

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика»

**Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»**

Форма здобуття освіти: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

**на тему «Аналіз впливу соціально-демографічних факторів на рівень
заробітної плати»**
(назва теми)

здобувача Гришко Єва Миколаївни
(ПІБ)

(підпис)

Науковий керівник: кандидат економічних наук,
доцент Ольга ОСИПОВА

(підпис)

**Робота допущена до захисту перед екзаменаційною комісією з
атестації здобувачів вищої освіти (ЕК)**

Завідувач кафедри: кандидат фізико-математичних наук,
професор Галина ВЕЛИКОІВАНЕНКО

(підпис)

Київ 2025

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»**

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика»

Галузь знань 05 «Соціальні та поведінкові науки»
Спеціальність 051 «Економіка»

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми
Ольга ПРИТОМАНОВА

(підпис)

_____ 2025 р.

ЗАТВЕРДЖУЮ

Завідувач кафедри математичного
моделювання та статистики
Галина ВЕЛИКОІВАНЕНКО

(підпис)

_____ 2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувача вищої освіти

Гришко Єви Миколаївни

(прізвище, ім'я, по батькові)

_____ очної (денної) _____

форми здобуття освіти

очної (денної), заочної, дистанційної

на підготовку кваліфікаційної бакалаврської роботи

на тему

«Аналіз впливу соціально-демографічних факторів на рівень заробітної плати»

Тему затверджено наказом ректора Університету від «___» _____ 20__ р. № _____

Кваліфікаційна бакалаврська робота виконується на матеріалах

Сайту Корейського інституту охорони здоров'я та соціальних питань, Korea Institute for Health and Social Affairs (KIHASA) <https://www.koweps.re.kr:442/eng/main.do>

План кваліфікаційної бакалаврської роботи

Розділ 1	Теоретичні та методичні засади впливу соціально-демографічних факторів на рівень заробітної плати
Розділ 2	Регресійний аналіз впливу соціально-демографічних факторів на рівень заробітної плати

Об'єкт дослідження:	Процес формування та диференціації заробітної плати в залежності від соціально-демографічних характеристик населення
Предмет дослідження:	Статистичні зв'язки між такими характеристиками, як вік, стать, освіта, професія, регіон проживання тощо, і рівнем заробітної плати працівників
Мета виконання кваліфікаційної бакалаврської роботи:	Побудова економетричних моделей для кількісного оцінювання впливу соціально-демографічних факторів на заробітну плату, а також порівняння точності алгоритмів прогнозування

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1: Теоретичні підходи до визначення заробітної плати та факторів її формування. Аналіз соціально-демографічних характеристик, які можуть впливати на рівень доходів. Огляд наукових досліджень щодо впливу соціально-демографічних факторів на заробітну плату в Україні та світі. Огляд методів економетричного аналізу впливу соціально-демографічних факторів на рівень заробітної плати. Визначення основних змінних для аналізу (залежна змінна – рівень заробітної плати, незалежні змінні – вік, стать, рівень освіти, досвід роботи, регіон, сімейний стан тощо). Вибір економетричних моделей (множинна лінійна регресія, логарифмічна модель, моделі з категорійними змінними).

У розділі 2: Попередня обробка та візуалізація даних (побудова графіків, розподілу заробітної плати за категоріями). Оцінка параметрів економетричних моделей та їх інтерпретація. Аналіз отриманих результатів щодо впливу віку, освіти, статі та інших факторів на рівень заробітної плати. Формулювання висновків та розробка рекомендацій щодо подолання нерівностей в оплаті праці.

Завдання підготував
науковий керівник

Ольга ОСИПОВА

(підпис)

(ім'я, прізвище)

«__» _____ 20__ р.

Завдання одержав
здобувач

Єва ГРИШКО

(підпис)

(ім'я, прізвище)

«__» _____ 20__ р.

Реферат

Кваліфікаційна бакалаврська робота складається зі вступу, двох розділів, висновків та переліку використаних джерел, що містить 32 найменувань на 3 сторінках. Загальний обсяг роботи складає 56 сторінку друкованого тексту, з яких 53 сторінок основної частини. Робота містить 19 рисунків та 2 таблиць.

«Аналіз впливу соціально-демографічних факторів на рівень заробітної плати»

Об'єктом дослідження є процес формування та диференціації заробітної плати в залежності від соціально-демографічних характеристик населення.

Предметом дослідження є статистичні зв'язки між такими характеристиками, як вік, стать, освіта, професія, регіон проживання тощо, і рівнем заробітної плати працівників.

Мета та завдання виконання дослідження. Метою виконання кваліфікаційної бакалаврської роботи є побудова економетричних моделей для кількісного оцінювання впливу соціально-демографічних факторів на заробітну плату, а також порівняння точності алгоритмів прогнозування.

У межах дослідження були поставлені такі *завдання*:

- розкрити теоретичні засади формування заробітної плати;
- визначити та обґрунтувати ключові соціально-демографічні фактори впливу;
- здійснити очищення, стандартизацію і візуалізацію даних;
- побудувати множинну лінійну регресію, Ridge та Lasso-регресії;
- застосувати методи машинного навчання (Random Forest, KNN, Decision Tree);
- порівняти моделі за точністю та інтерпретувати важливість факторів.

Теоретична, методична та практична значущість отриманих результатів. Під час дослідження було систематизовано підходи до аналізу заробітної плати як економічної категорії, розглянуто вплив соціально-демографічних чинників на її рівень, а також обґрунтовано доцільність поєднання економетричних методів з алгоритмами машинного навчання. Побудовано низку моделей, що дозволяють кількісно оцінити взаємозв'язок між доходами працівників та такими змінними, як вік, стать, освіта, професія, сімейний стан тощо.

Практичне значення роботи полягає у можливості використання побудованих моделей для виявлення основних факторів, що впливають на заробітну плату, а також для прогнозування доходів працівників залежно від їхніх соціально-демографічних характеристик. Результати дослідження можуть бути використані для розробки рекомендацій щодо зменшення гендерного розриву в оплаті праці, подолання регіональної нерівності та удосконалення державної політики у сфері праці та соціального захисту.

Рік виконання кваліфікаційної бакалаврської роботи – 2025.

Рік захисту роботи – 2025.

Ключові слова: заробітна плата, соціально-демографічні фактори, регресія, машинне навчання, економетрична модель.

В і д г у к
про кваліфікаційну бакалаврську роботу
здобувача освітньо-професійної програми «Економічна кібернетика»
навчально-наукового інституту
«Інститут інформаційних технологій в економіці»
Гришко Єви Миколаївни
на тему «Аналіз впливу соціально-демографічних факторів на рівень заробітної плати»

1. Актуальність теми обумовлена потребою у глибшому розумінні механізмів диференціації заробітної плати у зв'язку з соціально-демографічними характеристиками працівників. Такий аналіз є важливим для формування ефективної політики оплати праці, боротьби з дискримінацією та забезпечення соціальної справедливості.

2. Позитивні риси кваліфікаційної бакалаврської роботи: здобувачкою здійснено огляд теоретичних підходів до вивчення структури заробітної плати, виконано повноцінний цикл економетричного аналізу, включаючи регресійне моделювання та використання алгоритмів машинного навчання. У роботі використано реальний масив даних соціологічного дослідження, проведено передобробку, візуалізацію даних та побудовано моделі для оцінки впливу факторів на рівень заробітної плати. Робота засвідчує володіння базовими інструментами економетрії та машинного навчання, а також вміння інтерпретувати отримані результати.

3. Наявність самостійних розробок автора: вибір теми, підбір джерельної бази, підготовка даних та побудова моделей здійснені самостійно. Структура, логіка та оформлення роботи загалом відповідають вимогам до кваліфікаційних бакалаврських досліджень.

4. Цінність теоретичних висновків та практичних рекомендацій полягає у виявленні факторів, що найбільш суттєво впливають на рівень заробітної плати. Результати роботи можуть бути використані як підґрунтя для подальших досліджень проблем оплати праці, а також для обґрунтування рішень у сфері соціально-економічної політики.

5. Наявність недоліків:

- побудовані моделі машинного навчання демонструють ознаки перенавчання (overfitting), що знижує їх прогностичну здатність. Авторці слід було налаштувати гіперпараметри моделей для забезпечення балансу між точністю та узагальненістю;
- терміни виконання окремих етапів дослідження періодично порушувалися, що ускладнювало наукове керівництво та супровід бакалаврської роботи.

6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: рівень виконання кваліфікаційної бакалаврської роботи Гришко Є.М. оцінюю як *позитивний*. Робота відповідає вимогам освітньої програми «Економічна кібернетика» та може бути допущена до захисту перед екзаменаційною комісією. За умови успішного захисту здобувачка заслуговує на присвоєння освітньої кваліфікації «Бакалавр з економіки».

Науковий керівник: доцент кафедри
математичного моделювання та статистики,

кандидат економічних наук, доцент _____ Ольга ОСИПОВА «____» _____ 2025 р.

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1	6
ТЕОРЕТИЧНІ ТА МЕТОДИЧНІ ЗАСАДИ ВПЛИВУ СОЦІАЛЬНО-ДЕМОГРАФІЧНИХ ФАКТОРІВ НА РІВЕНЬ ЗАРОБІТНОЇ ПЛАТИ	6
1.1 Сутність та підходи до визначення поняття заробітної плати	6
1.2 Вплив соціально-демографічних факторів на рівень заробітної плати: економічний та соціальний контексти	8
1.3 Методичні підходи до аналізу взаємозв'язку між соціально-демографічними характеристиками та оплатою праці	15
РОЗДІЛ 2	25
РЕГРЕСІЙНИЙ АНАЛІЗ ВПЛИВУ СОЦІАЛЬНО-ДЕМОГРАФІЧНИХ ФАКТОРІВ НА РІВЕНЬ ЗАРОБІТНОЇ ПЛАТИ	25
2.1 Підготовка та візуальний аналіз основних соціально-демографічних факторів.....	25
2.2. Побудова та інтерпретація економетричної моделі залежності заробітної плати від соціально-демографічних факторів.....	43
2.3 Удосконалення економетричних моделей із застосуванням методів машинного навчання та регуляризації.....	47
ВИСНОВКИ	53
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	56

ВСТУП

Актуальність теми: рівень заробітної плати є одним з найважливіших соціально-економічних показників, що визначає добробут населення, рівень трудової мотивації та соціальну стабільність у суспільстві. В умовах глобалізації, регіональної нерівності та змін на ринку праці особливої актуальності набуває вивчення того, як соціально-демографічні чинники (вік, стать, освіта, сімейний стан тощо) впливають на формування доходів громадян. Зважаючи на сучасні виклики в економіці високий рівень тіньової зайнятості, гендерний розрив в оплаті праці, регіональні диспропорції, обрана тема є надзвичайно актуальною та має важливе прикладне значення для політики зайнятості, соціального забезпечення та розвитку людського капіталу.

Аналіз останніх досліджень і публікацій: проблематика заробітної плати та факторів, що впливають на її рівень, розглядалася у працях класичних економістів, зокрема А. Сміта, Д. Рікардо. У сучасному економічному дискурсі вагомий внесок зроблено такими вітчизняними науковцями, як Колот А.М., Грішнова О.А., Мочерний С.В., Калина А.В., які досліджували функції заробітної плати, її мотиваційний потенціал та соціальні аспекти. Також використано працю закордонного автору Wooldridge J. M для вивчення використання методів регресійних моделей та їх подальшої оцінки. Незважаючи на наявність значної кількості праць, недостатньо досліджено інтеграцію методів машинного навчання в аналіз впливу соціально-демографічних змінних на заробітну плату, що і зумовило вибір теми даної роботи.

Мета і завдання дослідження: мета кваліфікаційної роботи полягає в побудові економетричних моделей, що дозволяють виявити та кількісно оцінити вплив соціально-демографічних характеристик на рівень заробітної плати, а також здійснити порівняльну оцінку якості моделей.

Завдання дослідження:

- Розкрити теоретичну суть заробітної плати та її функцій у соціально-економічній системі.
- Систематизувати основні соціально-демографічні чинники, що впливають на оплату праці.
- Здійснити первинну обробку даних (очистку, стандартизацію, візуалізацію).
- Побудувати класичну модель множинної лінійної регресії.
- Оцінити ефективність регуляризованих моделей Ridge та Lasso.
- Реалізувати і проаналізувати моделі машинного навчання (Random Forest, KNN, Decision Tree).
- Визначити найвпливовіші змінні, що впливають на заробітну плату, та сформулювати практичні рекомендації.

Об'єкт дослідження: процеси формування та диференціації заробітної плати в залежності від соціально-демографічних характеристик населення.

Предмет дослідження: статистичні зв'язки між окремими соціально-демографічними характеристиками (вік, освіта, стать, професія, регіон тощо) та рівнем заробітної плати працівників.

Методи дослідження

У роботі використано:

- методи економетричного моделювання (лінійна регресія, Ridge, Lasso),
- методи машинного навчання (Random Forest, Decision Tree, K-Nearest Neighbors),
- методи стандартизації та видалення викидів (Z-score),
- візуалізація даних (гістограми, коробкові діаграми, теплові карти кореляцій),
- оцінювання якості моделей за метриками R^2 та RMSE.

Теоретична, методична та практична значущість отриманих результатів

Теоретична значущість роботи полягає в уточненні впливу соціально-демографічних характеристик на рівень оплати праці, а також у поєднанні класичних економетричних методів із сучасними алгоритмами машинного навчання в контексті аналізу доходів.

Методична значущість полягає у формуванні алгоритму побудови прогностичних моделей, що може бути використаний в інших дослідженнях ринку праці та соціальної політики.

Практична значущість роботи полягає в можливості використання моделей для оцінювання впливу окремих факторів на дохід працівників, прогнозування змін в оплаті праці, обґрунтування соціально-економічних рішень органами державного управління та підприємствами.

Інформаційна база дослідження

У якості інформаційної бази використано:

- набір реальних даних соціологічного опитування за 2005–2018 рр.,
- статистичні матеріали Євростату, Бюро статистики праці США, OECD,
- законодавчі акти України щодо праці та оплати праці,
- наукові праці вітчизняних та зарубіжних дослідників з економіки праці та економетрики.

Структура роботи: кваліфікаційна робота складається зі вступу, двох розділів, висновків та списку використаних джерел. Перший розділ присвячено теоретичним і методичним засадам дослідження, другий – побудові та оцінці економетричних моделей впливу соціально-демографічних факторів на заробітну плату.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ТА МЕТОДИЧНІ ЗАСАДИ ВПЛИВУ СОЦІАЛЬНО-ДЕМОГРАФІЧНИХ ФАКТОРІВ НА РІВЕНЬ ЗАРОБІТНОЇ ПЛАТИ

1.1 Сутність та підходи до визначення поняття заробітної плати

Заробітна плата відіграє ключову роль у соціально-економічному розвитку суспільства, оскільки є основним джерелом доходу для працівників та важливим елементом витрат для підприємств. Вона визначає рівень життя населення та впливає на економічну активність країни.

Згідно із законодавством України, «заробітна плата – це винагорода, обчислена, як правило, у грошовому виразі, яку за трудовим договором роботодавець виплачує працівникові за виконану ним роботу. Її розмір залежить від складності та умов виконуваної роботи, професійно-ділових якостей працівника, результатів його праці та господарської діяльності підприємства [1]».

Відомі економісти, зокрема А. Сміт і Д. Рікардо, по-різному трактували це поняття. А. Сміт вважав, що заробітна плата включає вартість засобів для існування, необхідних для підтримання працездатності людини. Д. Рікардо, своєю чергою, визначав заробітну плату як грошовий вираз мінімально необхідних засобів для життя. В. Лукашевич вважає, що терміни «заробітна плата» та «оплата праці» є рівнозначними, хоча між ними є певна різниця: заробітну плату виплачують із відповідного фонду, тоді як додаткові заохочувальні виплати можуть здійснюватися за рахунок інших фінансових ресурсів підприємства. Маркевич А.Є. розглядає заробітну плату через її функції у процесі суспільного відтворення, наголошуючи, що в науковій літературі існують різні підходи до їхнього визначення. Водночас

Чанишева Г. та Болотіна Н. зазначають, що термін «оплата праці» є ширшим і охоплює організацію, регулювання та систему правових норм у цій сфері, тоді як «заробітна плата» стосується безпосередньо права працівника на отримання грошової винагороди [2].

