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

Бакалаврська дипломна робота на тему:

«ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ВАРТОСТІ МЕДИЧНОГО СТРАХУВАННЯ»

Виконала: студентка 4 курсу, групи 2ІСТ-20б
спеціальності 126 «Інформаційні системи та технології»

_____ Катерина ІВАНОВА

Керівник: к.т.н., доц. каф. САІТ

_____ Ілона ВАРЧУК

«31» 05 2024 р.

Рецензент: к.т.н., доц. каф. КН

_____ Володимир ОЗЕРАНСЬКИЙ

«12» 06 2024 р.

Допущено до захисту

Завідувач кафедри САІТ

_____ д.т.н., проф. Віталій МОКІН

«03» 06 2024 р.

Вінниця ВНТУ – 2024 рік

Вінницький національний технічний університет
Факультет інтелектуальних інформаційних технологій та автоматизації
Кафедра системного аналізу та інформаційних технологій
Рівень вищої освіти – перший (бакалаврський)
Галузь знань – 12 Інформаційні технології
Спеціальність 126 – «Інформаційні системи та технології»
Освітньо-професійна програма – Прикладні інформаційні технології

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____ д.т.н., проф. Віталій МОКІН
"05" 05 2024 року

ЗАВДАННЯ
НА БАКАЛАВРСЬКУ ДИПЛОМНУ РОБОТУ СТУДЕНТЦІ
Івановій Катерині Олегівні

1. Тема роботи: «Технологія аналізу та прогнозування вартості медичного страхування»
керівник роботи: Варчук І.В., к.т.н., доц. каф. САІТ
затверджені наказом ВНТУ від «11» 05 2024 року №80
2. Термін подання студентом роботи 31.05
3. Вихідні дані до роботи:
1) Дані моніторингу вартості на медичне страхування.
4. Зміст текстової частини:
1) Загальна характеристика об'єкту дослідження;
2) Вибір оптимального рішення та налаштувань для розв'язання поставленої задачі;
3) Прогнозування вартості на медичне страхування.
5. Перелік ілюстративного матеріалу:
1) Кореляційна матриця даних вартості на медичне страхування;
2) Графік важливості ознак;
3) Таблиця порівняння точності моделей;
4) Графік порівняння прогнозованих даних найкращою моделлю та тренувальних даних;
5) Графік прогнозованих даних найкращою моделлю та реальних тестових даних.

6. Консультанти розділів роботи:

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

7. Дата видачі завдання 11.05

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва та зміст етапу	Термін виконання		Примітка
		початок	закінчення	
1	Аналіз предметної області	11.05	31.03	вик
2	Порівняльний аналіз проблематики	01.04	15.04	вик
3	Вибір оптимальних інформаційних технологій	16.04	30.04	вик
4	Прогнозування вартості медичного страхування	01.05	15.05	вик
5	Оформлення матеріалів до захисту БДР	16.05	21.05	вик

Студент _____ Катерина ІВАНОВА

Керівник роботи _____ Ілона ВАРЧУК

АНОТАЦІЯ

Бакалаврська дипломна робота складається з 72 сторінок формату А4, на яких є 53 рисунка, список використаних джерел містить 20 найменувань.

Метою роботи є аналіз та прогнозування ціни медичного страхування.

В бакалаврській роботі проведено аналіз ціноутворення медичного страхування. Проаналізовано сучасні підходи до обґрунтування ціни на медичну страховку у різних країнах. Проведено огляд інформаційних технологій, що застосовують страхові компанії для медичного страхування.

В роботі проведено аналіз даних ціни медичного страхування. За результатами аналізу проведено інженерію ознак та сформовано тренувальний та тестовий набори даних. Проаналізовано та обрано моделі машинного навчання для прогнозування ціни медичного страхування.

Натреновано та протестовано обрані моделі машинного навчання для прогнозування ціни медичного страхування. За результатами тестування відібрано оптимальну модель машинного навчання для прогнозування ціни медичного страхування. Для реалізації роботи використано мову Python.

Ключові слова: системний аналіз, прогнозування даних, машинне навчання, Python.

ABSTRACT

The bachelor's thesis consists of 72 pages of A4 format, containing 53 figures, and the list of used references includes 20 entries. The aim of the thesis is to analyze and forecast the price of medical insurance.

In the bachelor's thesis, an analysis of medical insurance pricing has been conducted. Modern approaches to justifying the price of medical insurance in different countries have been analyzed. A review of information technologies used by insurance companies for medical insurance has been carried out.

The thesis includes an analysis of medical insurance price data. Based on the analysis results, feature engineering was conducted, and training and testing datasets were formed. Models of machine learning for predicting the price of medical insurance were analyzed and selected.

The chosen models of machine learning have been trained and tested for predicting the price of medical insurance. Based on the test results, the optimal machine learning model for predicting the price of medical insurance was selected. The implementation of the work utilized the Python programming language.

Keywords: system analysis, data prediction, machine learning, Python.

ЗМІСТ

ВСТУП

Актуальність теми. В наш час прогнозування ціни медичного страхування стає все більш актуальним у контексті стрімкого зростання медичних витрат та потреби в ефективному управлінні фінансовими ресурсами у галузі охорони здоров'я. Різке збільшення цін на медичні послуги, нові медичні технології та зміни у демографічній структурі населення спонукають страхові компанії до розробки більш точних моделей для визначення страхових премій. Це важливо не тільки для підтримки фінансової стійкості страховиків, але й для забезпечення доступності медичного страхування для всіх верств населення.

Застосування передових технологій, таких як штучний інтелект і машинне навчання, дозволяє значно підвищити точність прогнозів вартості страхування, адаптуватися до змінних умов ринку та індивідуальних особливостей страхувальників. Це допомагає не тільки у формуванні більш справедливих і конкурентоспроможних тарифів, але й у вдосконаленні процесів управління ризиками.

Крім того, аналіз великих даних та прогнозування тенденцій дозволяють урядам і політикам краще планувати охорону здоров'я на національному рівні, розробляти програми громадської охорони здоров'я та визначати пріоритетні напрямки для інвестицій у медичну інфраструктуру. Це сприяє не тільки покращенню загального здоров'я населення, але й зниженню витрат на медичне обслуговування в довгостроковій перспективі.

Таким чином, актуальність теми прогнозування ціни медичного страхування обумовлена її впливом на оптимізацію витрат в галузі охорони здоров'я, забезпечення доступності медичних послуг та підвищення якості медичного обслуговування.

Мета і задачі дослідження. Метою дослідження є аналіз та прогнозування ціни медичного страхування.

Для досягнення поставленої мети необхідно розв’язати такі задачі:

– провести аналіз ціноутворення медичного страхування;

– здійснити вибір оптимальних технологій машинного навчання для прогнозування ціни медичного страхування;

– натренувати та протестувати обрані технології машинного навчання для прогнозування ціни медичного страхування.

Об’єктом дослідження є аналіз методів машинного навчання для прогнозування ціни медичного страхування.

Предметом дослідження є моделі машинного навчання для прогнозування ціни медичного страхування.

1. ЗАГАЛЬНА ХАРАКТЕРИСТИКА ОБ'ЄКТУ ДОСЛІДЖЕНЬ

1.1 Опис об'єкта досліджень

Медичне страхування є важливим елементом систем охорони здоров'я, яке забезпечує фінансовий захист та доступ до медичних послуг. Дослідження в різних країнах показують, що розширення державного медичного страхування може позитивно впливати на використання медичних послуг, фінансовий захист і загальний стан здоров'я населення, особливо в країнах із низьким і середнім рівнем доходу [1] (рис 1.1).

Рисунок 1.1 – Опис важливості медичного страхування

Наприклад, в Китаї виявлено, що різні схеми медичного страхування, такі як система страхування витрат на лікування раку легенів, сприяли зниженню загальних медичних витрат і витрат, які пацієнти сплачують з власної кишені. Це

важливо, оскільки зниження вартості лікування може забезпечити більший доступ до необхідних медичних послуг [2].

У В'єтнамі спостерігалась позитивна динаміка у використанні медичних послуг серед домогосподарств з низьким і середнім доходом після розширення обов'язкового державного медичного страхування. Зокрема, зросла кількість візитів до лікаря, що свідчить про підвищення доступності медичної допомоги для цих груп населення [3].

Також в Китаї проведено дослідження ефективності використання фондів базового медичного страхування для міських і сільських жителів, в результаті чого виявлено, що підвищення ефективності управління фондами може сприяти зниженню надмірних витрат і покращенню якості медичних послуг [4]. Ці результати підкреслюють важливість оптимізації систем медичного страхування для забезпечення їх сталого розвитку і вдосконалення доступу до медичних послуг.

Медичне страхування в Україні зазнало суттєвих змін протягом останніх років, особливо після запровадження реформ з метою покращення доступності та якості медичних послуг. Однак система все ще стикається з численними викликами. На рисунку 1.2 зображено лідерів у сфері медичного страхування в Україні.

	Страхова компанія	Премії за 9 місяців 2022 року, млн грн	Динаміка премій, % (9 міс. 2022-го до 9 міс. 2021-го)	Виплати за 9 місяців 2022 року, млн грн	Рівень виплат за 9 місяців 2022 року, %
1	Уніка	589,4	-18,7	304,1	51,6
2	Провідна	561,0	-17,2	302,1	53,9
3	ARX	365,5	-3,6	132,7	36,3
4	Інго	321,8	-25,2	187,7	58,3
5	Альфа Страхування	212,0	-37,8	159,6	75,3
6	УСГ	208,4	-20,9	119,8	57,5
7	СГ ТАС	153,6	-18,3	77,6	50,5
8	Країна	126,2	-20,2	91,5	72,6
9	ВУСО	118,5	93,0	49,7	42,0
10	Universalna	103,8	42,2	49,2	47,3

Рисунок 1.2 – Лідери у сфері медичного страхування

У 2018 році було створено Національну службу здоров'я України, що діє на засадах нових політик укладання договорів та закупівлі послуг з метою зміцнення первинної медичної допомоги. Незважаючи на реформи, велика частина медичних витрат у 2018 році покривалась за рахунок приватних витрат, зокрема виплат із власної кишені, що ставило під загрозу фінансову безпеку вразливих груп населення [5].

Загалом українська медична система показує відставання у порівнянні з іншими країнами пострадянського блоку та ЄС у показниках якості та доступності медичного обслуговування. Наприклад, відсоток вакцинації дітей проти кору значно знизився в останні роки, що свідчить про серйозні проблеми у системі громадського здоров'я [6]. На рисунку 1.3 зображена структура медичного страхування.

Рисунок 1.3 – Структура медичного страхування

Фінансування медичної системи є недостатнім, що призводить до того, що українці часто вимушені платити за медичні послуги з власної кишені, навіть за

ті, що офіційно мають бути безкоштовними за законом [7]. Це є частиною більш широкої проблеми фінансування та корупції в системі охорони здоров'я, що підкреслює потребу в подальших реформах і підвищенні прозорості.

Додатково, з початком війни в Україні, доступ до медичних послуг став ще більш обмеженим через знищення інфраструктури та логістичні проблеми, що вплинуло на здатність системи забезпечувати надання допомоги всім потребуючим [8].

На рисунку 1.4 зображено 4 моделі медичного страхування, які використовують у світі.

Рисунок 1.4 – 4 моделі медичного страхування

Модель Бісмарка – застосовується в країнах, таких як Німеччина, Австрія та Японія. Ця система заснована на страхуванні, яке фінансується спільно працівниками та роботодавцями через відрахування з заробітної плати. Страхові фонди, створені в рамках цієї моделі, відповідають за покриття медичних витрат своїх членів.

Модель Беверіджа – імплементована у Великобританії та інших країнах, ця модель фінансується з державного бюджету, що наповнюється за рахунок податків. У цій системі медичні послуги надаються безплатно в момент

отримання. Ця модель забезпечує широкий доступ до медичних послуг, однак може виникати проблема з тривалим очікуванням на отримання певних медичних послуг.

Модель національного медичного страхування – поєднує елементи двох попередніх моделей і використовується в таких країнах, як Канада та Південна Корея. В цій системі медичні послуги надаються приватними клініками, але фінансуються через державні страхові програми, до яких внески роблять усі громадяни.

Модель прямої оплати – ця модель не передбачає страхування. Пацієнти оплачують медичні послуги безпосередньо. Хоча це забезпечує високу якість обслуговування та широкий вибір, вона може бути недоступною для малозабезпечених верств населення через високі вартості послуг.

1.2 Аналіз аналогічних методів та інформаційних технологій

Аналіз сучасних інформаційних технологій у сфері медичного страхування вказує на значне вплив цих технологій на покращення обробки страхових вимог та клінічну ефективність. Використання електронних медичних записів (EMR) та інших цифрових технологій допомагає зменшити помилки у вимогах та покращує співпрацю між медичними спеціальностями, забезпечуючи кращу координацію догляду за пацієнтами [9,10]. На рисунку 1.5 зображена концепція «Лабораторна інформаційна система».

Рисунок 1.5 – Концепція «Лабораторна інформаційна система»

Оптимізація інформаційних технологій в охороні здоров'я також забезпечує значне покращення якості клінічної роботи, дозволяючи оцінювати та звітувати про клінічну якість, що стає ключовим аспектом в оплаті послуг охорони здоров'я [11]. Дослідження показують, що систематичне впровадження і використання цифрових інформаційних технологій сприяє підвищенню медичних результатів та ефективності медичних послуг.

Діджиталізація в медичному страхуванні та охороні здоров'я відкриває нові можливості для покращення доступу до медичних послуг та їх оплати, а також для захисту даних пацієнтів, що є особливо актуальним у світлі сучасних вимог до приватності та безпеки інформації.

Використання машинного навчання в медичному страхуванні прогресує, особливо у напрямку прогнозування витрат на медичне страхування та виявлення шахрайства. Основні напрямки досліджень включають використання регресійних ансамблевих моделей машинного навчання для прогнозування вартості медичного страхування, а також застосування зрозумілих методів штучного інтелекту для визначення ключових чинників, що впливають на ціни премій страхування [12, 13].

Дослідження показують, що машинне навчання може значно покращити точність прогнозів витрат, що дозволяє страховим компаніям краще управляти ризиками і встановлювати вартість страхових полісів. Наприклад, моделі, такі як екстремальне градієнтне підсилення, машина градієнтного підсилення та випадкові ліси, використовуються для аналізу медичних страхових витрат і демонструють високу точність [12, 13].

Крім того, використання машинного навчання для виявлення шахрайства у медичному страхуванні також зазнало вагомого прогресу. Дослідження показують, що інтеграція різноманітних даних і методів машинного навчання може підвищити здатність систем виявляти шахрайські дії, покращуючи загальну безпеку та ефективність медичних страхових систем [14].

Ці дослідження відкривають нові перспективи для застосування штучного інтелекту у медичному страхуванні, забезпечуючи не тільки покращення адміністративних процесів, але й збільшення доступності та якості страхових послуг для клієнтів.

Прогнозування ціни медичного страхування може використовувати різноманітні методики та підходи, серед яких домінують статистичні моделі та машинне навчання. Основні параметри, що впливають на вартість страховки, включають:

1. Демографічні фактори : вік, стать та сімейний стан часто враховуються, оскільки вони мають статистично підтверджений вплив на ризики здоров'я.

2. Медичний анамнез: історія хвороб, наявність хронічних захворювань, попередні медичні втручання.

3. Спосіб життя: куріння, рівень фізичної активності, харчові звички.

4. Географічні фактори: місце проживання може впливати на вартість страхування через варіативність медичних тарифів та рівень медичних послуг у різних регіонах.

Методики прогнозування:

– Статистичні моделі – лінійна регресія та логістична регресія часто використовуються для оцінки відносних ризиків на основі вищезгаданих параметрів.

– Машинне навчання – алгоритми, такі як випадкові ліси, градієнтний бустінг і нейронні мережі, дозволяють враховувати великі обсяги даних та їх складні взаємозалежності, забезпечуючи більш точні прогнози.

– Використання великих даних – дані про здоров'я, отримані з медичних записів, соціальних мереж і носимих пристроїв, можуть використовуватись для більш деталізованого аналізу поведінки та стану здоров'я.

Ці моделі та методології допомагають страховим компаніям:

– Оптимізувати страхові внески залежно від ризику.

– Управляти ризиками на основі передбачуваної вартості страхових виплат.

– Забезпечити справедливість у встановленні страхових премій для різних груп населення.

Прогнозування вартості медичного страхування є ключовим для забезпечення фінансової стійкості страхових програм і доступності медичних послуг для широких верств населення. Це також дозволяє регулювати і контролювати медичний ринок, сприяючи його більшій ефективності та транспарентності.

1.3 Висновки

У цьому розділі було проведено аналіз предметної області прогнозування ціни медичного страхування, а саме процес ціноутворення медичного страхування. Які фактори особливо впливають на розмір медичної страховки та які моделі медичного страхування використовують у світі та Україні.

Описано вплив інформаційних технологій на сферу охорони здоров’я. Проаналізували методи прогнозування ціни на медичне страхування.

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

2.1 Аналіз даних цін медичного страхування для прогнозування даних

Для аналізу та створення моделі прогнозування ціни на медичне страхування обрано дані з платформи Kaggle, а саме датасет «Medical Insurance Cost Prediction». Даний датасет містить дані що впливають на медичні витрати, такі як:

- Вік;
- Стать;
- ІМТ (індекс маси тіла);
- Кількість дітей;
- Чи палить клієнт;
- Регіон
- Вартість страхування

Опишемо кожну ознаку та її вплив на ціноутворення медичного страхування.

