

**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

Галузь знань

Спеціальність

Економічна кібернетика

05 Соціальні та поведінкові науки

051 Економіка

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «Аналіз та прогнозування стану ІТ-ринку України»

(назва теми)

здобувача Степаненко Юлії Олександрівни

(ПІБ, підпис)

Науковий керівник: кандидат економічних наук, доцент Осипова О. І.

(науковий ступінь, учене звання, ПІБ)

(підпис)

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

Завідувач кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.

(підпис)

Київ 2023

**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 О.М.Притоманова (підпис) 20__ р.	ЗАТВЕРДЖУЮ Завідувач кафедри математичного моделювання та статистики Г.І.Великоіваненко (підпис) 20__ р.
---	---

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
здобувачу вищої освіти Степаненко Юлії Олександрівні
очної (денної) форми навчання
на підготовку кваліфікаційної бакалаврської роботи
на тему «Аналіз та прогнозування стану ІТ-ринку України»
Тему затверджено наказом ректора Університету від _____

Кваліфікаційна бакалаврська робота виконується на матеріалах

DOU.ua та платформи Ukrainian Tech Ecosystem Overview
План кваліфікаційної бакалаврської роботи

Розділ 1	Аналітичний огляд стану ІТ-ринку в Україні (назва розділу)
Розділ 2	Теоретичні аспекти аналізу та прогнозування стану ІТ-ринку (назва розділу)
Розділ 3	Прогнозування заробітної плати на ІТ-ринку України (назва розділу)

Об'єкт дослідження:	ІТ-ринок України та заробітна плата ІТ-фахівців
Предмет дослідження:	Стан, процеси, тенденції, особливості функціонування ІТ-ринку України та фактори, які впливають на заробітну плату ІТ-спеціалістів
Мета кваліфікаційної бакалаврської роботи:	Визначення специфіки функціонування ІТ-ринку в Україні, його внеску в економіку країни та прогнозування заробітної плати ІТ-фахівців

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

У розділі 1	Оглянути основні теоретичні підходи та концепції, пов'язані з ІТ-ринком, розкрити ключові терміни, зібрати та проаналізувати актуальні дані про стан ІТ-ринку в Україні, включаючи його обсяг, вклад в економіку України, розподіл за секторами, рівень зайнятості та перспективи розвитку.
У розділі 2	Описати методи машинного навчання, необхідні для подальшого дослідження, висвітлити метрики оцінки точності та похибки моделей машинного навчання, провести SWOT-аналіз ІТ-ринку в Україні.
У розділі 3	Використати методи машинного навчання, такі як регресійні моделі дерева рішень, випадковий ліс, для прогнозування заробітної плати ІТ-фахівців України, враховуючи такі фактори, як досвід, знання англійської мови, освіта, посада, основна спеціалізація, стаж.

**Завдання підготував
науковий керівник**

О.І. Осипова

(підпис)

(ініціали, прізвище)

«_____» _____ 20__ р.

**Завдання одержав
здобувач**

Ю.О. Степаненко

(підпис)

(ініціали, прізвище)

«_____» _____ 20__ р.

РЕФЕРАТ

Кваліфікаційна бакалаврська робота містить 54 сторінки, 5 таблиць, 33 рисунки, список використаних джерел з 38 найменувань.

Об'єктом дослідження є ІТ-ринок України та заробітна плата ІТ-фахівців.

Предметом дослідження є стан, процеси, тенденції, особливості функціонування ІТ-ринку України та фактори, які впливають на заробітну плату ІТ-спеціалістів.

Мета кваліфікаційної бакалаврської роботи: визначення специфіки функціонування ІТ-ринку в Україні, його внеску в економіку країни та прогнозування заробітної плати ІТ-фахівців.

Відповідно до поставленої мети визначені такі завдання:

- проаналізувати сучасний стан ІТ-ринку в Україні;
- оцінити вплив ІТ-ринку на інші галузі економіки країни;
- здійснити SWOT-аналіз ІТ-ринку України;
- побудувати моделі машинного навчання для прогнозування заробітної плати ІТ-фахівців в Україні;
- перевірити ефективність побудованих моделей та їх прогностичну здатність.

Теоретична, методична та практична значущість отриманих результатів полягає в розширенні знань про розвиток та перспективи ІТ-ринку України, їх застосуванні для практичних потреб і розвитку цього ринку, а також у поліпшенні методологічного підходу до аналізу ринку та впровадженні нових аналітичних інструментів для прогнозування заробітної плати ІТ-фахівців.

Рік виконання кваліфікаційної бакалаврської роботи – 2023

Рік захисту кваліфікаційної бакалаврської роботи – 2023

Ключові слова: ІТ-ринок, ВВП, ІТ-послуги, інформаційні технології, ІТ-компанії, SWOT-аналіз, аналіз даних, прогнозування, лінійна регресія, дерево рішень, випадковий ліс, Python.

В і д г у к

про кваліфікаційну бакалаврську роботу
здобувача навчально-наукового інституту «Інститут інформаційних технологій в економіці»
освітньо-професійної програми «Економічна кібернетика»

Степаненко Ю.О.
(прізвище, ініціали)

на тему «Аналіз та прогнозування стану ІТ-ринку України»
(назва теми)

1. Актуальність теми: Тема роботи є вкрай актуальною в сучасному інформаційному суспільстві, оскільки ІТ-ринок постійно зростає і має значний вплив на економіку. Автор вдало обґрунтував актуальність теми шляхом аналізу статистичних показників, огляду ІТ-компаній та їхнього внеску у економіку України. Таке дослідження вирішує важливу проблему і може мати практичне значення для розвитку ІТ-галузі України.

2. Позитивні риси кваліфікаційної роботи: Кваліфікаційна робота відзначається чіткою структурою та систематичним підходом до дослідження. Автор вдало описав основні теоретичні аспекти дослідження ІТ-ринку, детально описав алгоритми машинного навчання, які використовувалися для дослідження заробітної плати ІТ-фахівців. Аналітичний огляд стану ІТ-ринку в Україні виконаний ґрунтовно та об'єктивно, з використанням відповідних статистичних даних та SWOT-аналізу. Автор ретельно описав процес розробки та адаптації моделей, що свідчить про його розуміння та вміння використовувати практичні інструменти для вирішення конкретних завдань.

3. Наявність самостійних розробок автора: Автор проявив творчий підхід та незалежність у своєму дослідженні, пропонуючи оригінальний підхід до аналізу галузі ІТ. Автор виявив глибоке розуміння теоретичних аспектів і практичних викликів, пов'язаних з ІТ-ринком, та вміло застосовував ці знання для досягнення своїх дослідницьких цілей. Всі результати, висновки та рекомендації отримані автором самостійно.

4. Цінність теоретичних висновків та практичних рекомендацій: Робота містить обґрунтовані теоретичні висновки та практичні рекомендації, які мають потенціал для використання державними органами та бізнесом. Дослідження стану ІТ-ринку в Україні, аналіз його внеску у економіку та SWOT-аналіз забезпечують цінну інформацію для прийняття стратегічних рішень в галузі ІТ. Крім того, побудовані моделі прогнозування заробітної плати на ІТ-ринку в Україні можуть бути використані як інструмент для розрахунку оптимальної заробітної плати та управління кадровими ресурсами в ІТ-компаніях та державних органах, що регулюють ринок праці.

5. Наявність недоліків: Робота включає досить широкий спектр тематик, але у деяких аспектах можна було б провести більш глибокий аналіз або розглянути додаткові фактори, що впливають на стан ІТ-ринку. Рекомендується розширити обсяг дослідження та детальніше розглянути деякі аспекти, які можуть бути корисні для поглибленого аналізу стану ІТ-ринку України.

6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: Загалом, робота демонструє високий рівень науково-дослідницьких здібностей автора, його глибоке розуміння теми та є завершеною науковою роботою. Вважаю, що кваліфікаційна бакалаврська робота Степаненко Ю.О. виконана на високому науковому та професійному рівні, відповідає встановленим вимогам, може бути допущена до захисту перед ЕК та заслуговує на високу оцінку.

Науковий керівник:

к.е.н., доцент кафедри математичного моделювання та статистики Осипова Ольга Ігорівна

О.І. Осипова

(підпис)

«___» _____ 2023 р.

3MICT

ВСТУП	3
РОЗДІЛ 1. АНАЛІТИЧНИЙ ОГЛЯД СТАНУ ІТ-РИНКУ В УКРАЇНІ.....	6
1.1 Основні теоретичні поняття ІТ-ринку	6
1.2 Внесок ІТ-ринку у економіку України: аналіз та статистичні показники....	7
1.3 Характеристика ІТ-компаній та кадрового потенціалу в ІТ-галузі.....	14
1.4 ІТ в інших галузях економіки України.....	18
РОЗДІЛ 2. ТЕОРЕТИЧНІ АСПЕКТИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ СТАНУ ІТ-РИНКУ	21
2.1 Опис методів машинного навчання для прогнозування заробітної плати ІТ-фахівців	21
2.2 Оцінка якості та прогностичної здатності моделей машинного навчання.	26
2.3 SWOT-аналіз ІТ-ринку в Україні	28
РОЗДІЛ 3. ПРОГНОЗУВАННЯ ЗАРОБІТНОЇ ПЛАТИ НА ІТ-РИНКУ УКРАЇНИ	29
3.1 Вибір інструментів для реалізації та аналіз даних.....	29
3.2 Побудова лінійної регресії для прогнозування заробітної плати.....	40
3.3 Застосування дерева рішень для прогнозування заробітної плати	43
3.4 Використання алгоритму «випадковий ліс» для прогнозування зарплати	45
3.5 Порівняння побудованих моделей прогнозування заробітної плати.....	46
ВИСНОВКИ.....	49
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	51
ДОДАТКИ.....	55

ВСТУП

Сучасний світ неможливо уявити без інформаційних технологій (ІТ), які проникли в усі сфери нашого життя і відіграють ключову роль у поліпшенні комунікації, ефективності бізнес-процесів, доступу до інформації та розвитку нових інновацій.

ІТ-галузь суттєво впливає не тільки на траєкторію розвитку світу, але й на соціальне благополуччя пересічних громадян, ефективність виробничої та комерційної діяльності підприємств. Дана галузь створює нові можливості для інновацій, ефективності та автоматизації в різних сферах, таких як бізнес, освіта, охорона здоров'я, комунікації та багато інших.

Україна, як країна з високим потенціалом в галузі інформаційних технологій, не є винятком. Адже ІТ-ринок є одним із ключових напрямів розвитку економіки України, оскільки в умовах загальної економічної та політичної кризи цей ринок стрімко розвивається та забезпечує функціонування багатьох суміжних галузей країни.

Впровадження ІТ-технологій допомагає підприємствам забезпечити ефективну обробку даних, покращити процеси управління, збільшити продуктивність та знизити витрати. Це також сприяє зростанню міжнародного співробітництва та підвищенню конкурентоспроможності країни на світовому ринку.

Актуальність даної теми полягає в тому, що для ефективного розвитку ІТ-ринку України необхідно мати чітке уявлення про його поточний стан та можливі майбутні тенденції. Це дослідження є необхідним для прийняття обґрунтованих рішень у сфері інвестицій, планування стратегій розвитку підприємств, розробки кадрової політики та встановлення конкурентних переваг.

Даний аналіз дозволить виявити сильні та слабкі сторони ІТ-ринку України, а також потенційні можливості та загрози, що можуть вплинути на його подальший розвиток. Крім того, розуміння стану та тенденцій розвитку ІТ-ринку України має

важливе значення для уряду, який може створити ефективну регуляторну політику, яка буде сприяти розвитку галузі, а також для академічного середовища, яке може використовувати отримані результати дослідження для покращення освітніх програм та підготовки кваліфікованих фахівців у сфері ІТ.

Питанням розвитку ринку ІТ-послуг займаються багато українських та закордонних учених. Зокрема, даній проблематиці присвячено праці вітчизняних дослідників, серед яких: Булкіна І.А. [1], Вікулова А. О., Савчук В.В. [2], Колобердянко І., Чернявська А. [3] та інші. Також дослідженню питань розвитку ІТ-ринку в Україні приділяється увага і в професійних колах, зокрема Асоціацією «ІТ Ukraine» та спільнотою програмістів dou.ua.

Мета роботи: визначення специфіки функціонування ІТ-ринку в Україні, його внеску в економіку країни та прогнозування заробітної плати ІТ-фахівців.

Для досягнення поставленої мети, були виконані такі завдання:

- проаналізувати сучасний стан ІТ-ринку в Україні;
- оцінити вплив ІТ-ринку на інші галузі економіки країни;
- здійснити SWOT-аналіз ІТ-ринку України;
- побудувати моделі машинного навчання для прогнозування заробітної плати ІТ-фахівців в Україні;
- перевірити ефективність побудованих моделей та їх прогностичну здатність.

Об'єкт дослідження: ІТ-ринок України та заробітна плата ІТ-фахівців.

Предмет дослідження: стан, процеси, тенденції, особливості функціонування ІТ-ринку України та фактори, які впливають на заробітну плату ІТ-спеціалістів.

В рамках дослідження було проведено аналіз літературних джерел і статистичних даних для отримання уявлення про стан ІТ-ринку в Україні. Використано SWOT-аналіз для оцінки сильних і слабких сторін ІТ-ринку та визначення можливостей та загроз. Застосовано методи машинного навчання, зокрема лінійну регресію, дерево рішень та алгоритм "випадковий ліс", для побудови моделей прогнозування заробітної плати ІТ-фахівців. Також, у

дослідженні використовувався кореляційний аналіз, який дозволив встановити зв'язок між різними факторами та заробітною платою ІТ-фахівців, оцінити силу цього зв'язку та визначити ступінь залежності між ними.

Теоретична значущість отриманих результатів полягає в розширенні знань про розвиток та майбутні перспективи ІТ-ринку України.

Практична значущість проявляється в застосуванні отриманих результатів для практичних потреб і розвитку ІТ-ринку України. Результати дослідження надають цінну інформацію та аналітичні інструменти для різних зацікавлених сторін, включаючи урядові органи, бізнес-спільноту, інвесторів та академічне оточення.

Методологічна значущість полягає в тому, що отримані результати сприяють вдосконаленню методологічного підходу до аналізу ІТ-ринку, а також встановленню зв'язків та закономірностей між різними факторами, що впливають на розвиток цього ринку. Використання методів машинного навчання, таких як лінійна регресія, дерева рішень та випадковий ліс, для прогнозування заробітної плати ІТ-фахівців додало новий інструментарій до аналітичного арсеналу дослідження.

До інформаційної бази дослідження відносяться праці вітчизняних та зарубіжних науковців, статистичні дані Державної служби статистики України, Національного банку України, Світового банку, української спільноти розробників DOU.ua та платформи Ukrainian Tech Ecosystem Overview, звіти Асоціації «IT Ukraine» та документація по бібліотекам для Python

Робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків.

РОЗДІЛ 1

АНАЛІТИЧНИЙ ОГЛЯД СТАНУ ІТ-РИНКУ В УКРАЇНІ

1.1 Основні теоретичні поняття ІТ-ринку

Український вчений С. В. Мочерний стверджує, що «Ринок – сукупність продавців і покупців, які вступають між собою в економічні відносини з приводу купівлі-продажу товарів і послуг за допомогою механізму узгодження цін на основі попиту і пропозиції» [4, с. 72]

Інформаційні технології (ІТ) являють собою:

- 1) процеси, методи пошуку, збирання, зберігання, обробки, надання, поширення інформації та способи здійснення таких процесів і методів;
- 2) прийоми, способи і методи застосування засобів обчислювальної техніки при виконанні функцій зберігання, обробки, передачі і використання даних;
- 3) ресурси, необхідні для збирання, обробки, зберігання і розповсюдження інформації [5].

Традиційно ІТ показують як частину сектору «технології, медіа та телекомунікації» (ТМТ) [6]. Для кращого розуміння структури даного сектору та його компонентів було створено рис. 1.1.



Рисунок 1.1 – Структура ринку технологій

Джерело: розроблено автором на основі [6]

У даній роботі основна увага була приділена аналізу ринку інформаційних технологій.

Структура ІТ-сектора сформувалася внаслідок еволюційного процесу і складається з трьох сегментів:

- 1) комп'ютерне обладнання (IT-hardware): персональні комп'ютери, сервери, робочі станції, обладнання передачі даних, принтери;
- 2) програмне забезпечення (Software), у тому числі системне та прикладне;
- 3) комп'ютерні послуги (IT-services): консалтингові послуги у сфері ІТ, впровадження ІТ, підтримка ІТ, поточне управління комп'ютерним обладнанням і ПО, тренінг і навчання в галузі використання ІТ [1].

Отже, ІТ-ринок - сегмент економічної діяльності, що охоплює продукти, послуги та технології, пов'язані з інформаційними технологіями, включаючи розробку, виробництво, дистрибуцію та використання комп'ютерного обладнання, програмного забезпечення та ІТ-послуг.

1.2 Внесок ІТ-ринку у економіку України: аналіз та статистичні показники

«Інформаційні технології стали важливим фактором економічного зростання, національної безпеки та міжнародної конкурентоспроможності. Дана сфера має глибокий вплив на світові тенденції та соціальний добробут звичайних громадян, їх якість життя та благополуччя. Розвиток інформаційних технологій прямо чи опосередковано впливає на всі сфери життя: від розподілу ресурсів до розподілу доходів» [3, с. 2-3].

Мельник Тетяна підкреслює: «Упродовж останніх років український ІТ-сектор впевнено зростає та перетворюється на одного з флагманів економіки України, а також стає неодмінною складовою її позитивного іміджу на міжнародній арені» [7, с. 18].

«Українська ІТ-галузь за останні 25 років зробила шалений ривок уперед. Стартувавши практично з нуля, вона перетворилася на високоінтелектуальну індустрію, де працює майже 300 тисяч фахівців і яка щороку зростає на 25-30%», - зазначає Костянтин Васюк, виконавчий директор Асоціації IT Ukraine [8].

Щоб оцінити внесок ІТ-ринку в економіку країни спочатку слід проаналізувати рівень номінального ВВП та обсяг експорту (табл. 1.1). Ці показники дадуть змогу зрозуміти, наскільки ІТ-ринок важливий для економіки України та який вплив він має на її зростання.

Таблиця 1.1 - Обсяг номінального ВВП та експорту в Україні протягом 2018-2022 рр. (млрд дол. США)

Показник	Роки				
	2018	2019	2020	2021	2022
Номінальний ВВП	130,89	153,88	156,62	200,09	189,39
Експорт товарів та послуг	59,177	63,556	60,707	60,707	57,018
Експорт товарів	43,341	46,091	45,143	63,113	40,912
Експорт послуг	15,836	17,465	15,564	18,391	16,106

Джерело: розроблено автором на основі [9], [10]

Для візуального представлення інформації, на основі даних, наведених в табл. 1.1, було побудовано гістограми (рис. 1.2, 1.3).

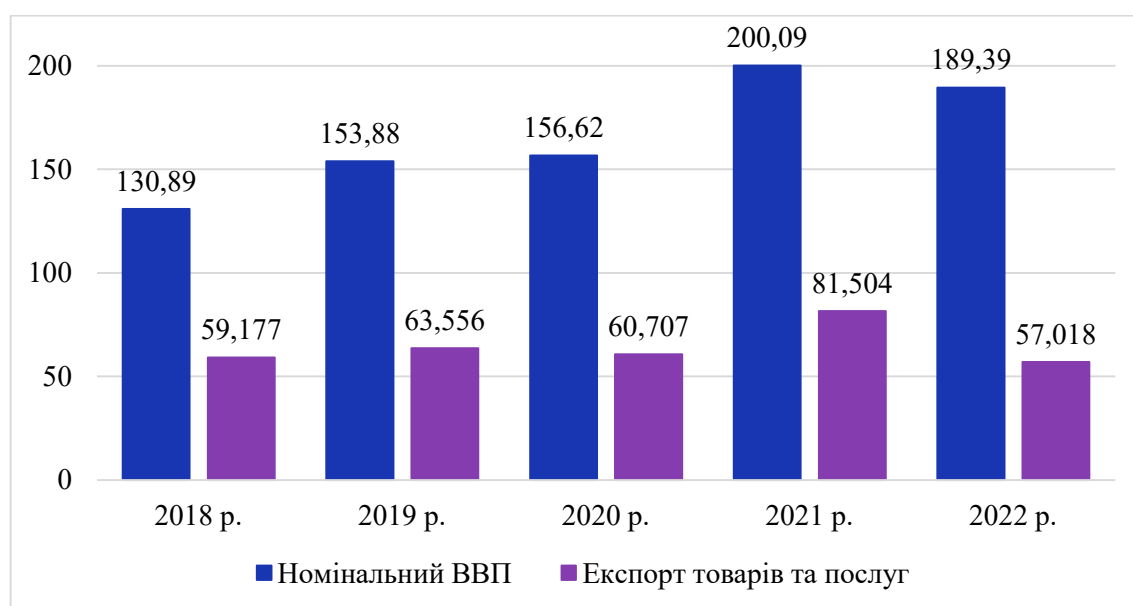


Рисунок 1.2 – Обсяг номінального ВВП та експорту в Україні (млрд дол.)