Існують основні 4 підходи для визначення поняття заробітна плата. З позиції вартісного підходу, заробітна плата розглядається з позиції ціни праці, яку отримує працівник. У межах ринкового підходу вона розглядається як ціна, за якою найманий працівник продає свою робочу силу, тобто вона виражає ринкову вартість використання найманої робочої сили.

У межах вартісного підходу заробітна плата розглядається як вартість або ціна праці, яку працівник отримує за виконання своїх обов'язків. Основний акцент робиться на грошовому вимірі праці, тобто оплата розглядається як результат оцінювання затрат робочої сили. Такий підхід зосереджується на еквівалентності праці та винагороди, підкреслюючи роль зарплати як індикатора економічної цінності праці на ринку. До представників цього підходу належать Брезицька К.Ф., Бутинець Ф.Ф. та Дубовська О.В. [3].

Згідно з ринковим підходом, заробітна плата є елементом ринку праці, а точніше — ціною, за якою працівник «продає» свою робочу силу роботодавцю. Це визначення підкреслює ринковий механізм формування заробітної плати, де попит і пропозиція на кваліфіковану робочу силу визначають рівень оплати праці. Отже, заробітна плата тут постає як результат економічної взаємодії між роботодавцем і працівником. Основними представниками цього підходу є Колот А.М. та Жидовська Н.М. [4].

У межах мотиваційного підходу заробітна плата трактується не лише як витрата виробництва, а передусім як ключовий стимул, який спонукає працівників до досягнення високих результатів. Вона розглядається як інструмент матеріального заохочення, який впливає на ефективність, ініціативність і продуктивність персоналу.

Мотиваційний підхід акцентує увагу на внутрішній функції зарплати – мотивації до праці. До представників цієї концепції належать Калина А.В., Мочерний С.В., Пашута Н.І., Швець Л.П. та Яременко Л.М. [5].

Юридичне трактування заробітної плати базується на нормативно-правових засадах, зокрема на визначеннях, поданих у законодавстві. Заробітна плата тут розглядається як грошова винагорода, яку власник або уповноважений орган виплачує працівникові відповідно до трудового договору за виконану роботу. Такий підхід фокусується на формальних ознаках трудових відносин, правилах нарахування та обов'язковості дотримання трудових норм. До науковців, які підтримують цей підхід, належать Васильчак С.В., Золотогоров В.Г., Петрова І.Л. та Усач В.Б. [3].

1.2 Вплив соціально-демографічних факторів на рівень заробітної плати: економічний та соціальний контексти

Заробітна плата виконує кілька ключових функцій, які впливають як на працівників, так і на роботодавців, формуючи економічні відносини в суспільстві. У сучасних умовах функції заробітної плати виходять далеко за межі простої грошової винагороди за виконану роботу. Різні науковці виокремлюють різну кількість функцій заробітної плати:

С.В. Мочерний класифікує функції на три основні: відтворювальну, стимулюючу та розподільчу [6]. О.І. Гадзевич, розширюючи попередню класифікацію, виділяє також соціальну функцію, яка пов'язана з роллю заробітної плати у забезпеченні гідного рівня життя, соціального захисту та стабільності працівників у суспільстві [7, с. 12].

Такі науковці як Калина А.В. [8, с. 164] та Завіновська Г.Т. [9, с. 149] розглядають чотири основні функції заробітної плати – стимулюючу, відтворювальну, соціальну та регулюючу.

О.А. Грішнова пропонує ще ширше трактування, виділяючи шість функцій: стимулююча, відтворювальна, соціальна, регулююча – встановлює співвідношення між попитом і пропозицією праці, впливає на міграційні процеси, структуру зайнятості та оптимізаційна – дозволяє узгодити інтереси працівників і роботодавців через ефективну систему мотивації [10]. Отже зміст функцій заробітної плати.

Відтворювальна функція: Заробітна плата є основним джерелом доходу для працівників, що дозволяє їм підтримувати свою фізичну та розумову працездатність. Вона покриває витрати на харчування, житло, освіту, медичне обслуговування та інші необхідні потреби. Оскільки робоча сила є особливим видом товару, який вимагає постійного відновлення, рівень заробітної плати має відповідати вартості її відтворення. Це означає, що мінімальна заробітна плата повинна забезпечувати гідний рівень життя, щоб працівник міг ефективно виконувати свої професійні обов'язки [11, с. 511].

Мотиваційна функція: Матеріальна зацікавленість є основою ефективного використання робочої сили. Працівники, які мають високу оплату праці, зазвичай демонструють вищу продуктивність, краще дотримуються трудової дисципліни та охочіше підвищують свою кваліфікацію. Таким чином, заробітна плата виступає стимулом до професійного розвитку, освоєння нових технологій і зростання загальної ефективності праці [12].

Регулювальна функція: заробітна плата відіграє важливу роль у ринку праці, впливаючи на співвідношення між попитом і пропозицією робочої сили. Вона визначає рівень зайнятості, формує структуру персоналу підприємств, а також впливає на міграційні процеси. Наприклад, високі зарплати в певних галузях можуть залучати більше спеціалістів, тоді як низькі – спричиняти нестачу кваліфікованих

кадрів. Регулювальна функція також впливає на диференціацію зарплат у різних регіонах та секторах економіки, зокрема між виробничою та бюджетною сферою обов'язки [11, с. 512].

Соціальна функція: Заробітна плата виконує важливу соціальну роль, оскільки вона є показником суспільного визнання трудових заслуг працівника. Справедлива система оплати праці сприяє зміцненню колективної єдності, вихованню відповідального ставлення до роботи та зменшенню соціальної напруги. Рівність оплати за однакову працю, незалежно від статі, віку чи інших характеристик, реалізує принцип соціальної справедливості [12].

Заробітна плата є важливим економічним показником, який відображає баланс між інтересами працівників і роботодавців. Для перших вона є основним джерелом доходу, а для других – елементом витрат на виробництво та фактором конкурентоспроможності підприємства. Підвищення зарплат може стимулювати покращення ефективності виробництва, впровадження нових технологій і розвиток бізнесу. Водночас неправильний розподіл зарплат може призводити до соціального невдоволення та економічних дисбалансів. Таким чином, заробітна плата виконує не лише економічну, а й соціальну роль, впливаючи на розвиток суспільства та якість життя населення [11, с. 512].

Оптимізаційна функція: сутність оптимізаційної функції полягає в тому, що заробітна плата як один із ключових елементів собівартості продукції стає важливим чинником для прийняття управлінських рішень. Роботодавець змушений постійно шукати оптимальні шляхи підвищення ефективності виробництва, зменшення витрат та максимізації прибутку. Це може включати такі заходи, як автоматизація виробничих процесів, впровадження нових технологій, скорочення надлишкових робочих місць або перенесення виробництва до країн із дешевшою робочою силою. Водночас, оптимізація не обов'язково означає зменшення зарплат чи звільнення працівників – вона може реалізовуватися шляхом підвищення кваліфікації персоналу,

покращення умов праці та впровадження прогресивних систем оплати праці [11, с. 513].

Заробітна плата має комплексну природу та є не лише джерелом доходу для працівників, а й важливим економічним інструментом, що впливає на розвиток підприємств, конкурентоспроможність національної економіки та рівень життя населення.

Рівень заробітної плати формується під впливом багатьох чинників, які взаємодіють між собою та відображають як індивідуальні характеристики працівника, так і структурні особливості ринку праці. Нижче наведено основні фактори, що впливають на диференціацію заробітної плати, з відповідними джерелами для кожного з них.

Освітній рівень є одним із найважливіших факторів, що впливають на диференціацію заробітної плати. Загалом, люди з вищою освітою заробляють більше, ніж ті, хто має нижчий рівень освіти. Це можна пояснити кількома причинами, такими як набуття спеціалізованих навичок і знань, що високо цінуються на ринку праці, кращі можливості для працевлаштування та доступ до високооплачуваних посад. Крім того, освіта забезпечує людям більшу впевненість і стабільність, що може спонукати їх до переговорів щодо вищої заробітної плати або пошуку більш оплачуваної роботи [13]. Прикладом є дослідження Бюро статистики праці США щодо середнього тижневого заробітку для громадян США, що показує суттєве збільшення заробітної оплати праці з підвищенням рівня освіти [14].

Досвід роботи є ще одним фактором, що впливає на диференціацію заробітної плати на ринку праці. Загалом, працівники з більшим досвідом мають вищу заробітну плату, ніж менш досвідчені працівники. Це може бути пов'язано з підвищеною продуктивністю, кращими показниками роботи та покращеними навичками вирішення проблем. Крім того, досвід часто забезпечує більшу стабільність роботи та можливості кар'єрного зростання. Проте взаємозв'язок між досвідом і заробітною

платою може відрізнятись залежно від галузі, розміру компанії та вимог конкретної посади. Наприклад, початкові посади можуть мати низькі зарплатні очікування незалежно від досвіду, тоді як високо спеціалізовані ролі можуть вимагати значного досвіду для отримання високої зарплати [15].

Дослідження підтверджують, що досвід роботи сприяє підвищенню заробітної плати, особливо в перші роки після завершення освіти. Наприклад, дослідження виявили, що досвід, набутий під час професійної освіти, призводить до значного збільшення заробітної плати на 7% до 19% через рік після закінчення вищої освіти [15].

Також стать має свій вплив на заробітну плату. Жінки, як правило, заробляють менше, ніж чоловіки, у більшості країн світу. Це можна пояснити такими факторами, як дискримінація, обмежений доступ до освіти та можливостей навчання, а також культурні норми, що віддають перевагу традиційним гендерним ролям. Прикладом є дослідження гендерного розриву у зарплаті в ЄС. У 2021 році гендерний розрив у погодинній оплаті праці в ЄС становив 12,7%, що означає, що жінки заробляють у середньому на 13% менше за чоловіків. Загальний розрив у доходах у 2018 році складав 36,7%, враховуючи відмінності в середній зарплаті, кількості оплачуваних годин та рівні зайнятості. У 2022 році рівень зайнятості жінок у ЄС був 69,3%, тоді як у чоловіків – 80% [16].

Розрив у зарплаті зумовлений не лише дискримінацією, а й структурними нерівностями. Сегрегація за секторами: 24% розриву пояснюється тим, що жінки переважно працюють у низькооплачуваних сферах (догляд, освіта, медицина). Нерівномірний розподіл оплачуваної та неоплачуваної праці: жінки виконують більше неоплачуваної роботи (догляд за дітьми, домашні справи), що обмежує їх кар'єрні можливості. "Скляна стеля": жінки рідше обіймають керівні посади (лише 10% топ-менеджерів – жінки). У середньому жінки-керівники заробляють на 23%

менше за чоловіків. Дискримінація в оплаті існує попри закріплений у ЄС принцип рівної оплати і жінки подекуди отримують менше за аналогічну роботу [16].

Також етнічна приналежність, може позначатися на доходах людини. Дослідження показали, що представники афроамериканської та латиноамериканської спільнот зазвичай заробляють менше, ніж їхні колеги іншої етнічності, виконуючи ту саму роботу. Це часто пояснюється історичними та системними бар'єрами, такими як дискримінація під час прийому на роботу, нерівний доступ до професійного навчання та обмежені можливості кар'єрного зростання. Крім того, у певних галузях можуть спостерігатися значні розбіжності у рівні заробітної плати залежно від етнічної приналежності. Вирішення цих нерівностей потребує державних ініціатив, таких як закони про рівні можливості працевлаштування та інвестиції в освіту і професійне навчання [17].

Професія є одним із ключових чинників диференціації заробітної плати. Зарплати працівників у різних професіях можуть суттєво відрізнятися через такі фактори, як рівень необхідних навичок, стабільність роботи та попит на ринку праці. Крім того, у деяких професіях заробітна плата може залишатися незмінною або навіть знижуватися з часом, тоді як в інших є більше можливостей для зростання. Наприклад, спеціалісти у високо затребуваних сферах, таких як технології чи медицина, зазвичай отримують вищу зарплату, ніж працівники, зайняті в малокваліфікованих галузях, таких як роздрібна торгівля або громадське харчування.

Галузь, у якій працює людина, суттєво впливає на її рівень доходів. Наприклад, деякі галузі вимагають високого рівня освіти або спеціальної підготовки, що призводить до вищих зарплат. Попит на робочу силу в певній галузі також може впливати на рівень заробітної плати, а ринкові умови та конкуренція між компаніями можуть визначати розмір оплати праці.

Також існує розрив в оплаті навіть в тих самих професіях та галузях. За дослідженням Бюро статистики праці США такі розриви в заробітній платі є в усіх

галузях, а найбільше в тих, що пов'язані з мистецтвом, розвагами та спортом, наприклад актори чи спортсмени [18].

У великих містах заробітні плати зазвичай вищі, ніж у сільській місцевості або передмістях. Це пояснюється вищою вартістю життя, більшою концентрацією бізнесів та можливостей для працевлаштування. Також є значні регіональні відмінності в оплаті праці. Це пов'язано з економічним розвитком регіону, наявністю певних галузей промисловості та попитом на конкретні професії. Наприклад, містах з високим рівнем життя пропонують вищі заробітні плати, але й вартість життя там відповідно вища. У той же час, у містах з нижчим рівнем життя зарплати нижчі.

Рівень заробітної плати також значно відрізняється між країнами. У 2023 році середня вартість години праці становила 31,8 євро в ЄС, коливаючись від 9,3 євро в Болгарії до 53,9 євро в Люксембурзі. Найбільш високооплачуваними ринками праці у 2023 році стали Данія, Люксембург, Ірландія та Бельгія. Ці відмінності можуть бути пов'язані з економічним розвитком, продуктивністю праці та іншими факторами [19].

Розмір компанії також може впливати на рівень заробітної плати. Великі компанії, як правило, пропонують вищі зарплати, оскільки мають більше ресурсів та можливостей для кар'єрного зростання. Водночас невеликі компанії можуть пропонувати конкурентні зарплати через меншу кількість ієрархічних рівнів або можливість більш гнучких умов роботи.

Стан ринку праці, зокрема рівень безробіття та кількість вакансій, безпосередньо впливає на рівень заробітної плати. У періоди високого безробіття конкуренція за робочі місця зростає, що може призводити до зниження зарплат. Навпаки, за низького рівня безробіття роботодавці змушені пропонувати вищу оплату, щоб залучити та утримати працівників. Таким чином, динаміка ринку праці визначає переговорну силу працівників та роботодавців у встановленні рівня оплати праці [20].

Заробітна плата також визначається співвідношенням попиту та пропозиції на робочу силу. Коли попит на певні професії перевищує пропозицію, заробітна плата

зростає, оскільки роботодавці змагаються за обмежену кількість кваліфікованих працівників. Навпаки, надлишок робочої сили може призводити до зниження зарплат. Цей механізм особливо помітний у галузях з високим попитом на спеціалістів, таких як ІТ чи медицина [21].

Економічні показники, такі як валовий внутрішній продукт (ВВП), рівень інфляції, безробіття, процентні ставки та фондові ринки, можуть впливати на рівень заробітної плати. Наприклад, економічний підйом може сприяти підвищенню зарплат, тоді як кризи чи рецесії можуть призводити до їхнього скорочення [22].