Вік має значний вплив на вартість медичного страхування з кількох причин:

1. Зростання ризику з віком – зі старінням збільшується ризик захворювань та потреба у медичному обслуговуванні. Страхові компанії враховують цей зростаючий ризик, встановлюючи вищі премії для старших осіб.

2. Статистичні дані про здоров'я. Дослідження та статистика показують, що з віком зростає частота візитів до лікарів, госпіталізацій, а також потреба в дорогих медичних процедурах та тривалому лікуванні. Відповідно, вищі страхові внески для літніх людей відображають більш високі очікувані витрати на охорону здоров'я.

3. Економічна модель страхування. Вартість страхування розраховується на основі прогнозованих витрат на медичне обслуговування. Страховики

використовують демографічні дані для оцінки середніх витрат на певні вікові групи.

На практиці, молодші особи зазвичай сплачують менші страхові премії, адже вони здоровіші та рідше потребують медичної допомоги. У той же час, старші особи сплачують вищі премії, оскільки їхній ризик та потреба в медичному обслуговуванні значно вищі.

Цей принцип діє у більшості країн, незалежно від конкретної моделі медичного страхування, чи то приватне страхування, чи державні медичні програми, як у США (Medicare для літніх осіб) або у країнах з універсальним охоронним страхуванням.

Вплив статі на ціну медичної страховки може варіюватися в залежності від країни, законодавства та конкретних політик страхових компаній. Ось деякі аспекти, які слід враховувати:

Законодавчі відмінності.

В деяких країнах, таких як Сполучені Штати до впровадження Закону про доступну медичну допомогу (Affordable Care Act), страхові компанії могли встановлювати різні тарифи залежно від статі, оскільки жінки, як правило, використовували більше медичних послуг, особливо у віці репродуктивної функції. Після 2014 року, це стало незаконним, і страховики більше не можуть встановлювати ціни виходячи зі статі у Сполучених Штатах.

Медичні витрати.

Статистично, жінки частіше відвідують лікарів, проходять медичні обстеження і витрачають більше на медичні послуги, що теоретично може вплинути на страхові тарифи. Це особливо актуально для послуг, пов'язаних з вагітністю та пологами.

Ризики і вартість полісів.

З іншого боку, чоловіки можуть мати більш високі ризики для деяких хвороб, що також може впливати на ціноутворення страховиками, особливо в країнах та системах, де дозволено враховувати статеві відмінності у вартості страхових полісів.

Вплив культури і політики.

У деяких країнах, таких як країни Європейського Союзу, дискримінація за статевою ознакою у страхових тарифах заборонена. Так, згідно з директивою ЄС 2004/113/ЕС, встановлення різних тарифів на основі статі вважається дискримінацією.

Враховуючи ці аспекти, вплив статі на ціну медичного страхування може значно відрізнятися в залежності від регіональних особливостей, законодавства і політик конкретної страхової компанії. Важливо знати місцеве законодавство та ринкові умови перед вибором страхового покриття.

Вплив індексу маси тіла (BMI) на вартість медичної страховки значний, особливо у контексті страхування здоров'я. Високий BMI, який вказує на надмірну вагу або ожиріння, може призвести до збільшення страхових премій, оскільки люди з високим BMI зазвичай стикаються з більшим ризиком розвитку хронічних захворювань, таких як діабет, серцево-судинні захворювання та інші умови, що потребують дорогого лікування.

Дослідження виявили, що ожиріння може становити до 21% від медичних витрат, залежно від даних і методології оцінки, з більш новими даними, що показують навіть більший вплив на витрати на охорону здоров'я. Це впливає не лише на індивідуальні витрати, але й загальну економічну навантаження на системи охорони здоров'я.

Щоб зменшити вплив високого BMI на вартість страховки, рекомендується підтримувати здоровий спосіб життя, що включає збалансоване харчування та регулярні фізичні вправи, що може допомогти зменшити вагу до здорового рівня BMI і таким чином знизити страхові премії.

Також, для людей з високим BMI корисно регулярно проходити медичні обстеження для раннього виявлення і лікування потенційних проблем зі здоров’ям, що також може допомогти знизити загальні витрати на страхування.

Кількість дітей може мати значний вплив на ціну медичної страховки, оскільки більша кількість дітей у сім'ї збільшує потенційні медичні витрати, що покриваються страхуванням. Ось основні аспекти впливу кількості дітей на вартість медичного страхування:

Вищі медичні витрати:

– Більша кількість дітей вимагає більшого покриття через збільшену кількість регулярних медичних оглядів, вакцинацій та потенційних медичних проблем.

– Діти частіше відвідують лікаря, що веде до збільшення витрат на страхові поліси.

Премії та покриття:

– Страхові компанії зазвичай підвищують премії для сімей із дітьми для компенсації збільшення ризику і витрат.

– Поліси, які покривають усю сім'ю, зазвичай дорожчі, але також надають краще покриття за різними аспектами охорони здоров'я дітей.

Законодавство та політика:

– В деяких країнах є законодавчі обмеження на те, наскільки страхові премії можуть зростати залежно від кількості дітей у полісі, щоб захистити сім'ї від дискримінаційного підвищення цін.

– В країнах з розвиненими системами соціального забезпечення можуть бути програми, які допомагають сім'ям з дітьми оплачувати страхові премії.

Сім'ям з багатьма дітьми може бути вигідно шукати страхові плани, які пропонують гнучкі умови покриття або спеціальні знижки для багатодітних сімей. Також важливо регулярно переглядати умови страхування та порівнювати пропозиції різних страхових компаній, щоб знайти найбільш вигідний варіант.

Вплив куріння на ціну медичної страховки є значним і відомим фактором, який страхові компанії беруть до уваги при розрахунку премій. Ось декілька ключових моментів:

Підвищені медичні ризики:

– Куріння сигарет значно збільшує ризик розвитку багатьох хронічних захворювань, включаючи серцево-судинні захворювання, рак легенів, хвороби дихальних шляхів, інсульти, а також інші форми раку, які не обмежуються лише легенями. Ці стани часто вимагають тривалого та дорогого лікування.

Збільшення витрат на страхування:

– Курці, як правило, платять значно вищі страхові премії порівняно з некурцями. Це пов'язано з тим, що страхові компанії прагнуть компенсувати вищий ризик витрат, які куріння накладає на медичні послуги.

– Страхові компанії використовують розрахунок ризиків для встановлення тарифів, і куріння є одним з ключових факторів у цьому процесі.

Законодавча регуляція:

– В деяких країнах, наприклад, у Сполучених Штатах, законодавство дозволяє страховикам встановлювати тарифи для курців до 50% вище, ніж для некурців у рамках індивідуальних і сімейних планів медичного страхування. Це відображено в положеннях Закону про доступну медичну допомогу (Affordable Care Act).

Соціально-економічний вплив:

– Курці частіше стикаються з медичними станами, які можуть призвести до втрати працездатності та інших соціально-економічних проблем, що також впливає на загальні витрати на охорону здоров'я в суспільстві.

– Програми, що спонукають людей відмовитися від куріння, можуть значно знизити страхові премії та загальні витрати на медичне страхування.

Куріння є одним з найбільш значущих індивідуальних факторів, що впливають на вартість медичного страхування через його сильний зв'язок з багатьма ризиками для здоров'я. Враховуючи це, багато страхових компаній і суспільств в цілому вкладають зусилля у профілактику куріння та підтримку програм зупинення куріння як засобу зниження витрат на охорону здоров'я.

Вплив регіону на ціну медичної страховки може бути значним, оскільки різні регіони мають різні медичні витрати, доступність послуг, рівень здоров'я населення та законодавчі рамки, які впливають на медичні страхові тарифи. Ось декілька ключових аспектів, через які регіон може впливати на вартість медичного страхування:

Витрати на медичні послуги можуть значно відрізнятися між регіонами через:

– В регіонах з високою вартістю життя, як правило, вищі й медичні витрати.

– У регіонах, де високі зарплати, можуть бути вищі витрати на медичне обслуговування.

– У місцевостях, де частіше вдаються до високовартісних медичних процедур, страхові премії можуть бути вищими.

Статистика здоров'я населення:

– Регіони з високим рівнем хронічних захворювань або поганими показниками здоров'я загалом можуть мати вищі страхові премії.

– Регіони, де населення веде більш здоровий спосіб життя, можуть мати нижчі страхові тарифи через нижчі медичні ризики.

Законодавство та регуляції:

– Регіональні закони можуть впливати на медичні тарифи, включаючи регуляції, що обмежують максимальний розмір страхових премій або закони, що вимагають певного рівня покриття.

– Регіональні політики з охорони здоров'я, такі як доступність страхових програм або державних медичних послуг, також впливають на вартість страхування.

Конкуренція на страховому ринку:

– У регіонах з більшою кількістю страхових компаній конкуренція може знижувати ціни страхових полісів.

– Велика кількість доступних страхових продуктів може дати споживачам кращі можливості для вибору і кращі ціни.

Розуміння впливу регіону на вартість медичного страхування важливе для страхових компаній та політиків для розробки ефективних стратегій зниження витрат і підвищення доступності медичного страхування. Споживачам важливо враховувати цей аспект при виборі страхового полісу та розгляді можливостей зниження вартості медичного страхування.

Візуалізуємо дані датасету «Medical Insurance Cost Prediction» на рисунку 2.1.

```
data.head(5)
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Рисунок 2.1 – Дані датасету «Medical Insurance Cost Prediction»

Відповідно дані датасету «Medical Insurance Cost Prediction» діляться на два типи: категоріальні та числові, не мають відсутніх даних. Об’єм даних складає 2772 рядки (рис. 2.2).

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       2772 non-null   int64
 1   sex       2772 non-null   object
 2   bmi       2772 non-null   float64
 3   children  2772 non-null   int64
 4   smoker    2772 non-null   object
 5   region    2772 non-null   object
 6   charges   2772 non-null   float64
```

Рисунок 2.2 – Інформація про типи даних

Описова статистика даних за допомогою describe (рис. 2.3).

	age	bmi	children	charges
count	2772.000000	2772.000000	2772.000000	2772.000000
mean	39.109668	30.701349	1.101732	13261.369959
std	14.081459	6.129449	1.214806	12151.768945
min	18.000000	15.960000	0.000000	1121.873900
25%	26.000000	26.220000	0.000000	4687.797000
50%	39.000000	30.447500	1.000000	9333.014350
75%	51.000000	34.770000	2.000000	16577.779500
max	64.000000	53.130000	5.000000	63770.428010

Рисунок 2.3 – Статистична оцінка даних

Оскільки існує велика різниця між мінімальним та максимальним значенням ціни медичної страховки, доцільно перевірити дані «charges» на викиди та аномалії. Для цього скористаємось графіком «ящик з вусиками», що зображено на рисунку 2.4.

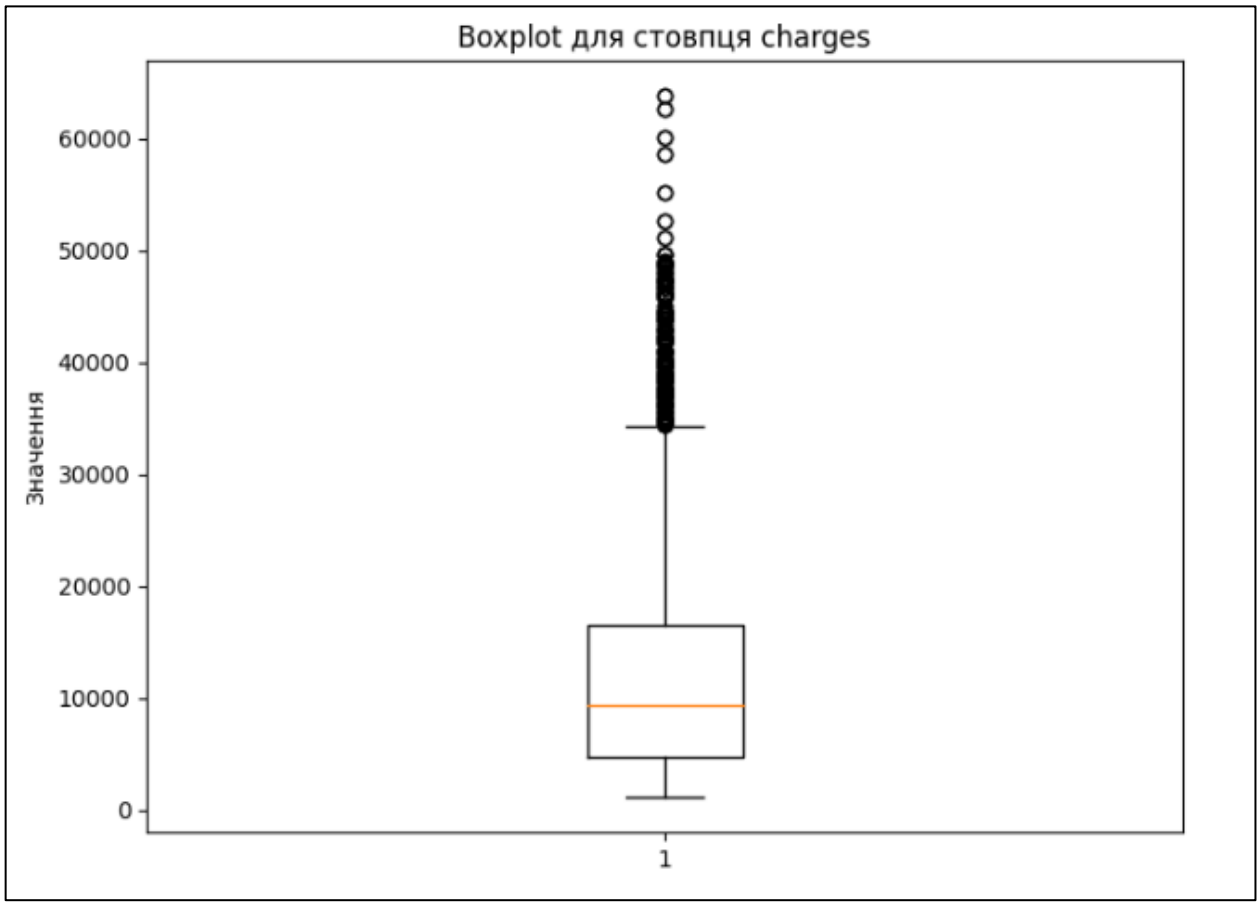


Рисунок 2.4 – Перевірка на аномалії

На рисунку 2.4 зображено графік боксплот (boxplot) для даних стовпця "charges". Ось його основні елементи та характеристики:

1. Медіана (середнє значення) позначена оранжевою лінією всередині "коробки". Це значення ділить ваші дані на дві рівні частини, де половина значень менша за медіану, а інша половина — більша.

2. Квартилі:

– Нижня межа "коробки" відповідає першому квартилю (25-й перцентиль), показуючи, що 25% значень менше за це.

– Верхня межа "коробки" відповідає третьому квартилю (75-й перцентиль), показуючи, що 75% значень менше за це.

3. Вуса (whiskers):

Вуса графіка вказують на мінімальні та максимальні значення даних, що не є викидами. Вони зазвичай охоплюють дані, які знаходяться в межах 1.5 міжквартильних розмахів від першого або третього квартилю.

4. Викиди (outliers):

Точки, розташовані над вусами, представляють викиди. Це значення, які відрізняються від інших даних, можливо, через помилки введення даних або через їхню природну варіабельність.

Відфільтруємо дані за допомогою Z-score, для ідентифікації викидів та їх видалення (рис. 2.5). Проведемо статистичну оцінку відфільтрованих даних (рис. 2.6).

```
data=data[np.abs(stats.zscore(data['charges'])) < 3]
```

Рисунок 2.5 – використання Z-score

	age	bmi	children	charges
count	2758.000000	2758.000000	2758.000000	2758.000000
mean	39.088470	30.672219	1.103698	13035.810151
std	14.089233	6.120764	1.215357	11757.338984
min	18.000000	15.960000	0.000000	1121.873900
25%	26.000000	26.205000	0.000000	4676.641325
50%	39.000000	30.400000	1.000000	9288.026700
75%	51.000000	34.752500	2.000000	16281.596250
max	64.000000	53.130000	5.000000	49577.662400

Рисунок 2.6 – Описова статистика відфільтрованих даних

Побудуємо графіки числових ознак для перевірки розподілу та побудуємо лінію тренду, щоб визначити тенденції змін у даних. Графік розподілу та тренду ознаки «age» (рис. 2.7).

