послуг і соціальної взаємодії, що підвищує рівень участі мешканців у житті громади.

Проблематика використання цифрових каналів комунікації в територіальних громадах України вже тривалий час є предметом наукових досліджень. Її актуальність особливо зросла внаслідок реформи децентралізації, у межах якої повноваження, ресурси та відповідальність були передані від центральної влади до органів місцевого самоврядування. Новостворені органи місцевого самоврядування отримали можливість самостійно налагоджувати комунікативні зв'язки з мешканцями громади.

Одним із ключових завдань стало створення офіційних цифрових каналів для полегшення й підвищення ефективності інформування громади та збору зворотного зв'язку від мешканців. Відтак у Стратегії розвитку Вінницької міської територіальної громади до 2030 року стратегічним пріоритетом №1 визначено «Цифровізацію муніципального простору». Цей пріоритет охоплює три стратегічні цілі та низку напрямів, спрямованих на підвищення цифрової участі та проінформованості громадян [2].

У контексті дослідження цифрової взаємодії мешканців Вінницької міської територіальної громади варто звернути увагу на напрям №3 Стратегії розвитку громади, а саме «Розбудова цифрової інфраструктури – Smart City» [3]. Серед його завдань передбачено вдосконалення цифрової системи оповіщення населення про небезпеки та загрози, що є особливо актуальним в умовах воєнного стану.

Станом на 01.07.2023 року рівень охоплення системи оповіщення становив близько 125 тисяч мешканців і це завдяки використанню соціальних мереж і месенджерів [2]. Водночас загальна чисельність населення громади у 2022 році становила 387 тисяч осіб [4]. Це свідчить про значний потенціал для подальшого розвитку цифрової взаємодії з представниками громади, що, своєю чергою,

підкреслює актуальність дослідження та передбачення популярних цифрових каналів для різних категорій населення Вінницької міської територіальної громади.

Офіційна інформація від Вінницької міської ради, міського голови, виконавчих органів та комунальних підприємств поширюється через такі цифрові канали, як Facebook, Telegram, Viber, Instagram, YouTube та радіо «Місто над Бугом» та інші.

Станом на початок 2023 року [5]:

- Сторінка Вінницької міської ради налічує 31 856 підписників у Facebook, 128 тисяч – у Telegram, 6 143 – у Viber.
- Міський голова активно комунікує з громадянами: 56 314 підписників у Facebook, 16 470 – у Telegram, 10 080 – в Instagram.

Такі показники свідчать про високий рівень цифрової присутності влади у медіапросторі громади та демонструють реальний потенціал для дослідження взаємодії через ці канали.

Розвиток інформатизації та цифрової трансформації триває, програма на 2024 – 2026 роки передбачає такі основні напрямки [5]:

- збільшення охоплення цифровою системою оповіщення цивільного захисту, зокрема через соціальні мережі, месенджери, чат-боти тощо;
- забезпечення безперебійної роботи цифрової системи оповіщення населення, включно з технічною підтримкою та розвитком цифрових каналів зв'язку;
- своєчасне оповіщення населення у разі повітряної тривоги, виникнення надзвичайних ситуацій або бойових дій;
- підтримка функціонування цифрових каналів сповіщення, зокрема інструментів у соціальних мережах, месенджерах, чат-ботах та інших цифрових сервісах[5].

У межах цього дослідження було враховано не лише аналітичні матеріали, що стосуються Вінницької міської територіальної громади, а й загальноукраїнські джерела – з метою ширшого розуміння предметної області.

Зокрема, у роботі Ганни Давиденко «Цифрова інклюзія та доступність: соціальна диджиталізація» [6] зазначається, що рівень цифровізації сфери послуг в Україні залишається одним із найвищих у Європі – навіть в умовах воєнного стану. Цифрові інструменти надають громадянам доступ до великого обсягу інформації: новин, наукових досліджень, навчальних матеріалів, мультимедійного та альтернативного контенту. Це допомагає користувачам розширювати кругозір та легко знаходити потрібну інформацію в мережі, ухвалювати зважені рішення, бути активними у соціальній взаємодії та розвивати навички критичного мислення. Цифрові технології також відкривають нові шляхи участі громадян у політичному процесі – зокрема через соціальні мережі та інші форми онлайн-активності, що сприяє зростанню громадянської активності та розвитку демократії.

Водночас авторка звертає увагу на важливу проблему: навіть за високого рівня цифровізації інструменти інформування не завжди є ефективними для донесення інформації до цільової аудиторії. Текстові матеріали (статті, наукові роботи тощо) з детальними інструкціями та рекомендаціями, попри якісне наповнення, часто не мають суттєвого цифрового чи соціального впливу. Основними причинами цього є інформаційне перенасичення сучасного медіапростору та психологічна схильність аудиторії віддавати перевагу простому, візуальному та лаконічному контенту [6].

Саме тому дослідження, спрямоване на виявлення ключових факторів, що впливають на вибір цифрових каналів різними групами громадян, актуальними. Такий підхід дає змогу не лише краще зрозуміти поведінку користувачів у цифровому середовищі, а й передбачити, які категорії населення схильні отримувати інформацію через ті чи інші канали (наприклад, соціальні мережі, месенджери, веб-сайти тощо), а також адаптувати контент до їхніх уподобань для привернення уваги цільової аудиторії та підвищення якості комунікації.

У статті «Особливості взаємодії з медіа в процесі адаптації спільноти до умов і наслідків воєнного конфлікту» наведено результати емпіричного дослідження медіаспоживання. Встановлено, що зростання кількості інтернет–

користувачів призводить до зниження популярності традиційних ЗМІ (телебачення, радіо, преси) на користь соціальних мереж та новинних сайтів. Найпопулярнішими джерелами інформації стали Facebook та агрегатори новин (зокрема ukr.net). Молодь до 35 років у понад 80% випадків переглядає новини саме онлайн. [7].

У роботі «Трансформація медіа в умовах багатопольного світу» [8] підкреслюється, що зі зростанням суспільства, роль медіа як одного з факторів формування громадської думки політичної свідомості, та навіть культури значно посилюється. Інтернет–телебачення, стрімінгові платформи, блоги, електронні версії книг, журналів, словників та соціальні мережі збільшили обсяги поглинання інформації сучасної людини в рази. Це спровокувало інформаційну революцію у суспільстві, яка має вплив на якість, швидкість і кількість отримання інформації індивідом [8].

Варто зауважити, що соціальна діяльність людини відбувається саме в частині її соціального простору – усталеної системи соціальних зв'язків. Як зазначає у своїй роботі [9] О. Бадюл: «Соціальний ареал життя особистості – це соціальний порядок її життєдіяльності у соціальному просторі, окреслений певними межами» [9]. А в рамках глобалізації і доступності інформації через сучасні медіа рамки простору стираються. В інформаційному відбувається постійне звуження соціального ареалу особистості і ускладнення взаємодії людей із зовнішнім середовищем. Це може спричинити негативні наслідки як для індивіду так і для соціуму загалом. На думку В. Воронкової, в «інформаційному суспільстві, де цілі особистості не збігаються із цілями соціуму, може викликати психоемоційні порушення й деструктивну поведінку» [8].

Таким чином для балансування необхідно впроваджувати більше нішевих спеціалізованих медіа, які будуть відповідати соціальним межам індивідів, які живуть на певній території. А також відповідати спеціалізаціям певних верств населення.

У сучасних умовах медіа стали інструментом соціальної трансформації, що формує громадську думку та поведінку. Дослідження нішевих цифрових каналів,

якими користуються різні соціальні групи, є важливим для ефективної та цільової взаємодії громади через цифрові засоби комунікації.

1.3 Аналіз середовища розробки та мови програмування для задачі передбачення взаємодії Вінницької міської територіальної громади через цифрові канали

У рамках даного дослідження для розробки моделей машинного навчання використано мову програмування Python, одну з найпоширеніших у сфері аналізу даних, штучного інтелекту та системного моделювання. Python – це об'єктно–орієнтована мова програмування. Програмне забезпечення на Python реалізується у вигляді моделей, які можуть бути організовані у пакети. Тип та структуру кожного об'єкта можна запитати під час виконання програми. Для кожного з об'єктів можна отримати всю інформацію щодо його внутрішньої структури.

Python має такі переваги:

- логічний і лаконічний синтаксис, завдяки чому вихідний код легко читати та розуміти;
- гнучкість та масштабованість, що дозволяє адаптувати високорівневу логіку та розширяти складні застосунки, як тільки виникне така необхідність;
- розробка на Python у більшості випадків проходить швидше, ніж на інших мовах програмування;
- Python – інтерпретована мова програмування. Це значить, що код можна написати у будь–якому текстовому файлі на будь–якій платформі, і потім успішно запустити;
- у Python – колосальна спільнота однодумців. Тож будь–які складнощі конкретних розробників вирішуються колективно [10].

Python має велику кількість бібліотек, орієнтованих на обробку даних та машинне навчання. У цьому дослідженні використано зокрема:

- `matplotlib`: комплексна бібліотека для створення статичних, анімованих та інтерактивних візуалізацій на Python. `Matplotlib` робить прості речі простими, а складні – можливими;
- `pandas`: програмна бібліотека написана для мови програмування Python призначена для маніпулювання та аналізу даних. Зокрема, вона пропонує структури даних та операції з числовими таблицями і часовими рядами;
- `numpy`: бібліотека, що додає підтримку великих багатовимірних масивів і матриць, а також колекцію високорівневих математичних функцій для роботи з цими масивами [11].

Для написання та виконання коду було обрано платформу `Google Colab`. `Collaboratory` – це потужне хмарне середовище, яке дозволяє писати та виконувати код на Python через веб-браузер. Також надає можливість працювати у середовищі `Jupyter Notebook` без жодних налаштувань чи інсталяції на локальному комп'ютері. `Colab` особливо корисний для розв'язання задач машинного навчання, аналізу даних та наукових обчислень, оскільки пропонує безкоштовний доступ до графічних процесорів та попередньо встановлених бібліотек Python, таких як `TensorFlow`, `Keras`, `PyTorch`, `NumPy`, `Pandas`, `Matplotlib`.

За допомогою `Google Colab` можливо:

- писати та виконувати код на Python;
- документувати свій код;
- створювати/завантажувати/поширювати `Jupyter` блокноти;
- імпортувати/зберігати блокноти з/на `Google Диск`;
- імпортувати/публікувати блокноти з `GitHub`;
- імпортувати зовнішні набори даних, наприклад, з `Kaggle`;
- інтегрувати `PyTorch`, `TensorFlow`, `Keras`, `OpenCV`;
- мати безкоштовний хмарний сервіс з графічним процесором [11].

Вибір Python та `Google Colab` зумовлений їхньою зручністю, широкою функціональністю та поширеністю у сфері обробки даних, що робить ці

інструменти оптимальними для моделювання соціальної взаємодії Вінницької територіальної громади через цифрові канали.

1.4 Висновки

У першому розділі проаналізовано загальні характеристики об'єкта дослідження – соціальної взаємодії Вінницької міської територіальної громади через цифрові канали. Визначено, що це складна соціально–комунікаційна система, на яку впливають численні фактори: вік, стать, освіта, місце проживання, працевлаштування, належність до соціальної категорії тощо.

На основі аналізу наукових джерел і офіційних документів визначено, що цифрові канали комунікації, такі як Facebook, Telegram, Viber, Instagram, відіграють важливу роль у комунікації органів влади з громадою. У Стратегії розвитку Вінницької міської територіальної громади до 2030 р. акцент зроблено на підвищенні цифрової участі громадян. Водночас виявлено, що не всі цифрові канали є однаково актуальними та ефективними на даний момент.

Здійснено обґрунтований вибір інструментів для реалізації дослідження. Обрано мову програмування Python як оптимальну для виконання задач аналізу даних і побудови моделей. Роботу реалізовано у хмарному середовищі Google Colab, яке забезпечує ефективну роботу з бібліотеками для машинного навчання, візуалізації та обробки інформації.

2 ВИБІР ОПТИМАЛЬНОГО РІШЕННЯ ТА НАЛАШТУВАНЬ ДЛЯ РОЗВ'ЯЗАННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ

2.1 Розвідувальний аналіз даних для задачі передбачення соціальної взаємодії через цифрові канали

Розвідувальний аналіз даних (EDA) – це аналіз основних властивостей даних, знаходження в них загальних закономірностей.

Метою EDA є наступне:

- максимальне вивчення і «розуміння» даних;
- виявлення основних структур та систематизація даних.

Методи EDA застосовуються як до усіх даних, так і до їх кластерів та окремо до тренувальних, валідаційних і тестових даних:

- аналіз імовірнісних розподілів змінних;
- побудова й аналіз кореляційних матриць;
- факторний аналіз;
- дискримінантний аналіз [12].

Для дослідження в рамках бакалаврської роботи використано дані соціологічного опитування, що охоплює 405 респондентів з Вінницької міської територіальної громади. Анкетування проводилося онлайн через Google Форми, та були проведенні фокус-групи, а отримані дані було збережено у форматі Google Таблиць.

Перелік змінних у датасеті:

1. num_male, num_female – кількість чоловіків та жінок;
2. age_young, age_middle, age_senior, age_elderly – кількість респондентів віком відповідно: 15–24 років, 25–45 років, 46–59 років, 60+ років;
3. edu_basic_higher, edu_incomplete_higher, edu_full_higher, edu_vocational, edu_basic_secondary, edu_academic_degree, edu_only_primary, edu_two_higher, edu_full_secondary – кількість респондентів, що мають: базову,

неповну, повну вищу, професійно технічну, базову середню, науковий ступінь, початкову, дві вищі, повну середню освіти;

4. loc_vinnytsia_city, loc_v_krushlyntsi, loc_havryshivka, loc_m_krushlyntsi, loc_vinnytski_khutory, loc_stadnytsia, loc_desna, loc_pysarivka, loc_shchitky – місце проживання у Вінницькій ОТГ (м. Вінниця, села: Великі Крушлинці, Гавришівка, Малі Крушлинці, Вінницькі Хутори, Стадниця, Десна, Писарівка, Щітки);

5. live_vin_1_3, live_vin_5_10, live_vin_always, live_vin_10+, live_vin_less_1 – кількість років, що прожили у Вінницькій ОТГ (від 1 до 3 років, від 5 до 10, завжди жили, понад 10+ років, менше одного року);

6. work_have, work_dont – працевлаштовані та не працевлаштовані респонденти;

7. categories_of_the_population_idp, categories_of_the_population_not_related, categories_of_the_population_retired, categories_of_the_population_handicapped_person, categories_of_the_population_student – приналежність респондентів до певних соціальних категорій населення (внутрішньо переміщені особи, не належать до жодної із зазначених категорій, пенсіонери, люди з інвалідністю, студенти);

8. channel_com_dig_email, channel_com_dig_official_site, channel_com_dig_internet, channel_media_mass, channel_media_social_ads, channel_app_viber, channel_app_instagram, channel_app_telegram, channel_app_facebook, channel_app_додаток WhatsApp – цифрові канали комунікації, якими користуються респонденти (електронна пошта, офіційний сайт громади або установ, інтернет у цілому, масові медіа, реклама у соціальних мережах мобільні додатки: Viber, Instagram, Telegram, Facebook, додаток WhatsApp).

Дані зібрані за допомогою опитування Вінницької територіальної громади представлено на рисунках 2.1. – 2.3.

1	num_male	num_female	age_young	age_middle	age_senior	age_elderly	edu_incomplete_higher	edu_full_higher	edu_vocational	edu_basic_higher
2		female		25-45				full_higher		
3		female		25-45				full_higher		
4		female		25-45				full_higher		
5		female			46-60				vocational	
6		female		25-45				full_higher		
7		female	15-24				incomplete_higher			
8		female	15-24				incomplete_higher			
9		female			46-60		incomplete_higher			
10		female				60+				basic_higher
11	male			25-45						basic_higher
12		female			46-60			full_higher		
13		female				60+		full_higher		
14		female		25-45						basic_higher
15		female				60+			vocational	
16		female		25-45			incomplete_higher	full_higher		
17		female		25-45						

Рисунок 2.1 – Дані опитування Вінницької територіальної громади (блок 1)

1	loc_vinnytsia_city	loc_havryshivka	loc_vinnytski_khutory	live_vin_1_3	live_vin_5_10	work_have	work_dont	categories_population	categories_population	categories_population
2	vinnytsia_city				5-10	have		idp		
3			vinnytski_khutory		5-10	have		idp		
4	vinnytsia_city			1-3		have			not_related	
5		havryshivka			5-10	have			not_related	
6	vinnytsia_city				5-10		dont_work			student
7		havryshivka			5-10	have				student
8		havryshivka			5-10		dont_work			student
9	vinnytsia_city			1-3				idp		
10			vinnytski_khutory		5-10	have			not_related	
11	vinnytsia_city			1-3		have			not_related	
12			vinnytski_khutory	1-3		have		idp		
13	vinnytsia_city				5-10		dont_work		not_related	
14			vinnytski_khutory	1-3		have		idp		
15		havryshivka		1-3		have		idp		
16			vinnytski_khutory	1-3			dont_work			student
17		havryshivka		1-3		have		idp		

Рисунок 2.2 – Дані опитування Вінницької територіальної громади (блок 2)

1	channel_app_viber	channel_app_instagram	channel_app_telegram	channel_app_facebook	channel_app_whatsapp
2	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
3	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
4	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
5		app_instagram	app_telegram	app_facebook	
6	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
7	app_viber	app_instagram		app_facebook	app_whatsapp
8	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
9	app_viber	app_instagram	app_telegram	app_facebook	
10	app_viber		app_telegram	app_facebook	app_whatsapp
11	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
12	app_viber	app_instagram	app_telegram	app_facebook	
13	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp
14		app_instagram	app_telegram	app_facebook	app_whatsapp
15	app_viber	app_instagram	app_telegram		app_whatsapp
16	app_viber	app_instagram	app_telegram	app_facebook	app_whatsapp

Рисунок 2.3 – Дані опитування Вінницької територіальної громади (блок 3)

Згідно з таблицею опитування Вінницької міської територіальної громади, датасет містить 7 числових та 36 категоріальних ознак. Першим етапом обробки стало перетворення категоріальних ознак на булеві значення (True або False). На рисунку Г.1 наведено код перетворення категоріальних ознак на булеві значення.

Такий підхід дозволяє уніфікувати структуру даних, особливо у випадках, коли кожна категорія представлена окремою змінною (one-hot encoding). Наприклад, категорії віку, освіти або зайнятості, що мають кілька дискретних варіантів, були представлені як окремі бінарні змінні, що позначають наявність або відсутність відповідної характеристики. Переглянемо зміни після one-hot кодування (рис. 2.4).

	num_male	num_female	age_young	age_middle	age_senior	age_elderly	edu_basic_higher	edu_incomplete_higher	edu_full_higher	edu_vocational
0	False	True	False	False	True	False	False	False	False	False
1	False	True	False	True	False	False	False	False	False	False
2	False	True	False	False	True	False	False	False	True	False
3	False	True	False	False	True	False	False	False	True	False
4	False	True	False	True	False	False	False	False	True	False

5 rows × 11 columns

Рисунок. 2.4 – Візуалізація змін після one-hot кодування

Далі булеві значення було перетворено у числову форму за допомогою функції `convert_bool`, яка кодує True як 1, False як 0, що забезпечує сумісність з алгоритмами машинного навчання. Код функції булевого кодування подано на рисунку Г.2. На рисунку Г.3 представлено код масштабування даних методом `StandardScaler`.

Таке перетворення дало змогу отримати єдину числову структуру, яку можна аналізувати за допомогою математичних методів.