1.3 Методичні підходи до аналізу взаємозв'язку між соціально-демографічними характеристиками та оплатою праці

Аналіз взаємозв'язку між соціально-демографічними характеристиками (зокрема такими, як вік, стать, освіта, досвід тощо) та рівнем заробітної плати ґрунтується на застосуванні сучасних економетричних інструментів. Економетрика, як міждисциплінарна наука, що поєднує математичну статистику та економічну теорію, спрямована на кількісне вимірювання економічних взаємозв'язків, перевірку гіпотез і оцінку ефективності соціально-економічної політики.

Одним із базових підходів, який використовується для аналізу впливу факторів на оплату праці, є регресійний аналіз. Його застосування дозволяє не лише встановити напрям і силу впливу окремих змінних, а й виявити наявність статистично значущих залежностей у спостережуваних даних. Важливо відзначити, що економетрика здебільшого працює з неекспериментальними, тобто спостережуваними даними, отриманими на основі реальних статистичних спостережень, що ускладнює аналіз

через можливі зворотні зв'язки, ефекти змішаного впливу та інші статистичні виклики [23, с. 20].

Перш ніж розпочати побудову економетричних моделей, вкрай важливим етапом є попередній візуальний аналіз даних. Його проведення дозволяє глибше зрозуміти структуру масиву спостережень, виявити внутрішні закономірності, перевірити наявність аномалій, а також оцінити доцільність тих чи інших припущень щодо форми розподілу змінних, які використовуються у моделюванні. Візуалізація надає змогу для кращого розуміння характеру даних, що є ключовим для правильної постановки моделі.

Одним із базових інструментів є гістограми, які дозволяють оцінити розподіл значень змінної. Гістограма відображає частоту (або щільність) появи значень у певних інтервалах, що дає змогу побачити симетрію або скіс розподілу, наявність мод (пік частот), а також довгі хвости чи інші відхилення від нормального розподілу. Наприклад, у випадку розподілу заробітної плати гістограма може продемонструвати асиметричність вправо, що характерно для економічних змінних із високою варіативністю та наявністю великого числа спостережень із низькими доходами й одиничними випадками дуже високих значень. Така інформація є вирішальною для вибору трансформацій змінних (логарифмування, стандартизація) або обґрунтування необхідності використання моделей, стійких до викидів.

Іншим важливим інструментом є boxplot-діаграма (або коробкова діаграма), яка дозволяє одночасно побачити медіану, квартилі, міжквартильний розмах та викиди. Вона особливо корисна при порівнянні числових змінних у розрізі категоріальних груп (наприклад, рівень доходу серед чоловіків і жінок, або серед респондентів із різною освітою). Boxplot не лише показує центральну тенденцію, але й наочно демонструє асиметрію розподілу, наявність екстремальних значень (outliers), а також відмінності в розсіянні між групами. У нашому контексті така візуалізація

дозволяє, наприклад, оцінити, чи справді існує гендерний розрив в оплаті праці, або наскільки дохід корелює з рівнем освіти.

Не менш важливою є візуалізація взаємозв'язків між змінними. Для цього застосовуються кореляційні матриці (heatmaps). Кореляційна матриця дозволяє оцінити ступінь лінійної залежності між усіма парами числових змінних, що є необхідним для попередження проблем мультиколінеарності у регресійних моделях.

Загалом, попередній візуальний аналіз виконує функцію "діагностики даних". Він допомагає виявити недоліки або аномалії у даних (наприклад, порожні значення, нетипові спостереження); надає підстави для обґрунтованого вибору типу моделі (лінійна, дерево рішень тощо); формує гіпотези, які можуть бути перевірені кількісно на подальших етапах.

Кореляція – це статистичний показник, що відображає ступінь і напрямок зв'язку між двома кількісними змінними. Якщо значення однієї змінної зростає (чи зменшується) одночасно зі зростанням (чи зменшенням) іншої змінної, то між ними існує кореляція. Найпоширенішим є коефіцієнт Пірсона, значення якого варіюється в межах від -1 до +1:

- значення +1 вказує на ідеальну пряму лінійну залежність;
- значення -1 вказує на ідеальну обернену лінійну залежність;
- значення 0 означає відсутність лінійної кореляції.

Кореляція не обов'язково вказує на причинно-наслідковий зв'язок, а лише на те, що зміни однієї змінної систематично пов'язані зі змінами іншої.

- У контексті моделювання заробітної плати кореляційний аналіз дозволяє:
- оцінити попередні зв'язки між змінними (наприклад, між освітою та доходом);
- виявити змінні, які можуть бути найперспективнішими предикторами;

- виявити надмірно скорельовані змінні, що може вплинути на стабільність моделі [23].

Автокореляція – це явище, коли значення залишків (похибок) у регресійній моделі є залежними між собою, особливо коли спостереження впорядковані в часі або просторі. Інакше кажучи, автокореляція означає, що помилки моделі мають внутрішню структуру, наприклад, залишки у поточному періоді залежать від залишків у попередньому.

Найчастіше автокореляція виникає в часових рядах, але може з'явитися і в перехресних даних, якщо є просторові зв'язки або ефекти згрупованих одиниць.

Наявність автокореляції порушує одне з ключових припущень класичної лінійної регресії (незалежність похибок) і може призводити до:

- заниження стандартних помилок;
- переоцінки статистичної значущості змінних;
- викривлення довірчих інтервалів та тестів.

Для виявлення автокореляції зазвичай використовується тест Дарбіна-Вотсона (Durbin-Watson test). Його значення близьке до 2 означає відсутність автокореляції; значення менше 1 або більше 3 свідчать про наявність позитивної або негативної автокореляції відповідно [23].

Переходячи до побудови регресії. Модель множинної регресії має вигляд:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (1.1)$$

де:

Y – залежна змінна (заробітна плата),

X_i – незалежні змінні (фактори),

β_i – коефіцієнти моделі, які оцінюють вплив відповідних змінних на Y ,

ε – випадкова похибка [24].

Однак якість побудови моделей значною мірою залежить від коректної попередньої обробки даних. Одним із критичних етапів є масштабування (scaling), яке дозволяє зробити дані співставними та уникнути ситуацій, коли змінні з великим

діапазоном значень домінують у моделі. У цьому контексті особливо важливим є застосування Z-стандартизації (z-score normalization).

Z-оцінка (або стандартне значення) – це статистичний метод, який визначає, наскільки певне значення відхиляється від середнього значення змінної в одиницях стандартного відхилення. Формула для обчислення Z-оцінки:

$$z = \frac{x - \mu}{\sigma} \quad (1.2)$$

де:

x – значення змінної;

μ – середнє значення вибірки;

σ – стандартне відхилення вибірки [25].

Z-стандартизація перетворює дані таким чином, що всі змінні мають середнє 0 та стандартне відхилення 1, що дає змогу працювати з ними на рівних умовах незалежно від початкових одиниць виміру. Цей підхід є надзвичайно важливим для алгоритмів, чутливих до масштабів (наприклад, градієнтного спуску або методів кластеризації). Крім того, Z-оцінки дозволяють виявляти викиди, спостереження, що суттєво відрізняються від решти. Типово, якщо значення Z-оцінки перевищує межі ± 3 , його розглядають як потенційний викид.

Продовжуючи підготовку даних, необхідно підкреслити важливість масштабування змінних. Стандартизація за допомогою Z-оцінок забезпечує рівноцінне трактування всіх предикторів незалежно від їхніх одиниць вимірювання чи діапазонів. Це особливо важливо, якщо модель включає ознаки різної природи: наприклад, вік, виміряний у роках, досвід у місяцях, а рівень освіти як порядкову категоріальну змінну, закодовану числовими значеннями.

Масштабування необхідне не лише для кращої роботи моделей, а й для правильного інтерпретування коефіцієнтів. Стандартизовані коефіцієнти дозволяють порівнювати відносний вплив змінних: наскільки змінюється заробітна плата при зміні кожного фактору на одну стандартну одиницю [26].

Також як альтернативний метод для покращення моделі в роботі буде використовуватись Ridge та Lasso-регресія.

Ridge-регресія (також відома як L2-регуляризація) модифікує стандартну лінійну регресію, додаючи штраф за велику величину коефіцієнтів. Це допомагає зменшити проблему мультиколінеарності – ситуації, коли незалежні змінні сильно корелюють між собою. Математично, Ridge-регресія мінімізує наступну функцію втрат:

$$L_{ridge}(\beta) = \left(\sum_{i=1}^n (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right), \quad (1.3)$$

де:

y_i – спостережувані значення залежної змінної;

X_i – вектор значень незалежних змінних для спостереження i ;

β – вектор коефіцієнтів;

λ – параметр регуляризації, що контролює силу регуляризації.

Цей метод не обнуляє коефіцієнти, але зменшує їх, що зменшує варіацію моделі та запобігає перенавчанню. Ridge-регресія особливо ефективна, коли всі змінні мають слабкий, але все ж наявний вплив [27].

Lasso-регресія (Least Absolute Shrinkage and Selection Operator) додає регуляризацію, пропорційну абсолютним значенням коефіцієнтів. Це призводить до того, що деякі коефіцієнти можуть стати точно нульовими, що ефективно виключає несуттєві змінні з моделі. Формально, Lasso-регресія мінімізує:

$$L_{lasso}(\beta) = \left(\sum_{i=1}^n (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right), \quad (1.4)$$

де:

y_i, X_i, β, λ мають ті ж значення, що й у Ridge-регресії

Цей метод особливо корисний при роботі з такими даними, де кількість змінних перевищує кількість спостережень, або коли потрібно ідентифікувати найбільш впливові фактори [27].

У дослідженнях, що аналізують вплив соціально-демографічних факторів на оплату праці, ці методи мають свої переваги. Ridge-регресія допомагає стабілізувати оцінки моделі при наявності мультиколінеарності, забезпечуючи більш надійні результати. Lasso-регресія дозволяє ідентифікувати найбільш впливові фактори, автоматично виключаючи менш значущі змінні, що спрощує інтерпретацію моделі. Наприклад, при аналізі впливу освіти, досвіду роботи, віку та інших факторів на заробітну плату, Lasso-регресія може виявити, що лише деякі з цих змінних мають значущий вплив, обнуляючи коефіцієнти інших [27].

Щоб об'єктивно оцінити якість побудованих моделей, використовують відповідні метрики. Корінь середньої квадратичної помилки (RMSE) – є квадратним коренем з MSE. Формула:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (1.5)$$

де:

n – кількість спостережень (вибірка),

x_i – фактичне (реальне) значення цільової змінної для i -го спостереження,

y_i – прогнозоване значення моделі для i -го спостереження.

RMSE має ті ж одиниці вимірювання, що й залежна змінна, що полегшує інтерпретацію. RMSE чутлива до великих помилок. RMSE дозволяє безпосередньо

оцінити середню величину помилки в тих же одиницях, що й залежна змінна. Але чутливість до викидів може призвести до переоцінки помилок при наявності аномальних значень.

Коефіцієнт детермінації (R^2) – вимірює частку варіації залежної змінної, яку пояснює модель. Формула:

$$R^2 = 1 - \frac{SSR}{SST} \quad (1.6)$$

де: SSR — сума квадратів залишків,

SST — загальна сума квадратів.

Значення R^2 , наближене до 1, свідчить про високу пояснювальну здатність моделі, але цей показник зростає навіть за наявності нерелевантних змінних, тому в складних моделях доцільно використовувати також скоригований R^2 .

Обидві метрики разом дають комплексну оцінку точності та надійності моделі: $RMSE$ з точки зору похибки прогнозу, R^2 з точки зору здатності моделі пояснювати змінність доходу. $RMSE$: Підходить для ситуацій, де великі помилки є особливо небажаними, оскільки ці метрики надають більшу вагу великим відхиленням. R^2 : Корисні для оцінки загальної якості моделі та порівняння моделей з різною кількістю незалежних змінних [28].

З метою підвищення точності прогнозування рівня заробітної плати, у роботі було використано низку методів машинного навчання. Серед них метод k найближчих сусідів (KNN), дерева рішень (Decision Trees) та випадковий ліс (Random Forest). Ці алгоритми є частиною так званого контрольованого навчання, оскільки вони працюють з мітками (значеннями цільової змінної) та вчаться на основі вже наявних прикладів.

Метод KNN є одним із найпростіших у реалізації, однак водночас надзвичайно ефективним у випадках, коли структура даних дозволяє виділити локальні закономірності. Головна ідея методу полягає в тому, що значення цільової змінної

нового об'єкта визначається на основі значень тих об'єктів навчальної вибірки, які найбільш подібні до нього.

У задачах регресії, до яких належить і прогнозування заробітної плати, передбачене значення розраховується як середнє (або зважене середнє) значень заробітної плати k найближчих «сусідів» — респондентів, які мають схожі соціально-демографічні характеристики (вік, освіта, професія тощо).

Перевагами цього методу є:

- відсутність припущень щодо форми залежності між змінними (непараметричність);
- гнучкість у застосуванні до різних типів даних;
- інтуїтивна інтерпретація результатів (аналогія до прийняття рішень на основі досвіду схожих випадків).

Проте метод KNN має і певні обмеження. Зокрема, він є чутливим до масштабів ознак (тому вимагає обов'язкового попереднього масштабування), а також до вибору гіперпараметра k . Якщо вибрати занадто мале значення k , модель буде надто чутливою до шуму; якщо занадто велике – вона втратить здатність уловлювати локальні особливості. У межах даного дослідження метод KNN дозволив оцінити дохід респондентів на основі схожості із іншими особами, що мають подібний соціально-демографічний профіль [29].

Дерева рішень – це один із найбільш популярних і зрозумілих алгоритмів у машинному навчанні. У його основі лежить ієрархічна структура, що імітує логіку прийняття рішень. Кожен внутрішній вузол дерева відповідає перевірці певної ознаки (наприклад, «вік > 40?»), гілки можливим результатам цієї перевірки (так/ні), а листові вузли – прогнозованим значенням цільової змінної (в нашому випадку дохід).

Побудова дерева здійснюється шляхом ітеративного розбиття навчальних даних на підмножини, таким чином, щоб на кожному етапі мінімізувати помилку передбачення (наприклад, за критерієм середньої квадратичної помилки або ентропії).

Основними перевагами дерев рішень є:

- інтерпретованість – модель можна легко візуалізувати та пояснити;
- можливість виявляти важливі змінні та взаємодії між ними;
- відсутність вимог до масштабування даних та їхнього розподілу.

Однак недоліком є висока ймовірність перенавчання (overfitting), особливо якщо дерево не обмежується за глибиною або мінімальною кількістю спостережень у листах. У дослідженні дерева рішень були використані для виявлення ключових факторів, що найбільше впливають на рівень заробітної плати, а також для побудови наочних правил, які показують, як комбінації змінних формують дохід [30].

Випадковий ліс є ансамблевим методом, який поєднує у собі велику кількість дерев рішень, побудованих на різних підвибірках даних. Його основною метою є підвищення точності прогнозування за рахунок зниження дисперсії та уникнення перенавчання, властивого окремим деревам.

Ключові етапи формування Random Forest:

- бутстреп-вибірка (bootstrap sampling): кожне дерево будується на основі випадково вибраної частини навчальних даних з поверненням;
- випадковий вибір ознак: при побудові кожного вузла дерево використовує випадкову підмножину ознак, що зменшує кореляцію між деревами;
- агрегація результатів: у випадку регресії прогноз обчислюється як середнє значення прогнозів усіх дерев у моделі.

Переваги Random Forest:

- висока точність і стабільність;
- стійкість до викидів і пропущених значень;
- можливість оцінки важливості ознак, що є надзвичайно корисним у соціально-економічному аналізі;
- гнучкість щодо нелінійних зв'язків і взаємодій між змінними [31].