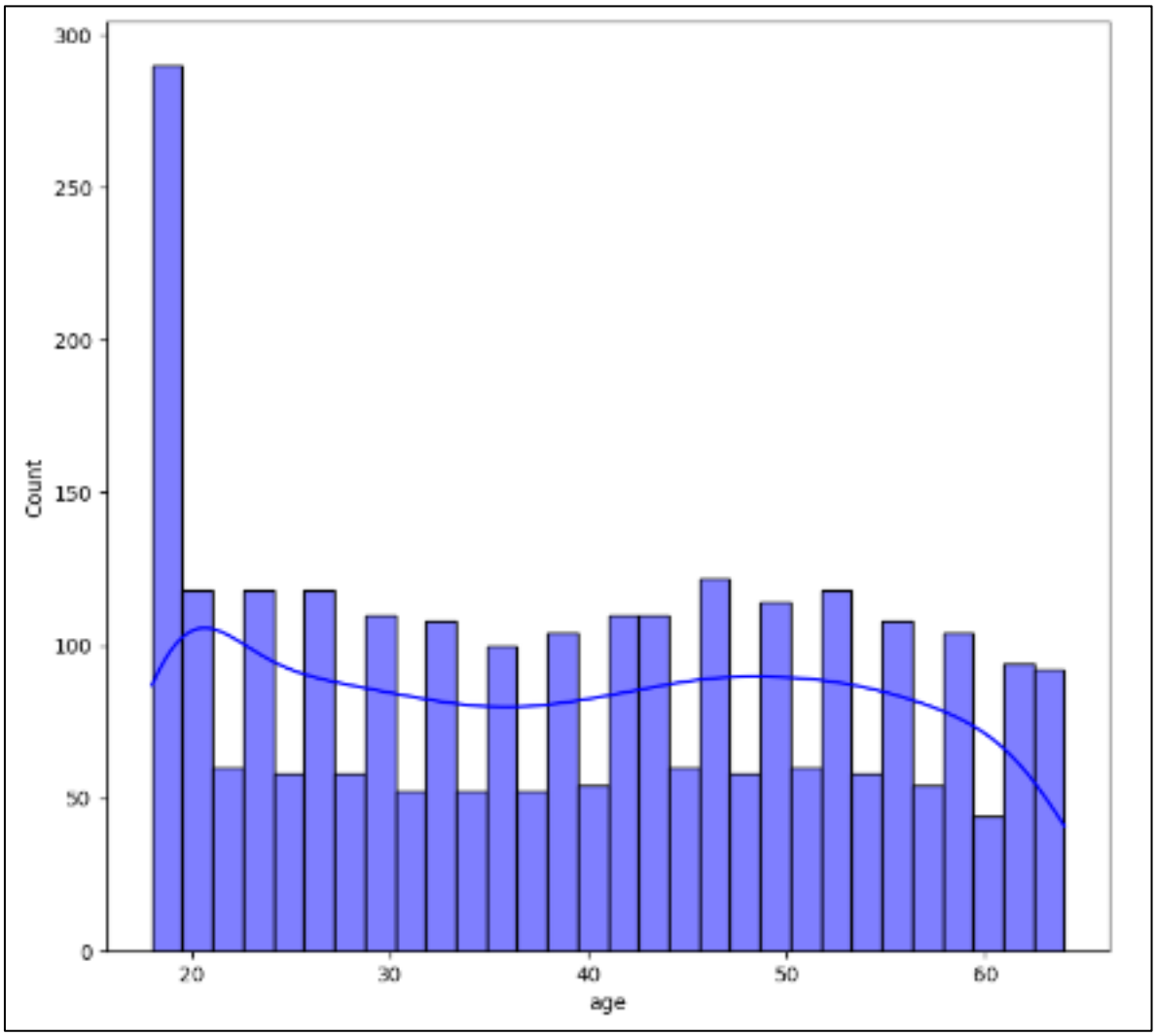


Рисунок 2.7 – Графік розподілу та тренду ознаки «age»

На графіку представлений розподіл кількості осіб за віком, зображений за допомогою стовпчикової діаграми, з накладеною кривою, яка показує загальний тренд. Основні спостереження з графіку такі:

– Спайк у молодому віці, видно значний сплеск у кількості осіб на початку 20-ти років. Це може свідчити про те, що дуже багато молодих людей, можливо студентів або молодих професіоналів, включено в цей набір даних.

– Зменшення кількості з віком, кількість осіб у кожному віковому інтервалі поступово зменшується після 20-х років, з невеликим підйомом у віці 50-60 років.

– Крива тренду, накладена крива показує загальний тренд зменшення кількості осіб з віком, з винятком вищезгаданого підвищення близько 50-ти років.

Графік розподілу та тренду ознаки «bmi» (рис 2.8)

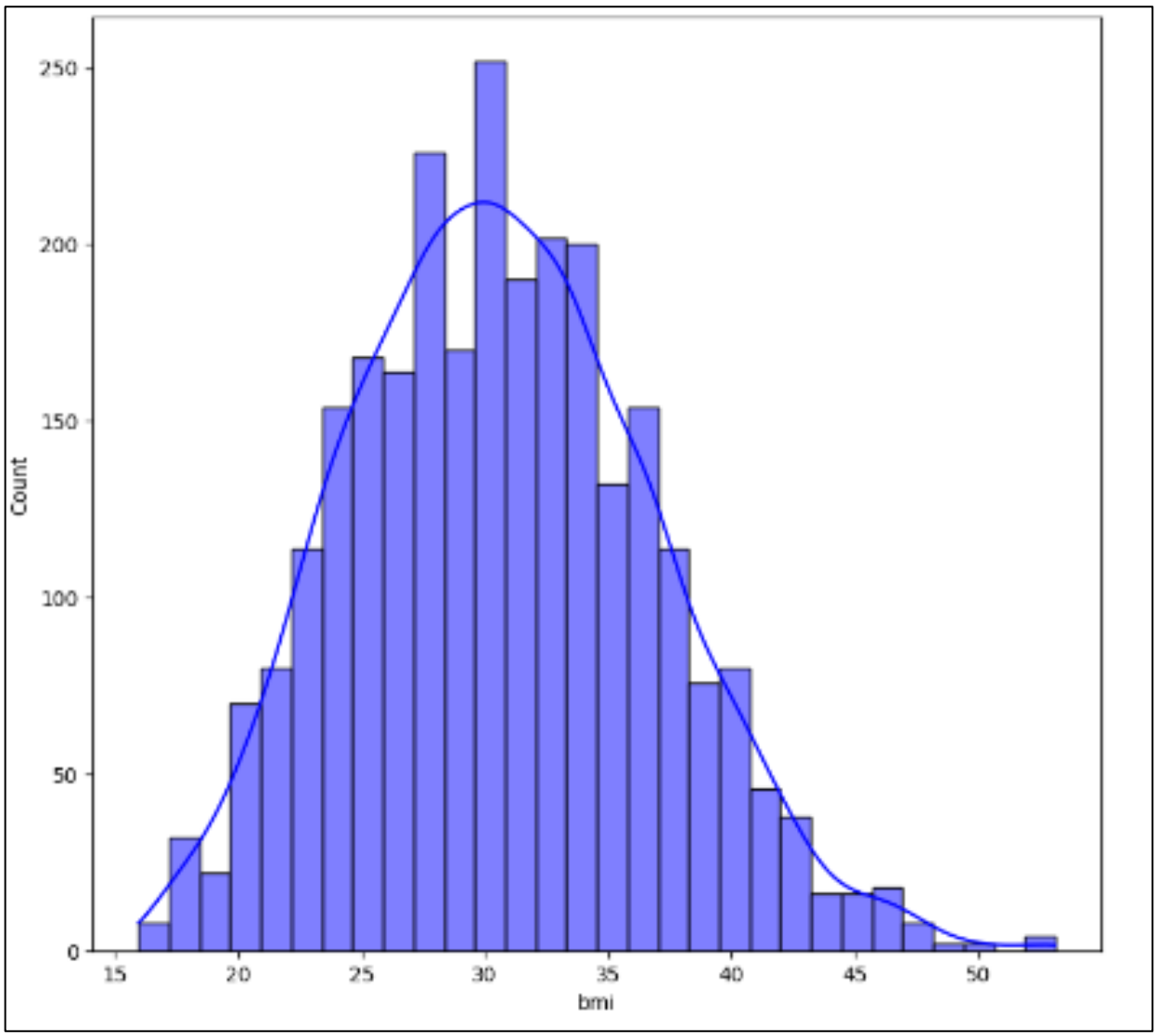


Рисунок 2.8 – Графік розподілу та тренду ознаки «bmi»

Графік, зображує розподіл індексу маси тіла (BMI) серед деякої групи людей. Видно, що дані утворюють дзвінообразний розподіл, який є характеристикою нормального розподілу. Нормальний розподіл, також відомий як розподіл Гаусса, часто використовується для моделювання фізичних, біологічних та соціальних явищ, коли дані згруповані навколо середнього значення.

Ось основні характеристики нормального розподілу, які можна визначити з графіка:

– Розподіл даних є симетричним відносно середнього значення, що вказує на нормальний розподіл.

– Графік має дзвінообразну форму, яка є типовою для нормального розподілу.

– Найбільша частота спостерігається в середньому діапазоні значень BMI, який зменшується по обидва боки від середини.

Накладена на графік лінія показує теоретичну криву нормального розподілу, яка відповідає емпіричним даним. Це використовується для візуального підтвердження того, що спостережувані дані можуть бути адекватно описані за допомогою нормального розподілу.

На рисунку 2.9 графік та тренд ознаки «children».

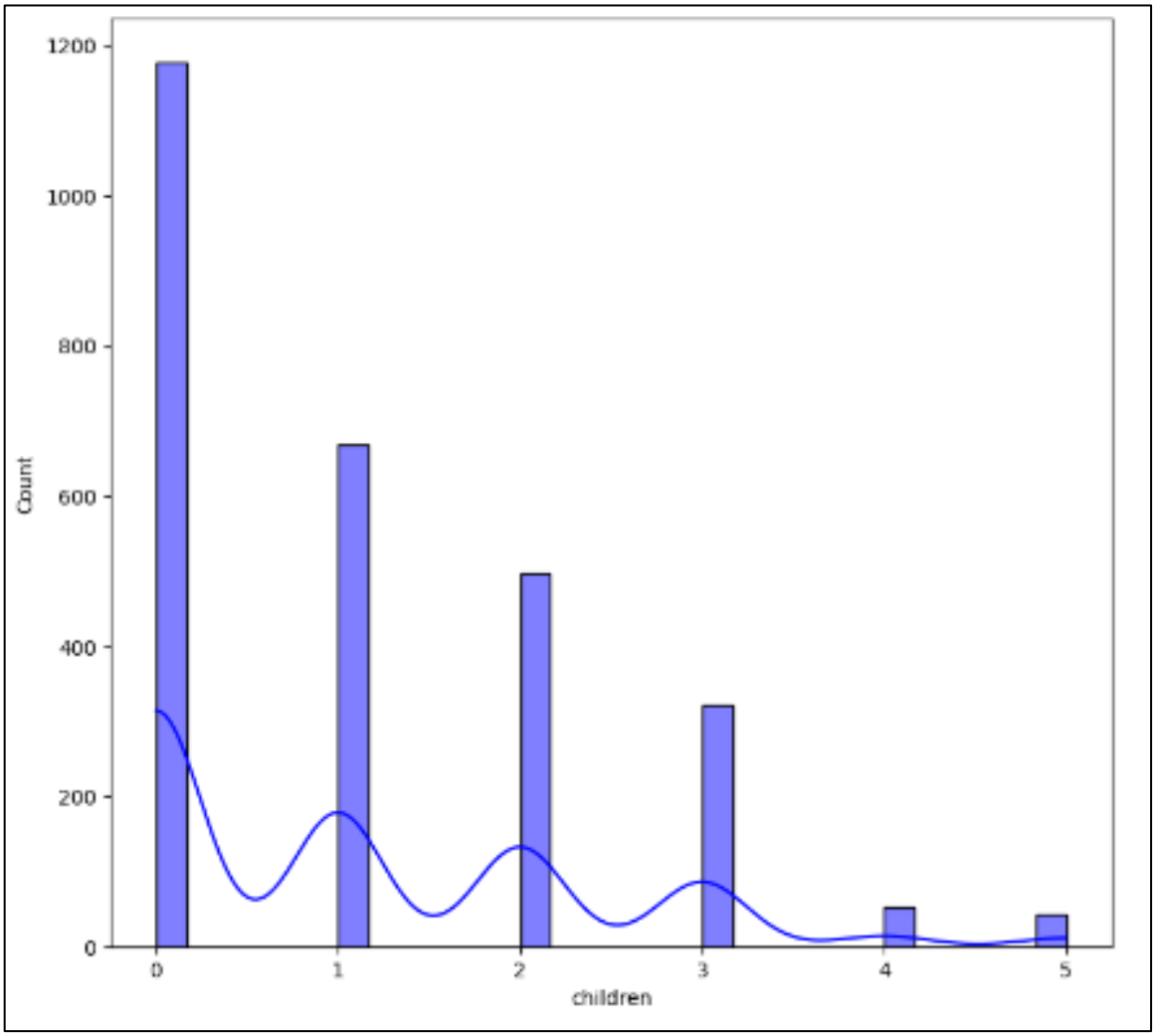


Рисунок 2.9 – Графік розподілу та тренду ознаки «children»

На графіку, показано розподіл кількості осіб за кількістю дітей. Цей розподіл не відповідає типовим неперервним розподілам, як нормальний чи експоненціальний, оскільки дані дискретні і представляють кількість дітей у сім'ях.

Графік має наступні характеристики:

– Найбільша кількість осіб має 0 дітей, після чого йде зменшення зі збільшенням кількості дітей.

– Чим більше дітей у сім'ї, тим менше таких сімей представлено в даному наборі даних.

– Кожне можливе значення (кількість дітей) представлене окремим стовпцем.

Цей тип розподілу нагадує розподіл Пуассона або геометричний розподіл, які часто використовуються для моделювання кількості настання певних подій протягом фіксованого періоду або в певній групі.

Дані вказують на те, що велика частина людей не має дітей, і з кожною додатковою дитиною кількість сімей різко знижується, що може бути типово для певних соціально-економічних умов або культурних особливостей вибірки.

На рисунку 2.10 графік та тренд ознаки «charges».

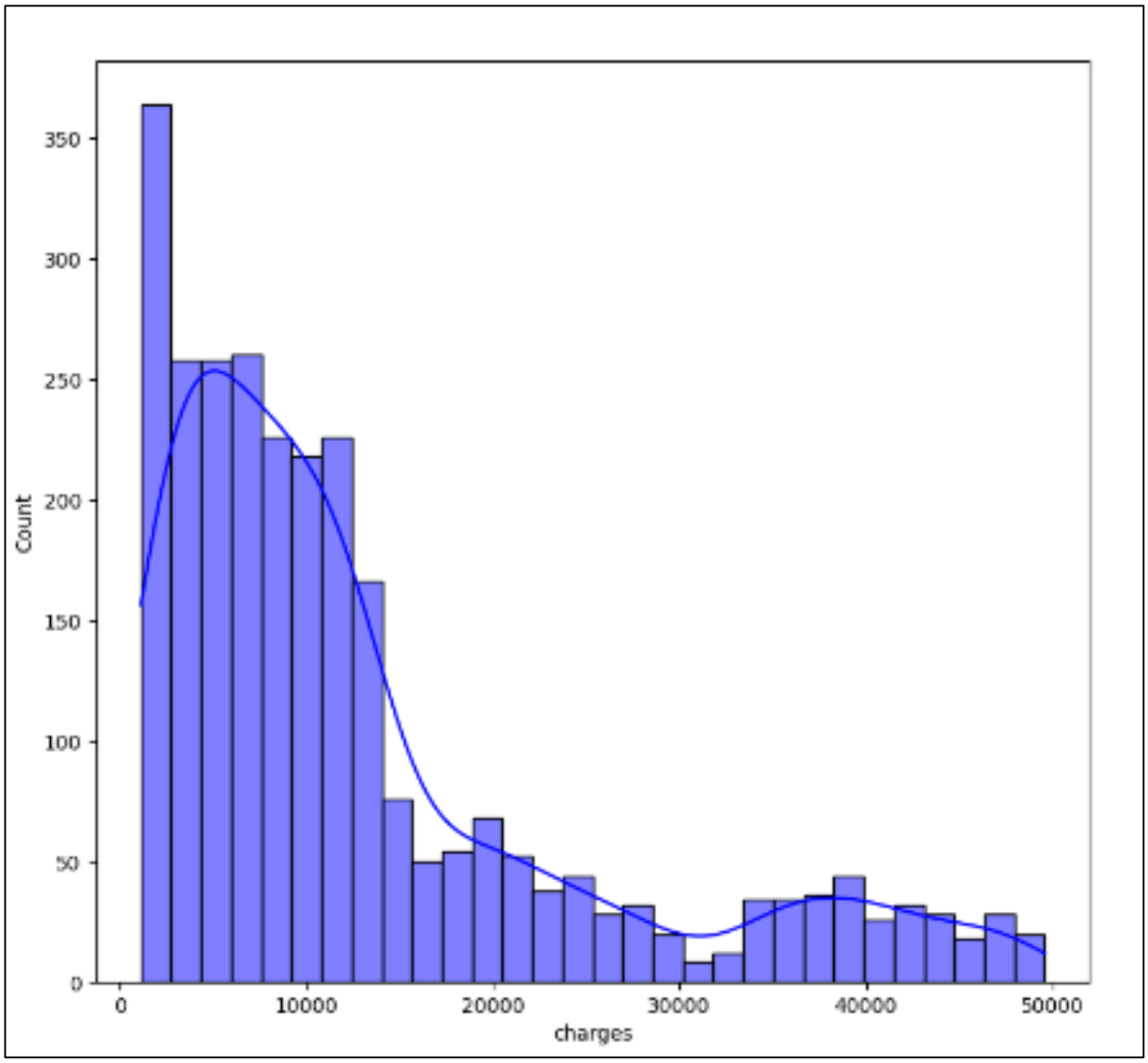


Рисунок 2.10 – Графік та тренд ознаки «charges»

На графіку зображено розподіл вартості медичного страхування, який є асиметричним із вираженою правою асиметрією (тобто з довгим правим хвостом). Цей тип розподілу вказує на те, що більшість спостережень зосереджена на менших вартостях, тоді як дуже високі вартості відносно рідкісні.