Джерело: розроблено автором на основі [9], [10]

З побудованої гістограми (рис. 1.2) бачимо, що за обраний період найбільше значення номінального ВВП спостерігалось в 2021 році, не зважаючи на пандемію COVID-19, даний показник перевищив відповідне значення в минулому році майже на 28%. У 2022 році обсяг номінального ВВП знизився на 5% порівняно з попереднім роком, причиною могли бути наслідки повномасштабного вторгнення російських військ в Україну.

Найбільше значення обсягу експорту товарів та послуг в Україні теж припадає на 2021 рік, відбулось зростання на 34% у порівнянні з попереднім роком, а ось у 2022 році даний показник навпаки знизився на 30%. Частка експорту товарів та послуг протягом 2018-2021 рр. становила близько 40-45% номінального ВВП країни, а у 2022 році – 30%.

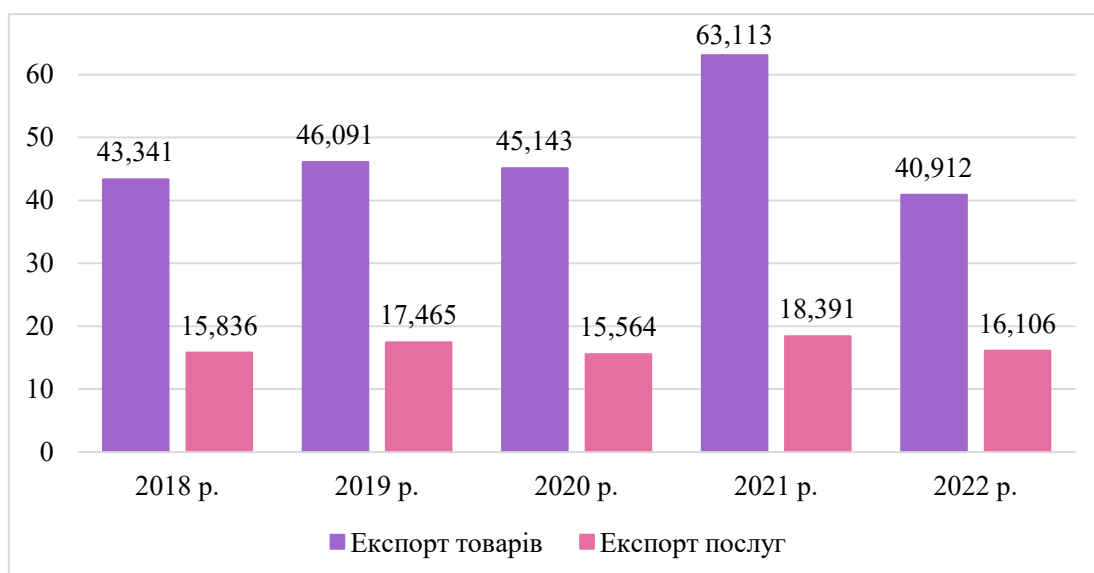


Рисунок 1.3 - Експорт товарів та послуг в Україні (млрд дол. США)

Джерело: розроблено автором на основі [10]

Експорт товарів значно перевищував обсяг експорту послуг в Україні протягом обраного періоду, як це видно з рис. 1.3. Найбільша різниця між даними показниками спостерігалась в 2021 році, тоді обсяг експорту товарів перевищував експорт послуг в 3,4 рази. В цьому ж році відбулось значне зростання обсягу експорту товарів, на 40% в порівнянні з попереднім роком, при цьому обсяг експорту послуг зріс на 18%, що може свідчити про позитивні зрушення в галузі торгівлі товарами та послугами. У 2022 році навпаки спостерігалось зниження обох

показників, обсяг експорту товарів знизився на 35%, а експорт послуг – на 12%, що могло бути спричиненим вторгненням російських військ в Україну.

Варто додати, що експорт послуг становив близько 9-12 % від обсягу номінального ВВП протягом обраного періоду.

У звіті Національного банку України про зовнішню торгівлю [10] експорт послуг включає послуги з переробки, транспорту (у т.ч. послуги з трубопровідного транспорту), подорожі, комп'ютерні послуги та інші послуги – дані, включені до платіжного балансу (за методологією КПБ6). В табл. 1.2 зібрано інформацію про експорт тих послуг, які займають найбільшу частку в загальному експорті послуг.

Таблиця 1.2 - Експорт послуг в Україні за категоріями (млрд дол. США)

Показник	Рік				
	2018	2019	2020	2021	2022
Комп'ютерні послуги	3,204	4,173	5,026	6,943	7,339
Послуги з переробки матеріалів	1,698	1,640	1,354	1,536	0,932
Транспорт	5,949	6,253	5,089	4,760	3,412
Інші послуги	3,540	3,779	3,739	4,202	3,649
Всього:	15,836	17,465	15,564	18,391	16,106

Джерело: розроблено автором на основі [10]

На підставі даних, наведених в табл. 1.2, було побудовано графіки (рис. 1.4) та гістограму (рис. 1.5), що візуально відображають подану інформацію.

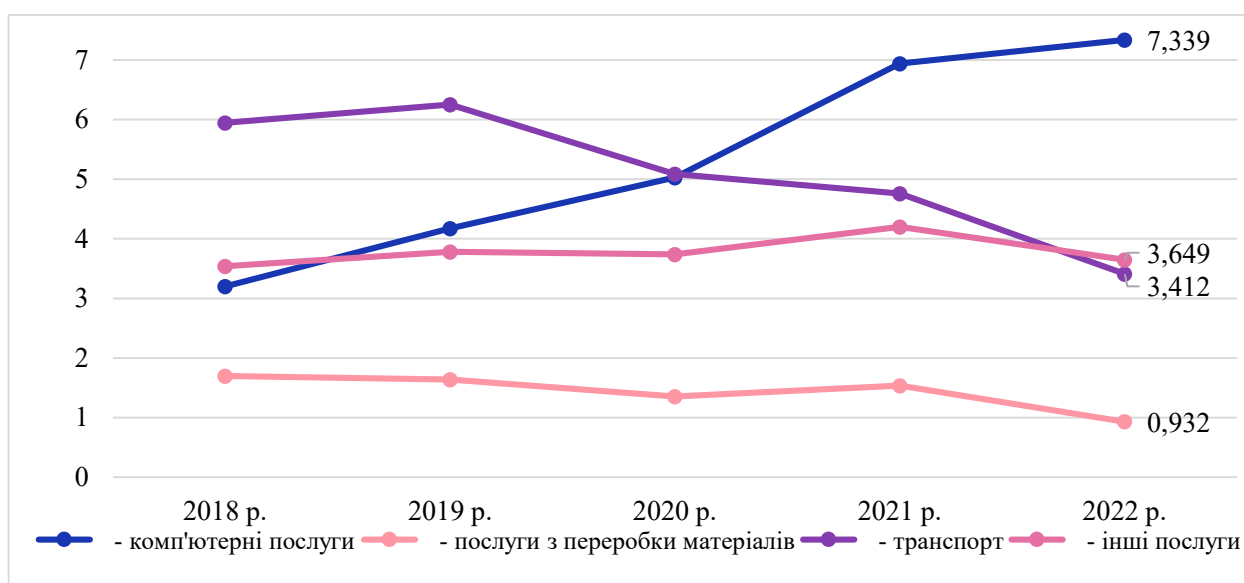


Рисунок 1.4 – Динаміка експорту послуг в Україні (млрд дол. США)

Джерело: розроблено автором на основі [НБУ]

Згідно з рис. 1.4 протягом 2018-2020 рр. найбільша частка експорту послуг в Україні припадала на транспортні послуги (близько третини). У 2021 році обсяг комп'ютерних послуг (ІТ-послуг) перевищив транспортні на 2 млрд дол. США (45,9%) і на даний час займає перше місце серед експорту послуг.

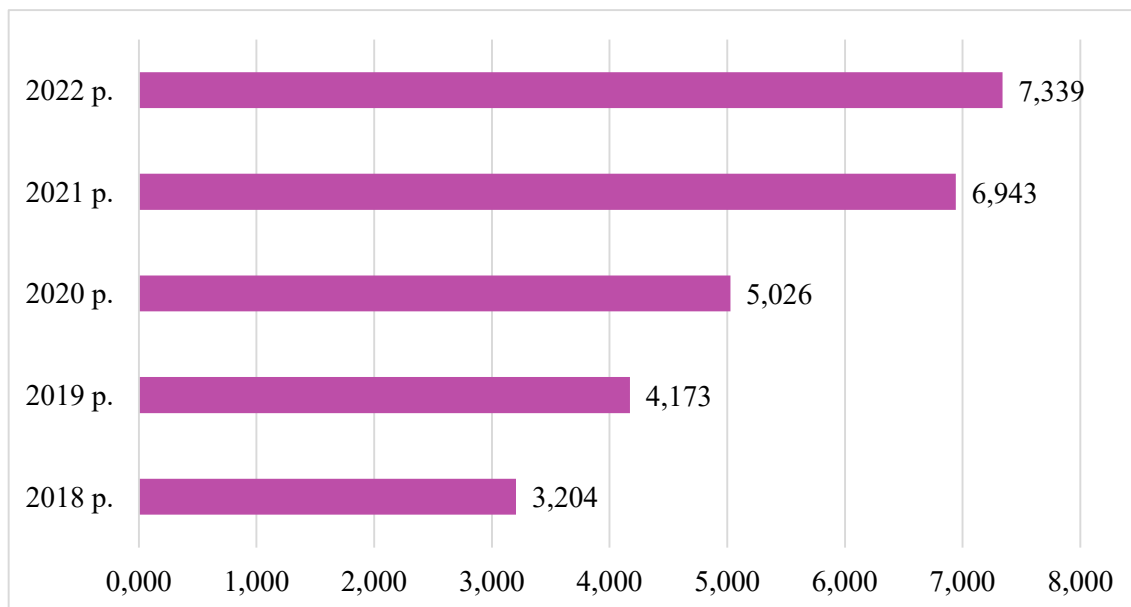


Рисунок 1.5 - Обсяг експорту ІТ-послуг в Україні (млрд дол. США)

Джерело: розроблено автором на основі [10]

Відповідно до рис. 1.5 протягом 2018-2023 рр. спостерігається постійне зростання експорту ІТ-послуг. У 2021 році даний показник зріс майже на 2 млрд дол. США порівняно з попереднім роком, а в 2022 році - майже на 400 млн дол. США.

На основі даних про обсяг номінального ВВП, загального експорту послуг та експорту ІТ-послуг було обчислено частки ІТ-послуг у ВВП країни та загальному обсязі експорту послуг протягом 2018-2022 рр., результати наведено в табл. 1.3.

Таблиця 1.3 – Частка експорту ІТ-послуг (%)

Показник	Рік				
	2018	2019	2020	2021	2022
Частка експорту ІТ-послуг у ВВП	2,4	2,7	3,2	3,5	3,9
Частка ІТ-послуг у загальному експорті послуг	20,2	23,9	32,3	37,8	45,6

Джерело: розроблено автором на основі [9], [10]

Для візуального відображення обчислених значень, наведених в табл. 1.3, було побудовано гістограму (рис. 1.6).

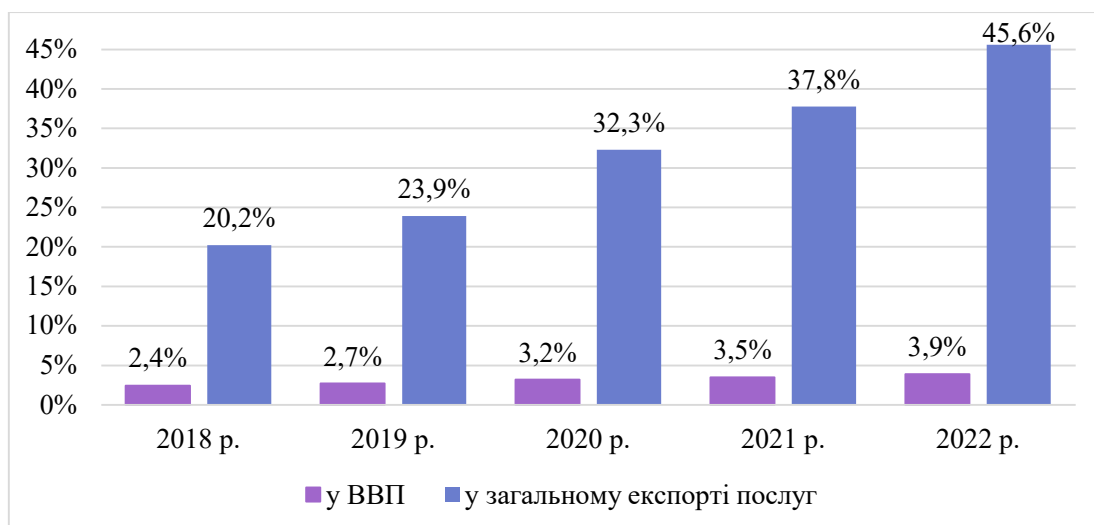


Рисунок 1.6 - Частка експорту ІТ-послуг

Джерело: розроблено автором на основі [9], [10]

З побудованої гістограми (рис. 1.6) бачимо, що у 2022 році експорт ІТ-послуг становив 45,6% від експорту послуг (зріс на 7,8% порівняно з попереднім роком) та складав майже 4% від обсягу ВВП – найбільші значення за обраний період.

До топ-10 партнерів України з експорту комп'ютерних послуг у 2021 році належали США, Велика Британія, Мальта, Ізраїль, Кіпр, Швейцарія, Німеччина, Естонія та Канада (рис. 1.7).

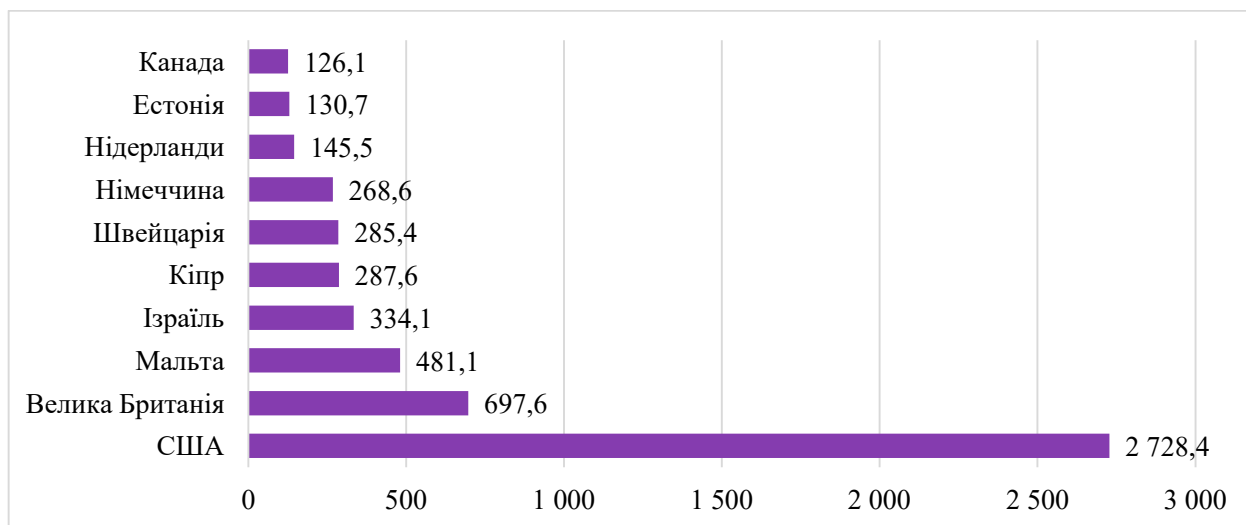


Рисунок 1.7 - Партнери України з експорту ІТ-послуг у 2021 році (млн дол)

Джерело: розроблено автором на основі [11]

ІТ-галузь – одна з небагатьох галузей української економіки, яка змогла не лише зберегти обсяги експорту послуг, а й наростити їх [11].

Кількість ІТ-компаній в Україні має значні відмінності в залежності від джерела даних. На рис. 1.8 наведено кількість активних підприємств з ІТ-КВЕДами в Україні за даними Держстату.

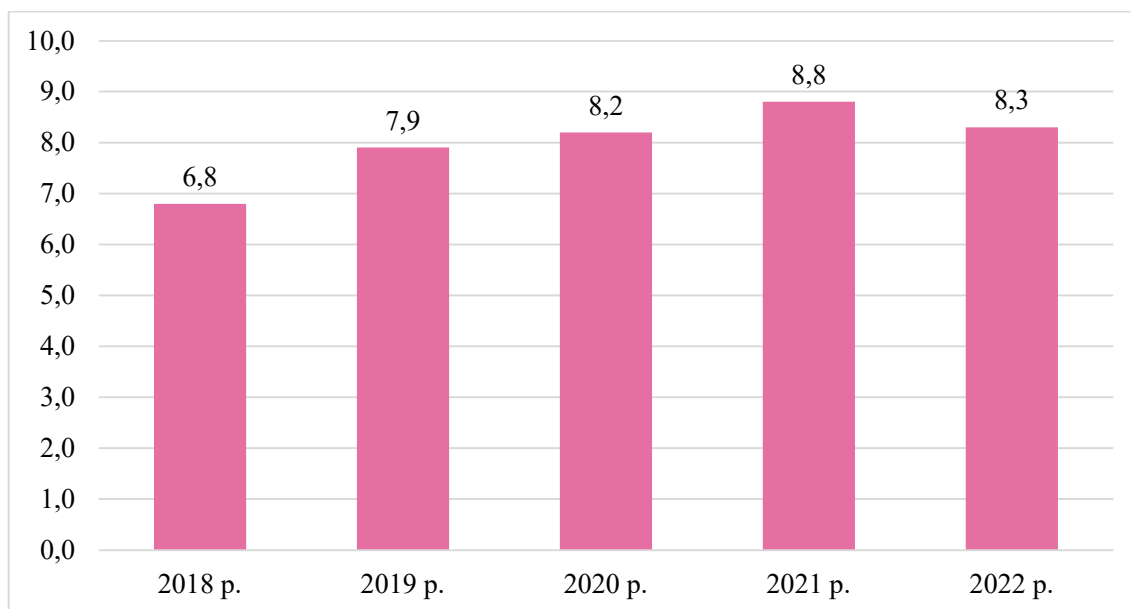


Рисунок 1.8 - Кількість активних підприємств з ІТ-КВЕДами (тис.)

Джерело: розроблено автором на основі [12]

Протягом 2018-2021 років відбувалось постійне збільшення кількості підприємств з ІТ-КВЕДами, а в 2022 році відбулось зменшення даних підприємств близько на півтисячі, після повномасштабного вторгнення.

В аналізі було враховано компанії, що класифікуються за такими КВЕДами-2010:

- 58.21 - Видання комп'ютерних ігор;
- 58.29 - Видання іншого програмного забезпечення;
- 62.01 - Комп'ютерне програмування;
- 62.02 - Консультування з питань інформатизації;
- 62.03 - Діяльність із керування комп'ютерним устаткуванням;
- 62.09 - Інша діяльність у сфері інформаційних технологій і комп'ютерних систем;

- 63.11 - Оброблення даних, розміщення інформації на вебвузлах і пов'язана з ними діяльність;
- 63.12 - Веб-портали [13].

Зокрема, за дослідженням Асоціації IT Ukraine з посиланням на дані порталу *Ukrainian Tech Ecosystem Overview*, станом на початок грудня 2022 року на ринку діє 2,4 тис. компаній. За експертними розрахунками dou.ua та Асоціації IT Ukraine 1,8-2 тис. з них є активними на ринку праці [11].