РОЗДІЛ 2

РЕГРЕСІЙНИЙ АНАЛІЗ ВПЛИВУ СОЦІАЛЬНО-ДЕМОГРАФІЧНИХ ФАКТОРІВ НА РІВЕНЬ ЗАРОБІТНОЇ ПЛАТИ

2.1 Підготовка та візуальний аналіз основних соціально-демографічних факторів

Для аналізу впливу соціально-демографічних факторів на рівень заробітної плати було використано узагальнений набір даних, сформований на основі результатів соціологічного опитування, що проводилося щорічно в період з 2005 до 2018 року. Джерелом інформації слугував масив даних, призначений для досліджень у соціальній економіці [32]. Вибір саме цього набору зумовлений його репрезентативністю, широким охопленням демографічних характеристик респондентів та доступністю для економетричного аналізу.

Загалом, структура даних містить ряд змінних, які відображають ключові соціальні та економічні параметри кожного респодента. Залежною змінною, тобто тією, вплив на яку аналізується, виступає річний дохід особи (income) у доларах США. У свою чергу, пояснювальні змінні охоплюють такі характеристики:

- region – регіон, де проживає опитувана особа. Присутні 7 регіонів, що позначені 1-7.
 - 1) Seoul – столичний мегаполіс, економічно найрозвинутіший регіон країни.
 - 2) Kyeong-gi – провінція навколо Сеула, що характеризується високим рівнем індустріалізації.

- 3) Kyoung-nam – південна частина регіону Кьонсан, відома машинобудівною промисловістю.
 - 4) Kyoung-buk – північна частина Кьонсану, з переважанням традиційної економіки.
 - 5) Chung-nam – південна частина регіону Чхунчхон.
 - 6) Gang-won & Chung-buk – об'єднані території двох переважно сільськогосподарських регіонів.
 - 7) Jeolla & Jeju – об'єднання південного західного регіону країни та острова Чеджу, що має туристичну спрямованість.
- family_member – кількість членів сім'ї.
 - gender – стать опитуваної особи. Чоловіча стать позначається 1, а жіноча позначається 2.
 - education_level – рівень освіти. Присутні 7 рівнів освіти такі, як 1 – відсутня освіта (менш ніж 7 років освіти), 2- відсутня освіта (7 та більше років освіти), 3 – початкова середня освіта, 4 – базова середня освіта, 5 - повна середня освіта, 6 – коледж, 7 – диплом бакалавра, 8 – диплом магістра, 9 – ступінь доктора наук.
 - marriage – сімейний стан особи. Присутні 5 станів: 1 – у шлюбі, 2 – розлучений(через смерть), 3 – розлучений, 4 – не в шлюбі, 5 – інші причини.
 - religion – приналежність особи до релігії. 1 – особа має приналежність до релігії, 2 – особа не має приналежності.
 - occupation - професія/галузь в якій працює особа, 242 унікальних значення
 - age – вік особи
 - income – річний дохід особи у валюті USD

Перед проведенням візуального аналізу соціально-демографічних чинників доцільно здійснити попередню перевірку якості даних. Спочатку робиться оцінка наявності пропущених значень. У процесі аналізу було виявлено, що частина

спостережень має відсутні дані. Оскільки частка таких записів є незначною, приймається рішення виключити їх із подальшого аналізу, щоб не викривляти результати моделей.

Крім того, для виявлення аномальних (екстремальних) значень, було проведено стандартизацію всіх числових змінних шляхом обчислення z-оцінок. Це дозволяє уніфікувати масштаб даних, зменшити вплив відхилень та покращити інтерпретованість графіків, які будуть побудовані в наступному підрозділі.

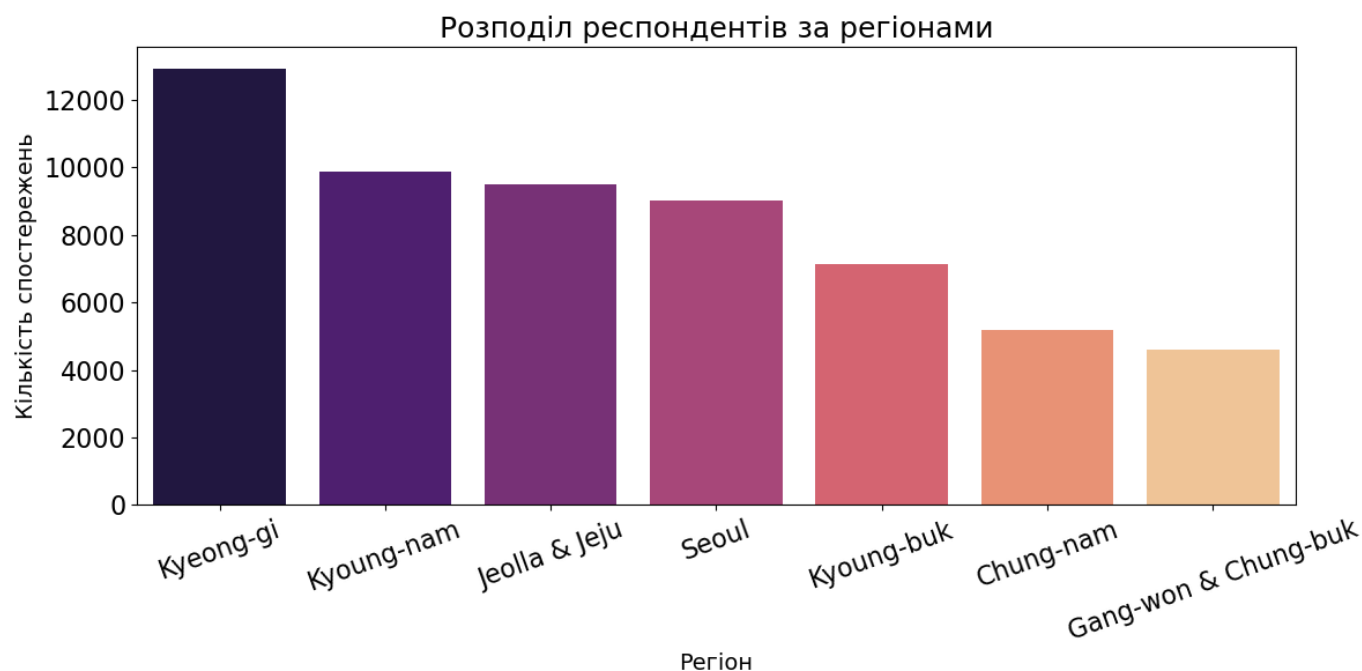


Рисунок 2.1 – Розподіл респондентів за регіонами

Джерело: розроблено автором на основі даних [32]

На рисунку 2.1 представлено кількість респондентів у кожному з регіонів. Як видно з графіка, найбільша кількість спостережень припадає на регіони Kyeong-gi та Seoul, що відповідає їхній високій щільності населення й економічному значенню. Це забезпечує високу репрезентативність вибірки для цих територій. Водночас у регіонах Gang-won & Chung-buk та Chung-nam кількість спостережень значно менша, що може

обмежити точність аналізу впливу регіону на рівень заробітної плати в цих частинах країни.

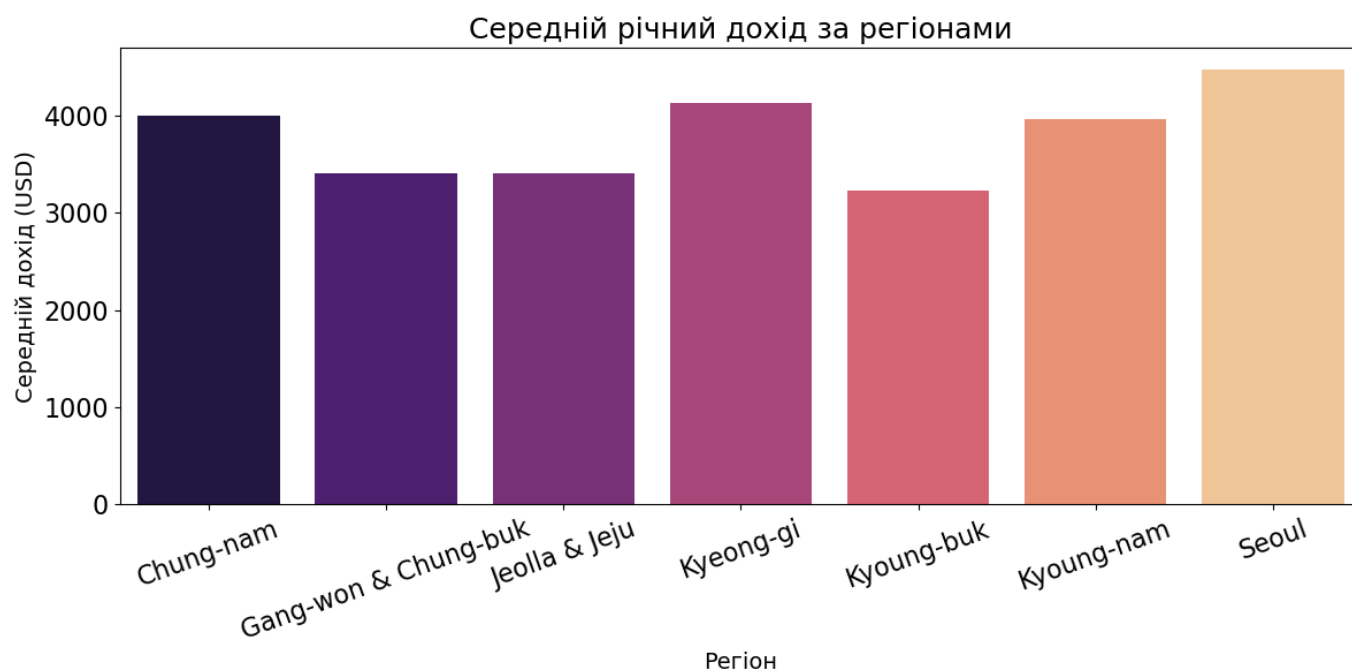


Рисунок 2.2 – Середній річний дохід за регіонами

Джерело: розроблено автором на основі даних [32]

Рисунок 2.2 демонструє середній річний дохід респондентів у розрізі регіонів. З графіка видно, що найвищі доходи спостерігаються в столичних районах, зокрема в Seoul та Kyeong-gi, що є логічним з огляду на рівень економічного розвитку, концентрацію бізнес-структур та більшу кількість високооплачуваних робочих місць. У менш розвинутих або переважно сільськогосподарських регіонах, таких як Gang-won & Chung-buk чи Jeolla & Jeju, спостерігаються нижчі середні доходи. Це свідчить про потенційний вплив регіональної належності на рівень заробітної плати, що доцільно враховувати при побудові економетричних моделей.

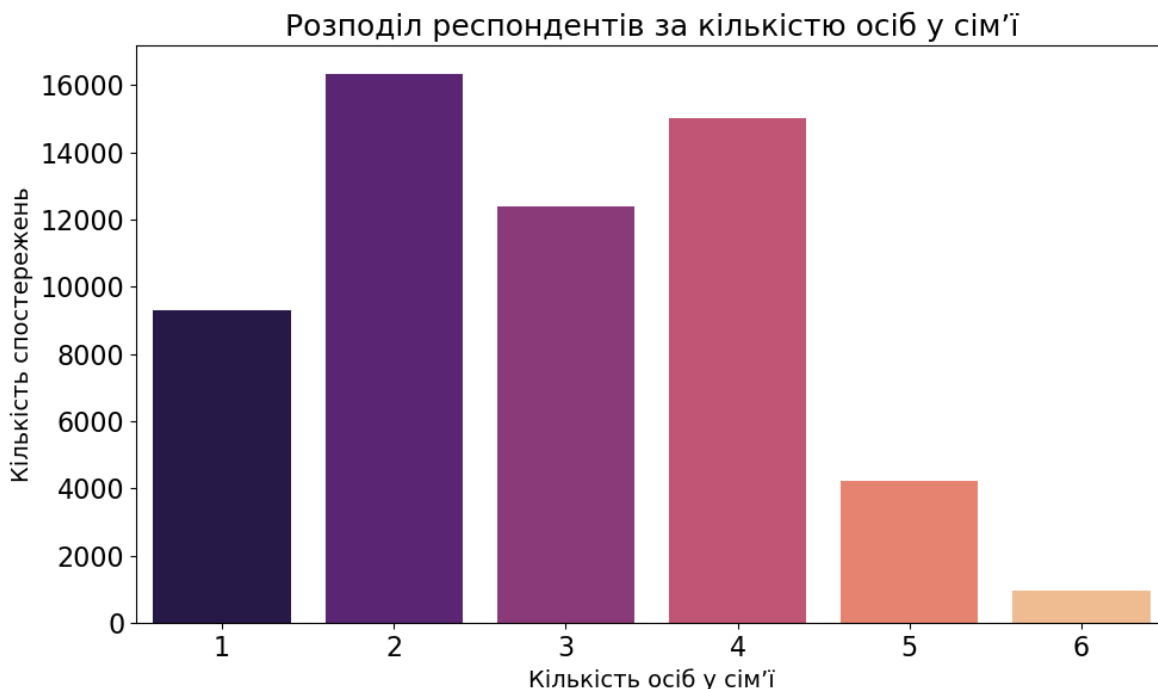


Рисунок 2.3 – Розподіл респондентів за кількістю осіб в сім'ї

Джерело: розроблено автором на основі даних [32]

На рисунку 2.3 зображено розподіл кількості респондентів за чисельністю домогосподарства. Найбільше опитаних припадає на сім'ї, що складаються з 2, 3 та 4 осіб. Зокрема, двоосібні родини становлять найчисельнішу групу, тоді як домогосподарства з п'яти та шести осіб трапляються значно рідше. Це свідчить про те, що малочисельні сім'ї є типовими для досліджуваного періоду та регіону, що загалом відповідає світовим урбанізаційним тенденціям внаслідок зростання індивідуалізму, мобільності та економічних міркувань.

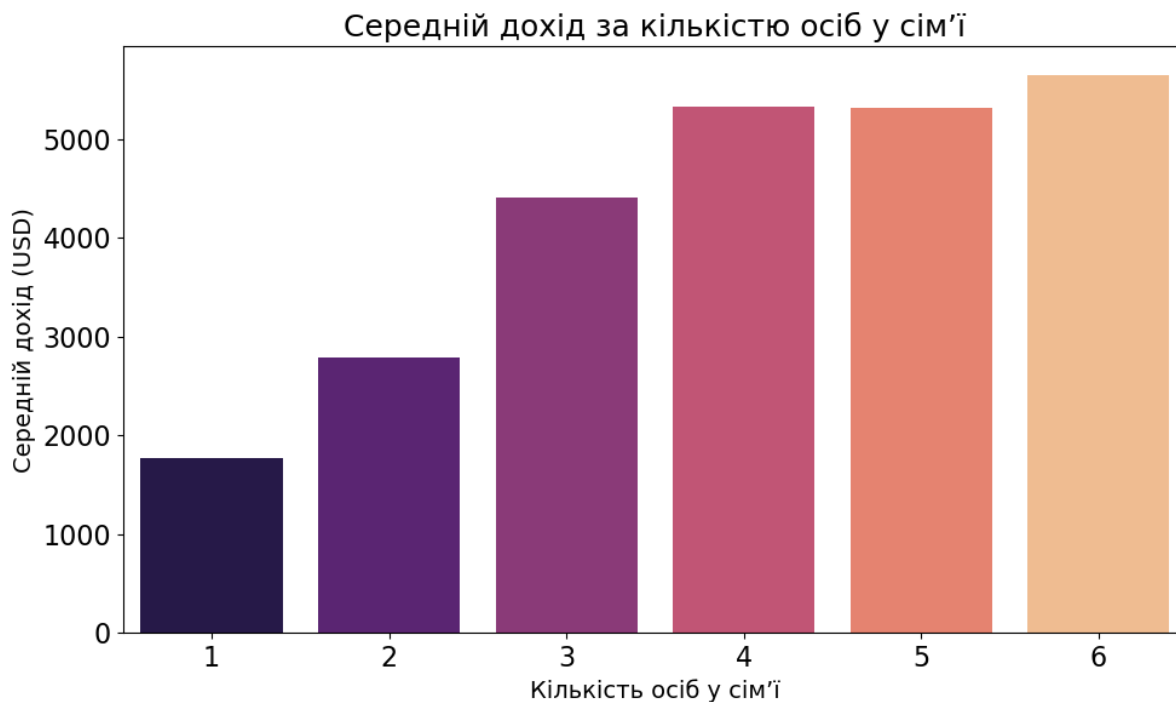


Рисунок 2.4 – Середній дохід за кількістю осіб в сім'ї

Джерело: розроблено автором на основі даних [32]

Графік 2.4 ілюструє залежність середнього річного доходу респондентів від кількості осіб у сім'ї. З графіку видно чітку тенденцію до зростання доходу зі збільшенням кількості членів родини. Зокрема, найнижчий дохід мають одноосібні домогосподарства, тоді як найбільший — сім'ї з шістьма особами. Така закономірність може свідчити як про наявність ефекту кількох працездатних членів у сім'ї, так і про дію соціальних механізмів підтримки, таких як податкові пільги чи інша соціальна підтримка для багатодітних сімей.