За допомогою функції `info()` переглянемо інформацію про дані (рис. 2.5). Датасет складається 45-ти числових ознак.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 404 entries, 0 to 403
Data columns (total 46 columns):
# Column      Non-Null Count  Dtype
---  -
0 num_male      404 non-null    int64
1 num_female    404 non-null    int64
2 age_young     404 non-null    int64
3 age_middle    404 non-null    int64
4 age_senior    404 non-null    int64
5 age_elderly   404 non-null    int64
6 edu_basic_higher  404 non-null    object
7 edu_incomplete_higher  404 non-null    int64
8 edu_full_higher  404 non-null    int64
9 edu_vocational  404 non-null    int64
10 edu_basic_secondary  404 non-null    int64
11 edu_academic_degree  404 non-null    int64
12 edu_only_primary  404 non-null    int64
13 edu_two_higher  404 non-null    int64
14 edu_full_secondary  404 non-null    int64
15 loc_vinnitsia_city  404 non-null    int64
16 loc_v_krushlyntsi  404 non-null    int64
17 loc_havryshivka  404 non-null    int64
18 loc_m_krushlyntsi  404 non-null    int64
19 loc_vinnitski_khutory  404 non-null    int64
20 loc_stadnitsia  404 non-null    int64
21 loc_desna     404 non-null    int64
22 loc_pysarivka  404 non-null    int64
23 loc_shchitky  404 non-null    int64
24 live_vin_1_3  404 non-null    int64
25 live_vin_5_10  404 non-null    int64
26 live_vin_always  404 non-null    int64
27 live_vin_10+  404 non-null    int64
28 live_vin_less_1  404 non-null    int64
29 work_have     403 non-null    float64
30 work_dont     404 non-null    int64
31 categories_of_the_population_idp  404 non-null    int64
32 categories_of_the_population_not_related  404 non-null    int64
33 categories_of_the_population_retired  404 non-null    int64
34 categories_of_the_population_handicapped_person  404 non-null    int64
35 categories_of_the_population_student  404 non-null    int64
36 channel_com_dig_email  404 non-null    int64
37 channel_com_dig_official_site  404 non-null    int64
38 channel_com_dig_internet  404 non-null    int64
39 channel_media_mass  404 non-null    int64
40 channel_media_social_ads  404 non-null    int64
41 channel_app_viber  404 non-null    int64
42 channel_app_instagram  404 non-null    int64
43 channel_app_telegram  404 non-null    int64
```

Рисунок 2.5 – Інформація про дані

Перевіримо дані на наявність аномалій на графіку (рис. 2.6).

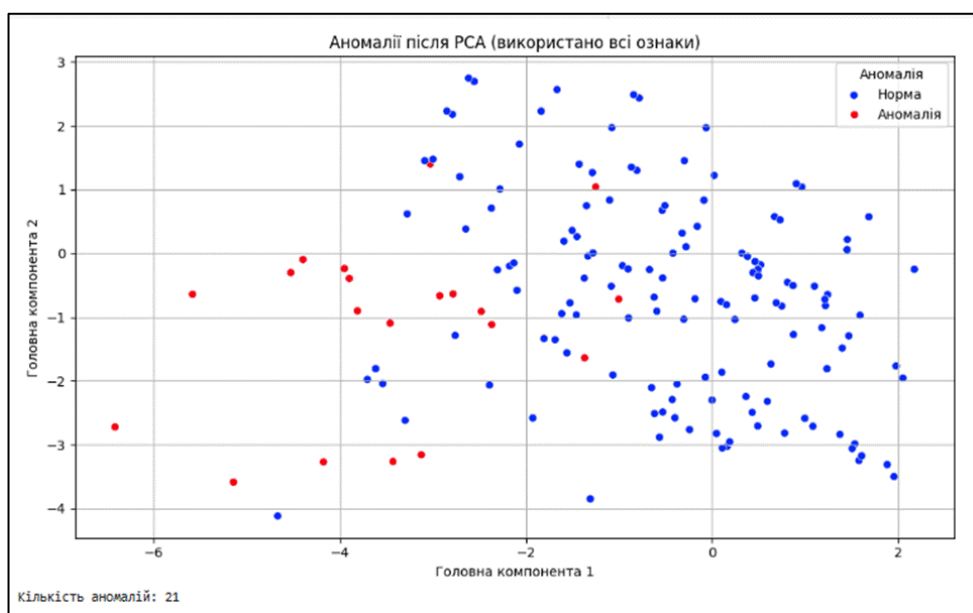


Рисунок 2.6 – Графік аномалій в даних

На рисунку 2.6 продемонстровано графік аномалій, отриманий після зниження розмірності даних методом головних компонент (Principal Component Analysis). Графік демонструє просторове розділення нормальних спостережень (сині точки) та аномальних (червоні точки). Аномалії чітко сконцентровані в лівій частині графіка, що вказує на наявність групи респондентів із характеристиками, які суттєво відрізняються від основної маси. Аналіз показав, що ці аномальні спостереження включають як жінок, так і чоловіків у приблизно рівних частках. Значна частина з них – люди середнього та старшого віку, здебільшого пенсіонери. Щодо рівня освіти, то третина мешканців мають професійно–технічну або повну вищу освіту. У цифровій поведінці користувачів переважає активне використання таких цифрових каналів як: Instagram (56%), Facebook (49%), Viber та Telegram (по 39%). Водночас офіційні цифрові платформи, такі як сайти громад або email, використовуються менш активно.

Далі побудуємо кореляційну матрицю на рисунку 2.7 для перевірки впливу ознак на досліджуванні характеристики. Кольорова шкала буде варіюватися від -1 (сильної негативної кореляції, темно–фіолетовий колір) до $+1$ (сильної позитивної кореляції, яскраво–жовтий). Значення близькі до 0 (бірюзовий колір) вказують на слабку або відсутню кореляцію.

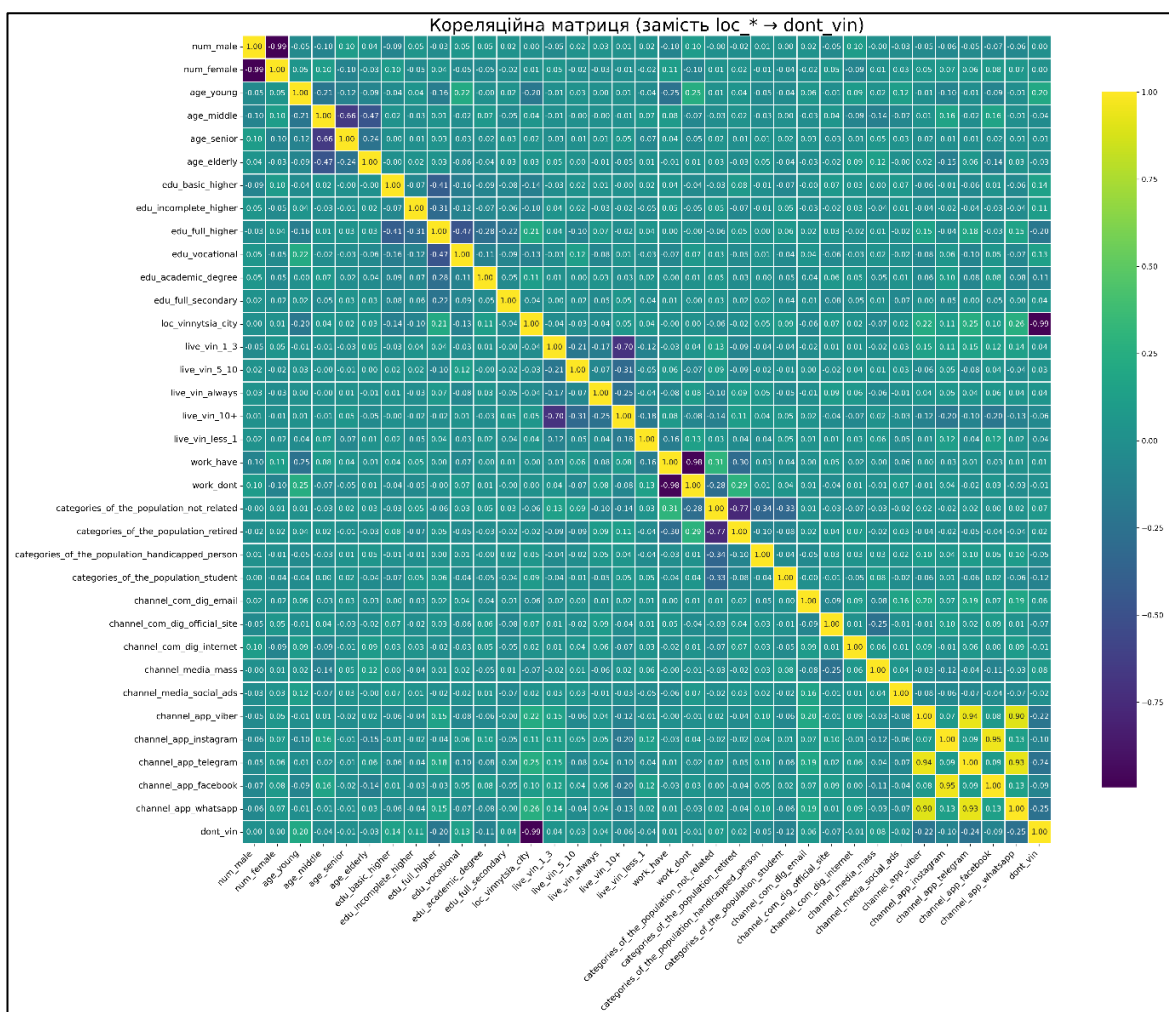


Рисунок 2.7 – Кореляційна матриця числових даних

На рисунку 2.7 зображена матриця кореляції між різними змінними, що впливають на передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали.

Матриця демонструє, що найвища позитивна кореляція (до 0,81) спостерігається між різними каналами комунікації, зокрема між месенджерами (Viber, WhatsApp, Facebook та Instagram). Це демонструє, що ці платформи часто використовуються одночасно одними й тими ж респондентами – що можна трактувати як наявність багатоканальної цифрової взаємодії.

Очікувано фіксуються сильні негативні кореляції між взаємовиключними категоріями, наприклад, між змінними стать – чоловіки та жінки (кореляція – 1.00), а також між змінними щодо працевлаштування – має роботу та безробітній. Це пояснюється тим, що відповідні категорії представляють альтернативні

варіанти одного і того ж фактора – статі або зайнятості, і в одному записі вони не можуть бути одночасно активними.

Більшість інших соціально–демографічних ознак, таких як вік, рівень освіти або тривалість проживання у Вінниці, демонструють слабкий кореляційний зв'язок із цифровою активністю. Це вказує на широку доступність цифрових каналів для різних користувачів.

2.2 Відбір ознак для передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали

Зростання можливостей інформаційно–комунікаційних технологій, зокрема штучного інтелекту (ШІ) [13], відкрило нові горизонти для збирання, обробки та візуалізації даних у сфері передбачення взаємодії респондентів через цифрові канали. Поєднання даних, отриманих за допомогою методів машинного навчання, з результатами традиційного соціологічного чи статистичного аналізу дозволяє більш обґрунтовано оцінити ключові чинники, що впливають на вибір цифрових каналів, виявляти тенденції у поведінці серед мешканців громади, а також ухвалювати ефективні рішення щодо впровадження комунікаційних стратегій, залучення цільової аудиторії та підвищення рівня соціальної взаємодії в межах територіальної громади.

Побудуємо графік розподілу за ознакою «Стать» (рис. 2.8). З діаграми бачимо, що кількість респондентів жіночої статі переважає над чоловічою.

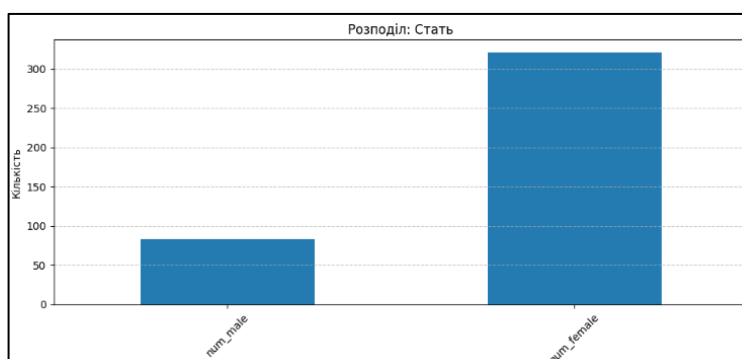


Рисунок 2.8 – Графік розподілу за ознакою «Стать»

На рисунку 2.9 представлено два боксплоти, які демонструють, що медіана вища у групі жінок й у середньому вони використовують більше цифрових каналів, ніж чоловіки.

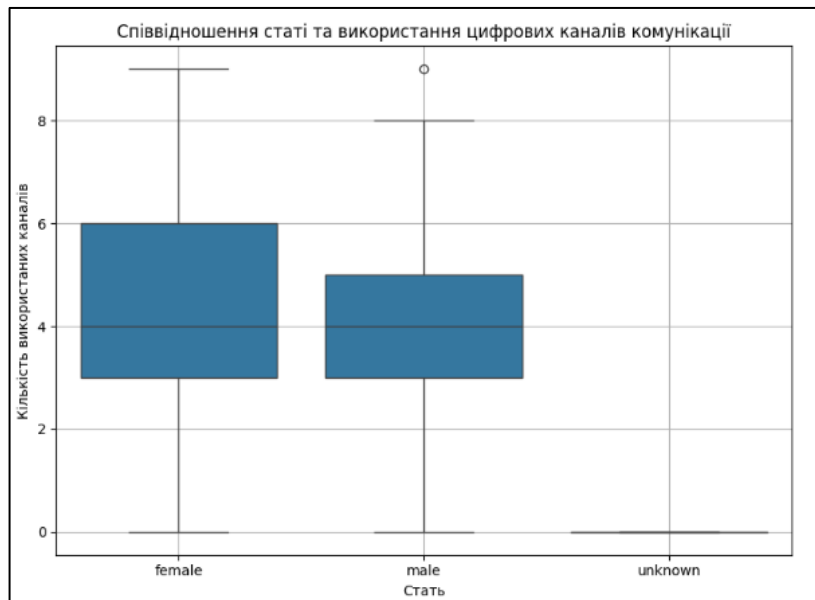


Рисунок 2.9 – Співвідношення статі та використання цифрових каналів

На рисунку 2.10 зображений графік розподілу за ознакою «Освіта». Видно, що більшість респондентів мають повну вищу або професійну освіту.

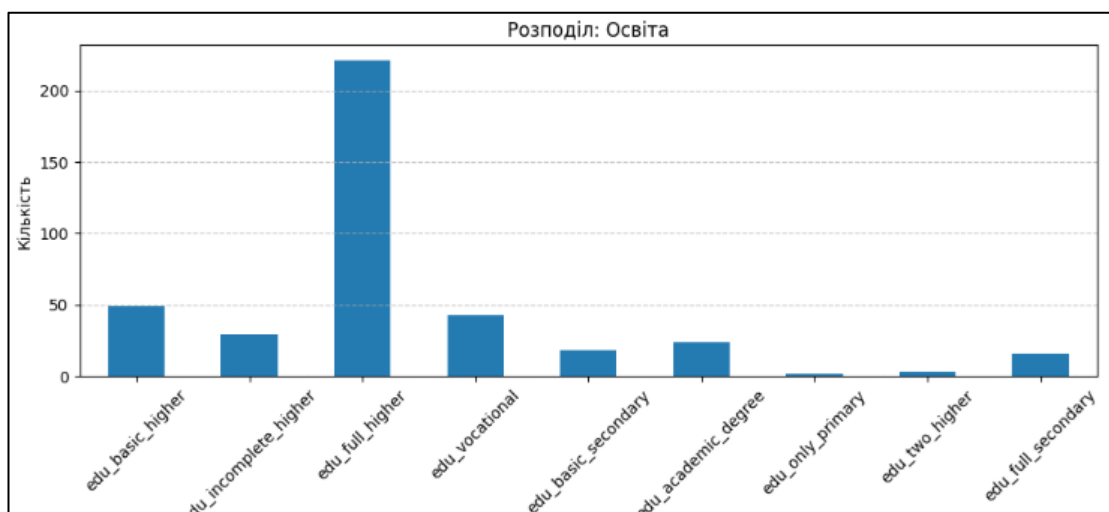


Рисунок 2.10 – Графік розподілу за ознакою «Освіта»

Згідно з боксплот представленим на рисунку 2.11, групи мають порівняно подібну медіану використаних каналів у більшості освітніх категорій. Найвищу варіативність у використанні спостерігаємо серед мешканців із повною загальною та середньою освітою. Респонденти з повною вищою освітою та академічним ступенем мають стабільно вищі медіанні показники, що говорить про зв'язок між рівнем освіти та інтенсивністю цифрової взаємодії.

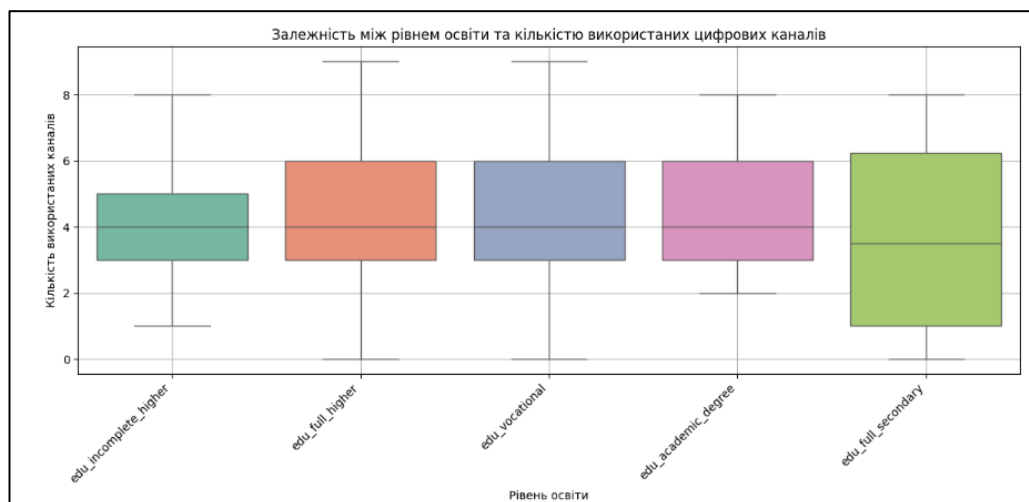


Рисунок 2.11 – Співвідношення між рівнем освіти та використання цифрових каналів

На рисунку 2.12 зображений графік розподілу за ознакою «Вік».

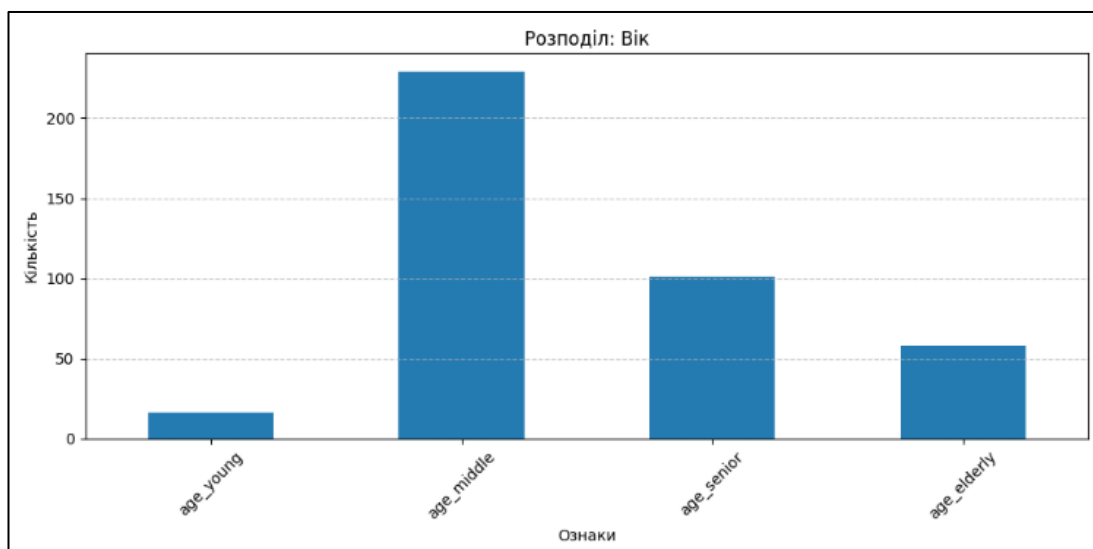


Рисунок 2.12 – Графік розподілу за ознакою «Вік»

З рисунка бачимо, що найбільшу частку респондентів становлять люди середнього віку (25–45 років).

Боксплот–графік, що представлений на рисунку 2.13, демонструє, що наймолодша група має найвищу медіану використання цифрових каналів та найширший інтерквартильний розмах. У старшій та середній групі рівень використання каналів є стабільним, з нижчою варіативністю. Найстарші респонденти мають найнижчу варіативність та найменшу кількість використовуваних каналів, що вказує на меншу цифрову залученість літніх людей. Отже, вікова група прямо впливає на рівень цифрової взаємодії.

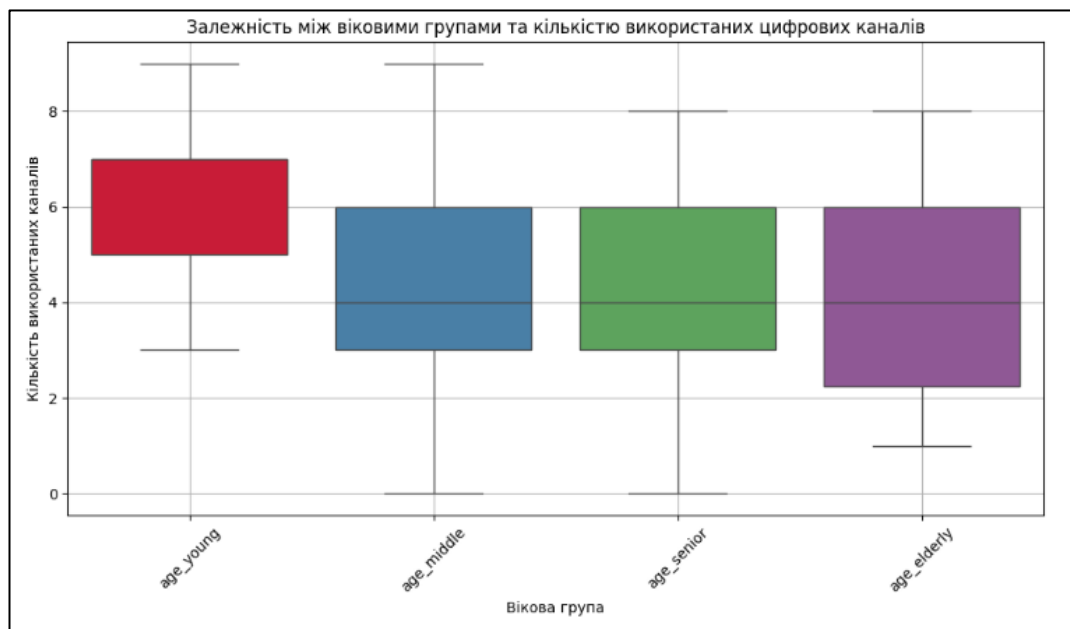


Рисунок 2.13 – Співвідношення між віковими групами та кількістю використаних цифрових каналів

На рисунку 2.14 зображено графік розподілу за ознаками «Місто проживання» та «Кількість років, прожитих у Вінницькій ОТГ». Проаналізувавши графік, можна зробити висновок, що основна частина респондентів проживає у Вінниці або прилеглих населених пунктах, зокрема у Вінницьких Хуторах і Стадниці. Більшість респондентів мешкає у громаді понад 10 років або з народження.

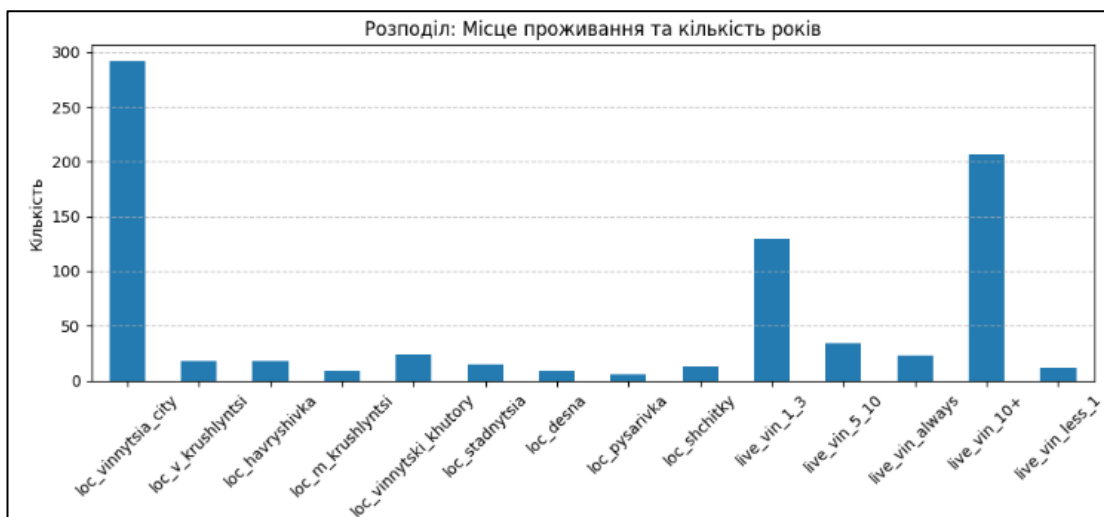


Рисунок 2.14 – Графік розподілу за ознаками «Місто проживання» та «Кількість років, прожитих у Вінницькій ОТГ»

На рисунку 2.15 представлено графік співвідношення між місцем проживання та кількістю використаних цифрових каналів.

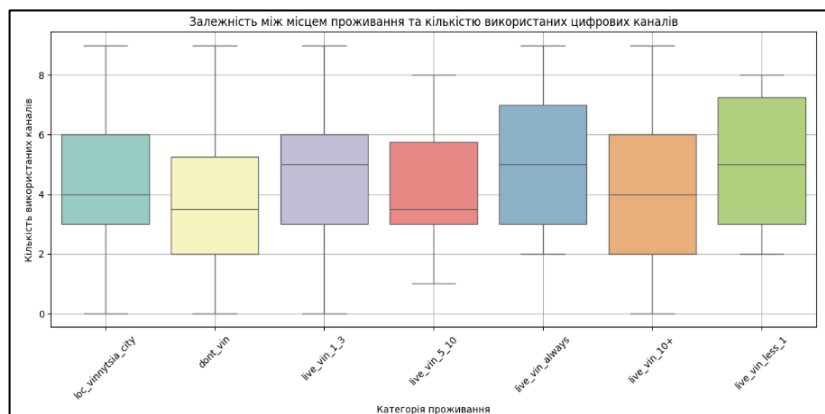


Рисунок 2.15 – Співвідношення між місцем проживання та кількістю використаних цифрових каналів

Боксплот графік демонструє, як місце та тривалість проживання у Вінницькій громаді впливають на кількість використаних цифрових каналів. Жителі міста Вінниця використовують трохи більше каналів у середньому, ніж мешканці навколишніх сіл (dont_vin). Це може вказувати на кращий доступ до цифрових каналів у міських жителів. Цікаво, що мешканці, що проживають у Вінницькій громаді від 1 до 3 років показують найвищу медіану цифрової

активності, ймовірно через необхідність активного пошуку інформації та взаємодії з місцевими службами. Ті, хто проживає у місті понад 10 років, навпаки, демонструють менший рівень активності, що може бути наслідком звички користуватись уже знайомими каналами або зниження потреби в активній комунікації. Респонденти з решти категорій демонструють помірний і стабільний рівень активності, без різких відхилень. Місце проживання (місто/село) та тривалість проживання в громаді впливають на цифрову залученість мешканців.

На рисунку 2.16 зображено розподіл за ознакою «Працевлаштування»

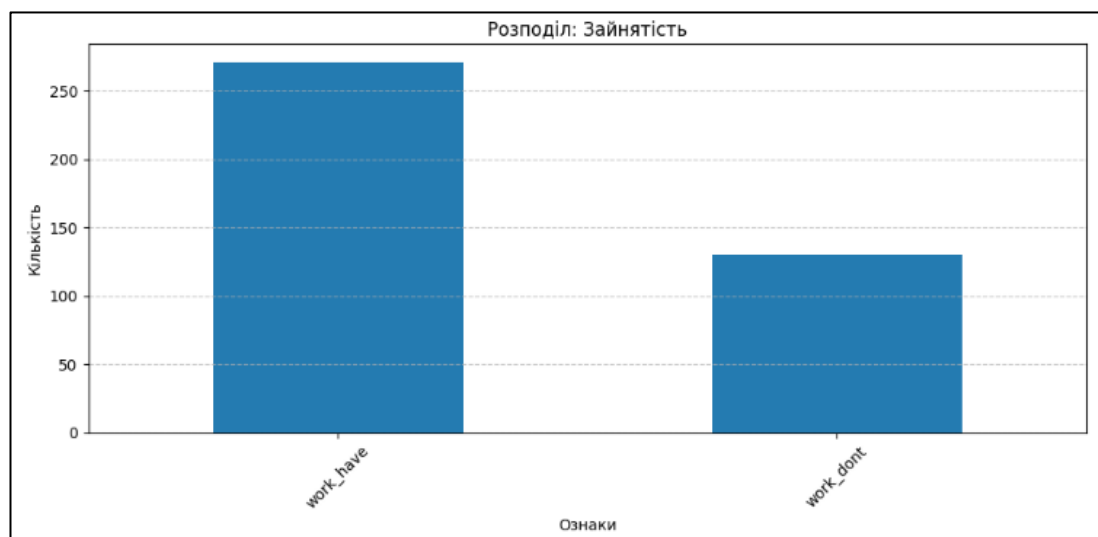


Рисунок 2.16 – Графік розподілу за ознакою «Працевлаштування»

Як видно у вибірці переважають працевлаштовані респонденти, але також представлені й ті, хто не працює.

На рисунку 2.17 представлено два боксплоти, що демонструють, як працевлаштованість впливає на активність використання цифровими каналами комунікації. Обидві групи мають однакову медіану використання. Розподіл значень та інтерквартильний розмах практично не відрізняються, тобто зайнятість не створює суттєвих відмінностей у цифровій активності користувачів. Люди активно використовують цифрові канали як у робочому, так і в неробочому середовищі.

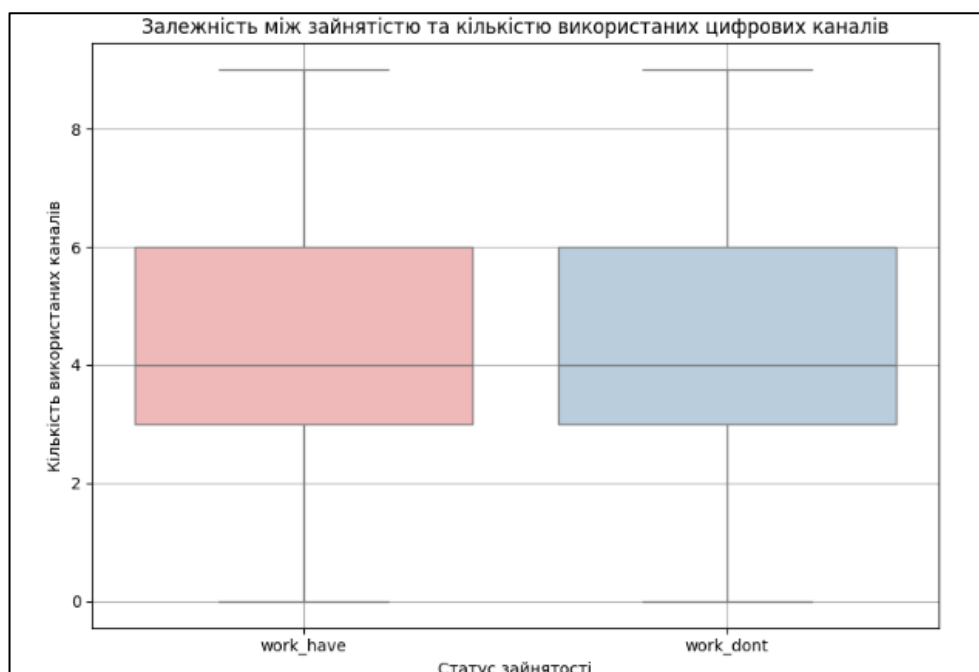


Рисунок 2.17 – Залежність між зайнятістю та кількістю використаних цифрових каналів

На рисунку 2.18 зображений розподіл респондентів за ознакою «Категорія населення». Серед учасників є студенти, пенсіонери, особи з інвалідністю та респонденти, що не належать до жодної з визначених соціальних груп.

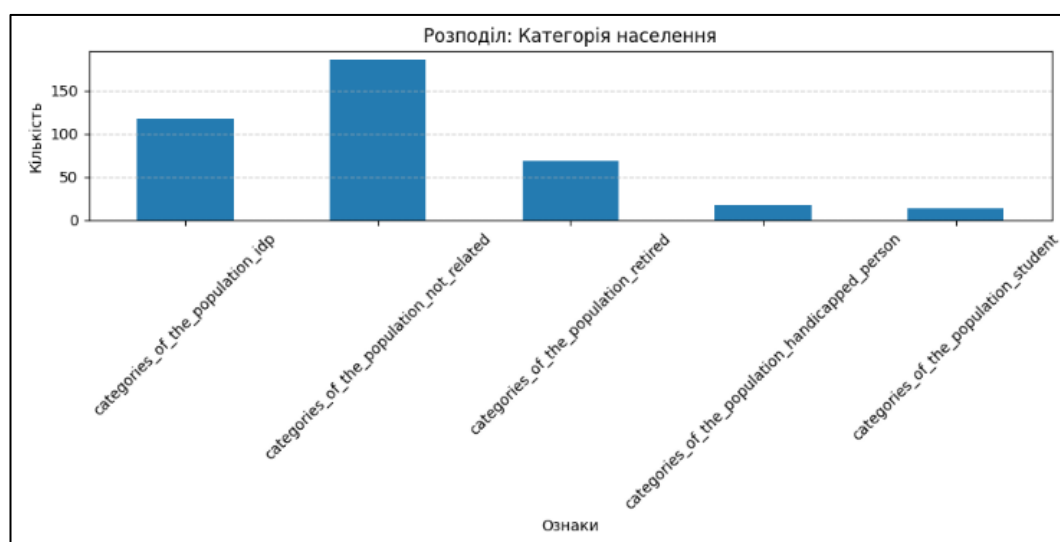


Рисунок 2.18 – Графік розподілу за ознакою «Категорія населення»

На рисунку 2.19 представлена залежність між категоріями населення та кількістю використаних цифрових каналів.

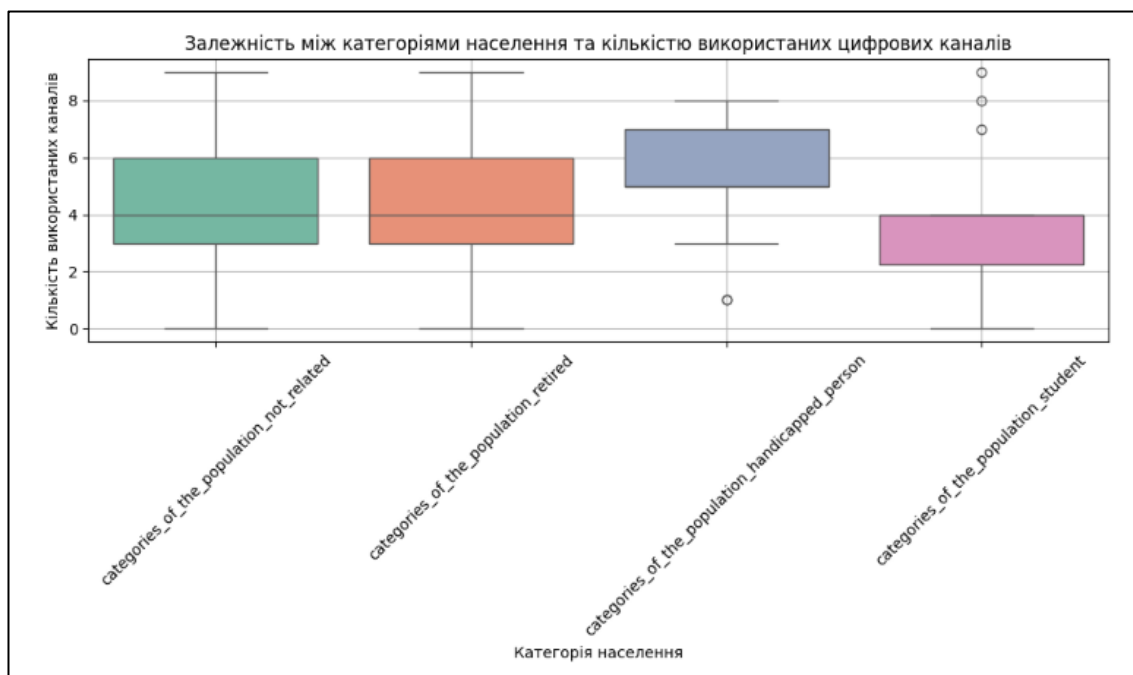


Рисунок 2.19 – Залежність між категоріями населення та кількістю використаних цифрових каналів

Графік ілюструє, як різні соціальні групи використовують цифрові канали комунікації. Найбільш активними користувачами є група люди з інвалідністю, які мають найвищу медіану та найменший розкид значень. Це вказує на сталість та регулярність використання цифрових каналів у цій групі.

Пенсіонери та респонденти, які непов'язані з іншими категоріями демонструють подібний один одному рівень цифрової активності, хоча мають трохи ширший діапазон і меншу медіану

Студенти мають найнижчу медіану, але наявність виражених верхніх викидів вказує на те, що наявні окремі студенти з дуже високим рівнем цифрової активності, в той час як більшість використовує менше каналів.

Побудуємо графік розподілу популярності цифрових каналів комунікації серед мешканців Вінницької міської територіальної громади (рис. 2.20).

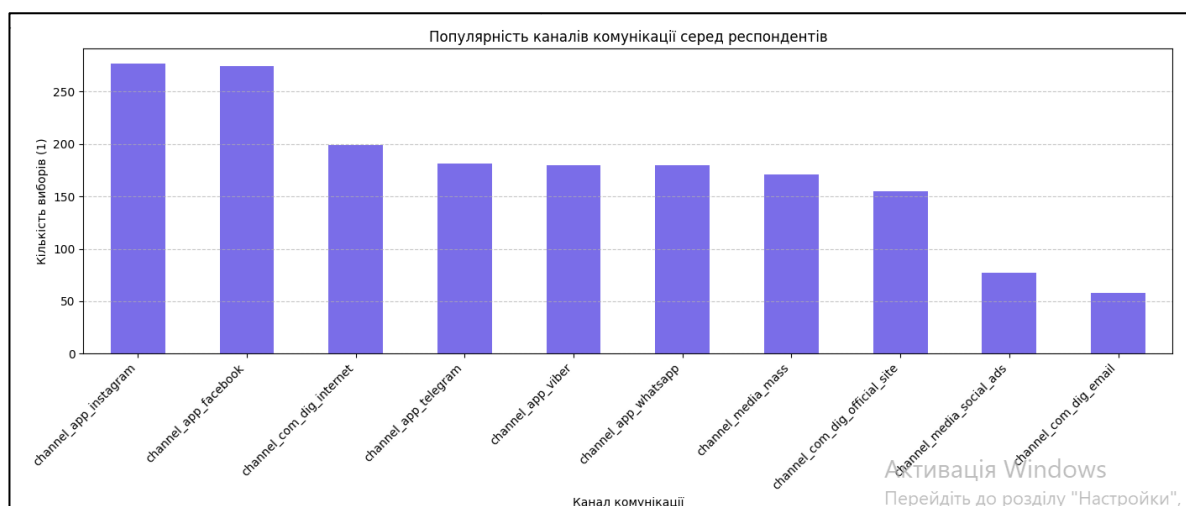


Рисунок 2.20 – Розподіл популярності цифрових каналів комунікації серед респондентів

Графік ілюструє рівень популярності різних каналів комунікації, які обирали мешканці Вінницької міської територіальної громади. Найпопулярніші канали – це мобільні застосунки Instagram та Facebook, які обрали понад 270 респондентів. Месенджери (Telegram, Viber, WhatsApp) мають майже однаковий рівень популярності кожен з них набрав близько 180 згадок. Мас-медіа та офіційні сайти ще залишаються актуальними, але поступаються за охопленнями цифровим платформам для миттєвого обміну месенджером. Рекламу у соцмережах та електронну пошту – найменш популярні. Громада значною мірою орієнтується на соціальні мережі та месенджери як на основні інструменти соціальної взаємодії.

Продовжимо аналіз даних і застосуємо метод `describe()`, щоб переглянути загальну статистичну характеристику основних змінних (рис. 2.21–2.22).

	num_male	num_female	age_young	age_middle	age_senior	age_elderly	edu_basic_higher	edu_incomplete_higher	edu_full_higher	edu_vocational	edu_academic_degree	edu_full_secondary	loc_vinnitsia_city
count	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000
mean	0.204938	0.792593	0.041975	0.565432	0.249383	0.14321	0.120988	0.071605	0.548148	0.155556	0.059259	0.039506	0.720988
std	0.404156	0.405951	0.200781	0.496313	0.433191	0.35072	0.326516	0.258152	0.498292	0.362882	0.236401	0.195037	0.449068
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Рисунок 2.21 – Описова статистика описуваних змінних

loc_m_krushlyntsi	loc_vinnytski_khutory	loc_stadnytsia	loc_desna	loc_pysarivka	loc_shchitky	live_vin_1_3	live_vin_5_10	live_vin_always	live_vin_10+	live_vin_less_1	work_have	work_dont
405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000
0.022222	0.059259	0.037037	0.022222	0.014815	0.032099	0.318519	0.083951	0.056790	0.511111	0.029630	0.669136	0.320988
0.147588	0.236401	0.189086	0.147588	0.120961	0.176480	0.466478	0.277657	0.231727	0.500495	0.169773	0.471106	0.467433
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Рисунок 2.22 – Описова статистика описуваних змінних

Статистичні дані, представлені на рисунках 2.21 – 2.22 надають детальний опис змінних, що впливають на соціальну взаємодію Вінницької міської територіальної громади через цифрові канали.

Аналіз за гендерним розподілом демонструє:

- жінки становлять переважну більшість опитаних – близько 79% (mean ≈ 0.79);
- чоловіків – респондентів всього 20% (mean ≈ 0.20).

За віковим розподілом:

- найбільшу частку становлять респонденти віком 25–45 років (≈ 0.57);
- молодь (15–25 років) – представлена найменше ($\approx 4\%$);
- старша вікова категорія (45–59 років) і особи пенсійного віку (60+) разом складають близько 39% опитаних.

За рівнем освіти:

- повну вищу освіту мають понад половина респондентів (54.8%);
- професійну технічну освіту – приблизно ($\approx 15\%$);
- неповну вищу освіту мають – близько ($\approx 7\%$);
- лише середню освіту мають ($\approx 4\%$).

За місцем проживання:

- абсолютна більшість респондентів мешкає у місті Вінниця (72%);
- інші населені пункти представлені незначною часткою респондентів (менше 5% кожен).

За кількістю років прожитих у Вінницькій громаді:

– найбільша група респондентів проживає у місті більше 10 років ($\approx 51\%$);

– 32% мешкають у Вінниці 1–3 роки;

За ознакою працевлаштування:

– працюючі респонденти становлять ($\approx 67\%$);

– не працюють ($\approx 32\%$).

За категорією населення:

– найпоширеніша відповідь – «не належу до жодної з категорій» ($\approx 74.8\%$);

– пенсіонери складають – ($\approx 17\%$);

– особи з інвалідністю та студенти – незначна частка (менше 5%).

Цифрові канали:

– інтернет є найпопулярнішим каналом – ним користуються 50,1% респондентів;

– офіційні сайти – 38,3%;

– email – лише 14,3%;

– мас-медіа – 42.2%;

– соціальна реклама є обмеженою – лише 19%;

– Instagram ($\approx 70\%$);

– Facebook ($\approx 67\%$);

– Telegram використовують ($\approx 44\%$) опитаних;

– Viber та WhatsApp – по $\approx 46\%$ кожен, отже месенджери є значущі як ключові канали комунікації.

Перед тренуванням моделей важливо попередньо й незалежно підготувати дані: оптимізувати їх, усунути шумові значення, об'єднати малі чи подібні за змістом категорії тощо. Це дозволить краще зрозуміти взаємозв'язки між характеристиками респондентів і використанням цифрових каналів.

На початковому етапі об'єднаємо малі категорії, які було виявлено під час статистичного аналізу даних. На рисунку Г.4. наведено код об'єднання ознак.

Ще однією важливою проблемою є наявність повторюваних, так званих корельованих ознак – змінних, що мають схожу або дубльовану інформацію. Наприклад, як ознаки `work_have` та `work_dont` є логічними інверсіями одна одної. Між змінними є сильна кореляція (коефіцієнт $> 0,9$), моделі важко, а іноді й неможливо визначити, яка з ознак є більш інформативною і це створює надмірність у даних, що в свою чергу може призвести до перенавчання моделі. Тому знайдемо та видалимо сильно корельовані пари, а саме менш значущу ознаку або ту, що дублює інформацію іншої.

На рисунку Г.5 представлено код автоматичного виявлення та видалення кореляційних пар ознак. Цільові змінні були позначені як недоторкані, оскільки мають важливе значення для дослідження і не підлягають видаленню.

На рисунку 2.23 наведено перелік ознак, які зберігають найбільшу інформативність і будуть використані для подальшого тренування моделей.