Ukrainian Tech Ecosystem Overview – це онлайн платформа для пошуку бізнес-інформації IT-компаній, людей, інвесторів та загалом всієї технологічної екосистеми України [14].

1.3 Характеристика IT-компаній та кадрового потенціалу в IT-галузі

За оцінкою порталу *Ukrainian Tech Ecosystem Overview*, станом на початок березня 2023 року в Україні налічувалось 1617 продуктових IT-компаній, 535 сервісних IT-компаній та 90 R&D центрів [15]. Розподіл IT-компаній за типом наведено на рис. 1.9.

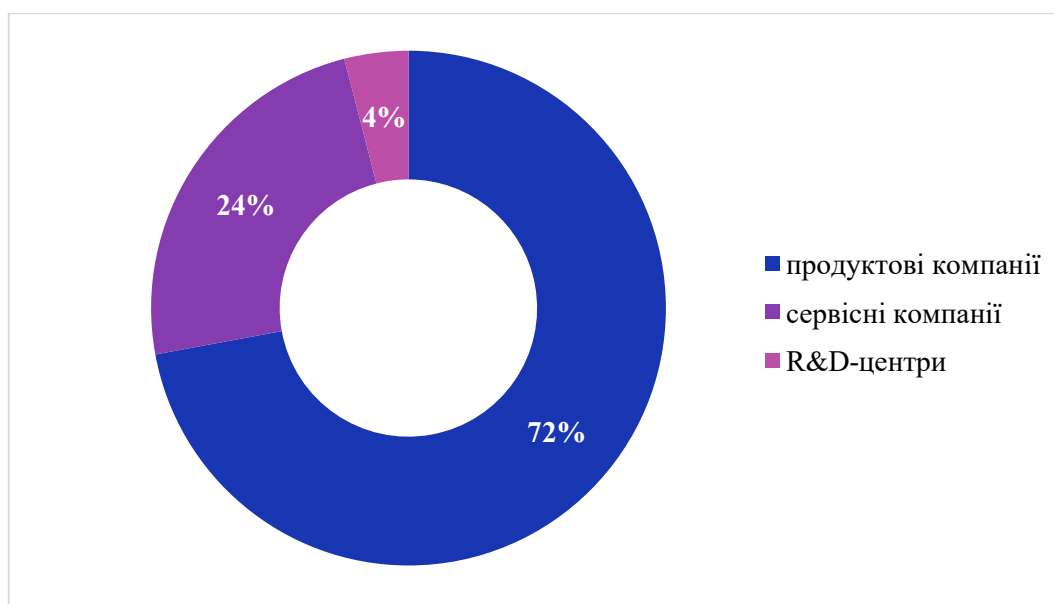


Рисунок 1.9 - Типізація IT-компаній

Джерело: розроблено автором на основі [15]

Отже, найбільшу частку серед ІТ-компаній в Україні займають продуктові компанії (72%), а найменшу – R&D центри (4%). Але варто зазначити, що в Україні також існують сервісно-продуктові ІТ-компанії.

На платформі Ukrainian Tech Ecosystem Overview є можливість переглянути основну інформацію про компанії, що відносяться до конкретного типу:

- назва компанії;
- посилання на сайт компанії;
- напрямки діяльності компанії (вертикалі);
- кількість персоналу;
- розташування;
- статус [16], [17].

Скориставшись зазначеною інформацією та можливістю фільтрувати компанії за кількістю персоналу, розділили продуктові та сервісні ІТ-компанії за кількістю працівників, на такі групи:

- до 50 працівників;
- 51-250 працівників;
- 251-1000 працівників;
- 1001-5000 працівників;
- 5001-10000 працівників (рис. 1.10).

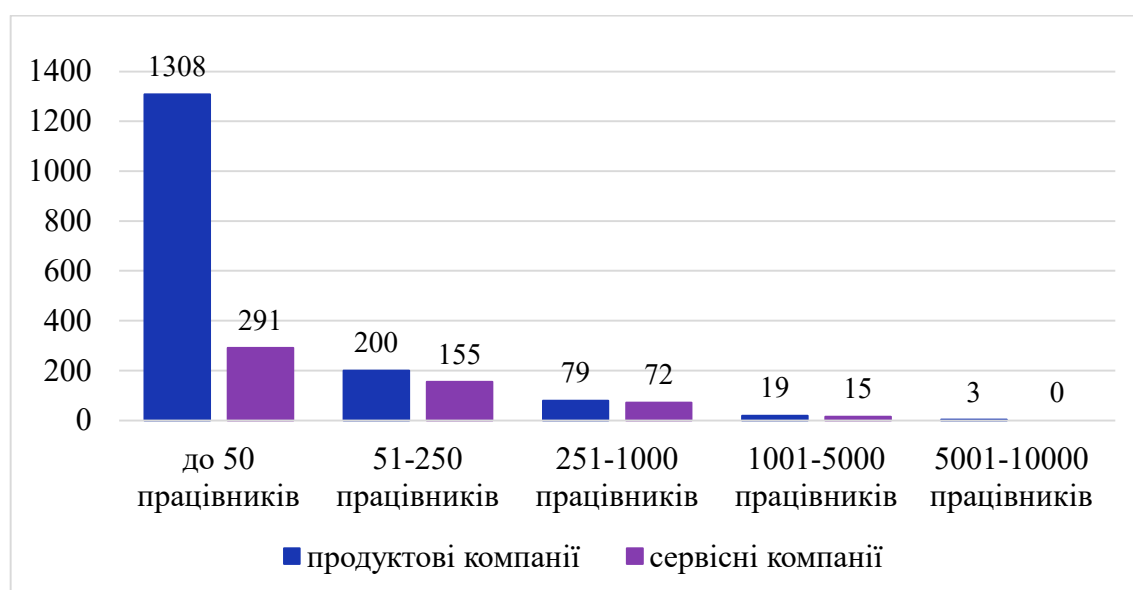


Рисунок 1.10 - Розподіл ІТ-компаній за кількістю працівників

Джерело: розроблено автором на основі [15]

Відповідно до побудованої гістограми (рис. 1.10) більшості продуктових (81,3%) та сервісних (18,1%) ІТ-компаній притаманна невелика кількість персоналу (до 50 працівників). В штаті 12,4% продуктових та 9,6% сервісних компаній працює 51-250 працівників. За поданою інформацією, в Україні немає сервісних ІТ-компаній з персоналом, який перевищує 5000 працівників, а серед продуктових ІТ-компаній, таких лише 3. Помітно, що зі збільшенням кількості працівників зменшується кількість компаній.

На сайті Державної служби статистики України міститься інформація про кількість найманих працівників у суб'єктів господарювання за видами економічної діяльності [18]. Серед наведених видів економічної діяльності було виокремлено лише ті, які відносяться до ІТ-галузі (мають ІТ-КВЕДи, зазначені раніше) та підраховано кількість найманих в них працівників протягом 2018-2021 рр. На основі отриманих даних було побудовано гістограму (рис. 1.11).

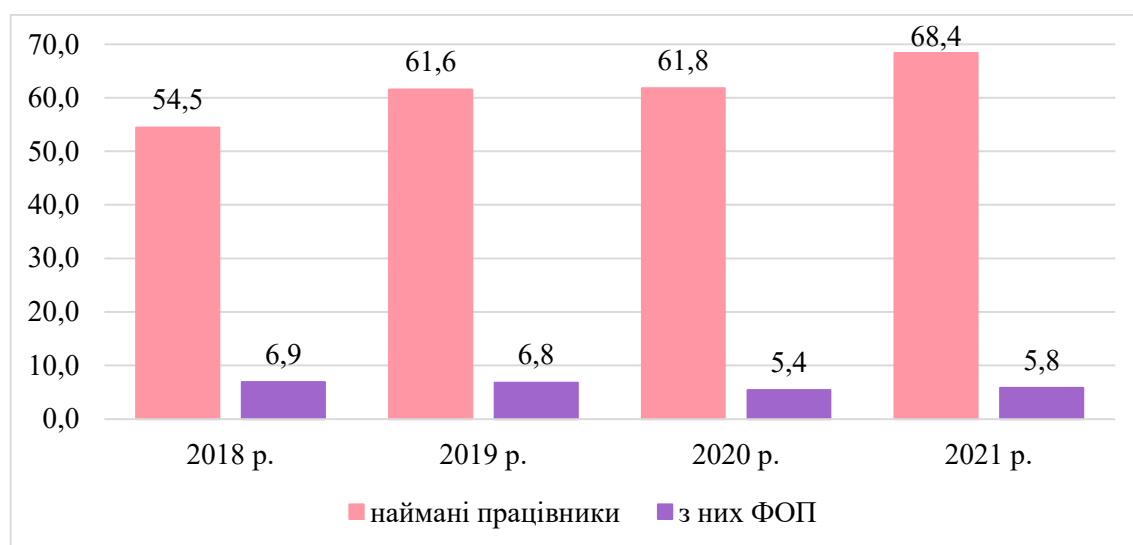


Рисунок 1.11 – Кількість найманих працівників у підприємствах з ІТ-КВЕДами в Україні (тис. осіб)

Джерело: розроблено автором на основі [18]

Згідно з рис. 1.11 кількість найманих працівників у ІТ-секторі зростала з року в рік, лише у 2020 році спостерігалось невелике зниження цього показника (на 0,3% порівняно з попереднім роком), причиною могли бути наслідки пандемії COVID-19, які призвели до скорочення робочих місць. У 2021 році кількість найманих співробітників навпаки підвищилась на 10,7% в порівнянні з попереднім

роком. А ось кількість фізичних-осіб підприємців протягом обраного періоду постійно зменшувалась, що може свідчити про зміни в структурі зайнятості в галузі.

Позитивна тенденція зростання кількості найманих працівників у ІТ-секторі країни вказує на його стабільний розвиток та сприятливі умови для розвитку ІТ-галузі в Україні.

Для оцінки кадрового потенціалу ІТ-галузі України потрібно проаналізувати кількість випускників ІТ-спеціальностей. За основу було взято дані дослідження «Аналіз ІТ-освіти у вишах України» Офісу ефективного регулювання BRDO (рис. 1.12).

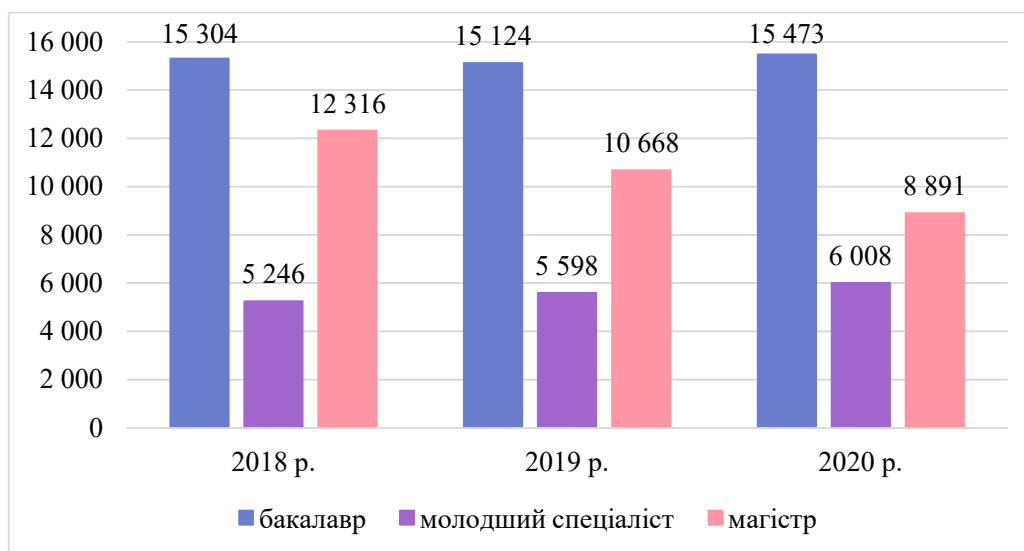


Рисунок 1.12 – Кількість випускників ІТ-спеціальностей за рівнями освіти

Джерело: розроблено автором на основі [19]

Протягом 2018-2020 рр. кількість випускників-бакалаврів майже не змінювалась і в середньому становила 5 300 студентів на рік.

Кількість випускників за освітньою програмою «молодший спеціаліст» має тенденцію до збільшення протягом обраного періоду. У 2020 році дипломи молодших спеціалістів за ІТ-напрямком отримали 6 008 студентів державних ЗВО, що на 14,5% більше, ніж у 2018 році.

Спостерігається значне зменшення (майже на 28%) кількості студентів, які здобули диплом магістра у 2020 році порівняно з 2018 роком.

1.4 ІТ в інших галузях економіки України

ІТ-послуги відіграють важливу роль у розвитку інших галузей економіки України. Вони стають необхідним інструментом для підвищення продуктивності, ефективності та конкурентоспроможності різних секторів.

У сучасному світі, де цифрові технології стають все більш поширеними, ІТ-послуги допомагають підприємствам автоматизувати свої процеси, управляти даними, впроваджувати нові технології та розвивати інноваційні рішення.

Використання інформаційних технологій (ІТ) в *агропромисловості* України останнім часом зростає і стає все більш важливим аспектом сучасного сільськогосподарського сектора. Вони мають значний вплив на всі аспекти агропромислового виробництва, включаючи управління господарством, вирощування рослин, тваринництво, механізацію процесів, моніторинг і прогнозування урожайності, аналіз даних, маркетинг та торгівлю.

Найбільш поширеними та ефективними способами впровадження інформаційних технологій у агропромисловості:

- *точне землеробство* – концепція організації рослинництва, яка використовує технологічні та інформаційні інновації, такі як інтернет речей, для поліпшення точності та ефективності виробничих процесів;
- *точне тваринництво* - застосування інформаційних технологій на основі штучного інтелекту для точного відслідковування потреб і стану тварин, а також для прийняття раціональних управлінських рішень стосовно виробництва продукції і розвитку галузі;
- *smart-менлици* - автоматизація, яка дозволяє контролювати параметри навколишнього середовища, відслідковувати мікроклімат та ефективно використовувати енергетичні ресурси. Отримані дані зберігаються на хмарних платформах, аналізуються штучним інтелектом та автоматично створюють сприятливі умови для виробничого процесу;

- використання *безпілотних літальних апаратів (БПЛА)* для відслідковування стану посівів, зрошення, комплексного відстеження посіву, обробки ріллі та виконання інших технологічних операцій з високою точністю та в найсприятливіші строки [20, с. 3-4].

Серед продуктових ІТ-компаній України, наведених на порталі Ukrainian Tech Ecosystem Overview, 40 відносяться до Agtech-компаній [16].

У сучасному світі використання інформаційних технологій, включаючи концепцію Індустрії 4.0, є важливим елементом розвитку *промисловості*.

Групою науковців та фахівців були окреслені дев'ять розробок (результатів) науково-технічного прогресу, які складають основу Індустрії 4.0 [21]. З огляду на це можна виділити базові технології, впровадження яких є основою розвитку Індустрії 4.0, а саме:

- *Інтернет речей (Internet of Things – IoT)* - це концепція, що використовує Інтернет для обміну інформацією між різними об'єктами, такими як машини, пристрої та датчики, не лише між людьми.
- *промисловий Інтернет речей (Industrial Internet of Things – IIoT)* - це використання Інтернету речей у промисловому середовищі, що дозволяє створювати повністю автоматизовані виробничі процеси.
- *цифрові платформи* – це складні інформаційні системи, що надають доступ клієнтам та партнерам і є логічним розвитком бізнесу в контексті комп'ютерної техніки та широкого поширення пристроїв для підключення до Інтернету;
- *цифрова екосистема* – це комплексний проект, що об'єднує різних учасників, таких як клієнти, розробники додатків, торговці та агенти, а також інформаційні сервіси і бізнес-процеси.;
- *аналітика великих даних (Data Driven Decision)* або просто *Великі дані (Big Data)* - це обробка і аналіз великого обсягу інформації, який накопичується внаслідок цифрового перетворення фізичного світу. Цей процес вимагає використання комп'ютерів з використанням хмарних обчислень та технологій штучного інтелекту (AI). [22, с. 4-6].

Усе більше українських промислових підприємств впроваджують у свою діяльність технології Індустрії 4.0.

Згідно з порталом Ukrainian Tech Ecosystem Overview, в Україні діє 40 Agtech-компаній [16].

На порталі Ukrainian Tech Ecosystem Overview зазначено, що серед продуктових ІТ-компаній України є 42 компанії зі сфери Advanced manufacturing (Industry 4.0) [16].

Застосування інформаційних технологій суттєво змінило традиційні методи ведення бізнесу і привело до зародження нової форми економічної діяльності - *електронного бізнесу* [23, с. 278]. ІТ відіграє важливу роль у ритейлі, сприяючи автоматизації та оптимізації бізнес-процесів.

Ритейл не просто адаптується до онлайн-середовища, а проходить трансформацію завдяки використанню інформаційних технологій:

- використання *технологій доповненої реальності (AR/VR/MR)* дозволяє забезпечити більш глибоку взаємодію з клієнтом. Наприклад, можна приміряти одяг на віртуальному аватарі, візуалізувати нові меблі в кімнаті та інші подібні можливості;
- застосування *технологій аналізу великих обсягів даних (Big Data)* дозволяє прогнозувати попит та продажі, розробляти стратегії ціноутворення, оптимізувати асортимент товарів згідно з потребами ринку та багато іншого;
- впровадження *роботизації* дозволяє збільшити ефективність виробництва, уникнути впливу людського фактору, підвищити конкурентоспроможність, оптимізувати витрати на операційну діяльність та багато іншого [11, с. 26].

У *фінансовому секторі*, інформаційні технології грають ключову роль у забезпеченні безпеки та надійності операцій, автоматизації банківських процесів та електронному обслуговуванні клієнтів.

У *сфері медицини та охорони здоров'я*, ІТ сприяє поліпшенню діагностики, лікування та управління медичними даними.

У галузі *освіти* ІТ розширюють можливості навчання та доступу до знань.

РОЗДІЛ 2

ТЕОРЕТИЧНІ АСПЕКТИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ СТАНУ ІТ-РИНКУ

2.1 Опис методів машинного навчання для прогнозування заробітної плати ІТ-фахівців

Далі було охарактеризовано методи машинного навчання, які в подальшому використовувалися для прогнозування заробітної плати: лінійна регресія, дерева рішень та випадковий ліс.

Регресійний аналіз – це статистичний метод для дослідження та моделювання зв'язку між змінними. Застосування регресії є численним та зустрічається майже в усіх галузях, включаючи техніку, фізичні та хімічні науки, економіку, менеджмент, біологію та соціальні науки [24, с. 1].

Регресійний аналіз широко використовується в інтелектуальному аналізі даних і є основним інструментом науки та аналітики даних. Завдяки широкому застосуванню до низки проблем регресійний аналіз може бути найпоширенішим статистичним методом [24, с. 1].

Рівняння лінійної регресії має такий вигляд:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (2.1)$$

де y – залежна змінна;

β_0 – точка перетину лінії регресії з віссю y ;

β_1 – нахил лінії регресії;

x – незалежна змінна;

ε – статистична похибка (випадкова змінна, яка пояснює нездатність моделі точно відповідати даним).

Поняття «залежна змінна» та «незалежна змінна» часто викликають плутанину з поняттям статистичної незалежності, тому x називаємо *предиктором* або *регресорною змінною*, а y – *змінною відгуку*. Оскільки рівняння (2.1) включає

лише одну регресорну змінну, його називають моделлю простої лінійної регресії [24, с. 2].

Параметри β_0 і β_1 зазвичай називаються коефіцієнтами регресії. Ці коефіцієнти мають просту і часто вживану інтерпретацію. Значення параметра β_0 вказує на середнє значення залежної змінної y , коли незалежна змінна x дорівнює нулю. Нахил β_1 — це зміна середнього значення залежної змінної y при змінні незалежної змінної x на одиницю [24, с. 13].

Іншими словами, коефіцієнт β_1 визначає, яким чином залежна змінна реагує на зміну незалежної змінної в регресійній моделі. Якщо β_1 позитивний, це означає, що збільшення значення x супроводжується збільшенням y . Зворотно, якщо β_1 від'ємний, збільшення значення x призведе до зменшення y .