Ось кілька характеристик такого розподілу:

– Найбільш часте значення (тобто пік гістограми) знаходиться близько найнижчих вартостей страхування.

– Ймовірно, середнє значення більше медіани через вплив великих значень у довгому хвості.

Цей розподіл може бути найкраще описаний за допомогою розподілу Парето або лог-нормального розподілу, оскільки обидва вони часто використовуються для опису економічних даних, де невелика частина даних має значно вищі значення порівняно з рештою.

Проведемо аналіз категоріальних даних, а саме «Стать», «Куріння», «Регіон». Побудовані кругові діаграми на рисунку 2.11.

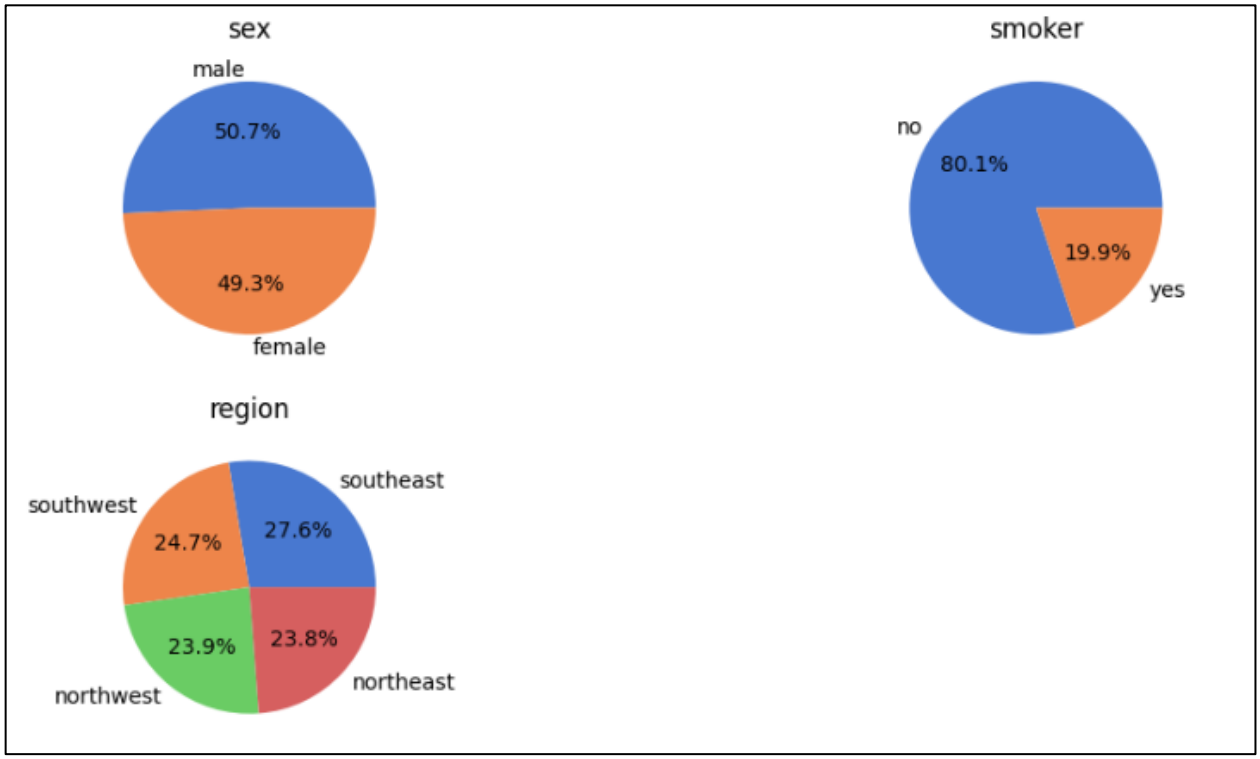


Рисунок 2.11 – Кругові діаграми категоріальних ознак

Графік "Sex", показує майже рівномірний розподіл між чоловіками (50.7%) та жінками (49.3%). Це свідчить про те, що вибірка є збалансованою з точки зору статевої приналежності.

Графік "Smoker", значна більшість (80.1%) опитаних не курять, тоді як 19.9% є курцями. Цей розподіл може впливати на вартість медичного страхування, оскільки куріння збільшує медичні ризики.

Графік "Region", географічний розподіл, показує, що опитувані розподілені майже рівномірно між чотирма регіонами: південно-східний (27.6%), південно-західний (24.7%), північно-східний (23.8%), та північно-

західний (23.9%). Цей рівномірний розподіл допомагає забезпечити, що аналіз є репрезентативним по всій території.

Для побудови кореляційної матриці, категоріальні ознаки перетворимо на числові (рис. 2.12, 2.13).

```
data_corr= data.copy()
label_encoder = LabelEncoder()
for col in ['sex','smoker','region']:
    data_corr[col] = label_encoder.fit_transform(data_corr[col])
data_corr.head()
```

Рисунок 2.12 – Код перетворення категоріальних ознак в числові

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

Рисунок 2.13 – Перетворені дані

Побудуємо кореляційну матрицю даних (рис. 2.14, 2.15).

```
plt.figure(figsize=(16, 6))
heatmap = sns.heatmap(data_corr.corr(), vmin=-1, vmax=1, annot=True, cmap='BrBG')
```

Рисунок 2.14 – Код кореляційної матриці

Для реалізації побудови кореляційної матриці використали бібліотеку Python – Seaborn, та функцію бібліотеки Pandas – .corr.

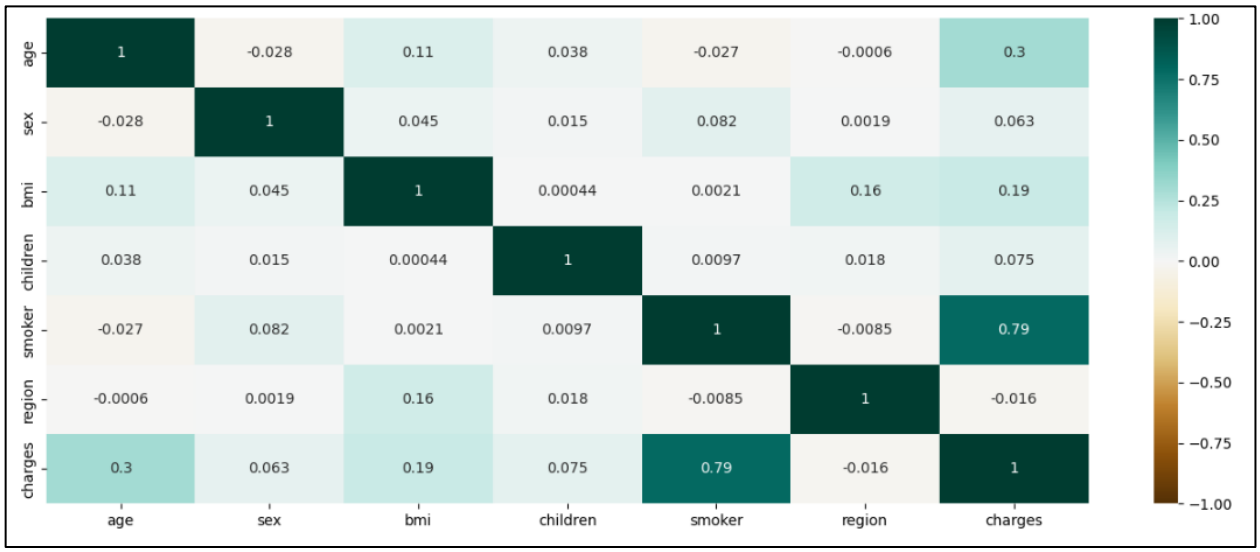


Рисунок 2.15 – Кореляційна матриця даних про медичне страхування

На даній кореляційній матриці відображені кореляційні зв'язки між різними змінними, що мають вплив на медичне страхування. Значення кореляції варіюються від -1 до 1, де значення ближче до 1 або -1 показують сильну позитивну або негативну кореляцію відповідно, а значення ближче до 0 вказують на слабкий або відсутній зв'язок.

Ось декілька ключових спостережень з цієї матриці:

– Кореляція між "smoker" та "charges" (0.79): Це вказує на дуже сильний позитивний зв'язок, що означає, що курці, як правило, мають значно вищі медичні витрати порівняно з некурцями.

– Кореляція між "bmi" та "charges" (0.19): Є позитивний зв'язок між індексом маси тіла та витратами на страхування, що може свідчити про вищі медичні витрати для осіб з вищим BMI.

– Кореляція між "age" та "charges" (0.3): Позитивний зв'язок вказує на те, що з віком витрати на медичне страхування зростають, що логічно, оскільки старші люди частіше потребують медичного обслуговування.

– Більшість інших кореляцій досить низькі, що вказує на відсутність сильного зв'язку між цими змінними. Наприклад, зв'язок між "region" і "charges" дуже малий (-0.016), що вказує на те, що географічний регіон має мало спільного з загальними витратами на страхування.

2.2 Інженерія ознак даних цін медичного страхування для прогнозування даних

Використовує бібліотеку Pandas у Python для трансформації категоріальних змінних у двійкові індикаторні (dummy) змінні та об’єднаємо отримані дані з основними числовими даними (рис. 2.16, 2.17).

```
binar = pd.get_dummies(data[['sex','region','smoker']],dtype=int)
data_model = pd.concat([data[['age','bmi','children','charges']],binar],axis=1)
data_model
```

Рисунок 2.16 – Код трансформація даних та об’єднання з числовими даними

	age	bmi	children	charges	sex_female	sex_male	region_northeast	region_northwest	region_southeast	region_southwest	smoker_no	smoker_yes
0	19	27.900	0	16884.92400	1	0	0	0	0	1	0	1
1	18	33.770	1	1725.55230	0	1	0	0	1	0	1	0
2	28	33.000	3	4449.46200	0	1	0	0	1	0	1	0
3	33	22.705	0	21984.47061	0	1	0	1	0	0	1	0
4	32	28.880	0	3866.85520	0	1	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
2767	47	45.320	1	8569.86180	1	0	0	0	1	0	1	0
2768	21	34.600	0	2020.17700	1	0	0	0	0	1	1	0
2769	19	26.030	1	16450.89470	0	1	0	1	0	0	0	1
2770	23	18.715	0	21595.38229	0	1	0	1	0	0	1	0
2771	54	31.600	0	9850.43200	0	1	0	0	0	1	1	0

Рисунок 2.17 – Представлення нового набору даних

Розіб’ємо основний датасет на X – ознаки та Y – основна ознака. Наступним кроком дослідження буде формування навчальної (X_train, y_train) та тестової (X_test, y_test) вибірки даних. Для цього скористаємось бібліотекою sklearn.model_selection та її функцією train_test_split (рис. 2.18).

```
X=data_model.drop(columns=['charges'])
y=data_model["charges"]


X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.25,random_state=7)
```

Рисунок 2.18 – Формування навчальної (X_train, y_train) та тестової (X_test, y_test) вибірки даних

Для перевірки важливості ознак, натренуємо модель XGBoost та використаємо її функцію plot_importance (рис. 2.19, 2.20).

```
model = xgb.XGBRegressor(objective ='reg:squarederror',
                         colsample_bytree = 0.3,
                         learning_rate = 0.1,
                         max_depth = 5,
                         alpha = 10,
                         n_estimators = 10)

# Тренуйте модель
model.fit(X_train, y_train)
```

Рисунок 2.19 – Код XGBoost

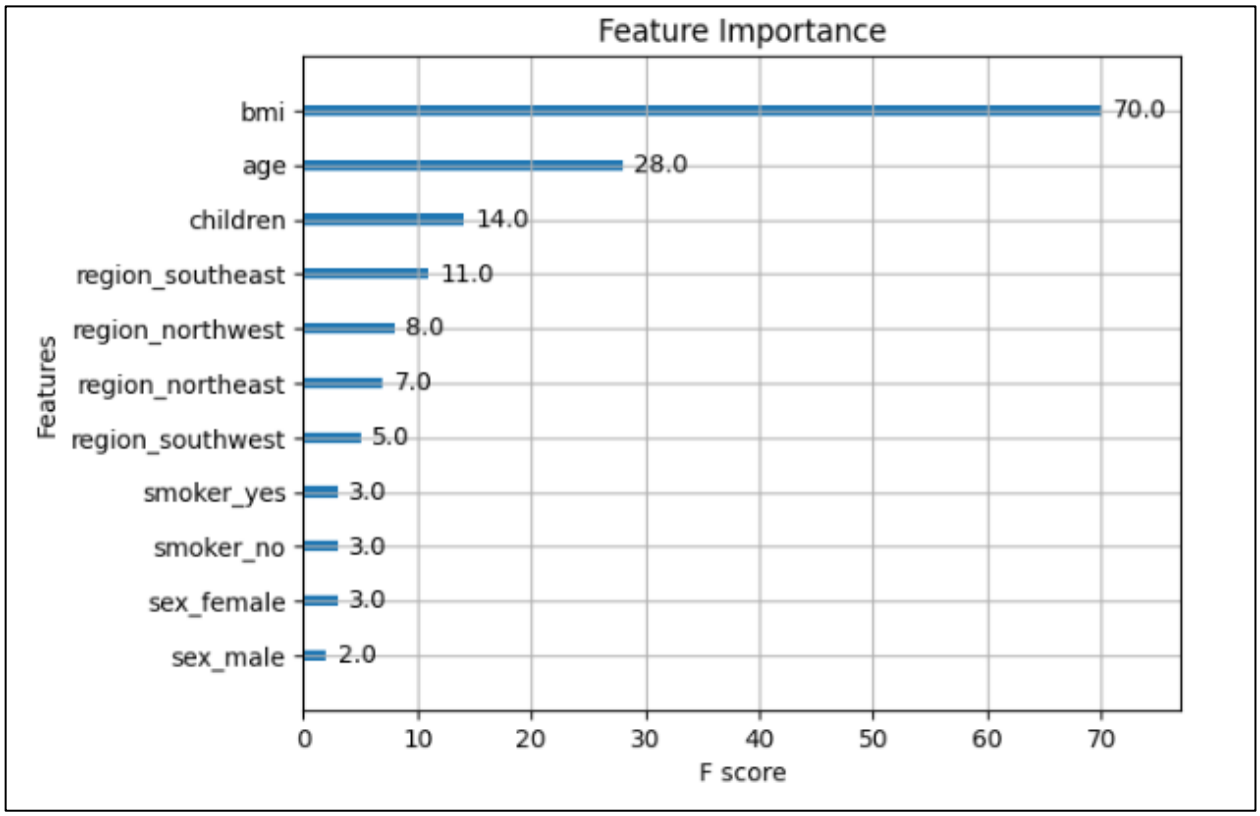


Рисунок 2.20 – Графік важливості ознак

На графіку важливості ознак, що представлений у вигляді стовпчикової діаграми, показано значення F-score для кожної з ознак у моделі машинного навчання. Ці значення вказують на те, як сильно кожна ознака впливає на прогнози моделі. Важливість ознак визначена за їх внесок у конструкцію моделі, наприклад, за кількість разів, коли ознака використовується для розгалуження у деревах, що є частиною ансамблю моделі.

Опис ознак на графіку:

– bmi (70.0) – ця ознака має найвищий F-score, що свідчить про її значний вплив на модель. Індекс маси тіла (BMI) є важливим фактором у медичному страхуванні, оскільки він пов'язаний із ризиком багатьох захворювань.

– age (28.0) – вік також має значний вплив на модель, оскільки він часто корелює зі здоров'ям та медичними потребами.

– children (14.0) – кількість дітей впливає на вартість страхування, можливо, через збільшення медичних витрат у більших сім'ях.

– region_southeast (11.0), region_northwest (8.0), region_northeast (7.0), region_southwest (5.0) – різні регіони мають різні впливи, що може бути зумовлено регіональними відмінностями у вартості та доступності медичних послуг.

– smoker_yes (3.0), smoker_no (3.0) – статус куріння має порівняно нижчий вплив на модель, що може бути несподіванкою, оскільки куріння сильно корелює з високими медичними витратами.

– sex_female (3.0), sex_male (2.0) – стать має відносно низький вплив на прогнози моделі, що може вказувати на більшу важливість інших ознак.

Графік підкреслює значення індексу маси тіла, як ключового предиктора медичних витрат у моделі, а вік та кількість дітей також грають важливу роль. Регіональні відмінності та статус куріння здаються менш важливими, хоча це може залежати від специфіки даних і моделі.

Для остаточного відбору ознак побудуємо кореляційну матрицю даних використаних для тренування моделі (рис. 2.21).