```
Список ознак у датасеті:
- num_female
- age_young
- age_middle
- age_senior
- age_elderly
- live_vin_5_10
- live_vin_always
- live_vin_10+
- work_dont
- categories_of_the_population_not_related
- categories_of_the_population_retired
- categories_of_the_population_handicapped_person
- categories_of_the_population_student
- channel_com_dig_email
- channel_com_dig_official_site
- channel_com_dig_internet
- channel_media_mass
- channel_media_social_ads
- channel_app_viber
- channel_app_instagram
- channel_app_telegram
- channel_app_facebook
- channel_app_whatsapp
- loc_non_vinnytsia
- edu_secondary
- live_vin_0_3
```

Рисунок 2.23 – Ключові ознаки

Результати аналізу та відбору ознак для передбачення соціальної взаємодії у Вінницькій міській територіальній громаді дозволив виявити та виокремити найінформативніші змінні, які будуть використанні для дослідження.

2.3 Тренування моделей передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали

Так як головна задача дослідження є передбачення, й для цього є певні розміщені дані з відомими правильними відповідями, підхід машинного навчання, який ми застосуємо – це навчання «з учителем». У межах цього підходу вирішується задача класифікації [16]. Також проаналізуємо переваги основних моделей машинного навчання, що застосовуються для розв'язання задач класифікації.

1. Дерево рішень (Decision Tree) – це спосіб представлення правил у вигляді ієрархічної, послідовної структури, де кожному об'єкту відповідає єдиний вузол, що дає рішення. Під правилом розуміється логічна конструкція, представлена у вигляді «якщо . то...». Дерева рішень відмінно справляються із завданнями класифікації, тобто віднесення об'єктів до одного із заздалегідь відомих класів, а цільова змінна повинна мати дискретні значення [17].

2. Метод опорних векторів – класифікаційний алгоритм, що добре працює з обмеженою кількістю навчальних даних. Знаходить гіперплощину, оптимально розділяє 2 класи з максимальним відступом. Завдяки високій точності й здатності узагальнювати, застосовується у складних задачах класифікації [18].

3. Логістична регресія (Logistic Regression) є ефективним інструментом для прогнозного моделювання й широко використовується в задачах класифікації, особливо для бінарних результатів. Вона дозволяє оцінити ймовірність настання певної події, враховуючи взаємозв'язок між залежною змінною та однією або кількома незалежними. Завдяки здатності працювати з категоріальними даними та простоті реалізації, логістична регресія є корисною моделлю в навчанні для прийняття рішень на основі обмежених, але інформативних ознак [19].

4. Випадковий ліс (Random Forest) – це потужний алгоритм класифікації, який використовує ансамблеве навчання для підвищення точності за рахунок об'єднання багатьох дерев рішень. Він усуває типові недоліки окремих дерев, зменшуючи ризик перенавчання та забезпечуючи стабільні результати навіть на великих наборах даних. Завдяки процедурі голосування більшості, випадковий ліс ефективно класифікує об'єкти, що робить його корисним інструментом для складних задач у сфері машинного навчання [20].

5. Нейронні мережі використовують попередньо розмічену навчальну множину, яка містить пари «вхід вихід». У процесі навчання мережа ітеративно коригує вагові коефіцієнти між нейронами для мінімізації похибки між прогнозованим і правильним виходом. Після досягнення прийнятного рівня точності, ваги фіксуються, і модель готова до класифікації нових даних [21].

6. Градієнтний бустинг – метод машинного навчання для задач класифікації та регресії, поєднує в собі ансамбль слабких моделей, зазвичай дерев рішень, для поступового покращення точності прогнозів. Він використовує градієнтний спуск для оптимізації функціоналу втрат, фокусуючись на прикладах, де попередні моделі давали найбільші похибки [22].

Кожен розглянутий метод має свої сильні сторони й обмеження, вибір конкретної моделі залежить від завдань дослідження, обсягу та якості доступних даних, вимог до інтерпретації результатів та наявних обчислювальних ресурсів.

Для реалізації моделювання імпортуємо необхідні бібліотеки машинного навчання в проєкт у Google Colab, а саме LogisticRegression , Random ForestClassifier і XGBClassifier. Для оцінки якості класифікаційних моделей використаємо такі метрики:

- recall_score – відображає здатність моделі виявляти всі позитивні класи у вибірці;
- f1_score – обчислює середнє гармонійне між точністю та повнотою;
- accuracy_score – визначає загальну точність класифікації;
- roc_auc_score – визначає площу під ROC-кривою, що характеризує здатність моделі розрізняти класи;

– `precision_score` – показує, наскільки точно модель передбачає позитивні класи [23].

У програму додаємо чотири метрики `accuracy_score`, `precision_score`, `recall_score`, `f1_score`, `roc_auc_score` для оцінювання якості роботи моделей.

Перш ніж почати тренувати моделі, важливо налаштувати їхні гіперпараметри, що дозволить підвищити ефективність та якість їхньої роботи. Для автоматизації цього процесу, використаємо функцію `GridSearchCV` [24] з бібліотеки `sklearn.model_selection`, що дозволить підібрати оптимальні параметри. Всі моделі мають свої параметри, тож для `XGBoost` налаштуємо такі:

– `max_depth` – глибина дерева, якщо буде значення занадто велике, модель може перенавчитись;

– `min_child_weight` – мінімальна вага в дочірньому вузлі. Вона контролює перенавчання;

– `gamma` – мінімальне зменшення втрат для додаткового поділу дерева. Чим більше її значення, тим обережніше модель розгалужує вузли;

– `learning_rate` – швидкість навчання. Контролює, наскільки сильно кожне нове дерево впливає на загальний результат. Рекомендовано значення від 0.01 до 0.2;

– `colsample_bytree` – частка ознак, яку використовує кожне дерево. Завеликі значення параметра можуть спричинити перенавчання;

– `subsample` – частина даних, що використовується для побудови кожного дерева. Занадто низьке значення може призвести до недонавчання;

– `n_estimators` – кількість дерев в ансамблі. Цей параметр варто балансувати з `learning_rate`, щоб уникнути перенавчання [25].

Ці параметри є ключовими для точного налаштування `XGBoost` для досягнення оптимальної якості класифікації.

На рисунку Г.6 продемонстрований код, у якому здійснено автоматизований підбір гіперпараметрів для моделі `XGBoost` й змінної `Telegram`. Після навчання моделі здійснено оцінку її точності на тренувальній вибірці. На

рисунку 2.24 переглянемо найкращі гіперпараметри моделі XGBoost й отриману точність на тренувальних даних для змінної Telegram.

Точність XGBoost на тренувальних даних: 0.9629629629629629
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0.1, 'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 3, 'n_estimators': 200, 'subsample': 0.8}

Рисунок 2.24 – Метричні показники Telegram

Для покращення якості моделі Logistic Regression підберемо оптимальні значення гіперпараметрів. Цей процес часто називають тюнінгом, тобто налаштуванням моделі з метою досягнення кращих результатів. У бібліотеці `scikit-learn` для LogisticRegression найбільш значущими параметрами є:

- `solver` – алгоритм, який використовується для оптимізації. Варто враховувати, що деякі алгоритми підтримують лише певні типи регуляризації;
- `penalty` – тип регуляризації, який застосовується до моделі;
- `C` – обернена сила регуляризації. Чим менше значення параметра, тим сильніша регуляризація впливає на модель;
- `max_iter` – максимальна кількість ітерацій, що необхідна для досягнення оптимального рішення [26].

На рисунку Г.7 показано код моделі Logistic Regression з автоматизованим підбором гіперпараметрів за допомогою `GridSearchCV`. Параметри `C`, `solver`, `penalty`, `max_iter` мають суттєвий вплив на результат моделі та включені до тюнінгу.

У результаті навчання на тренувальних даних було знайдено найкращу конфігурацію, що забезпечує високу якість класифікації. На рисунку 2.25 подано отриману точність моделі для змінної Telegram на тренувальних даних, а також значення підібраних оптимальних гіперпараметрів.

Точність Logistic Regression на тренувальних даних: 0.9691358024691358
Найкращі гіперпараметри Logistic Regression: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Рисунок 2.25 – Точність на тренувальних даних Telegram

Logistic Regression правильно класифікувала 96.91% прикладів з навчального набору.

Для моделі Random Forest Classifier також підберемо відповідні гіперпараметри, що впливають як на точність, так і на продуктивність:

- `max_depth` – визначає максимальну глибину кожного дерева. Занадто велика глибина може спричинити перенавчання;
- `min_samples_leaf` – мінімальна кількість вибірок, яка має бути в листовому вузлі. Цей параметр допомагає зменшити ризик перенавчання;
- `bootstrap` – визначає, чи використовуються бутстрап-вибірки при побудові дерев. За замовчуванням має значення `True`, що сприяє більшій варіативності моделі;
- `n_estimators` – кількість дерев у лісі. Більше дерев краще навчання, але й більше обчислювальне навантаження;
- `min_samples_split` – мінімальна кількість вибірок, необхідних для поділу вузла;
- `max_features` – кількість ознак, які враховуються при поділі [27].

На рисунку Г.8 продемонстровано приклад налаштованої моделі Random Forest Classifier. На рисунку 2.26 вписано найкращу конфігурацію та точність моделі на тренувальних даних.

Точність Random Forest на тренувальних даних: 0.9907407407407407 Найкращі гіперпараметри Random Forest: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}
--

Рисунок 2.26 – Точність моделі випадковий ліс на тренувальних даних Telegram

З рисунку 2.26 видно, що модель випадкового лісу правильно класифікує близько 99% навчальних прикладів.

Для подальшого моделювання використовуємо цільову змінну WhatsApp. Оскільки код для обробки змінної WhatsApp є аналогічним до використаного для змінної Telegram, надалі будемо зосереджувати уваги виключно на метричних показниках цільових змінних.

Результат навчання класифікаційних моделей для змінної WhatsApp показано на рисунку 2.27, а саме значення точності на тренувальній вибірці та перелік оптимальних гіперпараметрів, отриманих у результаті автоматизованого підбору.

```
Точність XGBoost на тренувальних даних: 0.9783950617283951
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 200, 'subsample': 0.8}

Точність Logistic Regression на тренувальних даних: 0.9598765432098766
Найкращі гіперпараметри Logistic Regression: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9629629629629629
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
```

Рисунок 2.27 – Метричні показники для змінної WhatsApp

Модель XGBoost для змінної WhatsApp правильно класифікує близько 98% навчальних прикладів, Random Forest 96% , а Logistic Regression 96%.

Такий підхід ми застосуємо до всіх ознак, щоб об'єктивно оцінити продуктивність кожної моделі на даних. Якщо для певної змінної, моделі будуть демонструвати високу точність класифікації то це буде означати про потенційну цінність цієї ознаки для передбачення соціальної взаємодії Вінницької ОТГ через цифрові канали. Натомість змінні, моделі яких матимуть надто низьку точність будуть виключатися з подальшого передбачення. Таким чином відсіємо слабкі ознаки, зосередивши увагу на цифрових каналах із практичною значущістю.

Наступна зміна для моделювання Facebook. Результати навчання моделей наведено на рисунку 2.28.

```
Точність XGBoost на тренувальних даних: 0.9876543209876543
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 1, 'n_estimators': 200, 'subsample': 0.8}

Точність Logistic Regression на тренувальних даних: 0.9814814814814815
Найкращі гіперпараметри Logistic Regression: {'C': 10, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9814814814814815
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
```

Рисунок 2.28 – Точність моделей на тренувальних даних для змінної Facebook

Модель XGBoost для змінної Facebook правильно класифікує близько 99% навчальних прикладів, Random Forest 98% та Logistic Regression 98%.

Далі розглядаємо Viber. Результати навчання моделей показано на рисунку 2.29.

```

Точність XGBoost на тренувальних даних: 0.9691358024691358
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 3, 'n_estimators': 200, 'subsample': 1.0}

Точність Logistic Regression на тренувальних даних: 0.9629629629629629
Найкращі гіперпараметри Logistic Regression: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9783950617283951
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

```

Рисунок 2.29 – Точність моделей на тренувальних даних для змінної Viber

Точність моделей для змінної Viber на тренувальних даних XGBoost близько 97%, Random Forest 98% та Logistic Regression 96% .

Розглянемо наступну змінну – Instagram. Результати навчання моделей показано наведено на рисунку 2.30.

```

Точність XGBoost на тренувальних даних: 0.9876543209876543
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 1.0}

Точність Logistic Regression на тренувальних даних: 0.9845679012345679
Найкращі гіперпараметри Logistic Regression: {'C': 10, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9907407407407407
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}

```

Рисунок 2.30 – Точність моделей на тренувальних даних для змінної Instagram

Модель XGBoost, Random Forest, Logistic Regression правильно класифікує близько 99% навчальних прикладів для змінної Instagram.

Наступна зміна для моделювання Social ads, результат навчання моделей показано на рисунку 2.31.

```

Точність XGBoost на тренувальних даних: 0.8055555555555556
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 0.8}

Точність Logistic Regression на тренувальних даних: 0.8055555555555556
Найкращі гіперпараметри Logistic Regression: {'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.8302469135802469
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}

```

Рисунок 2.31 – Точність моделей на тренувальних даних для змінної Social ads

Модель XGBoost правильно класифікує близько 80% навчальних прикладів, Random Forest 83% та Logistic Regression 80% для змінної Social ads.

Переглянемо результати навчання моделей для змінної Email (рис. 2.32).

```

Точність XGBoost на тренувальних даних: 0.8672839506172839
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 0.8}

Точність Logistic Regression на тренувальних даних: 0.8672839506172839
Найкращі гіперпараметри Logistic Regression: {'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.8703703703703703
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

```

Рисунок 2.32 – Точність моделей на тренувальних даних Email

Модель XGBoost правильно класифікує близько 87% навчальних прикладів, Random Forest 87% та Logistic Regression 87% для змінної Email.

Наступна зміна для моделювання Internet, результати навчання моделей показано на рисунку 2.33.

```

Точність XGBoost на тренувальних даних: 0.8240740740740741
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 7, 'min_child_weight': 1, 'n_estimators': 200, 'subsample': 1.0}

Точність Logistic Regression на тренувальних даних: 0.654320987654321
Найкращі гіперпараметри Logistic Regression: {'C': 1, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9506172839506173
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}

```

Рисунок 2.33 – Точність моделей на тренувальних даних для змінної Internet

Модель XGBoost правильно класифікує близько 82% навчальних прикладів, Random Forest Classifier 95% та Logistic Regression 65% для змінної Internet.

Далі зміна для моделювання Mass media. Результати навчання моделей показано на рисунку 2.34.

```

Точність XGBoost на тренувальних даних: 0.7314814814814815
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 7, 'min_child_weight': 3, 'n_estimators': 100, 'subsample': 0.8}

Точність Logistic Regression на тренувальних даних: 0.6574074074074074
Найкращі гіперпараметри Logistic Regression: {'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.9382716049382716
Найкращі гіперпараметри Random Forest: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}

```

Рисунок 2.34 – Точність моделей на тренувальних даних для змінної Mass media

Модель XGBoost правильно класифікує близько 73% навчальних прикладів, Random Forest 94% та Logistic Regression 66% для змінної Mass media.

Наступна зміна для моделювання Official site, результати навчання моделей показано на рисунку 2.35.

```

Точність XGBoost на тренувальних даних: 0.7407407407407407
Найкращі гіперпараметри XGBoost: {'colsample_bytree': 0.8, 'gamma': 0.1, 'learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 3, 'n_estimators': 200, 'subsample': 1.0}

Точність Logistic Regression на тренувальних даних: 0.6296296296296297
Найкращі гіперпараметри Logistic Regression: {'C': 0.01, 'max_iter': 100, 'penalty': 'l2', 'solver': 'lbfgs'}

Точність Random Forest на тренувальних даних: 0.7191358024691358
Найкращі гіперпараметри Random Forest: {'bootstrap': True, 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}

```

Рисунок 2.35 – Точність моделей на тренувальних даних для змінної Official site

Модель XGBoost правильно класифікує близько 74% навчальних прикладів, Random Forest 71% та Logistic Regression 63% для змінної Official site.

2.4 Висновки

У другому розділі виконано повний цикл підготовки даних, зібраних через соціологічне опитування мешканців Вінницької територіальної громади, для моделювання методами машинного навчання. Дані очищено від пропущених значень, категоріальні ознаки перетворено на числові за допомогою one-hot кодування, виявлено та усунено аномалії, проведено масштабування ознак (StandardScaler) та зменшено надлишковості шляхом виключення високо корельованих змінних. Також проведено розвідувальний аналіз даних (EDA), у межах якого побудовані діаграми розподілу, боксплоти та кореляційна матриця. Встановлено, що на цифрову активність найбільше впливають стать, вік, рівень освіти, тривалість проживання у громаді та соціальний статус. Найвищу цифрову залученість продемонстрували молодь, люди з інвалідністю та респонденти, які нещодавно оселилися у громаді.

Під час первинного моделювання оцінено точність класифікаційних моделей на тренувальних даних для різних каналів цифрової взаємодії. За результатами цього етапу деякі цільові змінні не продемонстрували задовільної якості передбачення в межах трьох застосованих моделей – Logistic Regression, Random Forest і XGBoost. Зокрема, змінні Mass media, Social ads, Internet, Email та Official site мали низьку точність (65–87%), особливо для моделей логістичної регресії, де в окремих випадках точність не перевищувала 70%, що вказує на слабку передбачуваність цих каналів.

Для уникнення хибних інтерпретацій і підвищення якості прогнозування ці змінні виключено з подальшого аналізу. Дослідження зосереджено на п'яти цільових змінних (Facebook, Instagram, Telegram, Viber, WhatsApp), які демонструють стабільно високу точність (понад 95%) та підтверджують надійність класифікації за допомогою обраних моделей.

На основі аналізу визначено оптимальні цільові змінні та сформовано набір ознак, які відображають соціально-демографічні особливості респондентів і будуть використані для тренування моделей передбачення цифрової взаємодії.

3 РЕЗУЛЬТАТИ ПЕРЕДБАЧЕННЯ СОЦІАЛЬНОЇ ВЗАЄМОДІЇ ВІННИЦЬКОЇ МІСЬКОЇ ТЕРИТОРІАЛЬНОЇ ГРОМАДИ ЧЕРЕЗ ЦИФРОВІ КАНАЛИ

3.1 Передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали

Після підготовки даних та навчання моделей проведемо передбачення соціальної взаємодії мешканців Вінницької міської територіальної громади через цифрові канали Viber, Instagram, Telegram, Facebook та WhatsApp.

Для моделювання застосуємо алгоритми машинного навчання: Logistic Regression, Random Forest та XGBoost. Першим об'єктом для поглибленого аналізу оберемо додаток Telegram. Побудуємо графік ознак, що впливають на вибір цифрового каналу Telegram (рис. 3.1).

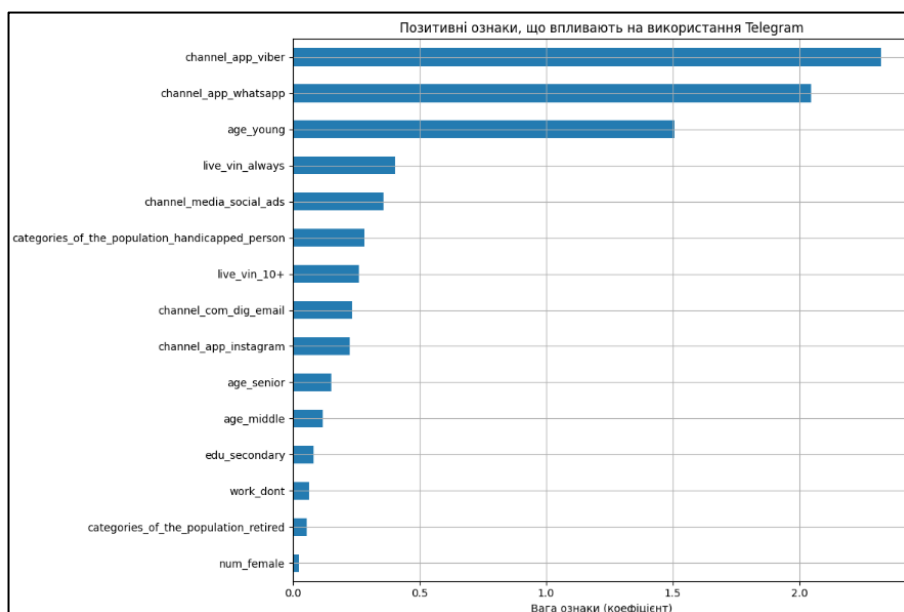


Рисунок 3.1 – Графік важливих ознак для цифрового каналу Telegram

На рисунку Г.9 показано код, що реалізує повний цикл побудови моделі машинного навчання для задачі передбачення – від підготовки даних, масштабування ознак і поділу вибірки до налаштування гіперпараметрів за

допомогою GridSearchCV та оцінювання ефективності моделей на тестових даних. На рисунку 3.2. представлено метрики трьох моделей застосованих для передбачення використання цифрового каналу Telegram на тестових даних.