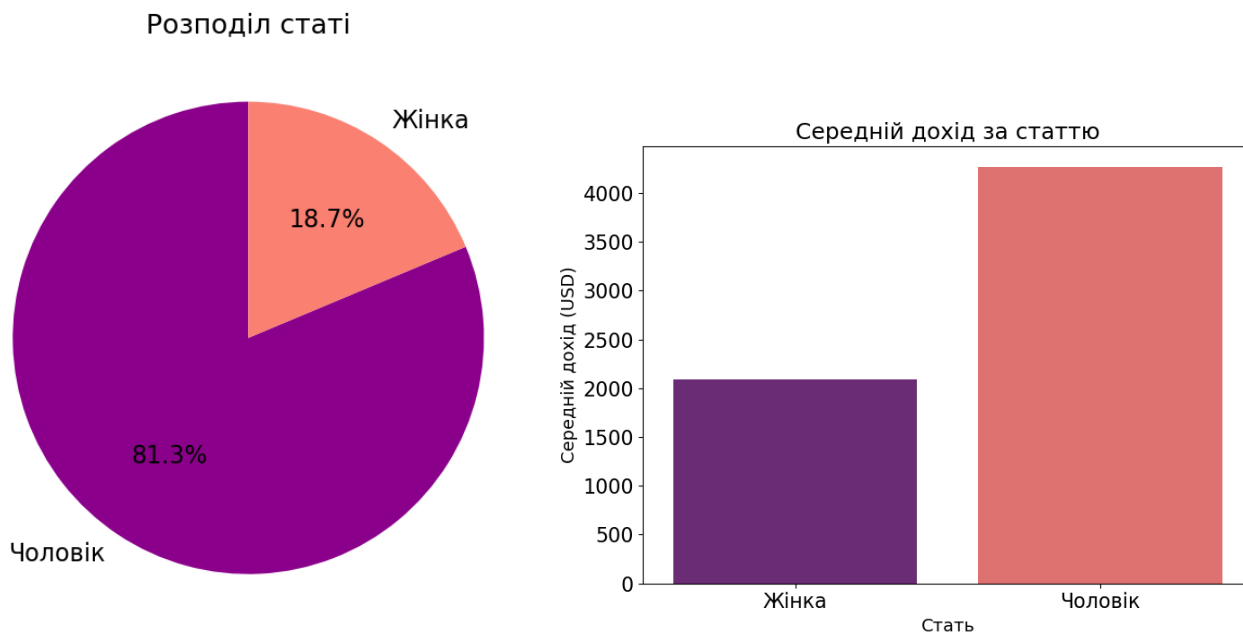


Рисунок 2.5 та Рисунок 2.6 – Розподіл респондентів за статтю та Середній дохід за статтю

Джерело: розроблено автором на основі даних [32]

На рисунку 2.5 представлено кругову діаграму, що демонструє розподіл респондентів за статтю. Як видно з діаграми, більшість опитаних становлять чоловіки (81,3%), тоді як жінки складають лише 18,7% вибірки. Такий дисбаланс у представленні може мати вплив на точність оцінки впливу статі на рівень заробітної плати. На рисунку 2.6 зображено середній дохід за статтю. Згідно з графіком, чоловіки в середньому заробляють суттєво більше, ніж жінки. Цей розрив у доходах може бути пов'язаний з різницею в займаних посадах, відмінностями у досвіді роботи, тривалістю зайнятості, а також із можливими проявами гендерної дискримінації. Таким чином, вплив статі на заробітну плату є візуально значущим.

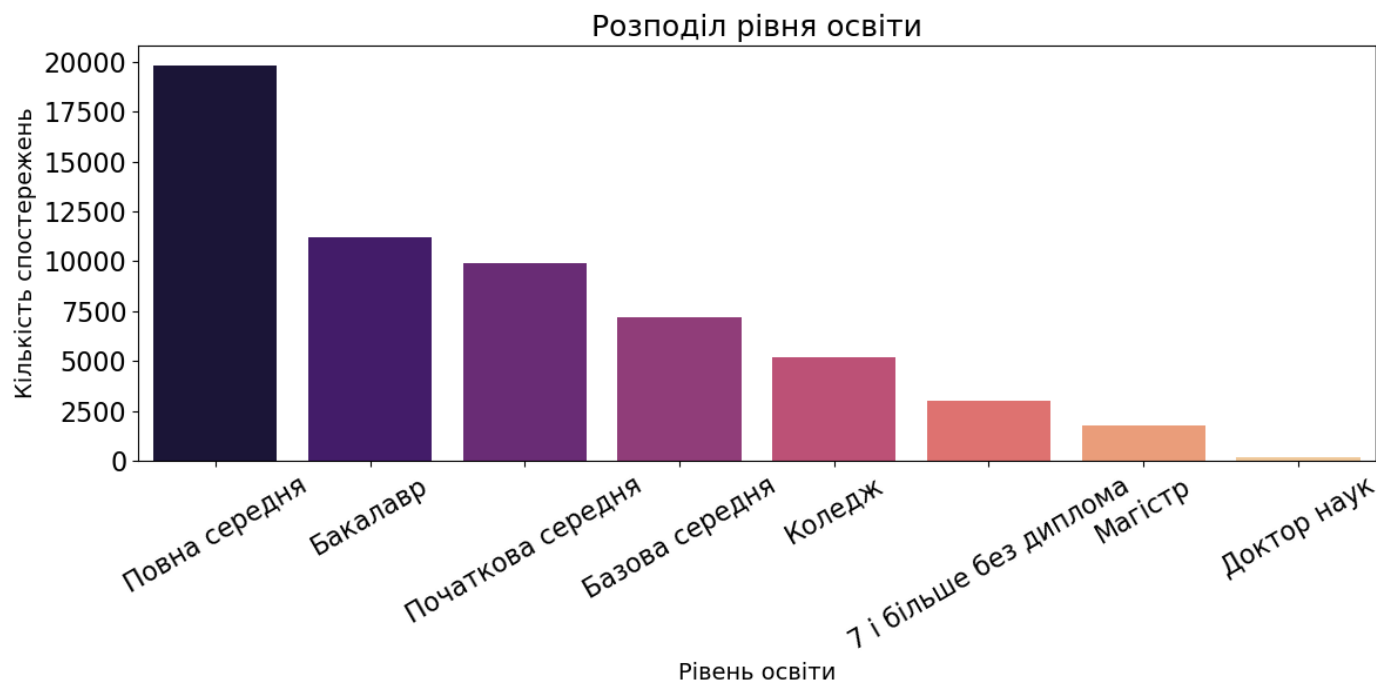


Рисунок 2.7 – Розподіл респондентів за рівнем освіти

Джерело: розроблено автором на основі даних [32]

Графік на рисунку 2.7 ілюструє структуру вибірки за рівнем освіти респондентів. Найбільш чисельну групу становлять особи з повною середньою освітою, кількість яких перевищує 19 тисяч. Наступними за чисельністю є бакалаври та респонденти з початковою середньою освітою, що свідчить про загальну тенденцію до базової або середньої освітньої підготовки серед учасників вибірки.

Натомість магістри та доктори наук представлені незначною кількістю спостережень, що може свідчити про обмежену кількість осіб з найвищим рівнем освіти або меншу доступність таких кваліфікацій у загальній популяції. Отже, освітня структура респондентів зміщена у бік середнього рівня, і це варто враховувати при аналізі впливу освіти на дохід, оскільки високі освітні рівні можуть бути недостатньо репрезентованими для статистично стійких висновків.

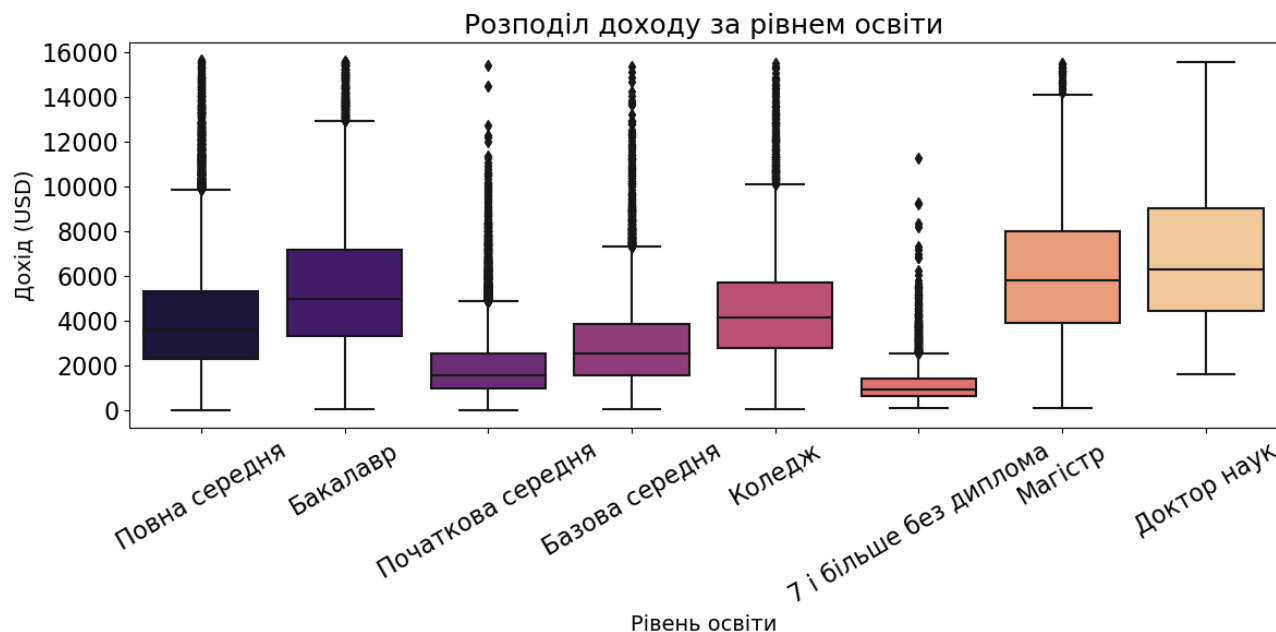


Рисунок 2.8 – Розподіл річного доходу за рівнем освіти

Джерело: розроблено автором на основі даних [32]

Рисунок 2.8 демонструє вплив рівня освіти на величину річного доходу. Спостерігається чітка позитивна тенденція: зі зростанням рівня освіти зростає і середній дохід. Це підтверджує теоретичне положення про те, що освіта є інвестицією в людський капітал, яка підвищує конкурентоспроможність на ринку праці. Водночас варто звернути увагу на значну кількість викидів (екстремальних значень) у межах усіх рівнів освіти — особливо серед середнього рівня: базової середньої освіти, повної середньої освіти та коледжу (4–6), що вказує на високий ступінь варіативності доходів у цих групах. Для найвищих освітніх категорій магістратури та докторату (8 і 9) розподіл доходів є більш стабільним і менш варіативним.

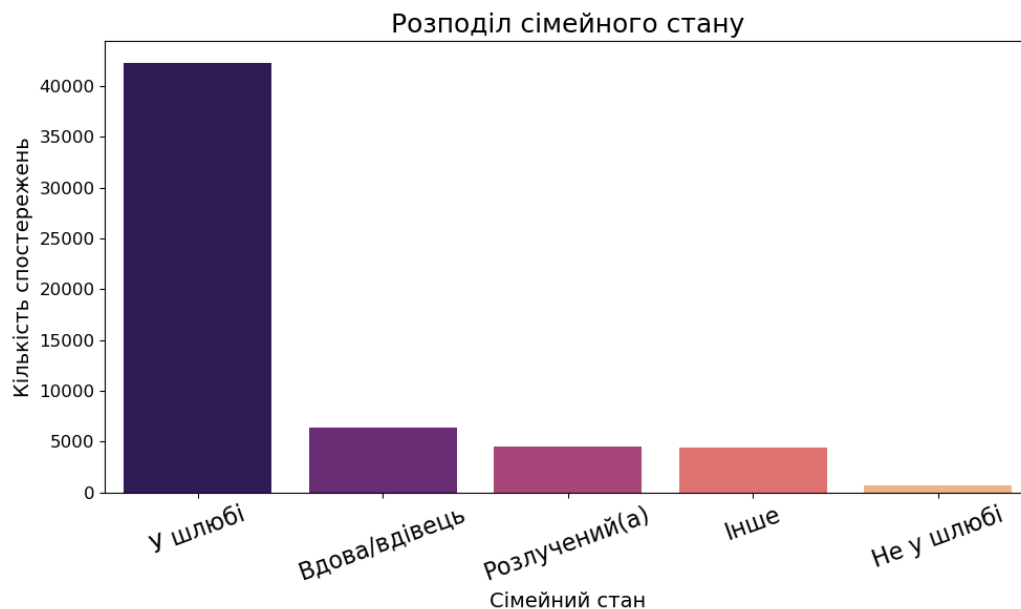


Рисунок 2.9 – Розподіл респондентів за сімейним станом

Джерело: розроблено автором на основі даних [32]

На рисунку 2.9 відображено частоти спостережень для кожної категорії змінної сімейний стан. Очевидно, що домінуючу частку респондентів становлять особи, які перебувають у шлюбі. Інші категорії — розлучені, не в шлюбі, вдівці та ті, що вказали інші причини — представлені значно меншою кількістю спостережень.

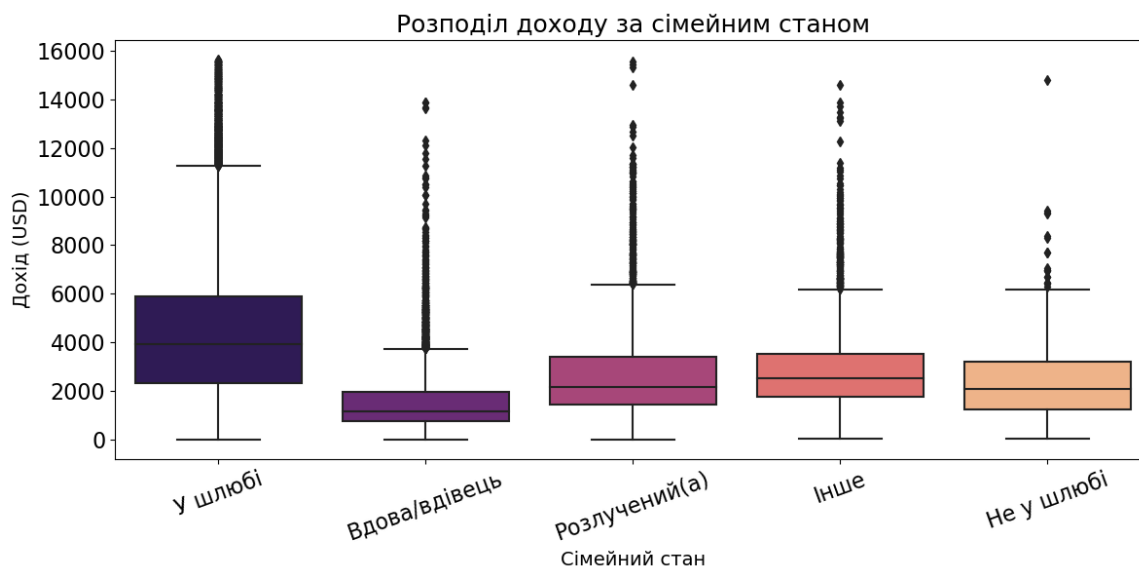


Рисунок 2.10 – Розподіл доходу за сімейним станом

Джерело: розроблено автором на основі даних [32]

На рисунку 2.10 зображено розподіл річного доходу в розрізі різних сімейних статусів. Візуалізація показує, що респонденти, які перебувають у шлюбі, мають найвищий медіанний дохід серед усіх груп. Натомість найнижчий середній дохід спостерігається у категорії вдівці. Інші статуси демонструють приблизно схожі рівні доходу. Варто зазначити, що в усіх категоріях присутні викиди, що свідчить про значну варіативність у межах кожної групи.