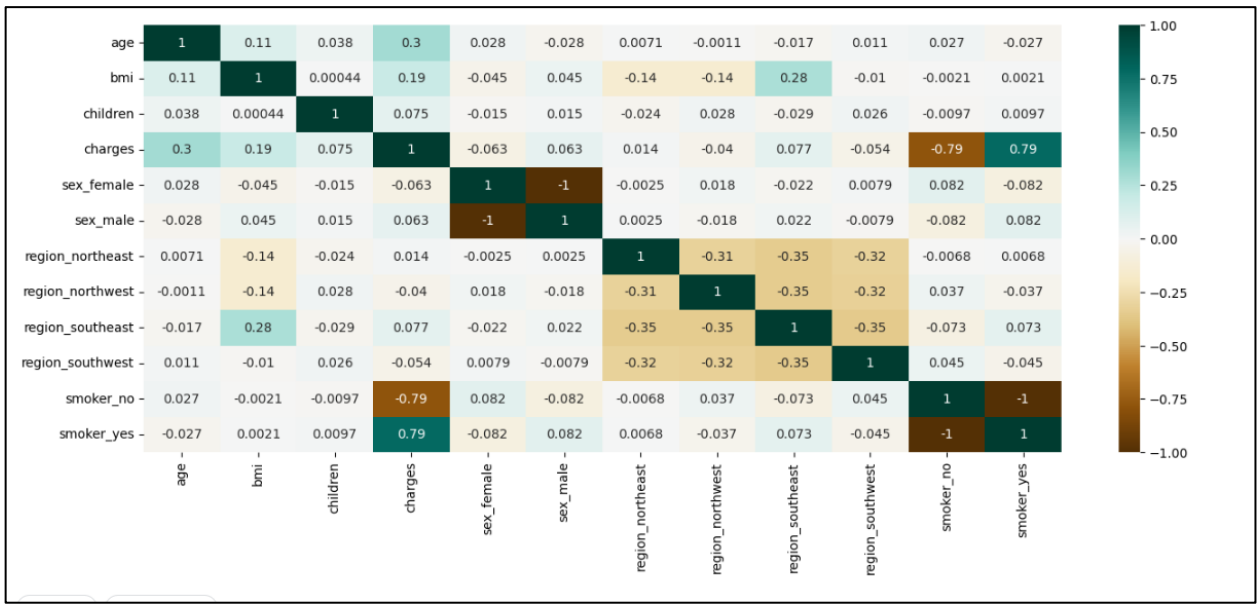


Рисунок 2.21 – Кореляційна матриця даних використаних для тренування моделі

Важливі кореляції:

– BMI та витрати (charges) – кореляція 0.19 вказує на позитивний зв'язок, хоча він і не дуже сильний, що означає, що зі збільшенням BMI зростають і медичні витрати.

– Вік та витрати (charges) – кореляція 0.3 також показує позитивний зв'язок, де зі збільшенням віку збільшуються і медичні витрати, хоча цей зв'язок не є дуже міцним.

– Статус курця (smoker_yes) та витрати (charges) – Значення кореляції 0.79 вказує на дуже сильний позитивний зв'язок, що є однією з найвищих кореляцій у матриці. Це підтверджує, що статус куріння має значний вплив на медичні витрати.

Отже, враховуючи графік важливості ознак та кореляційну матрицю, змінювати набір ознак не потрібно.

2.3 Тренування моделей машинного навчання для прогнозування ціни медичного страхування

Для початку потрібно обрати моделі для прогнозування вартості медичного страхування. Оскільки задача полягає саме в прогнозування, тому потрібно обирати з моделей регресорів.

Існує безліч моделей, які можна використовувати для прогнозування даних, і вибір конкретної моделі залежить від типу даних, їх характеристик, обсягу даних, змісту завдання прогнозування та багатьох інших факторів. Ось деякі з найпопулярніших моделей для прогнозування:

1. Лінійна регресія. Використовується для прогнозування неперервних числових значень, шукаючи лінійну залежність між вхідними змінними та вихідним значенням.

2. Дерева рішень. Створює дерево рішень на основі вхідних функцій, яке дозволяє виконувати прогнози на основі шляхів у дереві.

3. Випадковий ліс. Ансамбль дерев рішень, які працюють разом для покращення точності прогнозування.

4. Метод опорних векторів (Support Vector Machines, SVM). Використовується для класифікації та регресії, шукаючи гіперплощину, що найкращим чином розділяє дані.

5. Нейронні мережі. Моделі, які імітують нейронну структуру людського мозку та використовуються для вирішення широкого спектру задач, включаючи прогнозування.

6. Градієнтний бустінг. Ансамбль моделей, які будуються послідовно, кожна з яких намагається покращити прогноз попередньої моделі.

Використаємо: лінійна регресія, дерева рішень, випадковий ліс, градієнтний бустінг

Імпортуємо відповідні бібліотеки в програму (рис. 2.22).

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
```

Рисунок 2.22 – Імпорт бібліотек

Імпорт метрики для оцінки роботи моделей на рисунку 2.23.

```
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
```

Рисунок 2.23 – Метрики оцінки роботи моделей

Короткий опис основних метрик для задач регресії:

1. Mean Squared Error (MSE). MSE обчислює середньоквадратичну відстань між прогнозованими значеннями і фактичними значеннями цільової змінної. Чим менше значення MSE, тим краще модель.

2. Mean Absolute Error (MAE). MAE обчислює середню абсолютну різницю між прогнозованими і фактичними значеннями. Вона також вимірює точність прогнозів, проте не враховує квадратичні відмінності, як MSE.

3. R_2score. R_2score вимірює пропорцію дисперсії у залежній змінній, яка пояснюється моделлю. Значення може бути від 0 до 1, де 1 вказує на ідеальне узгодження між моделлю і даними, а 0 – на відсутність зв'язку.

4. Root Mean Squared Error (RMSE). RMSE обчислюється як квадратний корінь з MSE і є інтерпретованим, як стандартне відхилення прогнозів від фактичних значень.

Створимо обрані моделі прогнозування даних (рис. 2.24).

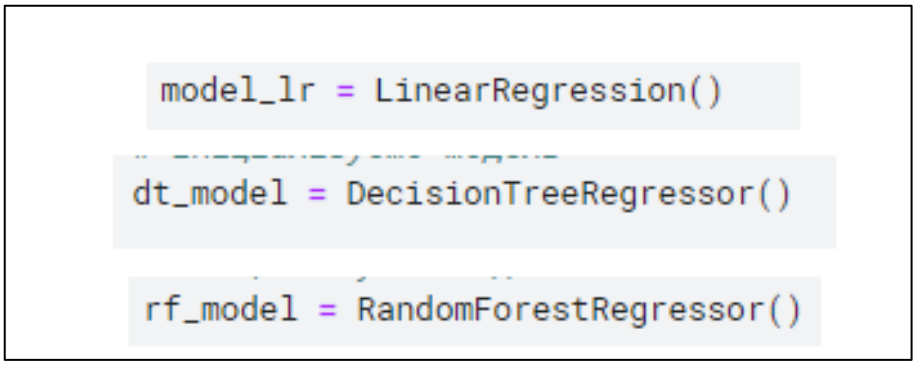


Рисунок 2.24 – Ініціалізація моделей машинного навчання

Сформовано чотири моделі машинного навчання для прогнозування ціни медичного страхування.

Наступним кроком натренуємо створені моделі (рис. 2.25).

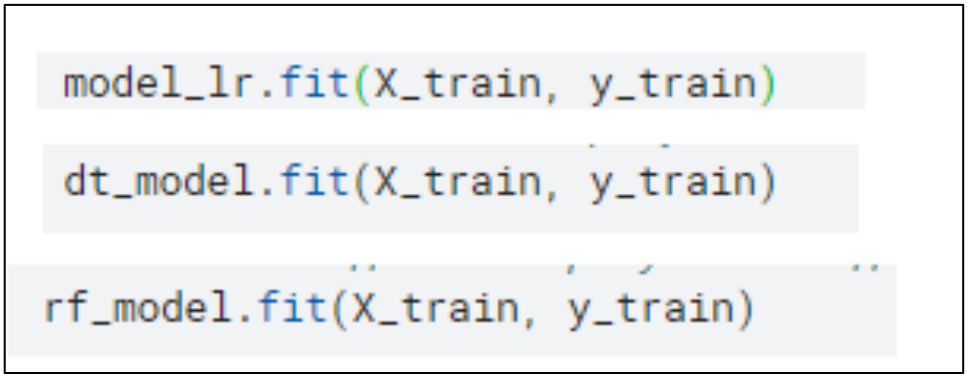


Рисунок 2.25 – Тренування моделей прогнозування

Натреновані моделі використаємо для прогнозування ціни медичного страхування та порівняємо їх точність. В 3-му розділі проведемо оптимізацію гіперпараметрів моделей машинного навчання для підвищення їх точності.

2.4 Висновки

В цьому розділі було проведено розвідувальний аналіз даних ціни медичного страхування. Побудовані графіки розподілу даних для числових ознак. Визначено та відфільтровано викиди та аномалії за допомогою Z-score.

Проаналізовано категоріальні дані. В наборі даних приблизний розподіл за статтю 50/50, 80% некурящих та рівномірний розподіл по регіонам.

В процесі інженерії ознак, категоріальні дані перетворені на числові вибрані оптимальні характеристики для прогнозування ціни медичного страхування. Обрано та створено неоптимізовані моделі машинного навчання для прогнозування ціни медичного страхування, а саме:

- лінійна регресія;
- дерево рішень;
- випадковий ліс.

3. ПРОГНОЗУВАННЯ ЦІНИ МЕДИЧНОГО СТРАХУВАННЯ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

3.1 Оптимізація параметрів моделей машинного навчання

Застосуємо модель лінійної регресії та оцінимо її точність. Оскільки в моделі лінійної регресії немає гіперпараметрів для налаштування, застосуємо її в стандартній формі (рис. 3.1).

```
# Ініціалізуємо модель
model_lr = LinearRegression()

# Навчаємо модель на тренувальних даних
model_lr.fit(X_train, y_train)


# Робимо прогнози на тренувальних даних
predictions_lr_train = model_lr.predict(X_train)
```

Рисунок 3.1 – Навчання моделі лінійної регресії на тренувальних даних

Оцінимо точність моделі лінійної регресії, код та результат зображено на рисунку 3.2 та 3.3.

```
# Оцінка моделі
mse_lr_train  = mean_squared_error(y_train, predictions_lr_train)
mae_lr_train  = mean_absolute_error(y_train, predictions_lr_train)
r2_lr_train  = r2_score(y_train, predictions_lr_train)
print('mse_lr_train =',mse_lr_train)
print('mae_lr_train =', mae_lr_train)
print('r2_lr_train =', r2_lr_train)
```

Рисунок 3.2 – Код оцінки точності роботи моделі лінійної регресії на тренувальних даних

```
Mean Squared Error Linear Regression Model =  34770004.94641928
Mean Absolute Error Linear Regression Model =  4059.9254965190858
R_2score Linear Regression Model =  0.74
```

Рисунок 3.3 – Оцінка точності моделі лінійної регресії на тренувальних даних

Точність моделі по метриці R2_score дорівнює 0,74, це значить, що модель прогнозує 74% даних і являється нормально натренованою.

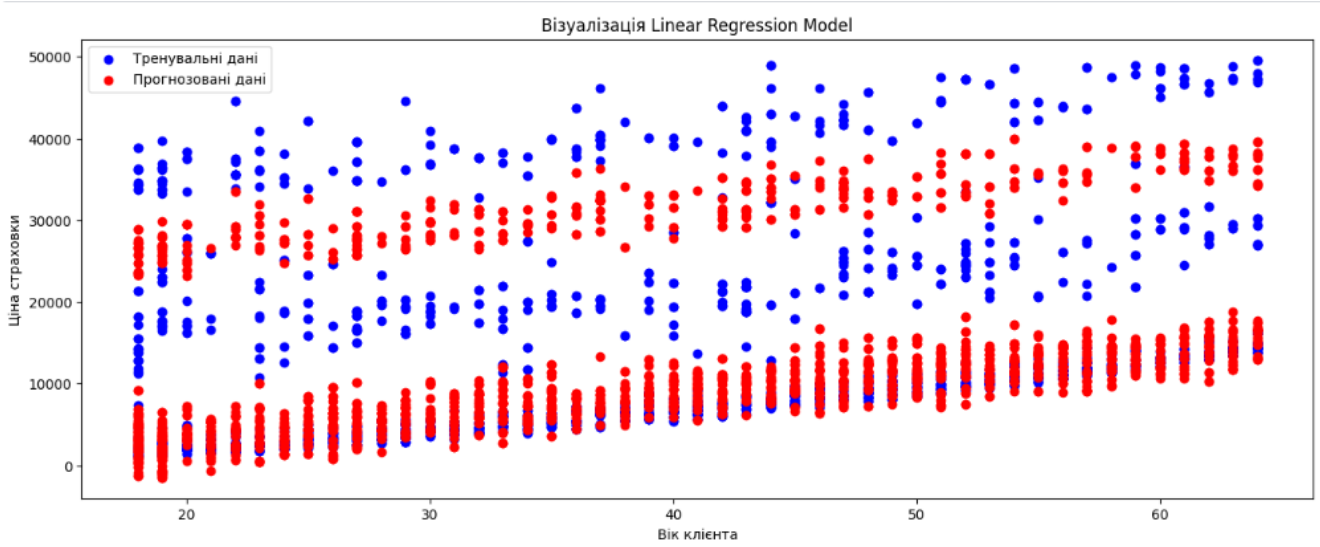
Побудуємо графік візуалізації роботи моделі лінійної регресії на тренувальних даних (рис. 3.4, 3.5).

```
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_lr_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Linear Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.4 – Код графіка візуалізації роботи моделі лінійної регресії на тренувальних даних



3.5 – Графік візуалізації роботи моделі лінійної регресії на тренувальних даних

Наступна модель — дерево рішень. Для автоматизованої оптимізації параметрів моделі можна використати GridSearchCV. Це потужний інструмент від scikit-learn, який систематично перевіряє ряд комбінацій параметрів, визначених користувачем, за допомогою крос-валідації для визначення тих, які дають найкращі результати згідно обраного критерію якості. На рисунку 3.6 зображено параметри моделі дерева рішень.

```
dt_model = DecisionTreeRegressor()
params = {
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 10, 20],
    'min_samples_leaf': [1, 5, 10],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_leaf_nodes': [ 5, 10, 20]
}
grid_search_dt = GridSearchCV(dt_model, params, cv=5, scoring='r2')
grid_search_dt.fit(X_train, y_train)

# Кращі параметри
print('Best parameters:', grid_search_dt.best_params_)
```

```
Best parameters: {'max_depth': 10, 'max_features': 'auto', 'max_leaf_nodes': 20, 'min_samples_leaf': 1, 'min_samples_split': 2}
```

Рисунок 3.6 – Налаштування та вибір найкращих параметрів моделі дерево рішень

Проведемо оцінку точності моделі за трьома метриками (рис. 3.7).

```
# Робимо прогнози на тренувальних даних
predictions_dt_train = grid_search_dt.predict(X_train)
# Оцінка моделі
mse_dt_train  = mean_squared_error(y_train, predictions_dt_train)
mae_dt_train  = mean_absolute_error(y_train, predictions_dt_train)
r2_dt_train  = r2_score(y_train, predictions_dt_train)
print('Mean Squared Error Decision Tree Regressor Model = ',mse_dt_train)
print('Mean Absolute Error Decision Tree Regressor Model = ', mae_dt_train)
print('R_2score Decision Tree egressor Model = ', round(r2_dt_train,2))
```

```
Mean Squared Error Decision Tree Regressor Model =  17025752.813225504
Mean Absolute Error Decision Tree Regressor Model =  2292.6481237907524
R_2score Decision Tree egressor Model =  0.87
```

Рисунок 3.7 – Оцінка точності натренованої моделі

Модель має точність 0,87 — це означає, що модель спрогнозувала 87% даних. Це є хорошою ознакою адекватності моделі. Потрібно перевірити її на тестових даних.

На рисунку 3.8 та 3.9 зобразимо код та графік реальні та прогнозовані дані.

```
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_dt_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Decision Tree Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.8 – Код графіка візуалізації роботи моделі дерева рішень на тренувальних даних

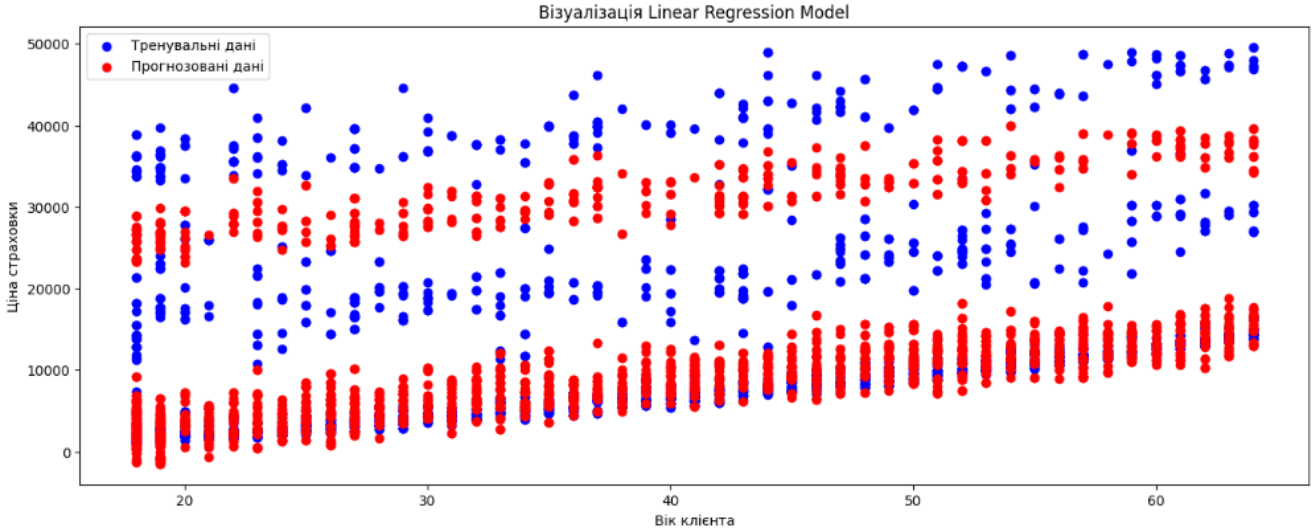


Рисунок 3.9 Графік візуалізації роботи моделі дерева рішень на тренувальних даних

Використаємо оптимізацію параметрів для моделі випадковий ліс. Для оптимізації застосуємо GridSearchCV. На рисунку 3.9 зображено код оптимізації даної моделі машинного навчання.