```
Метрики Logistic Regression на тестових даних:
Accuracy: 0.9259259259259259
Precision: 0.9655172413793104
Recall: 0.8484848484848485
F1 Score: 0.9032258064516129
ROC AUC Score: 0.9905303030303031

Метрики Random Forest на тестових даних:
Accuracy: 0.9382716049382716
Precision: 0.9375
Recall: 0.9090909090909091
F1 Score: 0.9230769230769231
ROC AUC Score: 0.9797979797979798

Метрики XGBoost на тестових даних:
Accuracy: 0.9382716049382716
Precision: 0.9375
Recall: 0.9090909090909091
F1 Score: 0.9230769230769231
ROC AUC Score: 0.9911616161616161
```

Рисунок 3.2 – Метрики ефективності моделей для цифрового каналу Telegram

На основі наведених на рисунку 3.2 метрик, можна зробити висновок, що всі три моделі демонструють високу якість класифікації на тестових даних.

Наступним кроком для більш кращого розуміння моделі є побудова матриці помилок [28]. Матриця плутанини зазвичай використовується для оцінки ефективності методів, використаних після класифікації. Має такі параметри:

- True Positives (TP) — це кількість випадків, які модель правильно класифікувала як позитивні;
- True Negatives (TN) — вказує на кількість екземплярів, які модель правильно класифікувала як негативні;
- False Positives (FP) — це випадки, які модель неправильно класифікувала як позитивні, хоча насправді вони належать до негативного класу;
- False Negatives (FN) — випадки, які були неправильно класифіковані як негативні моделі, хоча насправді належать до позитивного класу [28].

Для візуалізації результатів передбачення соціальної взаємодії мешканців Вінницької міської територіальної громади через цифровий канал Telegram застосуємо матрицю плутанини. Переглянемо на рисунку Г.10, код матриці для моделі XGBoost. Для кращого розуміння результатів передбачення розглянемо матрицю помилок з рисунка 3.3.

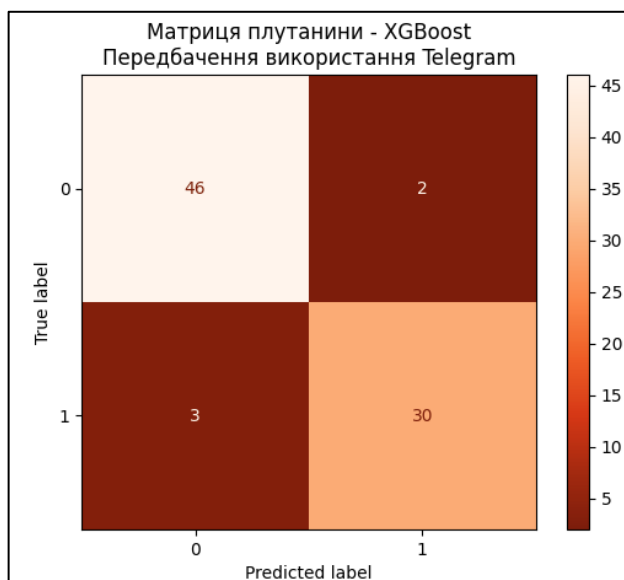


Рисунок 3.3 – Матриця плутанини результатів моделі XGBoost для цифрового каналу Telegram

Матриця плутанини дозволяє оцінити якість класифікації та ефективність моделі XGBoost:

- True Negative (TN) = 46 – кількість випадків, коли модель правильно ідентифікувала респондентів, що не використовують додаток Telegram;
- False Positive (FP) = 2 – кількість випадків, коли модель помилково класифікувала тих, хто не користується Telegram, як таких, що користуються;
- False Negative (FN) = 3 – кількість випадків, коли модель не змогла виявити справжніх користувачів Telegram;
- True Positive (TP) = 30 – кількість коректно передбачених випадків використання Telegram.

Модель XGBoost демонструє високу точність класифікації з незначною кількістю помилок, що підтверджує її ефективність у задачі передбачення.

На рисунку 3.4 розглянемо результат матриці моделі Logistic Regression.

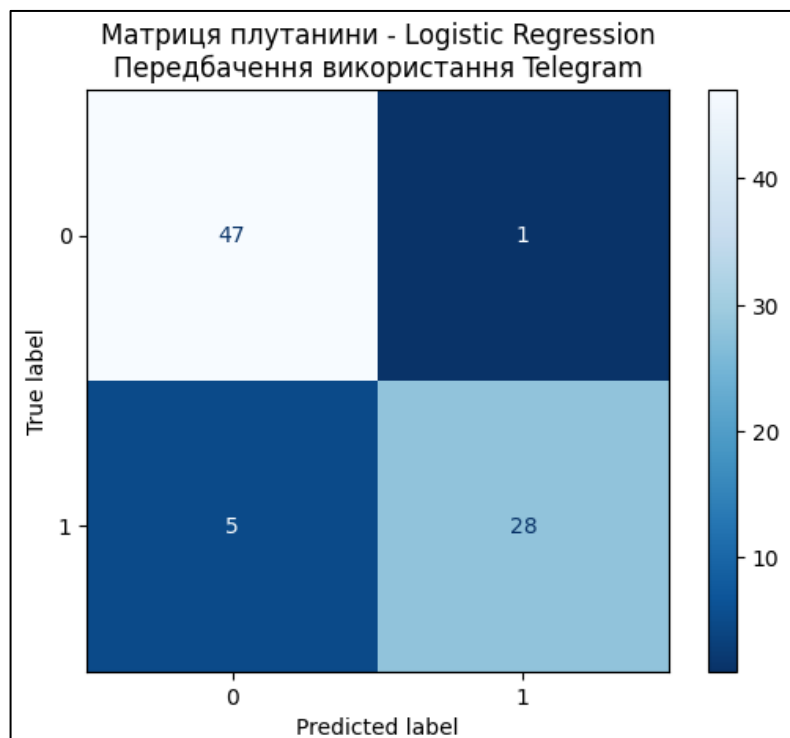


Рисунок 3.4 – Матриця плутанини результатів моделі Logistic Regression для цифрового каналу Telegram

На рисунку 3.4 подано результати матриці плутанини для моделі Logistic Regression, що дає змогу оцінити якість класифікації:

- True Negative (TN) = 47 – кількість випадків, коли модель правильно передбачила, що респондент не користується Telegram;
- False Positive (FP) = 1 – випадок, коли модель помилково передбачила використання Telegram там, де його не було;
- False Negative (FN) = 5 – випадки, коли справжні користувачі Telegram були пропущені;
- True Positive (TP) = 28 – кількість випадків, коли модель правильно розпізнала користувачів Telegram.

Модель Logistic Regression демонструє достатньо високу точність класифікації, коректно ідентифікуючи як користувачів так і протилежних.

Для подальшого аналізу результатів передбачення звернемося до матриці плутанини моделі Random Forest, поданої на рисунку 3.5.

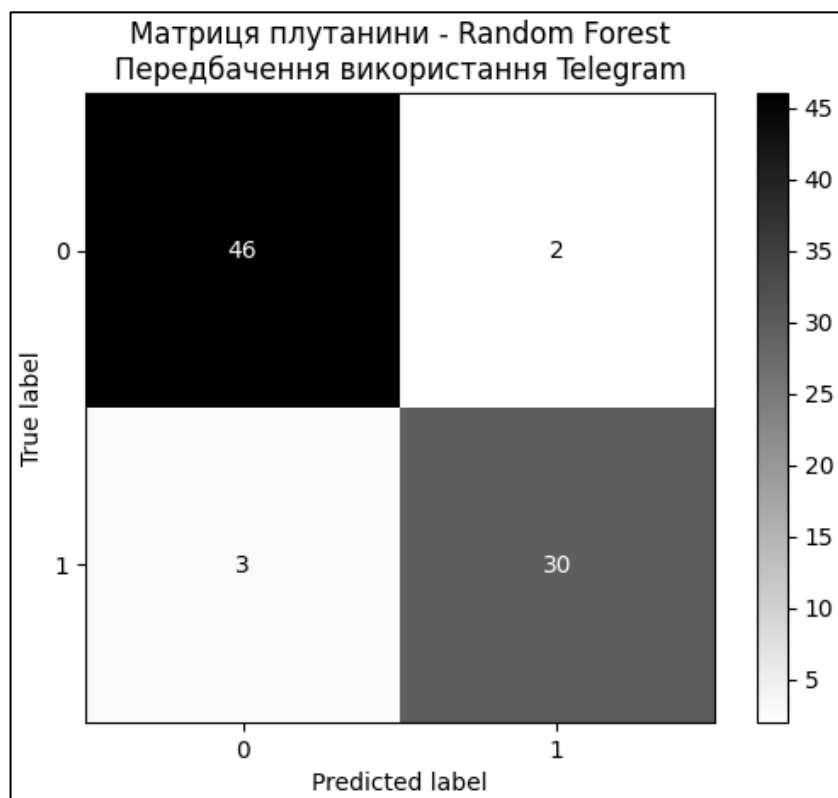


Рисунок 3.5 – Матриця плутанини результатів моделі Random Forest цифрового каналу Telegram

На рисунку 3.5 подано результати матриці плутанини для моделі Random Forest, за допомоги неї можливо оцінити якість класифікації:

- True Negative (TN) = 46 – кількість випадків, коли модель правильно класифікувала респондентів, що не використовують Telegram;
- False Positive (FP) = 2 – ситуації, коли модель неправильно розпізнала респондентів, які не використовують Telegram;
- False Negative (FN) = 3 – модель не розпізнала трьох справжніх користувачів;

– True Positive (TP) = 30 – модель правильно виявила 30 респондентів, що є користувачами месенджера.

Модель Random Forest показала добре збалансовану класифікацію з мінімальною кількістю помилок. Застосуємо метрику оцінки accuracy_score для кількісної оцінки якості класифікації. Результат представлено на рисунку 3.6.

```
LogisticRegression (channel_app_telegram):
Крос-валідація: середнє = 0.944, std = 0.008
Точність train: 0.969
Точність test: 0.926
Модель узагальнює добре

RandomForest (channel_app_telegram):
Крос-валідація: середнє = 0.938, std = 0.017
Точність train: 1.000
Точність test: 0.938
Модель узагальнює добре

XGBoost (channel_app_telegram):
Крос-валідація: середнє = 0.926, std = 0.018
Точність train: 0.994
Точність test: 0.951
Модель узагальнює добре
```

Рисунок 3.6 – Результати передбачення усіма моделями для цифрового каналу Telegram

Під час оцінювання моделей було проаналізовано їхню точність на тренувальних і тестових даних. Найвищу точність на тренувальному наборі продемонструвала модель Random Forest (1.000), модель XGBoost показала високий результат (0.994), LogisticRegression виявила нижчий показник порівнянні з іншими моделями (0.969).

На тестових даних найкращу точність продемонструвала модель XGBoost (0.951). Random Forest також забезпечила стабільну продуктивність (0.938). Logistic Regression продемонструвала нижчий рівень точності (0.926), але зберегла стабільність класифікації.

Далі досліджуємо цифровий канал WhatsApp. На рисунку 3.7 зображено графік важливих ознак, що впливають на вибір цього каналу.

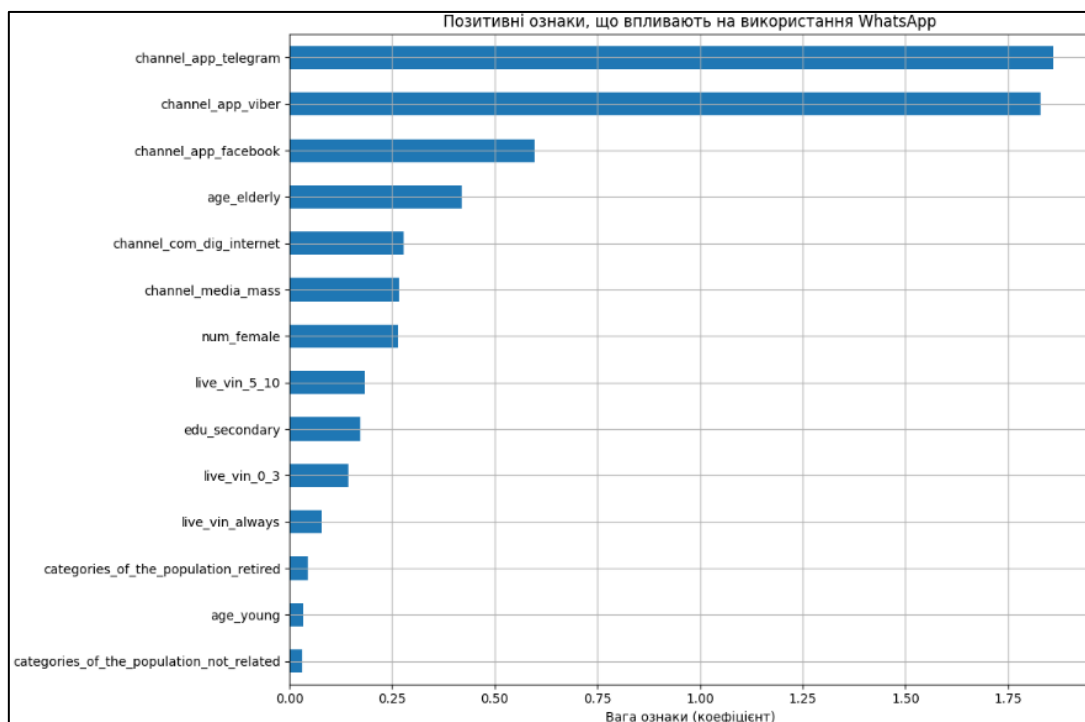


Рисунок 3.7 – Графік важливих ознак для цифрового каналу WhatsApp

На рисунку 3.8 подано дані для порівняння метрик трьох моделей машинного навчання, задіяних для вирішення задачі бінарної класифікації. Представлено значення таких показників, як точність, повнота, точність передбачення, F1-score та інші, що дають можливість комплексно оцінити ефективність класифікації кожної моделі. Також описано найкращі гіперпараметри для моделі XGBoost, і зазначено точність на тренувальних даних.

```

Logistic Regression Метрики:
Accuracy: 0.9753086419753086
Precision: 0.9714285714285714
Recall: 0.9714285714285714
F1 Score: 0.9714285714285714
ROC AUC Score: 0.993167701863354

Random Forest Метрики:
Accuracy: 0.9629629629629629
Precision: 0.9705882352941176
Recall: 0.9428571428571428
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9975155279503106

XGBoost Метрики:
Accuracy: 0.9753086419753086
Precision: 0.9714285714285714
Recall: 0.9714285714285714
F1 Score: 0.9714285714285714
ROC AUC Score: 0.9925465838509318

XGBoost найкращі параметри:
{'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 200, 'subsample': 0.8}
Точність XGBoost на тренуванні: 0.9783950617283951

```

Рисунок 3.8 – Метрики ефективності моделей для цифрового каналу WhatsApp

Усі три моделі демонструють високу якість класифікації. Для візуалізації результатів передбачення використаємо матрицю плутанини. На рисунку Г.11 зображено код побудови матриці плутанини для моделі XGBoost, а на рисунку Г.12 візуалізовано її. За допомоги матриці плутанини маємо змогу оцінити якість класифікації та продуктивності моделі XGBoost:

- True Negative (TN) = 45 – кількість випадків, коли модель правильно ідентифікувала респондентів, що не користуються додатком WhatsApp;
- False Positive (FP) = 1 – модель помилково класифікувала одного респондента як користувача WhatsApp, хоча він ним не є;
- False Negative (FN) = 1 – випадок, коли модель не змогла виявити істинного користувача додатку;
- True Positive (TP) = 34 – кількість коректно передбачених випадків використання месенджера.

Ці результати вказують на високу ефективність моделі в ідентифікації як користувачів, так і не користувачів WhatsApp.

На рисунку Г.13 зображено код побудови матриці плутанини для моделі Logistic Regression для цифрового каналу WhatsApp. На рисунку Г.14 зображено матрицю плутанини, результати якої наступні:

- True Negative (TN) = 45 модель вірно визначила, що 45 респондентів не є користувачами;
- False Positive (FP) = один респондент був помилково класифікований як користувач додатку, хоча ним не є;
- False Negative (FN) = один справжній користувач WhatsApp не був вірно ідентифікований моделлю;
- True Positive (TP) = 34 респонденти були правильно визначені як користувачі WhatsApp.

Ці результати вказують на високу збалансовану ефективність моделі Logistic Regression у задачі бінарної класифікації.

На рисунку Г.15 зображено код побудови матриці плутанини для моделі Random Forest, а на рисунку Г.16 візуалізовано її. Результати класифікації:

- True Negative (TN) = 45 кількість респондентів, яких модель вірно класифікувала як тих, хто не користується месенджером;
- False Positive (FP) = 1 респондент був помилково віднесений до користувачів WhatsApp, хоча ним не є;
- False Negative (FN) = один справжній користувач WhatsApp не був розпізнаний моделлю;
- True Positive (TP) = 34 кількість респондентів, яких модель правильно класифікувала як користувачів WhatsApp.

Загалом модель Random Forest демонструє високу точність та здатність ефективно розпізнавати обидва класи користувачів.

На рисунку 3.9 представлено результати кількісного порівняння результатів за допомоги метрики `accuracy_score`.

```
LogisticRegression (channel_app_whatsapp):
Крос-валідація: середнє = 0.951, std = 0.011
Точність train: 0.960
Точність test: 0.975
Модель узагальнює добре

RandomForest (channel_app_whatsapp):
Крос-валідація: середнє = 0.948, std = 0.012
Точність train: 0.994
Точність test: 0.975
Модель узагальнює добре

XGBoost (channel_app_whatsapp):
Крос-валідація: середнє = 0.954, std = 0.019
Точність train: 0.991
Точність test: 0.963
Модель узагальнює добре
```

Рисунок 3.9 – Результати передбачення моделей для цифрового каналу
WhatsApp

Усі моделі продемонстрували високу якість класифікації, з точністю на тестових даних вище 0.96, що вказує на їхню добру здатність до узагальнення без явних ознак перенавчання.

Найвищу точність на тренувальному наборі показала модель Random Forest (0.994), показуючи сильну здатність до навчання, XGBoost із точністю

(0.991) забезпечує збалансовану продуктивність, а Logistic Regression має трохи нижчий показник (0.960), але вирізняється стабільністю на тестовому наборі.

На тестових даних Logistic Regression і Random Forest досягли найкращого результату (0.975), демонструючи сильну узагальнювальну здатність, тоді як XGBoost трохи поступалася (0.963), але зберігала стабільність та ефективність.

Усі моделі підтвердили свою придатність для передбачення використання WhatsApp, забезпечуючи високу точність і стабільність результатів.

Наступний досліджуємо цифровий канал Viber. Виведемо графік важливих ознак, що впливають на вибір каналу Viber (рис. 3.10).

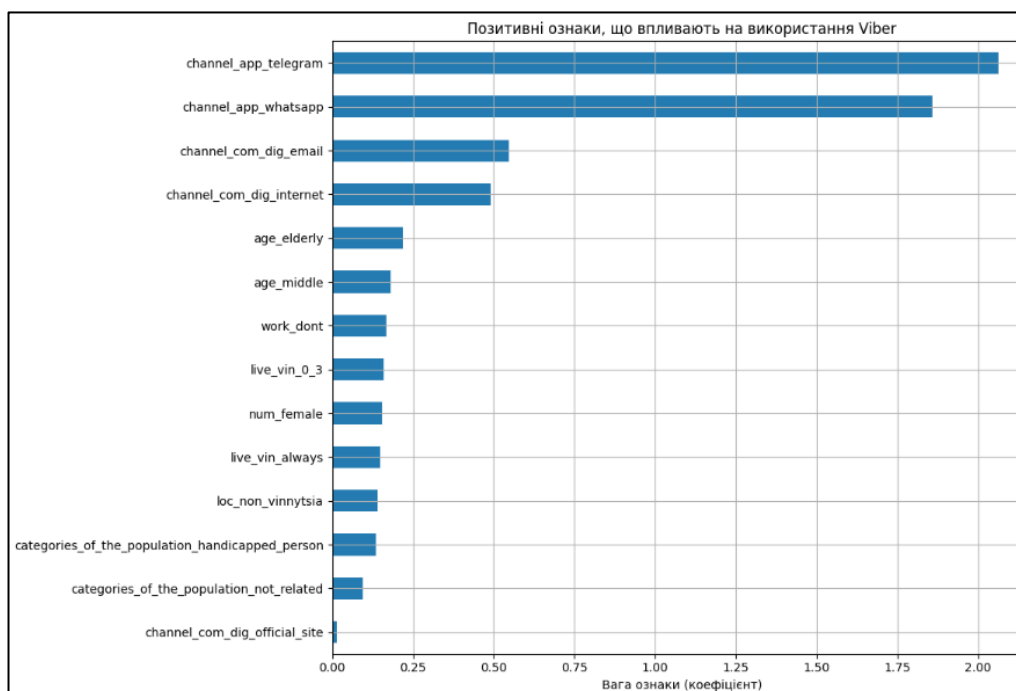


Рисунок 3.10 – Графік важливих ознак для цифрового каналу Viber

На рисунку 3.11 подано дані для порівняння метрик моделей машинного навчання та найкращі гіперпараметри для XGBoost, зазначимо точність моделі на тренувальних даних.

```

Logistic Regression Метрики:
Accuracy: 0.9629629629629629
Precision: 0.9428571428571428
Recall: 0.9705882352941176
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9799749687108887

Random Forest Метрики:
Accuracy: 0.9629629629629629
Precision: 0.9428571428571428
Recall: 0.9705882352941176
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9618272841051314

XGBoost Метрики:
Accuracy: 0.9629629629629629
Precision: 0.9696969696969697
Recall: 0.9411764705882353
F1 Score: 0.9552238805970149
ROC AUC Score: 0.9721526908635796

XGBoost найкращі параметри:
{'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 3, 'n_estimators': 200, 'subsample': 1.0}
Точність XGBoost на тренуванні: 0.9691358024691358
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [09:17:46] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

```

Рисунок 3.11 – Результати передбачення моделей для цифрового каналу Viber

Три моделі демонструють високу якість класифікації при розв’язанні задачі передбачення використання цифрового каналу Viber. Модель Logistic Regression забезпечує точність (0.962), precision (0.942), recall (0.970) та F1–score (0.956), це демонструє її стабільну роботу як моделі та здатність правильно класифікувати як позитивні, так і негативні випадки. Значення ROC AUC Score (0.979) додатково підтверджує ефективність.

Random Forest показує аналогічну точність (0.962), з precision (0.943) та recall (0.970), ідентичними recall (0,970), що забезпечує таку ж F1– score (0.956). ROC AUC Score становить (0.961), що підтверджує її надійність.

Модель XGBoost також досягає точності (0.962), демонструє precision (0.969), recall (0.941) та F1–score (0.955), що вказує про збалансованість моделі. Її ROC AUC Score (0.972) підтверджує здатність моделі до ефективного розрізнення класів. Всі моделі показують хорошу здатність до класифікації.

Для глибшого аналізу класифікаційних можливостей наведемо матрицю плутанини моделі XGBoost для цифрового каналу Viber (рис. Г.17). Маємо такі результати:

– True Negative (TN) = 46 модель правильно класифікувала 46 респондентів, що не є користувачами Viber;

- False Positive (FP) = 1 – виявлено лише один випадок помилкової позитивної класифікації, коли респондента помилково віднесло до користувачів Viber;

- False Negative (FN) = 2 – два випадки, коли модель не розпізнала фактичних користувачів месенджера;

- True Positive (TP) = 32 кількість респондентів, яких модель вірно ідентифікували як користувачів Viber.

Загалом модель XGBoost показує високу точність і здатність чітко розмежовувати класи.

Далі для аналізу класифікаційних можливостей наведемо матрицю плутанини моделі Logistic Regression для цифрового каналу Viber (рис. Г.18). Результати наступні:

- True Negative (TN) = 45 – модель вірно визначила, що 45 респондентів не користуються Viber;

- False Positive (FP) = 2 – два випадки помилкової класифікації респондентів до користувачів Viber;

- False Positive (FN) = 1 один справжній користувач не був розпізнаний;

- False Positive (TP) = 33 респонденти були правильно ідентифіковані як користувачі додатку.

Для глибшого аналізу класифікаційних можливостей наведемо матрицю плутанини моделі Random Forest для цифрового каналу Viber (рис. Г.19). Маємо такі результати:

- True Negative (TN) = 45 модель коректно класифікувала респондентів, які не користуються Viber;

- False Positive (FP) = 2 два випадки помилкової ідентифікації респондентів як користувачів месенджера;

- False Negative (FN) = 1 – один справжній користувача Viber не був розпізнаний;

– True Positive (TP) = 33 респонденти, яких модель вірно класифікувала як користувачів Viber.

Модель Random Forest показує високий рівень точності класифікації, зокрема у виявленні справжніх користувачів Viber. Низька кількість помилкових передбачень підтверджує її надійність та ефективність у задачах бінарної класифікації. Для оцінки ефективності моделей у передбаченні застосуємо метрику `assurasy_score`, результати наведено на рисунку 3.12.

LogisticRegression	
Крос-валідація:	середнє = 0.954, std = 0.014
Точність train:	0.963
Точність test:	0.963
Модель узагальнює добре	
RandomForest	
Крос-валідація:	середнє = 0.947, std = 0.019
Точність train:	0.994
Точність test:	0.963
Модель узагальнює добре	
XGBoost	
Крос-валідація:	середнє = 0.938, std = 0.026
Точність train:	0.991
Точність test:	0.963
Модель узагальнює добре	

Рисунок 3.12 – Оцінки перехресної перевірки для цифрового каналу Viber

На рисунку 3.12 наведено результати крос-валідації та точності трьох моделей. Вони демонструють однакову точність на тестовій вибірці (0.963), що свідчить про їхню стабільність та добру здатність до узагальнення.

Найвищу точність на тренувальних даних забезпечила модель Random Forest (0.994), тоді як XGBoost (0.991) і Logistic Regression (0.963) вказує на її збалансованість між тренувальними та тестовими результатами. Усі моделі є придатними для розв'язання задачі бінарної класифікації без ознак перенавчання.

Досліджуємо наступний цифровий канал Facebook. Виведемо графік важливих позитивних ознак, що впливають на вибір цифрового каналу Facebook (рис. 3.13).

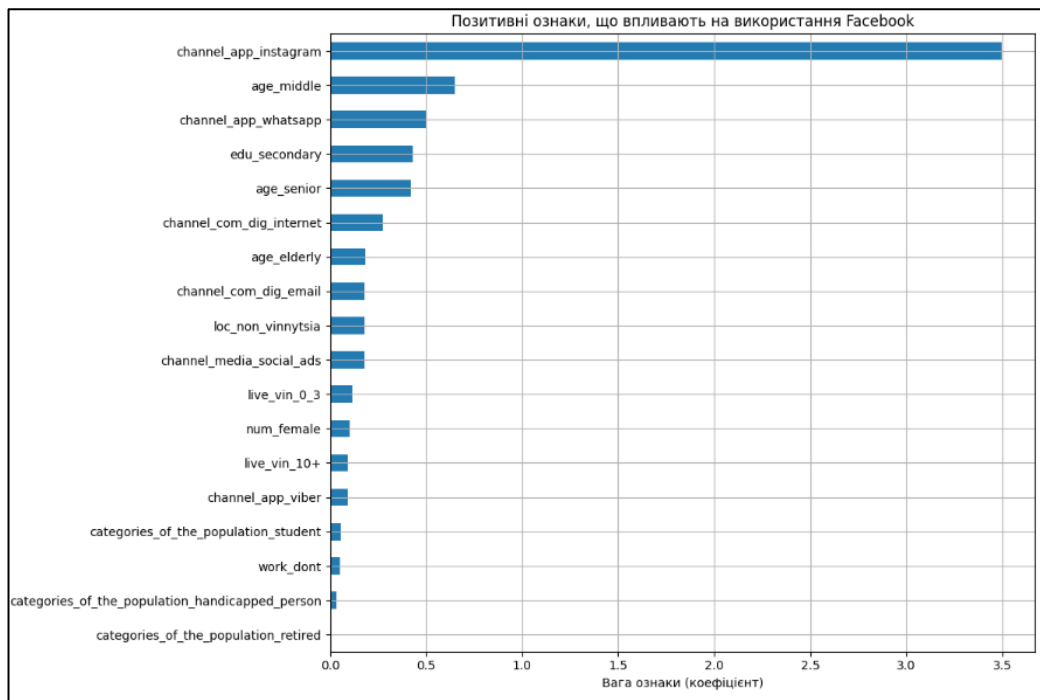


Рисунок 3.13 – Графік важливості ознак для цифрового каналу Facebook

На рисунку 3.14 подано дані для порівняння метрик моделей машинного навчання та найкращі гіперпараметри для XGBoost, зазначимо точність моделі на тренувальних даних.