Множинна лінійна регресія розширює просту лінійну регресію, щоб включити більше ніж одну пояснювальну змінну. В обох випадках все ще використовуємо термін «лінійний», оскільки припускаємо, що змінна відгуку безпосередньо пов'язана з лінійною комбінацією пояснювальних змінних [25, с. 10].

Рівняння для множинної лінійної регресії має таку ж форму, як і для простої лінійної регресії, але містить більше членів:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon, \quad (2.2)$$

де y — змінна відгуку;

β_0 — прогнозоване значення y , коли всі пояснювальні змінні дорівнюють 0;

$\beta_1, \beta_2, \dots, \beta_p$ — вплив кожного предиктора на змінну відгуку;

$x_1, x_2, \dots, x_p$ — предиктори або регресорні змінні;

ε — статистична похибка.

У моделі з p пояснювальними змінними кожна пояснювальна змінна має власний β -коефіцієнт [25, с. 11].

У інтелектуальному аналізі даних *дерево рішень* є прогностичною моделлю, яка може використовуватися для представлення як класифікаторів, так і регресійних моделей. У дослідженні операцій, з іншого боку, *дерева рішень*

відносяться до ієрархічної моделі рішень та їх наслідків. Особа, яка приймає рішення, використовує дерева рішень, щоб визначити стратегію, яка, швидше за все, досягне її мети [26, с. 33].

Коли дерево рішень використовується для завдань класифікації, його більш доречно називати деревом класифікації. Коли воно використовується для завдань регресії, воно називається деревом регресії [26, с. 33].

Дерево рішень складається з вузлів, які утворюють кореневе дерево, а саме, орієнтоване дерево з вузлом, який називається «коренем», і який не має вхідних ребер. Усі інші вузли мають рівно одне вхідне ребро. Вузол із вихідними ребрами називають «внутрішнім» вузлом або «тестовим» вузлом. Усі інші вузли називаються «листями» (також відомі як «кінцеві» вузли або вузли «рішення») [26, с. 35-36].

У дереві рішень кожен внутрішній вузол розділяє простір екземплярів на два або більше підпросторів відповідно до певної дискретної функції значень вхідних атрибутів. У найпростішому та найпоширенішому випадку кожен тест розглядає один атрибут, так що простір екземплярів розділено відповідно до значення атрибутів. У випадку числових атрибутів умова відноситься до діапазону [26, с. 36].

Для візуального представлення структури дерева рішень було побудовано блок-схему, яка відображає основні елементи моделі (рис. 2.1).

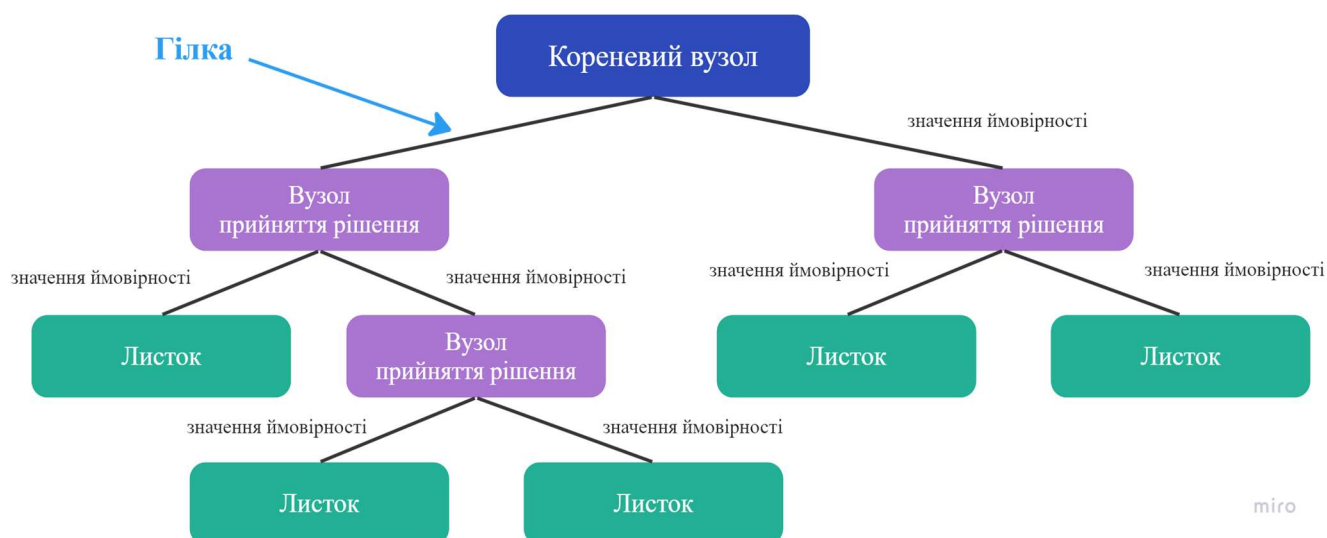


Рисунок 2.1 – Базова структура дерева рішень

Джерело: розроблено автором на основі [27]

Екземпляри класифікуються шляхом переходу від кореня дерева до листка відповідно до результатів тестів уздовж шляху. Класифікація починається з кореня дерева, потім розглядається ознака, яка відповідає кореню, і визначається, якій гілці відповідає спостережуване значення даної характеристики. Потім розглядаємо вузол, у якому з'являється дана гілка. Повторюємо ті ж операції для цього вузла, поки не дійдемо до листка [26, с. 36-37].

На рис. 2.2 зображено приклад використання дерева рішень для оцінки ризику виникнення серцевого нападу.

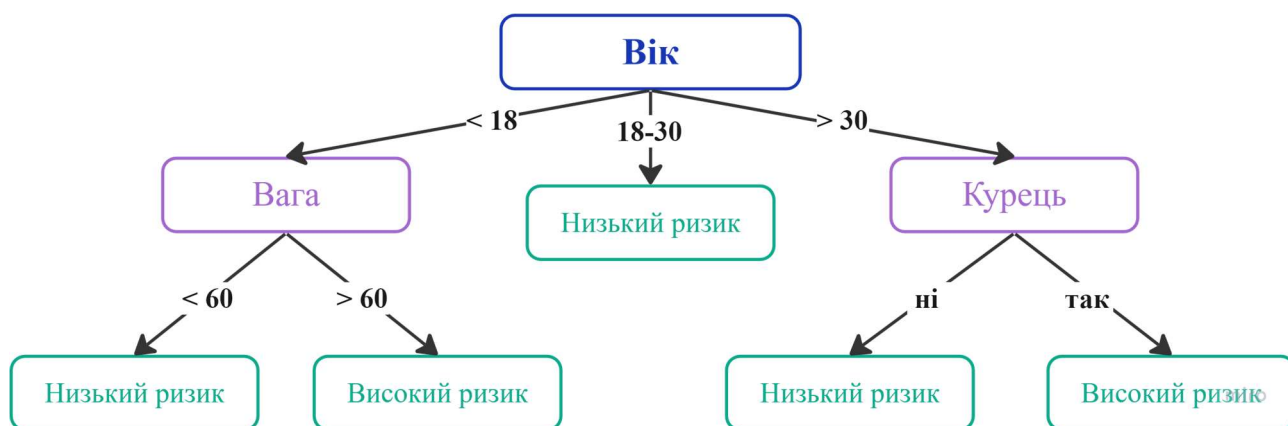


Рисунок 2.2 – Оцінка ризиків для запобігання серцевому нападу

Джерело: розроблено автором на основі [27]

Складність дерев рішень залежить від кількості записів і атрибутів у наданих даних. Однак, дерева рішень є вільними від розподілу або непараметричними методами, що означає, що вони не залежать від припущень щодо розподілу ймовірностей [27].

Випадковий ліс — це комбінація предикторів дерев, у яких кожне дерево залежить від значень випадкового вектора, відібраного незалежно та з однаковим розподілом для всіх дерев у лісі [28].

Випадковий ліс — контрольований алгоритм машинного навчання, який можна використовувати для вирішення проблем класифікації та регресії. Дерева рішень, створені за допомогою традиційних методів, зазвичай не можуть бути дуже складними, оскільки це може призвести до втрати точності узагальнення за межами навчальних даних. Алгоритм випадкового лісу намагається вирішити це [29].

Замість побудови єдиного дерева рішень, модель випадкового лісу будує безліч дерев за допомогою випадкового вибору підмножини ознак з набору даних. Завдяки такому підходу кожне дерево будується по-різному і не буде надто складним. Потім обчислюється середнє значення оцінок усіх дерев [29].

Для візуального відображення структури моделі «випадковий ліс» було побудовано відповідну блок-схему (рис. 2.3).

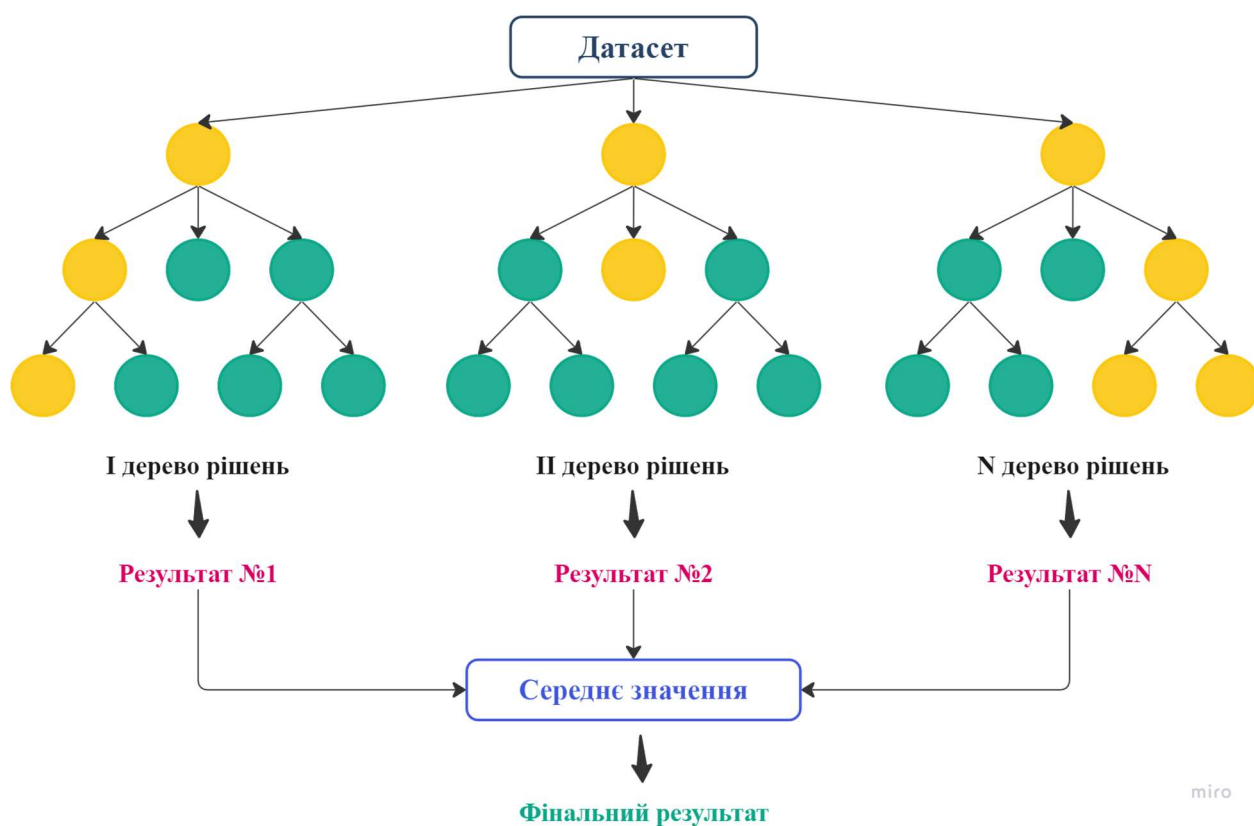


Рисунок 2.3 – Структура випадкового лісу

Джерело: розроблено автором на основі [27]

На рис 2.3 бачимо, що випадковий ліс складається з N -ї кількості дерев рішень, які працюють разом як так званий ансамбль. Кожне окреме дерево рішень робить прогноз, наприклад результат класифікації, і ліс використовує середнє значення отриманих результатів, тобто фінальний результат – той, який підтримується більшістю дерев рішень, як прогноз для всього ансамблю.

Секрет випадкового лісу полягає в так званому принципі «мудрості натовпу». Основна ідея полягає в тому, що рішення багатьох завжди краще, ніж рішення окремої людини чи одного дерева рішень.

2.2 Оцінка якості та прогностичної здатності моделей машинного навчання

Метод регресії передбачає елемент X_i для відповідного елемента Y_i наявного набору даних [30, с. 4].

Визначимо дві константи: $\bar{Y}$ - середнє значення реальних значень (1.3) та MST - середню загальну суму квадратів (1.4).

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i, \quad (1.3)$$

де m – кількість спостережень або значень змінної Y в вибірці;

Y_i – окреме спостереження або фактичне i -те значення змінної Y у вибірці.

$$MST = \frac{1}{m} \sum_{i=1}^m (Y_i - \bar{Y})^2 \quad (1.4)$$

Для оцінки точності моделі використовуємо значення коефіцієнта детермінації R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^m (X_i - Y_i)^2}{\sum_{i=1}^m (\bar{Y} - Y_i)^2}, \quad (1.5)$$

де X_i – передбачене i -те значення.

Коефіцієнт детермінації можна інтерпретувати як частку дисперсії залежної змінної, яку можна передбачити на основі незалежних змінних [30, с. 5]. Значення R^2 близьке до 1 вказує на вміле пояснення варіації залежної змінної.

При оцінці похибки моделі застосовуємо такі метрики:

- MSE – середня квадратична похибка:

$$MSE = \frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2 \quad (1.6)$$

Середню квадратичну похибку (MSE) можна використовувати, якщо є викиди, які необхідно виявити. Якщо модель врешті-решт видає дуже поганий прогноз, частина функції, що зводиться до квадрату, збільшує помилку [30, с. 5].

MSE може приймати значення від 0 до нескінченності. Чим менше значення даного показника, тим краще модель прогнозує дані.

- RMSE – квадратний корінь з середньої квадратичної похибки:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2} \quad (1.7)$$

Дві величини MSE і $RMSE$ пов'язані між собою (через квадратний корінь). Порядок регресійних моделей на основі MSE буде ідентичним порядку моделей на основі $RMSE$ [30, с. 5]. Чим ближче значення $RMSE$ до 0, тим кращою є модель.

- MAE – середня абсолютна похибка:

$$MAE = \frac{1}{m} \sum_{i=1}^m |X_i - Y_i| \quad (1.8)$$

Середню абсолютну похибку (MAE) можна використовувати, якщо викиди представляють пошкоджені частини даних. Якщо тестовий набір також має багато викидів, продуктивність моделі буде посередньою [30, с. 5-6].

Найкращим значенням середньої абсолютної похибки є 0, що вказує на повну точність моделі в узгодженні з даними. Чим більше значення даного показника, тим більшим є відхилення моделі від даних і менш точними результати моделювання.

- MAPE – середня абсолютна похибка у відсотках:

$$MAPE = \frac{1}{m} \sum_{i=1}^m \left| \frac{Y_i - X_i}{Y_i} \right| \quad (1.9)$$

Середня абсолютна похибка у відсотках (MAPE) є ще одним показником продуктивності для регресійних моделей, який має дуже інтуїтивну інтерпретацію відносної похибки [30, с. 6].

Однак він також має низку недоліків, найбільш критичними з яких є обмеження його використання виключно з додатними даними і упередженість до низьких прогнозів, що робить його непридатним для прогнозних моделей, де очікуються великі помилки [30, с. 6].

2.3 SWOT-аналіз ІТ-ринку в Україні

Для вивчення потенціалу та визначення перспектив розвитку ІТ-ринку в Україні, було проведено SWOT-аналіз, тобто виявлено сильні та слабкі сторони ринку, а також можливості та загрози, що можуть вплинути на його подальший розвиток.

Таблиця 2.1 – SWOT-аналіз ІТ-ринку України

Сильні сторони (S)	Слабкі сторони (W)
- висококваліфіковані фахівці; - велика кількість талановитих студентів в галузі ІТ-технологій; - низькі витрати на розробку програмного забезпечення та ІТ-послуги порівняно з країнами Європи та США; - сприятливе географічне розташування, що дозволяє працювати з клієнтами з різних часових зон; - наявність іноземних інвесторів та партнерів	- недостатня підтримка держави; - недостатня кількість кваліфікованих фахівців; - відтік висококваліфікованих кадрів; - низький рівень заробітної плати для ІТ-фахівців порівняно з іншими країнами; - нестабільність економічної ситуації в країні; - відсутність потужного внутрішнього ринку для продуктів та послуг вітчизняних ІТ-компаній.
Можливості (O)	Загрози (T)
- розвиток технологій штучного інтелекту та машинного навчання; - можливість привернення іноземних інвестицій та партнерств з іноземними компаніями; - збільшення використання хмарних технологій та Інтернету речей; - покращення освіти та підготовки кваліфікованого персоналу для ІТ-ринку.	- збільшення кількості кібератак та крадіжок даних; - відсутність світла, зв'язку та Інтернету через ракетні обстріли; - зменшення кількості фахівців у ІТ-сфері через еміграцію або призов на фронт; - зниження інвестицій в Україну через несприятливу політичну та економічну ситуацію

Джерело: розроблено автором

РОЗДІЛ 3

ПРОГНОЗУВАННЯ ЗАРОБІТНОЇ ПЛАТИ НА ІТ-РИНКУ УКРАЇНИ

3.1 Вибір інструментів для реалізації та аналіз даних

В даному розділі було застосовано мову програмування Python та інтерактивне середовище для програмування Jupyter Notebook, які дозволяють проводити аналіз даних, виконувати візуалізацію результатів та будувати моделі машинного навчання для прогнозування обраних показників.

В коді було використано такі бібліотеки та модулі:

- *pandas* – це бібліотека з відкритим вихідним кодом, ліцензована BSD, що надає високопродуктивні, прості у використанні структури даних та інструменти аналізу даних для мови програмування Python [31];
- *matplotlib* – це комплексна бібліотека для створення статичних, анімованих та інтерактивних візуалізацій на Python [32];
- *statsmodels* – це модуль Python, який надає класи та функції для оцінки багатьох різних статистичних моделей, а також для проведення статистичних тестів і дослідження статистичних даних [33];
- *numpy* – це бібліотека Python, яка надає об’єкт багатовимірного масиву, різні похідні об’єкти (такі як замасковані масиви та матриці), а також набір процедур для швидких операцій над масивами, включаючи математичні, логічні, основні статистичні операції, випадкове моделювання та багато іншого [34];
- *scikit-learn* – це бібліотека машинного навчання з відкритим кодом, яка підтримує контрольоване та неконтрольоване навчання [35];
- *seaborn* – це бібліотека візуалізації даних Python на основі matplotlib. Вона забезпечує інтерфейс високого рівня для малювання привабливої та інформативної статистичної графіки [36].

Для подальшого аналізу ІТ-ринку в Україні було використано результати анонімного опитування, яке проводилось в другій половині 2022 року фахівцями DOU.ua, де було зібрано 13 316 анкет від ІТ-фахівців усіх рівнів і спеціалізацій [37].

Дані були взяті з відкритого джерела - репозиторію devua/csv/salaries на Github [38], який належить українській спільноті розробників DOU.ua. Кожен рядок датасету містить відповіді на запитання конкретного респондента. Код програми з підготовкою та аналізом даних наведено у Додатку А.

Обраний датасет містить 13315 рядків та 28 стовпців, але не всі дані необхідні для подальшої роботи, тому було залишено лише такі стовпці: «В якій області ви зараз живете?», «Зарплата у \$\$\$ за місяць, лише ставка після сплати податків», «Ваша стать», «Ваш вік», «Яка у вас освіта?», «Знання англійської мови», «Ваш тайтл», «Основна мова програмування», «Вкажіть вашу основну спеціалізацію», «В якій сфері працюєте?», «Тип компанії» та «Загальний стаж роботи за спеціальністю».