```
# Ініціалізуємо модель
rf_model = RandomForestRegressor()

param_grid = {
    'n_estimators': [100, 200],
    'max_features': [1.0, 'sqrt', 'log2'],
    'max_depth': [4, 6],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}
grid_search_rf = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=

grid_search_rf.fit(X_train, y_train)

print("Best parameters:", grid_search_rf.best_params_)
print("Best score (R2):", grid_search_rf.best_score_)
```

```
Fitting 5 folds for each of 48 candidates, totalling 240 fits
Best parameters: {'max_depth': 6, 'max_features': 1.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best score (R2): 0.8754144269631283
```

Рисунок 3.9 – Налаштування та вибір найкращих параметрів моделі випадковий ліс

Оцінимо точність моделі випадковий ліс трьома метриками, що зображено на рисунку 3.10.

```
# Робимо прогнози на тренувальних даних
predictions_rf_train = grid_search_rf.predict(X_train)
# Оцінка моделі
mse_rf_train  = mean_squared_error(y_train, predictions_rf_train)
mae_rf_train  = mean_absolute_error(y_train, predictions_rf_train)
r2_rf_train  = r2_score(y_train, predictions_rf_train)
print('Mean Squared Error Random Forest Regressor Model = ',mse_rf_train)
print('Mean Absolute Error Random Forest Regressor Model = ', mae_rf_train)
print('R_2score Decision Random Forest Model = ', round(r2_rf_train,2))
```

```
Mean Squared Error Random Forest Regressor Model =  13415661.37767695
Mean Absolute Error Random Forest Regressor Model =  1913.4856115806222
R_2score Decision Random Forest Model =  0.9
```

Рисунок 3.10 – Оцінка точності моделі випадковий ліс на тренувальних даних

На рисунку 3.11 та 3.12 зображено код та графік візуалізацію тренувальних та прогнозованих даних.

```
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_rf_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Random Forest Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.11 – Код графіка візуалізації роботи моделі випадковий ліс на тренувальних даних

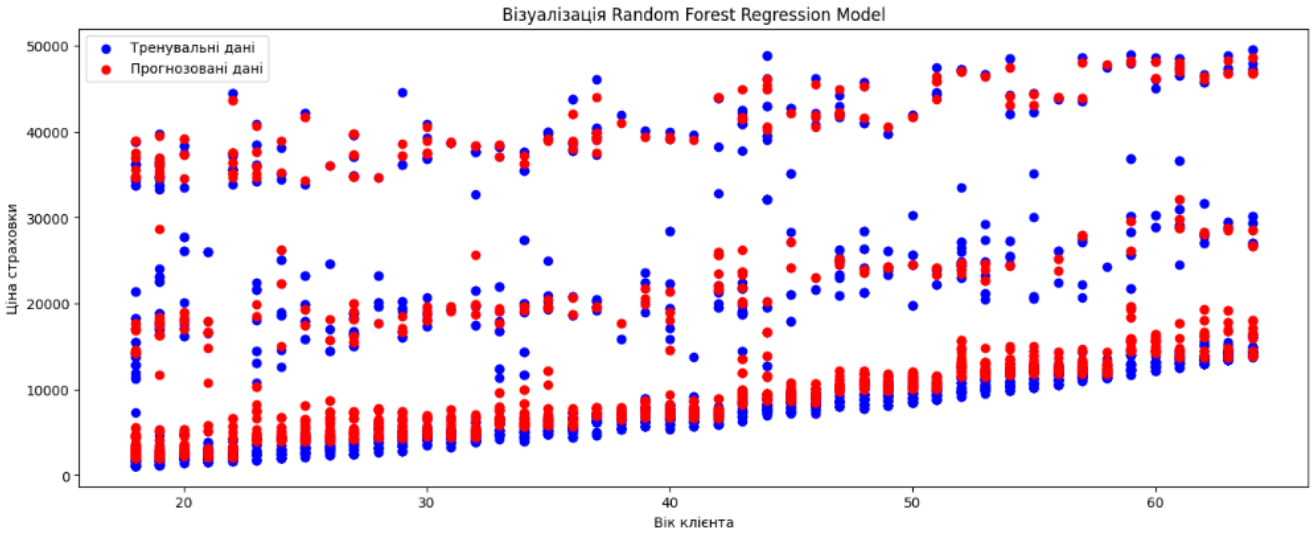


Рисунок 3.12 Графік візуалізації роботи моделі випадкового лісу на тренувальних даних

3.2 Застосування моделей на тестових даних

Для перевірки адекватності моделей перевіримо їх роботу на тестових даних та оцінимо їх точність.

На рисунку 3.13 зображено застосування моделі лінійної регресії на тестових даних.

```
# Робимо прогнози на тестових даних
predictions_lr_test = model_lr.predict(X_test)
# Оцінка моделі
mse_lr_Test = mean_squared_error(y_test, predictions_lr_test)
mae_lr_Test = mean_absolute_error(y_test, predictions_lr_test)
r2_lr_Test = r2_score(y_test, predictions_lr_test)

print('Mean Squared Error Linear Regression Model = ', mse_lr_Test)
print('Mean Absolute Error Linear Regression Model = ', mae_lr_Test)
print('R_2score Linear Regression Model = ', round(r2_lr_Test,2))
```

```
Mean Squared Error Linear Regression Model =  32525678.660072923
Mean Absolute Error Linear Regression Model =  4081.5749949052592
R_2score Linear Regression Model =  0.78
```

Рисунок 3.13 – Результат оцінки моделі лінійної регресії на тестових даних

На рисунку 3.14 та 3.15 зобразимо код та графік роботи даної моделі на тестових даних.

```
# Тестові дані
plt.figure(figsize=(16, 6))
plt.scatter(X_test['age'], y_test, color='blue', label='Тренувальні дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], predictions_lr_test, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Linear Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.14 – Код графіка візуалізації роботи моделі лінійна регресія на тестових даних

На рисунку 3.15 відображено 78% прогнозованих даних, що співпадають з тестовим набором даних

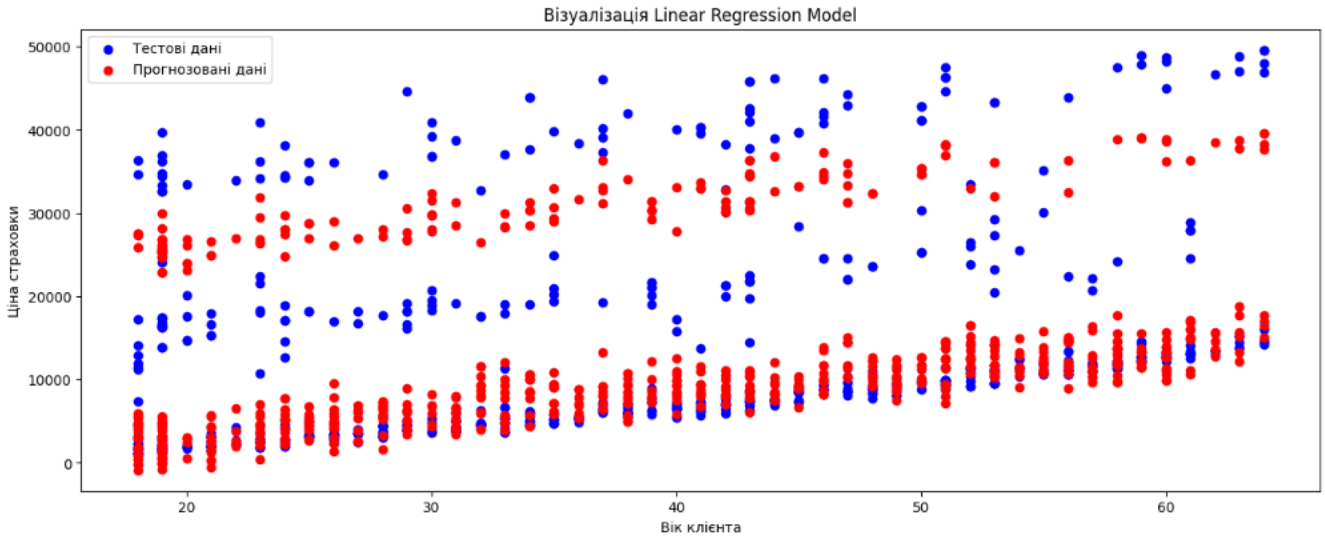


Рисунок 3.15 Графік візуалізації роботи моделі лінійної регресії на тестових даних

На рисунку 3.16 зображено застосування моделі дерева рішень на тестових даних.

```
dt_predictions = grid_search_dt.predict(X_test)

# Оцінка моделі

mse_dt = mean_squared_error(y_test, dt_predictions)
mae_dt = mean_absolute_error(y_test, dt_predictions)
r2_dt = r2_score(y_test, dt_predictions)
print('Mean Squared Error Decision Tree Regressor Model = ',mse_dt)
print('Mean Absolute Error Decision Tree Regressor Model = ', mae_dt)
print('R_2score Decision Tree egressor Model = ', round(r2_dt,2))
```

```
Mean Squared Error Decision Tree Regressor Model =  17242345.6399761
Mean Absolute Error Decision Tree Regressor Model =  2457.7306381526196
R_2score Decision Tree egressor Model =  0.88
```

Рисунок 3.16 – Застосування моделі дерева рішень на тестових даних та оцінка точності моделі

На рисунку 3.17 наведений код реалізації графіку порівняння тестових та прогнозованих даних моделлю дерева рішень.

```
# Тестові дані
plt.figure(figsize=(16, 6))

plt.scatter(X_test['age'], y_test, color='blue', label='Тренувальні дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], dt_predictions , color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Decision Tree Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.17 – Код візуалізації роботи моделі дерева рішень на тестових даних

На рисунку 3.18 зображено графік порівняння тестових та прогнозованих даних моделлю дерева рішень. На графіку видно покриття у 88% прогнозованими даними.

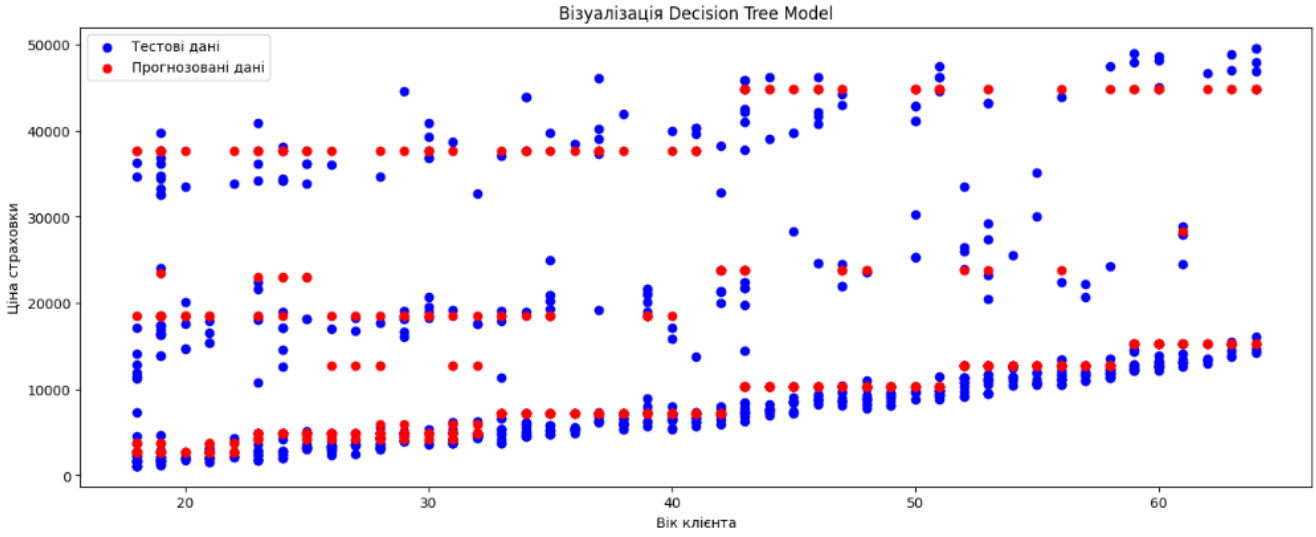


Рисунок 3.18 – Графік порівняння тестових та прогнозованих даних моделлю дерева рішень.

Застосуємо модель випадковий ліс на тестових даних для перевірки даної моделі на працездатність. На рисунку 3.19 зображено код прогнозування моделі випадковий ліс на тестових даних.

```
# Робимо прогнози на тестових даних
rf_predictions = grid_search_rf.predict(X_test)
# Оцінка моделі

# Оцінка моделі
mse_rf  = mean_squared_error(y_test, rf_predictions)
mae_rf  = mean_absolute_error(y_test, rf_predictions)
r2_rf  = r2_score(y_test, rf_predictions)
print('Mean Squared Error Random Forest Regressor Model = ',mse_rf)
print('Mean Absolute Error Random Forest Regressor Model = ', mae_rf)
print('R_2score  Random Forest Model = ', round(r2_rf,2))
```

```
Mean Squared Error Random Forest Regressor Model =  13560900.053032486
Mean Absolute Error Random Forest Regressor Model =  2020.544528785016
R_2score Decision Random Forest Model =  0.91
```

Рисунок 3.19 – Код роботи моделі на тестових даних

На рисунку 3.19 також видно, що точність моделі рівна 0.91. Це означає, що модель зробила прогноз 91% даних тестового набору даних.

На рисунку 3.20 зображено код візуалізації тестових та прогнозованих даних.

```
# Тестові дані
plt.figure(figsize=(16, 6))
plt.scatter(X_test['age'], y_test, color='blue', label='Тестові дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], rf_predictions , color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Random Forest Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
```

Рисунок 3.20 – Код візуалізації тестових та прогнозованих даних моделлю випадковий ліс

На рисунку 3.21 зображено графік тестових та прогнозованих даних моделлю випадковий ліс

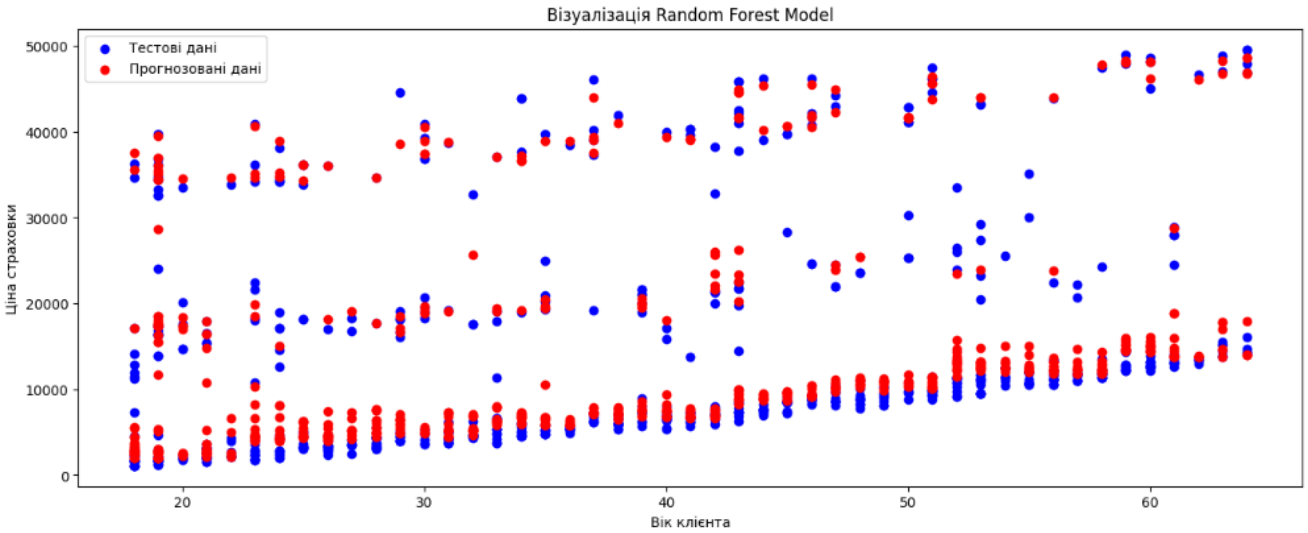


Рисунок 3.21 – Графік тестових та прогнозованих даних моделлю випадковий ліс

На графіку показано візуалізацію результатів моделі випадкового лісу для прогнозування медичних витрат залежно від віку клієнта. Сині точки представляють фактичні дані тестового набору, а червоні точки — прогнозовані дані моделлю. Вісь X представляє вік клієнта, а вісь Y — медичні витрати.

Аналіз графіку:

Розподіл даних. Обидва набори точок розподілені вздовж віку клієнта, зі збільшенням концентрації і розкиду витрат у старших вікових групах.