```

Logistic Regression Метрики:
Accuracy: 0.9135802469135802
Precision: 0.9298245614035088
Recall: 0.9464285714285714
F1 Score: 0.9380530973451328
ROC AUC Score: 0.8971428571428571

Random Forest Метрики:
Accuracy: 0.9382716049382716
Precision: 0.9322033898305084
Recall: 0.9821428571428571
F1 Score: 0.9565217391304348
ROC AUC Score: 0.932142857142857

XGBoost Метрики:
Accuracy: 0.9135802469135802
Precision: 0.9298245614035088
Recall: 0.9464285714285714
F1 Score: 0.9380530973451328
ROC AUC Score: 0.8935714285714285

XGBoost найкращі параметри:
{'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 1, 'n_estimators': 200, 'subsample': 0.8}
Точність XGBoost на тренуванні: 0.9876543209876543
  
```

Рисунок 3.14 – Метрики ефективності моделей для цифрового каналу Facebook

Усі три моделі демонструють високу якість класифікації Logistic Regression досягає точності (0.914), precision (0.929), recall (0.946) та F1-score

(0.938), що вказує на стабільну класифікацію, а ROC AUC (0.897) підтверджує хорошу дискримінацію класів.

Random Forest показує найвищу точність (0.938), precision (0.933), recall (0.982) та F1–score (0.956), з ROC AUC (0.932), що робить її ефективною моделю.

XGBoost має точність (0.914), precision (0.929), recall (0.946), F1 (0.938) і ROC AUC (0.893), що схоже на Logistic Regression, з тренувальною точністю (0.987) при оптимальних гіперпараметрах.

Усі моделі забезпечують надійну класифікацію, але Random Forest виділяється найвищою ефективністю.

Для кращого розуміння результатів передбачення розглянемо матрицю плутанини на рисунку Г.20 моделі XGBoost для цифрового каналу Facebook. Отримані результати:

- True Negative (TN) = 21 модель коректно класифікувала респондентів, які не є користувачами Facebook;
- False Positive (FP) = 4 чотири випадки помилкової класифікації респондентів як користувачів Facebook;
- False Negative (FN) = 3 три випадки, коли модель не ідентифікувала справжніх користувачів Facebook;
- True Positive (TP) = 53 модель вірно класифікувала 53 респондентів як користувачів Facebook.

Загалом модель XGBoost демонструє добрі результати класифікації з переважно точним розмежуванням між класами.

Для подальшої оцінки розглянемо матрицю плутанини моделі Logistic Regression для цифрового каналу Facebook, зображену на рисунку Г.21. Отримані результати такі:

- True Negative (TN) = 21 – модель вірно визначила респондентів, які не користуються Facebook;
- False Positive (FP) = 4 чотири випадки помилкової класифікації респондентів як користувачів Facebook;

- False Negative (FN) = 3 – три фактичних користувачі Facebook не були розпізнані моделлю;
- True Positive (TP) = 53 модель точно класифікація 53 респондентів як користувачів Facebook.

Модель Logistic Regression демонструє збалансовану продуктивність і надійність при виконанні задачі бінарної класифікації.

Наступним кроком розглянемо результати моделі Random Forest для Facebook подані на рисунку Г.22. Маємо такі результати:

- True Negative (TN) = 21 модель коректно класифікувати респондентів, які не користуються Facebook;
- False Positive (FP) = 4 чотири випадки помилкової класифікації респондентів як користувачів Facebook;
- False Negative (FN) = 1 один справжній користувач Facebook не був розпізнаний;
- True Positive (TP) = 55 – кількість респондентів, яких модель правильно класифікувала як користувачів платформи.

Модель Random Forest демонструє високу точність класифікації, особливо у виявленні позитивного класу. Мінімальна кількість помилок підтверджує її надійність та ефективність у задачах бінарного передбачення. Для оцінки ефективності моделей у передбаченні застосуємо метрику `accuracy_score`, результати наведено на рисунку 3.15.

```

LogisticRegression
Крос-валідація: середнє = 0.969, std = 0.014
Точність (train): 0.985
Точність (test): 0.914
Модель узагальнює добре

RandomForest
Крос-валідація: середнє = 0.963, std = 0.019
Точність (train): 0.994
Точність (test): 0.926
Модель узагальнює добре

XGBoost
Крос-валідація: середнє = 0.969, std = 0.017
Точність (train): 0.991
Точність (test): 0.914
Модель узагальнює добре

```

Рисунок 3.15 – Оцінки перехресної перевірки моделей для цифрового каналу Facebook

На рисунку 3.15 подано результати крос-валідації та точності трьох моделей у задачі передбачення використання Facebook. Моделі показали високі результати точності на тестових даних: Logistic Regression, XGBoost (0.914) та Random Forest (0.926), це вказує на їхню стабільність та здатність до узагальнення. Найвищу точність на тренувальних даних досягла модель Random Forest (0.994), тоді як XGBoost (0.991) і Logistic Regression (0.985) забезпечили хороший баланс між навчанням і тестуванням, без ознак перенавчання.

Досліджуємо наступний цифровий канал Instagram. Виведемо графік важливих ознак, що впливають на вибір цифрового каналу Instagram (рис. 3.16).

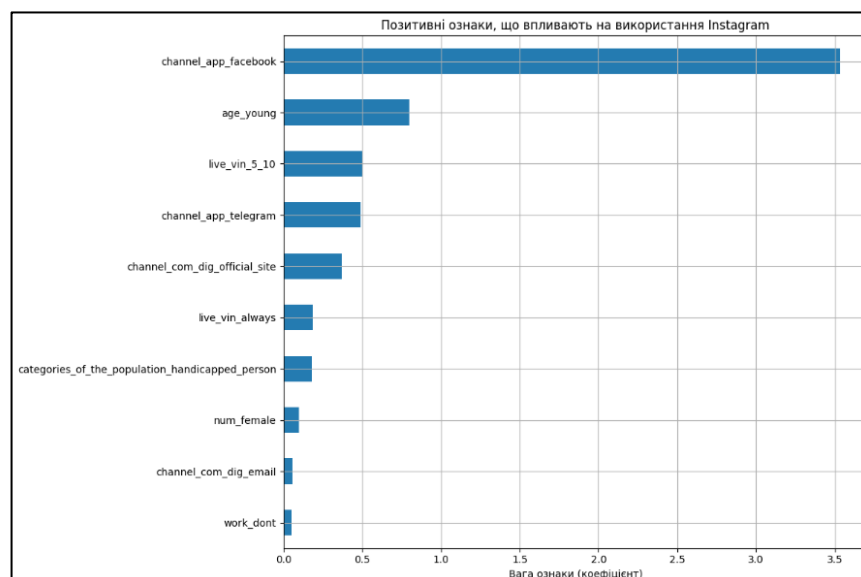


Рисунок 3.16 – Графік важливих ознак для цифрового каналу Instagram

На рисунку 3.17 подано порівняльні метрики моделей машинного навчання та зазначено найкращі гіперпараметри моделі XGBoost. Відображено також точність XGBoost на тренувальних даних.

```

Logistic Regression Метрики:
Accuracy: 0.9382716049382716
Precision: 0.9821428571428571
Recall: 0.9322033898305084
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9414483821263483

Random Forest Метрики:
Accuracy: 0.9382716049382716
Precision: 0.9821428571428571
Recall: 0.9322033898305084
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9291217257318952

XGBoost Метрики:
Accuracy: 0.9382716049382716
Precision: 0.9821428571428571
Recall: 0.9322033898305084
F1 Score: 0.9565217391304348
ROC AUC Score: 0.9406779661016951

XGBoost найкращі параметри:
{'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 1.0}
Точність XGBoost на тренуванні: 0.9876543209876543

```

Рисунок 3.17 – Метрики ефективності моделей для цифрового каналу Instagram

Усі три моделі демонструють високу якість класифікації у задачі передбачення використання цифрового каналу. Logistic Regression досягає точності (0.938), precision (0.922), recall (0.932) та F1–score (0.956), що вказує на стабільну класифікацію обох класів, а ROC AUC (0.941) підтверджує високу дискримінацію. Random Forest показує аналогічні значення точності precision, recall та F1–score, проте з дещо нижчими ROC AUC (0.929). Отже, усі моделі забезпечують надійну та збалансовану класифікацію, придатну для бінарного передбачення.

Для подальшої оцінки розглянемо матрицю плутанини моделі XGBoost для цифрового каналу Instagram, зображену на рисунку Г.23. Отримані результати такі:

- True Negative (TN) = 21 – правильна класифікація респондентів, які не користуються Instagram;
- False Positive (FP) = 1 – один випадок хибної класифікації;
- False Negative (FN) = 4 чотири випадки, коли модель не розпізнала справжніх користувачів Instagram;
- True Positive (TP) = 55 кількість респондентів, яких модель вірно визначила як користувачів Instagram.

Модель XGBoost демонструє високу точність класифікацію, хоча й припускається окремих помилок при виявленні позитивного класу.

Далі розглянемо матрицю плутанини моделі Logistic Regression для цифрового каналу Instagram, зображену на рисунку Г.24. Маємо такі результати:

- True Negative (TN) = 21 – модель правильно класифікувала респондентів, які не користуються Instagram;
- False Positive (FP) = 1 – один випадок хибної класифікації;
- False Negative (FN) = 4 – чотири справжні користувачі не були розпізнані;
- True Positive (TP) = 55 – правильна класифікація користувачів Instagram.

Попри кілька помилок у позитивному класі, модель демонструє збалансовану й ефективну роботу в задачі бінарної класифікації.

Для подальшої оцінки розглянемо матрицю плутанини моделі Random Forest для цифрового каналу Instagram, зображену на рисунку Г.25. Отримані результати такі:

- True Negative (TN) = 21 – модель правильно класифікує респондентів, які не користуються Instagram;
- False Positive (FP) = 1 – один випадок хибного передбачення;
- False Negative (FN) = 4 – чотири справжні користувачі не були розпізнані;
- True Positive (TP) = 55 – правильна класифікація користувачів Instagram.

Модель Random Forest демонструє високу точність, особливо у виявленні позитивного класу. Низька кількість помилок підтверджує її ефективність у задачах бінарної класифікації.

Для оцінки ефективності моделей класифікації використаємо метрику `accuracy_score`. Результати подано на рисунку 3.18.

```

LogisticRegression (Instagram):
Крос-валідація:   середнє = 0.972, std = 0.018
Точність train:   0.981
Точність test:    0.938
Модель узагальнює добре

RandomForest (Instagram):
Крос-валідація:   середнє = 0.966, std = 0.020
Точність train:   0.997
Точність test:    0.938
Модель узагальнює добре