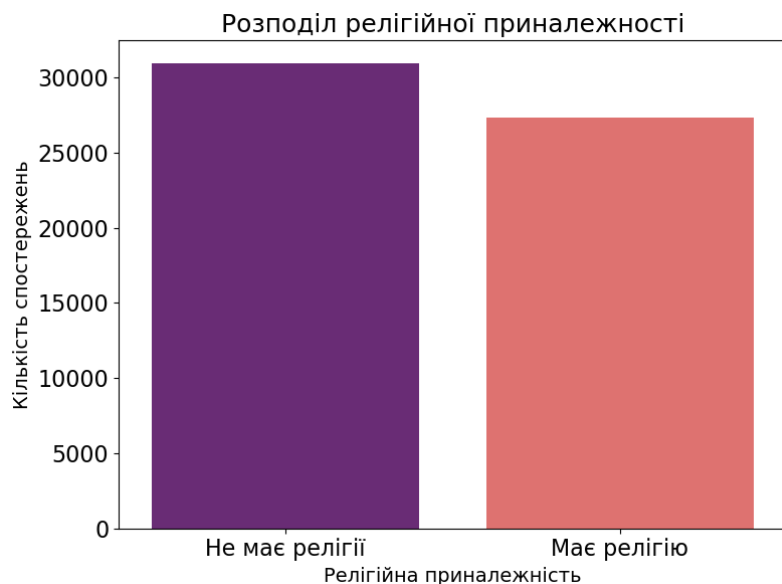


Рисунок 2.11 – Розподіл респондентів за релігійною приналежністю

Джерело: розроблено автором на основі даних [32]

На наведеному рисунку 2.11 зображено розподіл респондентів за релігійною приналежністю. Як видно з діаграми, дещо переважають особи, які не мають релігії, їхня кількість становить понад 31 тисячу спостережень. У той час як осіб із релігійною приналежністю трохи менше – близько 27 тисяч. Розрив між групами помірний, але все ж свідчить про те, що в межах вибірки нерелігійні респонденти складають більшість.

З погляду подальшого аналізу, зважаючи на відносну збалансованість груп, вплив релігійності на заробітну плату за цією вибіркою ймовірно не буде суттєвим, що також буде досліджено у наступних етапах.

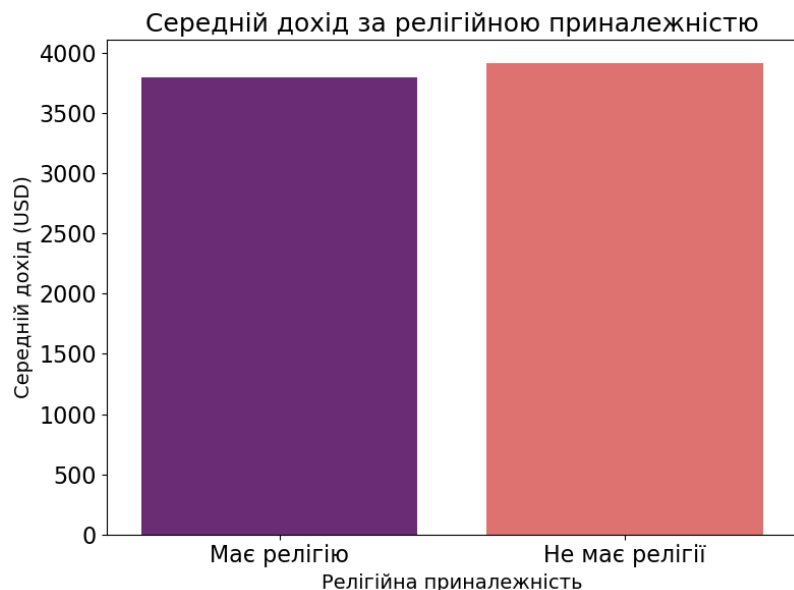


Рисунок 2.12 – Середній дохід за релігійною приналежністю

Джерело: розроблено автором на основі даних [32]

На цьому рисунку 2.12 зображено середній дохід респондентів залежно від їх релігійної приналежності. Видно, що особи, які не мають релігії, мають дещо вищий середній дохід, порівняно з тими, хто ототожнює себе з релігійною групою. Різниця є, проте вона невелика і не виглядає статистично суттєвою.



Рисунок 2.13 – Графік розподілу віку професій

Джерело: розроблено автором на основі даних [32]

Результат статистичного аналізу змінної «occupation»:

```

occupation
611      9946
941      2787
873      2577
312      2188
311      1622
...
1009       2
12         1
113        1
828        1
122        1
Name: count, Length: 242, dtype: int64

```

У даному етапі аналізу розглядається структура змінної «occupation», яка відображає професійну зайнятість респондентів на основі кодів. Оскільки у вибірці присутні 242 унікальні професійні коди, для наочності було обрано 10 найбільш поширених професій за кількістю респондентів.

Як видно з рисунка 2.13, найбільшу частку займають працівники з кодом 611 – працівники рослинництва (Crop cultivation workers). Ця категорія охоплює сільськогосподарських робітників, що займаються вирощуванням рослин, і налічує майже 10 000 осіб, що свідчить про значну представленість аграрного сектору в опитуванні.

Наступними за чисельністю є:

941 – прибиральники (Cleaners and environmental cleaners);

873 – водії автомобілів (Car drivers);

312 – офісні службовці, пов'язані з управлінням (Management-related clerks).

Також далі йдуть професії пов'язані з продажами, такі як продавці чи консультанти в магазині. Ці категорії також мають високу кількість спостережень, але помітно поступаються лідеру. Це свідчить про значну представленість низько- та середньокваліфікованої праці в структурі зайнятості респондентів.

Такий розподіл вказує на те, що переважає фізична праця та робітничі професії, а висококваліфіковані посади менш репрезентовані. Це варто враховувати при подальшому аналізі доходів, оскільки професійна структура може суттєво впливати на середній дохід та загальні висновки моделі.

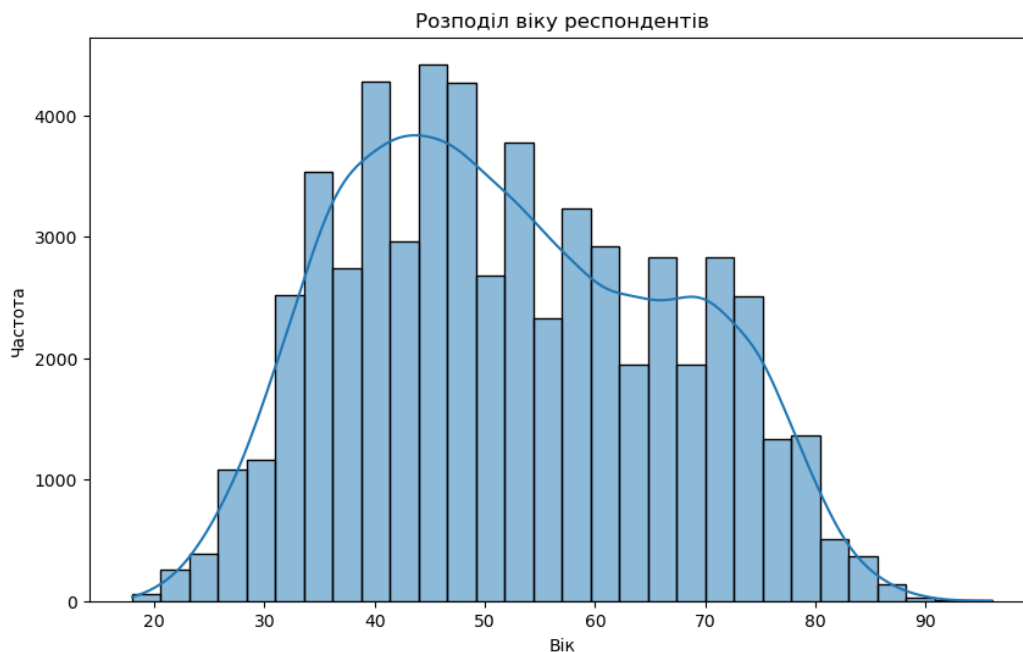


Рисунок 2.14 – Гістограма розподілу віку респондентів

Джерело: розроблено автором на основі даних [32]

На рисунка 2.14 подано розподіл віку респондентів у вибірці. Як видно з гістограми, найбільше представлено респондентів віком від 40 до 50 років. Кількість респондентів поступово зростає, починаючи з віку близько 20 років, досягаючи піку в 40-50 років. Це типовий розподіл для багатьох соціальних опитувань, де активна частина населення середнього віку є найбільш представленою. Опитування включає респондентів віком приблизно від 20 до 90 років, що свідчить про широке охоплення вікових груп.

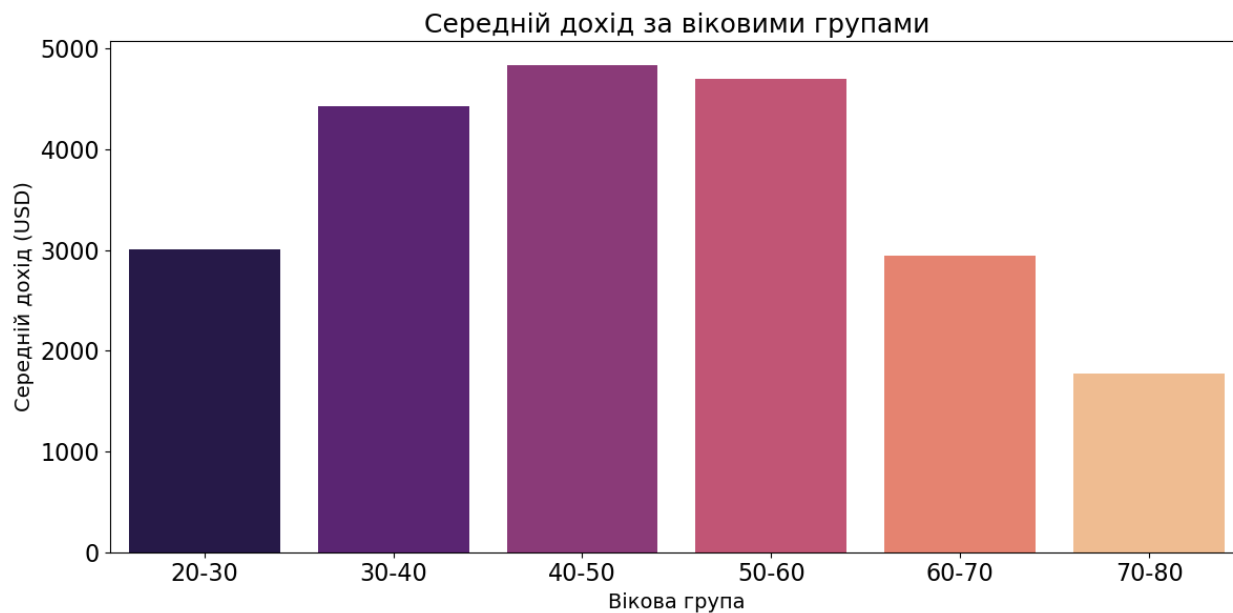


Рисунок 2.15 – Середній дохід за віковою групою

Джерело: розроблено автором на основі даних [32]

На рис. 2.15 зображено, як середній дохід змінюється залежно від вікової групи. Спостерігається чітка залежність доходу від віку: найнижчий рівень доходу мають респонденти у віці 20-30 років, тоді як пік доходів припадає на вікові групи 40-50 та 50-60 років. Після цього дохід поступово знижується, особливо помітно у групах 60-70 та 70-80 років.

Цей розподіл логічно відображає типову економічну динаміку життєвого циклу: у молодому віці особи часто перебувають на початкових позиціях з нижчим рівнем оплати, з віком та накопиченням досвіду дохід зростає, а у передпенсійному та пенсійному віці – знижується через зменшення економічної активності. Таким чином, вік є одним із важливих факторів, що впливають на рівень заробітної плати, і його доцільно включати до економетричних моделей оцінки доходу.

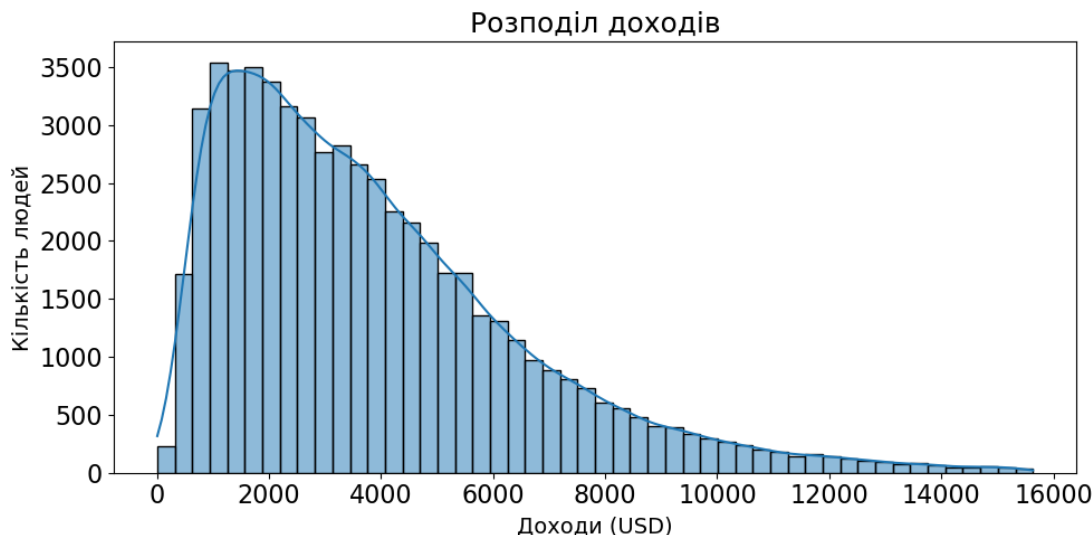


Рисунок 2.16 – Гістограма розподілу річного доходу респондентів

Джерело: розроблено автором на основі даних [32]

Далі йде залежна змінна дохід, або заробітна плата (income). Саме ця змінна є основною в подальшому дослідженні.

На рис. 2.16 найочевидніша характеристика, що більшість людей мають низькі або середні доходи, тоді як значно менша кількість людей отримує високі доходи. Хвіст розподілу тягнеться вправо. Найвища частота спостерігається у діапазоні доходів приблизно від 1500 до 2500 USD. Пік розподілу знаходиться близько 2000 USD. Після досягнення моди) кількість людей різко зменшується зі зростанням рівня доходу.

Даний розподіл доходів є типовим для більшості економік світу, де доходи розподіляються нерівномірно: велика частина населення має менші доходи, і лише невелика частина населення отримує значно більші доходи.