Зіставлення фактичних та прогнозованих значень. Прогнози моделі (червоні точки) загалом відтворюють тенденції фактичних даних (сині точки). Це вказує на здатність моделі влучно відтворювати закономірності в даних.

Варіації витрат з віком. З графіку видно, що медичні витрати загалом зростають з віком. Випадковий ліс досить добре передбачає ці зміни, хоча і з деякими відхиленнями, особливо в старших вікових групах.

Потенційні викиди. Є деякі аномальні значення у фактичних даних, які модель не передбачає з високою точністю, особливо у нижньому діапазоні витрат серед старших клієнтів.

Для повноцінного порівняння та вибору найточнішої моделі прогнозування ціни медичного страхування сформуємо таблицю з оцінками моделей на тренувальних та тестових наборах даних.

На рисунку 3.22 зображено код створення даної таблиці.

```
# Створюємо DataFrame з результатами оцінки моделей
results = pd.DataFrame({
    'Model': ['Linear Regression', 'Decision Tree', 'Random Forest'],
    'MSE_TR': [mse_lr_train, mse_dt_train, mse_rf_train],
    'MSE_TS': [mse_lr_Test, mse_dt, mse_rf],
    'MAE_TR': [mae_lr_train, mae_dt_train, mae_rf_train],
    'MAE_TS': [mae_lr_Test, mae_dt, mae_rf],
    'R-squared_TR': [round(r2_lr_train,2), round(r2_dt_train,2), round(r2_rf_train,2)],
    'R-squared_TS': [round(r2_lr_Test,2), round(r2_dt,2), round(r2_rf,2)]
})

# Виводимо результати
results
```

Рисунок 2.22 – Код таблиці результатів роботи моделей

На рисунку 2.23 наведена відповідна порівняльна таблиця точності моделей машинного навчання прогнозування ціни медичного страхування.

Model	MSE_TR	MSE_TS	MAE_TR	MAE_TS	R-squared_TR	R-squared_TS
Linear Regression	3.477000e+07	3.252568e+07	4059.925497	4081.574995	0.74	0.78
Decision Tree	1.702575e+07	1.724235e+07	2292.648124	2457.730638	0.87	0.88
Random Forest	1.341566e+07	1.356090e+07	1913.485612	2020.544529	0.90	0.91

Рисунок 2.23 – Таблиця точності моделей

Проаналізуємо результати оцінки точності моделей:

– Лінійна регресія показує помірні значення помилок (MSE та MAE) та коефіцієнт детермінації (R2_score), що є досить стабільними між тренувальними та тестовими наборами. Це свідчить про адекватність моделі без явного перенавчання.

– Дерево рішень має нижчі помилки на тренувальних даних порівняно з лінійною регресією, але показники на тестових даних також є досить хорошими, що свідчить про вищу здатність моделі до узагальнення порівняно з лінійною регресією.

– Випадковий ліс демонструє найкращі результати як по помилках (MSE та MAE), так і по коефіцієнту детермінації (R2_score) на обох наборах

даних. Це вказує на те, що модель випадкового лісу є найбільш ефективною серед трьох розглянутих моделей, забезпечуючи високу точність та відмінну здатність до узагальнення.

3.3. Висновки

В даному розділі застосовано три моделі машинного навчання для прогнозування вартості медичного страхування. Проведено оптимізацію параметрів навчання моделей машинного навчання за допомогою GridSearchCV.

Моделі натреновано на навчальних даних та оцінено їхню точність за допомогою метрик MSE, MAE та R2_score.

Натреновані моделі застосовано та тестовому наборі даних для перевірки адекватності моделей. Проведено оцінювання моделей та визначено оптимальну модель.

Отже, випадковий ліс є оптимальною моделлю в даному порівнянні, що робить її відмінним вибором для задач прогнозування ціни медичного страхування, де потрібна висока точність прогнозування та стійкість до перенавчання. Про оптимальність моделі свідчить результат оцінювання точності моделі. А саме за метрикою R2_score модель має точність 0.90 та 0.91 на тренувальному та тестовому наборах даних відповідно.

ВИСНОВКИ

В процесі виконання бакалаврської дипломної роботи на тему «Технологія аналізу та прогнозування вартості медичного страхування», проведено аналіз предметної області прогнозування вартості медичного страхування, а саме процес ціноутворення медичного страхування. Які фактори особливо впливають на розмір медичної страховки та які моделі медичного страхування використовують у світі та Україні.

Описано вплив інформаційних технологій на сферу охорони здоров'я. Проаналізувано методи прогнозування ціни на медичне страхування.

У другому розділі було проведено аналіз даних ціни медичного страхування. Побудовано графіки розподілу даних для числових ознак. Визначено та відфільтровано викиди та аномалії за допомогою Z-score.

Проаналізовано категоріальні дані. В наборі даних приблизний розподіл за статтю 50/50, 80% некурящих та рівномірний розподіл по регіонам.

В процесі інженерії ознак, категоріальні дані перетворено на числові вибрано оптимальні характеристики для прогнозування ціни медичного страхування. Обрано та створено неоптимізовані моделі машинного навчання для прогнозування ціни медичного страхування, а саме:

- лінійна регресія;
- дерево рішень;
- випадковий ліс.

В третьому розділі застосовано три моделі машинного навчання для прогнозування вартості медичного страхування. Проведено оптимізацію параметрів навчання моделей машинного навчання за допомогою GridSearchCV.

Моделі натреновано на навчальних даних та оцінено їхню точність за допомогою метрик MSE, MAE та R2_score.

Натреновані моделі застосовано та тестовому наборі даних для перевірки адекватності моделей. Проведено оцінювання моделей та визначено оптимальну модель випадковий ліс.

Випадковий ліс є оптимальною моделлю в даному порівнянні, що робить її відмінним вибором для задач прогнозування ціни медичного страхування, де потрібна висока точність прогнозування та стійкість до перенавчання. Про оптимальність моделі свідчить результат оцінювання точності моделі. А саме за метрикою R2_score модель має точність 0.90 та 0.91 на тренувальному та тестовому наборах даних відповідно.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Thapa, J. R., Xie, Y., Avery, M., & Veenstra, D. L. (2020). The impact of public health insurance on health care utilisation, financial protection and health status in low- and middle-income countries: A systematic review. PLOS ONE, 15(8), e0237770. https://doi.org/10.1371/journal.pone.0237770

2. Chen, Z., Fan, V. Y., Beliveau, Y., & Zhang, L. (2019). Medical insurance payment schemes and patient medical expenses: A cross-sectional study of lung cancer patients in urban China. BMC Health Services Research, 19(1), 1007. https://doi.org/10.1186/s12913-019-4781-6

3. Pham, L. T., Nguyen, H. T., Vu, D. M., Nguyen, T. X., Nguyen, H. T., & Pham, T. Q. (2021). How public health insurance expansion affects healthcare utilizations in middle and low-income households: An observational study from national cross-section surveys in Vietnam. BMC Public Health, 21(1), 736. https://doi.org/10.1186/s12889-021-10762-2

4. Liu, T., Gao, Y., Li, H., Zhang, L., & Sun, J. (2022). Analysis of the Operational Efficiency of Basic Medical Insurance for Urban and Rural Residents: Based on a Three-Stage DEA Model. International Journal of Environmental Research and Public Health, 19(21), 13831. https://doi.org/10.3390/ijerph192113831

5. Lekhan, V., Rudiy, V., & Richardson, E. (2015). Ukrainian health care system and its chances for successful transition from Soviet legacies. Globalization and Health, 11, 29. https://doi.org/10.1186/s12992-015-0117-x

6. World Health Organization. (2022). Accessing health care in Ukraine after 8 months of war: The health system remains resilient, but key health services and medicine are increasingly unaffordable. Retrieved from https://www.who.int

7. European Observatory on Health Systems and Policies. (2021). Health systems in action: Ukraine. Retrieved from https://eurohealthobservatory.who.int/publications/i/health-systems-in-action-ukraine

8. Wikipedia. (n.d.). Healthcare in Ukraine. Retrieved March 31, 2024, from https://en.wikipedia.org/wiki/Healthcare_in_Ukraine

9. Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Impact of digital health technology on health insurance claims rejection rate in Ghana: a quasi-experimental study. BMC Digital Health. https://doi.org/10.1186/s12911-022-00198-w

10. Van Der Heijden, F., Koppelaar, H., & Bijl, H. (2020). The influence of electronic health record use on collaboration among medical specialties. BMC Health Services Research, 20, 273. https://doi.org/10.1186/s12913-020-05176-0

11. Kruse, C. S., Beane, A. (2018). Health Information Technology Continues to Show Positive Effect on Medical Outcomes: Systematic Review. Journal of Medical Internet Research, 20(2), e41. https://doi.org/10.2196/jmir.8793

12. Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. International Journal of Environmental Research and Public Health, 19(13), 7898. https://doi.org/10.3390/ijerph19137898

13. Orji, U., & Ukwandu, E. (2023). Machine Learning For An Explainable Cost Prediction of Medical Insurance. Retrieved from https://arxiv.org/abs/2311.14139

14. Herland, M., Khoshgoftaar, T. M., & Wald, R. (2018). A review of data mining using big data in health informatics. Journal of Big Data, 1(1), 2. https://doi.org/10.1186/s40537-014-0007-7

15. Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. «International Journal of Environmental Research and Public Health», 19(13). https://doi.org/10.3390/ijerph19137898

16. Bharti, A., & Malik, L. (2022). Regression Analysis and Prediction of Medical Insurance Cost. «International Journal of Creative Research Thoughts (IJCRT)», 10(2). Retrieved from https://ijcrt.org/papers/IJCRT2203462.pdf

17. Reinhardt, U. E. (2006). The Pricing of US Hospital Services: Chaos Behind a Veil of Secrecy. «Health Affairs», 25(1), 57-69. https://doi.org/10.1377/hlthaff.25.1.57

18. Wharam, J. F., Ross-Degnan, D., & Rosenthal, M. B. (2013). The ACA and High-Deductible Insurance—Strategies for Sharpening a Blunt Instrument. «New

England Journal of Medicine», 369, 1481-1484. https://www.nejm.org/doi/full/10.1056/NEJMp1213488

19. Kaplan, R. S., & Anderson, S. R. (2004). Time-Driven Activity-Based Costing. «Harvard Business Review». Retrieved from https://hbr.org/2004/11/time-driven-activity-based-costing

20. Programme costs in the economic evaluation of health interventions (2021). «Cost Effectiveness and Resource Allocation». Retrieved from https://resource-allocation.biomedcentral.com/articles/10.1186/s12962-021-00298-w

Додаток А

(обов’язковий)

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

ЗАТВЕРДЖУЮ
Завідувач кафедри САІТ
_____ д.т.н., проф. Віталій МОКІН
“____”_____ 2024 р.

ТЕХНІЧНЕ ЗАВДАННЯ
на бакалаврську дипломну роботу
ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ВАРТОСТІ МЕДИЧНОГО СТРАХУВАННЯ
08-34.БДР.003.00.000 ТЗ

Керівник: к.т.н., доц.
_____ Ілона ВАРЧУК
«__» _____ 2024 р.
Розробила студентка гр. 2ІСТ-20б
_____ Катерина ІВАНОВА
«__» _____ 2024 р.

Вінниця 2024

1. Підстава для проведення робіт.

Підставою для виконання роботи є наказ №__ по ВНТУ від «__» _____2024р., та індивідуальне завдання на БДР, затверджене протоколом №__ засідання кафедри САІТ від «__» _____ 2024р.

2. Джерела розробки:

1) Wikipedia. (n.d.). Healthcare in Ukraine. Retrieved March 31, 2024, from https://en.wikipedia.org/wiki/Healthcare_in_Ukraine

2) Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. «International Journal of Environmental Research and Public Health», 19(13). https://doi.org/10.3390/ijerph19137898

3) Kaplan, R. S., & Anderson, S. R. (2004). Time-Driven Activity-Based Costing. «Harvard Business Review». Retrieved from https://hbr.org/2004/11/time-driven-activity-based-costing

3. Мета і призначення роботи.

Метою дослідження є аналіз та прогнозування ціни медичного страхування.

4. Вихідні дані для проведення робіт:

5. Kaggle Dataset «medical insurance cost prediction» https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction.

6. Методи дослідження:

Розвідувальний аналіз даних, моделі машинного навчання, мова програмування Python.

7. Етапи роботи і терміни їх виконання:

a) Аналіз предметної області _____ – _____
b) Порівняльний аналіз проблематики _____ – _____
c) Вибір оптимальних інформаційних технологій _____ – _____
d) Прогнозування вартості медичного страхування _____ – _____
e) Оформлення матеріалів до захисту БДР _____ – _____

7. Очікувані результати та порядок реалізації

Отримання оптимальної моделі машинного навчання для прогнозування ціни медичного страхування.

8. Вимоги до розробленої документації

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

9. Порядок приймання роботи

Публічний захист «___» _____ 2024 р.
Початок розробки «___» _____ 2024 р.
Граничні терміни виконання БДР «___» _____ 2024 р.

Розробив студент гр. 2ІСТ-20б _____ Катерина ІВАНОВА

Додаток Б
(обов'язковий)
Протокол перевірки бакалаврської дипломної роботи на наявність текстових запозичень

Назва роботи: «Технологія аналізу та прогнозування вартості медичного страхування»
Тип роботи: бакалаврська дипломна робота
Підрозділ: кафедра САІТ

Показники звіту подібності Unicheck

Оригінальність 97,1% Схожість 2,9%

Аналіз звіту подібності (відмітити потрібне)

● Запозичення, виявлені у роботі, оформлені коректно і не містять ознак плагіату.
○ Виявлені у роботі запозичення не мають ознак плагіату, але їх надмірна кількість викликає сумніви щодо цінності роботи і самостійності її автора. Роботу направити на розгляд експертної комісії кафедри.
○ Виявлені у роботі запозичення є недобросовісними і мають ознаки плагіату та/або в ній містяться навмисні спотворення тексту, що вказують на спроби приховування недобросовісних запозичень.

Особа, відповідальна за перевірку _____ Сергій ЖУКОВ
(підпис)

Ознайомлені з повним звітом подібності, який був згенерований системою Unicheck щодо роботи.

Автор роботи _____ Катерина ІВАНОВА
(підпис)

Керівник роботи _____ Ілона ВАРЧУК
(підпис)

Додаток В
(довідниковий)
Фрагмент лістингу програми

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import LabelEncoder
from scipy import stats
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the
input directory
import warnings

# Ігнорування всіх попереджень
warnings.filterwarnings('ignore', category=FutureWarning)

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

data = pd.read_csv('/kaggle/input/medical-insurance-cost-prediction/medical_insurance.csv')
data.head(5)
data.info()
data.describe()
plt.figure(figsize=(8,6))
plt.boxplot(data['charges'])
plt.title('Boxplot для стовпця charges')
plt.ylabel('Значення')
plt.show()
data=data[np.abs(stats.zscore(data['charges'])) < 3]
data.describe()
data.describe()
plt.figure(figsize=(30,28))
for i, col in enumerate( ['age','bmi','children','charges']):
    plt.subplot(3, 3, i+1)
    sns.histplot(data = data,
            x = col,
            kde = True,
            bins = 30,
            color = 'blue')

plt.show()
plt.figure(figsize=(12,9))
for i,col in enumerate(['sex','smoker','region']):
    plt.subplot(3,2,i+1)
    x=data[col].value_counts().reset_index()
    plt.title(col)
    plt.pie(x=x['count'],labels=x[col],autopct="%0.1f%%",colors=sns.color_palette('muted'))
```

```
data_corr= data.copy()
label_encoder = LabelEncoder()
for col in ['sex','smoker','region']:
   data_corr[col] = label_encoder.fit_transform(data_corr[col])
data_corr.head()
plt.figure(figsize=(16, 6))
heatmap = sns.heatmap(data_corr.corr(), vmin=-1, vmax=1, annot=True, cmap='BrBG')
from sklearn.model_selection import train_test_split
binar = pd.get_dummies(data[['sex','region','smoker']],dtype=int)
data_model = pd.concat([data[['age','bmi','children','charges']],binar],axis=1)
data_model
X=data_model.drop(columns=['charges'])
y=data_model["charges"]
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.25,random_state=7)
import xgboost as xgb
from sklearn.metrics import mean_squared_error
# Ініціалізуйте модель XGBRegressor
model = xgb.XGBRegressor(objective ='reg:squarederror',
                         colsample_bytree = 0.3,
                         learning_rate = 0.1,
                         max_depth = 5,
                         alpha = 10,
                         n_estimators = 10)

# Тренуйте модель
model.fit(X_train, y_train)

# Робіть прогнози
y_pred = model.predict(X_test)

# Вимірюйте точність моделі
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print("RMSE: %f" % (rmse))
xgb.plot_importance(model)
plt.title("Feature Importance")
plt.show()
plt.figure(figsize=(16, 6))
heatmap = sns.heatmap(data_model.corr(), vmin=-1, vmax=1, annot=True, cmap='BrBG')
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
# Ініціалізуємо модель
model_lr = LinearRegression()