XGBoost (Instagram):
Крос-валідація:   середнє = 0.978, std = 0.016
Точність train:   0.994
Точність test:    0.938
Модель узагальнює добре

```

Рисунок 3.18 – Оцінки перехресної перевірки для цифрового каналу Instagram

На рисунку 3.18 подано результати крос-валідації та точності трьох моделей для передбачення використання Instagram. Усі три моделі показали однакову точність на тестових даних (0.938), що вказує про їхню стабільність і здатність до узагальнення. Найвищу точність на тренувальних даних досягла Random Forest (0.997), дещо нижчі результати – у XGBoost (0.994) та Logistic Regression (0.981). Усі моделі продемонстрували збалансовану ефективність без ознак перенавчання, що підтверджує їхню придатність до бінарної класифікації.

Після проведеного моделювання для узагальнення результатів та оцінювання всіх моделей побудуємо зведену таблицю результатів моделей представлену на рисунку 3.19.

Channel	Model	CrossVal Mean	Train Accuracy	Test Accuracy
WhatsApp	Logistic Regression	0.954	0.963	0.963
WhatsApp	Random Forest	0.947	0.994	0.963
WhatsApp	XGBoost	0.938	0.991	0.963
Instagram	Logistic Regression	0.972	0.981	0.938
Instagram	Random Forest	0.966	0.997	0.938
Instagram	XGBoost	0.978	0.994	0.938
Facebook	Logistic Regression	0.969	0.985	0.914
Facebook	Random Forest	0.963	0.994	0.926
Facebook	XGBoost	0.969	0.991	0.914
Telegram	Logistic Regression	0.944	0.969	0.926
Telegram	Random Forest	0.938	1.0	0.938
Telegram	XGBoost	0.926	0.994	0.951
Viber	Logistic Regression	0.954	0.963	0.963
Viber	Random Forest	0.947	0.994	0.963
Viber	XGBoost	0.938	0.991	0.963

Рисунок 3.19 – Зведена таблиця результатів моделей для 5 цифрових каналів

У дослідженні застосовано три моделі класифікації: логістичну регресію, випадковий ліс та XGBoost. Таблиця містить основні метрики: середнє значення точності за крос-валідацією (CrossVal Mean), точність на тренувальній (Train Accuracy) і тестовій (Test Accuracy).

На основі аналізу рисунку 3.19 прослідковуємо, що для цифрового каналу WhastApp найкращою моделлю виявилася Random Forest з оцінкою метрики accuracy (0.963) на тестових даних.

Для цифрового додатку Instagram найкращою моделлю виявилася Instagram з оцінкою метрики accuracy (0.938) на тестових даних.

Для цифрового додатку Facebook найкращою моделлю виявилася Random Forest з оцінкою метрики accuracy (0.926) на тестових даних.

Для цифрового додатку Telegram найкращою моделлю виявилася XGBoost Forest з оцінкою метрики accuracy (0.951) на тестових даних.

Для цифрового додатку Viber найкращою моделлю виявилася Random Forest з оцінкою метрики accuracy (0.93) на тестових даних.

Для кожного цифрового каналу на основі оптимальної моделі машинного навчання створимо графік важливості ознак, що відображатиме, які вхідні характеристики найбільше впливають на рішення моделі.

Також побудуємо матрицю плутанини, що надасть можливість оцінити якість класифікації через кількість правильних і помилкових передбачень для кожного класу.

На рисунку 3.20 подано графік позитивних ознак, що найбільше впливають на вибір цифрового каналу Instagram.

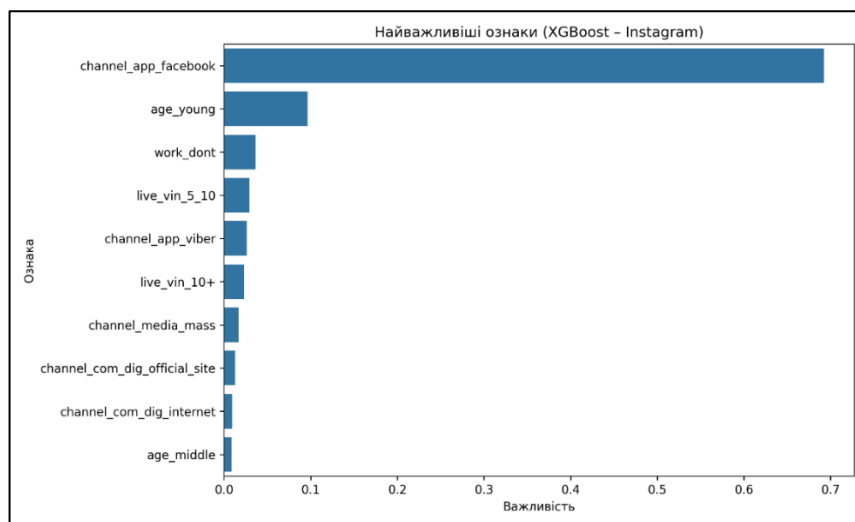


Рисунок 3.20 – Графік важливих ознак цифрового каналу Instagram для моделі XGBoost

З рисунка 3.20 видно, що найсуттєвіший вплив має ознака `channel_app_facebook`, що вказує на високу кореляцію між використанням Instagram і Facebook. Також вагомими є:

- `age_young` молоді респонденти значно частіше використовують Instagram, `age_middle` менше активно користуються додатком;
- `work_dont` активними користувачами є безробітні респонденти;
- `live_vin_5_10` та `live_vin_10+` респонденти, що живуть у громаді від 5 до 10 років та 10+ років є активними користувачами соціальної мережі;
- `channel_media_mass`, `channel_com_dig_official_site`, `channel_com_dig_internet`, `channel_app_viber` користувачі інших цифрових каналів також активно користуються Instagram;

Аналіз важливих ознак моделі показав, що на ймовірність використання цифрового каналу Instagram суттєво впливають такі групи чинників:

- вікові характеристики;
- соціальний статус;
- тривалість проживання в громаді;
- використання інших цифрових каналів.

На рисунку 3.21 подано матрицю плутанини для моделі XGBoost, навченої для класифікації використання Instagram:

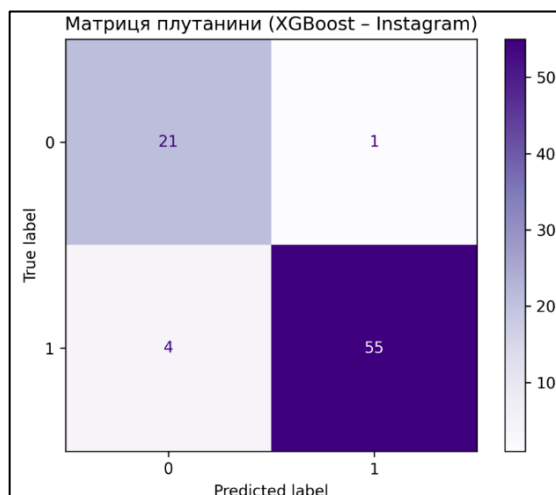


Рисунок 3.21 – Матриця плутанини моделі XGBoost для цифрового каналу Instagram

Результати матриці плутанини моделі XGBoost для цифрового каналу Instagram наступні:

- True Negative (TN) = 21 модель коректно ідентифікує респондентів, які не користуються Instagram;
- False Positive (FP) = 1 один випадок помилкової позитивної класифікації;
- False Negative (FN) = 4 модель пропустила чотирьох справжніх користувачів Instagram;
- True Positive (TP) = 55 точна класифікація більшості активних користувачів.

Загалом модель продемонструвала високу чутливість до позитивного класу, а низький рівень помилок підтверджує її ефективність для практичного застосування у виявленні користувачів Instagram.

На рисунку 3.22 наведено графік важливості ознак для моделі Random Forest, що продемонструвала найкращі результати при класифікації каналу WhatsApp.

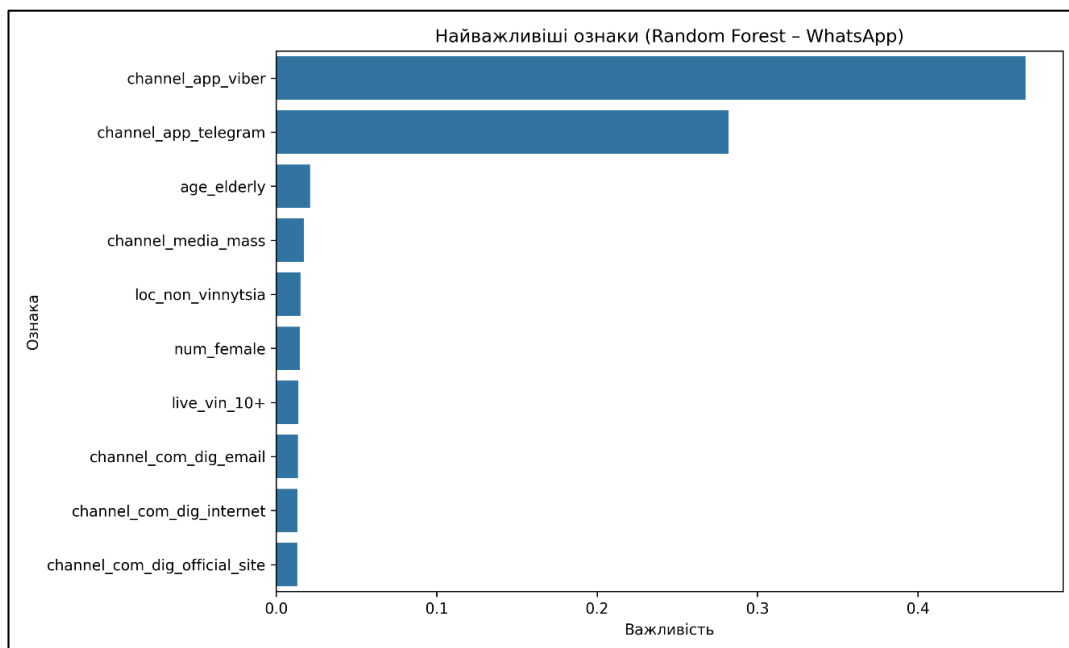


Рисунок 3.22 – Графік важливих ознак цифрового каналу WhatsApp для моделі Random Forest

Як видно з рисунка 3.22, найвищу вагу має ознака `channel_app_viber`, що вказує на схильність респондентів одночасно використовувати декілька месенджерів. Другу позицію займає `channel_app_telegram`, що підтверджує популярність мультимесенджерної поведінки серед респондентів. Інші вагомі чинники:

- `age_elderly` – старші респонденти частіше використовують WhatsApp, що відрізняє його від молодіжних платформ;
- `loc_non_vinnytsia` – мешканці з Вінницьких громад більш активно використовують цей канал;
- `num_female` та `live_vin_10+` – соціально-демографічні чинники (стать і тривалість проживання) впливають на залучення до додатку.
- `channel_media_mass`, `channel_com_dig_internet`, `channel_email`, `channel_com_dig_official_site` користувачі з високим рівнем цифрової активності також активно користуються й WhatsApp.

Аналіз важливих ознак моделі показав, що на ймовірність використання цифрового каналу WhatsApp суттєво впливають такі групи чинників:

- вікові характеристики;
- територіальний чинник;
- соціально–демографічні особливості;
- цифрова активність.

Оцінка точності за матрицею плутанини представлена на рисунку 3.23.

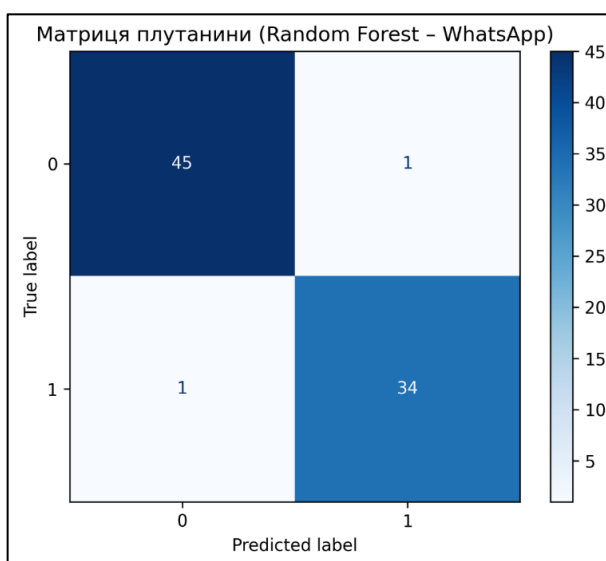


Рисунок 3.23 – Матриця плутанини моделі Random Forest для цифрового каналу WhatsApp

Результати матриці плутанини для моделі Random Forest для цифрового каналу WhatsApp наступні:

- True Negative (TN) = 45 правильно класифіковано респондентів, які не використовують WhatsApp;
- False Positive (FP) = 1 один випадок помилкової позитивної класифікації;
- False Negative (FN) = 1 один реальний користувач WhatsApp був класифікований як такий, що не користується додатком;
- True Positive (TP) = 34 точна класифікація користувачів WhatsApp.

Загалом модель правильно класифікує 98% усіх респондентів (79 із 81). Random Forest має високу ефективність та є надійним інструментом для прогнозування користування месенджером WhatsApp.

На рисунку 3.24 подано графік важливості ознак для моделі Random Forest, яка продемонструвала найкращі результати при класифікації використання Facebook.

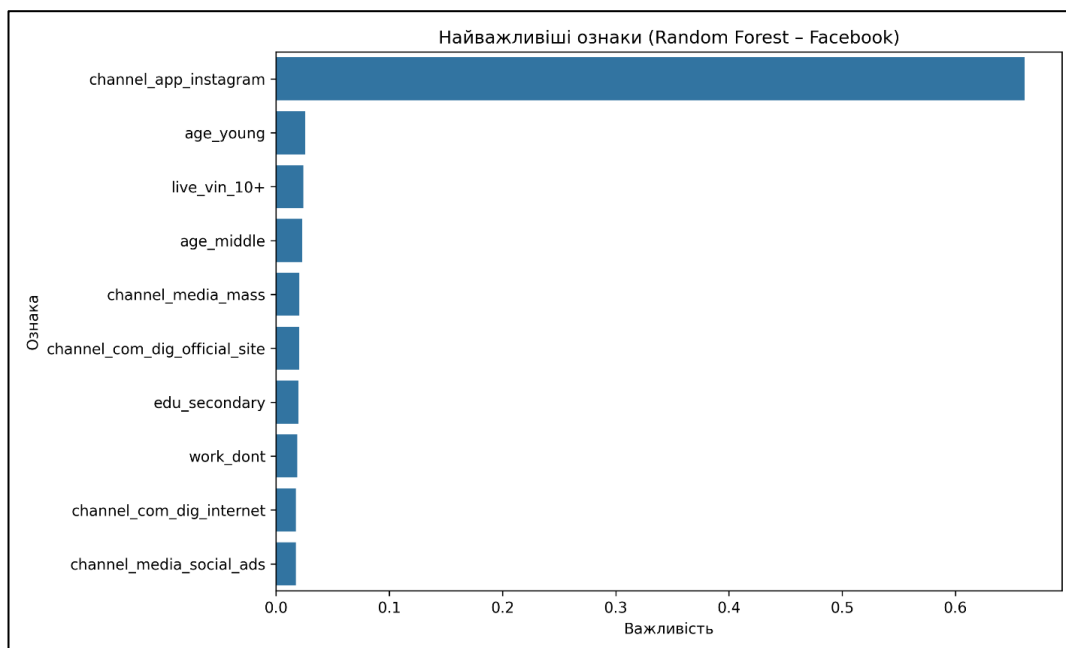


Рисунок 3.24 – Графік важливих ознак цифрового каналу Facebook для моделі Random Forest

На рисунку 3.24 представлено графік найважливіших ознак, що впливають на рішення моделі Random Forest щодо передбачення використання Facebook. Найсуттєвіший вплив має ознака `channel_app_instagram`, що свідчить про сильний зв'язок між користуванням Facebook і Instagram платформи часто використовуються паралельно. Серед інших ознак із помітним впливом маємо такі:

- `age_young` і `age_middle` Facebook; молоді та середнього віку респонденти більше користуються додатком;
- `live_vin_10+` тривалість проживання понад 10 років у громаді позитивно корелює з активністю у Facebook;

– `channel_media_mass`, `channel_com_dig_internet`, `channel_media_social_ads`, `channel_com_dig_official_site` користувачі з високим рівнем цифрової активності також активно використовують й Facebook;

– `edu_secondary`, `work_dont` соціально–освітні характеристики також мають помірний вплив на вибір користування додатком. Загалом модель враховує взаємопов'язані цифрові канали, вік, медіаспоживання та соціальні характеристики респондента, забезпечуючи комплексну оцінку ймовірності використання Facebook.

Аналіз важливих ознак моделі показав, що на ймовірність використання цифрового каналу Facebook суттєво впливають такі групи чинників:

- вікові характеристики;
- вік;
- соціально–освітні особливості;
- цифрова активність.

Оцінка точності за матрицею плутанини представлена на рисунку 3.25.

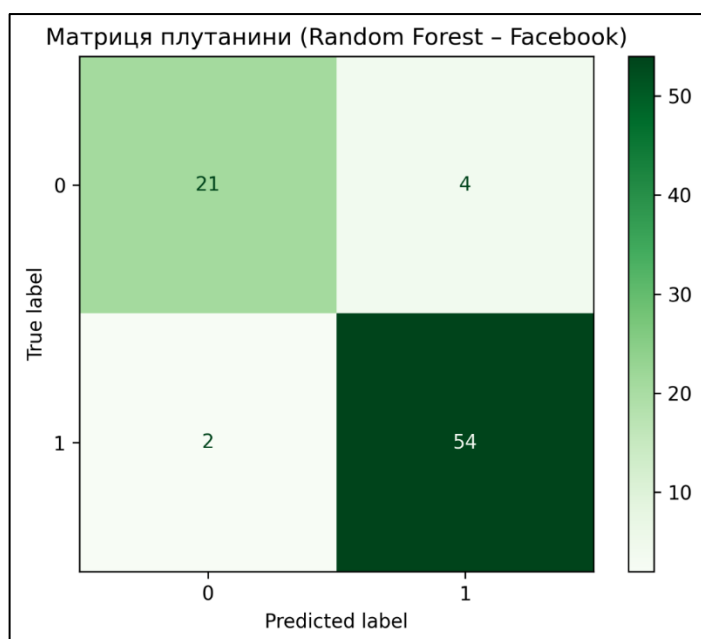


Рисунок 3.25 – Матриця плутанини моделі Random Forest для цифрового каналу Facebook

Результати матриці плутанини для моделі Random Forest для цифрового каналу Facebook такі:

- True Negative (TN) = 21 респонденти, які не використовують Facebook, правильно класифіковані;
- False Positive (FP) = 4 помилки позитивної класифікації;
- False Negative (FN) = 2 двоє реальних користувачів Facebook були класифіковані як такі, що не користуються;
- True Positive (TP) = 54 модель правильно ідентифікувала більшість користувачів Facebook.

Таким чином, модель правильно класифікує 75 із 81 прикладу, що відповідає загальній точності близько 93%. Модель демонструє стабільність та високу ефективність в умовах реальних даних, допускаючи незначну кількість помилок.

На рисунку 3.26 представлено графік найважливіших ознак, що впливають на рішення моделі XGBoost для передбачення використання Telegram.

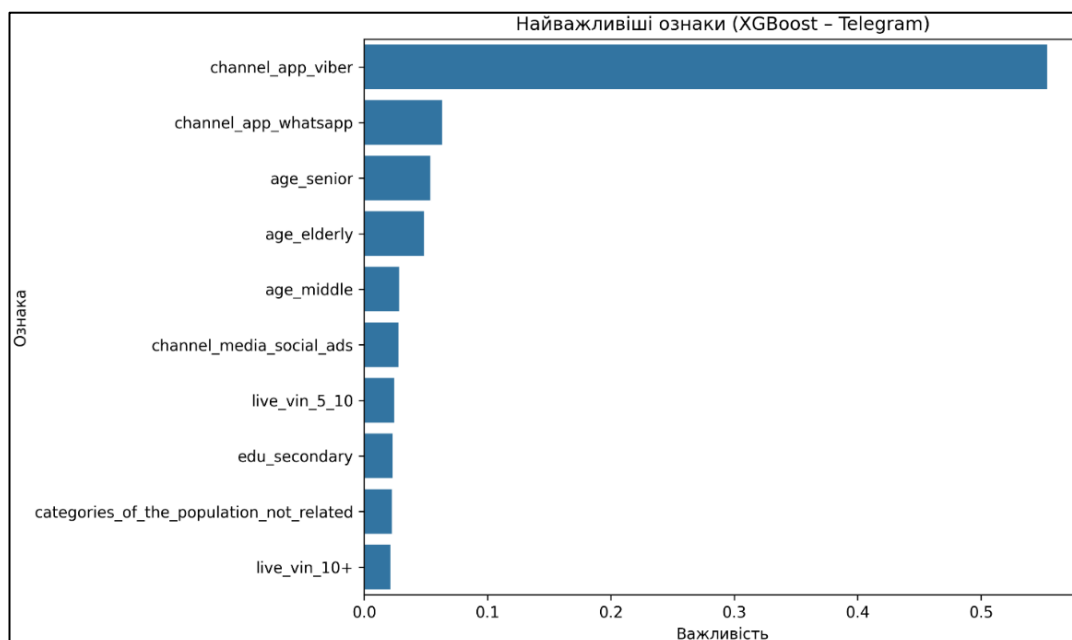


Рисунок 3.26 – Графік важливих ознак цифрового каналу Telegram для моделі XGBoost

З рисунка 3.26 бачимо, що найбільшу вагу має ознака `channel_app_viber`, що свідчить про високу ймовірність одночасного користування Viber та Telegram. Ця мультиканальність у межах месенджерів є типовою для цифрової поведінки активних користувачів. Інші впливові ознаки такі:

- `channel_app_whatsapp` ще один популярний месенджер, що доповнює Telegram у комунікації респондентів;
- `age_senior`, `age_elderly`, `age_middle` Telegram використовують представники старших вікових груп;
- `channel_media_social_ads` соціальна реклама також пов'язана з Telegram як каналом поширення інформації;
- `live_vin_5_10`, `live_vin_10+` – досвід тривалого проживання у громаді може впливати на рівень цифрової інтегрованості;
- `edu_secondary` та `categories_of_the_population_not_related` соціально-освітні змінні мають помірний, але наявний ефект.

Загалом Telegram виявляється частиною ширшого цифрового профілю, де взаємодія з іншими месенджерами є ключовим чинником.

Оцінка точності за матрицею плутанини представлено на рисунку 3.27.

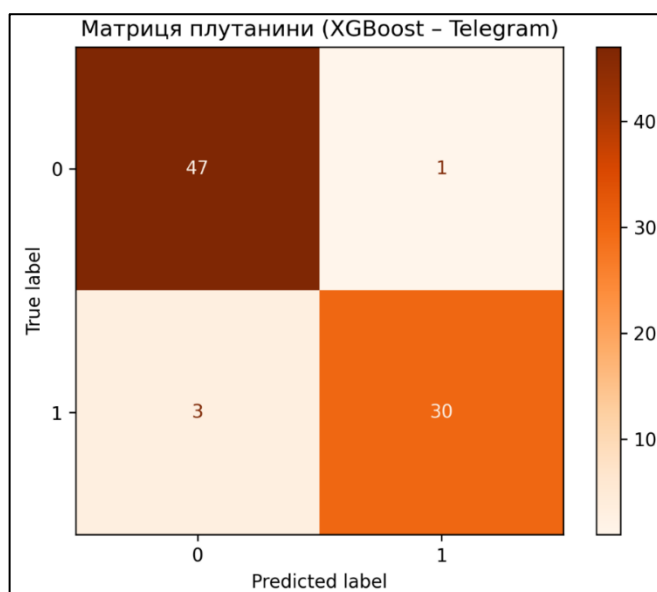


Рисунок 3.27 – Матриця плутанини моделі XGBoost для цифрового каналу Telegram

Результати матриці плутанини для моделі XGBoost для каналу Telegram наступні:

- True Negative (TN) = 47 модель правильно класифікує користувачів, які не використовують Telegram;
- False Positive (FP) = 1 лише один випадок хибного позитивного передбачення;
- False Negative (FN) = 3 – троє реальних користувачів Telegram були класифіковані помилково;
- True Positive (TP) = 30 – успішна ідентифікація більшості активних користувачів.

Загальна точність моделі є високою 77 з 81 респондента класифіковано правильно (~95%). Модель XGBoost демонструє стабільні результати, з особливою точністю у виявленні як позитивних, так і негативних класів.

На рисунку 3.28 представлено графік найважливіших ознак, що впливають на рішення моделі Random Forest для передбачення використання Viber.

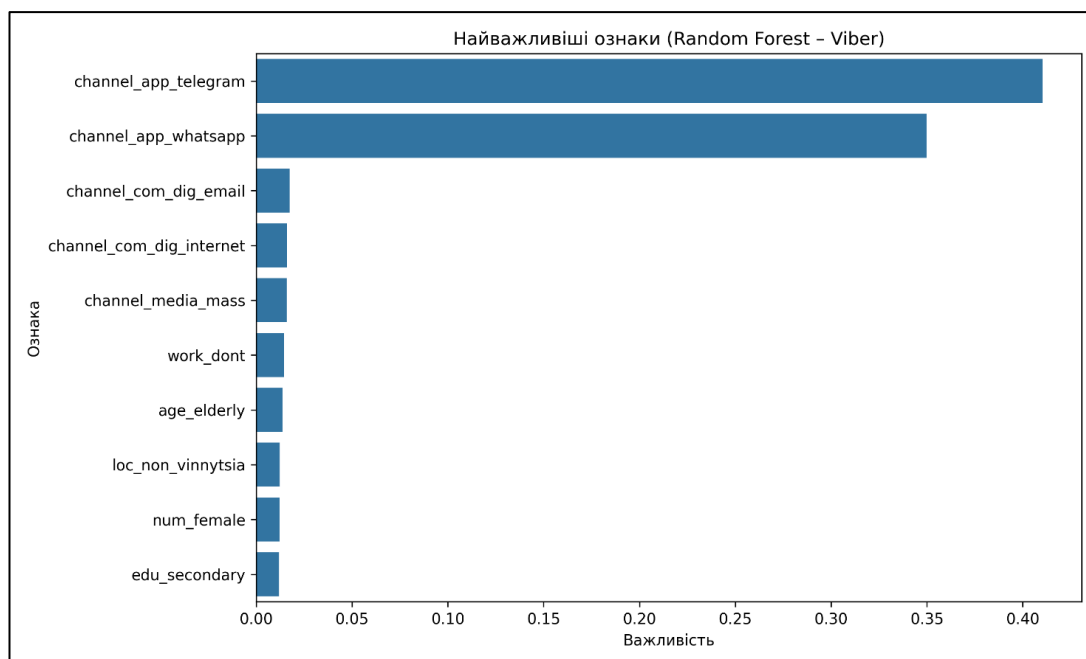


Рисунок 3.28 – Графік важливих ознак цифрового каналу Viber для моделі Random Forest

З рисунка 3.28 видно, що найбільшу вагу мають дві ознаки:

- `channel_app_telegram` найсильніший предиктор використання Viber, що вказує на тенденцію до паралельного користування кількома месенджерами;
- `channel_app_whatsapp` підтверджує мультиканальність респондентів, які активно користуються одразу кількома платформами для комунікації.

Інші ознаки з помітним, хоча й менш вираженим, впливом:

- `channel_com_dig_email`, `channel_com_dig_internet`, `channel_media_mass` між використанням Viber і загальною цифровою активністю;
- `work_dont`, `age_elderly`, `loc_non_vinnytsia` соціально-демографічні характеристики, що вказують на специфіку користувачів Viber серед певних категорій населення;
- `num_female`, `edu_secondary` освіта та стать мають незначний, але помітний вплив на поведінку.

Отже, Viber використовується здебільшого у поєднанні з іншими месенджерами, і його вибір зумовлений як цифровими звичками, так і соціальним профілем користувача.

Оцінка точності за матрицею плутанини представлена на рисунку 3.29.

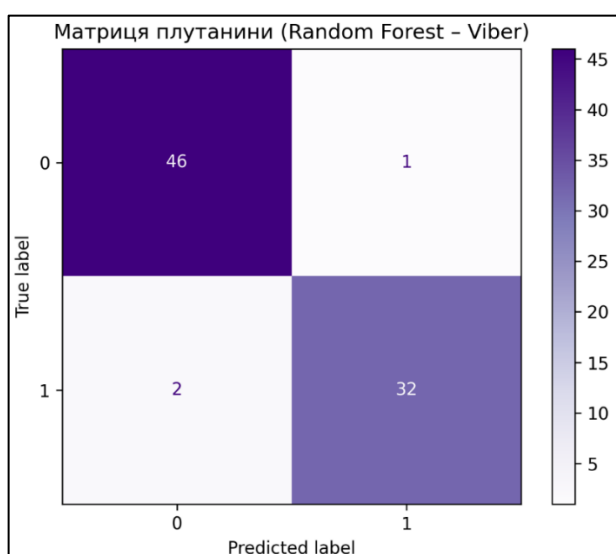


Рисунок 3.29 – Матриця плутанини моделі Random Forest для цифрового каналу Viber

Результати матриці плутанини для моделі Random Forest для каналу Viber:

- True Negative (TN) = 46 модель коректно класифікує респондентів, які не використовують Viber;
- False Positive (FP) = 1 один випадок помилкового позитивного передбачення;
- False Negative (FN) = 2 два користувачі Viber були помилково не розпізнані;
- True Positive (TP) = 32 правильна класифікація більшості активних користувачів.

Загальна кількість правильних передбачень 78 з 81 прикладу, що становить точність понад 96%.

Модель демонструє чудову збалансованість між виявленням позитивних та негативних класів, мінімальну кількість помилок і високу стабільність, що робить її оптимальним вибором для аналізу цифрового каналу Viber.

3.2 Висновки

У третьому розділі проведено передбачення використання цифрових каналів комунікації Вінницької міської територіальної громади на основі натренованих моделей машинного навчання: Logistic Regression, Random Forest та XGBoost, використовуючи тестові дані.

Для кожної моделі проведено налаштування гіперпараметрів із використанням GridSearchCV, тренування моделей і оцінку результатів на тестових вибірках.

Найвищі результати показали такі моделі:

- XGBoost показав найвищу точність для цифрових каналів Telegram (точність на тестових даних accuracy_score 0.951) та Instagram (0.938)
- Random Forest є найефективнішим для дослідження WhatsApp (0.975), Facebook (0.926) та Viber (0.963),

– Logistic Regression демонструє нижчу точність, однак є надійною базовою моделлю, яка забезпечує стабільні результати без перенавчання (наприклад, F1-score для Instagram – 0.956, ROC AUC – 0.941).

Також проведено аналіз важливих ознак кожної моделі, зокрема, виявлено ключові чинники, що впливають на використання каналів — вік, тривалість проживання у громаді, цифрова активність та використання інших платформ. Додатково побудовано матриці плутанини, які підтвердили високу чутливість моделей до позитивного класу та низький рівень хибних передбачень. Отримані результати свідчать про ефективність обраних моделей для розв'язання задачі передбачення цифрової взаємодії у межах Вінницької міської територіальної громади. Моделі на основі дерев рішень продемонстрували кращі результати в порівнянні з логістичною регресією, що може означати їх здатність краще моделювати нелінійні залежності у даних. Результати дослідження підтверджують доцільність використання моделей XGBoost та Random Forest у задачах передбачення цифрової взаємодії громади.

ВИСНОВКИ

У першому розділі бакалаврської кваліфікаційної роботи проаналізовано загальні характеристики об'єкта дослідження – соціальної взаємодії Вінницької міської територіальної громади через цифрові канали. Визначено, що це складна соціально–комунікаційна система, на яку впливають численні фактори: вік, стать, освіта, місце проживання, працевлаштування, належність до соціальної категорії тощо.

На основі аналізу наукових джерел і офіційних документів визначено, що цифрові канали комунікації, такі як Facebook, Telegram, Viber, Instagram, відіграють важливу роль у комунікації органів влади з громадою.

Здійснено обґрунтований вибір інструментів для реалізації дослідження. Обрано мову програмування Python як оптимальну для виконання задач аналізу даних і побудови моделей. Роботу реалізовано у хмарному середовищі Google Colab, яке забезпечує ефективну роботу з бібліотеками для машинного навчання, візуалізації та обробки інформації.

У другому розділі проведено розвідувальний аналіз даних, встановлено, що на цифрову активність найбільше впливають стать, вік, рівень освіти, тривалість проживання у громаді та соціальний статус. Найвищу цифрову залученість продемонстрували молодь, особи з інвалідністю та респонденти, які нещодавно оселилися у громаді. Під час первинного моделювання оцінено точність класифікаційних моделей на тренувальних даних для різних каналів цифрової взаємодії. Проведено оцінку моделей на тренувальних даних за різними метриками. За результатами цього етапу деякі цільові змінні не продемонстрували задовільної якості передбачення в межах трьох застосованих моделей – Logistic Regression, Random Forest і XGBoost, тому ці змінні були вилучені з подальшого аналізу. Дослідження було зосереджено на п'яти цільових змінних (Facebook, Instagram, Telegram, Viber, WhatsApp), які продемонстрували стабільну високу точність (понад 95%) та підтвердили надійність класифікації за допомогою обраних моделей.

У третьому розділі представлено результати передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали. На основі навчених моделей машинного навчання (Logistic Regression, Random Forest, XGBoost) виконано передбачення використання цифрових каналів, таких як Telegram, Viber, WhatsApp, Facebook та Instagram, з урахуванням соціально-демографічних характеристик респондентів.

Побудовано матриці плутанини для кожної моделі (Random Forest, XGBoost, Logistic Regression), що дало змогу оцінити для цифрових каналів, серед алгоритмів розмежування класів. Для цифрових каналів, серед зазначених моделей, оптимальними за метрикою `accuracy_score` на тестових даних виявилися наступні:

- WhatsApp: Random Forest — точність 0.975;
- Telegram: XGBoost — точність 0.951;
- Viber: Random Forest — точність 0.963;
- Instagram: XGBoost — точність 0.938;
- Facebook: Random Forest — точність 0.926.

Дослідження показало, що користувачі Instagram — це переважно молоді мешканці Вінницької міської територіальної громади віком 15–24 роки. Також вони активно користуються й іншими цифровими платформами — Facebook, Viber, офіційними сайтами та мас-медіа, що свідчить про високий рівень цифрової залученості. Більшість із них проживає в громаді понад 5 або 10 років, що вказує на їхню інтегрованість у місцеве середовище та інтерес до соціальних ініціатив.

Цифровим каналом Facebook користуються респонденти, які паралельно активно використовують Instagram. Це переважно молоді та середнього віку мешканці, здебільшого ті, хто проживає в громаді понад 10 років. Серед них часто трапляються особи з середньою освітою та тимчасово непрацевлаштовані. Користувачі Facebook також мають високий рівень цифрової активності — вони регулярно споживають контент мас-медіа, офіційних сайтів і соціальної реклами.

Користувачі Telegram активно використовують інші месенджери, зокрема Viber і WhatsApp. Цей канал популярний серед осіб старшого віку (45+ та 60+), які проживають у громаді понад 5 років. Часто це респонденти з середньою освітою або ті, що не належать до активних професійних і соціальних категорій. Telegram для них — це канал інформаційної підтримки та альтернативне джерело новин.

Viber найчастіше використовують мешканці громади, які водночас активні у Telegram і WhatsApp. Вони демонструють загальну цифрову активність, користуючись електронною поштою, інтернет-ресурсами та мас-медіа. Серед користувачів переважають старші респонденти, жінки, особи з середньою освітою, а також ті, хто мешкає за межами міста Вінниці. Частина з них є непрацевлаштованими, що може зумовлювати вищу частоту використання цифрових каналів.

Користувачі WhatsApp зазвичай паралельно користуються Viber і Telegram. Цей месенджер частіше обирають респонденти старшого віку, особливо ті, хто мешкає в навколишніх громадах і проживає у громаді понад 10 років. Вони характеризуються високим рівнем цифрової активності — активно споживають контент мас-медіа, офіційних сайтів, електронної пошти та інших онлайн-ресурсів.

В якості рекомендації для покращення соціальної взаємодії між місцевою владою та мешканцями громади варто адаптувати канали комунікації відповідно до цільових аудиторій: Instagram і Telegram — для молоді, Viber і Facebook — для старших вікових груп, WhatsApp — для старшої аудиторії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Туленков М. В., Лобанова А. С., Яремчук С. С. Системний аналіз у соціології : підручник. Чернівці : ЧНУ ім. Юрія Федьковича, 2023. 508 с. URL: <https://tinyurl.com/24rlbnfa> (дата звернення 06.05.2025).
2. Шпортко М. В. Цифровізація комунікативного простору Вінницької міської територіальної громади як стратегічний пріоритет її розвитку. *Соціально–економічні тенденції розвитку сучасної держави* : матеріали Міжнародної науково–практичної конференції, м. Черкаси, 13 листопада 2023 р. Черкаси, 2023. С.192-197. URL: <https://researcheurope.org/wp-content/uploads/2023/11/re-13.11.23.pdf> (дата звернення 06.05.2025).
3. Про затвердження Стратегії розвитку Вінницької міської територіальної громади до 2030 року – Стратегії 3.0 : Рішення Вінницької міської ради від 26.02.2021. № 194. 75 с. URL: <https://www.vmr.gov.ua/stratetiia> (дата звернення 08.05.2025).
4. Децентралізація в Україні. 2025. URL: https://decentralization.ua/newrayons?area_id=&%3Bpage=4&%3Bsort_by_otg_count=&%3Bsort_by_population=&%3Bsort_by_square=&sort_by_otg_count=&sort_by_villages_count=&sort_by_square=&sort_by_population= (дата звернення 08.05.2025).
5. Про проєкт рішення міської ради «Про затвердження Програми інформатизації та цифрової трансформації на 2024–2026 роки» : Рішення Вінницької міської ради Виконавчий комітет від 12.10.2023. № 2604. 73 с. URL: <https://tinyurl.com/2ycwktop> (дата звернення 09.05.2025).
6. Давиденко Г. В. Цифрова інклюзія та доступність: соціальна діджиталізація : монографія. Вінниця, 2023. 240 с. URL: https://www.vsei.vn.ua/images/Doc/Nauka/Inklusivna_osvita/cifrova-inklyuziya-ta-dostupnist-socialna-didzhitalizaciya.pdf (дата звернення 11.05.2025).

7. Гусев І. М. Особливості взаємодії з медіа в процесі адаптації спільноти до умов і наслідків воєнного конфлікту, *Наукові студії із соціальної та політичної психології*. 2019. Вип. 43. С. 72-81. URL: http://nbuv.gov.ua/UJRN/Nsspp_2019_43_10 (дата звернення 11.05.2025).

8. Сегень Я., Липін М. А. Трансформація медіа в умовах багатопольярного світу. *Соціокультурні трансформації та геополітичні виклики в умовах багатопольярного світу*: тези доповідей на Всеукраїнській науково-практичній конференції, м. Київ, 24 листопада 2022 року. Київ, 2022. С. 128-130. URL: <https://knute.edu.ua/file/MzEyMQ==/45d5c130f238f72bd31a477e5e1f3acc.pdf> (дата звернення 12.05.2025).

9. Бадюл О. С. Філософська природа аналізу духовності людини, *Наука. Релігія. Суспільство*. 2018. №2. С. 105–110. URL: <http://dspace.nbuv.gov.ua/bitstream/handle/123456789/29449/17-Badyul.pdf?sequence=1> (дата звернення 12.05.2025).

10. Wang H. Introduction to computer programming with Python : textbook. Canada. 2023. 504 p. DOI: <https://doi.org/10.15215/remix/9781998944088.01> (дата звернення 15.05.2025).

11. Соловйов В. М., Белінський А. О. Моделювання складних систем у python. : навчально–методичний посібник для самостійного вивчення дисципліни. Черкаси, 2024. 620 с. URL: <http://elibrary.kdpu.edu.ua/xmlui/handle/123456789/10508> (дата звернення 15.05.2025).

12. Мокін В. Б., Дратований М. В. Наука про дані: машинне навчання та інтелектуальний аналіз даних : навчальний посібник. Вінниця, 2024. 263 с. URL: <https://iq.vntu.edu.ua/repository/getfile.php/8163.pdf> (дата звернення 16.05.2025).

13. Зозуля Н. Ю., Трегуб А. М., Телелим І. В. Роль штучного інтелекту у розвитку соціальних комунікацій: виклики та перспективи, *Теорія та історія соціальних комунікацій*. 2025. №1. С. 141–145. DOI: <https://doi.org/10.32782/2710-4656/2025.1.2/23> (дата звернення 16.05.2025).

14. Lucas B.V. de Amorima,b, George D.C. Cavalcantia, Rafael M.O. Cruz. The choice of scaling technique matters for classification performance, *Soft Computing*. 2022. (Preprint submitted to Applied Soft Computing) P. 37. URL: <https://arxiv.org/pdf/2212.12343> (дата звернення 16.05.2025).

15. Tan J., Yang J., Wu S., Chen G., Zhao J. A critical look at the current train/test split in machine learning, *Machine Learning*. 2021 P. 1-21 URL: <http://arxiv.org/pdf/2106.04525> (дата звернення 16.05.2025).

16. Іванчук Я. В., Месюра В. І., Яровий А. А. Манжілевський О. Д. Інтелектуальний аналіз даних та машинне навчання частина 1 базові методи та засоби аналізу даних : навчальний посібник. Вінниця : ВНТУ, 2021 р. 680 с. URL: https://pdf.lib.vntu.edu.ua/books/2022/Ivanchuk_P1_2021_69.pdf (дата звернення 17.05.2025).

17. Ситник В. Ф., Ситник Н. В. Древа рішень в системах дейтамінгу, *Науково-технічна інформація*. 2021. №3. С. 442–454. URL: <https://ir.kneu.edu.ua/server/api/core/bitstreams/897117f9-492f-4710-bc40-d6755c070fb0/content> (дата звернення 17.05.2025).

18. Шеремет О. І., Садовий О. В. Метод опорних векторів (SVM), *Математичне моделювання*. 2023. №1. С. 13–18. URL: <https://www.dstu.dp.ua/Portal/Data/74/72/3st13-17.pdf> (дата звернення 19.05.2025).

19. Кравець П. О., Пасічник В. В., Проданюк М. М. Математична модель логістичної регресії для бінарної класифікації Ч.1. Регресійні моделі узагальнення даних, *Information systems and networks*. 2024. С. 290–321. DOI: <https://doi.org/10.23939/sisn2024.15.290> (дата звернення 19.05.2025).

20. Сташук Д., Дубровін В., Тарасова Ю. Випадковий ліс: пошук аномалій у комп'ютерних мережах, *Матеріали XXV Міжнародного науково-практичного семінару імені А. Я. Петренюка. Комбінаторні конфігурації та їхні застосування*. 2023. С. 190–195. URL: https://zp.edu.ua/uploads/dept_s&r/2023/conf/1.4/Petrenyuk_ISPS-25-proc.pdf (дата звернення 19.05.2025).

21. Гапонова В. О., Бученко І. А. Машинне навчання та нейронні мережі аналізу великих даних, на *III Всеукраїнській науково-практичній конференції «Сучасні інтелектуальні інформаційні технології в науці та освіті»*, 2023. С. 48–49. URL: https://duikt.edu.ua/uploads/n_11208_13331372.pdf?file=n_11208_13331372.pdf#page=48 (дата звернення 21.05.2025).
22. Ліп'яніна-Гончаренко Х., Юрків Х. Методи бустингового машинного навчання для нестационарних часових рядів, *Вимірювальна та обчислювальна техніка в технологічних процесах*. 2023. №3. С. 19–30. DOI: <https://doi.org/10.31891/2219-9365-2023-75-2> (дата звернення 22.05.2025).
23. Chukwura J.O. A Comparative Study of Several Classification Metrics and Their Performances on Data, *World Journal of Advanced Engineering Technology and Sciences*. 2023. №8(1). Р. 308–314. DOI: <https://doi.org/10.30574/wjaets.2023.8.1.0054> (дата звернення 24.05.2025).
24. Yang L., Shami A., On hyperparameter optimization of machine learning algorithms: theory and practice: Preprint submitted to Neurocomputing. 2022. 69 p. URL: <https://arxiv.org/pdf/2007.15745> (дата звернення 24.05.2025).
25. XGBoost Parameters Tuning: A Complete Guide with Python Codes, 2025. URL: <https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/> (дата звернення 25.05.2025).
26. Mastering Logistic Regression with Scikit-Learn: A Complete Guide, 2025. URL: <https://www.digitalocean.com/community/tutorials/logistic-regression-with-scikit-learn> (дата звернення 25.05.2025).
27. A Comprehensive Guide to Random Forest in Machine Learning, 2025. URL: <https://medium.com/@shridharpawar77/a-comprehensive-guide-to-random-forest-in-machine-learning-50ddaef79974> (дата звернення 26.05.2025).
28. Вакалюк Т. А., Янчук В. М., Морозов Д. С., Зубрицький В. В., Новіцька І. В. Прогнозне моделювання аналітики успішності студентів з використанням алгоритмів машинного навчання, *Інформатика, обчислювальна техніка та*

автоматизація. 2023. Том 34 (73) №.5. С. 108–116. DOI
<https://doi.org/10.32782/2663-5941/2023.5/18>. (дата звернення 26.05.2025).

Додаток А
(обов'язковий)

Міністерство освіти і науки України
Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації

ЗАТВЕРДЖУЮ

Завідувач кафедри САІТ

_____ д.т.н., проф. Віталій МОКІН

« ____ » _____ 2025 р.

ТЕХНІЧНЕ ЗАВДАННЯ

на бакалаврську кваліфікаційну роботу

СИСТЕМНИЙ АНАЛІЗ СОЦІАЛЬНОЇ ВЗАЄМОДІЇ ВІННИЦЬКОЇ МІСЬКОЇ
ТЕРИТОРІАЛЬНОЇ ГРОМАДИ ЧЕРЕЗ ЦИФРОВІ КАНАЛИ

08–34.БКР.001.00.000 ТЗ

Керівник: Ph.D., ст. викл. каф. САІТ

_____ Ольга ВОЙЦЕХОВСЬКА

« ____ » _____ 2025 р.

Розробила студентка гр. СА–216

_____ Анна БОЯНОВСЬКА

« ____ » _____ 2025 р.

Вінниця 2025

1. Підстава для проведення робіт.

Підставою для виконання роботи є наказ №__ по ВНТУ від «__» _____2025р., та індивідуальне завдання на БКР, затверджене протоколом №__ засідання кафедри САІТ від «__» _____ 2025р.

2. Джерела розробки:

1) Шпортко М. В. Цифровізація комунікативного простору Вінницької міської територіальної громади як стратегічний пріоритет її розвитку. *Соціально-економічні тенденції розвитку сучасної держави* : матеріали Міжнародної науково-практичної конференції, м. Черкаси, 13 листопада 2023 р. Черкаси, 2023. С.192-197. URL: <https://tinyurl.com/bdfb2skf>

2) Давиденко Г. В. Цифрова інклюзія та доступність: соціальна діджиталізація : монографія. Вінниця, 2023. 240 с. URL: <https://tinyurl.com/27yagguv>

3. Мета і призначення роботи.

Метою дослідження є системний аналіз соціальної взаємодії Вінницької міської територіальної громади та підвищення точності передбачення використання цифрових каналів.

4. Вихідні дані для проведення робіт:

Дані соціологічного опитування мешканців Вінницької міської територіальної громади.

5. Методи дослідження:

Розвідувальний аналіз, інженерія ознак, моделі машинного навчання.

6. Етапи роботи і терміни їх виконання:

а) Аналіз предметної області. _____ — _____

б) Аналіз наукових робіт дослідження соціальної взаємодії територіальних громад через цифрові канали. _____ — _____

с) Вибір оптимального рішення та налаштувань для розв'язання поставленої задачі. _____ — _____

д) Результати передбачення соціальної взаємодії Вінницької міської територіальної громади через цифрові канали. _____ — _____

е) Оформлення матеріалів до захисту БКР _____ — _____

7. Очікувані результати та порядок реалізації

Отримання оптимальної моделі машинного навчання для передбачення соціальної взаємодії мешканців Вінницької міської територіальної громади через цифрові канали.

8. Вимоги до розробленої документації

Текстова та ілюстративна частини роботи оформлені у відповідності до вимог «Методичних вказівок до виконання бакалаврських кваліфікаційних робіт для студентів спеціальностей: 124 «Системний аналіз», 126 «Інформаційні системи та технології» (освітня програма «Прикладні інформаційні технології»).

9. Порядок приймання роботи

Публічний захист «__» _____ 2025 р.

Початок розробки «__» _____ 2025 р.

Граничні терміни виконання БКР «__» _____ 2025 р.

Додаток Б
(обов'язковий)

ПРОТОКОЛ ПЕРЕВІРКИ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Назва роботи: «Системний аналіз соціальної взаємодії Вінницької міської територіальної громади через цифрові канали»

Тип роботи: бакалаврська кваліфікаційна робота

Підрозділ: кафедра САІТ, ФІІТА, гр. СА-216

Коефіцієнт подібності текстових запозичень, виявлених у роботі
системою StrikePlagiarism 5,17 %

Висновок щодо перевірки кваліфікаційної роботи (відмітити потрібне):

- Запозичення, виявлені у роботі, є законними і не містять ознак плагіату, фабрикації, фальсифікації. Роботу прийняти до захисту
- ☐ У роботі не виявлено ознак плагіату, фабрикації, фальсифікації, але надмірна кількість текстових запозичень та/або наявність типових розрахунків не дозволяють прийняти рішення про оригінальність та самостійність її виконання. Роботу направити на доопрацювання.
- ☐ У роботі виявлено ознаки плагіату та/або текстових маніпуляцій як спроб укриття плагіату, фабрикації, фальсифікації, що суперечить вимогам законодавства та нормам академічної доброчесності. Робота до захисту не приймається.

Експертна комісія:

Віталій МОКІН, зав. каф. САІТ

_____ (підпис)

Сергій ЖУКОВ, доц. каф. САІТ

_____ (підпис)

Особа, відповідальна за перевірку _____ (підпис)

Сергій ЖУКОВ

З висновком експертної комісії ознайомлений(-на)

Керівник _____ (підпис)

Ольга ВОЙЦЕХОВСЬКА, Ph.D., ст. викл. каф. САІТ

Здобувач _____ (підпис)

Анна БОЯНОВСЬКА

Додаток В (довідниковий)

Фрагмент лістингу програми

Код оптимізація даних ознак (локація, освіта, тривалість проживання Вінницькій ОТГ)