Потім було перейменовано стовпці датасету: «Current region of residence», «Net monthly salary in \$», «Gender», «Age», «Education level», «English level», «Title», «Primary programming language», «Main specialization», «Industry», «Company type» та «Total work experience in the field (years)». Вигляд датасету після перейменування стовпців став таким, як показано на рис. 3.1.

	Current region of residence	Net monthly salary in \$	Gender	Age	Education level	English level	Title	Primary programming language	Main specialization	Industry	Company type	Total work experience in the field (years)
0	Івано-Франківськ чи область	4000	Чоловік	29	Вища (бакалавр, спеціаліст, магістр)	Intermediate	Team Lead	NaN	NaN	Blockchain	Стартап	8 років
1	NaN	NaN	Чоловік	36	Вища (бакалавр, спеціаліст, магістр)	Advanced	Middle	Ruby	Full Stack розробка	Public services / Government	Продуктова	3 роки
2	Київ чи область	2500	Чоловік	34	Вища (бакалавр, спеціаліст, магістр)	Elementary	Middle	C# / .NET	Full Stack розробка	Adtech / Advertising, Security, Telecom	Продуктова	4 роки
3	Львів чи область	500	Жінка	23	Вища (бакалавр, спеціаліст, магістр)	Pre-Intermediate	Junior	NaN	NaN	Telecom	Продуктова	Пів року
4	Житомир чи область	2800	Жінка	33	Вища (бакалавр, спеціаліст, магістр)	Upper-Intermediate	Senior	NaN	NaN	Automotive	Продуктова	6 років
...	...	...	...	...	...	...	...	...	...	...	...	...

Рисунок 3.1 – Частина вихідних даних датасету [38]

Джерело: розроблено автором на основі [38]

Серед респондентів переважали чоловіки (76,8%), тоді як жінки становили всього 23,2% (рис. 3.2).

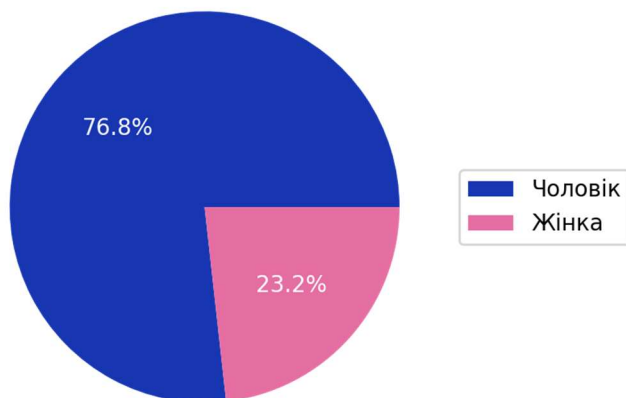


Рисунок 3.2 – Статевий розподіл респондентів

Джерело: розроблено автором на основі [38]

Далі було видалено NaN елементи в стовпцях «Net income per month in \$», «Main specialization», значення в стовпці «Total work experience in the field (years)» були змінені так, щоб можна було змінити тип даних на числовий

Для отримання загального уявлення про структуру та зв'язки між даними, було проведено детальний візуальний аналіз даних.

Спершу побудували гістограму вікового розподілу респондентів (рис. 3.3):

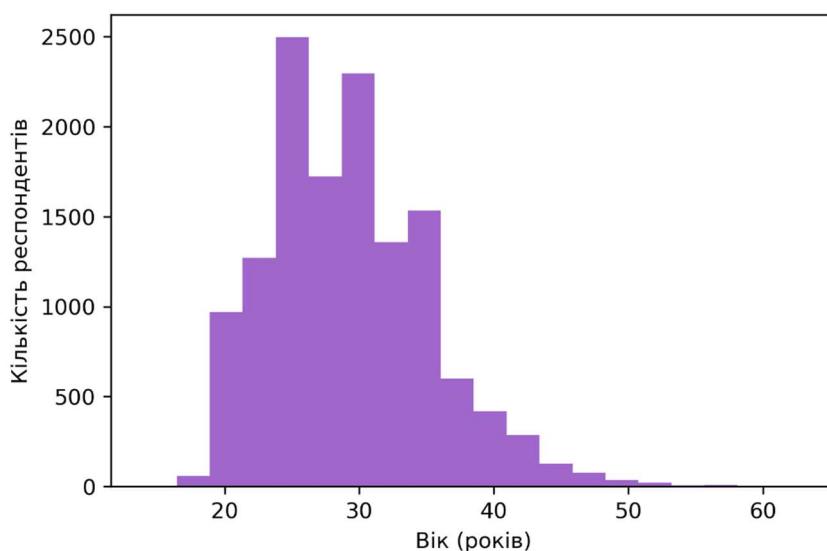


Рисунок 3.3 – Вікова структура респондентів

Джерело: розроблено автором на основі [38]

Побудована гістограма (3.3) свідчить про те, що віковий розподіл учасників опитування має схожість з нормальним розподілом, але спостерігається незначна правостороння асиметрія, що означає, що більшість респондентів мають вік, менший за середній. Найбільша кількість респондентів знаходиться у віковому діапазоні між 20 і 35 років, тоді як найменша кількість спостерігається у віковій групі 45 років і старше, що може бути пов'язано з різними факторами, такими як професійний розвиток, зміна кар'єри або пенсія.

Кількість респондентів зменшувалась після досягнення піку у середній віковій групі (30-35 років). В даних також є невелика частка опитаних старше 60 років, що свідчить про те, що серед опитаних є досвідчені фахівці, які продовжують працювати після пенсійного віку.

На підставі обчислених моди, середнього значення та медіани можна зробити такі висновки: найбільше серед респондентів було осіб у віці 25 років. Середній вік учасників дослідження склав 29,2 років, а медіанне значення віку становило 29 років, тобто половина респондентів має вік менше 29 років, а інша половина - більше 29 років.

Далі побудували гістограму розподілу місячної заробітної плати (рис. 3.4).

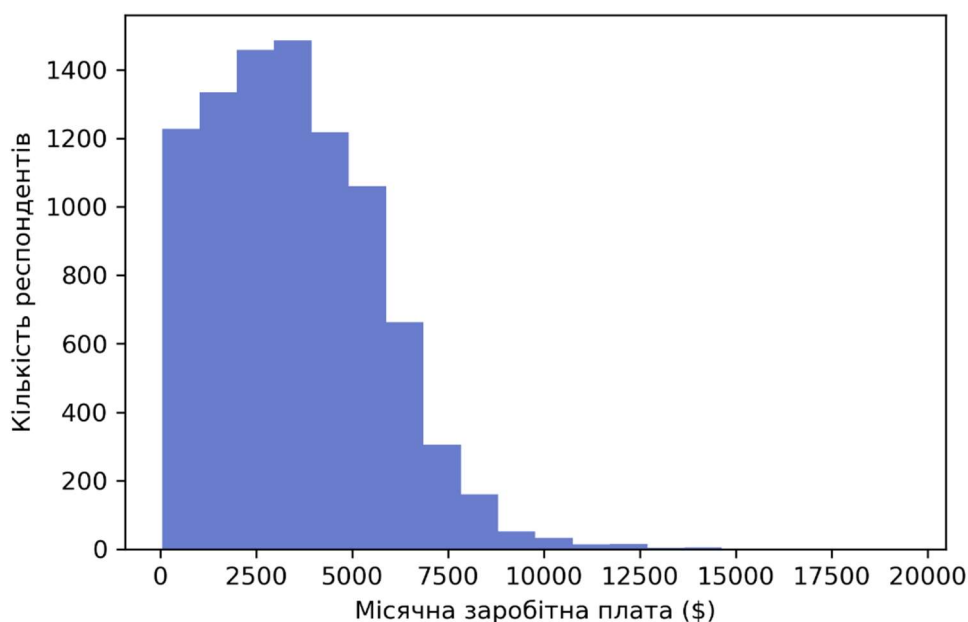


Рисунок 3.4 – Розподіл місячної заробітної плати респондентів

Джерело: розроблено автором на основі [38]

Відповідно до побудованої гістограми (рис. 3.4) більшість респондентів заробляють в місяць до 5000 доларів. Розподіл місячної заробітної плати відрізняється від нормального закону розподілу і спостерігається довгий хвіст праворуч від середнього значення, що свідчить про правосторонню асиметрію.

Високі значення заробітної плати, які спостерігаються на правому кінці гістограми, можуть вказувати на наявність групи високооплачуваних посад або окремих індивідуальних випадків з високими доходами (досвідчених працівників, які досягли високого рівня в своїй кар'єрі).

Велика кількість респондентів з низькими заробітними платами може бути пов'язаною з демографічними особливостями аудиторії, їх професійним рівнем або рівнем доходів у відповідних секторах ринку праці.

Найчастіше респонденти в якості місячної заробітної плати обирали значення 3000 доларів (мода вибірки). Середня місячна заробітна плата серед респондентів складала 3444,3 доларів, а медіана – 3200 доларів, що означає, що половина опитаних має значення заробітної плати нижче цієї величини.

Було побудовано гістограму розподілу стажу роботи учасників опитування (рис. 3.5).

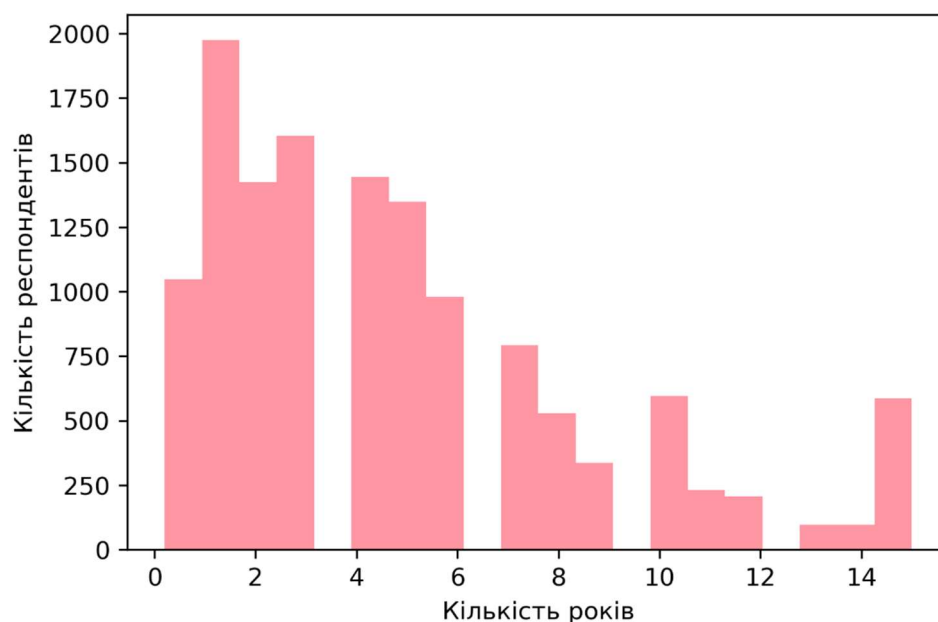


Рисунок 3.5 – Розподіл стажу роботи за спеціальністю респондентів

Джерело: розроблено автором на основі [38]

На рис. 3.5 перший стовпець відповідає за кількість учасників опитування зі стажем роботи менше 3 місяців, 3 місяці або пів року (в датасеті було замінено на 0.2, 0.25 та 0.5 відповідно), а другий стовпець - кількість респондентів, які мають стаж роботи за поточною спеціальністю тривалістю 1 або 1,5 року.

Значна кількість респондентів, які ще не мають значного досвіду або тільки почали свою кар'єру, вказує на те, що в IT-сфері є багато новачків і дана галузь не втрачає популярності.

Найбільша кількість опитаних мала стаж роботи, який становив 3 роки. Середній стаж за спеціальністю серед респондентів складав 4,8 років, а медіанне значення стажу роботи – 4 роки.

Дослідимо залежність середньої заробітної плати респондентів від стажу роботи (рис. 3.6).

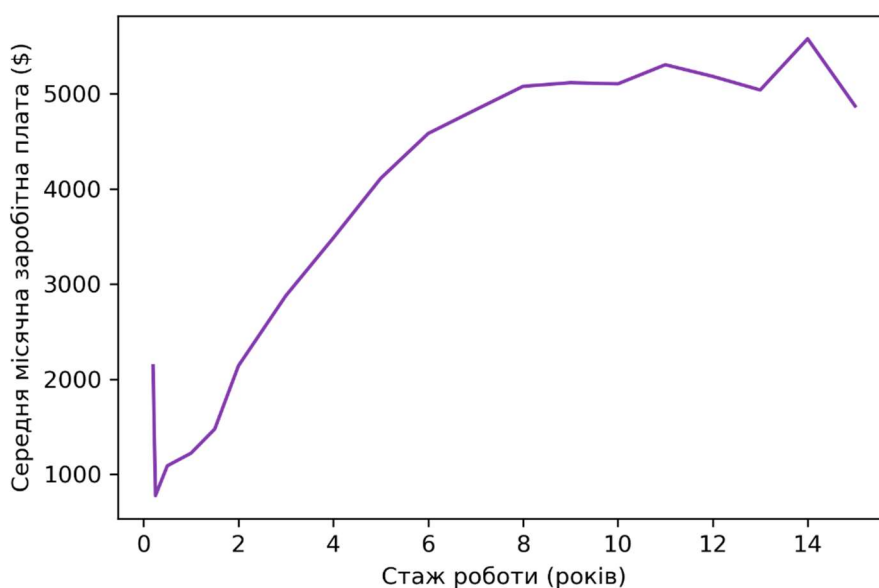


Рисунок 3.6 – Залежність середньої зарплати респондентів від стажу роботи

Джерело: розроблено автором на основі [38]

З графіка (рис. 3.6) видно, що середня заробітна плата респондентів мала тенденцію зростання зі збільшенням стажу роботи. Це може свідчити про те, що досвід роботи є важливим фактором для отримання високої зарплати. Однак цей висновок потребує подальшого аналізу і перевірки на інших даних. При стажі до 10 років спостерігається значний різкий ріст середньої заробітної плати, потім зростання середньої заробітної плати стає менш інтенсивним.

На розмір місячної заробітної плати також може впливати основна спеціалізація ІТ-фахівця, тому варто проаналізувати дану залежність (рис. 3.7).

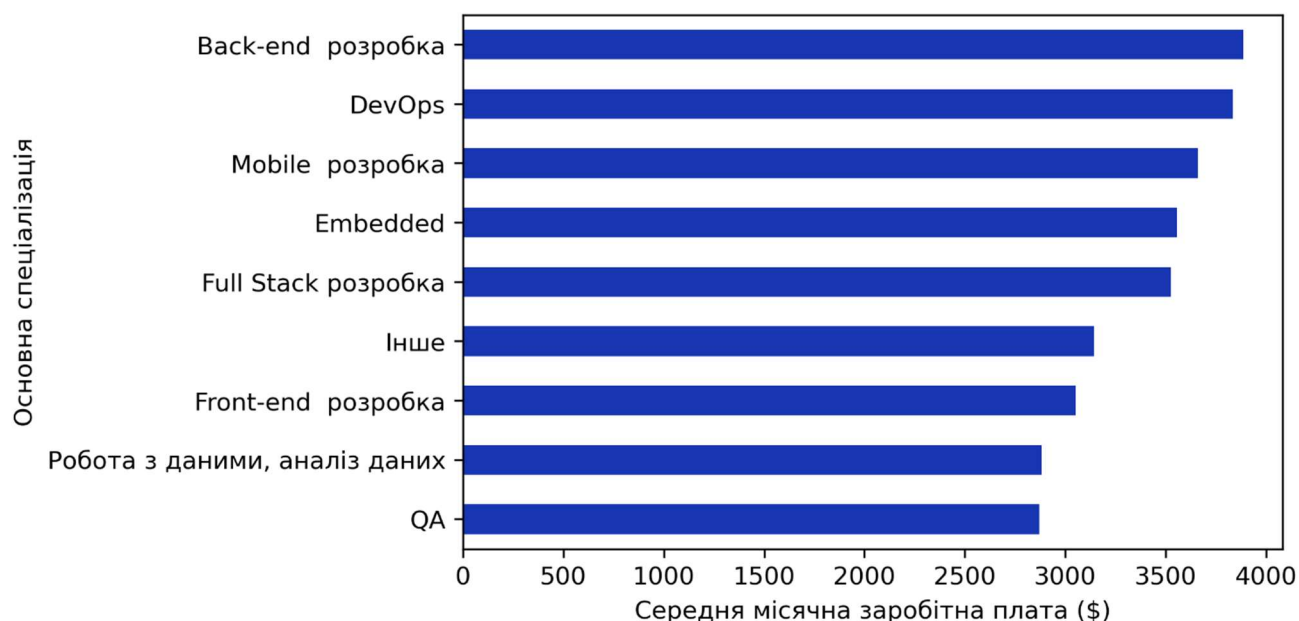


Рисунок 3.7 – Залежність середньої зарплати від основної спеціалізації

Джерело: розроблено автором на основі [38]

Відповідно до побудованої гістограми (рис. 3.7) середня заробітна плата ІТ-фахівців суттєво варіюється в залежності від основної спеціалізації. Найвищу середню заробітну плату мають респонденти, які працюють у сфері Back-end розробки та DevOps, що може бути пов'язано з високим попитом на фахівців у цих галузях або рівнем їх кваліфікації.

Середні заробітні плати працівників сфери QA та підтримки є нижчими порівняно з іншими спеціалізаціями, що може бути обумовлено меншою складністю роботи або нижчим попитом на фахівців у цих галузях.

Деякі спеціалізації, наприклад «Embedded» і «Full Stack розробка», мають помірну середню заробітну плату, що може відображати баланс між попитом на фахівців у цих галузях та рівнем вимог до компетенцій.

Побудований графік підкреслює значення основної спеціалізації при визначенні рівня заробітної плати.

Також варто дослідити залежність середньої місячної заробітної плати між різними регіонами України (рис. 3.8).

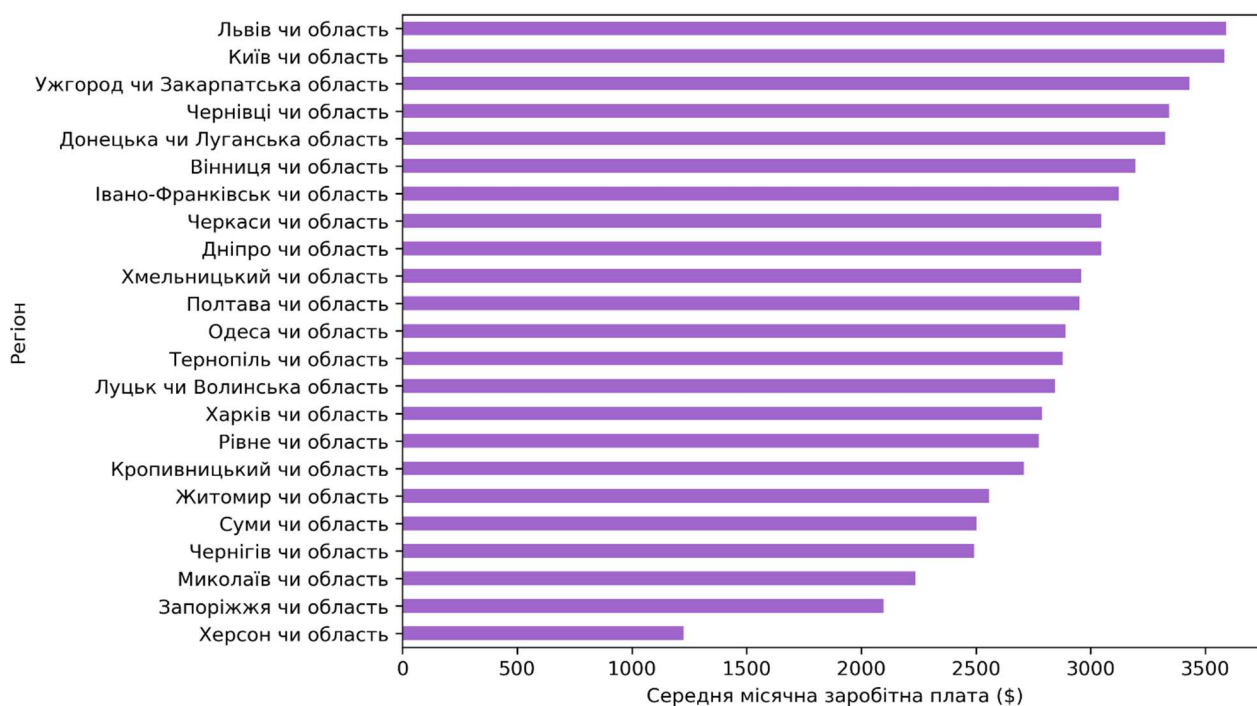


Рисунок 3.8 – Розподіл середньої зарплати між різними регіонами України

Джерело: розроблено автором на основі [38]

З побудованої гістограми (рис. 3.8) видно, найбільша частина областей має середню зарплату в межах 2-2,5 тисяч доларів.