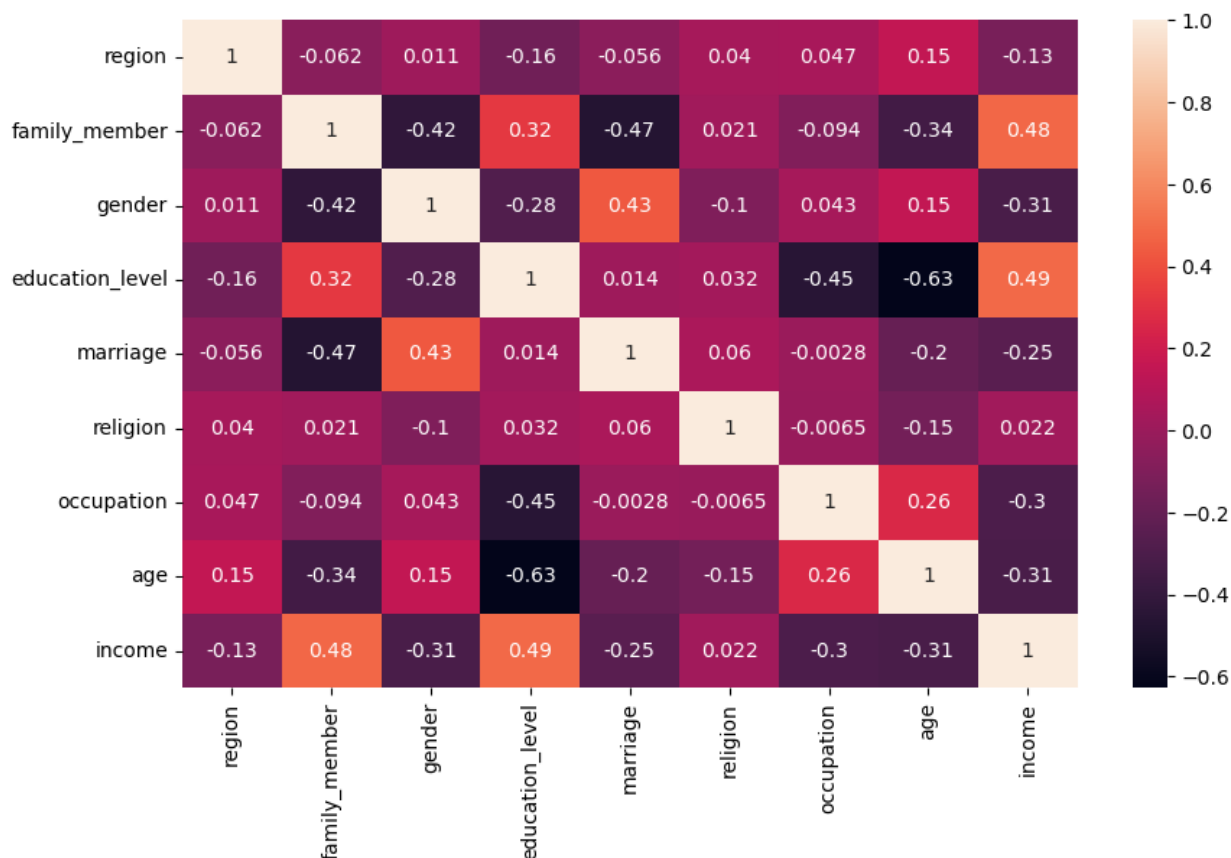


Рисунок 2.17 – Кореляційна матриця для визначення залежностей між змінними та змінною заробітної плати.

Джерело: розроблено автором на основі даних [32]

На графіку 2.17 зображена матриця кореляції між усіма числовими змінними в дослідженні. Значення варіюються в діапазоні від -1 (сильна обернена кореляція) до +1 (сильна пряма кореляція). Найвища позитивна кореляція з доходом (income):

- Рівень освіти (education_level) – 0.49, що підтверджує попередній візуальний аналіз: чим вищий рівень освіти, тим вищий дохід.
- Кількість членів сім'ї (family_member) – 0.48, що може свідчити про те, що особи з більшими родинами мають вищі доходи, можливо, через більшу кількість працюючих членів сім'ї.

- Регіон (region) – 0.13, що говорить про слабкий, але присутній вплив регіонального чинника на рівень доходу.

Негативна кореляція з доходом:

- Стать (gender) – -0.31, вказує на те, що жіноча стать пов'язана з нижчим середнім доходом (за умови кодування: 1 – чоловік, 2 – жінка).
- Вік (age) – -0.31, що може свідчити про зниження доходу після певного вікового порогу (як видно з попереднього графіку).

Сильна міжзмінна кореляція:

- Освіта та вік (education_level vs age) – -0.63, що логічно пояснюється тим, що старші покоління частіше мають нижчий рівень формальної освіти.
- Освіта та професія (education_level vs occupation) – -0.45, що вказує на зв'язок між рівнем освіти та типом професій, які обирають респонденти.
- Стать та шлюб (gender vs marriage) – 0.43, можливо пов'язано з гендерними відмінностями в сімейному стані (наприклад, жінки частіше вдови або розлучені в старших вікових групах).

Висновок:

Рівень освіти, кількість членів сім'ї та стать є найсильнішими предикторами доходу серед усіх соціально-демографічних змінних. Слабка або відсутня кореляція з релігією, шлюбом і професійним кодом вказує на менш очевидний вплив цих змінних у простій лінійній залежності, хоча вони можуть бути важливими у складніших моделях. Це підтверджує доцільність подальшого використання регресійного аналізу для виявлення взаємозв'язків.

2.2. Побудова та інтерпретація економетричної моделі залежності заробітної плати від соціально-демографічних факторів

Для побудови економетричних моделей залежності заробітної плати від соціально-демографічних характеристик було використано масив даних, що містить 58 963 спостереження. Дані включають такі ознаки, як регіон проживання, кількість членів сім'ї, стать, рівень освіти, сімейний стан, релігійна приналежність, професія, вік, а також річний дохід респондента.

На етапі попередньої обробки було застосовано метод z-оцінки для виявлення викидів. Зокрема, спостереження, які перевищували абсолютне значення 3 за будь-якою з числових змінних, були визнані викидами та видалені. Це дозволило зменшити вплив аномальних значень на результати моделювання. У підсумку було залишено 58 226 записів для подальшого аналізу.

Для забезпечення коректної роботи алгоритмів моделювання, особливо чутливих до масштабів ознак (наприклад, лінійної регресії), усі числові змінні були стандартизовані за допомогою `StandardScaler`. У результаті ознаки були приведені до одного масштабу (середнє значення 0, стандартне відхилення 1). Також було стандартизовано і цільову змінну річний дохід.

Дані були поділені на тренувальну вибірку (70%) та тестову (30%), що дозволяє об'єктивно оцінити якість моделей на нових даних. Фінальний розмір вибірок:

- X (ознаки): 58 226 спостережень, 8 змінних;
- y (дохід): 58 226 значень.

На першому етапі було побудовано базову лінійну регресійну модель за допомогою методу найменших квадратів (`LinearRegression` з бібліотеки `statsmodels`).

Оцінимо якість моделі на тренувальній та тестовій вибірках за допомогою статистики (функція `statsmodels`).

Результати:

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.388			
Model:	OLS	Adj. R-squared:	0.388			
Method:	Least Squares	F-statistic:	3230.			
Date:	Tue, 27 May 2025	Prob (F-statistic):	0.00			
Time:	00:06:27	Log-Likelihood:	-47950.			
No. Observations:	40758	AIC:	9.592e+04			
Df Residuals:	40749	BIC:	9.600e+04			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.0032	0.004	0.829	0.407	-0.004	0.011
region	-0.0617	0.004	-15.562	0.000	-0.069	-0.054
family_member	0.3194	0.005	61.506	0.000	0.309	0.330
gender	-0.0504	0.005	-10.768	0.000	-0.060	-0.041
education_level	0.3472	0.006	61.398	0.000	0.336	0.358
marriage	-0.0770	0.005	-14.816	0.000	-0.087	-0.067
religion	0.0119	0.004	2.983	0.003	0.004	0.020
occupation	-0.1188	0.004	-26.998	0.000	-0.127	-0.110
age	0.0527	0.006	9.339	0.000	0.042	0.064
Omnibus:	8885.992	Durbin-Watson:	2.010			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22515.274			
Skew:	1.195	Prob(JB):	0.00			
Kurtosis:	5.747	Cond. No.	2.96			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Рис. 2.18 – Статистика для моделі лінійної регресії

Джерело: розроблено автором на основі даних [32]

– R-squared (Коефіцієнт детермінації): 0.388

Цей показник, що показує, яку частку дисперсії залежної змінної (y) пояснюють незалежні змінні (предиктори), включені в модель. Значення 0.388 означає, що 38.8% варіабельності доходу може бути пояснено факторами, включеними у модель. Це помірний показник.

– Adj. R-squared (Скоригований R-squared): 0.388

Цей показник аналогічний R-squared, але він коригується з урахуванням кількості незалежних змінних у моделі та кількості спостережень. Він корисний для порівняння моделей з різною кількістю предикторів.

– Method (Метод): Least Squares

Вказує, що модель була побудована за методом найменших квадратів, який мінімізує суму квадратів різниць між спостережуваними та прогнозованими значеннями.

– F-statistic (F-статистика): 3230.

Це статистичний тест для перевірки загальної значущості моделі. Він перевіряє нульову гіпотезу про те, що всі коефіцієнти регресії (крім константи) одночасно дорівнюють нулю.

– Prob (F-statistic) (P-значення F-статистики): 0.00

P-значення F-статистики (0.00) є надзвичайно низьким, що дозволяє відхилити нульову гіпотезу про відсутність впливу предикторів. Це свідчить про те, що регресійна модель є статистично значущою в цілому, підтверджуючи, що принаймні одна з незалежних змінних має вагомий лінійний зв'язок з доходом.

– No. Observations (Кількість спостережень): 40758

– Df Residuals (Ступені свободи залишків): 40749

– Df Model (Ступені свободи моделі): 8

– coef (Коефіцієнт): Оцінка впливу кожної змінної на залежну змінну у.

Таблиця 2.1 – Аналіз коефіцієнтів в множинній лінійній регресії.

Змінна	Коефіцієнт	P-значення	Інтерпретація
const	0.0032	0.407	Константа незначуща.
region	-0.0617	<0.001	Доходи зменшуються зі зростанням номера регіону.
family_member	0.294	<0.001	Більше членів сім'ї, вищий дохід.

Змінна	Коефіцієнт	P-значення	Інтерпретація
gender	-0.0504	<0.001	Негативний вплив статі.
education_level	0.3472	<0.001	Один з найсильніших позитивних факторів.
marriage	-0.0770	<0.001	Одружені мають нижчий дохід.
religion	0.0119	0.003	Незначний позитивний вплив.
occupation	-0.1188	<0.001	Вплив залежить від ієрархії кодування професій.
age	0.0527	<0.001	Вік позитивно впливає на дохід, але помірно.

Джерело: розроблено автором на основі даних [32]

Підставляючи значення коефіцієнтів з нашого виводу statsmodels, отримуємо рівняння:

$$\begin{aligned}
 income = & 0.0032 - 0.0617 \cdot region + 0.3194 \cdot family_member - 0.0504 \cdot \\
 & gender + 0.3472 \cdot education_level - 0.0770 \cdot marriage + 0.0119 \cdot religion - \\
 & 0.1188 \cdot occupation + 0.0527 \cdot age
 \end{aligned}
 \quad (2.1)$$

std err (Стандартна похибка): Міра точності оцінки коефіцієнта. Менші значення означають більш точні оцінки.

t-статистика: відношення коефіцієнта до його стандартної помилки. Використовується для перевірки індивідуальної значущості кожного коефіцієнта.

P-значення: ймовірність відхилення або прийняття нульової гіпотези.

Значущість: якщо $P < 0.05$, коефіцієнт вважається статистично значущим, і ми можемо стверджувати, що змінна має статистично значущий лінійний вплив на y . У вашому випадку, майже всі змінні, крім константи, є сильно значущими ($P=0.000$ або 0.003).

Довірчий інтервал: Діапазон значень, у якому, з 95% ймовірністю, знаходиться справжній коефіцієнт. Якщо інтервал включає 0, то коефіцієнт не є статистично значущим.

Durbin-Watson: 2.010.

Тест на автокореляцію (залежність) залишків. Значення 2.010 дуже близьке до 2, що вказує на відсутність значної автокореляції у залишках.

Jarque-Bera(JB) : 22515.274. Prob(JB): 0.00

Тест на нормальність залишків, що враховує скіс (Skew) та ексцес (Kurtosis). Prob(JB) = 0.00 також підтверджує, що залишки не є нормально розподіленими.

Skew (Скіс): 1.195 та Kurtosis (Ексцес): 5.747.

Позитивне значення скісу вказує на те, що розподіл залишків є позитивно скошеним. Це означає, що є довший хвіст у позитивному напрямку (тобто, більше позитивних помилок). Значення ексцесу 5.747 (значно більше 3). Це також є ознакою ненормального розподілу залишків.

Cond. No. (Умовне число): 2.96

Цей показник оцінює наявність мультиколінеарності (сильної кореляції між незалежними змінними). Значення 2.96 є дуже низьким, отже проблема мультиколінеарності відсутня.

2.3 Удосконалення економетричних моделей із застосуванням методів машинного навчання та регуляризації

Для підвищення точності прогнозування річного доходу на основі соціально-демографічних характеристик було побудовано та протестовано низку

альтернативних моделей. Крім класичної лінійної регресії, розглядалися моделі з регуляризацією (Ridge та Lasso), а також алгоритми машинного навчання, зокрема дерева рішень, випадковий ліс (Random Forest) та метод k найближчих сусідів (KNN).

Для кожної моделі обчислювалися наступні показники:

- R^2 (коефіцієнт детермінації) – відображає частку варіації цільової змінної, що пояснюється моделлю;
- RMSE (корінь середньоквадратичної помилки) — вимірює середню відстань між прогнозованими і фактичними значеннями, чутливий до викидів.

Класична лінійна регресія

Базова модель лінійної регресії, побудована на стандартизованих даних, продемонструвала наступні результати:

- R^2 (навчальна вибірка): 38.80%
- R^2 (тестова вибірка): 37.67%
- RMSE (тестова вибірка): 0.78

Це свідчить про те, що модель пояснює близько 38% варіації річного доходу. Незважаючи на обмежену точність, ця модель демонструє прийнятну інтерпретованість та є відправною точкою для вдосконалення.

Ridge-регресія

Ridge-регресія включає L2-регуляризацію, що зменшує значення коефіцієнтів і запобігає їх переоціненню у разі мультиколінеарності. Однак за отриманими результатами:

- R^2 (тренувальна): 38.80%
- R^2 (тестова): 37.67%
- RMSE: 0.78

можна стверджувати, що модель не забезпечила жодного покращення порівняно з класичною лінійною регресією. Це свідчить про те, що проблема переобучення чи мультиколінеарності у даному випадку була не критичною.

Lasso-регресія

Lasso-регресія включає L1-регуляризацію, що дозволяє автоматично відкинути несуттєві ознаки, зануляючи відповідні коефіцієнти. Проте:

- R^2 (тренувальна): 35.75%
- R^2 (тестова): 34.97%
- RMSE: 0.80

Модель показала гірші результати порівняно з іншими лінійними підходами. Ймовірною причиною є те, що Lasso занадто агресивно обнулює деякі корисні предиктори, зменшуючи загальну пояснювальну здатність.

Загалом, регуляризаційні методи (Ridge, Lasso) не покращили якість прогнозу, а Lasso навіть погіршила. Це може свідчити про те, що всі змінні мають хоча б певний вплив, і видалення окремих з них з моделі зменшує її ефективність.