```
# 1. Об'єднання малих населених пунктів біля Вінниці в одну колонку
non_vinnytsia_cols = [
    'loc_v_krushlyntsi', 'loc_havryshivka', 'loc_m_krushlyntsi',
    'loc_vinnytski_khutory', 'loc_stadnytsia', 'loc_desna',
    'loc_pysarivka', 'loc_shchitky']
df['loc_non_vinnytsia'] = df[non_vinnytsia_cols].max(axis=1)
df.drop(columns=non_vinnytsia_cols, inplace=True)

# Списки ознак за категоріями освіти
higher_edu = ['edu_incomplete_higher', 'edu_full_higher', 'edu_academic_degree']
secondary_edu = ['edu_basic_higher', 'edu_vocational', 'edu_full_secondary']

# Створення нових бінарних ознак
df['edu_higher'] = df[higher_edu].max(axis=1)
df['edu_secondary'] = df[secondary_edu].max(axis=1)

# Видалення старих ознак освіти
df.drop(columns=higher_edu + secondary_edu, inplace=True)

# 3. Створення нової ознаки для проживання 0–3 роки
df['live_vin_0_3'] = df[['live_vin_1_3', 'live_vin_less_1']].max(axis=1)
df.drop(columns=['live_vin_1_3', 'live_vin_less_1'], inplace=True)
```

Код оптимізація даних зменшення надлишковості в датасеті шляхом виключення високо корельованих змінних

```
import pandas as pd
import numpy as np

# Матриця кореляцій
corr_matrix = df.corr(numeric_only=True)

# визначення списку важливих ознак
```

```

important_features = ['channel_app_telegram', 'channel_app_viber',
                      'channel_app_whatsapp', 'channel_app_facebook', 'channel_app_instagram']

# Пошук корельованих пар
high_corr_pairs = []

for i in range(len(corr_matrix.columns)):
    for j in range(i):
        corr_val = corr_matrix.iloc[i, j]
        if abs(corr_val) > 0.8:
            col1 = corr_matrix.columns[i]
            col2 = corr_matrix.columns[j]
            high_corr_pairs.append((col1, col2, corr_val))

# Визначення ознак, що видаляються так як зберігаються
cols_to_drop = set()

for col1, col2, _ in high_corr_pairs:
    if col1 in important_features and col2 in important_features:
        continue # обидві важливі – нічого не видаляти
    elif col2 in important_features:
        cols_to_drop.add(col1)
    elif col1 in important_features:
        cols_to_drop.add(col2)
    else:
        cols_to_drop.add(col2)

# Видалення вибраних ознак
df.drop(columns=list(cols_to_drop), inplace=True)

```

Код графік важливих ознак, що впливають на вибір цифрового каналу WhatsApp

```

# Відокремлення цільової змінної
y = df['channel_app_whatsapp']

X = df.drop(columns=['channel_app_whatsapp'])

# Масштабування ознак
scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)

# Розділення на тренувальні та тестові дані

```

```

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Побудова моделі логістичної регресії
model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

# Вибір ознак
coef_series = pd.Series(model.coef_[0], index=X.columns)
positive_features = coef_series[coef_series > 0].sort_values()

# Побудова та збереження графіка позитивних ознак
plt.figure(figsize=(12, 8))
positive_features.plot(kind='barh')

plt.title('Позитивні ознаки, що впливають на використання WhatsApp')
plt.xlabel('Бага ознаки (коефіцієнт)')
plt.grid(True)
plt.tight_layout()

```

Код автоматизованого підбору гіперпараметрів за допомогою GridSearchCV для трьох моделей машинного навчання

```

# Цільова змінна – WhatsApp
y = df['channel_app_whatsapp']
X = df.drop(columns=['channel_app_whatsapp'])

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Logistic Regression
log_params = {'C': [0.01, 0.1, 1, 10],
              'penalty': ['l2'],
              'solver': ['lbfgs'],
              'max_iter': [100, 200, 500]}

log_model = GridSearchCV(LogisticRegression(), log_params, cv=5, scoring='accuracy', n_jobs=-1)
log_model.fit(X_train, y_train)
log_best = log_model.best_estimator_

# Random Forest
rf_params = {'n_estimators': [100, 200],

```

```

'max_depth': [5, 10, None],
'max_features': ['sqrt', 'log2'],
'min_samples_split': [2, 5],
'min_samples_leaf': [1, 2],
'bootstrap': [True, False]}

rf_model = GridSearchCV(RandomForestClassifier(), rf_params, cv=5, scoring='accuracy', n_jobs=-1)

rf_model.fit(X_train, y_train)

rf_best = rf_model.best_estimator_

# XGBoost

xgb_params = {'max_depth': [3, 5, 7],
              'min_child_weight': [1, 3],
              'gamma': [0, 0.1],
              'subsample': [0.8, 1.0],
              'colsample_bytree': [0.8, 1.0],
              'learning_rate': [0.01, 0.1],
              'n_estimators': [100, 200]}

xgb_model = GridSearchCV(
    XGBClassifier(use_label_encoder=False, eval_metric='logloss'),
    xgb_params, cv=5, scoring='accuracy', n_jobs=-1)

xgb_model.fit(X_train, y_train)

xgb_best = xgb_model.best_estimator_

# Метрики

def print_metrics(name, model, X_test, y_test):
    y_pred = model.predict(X_test)
    y_proba = model.predict_proba(X_test)[:, 1] if hasattr(model, "predict_proba") else y_pred
    print(f"\n{name} Метрики:")
    print("Accuracy:", accuracy_score(y_test, y_pred))
    print("Precision:", precision_score(y_test, y_pred))
    print("Recall:", recall_score(y_test, y_pred))
    print("F1 Score:", f1_score(y_test, y_pred))
    print("ROC AUC Score:", roc_auc_score(y_test, y_proba))

```

Код побудова матриці плутанини

```

import matplotlib.pyplot as plt

from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, accuracy_score

# XGBoost
y_pred_xgb = xgb_best.predict(X_test)
acc_xgb = accuracy_score(y_test, y_pred_xgb)
plt.figure(figsize=(6, 5))
cm_xgb = confusion_matrix(y_test, y_pred_xgb)
ConfusionMatrixDisplay(cm_xgb).plot(cmap='Oranges_r', values_format='d')
plt.title('Матриця плутанини – XGBoost\nПередбачення використання Instagram')
plt.show()

# Logistic Regression
y_pred_log = log_best.predict(X_test)
acc_log = accuracy_score(y_test, y_pred_log)
plt.figure(figsize=(6, 5))
cm_log = confusion_matrix(y_test, y_pred_log)
ConfusionMatrixDisplay(cm_log).plot(cmap='Blues_r', values_format='d')
plt.title('Матриця плутанини – Logistic Regression\nПередбачення використання Instagram')
plt.show()

# Random Forest
y_pred_rf = rf_best.predict(X_test)
acc_rf = accuracy_score(y_test, y_pred_rf)
plt.figure(figsize=(6, 5))
cm_rf = confusion_matrix(y_test, y_pred_rf)
ConfusionMatrixDisplay(cm_rf).plot(cmap='binary', values_format='d')
plt.title('Матриця плутанини – Random Forest\nПередбачення використання Instagram')
plt.show()

# Підсумкова точність
print("Точність моделей на тестових даних (channel_app_instagram):")
print(f"XGBoost:      {acc_xgb:.4f}")
print(f"Logistic Regression: {acc_log:.4f}")
print(f"Random Forest:    {acc_rf:.4f}")

```

Код перехресної перевірки

```

# Словник моделей
models = {'LogisticRegression': LogisticRegression(max_iter=1000),
          'RandomForest': RandomForestClassifier(n_estimators=100, random_state=42),
          'XGBoost': XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)}

# Порівняння моделей
for name, model in models.items():
    # Крос-валідація
    cv_scores = cross_val_score(model, X_train, y_train, cv=5, scoring='accuracy')
    cv_mean = cv_scores.mean()
    cv_std = cv_scores.std()

    # Навчання і точність
    model.fit(X_train, y_train)
    train_acc = accuracy_score(y_train, model.predict(X_train))
    test_acc = accuracy_score(y_test, model.predict(X_test))

    # Виведення результатів
    print(f"\n{name} (Instagram):")
    print(f"Крос-валідація: середнє = {cv_mean:.3f}, std = {cv_std:.3f}")
    print(f"Точність train: {train_acc:.3f}")
    print(f"Точність test: {test_acc:.3f}")
    if train_acc - test_acc > 0.1:
        print("Можливе перенавчання (overfitting)")
    else:
        print("Модель узагальнює добре")

```

Додаток Г
(обов'язковий)

ІЛЮСТРАТИВНА ЧАСТИНА

**СИСТЕМНИЙ АНАЛІЗ СОЦІАЛЬНОЇ ВЗАЄМОДІЇ ВІННИЦЬКОЇ МІСЬКОЇ
ТЕРИТОРІАЛЬНОЇ ГРОМАДИ ЧЕРЕЗ ЦИФРОВІ КАНАЛИ**

Нормоконтроль: к.т.н., доцент

_____ Сергій ЖУКОВ

«___» _____ 2025 р.

```
# Список колонок, які потрібно перетворити на булеві значення (True – обрав, False – не обрав)
bool_columns = ['num_male', 'num_female',
                'age_young', 'age_middle', 'age_senior', 'age_elderly',
                'edu_basic_higher', 'edu_incomplete_higher', 'edu_full_higher', 'edu_vocational',
                'edu_basic_secondary', 'edu_academic_degree', 'edu_only_primary', 'edu_two_higher', 'edu_full_secondary',
                'loc_vinnitsia_city', 'loc_v_krushlyntsi', 'loc_havryshivka', 'loc_m_krushlyntsi',
                'loc_vinnytski_khutory', 'loc_stadnytsia', 'loc_desna', 'loc_pysarivka', 'loc_shchitky',
                'live_vin_1_3', 'live_vin_5_10', 'live_vin_always', 'live_vin_10+', 'live_vin_less_1',
                'work_have', 'work_dont', 'work_part_time_job', 'work_pensioner', 'work_fop',
                'work_mat_leave', 'work_ukrainian_army', 'work_looking_for_work',
                'categories_of_the_population_idp', 'categories_of_the_population_not_related',
                'categories_of_the_population_retired', 'categories_of_the_population_family_member_of_veteran_war',
                'categories_of_the_population_from_large_family', 'categories_of_the_population_family_member_of_the_deceased',
                'categories_of_the_population_handicapped_person', 'categories_of_the_population_student',
                'categories_of_the_population_veteran',
                'channel_com_dig_email', 'channel_com_dig_official_site', 'channel_com_dig_internet',
                'channel_media_mass', 'channel_media_tv', 'channel_media_social_ads',
                'channel_media_info_stands', 'channel_media_radio',
                'channel_app_viber', 'channel_app_instagram', 'channel_app_telegram',
                'channel_app_facebook', 'channel_app_whatsapp']

df[bool_columns] = df[bool_columns].notna() # Перетворення: якщо не NaN – True (обрано), якщо NaN – False (не обрано)
df[bool_columns].head() # Перевірка результату
```

Рисунок Г.1 – Код перетворення категоріальних ознак на булеві значення

```
# функція булевого кодування TRUE → 1, FALSE → 0
def convert_bool(val):
    val_str = str(val).strip().lower()
    if val_str in ['true', 'істина']:
        return 1
    elif val_str in ['false', 'хибність']:
        return 0
    elif val_str in ['1', '0']:
        return val
    else:
        return val
```

Рисунок Г.2 – Функція булевого кодування

```
# Масштабування ознак
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Рисунок Г.3 – Масштабування даних методом StandardScaler

```

# 1. Об'єднання малих населених пунктів біля Вінниці в одну колонку
non_vinnytsia_cols = [
    'loc_v_krushlyntsi', 'loc_havryshivka', 'loc_m_krushlyntsi',
    'loc_vinnytski_khutory', 'loc_stadnytsia', 'loc_desna',
    'loc_pysarivka', 'loc_shchitky'
]
df['loc_non_vinnytsia'] = df[non_vinnytsia_cols].max(axis=1)
df.drop(columns=non_vinnytsia_cols, inplace=True)

# Списки ознак за категоріями освіти
higher_edu = ['edu_incomplete_higher', 'edu_full_higher', 'edu_academic_degree']
secondary_edu = ['edu_basic_higher', 'edu_vocational', 'edu_full_secondary']

# Створення нових бінарних ознак
df['edu_higher'] = df[higher_edu].max(axis=1)
df['edu_secondary'] = df[secondary_edu].max(axis=1)

# Видалення старих ознак освіти
df.drop(columns=higher_edu + secondary_edu, inplace=True)

# 3. Створення нової ознаки для проживання 0-3 роки
df['live_vin_0_3'] = df[['live_vin_1_3', 'live_vin_less_1']].max(axis=1)
df.drop(columns=['live_vin_1_3', 'live_vin_less_1'], inplace=True)

```

Рисунок Г.4 – Об'єднання ознак

```

# 2. Визначимо список важливих ознак
important_features = [ 'channel_app_telegram', 'channel_app_viber',
    'channel_app_whatsapp', 'channel_app_facebook', 'channel_app_instagram' ]

# 3. Знайдемо корельовані пари
high_corr_pairs = []
for i in range(len(corr_matrix.columns)):
    for j in range(i):
        corr_val = corr_matrix.iloc[i, j]
        if abs(corr_val) > 0.8:
            col1 = corr_matrix.columns[i]
            col2 = corr_matrix.columns[j]
            high_corr_pairs.append((col1, col2, corr_val))

# 4. Визначимо, які ознаки видаляти, зберігаючи важливі
cols_to_drop = set()
for col1, col2, _ in high_corr_pairs:
    if col1 in important_features and col2 in important_features:
        continue # обидві важливі – нічого не видаляти
    elif col2 in important_features:
        cols_to_drop.add(col1)
    elif col1 in important_features:
        cols_to_drop.add(col2)
    else:
        cols_to_drop.add(col2) # за замовчуванням видаляємо другу

# 5. Видаляємо вибрані ознаки
df.drop(columns=list(cols_to_drop), inplace=True)

```

Рисунок Г.5 – Автоматичне виявлення та видалення кореляційних пар ознак

```

from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV

# Налаштування гіперпараметрів
params = {
    'max_depth': [3, 5, 7],
    'min_child_weight': [1, 3],
    'gamma': [0, 0.1],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0],
    'learning_rate': [0.01, 0.1],
    'n_estimators': [100, 200]
}

# Підбір параметрів і тренування моделі
model = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
model_xgb = GridSearchCV(model, param_grid=params, scoring='accuracy', cv=5, n_jobs=-1)
model_xgb.fit(X_train, y_train)

# Оцінка точності на тренувальних даних
y_pred_train = model_xgb.predict(X_train)
acc_train = accuracy_score(y_train, y_pred_train)
print("Точність XGBoost на тренувальних даних:", acc_train)

```

Рисунок Г.6 – Модель XGBoost

```

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV

# Налаштування гіперпараметрів
params = {
    'C': [0.01, 0.1, 1, 10],
    'penalty': ['l2'],
    'solver': ['lbfgs'],
    'max_iter': [100, 200, 500]
}

# Підбір параметрів і тренування моделі
model = LogisticRegression()
model_log = GridSearchCV(model, param_grid=params, scoring='accuracy', cv=5, n_jobs=-1)
model_log.fit(X_train, y_train)

# Оцінка точності на тренувальних даних
y_pred_train = model_log.predict(X_train)
acc_train = accuracy_score(y_train, y_pred_train)
print("Точність Logistic Regression на тренувальних даних:", acc_train)

```

Рисунок Г.7 – Модель Logistic Regression

```

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV

# Налаштування гіперпараметрів
params = {
    'n_estimators': [100, 200],
    'max_depth': [5, 10, None],
    'max_features': ['sqrt', 'log2'],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
    'bootstrap': [True, False]
}

# Підбір параметрів і тренування моделі
model = RandomForestClassifier()
model_rf = GridSearchCV(model, param_grid=params, scoring='accuracy', cv=5, n_jobs=-1)
model_rf.fit(X_train, y_train)

# Оцінка точності на тренувальних даних
y_pred_train = model_rf.predict(X_train)
acc_train = accuracy_score(y_train, y_pred_train)
print("Точність Random Forest на тренувальних даних:", acc_train)

```

Рисунок Г.8 – Модель Random Forest

```

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score

# Список моделей та їх назви
models = [
    ("Logistic Regression", log_model),
    ("Random Forest", rf_model),
    ("XGBoost", model_xgb)]

# Виведення метрик для кожної моделі
for name, model in models:
    y_pred = model.predict(X_test)
    y_proba = model.predict_proba(X_test)[:, 1] if hasattr(model, "predict_proba") else y_pred

    print(f"\nМетрики {name} на тестових даних:")
    print("Accuracy:", accuracy_score(y_test, y_pred))
    print("Precision:", precision_score(y_test, y_pred))
    print("Recall:", recall_score(y_test, y_pred))
    print("F1 Score:", f1_score(y_test, y_pred))
    print("ROC AUC Score:", roc_auc_score(y_test, y_proba))

```

Рисунок Г.9 – Код оцінки моделей для передбачення на тестових даних

```

#матриця
plt.figure(figsize=(6, 5))
cm_xgb = confusion_matrix(y_test, y_pred_xgb)
ConfusionMatrixDisplay(cm_xgb).plot(cmap='Oranges_r', values_format='d')
plt.title('Матриця плутанини - XGBoost\nПередбачення використання Telegram')
plt.show()

```

Рисунок Г.10 – Код матриці плутанини моделі XGBoost для цифрового каналу Telegram

```
# Матриця XGBoost
y_pred_xgb = xgb_best.predict(X_test)
acc_xgb = accuracy_score(y_test, y_pred_xgb)
plt.figure(figsize=(6, 5))
cm_xgb = confusion_matrix(y_test, y_pred_xgb)
ConfusionMatrixDisplay(cm_xgb).plot(cmap='Oranges_r', values_format='d')
plt.title('Матриця плутанини - XGBoost\nПередбачення використання WhatsApp')
plt.show()
```

Рисунок Г.11 – Код матриці плутанини для моделі XGBoost

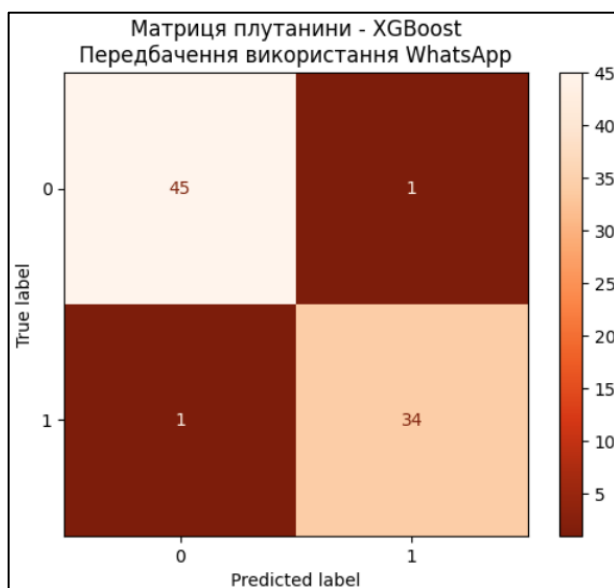


Рисунок Г.12 – Матриця плутанини результатів моделі XGBoost для цифрового каналу WhatsApp

```
# Матриця Logistic Regression
y_pred_log = log_best.predict(X_test)
acc_log = accuracy_score(y_test, y_pred_log)
plt.figure(figsize=(6, 5))
cm_log = confusion_matrix(y_test, y_pred_log)
ConfusionMatrixDisplay(cm_log).plot(cmap='Blues_r', values_format='d')
plt.title('Матриця плутанини - Logistic Regression\nПередбачення використання WhatsApp')
plt.show()
```

Рисунок Г.13 – Код матриці плутанини для моделі Logistic Regression для цифрового каналу WhatsApp

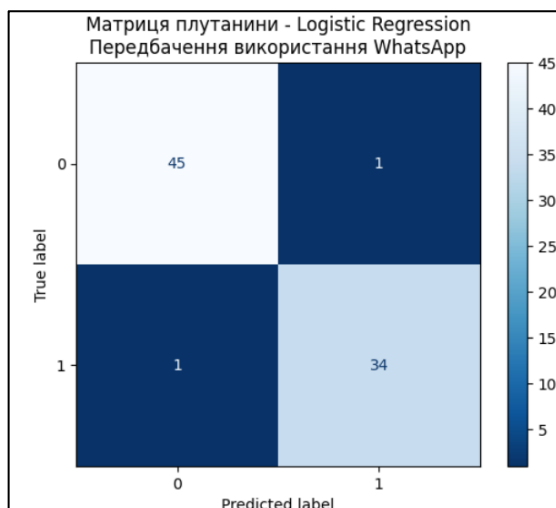


Рисунок Г.14 – Матриця плутанини результатів моделі Logistic Regression для цифрового каналу WhatsApp

```
# Матриця Random Forest
y_pred_rf = rf_best.predict(X_test)
acc_rf = accuracy_score(y_test, y_pred_rf)
plt.figure(figsize=(6, 5))
cm_rf = confusion_matrix(y_test, y_pred_rf)
ConfusionMatrixDisplay(cm_rf).plot(cmap='binary', values_format='d')
plt.title('Матриця плутанини - Random Forest\nПередбачення використання WhatsApp')
plt.show()
```

Рисунок Г.15 – Код матриці плутанини моделі Random Forest для цифрового каналу WhatsApp

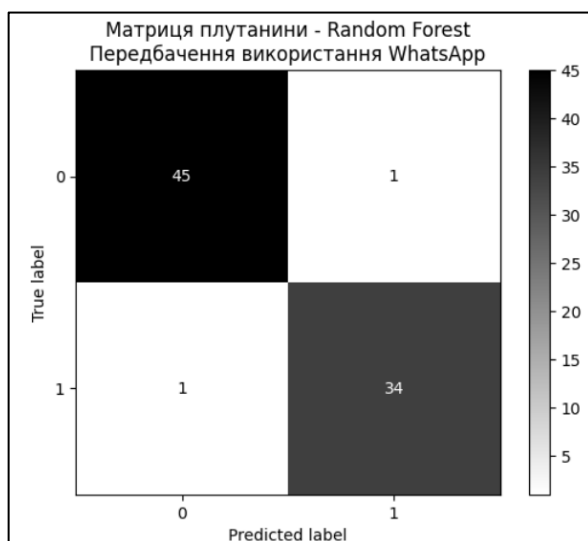


Рисунок Г.16 – Матриця плутанини результатів моделі Random Forest для цифрового каналу WhatsApp

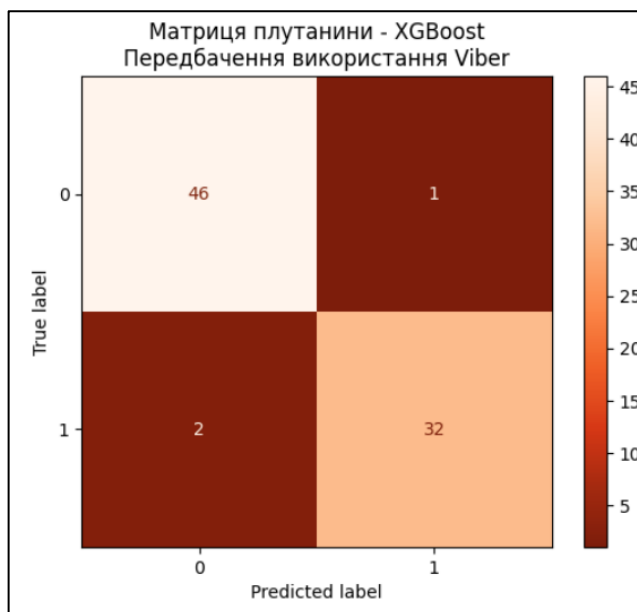


Рисунок Г.17 – Матриця плутанини результатів моделі XGBoost для цифрового каналу Viber