До топ-3 регіонів з найвищими середніми зарплатами серед ІТ-фахівців відносяться Київська, Львівська та Чернівецька області - області, які мають більш розвинену економіку або центри певних галузей (середня заробітна плата становить близько 3 тисяч доларів). Регіоном з найнижчими середніми зарплатами є Херсонська область (близько 1 тисячі доларів), причиною цього може слугувати окупація більшої частини області російськими військами, на момент анкетування.

Спостерігаємо значний регіональний розрив в українському ринку праці в ІТ-сфері, може бути пов'язано з рівнем розвитку регіону, економічними умовами, наявністю індустрій або рівнем життя.

Далі було проведено розвідувальний аналіз, який дає змогу проаналізувати основні властивості даних, знайти в них загальні закономірності і аномалії.

Спочатку за допомогою методу `describe()` переглянули основну статистику числових показників датасету (рис. 3.9).

	Net monthly salary in \$	Age	Total work experience in the field (years)
count	9033.000000	9033.000000	9033.000000
mean	3444.298990	29.224953	5.101716
std	2126.369224	6.236632	3.882178
min	50.000000	14.000000	0.200000
25%	1700.000000	25.000000	2.000000
50%	3200.000000	29.000000	4.000000
75%	5000.000000	33.000000	7.000000
max	19500.000000	63.000000	15.000000

Рисунок 3.9 – Основна статистика числових показників датасету

Джерело: розроблено автором на основі [38]

Згідно з отриманими результатами (рис. 3.9) межі 1-го, 2-го та 3-го квантилів всіх змінних виглядають цілком реалістично, але мінімальне та максимальне значення змінних «Net monthly salary in \$» та «Total work experience in the field (years)» – ні, тому варто провести діагностику викидів та нетипових спостережень, для цього було побудовано коробчасті діаграми для числових показників.

Проаналізувавши побудовані коробчасті діаграми, було вирішено видалити ті значення, які найбільше відрізнялись від інших – інформацію про респондентів, середня заробітна плата яких більша 12500 доларів, старших 50 років та стаж роботи яких понад 14 років. Після видалення зазначених значень основна статистика числових показників мала такий вигляд (рис. 3.10):

	Net monthly salary in \$	Age	Total work experience in the field (years)
count	8559.000000	8559.000000	8559.000000
mean	3357.953006	28.590606	4.577848
std	2057.098622	5.613844	3.232746
min	50.000000	14.000000	0.200000
25%	1650.000000	24.000000	2.000000
50%	3000.000000	28.000000	4.000000
75%	4800.000000	32.000000	7.000000
max	12500.000000	50.000000	14.000000

Рисунок 3.10 – Основна статистика числових показників датасету після видалення значень, які найбільше відрізнялись від інших

Джерело: розроблено автором на основі [38]

Для отримання більш точних результатів, було видалено рядки, в яких значення в стовпцях «Title», «Company type» та «Main specialization» відповідно дорівнювали «Немає тайтлу», «Інша» та «Інше». Після очищення даних в датасеті залишилось 8063 рядки.

Перед побудовою моделей прогнозування заробітної плати, всі категоріальні змінні, які мали бути задіяні в моделях, було перетворено на числові (рис. 3.11).

	Current region of residence	Net monthly salary in \$	Gender	Age	Education level	English level	Title	Primary programming language	Main specialization	Industry	Company type	Total work experience in the field (years)
2	Київ чи область	2500.0	1	34	4	1	3	C# / .NET	4	Adtech / Advertising, Security, Telecom	3	4.0
5	Київ чи область	5500.0	1	31	4	2	6	PHP	7	E-commerce, Fintech / Banking	3	8.0
7	Львів чи область	1200.0	1	24	4	4	2	C# / .NET	4	E-commerce, Logistics	2	1.5
8	Львів чи область	650.0	1	27	4	3	2	Python	1	Automotive	2	1.0
10	Дніпро чи область	6000.0	1	25	4	4	6	TypeScript	7	Dating, E-commerce, Travel	3	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...

Рисунок 3.11 – Вигляд датасету після внесених змін

Джерело: розроблено автором на основі [38]

При побудові лінійної регресії та інших моделей було використано такі змінні: «Net monthly salary in \$», «Gender», «Age», «Education level», «English level», «Title», «Main specialization», «Company type» та «Total work experience in the field (years)». При чому залежною змінною було обрано «Net monthly salary in \$» (Місячна заробітна плата після сплати податків), а всі інші змінні – незалежними.

Для оцінки взаємозв'язку між залежною за незалежними змінними було побудовано кореляційну матрицю (рис. 3.12).

Відповідно до побудованої кореляційної матриці найбільший позитивний вплив на розмір місячної заробітної плати (Net monthly salary in \$) мають змінні «Title», тобто посадове звання (коефіцієнт кореляції $r = 0,76$) та «Total work experience in the field (years)» - загальний стаж роботи за спеціальністю ($r = 0,63$).

І навпаки найменший вплив на розмір місячної заробітної плати мають «Gender» - стать ($r = 0,13$), «Company type» - тип компанії ($r = 0,18$) та «Main specialization» - основна спеціалізація ($r = 0,18$).

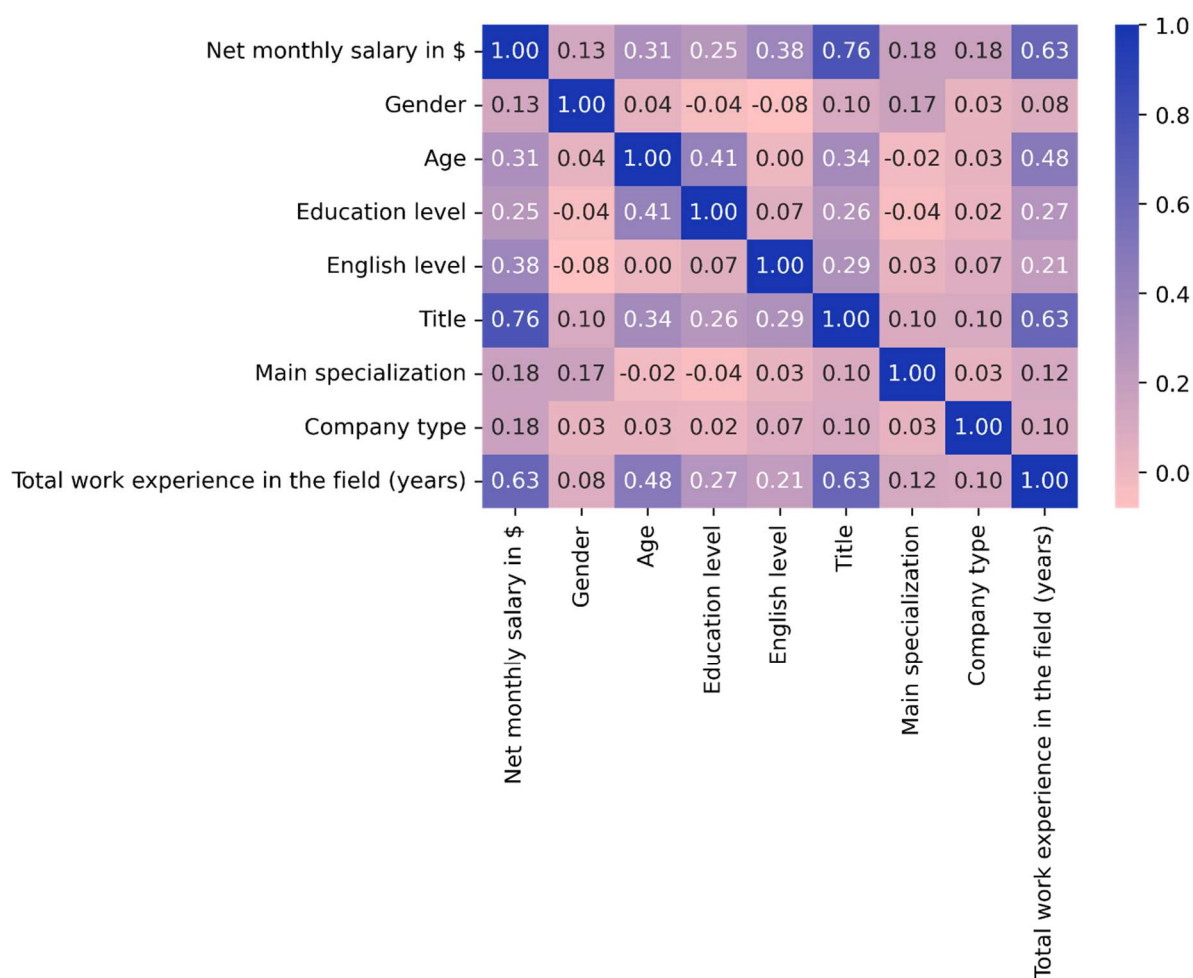


Рисунок 3.12 – Кореляційна матриця для ознак числового типу

Джерело: розроблено автором на основі [38]

Для подальшого моделювання дані було розбито на навчальну (train) та тестову (test) вибірки шляхом випадкового розподілу даних, де тестова вибірка складатиме 20% від загальної кількості даних (1613 спостережень), а навчальна - 80% даних (6450 спостережень).

При цьому кожний об'єкт має однакову ймовірність бути обраним в одну з груп:

- train – частина даних, на яких навчається модель;
- test – частина даних, яка не буде приймати участь в навчанні моделей, на цю частину не буде ніякого впливу. На цій частині даних необхідно валідувати нашу модель та робити фінальні висновки щодо предиктивних властивостей моделі.

3.2 Побудова лінійної регресії для прогнозування заробітної плати

За допомогою класу OLS бібліотеки statsmodels було побудовано модель лінійної регресії та оцінено її параметри за допомогою методу 1МНК. Результати основної регресійної статистики зображено на рис. 3.13.

OLS Regression Results

Dep. Variable:	Net monthly salary in \$	R-squared:	0.657				
Model:	OLS	Adj. R-squared:	0.657				
Method:	Least Squares	F-statistic:	1929.				
Date:	Thu, 01 Jun 2023	Prob (F-statistic):	0.00				
Time:	05:47:36	Log-Likelihood:	-68604.				
No. Observations:	8063	AIC:	1.372e+05				
Df Residuals:	8054	BIC:	1.373e+05				
Df Model:	8						
Covariance Type:	nonrobust						
		coef	std err	t	P> t	[0.025	0.975]
	const	-2753.8858	116.429	-23.653	0.000	-2982.117	-2525.655
	Gender	410.6736	44.853	9.156	0.000	322.750	498.597
	Age	-2.4366	2.946	-0.827	0.408	-8.211	3.337
	Education level	111.3434	16.352	6.809	0.000	79.290	143.397
	English level	413.5169	16.586	24.932	0.000	381.004	446.030
	Title	608.2518	10.024	60.680	0.000	588.602	627.901
	Main specialization	67.2426	5.318	12.645	0.000	56.819	77.666
	Company type	187.7100	15.438	12.159	0.000	157.448	217.972
	Total work experience in the field (years)	144.1646	5.770	24.986	0.000	132.854	155.475
Omnibus:	1445.086	Durbin-Watson:	1.958				
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5261.341				
Skew:	0.873	Prob(JB):	0.00				
Kurtosis:	6.552	Cond. No.	272.				

Рисунок 3.13 – Основна статистика для моделі лінійної регресії

Джерело: розроблено автором на основі [38]

Рівняння отриманої лінійної регресії має вигляд:

$$Y = -2753,886 + 410,674 \cdot X_1 - 2,437 \cdot X_2 + 111,343 \cdot X_3 + 413,517 \cdot X_4 + \\ + 608,252 \cdot X_5 + 67,243 \cdot X_6 + 187,710 \cdot X_7 + 144,165 \cdot X_8,$$

де Y – заробітна плата за місяць;

X_1 – стать;

X_2 – вік;

X_3 – рівень освіти;

X_4 – рівень англійської мови;

X_5 – посада (тайтл);

X_6 – основна спеціалізація;

X_7 – тип компанії;

X_8 – досвід роботи за спеціалізацією.

Якість отриманої моделі оцінимо за показником коефіцієнта детермінації. R^2 приймає значення 0,657 або 65,7%. Це означає, що розрахункові параметри моделі на 65,7% пояснюють залежність між параметрами, що досліджуються, і можуть існувати інші важливі фактори або змінні, не включені в модель, які також можуть вплинути на заробітну плату за місяць.

Рівень значимості F-критерію (Prob(F-statistic)) даної моделі дорівнює 0.00, це значення менше ніж 0.05, тобто модель в цілому є адекватною, достовірною і її можна використовувати, також можна довіряти її результатам.

Коефіцієнт -2753.886 показує, яким буде значення змінної «Net monthly salary», якщо всі незалежні змінні в моделі, що розглядаються, будуть рівні 0.

Згідно з отриманою описовою статистикою (рис. 3.13) між усіма незалежними змінними, крім «Age» існує прямий зв'язок зі змінною «Net monthly salary». Також слід зауважити, що змінна «Age» не є статистично значущою, оскільки p-value ($P > |t|$) дорівнює 0.408, що перевищує звичайний рівень значущості 0.05. Це означає, що вік не має значимого впливу на місячну заробітну плату, і в подальшому дана змінна може бути виключеною з моделі.

Усі інші змінні, такі як «Education level», «English level», «Title», «Main specialization», «Company type» і «Total work experience in the field (years)», є статистично значущими, оскільки їх p-value дорівнює 0.000 (менше за 0.05). Це означає, що ці змінні мають вплив на рівень місячної заробітної плати.

Також на основі отриманих даних можна зробити такі висновки:

- чоловіки мають на 410,674 доларів вищу очікувану заробітну плату в порівнянні з жінками;
- здобуття вищого рівня освіти призводить до збільшення заробітної плати на 111,343 доларів;
- підвищення рівня володіння англійською мовою призводить до збільшення заробітної плати на 413,517 доларів;
- вищий рівень посади (тайтлу) супроводжується збільшенням заробітної плати на 608,252 доларів;

- основна спеціалізація не суттєво впливає на розмір заробітної плати, адже перехід до більш складної спеціалізації забезпечує її підвищення лише на 67,243 доларів;
- збільшення рівня типу компанії призводить до збільшення заробітної плати на 187,710 доларів;
- кожен додатковий рік в загальному стажі роботи забезпечує збільшення заробітної плати на 144,165 доларів.

Отже, на рівень місячної заробітної плати з поміж обраних факторів дійсно суттєво впливають стать, рівень англійської мови, посада та досвід роботи за спеціальністю, а вік навпаки суттєво не впливає на розмір заробітної плати.

Середня абсолютна похибка (Mean Absolute Error) для побудованої лінійної регресії становить 856.78, що можна вважати помірним значенням, а середня абсолютна відсоткова помилка (Mean Absolute Percentage Error) – 38,81%, тобто прогнозовані значення в середньому відхиляються від реальних на 38,81%.

Отримане високе значення MAPE вказує на значну відсоткову похибку прогнозування лінійної регресійної моделі. Таке значення може свідчити про недостатню точність моделі або про наявність викидів та впливових спостережень, які впливають на якість прогнозу.

Середньоквадратична похибка (Mean Squared Error) для даної моделі дорівнює 1309749.4568, що вказує на велику величину похибки прогнозування.

Порівняємо тестові значення з прогнозованими. Цей порівняльний аналіз дозволяє оцінити точність моделі та її здатність узагальнювати та прогнозувати значення на основі незалежних змінних.

На рис. 3.14 представлено графіки значень місячної заробітної плати з тестової вибірки та спрогнозованих значень за допомогою лінійної регресії. Модель місцями добре описує закономірності, але є багато точок, в яких спостерігаються великі розбіжності між реальними та прогнозованими значеннями. Це може свідчити про недосконалість побудованої моделі і може бути результатом нелінійних залежностей між змінними, які модель не врахувала, або впливу інших факторів, які не були включені до аналізу.

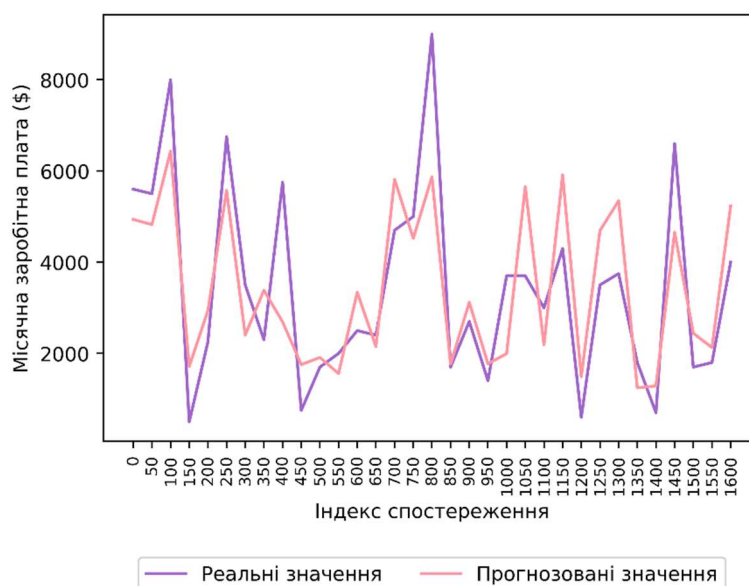


Рисунок 3.14 – Порівняння реальних (тестових) та прогнозованих значень для лінійної регресії

Джерело: розроблено автором на основі [38]

3.3 Застосування дерева рішень для прогнозування заробітної плати

Побудова моделі дерева рішень здійснювалась за допомогою класу DecisionTreeRegressor бібліотеки sklearn. Було визначено основні показники якості моделі та точності прогнозування. R^2 для отриманої моделі становить 0,70 (70%), що є досить непоганим результатом, але недостатнім для точного прогнозування.

Значення середньоквадратичної похибки для даної моделі становить 1242334,2957 і є меншим, порівняно з показником лінійної регресії.

Відповідно до обчисленого значення MAE, прогнозовані значення моделі дерева рішень в середньому відрізняються від фактичних значень на 790.17 одиниць, цей показник також є меншим, в порівнянні з відповідним показником лінійної регресії. Варто дослідити важливість факторів для даної моделі (рис. 3.15).

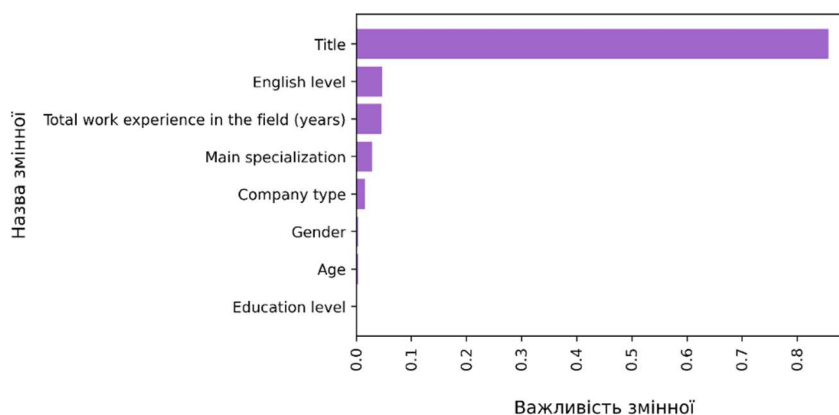


Рисунок 3.15 – Візуалізація важливих факторів для моделі дерева рішень

Джерело: розроблено автором на основі [38]

Згідно з рис. 3.15 найважливішим фактором для моделі дерева рішень є посада (важливість становить 0,86), на другому місці – рівень англійської мови та досвід роботи (0,05), а на третьому – основна спеціалізація з важливістю 0,03.

Важливість змінної «Education level» (рівень освіти) дорівнює 0, тобто дана змінна не є значимою для цієї моделі.

Слід теж порівняти реальні значення з прогнозованими, відповідний графік зображений на рис. 3.16.