# Навчаємо модель на тренувальних даних
model_lr.fit(X_train, y_train)

# Робимо прогнози на тренувальних даних
predictions_lr_train = model_lr.predict(X_train)
# Оцінка моделі
```

```
mse_lr_train  = mean_squared_error(y_train, predictions_lr_train)
mae_lr_train  = mean_absolute_error(y_train, predictions_lr_train)
r2_lr_train  = r2_score(y_train, predictions_lr_train)
print('Mean Squared Error Linear Regression Model = ',mse_lr_train)
print('Mean Absolute Error Linear Regression Model = ', mae_lr_train)
print('R_2score Linear Regression Model = ', round(r2_lr_train,2))
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_lr_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Linear Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
# Робимо прогнози на тестових даних
predictions_lr_test = model_lr.predict(X_test)
# Оцінка моделі
mse_lr_Test = mean_squared_error(y_test, predictions_lr_test)
mae_lr_Test = mean_absolute_error(y_test, predictions_lr_test)
r2_lr_Test = r2_score(y_test, predictions_lr_test)

print('Mean Squared Error Linear Regression Model = ', mse_lr_Test)
print('Mean Absolute Error Linear Regression Model = ', mae_lr_Test)
print('R_2score Linear Regression Model = ', round(r2_lr_Test,2))

# Тестові дані
plt.figure(figsize=(16, 6))
plt.scatter(X_test['age'], y_test, color='blue', label='Тестові дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], predictions_lr_test, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Linear Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
from sklearn.model_selection import GridSearchCV
# Ініціалізуємо модель
dt_model = DecisionTreeRegressor()

# Навчаємо модель на тренувальних даних
dt_model.fit(X_train, y_train)

# Робимо прогнози на тестових даних
dt_predictions = dt_model.predict(X_test)
# Оцінка моделі

mse_dt = mean_squared_error(y_test, dt_predictions)
mae_dt = mean_absolute_error(y_test, dt_predictions)
```

```
r2_dt = r2_score(y_test, dt_predictions)
dt_model = DecisionTreeRegressor()
params = {
   'max_depth': [10, 20, 30],
   'min_samples_split': [2, 10, 20],
   'min_samples_leaf': [1, 5, 10],
   'max_features': ['auto', 'sqrt', 'log2'],
   'max_leaf_nodes': [ 5, 10, 20]
}
grid_search_dt = GridSearchCV(dt_model, params, cv=5, scoring='r2')
grid_search_dt.fit(X_train, y_train)

# Кращі параметри
print('Best parameters:', grid_search_dt.best_params_)
# Робимо прогнози на тренувальних даних
predictions_dt_train = grid_search_dt.predict(X_train)
# Оцінка моделі
mse_dt_train  = mean_squared_error(y_train, predictions_dt_train)
mae_dt_train  = mean_absolute_error(y_train, predictions_dt_train)
r2_dt_train  = r2_score(y_train, predictions_dt_train)
print('Mean Squared Error Decision Tree Regressor Model = ',mse_dt_train)
print('Mean Absolute Error Decision Tree Regressor Model = ', mae_dt_train)
print('R_2score Decision Tree egressor Model = ', round(r2_dt_train,2))
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_dt_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Decision Tree Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
dt_predictions = grid_search_dt.predict(X_test)

# Оцінка моделі

mse_dt = mean_squared_error(y_test, dt_predictions)
mae_dt = mean_absolute_error(y_test, dt_predictions)
r2_dt = r2_score(y_test, dt_predictions)
print('Mean Squared Error Decision Tree Regressor Model = ',mse_dt)
print('Mean Absolute Error Decision Tree Regressor Model = ', mae_dt)
print('R_2score Decision Tree egressor Model = ', round(r2_dt,2))
# Тестові дані
plt.figure(figsize=(16, 6))

plt.scatter(X_test['age'], y_test, color='blue', label='Тестові дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], dt_predictions , color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Decision Tree Model')
```

```
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
# Ініціалізуємо модель
rf_model = RandomForestRegressor()

param_grid = {
   'n_estimators': [100, 200],
   'max_features': [1.0, 'sqrt', 'log2'],
   'max_depth': [4, 6],
   'min_samples_split': [2, 5],
   'min_samples_leaf': [1, 2]
}
grid_search_rf = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, scoring='r2',
verbose=1, n_jobs=-1)

grid_search_rf.fit(X_train, y_train)

print("Best parameters:", grid_search_rf.best_params_)
print("Best score (R2):", grid_search_rf.best_score_)
# Робимо прогнози на тренувальних даних
predictions_rf_train = grid_search_rf.predict(X_train)
# Оцінка моделі
mse_rf_train  = mean_squared_error(y_train, predictions_rf_train)
mae_rf_train  = mean_absolute_error(y_train, predictions_rf_train)
r2_rf_train  = r2_score(y_train, predictions_rf_train)
print('Mean Squared Error Random Forest Regressor Model = ',mse_rf_train)
print('Mean Absolute Error Random Forest Regressor Model = ', mae_rf_train)
print('R_2score  Random Forest Model = ', round(r2_rf_train,2))
# Тренувальні дані
plt.figure(figsize=(16, 6))
plt.scatter(X_train['age'], y_train, color='blue', label='Тренувальні дані')

# Прогнози на тренувальних даних
plt.scatter(X_train['age'], predictions_rf_train, color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Random Forest Regression Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
# Робимо прогнози на тестових даних
rf_predictions = grid_search_rf.predict(X_test)
# Оцінка моделі

# Оцінка моделі
mse_rf  = mean_squared_error(y_test, rf_predictions)
mae_rf  = mean_absolute_error(y_test, rf_predictions)
r2_rf  = r2_score(y_test, rf_predictions)
print('Mean Squared Error Random Forest Regressor Model = ',mse_rf)
print('Mean Absolute Error Random Forest Regressor Model = ', mae_rf)
print('R_2score  Random Forest Model = ', round(r2_rf,2))
```

```
# Тестові дані
plt.figure(figsize=(16, 6))
plt.scatter(X_test['age'], y_test, color='blue', label='Тестові дані')

# Прогнози на тестових даних
plt.scatter(X_test['age'], rf_predictions , color='red',  label='Прогнозовані дані')

plt.title('Візуалізація Random Forest Model')
plt.xlabel('Вік клієнта')
plt.ylabel('Ціна страховки')
plt.legend()
plt.show()
# Створюємо DataFrame з результатами оцінки моделей
results = pd.DataFrame({
    'Model': ['Linear Regression', 'Decision Tree', 'Random Forest'],
    'MSE_TR': [mse_lr_train, mse_dt_train, mse_rf_train],
    'MSE_TS': [mse_lr_Test, mse_dt, mse_rf],
    'MAE_TR': [mae_lr_train, mae_dt_train, mae_rf_train],
    'MAE_TS': [mae_lr_Test, mae_dt, mae_rf],
    'R-squared_TR': [round(r2_lr_train,2), round(r2_dt_train,2), round(r2_rf_train,2)],
    'R-squared_TS': [round(r2_lr_Test,2), round(r2_dt,2), round(r2_rf,2)]
})

# Виводимо результати
results
```

Додаток Г

(обов’язковий)

ІЛЮСТРАТИВНА ЧАСТИНА

ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ВАРТОСТІ МЕДИЧНОГО СТРАХУВАННЯ

Нормоконтроль: к.т.н., доцент

_____ Сергій ЖУКОВ

«___» _____ 2024 р.

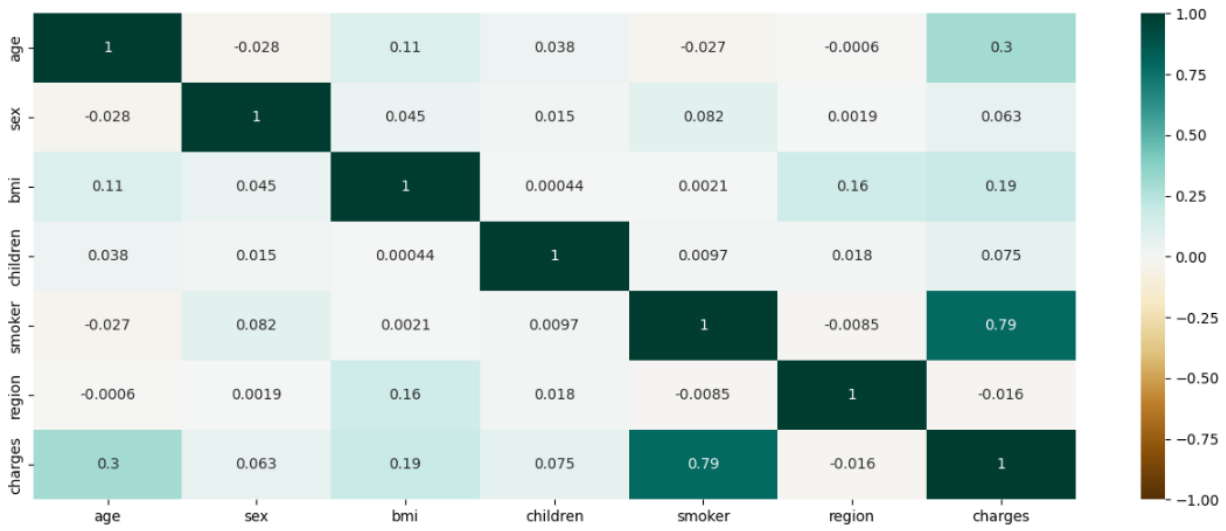


Рисунок Г.1 – Кореляційна матриця даних вартості на медичне страхування

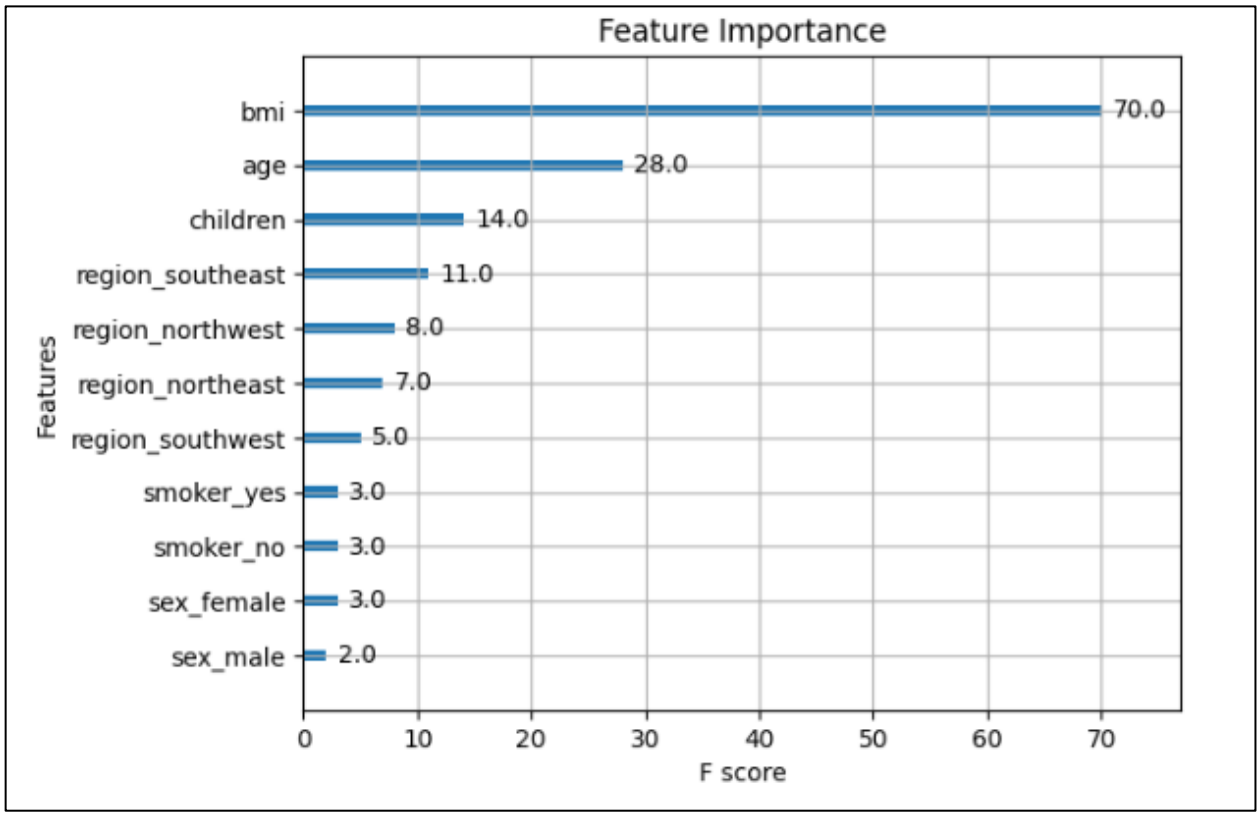


Рисунок Г.2 – Графік важливості ознак

Model	MSE_TR	MSE_TS	MAE_TR	MAE_TS	R-squared_TR	R-squared_TS
Linear Regression	3.477000e+07	3.252568e+07	4059.925497	4081.574995	0.74	0.78
Decision Tree	1.702575e+07	1.724235e+07	2292.648124	2457.730638	0.87	0.88
Random Forest	1.341566e+07	1.356090e+07	1913.485612	2020.544529	0.90	0.91

Рисунок Г.3 – Таблиця порівняння точності моделей

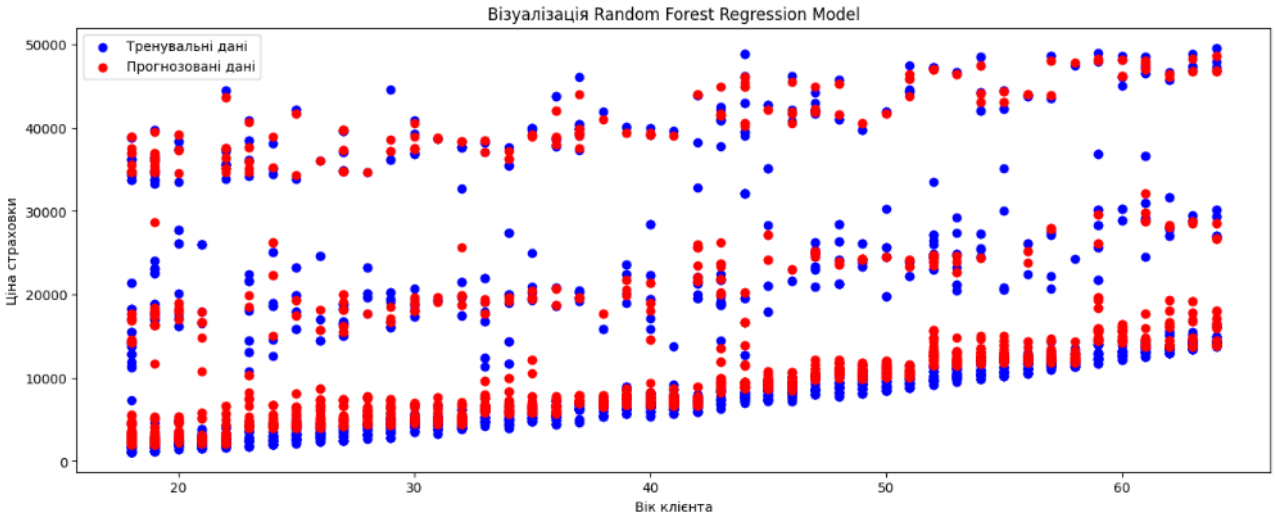


Рисунок Г.4 – Графік порівняння прогнозованих даних найкращою моделлю та тренувальних даних

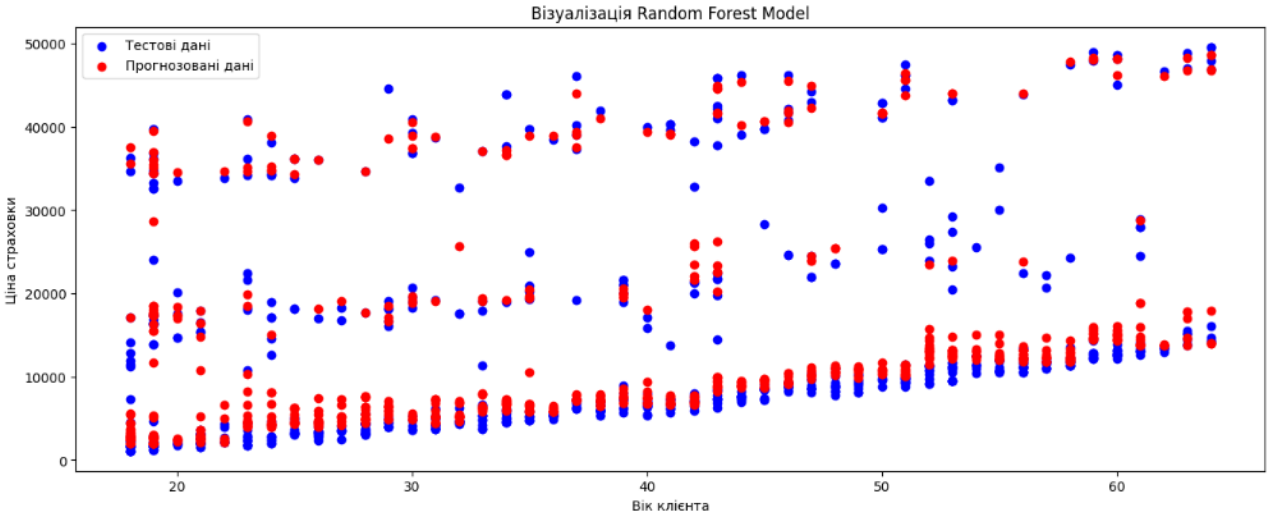


Рисунок Г.5 – Графік прогнозованих даних найкращою моделлю та реальних тестових даних