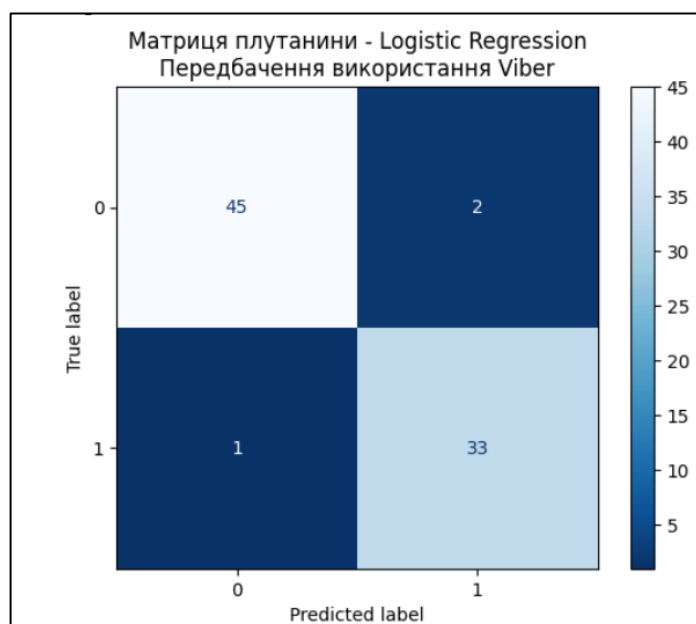


Рисунок Г.18 – Матриця плутанини результатів моделі Logistic Regression для цифрового каналу Viber

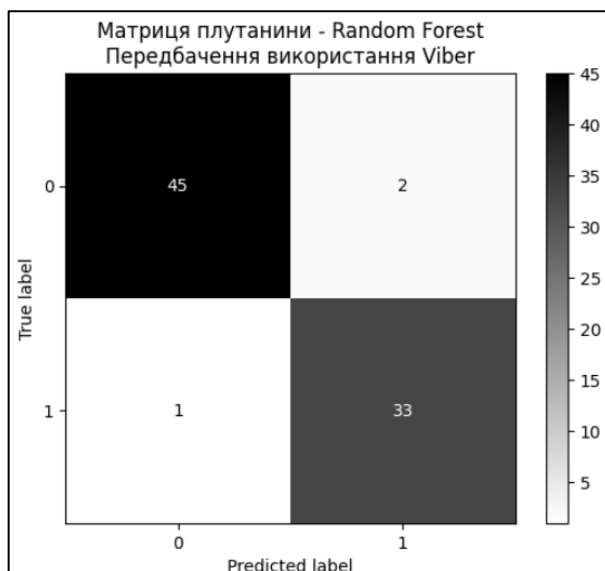


Рисунок Г.19 – Матриця плутанини результатів моделі Random Forest для цифрового каналу Viber

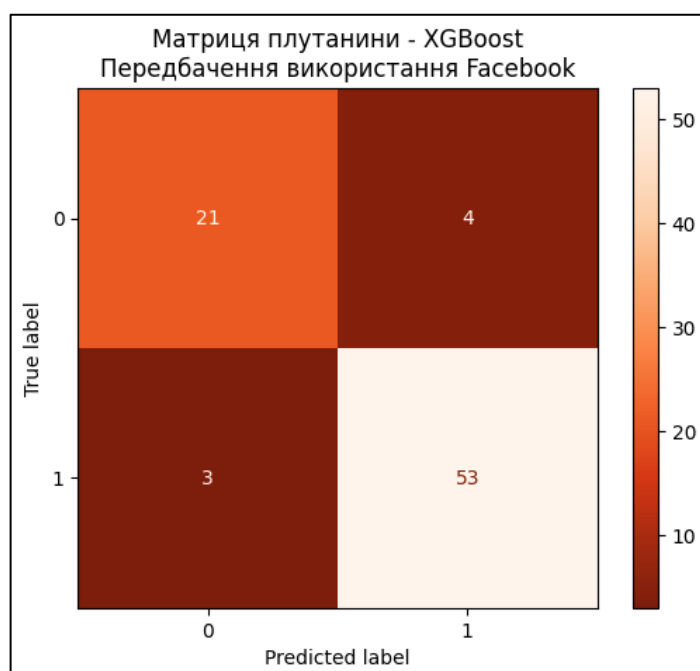


Рисунок Г.20 – Матриця плутанини результатів моделі XGBoost для цифрового каналу Facebook

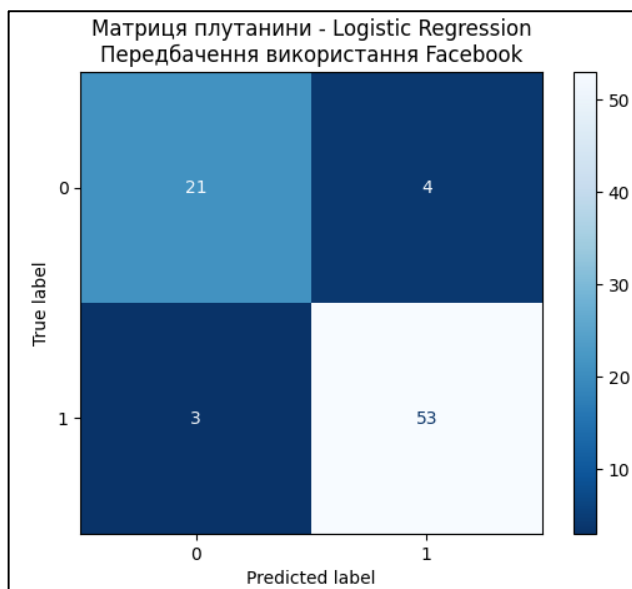


Рисунок Г.21 – Матриця плутанини результатів моделі Logistic Regression для цифрового каналу Facebook

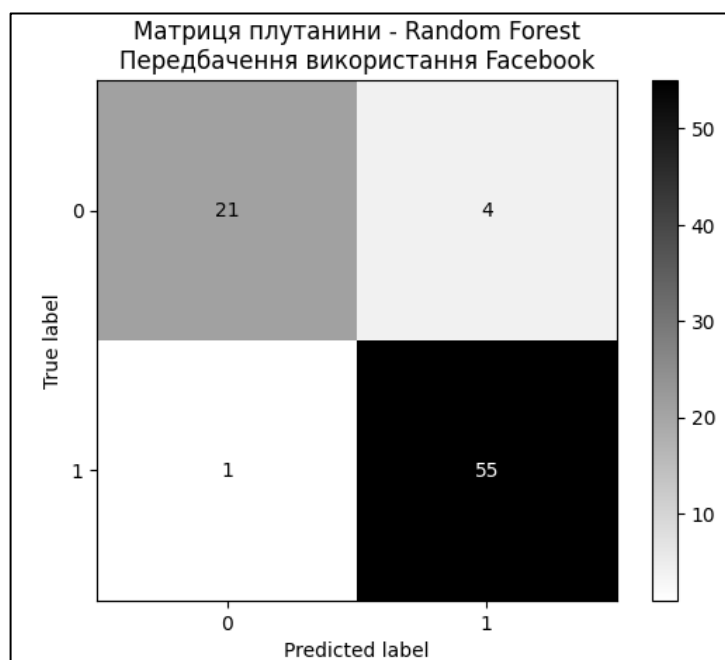


Рисунок Г.22 – Матриця плутанини результатів моделі Random Forest для цифрового каналу Facebook

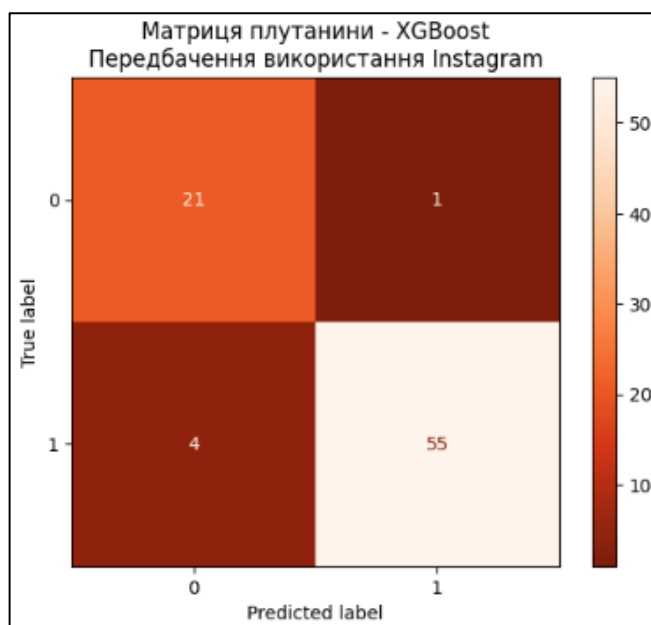


Рисунок Г.23 – Матриця плутанини результатів моделі XGBoost для цифрового каналу Instagram

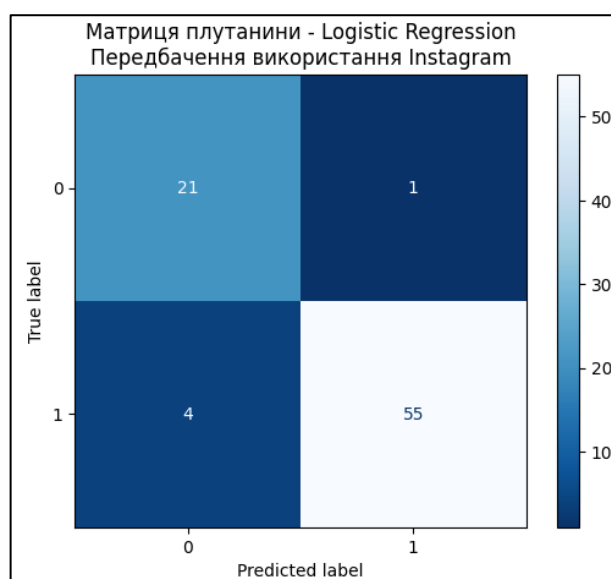


Рисунок Г.24 – Матриця плутанини результатів моделі Logistic Regression для цифрового каналу Instagram

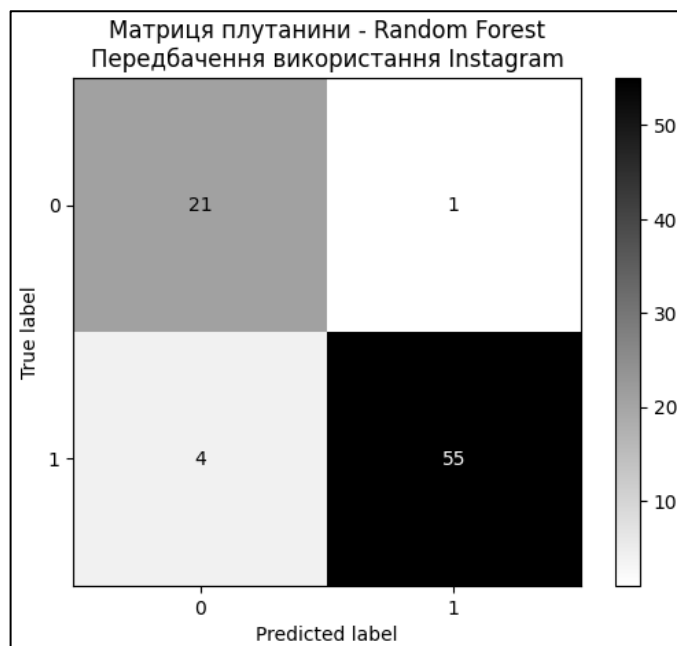


Рисунок Г.25 – Матриця плутанини результатів моделі Random Forest для цифрового каналу Instagram

```
<class 'pandas.core.frame.DataFrame'>
```

RangeIndex: 404 entries, 0 to 403

Data columns (total 46 columns):

#	Column	Non-Null Count	Dtype
0	num_male	404 non-null	int64
1	num_female	404 non-null	int64
2	age_young	404 non-null	int64
3	age_middle	404 non-null	int64
4	age_senior	404 non-null	int64
5	age_elderly	404 non-null	int64
6	edu_basic_higher	404 non-null	object
7	edu_incomplete_higher	404 non-null	int64
8	edu_full_higher	404 non-null	int64
9	edu_vocational	404 non-null	int64
10	edu_basic_secondary	404 non-null	int64
11	edu_academic_degree	404 non-null	int64
12	edu_only_primary	404 non-null	int64
13	edu_two_higher	404 non-null	int64
14	edu_full_secondary	404 non-null	int64
15	loc_vinnytsia_city	404 non-null	int64
16	loc_v_krushlyntsi	404 non-null	int64
17	loc_havryshivka	404 non-null	int64
18	loc_m_krushlyntsi	404 non-null	int64
19	loc_vinnytski_khutory	404 non-null	int64
20	loc_stadnytsia	404 non-null	int64
21	loc_desna	404 non-null	int64
22	loc_pysarivka	404 non-null	int64
23	loc_shchitky	404 non-null	int64
24	live_vin_1_3	404 non-null	int64
25	live_vin_5_10	404 non-null	int64
26	live_vin_always	404 non-null	int64
27	live_vin_10+	404 non-null	int64
28	live_vin_less_1	404 non-null	int64
29	work_have	403 non-null	float64
30	work_dont	404 non-null	int64
31	categories_of_the_population_idp	404 non-null	int64
32	categories_of_the_population_not_related	404 non-null	int64
33	categories_of_the_population_retired	404 non-null	int64
34	categories_of_the_population_handicapped_person	404 non-null	int64
35	categories_of_the_population_student	404 non-null	int64
36	channel_com_dig_email	404 non-null	int64
37	channel_com_dig_official_site	404 non-null	int64
38	channel_com_dig_internet	404 non-null	int64
39	channel_media_mass	404 non-null	int64
40	channel_media_social_ads	404 non-null	int64
41	channel_app_viber	404 non-null	int64
42	channel_app_instagram	404 non-null	int64
43	channel_app_telegram	404 non-null	int64

Рисунок Г.26 – Інформація про дані

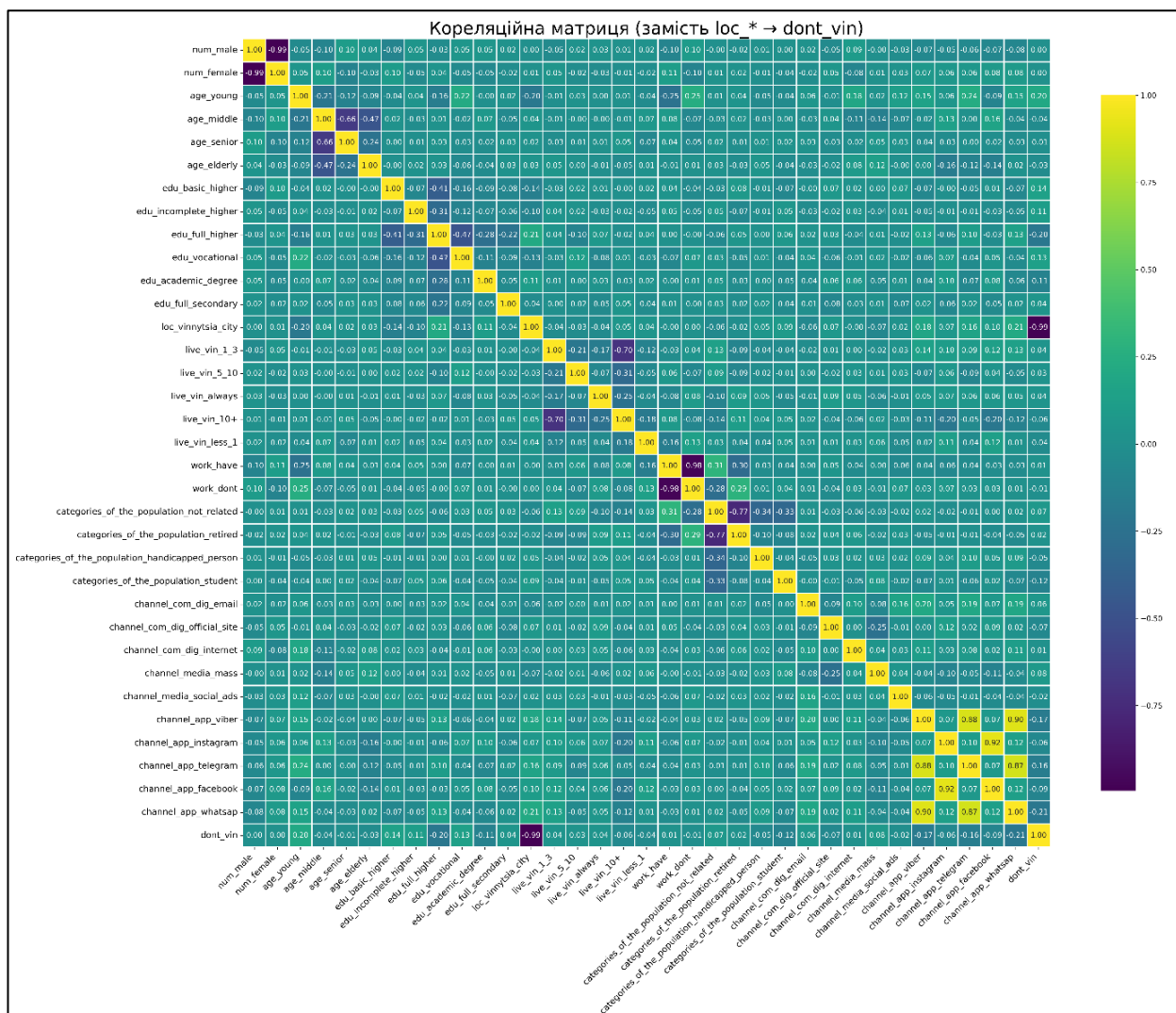


Рисунок Г.27 – Кореляційна матриця числових даних

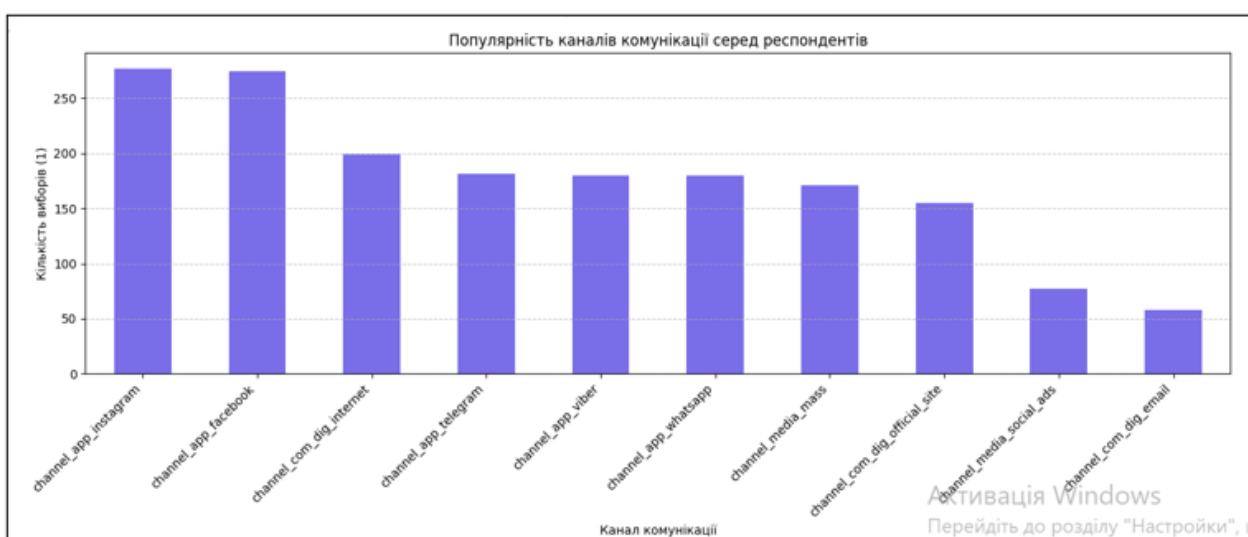


Рисунок Г.28 – Розподіл популярності цифрових каналів комунікації серед респондентів

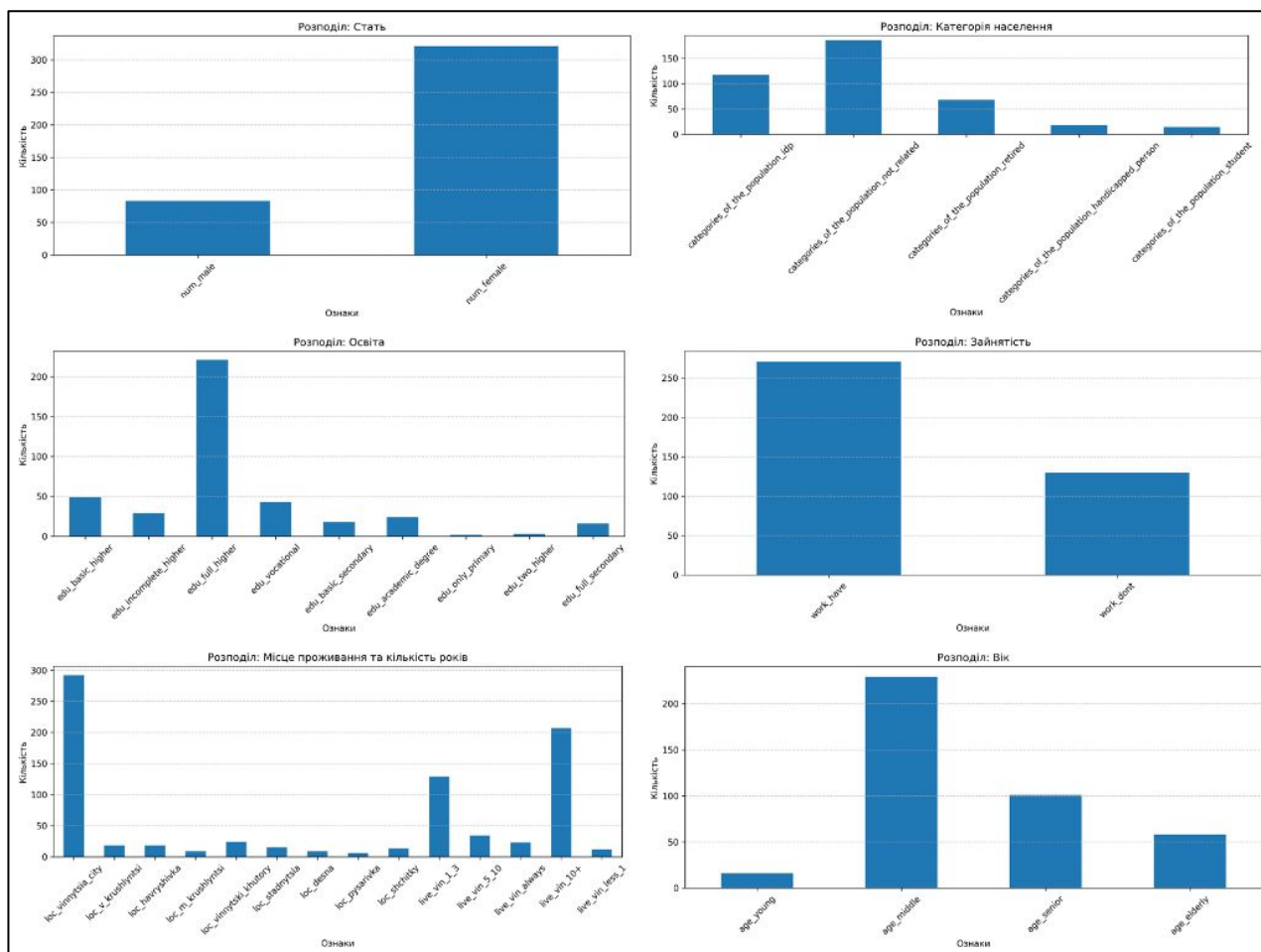


Рисунок Г.29 – Графік розподілу за ознаками «Стать», «Освіта», «Вік», «Місто проживання» та «Кількість років, прожитих у Вінницькій ОТГ», «Працевлаштування», «Категорія населення»

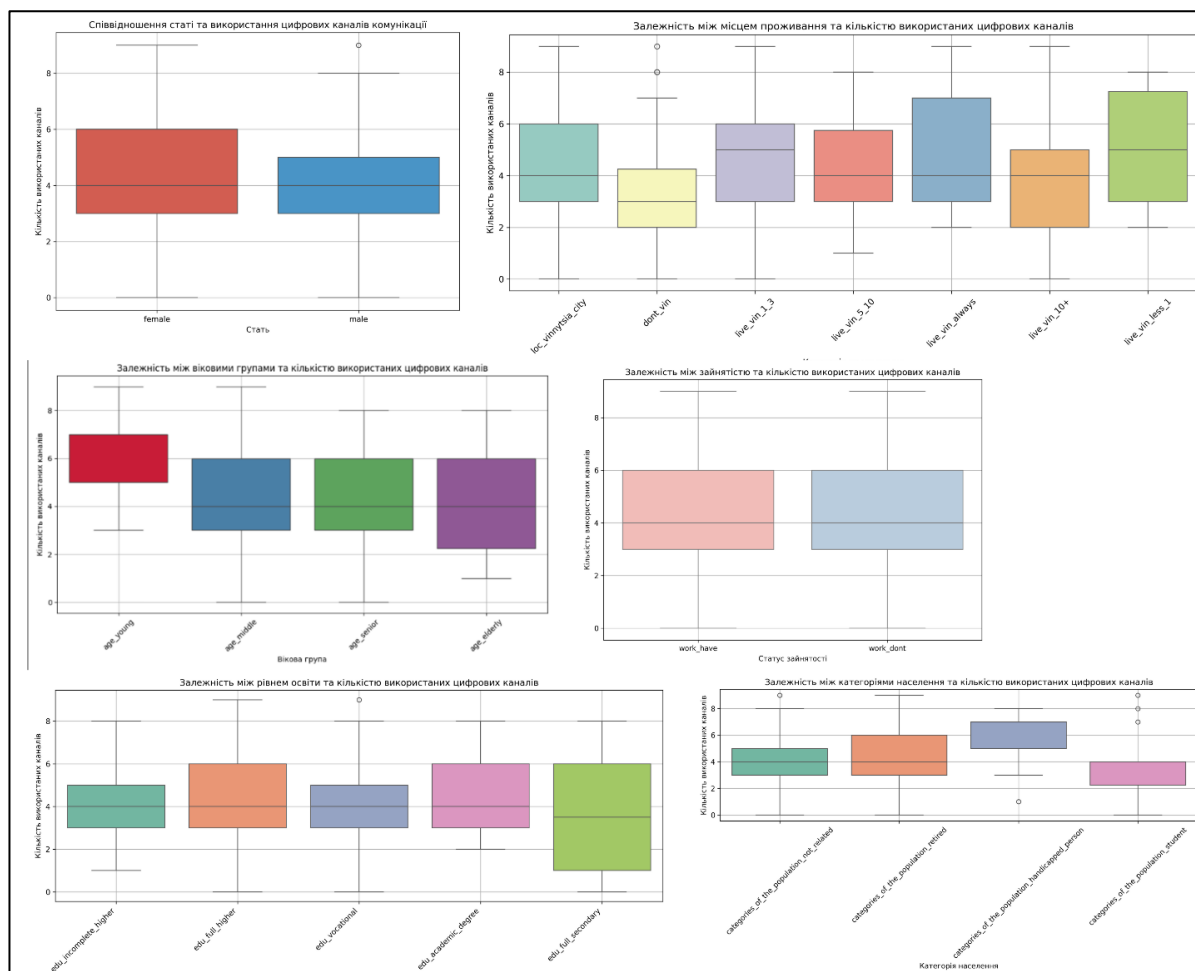


Рисунок Г.30 – Графіки «Співвідношення між статтю, освітою, віком, містом проживання та кількістю років, прожитих у Вінницькій ОТГ, працевлаштуванням, категорією населення та використанням цифрових каналів»

Channel	Model	CrossVal Mean	Train Accuracy	Test Accuracy
WhatsApp	Logistic Regression	0.954	0.963	0.963
WhatsApp	Random Forest	0.947	0.994	0.963
WhatsApp	XGBoost	0.938	0.991	0.963
Instagram	Logistic Regression	0.972	0.981	0.938
Instagram	Random Forest	0.966	0.997	0.938
Instagram	XGBoost	0.978	0.994	0.938
Facebook	Logistic Regression	0.969	0.985	0.914
Facebook	Random Forest	0.963	0.994	0.926
Facebook	XGBoost	0.969	0.991	0.914
Telegram	Logistic Regression	0.944	0.969	0.926
Telegram	Random Forest	0.938	1.0	0.938
Telegram	XGBoost	0.926	0.994	0.951
Viber	Logistic Regression	0.954	0.963	0.963
Viber	Random Forest	0.947	0.994	0.963
Viber	XGBoost	0.938	0.991	0.963

Рисунок Г.31 – Оцінка точності моделей