З метою виявлення нелінійних взаємозв'язків і підвищення точності прогнозування були застосовані сучасні алгоритми машинного навчання: дерево рішень, випадковий ліс (Random Forest) та метод найближчих сусідів (K-Nearest Neighbors).

Decision Tree

- R^2 (тренувальна): 96.03%
- R^2 (тестова): 30.90%
- RMSE: 0.83

Модель дерева рішень показала перенавчання (overfitting): дуже висока точність на тренувальній вибірці і значне погіршення на тестовій. Це свідчить про низьку узагальнювальну здатність.

K-Nearest Neighbors (KNN)

- R^2 (тренувальна): 68.20%
- R^2 (тестова): 49.23%
- RMSE: 0.71

Метод найближчих сусідів показав кращі результати за класичні моделі, але поступається Random Forest. Він також досить чутливий до розмірності та структури даних.

Random Forest (найкраща модель)

- R^2 (тренувальна): 91.55%
- R^2 (тестова): 57.82%
- RMSE: 0.64

Random Forest Regression став найкращою моделлю в дослідженні. Завдяки поєднанню великої кількості дерев рішень, ця модель добре справляється з шумами, викидами та нелінійними залежностями. Високий показник R^2 на тестовій вибірці ($\approx 58\%$) свідчить про її здатність до узагальнення.

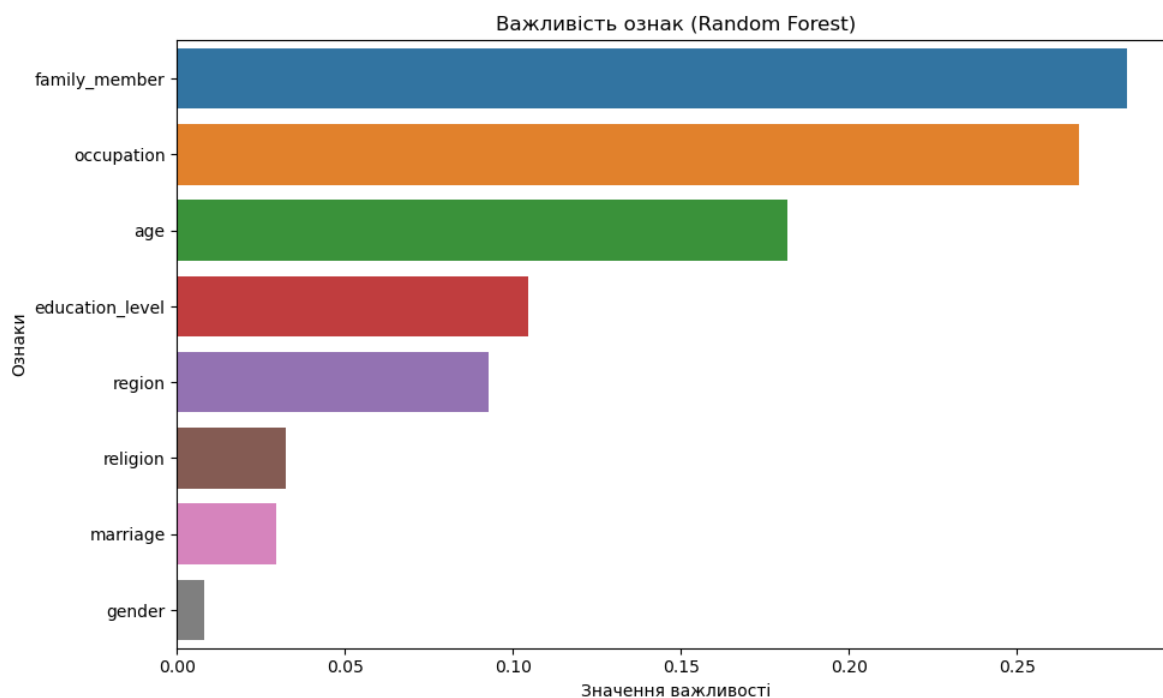


Рисунок 2.19 – Важливість ознак у моделі Random Forest

Джерело: розроблено автором на основі даних [32]

На рисунку 2.19 зображено горизонтальну гістограму важливості ознак. Очевидно, що найбільший вплив на рівень доходу мають:

- Кількість членів сім'ї – 28,3%
- Професійна сфера (occupation) – 26,8%
- Вік – 18,2%

Ці змінні суттєво випереджають інші за впливом. Натомість стать (gender), всупереч очікуванням, має лише 0.8% важливості, що може свідчити або про ефект стандартизації, або про більш опосередкований вплив через інші змінні.

Ці результати підтверджують, що дохід формується як результат взаємодії декількох соціальних та демографічних характеристик, а не лише традиційних факторів, таких як освіта чи стать.

Таблиця 2.2 – Коефіцієнти ознак у моделі Random Forest

Ознака	Важливість
family_member	0.2828
occupation	0.2684
age	0.1818
education_level	0.1043
region	0.0927
religion	0.0323
marriage	0.0296
gender	0.0081

Джерело: розроблено автором на основі даних [32]

Порівнюючи з лінійною регресією, Random Forest виявив сильну роль ознаки occupation, хоча лінійна модель її практично ігнорувала. Це може свідчити про нелінійний вплив професії: можливо, для певних груп вона суттєво впливає, але цей ефект "середньо нульовий" у лінійній моделі.

gender має вагомий негативний вплив у лінійній регресії, але Random Forest вважає її майже неважливою — це може означати, що взаємодії або кореляції з іншими змінними згладжують її ефект.

Random Forest краще виявляє складні, нелінійні зв'язки, тоді як лінійна регресія дає просту й пряму інтерпретацію через коефіцієнти.

ВИСНОВКИ

У результаті проведеного дослідження було реалізовано повний цикл аналітичної роботи від теоретичного обґрунтування проблематики впливу соціально-демографічних факторів на рівень заробітної плати до побудови та порівняння низки економетричних та моделей машинного навчання. Проведений аналіз дозволив глибше зрозуміти структуру формування доходів населення і виявити ті чинники, що справді визначають фінансову ситуацію працівників у сучасному суспільстві.

На першому етапі проведено ретельне дослідження теоретичних засад оплати праці, її функцій та підходів до вимірювання. Окреслено ключові соціально-демографічні фактори, що згідно з сучасною науковою літературою мають вплив на рівень заробітної плати. Зокрема, було систематизовано роль таких змінних, як вік, стать, освіта, сімейний стан, професія, регіон проживання тощо. У межах цього блоку було обґрунтовано актуальність поєднання класичних економетричних підходів з сучасними алгоритмами машинного навчання, що дозволяє комплексно охопити як лінійні, так і складні нелінійні взаємозв'язки в даних.

У ході практичної частини дослідження здійснено глибоку попередню обробку даних, що включала видалення пропусків, очищення від викидів методом Z-оцінок, стандартизацію числових змінних та кодування ознак. Це дозволило отримати чистий і уніфікований масив обсягом понад 58 тисяч спостережень, репрезентативний за соціально-демографічними параметрами.

Особливе значення мав етап візуального аналізу даних, який дозволив виявити первинні залежності між окремими змінними та річним доходом. На цьому етапі було встановлено:

- істотний вплив регіону проживання: у найбільш розвинених регіонах (Seoul, Kyeong-gi) рівень доходу суттєво вищий;
- позитивну залежність доходу від кількості членів сім'ї, що, ймовірно, пов'язано з ефектом декількох доходів у домогосподарстві;

- суттєвий гендерний розрив у середньому доході: чоловіки заробляють значно більше;
- зростання доходу разом із підвищенням рівня освіти — наочна демонстрація впливу людського капіталу;
- найвищі доходи спостерігаються у віковій групі 40–50 років, після чого рівень доходу знижується.

Кореляційний аналіз підтвердив, що найсильнішими предикторами доходу в лінійних зв'язках є рівень освіти, кількість членів сім'ї, а також стать. Водночас, з боку віку та професії виявлені слабші або нелінійні зв'язки, що потребують більш гнучких моделей.

У рамках економетричного моделювання було побудовано кілька моделей: базова лінійна регресія, Ridge-регресія, Lasso-регресія, а також алгоритми машинного навчання — Decision Tree, K-Nearest Neighbors та Random Forest.

Модель класичної множинної лінійної регресії дозволила пояснити приблизно 38,8% варіації доходу. Значущими факторами виявились: рівень освіти, кількість членів сім'ї, стать, шлюбний стан та вік. При цьому коефіцієнти регресії дозволили зробити важливі висновки: наприклад, негативний вплив жіночої статі, позитивний вплив вищої освіти та старшого віку до певного порогу. Лінійна модель продемонструвала хорошу інтерпретованість, але обмежену точність.

Регуляризовані моделі (Ridge і Lasso) не покращили якість прогнозу: R^2 залишився незмінним або навіть зменшився. Це свідчить про відсутність критичної мультиколінеарності та важливість усіх змінних у моделі.

Найбільш помітні результати дали моделі машинного навчання, особливо Random Forest. Саме ця модель продемонструвала найкращу здатність до узагальнення: R^2 на тестовій вибірці становив понад 57%, що суттєво вище за класичні моделі. Крім того, вона виявила нелінійні ефекти, які не могли бути ідентифіковані простими лінійними методами.

Оцінка важливості ознак у Random Forest Regression засвідчила суттєві зрушення в порівнянні з лінійною моделлю. Найважливішими виявились:

- кількість членів сім'ї (28,3% впливу),
- професія (26,8%),
- вік (18,2%).

Натомість вплив статі, який був значущим у лінійній моделі, в Random Forest знизився до менш ніж 1%, що може вказувати на те, що гендерні ефекти опосередковано враховуються через професію, вік чи освіту.

Порівняльний аналіз моделей дозволив дійти висновку, що лінійна регресія — може бути базою для початку інтерпретації, однак для точного прогнозування та врахування складних взаємозв'язків краще використовувати інші моделі, зокрема Random Forest.

Отже, в рамках дослідження було не лише побудовано прогностичні моделі, а й показано цінність сучасних аналітичних методів у розв'язанні прикладних соціально-економічних задач. Результати можуть бути корисними як для теоретиків, так і для практиків — державних органів, роботодавців, аналітичних служб, що займаються питаннями оплати праці, гендерної рівності та соціальної політики.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Закон України «Про оплату праці» від 24.03.1995 № 108/95-ВР. URL: <https://zakon.rada.gov.ua/laws/show/108/95-%D0%B2%D1%80#Text> (дата звернення: 20.05.2025)
2. Машевська А. Економічна сутність та підходи до визначення «заробітна плата». *Молодий вчений*, 2021. № 92. С. 308-313.
3. Глухова С. В., Зотова А. А. Підходи до сутності заробітної плати. *Молодий вчений*, 2016. № 1. С. 698-701.
4. Тимош І.М. Економіка праці. Тернопіль: Астон, 2010. 347 с
5. Мочерний С. Економічна енциклопедія. У 3-х т. Т. 1. Вид-во «Академія», 2008. 864 с.
6. Мочерний С.В. Політична економія: навч. посіб. С.В. Мочерний. *Знання-Прес*, 2006. 687 с
7. Гадзевич О.І. Оплата праці в умовах ринку: теорія, практика. К. КОНДОР, 2008. 400 с
8. Калина А.В. Економіка праці. МАУП, 2009. 272 с.
9. Завіновська Г.Т. Економіка праці: Навчальний посібник, КНЕУ, 2008. 300 с.
10. Грішнова О. А. Людський розвиток. КНЕУ, 2009. 308 с.
11. А. М. Колот, О. А. Грішнова, Е. А. Гришнова та ін. за заг. ред. Економіка праці та соціально-трудові відносини : підручник, КНЕУ, 2015. 711 с.
12. Лепьохіна, І. О. Теоретичні аспекти формування заробітної плати та її функції. *Вісник ЖДТУ: Економіка, управління та адміністрування*. Ч. 2: 76-78.
13. Education and training – advantages and impact. Investopedia. 2009. URL: <https://www.investopedia.com/articles/economics/09/education-training-advantages.asp> (дата звернення: 20.05.2025)
14. Chart: Education pays, 2024. URL: <https://www.bls.gov/emp/chart-unemployment-earnings-education.htm> (дата звернення: 20.05.2025)

- 15.No experience, no employment. ScienceDirect. 2020. URL: <https://www.sciencedirect.com/science/article/pii/S0272775720305513> (дата звернення: 20.05.2025)
- 16.Gender pay gap statistics. Eurostat. 2023. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Gender_pay_gap_statistics (дата звернення: 20.05.2025)
- 17.Wealth gaps across racial and ethnic groups. Pew Research Center. 2023. URL: <https://www.pewresearch.org/2023/12/04/wealth-gaps-across-racial-and-ethnic-groups/> (дата звернення: 20.05.2025)
- 18.Same occupation, different pay: How wages vary. U.S. Bureau of Labor Statistics. 2015. URL: <https://www.bls.gov/careeroutlook/2015/article/wage-differences.htm> (дата звернення: 20.05.2025)
- 19.Wages and labour costs. Eurostat. 2023. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Wages_and_labour_costs (дата звернення: 20.05.2025)
- 20.Effect of job market on the economy. U.S. Bank. 2024. URL: <https://www.usbank.com/investing/financial-perspectives/market-news/effect-of-job-market-on-the-economy.html> (дата звернення: 20.05.2025)
- 21.Demand and supply at work in labor markets. Lumen Learning. 2022. URL: <https://courses.lumenlearning.com/suny-macroeconomics2/chapter/demand-and-supply-at-work-in-labor-markets/> (дата звернення: 20.05.2025)
- 22.Economic indicator. Investopedia, 2023. URL: https://www.investopedia.com/terms/e/economic_indicator.asp (дата звернення: 20.05.2025)
- 23.Wooldridge J. M. Introductory Econometrics: A Modern Approach. — Cengage Learning, 2016. 826 p.

24. Multiple linear regression (MLR). Investopedia. 2023. URL: <https://www.investopedia.com/terms/m/mlr.asp> (дата звернення: 20.05.2025)
25. Z-score normalization: Definition and examples, 2023. URL: <https://www.geeksforgeeks.org/z-score-normalization-definition-and-examples/> (дата звернення: 20.05.2025)
26. Data normalization, 2024. URL: <https://developers.google.com/machine-learning/crash-course/numerical-data/normalization?hl=uk> (дата звернення: 20.05.2025)
27. Lasso and Ridge Regression: A tutorial. DataCamp, 2023. URL: <https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression> (дата звернення: 20.05.2025)
28. Regression metrics, 2023. URL: <https://www.geeksforgeeks.org/regression-metrics/> (дата звернення: 20.05.2025)
29. K-Nearest Neighbors (KNN) Regression with Scikit-Learn, 2024. <https://www.geeksforgeeks.org/k-nearest-neighbors-knn-regression-with-scikit-learn/>
30. Decision tree: Introduction and example, 2023. URL: <https://www.geeksforgeeks.org/decision-tree-introduction-example/> (дата звернення: 20.05.2025)
31. Random forest algorithm in machine learning, 2023. URL: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/> (дата звернення: 20.05.2025)
32. Korean Welfare Panel Study (KOWEPS). Korea Institute for Health and Social Affairs (KIHASA). 2024. URL: <https://www.koweps.re.kr:442/eng/main.do> (дата звернення: 20.05.2025)



Дата звіту 5/28/2025

Дата редагування ---



Звіт не був оцінений

Звіт подібності

метадані

Назва організації

Kyiv National Economic University named after Vadym Hetman KNEU

Заголовок

Гришко_Осипова_плагіат

Автор

Науковий керівник / Експерт

ГришкоОсипова

підрозділ

кафедра математичного моделювання та статистики

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.



КП 1

25

Довжина фрази для коефіцієнта подібності 2



КП 2

9174

Кількість слів



КЦ

69209

Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв		1
Інтервали		0
Мікропробіли		1
Білі знаки		0
Парафрази (SmartMarks)		6