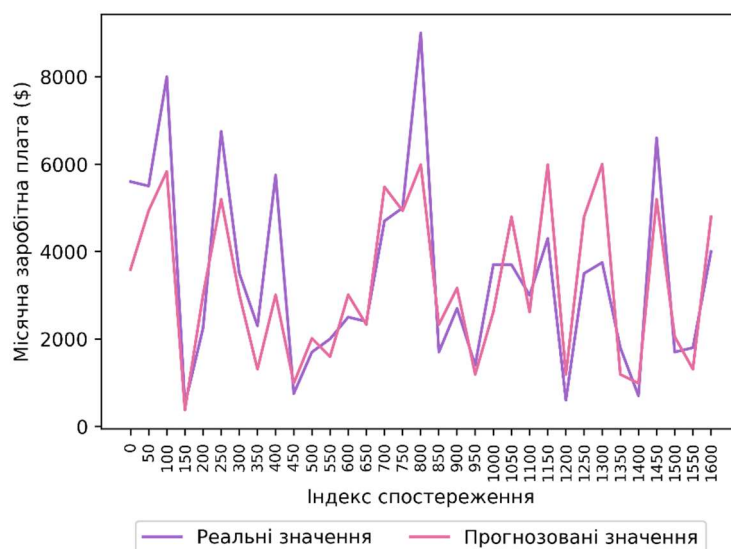


Рисунок 3.16 – Порівняння реальних (тестових) та прогнозованих значень для моделі дерева рішень

Джерело: розроблено автором на основі [38]

Дана модель краще описує закономірності в даних, але все одно недостатньо, для отримання якомога точного прогнозу.

3.4 Використання алгоритму «випадковий ліс» для прогнозування зарплати

Далі було побудовано модель «випадковий ліс» за допомогою класу RandomForestRegressor бібліотеки sklearn. R^2 для тренувальних даних цієї моделі становить 0,93 (93%), але для тестових даних він дорівнював лише 0,66 (66%), що може вказувати на перенавчання моделі.

Середня абсолютна похибка (MAE) для побудованої моделі дорівнює 837.81, а середньоквадратична похибка (MSE) для даної моделі дорівнює 1375663.5157, що вказує на велику величину похибки прогнозування. Важливість факторів для моделі «випадковий ліс» зображено на рис. 3.17.

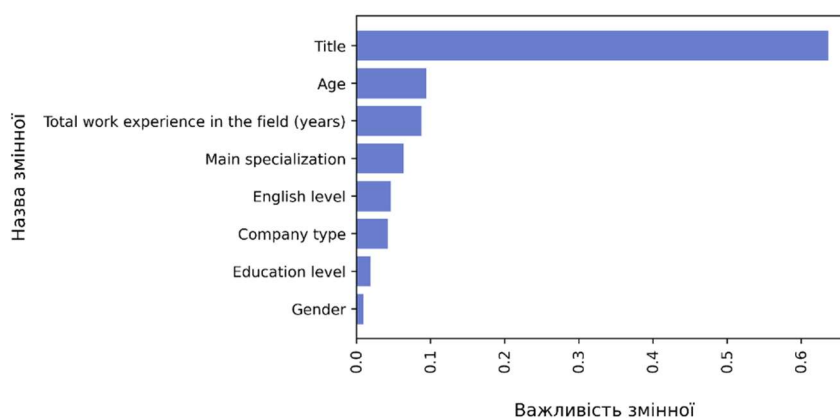


Рисунок 3.17 – Візуалізація важливих факторів для моделі дерева рішень

Джерело: розроблено автором на основі [38]

Відповідно до побудованої гістограми (рис. 3.17) найважливішим фактором для моделі «випадковий ліс» теж є посада (важливість становить 0,64), на другому місці – вік та досвід роботи (0,09), а на третьому – основна спеціалізація з важливістю 0,06.

Важливість змінної «Gender» (стать) дорівнює 0,01, тобто дана змінна є найменш значимою для цієї моделі.

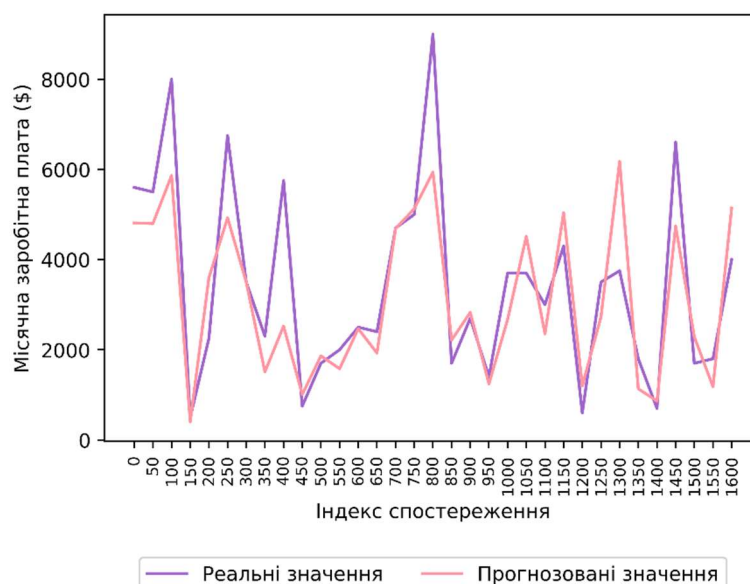


Рисунок 3.18 – Порівняння реальних (тестових) та прогнозованих значень для моделі «випадковий ліс»

Джерело: розроблено автором на основі [38]

На графіку видно, що модель розуміє деякі закономірності, але недостатньо добре прогнозує місячну заробітну плату.

3.5 Порівняння побудованих моделей прогнозування заробітної плати

Для наочності обчислені показники ефективності та точності прогнозу побудованих моделей було розміщено в таблиці (табл. 3.1).

Таблиця 3.1 – Порівняльна таблиця алгоритмів на тренувальній вибірці

Метрика	Алгоритм		
	Лінійна регресія	Дерево рішень	Випадковий ліс
MAE	856,78	790,17	838,19
MSE	1309749,46	1242334,3	1376066,83
RMSE	1144,44	1114,6	1173,06
R^2	0,67	0,69	0,66

Джерело: розроблено автором на основі [38]

На основі наведеної таблиці можна зробити такі висновки про результати побудованих моделей на тренувальній вибірці:

1. З точки зору середньої абсолютної похибки (MAE), найкращий результат показує дерево рішень зі значенням 790,17, проти 856,78 у лінійної регресії та 838,19 у випадковому лісі. Чим нижче значення MAE, тим краще модель прогнозує результати.

2. У випадку середньої квадратичної похибки (MSE), також найкращий результат показує дерево рішень зі значенням 1242334,3, проти 1309749,46 та 1376066,83 у лінійної регресії та випадковому лісі відповідно. Менше значення MSE вказує на більш точні прогнози моделі.

3. Корінь середньоквадратичної похибки (RMSE) також показує найкращі результати для дерева рішень зі значенням, адже чим нижче значення RMSE, тим краще модель прогнозує дані.

4. У випадку коефіцієнта детермінації (R^2), дерево рішень та дерево рішень показують схожу ефективність зі значеннями 0,69 та 0,66 відповідно, тоді як лінійна регресія має значення 0,67. Коефіцієнт детермінації вказує на якість пояснення моделлю варіації вхідних даних.

Отже, з урахуванням обраних метрик, дерево рішень демонструє найкращі результати на тренувальній вибірці, з меншими значеннями MAE, MSE та RMSE, що свідчить про його кращу прогностичну здатність порівняно з лінійною регресією та випадковим лісом.

Модель дерева рішень має певну прогностичну здатність і може бути використана для прогнозування, однак, залишається ще 31% варіації, яку модель не може пояснити.

З метою покращення точності та прогностичної здатності даної моделі може бути доречним додати додаткові незалежні змінні до аналізу, наприклад, мова програмування, сфера праці та регіон розташування компанії, в якій працює респондент. Ці змінні можуть включати інші фактори, які можуть впливати на прогнозовану залежну змінну.

Далі були розглянуті фактори, які кожна модель вважає важливими, результати цього аналізу подані в таблиці 3.2.

Таблиця 3.2 - Порівняння важливих факторів для використаних алгоритмів

Пріоритет фактора	Алгоритм		
	Лінійна регресія	Дерево рішень	Випадковий ліс
1	Title	Title	Title
2	English level	English level	Age
3	Gender	Total work experience	Total work experience
4	Company type	Main Specialization	Main Specialization
5	Total work experience	Company type	English level
6	Education level	Gender	Company type
7	Main Specialization	Age	Education level
8	Age	Education level	Gender

Джерело: розроблено автором на основі [38]

Відповідно до табл 3.2 змінна «Title» має найвищий пріоритет серед всіх побудованих моделей. Це означає, що посада, ймовірно, має суттєве значення для визначення результатів моделей і найбільше впливає на рівень заробітної плати ІТ-спеціалістів.

Для лінійної регресії та дерева рішень змінна «English level» є важливим фактором, який впливає на зарплату, тоді як для випадкового лісу цим фактором є «Age». Це вказує на різні аспекти, які можуть впливати на результати цих моделей.

«Total work experience» є третім важливим фактором для двох моделей з трьох побудованих. Це означає, що досвід роботи може мати певний вплив на прогнозування заробітної плати.

За даними таблиці освітній рівень може мати різний вплив на прогнозування в залежності від використаного алгоритму. Пріоритет інших факторів варіюється в залежності від алгоритму.

ВИСНОВКИ

Розвиток інформаційних технологій відкриває широкі перспективи для економічного росту та підвищення якості життя населення. Ці технології надають неповторні можливості для стимулювання нашої економіки і сприяють підвищенню продуктивності різних галузей, а також сприяють розвитку інновацій та підприємництва.

Крім того, інформаційні технології забезпечують доступ до глобального ринку, що дозволяє підприємцям розширювати свої бізнес-можливості та взаємодіяти з клієнтами з усього світу. Це стимулює експорт, привертає іноземні інвестиції та сприяє зміцненню позиції країни на світовій арені.

У першому розділі роботи було розглянуто основні теоретичні підходи та концепції, пов'язані з ІТ-ринком, розкрито ключові терміни та проведено аналіз сучасного стану ІТ-ринку в Україні, висвітлено внесок ІТ-ринку в економіку країни, охарактеризовано типи та кількість ІТ-компаній, описано кадровий потенціал в ІТ-галузі України та роль інформаційних технологій в інших сферах економіки країни.

Згідно з результатами аналізу комп'ютерні послуги даний час займають перше місце серед експорту послуг і у 2022 році складала майже 4% від обсягу ВВП. До топ-5 партнерів України з експорту комп'ютерних послуг у 2021 році належали США, Велика Британія, Мальта, Ізраїль та Кіпр. У період з 2018 по 2021 рік спостерігалось стабільне зростання кількості підприємств, які відносяться до ІТ-галузі. Однак, у 2022 році відбулося зменшення цього показника на приблизно півтисячі підприємств внаслідок повномасштабного вторгнення російських військ.

Найбільшу частку серед ІТ-компаній в Україні займають продуктові компанії (72%), тоді як R&D центри становлять найменшу частку (4%). Більшості продуктових та сервісних ІТ-компаній притаманна невелика кількість персоналу (до 50 працівників).

Другий розділ містить опис методів машинного навчання, необхідних для подальшого дослідження. Також в даному розділі було наведено опис основних

метрик, за якими проводиться оцінка точності та похибки побудованих моделей. Ще в другому розділі було здійснено SWOT-аналіз.

В результаті SWOT-аналізу було визначено фактори, що стримують розвиток ІТ-ринку в Україні. Серед них недостатня підтримка держави, відтік висококваліфікованих кадрів, низький рівень заробітної плати для ІТ-фахівців порівняно з іншими країнами, нестабільність економічної ситуації в країні та відсутність потужного внутрішнього ринку для продуктів та послуг вітчизняних ІТ-компаній.

Третій розділ роботи присвячено використанню методів машинного навчання, таких як лінійна регресія, дерева рішень, випадковий ліс, для прогнозування заробітної плати ІТ-фахівців України, враховуючи такі фактори, як стать, вік, освітній рівень, рівень англійської мови, посада, основна спеціалізація, тип компанії та досвід роботи за спеціальністю.

В результаті побудови та аналізу лінійної регресії, моделі дерева рішень та випадкового лісу, виявили, що найкращі результати за обраними метриками ефективності та похибки на тренувальній вибірці отримала модель дерева рішень, що свідчить про кращу здатність даної моделі до прогнозування місячної заробітної плати на ІТ-ринку України, порівняно з двома іншими моделями.

На основі побудованих моделей виявлено, що серед факторів, які найсуттєвіше впливають на зарплату ІТ-фахівців, можна виділити посаду, рівень англійської мови, досвід роботи та основну спеціалізацію. З метою поліпшення точності та прогностичної здатності побудованих моделей можна розглянути можливість додати інші незалежні змінні до аналізу. Наприклад, можна врахувати такі фактори, як мова програмування, сфера праці та регіон розташування компанії, в якій працює респондент.

Отже, дана робота має вагомий внесок у розуміння впливу інформаційних технологій на економічний розвиток України, а також може бути використана в подальших дослідженнях та стратегічному плануванні для розвитку сучасного ІТ-ринку країни.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Булкіна І. А. Внесок України в розвиток міжнародного ринку ІТ-послуг. *Бізнес Інформ*. 2020. № 11. С. 37-42. URL: <https://doi.org/10.32983/2222-4459-2020-11-37-42>.
2. Вікулова А. О., Савчук В. В. Перспективи розвитку ринку ІТ-послуг в Україні. *Причорноморські економічні студії*. 2020. Вип. 51. С. 27-32. URL: <https://doi.org/10.32843/bses.51-4>.
3. Колобердянко І., Чернявська А. Стан і тенденції міжнародного ринку торгівлі ІТ-послугами. *Економіка та суспільство*. 2021. № 32. URL: <https://doi.org/10.32782/2524-0072/2021-32-63>.
4. Мочерний С. В. Основи економічних знань : підручник. Київ: Академія, 2000. 304 с. (Альма-матер).
5. Полякова О. М. Сучасні тенденції розвитку ІТ-індустрії в Україні і світі. *Вісник економіки транспорту та промисловості*. 2018. № 64. С. 179-186. URL: <http://lib.kart.edu.ua/handle/123456789/1726>.
6. Top Lead. Information technology industry in Ukraine. Infographic atlas 2017. URL: <http://publications.chamber.ua/2018/IT/the-infographics-report-it-industry-of-ukraine-2017.pdf>.
7. Мельник Т., Завгородня Є. ІТ-сектор України на світовому ринку. *Зовнішня торгівля: економіка, фінанси, право*. 2022. № 6. С. 17-36. URL: [https://doi.org/10.31617/3.2022\(125\)02](https://doi.org/10.31617/3.2022(125)02).
8. IT Ukraine Association. Ukraine IT report 2021. URL: <https://drive.google.com/file/d/1LujaT9pHEGhgpRRojfnlZgQikkyillbE/view> (дата звернення: 24.05.2023).
9. GDP (current US\$) – Ukraine. *World Bank Open Data*. URL: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=UA> (accessed: 15.04.2023).

10. Зовнішня торгівля. *Національний банк України*. URL: https://bank.gov.ua/files/ES/Tov_u.pdf (дата звернення: 15.04.2023).
11. IT Ukraine Association. Do IT like Ukraine 2022. URL: <https://itukraine.org.ua/files/reports/2022/DoITLikeUkraine2022.pdf> (дата звернення: 20.02.2022).
12. Кількість активних підприємств за регіонами України та видами економічної діяльності. *Державна служба статистики України*. URL: https://www.ukrstat.gov.ua/operativ/operativ2014/kap/kap_u/arh_kap_u.html.
13. Класифікація видів економічної діяльності (КВЕД-2010). Секція J. Інформація та телекомунікації. *Державна служба статистики України*. URL: http://kved.ukrstat.gov.ua/KVED2010/SECT/KVED10_J.html.
14. Ukrainian Tech Ecosystem Overview. *Команда Міністерства цифрової трансформації*. URL: <https://uatechecosystem.com> (дата звернення: 06.03.2023).
15. Дашборд технологічної екосистеми України. *Команда Міністерства цифрової трансформації*. URL: <https://uatechecosystem.com/dashboard> (дата звернення: 06.03.2023).
16. Продуктові компанії. *Ukrainian Tech Ecosystem Overview*. URL: <https://uatechecosystem.com/product-companies> (дата звернення: 06.03.2023).
17. Сервісні компанії. *Ukrainian Tech Ecosystem Overview*. URL: <https://uatechecosystem.com/service-companies> (дата звернення: 06.03.2023).
18. Кількість найманих працівників у суб'єктів господарювання за видами економічної діяльності. *Державна служба статистики України*. URL: https://www.ukrstat.gov.ua/operativ/operativ2012/fin/osp/knp/knp_u/arch_knp_u.htm.
19. Лебедєв Д., Самоходський І. Аналіз ІТ-освіти у вишах України. *Офіс ефективного регулювання BRDO*. 2021. URL: https://brdo.com.ua/wp-content/uploads/2021/02/Analiz_IT_osvity_u_vyshah_Ukrai-ny_Print.pdf.
20. Степаненко С. В., Дяченко О. П. Інформаційні технології як інструмент розвитку інклюзивного аграрного бізнесу. *Таврійський науковий вісник*. Серія: Економіка. 2023. №15. С. 116-122. URL: <http://tnv-econom.ksauniv.ks.ua/index.php/journal/article/view/330/309>.

21. Rußmann M. Industry 4.0: the future of productivity and growth in manufacturing industries. URL: https://www.bcg.com/publications/2015/engineered_products_project_business_industry_4_future_productivity_growth_manufacturing_industries (accessed: 15.03.2023).
22. Сосновська О. О., Вакофян В. Г. Індустрія 4.0: сутність і тенденції розвитку. *Бізнес Інформ*. 2022. №1. С. 137–144. URL: <https://doi.org/10.32983/2222-4459-2022-1-137-144>.
23. Мороз О. В., Даценко І. В. Огляд ринку електронної торгівлі у світі та в Україні. *Східна Європа: економіка, бізнес та управління*. 2018. № 6 (17). С. 278-283. URL: http://www.easterneurope-ebm.in.ua/journal/17_2018/17_2018.pdf.
24. Montgomery D. C, Peck E. A., Vining G. G. Introduction to linear regression analysis. 6th ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2021. P. 704.
25. Tranmer M., Murphy J., Elliot M., Pampaka M. Multiple linear regression. 2nd ed. *Cathie Marsh Institute Working Paper*. 2020. P. 59.
26. Rokach L., Maimon O. Data mining with decision trees: theory and applications. 2nd ed. *Series in Machine Perception and Artificial Intelligence*. World Scientific, 2014. Vol. 81. P. 328. URL: <https://doi.org/10.1142/9097>.
27. Navlani A. Decision tree classification in Python tutorial. *DataCamp*. URL: <https://www.datacamp.com/tutorial/decision-tree-classification-python>.
28. Breiman L. Random Forests. *Machine Learning*. 2001. Vol. 45. P. 5-32. URL: <https://doi.org/10.1023/A:1010933404324>.
29. Ho T. K. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, 1995. Vol. 1. P. 278-282. URL: <https://doi.org/10.1109/ICDAR.1995.598994>.
30. Chicco D., Warrens M. J., Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*. 2021. Vol. 7. P. 24. URL: <https://doi.org/10.7717/peerj-cs.623>.

31. Pandas documentation – pandas 2.0.1 documentation. *Pandas - Python Data Analysis Library*. URL: <https://pandas.pydata.org/docs/> (accessed: 25.05.2023).
32. Matplotlib documentation – Matplotlib 3.7.1 documentation. *Matplotlib – Visualization with Python*. URL: <https://matplotlib.org/stable/index.html> (accessed: 25.05.2023).
33. Statsmodels documentation - statsmodels 0.14.0. *Statsmodels*. URL: <https://www.statsmodels.org/stable/index.html> (accessed: 25.05.2023).
34. NumPy documentation – NumPy v1.24 Manual. *NumPy*. URL: <https://numpy.org/doc/stable/> (accessed: 25.05.2023).
35. Scikit-learn: machine learning in Python – scikit-learn 1.2.2 documentation. *Scikit-learn: machine learning in Python*. URL: <https://scikit-learn.org/stable/index.html> (accessed: 25.05.2023).
36. Seaborn: statistical data visualization – seaborn 0.12.2 documentation. *Seaborn*. URL: <https://seaborn.pydata.org/> (accessed: 25.05.2023).
37. Зарплати українських розробників — зима 2023. *Редакція DOU*. URL: <https://dou.ua/lenta/articles/salary-report-devs-winter-2023/> (дата звернення: 25.02.2023).
38. Волошин Сергій. Датасет 2022_dec_raw.csv. *GitHub*. URL: https://github.com/devua/csv/blob/master/salaries/2022_dec_raw.csv (дата звернення: 25.02.2023).

ДОДАТКИ

Додаток А

Код програми для підготовки та аналізу тестових даних

Імпортуємо потрібні бібліотеки

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import matplotlib.colors as mcolors
```

Відключаємо попередження Anaconda

```
import warnings
warnings.simplefilter('ignore')
```

Зчитуємо файл з початковими даними

```
df = pd.read_csv('2022_dec_raw.csv', sep=';')
df.head()
```

Проглядаємо коротку довідку про датафрейм

```
df.info()
```

Видаляємо зайві стовпці

```
df = df.loc[:, ['В якій області ви зараз живете?', 'Зарплата у $$$ за місяць, лише ставка після сплати
податків', 'Ваша стат'ь', 'Ваш вік', 'Яка у вас освіта?', 'Знання англійської мови', 'Ваш тайтл',
'Основна мова програмування', 'Вкажіть вашу основну спеціалізацію', 'В якій сфері працюєте?',
'Тип компанії', 'Загальний стаж роботи за спеціальністю']]
df.head()
```

Для зручності перейменуємо стовпці датасету

```
df.rename(columns={'Ви працюєте зараз в IT?': 'IT employment status',
                  'Де ви зараз живете? ': 'Place of current residence',
                  'В якій області ви зараз живете?': 'Current region of residence',
                  'Зарплата у $$$ за місяць, лише ставка після сплати податків': 'Net monthly salary in $',
                  'Ваша стат'ь': 'Gender',
```

```

'Ваш вік': 'Age',
'Яка у вас освіта?': 'Education level',
'Чи є у вас науковий ступінь? ': 'Academic degree status',
'Знання англійської мови': 'English level',
'Оберіть вашу посаду': 'Job position',
'Ваш тайтл': 'Title',
'Основна мова програмування': 'Primary programming language',
'Вкажіть вашу основну спеціалізацію': 'Main specialization',
'В якій сфері працюєте?': 'Industry',
'Тип компанії': 'Company type',
'Загальний стаж роботи за спеціальністю': 'Total work experience in the field (years)',
}, inplace=True)

```

df

Побудуємо кругову діаграму статевого розподілу

```

gender_counts = df['Gender'].value_counts()
colors = ['#1836B2', '#E56EA2']
gender_counts.plot(kind='pie', autopct='%1.1f%%', colors = colors, textprops={'color': 'white'},
ylabel='')
plt.title('Статевий розподіл респондентів')
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()

```

Видалимо NaN елементи в стовпцях «Net monthly salary in \$» і «Main specialization»

```

df = df.dropna(subset=['Net monthly salary in $'])
df = df.dropna(subset=['Main specialization'])

```

Видалимо зі стовпця "Age" значення "29,5"

```

df = df.drop(df[df['Age'] == '29,5'].index)

```

Змінимо дані в стовпці «Total work experience in the field (years)» так, щоб можна було змінити тип даних на числовий:

```

df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].str.replace('Пів року', '0.5', regex=True)

```

```
df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].str.replace('рік|років|роки|року', '', regex=True)
```

```
df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].str.replace('Менше як 3 місяці', '0.2', regex=True)
```

```
df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].str.replace('3 місяці', '0.25', regex=True)
```

```
df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].str.replace('15 і більше', '15', regex=True)
```

Змінимо тип даних стовпців "Net income per month in \$", "Age", "Total work experience in the field (years)" і замінимо значення ",", " на "."

```
df['Net monthly salary in $'] = df['Net monthly salary in $'].replace(',', '.', regex=True).astype(float)
```

```
df['Age'] = df['Age'].astype(int)
```

```
df['Total work experience in the field (years)'] = df['Total work experience in the field (years)'].replace(',', '.', regex=True).astype(float)
```

Замінимо значення "Embedded ", "embedded", "embedded " в стовпці Main specialization на "Embedded"

```
df['Main specialization'] = df['Main specialization'].replace({"Embedded ": "Embedded",
                                                                "embedded": "Embedded",
                                                                "embedded ": "Embedded"})
```

Побудуємо гістограму вікового розподілу респондентів:

```
plt.hist(df['Age'], bins=20, color='#A066CB')
```

```
plt.title('Віковий розподіл респондентів')
```

```
plt.xlabel('Вік (років)')
```

```
plt.ylabel('Кількість респондентів')
```

```
plt.show()
```

Обчислимо значення основних числових характеристик (середнє значення, мода, медіана) віку

```
average_age = df['Age'].mean()
print('Середній вік:', average_age)
```

```
moda = df['Age'].mode()[0]
print('Мода:', moda)
```

```
median_age = df['Age'].median()
print('Медіана:', median_age)
```

Виведемо статистику розподілу місячної заробітної плати за допомогою гістограми:

```
plt.hist(df['Net monthly salary in $'], bins=20, color='#6A7DCD')
plt.title('Розподіл місячної заробітної плати')
plt.xlabel('Місячна заробітна плата ($)')
plt.ylabel('Кількість респондентів')
plt.show()
```

Обчислимо значення основних числових характеристик (середнє значення, мода, медіана) заробітної плати

```
average_age = df['Net monthly salary in $'].mean()
print('Середня заробітна плата:', average_age)
```

```
moda = df['Net monthly salary in $'].mode()[0]
print('Мода:', moda)
```

```
median_age = df['Net monthly salary in $'].median()
print('Медіана:', median_age)
```

Побудуємо гістограму розподілу стажу роботи респондентів:

```
plt.hist(df['Total work experience in the field (years)'], bins=20, color='#FF96A3')
plt.title('Розподіл стажу роботи респондентів')
plt.xlabel('Кількість років')
plt.ylabel('Кількість респондентів')
plt.show()
```

Обчислимо значення основних числових характеристик (середнє значення, мода, медіана) стажу роботи

```
average_age = df['Total work experience in the field (years)'].mean()
print('Середній стаж:', average_age)
moda = df['Total work experience in the field (years)'].mode()[0]
print('Мода:', moda)
```

```
median_age = df['Total work experience in the field (years)'].median()
print('Медіана:', median_age)
```

Побудуємо графік, що продемонструє залежність середньої зарплати від стажу роботи

```
exp_salary = df.groupby('Total work experience in the field (years)')['Net monthly salary in $'].mean()
plt.plot(exp_salary.index, exp_salary.values, color='#853CAF')
plt.title('Залежність середньої зарплати від стажу роботи')
plt.xlabel('Стаж роботи (років)')
plt.ylabel('Середня місячна заробітна плата ($)')
plt.show()
```

Побудуємо графік, що продемонструє залежність середньої зарплати від основної спеціалізації

```
avg_salary = df.groupby('Main specialization')['Net monthly salary in $'].median()
avg_salary = avg_salary.sort_values(ascending = True)
avg_salary.plot.barh(color='#1836B2')
plt.xlabel('Середня місячна заробітна плата ($)')
plt.ylabel('Основна спеціалізація')
plt.show()
```

Побудуємо графік розподілу середньої зарплати між різними областями

```
region_salary = df.groupby('Current region of residence')['Net monthly salary in $'].mean()
region_salary = region_salary.sort_values(ascending = True)
region_salary.plot.barh(figsize=(8, 6), color='#A066CB')
plt.title('Розподіл середньої зарплати між різними областями України')
plt.xlabel('Середня місячна заробітна плата ($)')
plt.ylabel('Region')
plt.show()
```

Проглянемо основну статистику числових показників:

```
df.describe()
```

Для діагностики викидів та нетипових спостережень побудуємо коробчасті діаграми:

Будуємо коробчасту діаграму для показника "Net monthly salary in \$"

```
df[['Net monthly salary in $']].boxplot(figsize=(8, 8));
```

Видалимо інформацію про всіх респондентів, середня заробітна плата яких більша 12500 доларів (значення найбільше відрізняються від інших)

```
df.drop(df[df['Net monthly salary in $'] > 12500].index, inplace=True)
```

Будуємо коробчасту діаграму для показника "Age"

```
df[['Age']].boxplot(figsize=(8, 8));
```

Видалимо інформацію про респондентів, старших 50 років (значення найбільше відрізняються від інших)

```
df.drop(df[df['Age'] > 50].index, inplace=True)
```

Будуємо коробчасту діаграму для показника "Total work experience in the field (years)"

```
df[['Total work experience in the field (years)']].boxplot(figsize=(8, 8));
```

Видалимо інформацію про респондентів, стаж роботи яких понад 14 років (значення найбільше відрізняються від інших)

```
df.drop(df[df['Total work experience in the field (years)'] > 14].index, inplace=True)
```

Перекодування змінних:

```
df = df.drop(df[df['Title'] == 'Немає тайтлу'].index)
```

```
df = df.drop(df[df['Company type'] == 'Інша'].index)
```

```
df = df.drop(df[df['Main specialization'] == 'Інше'].index)
```

Перекодуємо дані в стовпці "Gender" в числові

```
df['Gender'] = df['Gender'].replace({"Жінка": "0",  
                                     "Чоловік": "1"})
```


Перекодуємо дані в стовпці "Company type" в числові

```
df['Company type'] = df['Company type'].replace({"Фріланс": "1",
                                                "Аутсорсингова": "2",
                                                "Продуктова": "3",
                                                "Аутстафінгова": "4",
                                                "Стартап": "5"})
```

Змінимо тип даних стовпців на числовий, замінімо значення "NaN" на "0" та "," на "."

```
df['Education level'] = df['Education level'].astype(int)
df['Gender'] = df['Gender'].astype(int)
df['English level'] = df['English level'].astype(int)
df['Title'] = df['Title'].astype(int)
df['Main specialization'] = df['Main specialization'].astype(int)
df['Company type'] = df['Company type'].astype(int)
```

Створимо датасет, в який додамо змінні, необхідні для побудови моделей

```
data = df.loc[:, ['Net monthly salary in $', 'Gender', 'Age', 'Education level', 'English level', 'Title', 'Main specialization', 'Company type', 'Total work experience in the field (years)']]
```

Будуємо кореляційну матрицю

```
corr_matrix = data.corr()
colors = ['#FFC2C2', '#1836B2']
positions = [0, 1]
cmap = mcolors.LinearSegmentedColormap.from_list('CustomCmap', list(zip(positions, colors)))
sns.heatmap(corr_matrix, annot=True, cmap=cmap, fmt=".2f")
```

Додаток Б

Код програми для побудови регресійних моделей

Імпортуємо необхідні бібліотеки

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
```

X - матриця незалежних змінних, яка складатиметься зі змінних Age, Gender, English level, Total work experience in the field (years)

```
X = data.drop('Net monthly salary in $', axis=1)
X.head()
```

Вектор залежної змінної Y - стовпчик "Net income per month in \$"

```
Y = data[['Net monthly salary in $']]
Y.head()
```

Розділимо дані на тестові та тренувальні

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Побудуємо модель лінійної регресії

```
from sklearn.linear_model import LinearRegression
lr_model = LinearRegression(normalize=True)
lr_model.fit(X_train, Y_train)
```

```
Y_pred_lr = lr_model.predict(X_test)
indexs = np.arange(0, len(Y_pred_lr), 50)
```

Побудуємо графік для порівняння прогнозованих і реальних значень для лінійної регресії

```
Y_pred_lr = lr_model.predict(X_test)
indexs = np.arange(0, len(Y_pred_lr), 50)
```

```
plt.plot(indexs, Y_test.values[indexs], label='Реальні значення', color = '#A066CB')
plt.plot(indexs, Y_pred_lr[indexs], label='Прогнозовані значення', color = '#FF96A3')
```

```
plt.xlabel("Індекс спостереження")
plt.xticks(indexs, rotation='vertical', fontsize=8)
plt.ylabel("Місячна заробітна плата ($)")

plt.title("Порівняння прогнозованих і реальних значень для лінійної регресії")
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.25), ncol=2)
plt.show()
```

Оцінимо параметри лінійної регресії за допомогою методу ІМНК

```
Model_1 = sm.OLS(Y, sm.add_constant(X)).fit()
```

Виведемо на екран основну регресійну статистику

```
Model_1.summary()
```

Обчислимо MAE

```
mae_lr = mean_absolute_error(Y_test, Y_pred_lr)
print("Mean Absolute Error:", mae_lr)
```

Обчислимо MSE та RMSE

```
mse_lr = mean_squared_error(Y_test, Y_pred_lr)
print("Mean Squared Error:", mse_lr)
print("RMSE: ", mse_lr**(1/2.0))
```

Обчислимо R²

```
r2_lr = r2_score(Y_test, Y_pred_lr)
print("R-squared:", r2_lr)
```

Обчислимо MAPE

```
def mean_absolute_percentage_error(Y_test, Y_pred_lr):
```

```
    Y_test = np.array(Y_test)
```

```
    Y_pred_lr = np.array(Y_pred_lr)
```

```
    # Обчислюємо абсолютну відсоткову помилку
```

```
    absolute_percentage_error = np.abs((Y_test - Y_pred_lr) / Y_test) * 100
```

```

# Виключаємо недійсні значення (нулі або дуже близькі до нуля)
absolute_percentage_error = absolute_percentage_error[np.isfinite(absolute_percentage_error)]

# Обчислюємо середнє значення абсолютної відсоткової помилки
mape = np.mean(absolute_percentage_error)

return mape

mape_lr = mean_absolute_percentage_error(Y_test, Y_pred_lr)
print("MAPE:", mape_lr)

# Побудуємо модель дерева рішень
dt_model = DecisionTreeRegressor(max_depth=6, min_samples_split=4, min_samples_leaf=6)
dt_model.fit(X_train, Y_train)

score_dt = dt_model.score(X, Y)
print("R-squared:", score_dt)

Y_pred_dt = dt_model.predict(X_test)

# Обчислимо MSE та RMSE
mse_dt = mean_squared_error(Y_test, Y_pred_dt)
print("Mean Squared Error:", mse_dt)
print("RMSE: ", mse_dt**(1/2.0))

# Обчислимо R2
r2_dt = r2_score(Y_test, Y_pred_dt)
print("R-squared:", r2_dt)

# Обчислимо MAE
mae_dt = mean_absolute_error(Y_test, Y_pred_dt)
print("Mean Absolute Error:", mae_dt)

```

Обчислимо MAPE

```
def mean_absolute_percentage_error(Y_test, Y_pred_dt):
    Y_test = np.array(Y_test)
    Y_pred_dt = np.array(Y_pred_dt)

    # Обчислимо відсоткову помилку
    absolute_percentage_error = np.abs((Y_test - Y_pred_dt) / Y_test) * 100

    # Виключимо недійсні значення (нулі або дуже близькі до нуля)
    absolute_percentage_error = absolute_percentage_error[np.isfinite(absolute_percentage_error)]

    # Обчислимо середнє значення абсолютної відсоткової помилки
    mape = np.mean(absolute_percentage_error)

    return mape

mape_dt = mean_absolute_percentage_error(Y_test, Y_pred_dt)
print("MAPE:", mape_dt)
```

Побудуємо графік для порівняння прогнозованих і реальних значень для моделі дерева рішень

```
indexs = np.arange(0, len(Y_pred_dt), 100)
plt.plot(indexs, Y_pred[indexs], label='Прогнозовані значення')
plt.plot(indexs, Y_test.values[indexs], label='Реальні значення')
plt.xlabel("Індекс спостереження")
plt.xticks(indexs, rotation='vertical')
plt.ylabel("Місячна заробітна плата ($)")

plt.title("Порівняння прогнозованих і реальних значень для моделі дерева рішень ")
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.25), ncol=2)
plt.show()
```

Визначимо важливі фактори і виведемо їх у вигляді гістограми

```
importance_list = []
name_list = []
```

```
for importance, name in sorted(zip(dt_model.feature_importances_, X_train.columns), reverse=False):
    importance_list.append(importance)
    name_list.append(name)
```

```
plt.figure(dpi=500)
```

```
plt.barh(name_list, importance_list, color='#A066CB')
plt.xlabel('Важливість змінної', labelpad=20, fontsize=12)
plt.ylabel('Назва змінної', labelpad=10, fontsize=12)
```

```
plt.title('Важливість змінних за допомогою моделі дерева рішень')
plt.xticks(rotation='vertical')
plt.show()
```

Побудуємо модель «випадковий ліс»

```
rf_model = DecisionTreeRegressor(max_depth=6, min_samples_split=4, min_samples_leaf=6)
rf_model.fit(X_train, Y_train)
```

```
score_rf = rf_model.score(X, Y)
print("R-squared:", score_rf)
```

```
Y_pred_rf = rf_model.predict(X_test)
```

Обчислимо MSE та RMSE

```
mse_rf = mean_squared_error(Y_test, Y_pred_rf)
print("Mean Squared Error:", mse_rf)
print("RMSE: ", mse_rf**(1/2.0))
```

Обчислимо R^2

```
r2_rf = r2_score(Y_test, Y_pred_rf)
print("R-squared:", r2_rf)
```

Обчислимо MAE

```
mae_rf = mean_absolute_error(Y_test, Y_pred_rf)
print("Mean Absolute Error:", mae_rf)
```

Обчислимо MAPE

```
def mean_absolute_percentage_error(Y_test, Y_pred_rf):
    Y_test = np.array(Y_test)
    Y_pred_rf = np.array(Y_pred_rf)

    # Обчислимо відсоткову помилку
    absolute_percentage_error = np.abs((Y_test - Y_pred_rf) / Y_test) * 100

    # Виключимо недійсні значення (нулі або дуже близькі до нуля)
    absolute_percentage_error = absolute_percentage_error[np.isfinite(absolute_percentage_error)]

    # Обчислимо середнє значення абсолютної відсоткової помилки
    mape = np.mean(absolute_percentage_error)

    return mape
```

```
mape_rf = mean_absolute_percentage_error(Y_test, Y_pred_rf)
print("MAPE:", mape_rf)
```

Побудуємо графік для порівняння прогнозованих і реальних значень для моделі дерева рішень

```
indexs = np.arange(0, len(Y_pred_rf), 100)
plt.plot(indexs, Y_pred_rf[indexs], label='Прогнозовані значення')
plt.plot(indexs, Y_test.values[indexs], label='Реальні значення')
plt.xlabel("Індекс спостереження")
plt.xticks(indexs, rotation='vertical')
plt.ylabel("Місячна заробітна плата ($)")

plt.title("Порівняння прогнозованих і реальних значень для моделі випадковий ліс")
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.25), ncol=2)
plt.show()
```

Визначимо важливі фактори і виведемо їх у вигляді гістограм

```
importance_list = []
name_list = []
```

```
for importance, name in sorted(zip(rf_model.feature_importances_, X_train.columns), reverse=False):
    importance_list.append(importance)
    name_list.append(name)

plt.figure(dpi=500)

plt.barh(name_list, importance_list, color='#A066CB')
plt.xlabel('Важливість змінної', labelpad=20, fontsize=12)
plt.ylabel('Назва змінної', labelpad=10, fontsize=12)

plt.title('Важливість змінних за допомогою моделі випадковий ліс')
plt.xticks(rotation='vertical')
plt.show()
```

Ім'я користувача:
Математичного моделювання та статистики Даценк...

ID перевірки:
1015410363

Дата перевірки:
04.06.2023 11:07:48 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
04.06.2023 11:12:57 EEST

ID користувача:
100005744

Назва документа: Степаненко_Осипова_Плагіат

Кількість сторінок: 43 Кількість слів: 9134 Кількість символів: 68771 Розмір файлу: 1.78 MB ID файлу: 1015073701

13.4% Схожість

Найбільша схожість: 0.88% з джерелом з Бібліотеки (ID файлу: 1000808210)

10.1% Джерела з Інтернету

263

Сторінка 45

10.7% Джерела з Бібліотеки

579

Сторінка 48

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0% Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи

3